Final Project on STAT 2332, Fall 2021

School of Data Science and Analytics

Kennesaw State University

Write a report in one overleaf file together with R and Python code for each problem.

This project will carry 50% of the project.

You can complete the project by R or Python.

If complete the project by R and Python, you will receive 7% extra credit.

3% extra credit to complete the Project if you typed the project in overleaf.

Each problem will carry 1 percentage point.

DUE DATE: 06/DEC/2021

Send the final pdf file via email at [mchowd10@kennesaw.edu](mailto:mchowd10@kennesaw.edu)

1. Read the Final.csv data from D2L to R and Python and denote this data by d1.
2. How many observations (number of rows) and Variables (columns) in the d1 data?
3. How many variables are numerical/continuous and how many are them are integers/discrete?
4. Delete ID variable from the d1 data
5. Report the number of missing values for the variables MOFB, YOB, and AOR.
6. Create d2 data from d1 data by selecting variables RMOB, WI, RCA, Religion, Region, AOR, HEL, DOBCMC, DOFBCMC, MTFBI, RW, RH, and RBMI variables.
7. Delete rows that have missing values for any variable in the d2 data and denote this new data by d3.
8. Find the summary statistics of the d3 data.
9. Add a new variable in the d3 data by finding the average of DOBCMC, DOFBCMC and MTFBI.
10. Create a new variable named “Newreligion” by recoding ‘1’ as ‘1’ and rest as ‘2’ from the Religion Variable.
11. Find the frequency table for the Region variable
12. Find the joint frequency table for the variables Region and Religion.
13. Find the mean values of AOR variable corresponding to each label of Region variable.
14. Find the variances of AOR variable corresponding to each label of Religion variable.
15. Draw a boxplot for the MTFBI variable.
16. Draw a histogram for the RCA variable.
17. Draw a bar chart for the Region variable
18. Draw a pie chart for the Region variable
19. Put above four figures (question 15 to question 18) in a 2 by 2 grid
20. Split the d3 data by WI variable and denote it by d4
21. For each split data in d4 write a single loop to find the mean, minimum, maximum, standard deviation of MTFBI.
22. Conduct a one sample mean test of hypothesis to check whether MTFBI has a mean of 30 or not.

﻿d1=pd.read\_csv('/Users/mchowd10/Desktop/Projects/classdata.csv')

﻿d2=d1.dropna()

﻿d3=d2

﻿

import numpy as np

import pandas as pd

import scipy.stats as stats

import matplotlib.pyplot as plt

import math

import statistics as ss

import scipy

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import matplotlib as mpl

import seaborn as sns

import scipy.stats as stats

import statsmodels.stats as sms

from statsmodels.stats import api as sms\_api

from textwrap import fill

﻿scipy.stats.ttest\_1samp(d3.SBP, 100)

1. Conduct a normality test of the MTFBI variable

﻿scipy.stats.shapiro(d3.SBP)

1. Check the equality of mean for MTFBI variable corresponding to two labels of “Newreligion” variable.

﻿scipy.stats.ttest\_ind(d3[d3.RACE==1].DBP,d3[d3.RACE==2].DBP)

1. Find the correlation matrix of the variables DOBCMC, DOFBCMC, AOR, MTFBI, RW, RH and RBMI from the d3 data.

﻿columns = ["SBP","DBP","HT","WT","WM","BMI","TC","TG","HDL","LDL"]

c1=d3[columns]

import matplotlib.pyplot as plt

plt.matshow(c1.corr())

# correltion matrix

c1.corr()

1. PLEASE DO NOT HELP EACH OTHER TO COMPLETE THE PROJECT. IF YOU HELP YOU ARE DESTROYING YOUR FRIEND LIFE AS WELL AS YOU ARE COMPROMISING WITH ACADEMIC INTEGRITY.
2. Fit a multiple regression model by considering MTFBI as dependent variable and AOR, RW, Region as independent variables

﻿from statsmodels.api import OLS

y=d3.SBP

x=d3[['WT','BMI','AGE','TC']]

x1=sm.add\_constant(x)

model = OLS(y, x1).fit()

model.summary()

1. Simulate one data from the following equation . Where X is binomial with n=20, p=.70. U is uniform between 15 and 30 (inclusive). N is normal with mean 0 and standard deviation 5. E is random uniform between -1 and 1. True mean is 640. True variance is 257920.
2. Repeat the procedure 100 times and check the true mean with the simulated mean.
3. Repeat the procedure 100 times and check the true variance with the simulated variance.
4. Repeat the procedure 500 times and check the true mean with the simulated mean.
5. Repeat the procedure 500 times and check the true variance with the simulated variance.

﻿mport numpy as np

import pandas as pd

import scipy.stats as stats

import statsmodels.stats.api as sms

def sim(n):

x1=np.random.binomial(20,.65,n)

x2=np.random.exponential(1/(1/5),n)

x3=np.random.normal(20,3,n)

u1=np.random.uniform(0,10,n)

y= 5 + 3\*x1 + np.log(34) + np.exp(-x2) + u1\*np.exp(5) + (np.pi)\*x3

y1=pd.DataFrame(y)

return y1

a=sim(1000)

# mean and variance

np.mean(a)

np.var(a)

# single or specified mean test

stats.ttest\_1samp(a,1000)

# repeating 100 times

B=100

repeat=[sim(1000) for i in range(B)]

alldata=pd.concat(repeat)

# computing mean and variance from simulated data

simulated\_mean=np.mean(alldata)

simulated\_variance=np.var(alldata)

# Theoretical Mean and Variance

theoretical\_mean=5+3\*20\*.65+np.log(34)+1/6+5\*np.exp(5)+20\*(np.pi)

theoretical\_mean

theoretical\_variance=9\*20\*.65\*.35+(1/11-1/36)+(100/12)\*np.exp(10)+(np.pi\*np.pi)\*9

theoretical\_variance

# difference between them

abs(simulated\_mean-theoretical\_mean)

abs(simulated\_variance-theoretical\_variance)

1. For five values of x=1:5, y=2:6, and z=3:7, compute 5 values for .
2. Solve the following system of linear equations:

70x+100y+40z=900; 120x+450y+340z=1000; 230x+230y+1230z=3000

1. Find the inverse of the following matrix: A=
2. Suppose b=. Then find . Here A’ means A transpose.
3. Draw the graph for the function f(x)=
4. Draw the graph for the step functions Consider the continuous function
5. Find the areas of 10 circles, which have radii 10:19. The Area of a circle is given .
6. Find .
7. Compute
8. Compute the integral .
9. Compute the integral .
10. For five values of x=1:5, y=2:6, and z=3:7, compute 5 values for .
11. Solve the equation .
12. From the word “squelched”, write the python code to get the output of word ‘seed’.
13. If $40 is invested today for 50 years with interest rate .10, the find the total amount of money in 50 years. The formula is p\*(1+r)^t. p=40, t=50, and r=.10.
14. Fit a simple regression model by using MTFBI as dependent variable and AOR as independent variable.

import statsmodels.api as sm

# regression without intercept

model = sm.OLS(d3.SBP, d3.WT).fit()

predictions = model.predict(d3.WT) # make the predictions by the model

model.summary()

1. Check whether AOR and MTFBI are correlated or not.

﻿pearsonr(d3.SBP, d3.DBP)

1. Check whether variance of AOR is 10 or not.

R code for q50

install.packages("EnvStats")

library(EnvStats)

varTest(cars$speed, alternative = "two.sided", conf.level = 0.95,

sigma.squared = 10, data.name = NULL)

Python code for q50

﻿ ﻿def chi\_sq\_test\_for\_variance(variable,h0): #function for the chi squared distribution associated with the above formula

sample\_variance = variable.var() # Find the variance of the sample

n = variable.notnull().sum() # Take the sum of the number of values that are not missing

# the actual number of observations for the variable where

# True = 1, False = 0

degrees\_of\_freedom = n-1 # Find the degrees of freedom

x\_sq\_stat = (n - 1) \* sample\_variance / h0 # Using the formula above to calculate the X^2 statistic

p = stats.chi2.cdf(x\_sq\_stat,degrees\_of\_freedom) # Here, a cumulative distribution function is used to determine

# the significance of the variance using the X^2 statistic.

# If a chi square test statistic is over the 99th percentile,

# we'd have reason to suspect significance at alpha = .05.

# We need to account for circumstance where the p value is greater

# than .05, however:

if p > .05:

p = 1 - p

return (x\_sq\_stat, p,degrees\_of\_freedom) # End of function

dbp\_variance = round(d3["DBP"].var(),2)

x\_sq\_stat, pval, dof = chi\_sq\_test\_for\_variance(d3["DBP"],h0=60)