iubh INTERNATIONALE HOCHSCHULE FERNSTUDIUM

Advanced Statistics

Samyak Anand

Registration number: 321149613

Parameters generated on 15-12-2022
with signature 20517e351e4c1f647020f3b0458bbf5ca371d670

Date 01-01-2023

Place: India

Table of Content

Content…………………………………………………………………………………….1

List if figures……………………………………………………………………………2

Task 1……………………………………………………………………………………3

Task 2……………………………………………………………………………………7

Task 3……………………………………………………………………………………11

Task 4……………………………………………………………………………………14

Task 5……………………………………………………………………………………16

Task 6……………………………………………………………………………………25

List of Figure

No	Name	Page no
Fig 1	Bar Graph: Proportion of "for" and "against"	5
Fig 2	Pie Chart: Proportion of "for" and "against"	6
Fig 3	CDF (Cumulative distribution function)	10
Fig 4	*PDF (Probability density function)*	10
Fig 5	Histogram	10
Fig 6	Data Plot for given data set	18
Fig 7	Plot for Linear Model	19
Fig 8	Plot for Prediction Model	20
Fig 9	*Plotting lower and upper confidence intervals*	21
Fig 10	OLS Regression Results	22

Task 1:

Please provide the requested visualization as well as the numeric results. In both cases, please provide how you realized these (calculations, code, steps…) and why it is the appropriate tools. Do not forget to include the scale of each graphics so a reader can read the numbers represented.

A) **If ξ_1 is 0: A vote with outcome for or $against$ follows a Bernoulli distribution where P(vote="for") =ξ_2. Represent the proportion of "for" and "against" in this single Bernoulli trial using a graphic and a percentage. Can an expectation be calculated? Justify your answer by all necessary hypotheses.**

B) If ξ_1 is between 1 and 3: The number of meteorites falling on an ocean in a given year can be modelled by one of the following distributions. Give a graphic showing the probability of one, two, three… meteorites falling (until the probability remains provably less than 0.5% for any bigger number of meteorites). Calculate the expectation and median and show them graphically on this graphic:

   a. If ξ_1 is 1: a Poisson distribution with an expectation of λ=ξ_2

   b. If ξ_1 is 2: a negative binomial distribution with an expectation of k=ξ_2 and p=ξ_3

   c. If ξ_1 is 3: a geometric distribution counting the number of Bernoulli trials with p=ξ_2 until it succeeds.

Given Parameters

- ξ_1: 0
- ξ_2: 0.74

Given that outcome follows the Bernoulli distribution (X~Bernouli(p)), therefore PMF of Bernoulli

distribution is given by

$$f(x) = \begin{cases} p & ,x = 1 \\ 1 - P & \quad ,x = 0 \end{cases}$$

Here, A vote with outcome for is taken as success and against is taken as failure, such that X defined on this as X(success) = 1 and X(failure) = 0. Bernoulli random variable with parameter, written as X ~ Bernoulli (0.74)

P (vote = "for") =0.74 then,(vote="against") =1- P (vote = "for") =0.24

Python Script:

```
from scipy.stats import bernoulli
import matplotlib.pyplot as plt


#Instance of Bernoulli distribution with parameter p=0.74

bd=bernoulli(0.74)

#for the visualization of thr bar plot of Bernoulli's distribution

plt.figure(figsize=(10,10))
plt.xlim(-2,2)
plt.bar(x,bd.pmf(x),color='blue')

#for labeling the Bar Plot

plt.title('Proportion of “for” and “against” using a bar graph)', fontsize='12')
plt.xlabel('Values of random variable x (0, 1)', fontsize='12')
plt.ylabel('Probability', fontsize='12')

plt.show()

#for labeling the pie chart
size=[0.74,1-0.74]
plt.figure(figsize=(10,10))
plt.title('Proportion of “for” and “against” using a percentage', fontsize='12')
Mlabels = 'Vote for', 'Vote Against'
mcolor='Green','Red'
plt.pie(size,colors=mcolor, labels = Mlabels,autopct='%1.1f%%',shadow=True, startangle=90)
plt.show()
```

Output:

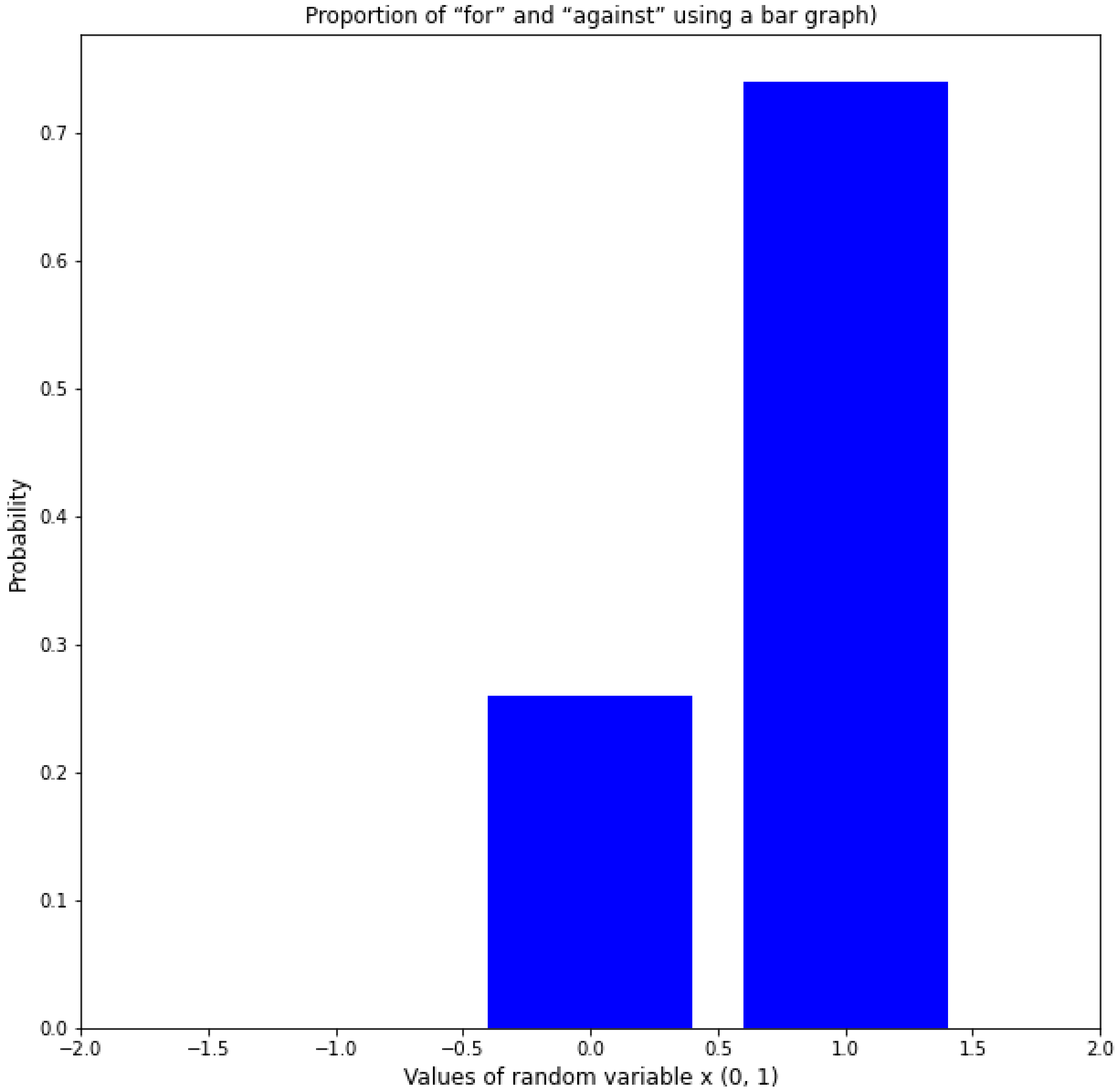


Fig 1: Bar Graph: Proportion of "for" and "against"

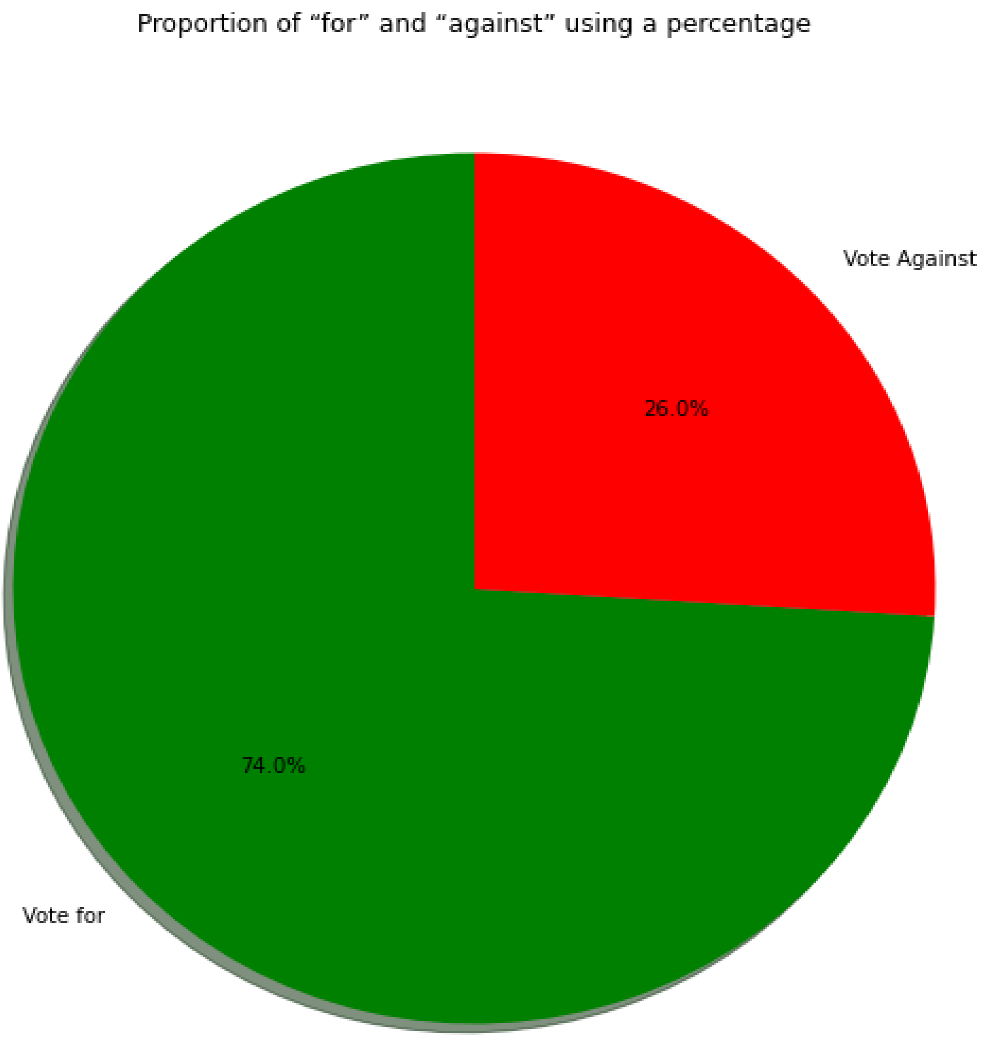


Fig 2 Pie Chart: Proportion of "for" and "against"

Where,

1. Probability of voting "for"
2. Probability of voting "against"

Yes, expectation can be calculated

Justification, The expected value for the Bernoulli distribution can be calculated as follows:

E[X] =∑ x ∗ p(x) x=0 E[x]=0*P(x)+1*P(x)

=0*P(x=0) +1*P(x=1) = p

Therefore **E[x]=0.74**

Task 2: Basic Probabilities and Visualizations

Let Y be the random variable with the time to hear an owl from your room's open window (in hours). Assume that the probability that you still need to wait to hear the owl after y hours is one of the following:

a) If $\xi 4$ is 0: the probability is given by $\xi_5 e^{-\xi_6 y}+\xi_7 e^{-\xi_8 y}$

b) If $\xi 4$ is 1: the probability is given by $\xi_5 e^{-\xi_6 y^2}+\xi_7 e^{-\xi_8 y^8}$

c) If $\xi 4$ is 2: the probability is given by $\xi 5e^{-\xi_6 \sqrt{y}}+\xi_7\ e^{-\xi_8}\sqrt[3]{y}$

d) If $\xi 4$ is 3: the probability is given by $\boldsymbol{\xi_5 e^{-\xi_6 y^2}+\xi_7 e^{-\xi_8 y^2}}$

Find the probability that you need to wait between 2 and 4 hours to hear the owl, compute and display the probability density function graph as well as a histogram by the minute. Compute and display in the graphics the mean, variance, and quartiles of the waiting times. Please pay attention to the various units of time!

Given Parameters

ξ_4: 3 ξ_5: $0.\boxed{41}$

ξ_6: 4 ξ_7: $0.\boxed{58}$

ξ_8: 3

Since, $\xi_4 = 3$, The probability is given by $\boldsymbol{\xi_5 e^{-\xi_6 y^2}+\xi_7 e^{-\xi_8 y^2}}$ **----- (i)**

Since ξ_5 and ξ_7 are recurring decimals, we need to convert the decimals.

ξ_5: $0.\boxed{41}$

will be 41/99

similarly ξ_7: $0.\boxed{58}$

will be 58/99

Let Y be probability that we still need to wait to hear the owl y hours

$\underline{P(Y>y) = (41/99e^{-4y*y}+ 58/99e^{-3y*y})}$

$P(Y \le y) = 1- P(Y>y)$

$= 1- (41/99e^{-4y*y}+ 58/99e^{-3y*y})$

$\underline{f(y)= 1-(41/99e^{-4y*y}+ 58/99e^{-3y*y})}$

The function f(y) represents the Cumulative Distribution Function (CDF)

Probability that needs to wait between 2 and 4 hours to hear the owl,

$P(2\le y\le 4) = [1-(41/99e^{-4y*y}+ 58/99e^{-3y*y})]_{y=4} - [1-(41/99e^{-4y*y}+ 58/99e^{-3y*y})]_{y=2}$

[1-(41/99*1.60381089 $\times10^{-28}$) + (58/99*1.60381089 $\times10^{-28}$)]-[1-(41/99*0.0000061442)+(58/99* 1.125351747 $\times10^{-7}$)]

P(2≤y≤4) = -0.005658483377780854

To find Probability density we need to differentiate equation with respect to y

$f(y) = \frac{df(y)}{dy}$ (1-(41/99e^{-4y*y}+ 58/99e^{-3y*y})

$$f\,'(y) = 3.31e^{-4y^2} + 3.51e^{-3y^2}$$

Quartiles
f(y)= 41/99e^{-4y*y}+ 58/99e^{-3y*y})
The quantile function (Q) for CDF is delivered by finding Q for which
1-0.41e^{-10Q}-0.58e^{-6Q}=p
ln(1-p) =-4.1Q-3.48Q
-7.58Q(p)=ln(1-p)
=> $Q(p) = \frac{ln(1-p)}{-7.58}$

first quartile (p = 1/4)

=> $Q(p) = \frac{ln(1-1/4)}{-7.58}$

$$Q(p) = \frac{-0.2876}{-7.58}$$

Q(p=1/4) = 0.03795

Second quartile (p = 1/2)

Python Script

```
import numpy as np
from matplotlib import pyplot as plt


def cdf(x):
   return 1 - (np.exp(-4*x**2) * 41/99 + np.exp(-3*x**2) * 58 / 99)
def pdf(x):
   return (np.exp(-4*x**2) * 3.31 + np.exp(-3*x**2) *3.51)

#CDF
x = np.linspace(0, 1, 200)
fig, ax = plt.subplots(figsize=(10, 4))
```

```
ax.plot(x, cdf(x), color='y')
ax.axvline(0.03795, color='b', label=f'25th percentile')
ax.axvline(0.0914, color='r', label=f'50th percentile')
ax.axvline(0.18288, color='g', label=f'75th percentile')
plt.xlabel('Count', fontsize='10')
plt.ylabel('Value', fontsize='10')
plt.title("CDF")
ax.legend()

plt.show()


#PDF
x = np.linspace(0, 1, 200)
fig, ax = plt.subplots(figsize=(10, 4))
ax.plot(x, pdf(x), color='y')
ax.axvline(0.03795, color='b', label=f'25th percentile')
ax.axvline(0.0914, color='r', label=f'50th percentile')
ax.axvline(0.18288, color='g', label=f'75th percentile')
plt.xlabel('Count', fontsize='10')
plt.ylabel('Value', fontsize='10')
plt.title("PDF")
ax.legend()

plt.show()



cdf_=cdf(4) - cdf(2)
print ("CDF is:",cdf_ )

data = cdf(x)
fig, ax = plt.subplots(figsize=(10, 4))
plt.hist(data, bins=np.arange(min(data), max(data) + 1/60, 1/60))
plt.xlabel('Count', fontsize='10')
plt.ylabel('Value', fontsize='10')
plt.title("Histogram")
plt.show()
```

Output:

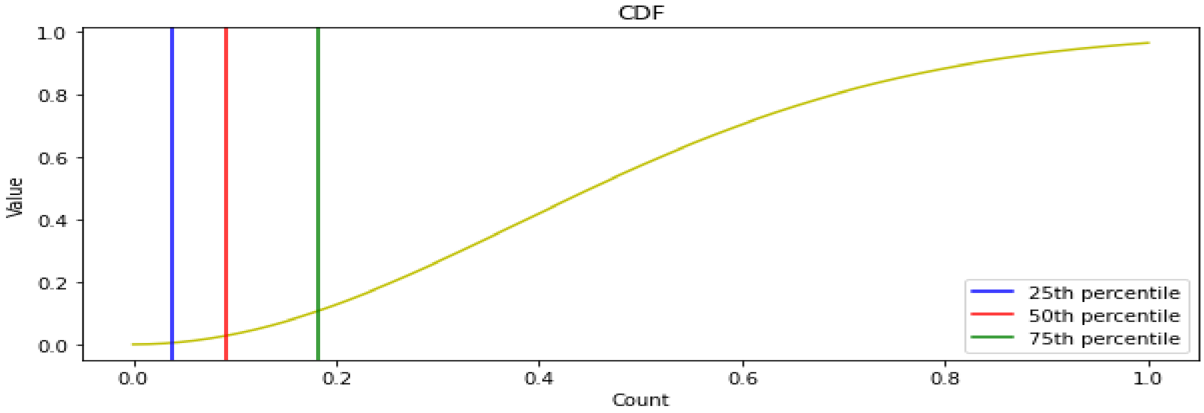


Fig 3 CDF (Cumulative distribution function)

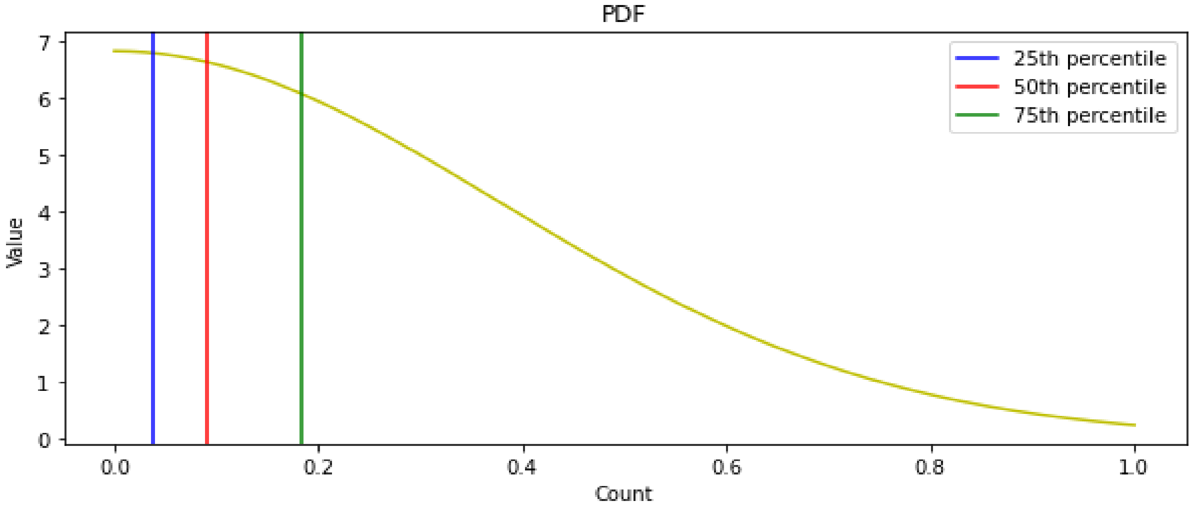


Fig 4: PDF (Probability density function)

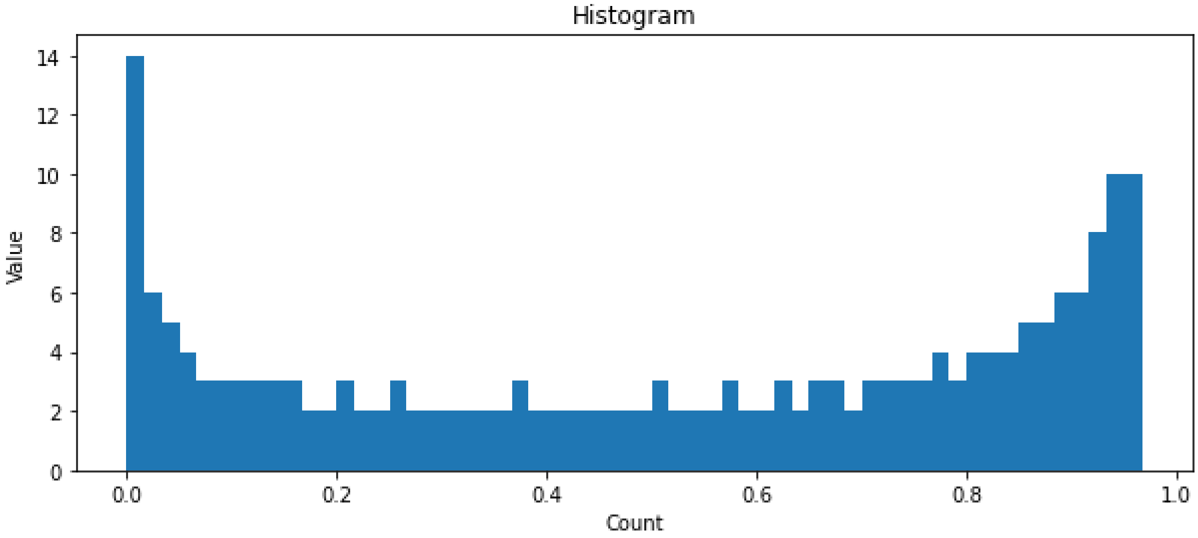


Fig 5: Histogram

Task 3: Transformed Random Variables

A type of network router has a bandwidth total to first hardware failure called S expressed in terabytes. The random variable S is modelled by an exponential distribution whose density is given by one of the following functions:

- (if ξ_9=0): $f_S(s)$=1/θ*$e^{-s/\theta}$
- (if ξ_9=1): $fS(s)$=1/24θ*$s^4 e^{-s/\theta}$
- (if ξ_9=2): $fS(s)$=1/θ for s∈[0,θ]

with a single parameter θ. Consider the bandwidth total to failure T of the sequence of the two routers of the same type (one being brought up automatically when the first is broken).
Express T in terms of the bandwidth total to failure of single routers $S1$ and $S2$. Formulate realistic assumptions about these random variables. Calculate the density function of the variable T.
Given an experiment with the dual-router-system yielding a sample $T1$, $T2$, …, Tn, calculate the likelihood function for θ. Propose a transformation of this likelihood function whose maximum is the same and can be computed easily.
An actual experiment is performed, the infrastructure team has obtained the bandwidth totals to failure given by the sequence ξ_9 of numbers. Estimate the model-parameter with the maximum likelihood and compute the expectation of the bandwidth total to failure of the dual-router-system.

Given Parameters:

ξ_9: 0, ξ_{10}: 4, 4, 63, 78, 0

So, $f(s) = \frac{1}{\theta} e^{\frac{-s}{\theta}}$ --------(i)

T in the term of bandwidth total to failure of single router S1 and S2,

$S_1(T_1) = \frac{1}{\theta} e^{\frac{-T_1}{\theta}}$

$S_2(T_2) = \frac{1}{\theta} e^{\frac{-T_2}{\theta}}$

Realistic assumption about these random variables are as follows:

- We need a sample of n measurement of same random variable T_1, T_2,……,T_n where Ti can either be a single variable or a vector of variables.
- We assume that we know We assume that we know the underlying probability density distribution (T|θ) but not the value of θ. This means that there is a data-generating process that can be described sing a probability density distribution (PDF) for f (T| θ). Which maps measurements T to a number yielding the probability of sets of values. This function describes how the values of the measurements are distributed, and each measurement is a so-called “realization” of this PDF. The functional form of the density function depends on parameter θ. In the maximum likelihood approach, we estimate the best numerical value of the parameter θ that maximizes the probability to observe the

data, but we assume that the choice of the underlying PDF f(T| θ) is correct. This implies that if we make the wrong assumption about it, meaning that we choose the wrong type of probability distribution, the result will also be wrong, even if all subsequent numerical steps and estimates of the parameter are done correctly.

The density function of the variable T

From equation 3.1, density function of the variable T given by

$f_T(T)= \frac{1}{\theta} e^{\frac{-T}{\theta}}$ --------------------(ii)

Given that the experiment with the dual-router system yield a sample T_1 and T_2

Eq(ii)can be written as for sample T_1 and T_2

$f_T(T_1) = \frac{1}{\theta} e^{\frac{-T1}{\theta}}$ $f_T(T_2) = \frac{1}{\theta} e^{\frac{-T2}{\theta}}$ $f_T(T_3) = \frac{1}{\theta} e^{\frac{-T3}{\theta}}$ …… $f_T(T_n) = \frac{1}{\theta} e^{\frac{-Tn}{\theta}}$

for 2 routers:

$f_T(T_1) = \frac{1}{\theta} e^{\frac{-T1}{\theta}}$ $f_T(T_2) = \frac{1}{\theta} e^{\frac{-T2}{\theta}}$

Likelihood function L(θ)

$$L[\theta] = \frac{1}{\theta} e^{\frac{-T1}{\theta}} * \frac{1}{\theta} e^{\frac{-T2}{\theta}} * \frac{1}{\theta} e^{\frac{-T3}{\theta}} * \ldots\ldots\ldots\ldots\ldots * \frac{1}{\theta} e^{\frac{-Tn}{\theta}}$$

$$= \frac{1}{\theta^n} (e^{\frac{-T1}{\theta} - \frac{T2}{\theta} - \frac{T3}{\theta} - \ldots\ldots - \frac{Tn}{\theta}})$$

$$\boxed{L[\theta] = \frac{1}{\theta^n} (e^{\frac{-1\sum_{i=1}^{n} Ti}{\theta}})}$$ -------(iii)

The above equation represents likelihood function for θ

$$L[\theta] = \frac{1}{\theta^n} (e^{\frac{-1\sum_{i=1}^{n} Ti}{\theta}})$$

Taking ln on both sides,

$$\ln(L[\theta]) = \frac{1}{\theta^n} (e^{\frac{-1\sum_{i=1}^{n} Ti}{\theta}})$$

$$\frac{dL[\theta]}{d\theta} = \frac{d}{d\theta} (\frac{1}{\theta^n} e^{\frac{-1\sum_{i=1}^{n} Ti}{\theta}})$$

$= \theta^{-n-1}. e^{\frac{\sum_{i=1}^{n} Ti}{\theta}} \theta^{-1} \sum_{i=1}^{n} Ti = 0$

$= \theta^{-n-1}. e^{\frac{\sum_{i=1}^{n} Ti}{\theta}} = 0$

$\theta = \frac{\sum_{i=1}^{n} Ti}{n}$ ----------(iv)

Using equation (iv) which represent likelihood equation.

Since given parameters: 4, 4, 63, 78, 0 where the infrastructure team has obtained the bandwidth failure

The model parameters (θ) for the maximum likelihood are given by:

$\theta = \frac{\sum_{i=1}^{n} Ti}{n} = \frac{4+4+63+78+0}{5} = \mathbf{29.8}$

θ = 29.8

Expectation of the bandwidth total to failure of the dual-router system

$F(x) = \lambda e^{-\lambda x}$

$E[x] = 1\backslash \lambda$

On comparing expectation for the given exponential distribution is given by

$\boldsymbol{\lambda = \frac{1}{\theta} = \theta = 29.8}$

bandwidth total to failure of the dual-router-system is <u>29.8</u>

Task 4: Hypothesis Test

Over a long period of time, the production of 1000 high-quality hammers in a factory seems to have reached a weight with an average of ξ_{11} (in g) and standard deviation of ξ_{12} (in g). Propose a model for the weight of the hammers including a probability distribution for the weight. Provide all the assumptions needed for this model to hold (even the uncertain ones)? What parameters does this model have?

One aims at answering one of the following questions about a new production system:

- (if ξ_{13}=0): Does the new system make *more constant* weights?
- (if ξ_{13}=1): Does the new system make *lower* weights?
- (if ξ_{13}=2): Does the new system make *higher* weights?
- (if ξ_{13}=3): Does the new system make *less constant* weights?

To answer this question a random sample of newly produced hammers is evaluated yielding the weights in ξ_{14}.

What hypotheses can you propose to test the question? What test and decision rule can you make to estimate if the new system answers the given question? Express the decision rules as logical statements involving critical values. What error probabilities can you suggest and why? Perform the test and draw the conclusion to answer the question.

Given parameters,

- ξ_{11}: 813
- ξ_{12}: 67.6
- ξ_{13}: 2
- ξ_{14}: 805, 842, 843, 805, 870, 857, 745, 811, 755, 838

Hypotheses that can be proposed to test the question. The hypotheses test is rule that can be used to (Casella & Berger, 2002, p. 374): fail to reject the null Hypotheses H_0 or reject the null hypotheses H_0 and accept the alternative hypotheses H_1 as true. Test and decision rule that we can make to estimate if new system answer the given question. Decision rules as logical statements involving critical values. To distinguish between the and alternative hypotheses, it is necessary to compare the mean of the two groups. Which is approx. by the sample mean for the measured values using the central limit theorem. which is used to calculate test statistics, symbolized by the letter 'Z'. Further comparing the test statistics and critical value ($Z_{\alpha/2}$) with 5% level of significance(α) that is

For, **two tailed**

If $Z<Z_{\alpha/2}$ accept null hypothesis and reject alternative hypothesis otherwise

If $Z>Z_{\alpha/2}$ accept alternative hypothesis and reject null hypothesis.

Right tailed

If $Z<Z_{\alpha}$ accept null hypothesis and reject alternative hypothesis otherwise

If $Z > Z_{\alpha}$ accept alternative hypothesis and reject null hypothesis

Left tailed

If $Z < Z_{-\alpha}$ accept alternative hypothesis and reject null hypothesis otherwise

If $Z > Z_{-\alpha}$ accept null hypothesis and reject alternative hypothesis

Sample mean is given by,

$$\boldsymbol{\bar{x}} = \frac{\boldsymbol{x1+x2+x3+\dots\dots\dots+xn}}{\boldsymbol{n}}$$

<u>Sample space</u>:

805	842	843	805	870	857	745	811	755	838

Table 1: Sample Space

$$\bar{x} = \frac{805+842+843+805+870+857+745+811+755+838}{10}$$

$$\bar{x} = \frac{805+842+843+805+870+857+745+811+755+838}{10} = \frac{8171}{10}$$

$$\bar{x} = 817.1$$

Now, we need to calculate the test statistics using Central limit theorem,

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

From given parameters, S.D(σ) = 67.6

$$z = \frac{817.1 - 813}{\frac{67.6}{\sqrt{10}}} = \frac{4.1}{21.3775} = 0.19179$$

From Z table we get the Z = 0.57535

H_1: system makes higher weight. $\mu<813$

For $\alpha = 5\%$ the critical value is $Z_{\alpha} = 1.960$

Therefore, $Z<Z\alpha$ which implies we must accept H_0 and reject H_1.

<u>Conclusion</u>: <u>the average weight of new production system to produce hammers is 812g hence new system does make higher weight</u>.

Task 5:

Given the values of an unknown function f:ℝ→ℝ at some selected points, we try to calculate the parameters of a model function using OLS as a distance and a ridge regularization:

A. (if $\xi 15=0$): a polynomial model function of twelve αi parameters: $f(x)=\alpha 0+\alpha 1x+\alpha 2x2+\cdots+\alpha 12x12$

B. (if $\xi 1\mathbf{5}=1$): a Fourier series model function of 8 parameters $A0,\ldots,A3,P,\varphi 1,\varphi 2,\varphi 3$: $f(x)=A0+A1\cdot\cos(2\pi Px-\varphi 1)+A2\cdot\cos(2\pi P2x-\varphi 2)+A3\cdot\cos(2\pi P3x-\varphi 3)$

C. (if $\xi 15=2$): a polynomial model function of ten αi parameters: $f(x)=\alpha 0+\alpha 1x+\alpha 2x2+\cdots+\alpha 10x10$

Calculate the OLS estimate, and the OLS ridge-regularized estimates for the parameters given the sample points of the graph of f given that the values are y = $\xi 16$.
Remember to include the steps of your computation which are more important than the actual computations.

<u>Given parameters</u>:

- ξ_{15}: 2
- ξ_{16}: (-24, -12934761147103.45), (-22, -6477551847353.7), (-12, -27946382177.67), (-21, -3980139750637.01), (-38, -801537941653295.6), (71, 234646514645107900), (57, 31528272676895520), (-6, -55679336.64), (-43, -2458047384412330), (48, 6538907454406330), (83, 916789869083713900), (-28, -52870355125109.15), (13, 49345024422.9), (28, 53701358035221.88), (91, 2203202745851523600), (-24, -13077220619795.37), (-77, -463581888527314000), (89, 1788647976728231000), (-23, -9546000456603.61), (-54, -19964988879122164)

From the given parameters,

$\boldsymbol{f(x)=\alpha_0+\alpha_1x+\alpha_2x^2+\cdots+\alpha_{10}x^{10}}$

Polynomial model in one variable. The nth order polynomial model is given by
$Y = \alpha_0 + \alpha_1x+ \alpha_2x^2+ \alpha_3x^3+\ldots\ldots\ldots\ldots+ \alpha_nx^n$ ………(i)

If $x^j=x_j$, j=1,2,3,4…...n, then the model is multiple linear regression model in n explanatory variables $x_1,x_2,x_3,\ldots\ldots x_n$. So, the linear regression model $y=X\alpha+\varepsilon$ includes the polynomial regression model.

Given, $f(x)= \alpha_0 + \alpha_1x+ \alpha_2x^2+ \alpha_3x^3+\ldots\ldots\ldots\ldots+ \alpha_{10}x^{10}$ …….. (ii)

Therefore, deriving OLS estimate α for equation (ii)

Python Script:

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

# import the excel file.
df=pd.read_excel(r'E:\task5.xlsx')
df.head() # print the top 5 row
```

x	y
-77	-4.64E+17
-54	-2.00E+16
-43	-2.46E+15
-38	-8.02E+14
-28	-5.29E+13

Table 2: Output for imported file for (Top 5) row

```
#initializing variables
x=df.iloc[:,0:-1].values
y=df.iloc[:,1].values
#create single dimendion
x=x[:,np.newaxis]
y=y[:,np.newaxis]

#sort x values and get index
inds=x.ravel().argsort()
x=x.ravel()[inds].reshape(-1,1)
#sort y according to x sorted index
y= y[inds]

print(x.shape)
print(y.shape)
```

```
(20, 1) (20, 1)
```

```
#plotting the Polynomial function
fig, ax = plt.subplots(figsize=(15, 8))
plt.scatter(x,y, color='red')
plt.xlabel('y-value', fontsize='12')
plt.ylabel('x-value', fontsize='12')
plt.title("PDF",fontsize=12)
plt.show()
```

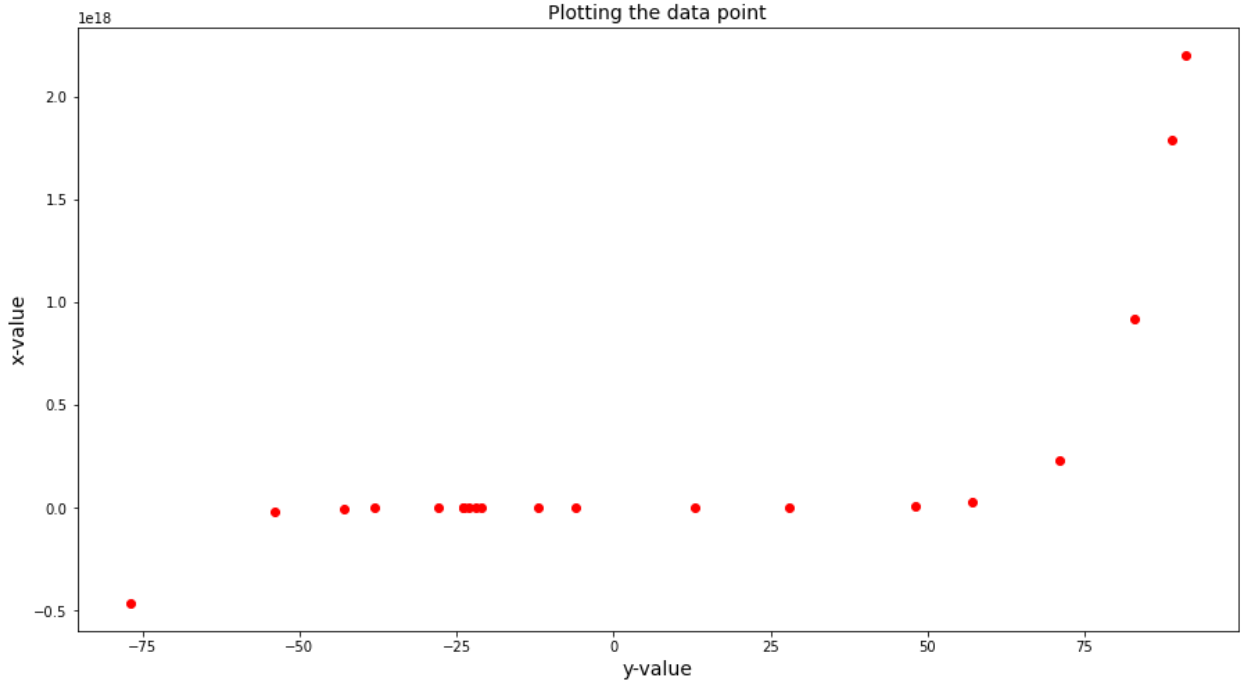


Fig 6: Data Plot for given data set

```
#OLS Regression Model
import statsmodels.api as sm
model =sm.OLS(y,x).fit()
#predected variable
ypred=model.predict(x)
```

```
fig, ax = plt.subplots(figsize=(15, 8))
plt.scatter(x,y,color='red')
plt.plot(x,ypred,color='green')
plt.xlabel('y-value', fontsize='12')
plt.ylabel('x-value', fontsize='12')
plt.title("Linear Model",fontsize=12)
plt.show()
```

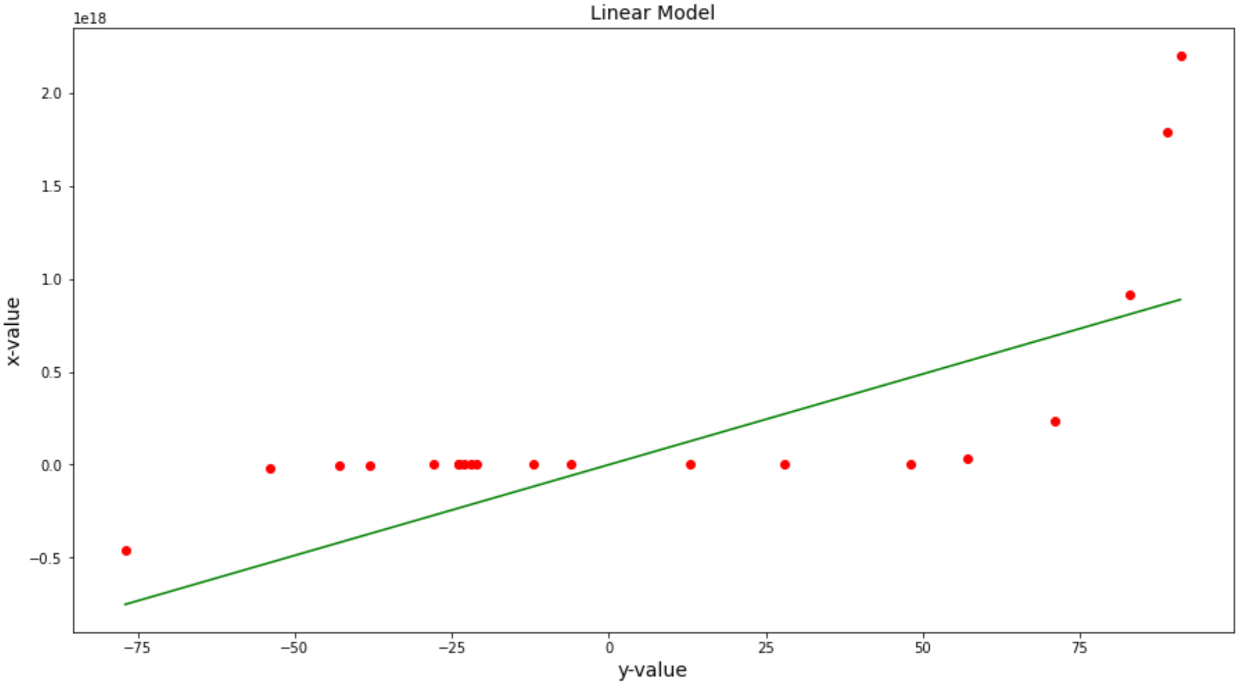


Fig 7: Plot for Linear Model

```
#genrate polynomial
from sklearn.preprocessing import PolynomialFeatures
polynomial_features= PolynomialFeatures (degree=10)
xp=polynomial_features.fit_transform(x)
xp.shape
```

```
(20, 11)
```

```
#running regression on polynomials using statsmodels ols
model= sm.OLS(y,xp).fit()
ypred=model.predict(xp)

ypred.shape
```

```
(20,)
```

```
#plot the predection model
fig, ax = plt.subplots(figsize=(15, 8))
plt.scatter(x,y,color='red')
plt.plot(x,ypred,color='green')
plt.xlabel('y-value', fontsize='12')
plt.ylabel('x-value', fontsize='12')
plt.title("Prediction model",fontsize=12)
plt.show()
```

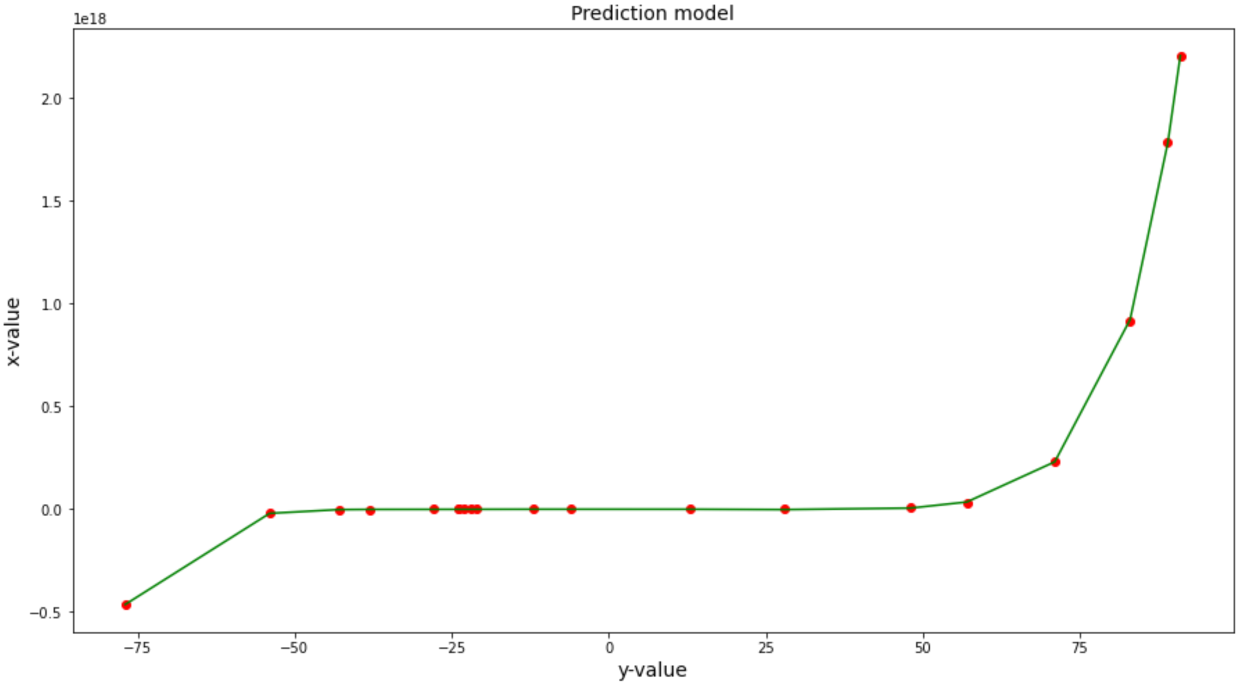


Fig 8: Plot for Prediction Model

```
#Plotting lower and upper confidance intervals
from statsmodels.sandbox.regression.predstd import wls_prediction_std
_, upper,lower = wls_prediction_std(model)
fig, ax = plt.subplots(figsize=(15, 8))
plt.scatter(x,y,label='data',color='red')
plt.plot(x,ypred,label='Predicted Plot',color='green')
plt.plot(x,upper,'--',label="Upper",color='black') # confid. intrvl
plt.plot(x,lower,':',label="lower",color='orange')
plt.legend(loc='upper left')
plt.xlabel('y-value', fontsize='12')
plt.ylabel('x-value', fontsize='12')
plt.title("Plotting lower and upper confidence intervals",fontsize=12)
plt.show()
```

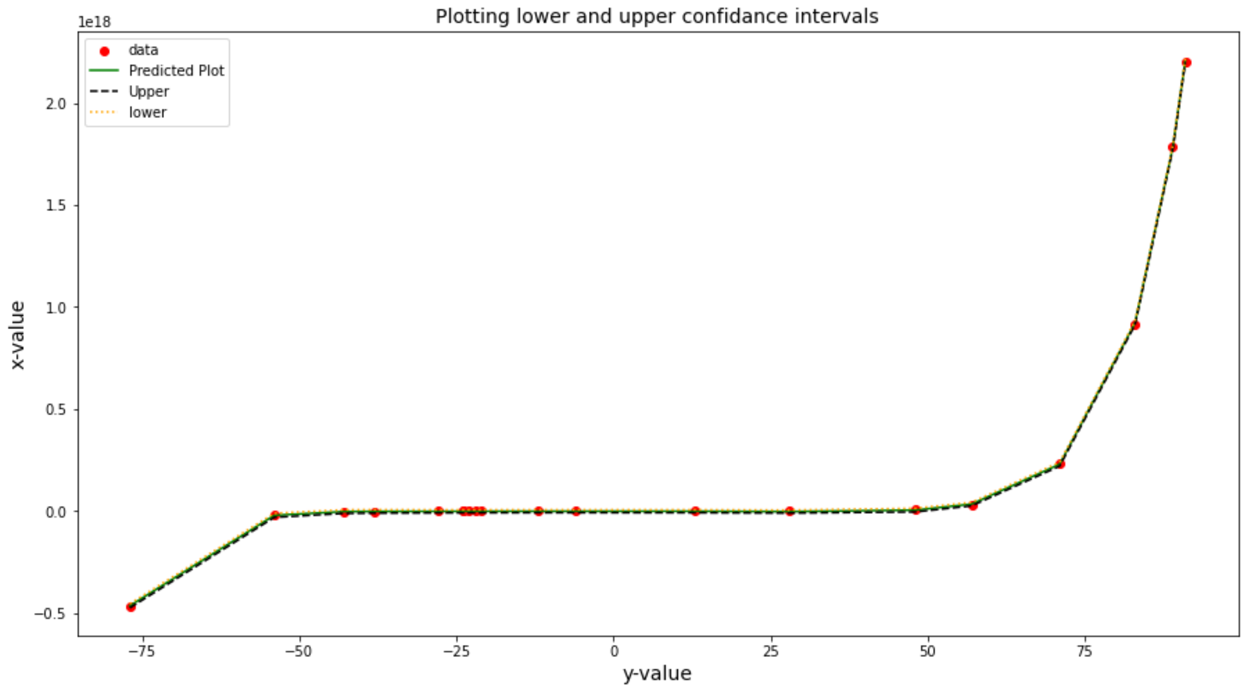


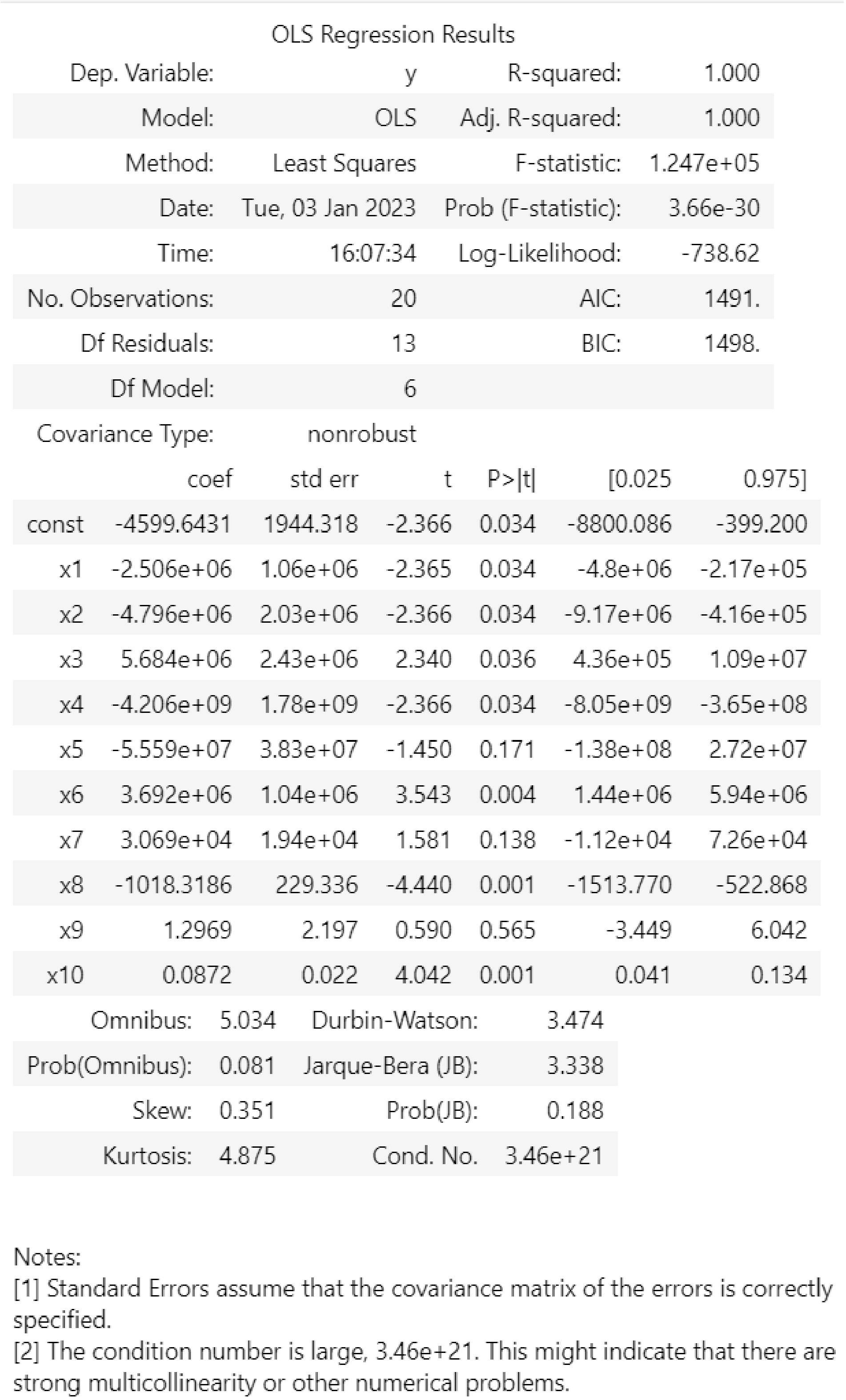
Fig 9: Plotting lower and upper confidence intervals

```
model.summary()

# Apply ridge regularization with alpha = 0.1
from sklearn.linear_model import Ridge
reg = Ridge(alpha=0.1)
reg.fit(x, y)

# Print the coefficients of the model
print(reg.coef_)
```

```
[9.36660384e+15]
```



```
                    OLS Regression Results
   Dep. Variable:                 y         R-squared:      1.000
           Model:               OLS    Adj. R-squared:      1.000
          Method:     Least Squares        F-statistic:  1.247e+05
            Date:  Tue, 03 Jan 2023  Prob (F-statistic):   3.66e-30
            Time:          16:07:34     Log-Likelihood:    -738.62
No. Observations:                20                AIC:      1491.
    Df Residuals:                13                BIC:      1498.
        Df Model:                 6
 Covariance Type:         nonrobust
                coef    std err       t    P>|t|      [0.025      0.975]
const     -4599.6431   1944.318   -2.366   0.034   -8800.086    -399.200
   x1     -2.506e+06   1.06e+06   -2.365   0.034    -4.8e+06   -2.17e+05
   x2     -4.796e+06   2.03e+06   -2.366   0.034   -9.17e+06   -4.16e+05
   x3      5.684e+06   2.43e+06    2.340   0.036    4.36e+05    1.09e+07
   x4     -4.206e+09   1.78e+09   -2.366   0.034   -8.05e+09   -3.65e+08
   x5     -5.559e+07   3.83e+07   -1.450   0.171   -1.38e+08    2.72e+07
   x6      3.692e+06   1.04e+06    3.543   0.004    1.44e+06    5.94e+06
   x7      3.069e+04   1.94e+04    1.581   0.138   -1.12e+04    7.26e+04
   x8     -1018.3186    229.336   -4.440   0.001   -1513.770    -522.868
   x9         1.2969      2.197    0.590   0.565      -3.449       6.042
  x10         0.0872      0.022    4.042   0.001       0.041       0.134
      Omnibus:  5.034     Durbin-Watson:     3.474
Prob(Omnibus):  0.081  Jarque-Bera (JB):     3.338
         Skew:  0.351          Prob(JB):     0.188
     Kurtosis:  4.875          Cond. No.  3.46e+21


Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 3.46e+21. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Fig 10: OLS Regression Results

Task 6: Bayesian Estimates

(Following Hogg, McKean & Craig, exercise 11.2.2)

Let $X1$, $X2$, …, $X10$ be a random sample from a gamma distribution with α=3 and β=1/θ. Suppose we believe that θ follows a gamma-distribution with α=ξ17 and β=ξ18 and suppose we have a trial (x1,…,xn) with an observed $\bar{x}$=ξ19.

a) Find the posterior distribution of θ.

b) What is the Bayes point estimate of θ associated with the square-error loss function?

c) What is the Bayes point estimate of θ using the mode of the posterior distribution?

Given Parameters:

- ξ_{17}: 77
- ξ_{18}: 56
- ξ_{19}: 29.17

α=3 and β=1/θ, that θ follows a gamma-distribution with α=77 and β=56

μ=29.17

Random distribution: $X1$, $X2$, …, $X10$

n =10 (no of sample events)

For continuous probability distribution, posterior distribution is given by

$$\boldsymbol{f(\theta|data)=\frac{f(data|\theta)f(\theta)}{\int_0^\infty f(data|\theta)f(\theta)d\theta}} \text{ ------------------}\textbf{(i)}$$

Here θ is the parameter of the distribution that we need to determine f(θ) is the prior that depends on this parameter, x are the observed data, the likelihood is given by f(data| θ) and the posterior distribution is given by f(θ|data).

Suppose we believe that θ follow gamma distribution

As, we know Gamma distribution

$$\textbf{Gamma Distribution = }\boldsymbol{\frac{x^{\alpha-1}e^{-\frac{\dot{x}}{\theta}}}{\Gamma(\alpha)\theta^{\alpha}}} \text{ -----------------------}\textbf{(ii)}$$

To find the prior f(θ), as θ follows gamma distribution given that α=77, β=56

$$\text{f(x:α,β) = }\boldsymbol{\frac{x^{\alpha-1}e^{-\frac{\dot{x}}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}}$$

Substituting the α and β vales to equation (ii)

$$=\boldsymbol{\frac{x^{77-1}e^{-\frac{\dot{x}}{56}}}{\Gamma(77)56^{77}}}$$

$$f(x{:}\,77{,}56) = \boldsymbol{\frac{x^{77-1}e^{-\frac{\dot{x}}{56}}}{\Gamma(77)56^{77}}}$$

$$f(x{:}\,77{,}56) = \boldsymbol{\frac{x^{77-1}e^{-\frac{\dot{x}}{56}}}{\Gamma(77)56^{77}}}$$

To find the likelihood function f (x| θ), sample data follows gamma distribution α=3 and β=1/ θ

f(data |θ) = f(x:α,β) = $\boldsymbol{\frac{x^{\alpha-1}e^{-\frac{\dot{x}}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}}$

f(data |θ) = f(x:α,β) = $\boldsymbol{\frac{x^{3-1}e^{-x\theta}}{\Gamma(3)\beta^{\frac{1}{\theta^{3}}}}}$

The posterior distribution of θ is given by $\boldsymbol{f(\theta|data)} = \boldsymbol{\frac{f(data|\theta)f(\theta)}{\int_0^\infty f(data|\theta)f(\theta)d\theta}}$

$$\frac{f(data|\theta)f(\theta)}{\int_0^\infty f(data|\theta)f(\theta)d\theta} = \frac{\left(\prod_{i=1}^{n}\frac{{x_i}^{77-1}e^{x_i\theta}}{\Gamma(3)\left(\frac{1}{\theta}\right)^3}\right)\times\left(\frac{{x_i}^{77-1}e^{-0.17\theta}}{\Gamma(77)\times(56)^{77}}\right)}{\prod_{i=1}^{n}\frac{{x_i}^{77-1}e^{x_i\theta}}{\Gamma(3)\left(\frac{1}{\theta}\right)^3} * \left(\frac{{x_i}^{77-1}e^{-0.17\theta\cdot}}{\Gamma(77)\times(56)^{77}}\right)d\theta}$$

Removing the non θ terms from above equation

α-1 =3n+ 76 1/β = $\sum_{i=1}^{n} x_i + \frac{1}{56}$

the posterior distribution of θ, **α-1 =3n+ 76 1/β =** $\boldsymbol{\sum_{i=1}^{n} x_i + \frac{1}{56}}$

b) Bayes point estimate of θ associated with the square-error loss function square-error loss function.

Mode = $(3n + 77)x(\frac{1}{\overline{n.x}0.017})$

Mode = (3x10+77) *(1/ (291.7+0.017))

Mode = 0.3762

c) What is the Bayes point estimate of θ using the mode of the posterior distribution?

Mode = $(3n + 77 - 1)x(\frac{1}{\overline{n.x}0.017})$

Mode = (3x10+76) *(1/ (291.7+0.017))

Mode =0.37278

Reference

1. Casella, G., & Berger, R. L. (2002). Statistical inference (No. 2). Duxbury. Casscells, W., Schoen berger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. NewEngland Journal of Medicine, 299 (18), 999–1001.
2. Microsoft Excel as a tool for Data Analytics—Skillfin Learning. (n.d.). Retrieved December 28, 2022, from https://www.skillfinlearning.com/blog/microsoft-excel-as-a-tool-for-data-analytics Accessed from https://www.skillfinlearning.com/blog/microsoft-excel-as-a-tool-for-data-analytics#:~:text=Second%2C%20Excel%20has%20some%20great,apt%20as%20a%20visualization%20tool. (last ac- cess 28-12-2022)
3. Quantile function. (2022). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Quantile_function&oldid=1127668602 Accessedfromhttps://en.wikipedia.org/w/index.php?title=Quantile_function&oldid=1127668602 (last access 28-12-2022)
4. stat.cmu.edu/~brian/463-663/week09/Chapter%2003 (2003) Basics of Bayesian Statistics, Ac-cessed from https://www.stat.cmu.edu/~brian/463-663/week09/Chapter%2003.pdf(last access 28-12-2022)
5. https://ostwalprasad.github.io/machine-learning/Polynomial-Regression-using-statsmodel.html
6. https://stackoverflow.com/questions/61899474/polynomial-regression-using-statsmodels-formula-api
7. https://www.geeksforgeeks.org/ordinary-least-squares-ols-using-statsmodels/
8. https://towardsdatascience.com/polynomial-regression-in-python-b69ab7df6105
9. Bishop, C. (2007). *Pattern recognition and machine learning* (2nd ed.). Springer.
10. Bruce, P., & Bruce, A. (2017). *Statistics for data scientists: 50 essential concepts*. O'Reilly Publishing.
11. Hogg, R., McKean, J., Craig, A., (2020). *Introduction to mathematical statistics*, Pearson Education Canada.
12. Harvard Department of Physics. (2007, Fall). A summary of error propagation. Retrieved June 25, 2020
13. Illowsky, B., Dean, S. (2020). Introductory Statistics, OpenStax Textbook https://openstax.org/details/introductory-statistics
14. Liu, Y., & Abeyratne, A. I. (2019). *Practical applications of Bayesian reliability*. Wiley.
15. MacKay, D. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.