UNIVERSITY OF APPLIED SCIENCES

ADVANCED STATISTICS

Author: Anu Lekshmikutty Sasidharan

Matriculation: 32110119

Tutor: Paul Libbrecht

Parameters generated on 2022-11-27

with signature f4ec0f0a38bea4d9ea9860fb16775655dc22aab0

Date: 2022-12-13

Place: INDIA

Table of Contents

IMAGE DIRECTORY

ABBREVIATIONS

pmf: probability mass function
pdf: Probability density function
OLS: Ordinary Least Squares

SECTION 1 (Basic Probabilities and Visualizations - 1)

Task 1

Let's take a random sample of 100 people from the total voting population. Let X be the total number of "for" in the sample. The possible value of X1 (or the number of successes as a Bernoulli theorem (Hogg, 2020)) is 1,2, 3….100. And from the problem statement we know that p('for') =66, which means there are 66 people who have voted 'for'. Hence, we can evaluate the parameters.

Let us consider a sample with 100 outcomes: n = 100

The probability of voting 'for', p=p('for') = 0.66

Also, probability of voting 'against', q = 1-p = 1-0.66 = 0.34

Therefore, ξ1 can be approximated to a Bernoulli distribution with a probability of p=0.66

And the probability mass function(pmf) can be calculated as follows.

$f(x) = P(X = x) = p^x(1-p)^{(1-x)}\ ;\ x = 0,1$

Here,

x=0 represents the people who voted '*for*'

and

x=1 represents the people who voted '*against*'

Therefore,

$f(x) = 0.66x * 0.341 - x\ ;\ x = 0,1$

$f(0) = 0.660 * 0.341 - 0 = 1 * 0.34 = 0.34\ (34\%\ voted\ 'for')$

$f(1) = 0.661 * 0.341 - 1 = 0.66 * 1 = 0.66\ (66\%\ voted\ 'against')$

Also, the expectation can be calculated as follows.

$E(X) = \sum X * P(X = x) = 0 * 0.34 + 1 * 0.66 = 0.66$

Visualization Using Python Code:

```
#Import the library for the Bernoulli plot
from scipy.stats import bernoulli
import matplotlib.pyplot as plt

#Declare the Bernoulli distribution parameter p=0.66
bd=bernoulli(0.66)

# Probable x values; 0--> for, 1--> against
x=[0,1]

# To visualize the bar plot
plt.figure(figsize=(8,8))
plt.xlim(-2, 2)
plt.bar(x, bd.pmf(x), color='blue')

# Plot labelling
plt.title('Bernoulli distribution (p=0.66)', fontsize='15')
plt.xlabel('Voting Distribution: 0--> for; 1--> Against', fontsize='15')
plt.ylabel('Probability P(X=x)', fontsize='15')

plt.show()
```

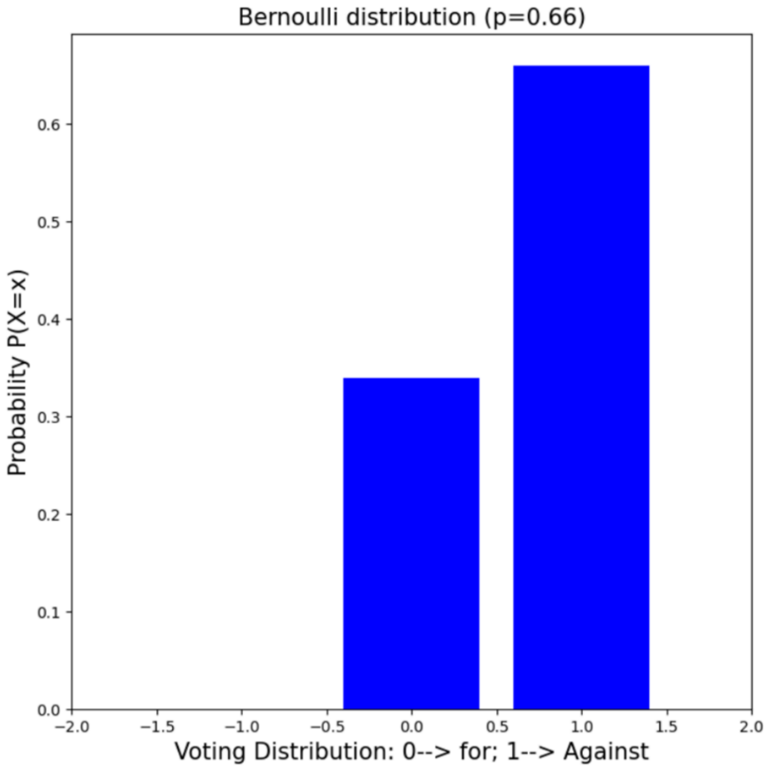

Figure 1: Probability of votes using Bernoulli Distribution

Task 2

Here the Poisson Distribution (Illowsky, 2013) can be used because it is a discrete distribution function describing the probability that an event will occur a certain number of times in a fixed time (or space) interval.

$$\boldsymbol{Poi(k;\lambda) = \left(\frac{\lambda^k}{k!}\right) e^{-\lambda}}$$

Also, we can take many trials n for the Poison distribution while keeping the mean fixed. That means, Expectation is the same as that of variance.

$$\boldsymbol{E(X) = \lambda = \sigma^2 = np}$$

If X1 is 1: a Poisson distribution with an expectation λ = 66

Here

$$\boldsymbol{E(X) = \lambda = \sigma^2 = np = 66}$$

That means the average number of meteorites falling in a year = 66

Median:	66.0
Variance:	66.0
Mean:	66.0

For example:

Probability of 1 meteorite falling in the ocean in a year

$$\mathrm{P(X = 1)} = \left(\frac{(66)^1}{1!}\right) * e^{-66}$$

Probability of 2 meteorites falling into the ocean in a year

$$\mathrm{P(X = 2)} = \left(\frac{(66)^2}{2!}\right) * e^{-66}$$

Visualization Using Python Code:

```
#Import required libraries
from scipy.stats import poisson
from matplotlib import pyplot as plt

lamda=66
# Calculating the median and variance
mean=poisson.mean(lamda, loc=0)
median=poisson.median(lamda, loc=0)
variance=poisson.var(lamda, loc=0)

print("-"*60)
print("The mean, median and variance of Poisson Distribution data")
print("-"*60)
print("Mean:", mean)
print("Median:", median)
print("Variance:", variance)
print("-"*60)
print("Poisson Distribution Plot along with its mean and variance")
print("-"*60)

# To plot the Poisson Distribution, we need x and y axis data.
# We can generate it and store them in separate arrays for plotting
# Let y axis represent the probability
y_poi_values=[]
# And the x-axis represents the number of meteorites falling on earth.
x_falling_meteorites=[]
#Generate Poisson distribution with sample size 100
for i in range(100):
    y=poisson.pmf(k=i,mu=lamda)
    if y>0.005: #Condition given is "probability remains provably less than
0.5%""
        y_poi_values.append(y)
        x_falling_meteorites.append(i)

# Poisson Distribution plot for falling meteorites along with its mean and
variance.
plt.figure(figsize=(7,7))
plt.plot(x_falling_meteorites,y_poi_values,color="blue",linestyle='dashed',
label="Poisson Destribution")
plt.scatter(median,max(y_poi_values), color="red",label="Median",s=75)
plt.scatter(variance,max(y_poi_values), color="green",marker="x",label="Vari-
ance" ,s=150)
plt.title("Poisson Distribution of meteorites shower in a year")
plt.xlabel("Number of meteorites falling on earth in a year")
plt.ylabel("Probability(in %)")

plt.legend(loc=1)
plt.show()
```

```
----------------------------------------------------------
The mean, median and variance of Poisson Distribution data
----------------------------------------------------------
Mean: 66.0
Median: 66.0
Variance: 66.0
----------------------------------------------------------
Poisson Distribution Plot along with its mean and variance
----------------------------------------------------------
```

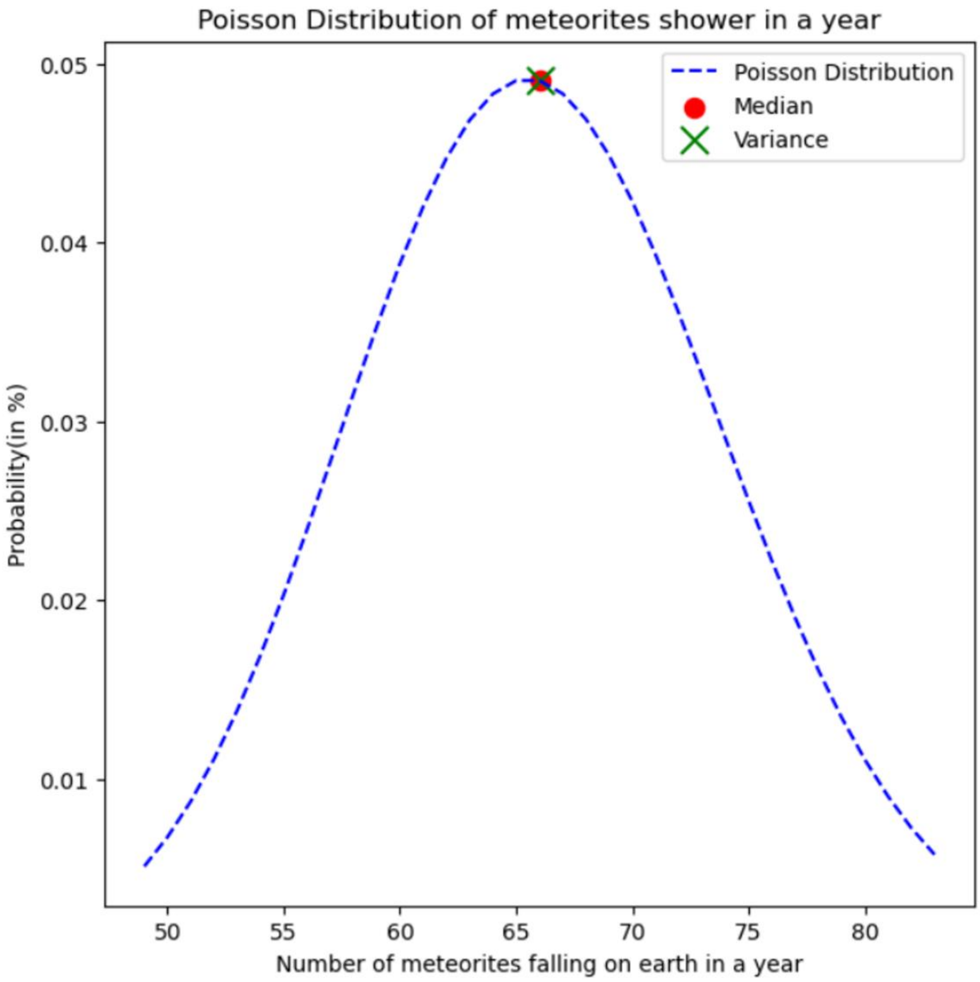

Figure 2: Poisson Distribution of meteorites shower per year

SECTION 2 (Basic Probabilities and Visualizations - 2)

Let us Consider ‘Y’ as the random variable that denotes the time to hear the owl. Furthermore, the probability that we need to wait for 2 to 4 hours can be derived as follows.

$P[Y > y] \ = \frac{77}{99}e^{-7\sqrt{y}} \ + \frac{22}{99}e^{-4\sqrt[3]{y}}$

And P [Y ≤ y] is given by

$$P[Y \leq y] \ = \ 1 \ - \ P[Y > y] \ = \ 1 \ - \ \frac{77}{99}e^{-7\sqrt{y}} \ - \ \frac{22}{99}e^{-4\sqrt[3]{y}}$$

Therefore, the probability that we need to wait between 2 to 4 hours is given by

$$\begin{aligned}
P[2 \leq Y \leq 4] \ &= P[Y \leq 4] \ - P[Y \leq 2] \\
&= \left(1 \ - \ \frac{77}{99}e^{-7\sqrt{y}} \ - \ \frac{22}{99}e^{-4\sqrt[3]{y}}\right)_{y=4} - \left(1 \ - \ \frac{77}{99}e^{-7\sqrt{y}} \ - \ \frac{22}{99}e^{-4\sqrt[3]{y}}\right)_{y=2} \\
&= \left(1 \ - \ \frac{77}{99}e^{-7\sqrt{4}} \ - \ \frac{22}{99}e^{-4\sqrt[3]{4}}\right) - \left(1 \ - \ \frac{77}{99}e^{-7\sqrt{2}} \ - \ \frac{22}{99}e^{-4\sqrt[3]{2}}\right) \\
&= \frac{77}{99}e^{-7\sqrt{2}} \ + \ \frac{22}{99}e^{-4\sqrt[3]{2}} \ - \ \frac{77}{99}e^{-7\sqrt{4}} \ - \ \frac{22}{99}e^{-4\sqrt[3]{4}} \\
&= \frac{77}{99}e^{-9.899} \ + \ \frac{22}{99}e^{-5.038} \ - \ \frac{77}{99}e^{-14} \ - \ \frac{22}{99}e^{-6.347} \\
&= \frac{77}{99 * e^{9.899}} + \frac{22}{99 * e^{5.038}} - \frac{77}{99 * e^{14}} - \frac{22}{99 * e^{6.347}} \\
&= \frac{77}{1971134} + \frac{22}{15261} - \frac{77}{119057824} - \frac{22}{56507}
\end{aligned}$$

Therefore,

$P[2 \leq Y \leq 4]$ =0.001106455369677151

Probability density calculation:

$$\boldsymbol{PDF = \int_{119}^{241} \frac{77}{99}e^{-7\sqrt{y}} \ + \ \frac{22}{99}e^{-4\sqrt[3]{y}}}$$

Python code for visualization of probability density function graph as well as a histogram by the minute as well as the mean, variance, and quartiles (Hogg, 2020) of the waiting times.

Visualization Using Python Code:

```
import numpy as np
import scipy.integrate as spi
from matplotlib import pyplot as plt


"""
First, we need to calculate the probability density using the given equation
for doing the integration of the function to find the area under the curve we
need to use the scipy module
From the equation, "y= a*np.exp(-7*y**0.5) + b*np.exp(-4*y**0.33)"
we can see the value of a and b are repeating decimals (a=0.7777... and
b=0.2222....),
which can be denoted as the fraction of numbers as well. (a=77/99, b=22/99)
Define the probability density of variables a and b
"""

def probability_density(a,b):
    a=(a*100)/99
    b=(b*100)/99
    return lambda y: a*np.exp(-7*y**0.5) + b*np.exp(-4*y**0.33)

"""
Using the probability density calculated on the above step, we can find out
the
probability density function graph parameters like the probability of each
minute.
our waiting period is 2hrs (120 minutes) to 4 hours(240minutes)
So we need to find the probability of each minute from 120 min to 240 minutes
To cover both starting and ending minutes we must calculate the pdf from 119
to 241

"""
def pdf(a,b):
    probability_per_minute=[]
    for i in range(119, 241):
        result,none=spi.fixed_quad(probability_density(a,b), i/60, (i+1)/60)
        probability_per_minute.append(result*100)

    """
    In the for loop: we are considering the starting and ending minutes as
well for calculation
    integrate over the limit 119,241 using the fixed quad function in SciPy
to get the Gaussian quadrature
    which will give two values one is the integral value and the other is the
error estimate (None) which can be discarded

    """
    # probability_per_hour(probability that we need to wait between 2 to 4
hours is given by)
    probability_per_hour, none = spi.fixed_quad(probability_density(a,b),2,4)
    print("-"*100)
    print("The probability that we need to wait between 2 to 4 hours to hear
the owl:",probability_per_hour*100)

    # Calculation of Mean, and variations
    mean = np.mean(probability_per_minute)
    variance = (np.std(probability_per_minute))**2
```

```
    print("The mean values is:" ,mean)
    print("And the variance is: ",variance)
    print("-"*100)

    # Graphical representation of the pdf, mean, variance, and quartiles
    print("Graphical representation of the pdf, mean, variance, and
quartiles")
    print("-"*100)
    plt.figure(figsize=(10,7))
    plt.rcParams.update({'font.size': 13})
    #pdf

plt.plot(np.arange(119,241,1),probability_per_minute,color='#34495E',label="P
DF Plot" )
    #Histogram
    plt.bar(np.arange(119,241,1), probability_per_minute,color="#AED6F1",
label="Histogram Plot" )
    #Q1,Q2,Q3 Quartiles
    """
    For the Q1, Q2, and Q3 quartiles, we can divide the total time frame into
30 minutes divisions, 0,30,60,90,120
    here we can consider 30 as Q1 point, 60 as Q2 point and 90 as Q3 point,
    When plotting on the 120-240 minutes x-axis graph, we will get the
corresponding points as 150, 180, and 210-minute points.
    """
    plt.scatter(150, probability_per_minute[30], marker="H",label="Q1"
,s=150)
    plt.scatter(180, probability_per_minute[60], marker="H",label="Q2"
,s=150)
    plt.scatter(210, probability_per_minute[90], marker="H",label="Q3"
,s=150)
    #plot labels and legends declarations:
    plt.legend(loc=1)
    plt.title("Graphical representation of the pdf, histogram and quartiles")
    plt.xlabel("Time between 2-4hrs in minutes")
    plt.ylabel("Probability of hearing the owl")
    plt.show()


pdf(0.77,0.22)
```

```
-----------------------------------------------------------------------------------------------
The probability that we need to wait between 2 to 4 hours to hear the owl: 0.15729086876079
The mean values is: 0.001315289534293259
And the variance is:  2.641393632630656e-07
-----------------------------------------------------------------------------------------------
Graphical representation of the pdf, mean, variance, and quartiles
-----------------------------------------------------------------------------------------------
```

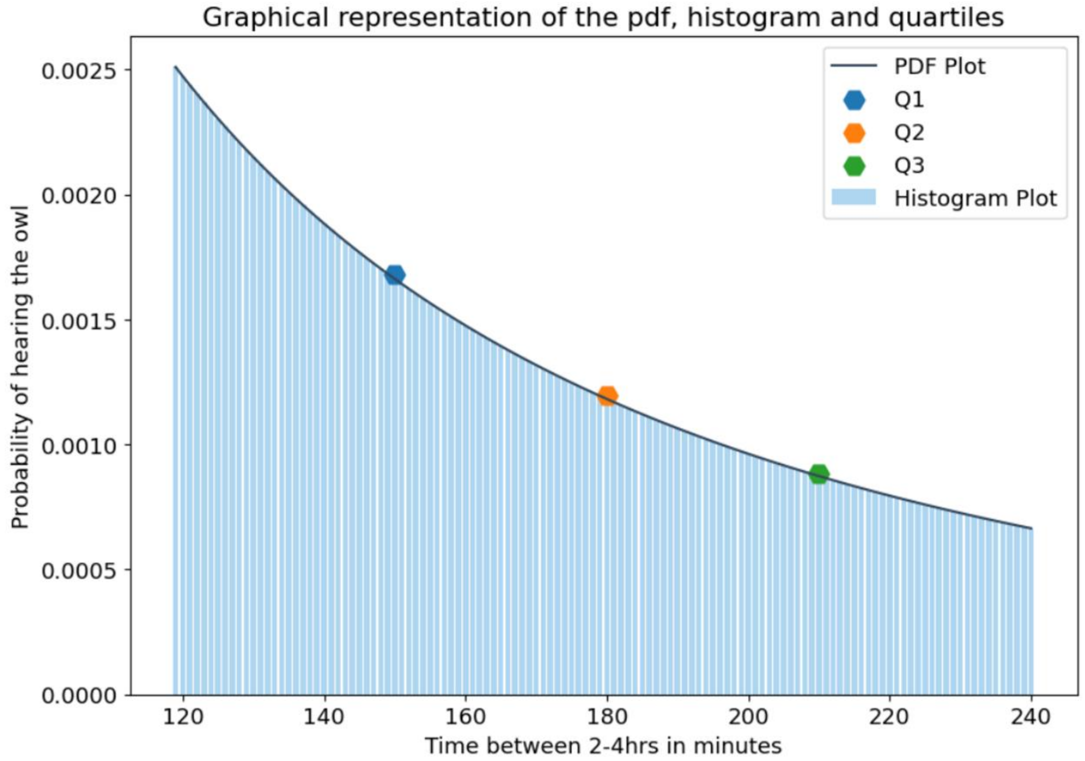

Figure 3: Graphical representation of PDF, Quartiles, and Histogram in a single plot

SECTION 3 (Transformed Random Variables)

Assume that a network router has a total bandwidth (T) to the first hardware failure, which is modeled by the exponential distribution function given below.

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

Task 1

Let f(S1) & f(S2) denote T (hardware failure), the total bandwidth over which individual routers S1 and S2 failed, and compute the density function of the random variable T.

Let us assume that the pre-failure bandwidth (T) of the two routers is independent of each other,

T = sum of 2 independent and identically distributed random variables (Hogg, 2020)

$$T = \gamma(shape = 2, scale = \theta)$$

$$T = f(S1) + f(S2)$$

Which is the given probability density function.

$$f(S1) = \frac{1}{\theta}\, e^{\frac{-t1}{\theta}}$$

$$f(S2) = \frac{1}{\theta}\, e^{\frac{-t2}{\theta}}$$

Let us calculate the density function on sample T which will be the density function in terms of $\boldsymbol{f(S)}$

$$f(t) = \int \frac{1}{\theta}\, e^{\frac{-t}{\theta}}\, dt$$

$$f(t) = e^{-\frac{t}{\theta}} + c$$

Task 2

Calculate the maximum likelihood function for the samples T1, T2, ... Tn, used in the actual experiment, with the following values.

- ξ_{10}: 39, 74, 126, 5, 156

Estimate the model parameters with maximum likelihood and calculate the expected value of the total bandwidth to failure of the dual-router system.

Given that

$$f(T) = \frac{1}{\theta}\, e^{-\frac{T}{\theta}}$$

To find the maximum likelihood function T_i (Hogg, 2020)

$$L(T1, T2, .. Tn) = \prod_{I=1}^{n} \frac{1}{\theta} e^{-\frac{Ti}{\theta}}$$

$$= \frac{1}{\theta}\, e^{-\frac{T1}{\theta}} * \frac{1}{\theta}\, e^{-\frac{T2}{\theta}} * \ldots .* \frac{1}{\theta}\, e^{-\frac{Tn}{\theta}}$$

$$= \frac{1}{\theta^n}\, e^{-\frac{1}{\theta}[T1+T2+\cdots .+Tn]}$$

Let's use log function to find the likelihood function.

$$Log\ L = \log\left(\frac{1}{\theta^n}\right) + \log e^{-\frac{1}{\theta}[T1+T2+\cdots+Tn]}$$

In order to estimate the model parameters with the maximum likelihood function, we need to make the derivative of the above equation zero, which means we need to make the fraction function zero.

$$\frac{d[LogL]}{d\theta} = -\frac{n}{\theta} + \frac{T1 + T2 + \cdots + Tn}{\theta^2} = 0$$

$$\frac{n}{\theta} = \frac{T1 + T2 + \cdots + Tn}{\theta^2}$$

$$\theta = \frac{T1 + T2 + \cdots + Tn}{n}$$

Bu substituting the values for T1, T2, …Tn

$$\hat{\theta} = \frac{39 + 74 + 126 + 5 + 156}{5} = 80$$

Therefore, the Expectation value = 2* $\hat{\theta}$ = 2*80 = 160

SECTION 4 (Hypothesis Test)

The population statistical data is given below.

The Population mean, $\mu_0 = 937$

Population Standard deviation, $\sigma = 11.5$

Total number of hammers, N = 1000

Sample statistics data:

Sample data (x): 912, 987, 812, 856, 1018, 935, 960, 925, 902, 969

Number of sample data, n=10

The sum of sample data, $\sum x = 9276$

The sample mean, $\bar{x} = \frac{\sum x}{n} = \frac{9276}{10} = 927.6$

To find the standard deviation (Illowsky, 2013) of the sample data,

$$S = \sqrt{\frac{1}{n-1}\sum(x-\bar{x})^2}$$

$$S = \sqrt{\frac{1}{9}\sum(x-927.6)^2}$$

Were, x: 912, 987, 812, 856, 1018, 935, 960, 925, 902, 969

Therefore, $S = 61.3862$

Here we must test the following condition for the null hypothesis as well as the alternate hypothesis (Hogg, 2020).

H_0: μ=937 against H_1: μ > 937(New system)

Here we must rely on the z-test (Illowsky, 2013) (Jim, n.d.) to draw a conclusion because we already know the population standard deviation.

The test statistics, $z = \frac{(\bar{x} - \mu_0)^2}{\frac{\sigma}{\sqrt{n}}}$

$$z = \frac{(927.6 - 937)^2}{\frac{11.5}{\sqrt{10}}}$$

$$z = \frac{-9.4}{3.6366} = -2.5848$$

Therefore, the test statistic z = -2.5848

Let us assume $\alpha = 0.05$ as the significance level. From the z-table (Jim, n.d.), it can be deduced that the corresponding two tail p-values are 0.0049, because p-value (0.0049) < α ($\alpha = 0.05$) we no longer need to hold the null hypothesis and can reject the null hypothesis, H_0.

Conclusion:

From the evaluation, we can conclude that the new system (H_1) will produce more consistent results. That means the new system is more consistent.

SECTION 5 (Regularized Regression)

Task 1

OLS estimate (Bruce, 2017):

Here, our observed data (ξ_{16}) is denoted as (x_i,y_i); where i ranges from 1 to N. x and y are the independent and dependent variables, respectively. Let us assume that the function f(.|θ) is linear in the parameters θ.

The OLS estimate is the sum of the intercept and slope to reduce the squared error.

The OLS estimate, $\theta_{OLS} = \frac{\sum_{i=1}^{N} X_i Y_i}{\sum_{i=1}^{N} X_i^2}$

From the above data the following data can be derived.

Sum	**ΣX**	670	**ΣY**	2.50626E+23
Number of observed data(N)	24			
Average	$\bar{X}$	27.9166667	$\bar{Y}$	1.04427E+22

Also,

Sum of squares and Products	**ΣX²**	84782	**ΣY²**	1.25787E+46	**ΣXY**	2.33637E+25

There for the OLS estimate is evaluated as follows.

$$\theta_{OLS} = \frac{2.33637\text{E+}25}{84782} = 2.75573\text{E+}20$$

Task 2

OLS ridge-regularized estimates (Bruce, 2017) (KENNARD, 1970):

Ridge regularized estimation is used to reduce the residuals of the estimated regression line.

Let's consider a polynomial function for evaluating ridge regularization.

We can calculate the cost function for a linear regression line as follows.

$$C(\theta) = \sum_{i=1}^{n} (y - \hat{y})^2 + \lambda (slope)^2$$

In the case of ridge regularization, we need to reduce this cost function so that the predictions fit the regression line. Assume that the X and Y values are centered so that we can ignore constant values from the slope intersection formula Consider X as an n x n probability-matrix and Y as n-vectors.

$$\hat{\beta} = (\acute{X}X)^{-1} \acute{X}Y$$

This can be optimized by adding a small constant value λ to the diagonal elements of the matrix.

$$\widehat{\beta_{ridge}} = (\acute{X}X + \lambda I_p)^{-1} \acute{X}Y$$

Here, the ridge regularization adds a constant to reduce the penalized sum of the squares in the calculation in order to fit the best line. Therefore, the cost function can be rewritten as follows.

$$C(\theta) = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^{\,2}$$

This means that the slope multiplied by the factor lambda (λ), the penalty term. To reduce the difference between the actual regression line and the squared error, we need to reduce the slope so that the OLS estimate is the best fit. This is the use of ridge-regularized estimates in linear regressions. Let's consider that the penalty is very small and negligible, so the cost estimator can be written as follows.

$$C(\theta) = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

SECTION 6 (Bayesian Estimates)

Task 1

The posterior distribution of θ is given by (Hogg, 2020)

$$f(\theta|data) = \frac{f(data|\theta).f(\theta)}{\int_0^\infty f(data|\theta).f(\theta)d\theta}$$

Here it is given from the problem that the value of α = 3 and β = 1/θ

Therefore,

$$X_i|\theta = \gamma(3, \frac{1}{\theta})$$

The gamma distribution with both X and θ is given below.

$$f(x) = \frac{x^{\alpha-1}.e^{\frac{-x}{\beta}}}{\gamma(\alpha).\beta^\alpha}$$

Generally, the posterior pdf k(θ|X) is given by

$$k(\theta|x) = \frac{L(x|\theta).h(\theta)}{g_1(x)}$$

Where, $g_1(x)$ is given as follows

$$g_1(x) = \int_{-\infty}^{\infty} L(x|\theta).h(\theta)d\theta$$

That means,

$$k(\theta|x) \propto L(x|\theta).h(\theta)$$

Therefore, $g_1(x)$ can be dropped.

$$L(x|\theta) = \prod_{i=1}^{n} \frac{x_i^{3-1}e^{\frac{-x_i}{1/\theta}}}{\gamma(3).(1/\theta)^3}$$

$$L(x|\theta) = \prod_{i=1}^{n} \frac{x_i^2.\theta^3e^{-x_i\theta}}{\gamma(3)}$$

Applying the limits,

$$L(x|\theta) = \frac{x^{2n}.\theta^{3n}.e^{-\theta}.\sum_{i=1}^{n} x_i}{\gamma(3)^n}$$

Also, the prior pdf is given by,

$$h(\theta) = \frac{\theta^2.e^{-\theta/2}}{2^3.\gamma(3)}$$

Combining, we can rewrite the k(θ|x)

$$k(\theta|x) \propto L(x|\theta).h(\theta)$$

$$k(\theta|x) \propto \frac{x^{2n}.\theta^{3n}.e^{-\theta}.\sum_{i=1}^{n} x_i}{\gamma(3)^n} * \frac{\theta^2.e^{-\theta/2}}{2^3.\gamma(3)}$$

For the range, 0<θ<∞,

$$k(\theta|x) \propto \theta^{3n}.\theta^3.e^{-\theta}.\sum_{i=1}^{n} x_i \; e^{-\theta/2}$$

$$k(\theta|x) \propto \theta^{3n+3}.e^{-\theta(\frac{1}{2}+\sum_{i=1}^{n} x_i)}$$

$$\sim\gamma(\alpha = 3n+3,\; \beta = \frac{1}{\sum x_i + 0.5})$$

The posterior distribution is given by

$$\sim\gamma(\alpha = 3n+3,\; \beta = \frac{1}{\sum x_i + 0.5})$$

Task 2

Given $\bar{x} = 38.1$

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

$$mode = \alpha\beta$$

$$mode = (3n+3) * (\frac{1}{\sum x_i + 0.5})$$

From the problem it is given that the number of observations equals 10,

Therefore,

$$mode = (3*10+3)*\left(\frac{1}{38.1*10+0.5}\right) = 33 * \frac{1}{382.3} = 0.08631$$

Task 3

Bayes point estimate of θ using the mode of the posterior distribution

$$mode = (\alpha - 1)\beta$$

$$mode = 32 * \frac{1}{382.3} = 0.8370$$

Bibliography

Bruce, P. (2017). *Practical statistics for data scientists.* O'REILLY.

Hogg, R. M. (2020). *Introduction to mathematical statistics.* Harlow, England Pearson, [2020].

Illowsky, B. (2013). *Introductory statistics.* Houston, TX: Victoria: OpenStax CollegeBCcampus, BC Open Textbook Project, 2013. Revision ST-1-000-RS.

Jim. (n.d.). *Z-table*. Retrieved from Statistics By Jim: https://statisticsbyjim.com/hypothesis-testing/z-table/

KENNARD, E. H. (1970). *Ridge Regression: Biased Estimation for Nonorthogonal Problems.*