

**Advanced Statistics Workbook**

(DLMDSAS01)

Author - Vinay Tula

Registration number - 92113281

Parameters generated on 30th Jan 2022with signature 5c96e5906fee9f5ef5506d27e1fb9c73a9853f8e

Date - 30th Jan 2022

Place - Hyderabad, India

## Table of contents

[BASIC PROBABILITIES AND VISUALIZATIONS (1)… 3](#_Toc47428176)

[Section 1a 3](#_Toc47428177)

[Section 1b](#_Toc47428177) 4

[Section 1c](#_Toc47428177) 6

[BASIC PROBABILITIES AND VISUALIZATIONS (2)…](#_Toc47428176) 9

[Section 2a](#_Toc47428177) 9

[Section 2b](#_Toc47428177) 11

[A SIMPLE PARAMETER ESTIMATION… 1](#_Toc47428178)5

[HYPOTHESIS TEST… 1](#_Toc47428178)8

[SUFFICIENT STATISTICS… 2](#_Toc47428178)0

[BAYESIAN ESTIMATES…](#_Toc47428176) 21

[Section 6a](#_Toc47428177) 21

[Section 6b](#_Toc47428177) 21

[Section 6c](#_Toc47428177) 22

[Literature](#_Toc47428179) 23

## Section 1

BASIC PROBABILITIES AND VISUALIZATIONS (1)

### Section 1a

A vote with outcome 𝑓𝑜𝑟 or 𝑎𝑔𝑎𝑖𝑛𝑠𝑡 follows a Bernoulli distribution where p𝑃(vote = “for”) = 0.07 . Represent the proportion of “for” and “against” in this single Bernoulli trial using a graphics and a percentage. Can an expectation be calculated? Justify your answer.

Given

p(vote = “for”) = 0.07

From Bernoulli Distribution we know that

p(vote="against") = (1-p) = (1 - p(vote="for")) = 0.61

To represent the above probability we are using python scipy.stats library and we get

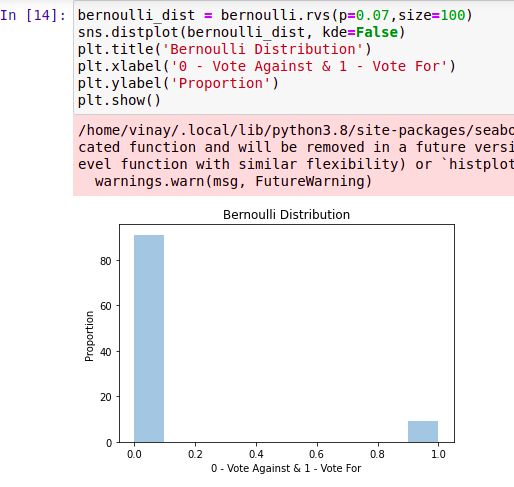


Figure 1 Bernoulli Distribution

Getting the value in terms of percentage is:

Votes against is calculated as (1-0.07)X100 = 93.0%

Votes for is (0.07)X100 = 7%

### Section 1b

The number of meteorites falling on an ocean in a given year can be modelled by a Poisson

distribution with an expectation of 𝜆= 44 . Explain why a Poisson distribution is a natural

candidate for this phenomenon. Give a graphic showing the probability of one, two, three…

meteorites falling (until the probability is less than 0.5%). Calculate the median and variance

and show them graphically on this graphic.

Given

𝜆 = 44, So the veriance will also be 44

The median for each position Distribution will be as

= ~𝜆 + 1/3 -0.02/𝜆 = 44

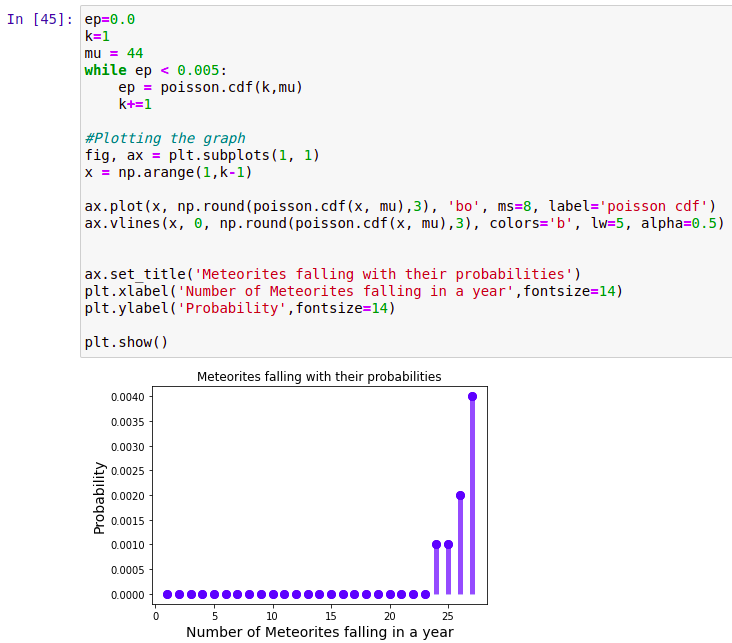


Figure 2 Probability of falling meteorites

The position distribution is given by

Predicting the probability of falling meteorites by taking samples of 100 as

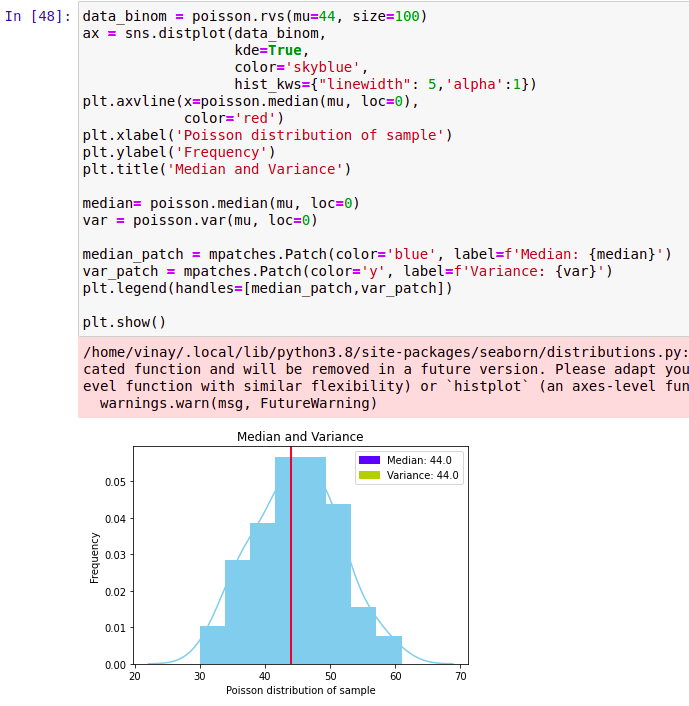


Figure 3 Showing median and variance

### Section 1c

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:

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.

P(Y>2) =G(2)= 1-P(2) = 1-(

0.5330 = 0.4669

The probability to wait of 2 hours is 0.4669

P(Y>4) =G(4)= 1-P(4) = 1-(

1-0.2969 = 0.7030

The probability to wait of 4 hours is 0.7030

To find the probability of waiting time in between the 2 and 4 hours is

P(2<y<4) = P(2) - P(4)

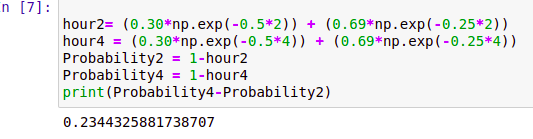


Figure 4 Probability wait time in between 2 and 4 hour

Now splitting the hours into equal amount of minutes and finding the probability of each minute by using

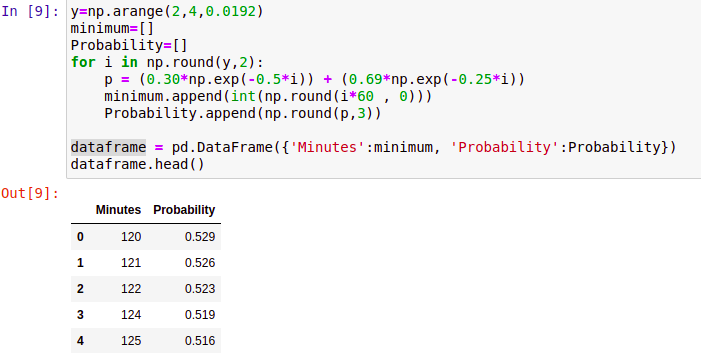


Figure 5 Splitting hours into minutes

Plotting the data-frame into line plot to know the probability

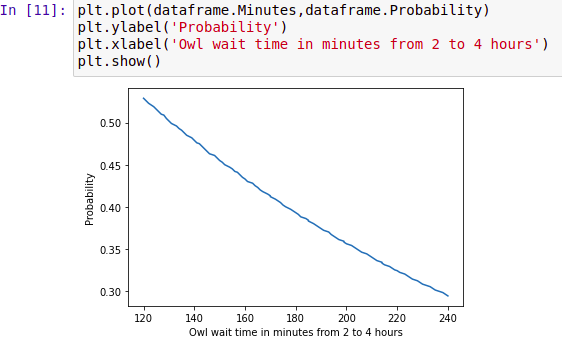


Figure 6 plot based on the data-frame

The below histogram gives the wait time probability of 2 to 4 hours

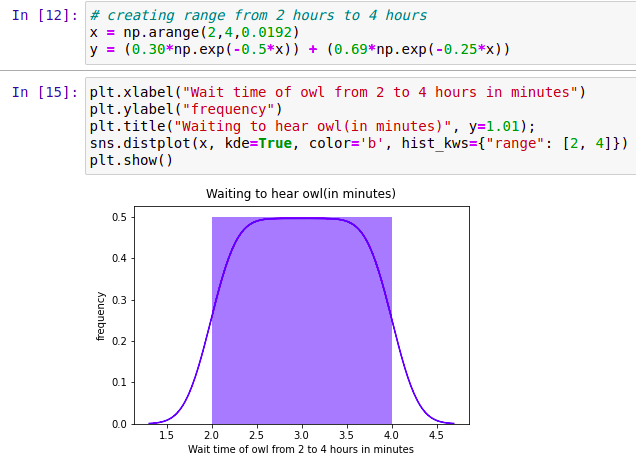


Figure 7 Wait time in minutes

## Section 2

**BASIC PROBABILITIES AND VISUALIZATIONS (2)**

### Section 2a

Consider the variables 𝑋 and 𝑌. The realization of a sample of size 20 is given below (where 𝑋 is

the first variable and 𝑌 is the second):

(0.531, 185.25), (-1.711, -14.3), (2.784, 239.72), (-3.33, -284.38), (4.206, 150.2), (4.816, 437.65), (7.785, 237.82), (-3.473, -349.61), (4.062, 671.91), (0.369, 227.18), (4.632, -231.02), (-9.556, 553.03), (5.092, -695.55), (-10.452, 142.98), (-3.413, -104.25), (-3.007, 506.92), (3.899, 428.88), (1.97, 777.71), (0.607, -349.13), (-4.86, 422.26) Sketch an appropriate plot that displays the values of these points. Calculate the sample covariance as well as the sample’s expectations and variances of 𝑋 X and Y𝑌.

Importing the required libraries in the jupyter notebook as

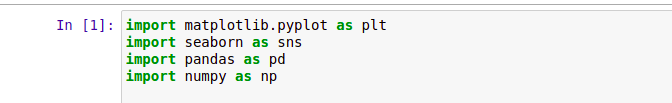


Figure 8 Adding required libraries

Now creating a Data-frame from the given data and it is assigned to a variable as

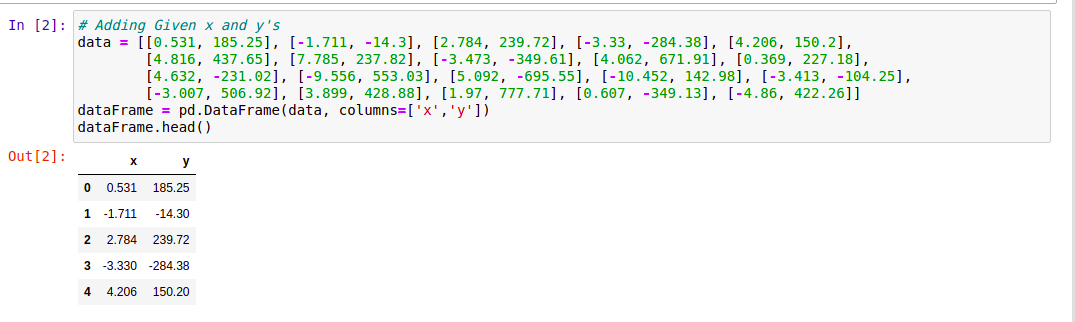


Figure 9 Creating a Data-frame

Plotting the data in a scatter plot using seaborn which we will be seeing as

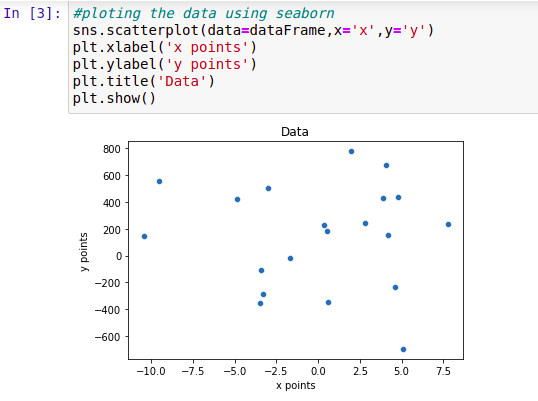


Figure 10 plotting points

Creating a function for calculating the mean of x and y points



Figure 11 Mean function

By using the mean function we get the mean of x and y as

= 0.0475, = 147.6634

We know the formula for covariance which is defined as

covariance(x,y) =

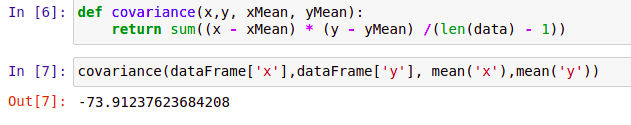


Figure 12 Finding covariance

The covariance of the given data (x, y) is -73.9123

Now lets find the variance

The variance is given by the

variance =

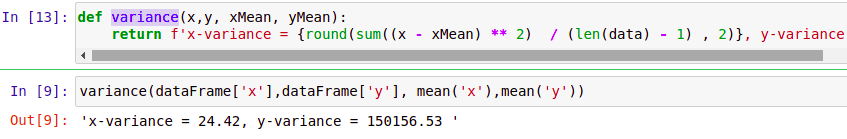


Figure 13 variance of x and y

The variance of the given data x is 24.42 and y is 150156.53.

In pandas library we are having some in build methods using that we can verify the x and y values as

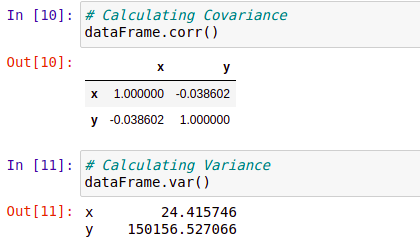


Figure 14 Verifying

### Section 2b

Consider that a ball is thrown with a random angle 𝜃 ∈ [0, 360) (in degrees) and a random

radius 𝑟 ∈ [0, 1] (in meters) both independent and uniform. Calculate the density of the

variables 𝑋 and 𝑌 (the cartesian coordinates of the point at angle 𝜃 and radius 𝑟) as well as their

expectation and variance.

Given,

where 360 = 2

r

We need to find pdf of x and y

where

x = rcos (Equation1)

y = rsin (Equation2)

() = where r

Finding inverse transform given

tan =

=

Squaring on both equations 1 and 2 adding given

+ = ( + )

r = =

We may readily check that the transformation is one to one. As a result, we can quickly apply the transformation formula.

Applying transformation of random variable formula

we get,

where | =

We get = ) |

| =

=

=

From above we can say that x and y are not independent

Marginal value can be calculated as

=

put gives dy = x dt

=

=

We don't need to calculate marginal pdf for both x and y since the distribution is symi in both.So, x and y are dependent random variable with pdf

, x

, y

Now,

=

=

Putting x = rcos

Given dx = -sin

(x) =

=

=

=

(y) =

Now,

var(x) = () -

=

putting x = sin and dx = cos

var(x) =

=

=

=

var(x) var(y) =

## Section 3

**A SIMPLE PARAMETER ESTIMATION**

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:

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 and . 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 𝑇 , 𝑇 , …, 𝑇, calculate the likelihood function for 𝜃𝜃𝜃. 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 following bandwidth totals to failure: 80, 0, 56, 62, 176 Estimate the model-parameter with the maximum likelihood and compute the expectation of the bandwidth total to failure of the dual-router-system.

Let and be bandwidth of the first and second router respectively.

So that the gives the relation between a total bandwidth failure and failure of individual router.

The Cumulative Distribution Function of the can be derived as the:

By assuming that S1 and S2 are independent, the Periodic Distribution Function of a T may now be calculated. This is a reasonable assumption because one router's bandwidth loss should not be dependent on the other's loss.

Using the derivative of the Periodic Distribution Function yields the Probability Density Function of T:

The Probability Density function of variable T is given by:

Given an experiment with the dual-router-system yielding a sample 𝑇1, 𝑇2, …, 𝑇𝑛 the likelihood function is given as follows:

A log transformation of the above likelihood function will yield the same maxima and is easier to compute. The log likelihood function is given by:

For finding the Maximum Likelihood Estimate of we solve the following maximization exercise:

The First Order Condition of Optimization is given by:

So we have to solve the following equation:

Now the sample is given by 80, 0, 56, 62, 176 the sample mean is given by 74.8

is the sample expectation = 74.8

is the Maximum Likelihood Estimate = 37.4

## Section 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 971𝑔 and standard deviation of 29 𝑔. Propose a model for the weight of the hammers including a probability distribution for the weight. What are the assumptions for this model to hold? What parameters does this model have?

A new production system is configured, and one wants to evaluate if the new system makes more constant weights. For this a random sample of newly produced hammers is evaluated yielding the following weights:

941, 1011, 930, 911, 971, 886, 125, 953, 961, 879

What hypothesis can you formulate and what test and decision rule can you make to estimate if the new system produces a more constant weight? Express these assertions as logical statements involving critical values. What error probabilities can you suggest and why? Calculate the 𝑝-value. Perform the test and express conclusions.

Here,

Population mean = 971

Population standard derivation

Total hammers, n = 1000

Then, We take 10 Samples

= 941+ 1011+ 930+ 911+ 971+ 886+ 1125+ 953+ 961+ 879

= = 956.8

mean

We know the Variance as

5065.95

s = = 71.1755

We also know the Standard deviation is

= = 841

We have to test Null Hypothesis as follows:

We have assumed that the samples are derived from a population which follows a Normal Distribution, As per [Illowsky, B., Dean, S. (2020). Introductory Statistics, OpenStax Textbook](https://openstax.org/books/introductory-statistics/pages/11-6-test-of-a-single-variance) we can use a Chi-Squared test to test.

= 0.5421

The critical value for for 9 degrees of freedom is 2.088

Since we reject the Null Hypothesis and conclude that the new procedure produces more constant weights.

The p-value or the observed level of significance is 0. So the probability of Type – I error, i.e., Probability of rejecting the Null Hypothesis when it is actually True is negligible.

## Section 5

**SUFFICIENT STATISTICS**

Let 𝑋 , 𝑋 , … , 𝑋 𝑛 be the random sample of a positive random variable 𝑋 having the density function:

𝑥 =

with one parameter 𝜃 ∈ ℝ + . Find an estimator for 𝜃 that is sufficient.

Given, , 𝑋 , … , ͠ 𝑥 =

The joint density of x=( , , … ) is given by

f(x,0) =

From the Neyman theorem we know that the f(x;θ) is given by

= g(T(X),θ)h(x)

= G(T,) h(x) { By Neyman factorization theorem}

Where, G(T,) = , T(x)=

And h(x) =

Then by Neyman factorization theorem T= is sufficient statistic for

## Section 6

**BAYESIAN ESTIMATES**

Let 𝑋 1 , 𝑋 2 , …, 𝑋 10 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 = 3 and 𝛽 = 2:

### Section 6a

### Find the posterior distribution of θ.

Appropriately gamma function is

### Section 6b

If the observed = 17.5, what is the Bayes point estimate associated with the square-error loss function?

Under square error loss,The bayes estimate is the posterior mean

Mode = (3n +3) X ()

= (3 x 10 + 3) X ()

= (30 + 3) X ()

= 33 X = 0.1880

### 

### Section 6c

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

Mode = (3n +3-1) X ()

= (3 x 10 + 3-1) X ()

= (30 + 2) X ()

= 32 X = 0.1823

Mode = 0.1823

## Literature

All the code related to the workbook is present in my git repository which is

<https://github.com/vikasvinay/DLMDSAS01-Advanced-Statistics.git>