Project Component 1 - EE201

Omkar Jadhav , 190010029
October 20, 2020

1 Understanding the problem

We are given a dataset (100000 samples) of a random variable Z, Such that $Z = X + 10Y$, where X is a uniform random variable between -3 and 3. We also know that:

$$Y = \sum_{i=1}^{k} W_i \tag{1.1}$$

Where k could be 2,3 or 4 and W_i's are independent and identically distributed. Our goal is to find the distribution of W_i (it is among the following distributions)

- Exponential
- Rayleigh
- Half Normal

and other parameters like k and distribution specific parameters (λ for exponential and σ for others).

2 Basic Approach

Since the size of the dataset is huge, we could calculate average and variance of the data and somehow relate it with expectation formulae. Expectation could be thought of as an average of the random variable when the experiment is conducted infinite times.

The size of our dataset (10^5) is huge enough, using the weak law of large numbers (it states that the mean for very large number of outcomes will converge to the expected value), we could expect that the average of the dataset would give us a value very close to expectation of Z.

3 IDEA AND WORKING MECHANISM

The idea is to calculate average and variance of dataset and relate it to X and Y using following equations. From now on, let us denote mean of the dataset by M and variance by V. :

$$\mathbb{E}[Z] = \mathbb{E}[X + 10Y] \tag{3.1}$$

$$\mathbb{E}[Z] = \mathbb{E}[X] + 10\mathbb{E}[Y] \tag{3.2}$$

$$M \approx \mathbb{E}[Z] \tag{3.3}$$

Here, $\mathbb{E}[Y] is:$

$$\mathbb{E}[Y] = \mathbb{E}[\sum_{i=1}^{k} W_i] \tag{3.4}$$

$$\mathbb{E}[Y] = \sum_{i=1}^{k} \mathbb{E}[W_i] \tag{3.5}$$

Here, each W_i has same distribution parameters as they are identically distributed, therefore each of W_i has same mean and variance as these properties depend only upon distribution parameters. Therefor $\mathbb{E}[Y]$ could be further simplified to:

$$\mathbb{E}[Y] = k\mathbb{E}[W_1] \tag{3.6}$$

$$\tag{3.7}$$

Let us calculate expectation and variance of X since we know its distribution. For a uniformly distributed random variable X with upper bound b and lower bound a, The expectation and variance are given by:

$$\mathbb{E}[X] = (a + b)/2 \tag{3.8}$$

$$Var(X) = (b - a)^2/12 \tag{3.9}$$

Here, in our case b=3 and a=-3 . Therefore,

$$\mathbb{E}[X] = 0 \tag{3.10}$$

$$Var(X) = 3 \tag{3.11}$$

Using above required values, finally our M boils down to:

$$M \approx 10k\mathbb{E}[W_1] \tag{3.12}$$

$$\tag{3.13}$$

Similar Analysis could be done for variance:

$$var(Z) = var(X + 10Y) \tag{3.14}$$

$$var(Z) = var(X) + var(10Y) + 2cov(X, 10Y) \tag{3.15}$$

$$var(Z) = 3 + 100var(Y) + 0 \tag{3.16}$$

Covariance of X and Y is 0 since any value taken by Y doesnt influence values of X and vice versa. For $W_i^{\prime}s$ having i.i.d the covariance is 0 as well. Finally, we have:

$$var(Y) = var(\sum_{i=1}^{k} W_i) \tag{3.17}$$

$$var(Y) = \sum_{i=1}^{k} var(W_i) \tag{3.18}$$

$$V \approx 3 + 100k var(W_1) \tag{3.19}$$

Further, we notice that $\mathbb{E}[W_1]$ and $var(W_1)$ are functions of a single parameter which differs according to the distribution. Therefore, we can see that Equations 3.12 and 3.19 are 2 equations in 2 variables. If we choose a distribution, we will get a value of k corresponding to that distribution.

This value of k won't be exactly integral (as we are dealing with an approximation) but we could round off to closest integral value. If this integral value is among 2,3 and 4, we say the corresponding distribution is a valid candidate to be the actual distribution which was used to generate $W_i^{\prime}s$.

To confirm this, we generate 10 more datasets of the same size using this k and corresponding distribution parameter and see if the mean and variance is close to that of original dataset. If that is the case, we conclude we chose the right k and right distribution, if not we check for other valid candidate and repeat the process.

Finally, when the above tests are passed we generate another dataset with concluded k and distribution and plot the histograms of original dataset and generated dataset. We shall get almost same plot for both of them.

I have referred wikipedia pages to get the expectation and variance of each distribution to solve equations 3.12 and 3.19. Following are the results which were hardcoded in the python program.

- For Exponential Distribution:

$$PDF = \lambda e^{-\lambda x} \tag{3.20}$$

$$Expectation = \frac{1}{\lambda} \tag{3.21}$$

$$Variance = \frac{1}{\lambda^2} \tag{3.22}$$

$$k = \frac{M^2}{V - 3} \tag{3.23}$$

$$\lambda = \frac{10M}{V - 3} \tag{3.24}$$

- For Rayleigh Distribution:

$$PDF = \frac{x}{\sigma^2} e^{\frac{-x^2}{2\sigma^2}} \tag{3.25}$$

$$Expectation = \sigma\sqrt{\frac{\pi}{2}} \tag{3.26}$$

$$Variance = \frac{4-\pi}{2}\sigma^2 \tag{3.27}$$

$$k = \frac{M^2}{V-3}\frac{4-\pi}{4} \tag{3.28}$$

$$\sigma = \frac{(V-3)}{(4-\pi)}\frac{\sqrt{\pi}}{5\sqrt{2}M} \tag{3.29}$$

- For Half Normal Distribution:

$$PDF = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} e^{\frac{-x^2}{2\sigma^2}} \tag{3.30}$$

$$Expectation = \sigma\sqrt{\frac{2}{\pi}} \tag{3.31}$$

$$Variance = \sigma^2(1-\frac{2}{\pi}) \tag{3.32}$$

$$k = \frac{M^2}{V-3}\frac{\pi-2}{2} \tag{3.33}$$

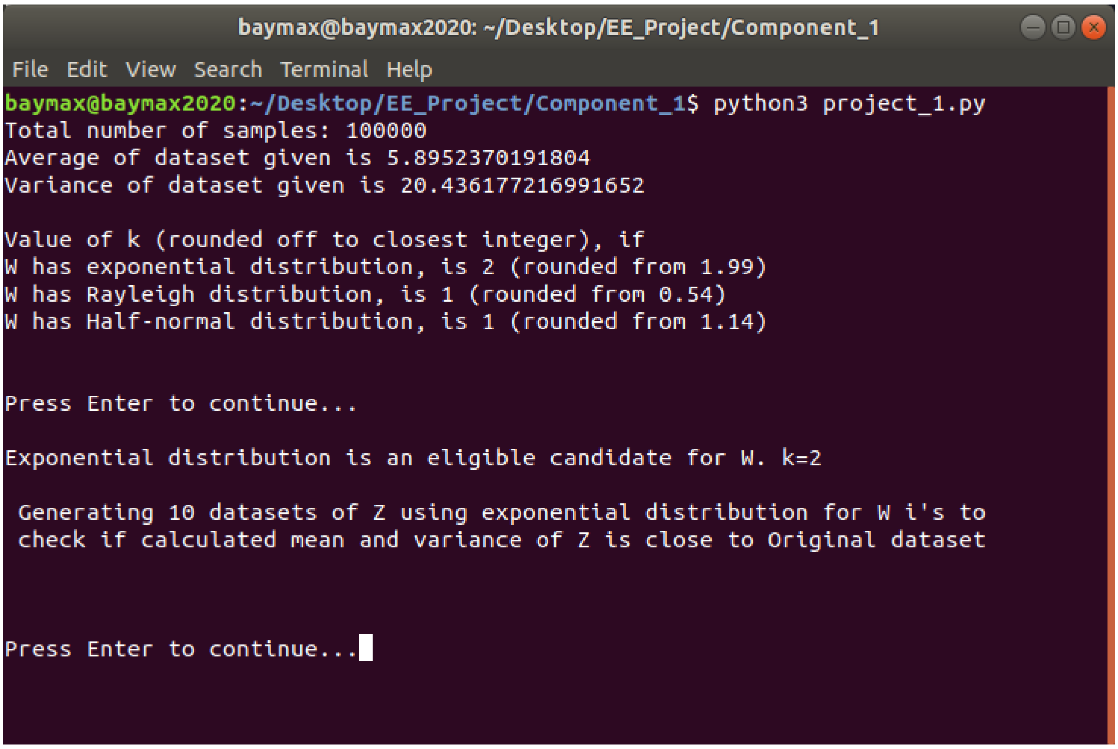
$$\sigma = \frac{(V-3)}{(4-\pi)}\frac{\sqrt{2\pi}}{10M} \tag{3.34}$$

- Resources:
  https://en.wikipedia.org/wiki/Exponential_distribution
  https://en.wikipedia.org/wiki/Rayleigh_distribution
  https://en.wikipedia.org/wiki/Half-normal_distribution

- General formulae/rules used:
  $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
  $\mathbb{E}[aX] = a\mathbb{E}[X]$
  $Var(aX+bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$

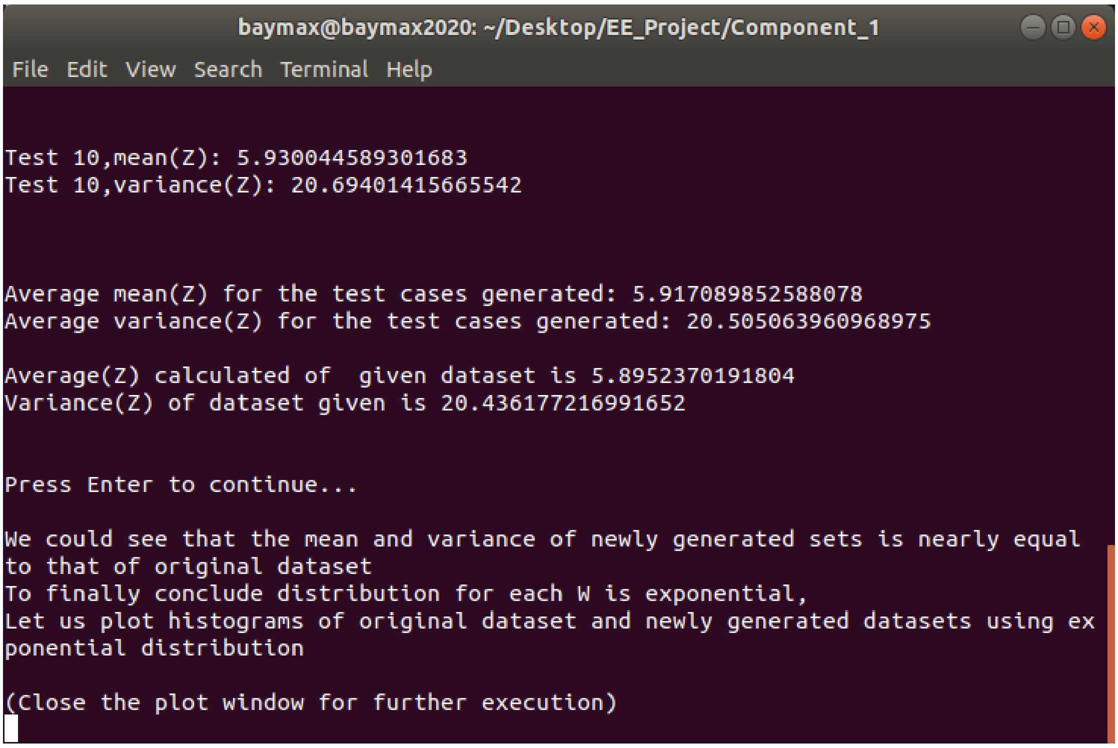
4 RESULTS

After implementing the above procedure in python with the help of libraries like numpy and mathplotlib, We get the following results: (screenshots attached from next page)



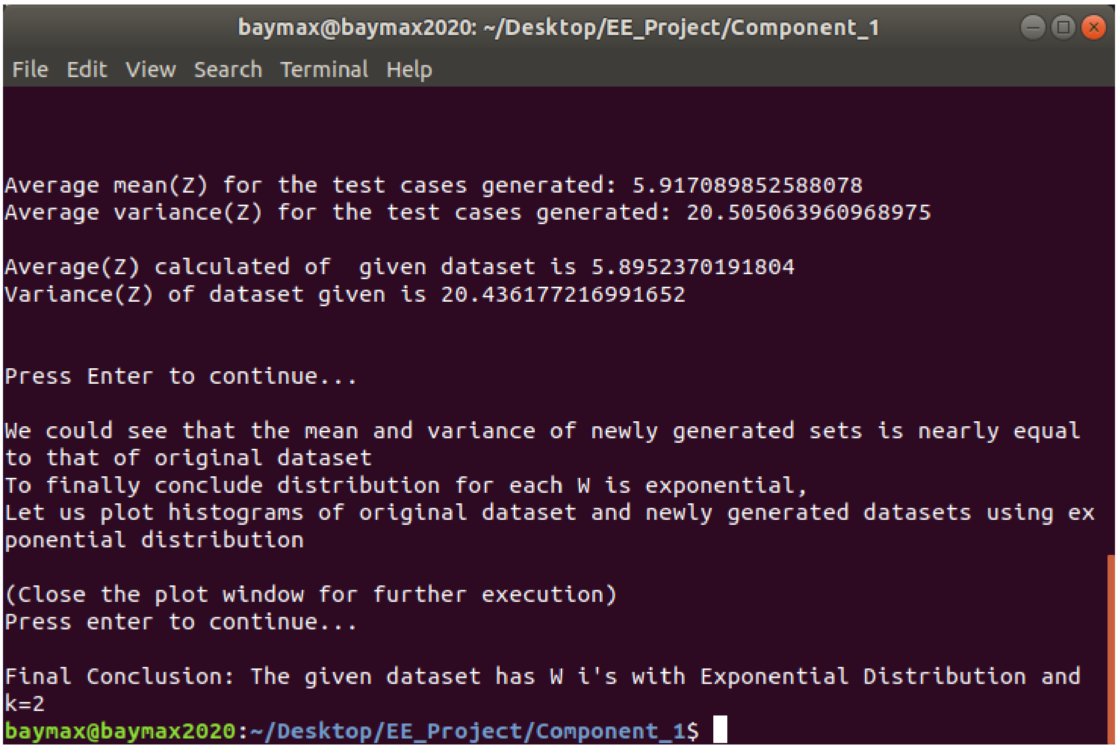
baymax@baymax2020: ~/Desktop/EE_Project/Component_1
File Edit View Search Terminal Help

```
baymax@baymax2020:~/Desktop/EE_Project/Component_1$ python3 project_1.py
Total number of samples: 100000
Average of dataset given is 5.8952370191804
Variance of dataset given is 20.436177216991652

Value of k (rounded off to closest integer), if
W has exponential distribution, is 2 (rounded from 1.99)
W has Rayleigh distribution, is 1 (rounded from 0.54)
W has Half-normal distribution, is 1 (rounded from 1.14)


Press Enter to continue...

Exponential distribution is an eligible candidate for W. k=2

 Generating 10 datasets of Z using exponential distribution for W i's to
 check if calculated mean and variance of Z is close to Original dataset



Press Enter to continue...
```

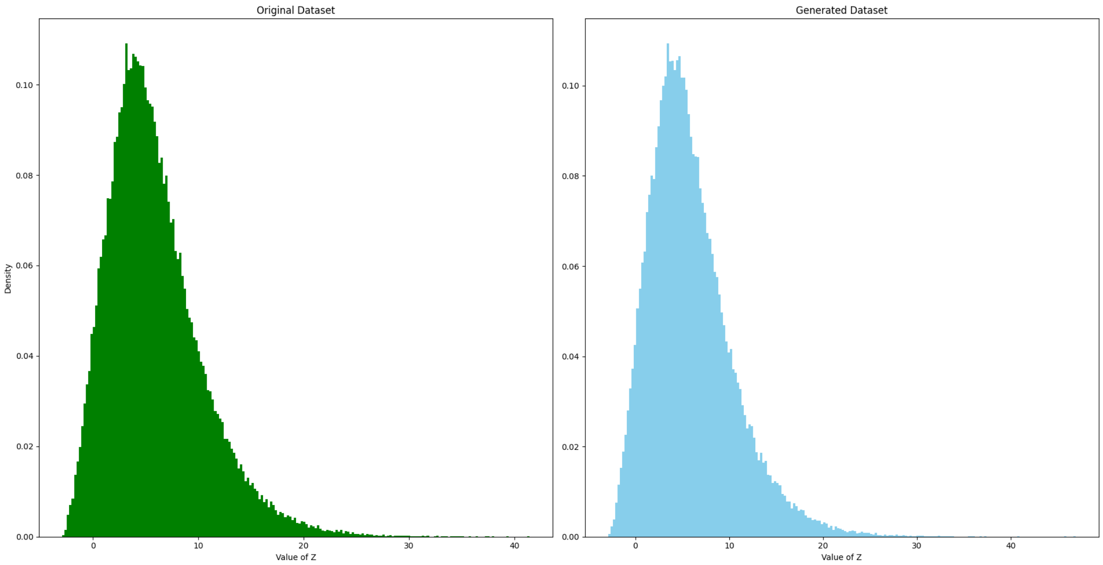


baymax@baymax2020: ~/Desktop/EE_Project/Component_1
File Edit View Search Terminal Help

```


Test 10,mean(Z): 5.930044589301683
Test 10,variance(Z): 20.69401415665542



Average mean(Z) for the test cases generated: 5.917089852588078
Average variance(Z) for the test cases generated: 20.505063960968975

Average(Z) calculated of  given dataset is 5.8952370191804
Variance(Z) of dataset given is 20.436177216991652


Press Enter to continue...

We could see that the mean and variance of newly generated sets is nearly equal
to that of original dataset
To finally conclude distribution for each W is exponential,
Let us plot histograms of original dataset and newly generated datasets using ex
ponential distribution

(Close the plot window for further execution)
```


baymax@baymax2020: ~/Desktop/EE_Project/Component_1
File Edit View Search Terminal Help

```
Average mean(Z) for the test cases generated: 5.917089852588078
Average variance(Z) for the test cases generated: 20.505063960968975

Average(Z) calculated of  given dataset is 5.8952370191804
Variance(Z) of dataset given is 20.436177216991652


Press Enter to continue...

We could see that the mean and variance of newly generated sets is nearly equal
to that of original dataset
To finally conclude distribution for each W is exponential,
Let us plot histograms of original dataset and newly generated datasets using ex
ponential distribution

(Close the plot window for further execution)
Press enter to continue...

Final Conclusion: The given dataset has W i's with Exponential Distribution and
k=2
baymax@baymax2020:~/Desktop/EE_Project/Component_1$
```



Here, We could see that the histogram of original dataset and newly generated dataset is almost the same.

Final conclusion: The dataset of data_190010029.csv was generated using Exponential distribution for each W_i using k=2 and the distribution parameter $\lambda \approx 3.381$