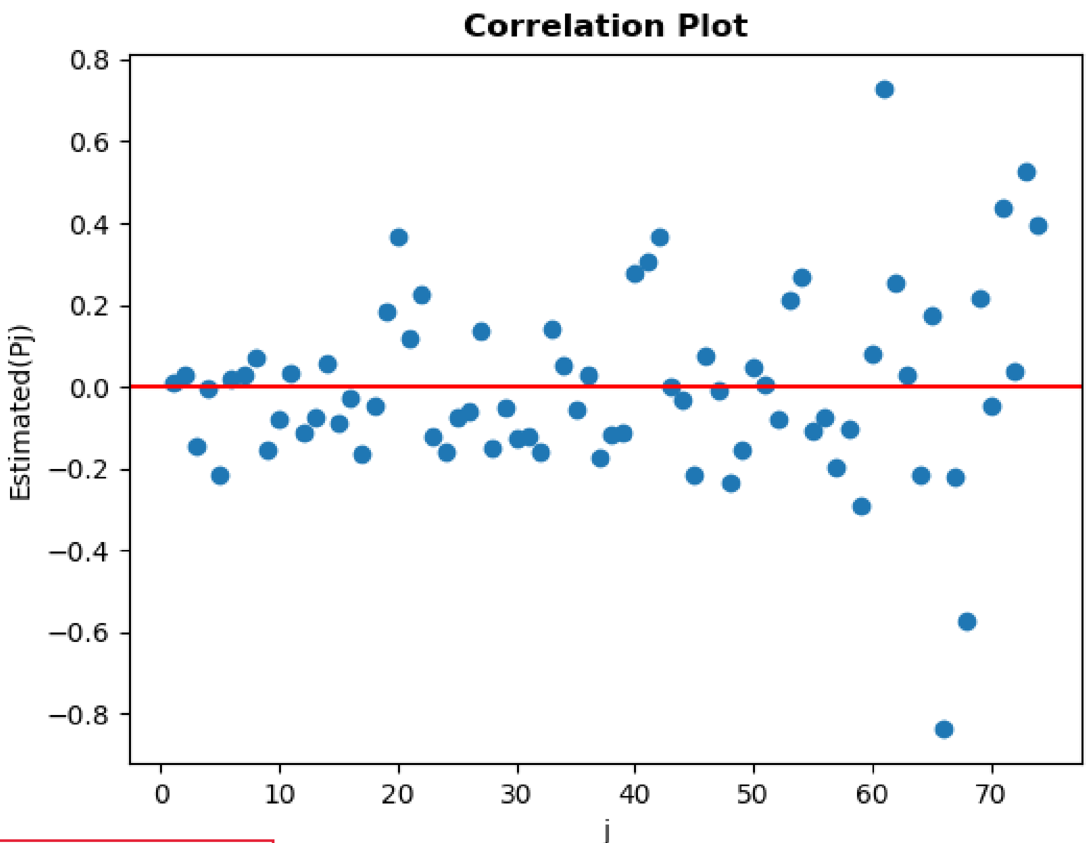
CSE 411 ASSIGNMENT 3

Submitted by:
Md. Tawkat Islam ~~Shawon~~ Khondokar
Student Id: 1405036
Md. Hasin Abrar
Student Id: 1405048

We are using the data of student id: 1405036

Sample Independence Assessment:

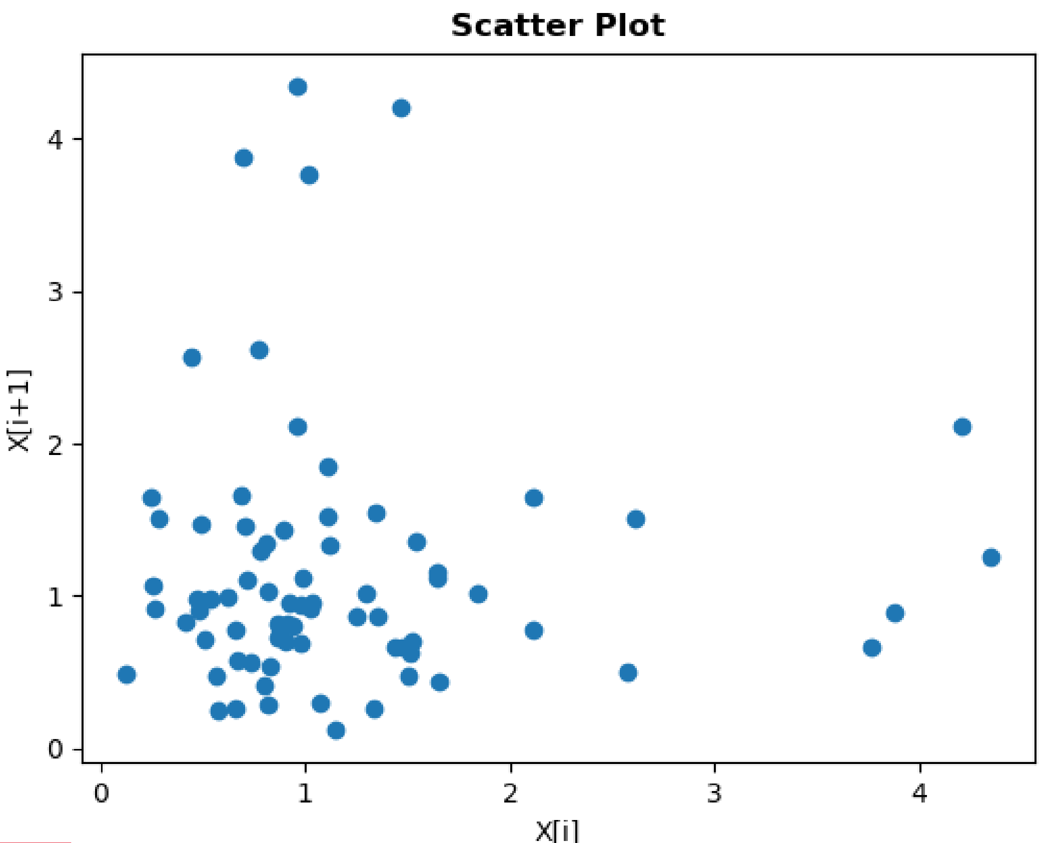
- <u>Correlation Plot:</u>



1

Most of the plotted points are near the horizontal line regardless of some outliers that is situated far from the horizontal line. So, we can say observations are independent.

- Scatter Diagram:



Here we can see this graph is sparse and does not show any positive or negative slop. So, we can say observations are independent.

Activity I: Hypothesizing Families of Distributions:

- Summary Statistic:

Minimum: 0.126

Maximum: 4.345

Mean: 1.1599466666666667

Median: 0.943

Variance: 0.7328737809009009

Coefficient of Variance: 0.7380343427128377

Lexis Ratio: 0.6318167912038205

Skewness: 2.1265552574071966

Result:

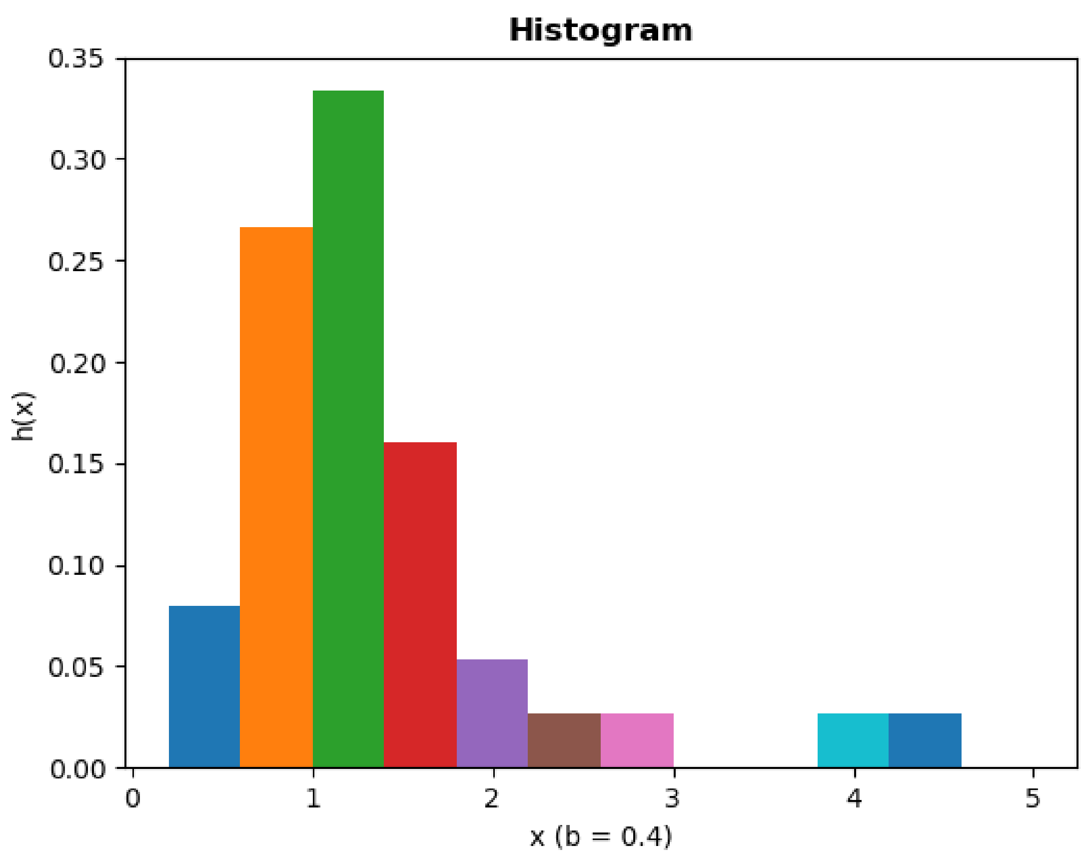
There is no specific domain for generated data. So, sample dataset is continuous.

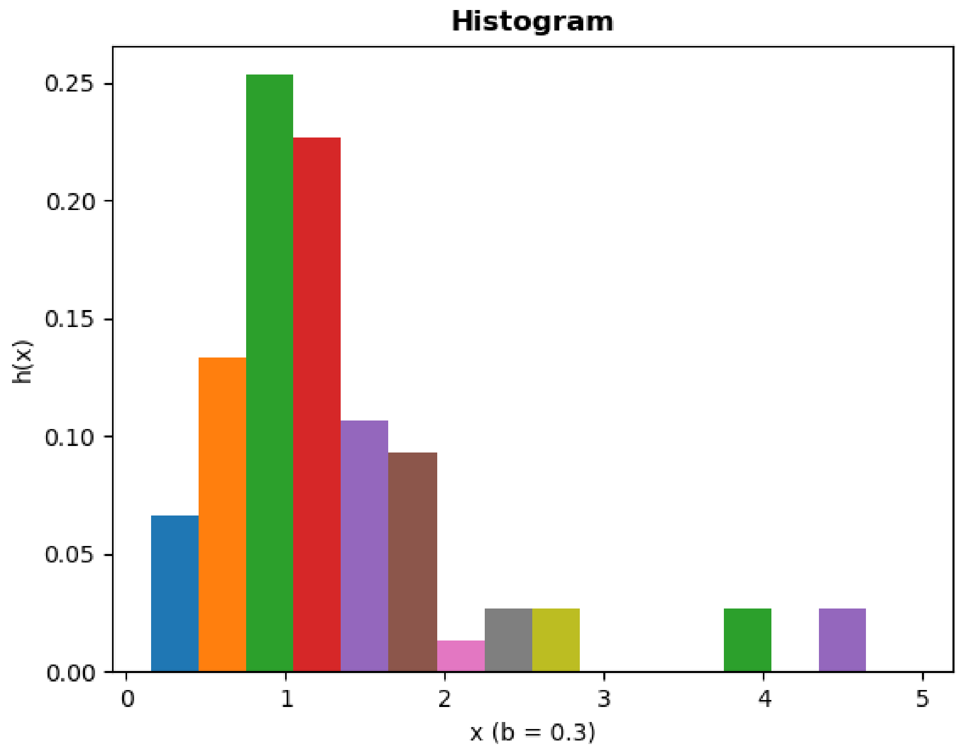
Mean > Median: Not Symmetric.

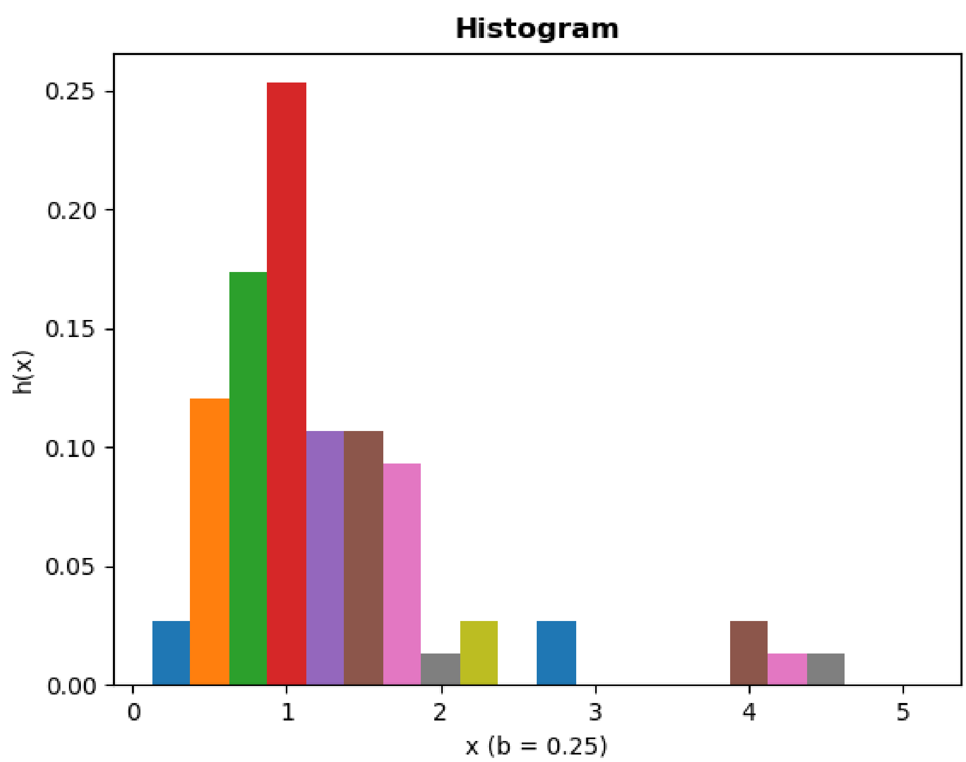
Skewness > 0: Skewed to Right.

Coefficient of variation < 0: Weibull or Gamma.

- Histogram:




Histogram
h(x)
0.25
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
5
x (b = 0.3)


Histogram
h(x)
0.25
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
5
x (b = 0.25)

4

We can see relatively smooth-looking shape occurs at Del(b)= 0.3, with k=17 number of intervals.

Moreover, the shape of the Histogram resembles with that of a Weibull density.

- <u>Quantile Summaries and Box Plot:</u>

5

Median: 0.943

Quantile[0]: 0.664

Quantile[1]: 1.436

Quantile Midpoint: 1.05

Octile[0]: 0.477

Octile[1]: 1.657

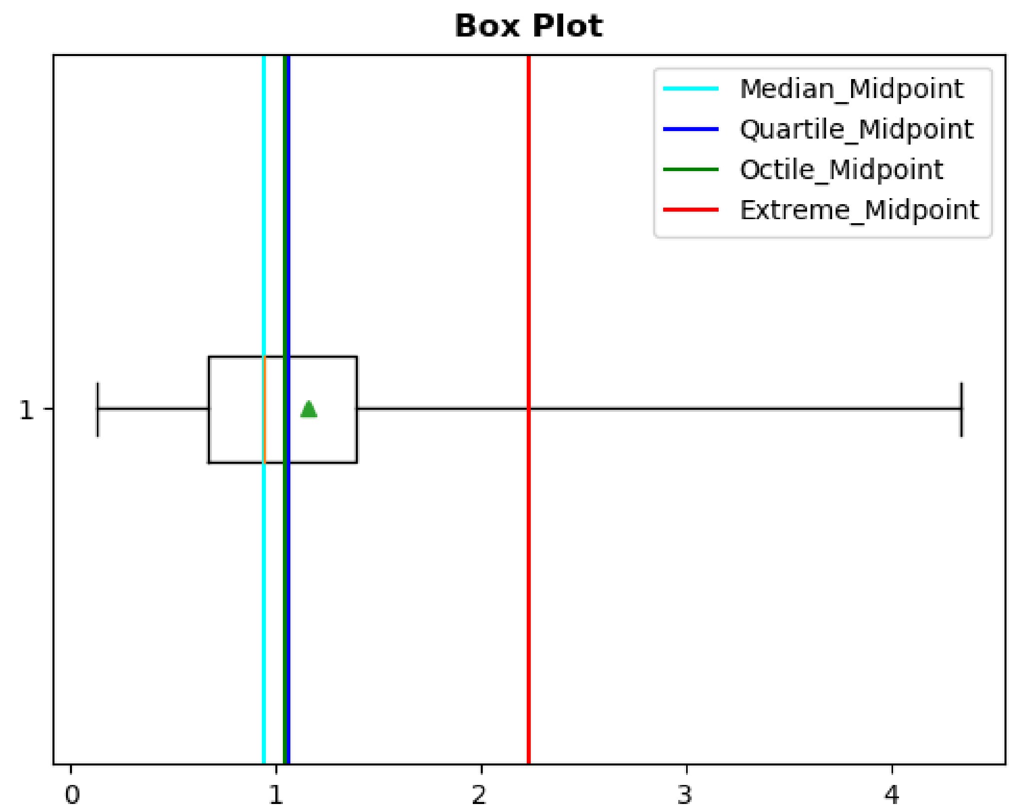
Octile Midpoint: 1.067

Extreme[0]: 0.126

Extreme[1]: 4.345

Extreme Midpoint: 2.2355

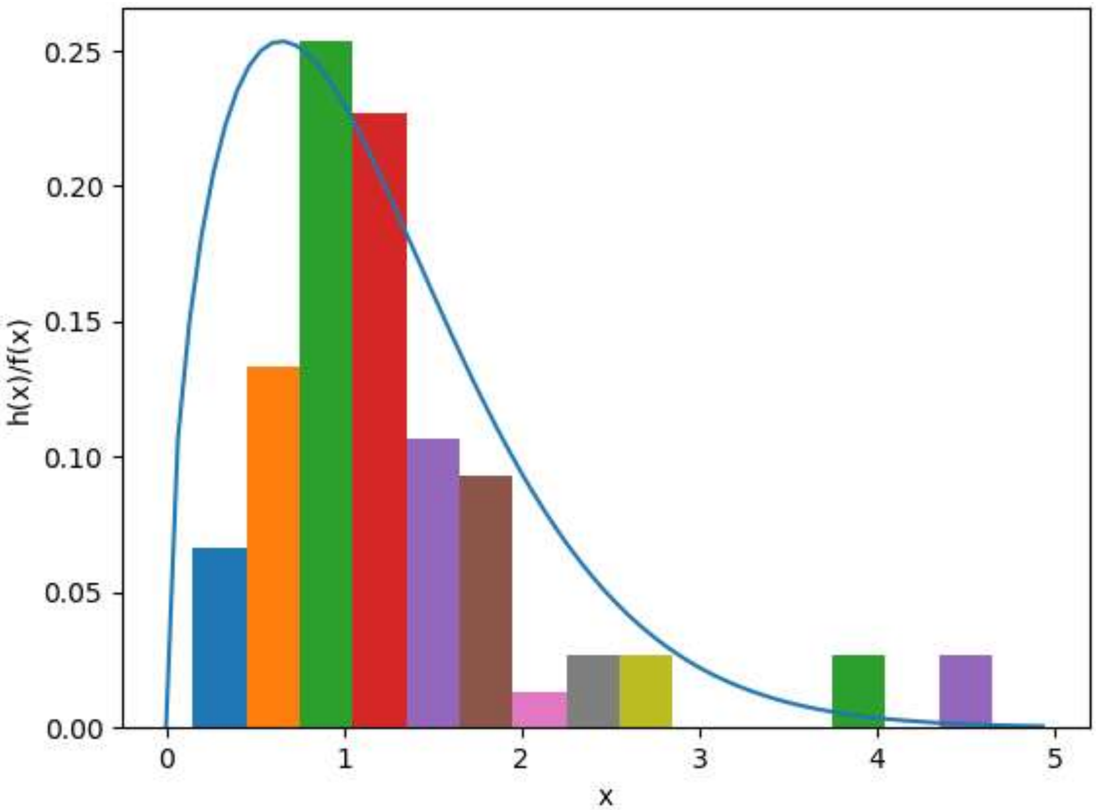
We can see midpoints are gradually increasing. So, underlying Distribution is Right Skewed.



The elongated nature of the right side of the box plot reaffirms our hypothesis that it is right skewed.

Activity II: Estimation of Parameters:

- <u>Estimation of Parameter:</u>



As our histogram resembles mostly with the density function of Weibull distribution, we will estimate parameters of Weibull(alpjha, beta), using Maximum-Likelihood Estimators(MLEs).

MLE for Weibull distribution:

Density function for Weibull distribution;

$$f(x) = \begin{cases} \alpha\beta^{-\alpha} x^{\alpha-1} e^{-(x/\beta)^{\alpha}} & ; \text{if } x>0 \\ 0 & ; \text{otherwise} \end{cases}$$

Here, α = shape parameter $(\alpha>0)$

β = scale parameter $(\beta>0)$

So, the likelihood function is,

$$L(\alpha,\beta) = \prod_{i=1}^{n} \alpha\beta^{-\alpha} x_i^{\alpha-1} e^{-(x_i/\beta)^{\alpha}}$$

$$= \prod_{i=1}^{n} \frac{\alpha}{\beta}\left(\frac{x_i}{\beta}\right)^{\alpha-1} e^{-(x_i/\beta)^{\alpha}}$$

Maximizing $L(\alpha,\beta)$ is equivalent to maximizing $LL(\alpha,\beta) = \ln L(\alpha,\beta)$, which is the log-likelihood function.

$$\therefore LL(\alpha,\beta) = \sum_{i=1}^{n} \ln\left[\frac{\alpha}{\beta}\left(\frac{x_i}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x_i}{\beta}\right)^{\alpha}}\right]$$

$$= \sum_{i=1}^{n} \left[\ln\alpha - \ln\beta + (\alpha-1)\ln x_i - (\alpha-1)\ln\beta - \left(\frac{x_i}{\beta}\right)^{\alpha} \right]$$

$$= n\left[\ln\alpha - \alpha\ln\beta\right] + (\alpha-1)\sum_{i=1}^{n}\ln x_i - \sum_{i=1}^{n}\left(\frac{x_i}{\beta}\right)^{\alpha}$$

As, the logarithm function is strictly increasing, maximizing $L(\alpha,\beta)$ is equivalent to maximizing $LL(\alpha,\beta)$ $\Rightarrow$ (α,β) maximizes $L(\alpha,\beta)$ if and only if (α,β)

maximizes $LL(\alpha, \beta)$. $LL(\alpha, \beta)$ will be maximized when

$$\frac{dLL(\alpha,\beta)}{d(\alpha,\beta)} = 0$$

Solving this, we get:

$$\frac{\sum_{i=1}^{n} x_i^{\hat{\alpha}} \ln x_i}{\sum_{i=1}^{n} x_i^{\hat{\alpha}}} - \frac{1}{\hat{\alpha}} = \frac{\sum_{i=1}^{n} \ln x_i}{n}$$

$$\hat{\beta} = \left(\frac{\sum_{i=1}^{n} x_i^{\hat{\alpha}}}{n} \right)^{1/\hat{\alpha}}$$

Using recursion for Newton iteration:

$$\hat{\alpha}_{k+1} = \hat{\alpha}_k + \frac{A + 1/\hat{\alpha}_k - C_k/B_k}{1/\hat{\alpha}_k + (B_k H_k - C_k^2)/B_k^2}$$

Where, $A = \frac{\sum_{i=1}^{n} \ln x_i}{n}$

$$B_k = \sum_{i=1}^{n} x_i^{\hat{\alpha}_k}$$

$$C_k = \sum_{i=1}^{n} x_i^{\hat{\alpha}_k} \ln x_i$$

$$H_k = \sum_{i=1}^{n} x_i^{\hat{\alpha}_k} (\ln x_i)^2$$

And initial value for $\hat{\alpha}$,

$$\hat{\alpha}_0 = \left\{ \frac{(6/\pi^2)\left[\sum_{i=1}^{n}(\ln x_i) - \left(\sum_{i=1}^{n} \ln x_i\right)^2 / n\right]}{n-1} \right\}^{-1/2}$$

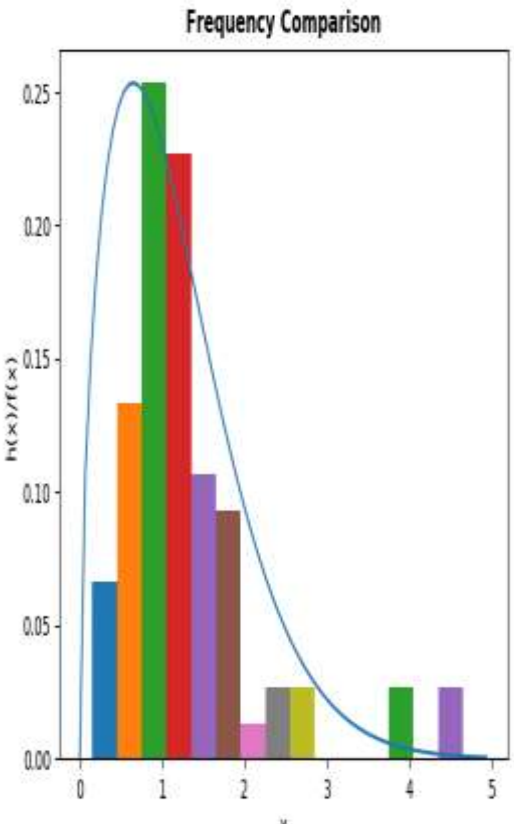
Then we can solve these equations using Excel solver.

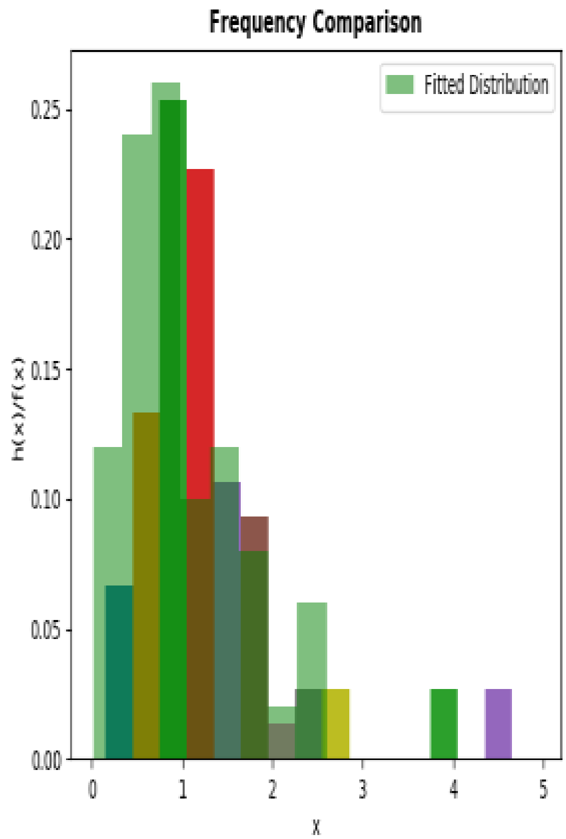
6

From this we estimate, Alpha= 1.5278996202501096, Beta= 1.2999425079066453

Activity III: Determining How Representative the Fitted Distributions are:

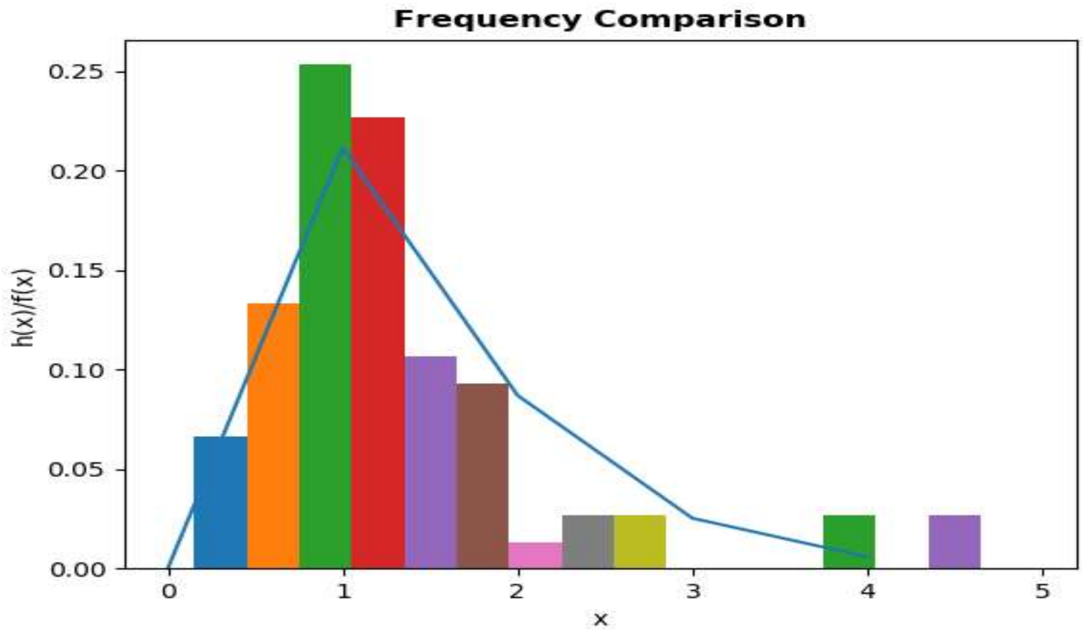
- Frequency Comparisons:





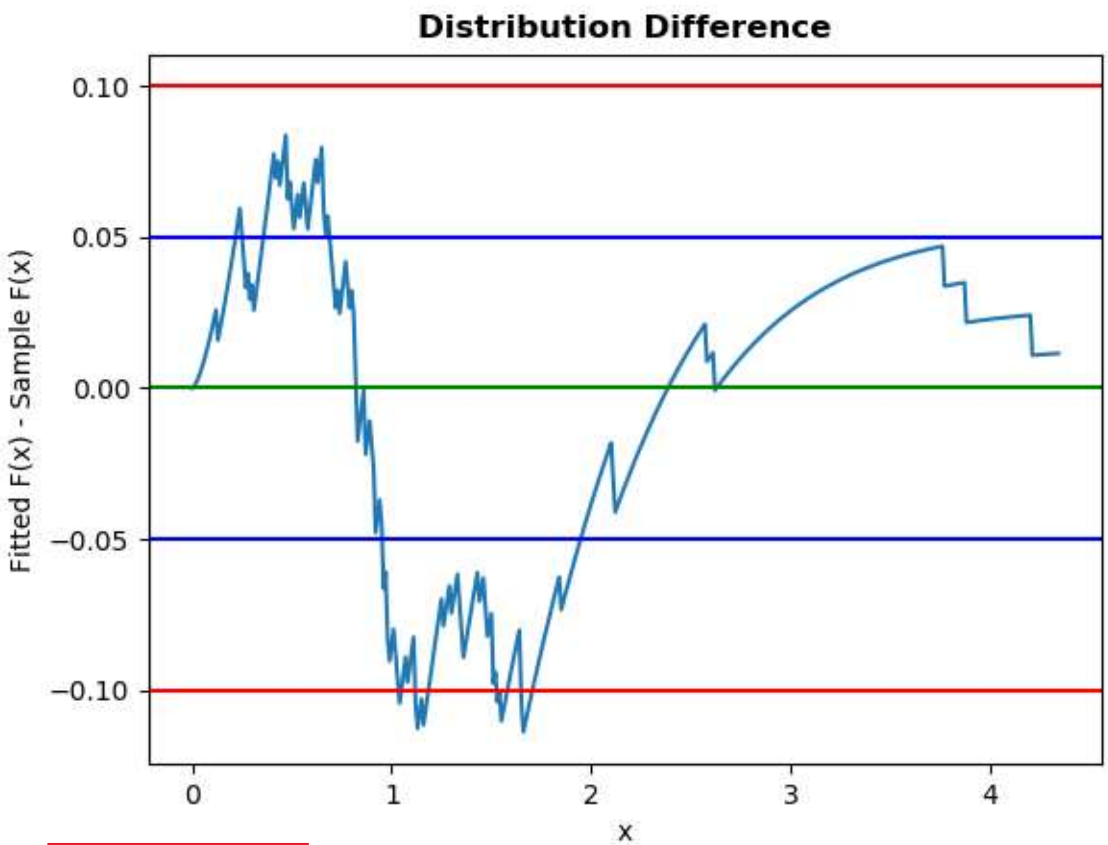
7

For Frequency Comparison, the fitted distribution is not a good representation of the true underlying distribution of the data because sample size(=75) is not sufficiently large, on the other hand fitted distribution is generated for sufficiently huge data.



But if we take the fitted distribution for small size of data we can see it almost follows the underlying distribution. So we can say, Frequency Comparisons almost represents the underlying distribution.

- <u>Distribution –Function –Difference:</u>

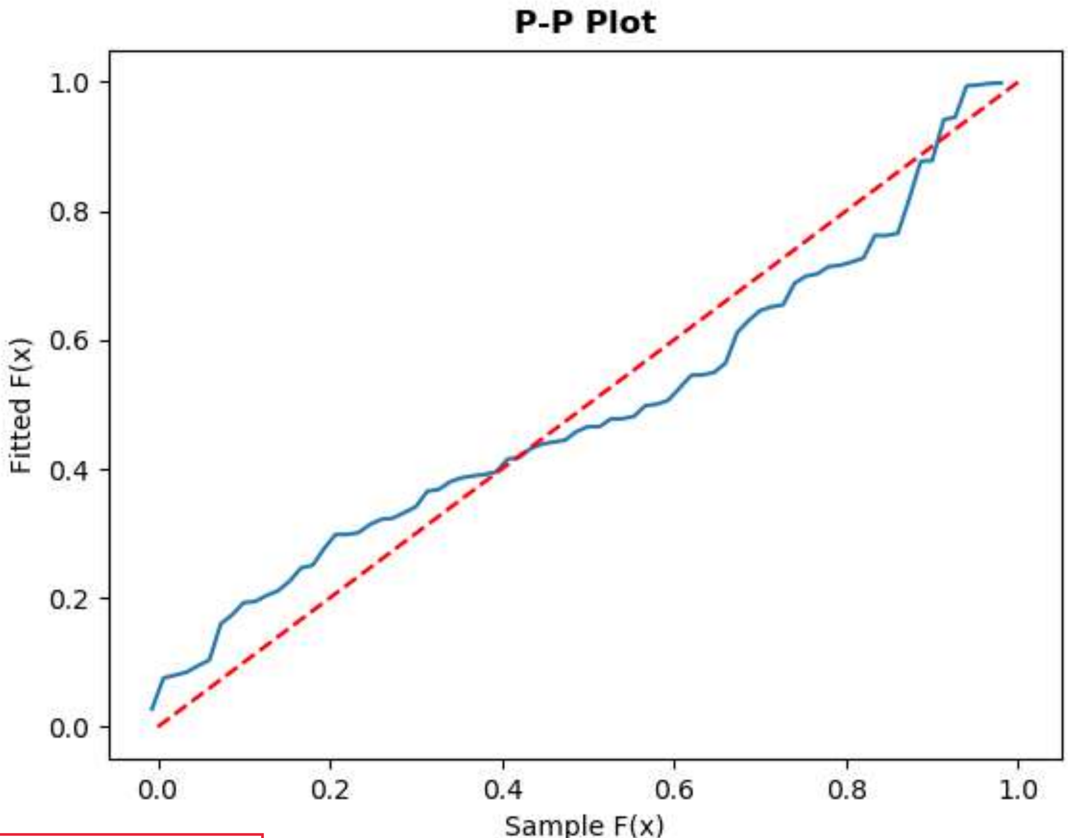


8

We can see the difference plot crosses blue and red lines in several cases. It does remain stable near 0.0. So we can say it is not good fitted in this fitted model due to small sample size(=75). However if sample size were large enough we would expect a good fit.

?????????????????????????

- P-P Plot:

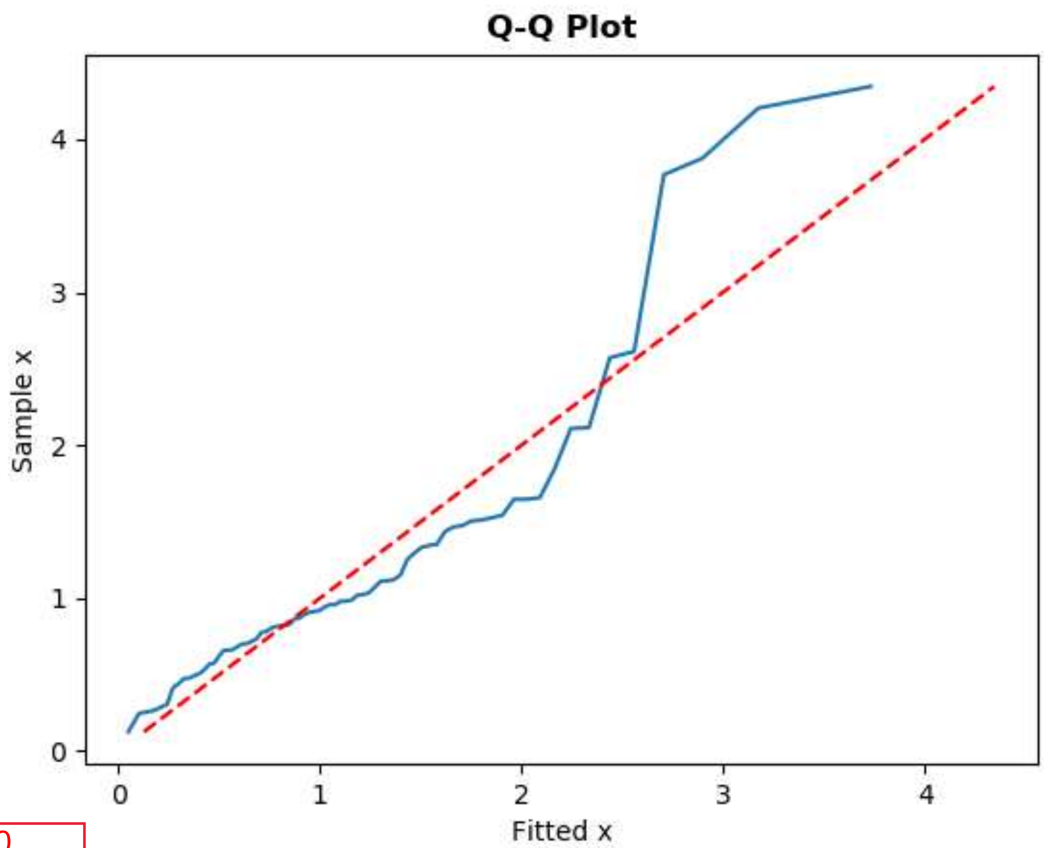


9

We know P-P plot amplifies difference between the middle of the fitted $\hat{F}(x)$ and sample $F_n(x)$ but fails to amplify difference between the right tails.

Here, Sample distribution almost follows Fitted distribution with some deviation in the middle due to small dataset(=75). So, we can say this fitted model is a good representation of the sample distribution.

- <u>Q-Q Plot:</u>



10

Q-Q plot amplifies difference between the right tails of the fitted distribution $\hat{F}$(x) and our sample distribution F_n(x).

Here, Sample distribution almost follows Fitted distribution with some deviation in the right tails due to small dataset(=75). So, we can say this fitted model is a good representation of the sample distribution.

- Chi-Square Test:

11

We choose No. of Intervals k=15.

So n*Pj= 75*(1/15)

= 5

So the conditions of equiprobable intervals k>=3 and n*Pj>=5 for all j is satisfied.

Simulation Result:

j: 1 Interval= [0.000000, 0.225907) nPj= 5.000000 Nj= 1 ((Nj-nPj)^2)/nPj= 3.200000

j: 2 Interval= [0.225907, 0.364165) nPj= 5.000000 Nj= 5 ((Nj-nPj)^2)/nPj= 0.000000

j: 3 Interval= [0.364165, 0.487054) nPj= 5.000000 Nj= 4 ((Nj-nPj)^2)/nPj= 0.200000

j: 4 Interval= [0.487054, 0.604179) nPj= 5.000000 Nj= 5 ((Nj-nPj)^2)/nPj= 0.000000

j: 5 Interval= [0.604179, 0.720000) nPj= 5.000000 Nj= 8 ((Nj-nPj)^2)/nPj= 1.800000

j: 6 Interval= [0.720000, 0.837511) nPj= 5.000000 Nj= 8 ((Nj-nPj)^2)/nPj= 1.800000

j: 7 Interval= [0.837511, 0.959324) nPj= 5.000000 Nj= 9 ((Nj-nPj)^2)/nPj= 3.200000

j: 8 Interval= [0.959324, 1.088220) nPj= 5.000000 Nj= 7 ((Nj-nPj)^2)/nPj= 0.800000

j: 9 Interval= [1.088220, 1.227652) nPj= 5.000000 Nj= 4 ((Nj-nPj)^2)/nPj= 0.200000

j: 10 Interval= [1.227652, 1.382473) nPj= 5.000000 Nj= 5 ((Nj-nPj)^2)/nPj= 0.000000 j: 11 Interval= [1.382473, 1.560331) nPj= 5.000000 Nj= 7 ((Nj-nPj)^2)/nPj= 0.800000

j: 12 Interval= [1.560331, 1.774970) nPj= 5.000000 Nj= 3 ((Nj-nPj)^2)/nPj= 0.800000

j: 13 Interval= [1.774970, 2.056158) nPj= 5.000000 Nj= 1 ((Nj-nPj)^2)/nPj= 3.200000

j: 14 Interval= [2.056158, 2.495141) nPj= 5.000000 Nj= 2 ((Nj-nPj)^2)/nPj= 1.800000

j: 15 Interval= [2.495141, 4.445000) nPj= 5.000000 Nj= 6 ((Nj-nPj)^2)/nPj= 0.200000

No. of Intervals k: 15

X2: 18.0

X2(15-1,1-0.05): 23.685

X2(15-1,1-0.10): 21.064

Cannot Reject the Hypothesis at alpha=0.05

Cannot Reject the Hypothesis at alpha=0.10

So, Chi-Square Test gives us no reason to conclude that our sample data is poorly fitted by Weibull(alpha=1.5278996202501096,beta=1.2999425079066453)distribution.