CE605A:Probability & Statistics for Civil Engineers

PROJECT-2

NILIMA PAUL

Roll no: 20103072

M.tech, Geoinformatics

Department of Civil Engineering

Indian Institute of Technology, Kanpur

CE605A: PROJECT-2

EXECUTIVE SUMMARY:

In this project, as a civil engineer we will analyze discharge data (X) collected independently at four sites.
For each site we will find:

(i) The magnitude of discharge xk such that $P(X \geq xk) = 0.01$ along with 90% confidence interval for xk.
(ii) Plot the histogram for each dataset to observe the distribution.
(iii) Test the hypothesis that the mean discharge is equal to 2500 units.
(iv) Decide which curve id best fit for a given distribution, we will perform Goodness of fit test.

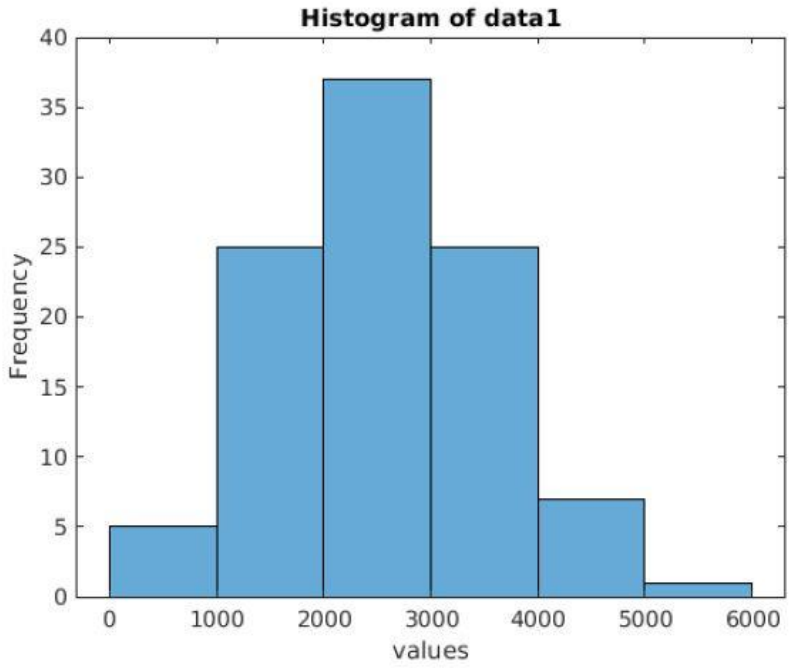
METHODOLOGY:

(i) Import the data into Excel file
(ii) Read the data (into MATLAB)
(iii) Plot histogram of each dataset and observe the distribution
(iv) Calculate Mean and Standard Deviation of each dataset
(v) Perform Hypothesis Testing to decide if the given mean discharge is equal to 2500 units or not.
(vi) Perform Goodness of fit test to decide which distribution is best fit for the given datasets.

CALCULATIONS & RESULTS:

DATA-1

From the given dataset, histogram plotted for Data-1 is given below:



From data-1 we calculate mean and standard deviation as follows:

Mean (μ) = 2602.41

Standard deviation (σ) = 1007.079

$$P(X \geq xk) = 0.01$$

$$1 - P(X \leq xk) = 0.01$$

$$P(X \leq xk) = 0.99$$

$$\varphi(xk) = 0.99$$

$$\frac{Xk-\mu}{\sigma} = 2.33$$

$$Xk = 2.33 * \sigma + \mu$$

$$Xk = 2.33 * 1007.079 + 2602.41$$

$$\boldsymbol{Xk = 4948.904}$$

Goodness of fit test:

(i) Null hypothesis, H0: Data follow the normal distribution.

(ii) Alternative hypothesis, Ha: Data does not follow the normal distribution.

(iii) Level of significance (α) = 5%

(iv) We are performing Chi-square distribution.

(v) Degree of freedom (DOF) = K-1-m = 6-1-2 = 3, where k = no. of intervals, m = no. of parameters.

(vi) **Test statistics**: n = 100, Oi = Frequency in each interval, Zi = Normal variate interval, Pi = Probability, ei = expected no. of observations.

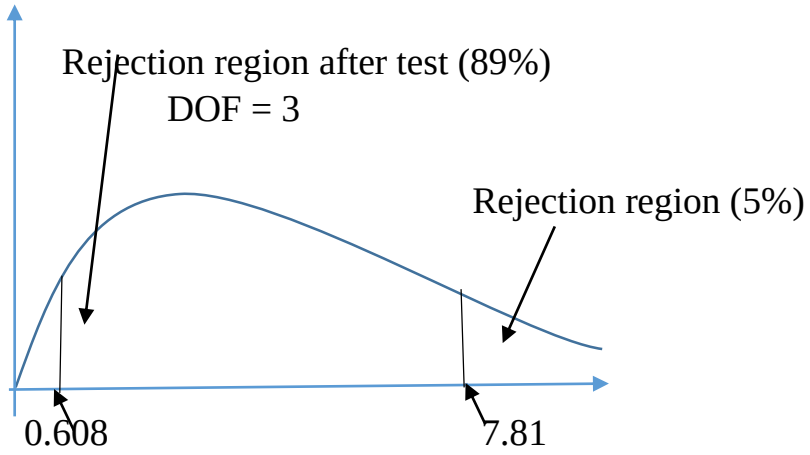
Sl. no.	Interval	Oi	Zi	Pi	ei = nPi	$\frac{Oi^2}{ei}$
1	<1000	5	<-1.5911	0.056	5.6	4.4643
2	1000-2000	25	-1.5911 -0.5982	0.22	22.0	28.4091
3	2000-3000	37	-0.5982 0.3948	0.3765	37.65	36.3612
4	3000-4000	25	0.3948 1.3878	0.2645	26.45	23.6295
5	4000-5000	7	1.3878 2.3807	0.0743	7.43	6.5949
6	>5000	1	>2.3807	0.0087	0.87	1.1494
		$\sum Oi = 100$				$\sum\left(\frac{Oi^2}{ei}\right) = 100.608$

Test statistics,

$$\sum\left(\frac{Oi^2}{ei}\right) - n = 100.608 - 100 = 0.608$$

Hence, H0 is not rejected. So the distribution is a **Normal distribution**.

(vii) P-value = 89%



Hypothesis testing:

Given, Mean discharge = 2500 units

(i) Null hypothesis, H0 : $\mu = \mu 0 = 2500$

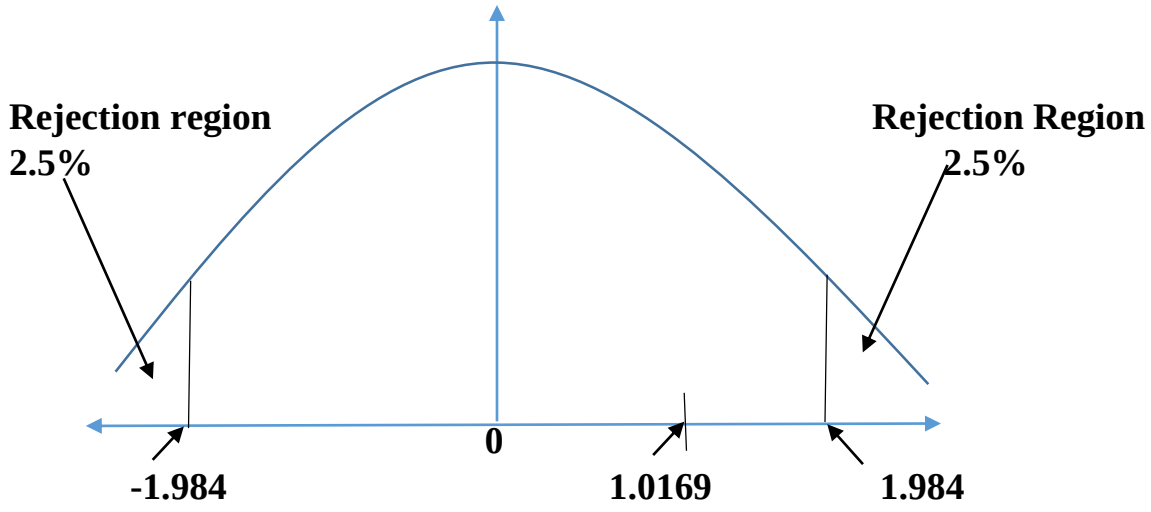
(ii) Alternative hypothesis, Ha: $\mu \neq \mu 0$

(iii) Level of significance (α) = 5%

(iv) At level of significance =0.05 corresponding LCL and UCL = 1.984

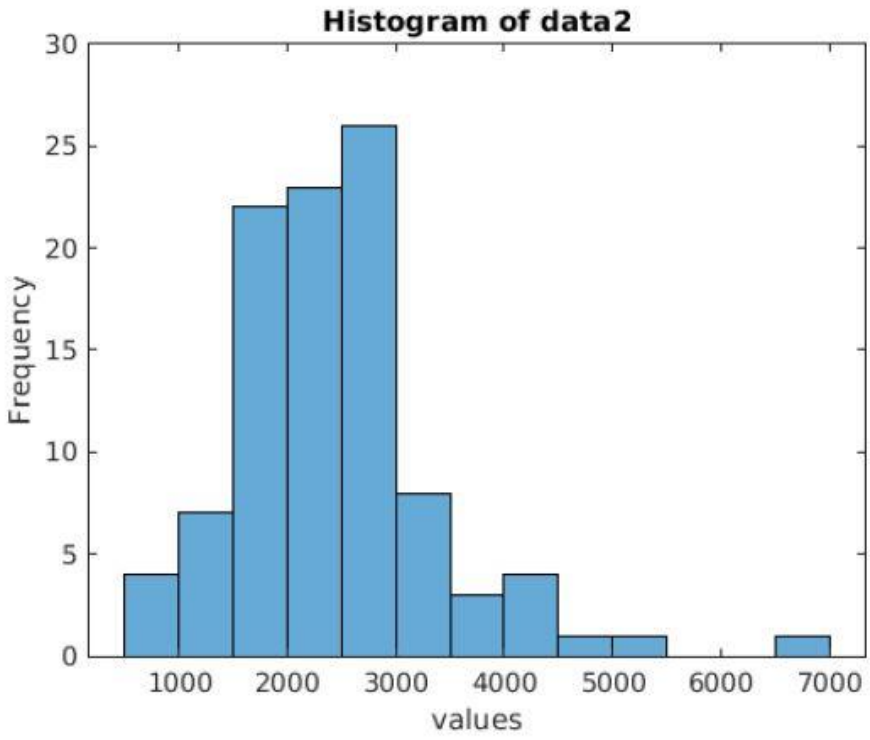
(v) Test statistics: $T = \frac{\mu - \mu 0}{\frac{\sigma}{\sqrt{n}}} = \frac{2602.41 - 2500}{\frac{1007.079}{\sqrt{100}}} = 1.0169$

(vi) As 1.0169 < 1.984, hence, **H0 cannot be rejected.**



DATA-2

From the given dataset, histogram plotted for Data-2 is given below:



From data-2 we calculate mean and standard deviation as follows:

Mean (μ) = 2451.497

Standard deviation (σ) = 945.11

From observing the histogram, we assume it as log normal distribution.

Now,

$$P(X \leq xk) = P[\ln(X) \leq \ln(xk)]$$

$$P(X \leq xk) = P[\mathrm{Y} \leq \mathrm{yk}]$$

$$P(X \leq xk) = P\left[\frac{\mathrm{Y} - \mu\mathrm{y}}{\sigma\mathrm{y}} \leq \frac{\mathrm{yk} - \mu\mathrm{y}}{\sigma\mathrm{y}}\right]$$

$$\text{Let, } P(X \leq xk) = P[\mathrm{Z} \leq \mathrm{zk}]$$

Given,

$$P[\mathrm{Z} \geq \mathrm{zk}] = 0.01$$

$$1 - P[Z \leq zk] = 0.01$$

$$P[Z \leq zk] = 0.99$$

Taking log values of the observations,

Yi = log (xi)

Calculating mean and standard deviation of the log values of observations,

$$\mu y = 3.3589$$

$$\sigma y = 0.1656$$

From table we get, zk = 2.33

Now substituting these values, calculate xk,

$$xk = e^{(\sigma y * zk + \mu y)}$$

$$xk = e^{(0.1656 * 2.33 + 3.3589)}$$

$$xk = 42.298$$

Goodness of fit test:

(i) Null hypothesis, H0: Data follow the log normal distribution.
(ii) Alternative hypothesis, Ha: Data does not follow the log normal distribution.
(iii) Level of significance (α) = 5%
(iv) We are performing Chi-square distribution.
(v) Degree of freedom (DOF) = K-1-m = 11-1-2 = 8, where k = no. of intervals, m = no. of parameters.
(vi) **Test statistics**:
n = 100, Oi = Frequency in each interval, Zi = Normal variate interval, Pi = Probability, ei = expected no. of observations.

Sl. no.	**Interval**	**Oi**	**Yi=log(xi)**	$\boldsymbol{Zi = \frac{Yi - \mu y}{\sigma y}}$	**Pi**	**ei = nPi**	$\boldsymbol{\frac{Oi^2}{ei}}$
1	<1000	4	<6.907	-2.1688	0.01505	1.5047	10.63
2	1000-1500	7	6.907 7.3132	-2.1688 -1.1038	11.978	11.978	4.0908
3	1500-2000	22	7.3132 7.6009	-1.1038 -0.3495	0.2285	22.85	21.1816
4	2000-2500	23	7.6009 7.8240	-0.3495 0.2354	0.2297	22.97	23.03

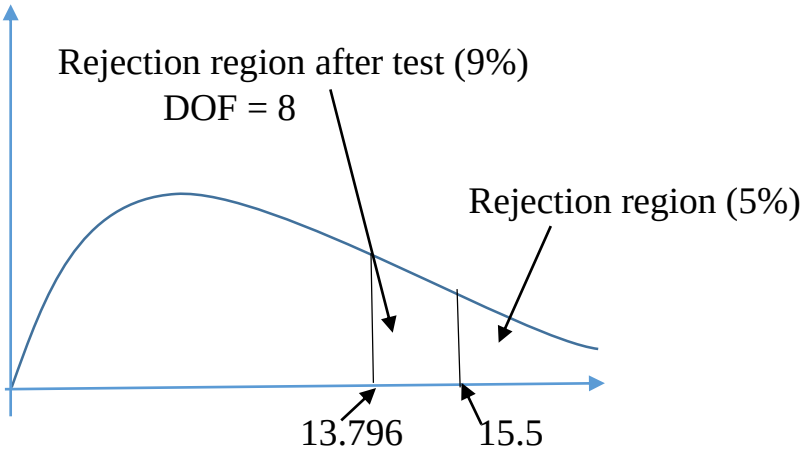
5	2500-3000	26	7.8240 8.0064	0.2354 0.7135	0.16918	16.9187	39.955
6	3000-3500	8	8.0064 8.1605	0.7135 1.1177	0.10589	10.589	6.044
7	3500-4000	3	8.1605 8.2940	1.1177 1.9677	0.06076	6.076	1.4812
8	4000-4500	4	8.2940 8.4118	1.9677 1.7766	0.03327	3.327	4.809
9	4500-5000	1	8.4118 8.5179	1.7766 2.0548	0.01786	1.786	0.5599
10	5000-5500	1	8.5179 8.6125	2.0548 2.3028	0.0093	0.93	1.0752
11	>5000	1	>8.6125	>2.3028	0.01064	1.064	0.9398
		$\sum Oi = 100$					$\sum\left(\frac{Oi^2}{ei}\right) = 113.7965$

Test statistics,

$$\sum\left(\frac{Oi^2}{ei}\right) - n = 113.7965 - 100 = 13.7965$$

Hence, H0 is not rejected. So the distribution is a **Log normal distribution**.

(i) P-value = 9%



Hypothesis testing:

Given, Mean discharge = 2500 units

(i) Null hypothesis, H0 : $\mu = \mu 0 = 2500$

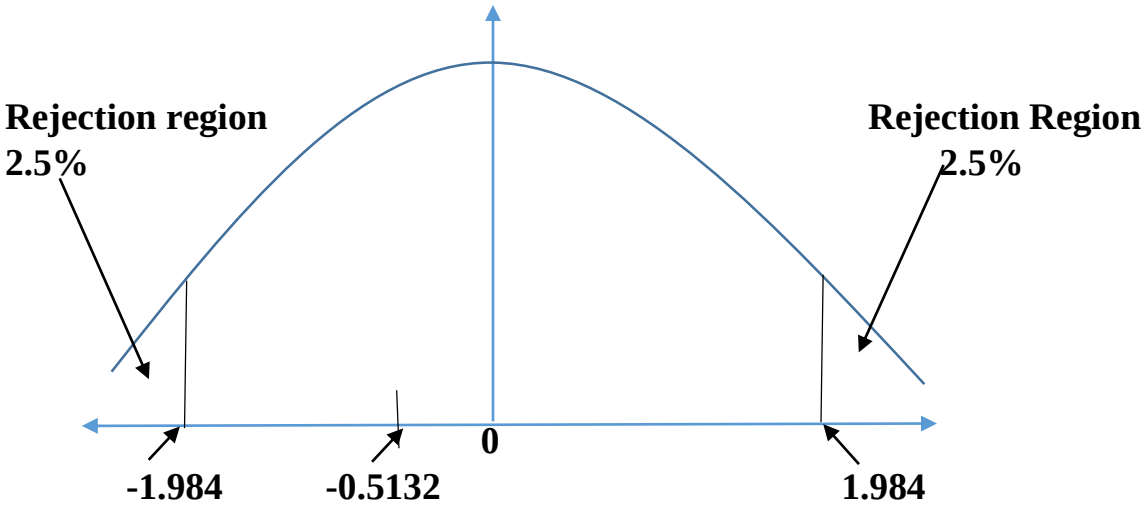
(ii) Alternative hypothesis, Ha: $\mu \neq \mu 0$

(iii) Level of significance (α) = 5%

(iv) At level of significance =0.05 corresponding LCL = -1.984 and UCL = 1.984

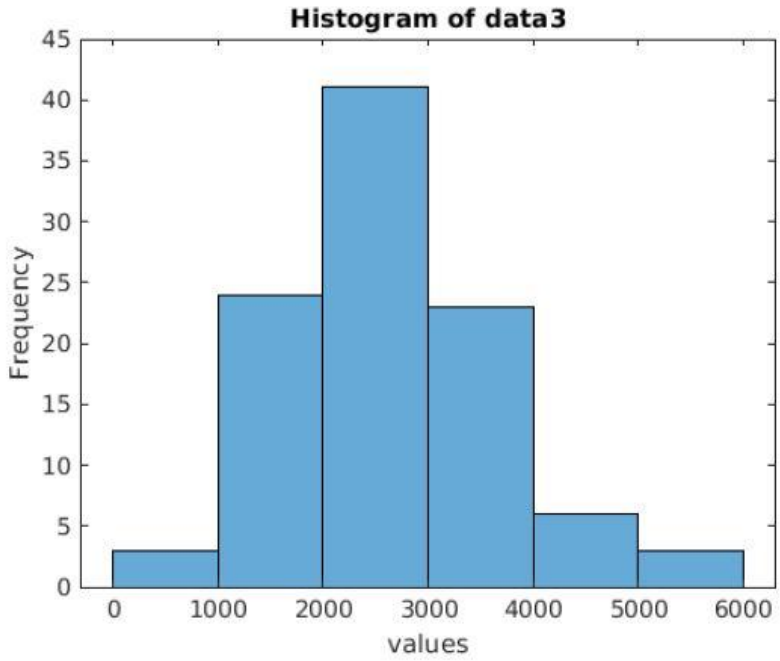
(v) Test statistics: $T = \frac{\mu - \mu 0}{\frac{\sigma}{\sqrt{n}}} = \frac{2451.497 - 2500}{\frac{945.11}{\sqrt{100}}} = -0.5132$

(vi) As $-0.5132 < 1.984$, hence, **Ho cannot be rejected.**



DATA-3

From the given dataset, histogram plotted for Data-3 is given below:



From data-3 we calculate mean and standard deviation as follows:

Mean (μ) = 2654.036

Standard deviation (σ) = 1001.222

$$P(X \geq xk) = 0.01$$

$$1 - P(X \leq xk) = 0.01$$

$$P(X \leq xk) = 0.99$$

$$\varphi(xk) = 0.99$$

$$\frac{Xk-\mu}{\sigma} = 2.33$$

$$Xk = 2.33 * \sigma + \mu$$

$$Xk = 2.33 * 1001.22 + 2654.036$$

$$\boldsymbol{Xk = 4986.8786}$$

Goodness of fit test:

(i) Null hypothesis, H0: Data follow the normal distribution.

(ii) Alternative hypothesis, Ha: Data does not follow the normal distribution.

(iii) Level of significance (α) = 5%

(iv) We are performing Chi-square distribution.

(v) Degree of freedom (DOF) = K-1-m = 6-1-2 = 3, where k = no. of intervals, m = no. of parameters.

(vi) **Test statistics**:
n = 100, Oi = Frequency in each interval, Zi = Normal variate interval, Pi = Probability, ei = expected no. of observations,

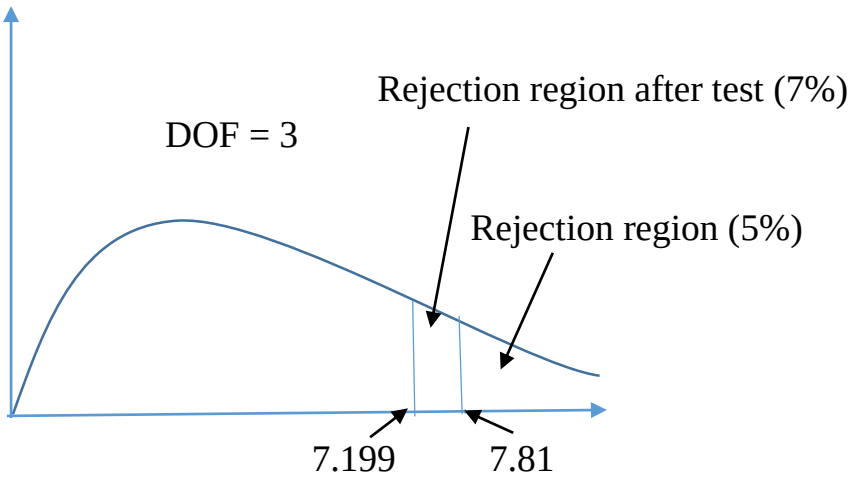
Sl. no.	Interval	Oi	Zi	Pi	ei = nPi	$\frac{Oi^2}{ei}$
1	<1000	3	<-1.652	0.0495	4.95	1.818
2	1000-2000	24	-1.652 -0.653	0.2045	20.45	28.17
3	2000-3000	41	-0.653 0.3455	0.3808	38.08	44.144
4	3000-4000	23	0.3455 1.3443	0.2757	27.57	19.187
5	4000-5000	6	1.3443 2.3431	0.0799	7.99	4.505
6	>5000	3	>2.3431	0.0096	0.96	9.375
		$\sum Oi = 100$				$\sum\left(\frac{Oi^2}{ei}\right)$ $= 107.199$

Test statistics,

$$\sum\left(\frac{Oi^2}{ei}\right) - n = 107.199 - 100 = 7.199$$

Hence, H0 is not rejected. So the distribution is a **Normal distribution**.

vii) P-value = 7%



Hypothesis testing:

Given, Mean discharge = 2500 units

(i) Null hypothesis, H0 : $\mu = \mu 0 = 2500$

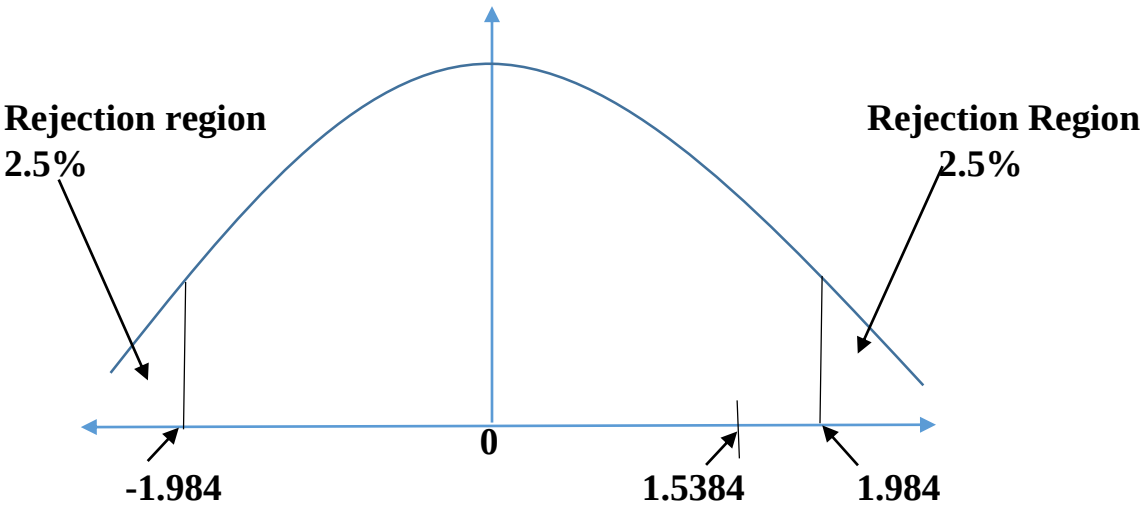
(ii) Alternative hypothesis, Ha: $\mu \neq \mu 0$

(iii) Level of significance (α) = 5%

(iv) At level of significance =0.05 corresponding LCL= -1.984 and UCL = 1.984

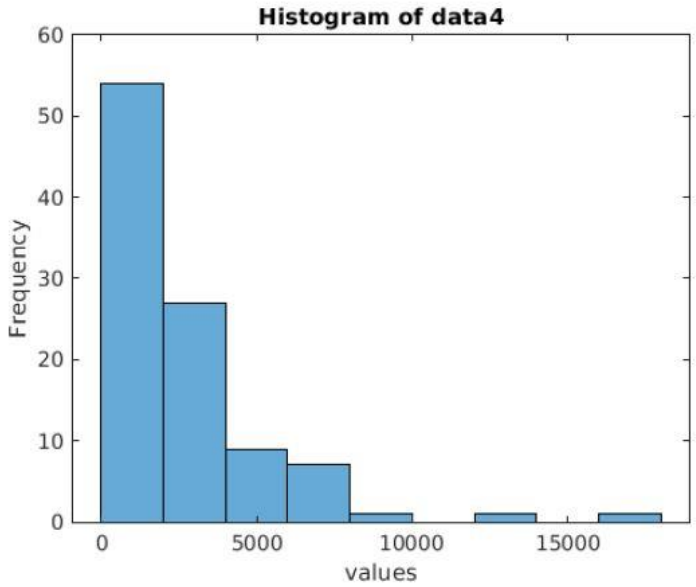
(v) Test statistics: $T = \frac{\mu - \mu 0}{\frac{\sigma}{\sqrt{n}}} = \frac{2654.036 - 2500}{\frac{1001.222}{\sqrt{100}}} = 1.5384$

(vi) As 1.5384< 1.984, hence, **Ho cannot be rejected.**



DATA-4

From the given dataset, histogram plotted for Data-4 is given below:



From data-4 we calculate mean and standard deviation as follows:

Mean (μ) = 2677.399

Standard deviation (σ) = 2701.066

Parameter, $\lambda = \frac{1}{\mu} = \frac{1}{2677.399} = 3.7349 * 10^{-4}$

$$P(X \geq xk) = 0.01$$

$$1 - P(X \leq xk) = 0.01$$

$$P(X \leq xk) = 0.99$$

$$\int_0^{Xk} \lambda e^{-\lambda x} = 0.99$$

$$\left[\frac{\lambda e^{-\lambda x}}{-\lambda}\right] = 0.99$$

$$1- \ e^{-\lambda Xk} = 0.99$$

$$e^{-\lambda Xk} = 0.01$$

$$-\lambda Xk = \ln(0.01)$$

$$-\lambda Xk = -4.605$$

$$Xk = \frac{4.605 * 10^4}{3.7349}$$

Xk=12329.64

Goodness of fit test:

(i) Null hypothesis, H0: Data follow the Exponential distribution.
(ii) Alternative hypothesis, Ha: Data does not follow the Exponential distribution.
(iii) Level of significance (α) = 5%
(iv) We are performing Chi-square distribution.
(v) Degree of freedom (DOF) = K-1-m = 7-1-1 = 5, where k = no. of intervals, m = no. of parameters (here parameter is only λ)
(vi) **Test statistics**:

n = 100, Oi = Frequency in each interval, Zi = Normal variate interval, Pi = Probability, ei = expected no. of observations,

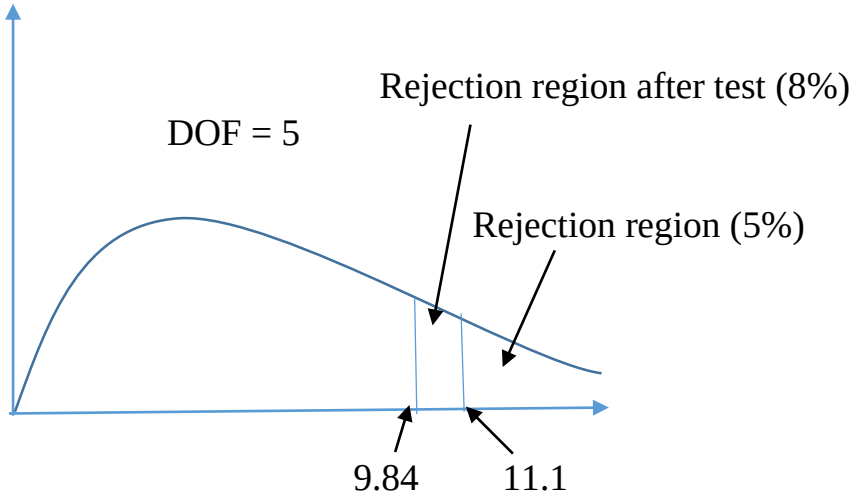
Sl. no.	**Interval**	**Oi**	**Pi**	**ei = nPi**	$\frac{Oi^2}{ei}$
1	0-2000	54	0.5262	52.62	55.4762
2	2000-4000	27	0.2493	24.93	29.2418
3	4000-6000	9	0.1181	11.81	6.8586
4	6000-8000	7	0.05596	5.596	8.7562
5	8000-10000	1	0.02651	2.651	0.3772
6	12000-14000	1	0.00595	0.595	1.6807
7	16000-18000	1	0.00133	0.133	7.5187
		$\sum Oi = 100$			$\sum\left(\frac{Oi^2}{ei}\right) = 109.8496$

Test statistics,

$$\sum\left(\frac{Oi^2}{ei}\right) - n = 109.8496 - 100 = 9.8496$$

Hence, H0 is not rejected. So the distribution is **Exponential distribution**.

(vii) P-value = 8%



Hypothesis testing:

Given, Mean discharge = 2500 units

(i) Null hypothesis, H0 : $\mu = \mu 0 = 2500$

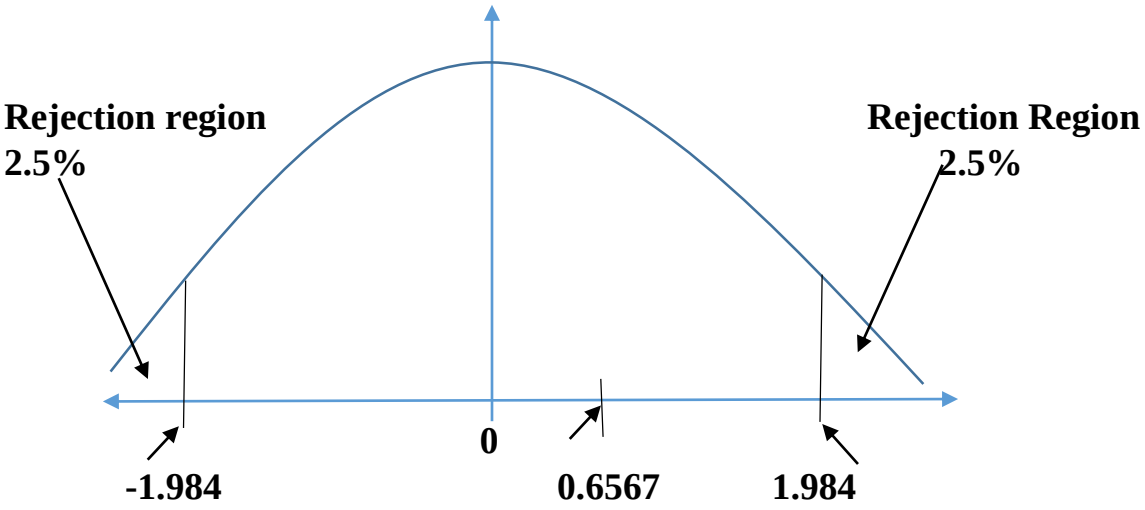
(ii) Alternative hypothesis, Ha: $\mu \neq \mu 0$

(iii) Level of significance (α) = 5%

(iv) At level of significance =0.05 corresponding LCL and UCL = 1.984

(v) Test statistics: $T = \frac{\mu - \mu 0}{\frac{\sigma}{\sqrt{n}}} = \frac{2677.399 - 2500}{\frac{2701.066}{\sqrt{100}}} = 0.6567$

(vi) As 0.6567< 1.984, hence, **Ho cannot be rejected.**



CONCLUSION:

In this project we have analyzed all of the datasets of each site and performed different operations on the data. We have plotted histograms for each dataset and observing the nature of histogram we find 2 to them are normal distribution, one log normal and another is exponential distribution. We have performed Goodness of fit test to accept or reject our assumptions of the distributions. We have also performed Hypothesis testing to see if the mean discharge equals the given value. We have used several tools (ex. MATLAB, Excel sheet etc) to work on this project. This project helped us to clears our concept on this wide area of probability and statistics.

REFERENCES:

Lecture notes by Dr Shivam Tripathi (Professor)