ST. XAVIERS’S COLLEGE (AUTONOMOUS), KOLKATA

DEPARTMENT OF STATISTICS

BRIEF ANALYSIS OF NORMALITY ASSUMPTION ON NON-NORMAL (EXPONENTIAL) QUALITY CHARACTERISTICS IN STATISTICAL QUALITY CONTROL

NAME:	*SAMAPAN KAR*
ROLL NUMBER:	*418*
REGISTRATION NUMBER:	*A01-1112-0844-20*
SUPERVISOR’S NAME:	*DR. SURABHI DASGUPTA*
SESSION:	*2020-2023*

DECLARATION:

”I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.”

Signature: ________________________________

CONTENTS

Topic **Page no.**

Introduction:

In the studies related to Statistical Quality Control industrial organizations generally use Shewhart control chart for quality control purpose. Xbar chart is one of the mostly used method which helps to detect the change in the values of parameters over time. Now while constructing those control charts we often hear that the data have to be Normally distributed or should at least have tendency towards Normality. But that does not have to be necessarily true at all. In real life there are lots of other phenomenon and those phenomenon follows absolutely Non-Normal distributions such as Exponential, Gamma distribution, Uniform distribution etc.

For example the prediction of time when an earthquake might occur, duration of calls made by workers in call centre, lifespan of electronic devices etc. follows Exponential distribution.

Now if we use those control charts for data following Non-Normal distributions then it will affect the control limits and that will lead to different conclusions about the analysis of the dataset and sometimes those can be misleading.

Objective of study:

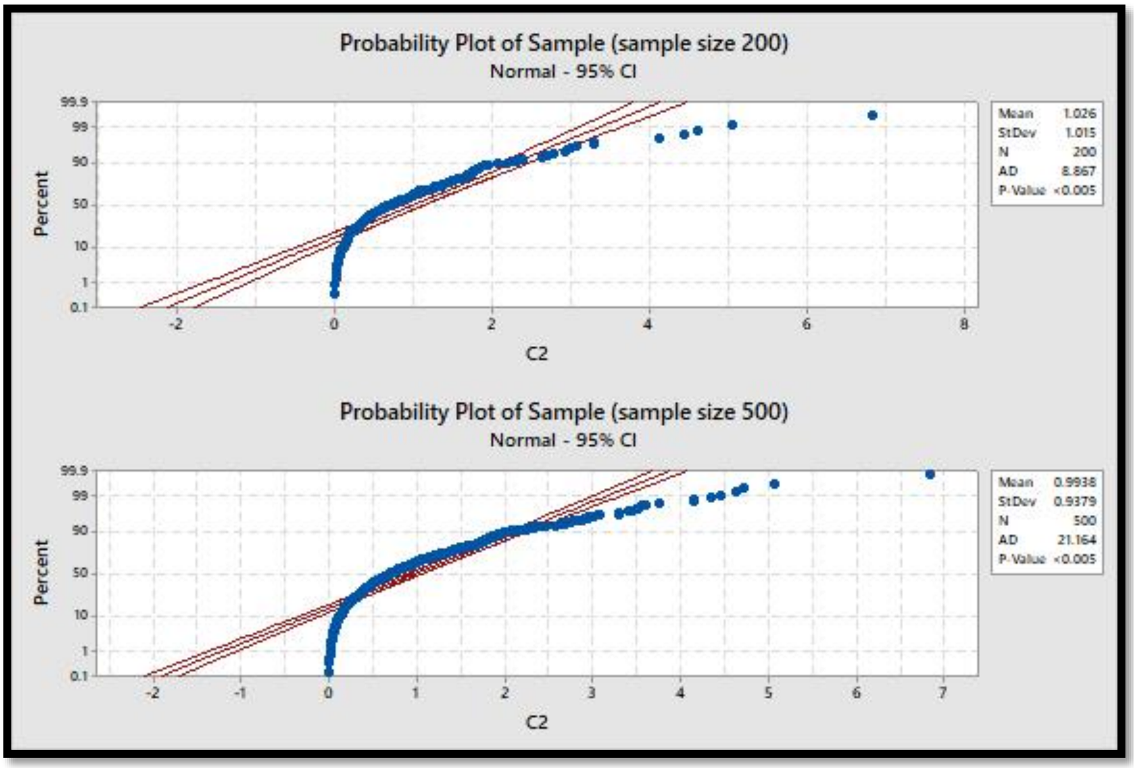
The objective of this study is to construct the control charts for datasets of different sample sizes and subgroups as well as different parameters and observe the robustness of the control limits for using Normal assumption and also some other Quality Control related factors.

Data description:

The data used for this dissertation contains generated random samples of sample size 200 and 500. Both of the random sample follow **Exponential distribution with rate parameter 1** i.e. they have mean 1.

The data is generated using **R (4.2.2) software**. A seed value of **987654321** has been assigned for the Exp (1) dataset i.e. **‘set.seed(987654321)’** for maintaining uniformity of the results obtained.

The following Normal probability plot on the Exponential dataset shows the deviation from Normality of the dataset:



Methodology:

There are many different types of quality control charts present and they are used depending on the data types and sizes. For example, if the data is continuous I-mR chart, Xbar-R chart, Xbar-S chart are generally used. Now as the data used for this study is Exponential i.e. continuous and according to the size of the data, Xbar-R chart is the suitable control chart to work with.

Now in Xbar-R chart there are two components, Xbar which denotes the mean of the sample and R which denotes the range of the sample. Both of the control charts have three different levels such as CL, UCL and LCL. CL denotes the central line which is the average value of the parameter or range. UCL denotes the upper control limit and LCL denotes the lower control limit. Both of them are defined as the natural boundaries of the process.

Also another important factor about the control limits is the deviation of UCL and LCL from the central line. Generally for most of the cases 3-σ limit is used as it is empirically observed that it balances over two types of errors namely Type-I error and Type-II error.

Now let X be the random variable denoting the quality characteristics. The number of sample observation taken is n. If the parameters of the distribution which X follows are known such as mean μ and standard deviation σ then the control limits can be calculated by the following equations:

For Xbar chart:

$$\boldsymbol{UCL = E(\bar{X}) + 3\sqrt{V(\bar{X})} = \mu + 3\sigma}$$

$$\boldsymbol{CL = E(\bar{X}) = \mu}$$

$$\boldsymbol{LCL = E(\bar{X}) - 3\sqrt{V(\bar{X})} = \mu - 3\sigma}$$

For R chart:

$$\boldsymbol{UCL = E(R) + 3\sqrt{V(R)} = \ d_2\sigma + 3D\sigma = \ D_2\sigma}$$

$$\boldsymbol{CL = E(R) = \ d_2\sigma}$$

$$\boldsymbol{LCL = E(R) - 3\sqrt{V(R)} = \ d_2\sigma - 3D\sigma = \ D_1\sigma}$$

Here d$_2$ and D are both functions of sample size n. The values of d$_2$, D$_1$, D$_2$ are also given according to the sample size.

Now if the values of the parameters are unknown then we have to estimate it from the sample itself. In that case we have to divide the sample into subgroups using the principle of rational subgrouping. Let the data is divided into m subgroups and each subgroup has n number of observations.

$\boldsymbol{\overline{X} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}}$, where X$_{ij}$ denotes the observation corresponding to the j-th sample of i-th subgroup. $\boldsymbol{\overline{R} = \frac{1}{m}\sum_{i=1}^{m} R_i}$, where $\boldsymbol{R_i = X_{i(n)} - X_{i(1)}}$, the sample range of j-th subgroup. Then the control limits for the Xbar and R chart can be calculated by the following equations:

For Xbar chart:

$$\boldsymbol{UCL = \hat{\mu} + 3\frac{\hat{\sigma}}{\sqrt{n}} = \overline{X} + \frac{3}{\sqrt{n}}\frac{\overline{R}}{d_2} = \overline{X} + A_2\overline{R}}$$

$$\boldsymbol{CL = \hat{\mu} = \overline{X}}$$

$$\boldsymbol{UCL = \hat{\mu} - 3\frac{\hat{\sigma}}{\sqrt{n}} = \overline{X} - \frac{3}{\sqrt{n}}\frac{\overline{R}}{d_2} = \overline{X} - A_2\overline{R}}$$

For R chart:

$$\boldsymbol{UCL = D_2\hat{\sigma} = D_2\frac{\overline{R}}{d_2} = D_4\overline{R}}$$

$$\boldsymbol{CL = d_2\hat{\sigma} = \overline{R}}$$

$$\boldsymbol{UCL = D_1\hat{\sigma} = D_1\frac{\overline{R}}{d_2} = D_3\overline{R}}$$

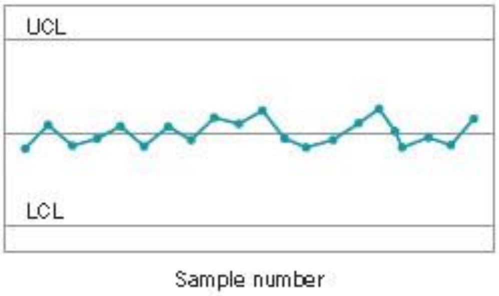
Results & Discussion:

Here we have the two datasets of sample size 200 and 500. For calculating the control charts we are using the subgroups of size 5 and 10 for both of the sample.

Now for the use of control chart the use of Normality assumption causes different types of erroneous features that can be easily observed. Some of them are stated below:

Stratification:

The tendency for the points to cluster artificially around the centre line is known as Stratification. It can be observed in the Xbar chart. The reason behind stratification is the incorrect calculation of the control limits. It can also happen if the sample within a subgroup follow different probability distributions.



■ FIGURE A stratification pattern.

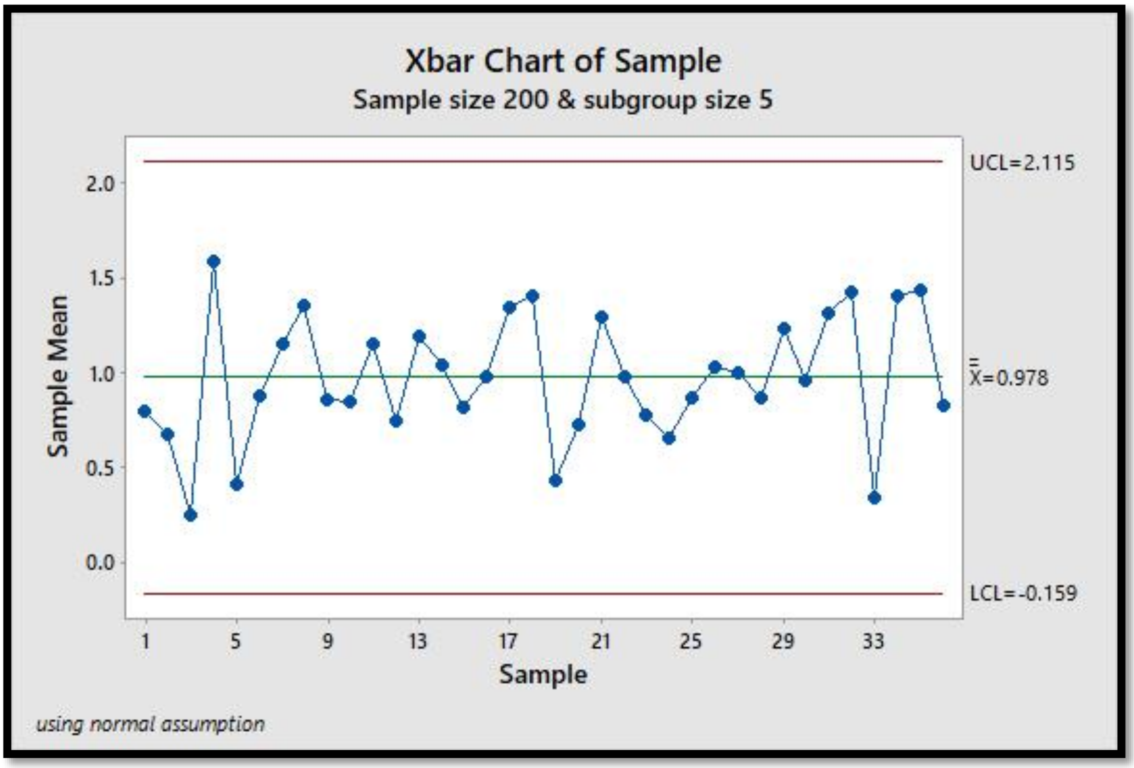
While observing Xbar chart we should examine the R chart first. If the R chart is not under control then there is no meaning of interpreting the Xbar chart. When both of the charts show out of control observations then it is better to remove those out of control observations. This will automatically eliminate the non-randomness of the data.

Now for our dataset we have observations from Exponential distribution and as it is a random generated observation we can ignore the cause of sampling from different probability distribution.

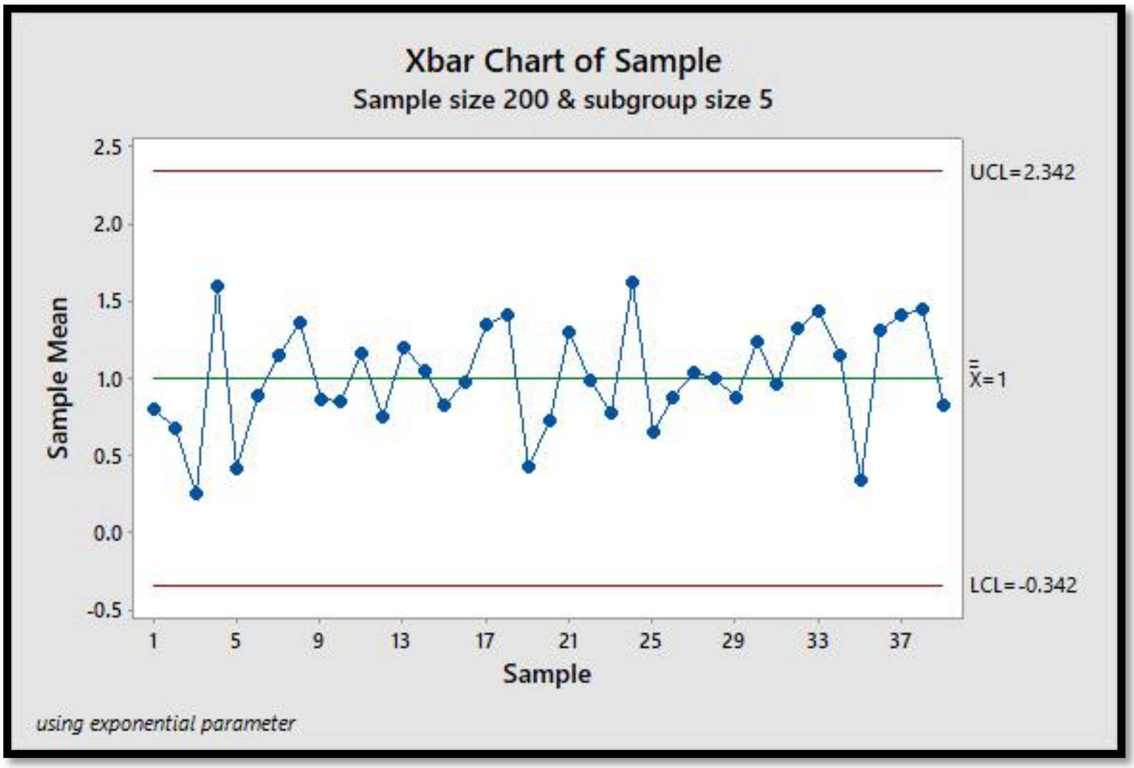
The Xbar charts for the two datasets using Normality assumption and using the known Exponential distribution parameter are given below:

For sample of 200 Exponential (1) observations and subgroup size 5:

Using Normality assumption:

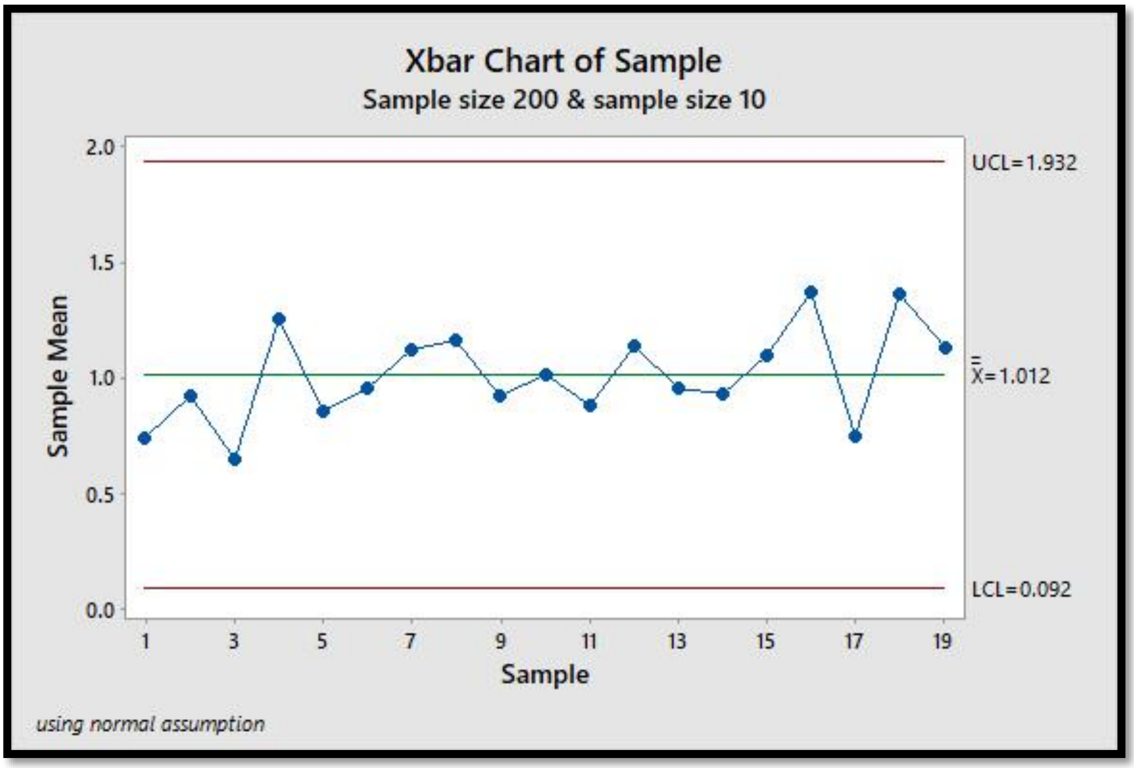


Using Exponential Parameter:

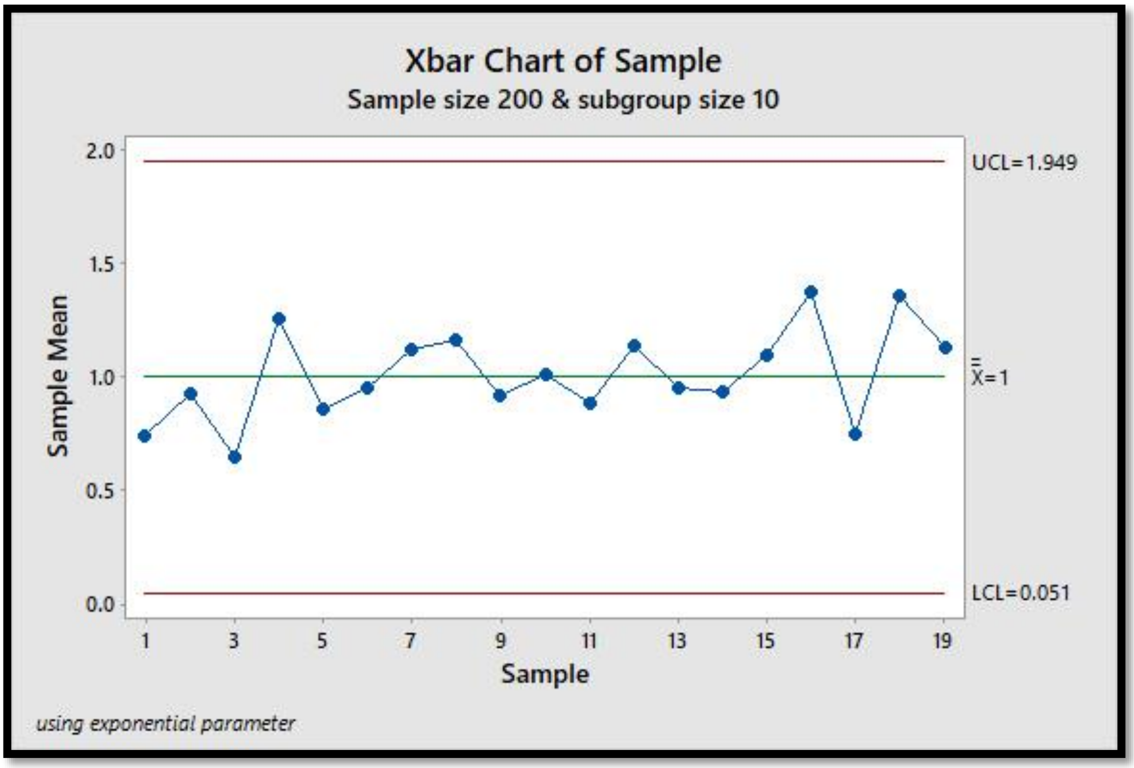


For sample of 200 Exponential (1) observations with subgroup size 10:

Using Normality assumption:

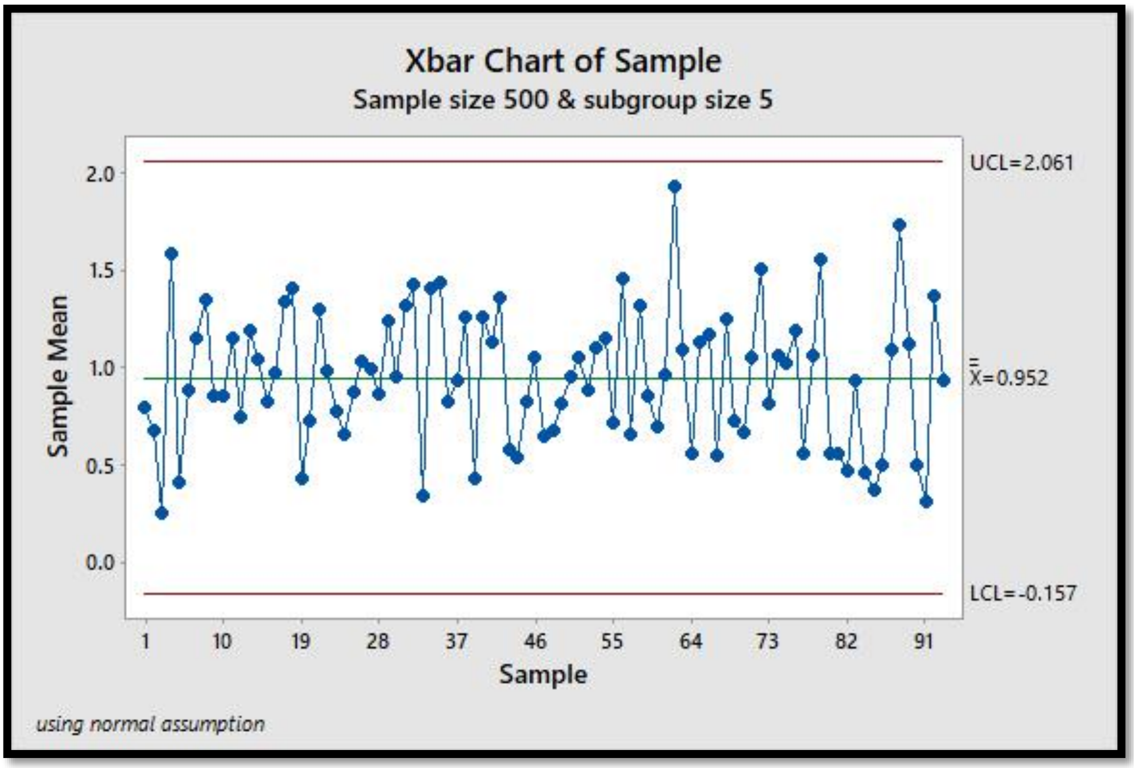


Using Exponential parameter:

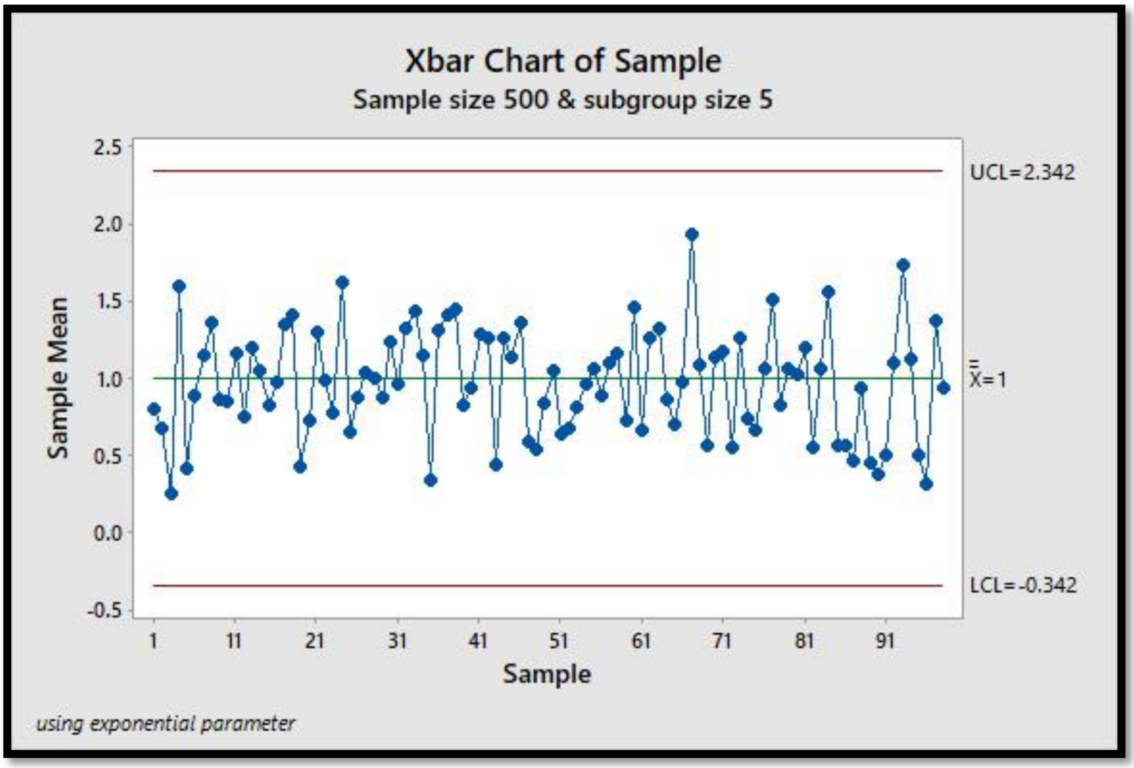


For sample of 500 Exponential (1) observations with subgroup size 5:

Using Normality assumption:

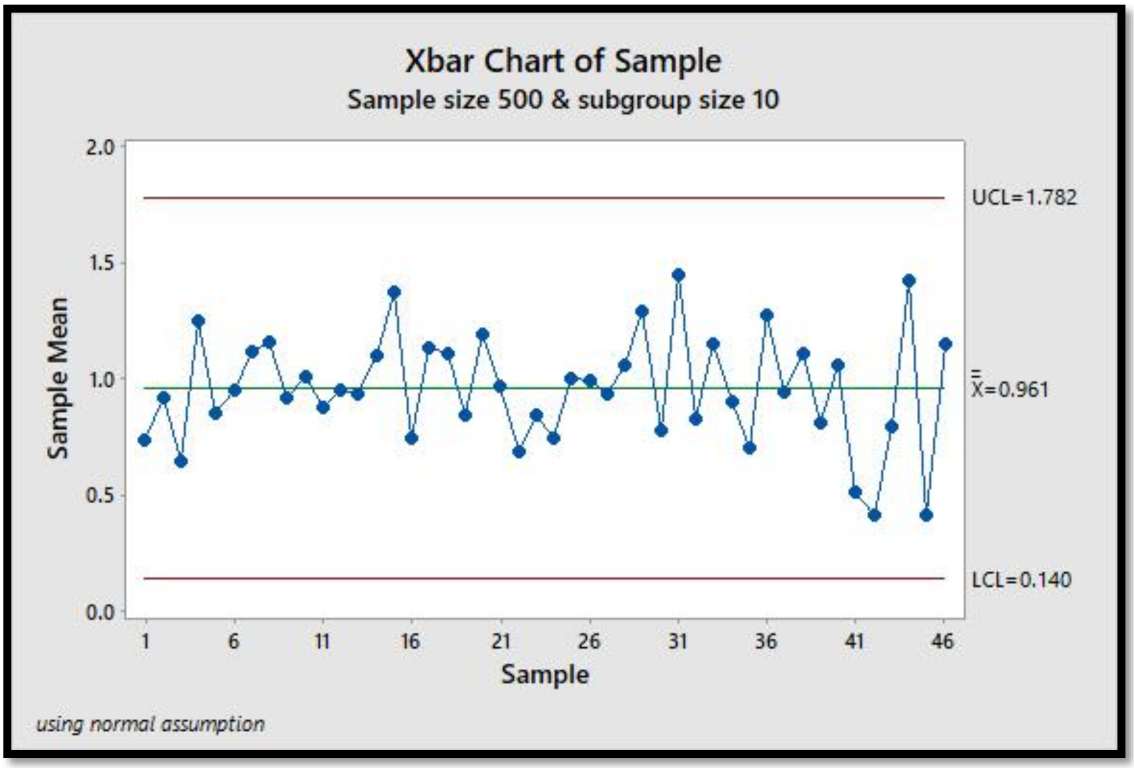


Using Exponential parameter:

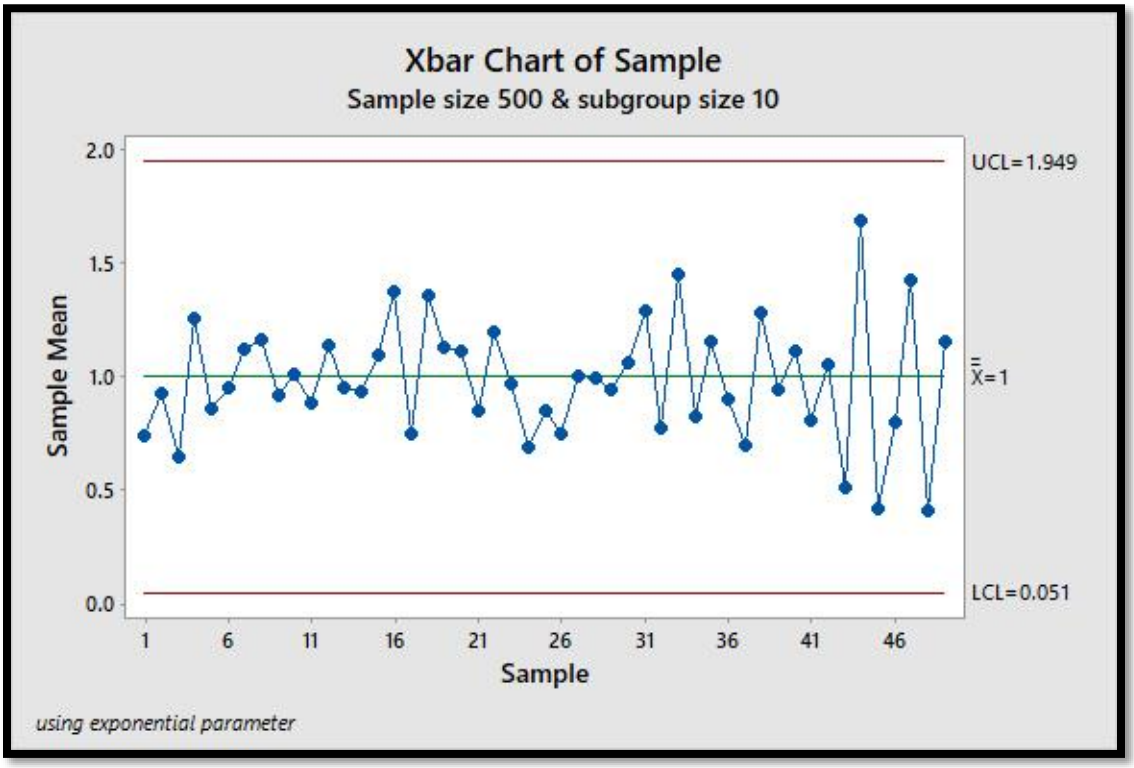


For sample of 500 Exponential (1) observations with subgroup size 10:

Using Normal assumption:



Using Exponential parameter:



Now from those diagrams of Xbar charts we see that:

- For most of the cases the points are tend to fall near the central line and away from UCL and LCL which clearly indicates that stratification is present throughout the observations.

- For small sample size the most of the points are present near central line. As the sample size increases the total number of subgroups increases and for that reason the number of stratified points increases.

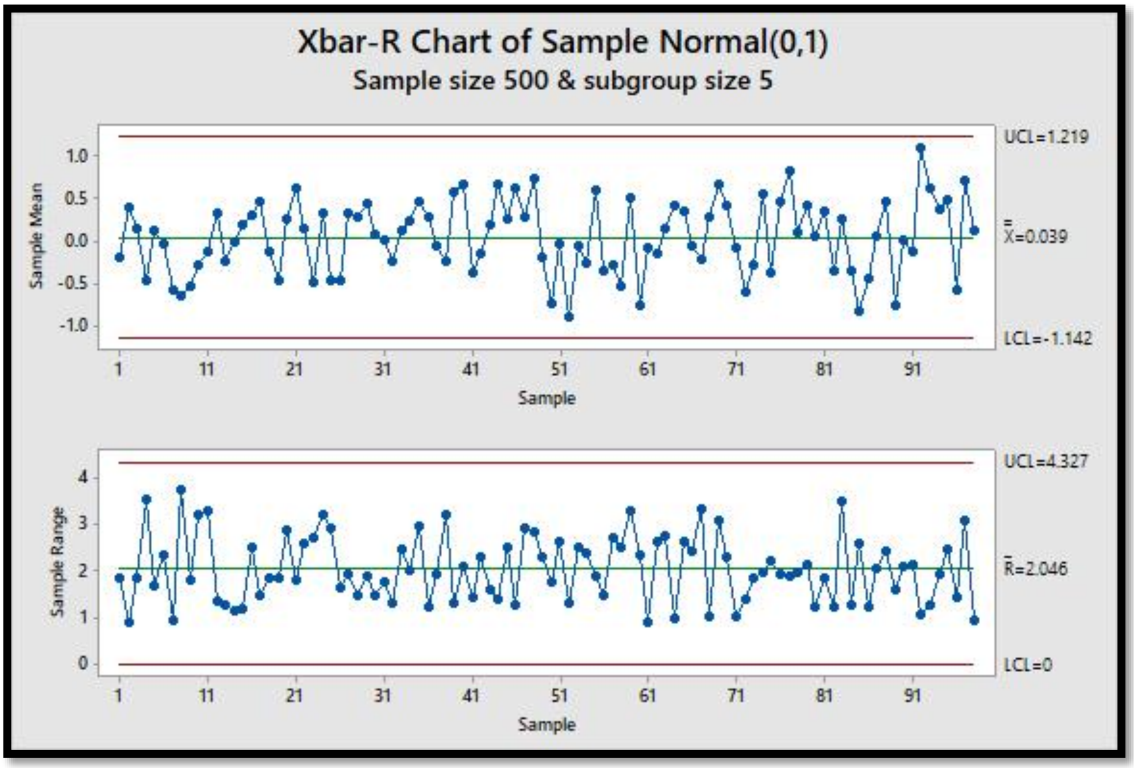
- For different subgroup sizes a change of stratification can be observed but the change is not that much significant. For small subgroup sizes the number of total subgroups is bigger so more of the points are stratified than the bigger subgroup sizes with smaller number of total subgroups.

But here we can observe that the graphs for Normality assumption is more or less same with the graphs for using Exponential distribution parameters. Though there are presence of upward or downward shifts but that does not affect much in the stratification. For a large number of data or for using other parameter values we may get significantly different Xbar charts for both cases but for our study it is not a strong point to make any such conclusion. So we proceed to the next point.

Correlation between the Xbar and R chart:

When we are interpreting the patterns on Xbar and R chart it is better to consider two of them Jointly. Now while comparing two of them there is a systematic way to check if the two random variables Xbar and R are statistically independent. If the probability distribution that X follows is Normal then the random variables Xbar and R which are calculated from the same sample have to be statistically independent. Therefore on the control chart the points will behave independently for the both Xbar and R chart i.e. the pattern for both of the graphs will be different. But if the underlying distribution is Non-Normal then there will be presence of correlation between Xbar and R values. So we can observe that the points on the Xbar and R chart will follow each other i.e. there will be a similar pattern in both of the graphs.

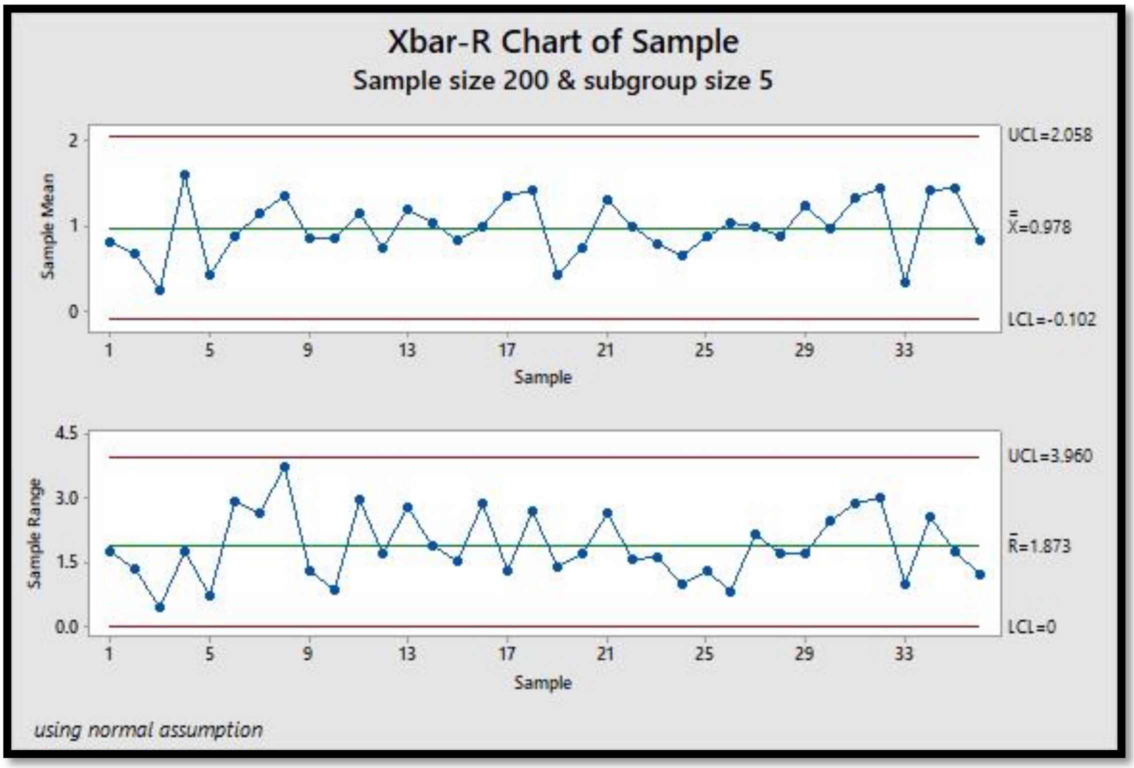
The Xbar-R chart of a random sample of 500 observations from Normal (0,1) distribution is given below:



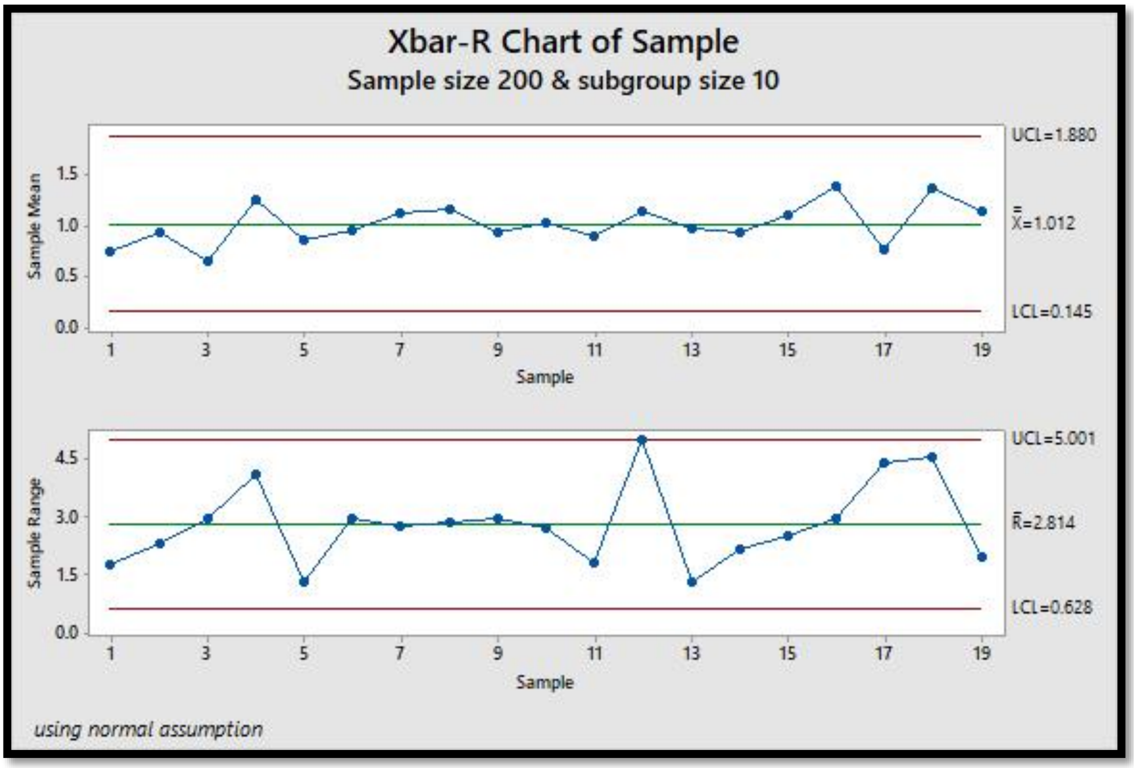
Here we can see that the two graph has no similarity between them. Two of them are completely random.

Now for our two datasets the Xbar-R charts are given below:

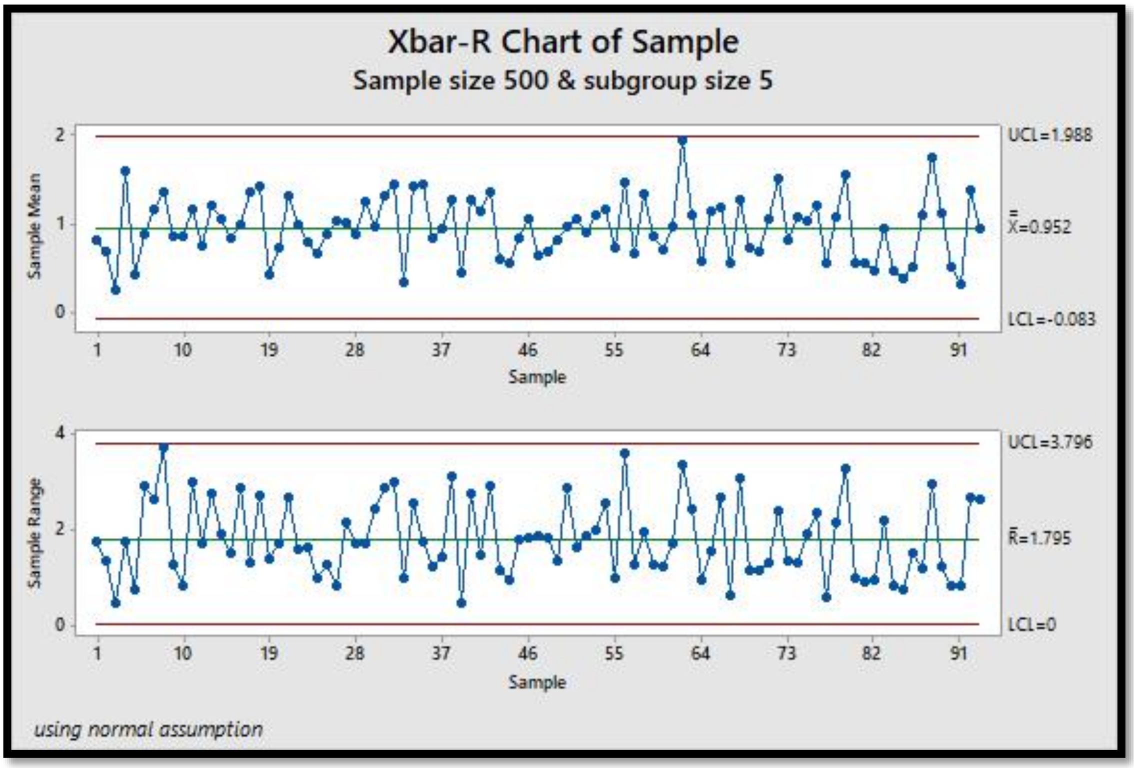
For sample of 200 Exponential (1) observations and subgroup size 5:



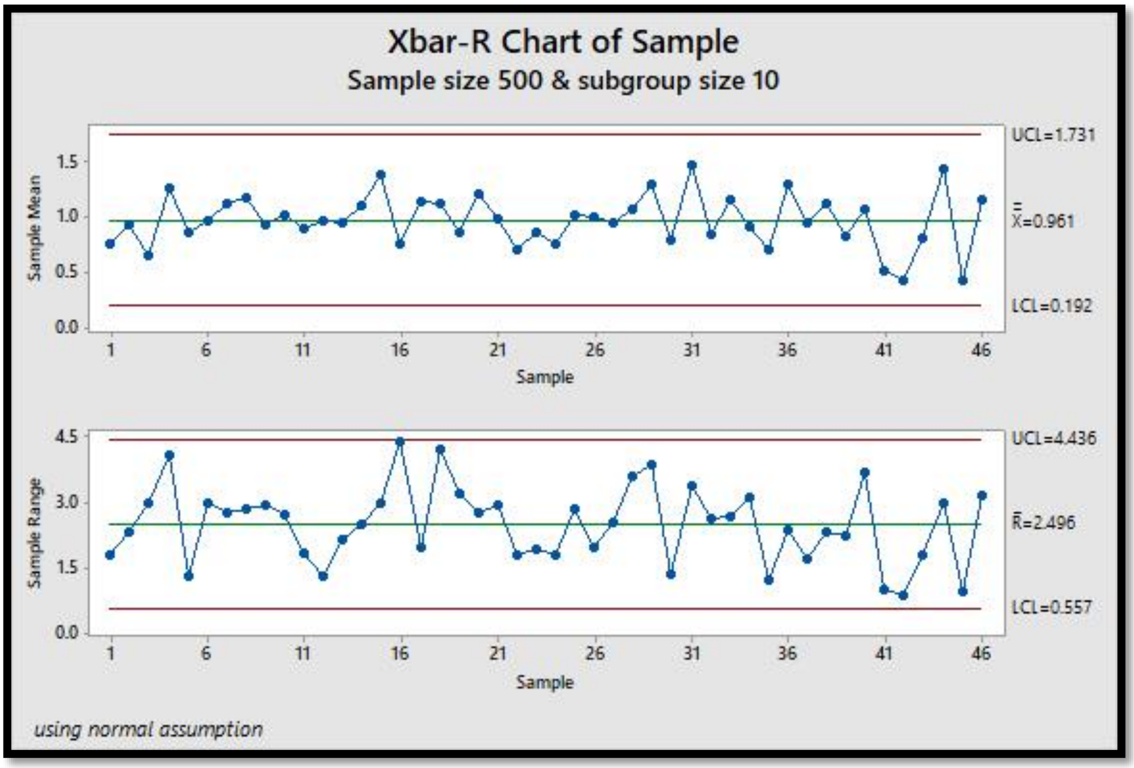
For sample of 200 Exponential (1) observations and subgroup size 10:



For sample of 500 Exponential (1) observations and subgroup size 5:



For sample of 500 Exponential (1) observations and subgroup size 10:



Now from those Xbar-R charts we can say that compared to the charts for the data from Normal distribution there are presence of similarities between the Xbar and R chart for the Exponential datasets. We can also observe that:

- For small sample size the similarity between the Xbar and R chart is not that much prominent. As we increase the sample size the two graphs become more similar to each other.

- For different subgroup sizes we have same kind of interpretation. The similarities between the two graphs become more prominent as we increase the subgroup size.

So from these observations we can say that the similarity in those Xbar-R charts is due to the sample taken from a Non-Normal distribution and as Exponential (1) distribution is highly skewed the normality assumption on those datasets are not much effective to reduce their Non-Normality criteria. Also increasing the sample size can reduce that effect. So Normality assumption on this type of highly skewed data is not a good choice and the analysis can lead to a wrong conclusion.

So we proceed to the third point.

α-risk (Type-I error):

In the context of Statistical Quality Control, α-risk (also known as Type I error) refers to the probability of rejecting a true null hypothesis. In SQC, the null hypothesis is often that a process is under control and the alternative hypothesis is that it is out of control.

If the process is actually under control but a SQC test incorrectly identifies it as being out of control, this is a Type I error, and the probability of this occurring is α. In other words, α represents the level of significance at which the null hypothesis is rejected.

In SQC, controlling the α-risk is important because it helps ensure that the conclusions drawn from statistical analyses are reliable and accurate. To reduce the α-risk, researchers can increase the sample size, use a more conservative significance level, or perform a power analysis to determine the minimum sample size needed to detect a true difference or relationship with a desired level of confidence.

Now for our dataset we have to calculate the α-risk for both the cases i.e. for the Normality assumption and for using the Exponential distribution parameter. Also we are going to calculate it for both Xbar and R chart.

For calculating the α-risk we have to use the following equation:

For Xbar chart:

$$\boldsymbol{\alpha = 1 - P(\ a\ randomly\ selected\ point\ will\ fall\ inside\ the\ limits)}$$

$$\Rightarrow \boldsymbol{\alpha = 1 - P(LCL < \overline{X} < UCL)}$$

$$\Rightarrow \boldsymbol{\alpha = 1 - P\left(LCL < \frac{1}{n}\sum_{i=1}^{n} X_i < UCL\right)}$$

Here **Xᵢ** is are the random variable following Exponential (1) distribution, **n** is the subgroup size and UCL and LCL are the control limits of the Xbar chart which are collected from the in-control Xbar-R charts for both Normality assumption and Exponential parameter cases.

<u>For R chart:</u>

$$\boldsymbol{\alpha = 1 - P(\,a\ randomly\ selected\ point\ will\ fall\ inside\ the\ limits)}$$

$$\Rightarrow \boldsymbol{\alpha = 1 - P(LCL < R < UCL)}$$

$$\Rightarrow \boldsymbol{\alpha = 1 - P(LCL < X_{(n)} - X_{(1)} < UCL)}$$

Here $\mathbf{X_{(n)}}$ and $\mathbf{X_{(1)}}$ are the maxima and minima for the n-th subgroup and UCL and LCL are the control limits of the Xbar chart which are collected from the in-control Xbar-R charts for both Normality assumption and Exponential parameter cases. **R** has the pdf $f_X(x) = (n-1)\lambda e^{-\lambda x}(1 - e^{-\lambda x})^{n-2}$, where λ is the Exponential distribution parameter i.e. 1 for our dataset and n is the size of subgroup.

The α-risk values are given in the following table:

Sample size	**Subgroup size**	**Assumption type**	**α-risk for Xbar chart**	**α-risk for R chart**
200	5	Normality assumption	0.0242	0.0740
		Using Exponential parameter	0.0093	0.0289
200	10	Normality assumption	0.0099	0.06
		Using Exponential parameter	0.0067	0.0391
500	5	Normality assumption	0.0304	0.08686
		Using Exponential parameter	0.0093	0.0289
500	10	Normality assumption	0.0223	0.1021
		Using Exponential parameter	0.0067	0.0391

Now in a controlled process, it is important to keep the α-risk small because it helps ensure that any observed differences or effects are real and not simply due to chance. A small α-risk means that the probability of falsely rejecting a true null hypothesis is low, which helps to reduce the likelihood of making incorrect decisions based on faulty conclusions.

From the table we can observe that:

- For small sample size the α-risk is relatively smaller. Increasing the sample size also increases the value of α-risk. For both Xbar chart control limits and R chart control limits we can observe the same thing.

- For both Xbar and R chart control limits the α-risk is larger for small subgroup size. As we increase the subgroup sizes the α-risk is getting smaller.

- For Normality assumption cases the α-risk values are coming as larger for both Xbar and R chart control limits considering all sample sizes, subgroup sizes than using the Exponential parameter value.

So we can conclude that using Normality assumption we are getting higher α-risk values which implies that the data must has more out of control points. But actually the points are not much out of control and Normality assumption leads us to an incorrect decision about the dataset.

Some more information supporting this are given in the next point.

Average Run Length (ARL_0):

In Statistical Quality Control, the Average Run Length (ARL) is a measure of the expected number of samples that will be taken before a signal indicating a change in the process mean or variance is detected.

The ARL depends on the control chart used, the type and magnitude of the shift in the process, and the sample size. Generally speaking, the ARL increases as the magnitude of the shift decreases and as the sample size increases. When the process is in-control, ARL is denoted as ARL_0.

In the following table the ARL_0 values for different subgroup and sample sizes are given:

Sample size	Subgroup size	Assumption type	ARL_0
200	5	Normality assumption	42
		Using Exponential parameter	108
200	10	Normality assumption	101
		Using Exponential parameter	149
500	5	Normality assumption	33
		Using Exponential parameter	108
500	10	Normality assumption	45
		Using Exponential parameter	149

Now from the table we can see that increasing the sample size the ARL_0 value decreases for the Normality assumption cases which opposes the true nature of ARL_0. Also we can see that the ARL_0 values are lower for Normality assumption cases than the using of Exponential parameter cases which also indicates that Normality assumption is not suitable in this cases.

Conclusion:

The normality assumption is an important aspect of many statistical methods, including those used in Statistical Quality Control. However, in practice, many datasets do not follow a normal distribution. When data is non-normal, it may be tempting to assume normality in order to use familiar statistical methods, such as control charts or process capability analysis.

However, assuming normality when data is not actually normally distributed can lead to incorrect conclusions and poor decisions. For example, if a control chart assumes normality but the data is actually highly skewed or has heavy tails, it may result in false alarms or failure to detect actual process changes. For our dataset we have used ransom sample which is generated from Exponential distribution with rate parameter 1 which actually is a highly skewed data. And we also have shown through the above stated four points that Normality assumption in all cases is not a very good choice as it causes typical errors leading towards the wrong conclusions.

Therefore, it is important to assess the normality of the data before applying any statistical method that assumes normality. In conclusion, while the normality assumption is important in many statistical methods, it is crucial to properly assess and handle non-normal data in SQC to ensure accurate results and effective quality control.

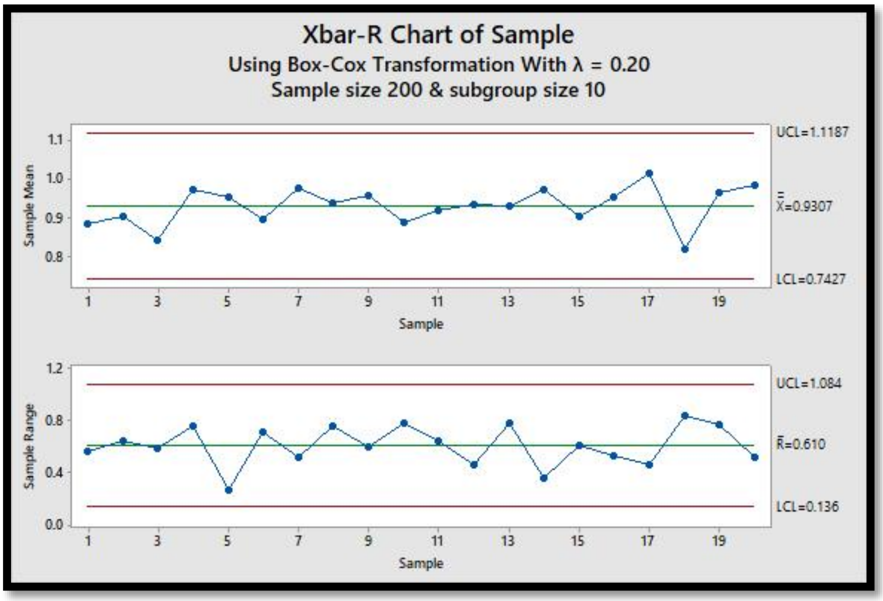
If the data is found to be non-normal, there are alternative methods that can be used in SQC, such as **Box-Cox transformation.** The following process is stated below:

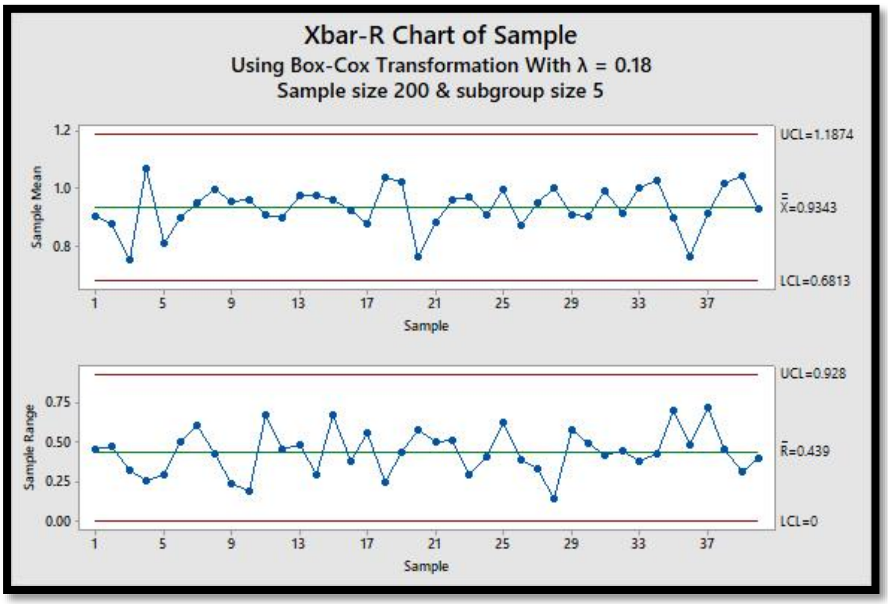
Box-Cox transformation:

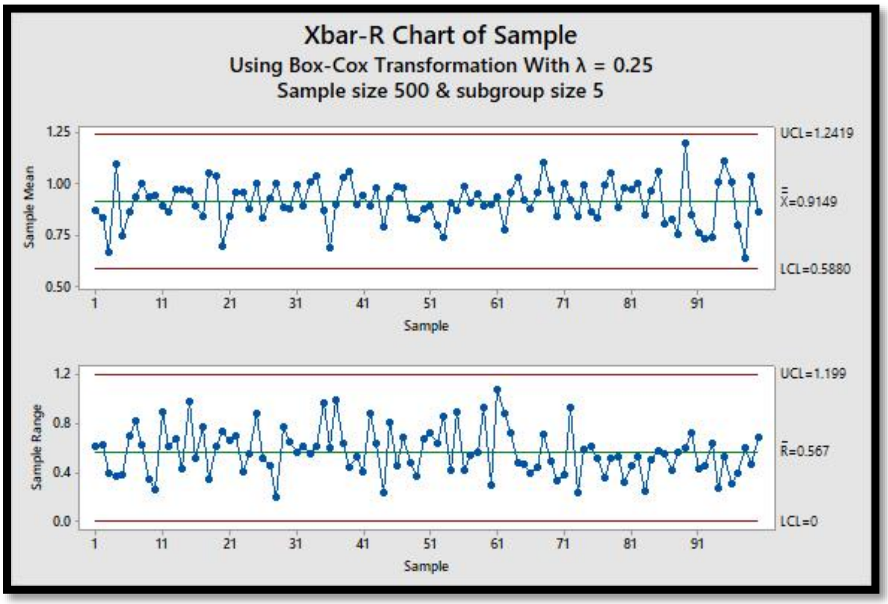
Box-Cox transformation is a statistical technique used to transform a non-normal dependent variable into a normal distribution. The transformation involves raising the data to a power, which can be optimized to find the best fit to a normal distribution.

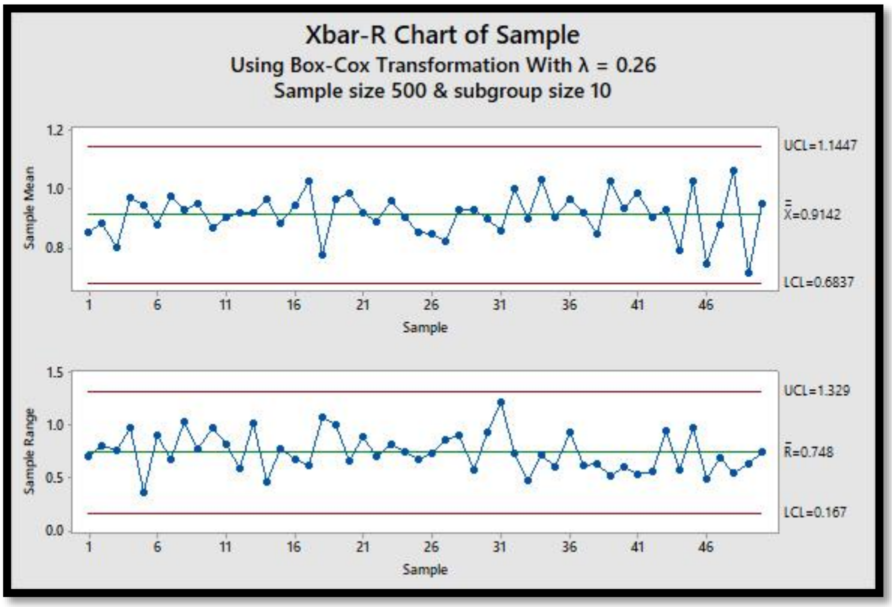
Now in Box-Cox transformation we have a target variable which we intend to estimate and for our dataset it is the random variable following Exponential distribution. If $y(\lambda)$ is denoted as the transformed variable and y is our target variable then the transform equation looks like: $y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & if\ \lambda \neq 0 \\ \log(y), & if\ \lambda = 0 \end{cases}$. Here we choose the λ according to the best approximation for the Normal distribution of our response variable.

The Xbar-R chart created using the Box-Cox transformation are given below. Here the optimal λ value is used to set up the control charts.









However for interpretation purpose this method sometimes is not a good choice. For non-zero λ values the transformed variable sometimes becomes more difficult to interpret. But overall Box-Cox transformation can be useful for analysing data that violate the normality assumption, which is often assumed in many statistical tests and control charts. By transforming the data to approximate normality, SQC techniques can be more accurately applied to the data, leading to better decisions and more effective process control.

Scopes of further work:

For our dissertation purpose we have limited our work only to the analysis of Xbar and R chart for assumption of Normality to the Non-Normal Continuous type dataset. But we have not studied about the Xbar and S chart for the same assumption. So the study of Xbar-S chart can be done as a further work. Also there are other Non-Normal continuous type distributions which are not much highly skewed or tends to become Normal distribution with increasing sample size or degrees of freedom such as t-distribution. So for that types of datasets the similar analysis work can be done.

References:

● Douglas C. Montgomery. Introduction to Statistical Quality Control (Sixth Edition)

● https://www.spcforexcel.com/knowledge/variable-control-charts/control-charts-and-non-normal-data

● https://builtin.com/data-science/box-cox-transformation-target-variable

Acknowledgement:

This project would not have been possible without many people’s generous support and assistance. I want to express my heartfelt gratitude to every one of them.

First and foremost, it has been a tremendous honor and privilege to work under my research supervisor, Dr. Surabhi Dasgupta. I am grateful to her for her continuous motivation, guidance and enthusiasm throughout the project. Her expertise, valuable insights, and continual advice have been instrumental in shaping the ideas and approaches to the project.

I would also like to thank all the other professors of the Department of Statistics, St. Xavier’s College (Autonomous), Kolkata, for sharing their pearls of wisdom with me during the entire course of the project, as well as St. Xavier’s College itself for giving the chance and providing me with necessary resources and facilities to carry out this project.

My thanks and appreciation also go to my classmates and peers for their help, valuable feedbacks and suggestions.

Last but not the least, I am also immensely grateful to my family for their care, motivation and guidance throughout the project.