A Simulation Study of Sample Size Determination for Achieving Normality of Beta Distribution

Brishti Sarkar, Ishani Karmakar, Soham Biswas, Srijan Sen

UG Students

Department of Statistics

St. Xavier's College (Autonomous)

30 Mother Teresa Sarani

Kolkata 700016

West Bengal, India

Email Id: biswassoham434@gmail.com

Keywords: Central Limit Theorem, Shapiro-Wilk Test, Beta distribution, Mean, Standard Deviation, Coefficient of Skewness, Coefficient of Kurtosis, Sensitivity Study.

Abstract

Theoretically, we know that the distribution of numerous statistics converges to normal distribution.

But, the achievement of normality in their distribution is influenced by various factors including the distribution of the parent population, values of the population parameters and the statistic under consideration, so as to determine the sample size required to achieve normality.

In the following simulation study, we determine the sample size required to achieve normality for the statistics mean, standard deviation, coefficient of skewness and coefficient of kurtosis where the sample has been drawn from Beta populations with varying parameters.

Introduction

For testing of hypothesis and interval estimation of parameters, the parent population is usually considered to be Normal. However, in practical situations this phenomenon seldom holds true. In this situation if one uses the normality assumption, the inference about the parameters is bound to become erroneous, unless, the sample size is large enough.

A natural question then arises that how large a sample should be considered so that it may be regarded to be drawn from a Normal population.

In this paper, we have considered samples from Beta population (with varying parameters), and studied the size of the sample required for achieving Normality of a few statistics, viz. sample mean, sample standard deviation, sample coefficients of skewness and kurtosis. The relative magnitude of the parameters has been varied in such a way that the sample size may be determined for symmetric, as well as skewed situations. This in another way may be looked upon as a sensitivity study of the sample size with respect to the value of the parameters.

Statistical Methodologies

Central Limit Theorem

The central limit theorem in the discipline of Statistics is a statistical theory which states that when we are given a sufficiently large sample size from a population possessing a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population.

Central Limit Theorem is basically the shape of the distributions of means which we will get if we draw samples repeatedly from a given population.

To be specific, as the sample sizes get larger, the distribution of means that we calculate from repeated samplings approach normality. This is what we get to know from the central limit theorem.

Thus, from central limit theorem we get to understand that when we successively take an infinite number of samples randomly from a population, we will find that the sampling distributions of the means of these samples will approximately follow a normal distribution with some mean and standard deviation, irrespective of the shape of the population distribution.

Thus, the central limit theorem can be broken down to 3 components namely,

(i) Successive sampling from a population
(ii) Increasing sample size
(iii) Population Distribution

The central limit theorem is useful when examining returns for a particular stock or some index for the very reason that it simplifies many analysis procedures. When an appropriate sample size is achieved, which is relatively easy to generate for financial data as they are capable of producing large sample sizes central limit theorem can be applied on the data.

The central limit theorem being the basis for sampling in statistics, holds the foundation for sampling and statistical analysis in finance too. Investors rely heavily on the central limit theorem to analyze stock returns, construct portfolios and manage various financial risks.

Shapiro-Wilk Test

The Shapiro-Wilk Test is used for testing normal distribution and exhibits high probability of correctly rejecting a false null hypothesis (the normal assumption in statistics).

Even with less no. of observations Shapiro-Wilk test leads to good results but in contrast to other tests it is only applicable to tests for normality.

The Shapiro-Wilk test calculates a W statistic, which is used to test whether a random sample x_1, x_2, …,x_n comes from a normal distribution (specifically).

The W statistic is calculated as:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where the $x_{(i)}$ are the ordered sample values ($x_{(1)}$ being the smallest)a_i's are the constants generated from the means, variances and co – variances of the order statistics of a sample size n from a normal distribution.

Small values of W statistic are evidence of the departure from normality.

Shapiro-Wilk test is quite effective as compared to other goodness of fit tests.

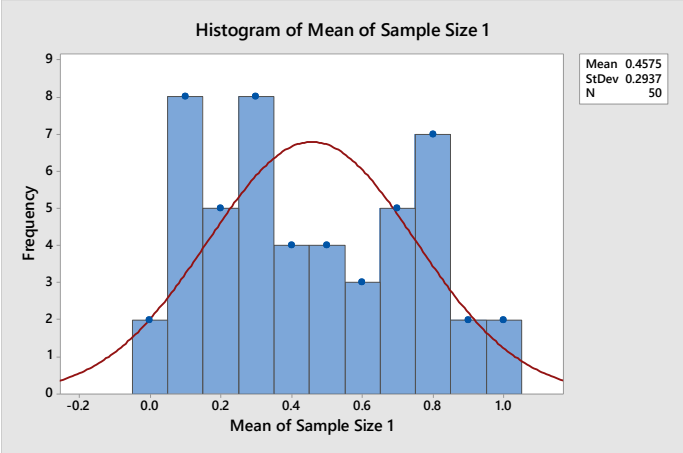
The Shapiro-Wilk test was published in 1965 by Samuel Stanford Shapiro and Martin Wilk.

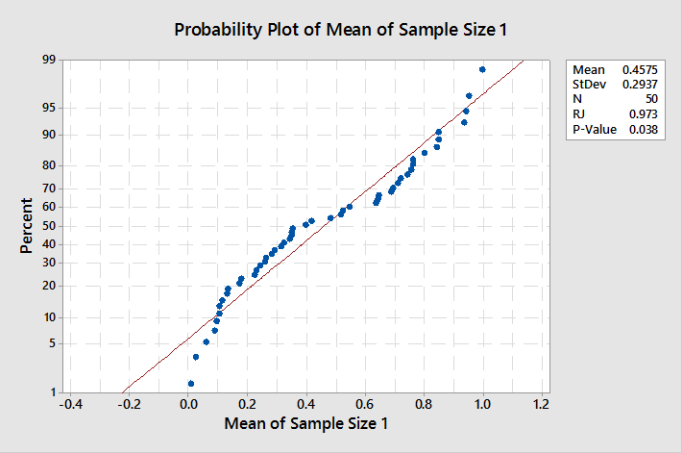
RESULTS

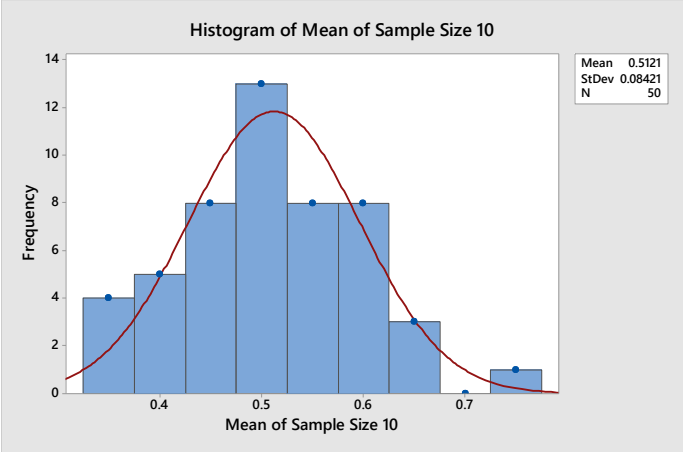
Let us consider the parent distribution as Beta with varying values of parameters m and n, i.e. X~Beta(m,n).

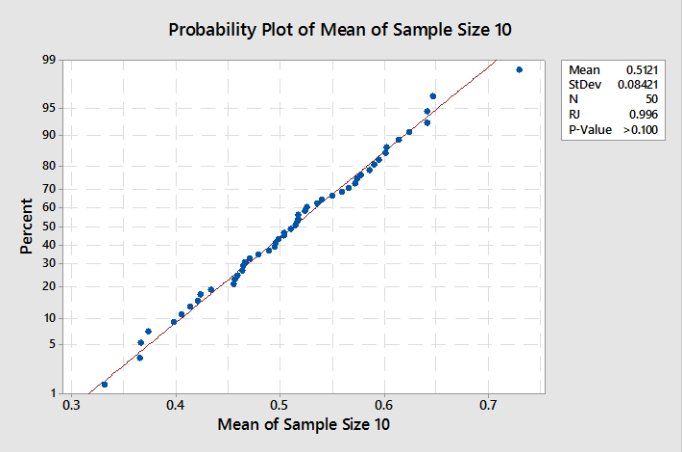
Beta(1,1)	**Normality achieved at sample size**	**p-value**	**Normality not achieved till sample size**	**p-value**
Mean	10	>0.1	1	0.038
Standard Deviation	8	>0.1	2 3	<0.01 0.041
Coefficient of Skewness	19	>0.1	3 9	<0.01 0.029
Coefficient of Kurtosis	41	>0.1	7 35	<0.01 <0.01

The histograms of the distribution of the statistic Mean along with the Normal density curve is given below:

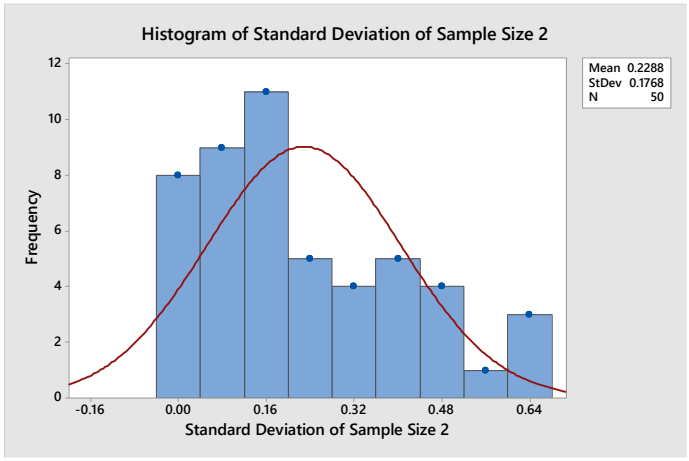


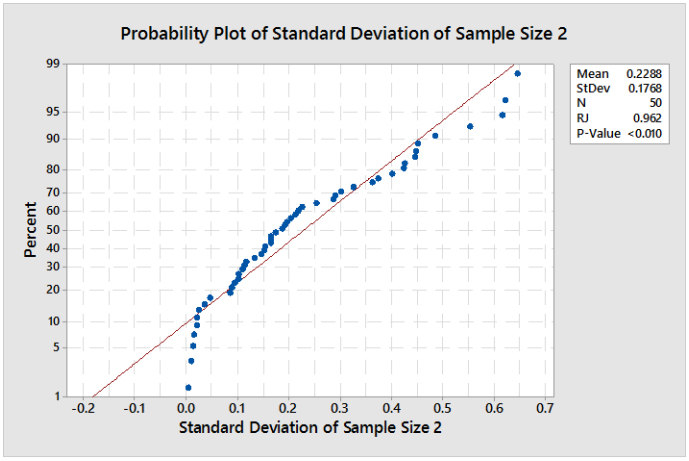


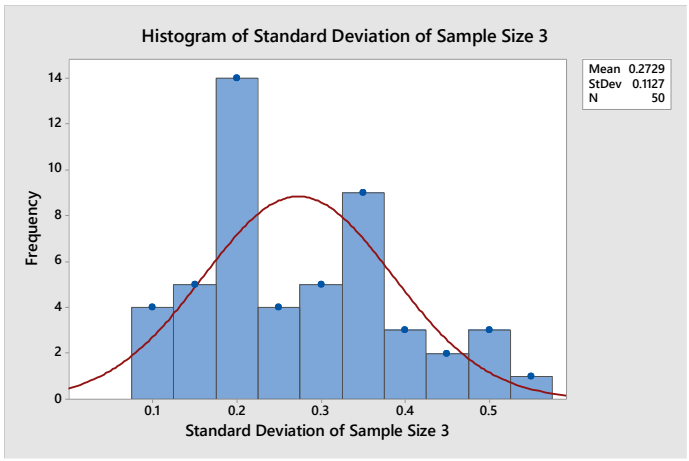


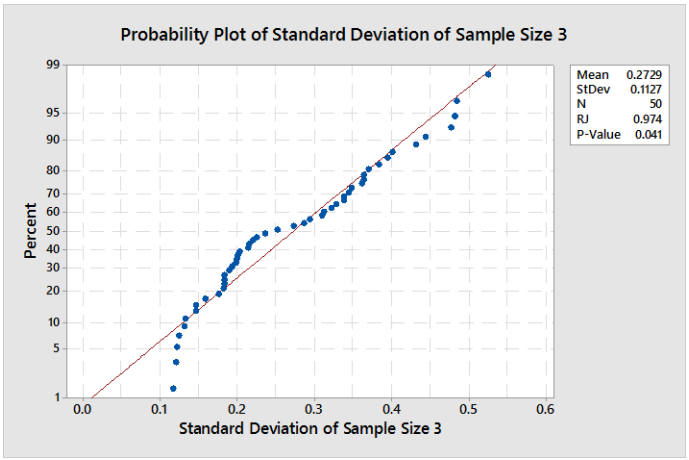


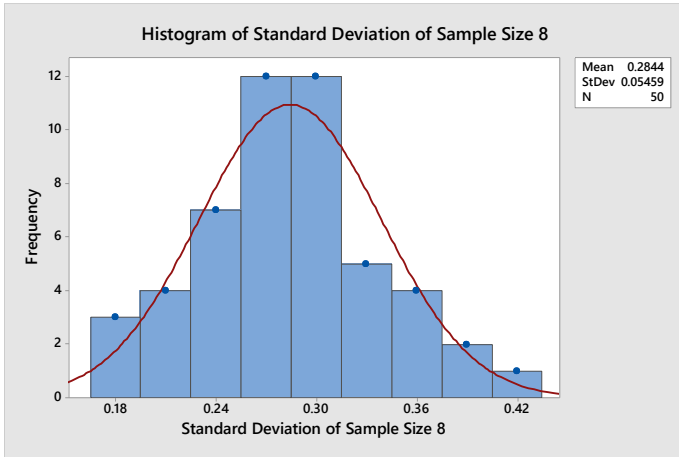
The histograms of the distribution of the statistic Standard Deviation along with the Normal density curve is given below:

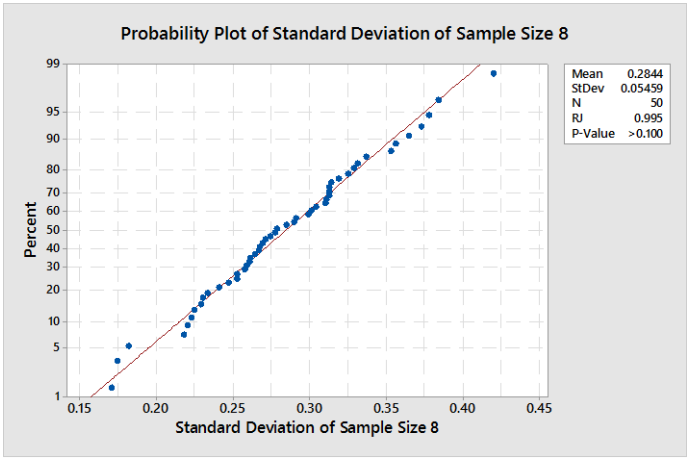




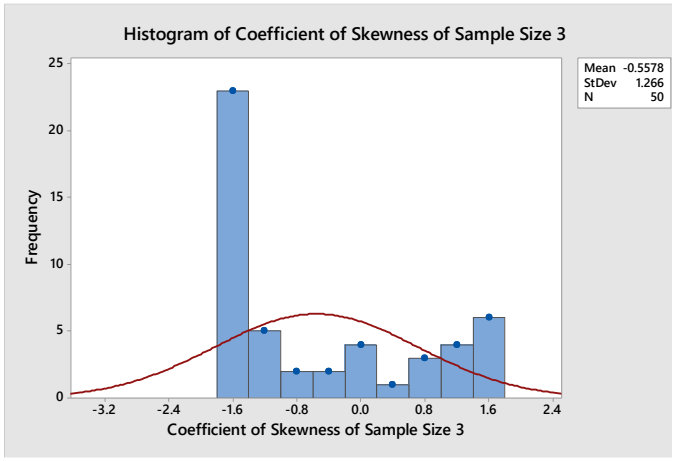


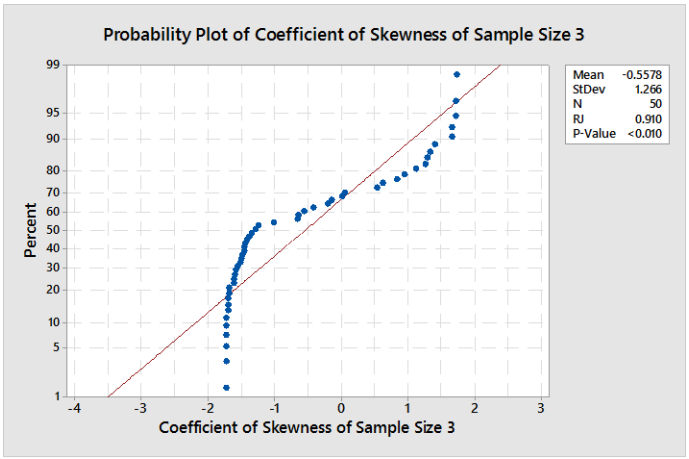


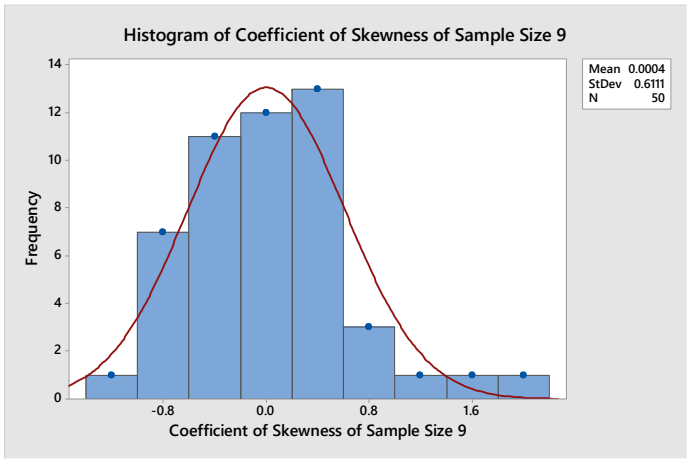


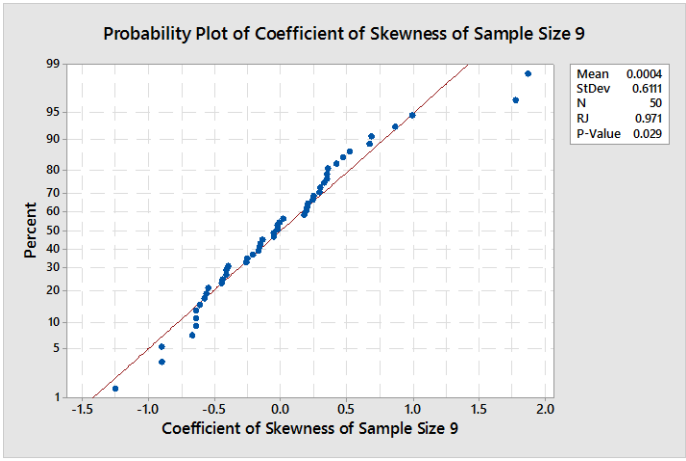


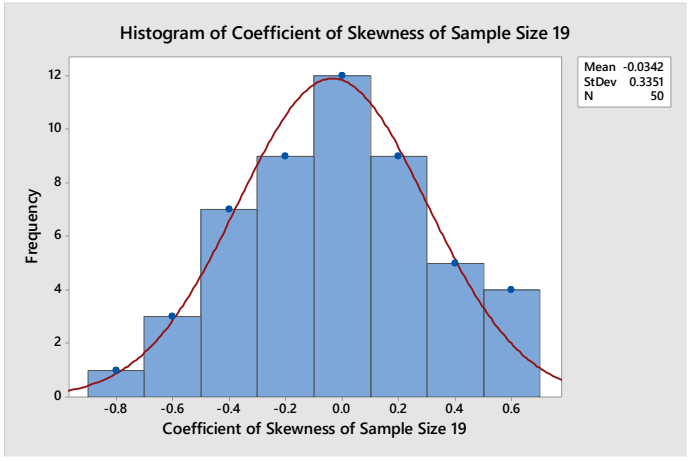
The histograms of the distribution of the statistic Coefficient of Skewness along with the Normal density curve is given below:

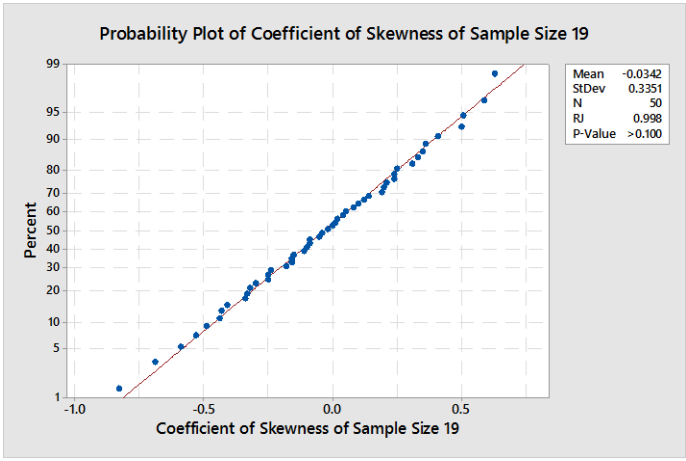




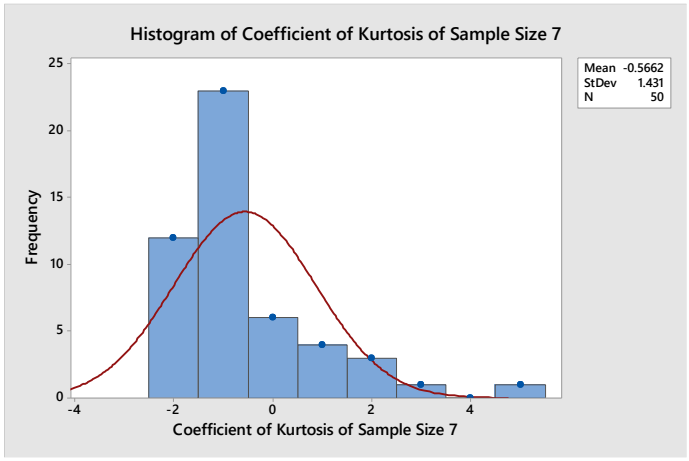


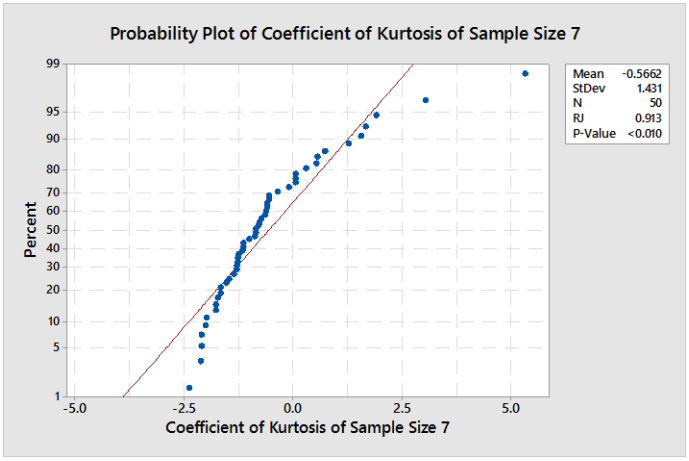


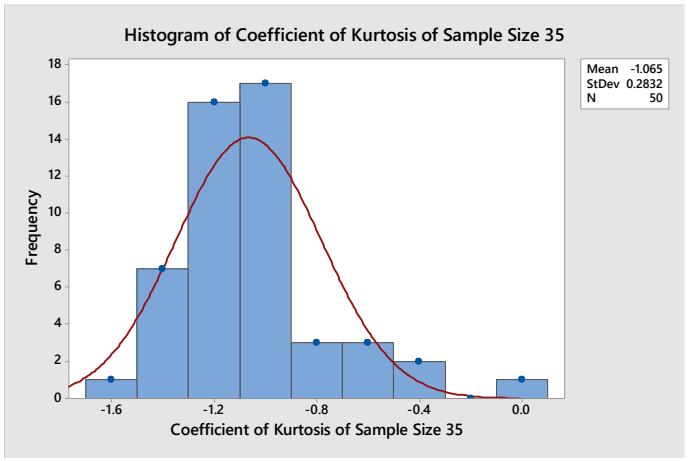


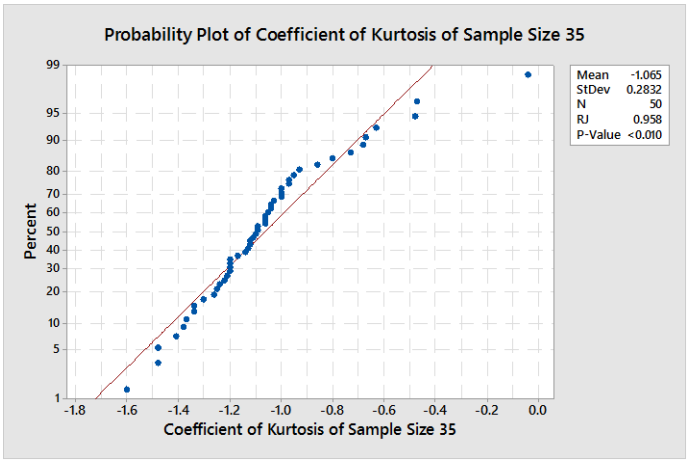


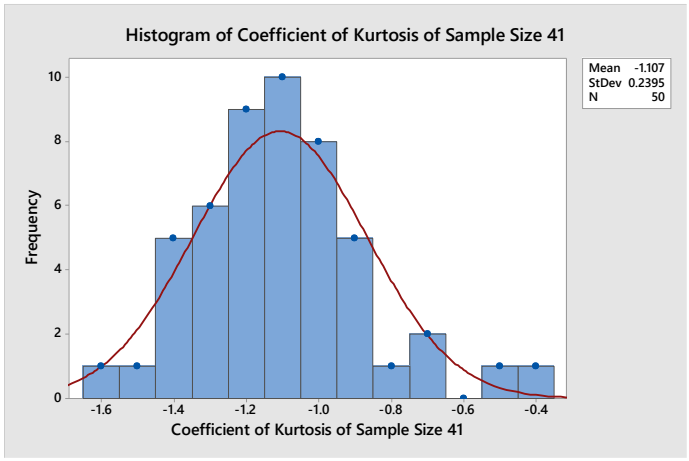
The histograms of the distribution of the statistic Coefficient of Kurtosis along with the Normal density curve is given below:

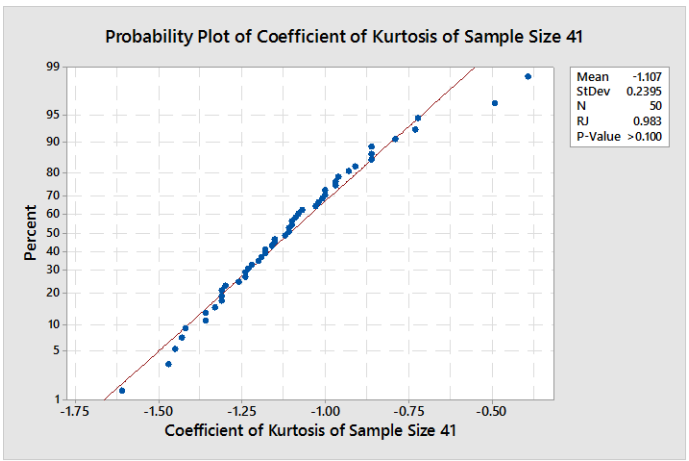






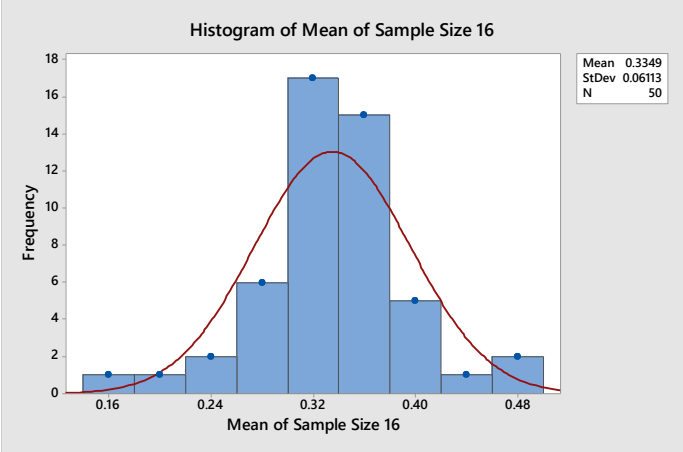


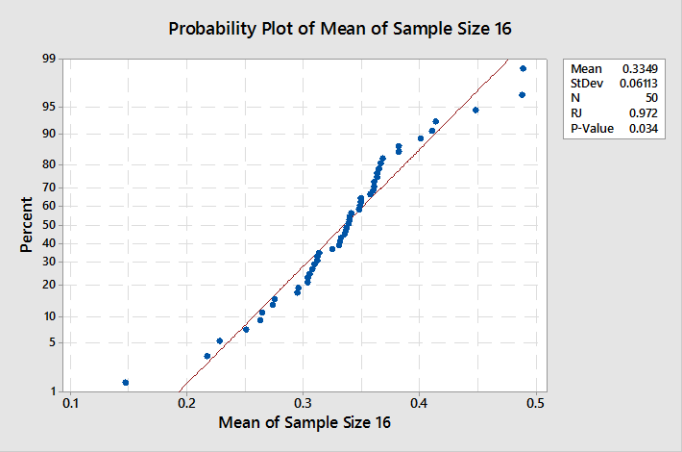


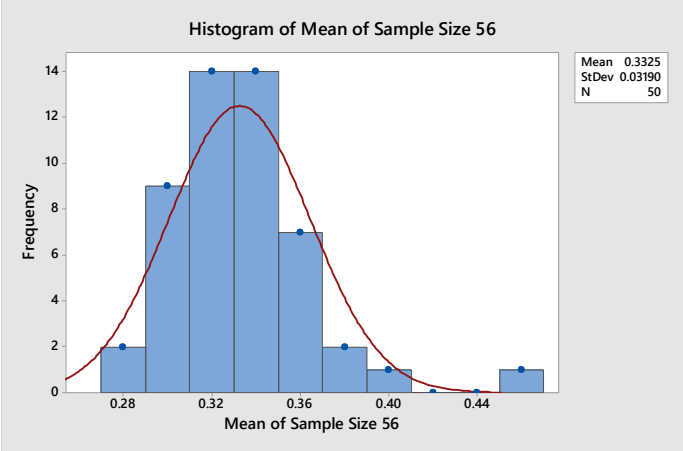


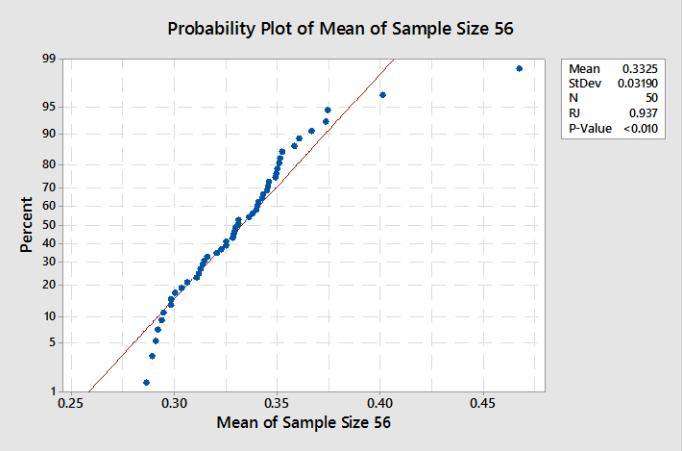
Beta(1,2)	Normality achieved at sample size	p-value	Normality not achieved till sample size	p-value
Mean	60	>0.1	16 56	0.034 <0.01
Standard Deviation	35	>0.1	2 24	<0.01 0.012
Coefficient of Skewness	80	>0.1	15 77	<0.01 0.026
Coefficient of Kurtosis	80	>0.1	15 77	<0.01 <0.01

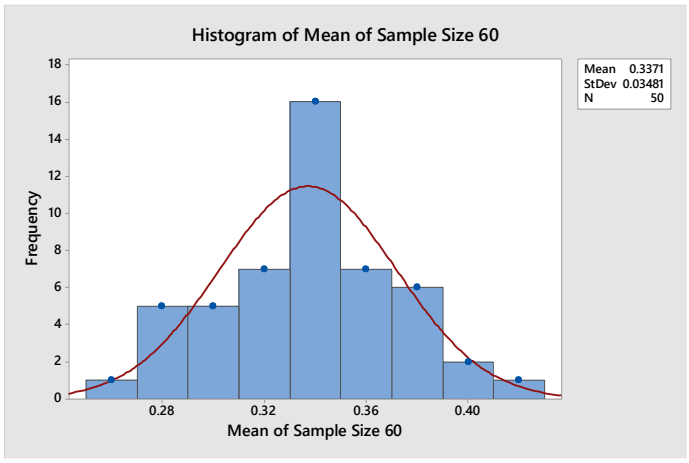
The histograms of the distribution of the statistic Mean along with the Normal density curve is given below:

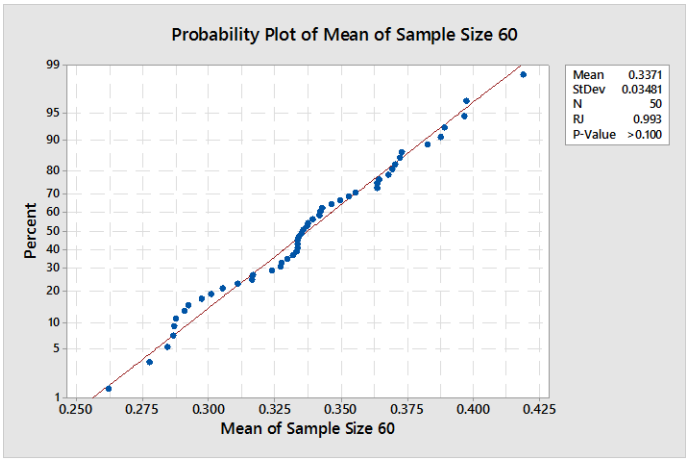




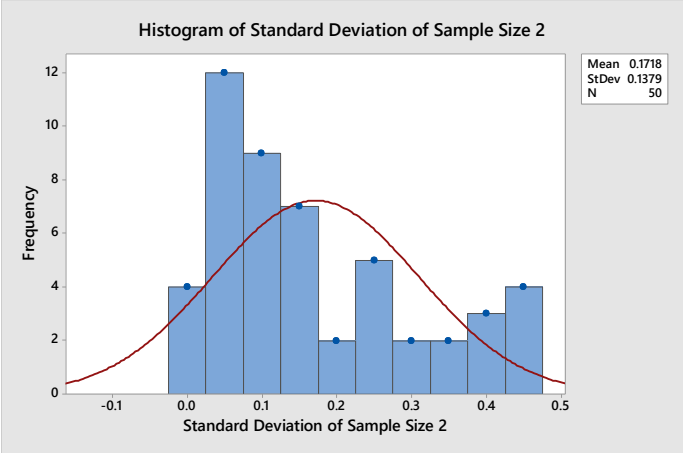


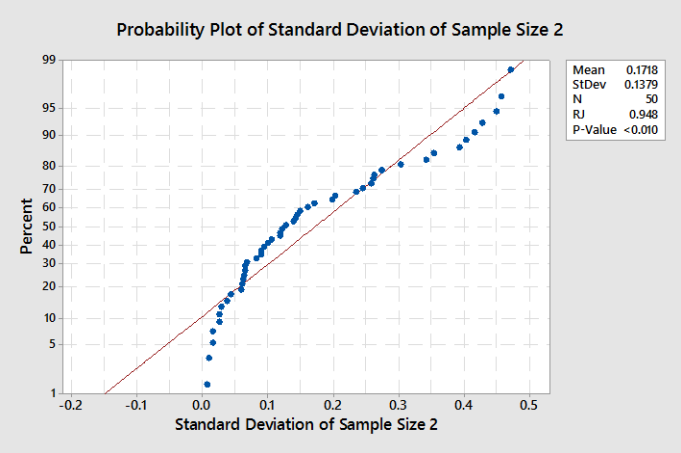


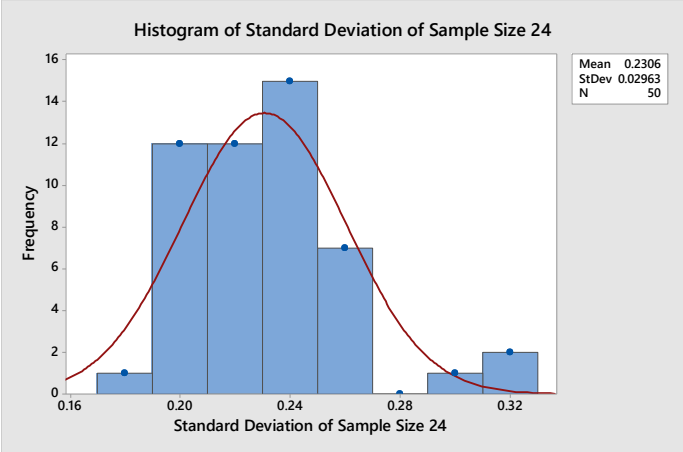

Histogram of Mean of Sample Size 60
Mean 0.3371
StDev 0.03481
N 50
Frequency
Mean of Sample Size 60
0
2
4
6
8
10
12
14
16
18
0.28
0.32
0.36
0.40

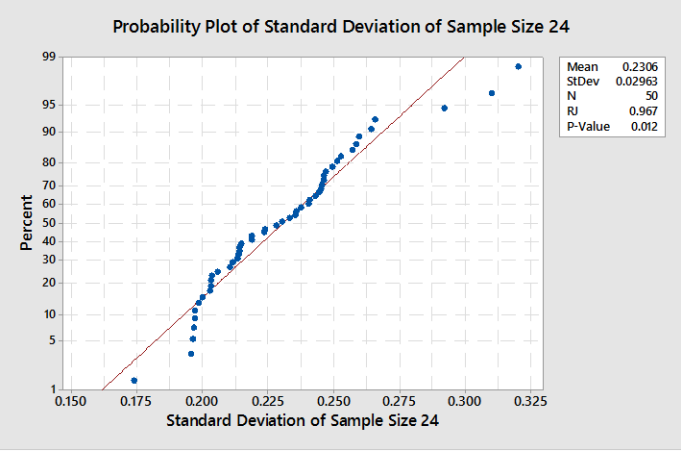

Probability Plot of Mean of Sample Size 60
Mean 0.3371
StDev 0.03481
N 50
RJ 0.993
P-Value >0.100
Percent
Mean of Sample Size 60
1
5
10
20
30
40
50
60
70
80
90
95
99
0.250
0.275
0.300
0.325
0.350
0.375
0.400
0.425

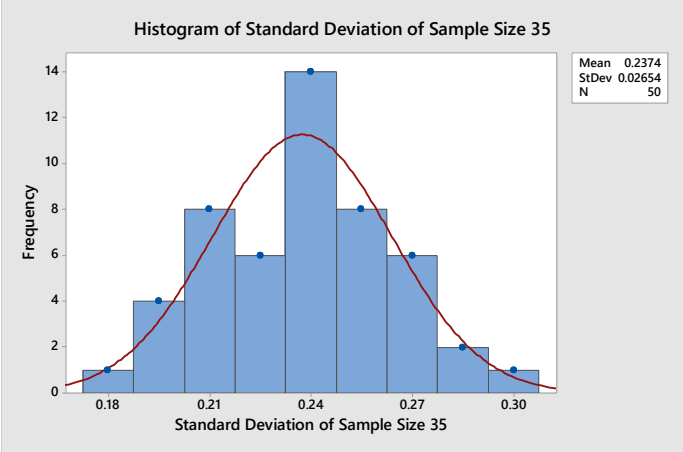
The histograms of the distribution of the statistic Standard Deviation along with the Normal density curve is given below:

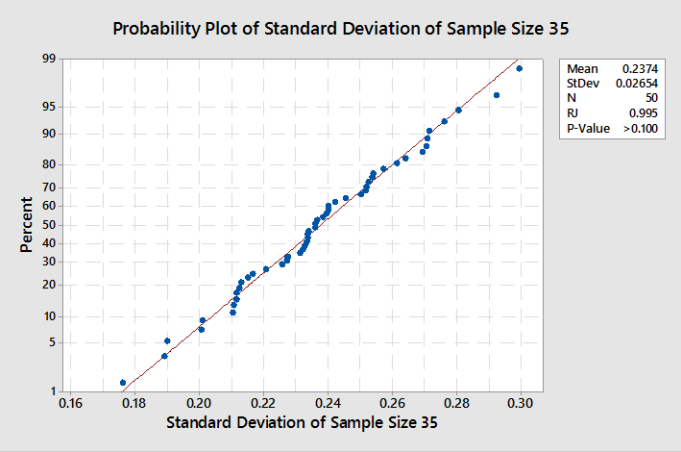




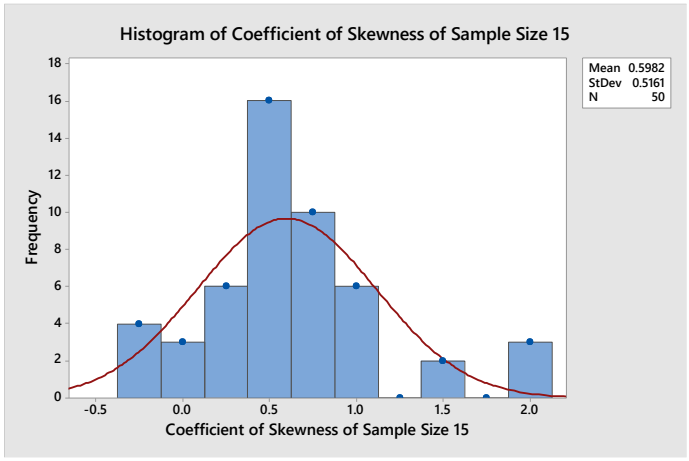


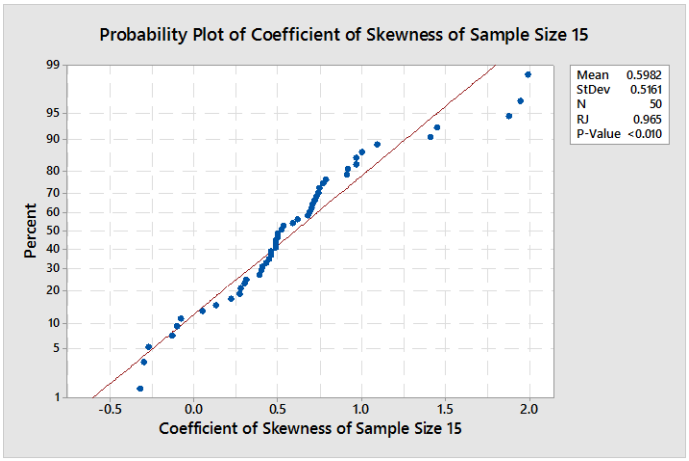


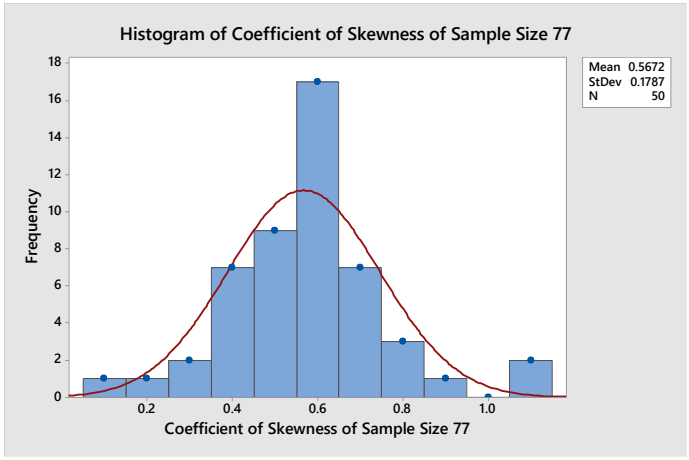


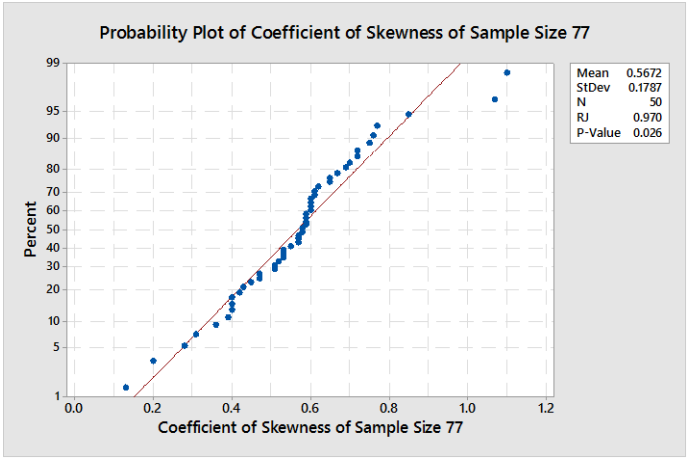


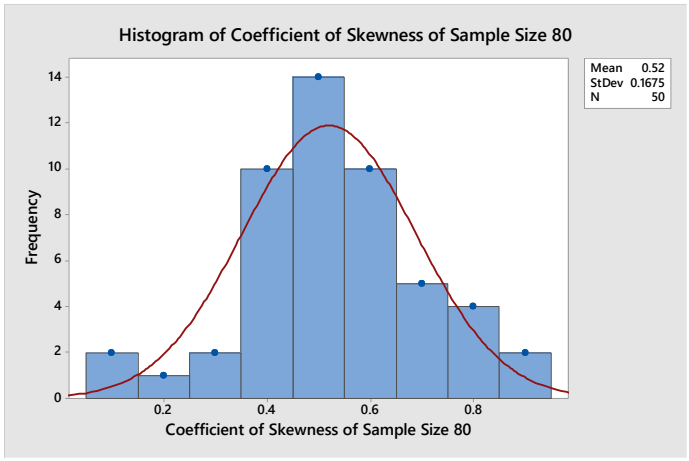
The histograms of the distribution of the statistic Coefficient of Skewness along with the Normal density curve is given below:

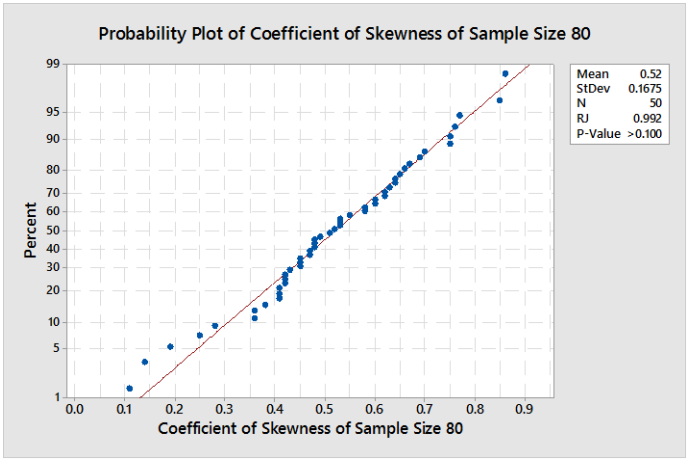




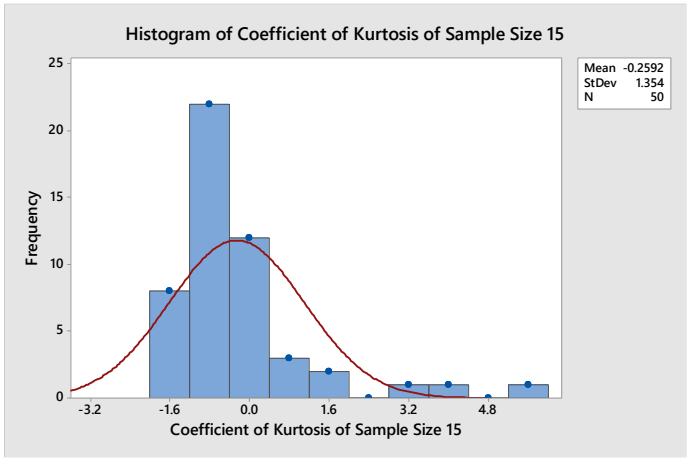


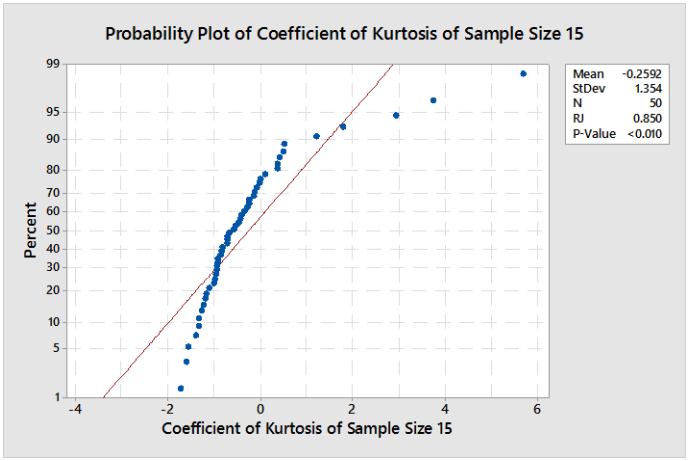


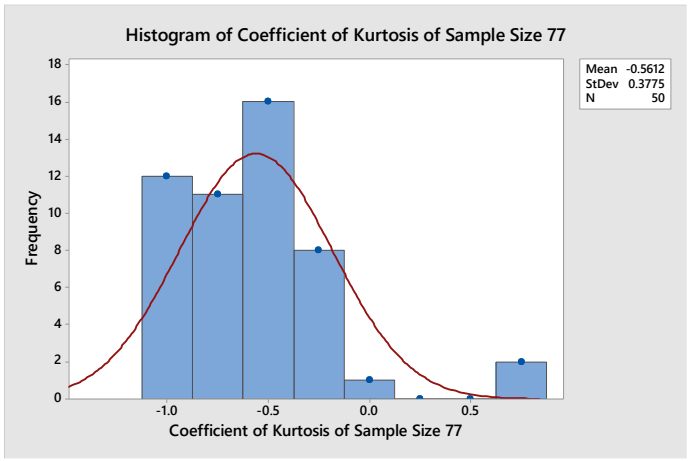


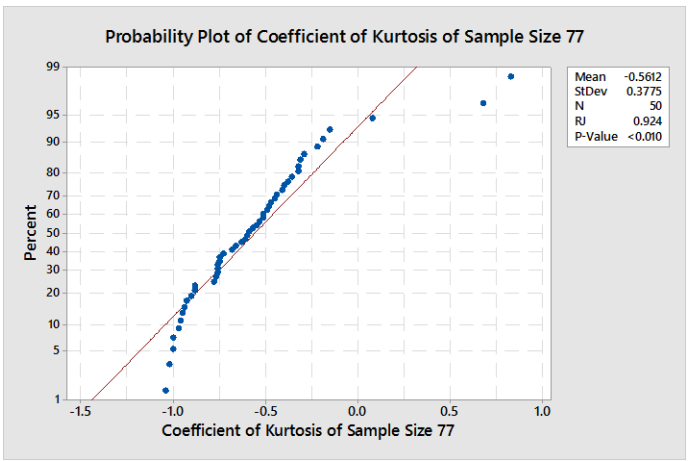


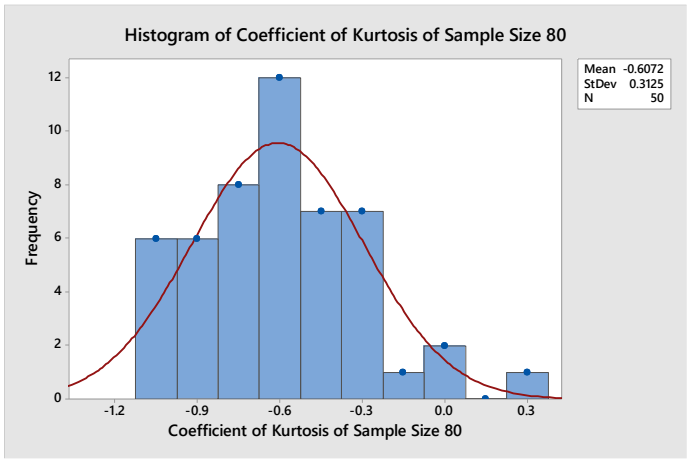
The histograms of the distribution of the statistic Coefficient of Kurtosis along with the Normal density curve is given below:

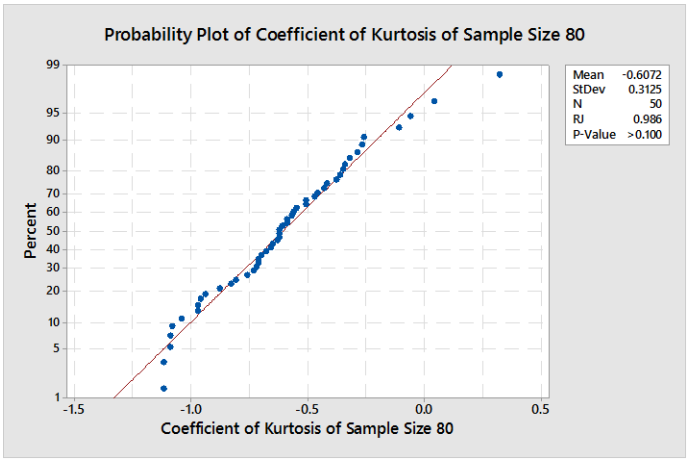






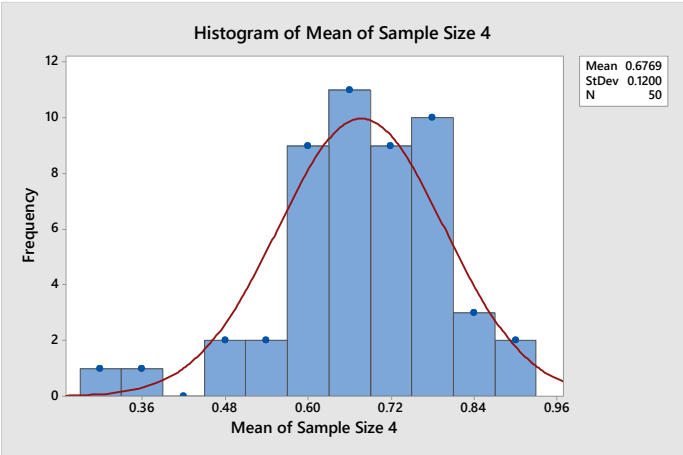


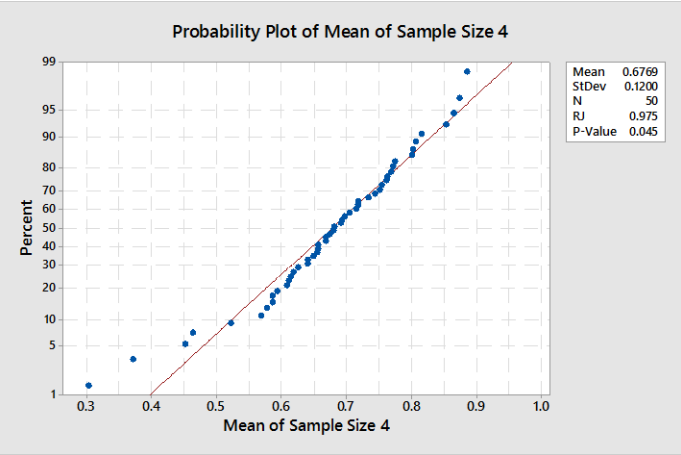


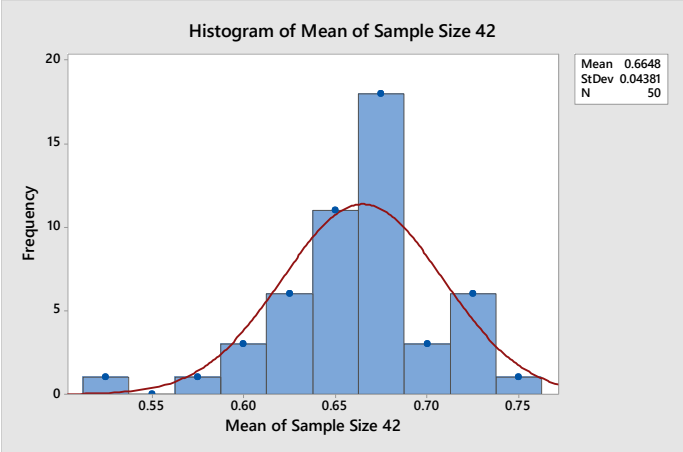


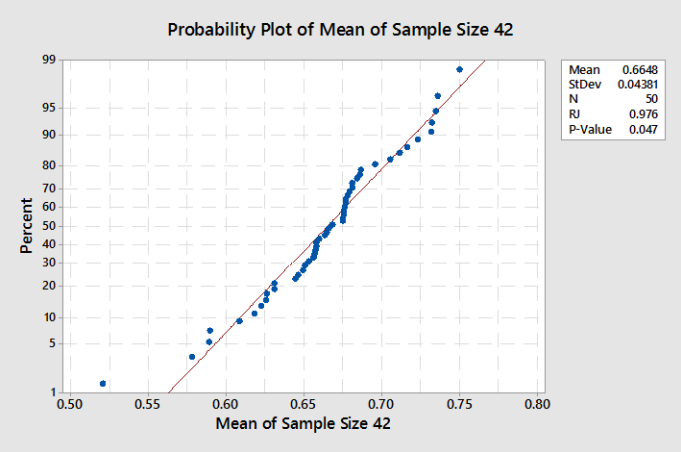
Beta(2,1)	**Normality achieved at sample size**	**p-value**	**Normality not achieved till sample size**	**p-value**
Mean	70	>0.1	4 42	0.045 0.047
Standard Deviation	20	>0.1	2 19	<0.01 0.015
Coefficient of Skewness	10	>0.1	3 4	<0.01 0.029
Coefficient of Kurtosis	70	>0.1	10 60	<0.01 <0.01

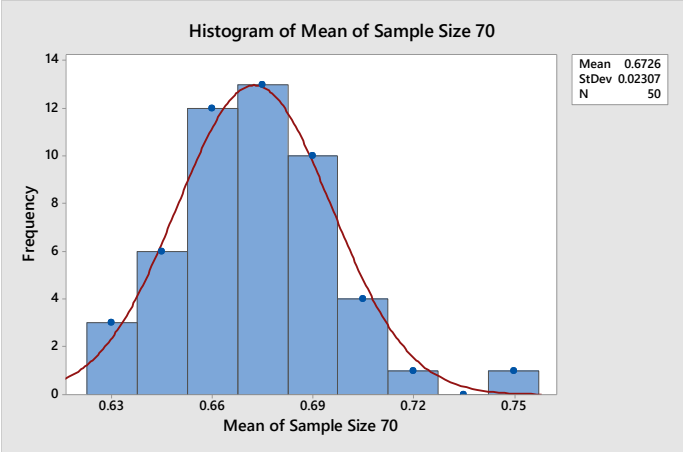
The histograms of the distribution of the statistic Mean along with the Normal density curve is given below:

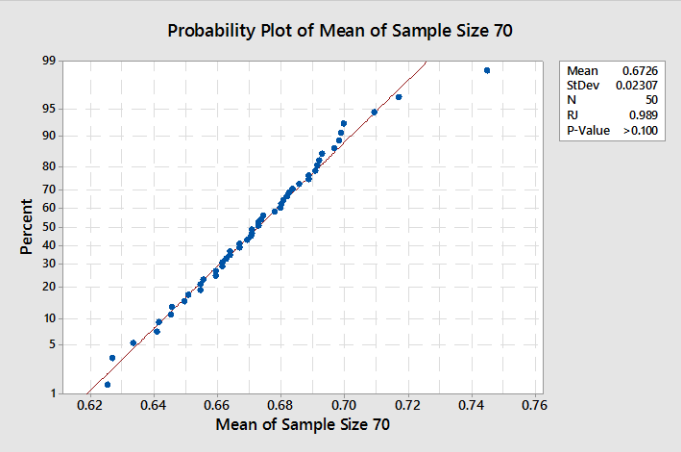




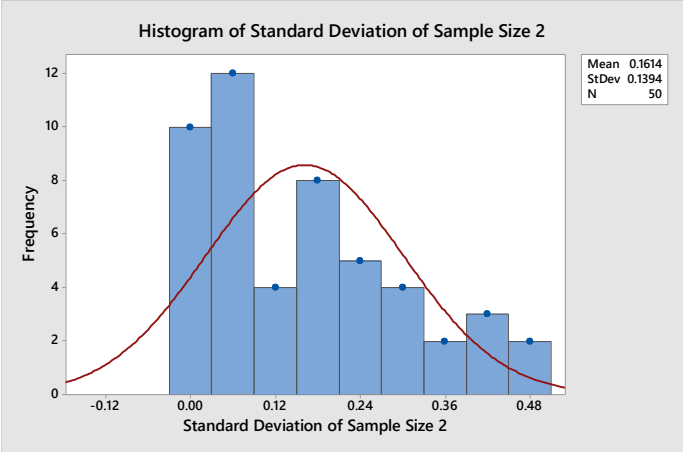


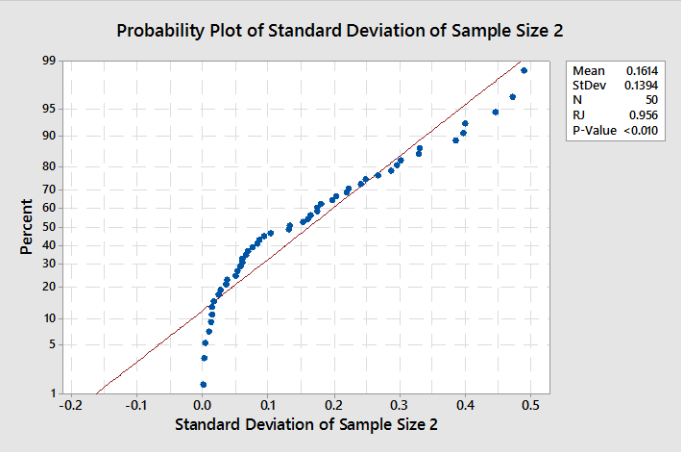


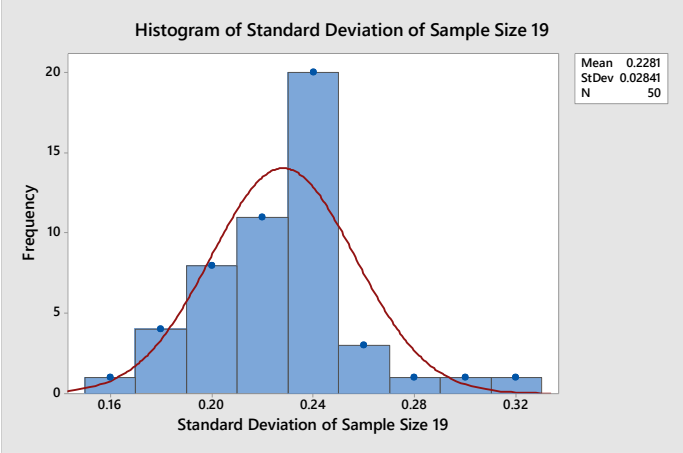


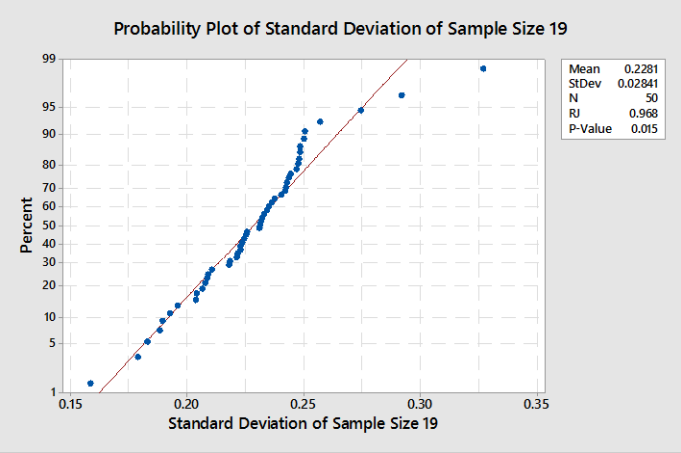


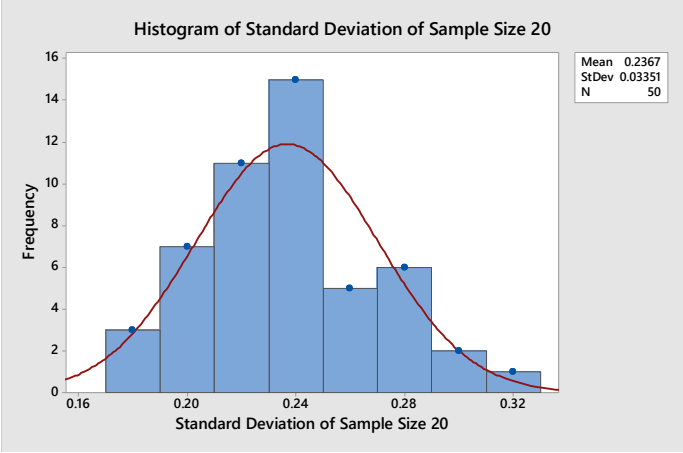
The histograms of the distribution of the statistic Standard Deviation along with the Normal density curve is given below:

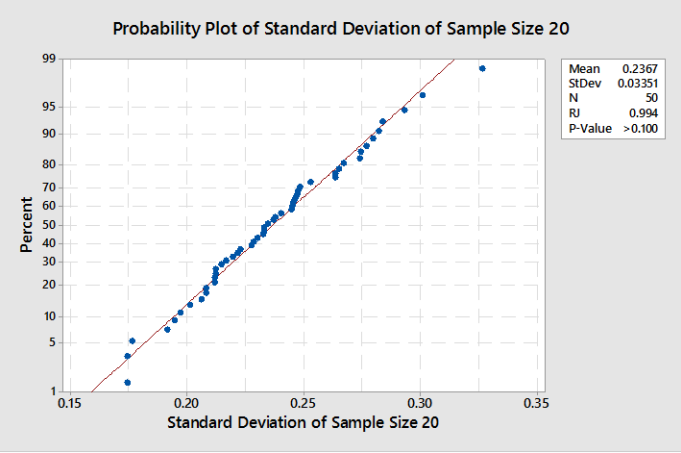




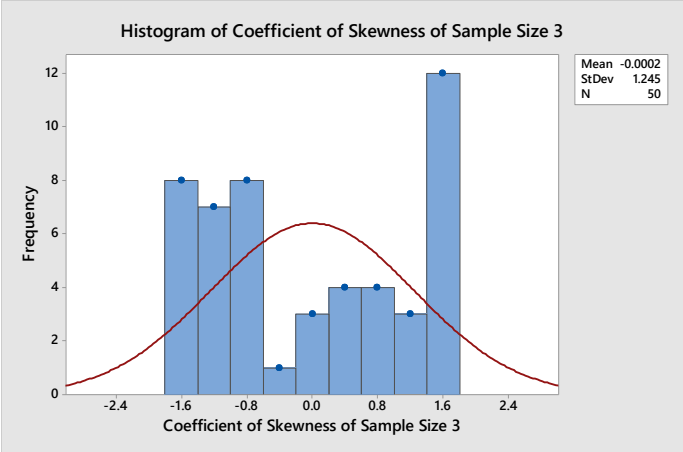


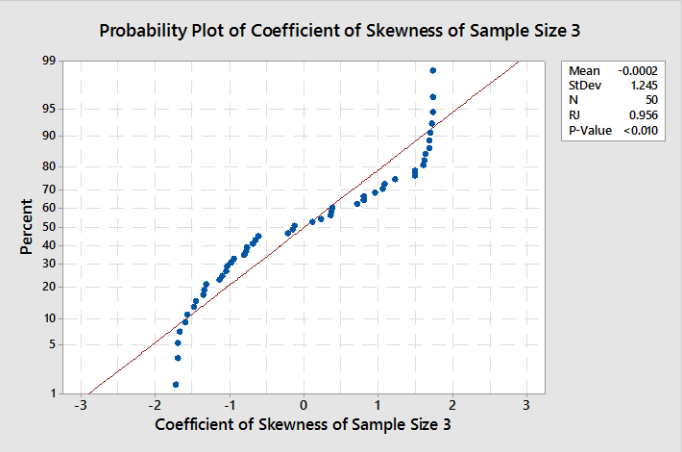


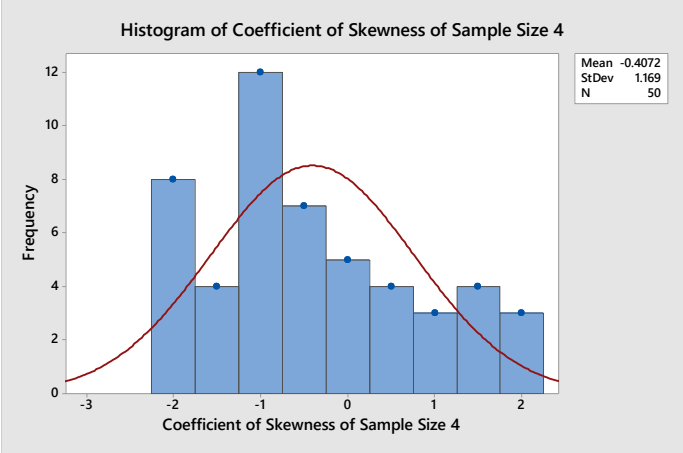


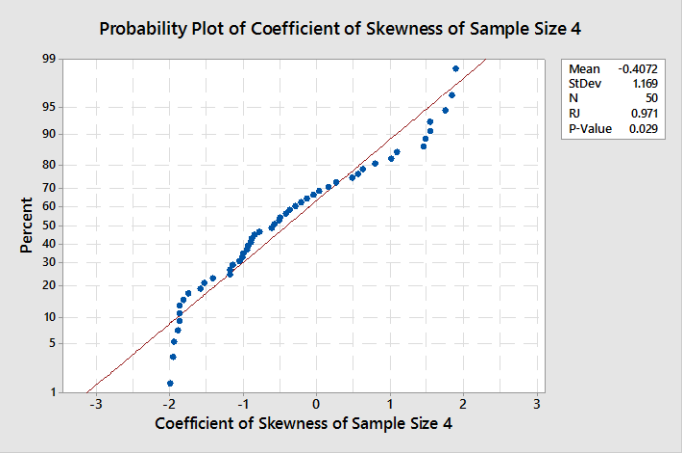


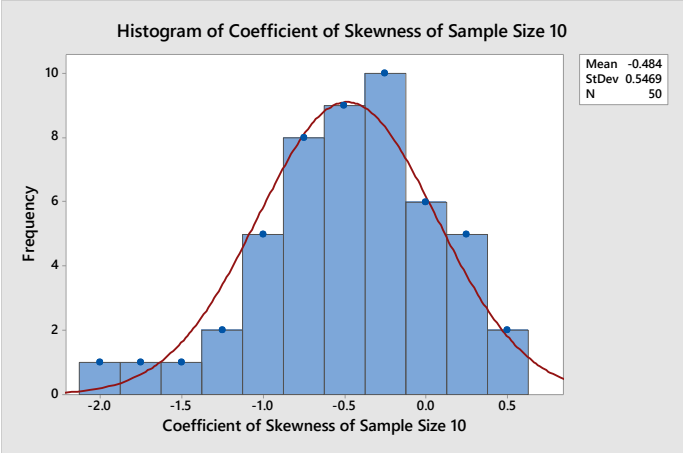
The histograms of the distribution of the statistic Coefficient of Skewness along with the Normal density curve is given below:

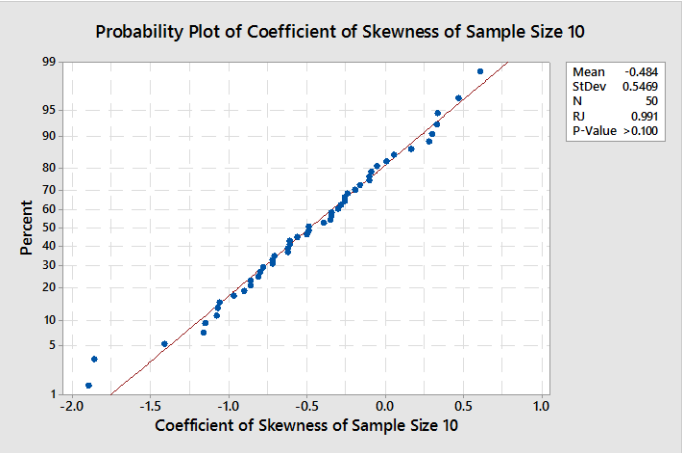




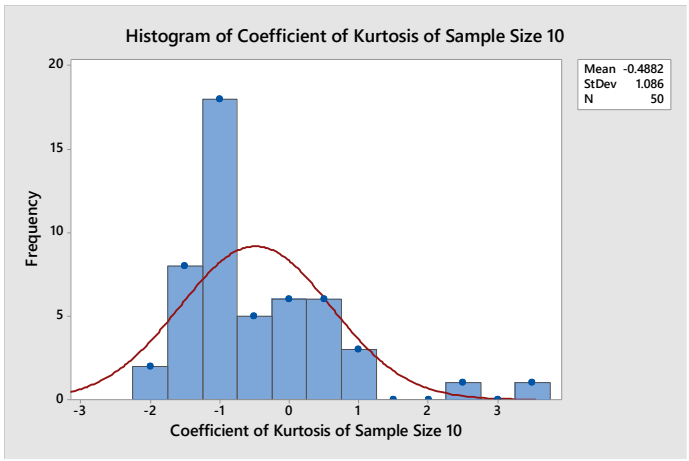


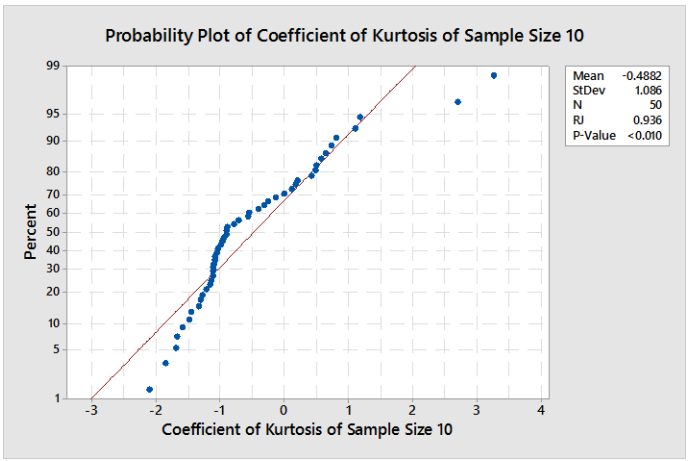


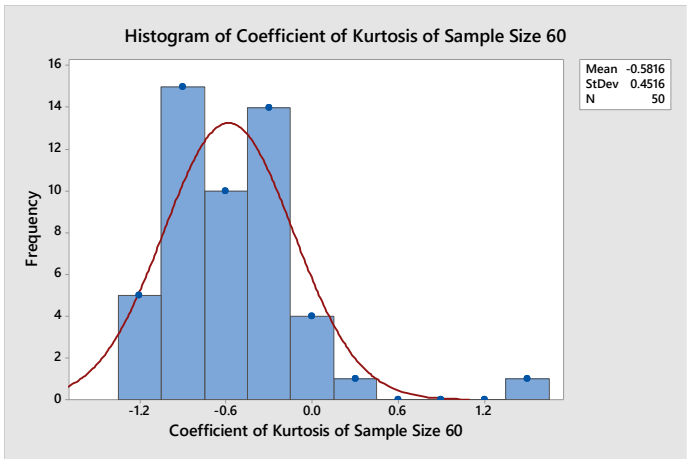


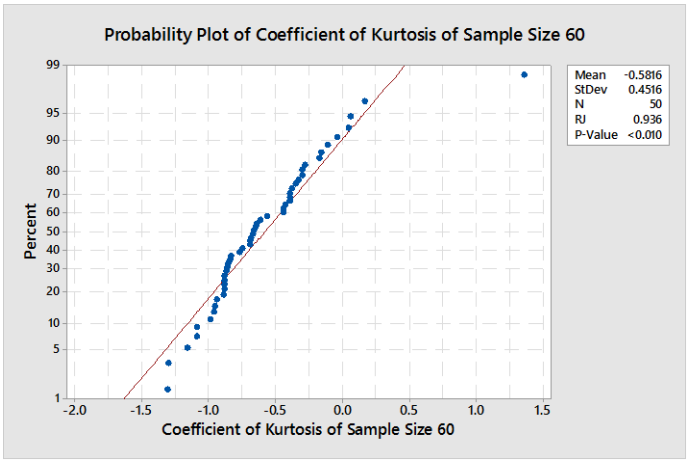


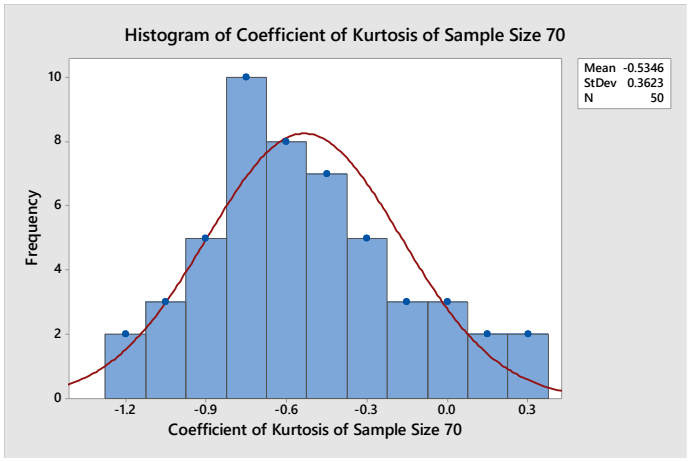
The histograms of the distribution of the statistic Coefficient of Kurtosis along with the Normal density curve is given below:

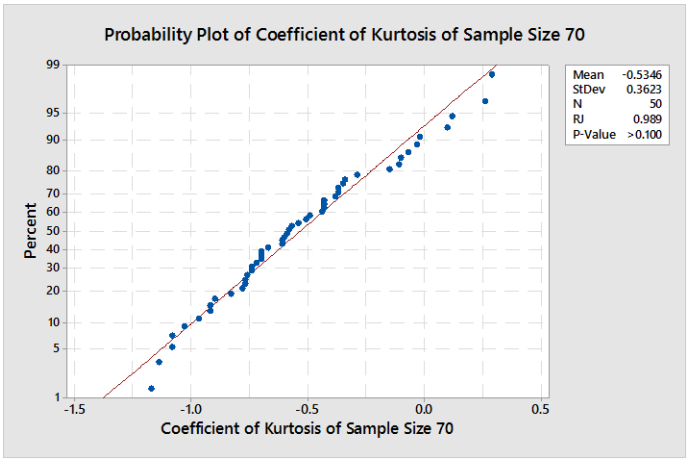

Histogram of Coefficient of Kurtosis of Sample Size 10
Mean -0.4882
StDev 1.086
N 50
Frequency
Coefficient of Kurtosis of Sample Size 10


Probability Plot of Coefficient of Kurtosis of Sample Size 10
Mean -0.4882
StDev 1.086
N 50
RJ 0.936
P-Value <0.010
Percent
Coefficient of Kurtosis of Sample Size 10


Histogram of Coefficient of Kurtosis of Sample Size 60
Mean -0.5816
StDev 0.4516
N 50
Frequency
Coefficient of Kurtosis of Sample Size 60

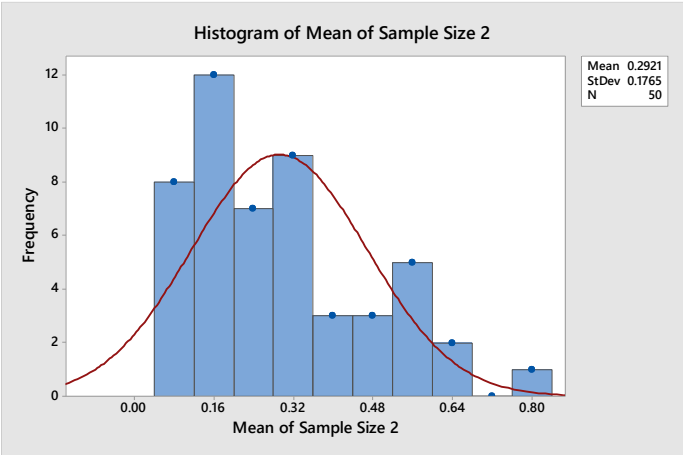

Probability Plot of Coefficient of Kurtosis of Sample Size 60
Mean -0.5816
StDev 0.4516
N 50
RJ 0.936
P-Value <0.010
Percent
Coefficient of Kurtosis of Sample Size 60

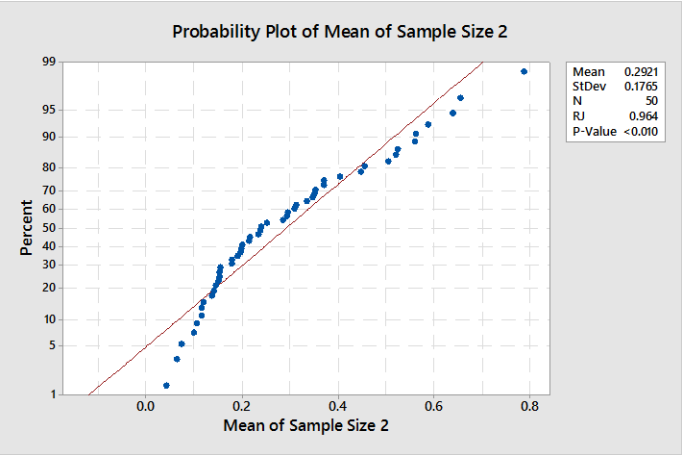

Histogram of Coefficient of Kurtosis of Sample Size 70
Mean -0.5346
StDev 0.3623
N 50
Frequency
Coefficient of Kurtosis of Sample Size 70

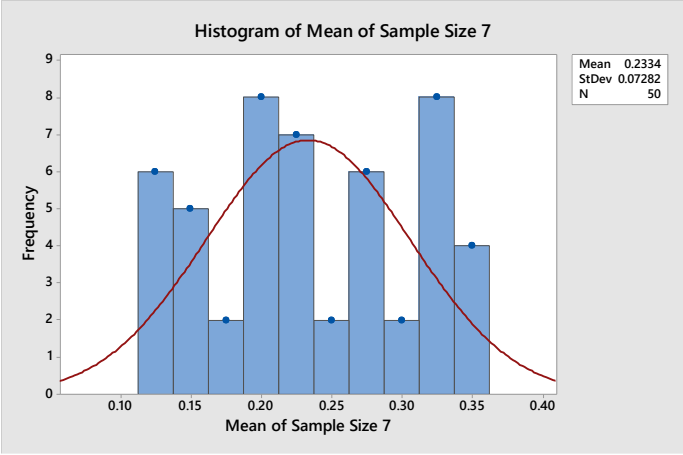

Probability Plot of Coefficient of Kurtosis of Sample Size 70
Mean -0.5346
StDev 0.3623
N 50
RJ 0.989
P-Value >0.100
Percent
Coefficient of Kurtosis of Sample Size 70

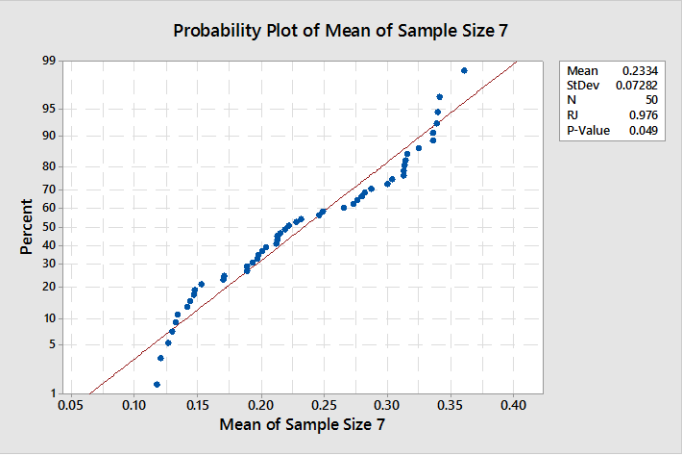
Beta(1,3)	**Normality achieved at sample size**	**p-value**	**Normality not achieved till sample size**	**p-value**
Mean	30	>0.1	2 7	<0.01 0.049
Standard Deviation	16	>0.1	2	<0.01
Coefficient of Skewness	22	>0.1	3 6	<0.01 <0.01
Coefficient of Kurtosis	132	0.097	10 130	<0.01 <0.01

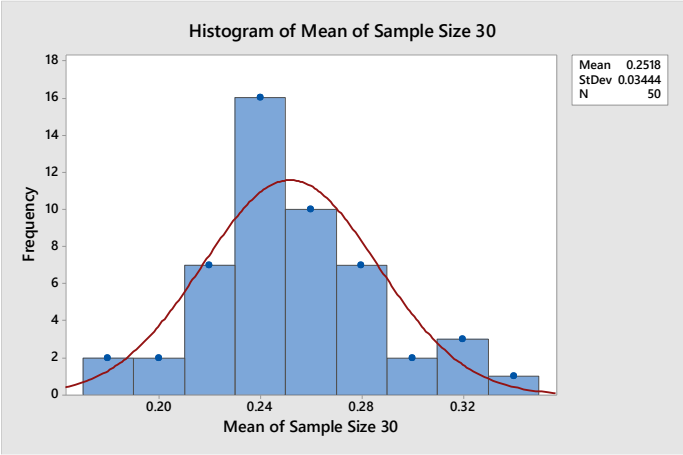
The histograms of the distribution of the statistic Mean along with the Normal density curve is given below:

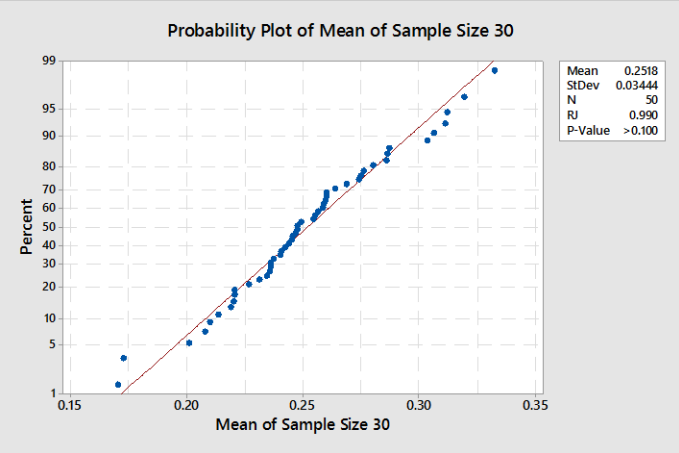




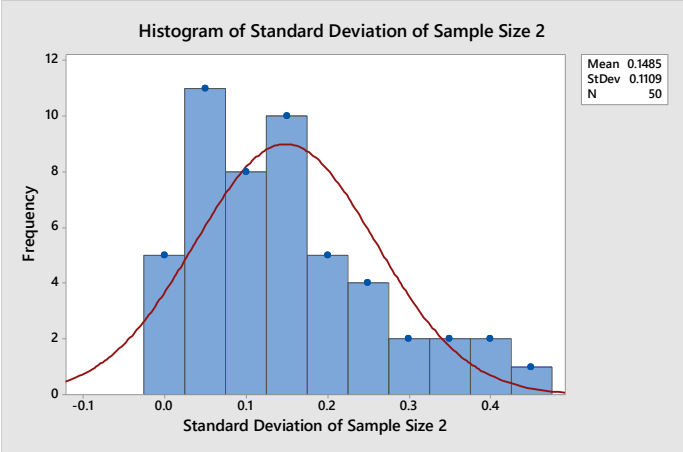


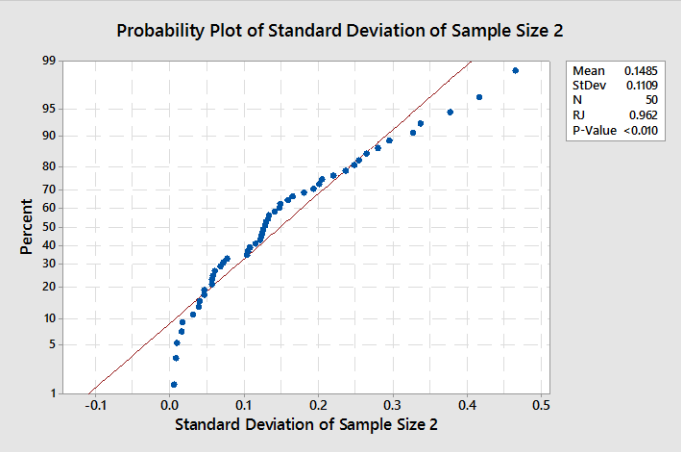


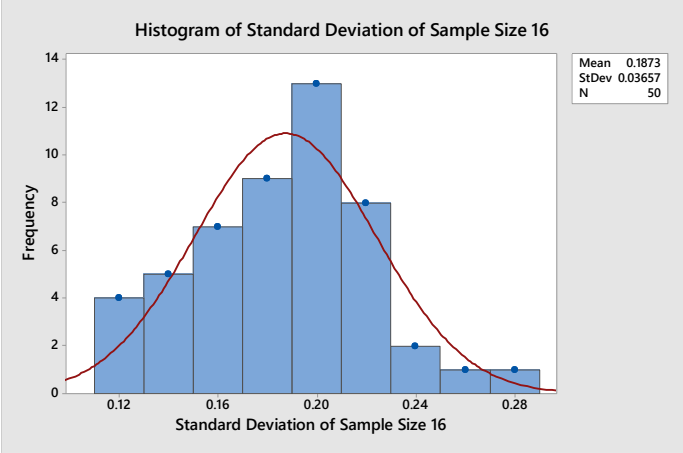


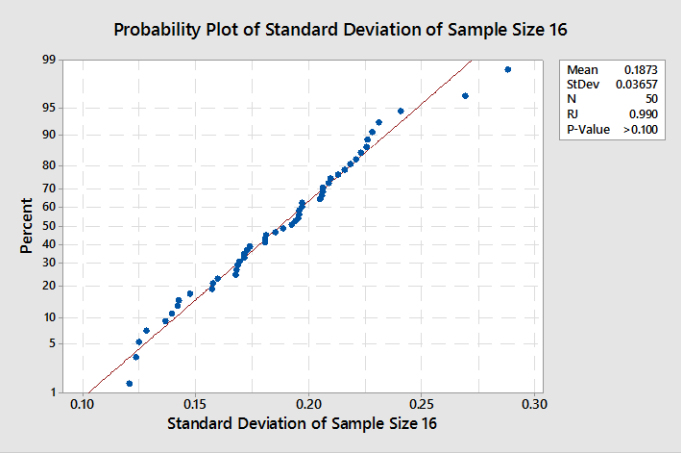


The histograms of the distribution of the statistic Standard Deviation along with the Normal density curve is given below:

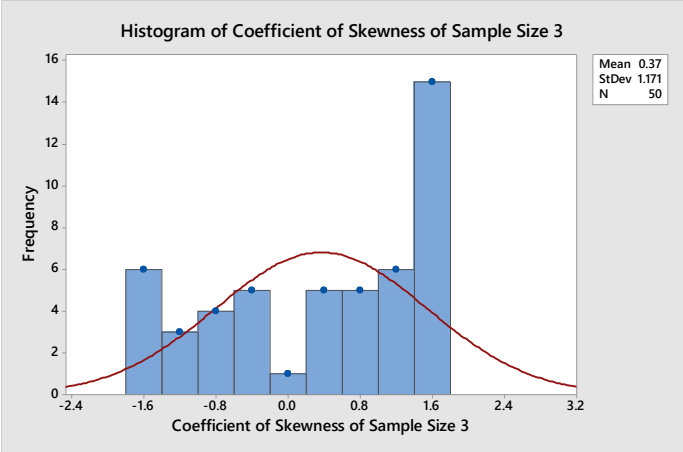


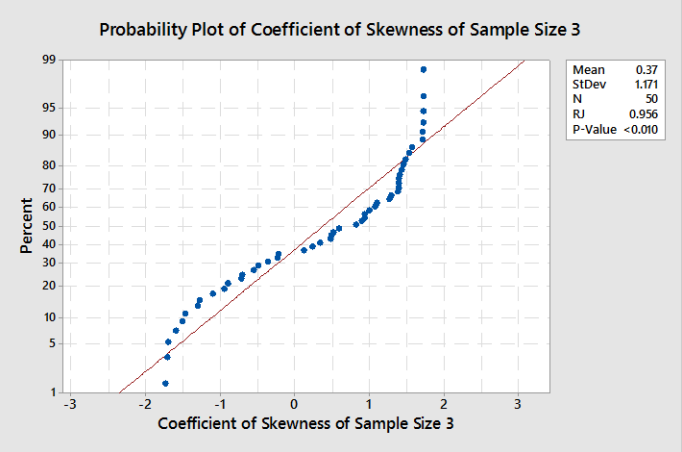


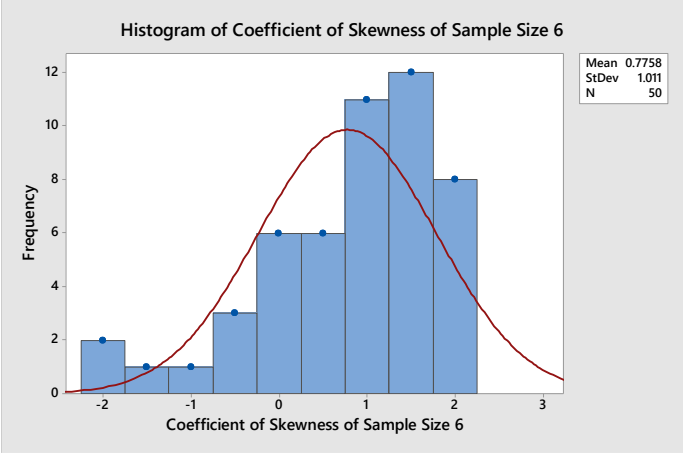


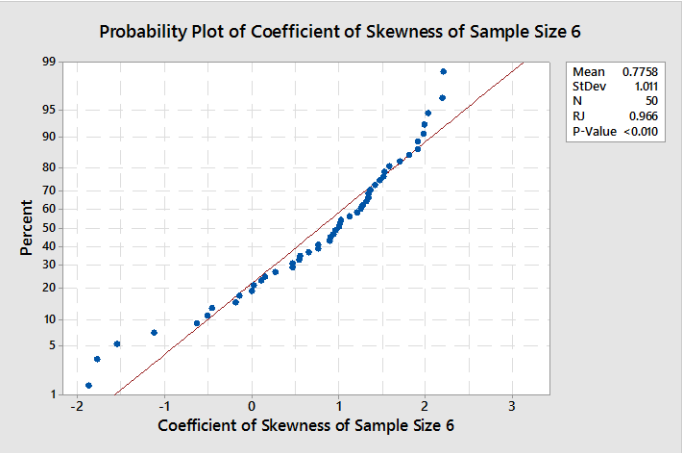


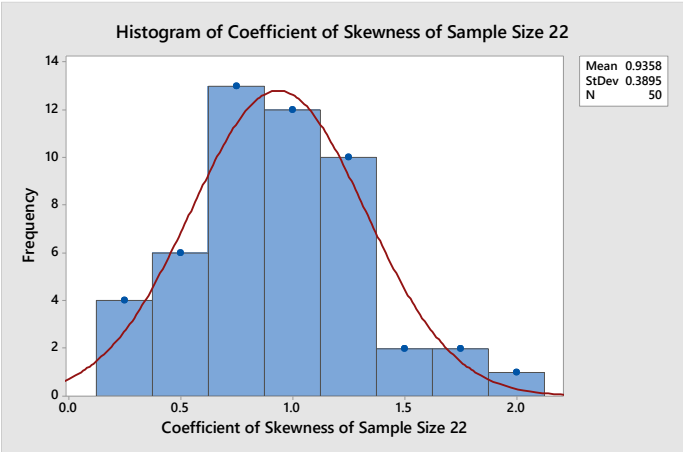
The histograms of the distribution of the statistic Coefficient of Skewness along with the Normal density curve is given below:

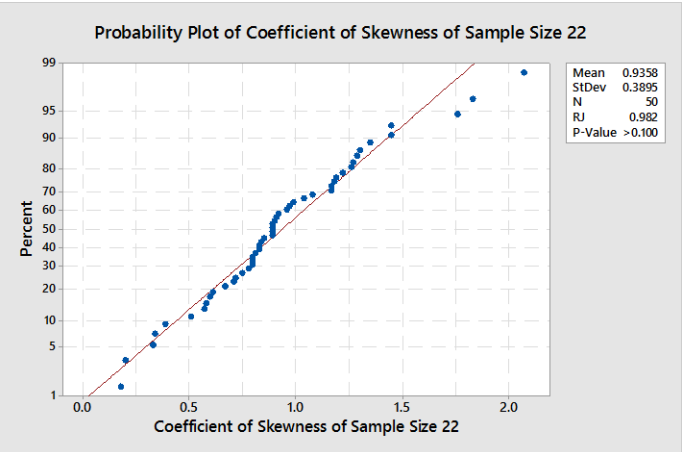




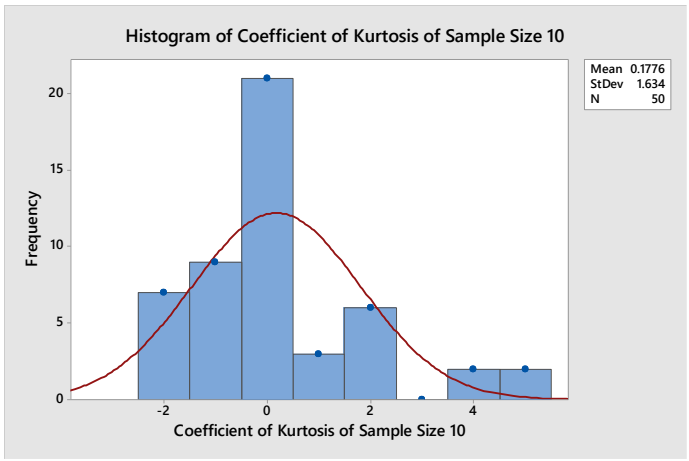


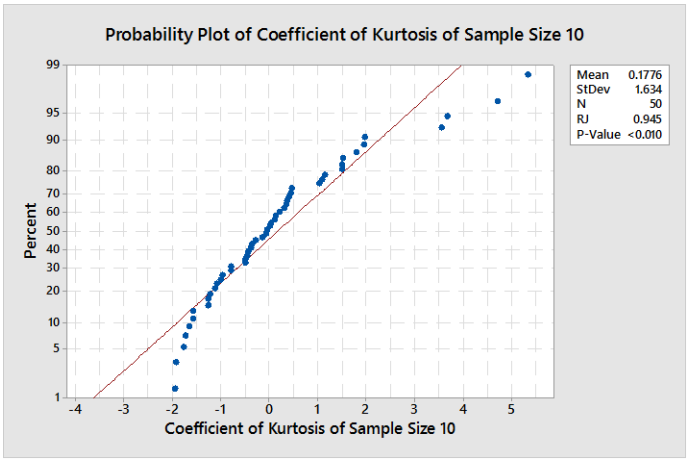


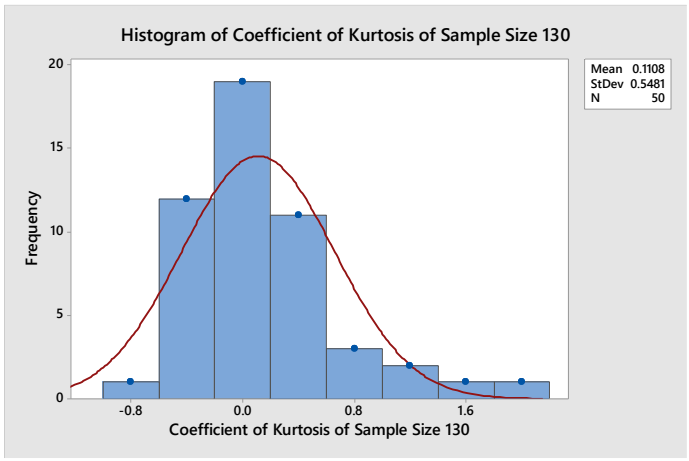


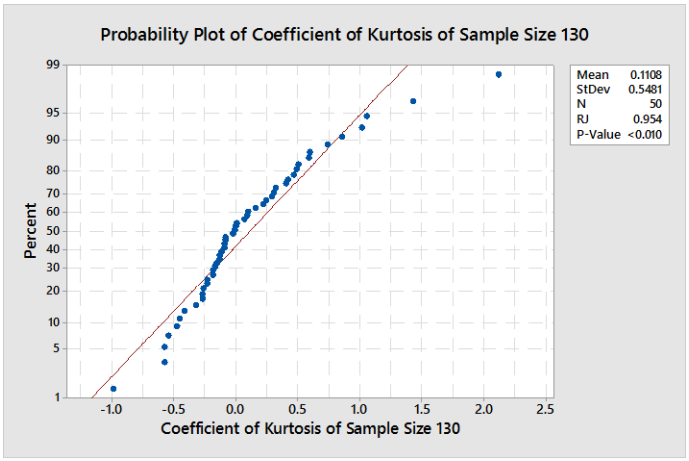


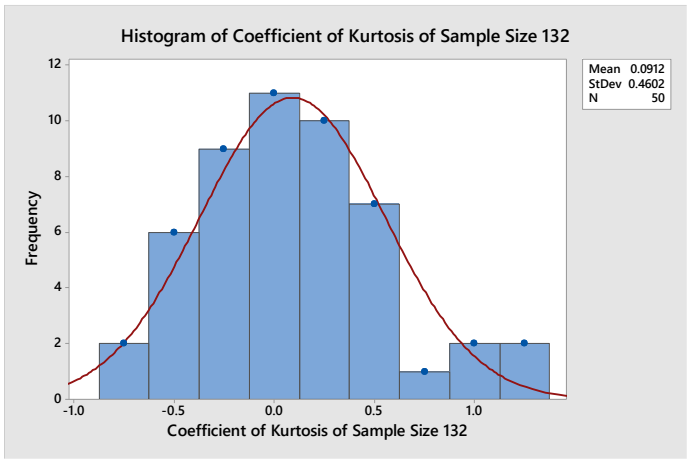
The histograms of the distribution of the statistic Coefficient of Kurtosis along with the Normal density curve is given below:

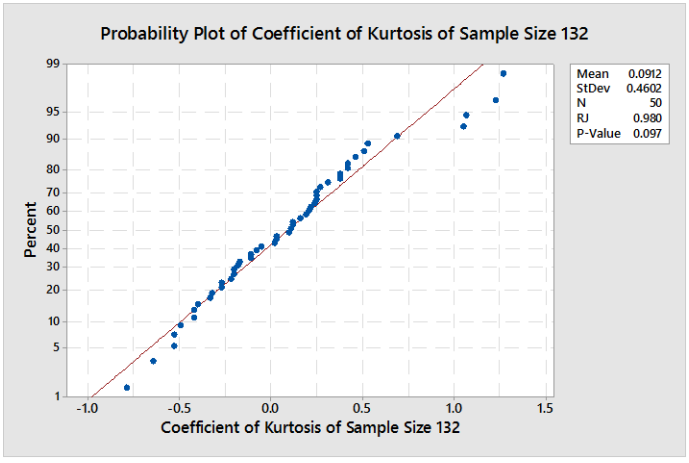

Histogram of Coefficient of Kurtosis of Sample Size 10
Mean 0.1776
StDev 1.634
N 50
Frequency
Coefficient of Kurtosis of Sample Size 10


Probability Plot of Coefficient of Kurtosis of Sample Size 10
Mean 0.1776
StDev 1.634
N 50
RJ 0.945
P-Value <0.010
Percent
Coefficient of Kurtosis of Sample Size 10


Histogram of Coefficient of Kurtosis of Sample Size 130
Mean 0.1108
StDev 0.5481
N 50
Frequency
Coefficient of Kurtosis of Sample Size 130

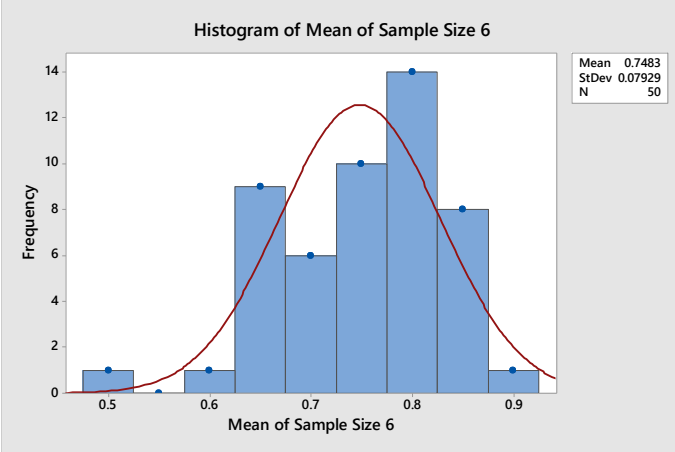

Probability Plot of Coefficient of Kurtosis of Sample Size 130
Mean 0.1108
StDev 0.5481
N 50
RJ 0.954
P-Value <0.010
Percent
Coefficient of Kurtosis of Sample Size 130

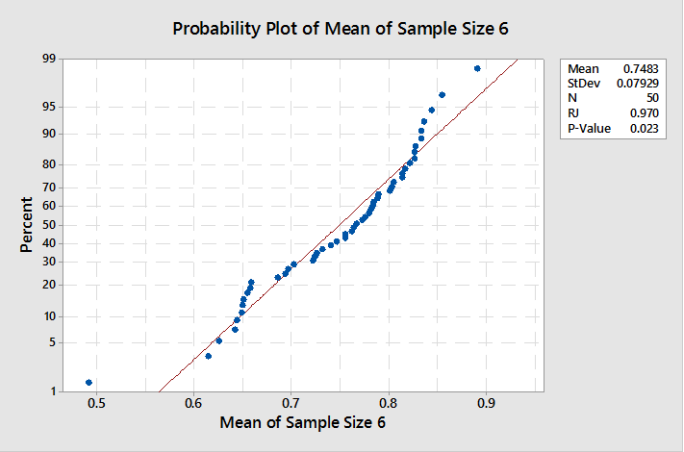

Histogram of Coefficient of Kurtosis of Sample Size 132
Mean 0.0912
StDev 0.4602
N 50
Frequency
Coefficient of Kurtosis of Sample Size 132

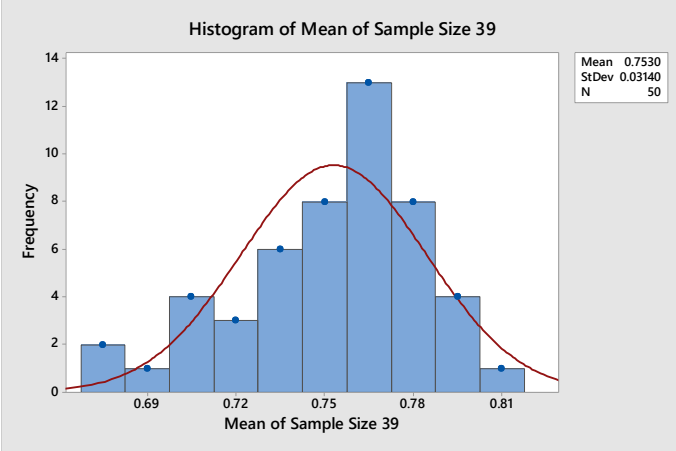

Probability Plot of Coefficient of Kurtosis of Sample Size 132
Mean 0.0912
StDev 0.4602
N 50
RJ 0.980
P-Value 0.097
Percent
Coefficient of Kurtosis of Sample Size 132

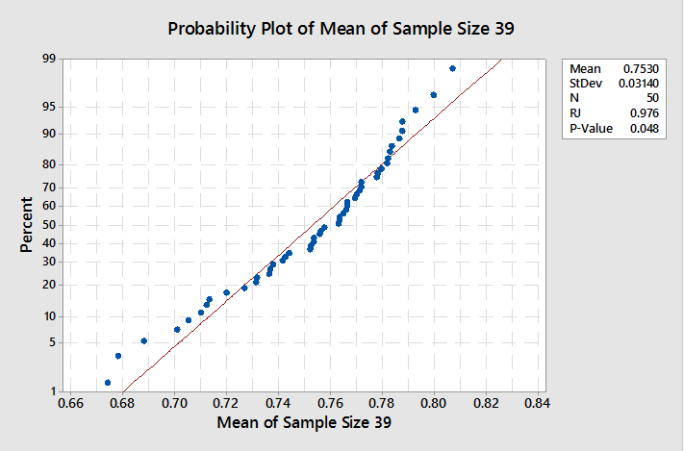
Beta(3,1)	**Normality achieved at sample size**	**p-value**	**Normality not achieved till sample size**	**p-value**
Mean	40	>0.1	6 39	0.023 0.048
Standard Deviation	20	>0.1	2	0.012
Coefficient of Skewness	40	0.066	11 38	0.036 0.045
Coefficient of Kurtosis	100	>0.1	10 95	<0.01 <0.01

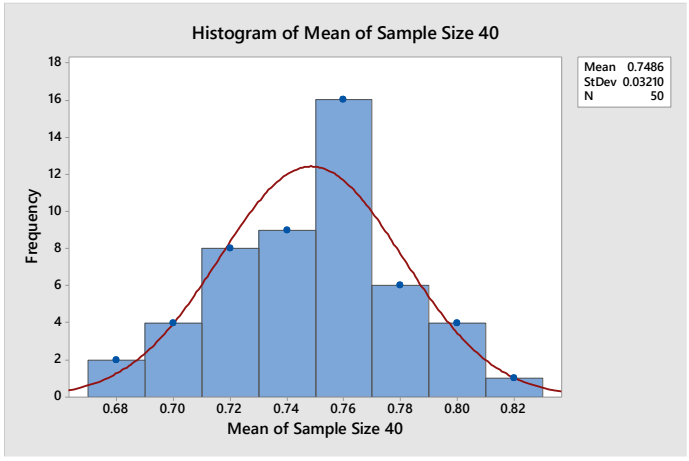
The histograms of the distribution of the statistic Mean along with the Normal density curve is given below:

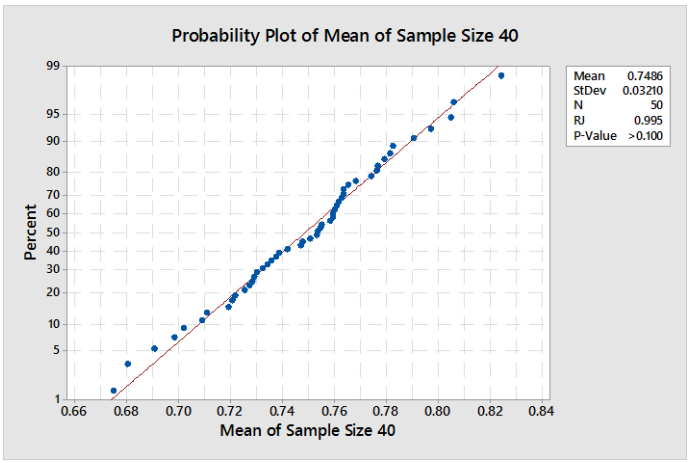




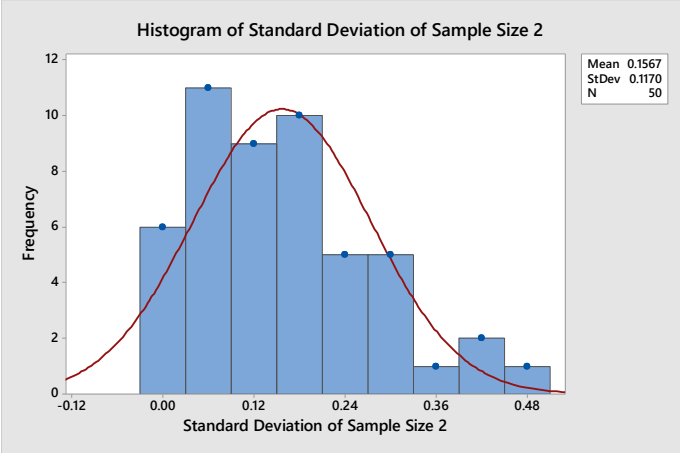


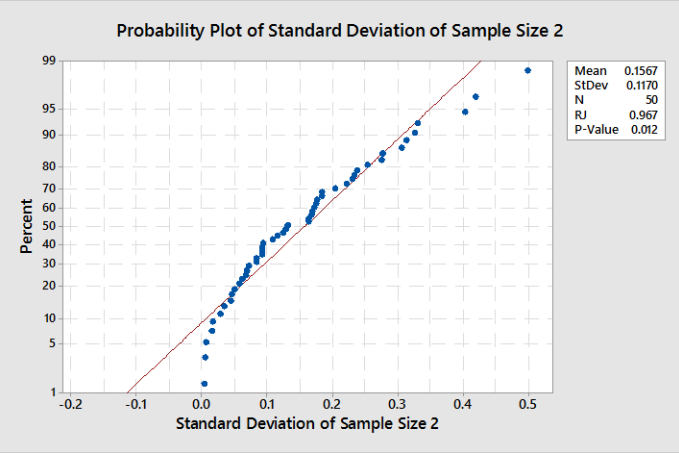


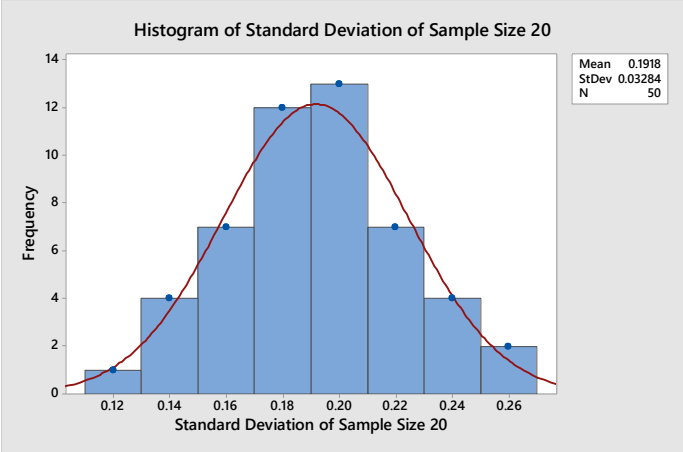

Histogram of Mean of Sample Size 40
Mean 0.7486
StDev 0.03210
N 50
Frequency
Mean of Sample Size 40
0
2
4
6
8
10
12
14
16
18
0.68
0.70
0.72
0.74
0.76
0.78
0.80
0.82

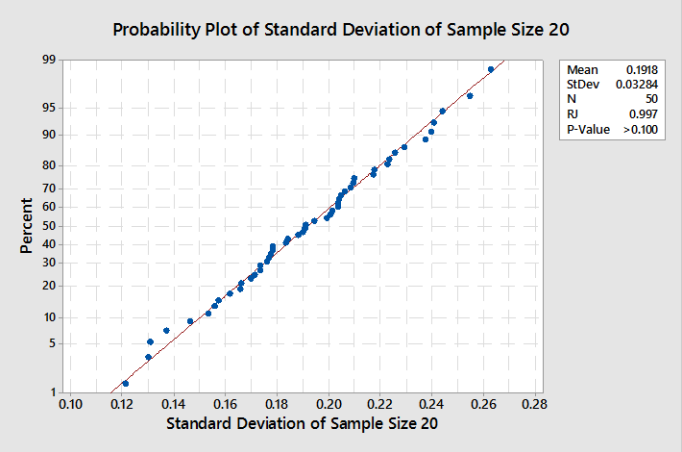

Probability Plot of Mean of Sample Size 40
Mean 0.7486
StDev 0.03210
N 50
RJ 0.995
P-Value >0.100
Percent
Mean of Sample Size 40
1
5
10
20
30
40
50
60
70
80
90
95
99
0.66
0.68
0.70
0.72
0.74
0.76
0.78
0.80
0.82
0.84

The histograms of the distribution of the statistic Standard Deviation along with the Normal density curve is given below:

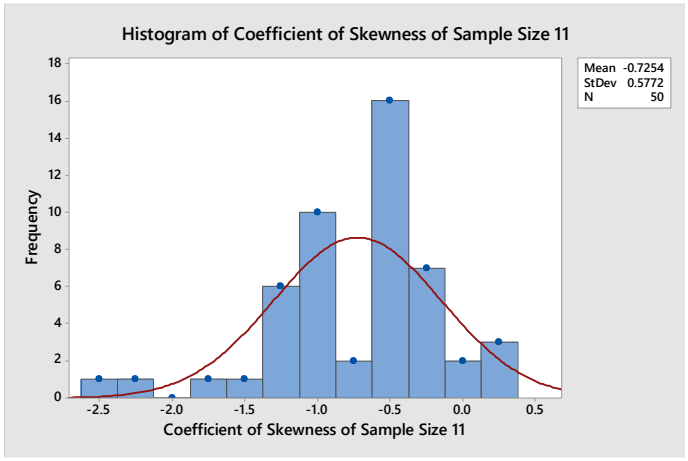


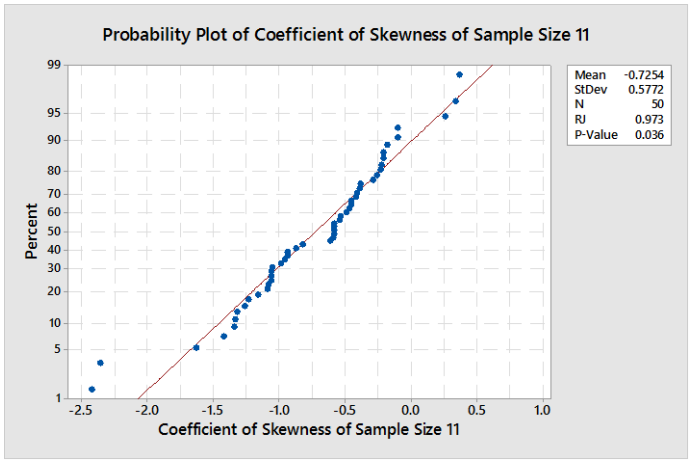


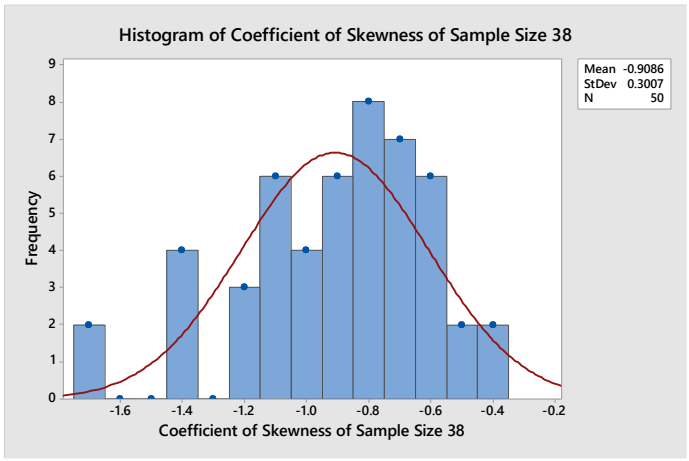


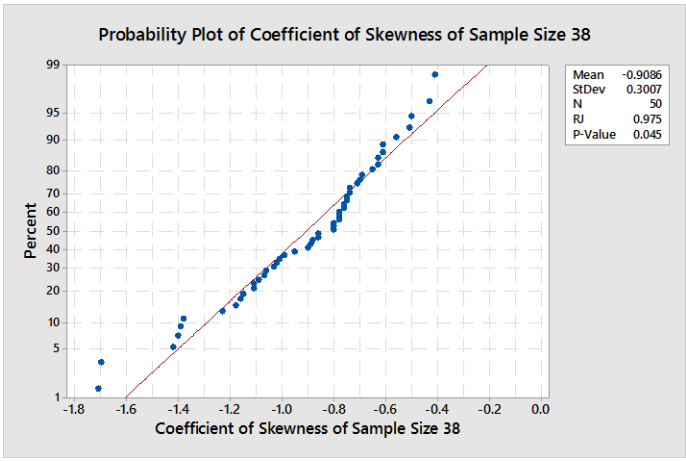


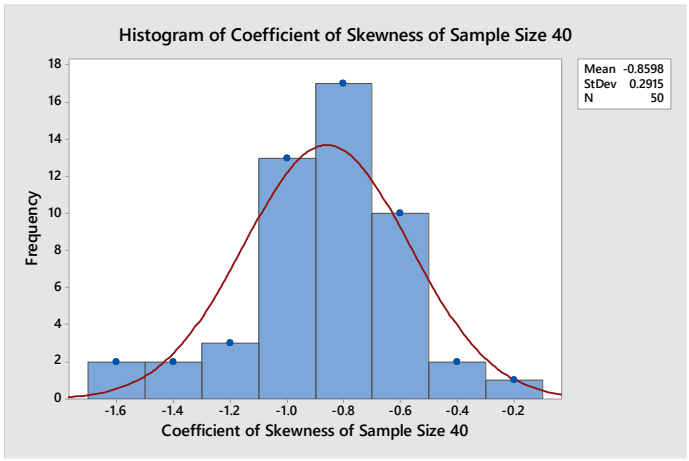
The histograms of the distribution of the statistic Coefficient of Skewness along with the Normal density curve is given below:

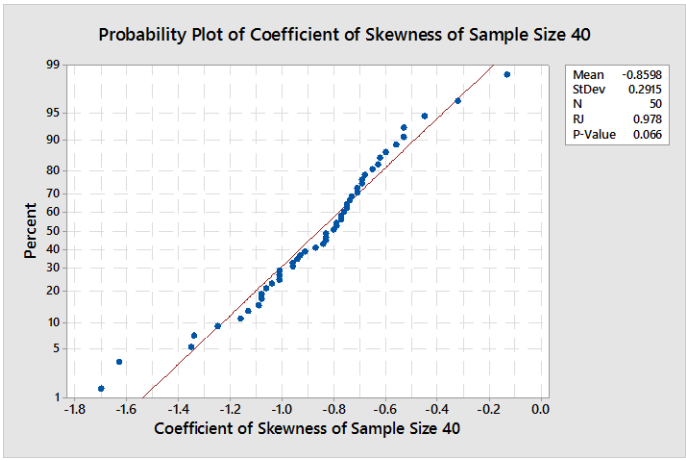




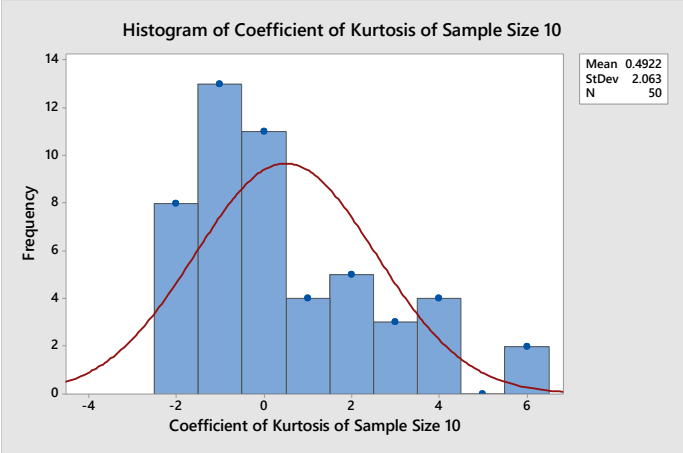


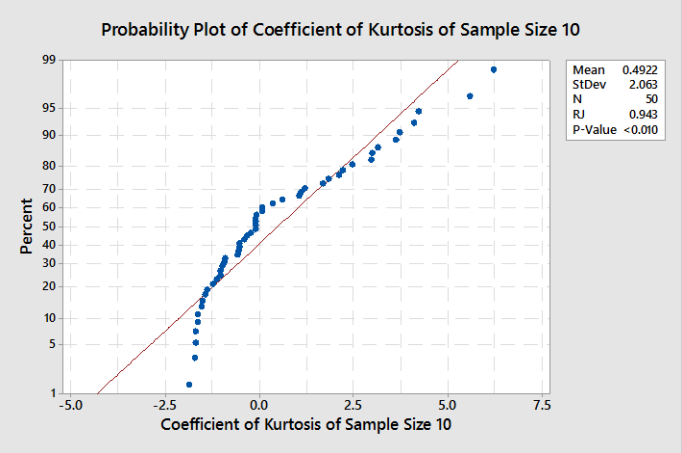


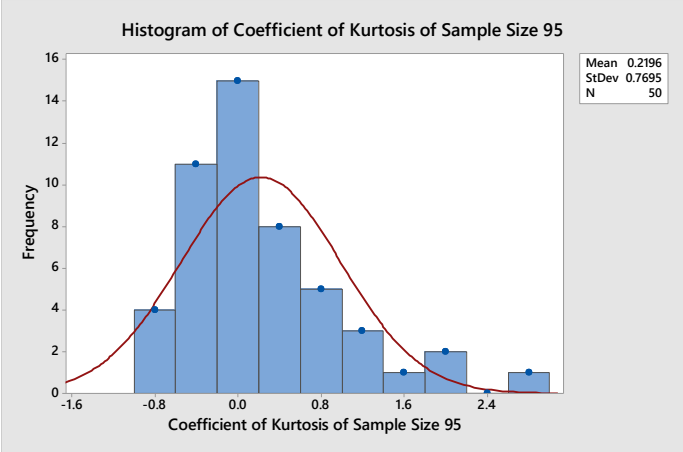


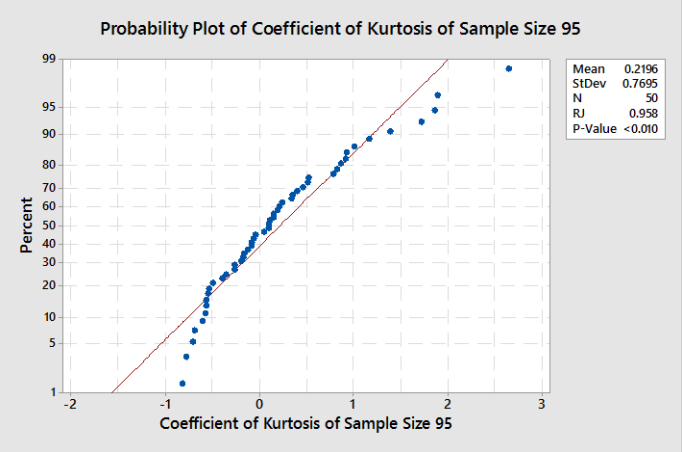


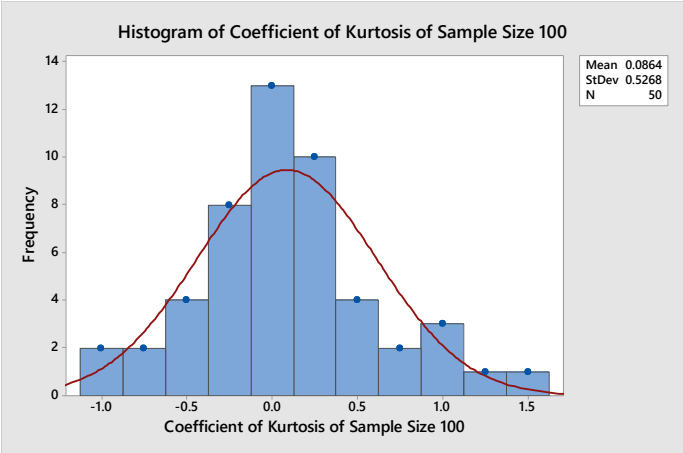
The histograms of the distribution of the statistic Coefficient of Kurtosis along with the Normal density curve is given below:

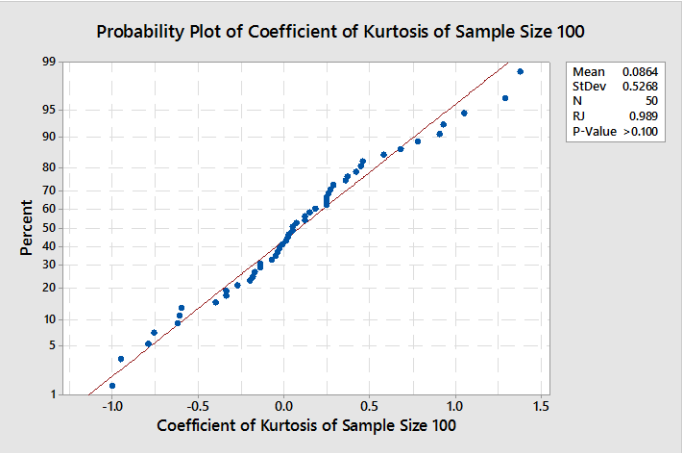












Conclusion

Beta(1,1)

For the parent distribution beta(1,1), the statistic **Mean** attains normality at sample size **10**, the statistic **Standard Deviation** attains normality at sample size **8**, the statistic **Coefficient of Skewness** attains normality at sample size **19**, the statistic **Coefficient of Kurtosis** attains normality at sample size **41**.

Beta(1,2)

For the parent distribution beta(1,2), thestatistic **Mean** attains normality at sample size **60**, the statistic **Standard Deviation** attains normality at sample size **35**, the statistic **Coefficient of Skewness** attains normality at sample size **80**, the statistic **Coefficient of Kurtosis** attains normality at sample size **80**.

Beta(2,1)

For the parent distribution beta(2,1), the statistic **Mean** attains normality at sample size **70**, the statistic **Standard Deviation** attains normality at sample size **20**, the statistic **Coefficient of Skewness** attains normality at sample size **10**, the statistic **Coefficient of Kurtosis** attains normality at sample size **70**.

Beta(1,3)

For the parent distribution beta(1,3), the statistic **Mean** attains normality at sample size **30**, the statistic **Standard Deviation** attains normality at sample size **16**, the statistic **Coefficient of Skewness** attains normality at sample size **22**, the statistic **Coefficient of Kurtosis** attains normality at sample size **132**.

Beta(3,1)

For the parent distribution beta(3,1), the statistic **Mean** attains normality at sample size **40**, the statistic **Standard Deviation** attains normality at sample size **20**, the statistic **Coefficient of Skewness** attains normality at sample size **40**, the statistic **Coefficient of Kurtosis** attains normality at sample size **100**.

We also observe that for the statistic Mean, Beta(1,1) attains normality at a sample size of 10, while the sample size goes on increasing to 60 and then 70 for Beta(1,2) and Beta(2,1) respectively, while it again falls back to 30 and 40 for Beta(1,3) and Beta(3,1) respectively.

For the statistic Standard Deviation we observe that a sample size of 8 attains normality for Beta(1,1) after which the sample size required to attain normality rises to 35 for Beta(1,2) and further 20 and 16 for Beta(2,1) and Beta(1,3) respectively and reaches 20 for Beta(3,1).

For the statistic of Coefficient of Skewness we find it achieving normality at a sample size of 19 for Beta (1,1) and then the required sample size rises to 80 for Beta(1,2) and again falls back to 10 for Beta(2,1). We again find the sample size required to achieve normality to be 22 and further it rises to 40 for Beta(1,3) and Beta(3,1) respectively.

As for the statistic of Coefficient of Kurtosis, it attains normality at a sample size of 41 for Beta(1,1), the required sample size to attain normality rises to 80 for Beta(1,2) moves on to 70 for Beta(2,1) and again rises to 132 for Beta(1,3) and finally falls to 100 for Beta(3,1).

From the above discussion it clear that the sample size required for attaining normality varies with values of the parameters leading to different shape and spread. The same is also true for thestatistics taken into considertion. Thus, before making normality approximation one has to be particularly careful about the parameters of the parent distribution and also on the statistic beingused.

Reference

- An Introduction to Probability and Statistics, Rohatgi, V.K. and Saleh, A.K.M.E., John Wiley & Sons, 1976

- An Outline of Statistical Theory Volume-I, Gun, A.M., Gupta, M.K., Dasgupta, B., The World Press Private Limited, 2003

- Laws of Large Numbers, Chandra T.K., Alpha Science International Private Limited, 2012