SYSC 4005 A

Winter 2023

Project Deliverable #2

Completed By:

Favour Olotu	101130753
Ray Agha	101108060
Joseph Anyia	101117261

Due: March 17, 2023

1. Data Collection and Input Modeling

The distribution of each of the data sets for the simulation entities, based on the shape of their histograms, is an exponential distribution as it starts at the peak and observes a continuous downward trend. Refer to Appendix A for the histograms. The histograms were plotted with a variable bin size. Although it is recommended to use the square root of the sample size for the bin width (from the textbook), it was difficult to get reasonable information with that bin width.

QQ Testing:

The shape of a histogram is necessary for selecting a distribution but not enough to justify the fit of a particular distribution. Knowing that information, the QQ plots are used to evaluate the distribution fit, and the Chi-squared test is used to check the goodness of the fit.

QQ and Chi-squared testing will be conducted using an exponential distribution to test the fit. If the exponential distribution is unsuccessful, a different distribution will be attempted. Then the inverse transform method of random variate generation will be conducted on the distribution.

A QQ or quantile-quantile plot is a graphical method of comparing two probability distributions by plotting their quantiles against each other. It provides a method to evaluate how well a distribution represents a data set. In the "Statistical_Analysis" excel workbook, we currently have implemented QQ testing for Exponential distributions because of the initial observations of the histograms. For the exponential distribution, the following equation is used as the inverse function:

$$F^{-1}(p;\lambda) = \frac{-\ln(1-p)}{\lambda}, \qquad 0 \le p < 1$$

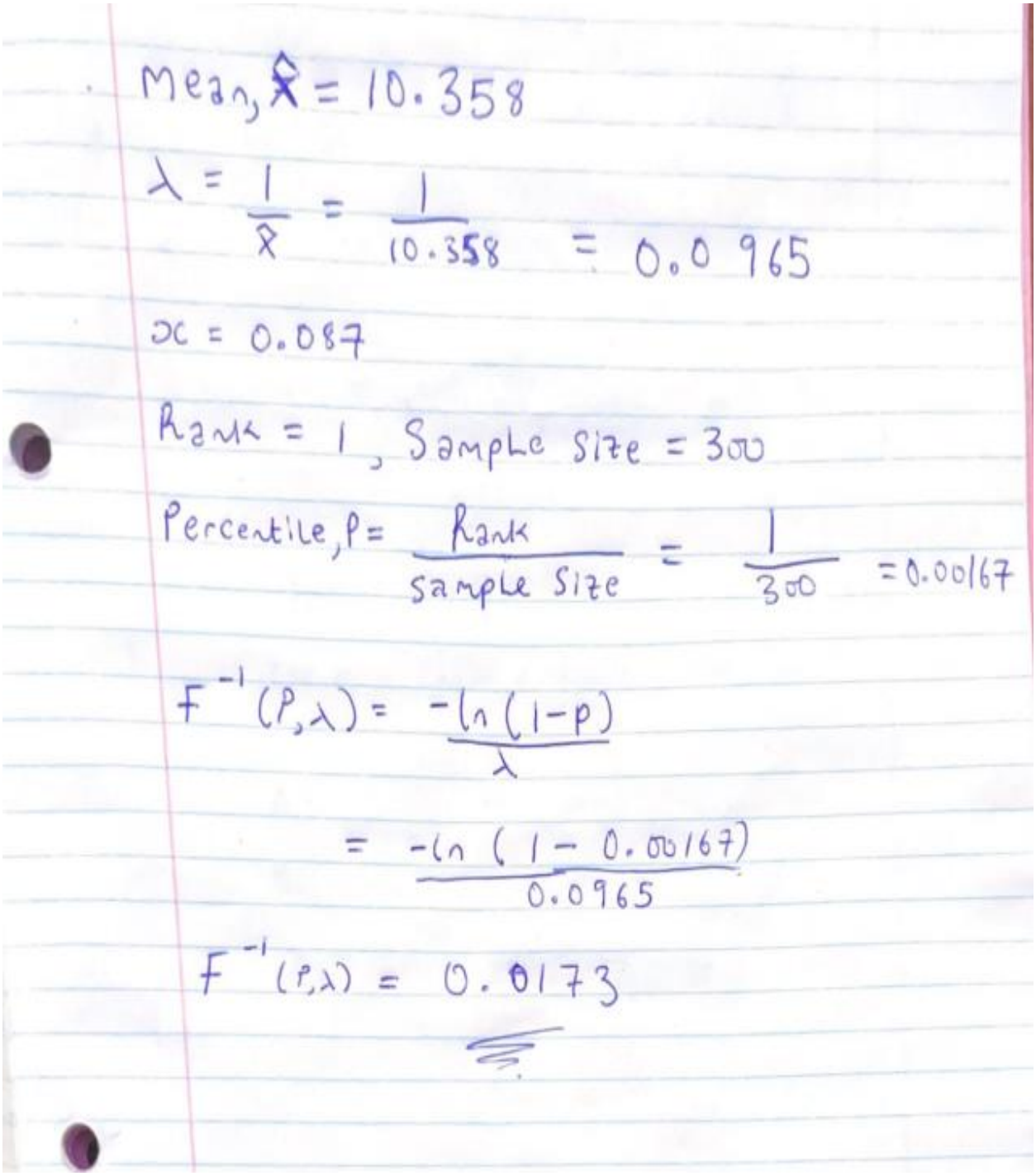
In this function, p represents the percentile varied between 0 and 1.

The value of lambda is 1 divided by the sample mean, as seen in the equation below:

$$\hat{\lambda} = \frac{1}{\bar{X}}$$

(From Table 3 in Chapter 9 of the Textbook)

Below is a sample calculation for the input from the inspector 1 data set:



The fit of the distribution is evaluated by how close data points are to a straight line on the graph. The closer the fit is to a straight line, the better the fit of the distribution to the data, and if, through visual inspection, we determine that the line on the QQ plot is sufficiently straight, then we will proceed to test the goodness of fit with Chi-Squared testing. The QQ plots made can be found in the appendix. Refer to the "Statistical_Analysis.xlsx" file, which contains the detailed calculations of the described steps above for all the sample data sets.

In conclusion, The QQ plots have the shape of a straight line with a slight tail showing that the hypothesis of an exponential distribution is correct. The slight tail's presence is because the extreme values' variance is higher than the middle's (Collected From Lecture Material Chapter 6 Slide 16).

Chi-Squared Testing:

A Chi-Squared test presents a null hypothesis stating that the distribution of a set of sample data points is the same as a specified hypothesized distribution, and the alternate hypothesis states that they are different. Testing if the input data fits a distribution using a chi-squared test is done by first sorting the input data from smallest to largest and then determining how many bins of what size to use to categorize the sample data. Once the bin count and sizes are known, the next step is to tabulate the observed frequencies in each bin in the same way it is done for making histograms. Then, an expected frequency number will need to be determined for each bin. This expected frequency number will be calculated by determining the difference between the CDFs of the upper and lower bounds of the current bin and multiplying the difference by the size of the dataset. Then for each bin, the formula below should be applied:

$$\frac{(O_i - E_i)^2}{E_i}$$

The Chi-Squared value is calculated using the formula:

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

This Chi-Squared value will be compared to a reference Chi-Squared value from a table whose value will depend on the number of bins and the alpha level of significance. If the table value is smaller than the calculated value, then the null hypothesis is rejected, and the distribution is not the same as the sample data. If the table value is larger than the calculated value, the null hypothesis will not be rejected. We can assume that the selected distribution is the same as the input data.

Assuming the distribution to be tested is continuous, the pdf or cdf can be computed as:

$$p_i = \int_{a_{i-1}}^{a_i} f(x)\,dx = F(a_i) - F(a_{i-1})$$

For the Chi-squared test for the exponential distribution with intervals of equal probability, selecting the class intervals is the first step. The recommendation from the textbook for a sample

size greater than 100 should be between $\sqrt{n}$ to $n/5$. Given the sample size of 300 the number of class intervals, k will be between 17 and 60 (Table 5 Chapter 9). Based on that information let k = 20 and then each interval will have a probability, p = 0.05.

The cdf for the exponential distribution is given as:

$$F(a_i) = 1 - e^{-\lambda a_i}$$

Where ai is the endpoint at the ith interval, the value of lambda is 1 divided by the sample mean. The equation can be re-written as seen below because the cumulative area from zero to ai is eqaul to ip.

$$e^{-\lambda a_i} = 1 - ip$$

And ai can be calculated by taking the log of both sides.

$$a_i = -\frac{1}{\lambda}\ln(1 - ip), \quad i = 0, 1, \ldots, k$$

Regardless of the value of lambda a0 will always be zero and ak infinity.

The tables below correspond to the chi-squared testing tables for all 6 data files.

1. Inspector 1 Table

F	G	H	I	J
interval start	interval finish	Observed Frequency	Expected Frequency	Chi-squared
0	0.531291327	16	15	0.066666667
0.531291327	1.091314739	7	15	4.266666667
1.091314739	1.683356445	11	15	1.066666667
1.683356445	2.311300822	10	15	1.666666667
2.311300822	2.979785015	15	15	0
2.979785015	3.694406969	14	15	0.066666667
3.694406969	4.462010674	18	15	0.6
4.462010674	5.291085837	16	15	0.066666667
5.291085837	6.192341848	19	15	1.066666667
6.192341848	7.179556113	17	15	0.266666667
7.179556113	8.270870852	15	15	0
8.270870852	9.490856935	19	15	1.066666667
9.490856935	10.87396308	18	15	0.6
10.87396308	12.47064195	14	15	0.066666667
12.47064195	14.35911223	21	15	2.4
14.35911223	16.67041305	13	15	0.266666667
16.67041305	19.65019806	13	15	0.266666667
19.65019806	23.84996916	20	15	1.666666667
23.84996916	31.02952527	12	15	0.6
31.02952527	infinity	12	15	0.6
		300	300	16.66666667

As shown in the table above x^2(0) = 16.67,

alpha = 0.05, s = 1, k - s - 1 = 18

x^2 @(0.05,18) = 28.869

x^2 @(0.05,18) > x^2(0), at 0.05 significance level the hypothesis is not rejected.

2. Inspector 2 Component 2 Table

	interval start	interval finish	Observed Frequency	Expected Frequency	Chi-squared
)	0	0.796938957	16	15	0.066667
L	0.796938957	1.636976147	7	15	4.266667
?	1.636976147	2.525040897	11	15	1.066667
3	2.525040897	3.466959786	10	15	1.666667
4	3.466959786	4.46968855	15	15	0
5	4.46968855	5.541624125	11	15	1.066667
5	5.541624125	6.693032525	21	15	2.4
7	6.693032525	7.936648337	16	15	0.066667
3	7.936648337	9.28853569	19	15	1.066667
9	9.28853569	10.76936074	17	15	0.266667
)	10.76936074	12.40633689	15	15	0
L	12.40633689	14.23632053	19	15	1.066667
?	14.23632053	16.31098487	18	15	0.6
3	16.31098487	18.70600908	14	15	0.066667
4	18.70600908	21.53872148	21	15	2.4
5	21.53872148	25.00568127	13	15	0.266667
5	25.00568127	29.47536982	13	15	0.266667
7	29.47536982	35.77504201	21	15	2.4
3	35.77504201	46.54440275	11	15	1.066667
9	46.54440275	infinity	12	15	0.6
			300	300	20.66667

As shown in the table above x^2(0) = 20.67,

alpha = 0.05, s = 1, k - s - 1 = 18

x^2 @(0.05,18) = 28.869

x^2 @(0.05,18) > x^2(0), at 0.05 significance level the hypothesis is not rejected.

3. Inspector 2 Component 3 Table

interval start	interval finish	Observed Frequency	Expected Frequency	Chi-squared
0	1.058322062	15	15	0
1.058322062	2.173877882	12	15	0.6
2.173877882	3.353213526	15	15	0
3.353213526	4.604066596	15	15	0
4.604066596	5.935674198	15	15	0
5.935674198	7.359187327	9	15	2.4
7.359187327	8.888239084	23	15	4.266667
8.888239084	10.53974079	14	15	0.066667
10.53974079	12.33502536	9	15	2.4
12.33502536	14.30153711	14	15	0.066667
14.30153711	16.47541499	18	15	0.6
16.47541499	18.90560371	18	15	0.6
18.90560371	21.66072444	21	15	2.4
21.66072444	24.8412779	14	15	0.066667
24.8412779	28.60307422	10	15	1.666667
28.60307422	33.20714082	20	15	1.666667
33.20714082	39.14281502	15	15	0
39.14281502	47.50867793	14	15	0.066667
47.50867793	61.81021504	14	15	0.066667
61.81021504	infinity	15	15	0
		300	300	16.93333

As shown in the table above $x^2(0) = 16.93$,

alpha = 0.05, s = 1, k - s - 1 = 18

x^2 @(0.05,18) = 28.869

x^2 @(0.05,18) > $x^2(0)$, at 0.05 significance level the hypothesis is not rejected.

4. Workstation 1 Table

interval start	interval finish	Observed Frequency	Expected Frequency	Chi-squared
0	0.2361757	6	15	5.4
0.2361757	0.48512371	12	15	0.6
0.485123714	0.74830487	18	15	0.6
0.748304868	1.02744589	21	15	2.4
1.027445887	1.32460813	18	15	0.6
1.324608129	1.64228006	17	15	0.266667
1.642280056	1.98350404	19	15	1.066667
1.983504039	2.35205402	16	15	0.066667
2.352054016	2.75269065	12	15	0.6
2.75269065	3.19153843	16	15	0.066667
3.191538431	3.67666214	23	15	4.266667
3.676662145	4.21898432	11	15	1.066667
4.218984317	4.83381849	13	15	0.266667
4.833818487	5.54359245	11	15	1.066667
5.543592446	6.38307686	10	15	1.666667
6.383076861	7.41052275	20	15	1.666667
7.410522748	8.73513088	9	15	2.4
8.735130877	10.6020612	17	15	0.266667
10.60206118	13.7935996	13	15	0.266667
13.79359961	infinity	18	15	0.6
		300	300	25.2

As shown in the table above x^2(0) = 25.2,

alpha = 0.05,

s = 1

k - s - 1 = 18

x^2 @(0.05,18) = 28.869

x^2 @(0.05,18) > x^2(0), at 0.05 significance level the hypothesis is not rejected.

5. Workstation 2 Table

interval sta	interval finish	Observed Frequency	Expected Frequency	Chi-squared
0	0.568976339	19	15	1.066667
0.568976	1.168722758	20	15	1.666667
1.168723	1.802758561	17	15	0.266667
1.802759	2.475243645	17	15	0.266667
2.475244	3.191144075	8	15	3.266667
3.191144	3.956454861	10	15	1.666667
3.956455	4.778505447	18	15	0.6
4.778505	5.66638772	17	15	0.266667
5.666388	6.6315707	15	15	0
6.631571	7.688809036	13	15	0.266667
7.688809	8.857531794	19	15	1.066667
8.857532	10.16405268	16	15	0.066667
10.16405	11.6452639	15	15	0
11.64526	13.35519676	14	15	0.066667
13.3552	15.37761807	17	15	0.266667
15.37762	17.85286172	9	15	2.4
17.85286	21.04400579	8	15	3.266667
21.04401	25.54167075	11	15	1.066667
25.54167	33.23047979	15	15	0
33.23048	infinity	22	15	3.266667
		300	300	20.8

As shown in the table above x^2(0) = 20.8,

alpha = 0.05,

s = 1

k - s - 1 = 18

x^2 @(0.05,18) = 28.869

x^2 @(0.05,18) > x^2(0), at 0.05 significance level the hypothesis is not rejected.

6. Workstation 3 Table

interval start	interval finish	Observed Frequency	Expected Frequency	Chi-squared
0	0.451154274	7	15	4.266667
0.451154274	0.926706844	22	15	3.266667
0.926706844	1.429448246	13	15	0.266667
1.429448246	1.962676957	16	15	0.066667
1.962676957	2.530330683	11	15	1.066667
2.530330683	3.137163003	11	15	1.066667
3.137163003	3.788985601	16	15	0.066667
3.788985601	4.49300764	21	15	2.4
4.49300764	5.258323167	18	15	0.6
5.258323167	6.096631478	22	15	3.266667
6.096631478	7.023338323	15	15	0
7.023338323	8.059308435	18	15	0.6
8.059308435	9.233794482	13	15	0.266667
9.233794482	10.58963912	13	15	0.266667
10.58963912	12.19326296	9	15	2.4
12.19326296	14.15593991	11	15	1.066667
14.15593991	16.6862706	16	15	0.066667
16.6862706	20.25257139	16	15	0.066667
20.25257139	26.34920287	17	15	0.266667
26.34920287	infinity	15	15	0
		300	300	21.33333

As shown in the table above x^2(0) = 21.33,

alpha = 0.05,

s = 1

k - s - 1 = 18

x^2 @(0.05,18) = 28.869

x^2 @(0.05,18) > x^2(0), at 0.05 significance level the hypothesis is not rejected.

The following table shows a summary of the chi-squared tests.:

	Lambda	Chi Total	Fail to Reject Null Hypothesis?	Distribution Type
Inspector 1	0.096545	16.67	Yes	Exponential
Inspector 2_2	0.064363	20.67	Yes	Exponential
Inspector 2_3	0.048467	16.93	Yes	Exponential
Workstation 1	0.217183	25.2	Yes	Exponential
Workstation 2	0.09015	20.8	Yes	Exponential
Workstation 3	0.113693	21.33	Yes	Exponential

Testing all six of the data sets results in the Null hypothesis not being rejected, and we can assume that the distribution does match. This list of 300 random variables using the parameters was compared to the distribution using the same parameters. These tests were also successful in confirming and reinforcing our distribution parameter selections.

2. Input Generation

Generating Random Numbers

To generate randomized numbers for the simulation's component inspection times and workstation build times, a random number generation system was manually implemented and tested. For a generated number to be considered random, it should be independent and uniformly distributed. The random number generation model would be the multiplicative congruential model (MCM), based on the linear congruential model with a zero incrementor.

Multiplicative Congruential Model:

Random number generation with the multiplicative congruential model is done with a recursive method as shown below:

$$X \text{ i} = (aX \text{ i-1}) \% m$$

Where X represents a randomly selected number, a is the multiplier parameter, m is the modulus parameter, and % is the modulus operator. Since the current random number to generate depends on the previously generated number, the user must define an initial seed value to set as the initial '*X* i-1' value so subsequent numbers can be generated. Once a number is generated, the value for '*X* i-1' will be changed to *X* i. The numbers generated are integers in the range of [0, m-1].

The integers are converted to random numbers using the equation below:

$$R_i = \frac{X_i}{m}$$

To ensure this generator is good, a large integer for modulus m will be chosen. Refer to the "Multiplicative_Congruential_Model.py" module in the source code for implementation.

Testing Random Number Generation

To ensure that the randomly generated numbers are truly random, we shall test for uniformity and independence with 95% confidence. Independence will be tested using an Autocorrelation test, and uniformity will be tested using a Kolmogorov-Smirnov test.

Autocorrelation Testing:

Autocorrelation is used to determine if the random number generation process produces independent numbers. The test assesses if subsequently selected random numbers tend to follow a pattern. For example, are low numbers followed by higher numbers, or check if high numbers follow high numbers, etc. The null hypothesis of this test states that the numbers are independent, and the alternate hypothesis is that they are not. This process first involves setting and calculating the below variables, and the values used for testing this project are shown below:

- i = 1, the $starting\ index\ in\ the\ set\ of\ random\ numbers$
- m = 1 (lag value from lecture), $how\ many\ numbers\ to\ skip\ between\ each\ choice$
- M = 898, $Largest\ integer\ that\ makes\ i + (M + 1)m \leq N\ true$
- $\alpha = 0.05$, $confidence$. ($\alpha/2$) from $Z\text{-}Table$ equals the $value\ to\ beat$
- N = 900, $Total\ number\ of\ random\ numbers\ to\ generate$

The formula retrieved from the textbook used is given below:

$$\widehat{\rho}_{i\ell} = \frac{1}{M+1}\left[\sum_{k=0}^{M} R_{i+k\ell}R_{i+(k+1)\ell}\right] - 0.25$$

$$\sigma_{\hat{\rho}_{i\ell}} = \frac{\sqrt{13M+7}}{12(M+1)}$$

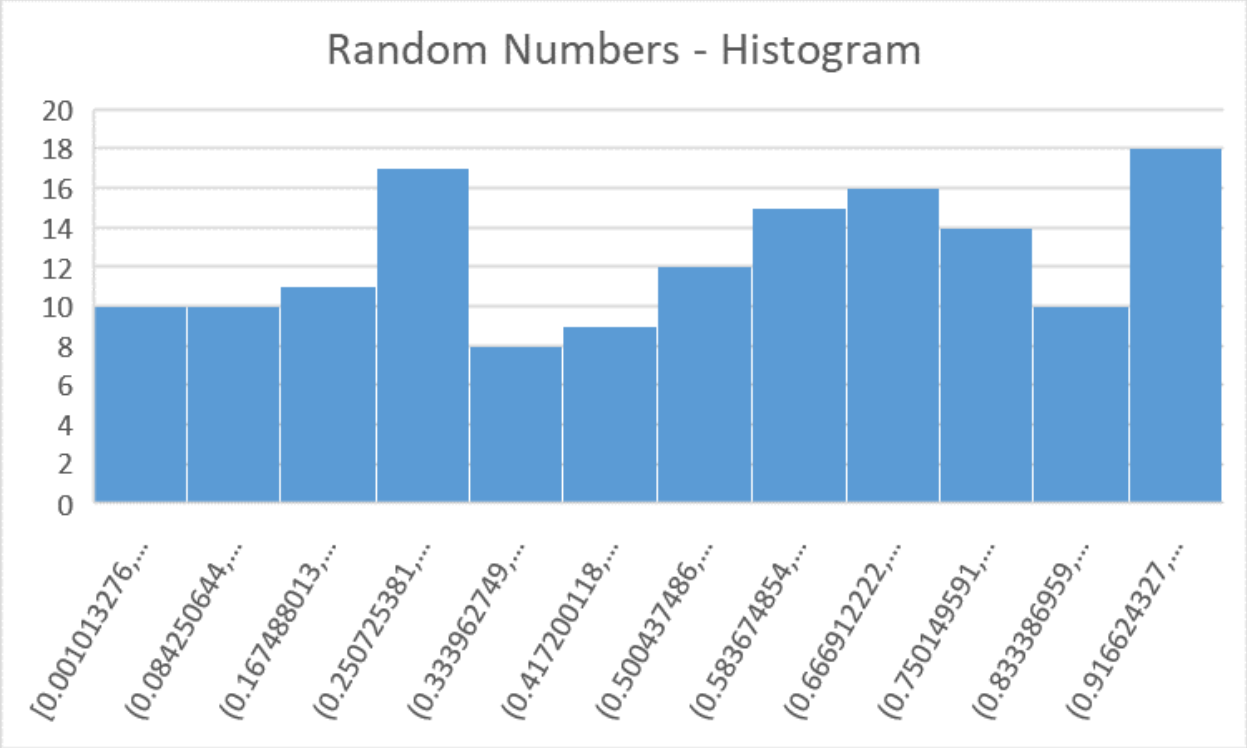
$$\frac{\rho}{\sigma} \leq Z_{table} = 1.96$$

If the value of calculated above, is less than the value from the z table then the null hypothesis is not rejected and then we can conclude the numbers are independent with a confidence of 95%. Alternatively, the null hypothesis is rejected, and the numbers are dependent. Refer to the "Multiplicative_Congruential_Model.py" module in the source code for the implementation of the tests.

Kolmogorov-Smirnov Testing:

To determine if the random number generation process produces uniform values, a Kolmogorov-Smirnov test will compare a set of randomly generated numbers to a uniform distribution. This test will compare a uniform distribution's continuous cumulative distribution function to the empirical cumulative distribution function observed from a set of random samples. The null hypothesis of this test states that there is no difference between the distribution of the random numbers and uniform distribution, and the alternate hypothesis states that they are different. The first step is to generate a set of random numbers. 150 random numbers will be generated for the testing in this project, and this set of random numbers shall be sorted from smallest to largest. Another set of 150 numbers will represent the expected cumulative distribution function values, ordered from smallest to largest. Given that there are 150 random samples to match, this list will also have 200 values, starting at 0.01, with each element being 0.01 larger than the previous, [0.01, 0.02, 0.03, . . . 0.98, 0.99, 1.0].

The figure below is a histogram of randomly generated variables.



Given these sorted lists of random numbers and expected cumulative distributions, two more lists will be made, taking in corresponding values from the two lists. The first list will be called the D_plus list and each element will be the calculation result of the

(current expected value - current sorted random number).

The second list will be called D_minus, and each element of the list will be calculated as

(current random number - previous expected value).

The first element of D-minus will use previousExpectedValue=0, and the second element in D_minus will be

(2nd smallest random number) - (0.01 * previous expected value)

Once the D_plus and D_minus lists are made, the most significant contained value between both of them will be retrieved and will be referred to as D_max. This D_max value will then be compared to the table based D_alpha value and using alpha=0.05, D_alpha=1.36/sqrt(150)

If the D_max value is less than the D_alpha table value, then the null hypothesis will not be rejected and it can be assumed that the generated numbers are uniformly distributed. Otherwise, the null hypothesis will be rejected and it can be concluded that the randomly generated values are not uniformly distributed.

After performing the test on the random number generator, the null hypothesis was not rejected, meaning the generator is uniform. Refer to the “Multiplicative_Congruential_Model.py” module in the source code for the implementation of the test.

Inverse-Transform Technique:

For generating random variates, the preferred method is the inverse-transform technique. The inverse-transform technique requires an invertible cumulative distribution function for the distribution to generate from. Once the inverse of a distribution’s cumulative distribution function is determined, a randomly selected number from a uniform distribution will become the input parameter, and the calculation output will be a random variate. This technique will be used for uniform, exponential, Weibull, and triangular distributions, and given that all distributions were determined to be exponential, this technique will be used throughout.

The implication to Simulation Code:

Upon the completion of all these analyses, The randomly generated service times for each entity of the system to be simulated the inverse transform techniques for generating random variates will be adopted. The “numpy” library in python has a built-in function to aid with this known as “`numpy.random.exponential()`” for generating random exponential numbers. This functionality is applied to the inspectors and workstation modules for generating a random delay time. They are biased by the mean from the sample data set. Refer to lines 51, 116 and 118 of the “Inspector.py” module and line 54 of the “Workstation.py” module for implementation.

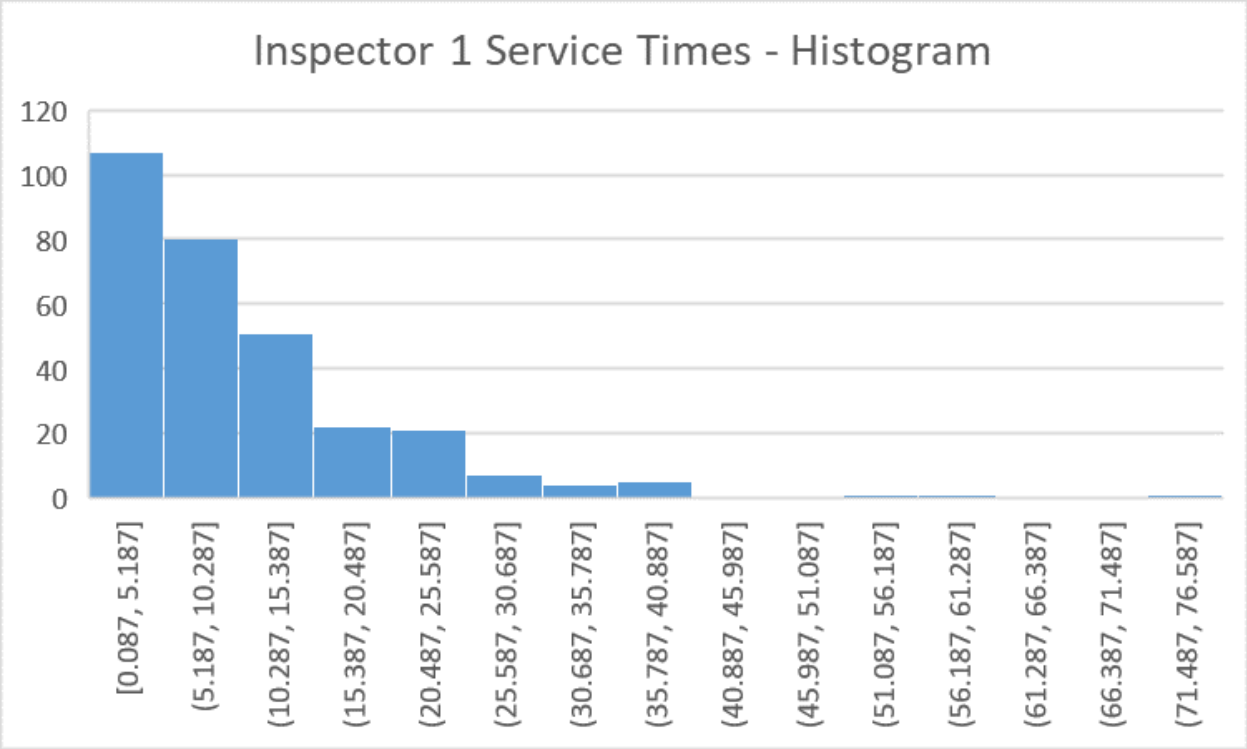
Appendices:

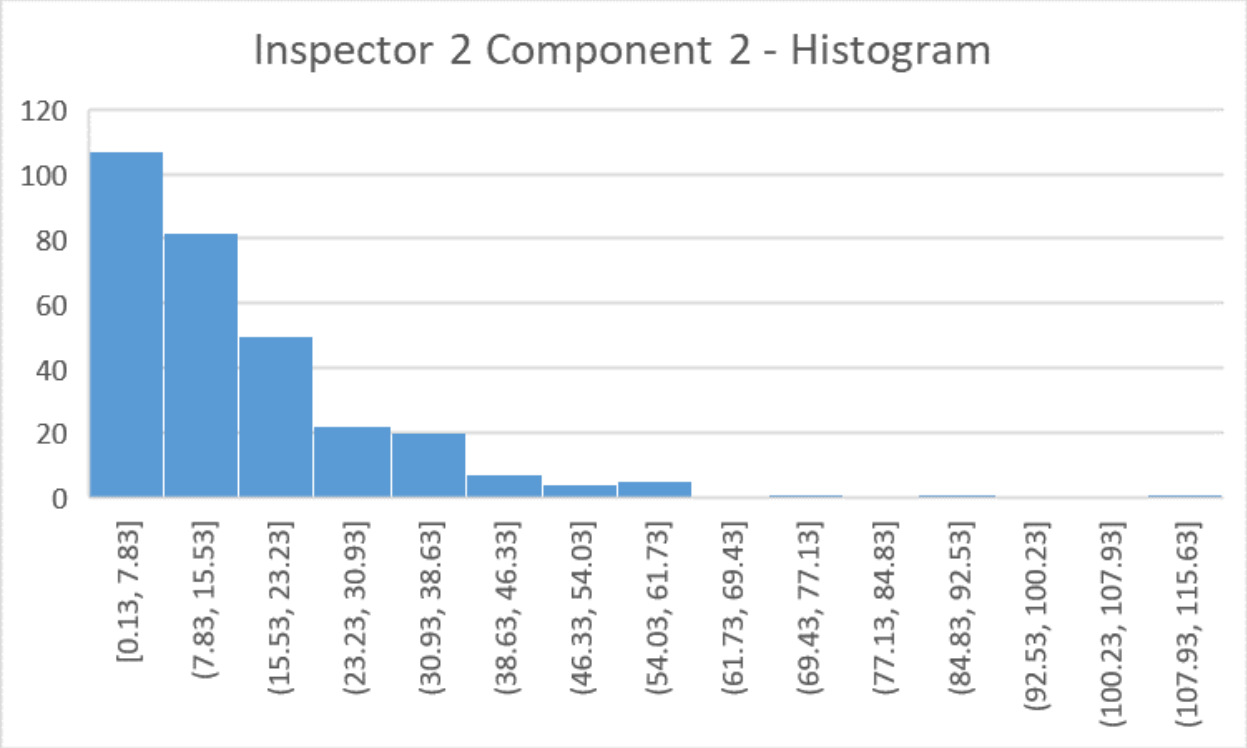
Appendix A (Analysis of Data Files for Simulation Entities)

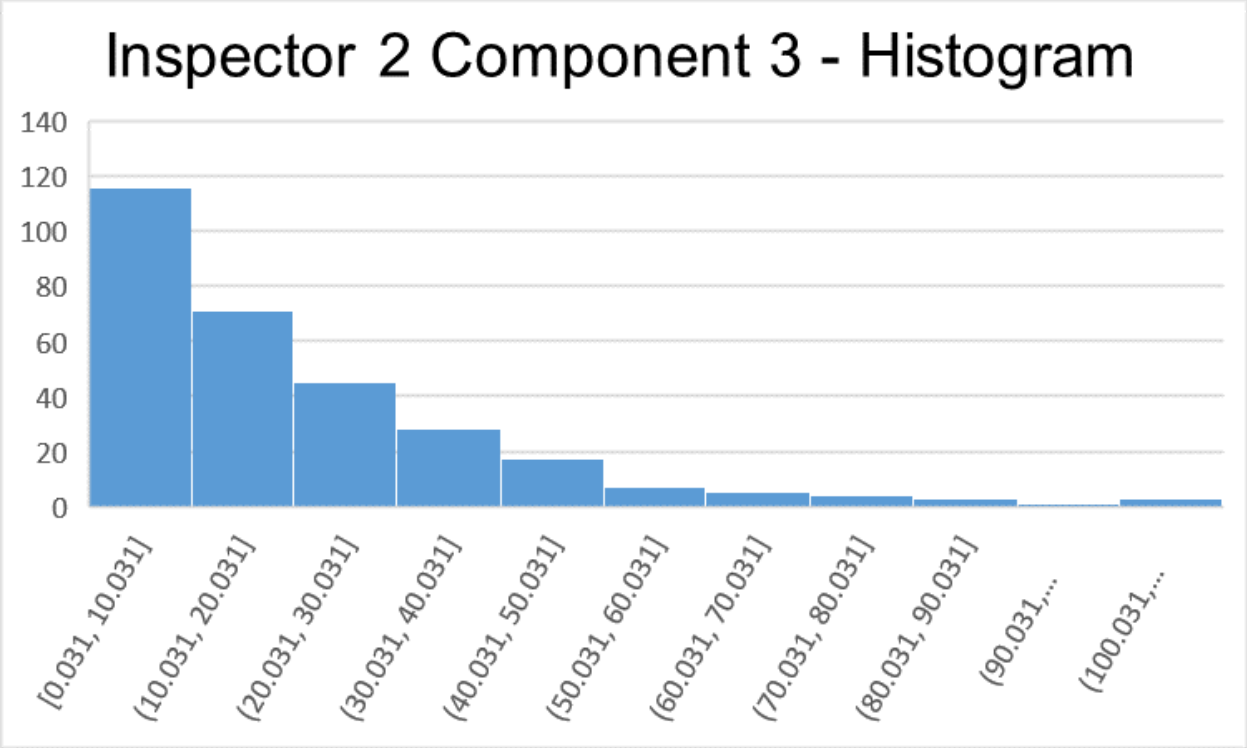
The following information corresponds to the histogram and Q-Q plots from the sample data for each simulation entity as part of the input modelling process.

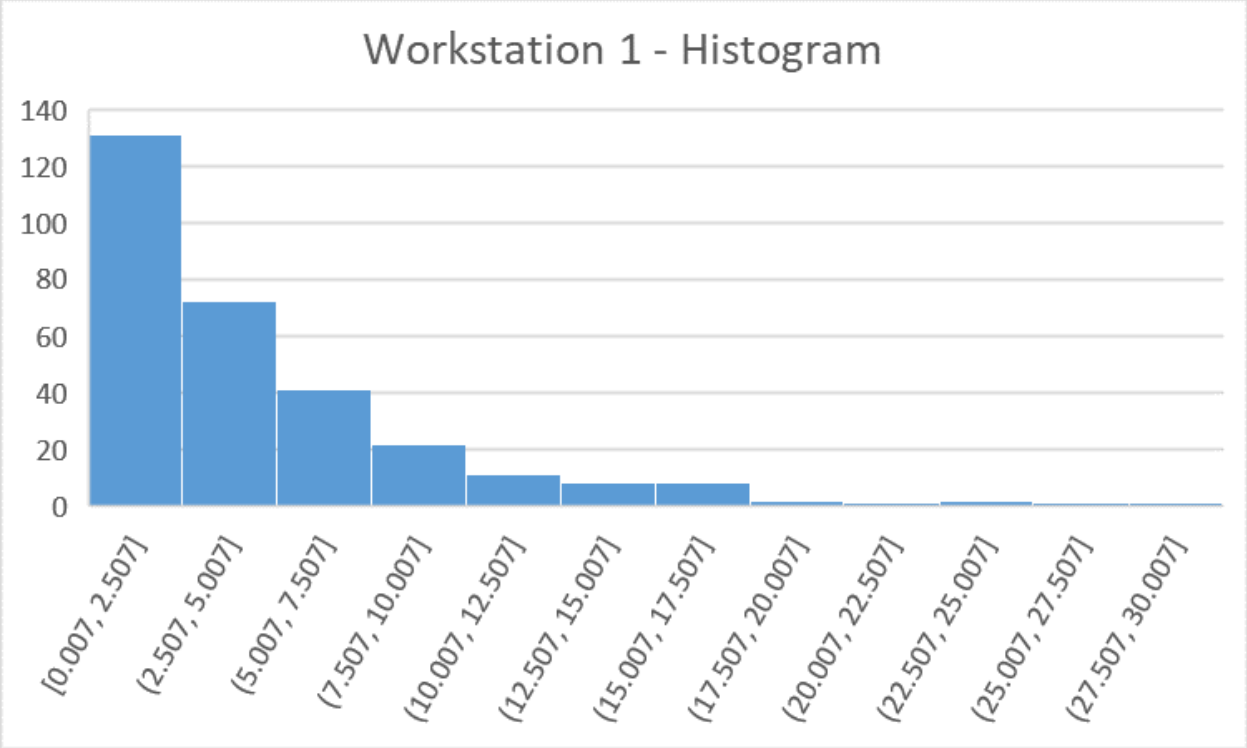
Histogram Figures

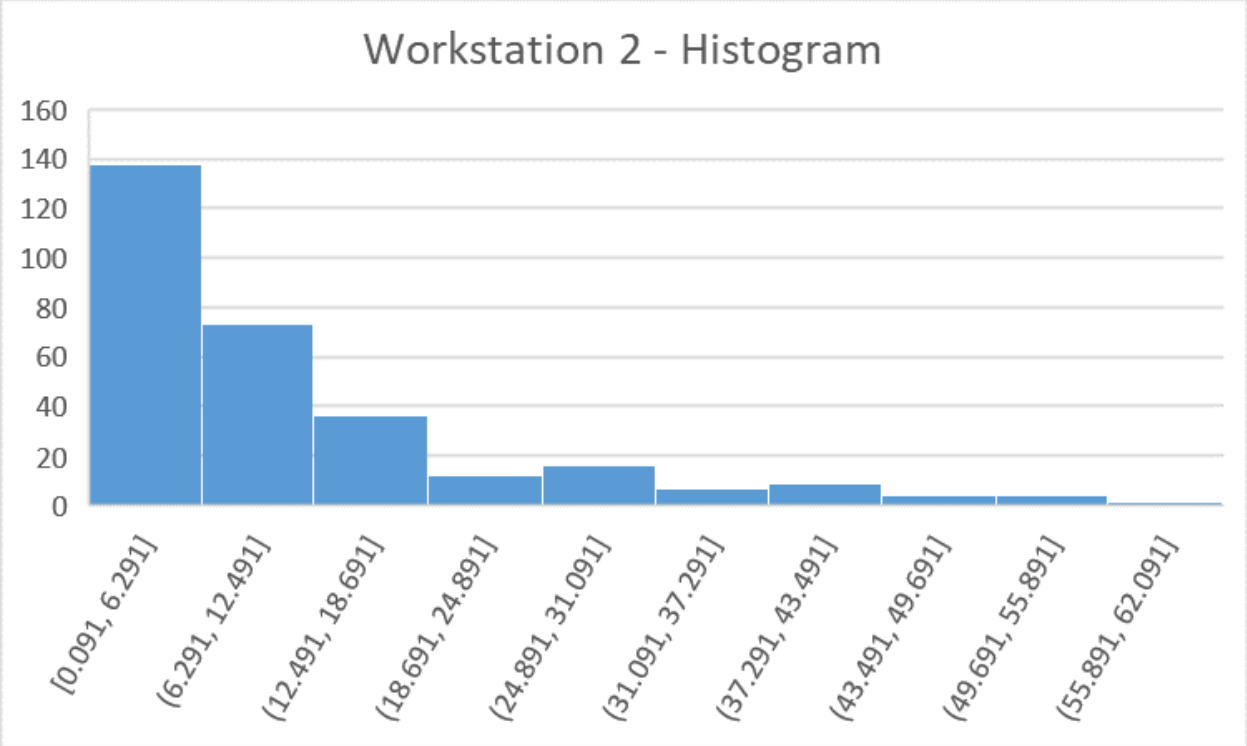
As shown in the histograms below, the distributions displayed follow an exponential distribution as it starts at the peak and observes a continuous downward trend.

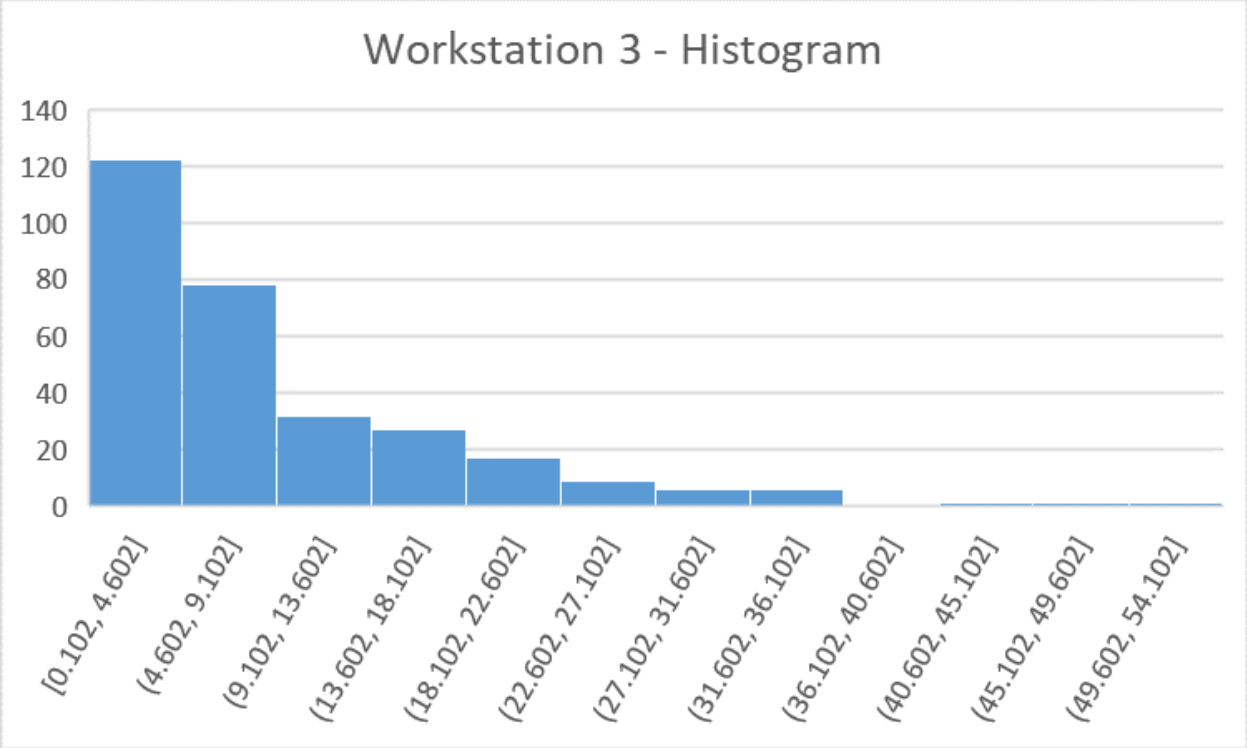



Inspector 2 Component 2 - Histogram
120
100
80
60
40
20
0
[0.13, 7.83]
(7.83, 15.53]
(15.53, 23.23]
(23.23, 30.93]
(30.93, 38.63]
(38.63, 46.33]
(46.33, 54.03]
(54.03, 61.73]
(61.73, 69.43]
(69.43, 77.13]
(77.13, 84.83]
(84.83, 92.53]
(92.53, 100.23]
(100.23, 107.93]
(107.93, 115.63]


Inspector 2 Component 3 - Histogram
140
120
100
80
60
40
20
0
[0.031, 10.031]
(10.031, 20.031]
(20.031, 30.031]
(30.031, 40.031]
(40.031, 50.031]
(50.031, 60.031]
(60.031, 70.031]
(70.031, 80.031]
(80.031, 90.031]
(90.031,...
(100.031,...

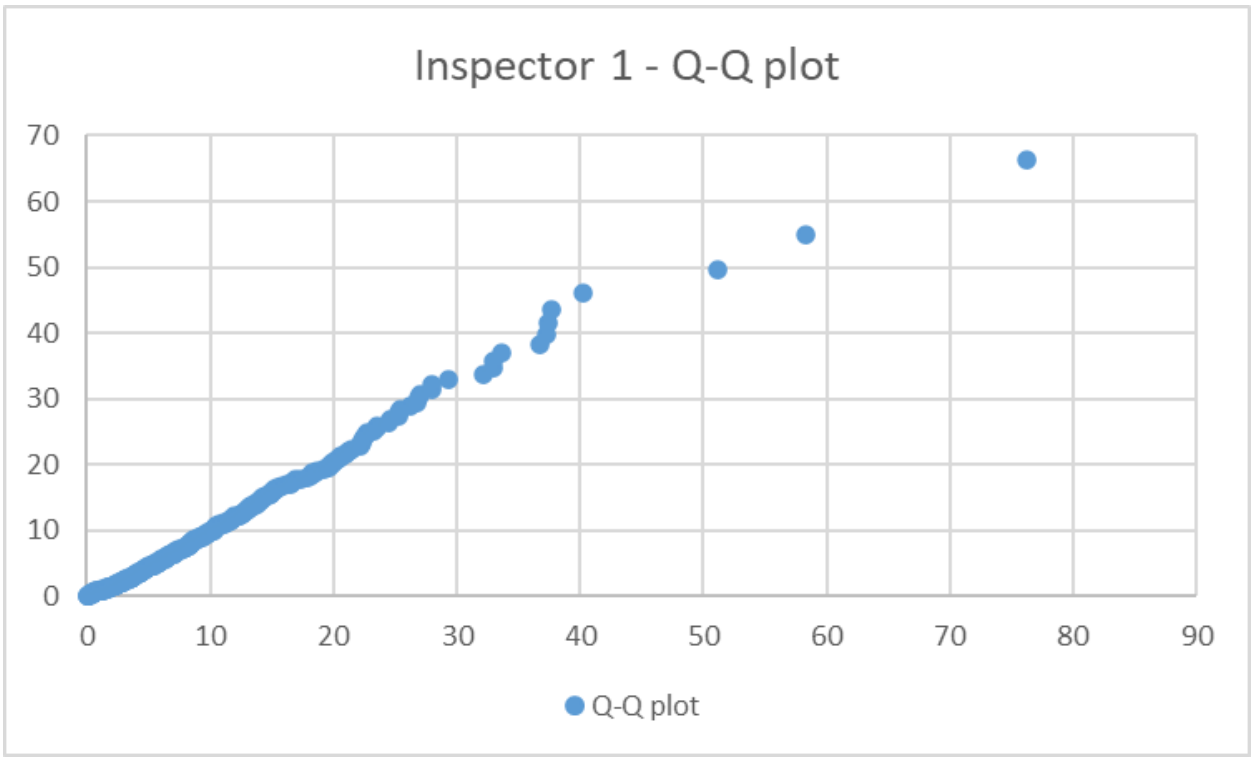

Workstation 1 - Histogram
140
120
100
80
60
40
20
0
[0.007, 2.507]
(2.507, 5.007]
(5.007, 7.507]
(7.507, 10.007]
(10.007, 12.507]
(12.507, 15.007]
(15.007, 17.507]
(17.507, 20.007]
(20.007, 22.507]
(22.507, 25.007]
(25.007, 27.507]
(27.507, 30.007]

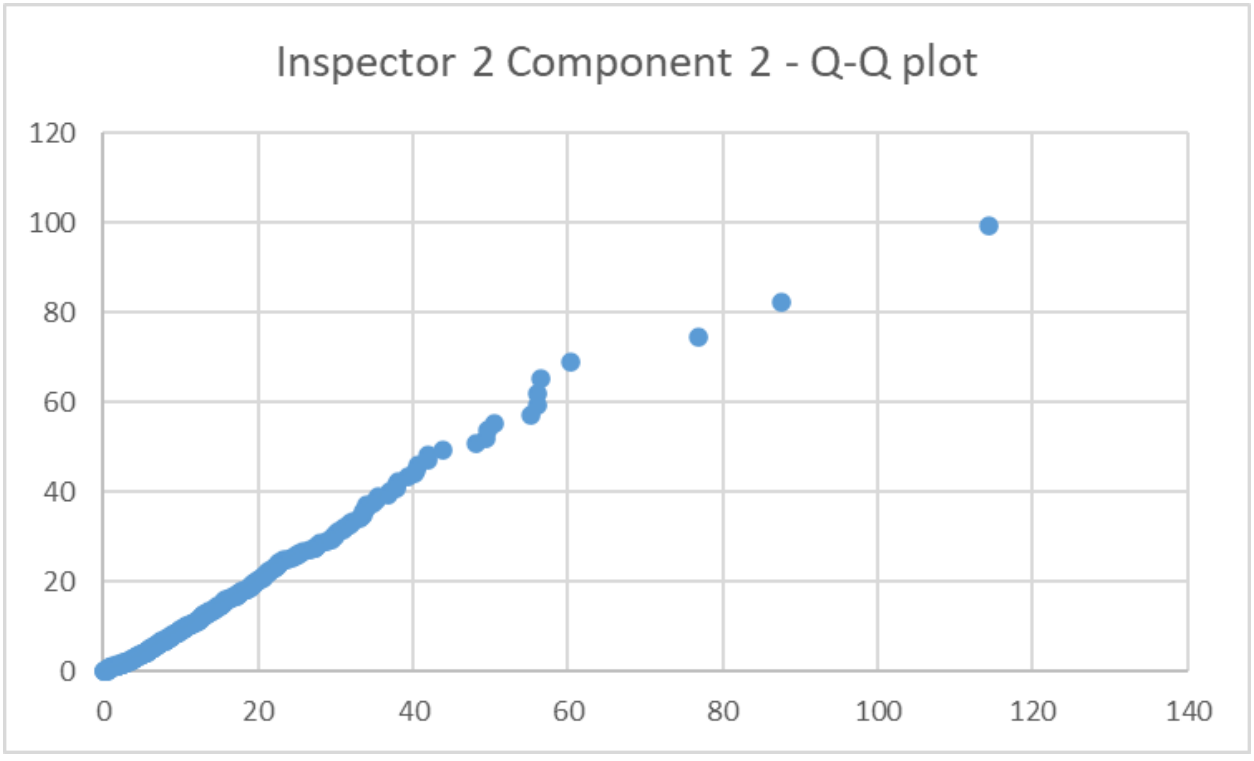

Workstation 2 - Histogram
160
140
120
100
80
60
40
20
0
[0.091, 6.291]
(6.291, 12.491]
(12.491, 18.691]
(18.691, 24.891]
(24.891, 31.091]
(31.091, 37.291]
(37.291, 43.491]
(43.491, 49.691]
(49.691, 55.891]
(55.891, 62.091]

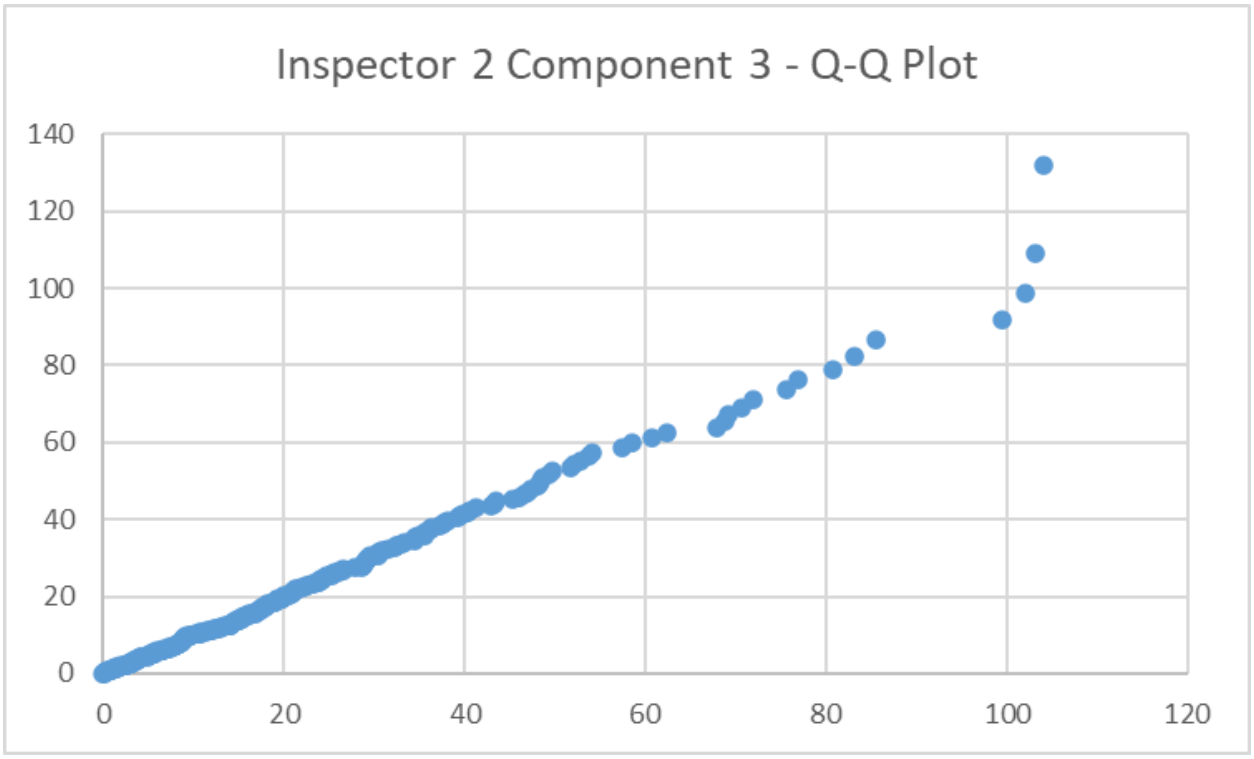


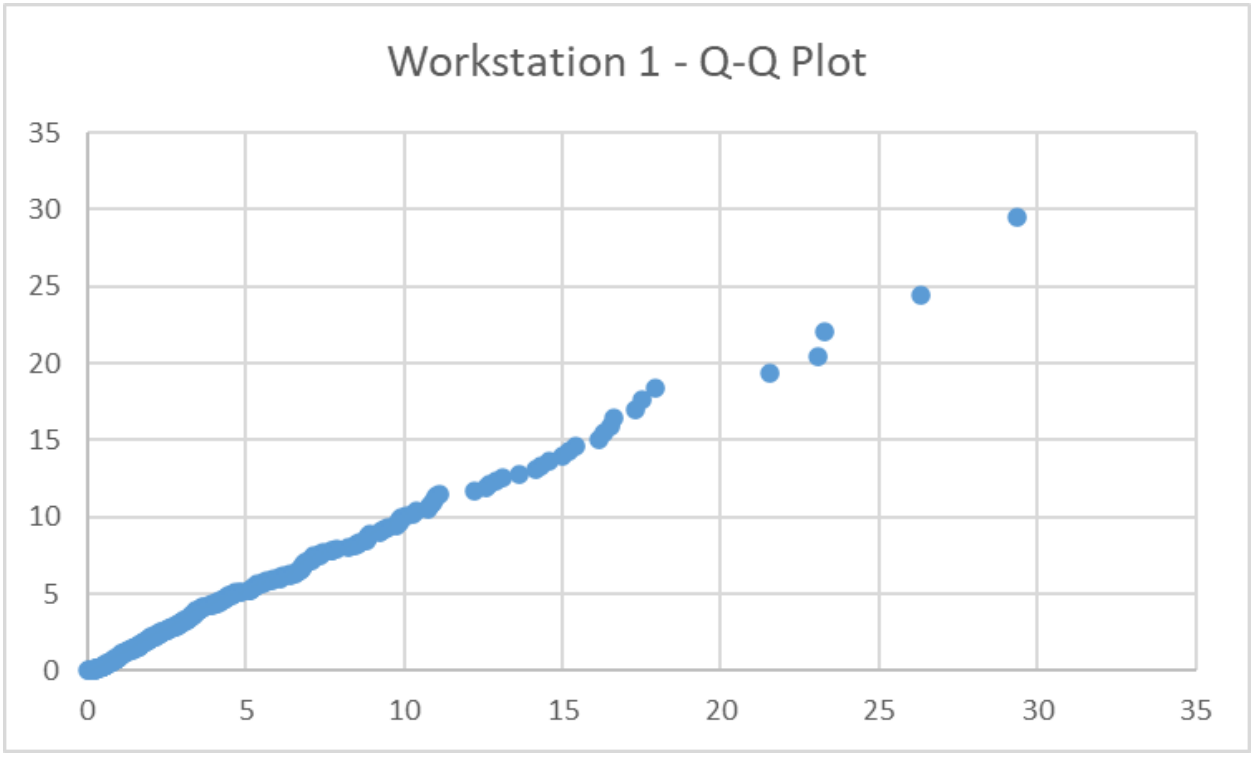
Exponential Q-Q Plots

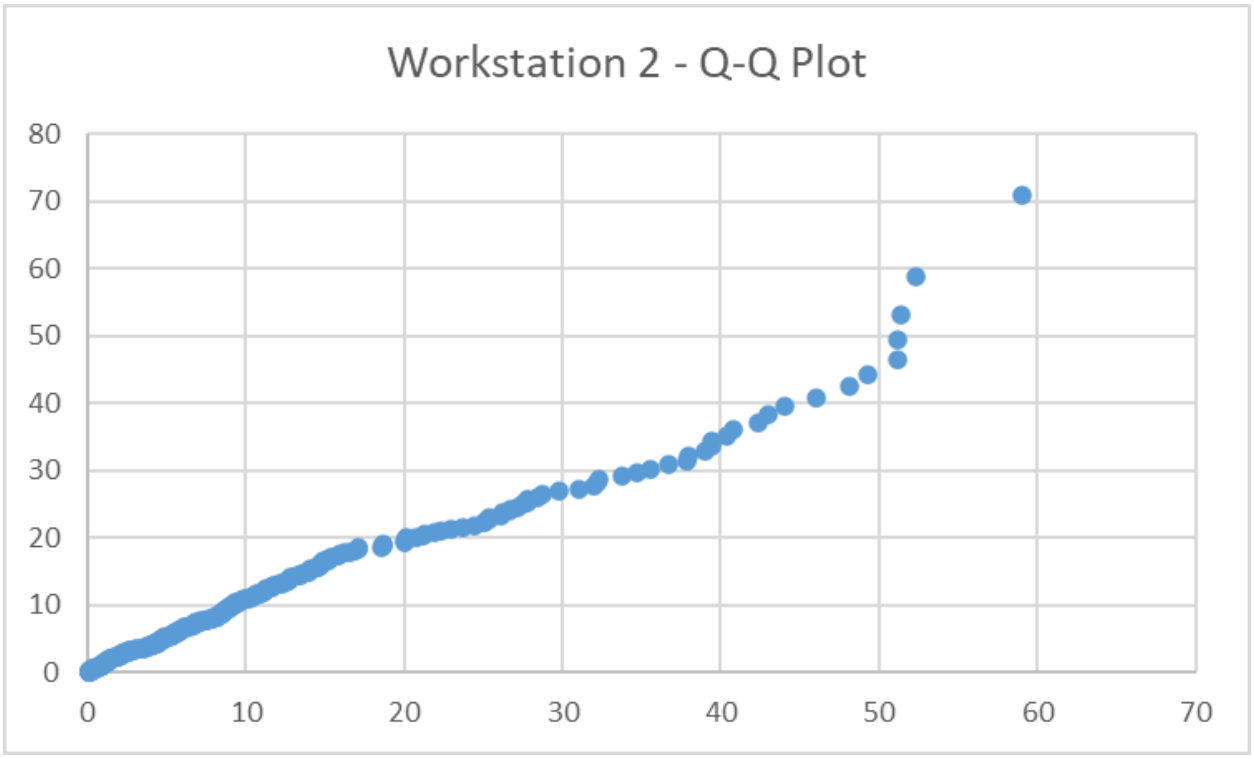
QQ and Chi-squared testing will be conducted using an exponential distribution to test the fit. If the exponential distribution is unsuccessful, a different distribution will be attempted. Then the inverse transform method of random variate generation will be conducted on the distribution. Below is a QQ plot of the sample data compared to an exponential distribution. This QQ plot is a straight line with a slight tail showing that the correct distribution was selected. The figures below represent the exponential Q-Q plot of the system entity's sample data.

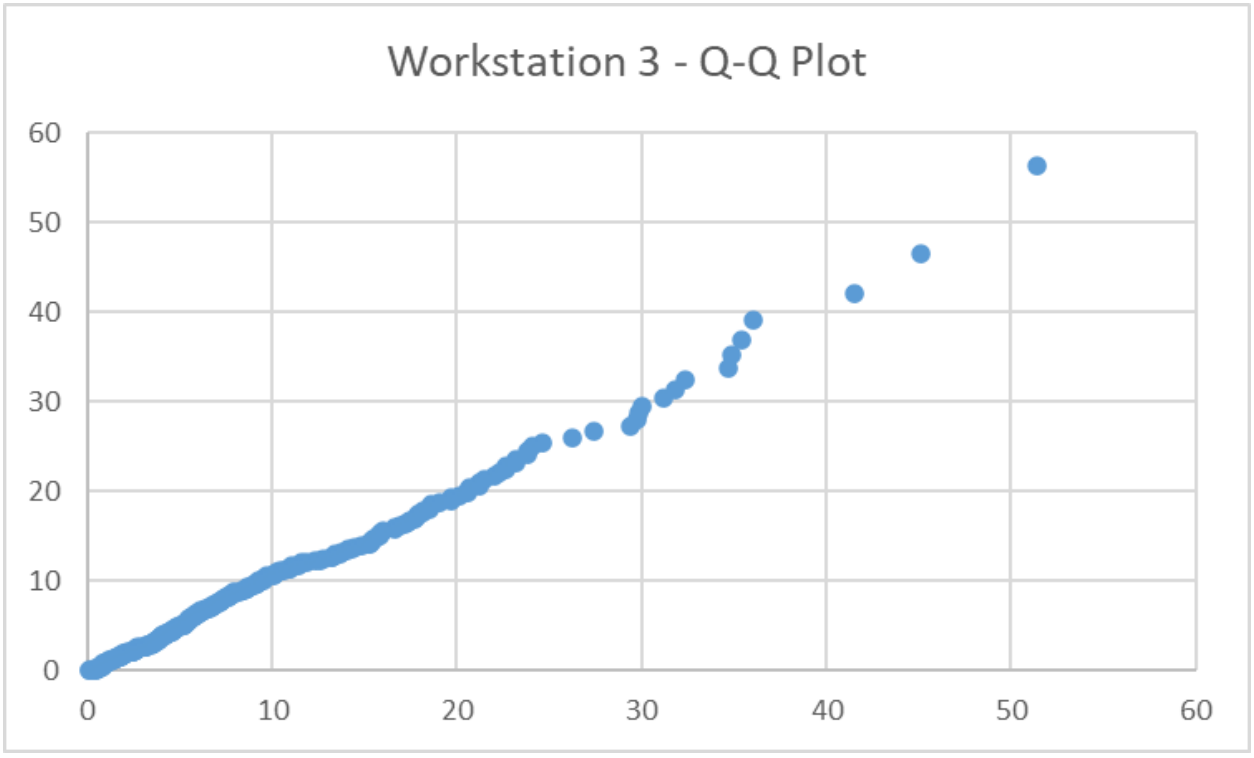

Inspector 1 - Q-Q plot
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
Q-Q plot


Inspector 2 Component 2 - Q-Q plot
120
100
80
60
40
20
0
0
20
40
60
80
100
120
140


Inspector 2 Component 3 - Q-Q Plot
140
120
100
80
60
40
20
0
0
20
40
60
80
100
120


Workstation 1 - Q-Q Plot
35
30
25
20
15
10
5
0
0
5
10
15
20
25
30
35


Workstation 2 - Q-Q Plot
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70


Workstation 3 - Q-Q Plot
60
50
40
30
20
10
0
0
10
20
30
40
50
60