BOĞAZİÇİ UNIVERSITY

IE 306 - SYSTEMS SIMULATION

Project 2

Aybek KORUGAN

Group 1

Baran Deniz KORKMAZ - 2015400183

Doğukan KALKAN - 2015400132

Mustafa Alparslan - 2009400189

SPRING 2020

Contents

1 Introduction

In real-life applications, choosing the representative distributions for input data is a challenging task. Yielding correct observations guided by outputs from the model will be possible only if the input data is analyzed appropriately.

In the project, we are asked to determine the inter-arrival time process that will later be used as input in building a model.

To achieve our goal, we are given a data collected from the real system of interest. The collection of data from a real system, if possible, constructs the first step of the development of a useful model of input data. Input modeling continues with the following steps:

1. Identification of a probability distribution to represent the input. Since a real-system data are available, this step includes the development of histogram of the data. The histogram allows the construction of a recognized distribution.

2. Choosing the values of parameters that defines the predetermined distribution. The sample statistics are used, whenever the real-system data are available.

3. Evaluation of the chosen distribution and the associated parameters. The goodness-of-fit tests including the Kolmogorov-Smirnov Test and Chi-Square Test enables testing the null-hypothesis. The rejectance must be followed by the repetition of the entire procedure declared.

In the project, we will be applying the procedure stated above for input modeling by using R language.

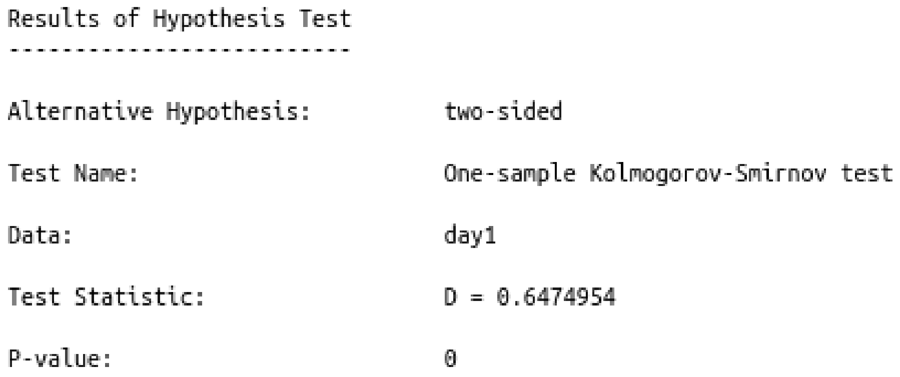
2 Task 1: Kolmogorov-Smirnov Test

We can deal with testing whether the given sample of input has the desired properties in terms of two perspectives. Accoringly, the tests can be given in two categories according to the properties of interest. These categories are as follows:

1. Frequency Tests: Two different methods of testing in this category are Kolmogorov-Smirnov and Chi-Square test. Both tests measure the compliance between the actual distribution of sample and the theoretical distribution.

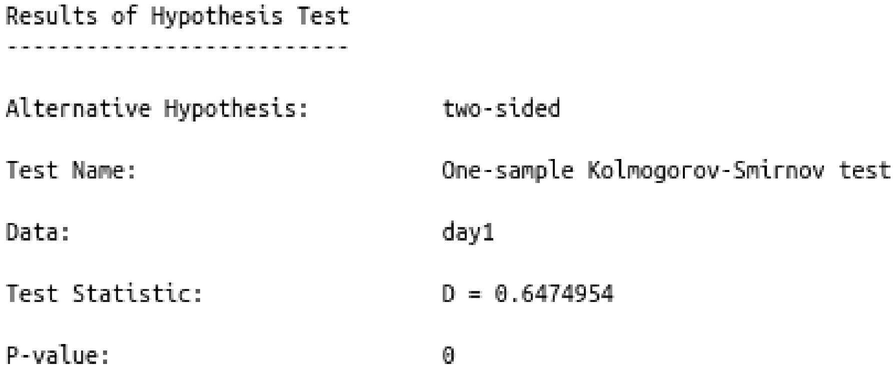
2. Autocorrelation Test: Tests the correlation between the numbers and compares the sample correlation to the desired correlation, zero.

In this section, we begin with testing whether the inter-arrival times are distributed uniformly between 0 and 400 seconds by using Kolmogorov-Smirnov test with the significance level of 0.05. The Kolmogorov-Smirnov test can be done by using **ks.test** built-in function which takes the data that we want to test, distribution type, and distribution-related parameters as input. Below, you can see the critical values (D) obtained by Kolmogorov-Smirnov tests on data samples named as ’day1’ and ’day2’ which stands for given real-system data for two days in excel file.



```
Results of Hypothesis Test
--------------------------

Alternative Hypothesis:         two-sided

Test Name:                      One-sample Kolmogorov-Smirnov test

Data:                           day1

Test Statistic:                 D = 0.6474954

P-value:                        0
```

Figure 1: Kolmogorov-Smirnov Test: Day 1 ($\alpha = 0.05$)



```
Results of Hypothesis Test
--------------------------

Alternative Hypothesis:         two-sided

Test Name:                      One-sample Kolmogorov-Smirnov test

Data:                           day1

Test Statistic:                 D = 0.6474954

P-value:                        0
```

Figure 2: Kolmogorov-Smirnov Test: Day 2 ($\alpha = 0.05$)

Below, you can find the table for Kolmogorov-Smirnov critical values. The degree of freedom in Kolmogorov-Smirnov test is defined by N-1 where N is the number of data samples. The size of both data representing the real-life system is 488, which means the degree of freedom is greater than 35. Observe that, for degree of freedoms over 35, we approximate the critical values by the formula $\frac{1.36}{\sqrt{N}}$ for $\alpha = 0.05$. Substituting N by 488, we obtain the critical value $D_{0.05} =$ 0.061564307. By Kolmogorov-Smirnov test, any test statistic D above critical value will result in the rejectance of null-hypothesis. We define the testing hypotheses as follows:

H_0 : The inter-arrival times are distributed uniformly between 0 and 400 seconds.,
H_1 : The inter-arrival times are not distributed uniformly between 0 and 400 seconds.

Observing the results obtained from the Kolmogorov-Smirnov test, we see that for both data samples, Day 1 and Day 2, we reject the null hypothesis. We conclude that both data samples are not distributed uniformly between 0 and 400 seconds, given that the test statistics D are above the critical value $D_{0.05}$ = 0.061564307.

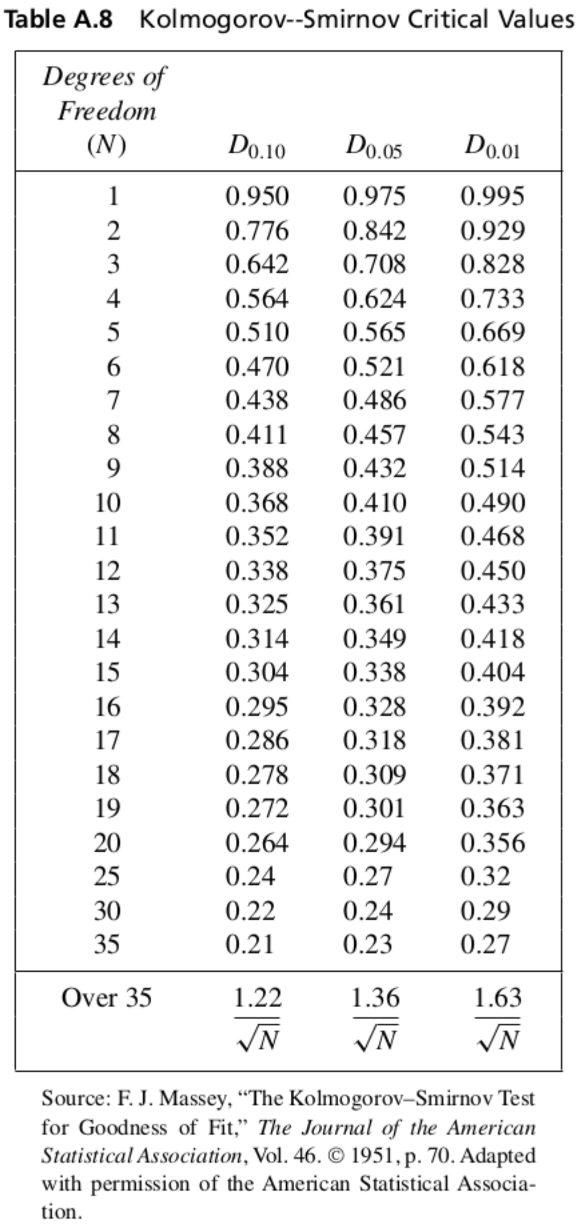


Table A.8 Kolmogorov--Smirnov Critical Values

Degrees of Freedom (N)	$D_{0.10}$	$D_{0.05}$	$D_{0.01}$
1	0.950	0.975	0.995
2	0.776	0.842	0.929
3	0.642	0.708	0.828
4	0.564	0.624	0.733
5	0.510	0.565	0.669
6	0.470	0.521	0.618
7	0.438	0.486	0.577
8	0.411	0.457	0.543
9	0.388	0.432	0.514
10	0.368	0.410	0.490
11	0.352	0.391	0.468
12	0.338	0.375	0.450
13	0.325	0.361	0.433
14	0.314	0.349	0.418
15	0.304	0.338	0.404
16	0.295	0.328	0.392
17	0.286	0.318	0.381
18	0.278	0.309	0.371
19	0.272	0.301	0.363
20	0.264	0.294	0.356
25	0.24	0.27	0.32
30	0.22	0.24	0.29
35	0.21	0.23	0.27
Over 35	$\frac{1.22}{\sqrt{N}}$	$\frac{1.36}{\sqrt{N}}$	$\frac{1.63}{\sqrt{N}}$

Source: F. J. Massey, "The Kolmogorov–Smirnov Test for Goodness of Fit," *The Journal of the American Statistical Association*, Vol. 46. © 1951, p. 70. Adapted with permission of the American Statistical Association.

Figure 3: Table: Kolmogorov-Smirnov test Critical Values

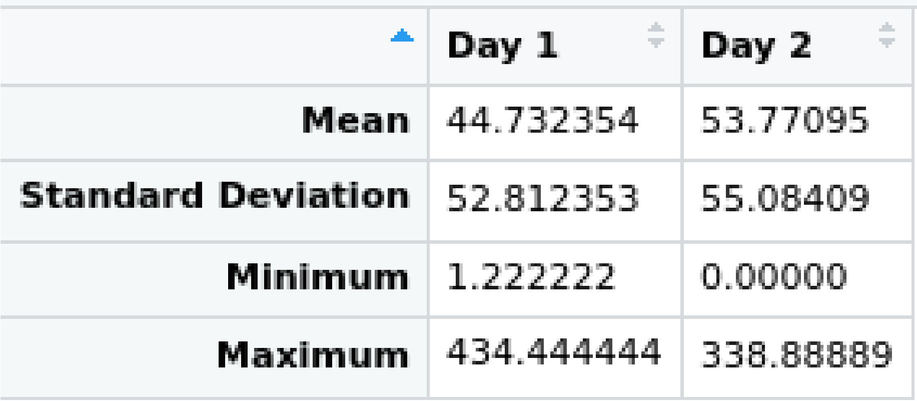
3 Task 2: Sample Statistics

As stated above, for any model to function properly meaning producing correct output, the input that is supplied to the model must be analyzed thoroughly. In real life, the analysis requires us to have different samples if possible, since analyzing more input samples always yields more result which helps us infer from the data in a more accurate way.

Ideally, we would like to know the exact values of mean, standard deviation and other statistics. In real life, this is always impossible. However, we can analyze different samples and make conclusions about them. These inferences and conclusions reflect the nature of our input.

Even though, sample statistics are not the same as population statistics, they are quite helpful in regards to making predictions. Eventually, our decision about the input and its properties is done by using sample statistics.

Thus, looking at Figure 4, we can see several statistics of our input. These statistics, especially mean and standard deviation, give us a slight idea about the true nature of our data. And, building upon this idea, our goal is to find a distribution that fits our input.



	Day 1	Day 2
Mean	44.732354	53.77095
Standard Deviation	52.812353	55.08409
Minimum	1.222222	0.00000
Maximum	434.444444	338.88889

Figure 4: Descriptive Statistics For Day 1 and Day 2

4 Task 3: Frequency Histograms

One technique is to visually predict the correct distribution of the input is to create frequency histograms with different number of intervals. As the number of intervals increase, the correct shape of the input becomes clear. For each interval, there is a corresponding frequency which might replicate a density function. Looking at the values, we are supposed to come up with a distribution that visually fits the histogram.

In **Task 1: Kolmogorov-Smirnov Test**, we observed that the inputs, both Day 1 and Day 2 do not fit a uniform distribution, which can be also visually inferred.

Looking at the histograms in Figure 5 and Figure 6 , on the other hand, the inputs resemble an exponential distribution. For 20 second intervals, this resemblance is much more obvious than the other two. Since as made clear above, as the number of intervals increase, the correct shape of the input becomes clear. Nonetheless, the inputs for Day 1 and Day 2, at least visually, fit an exponential distribution.

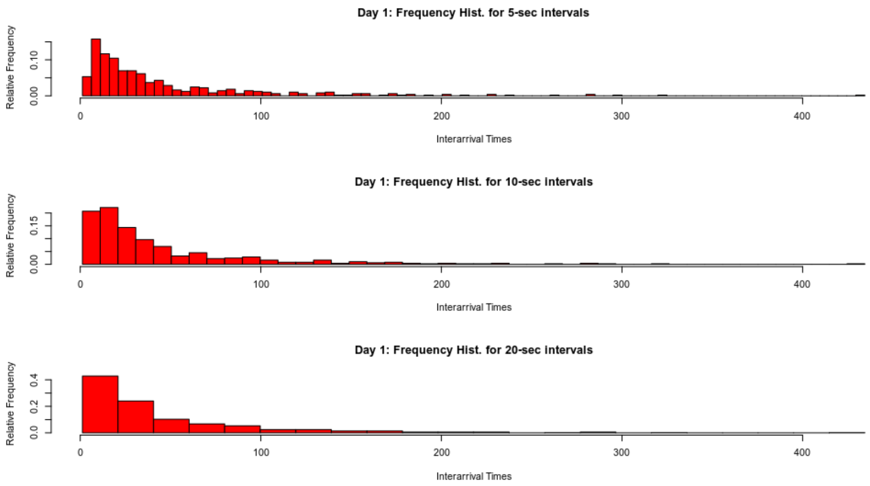
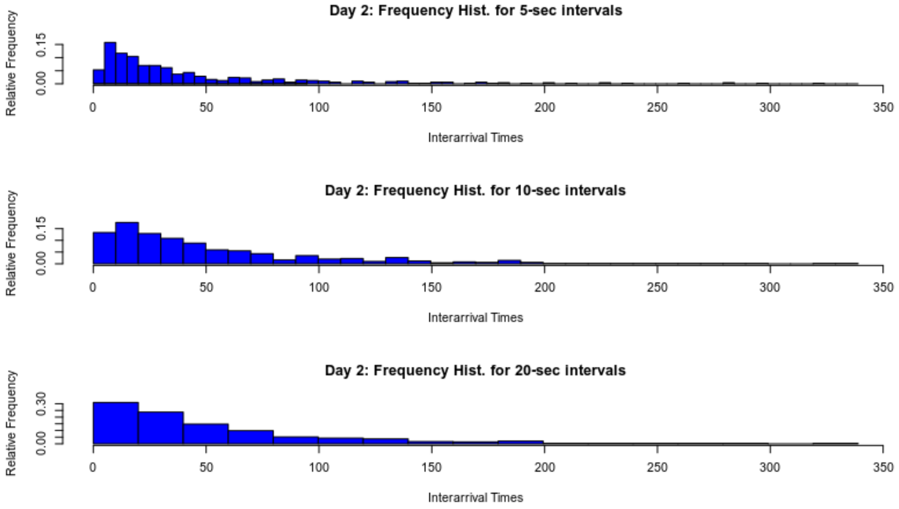


Figure 5: Frequency Histograms of Day 1 for 5, 10 and 20 Second Intervals


Day 2: Frequency Hist. for 5-sec intervals

Day 2: Frequency Hist. for 10-sec intervals

Day 2: Frequency Hist. for 20-sec intervals

Figure 6: Frequency Histograms of Day 2 for 5, 10 and 20 Second Intervals

5 Task 4: Chi-Square Test

As stated in the previous section **Task 3: Frequency Histograms**, the observations on histograms lead us into the idea of exponentially distributed inter-arrival time process with a mean equals to sample mean.

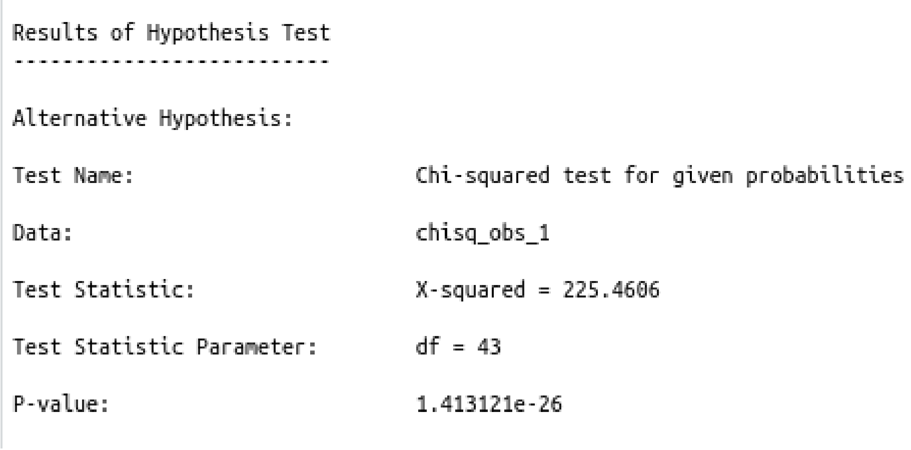
This section provides the evaluation of the idea that the inter-arrival times are distributed exponentially with a mean equals to sample mean, by using Chi-Square Test. As denoted in **Task 1: Kolmogorov-Smirnov Test**, another test used to measure the compliance between the actual distribution of sample and the theoretical distribution is Chi-Square test. This test formalizes the intuitive idea we obtained by observing the histogram by comparing the intervals divided within the histogram to the candidate density function. The test statistic is given by

$$X_0^2 = \textstyle\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency in the ith class and E_i is the expected frequency in that class interval.

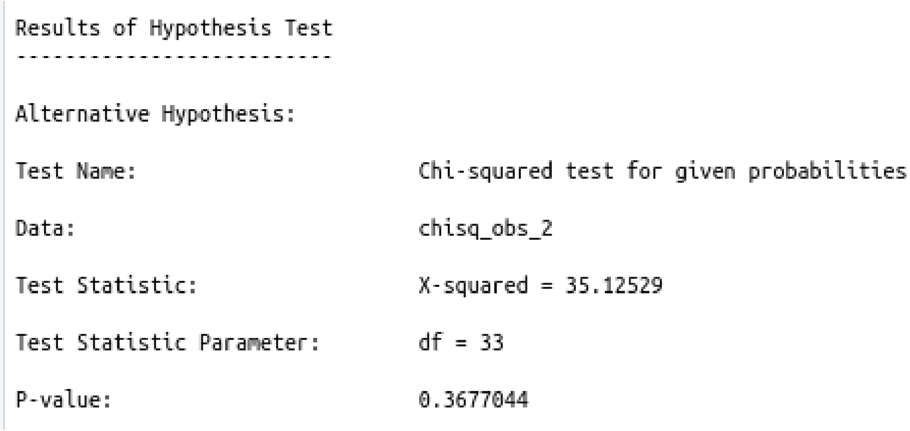
Applying Chi-Square test requires the explicit declaration of intervals that divide the data samples, and the number of observed data samples within that intervals. When requirements are satisfied, the built-in function **chisq.test** which takes two lists specifying the interval boundaries and number of observed instances within those classes as parameters, applies Chi-Square test. An option argument that enables rescale of given lists to fit these intervals into the cumulative density function is set into TRUE as well.

Below, you can see the results obtained by Chi-Square test for both data samples, i.e. 'Day 1' and 'Day 2'.



```
Results of Hypothesis Test
--------------------------

Alternative Hypothesis:

Test Name:                       Chi-squared test for given probabilities

Data:                            chisq_obs_1

Test Statistic:                  X-squared = 225.4606

Test Statistic Parameter:        df = 43

P-value:                         1.413121e-26
```

Figure 7: Chi-Square Test: Day 1



```
Results of Hypothesis Test
--------------------------

Alternative Hypothesis:

Test Name:                        Chi-squared test for given probabilities

Data:                             chisq_obs_2

Test Statistic:                   X-squared = 35.12529

Test Statistic Parameter:         df = 33

P-value:                          0.3677044
```

Figure 8: Chi-Square Test: Day 2

By Chi-Square test, any test statistic X_0^2 above critical value will result in the rejectance of null-hypothesis. We define the testing hypotheses as follows:

H_0 : The inter-arrival times are distributed exponentially with sample mean.,
H_1 : The inter-arrival times are not distributed exponentially with sample mean.

Given that the degrees of freedom is calculated by k-s-1 where k is number of intervals, and s is the number of parameters of the hypothesized distribution estimated by the sample statistics, in our case s equals 1 since the exponential distribution requires estimation for mean only, the degrees of freedom are 43 and 33, respectively.

Observing the results obtained from the Chi-Square test, we see that the values of test statistic are 225.4606 and 35.12529 with degree of freedom 43 and 33 respectively for Day 1 and Day 2. Observing the table for percentage points of the Chi-Square distribution, we conclude that we reject the null hypothesis for Day 1, and fail to reject the null hypothesis for Day 2. Also by observing the p-values derived in tests, we reject the null hypothesis for Day 1 since it is close to 0, and fail to reject the null hypothesis for Day 2 since the p-value is above 0.05.

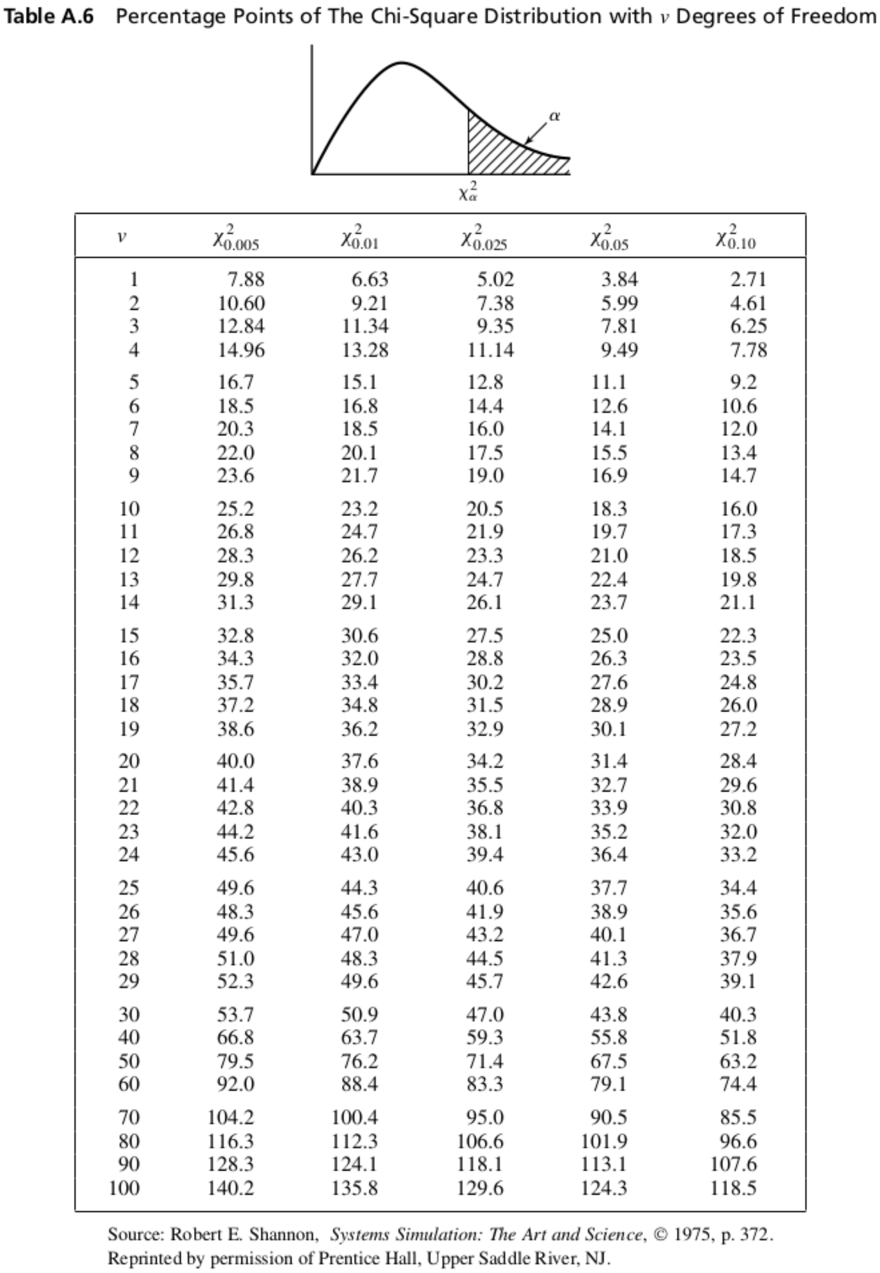


Table A.6 Percentage Points of The Chi-Square Distribution with v Degrees of Freedom

α
χ^2_α

v	$\chi^2_{0.005}$	$\chi^2_{0.01}$	$\chi^2_{0.025}$	$\chi^2_{0.05}$	$\chi^2_{0.10}$
1	7.88	6.63	5.02	3.84	2.71
2	10.60	9.21	7.38	5.99	4.61
3	12.84	11.34	9.35	7.81	6.25
4	14.96	13.28	11.14	9.49	7.78
5	16.7	15.1	12.8	11.1	9.2
6	18.5	16.8	14.4	12.6	10.6
7	20.3	18.5	16.0	14.1	12.0
8	22.0	20.1	17.5	15.5	13.4
9	23.6	21.7	19.0	16.9	14.7
10	25.2	23.2	20.5	18.3	16.0
11	26.8	24.7	21.9	19.7	17.3
12	28.3	26.2	23.3	21.0	18.5
13	29.8	27.7	24.7	22.4	19.8
14	31.3	29.1	26.1	23.7	21.1
15	32.8	30.6	27.5	25.0	22.3
16	34.3	32.0	28.8	26.3	23.5
17	35.7	33.4	30.2	27.6	24.8
18	37.2	34.8	31.5	28.9	26.0
19	38.6	36.2	32.9	30.1	27.2
20	40.0	37.6	34.2	31.4	28.4
21	41.4	38.9	35.5	32.7	29.6
22	42.8	40.3	36.8	33.9	30.8
23	44.2	41.6	38.1	35.2	32.0
24	45.6	43.0	39.4	36.4	33.2
25	49.6	44.3	40.6	37.7	34.4
26	48.3	45.6	41.9	38.9	35.6
27	49.6	47.0	43.2	40.1	36.7
28	51.0	48.3	44.5	41.3	37.9
29	52.3	49.6	45.7	42.6	39.1
30	53.7	50.9	47.0	43.8	40.3
40	66.8	63.7	59.3	55.8	51.8
50	79.5	76.2	71.4	67.5	63.2
60	92.0	88.4	83.3	79.1	74.4
70	104.2	100.4	95.0	90.5	85.5
80	116.3	112.3	106.6	101.9	96.6
90	128.3	124.1	118.1	113.1	107.6
100	140.2	135.8	129.6	124.3	118.5

Source: Robert E. Shannon, *Systems Simulation: The Art and Science*, © 1975, p. 372. Reprinted by permission of Prentice Hall, Upper Saddle River, NJ.

Figure 9: Table: Chi-Square test Critical Values

6 Task 5: Quantile-Quantile Plots

In statistics, a Q–Q plot is a useful tool for evaluating distribution fit. The procedure for obtaining Q-Q plots is as follows. The sample data must be sorted in ascending order first, then we calculate the theoretical intervals that comes from the assumed distribution and compare these theoretical intervals to the actual intervals we have observed in our sample data.

Below, you can see the quantile-quantile plots we have obtained. We used both R and Excel for double-checking the Q-Q plots, since our interpretations on the figures will be deterministic on the determination of appropriate distribution.

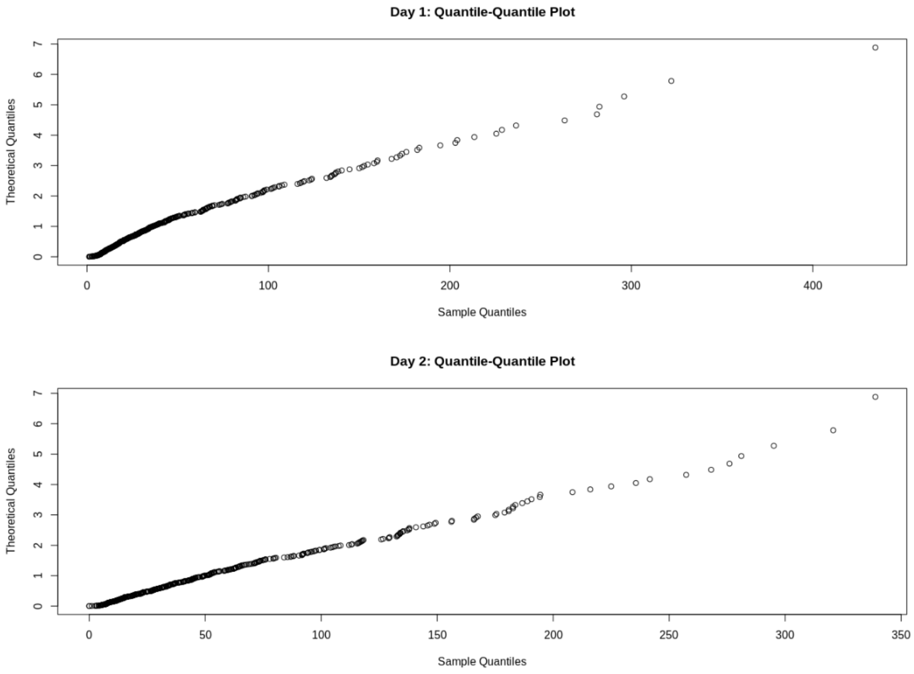


Figure 10: Q-Q Plots by R

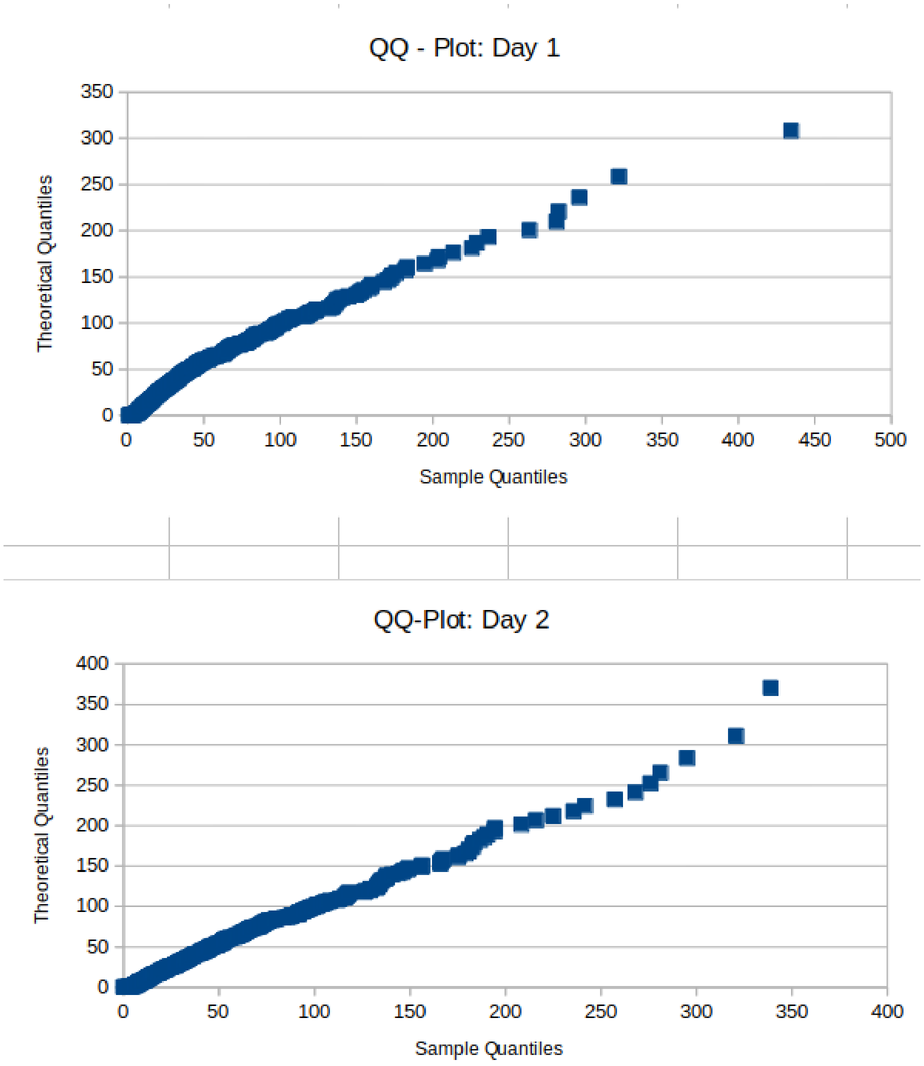


Figure 11: Q-Q Plots by Excel

We see that the shape of figures derived by both resources are exactly the same. Let us briefly explain the logic behind Q-Q plots. By dividing N many intervals, we imitate a sample that is exactly derived from the assumed distribution, which is exponential distribution in our case, by inverse of cumulative density function. Then we measure the degree of agreement between the actual and theoretical quantiles. In case of a perfect match, we must obtain a linear line with slope 1. However, any nonlinear trend or deviations imply non-appropriately assumed distribution. Looking at our figures, we see a nonlinearity in the left-hand side of day 1. However, day 2 presents much linear shape than the first one. We must also note that the observations on q-q plots are more subjective than the rest of our work. We conclude that day 2 fits much better than day 1 into the assumed exponential distribution, in the light the results we have obtained from the chi-square test.

7 Task 6: Analysis of Observation Times

One of the most important aspects in input analysis is to check the stationarity of the input. Ideally, we want our input to be stationary. A stationary data has a property that the mean, standard deviation and autocorrelation do not change over time. In input analysis, stationary data is always desired. Using non-stationary data in almost all models produces unreliable results. Also, A non-stationary process can be transformed to a stationary process by performing different techniques according to the nature of the non-stationarity.

Observing the results in Figure 7, we do not see any large fluctuation in the time series data for both Day 1 and Day 2. Also, the changes in time seem to be stable which means both data follow a pattern. Thus, we can safely state that the data is stationary.

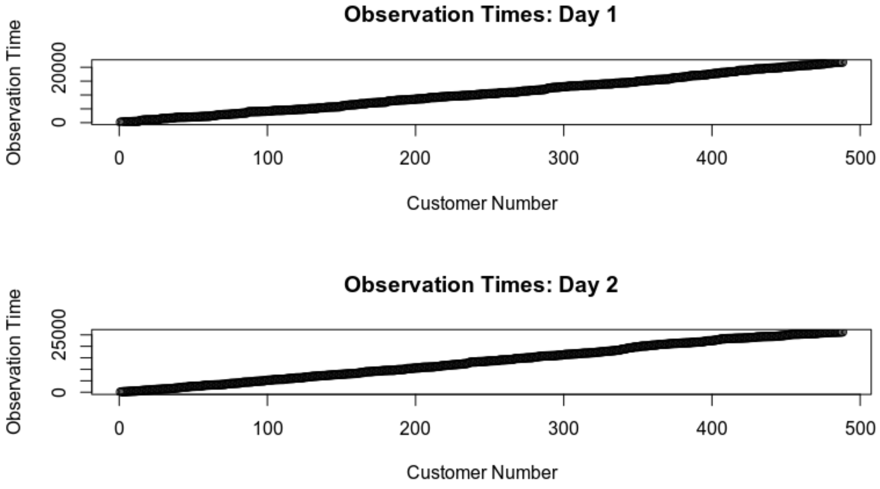


Figure 12: Inter-arrival Times with Respect to Observation Times for Day 1 and Day 2

8 Task 7: Autocorrelation

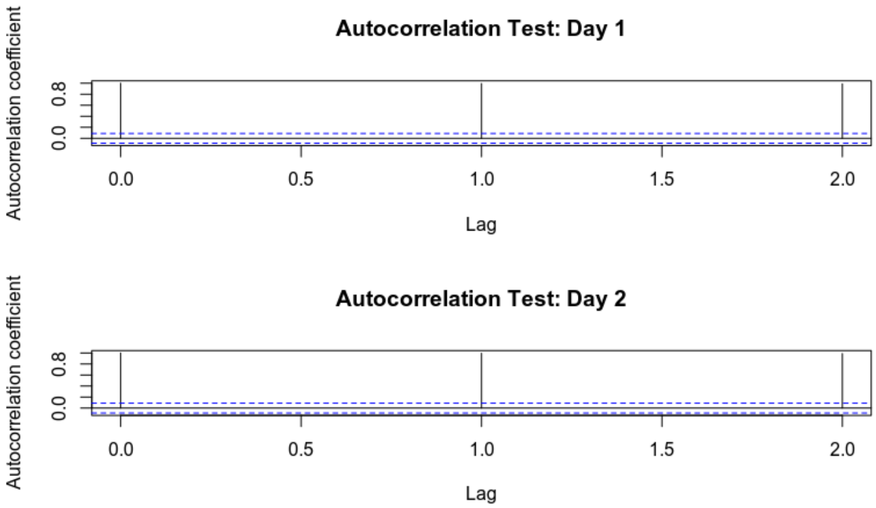
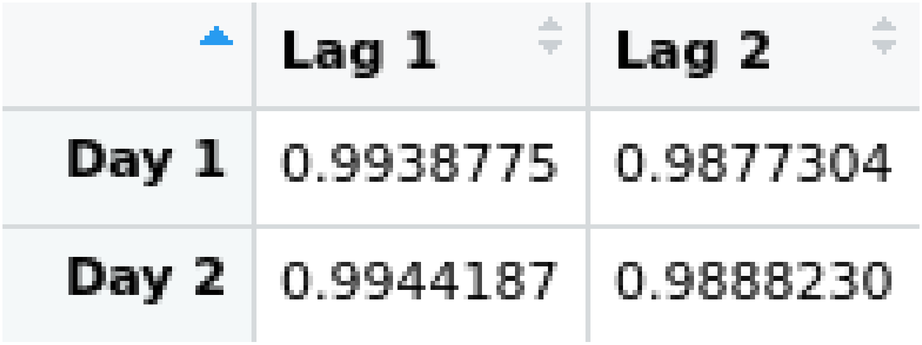


Figure 13: Lags versus Autocorrelation Coefficients for Day 1 and Day2-Plot



	Lag 1	Lag 2
Day 1	0.9938775	0.9877304
Day 2	0.9944187	0.9888230

Figure 14: Autocorrelation Test for Day 1 and Day 2-Table

As explained in **Task 6: Analysis of Observation Times**, one condition for an input to be stationary is that it should be autocorrelated. Autocorrelation means observations done in one period of time will be also done later in time which means the data will be autocorrelated. Looking at Figure 12, for both Day 1 and Day 2, the autocorrelation coefficients are very close to 1 and the differences between lag 1 and lag 2 autocorrelation coefficients are so small, which suggests that the input is autocorrelated.

9 Auxiliary Functions

1. ks.test
   Description: Performs one or two sample Kolmogorov-Smirnov tests.
   Parameters:

   - x: A numeric vector of data values.
   - y: A numeric vector of data values, or a character string naming a cumulative distribution function or an actual cumulative distribution function such as pnorm. Alternatively, y can be an ecdf function (or an object of class stepfun) for specifying a discrete distribution.
   - Parameters of the distribution specified (as a character string) by y.

2. chisq.test
   Description: Performs chi-squared test.
   Parameters:

   - x: A numeric vector for actual observations in intervals.
   - y: A numeric vector for expected observations in intervals.

3. qexp
   Description: Quantile function for the exponential distribution with rate rate (i.e., mean 1/rate).
   Parameters:

   - q: Vector of quantiles.

4. ppoints
   Description: Generates the sequence of probability points.

5. hist
   Description: The generic function hist computes a histogram of the given data values.
   Parameters:

   - x: A vector of values for which the histogram is desired.
   - breaks: A vector giving the breakpoints between histogram cells,

6. acf
   Description: The function acf computes (and by default plots) estimates of the autocovariance or autocorrelation function.
   Parameters:

   - x: A univariate or multivariate (not ccf) numeric time series object or a numeric vector or matrix, or an ”acf” object.

- lag.max: Maximum lag at which to calculate the acf. Default is $10log_{10}(N/m)$ where N is the number of observations and m the number of series. Will be automatically limited to one less than the number of observations in the series.

10 References

- Text Book: Discrete-Event System Simulation(5th Edition) [Jerry Banks, John S. Carson II, Barry L. Nelson, David M. Nicol]