IE 306.02 - System Simulation

Homework 2

Aslı Aykan, Algı Kanar, Ömer Faruk Deniz

2016400222, 2016400123, 2016400003

Computer Engineering Department
Boğaziçi University

June 6, 2021

Contents

1 Answers

1.1 Question 1

In this question, we are asked to calculate general descriptive statistics such as mean, standard deviation etc. We have also calculated Kurtosis and Skewness in addition to the usual statistics. Skewness is used to determine whether the data is symmetric or not. Kurtosis, on the other hand, is used to see whether the data is heavy-tailed or not compared to a normal distribution.

Mean:	217
Max:	1141
Min:	1
Count(N):	133
Sum:	28807
Standard Error:	20,03910108
Median:	137
Mode:	159,6342771
Standard Deviation:	231,1021875
Sample Variance:	53408,22109
Kurtosis:	4,047394918
Skewness	1,890216849
Range:	1140

Figure 1.1: Q1: Descriptive Statistics for Day 1

Mean:	234
Max:	1066
Min:	1
Count(N):	122
Sum:	28607
Standard Error:	19,27274817
Median:	174
Mode:	#N/A
Standard Deviation:	212,8744613
Sample Variance:	45315,53628
Kurtosis:	3,002250122
Skewness	1,595853956
Range:	1065

Figure 1.2: Q1: Descriptive Statistics for Day 2

1.2 Question 2

In this question, we are expected to perform a Kolmogorov-Smirnov test with $\alpha = 0.05$ to test the uniform distribution of the data. Since the data has values greater than 400 it is impossible to be distributed between 0 and 400. However, this test can be applied to the values between 0 and 400 in the data. To apply the test to that range, first, we have selected values between 0 and 400 and sorted them. Then we have scaled them into 0-1 range. Then we have calculated the R_i, i/N, $i/N - R_i$ and $R_i - (i-1)/N$ values for each cell. Then we have calculated $D+ = max(i/N - R_i)$ and $D- = max(R_i - (i-1)/N)$. Then we have calculated $D = max(D+, D-)$ and if it is bigger than the $D_{\alpha,N}$, we reject the uniformity hypothesis. Let's compare these statistics for Day 1 and Day 2.

- For Day 1, there are 112 data points in the 0-400 range. $D = 0,289$ and $D_{0.05,112} = 0,1285$. Since $D > D_{0.05,112}$, we reject the uniformity hypothesis in the 0-400 range.
- For Day 2, there are 102 data points in the 0-400 range. $D = 0,1997$ and $D_{0.05,102} = 0,1346$. Since $D > D_{0.05,102}$, we reject the uniformity hypothesis in the 0-400 range.

1.3 Question 3

We can see from the histograms that there is a downtrend in the frequencies along the histogram. As bin size increases, similar looking bins are combined and the trend in the histograms become clearer. Since this downtrend resembles an exponential decay in the frequencies, we can see the density pattern of an exponential distribution in the histograms.

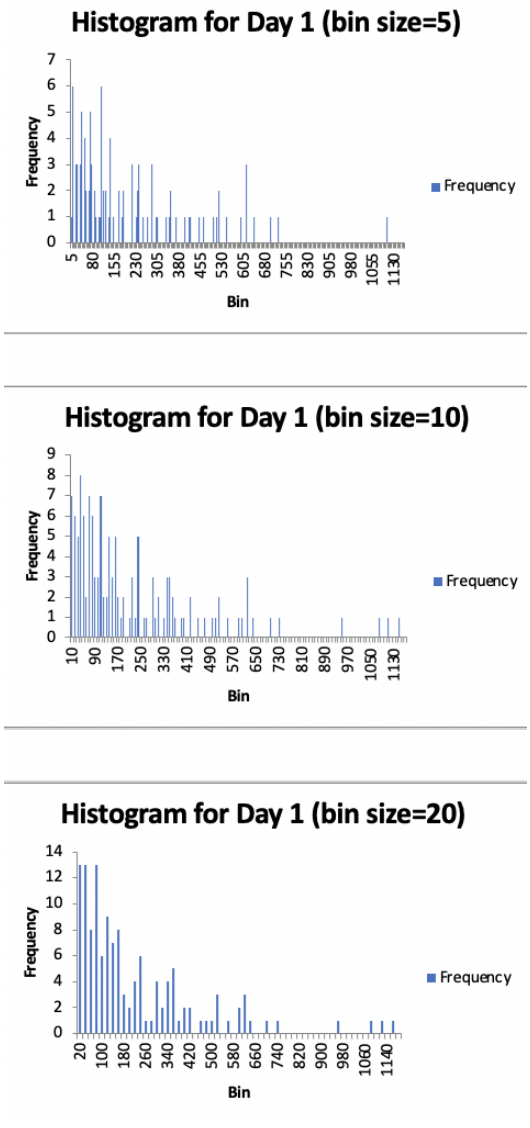

Figure 1.3: Q3: Histograms for Day 1

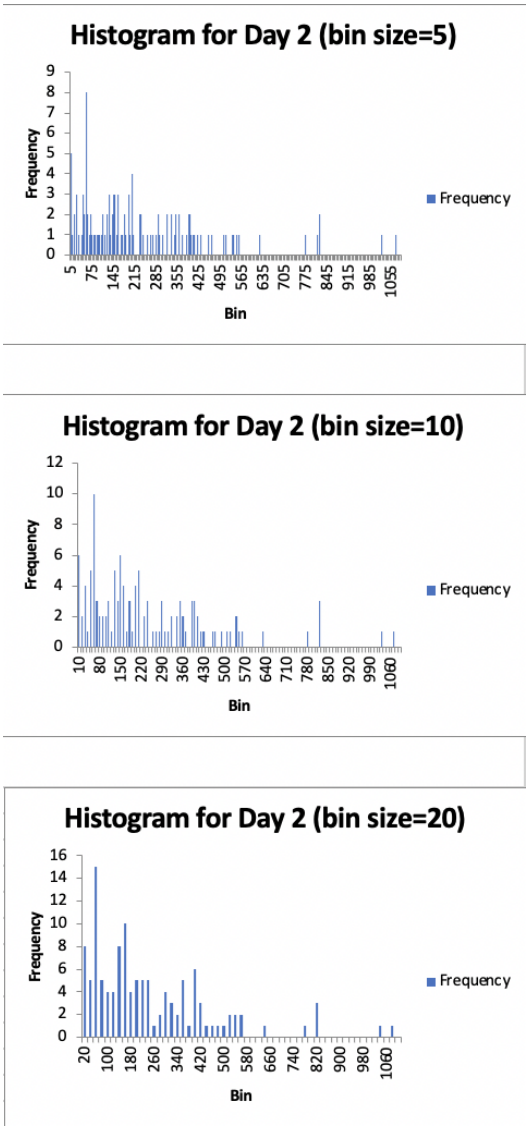

Figure 1.4: Q3: Histograms for Day 2

1.4 Question 4

In this question, we are expected to perform a Chi-Square test for exponential distribution with bin size 10 and $\alpha = 0.05$. In order to do that, we have to find observed frequencies, expected frequencies and $(O_i - E_i)^2/E_i$ for each bin. Excel will easily handle the first and the third but we need to derive the formula for the second one. To find the expected frequency, we can multiply the probability of being in that interval with total number of arrivals. We can calculate the probability of being in an interval (a,b) as F(b)-F(a) where F is the Cumulative Density Function(CDF) of the exponential distribution. The formula of this CDF function is $F(x) = 1 - e^{-\lambda * x}$ where $\lambda = 1/mean$ and the $mean$ is the value we have calculated from the data in the first question. After calculating the observed and expected frequencies, we can easily calculate $(O_i - E_i)^2/E_i$ for each bin and sum of these values will give us the χ^2 test statistic. For both days, we have calculated the degree of freedom for the test statistic from the $k - s - 1$ formula where k is the # of bins and $s = 1$ since λ is estimated from the data.

Day 1

When we calculate the χ^2 for Day 1, we obtain 170,4113125 where $\chi^2_{0.05,114} = 140$. Since $\chi^2 > \chi^2_{0.05,114}$, we should reject the exponential hypothesis. However, running the Chi-Square test as explained above has a deficiency which is that it has created bins that have expected frequency less than 5. This situation harms the reliability of the test and one way to improve the accuracy of the test is merging the cells having values less than 5. After this merge, $\chi^2 = 9.764$ and $\chi^2_{\alpha,dof} = 26.3$, **therefore we fail to reject the exponential hypothesis for Day 1.**

Day 2

When it comes to Day 2, we obtain $\chi^2 = 155.185$ and $\chi^2_{\alpha,dof} = 131$ and should reject exponential hypothesis. However, same problem also appears therefore we merged cells having expected frequencies less than 5. After that, we obtain $\chi^2 = 19735$ and $\chi^2_{\alpha,dof} = 26.3$ **therefore we fail to reject the exponential hypothesis for Day 2.**

1.5 Question 5

As we learned in class, a QQ (quantile-quantile) plot is needed to understand whether our data fits a distribution such as normal distribution or exponential distribution. For the exponential distribution case, our hypothesis is our data comes from an exponential distribution. To test whether our data is in exponential distribution, we need to use $F(x) = 1 - e^{\lambda * x}$ formula, which is the formula for exponential cumulative distribution function and calculate $F^{-1}(x)$, which is equal to $-(1/\lambda)ln(1 - Ri)$. This value is called the expected value and should be equal to the observed value. To compare these values, QQ plot is used, in which x-line corresponds to the observed values and y-line corresponds to the expected(hypothesized) values. If it is exact, it is expected to be a linear line whose slope is equal to 1. We compare the points with this line, and if the points are far away from the line, we can say that the hypothesis is not true.

As the first step of creating a QQ plot, we sorted our data in ascending order (ordered statistic). Each point in our data should be in equal intervals. Then we found , which is equal to 1/mean.

To find the percentile value, we used the formula that we learned in class, which is $(j - 0.5)/n$. (We also used our rank column to take j's). With $F^{-1}(x)$ formula, we calculated the inverse exponential cumulative distribution function of each percentile values and compared them with our data points.

As can be seen from our QQ plots for day 1 and day 2, since the points are not away from the line, **we fail to reject our hypothesis, which was "our data comes from the exponential distribution"**.

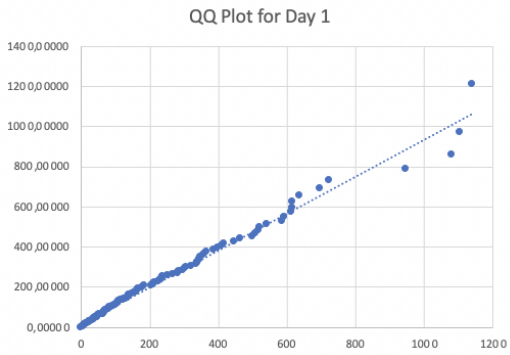

Figure 1.5: Q5: QQ Plot Day 1

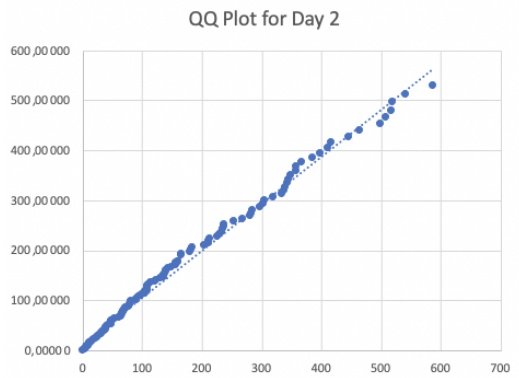

Figure 1.6: Q5: QQ Plot for Day 2

1.6 Question 6

In order to analyze the stationarity of the data, we have created histograms that shows the # of arrivals in a portion of time which are 30-minutes, 1-hour and 2-hours in our case. We have selected these values since each day's data is collected in the 8-hour period.

Considering the 30-min and 1-hour histograms of both days, we can see that there are increases and decreases in the # of arrivals but they do not form a trend. Therefore, there seems no repeating change in the data with respect to time and these histograms show stationary behaviour. However, when we look at the histograms with 2-hour bins for both days, we can observe a downtrend in the # of arrivals. Therefore, we can say that histograms with 2-hour bins shows a non-stationary behaviour.

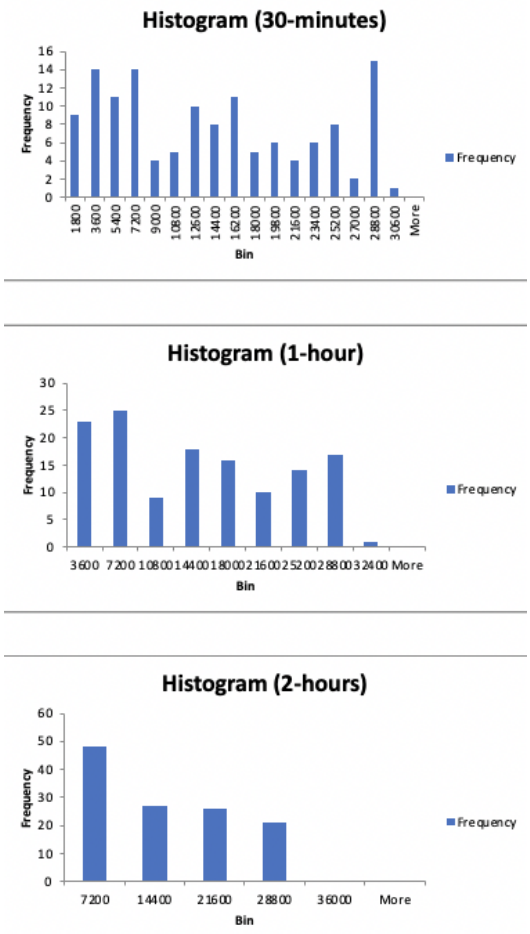

Figure 1.7: Q6: Interarrival Time vs. Observation Time for Day 1

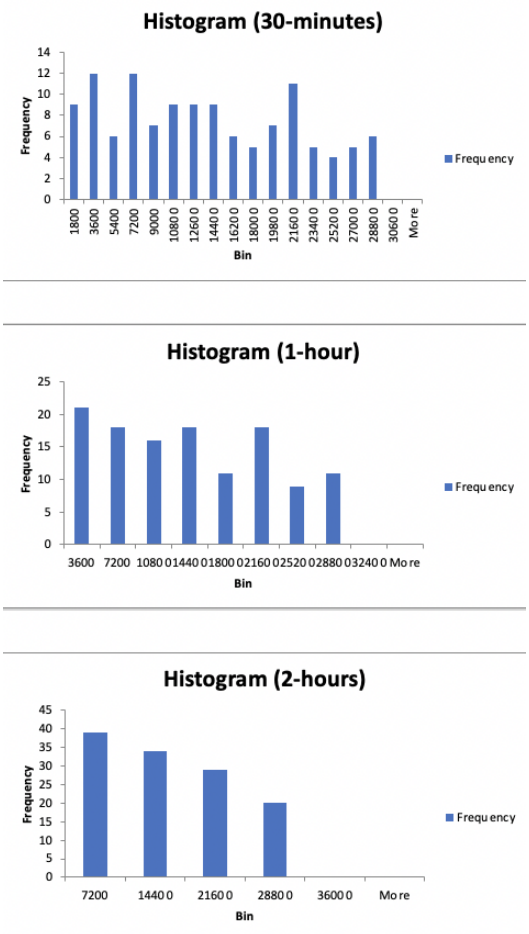

Figure 1.8: Q6: Interarrival Time vs. Observation Time for Day 2

1.7 Question 7

As we learned in class, we created two lag columns for lag 1 and lag 2 by shifting the data down by one cell and two cells. To plot the lag differences, we subtracted these values from each points in a day data.

To test whether the data is auto-correlated, we need to find correlation values. We used this formula to calculate lag 1 correlation: $\frac{\sum_{i=1}^{i-1}(r_i - mean)*(r_{i+1} - mean)}{\#\ of\ data\ points * variance}$

For lag 2 correlation, instead of i & i+1 pair in the formula, we used i & i+2 pair. The results are as follows:

The Correlation Value for Day 1 & Lag 1 : 0,1837027
The Correlation Value for Day 1 & Lag 2 : -0,0080518
The Correlation Value for Day 2 & Lag 1 : 0,24917
The Correlation Value for Day 2 & Lag 2 : 0,05713337

As can be seen, all values are between -1 and 1 as expected. However, the range for meaningful auto-correlation values is between -2σ and +2σ, in which σ is equal to $1/\sqrt{num.of observations}$.

For day 1, $|2\sigma| = 0.17342$
For day 2, $|2\sigma| = 0.18107$

By looking at these values, since the correlation values for day 1 & lag 1 and day 2 & lag 1 are greater than the related upper range value for auto-correlation, we can say there exist auto-correlation for these values.

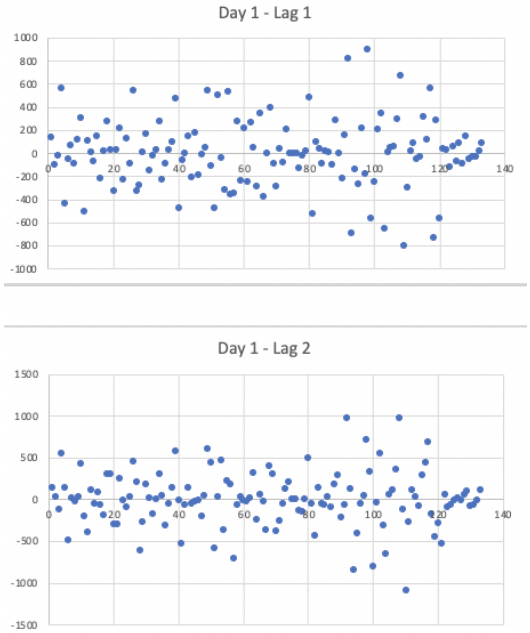

Figure 1.9: Q7: Differences for Day 1

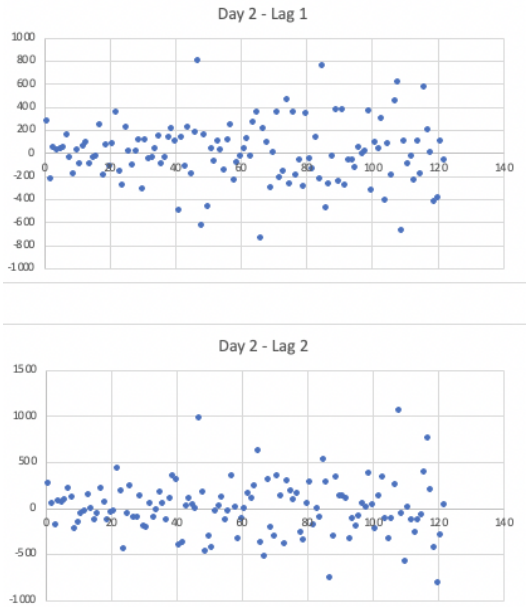

Figure 1.10: Q7: Differences for Day 2

1.8 Generating Random Variates

In this section, we are expected to generate random variates after fitting a distribution to the data. For both days, we have rejected the uniformity hypothesis on the Kolmogorov-Smirnov test for the 0-400 range, failed to reject the exponential hypothesis on the Chi-Square test and QQ plot and observed density pattern of the exponential distribution in the histograms of Question 3. Therefore, we have determined that arrival process is a Poisson Process and interarrival times are exponentially distributed for both days where $\lambda_1 = 0,00461$ and $\lambda_2 = 0,00426$.

When it comes to generating the exponential random variates, we will use the Inverse-Transform Technique and find these variates by mapping the CDF of the exponential distribution to the random numbers between 0-1. For that purpose, we have used the $X_i = F^{-1}(R_i) = \frac{-ln(1-R_i)}{\lambda}$ formula to generate the random variates after generating 100 random numbers for each day.