Karahan Sarıtaş - 2018400174
Yılmaz Burak - 2018400237
Mustafa Atay - 2020400333

IE 306 Group 11 - HW2 Report

1) **Find sample mean, standard deviation and other descriptive statistics that you deem appropriate.**

Day 1 - Statistics	Day 2 - Statistics	2 Days - Statistics
Sample size: 109	Sample size: 111	Sample size: 220
Sample mean: 265	Sample mean: 257	Sample mean: 261
Sample variance: 85328.929	Sample variance: 57070.498	Sample variance: 70759.8936
Sample standard deviation: 292.11116	Sample standard deviation: 238.89432	Sample standard deviation: 266.0073187
Sample coefficient of variation: 1.103794	Sample coefficient of variation: 0.929226	Sample coefficient of variation: 1.019842
Sample median:179	Sample median:182	Sample median: 179
Minimum: 1	Minimum: 4	Minimum: 1
Maximum: 1869	Maximum: 1297	Maximum: 1869
Standard Error: 27.97917	Standard Error: 22.674846	Standard Error: 17.9420977

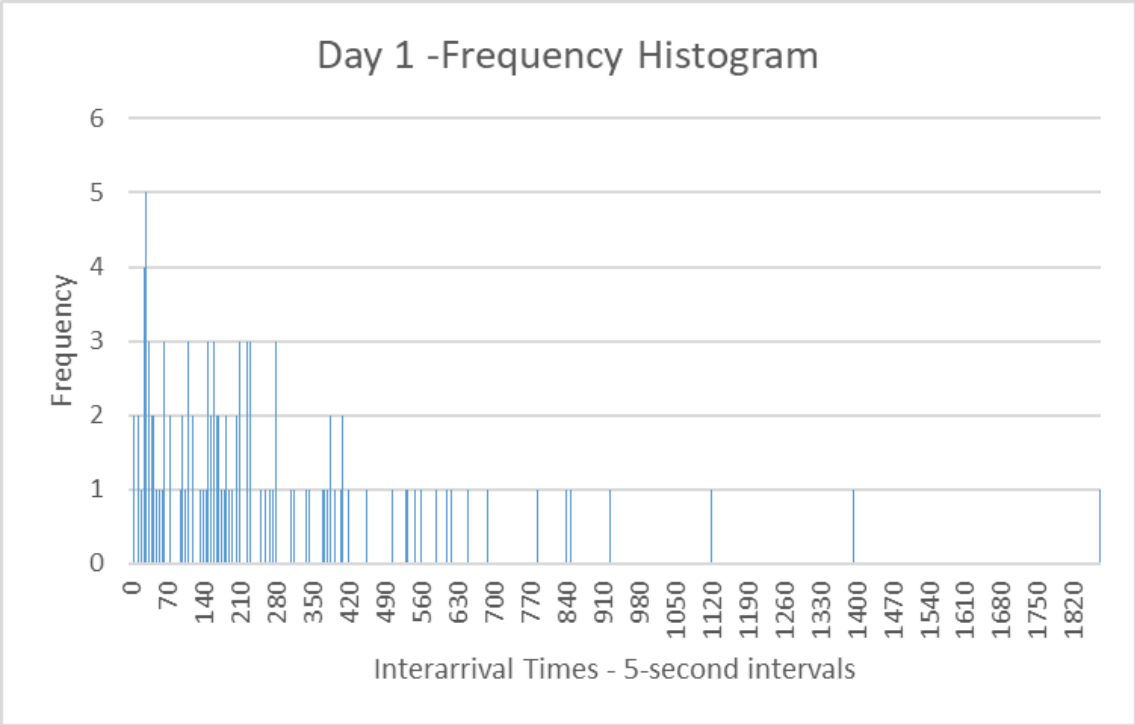
2) **One of the managers claims that it is safe to assume that inter-arrival times are distributed normally with mean 200 seconds and standard deviation 50 seconds. Test the validity of this claim using the Kolmogorov-Smirnov test with a significance level of 0.05.**

From the Kolmogorov-Smirnov test, we found the D value is 0.309744049.
Critical value is calculated as D_0.05 = 1.36 / SQRT(sample_size) = 1.36 / SQRT(220)= 0.09169118.
Because D > D_0.05, we reject the hypothesis that inter-arrival times are distributed normally with mean 200 seconds and standard deviation 50 seconds. Details of the calculations are shown in the excel file.

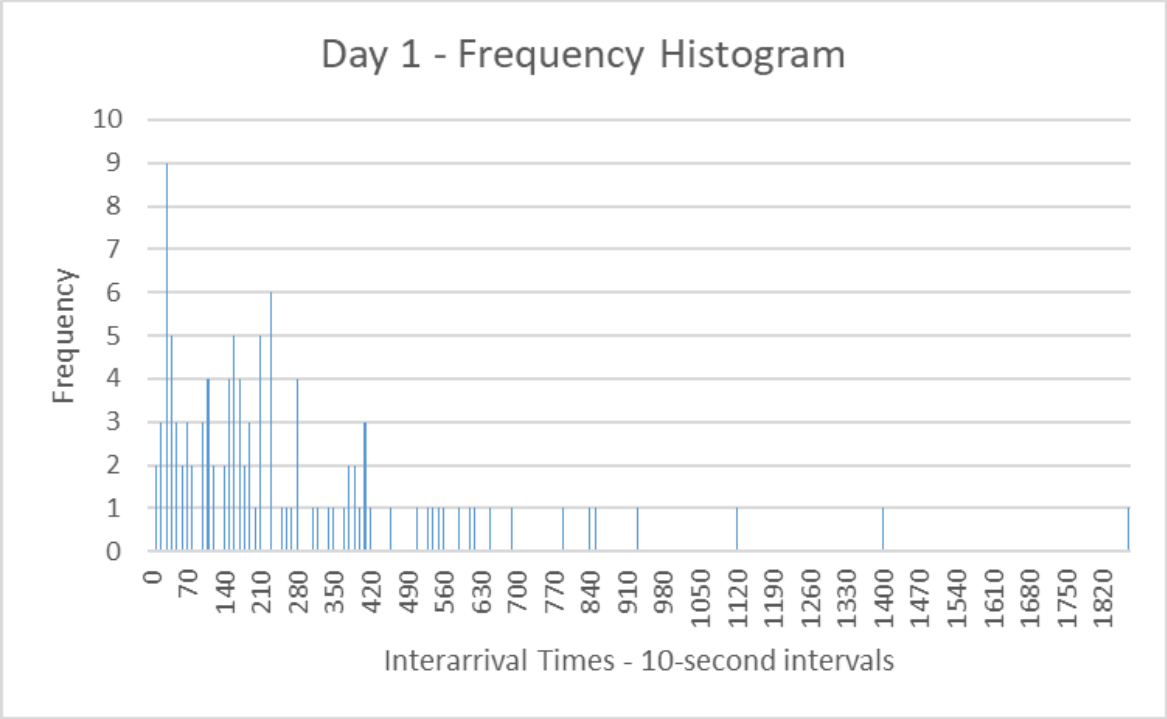
3) **Draw frequency histograms of the data for 5, 10 and 20 second intervals. Comment on the shape of the histograms.**

3.3) Frequency Histogram for Day - 1

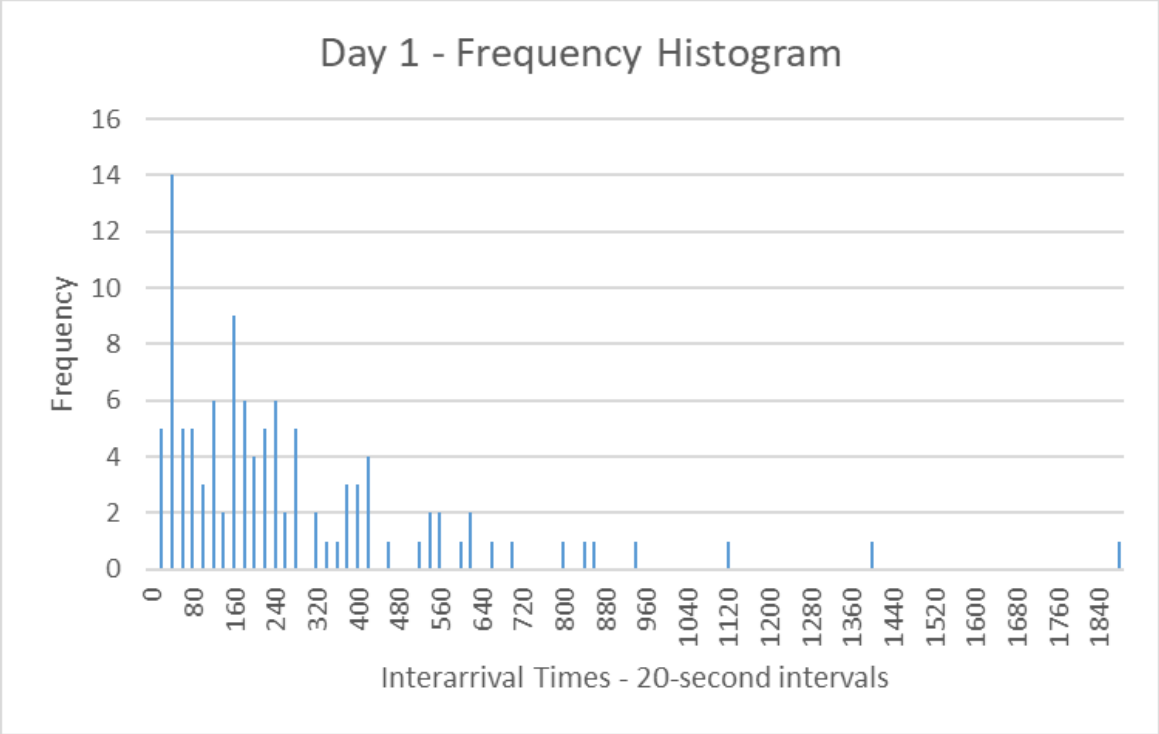
a) Frequency histogram with 5-second intervals (Day-1):



b) Frequency histogram with 10-second intervals (Day-1):

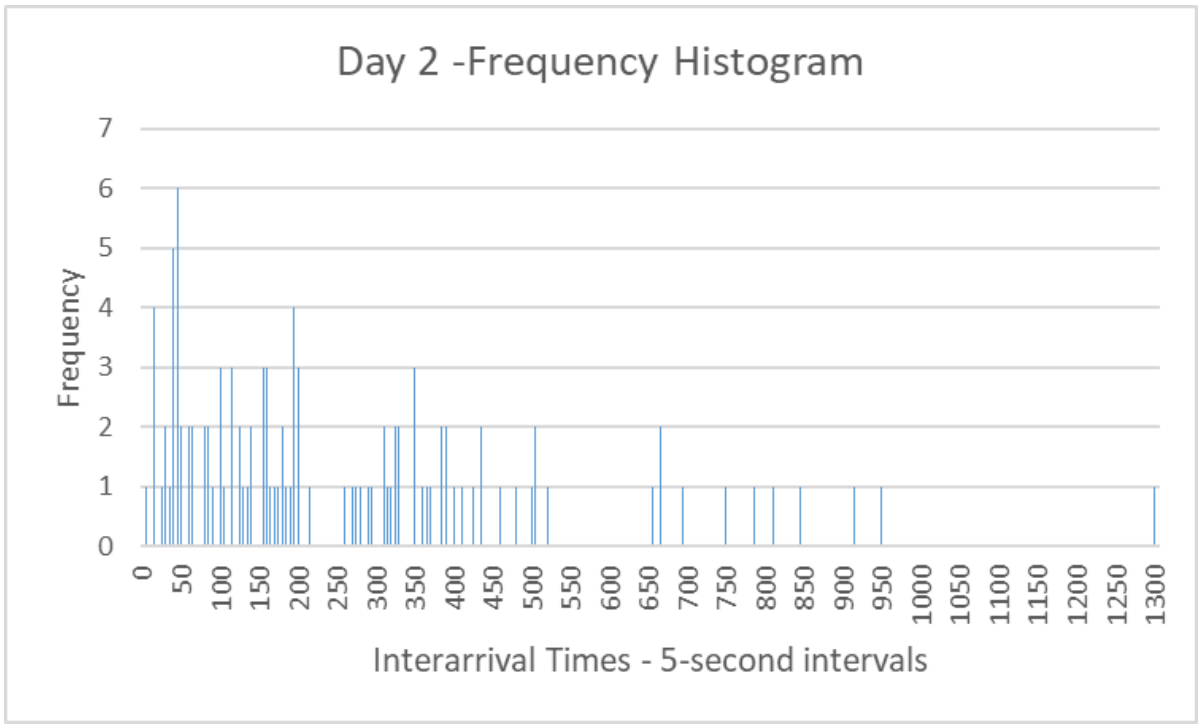


c) Frequency histogram with 20-second intervals (Day-1):

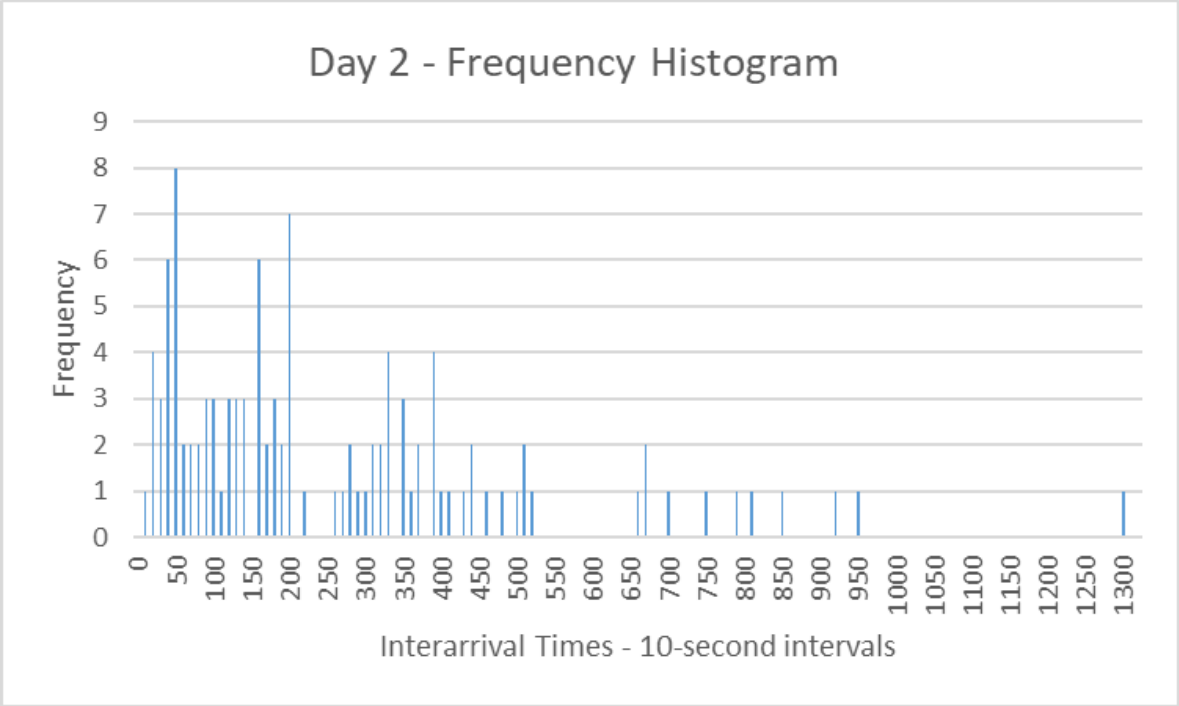


3.3) Frequency Histogram for Day - 2

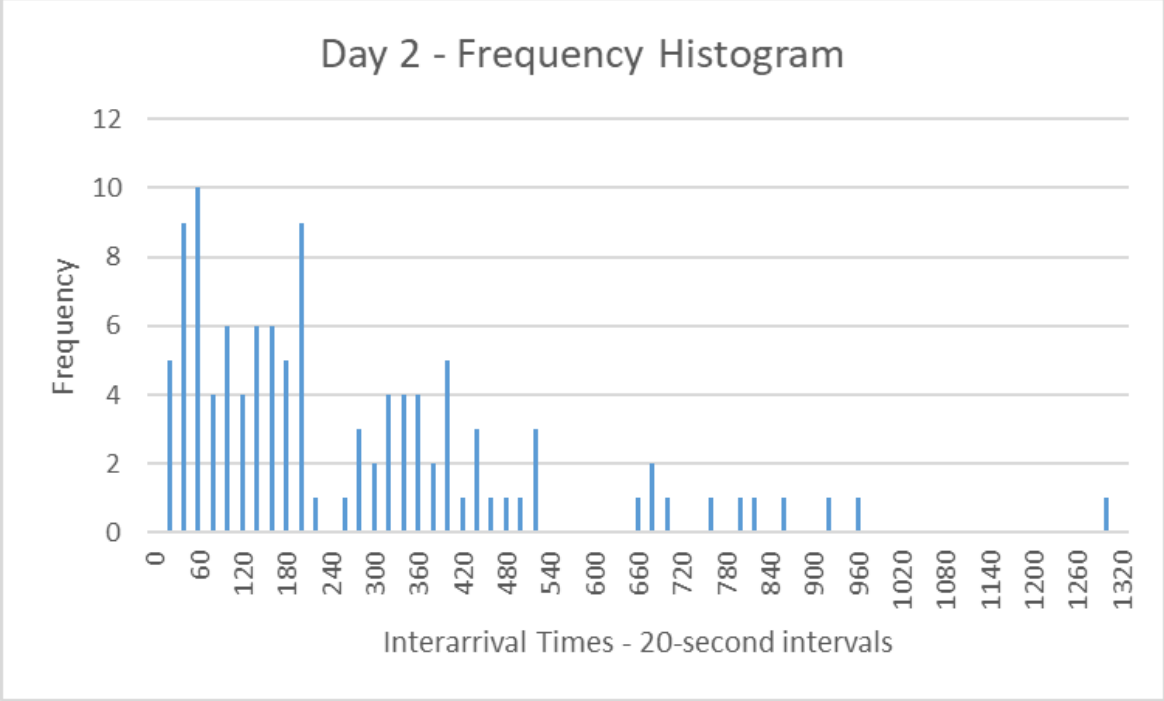
d) Frequency histogram with 5-second intervals (Day-2):



e) Frequency histogram with 10-second intervals (Day-2):

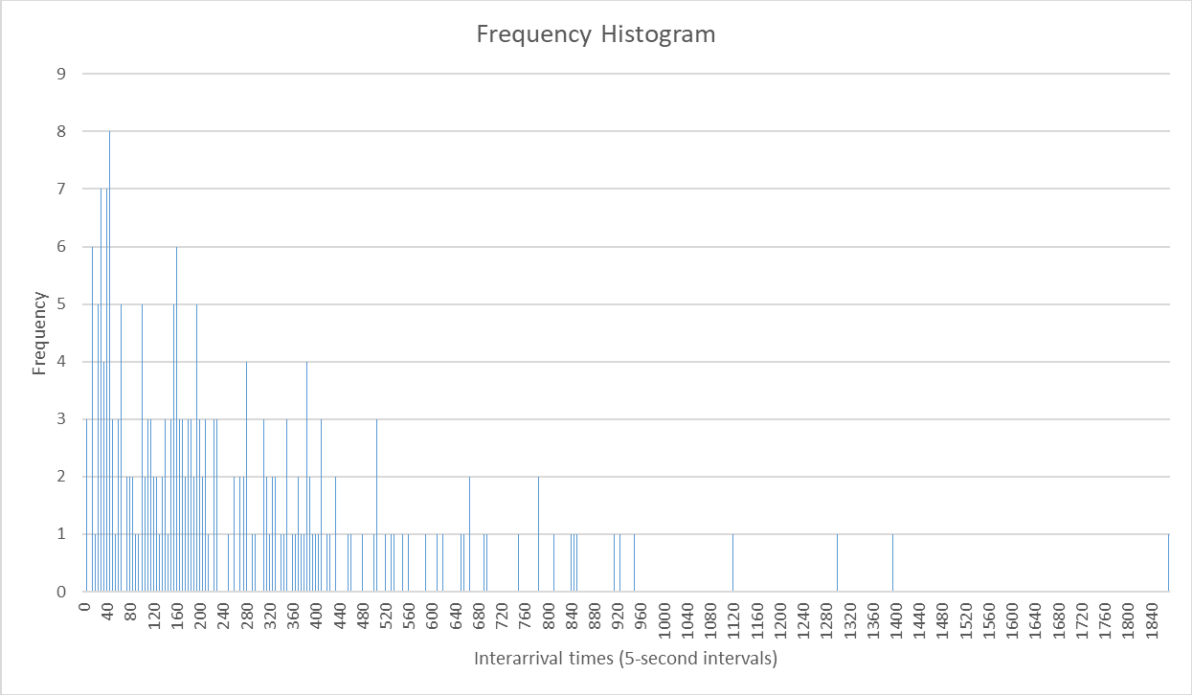


f) Frequency histogram with 20-second intervals (Day-2):

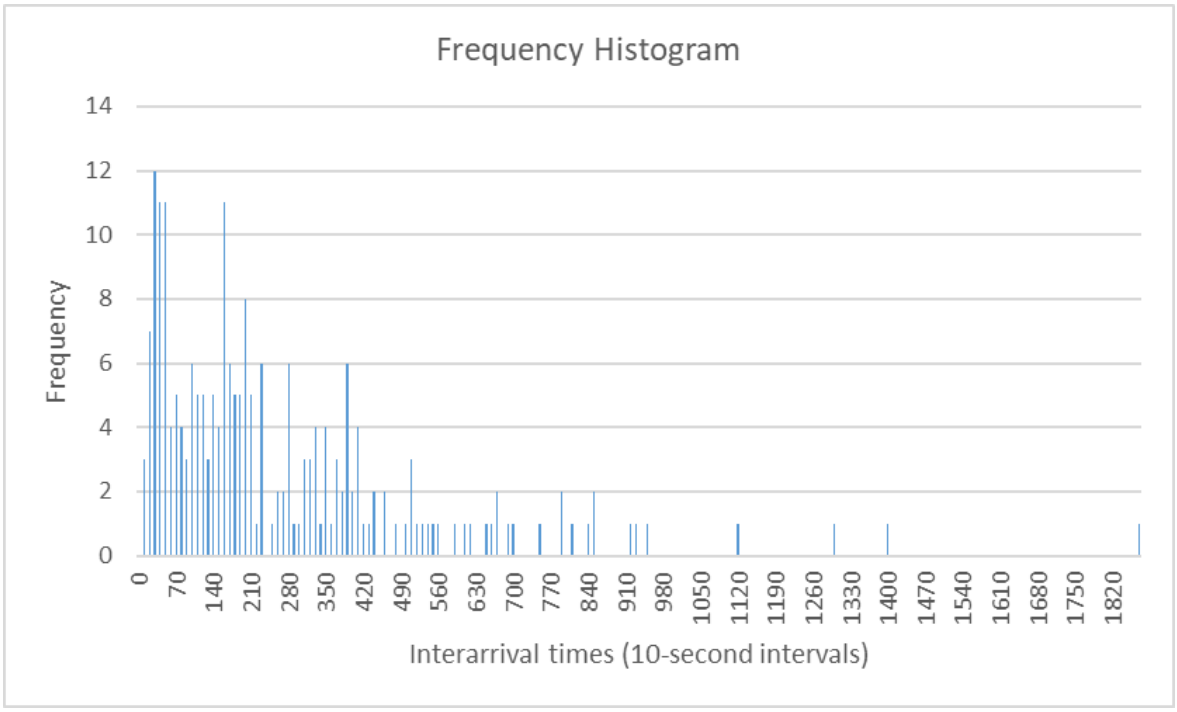


3.3) Frequency Histogram for combined data

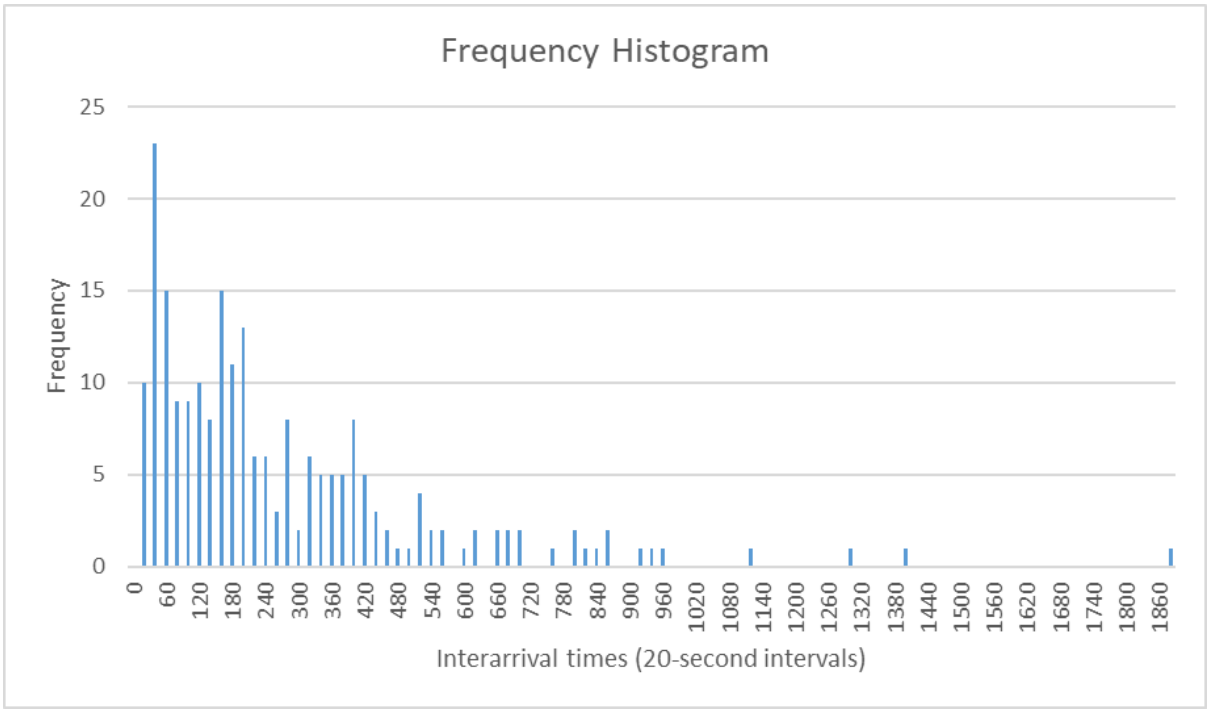
g) Frequency histogram with 5-second intervals (combined):



h) Frequency histogram with 10-second intervals (combined):



i) Frequency histogram with 20-second intervals (combined):



- As can be seen from the frequency histograms, distribution of the interarrival times are not symmetric meaning that it's not likely that the data is normally distributed. It cannot be a uniform or cauchy distribution (which is symmetric about a parameter

a) as well. Cauchy distribution is the ratio of two normal distributions. Shapes of the histograms are similar to exponential distribution, however there are fluctuations at the beginning. More tests are required to determine if the data comes from an exponential distribution. It seems like interarrival times have bimodal distribution since there are two distinct peaks in the histogram.

- Notice that as we increase the interval length, we start losing some information as to the nature of the distribution. We are able to observe the fluctuations more detailed in the histogram with 5-second intervals compared to the histogram with 20-second intervals. Obviously, as we increase the interval length we get less number of bins and therefore more data in one bin.

4) Perform a chi-square test at a significance level of 0.05 with 10 second intervals to test whether the data comes from an exponential distribution where the mean is as found in step 1.

Using the 10-second interval frequency histogram, 187 bins are generated including the intervals starting from 1-10 to 1861-1879.

Sample mean: 261

Exponential distribution: $f(x) = \lambda \cdot e^{-\lambda x}$ $\lambda = \frac{1}{261} = 0.003833887$

In order to calculate the expected frequency for each bin, first p_i, probability of a data value falling in the i^{th} interval under the hypothesized distribution is calculated.

For each bin, it's computed using the formula $cdf(x_0 + 9) - cdf(x_0)$ with the help of EXPON.DIST(x, lambda, cumulative) function in EXCEL.

To get the expected frequency, we multiplied the results with the sample size for each bin. Then we applied the Goodness-of-Fit test:

$$\chi^2 = \sum_{i=1}^{187} \frac{(o_i - e_i)^2}{e_i}$$

Significance level: 0.05

$\nu = k - s - 1$ degrees of freedom = $187 - 1 - 1 = 185$

$s = 1$ since exponential distribution has only one parameter, rate = λ.

$$\chi^2_{0.05,\,185} = 217.73498$$

Value of our Chi-Squared test statistic is $\chi^2 = 306.083$. Since $\chi^2 > \chi^2_{0.05,\,185}$ null

hypothesis must be rejected. However, an important observation must be made at this point. One of the interarrival times in our sample data is equal to 1869, which is extremely unlikely according to our hypothesis distribution. If we replace it with our mean value 261, our test static becomes $\chi^2 = 164.987$, which is less than the critical value. Therefore we don't reject the null hypothesis if that's the case. It's reasonable to suggest at this point that 1869 is an outlier and it causes a serious problem in our statistical analysis. It's likely that we could have a chi-squared statistic less than the critical value if we are provided more data in our sample since increasing the sample size decreases the probability of falling in the critical region. Let's apply other tests to make sure that the data comes from exponential distribution.

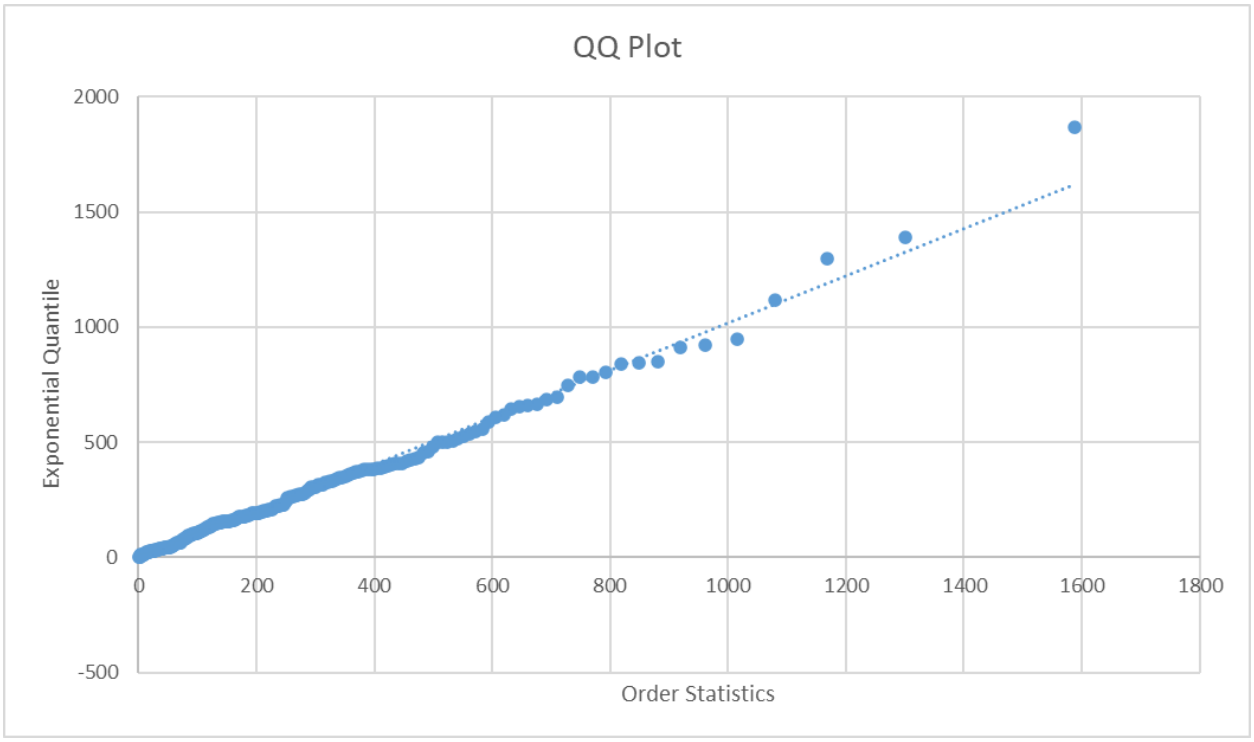
5) Draw the QQ-plot to test whether the data comes from an exponential distribution.

First, **order statistics** are prepared by sorting the sample data. Then two columns are created for rank (j) and $(j - 0.5)/n$ respectively. Then **exponential quantiles** are calculated using the inverse of exponential distribution with $\lambda = \frac{1}{261}$.

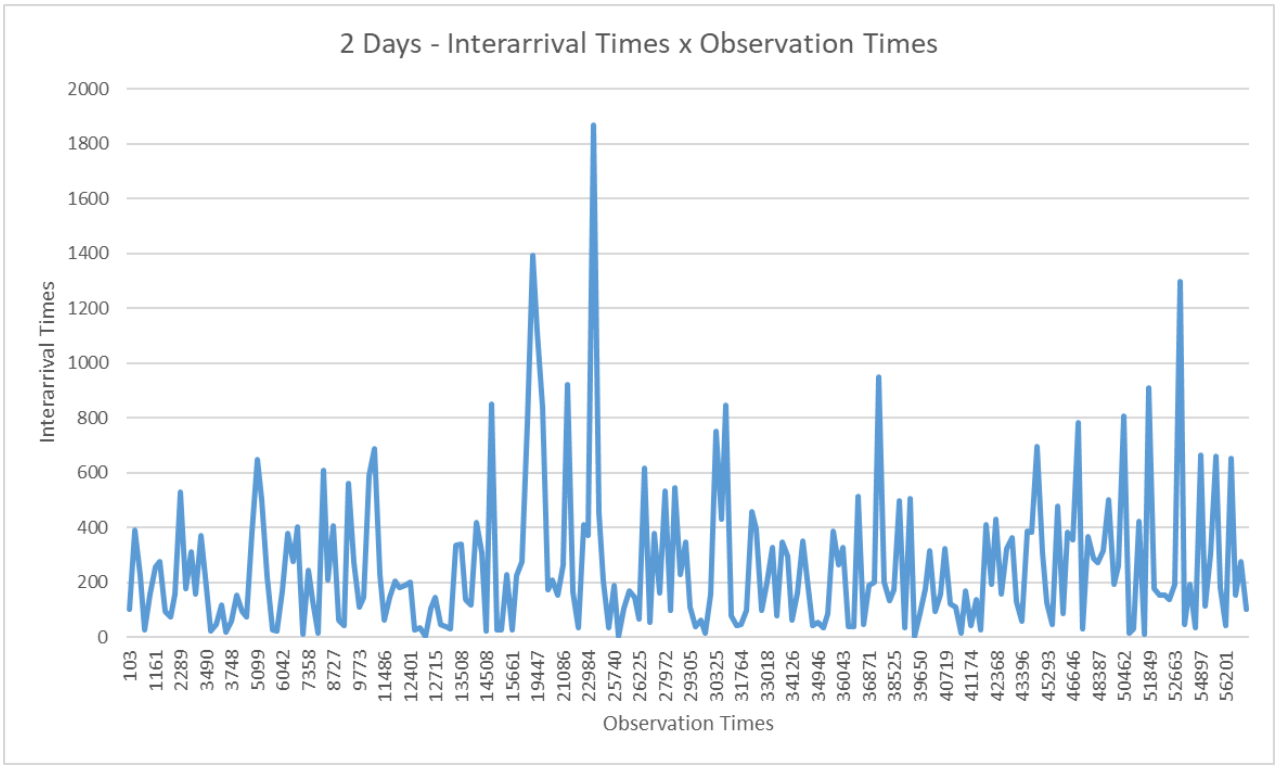
$$F^{-1}((j - 0.5)/n) = -\frac{1}{\lambda} \cdot ln(1 - (j - 0.5)/n)$$

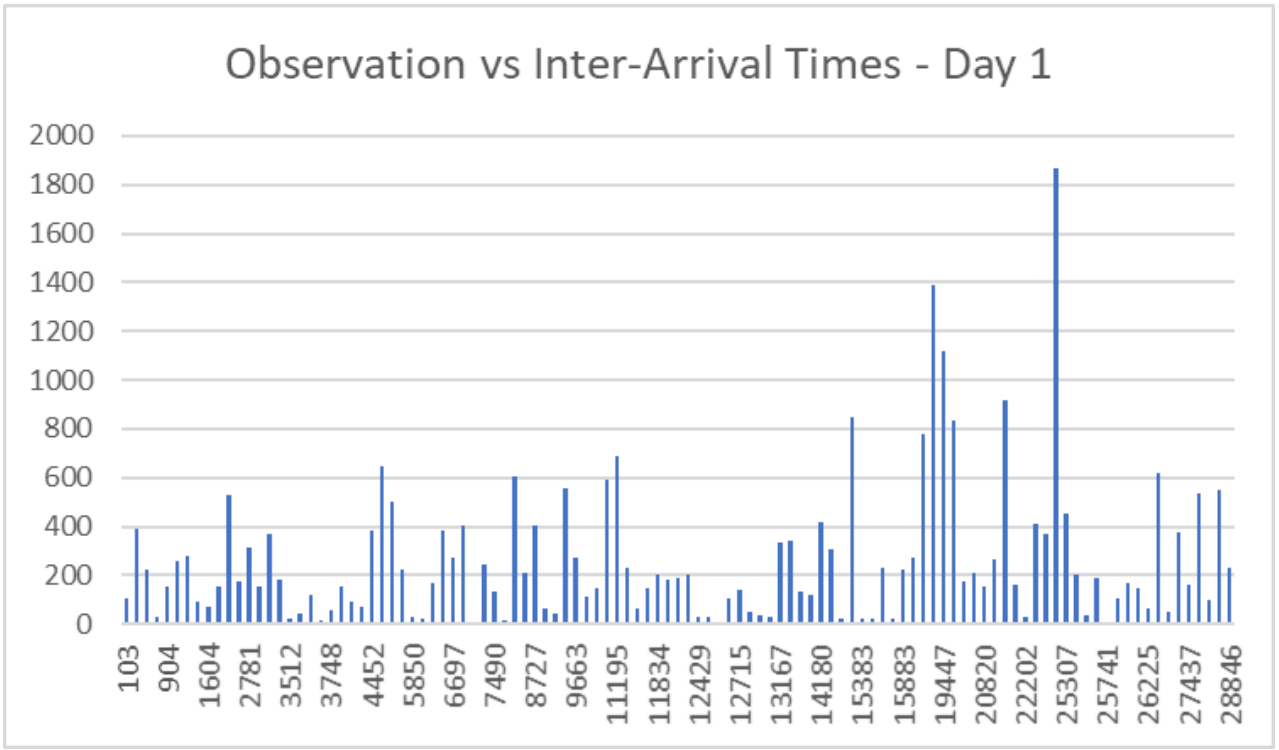
Then plotted each order statistic and exponential quantile pair.

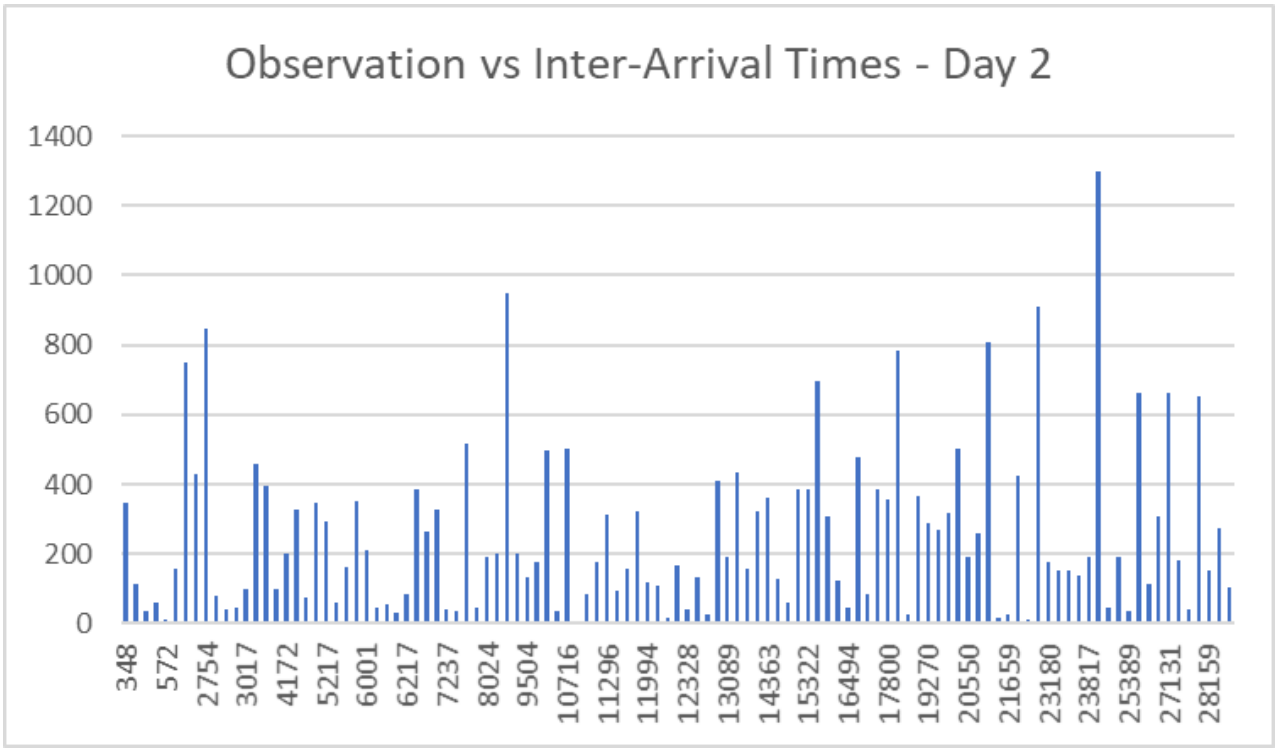
A trendline is created and it can be seen that the plot gives a straight line although there are a few outliers at the higher values.

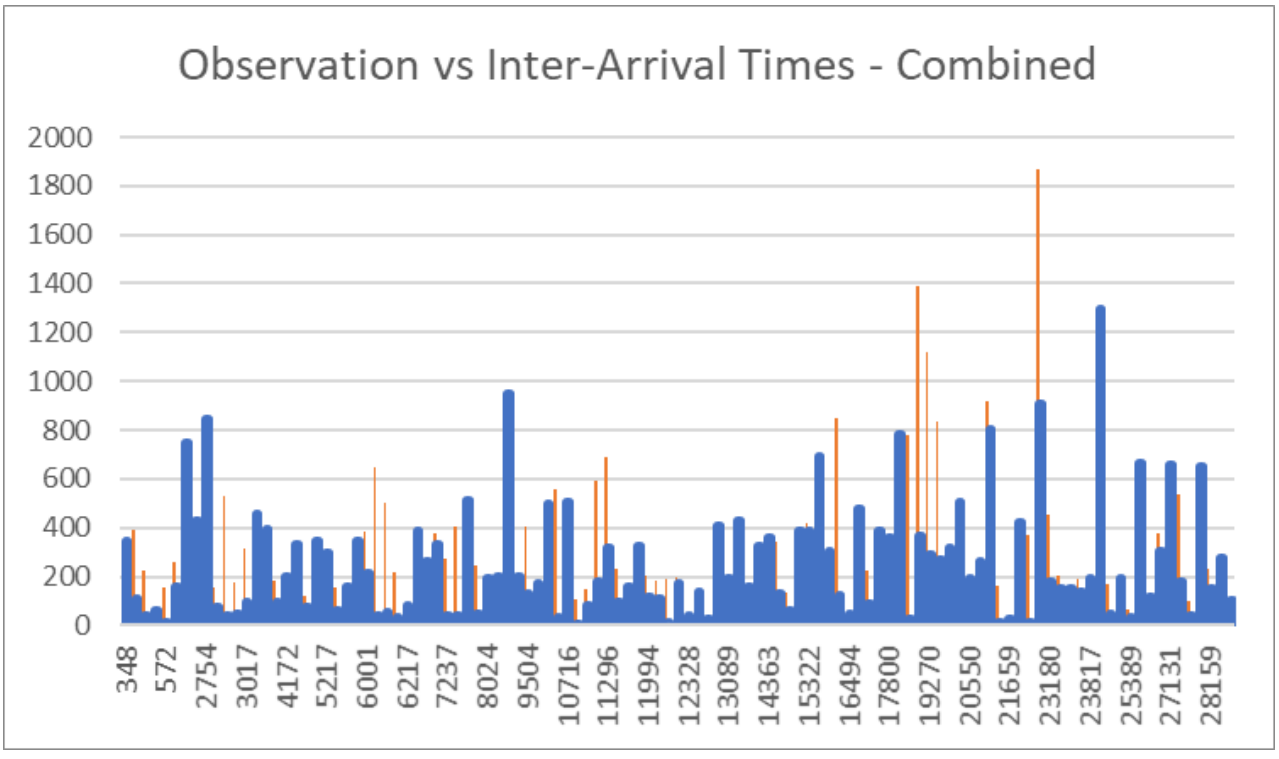


6) Plot the inter-arrival times with respect to observation times. Is there an obvious pattern? Analyze visually if the data is stationary or not.


2 Days - Interarrival Times x Observation Times
Interarrival Times
2000
1800
1600
1400
1200
1000
800
600
400
200
0
103 1161 2289 3490 3748 5099 6042 7358 8727 9773 11486 12401 12715 13508 14508 15661 19447 21086 22984 25740 26225 27972 29305 30325 31764 33018 34126 34946 36043 36871 38525 39650 40719 41174 42368 43396 45293 46646 48387 50462 51849 52663 54897 56201
Observation Times


Observation vs Inter-Arrival Times - Day 1
2000
1800
1600
1400
1200
1000
800
600
400
200
0
103 904 1604 2781 3512 3748 4452 5850 6697 7490 8727 9663 11195 11834 12429 12715 13167 14180 15383 15883 19447 20820 22202 25307 25741 26225 27437 28846



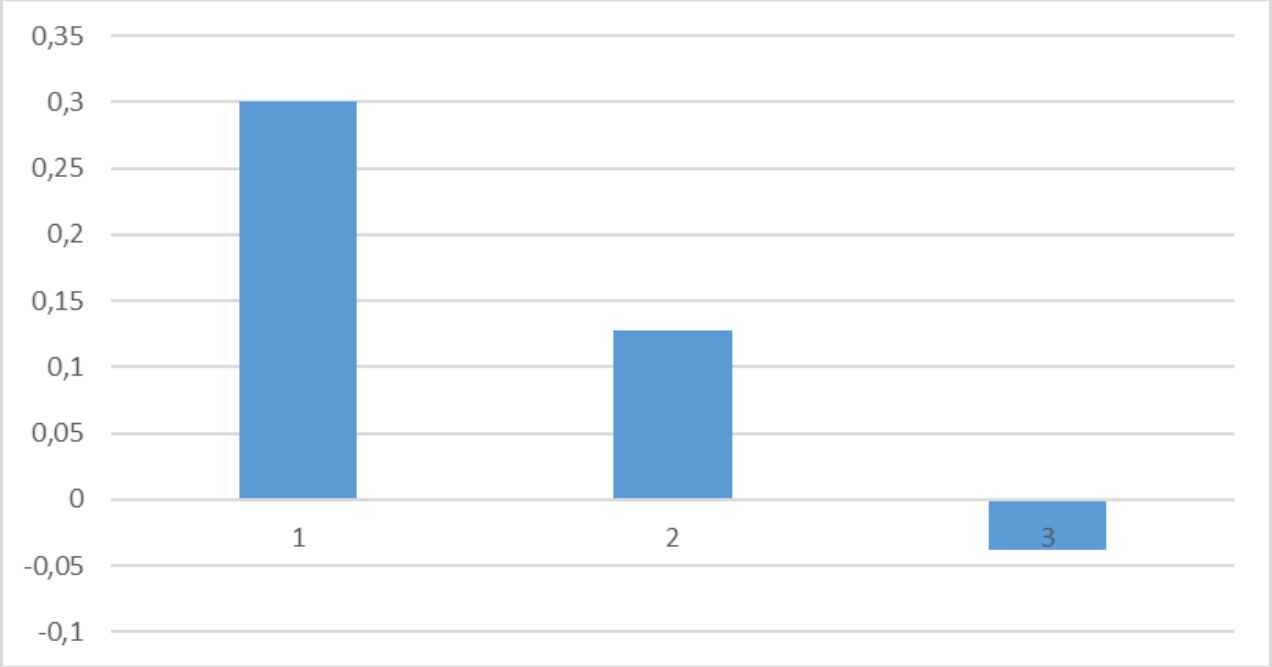


When we look into these graphs (first one represents the two consecutive days combined as one dataset), we see that if we exclude some extreme inter-arrival times, we have a stationary structure, having a somewhat similar mean and variance. Those extreme values are also the ones that caused problems in our Chi-Square tests.

7) Test whether the data is autocorrelated. Plot the lag 1, lag 2 and lag 3 differences. Find and report the correlation for lag 1,2 and 3 differences. Comment on the results.

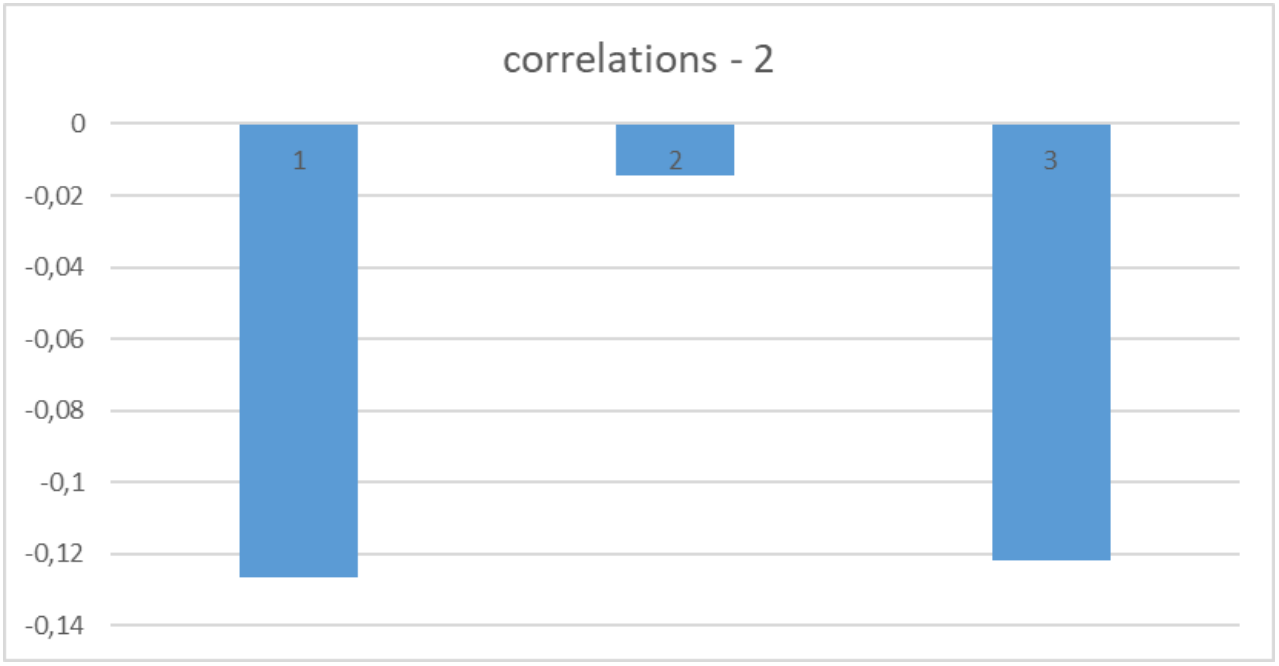
Correlation 1 - 1 → 0,299942844
Correlation 1 - 2 → 0,127089647
Correlation 1 - 3 → -0,037492208
Standard Deviation - 1 → 0,095782629

(Correlations - Day 1 Graph)



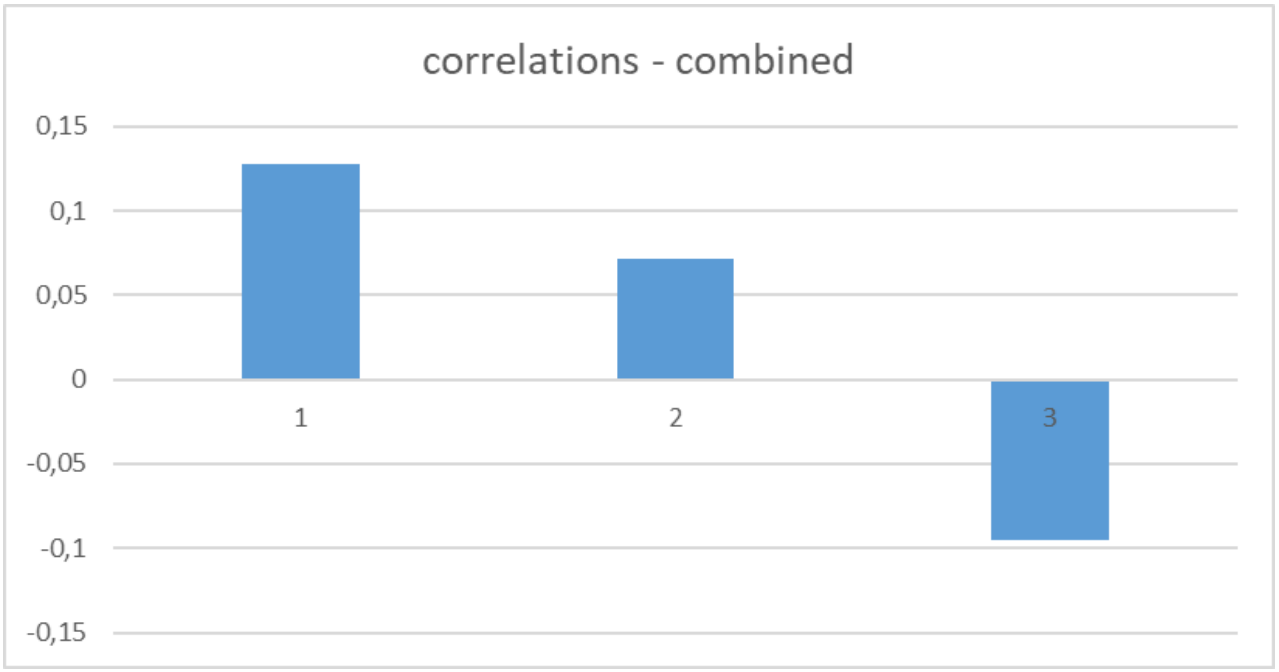
Correlation 2 - 1 → -0,126491054
Correlation 2 - 2 → -0,014565275
Correlation 2 - 3 → 0,0949158
Standard Deviation - 2 → 0,095782629

(Correlations - Day 2 Graph)



Correlation combined - 1 → 0,127241428
Correlation combined - 2 → 0,071372692
Correlation combined - 3 → -0,095617172
Standard Deviation - combined → 0,067419986

(Correlations - Days Combined Graph)

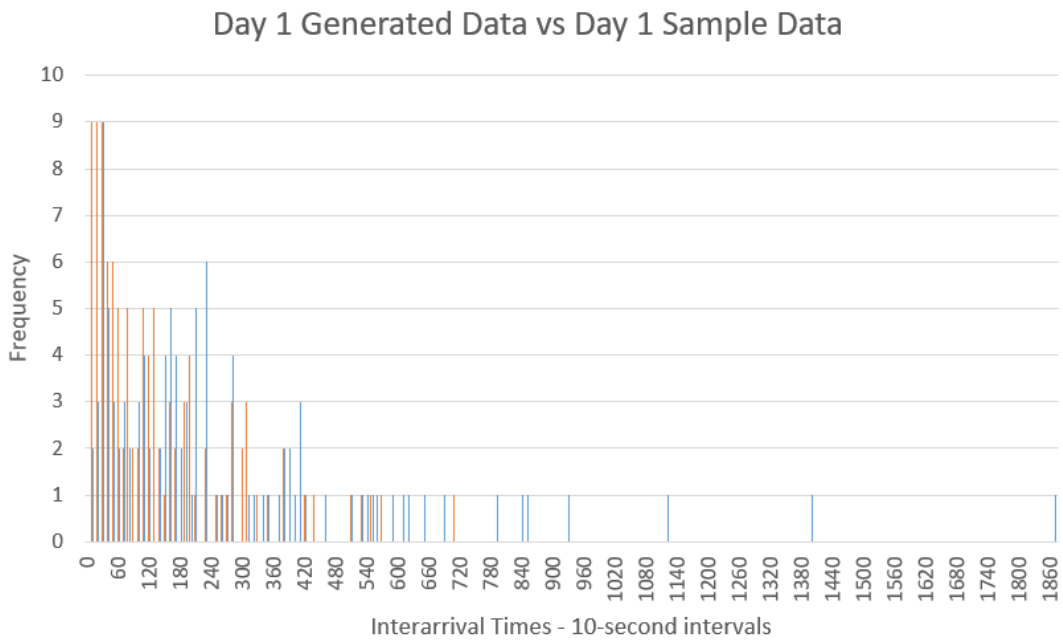


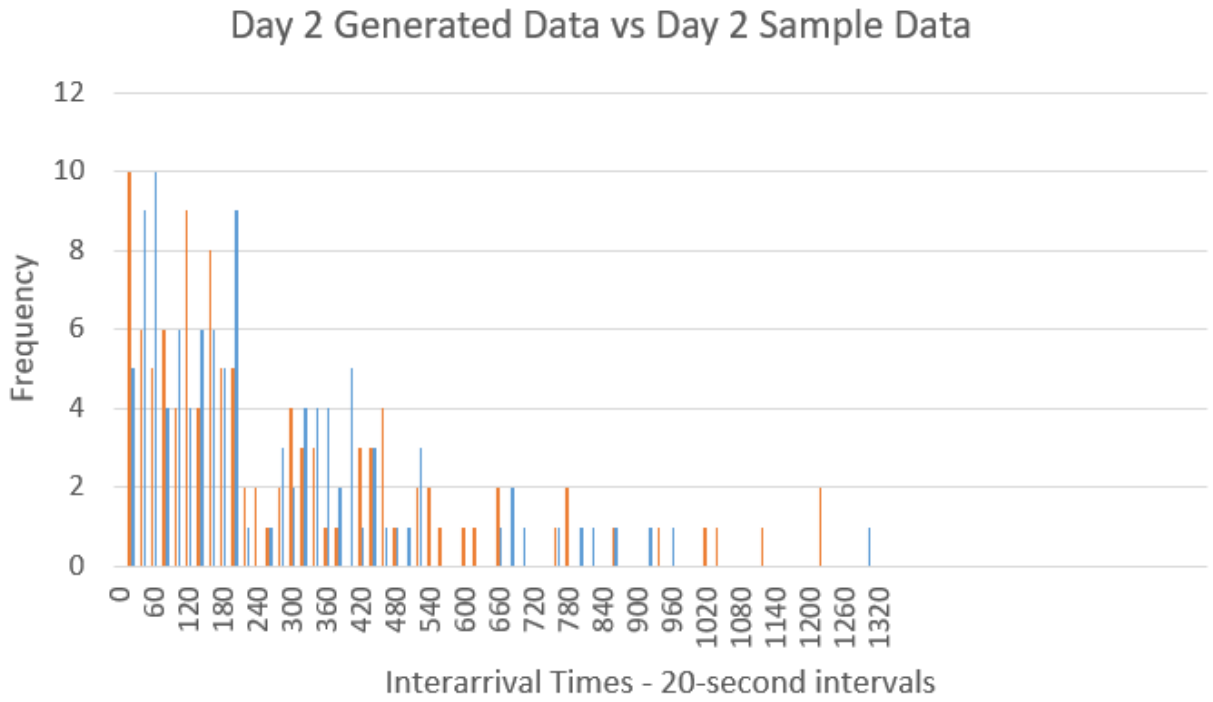
For the first day with lag 1, we see that correlation value exceeds 2 sigmas, however if we take a collective look, we can see that the range stays within the 2 sigma range. From that we can conclude that the data is not auto-correlated. This seems to again come to those extreme values that caused us trouble in the chi-square test and the stationary visualization test.

Overall, we can assume that our values come from an exponential distribution with rate $\lambda = \frac{1}{261} = 0.003833887$.

8) Conclusion and Simulation

We have simulated arrival processes using an exponential distribution with the rate $\lambda = \frac{1}{261} = 0.003833887$, we have generated an excel file named “Q8 - Conlusion.xlsx”. You can examine the frequency histograms and observation time x interarrival time graphs to see the similarities & differences between sample data and generated data.




Day 2 Generated Data vs Day 2 Sample Data
Frequency
0
2
4
6
8
10
12
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960
1020
1080
1140
1200
1260
1320
Interarrival Times - 20-second intervals