Traffic key

Data set dec-pkt-1.tcp has the following format

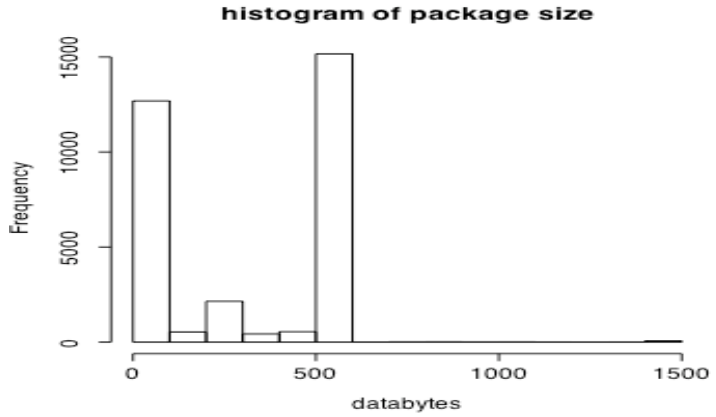
	timestamp	source	destination	sourceport	destport	databytes
1	0.416754	1	2	1223	2046	0
2	0.418705	2	3	1985	20	0
3	0.420657	4	5	119	3849	5
4	0.426512	3	2	20	1985	512
5	0.427488	3	2	20	1985	512

……
……
…..

We read 50000 lines for this assignment.

Question 1.

After removing the packages with 0 databytes, the sample size becomes n=31656



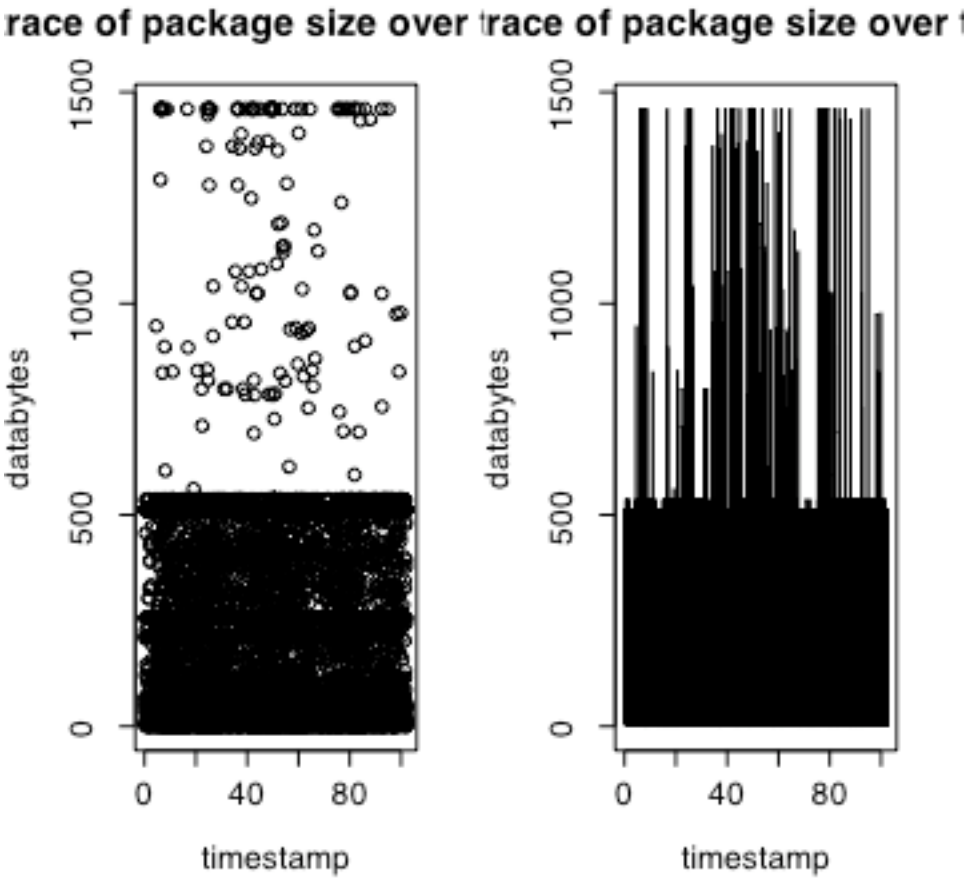
The histogram reveals that the package size has a bimodal distribution, with most packages being either around 0-100 or 500-600, and a few packages being in between and bigger than 600 all the way to 1400. Summary statistics for the whole data reveal that the minimum size is 1 and the maximum size is 1460. But the summary statistics of the whole data set are not representative of what is going on because of the bimodality, so I cut the summary statistics in two parts, summary for size < 300 and summary for size >= 300, which is somewhat arbitrary.

part	min	1st quartile	Median	Mean	3rd quartile	Max
Databytes <300	1	5	27	57.61	51	299
Databytes 300 or more	300	512	512	512.3	512	1460
All	1	31	407	291.7	512	1460

There is much more variability (as reflected in the IQR) between the small packages than between the large ones. Smaller packages are skewed right, larger packages are more symmetric. If you inspect the data further, you will see that there is a good 130-140 observations that are larger than 600. These are not outliers, they are bigger packages.

The package size depends on several factors, such as the network bandwidth (packages admitted per second), the specific implementation of the TCP protocol and the size of what is being transmmitted. Some networks don't allow big packets to go through and fragment bigger packages into smaller sizes. Other networks do allow, and therefore big packages can go through. If there is a lot of traffic at a traffic router, the packages get split further..

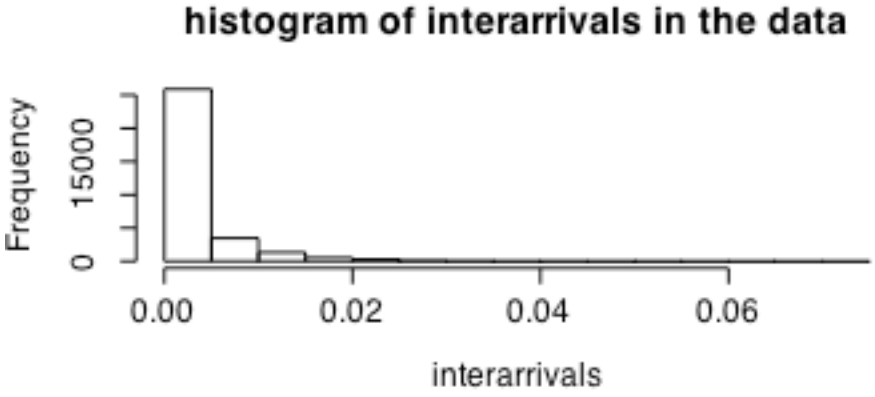
Question 2.-

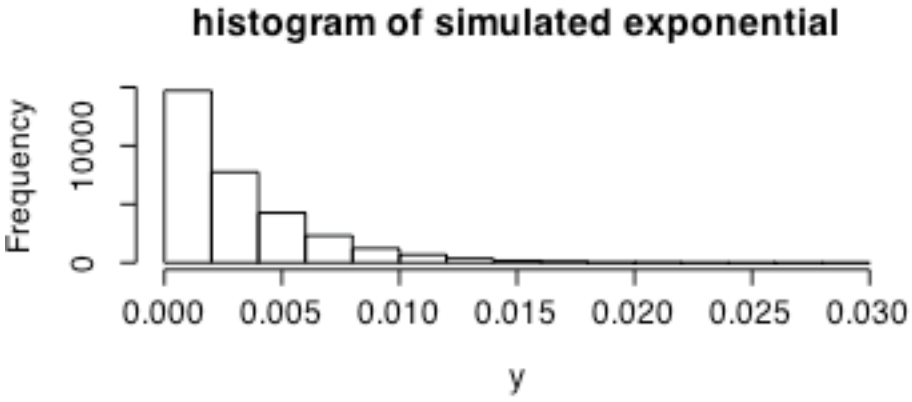


The plot shows that the behavior followed by the size of the packet is pretty constant over time for smaller packets, i.e., those of size less than 600, because the variability between 0 and 550 or so tends to be the same at all time points. At all times there are packets received of that size, with concentrations around 500 and less than 100 (if you make the plot bigger, which is the case if you don't plot them together). But bigger packets tend to

occur in some time intervals but not others as illustrated by the line plot. Of course, this is not a very long time period, only 120 seconds, but we can see that the time of arrival of the package could be a determinant of the size of the package and therefore, of the distribution of package size. Networks get more congested at some times than at others, so the big packages don't go through during those congested periods.
The histogram we got then, would be the histogram for those time intervals where we have tall lines in the right hand side plot. But it would not be the histogram for those time intervals where there are no lines in the line plot above. So during the intervals of time in which there is an empty space in the line plot,our bimodal histogram would not be an accurate picture of the distribution (no packets around the 800-1400 range.

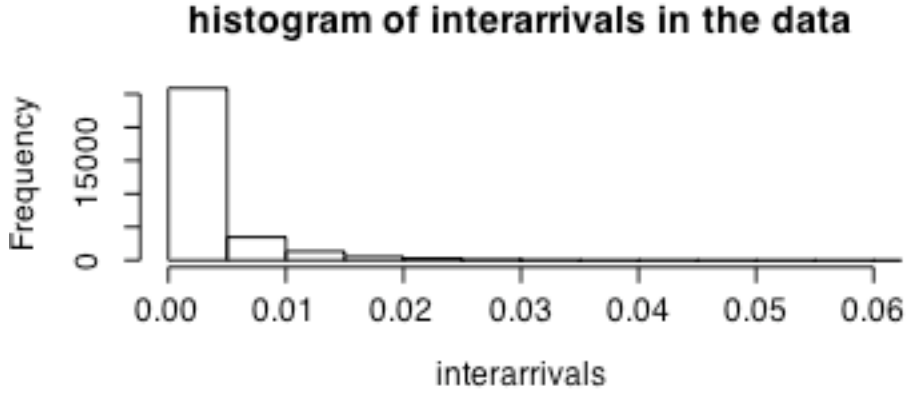
Question 3.-

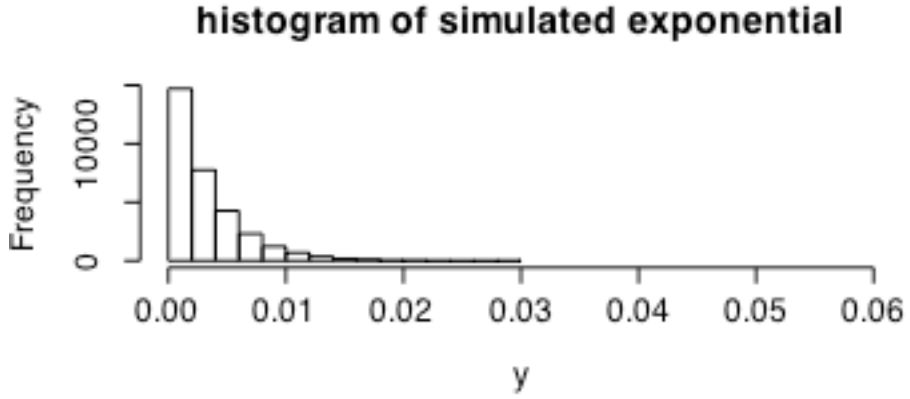




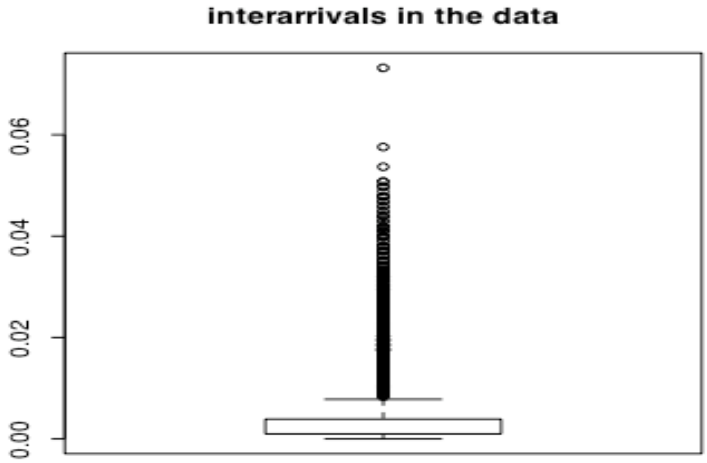
The first thing to notice is that an exponential with the same mean and variance as our data has most of the values appearing between 0 and 0.03. In our data, we have a thick tail. The range of values goes all the way to 0.06. So the exponential model is not a good model for our data in the upper tail.

If we compared the two histograms using the same scale on the horizontal axis, the above would be even more obvious. All we have to do is add the xlim=c(0,0.6) in both histogram commands.

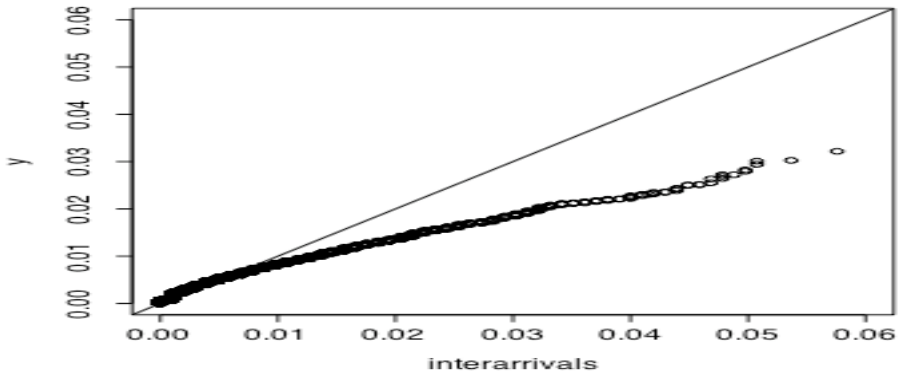




In this last graph it is more clear how the distributions differ. The data has thicker tail. It gives positive probability to rare events, such as large interarrival times.
There is the possibility that the effect in the data is due just to one outlier, so we should do a box plot to check. This reveals that there are way too many outliers to call them outliers. The distribution has thick tails. It is not exponential.



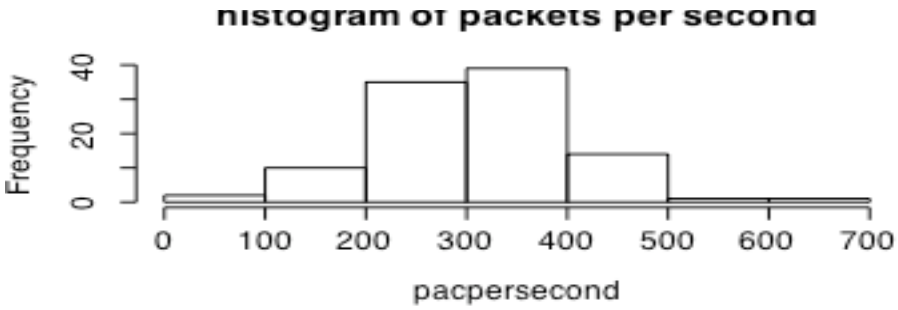
We check now the qqplot to get a better idea about this discrepancies in the tail of the distribution.

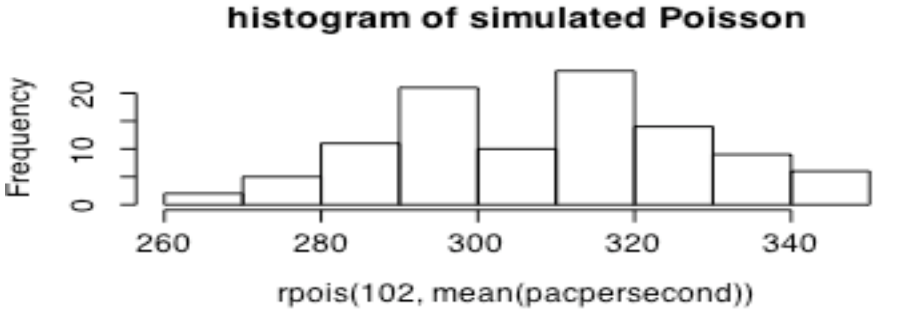


It is clear from the qq-plot (to which I added abline(0,1) that the quantiles of the data do not happen at the same value of the quantiles of the random numbers. The data do not follow an exponential distribution. We look for straight lines in the qq-plot, but we also want that the quantiles happen at similar values. So even without the line, one can tell that the quantiles do not happen just at the same level.

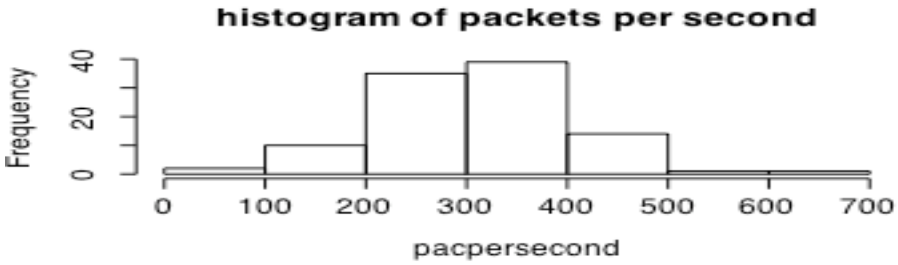
Question 4.-

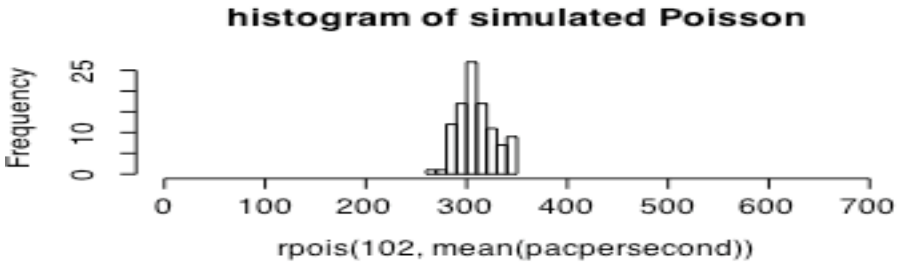
Again, we see that the range of the distributions is very different. The data goes all the way to 700 and has highest frequency 40, whereas the simulated Poisson goes only to 350 and doesn't go to 0. The rate of packages per second does not follow a Poisson distribution. To see more clearly what is going on, let's use the same range on the two axis for the two distributions.





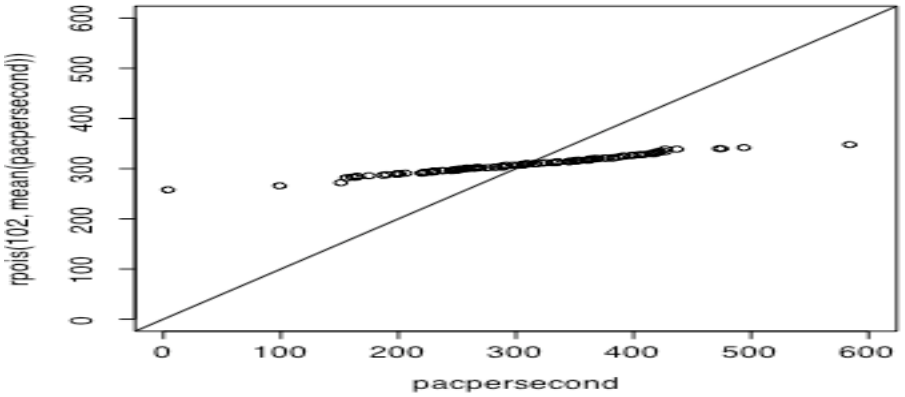
The plots on the same scale, obtained by adding the option xlim=c(0,700) to both histograms, show that the distribution of the data has thick tails but the simulated poisson doesn't.





The qq-plot provides strong support to what we see in the histogram. Clearly, the quantiles of the data do not correspond to the same quantile values in the simulated

Poisson random numbers with the same mean. Therefore, the number of pacs per second is not a Poisson random variable.



Question 5.-

The plot is below. I is hard to see much in such a short time window. However, we can see that most of counts of packets per second are smaller than 400 and there is an episode that they increase beyond that (the highest peak in the data. That is pretty typical of traffic, burstiness, but it is more easily seen with more data at different time scales. The phenomenon we are talking about is the one referred to in the article "Where Mathematics Meets the Internet" published in Notices of the AMS, Vol. 45, no.8, p 961-970 (1998). Figure 1 describes the phenomenon that happens when you change the scale if you have Poisson counts, and what happens when you change the scale with data like ours. As a matter of fact, this whole handout brings support to the comments in that article, that Poisson counts and exponential time between arrivals are NOT good models for real internet traffic data.