Rick Boelen

Hwang Min Yu

Wen Zeng

MATH 189

05/17/2020

**Homework 3 Data Analysis Report**

# 

# **Introduction**

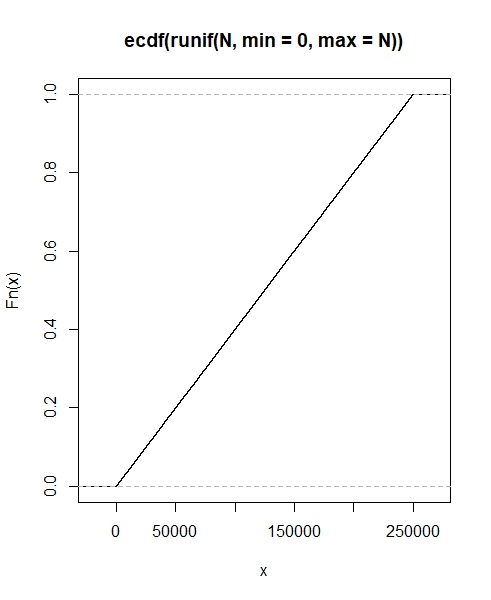
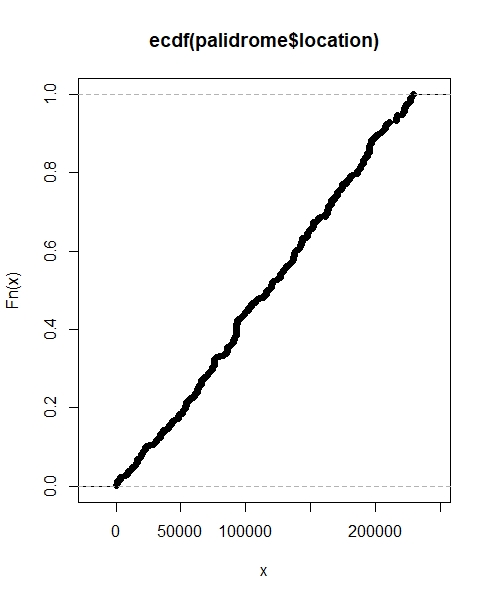
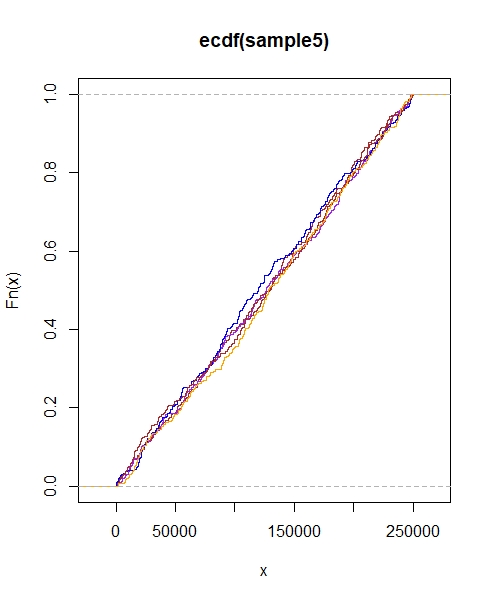
The human cytomegalovirus, CMV, is a deadly virus especially for those who have a suppressed or deficient immune system. The virus initializes harm when it starts to enter a production phase where it replicates its own cell. Our analysis will look into the virus’s DNA to analyze the patterns between the code and find the source of the code that relates to the replication of the virus.

Our process will be looking at specifically complementary palindromes, which is a palindrome that is based on the same letters being read reversed as the complements which correspond to letters being complements to one letter to another. The data will be using a DNA code sequence of CMV that is 229,354 letters long and was published in 1990. 296 palindromes were found from this sequence and the location of these palindromes were given. The length of the palindromes were between 10 to 18 base pairs, if the palindrome was under 10 it was ignored.

Some framing issues with the data is that assumptions are being made that complementary palindromes serve the same role that have been found in other groups of the herpes family. This is a relation that we have to keep into account when concluding our analysis and further analysis. Another framing issue is the fact that we ignored palindromes that had base pairs less than 10. Epstein-Barr virus is a virus from the herpes family and contains several short palindromes that are clustered at the origin of replication. This demonstrates that we are omitting a part of data that could be insightful based on other viruses. This could also be considered for the next analysis. Finally, the time of publication could be an issue, advancement of technology could lead to show that there was an improvement in accuracy of DNA code or even being able to utilize those palindromes smaller than 10. This is another aspect to consider when looking for further analysis or checking our analysis with other studies.

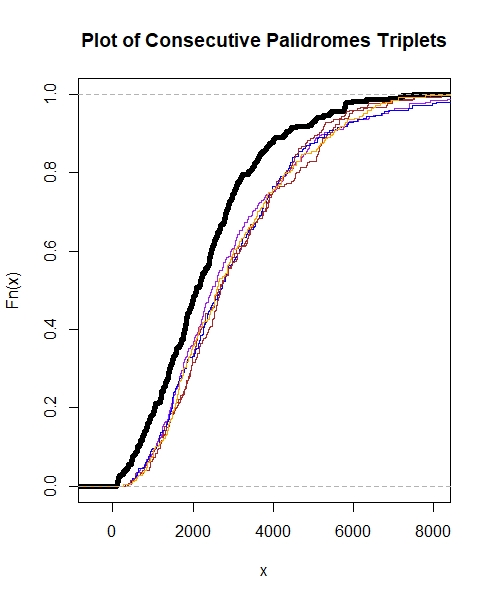
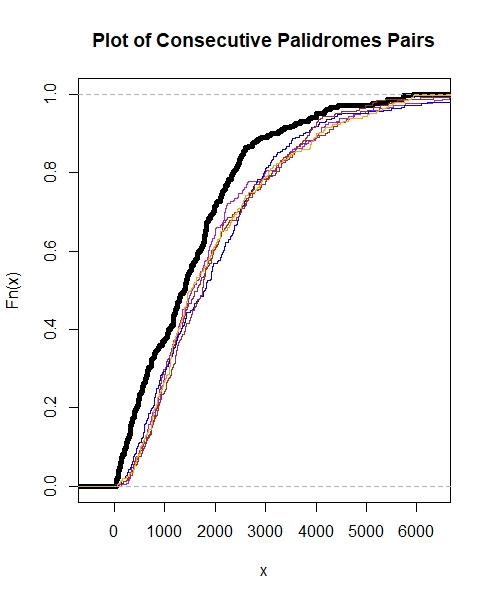
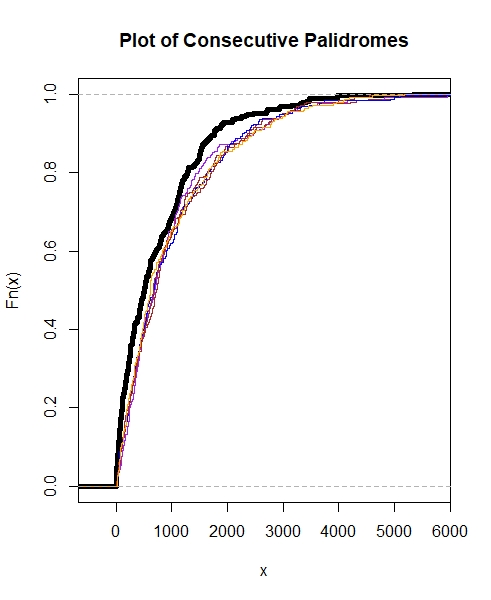
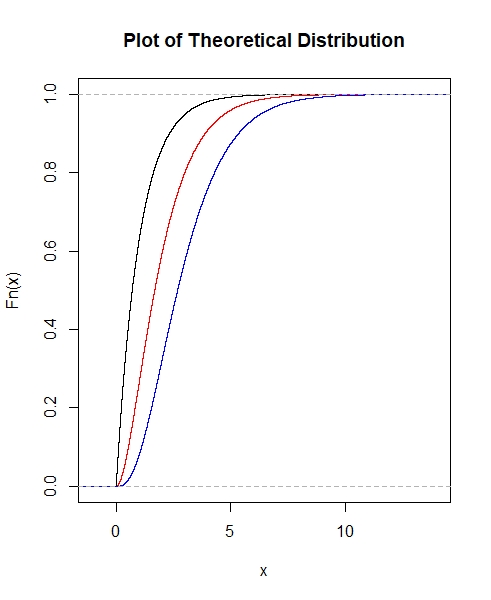
The process behind our investigation is to first look at histograms and compare how randomly scattered our real data is to our simulated data based on a discrete uniform distribution. Then we will proceed to observe the spacing between consecutive palindromes based on consecutive pairs and triplets. From there, we will look at splitting the DNA into intervals and analyzing their counts and checking for the largest count to check whether the result is unusual and a source of replication. This will help us understand the processes behind searching for the origin of replication for similar experiments.

# **Methods/Analysis**

*Locations:*   
*Method:* We plotted an empirical cumulative distribution function for the palindrome original dataset (Left Graph) and its location between the DNA code. From there, we decided to develop 5 samples of sample size 296 from a discrete uniform distribution (Middle Graph) based on the total number of locations in the original DNA code, which was 229,354. This was in order to compare the random uniform scatter with our original dataset and to observe potential differences. Moreover, we applied the theoretical uniform distribution (Right Graph) to analyze a comparison of both.

*Analysis:* The ecdf for the palindrome original dataset reveals a roughly random uniform scatter. In comparison with the theoretical uniform distribution, it is very similar. Therefore, the data appears to not have any places that have an unusual cluster of palindromes yet. There are observations of spikes in the original dataset from the ecdf, but these are small increases and are difficult to allude to anything before further analysis.

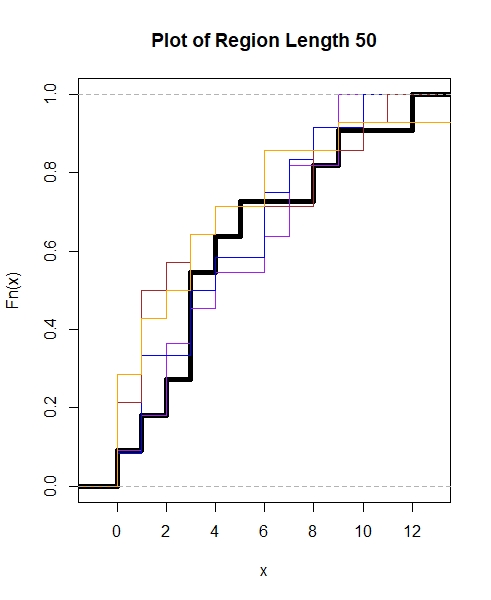
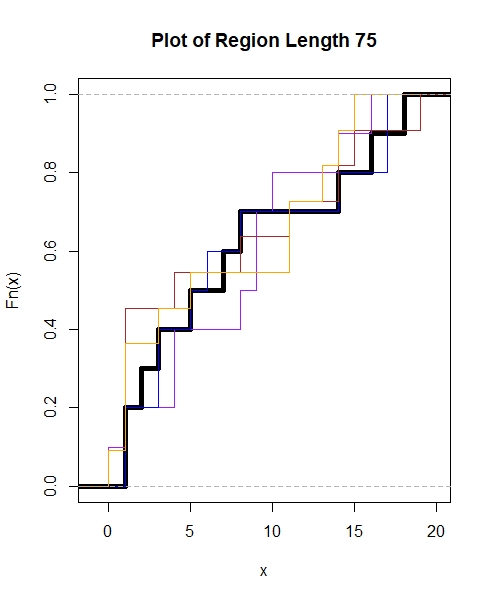
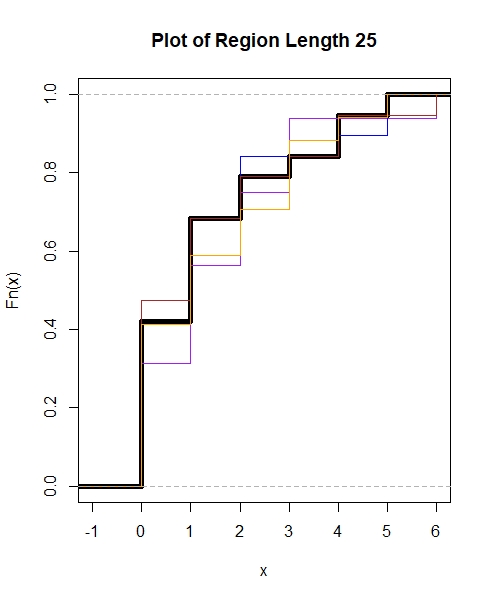
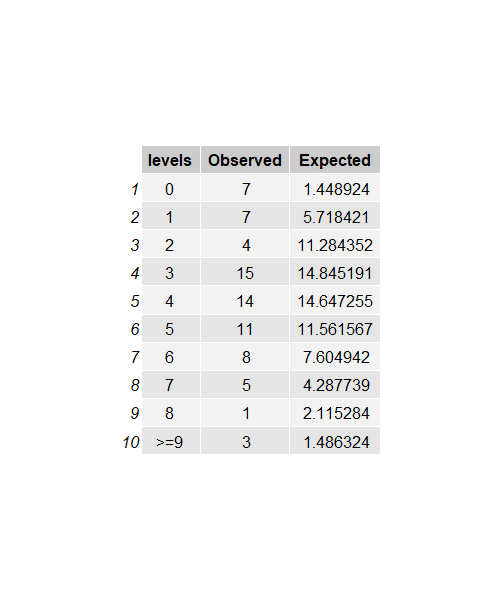
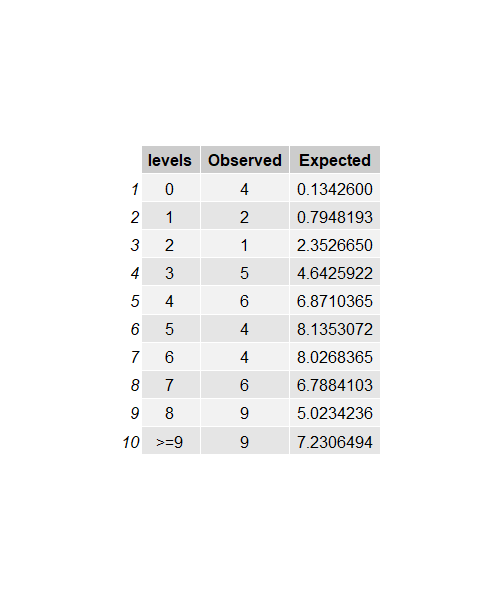
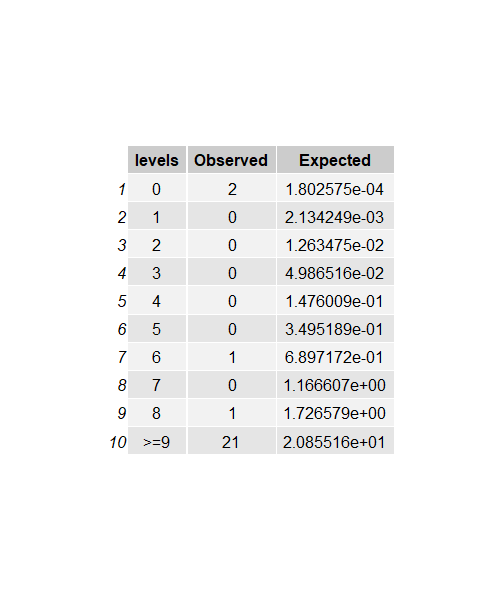
*Spacings:*

**

*Method:* We first got the theoretical distribution of consecutive palindromes, pairs of consecutive palindromes, and triplets of consecutive palindromes through the usage of gamma distribution.Then we plotted the theoretical distribution in empirical cdf plot. We sorted the palindrome dataset and each sample so that we could get the spacing of each consecutive palindromes. The theoretical distribution was from the gamma distribution, when distances are based on consecutive pairs, and exponential distribution, when distances are consecutive. The sorted dataset and samples are our consecutive palindromes. Finally we plotted all the dataset and samples in different plots of consecutive palindromes, pairs of consecutive palindromes, and triplets of consecutive palindromes.

*Analysis:* In each of these graphs, the plot of the theoretical distribution’s shape is similar to the actual plot of consecutive palindromes. In other words, the plots’ curves become progressively less steep as one increases the amount of consecutive palindromes to analyze. Additionally, the sampled curves on each plot are less steep than the plot of the original data. This is shown for all three graphs, and this demonstrates that there are some irregularities in the spacing between palindromes in certain areas leading to certain possibilities of clustering.

*Counts:*

**

*Method:* When counting the number of palindromes in a cluster, we can use this data to determine if the count is in line with a uniformly distributed Poisson distribution or if it indicates something unusual. Therefore we used charts above to compare the number of intervals expected vs. observed for each count of consecutive palindromes. Then, we can use a hypothesis test for the maximum number of palindromes for an interval. We can apply a chi-squared goodness of fit test with 8 degrees of freedom to see if deviations as large or larger than the ones observed in our data set is likely or not (1.2).

*Analysis:* Our null hypothesis is that the number of palindromes in each interval of this data set is equal to or less than the number of palindromes in a theoretically uniformly distributed data set.

The p-values for the chi-squared goodness of fit tests are 0.0005, 0.0005, and 0.0030 for region lengths of 25, 50, and 75 respectively (1.1). At a confidence level of .05, we find that the likelihood of having deviations as large or larger than the ones observed in our data is not very likely and therefore we reject our null hypothesis that the number of palindromes in each interval of this data set is equal to or larger than the number of palindromes in a theoretically uniformly distributed data set. Rejecting the null hypothesis suggests that there are more palindromes than expected within the intervals of this data set and suggesting that the data does not align with the Poisson distribution.

*Biggest Cluster:*

*Method:* Our objective for this was to analyze the largest cluster of palindromes found in a sub interval. We again proceeded with splitting the DNA sequence code into equal width intervals. Our interval sizes were 25, 50, and 75 in order to see notable similarities and differences. Using our original data set we found the observed values from the split region factor that took the counts from each interval. From there we followed a hypothesis test with the test statistic being the largest cluster of palindromes in a sub interval. The null hypothesis states that given a random sample there would be a chance to produce the observed value of the biggest cluster within the confidence interval of 5%. We created samples that were randomly scattered and followed the same procedure and compared the values with the observed values in order to produce the p-value for the test.

*Analysis:* Our experiments found a p-value of 0.0 for all interval widths and therefore, we reject the null hypothesis within the confidence intervals. This is because the observed values for all widths were so extreme that when randomly sampled, it never reached a value that high. The observed value for the corresponding 25, 50, 75 intervals were 24, 19, 16. This indicates that the largest clustering of palindromes was likely not a random occurrence in this case, and therefore there is a pattern behind it that can be further investigated. Through this analysis, we are able to locate a cluster of palindromes that could potentially reach the source of replication.

*Additional Analysis:*

*Method:* We decided that another way to observe patterns of the location of the palindromes is the number of counts that we observe with the mode of number of palindromes in an interval. If this value was higher on average than if it was randomly scattered than we can raise a claim that clusters are being formed around the DNA sequence code. Therefore, we produced a hypothesis test and sampled to check the test statistic, which was the value of the mode of the number of palindromes in an interval. This test was replicated 500 times. The null hypothesis being that the mode of the number of palindromes is found by chance.

*Analysis:*Our observed values for the mode of number of palindromes in an interval was 5, 9, 15 for the same intervals of 25,50 and 75. Using this and comparing to the array of values we found from the randomly scattered samples, we found a p-value of 0.31 for 25 intervals,failing to reject the null hypothesis, and 0 for the other two, rejecting the null hypothesis. This could be the case where there were not enough intervals to find any real insight on the commonality between the number of palindromes between different intervals.

# **Conclusion**

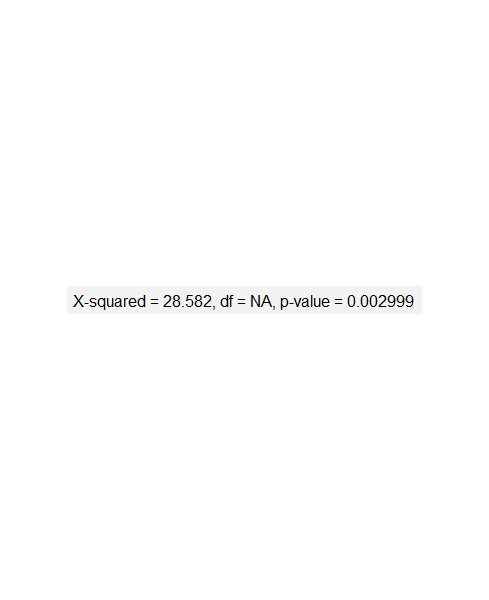
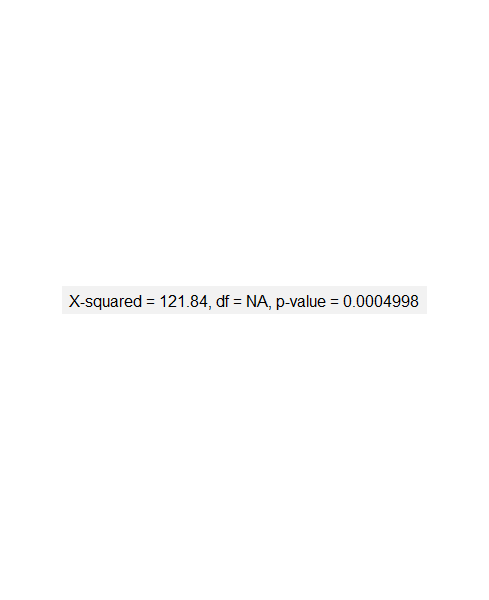
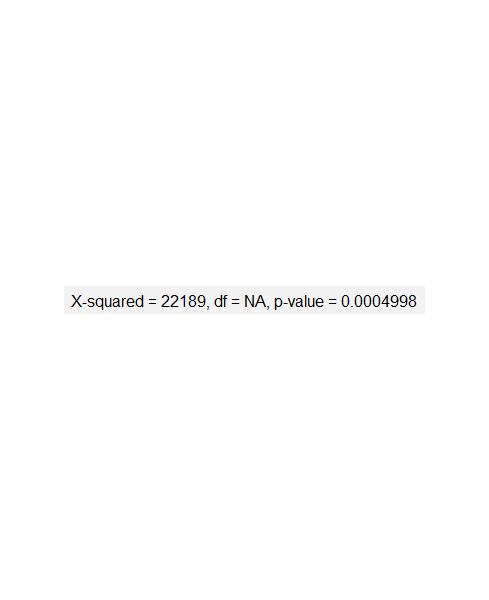
Our findings found that graphically we were able to find small differences between the randomly scattered plots and the different characteristics shown in the locations of DNA sequence code for CMV. The main characteristics we observed were location, spacing and sub-interval clusterings. However, by running tests against the Poisson model and utilizing hypothesis testing, we were able to gather that there were some patterns within the locations of the complementary palindromes that were not explained by a random scatter plot. By finding that there are certain clustering of palindromes, we gain an understanding that if we were to look at the source of replication, we could cut down time by identifying these unusually high clustered regions.

Our recommendation would be to firstly understand the procedure we took to identify these large clusterings of palindromes. The history of the patterns and how these patterns associate with other viruses and their notable patterned findings. From there, I would split the intervals similarly to what we did and analyze those with high clusterings. Moreover, it is critical to experiment with the number of intervals you input due to the issue of granularity. This is because we could potentially say that a high clustering is not due to chance and these are dependent on other factors. Utilizing those intervals, I would start specifically analyzing the shape, symmetry and length of base pairs between the intervals in order to seek other patterns that might correlate with other groups that are within the herpes group. This form of shortcutting the entire sequence code would help to alleviate time and pressure on the biologists and can focus on the patterns within the intervals.

For further analysis, our group would investigate the length of the palindromes itself and the patterns within these palindromes. This would allow us to assess the similarities between the other viruses within the same group of CMV. CMV would contain similar characteristics to their family group members and so another method would be testing this process out on other viruses within its group.

# **Appendix**

1.1) Chi Square Test Output (in order): Region Size 25, 50, 75



1.2) Chi-Square Goodness of Fit Test: A type of hypothesis test used to compare the differences between the observed and the expected distributions. It is a method that assesses how well the model fits the data. The null hypothesis suggests that there is no relation between the model and the data. In this case, rejecting the null hypothesis suggests that there is an abnormality in the distribution of palindromes.

**Contribution Statement**

Introduction: Rick

Analysis Writeup: Wen

Analysis: Everyone

Conclusion: Rick

Code: Hwang