Leomart Crisostomo

3033311439

Stat 135 Lab 2 Memo

**Introduction**

The molecule of life known as deoxyribonucleic acid (DNA) has been one of the important discoveries because DNA brings the genetic instruction for controlling life processes, cell’s development, reproduction as well as its death. DNA is a long, coded message and has many sequences of pattern that may contain the origin of replication.

Human cytomegalovirus (CMV) is a severe disease that may cause death to people with deficient immune system and is a member of Herpes virus family including chicken pox and Epstein Barr-virus. About 30% to 80% of the human population is infected with CMV, and once infected, the virus typically does not grow actively. However, CMV becomes life threatening when it enters a productive cycle that replicates rapidly.

The origin of replication of CMV is marked by complementary palindromes which is a sequence of letters that reads in reverse as the complement of forward sequence. Finding the origin of replication helps scientist to fight the virus but testing each segment of DNA to find the location is inefficient and time consuming. Where do we begin our search to find the origin of replication of CMV? How do we examine if a cluster of palindromes may be an origin of replication? We will test to see if the palindromes across the DNA of CMV comes from a random scatter.

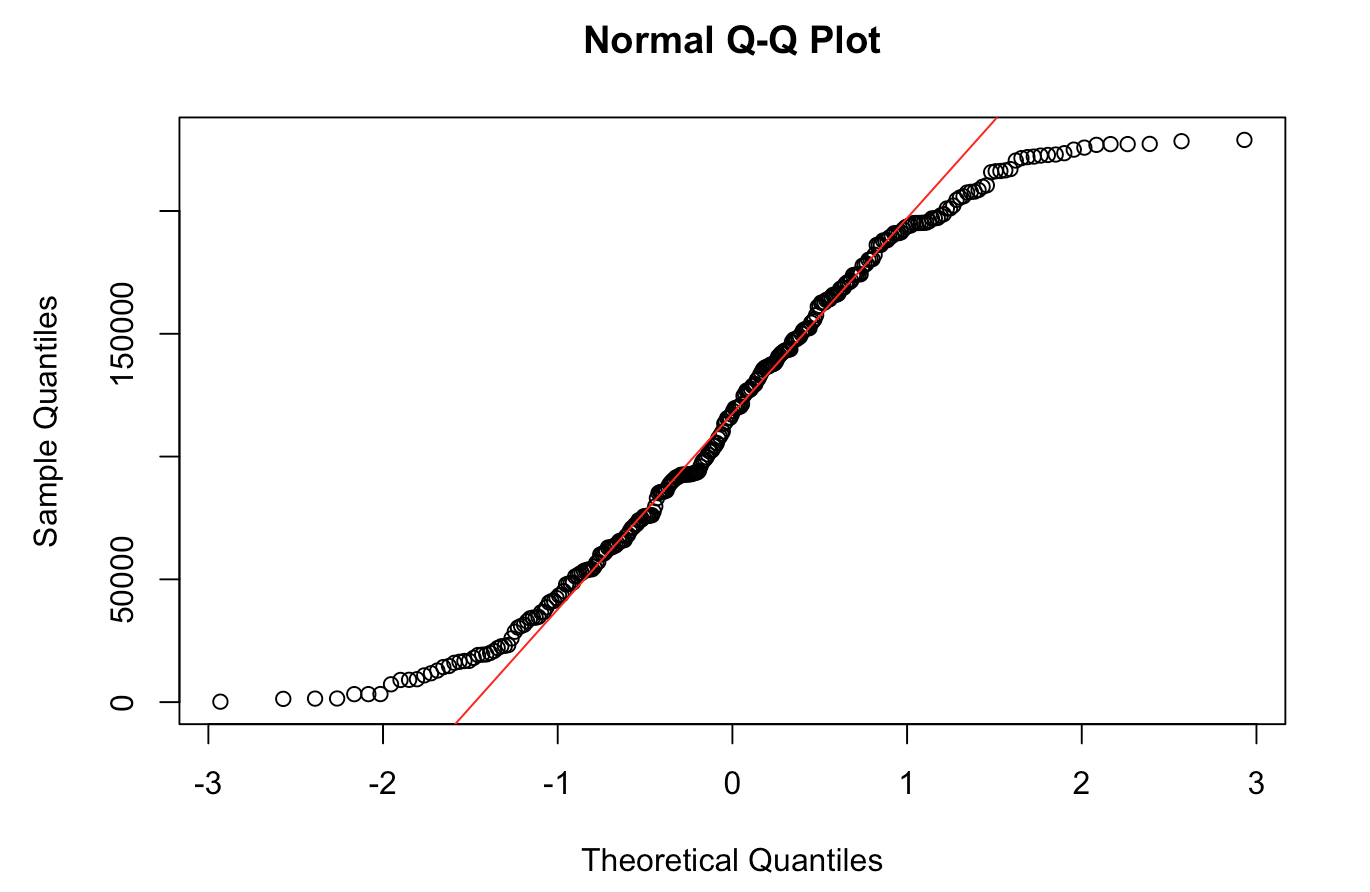
**Methodology**

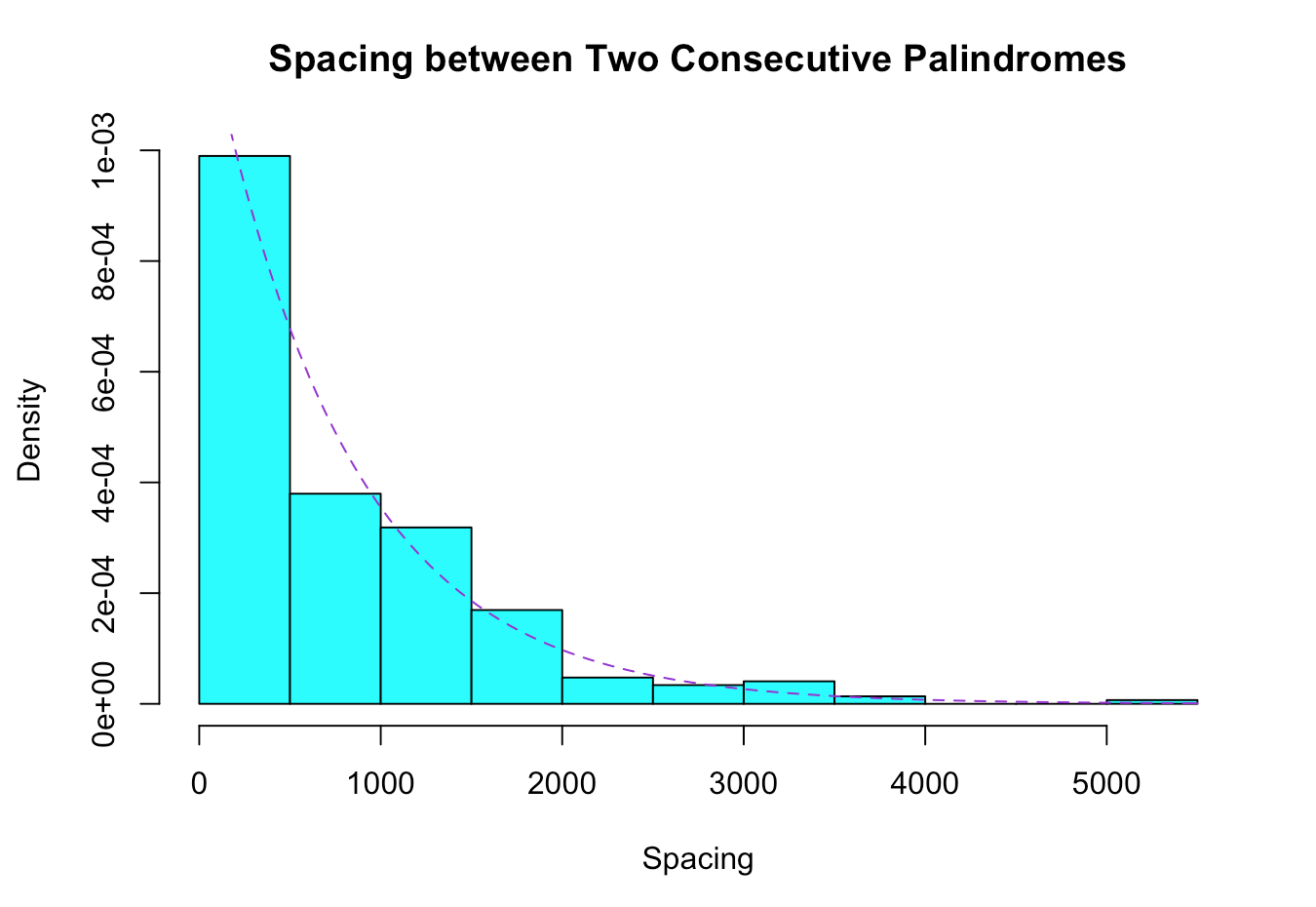
The CMV DNA molecule contains 229,354 complementary pairs of letter, but the data will be focusing on are the different locations of the 296 palindromes only in the DNA sequence of Cytomegalovirus that are at least ten base pairs long which were published by Chee et al. ([CBB+90]). A glimpse of the data is shown below.



Method I:

We first look at the qqplot (shown below) of all the locations of palindromes and see if it follows a normal distribution, but the plot has a long right tail and short left tail. Hence, it does not follow a normal distribution. Now we investigate at the spacing of two consecutive palindromes. For every consecutive pair of palindromes, we take the difference and plot each difference, which roughly shows an exponential distribution. We find the corresponding exponential distribution by calculating the rate as the total number of palindromes divided by the range of the first and last palindromes. Then we plot the corresponding exponential distribution on the histogram and it can be seen on the graph below that the exponential distribution mostly fits the histogram. Thus, the structure in the data may come from a uniform scatter of palindromes across the DNA.





If our findings from the previous method is true, then if we are going to look at all the locations as a whole, we would expect them to have a poison distribution -- palindromes would be scattered randomly and uniformly across the DNA and the number of palindromes across the DNA is independent of each other. We would test this method by using chi-squared goodness-of-fit test.

Let the null hypothesis be the number of segments for each count of palindrome follows a poisson distribution and the alternative would conclude that it does not fit the distribution. First, we divide the CMV DNA into 46 non-overlapping intervals of equal length of 5000 bases and count the number of complementary palindromes in each interval. Then we find the distribution of these counts of palindromes and out of 46 segments, 3 has 0, 1, or 2 palindromes, 5 has 3 palindromes,… and 4 has at least 9 palindromes in them as can be shown in the table. These are the observed number of segments containing the corresponding number of palindromes.

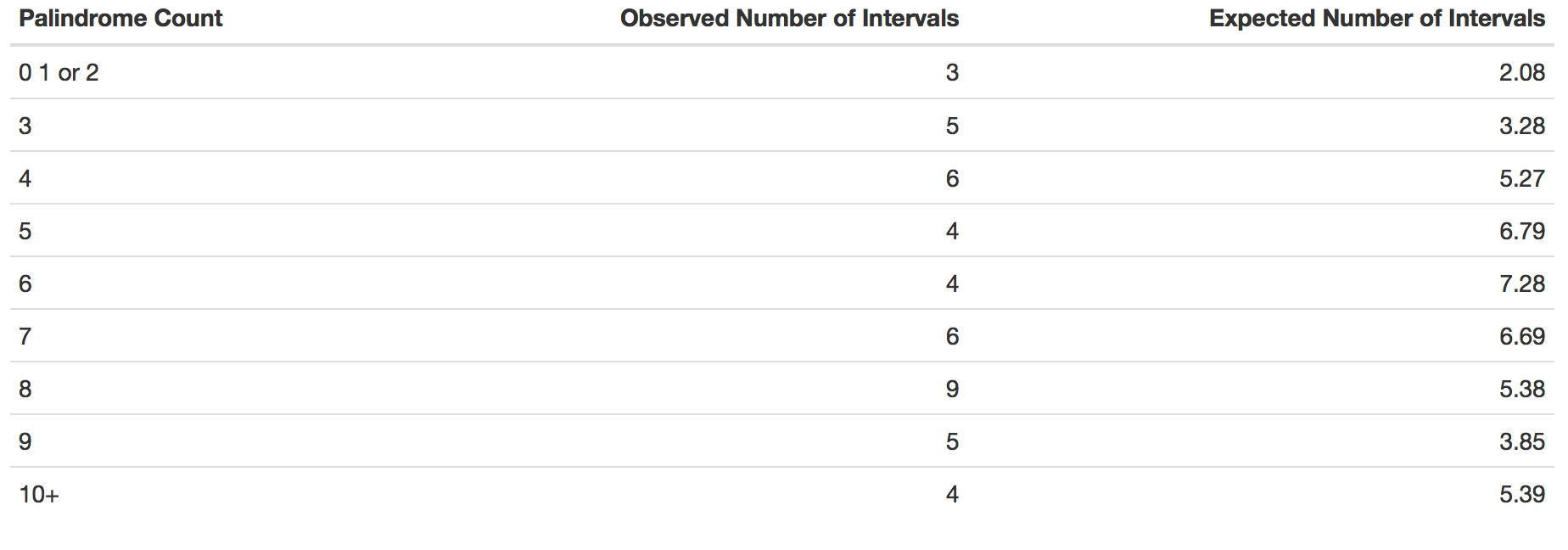
Method II:

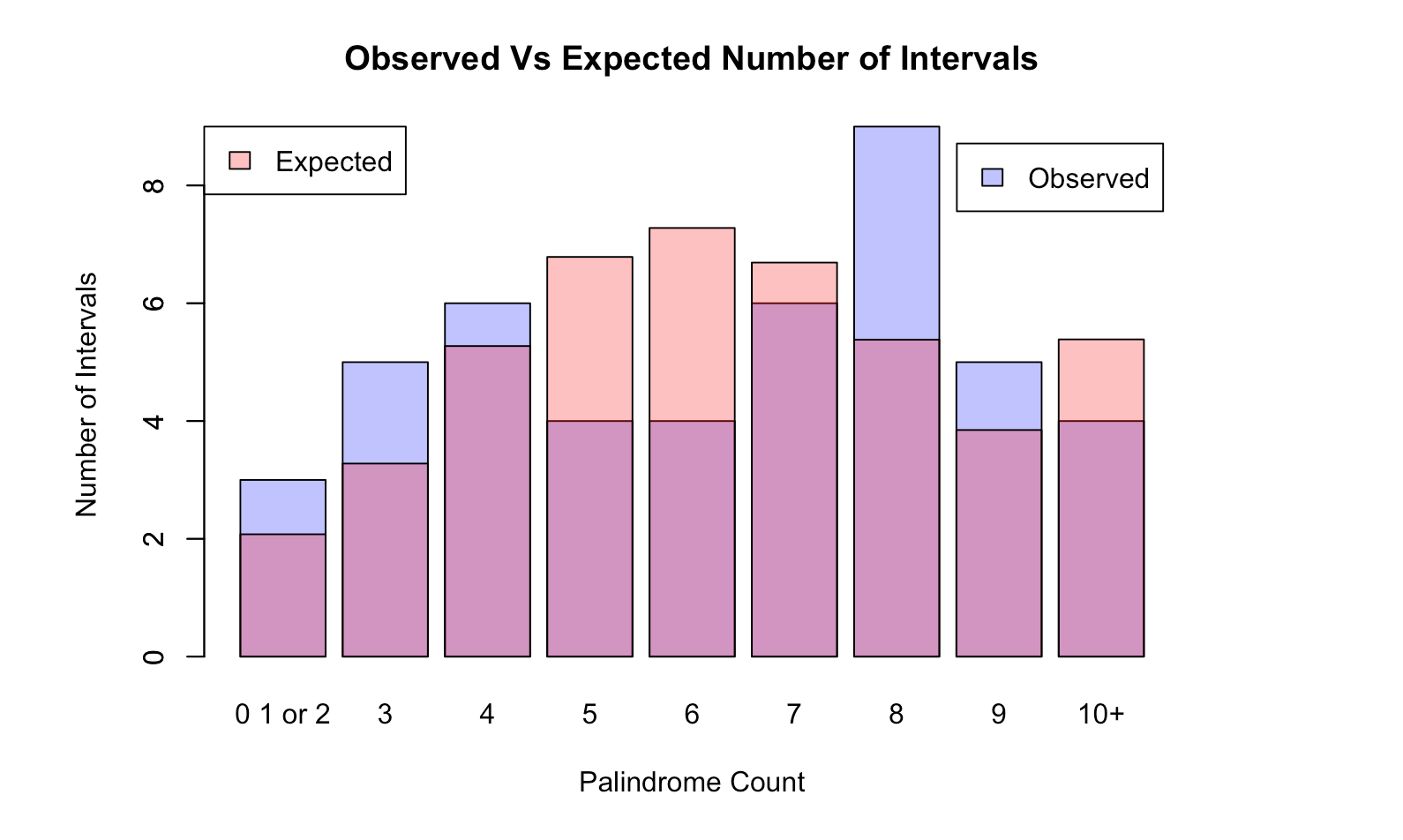
We can use the maximum palindrome count to see if it is as extreme as what we would expect from the maximum hit of the poisson process especially if the chance is small. The largest number of palindrome counts in the 46 segments of CMV DNA would behave as the maximum of independent poisson random variables. Under the poisson process model, the probability of maximum count of palindromes is at least as the observed largest number of hits can be used as the p-value for the test statistic of the chi squared test we are performing.

**Results**

Result I:

We plot the observed and expected number of intervals for each category of palindromes count on a bar graph, and this does not really tell us apart if both distributions are the same. We investigate further using chi-squared test. The rate of the expected poisson distribution is the total number of palindromes, 296, divided by the total number of segments, 46, so the rate is 6.43 per 5000 base pair. We find the expected number of segments by multiplying the chance of each hits under poisson distribution with the total number of segments, and the results are shown on the third column of the table.



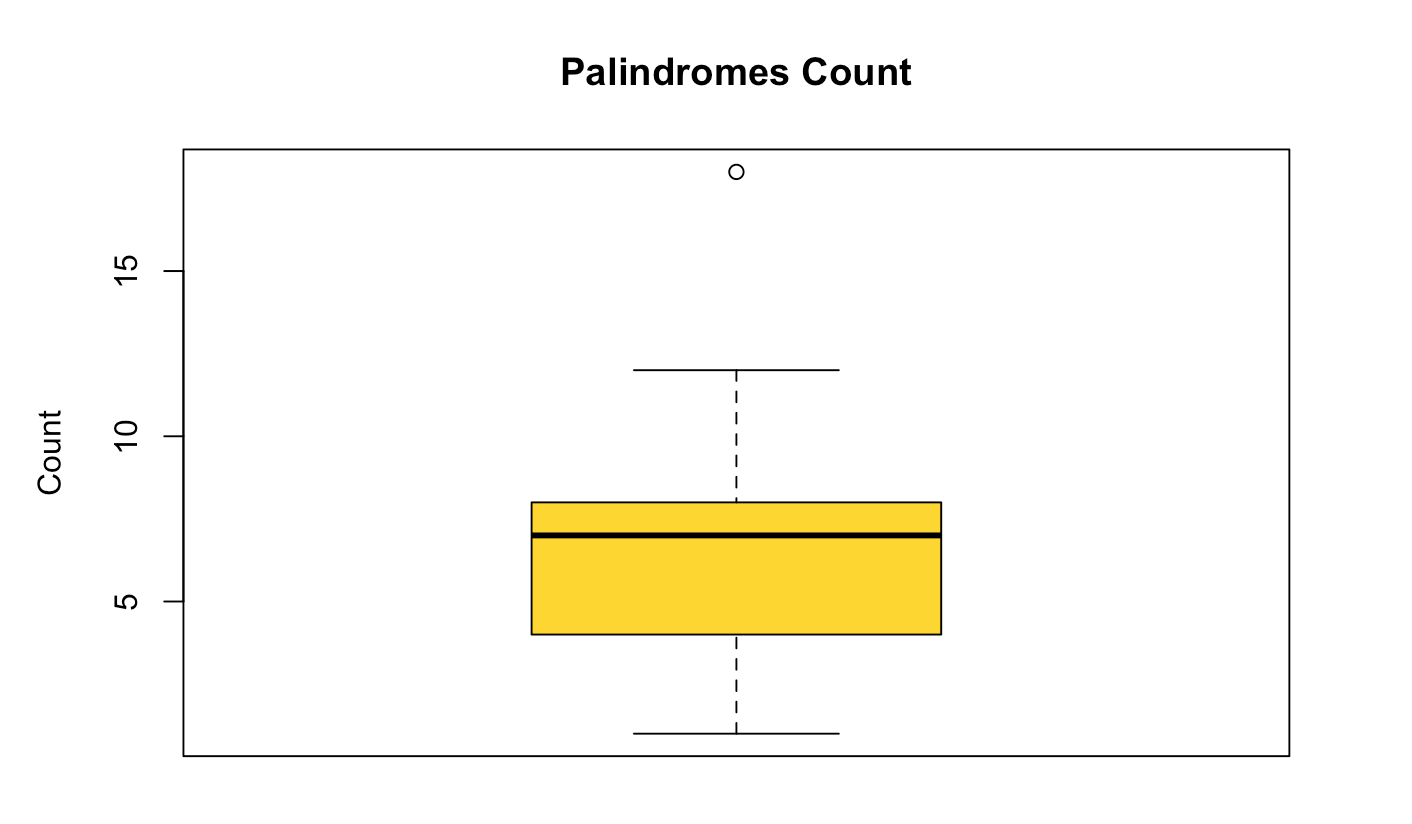
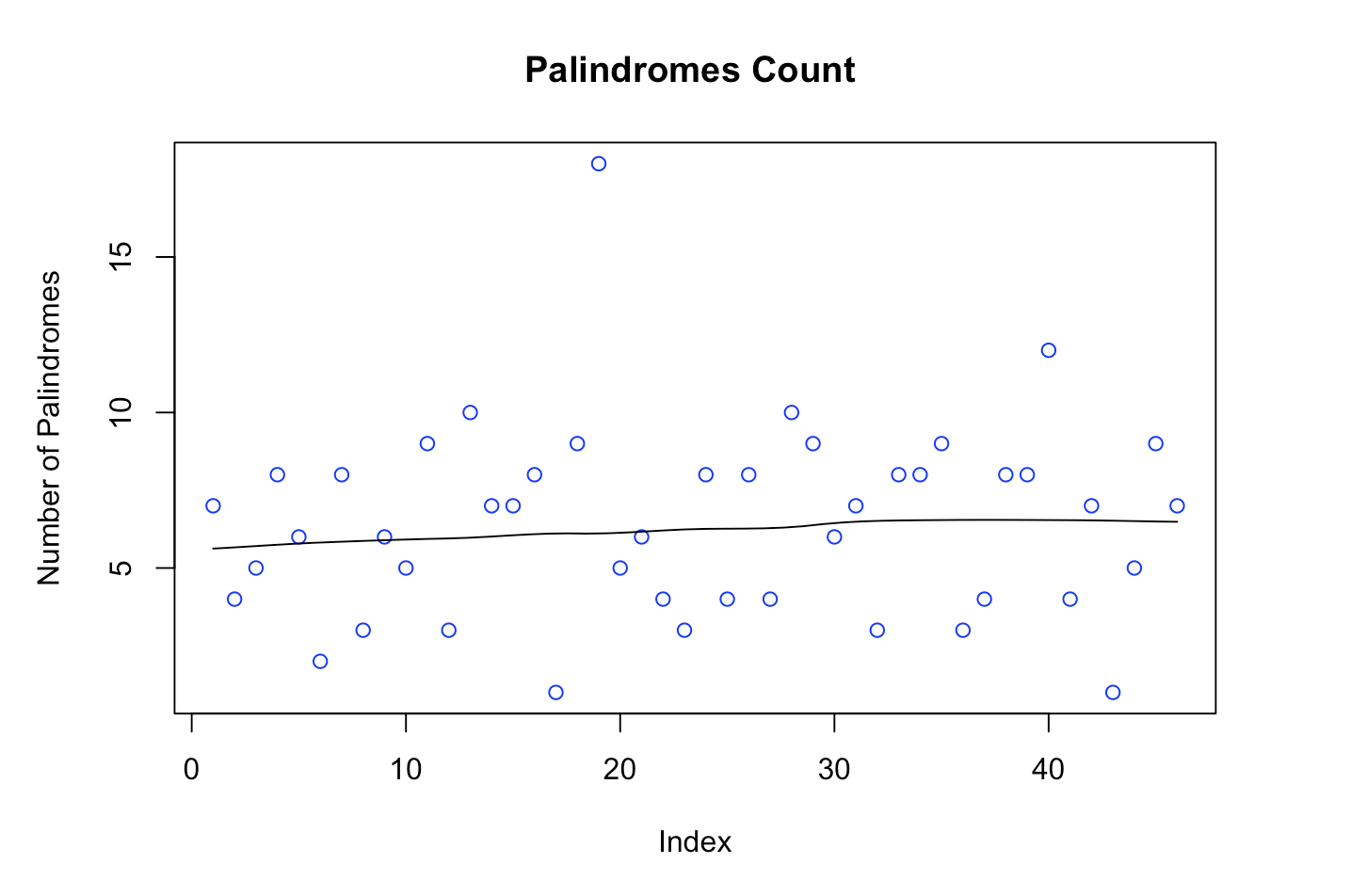


The test statistic using chi-squared goodness of fit test will be the summation of the squared difference of observed and expected number of segment from our table divided by the expected number of segments which is 7.24 and the corresponding p-value is 0.612. At 5% significance level, we do not reject the null hypothesis because the p-value is more than the 5%.

Result II:

The observed largest number of palindrome count is 18. Plotting the count of palindromes into a scatter plot gives a visualization that the maximum count has a significant difference from the rest of the counts. We would think that the maximum count is an outlier in our data and drawing out the box plot for this data proves this assumption even further. However, we can’t just assume this without doing hypothesis testing. Under the poisson process, we calculate the probability that the count is at least 18 and this yields to a very small number which is close to zero.

Therefore, at 5% significance level, we reject the null hypothesis that the palindromes are from a random scatter and we have strong evidence to conclude that the palindromes across the CMV DNA does not come from a poisson distribution.



**Discussion**

In the first method, we divided the locations of palindromes into 5000 non-overlapping segments and for each segment we counted the number of palindromes. Then, we use chi-squared test if these counts have the same distribution as the expected counts under the possion process. We concluded that it does indeed follow the distribution of poisson process which implies that the hits of palindromes are scattered randomly and uniformly. However, the disadvantage of this method is we just chose any number for the length of each interval without thinking any statistical process. This may arise faulty on this experiment since if the segments are very small, the counts of palindromes will be split between the segments and won’t appear as higher counts, and others might appear as high. So how big or how small should be the size of each interval? We need to repeat the process by choosing other size of interval and compare if there is a significance difference between our results. Since this method concludes that the hits of palindromes come from a random scatter, this does not give us information as where to begin our search for the origin of replication of the DNA of CMV.

While the first method states that the counts of palindromes follow an exponential distribution, the second method conclude the opposite. The observed maximum hit is very unlikely to happen under poisson process, so we concluded that the it does not follow a random process – the counts of palindromes in each segment is not due to chance, but there are other factors other than randomness that makes the count of palindromes higher. Since the maximum count of palindrome is the only outlier out of all the other count of palindromes, this would be a good place to start searching whether if this specific segment has the origin of replication or not.

The maximum counts of palindrome can be found in the interval (90000, 95000), which may contain the origin of reproduction of the DNA of CMV.