Palindromes in DNA sequences

Nick Collins

UMass Dartmouth

April 3, 2022

Abstract-

DNA gives scientists some great insight into how and what is happening in the human body. Looking at DNA can give vital predictions for diseases, hereditary genes and physical aspects of a person. The data for this assignment was all about DNA and DNA palindromes. The raw data was the locations of the palindromes. Using these locations, we break them down into certain intervals based on the distribution we want to examine. Using this data we can apply some statistical tools and methods to see the distribution of the data. We want to see exactly how these palindromes are spread out and if they follow a specific distribution.

Introduction-

The discovery of DNA has changed the way scientists view and understand the basics of the human body. Scientists often look into the DNA of a person to see the exactly what is going on. In this instance we are investigating human cytomegalovirus, of genetic palindromes: these are sequences that form complementary palindromes. We have the positions of these palindromes and we want to see if they are "random." If they were random, we would expect to see them evenly spread across the interval. To do this we need to use some statistical methods to see exactly how these locations are distributed.

Methods-

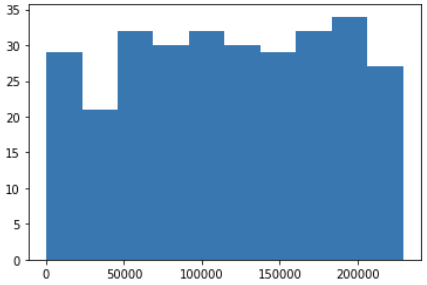
The first thing to do was to get the data from the course textbook. After importing the data and making it a list it was easy to split the data into different intervals. We can use some simple plotting functions to see what the data looks like. This will give us a little insight into how the data is acting.

As the book mentions we split the data into intervals of size 4,000 to start. We can change the size of the interval to see different results in the data. Using intervals of 4,000 we get 58 different intervals with the count of how many palindromes were in each step size of 4,000. The next step was to use this list and break it down into smaller groups based on the count. For example, the first group had counts zero through two, then it went three, four, five, six, seven, eight and nine plus. For a total of eight groups, which we can then use for the Poisson distribution.

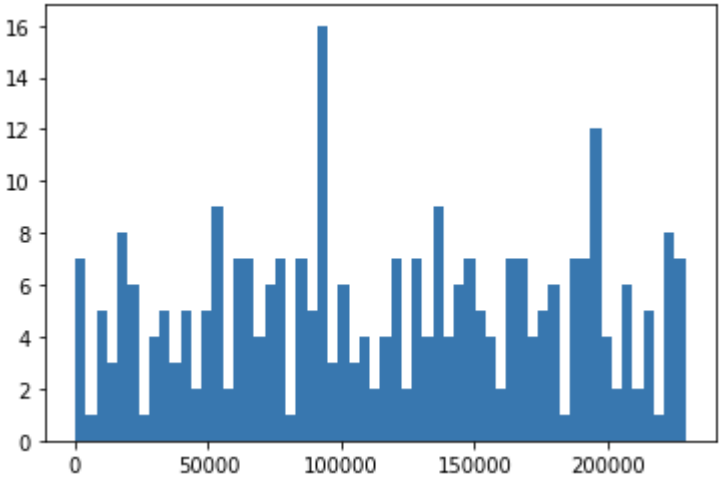
The next task was to see if the distribution follows a normal/uniform distribution. To use this method, we have to use a different interval size. For this I split the data into ten intervals, which gave a step size of 22,900. For this the expected counts in the interval would be 29.6, because there are a total of 296 palindromes divided among the ten intervals. After this we are able to use the observed counts to compare to the uniform distribution.

Results-

Plotting the list of positions of the palindromes we see the distribution below.



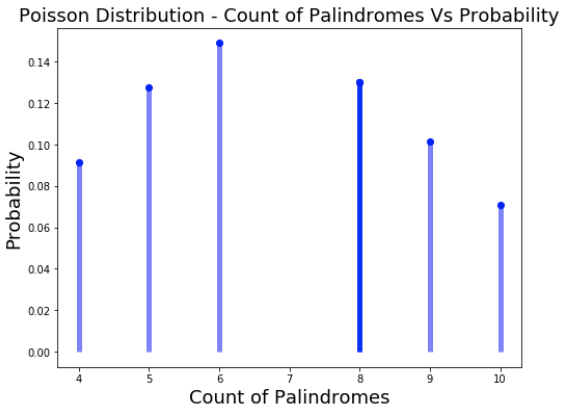
After changing the bin size to 58 we see a better representation of the data.



The second graph is being broken down into 58 different groups of step size 4,000. After breaking down the data into step sizes of 4,000 we get the following 58 groups.

```
[7, 1, 5, 3, 8, 6, 1, 4, 5, 3, 6, 2, 5, 8, 2, 9, 6,
4, 9, 4, 1, 7, 7, 14, 4, 4, 4, 3, 5, 5, 3, 6, 5, 3,
9, 9, 4, 5, 6, 1, 7, 6, 7, 5, 3, 4, 4, 8, 11, 5, 3,
6, 3, 1, 4, 8, 6, 2]
```

Once we have this breakdown we can organize the data into similar groups.



This graph shows the Poisson distribution for the count of palindromes compared to the probability of each occurrence. We get the counts from the original 58 groups, and we see the table below shows the breakdown of the counts.

The following table shows our observed and expected counts, like the one in the textbook.

Palindrome Count	Observed	Expected
0-2	8	7.4
3	8	6.5
4	10	9.7
5	9	10
6	8	8.6
7	5	6.3
8	4	4.1
9+	6	4.5
Total	58	58

We then use these observed and expected values to compute a test statistic. The test statistic computed was 1.003814. This value is our chi-squared goodness of fit test score.

The following table was produced with interval sizes of 22,900. Over the ten intervals we'd expect to see 29.6 palindromes in each group.

Group	Observed	Expected
1	28	29.6
2	22	29.6
3	32	29.6
4	29	29.6
5	33	29.6
6	30	29.6
7	29	29.6
8	32	29.6
9	34	29.6
10	27	29.6
Total	296	296

Computing a test statistic like we did above for this breakdown of the data we get a value of 4.277129. This is our chi-squared goodness of fit test score for the uniform distribution. The first histogram plot in the results section is the visual of the above table. The first histogram was

broken down into ten bins, and we see ten intervals in the table. This table and histogram can help to see if the data is a uniform/normal distribution. For this to be normal we would expect to see a peak around the middle with both tails slowly growing towards the middle. We don't see that in our distribution. It stays pretty much the same throughout the whole plot.

Conclusion-

After doing some investigations on the data to see what kind of distribution it follows we can say that it doesn't follow a Poisson or normal distribution. Of the two distributions it is closer to a Poisson distribution, but we can't say with enough evidence that it follows it well enough. I believe that the data is more "random" than it is a Poisson or normal distribution.

Although our Poisson distribution doesn't exactly fit the data, it is better than our normal distribution. When looking at the histogram with 58 bins, we get a better graph. There are some outliers that have a higher than average count, but these are only about two/three groups. If these groups were split into more groups our praph would be much better. In the graph we also see the randomness of the data, and the graph is scattered around, with peaks and dips in no particular order. When graphing the probability compared to the counts table, we see some interesting results. The graph looks promising as we see the counts look to center around six to eight, but then we see a probability of zero for count seven. We would expect to see count seven have around the same or an even higher probability than counts six and eight. This was a surprising result and would have been better if we saw some probability in count seven. However, when we compute the test statistic for the Poisson model we get a value of 1.003814. This value tells us how well our expected data fits our observed data. The higher the value the worse the observed data fits the expected data. Since our value is a little over one, this tells us our observed data fit the expected data pretty well.

In a normal/uniform distribution we would expect a histogram to start low on the minimum and maximum sides and grow until the peak. We don't see this in the first histogram in the results section. This histogram doesn't peak or dip throughout the ten bins. It stays roughly the same, so just looking at the data we wouldn't expect a normal distribution. To back up these claims we can look at our test statistic. This value was 4.277129, which was much higher than our test statistic value for the Poisson model. With this value being above four we can see that our observed data doesn't fit the predicted values.

References-

MTH 332 Mathematical Statistics Spring 2022

https://mth332.files.wordpress.com/2022/01/mth_332_spring_2022.pdf

Stat Labs: Mathematical Statistics Through Applications, by Deborah Nolan and Terry Speed

https://mth332.files.wordpress.com/2020/01/d.nolan-t.speed-stat-labs-mathematical-statistics-through-applications.pdf