Aryaman Singh, Hiroki Hoshida, Eric Manirasith
Prof. Armin Schwartzman
MATH189
May 17, 2020

MATH189 Homework 3

Contribution Statement

Aryaman Singh

- Collaborated on a Zoom conference with group members to brainstorm ideas for the project and helped finalize the steps for each scenario.
- Wrote out Section 4, Conclusion, and did analysis code.

Hiroki Hoshida

- Collaborated on a Zoom conference with group members to brainstorm ideas for the project and helped finalize the steps for each scenario.
- Wrote out Introduction, Section 1, and 2.

Eric Manirasith

- Collaborated on a Zoom conference with group members to brainstorm ideas for the project and helped finalize the steps for each scenario.
- Compiled the Appendix and Section 3.

Introduction

The human cytomegalovirus, known as CMV, is a widespread virus that infects humans, and can potentially be life-threatening for those with compromised immune systems. CMV lies dormant for the majority of the time, but becomes harmful when it enters a productive cycle, in which it replicates itself tens of thousands of times. DNA initiates replication on certain locations on the strand, known as origins of replication. By studying the origins of replications of CMV, scientists can create drugs and vaccines to better target CMV to protect those with insufficient immune systems.

Finding potential origins of replication can be done by looking for abnormalities in the randomness of the DNA strand. One key feature of the origin of replications for viruses in the

family of CMV are the complimentary palindromes. *Herpes simplex*, one virus in the family, has its origin of replication marked by a long complimentary palindrome. Another, the Epstein-Barr virus, has its origin of replication marked by multiple short complimentary palindromes. CMV altogether contains 296 palindromes between 10 and 18 base pairs long. Because of the relative shortness of these palindromes, it is conjectured that the origin of replication for CMV is likely marked in a similar method as the Epstein-Barr virus, in which the origin of replication contains a cluster of shorter complimentary palindromes.

The aim of our analysis was to determine whether or not the CMV DNA was a random uniform distribution, and in the case that it wasn't - to determine the base locations at which clusters of palindromes were present so that scientists could then focus their research on these parts. We approached the analysis in four different ways- comparing the locations of palindromes to randomly generated samples, the spacings between consecutive palindromes, the counts of palindromes in each interval and the intervals with the biggest clusters of palindromes. On the outset it seemed highly likely that the CMV DNA was taken from a random uniform sample, but comparing the spaces between consecutive palindromes with two and three in between to a random sample suggested that there existed more clusters in the DNA than in any random sample. Finally, by zooming onto these intervals with the perceived clusters, we ascertained locations with a high density of palindromes.

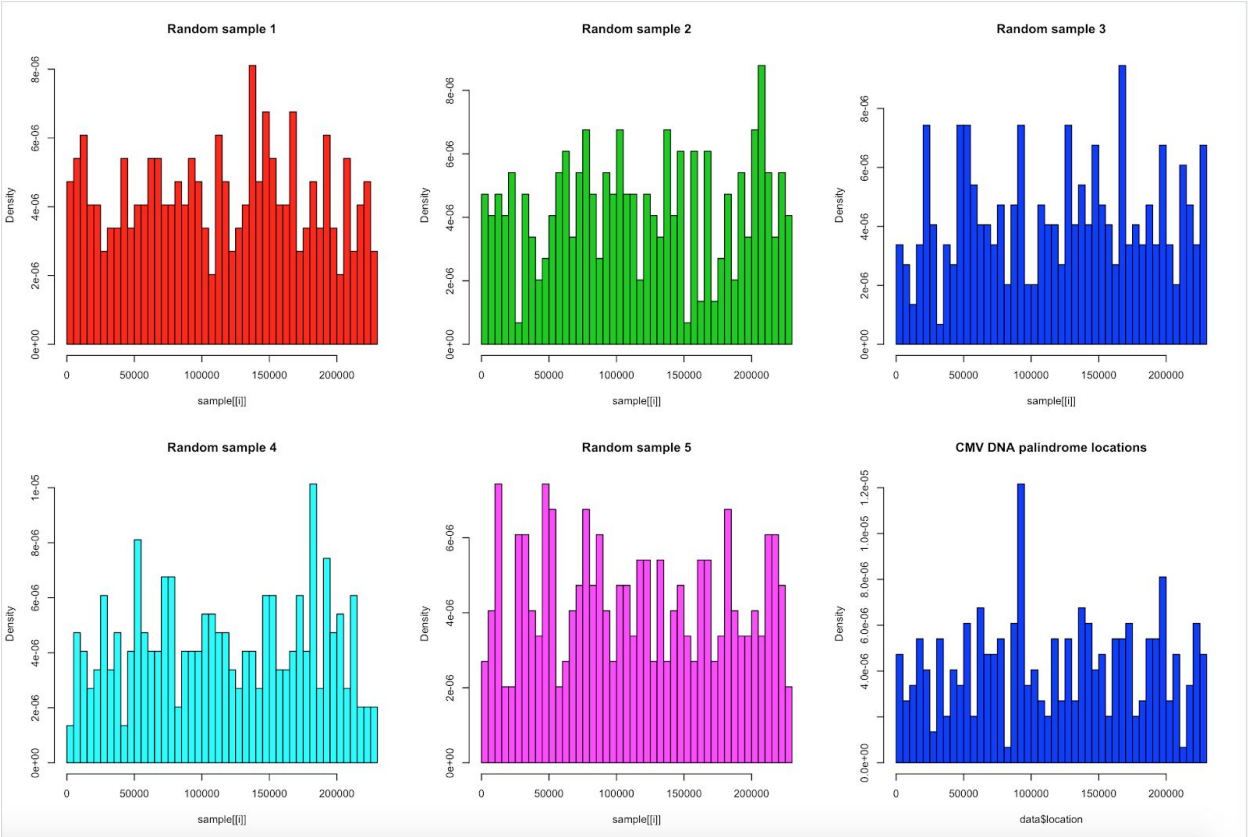
Analysis

We begin our analysis with the null hypothesis and alternate hypothesis.

H_0 = The CMV DNA sample comes from a uniform random scatter.
H_1 = The CMV DNA sample does not come from a uniform random scatter.

1. Locations

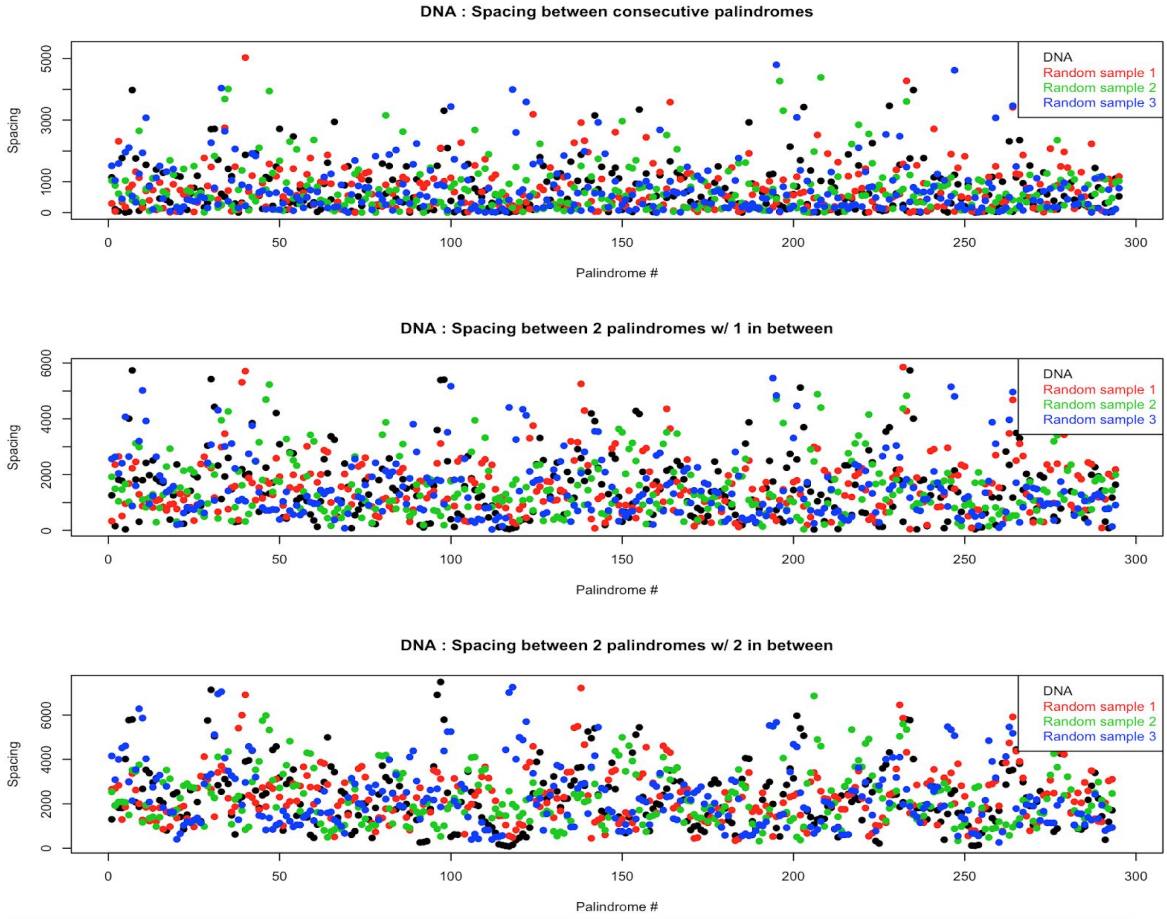
We attempted to compare the locations at which palindromes occur in the CMV DNA with locations at which palindromes appeared in a random sample with uniform distribution. We generated five different random samples in our R code and mapped them against the DNA sample. From the six distributions below we cannot see clearly any differences between the five random distributions and the sample distribution. Thus, we fail to reject the null hypothesis H_0



2. Spacings

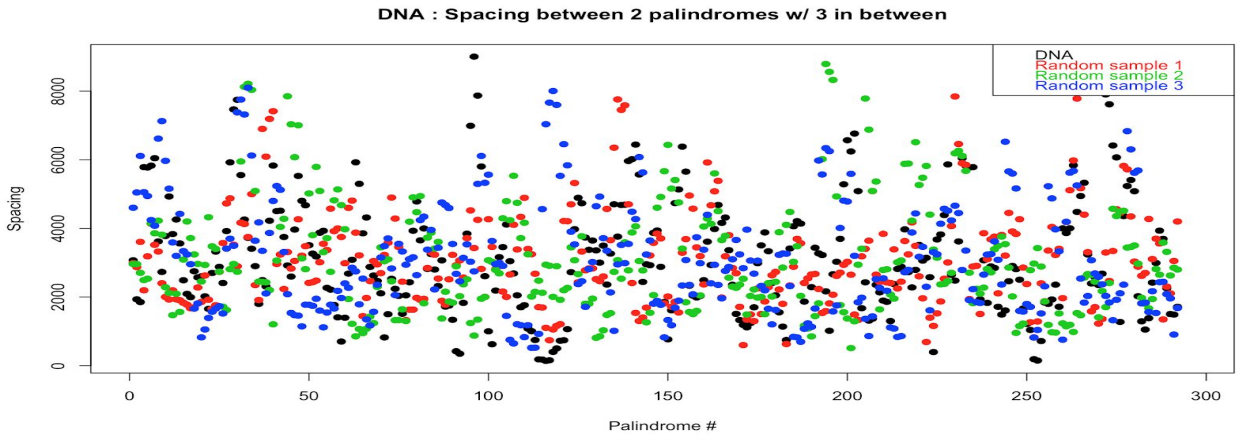
Next, we attempt to analyze the spacings of the palindromes in order to see if any regions of the DNA might have clusters of palindromes very close to each other. For this, we look at 3 types of spacings between the data -

I. Gap between consecutive palindromes
II. Gap between consecutive palindromes with one in between them (ie. palindrome i & i+2)
III. Gap between consecutive palindromes with two in between them (ie. palindrome i & i+3)



Once again, we compared these different spacings to those in the generated random samples using scatter plots. In cases (I) & (II) we don't see any difference between the spacings in the DNA sample and random samples. However, in case (III) we see that the DNA sample has certain palindromes clustered very close to each other (near palindrome #120 and #255).

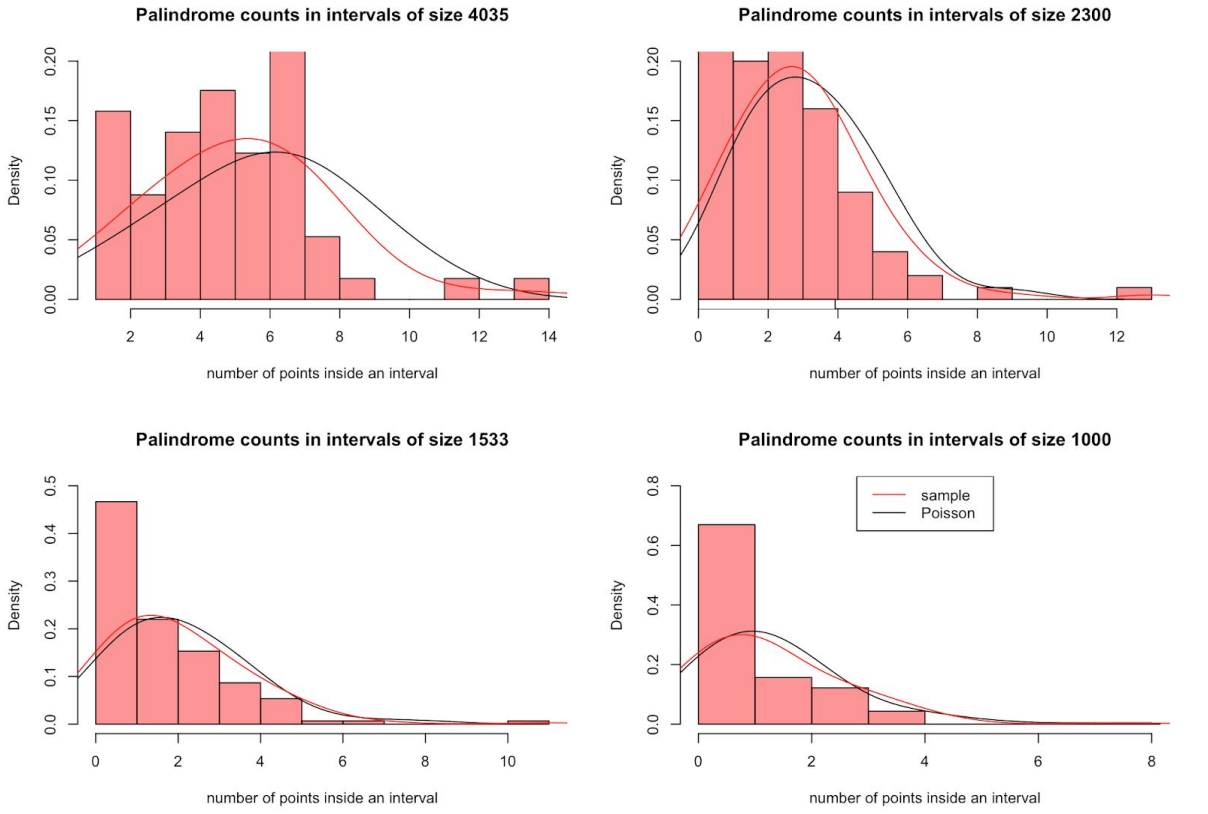
This would suggest that if we checked for case (IV) - Gap between consecutive palindromes with three in between them (ie. palindrome i & i+4), there would be much fewer clusters in the random samples, while these clusters would be more prominently visible. On checking the scatter plot obtained in this case, we see that this is indeed true.



Although the spacing of the palindromes does suggest that some of them are clustered together much closer than we could expect from a random uniform distribution, this is not enough evidence to reject H_0. The intervals with the clusters need to be analyzed further.

3. Counts

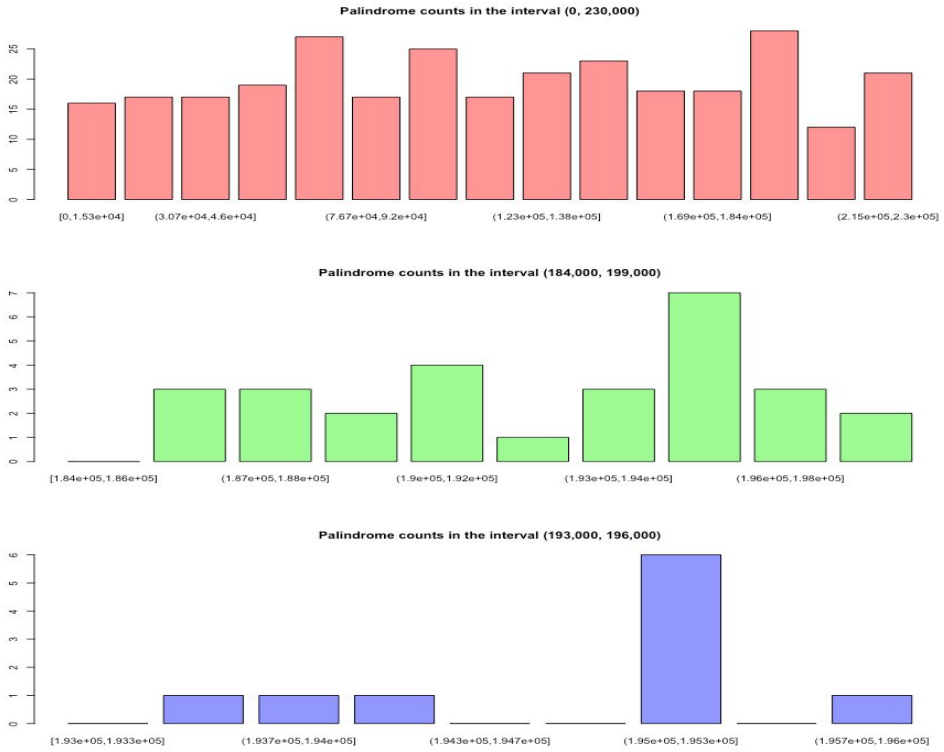
We now divide the palindrome locations into subintervals and observe the number of palindromes in each. We then compare the table of frequency of each count obtained to that for a Poisson distribution, both graphically and by using a Chi-square Goodness of Fit Test.



The graphs seem to be in line with what we would expect from a Poisson distribution, as is also evidenced by the density plot for the Poisson probabilities versus the sample. Moreover, on performing a Chi-squared test we obtained a p-value of 0.31 for 57 intervals, and a p-value of 0.93 for 100 intervals. Thus, by looking at the counts of palindromes by interval it seems likely that the null hypothesis is true.

4. Largest Cluster

Using the results from Section 2, we decided to look specifically at the intervals in which the palindromes appeared to be clustered. Repeatedly zooming in on intervals, we arrived at the interval between 193,000 and 196,000. Once more, we zoomed in on this interval, dividing it into 5 sub-intervals. We observed that out of the 10 palindromes in this interval, 6 of them were in the sub-interval (194,800, 195,400). Seeing this anomaly, we ran a chi-square goodness of fit test on this interval, with the expected values we would get from a Poisson distribution. This test gave us a X-value of 12.86 and p-value of 0.02472. We tried the same process by zooming onto the interval between 93,000 and 96,000, which appeared to show clusters of palindromes in Section 2. Again, we found that the sub-interval (92600,93000) had 6 palindromes while the remaining sub-intervals had a maximum of 2. Applying the Chi-square test we obtained an X-value of 13.471 and a p-value of 0.06144.



This implies that there was a 2.4% and 6.1% chance respectively that these sub-intervals were a part of a Poisson distribution. This is a significantly low chance, and so it is **enough evidence for us to reject the null hypothesis H_0**

Conclusion

We began our study with the Null Hypothesis. Our aim was to see whether we could reject the null hypothesis H_0 : The CMV DNA sample comes from a uniform random scatter. For this, we analyzed the data using 4 different approaches.

1. **Palindrome Locations**
    - We generated 5 random samples with uniform distribution of size 296 with possible values between 0 and 230,000. On comparing their distribution to that of our sample we did not find enough differences to reject H_0 .
2. **Spacings**
    - When comparing the gap between consecutive palindromes, the CMV spacings do not appear very different from the random spacings.
    - However, when comparing sums of spacings of pairs and triplets of palindromes, we begin to see clusters of palindromes in the CMV data with very little gap.
    - Still, this doesn't serve as significant evidence that the data doesn't follow Poisson distribution. Hence we failed to reject H_0.
3. **Palindrome counts by interval size**
    - We checked the distribution of palindrome counts for different interval sizes. All of the graphs we obtained appeared to closely follow a Poisson distribution.
    - Moreover, we ran a Chi-squared Goodness of Fit Test which gave us a minimum p-value of 0.31 (for 57 intervals), and a p-value of 0.93 for 100 intervals.
    - This seems to point very strongly to the fact that the data follows a Poisson distribution, so the null hypothesis H_0 could not be rejected.
4. **Biggest Clusters**
    - We investigated further into the intervals which appeared to contain clusters in section 2.
    - We zoomed onto these intervals by further dividing them into sub-intervals. Eventually we found sub-intervals with disproportionately high numbers of palindromes when compared to other sub-intervals in that particular interval.

- In particular, we found 6 palindromes between locations 92,600 and 93,000 and 6 palindromes between locations 194,800 and 195,400.
- Running Chi-squared Goodness of Fit Tests on the intervals containing these sub-intervals, we received p-values of 0.02 and 0.06. Thus, we considered this to be significant evidence that the data did not follow a Poisson distribution. Hence, **we rejected the null hypothesis H_0**

We conclude that the regions between 92,600 and 92,800 and between 194,800 and 195,400 contain clusters of palindromes, and would thus be likely to be the most prone to contain data about the replication of the virus. Thus, any scientists looking to study the virus are recommended to begin in these locations.

Appendix

```
#Generating 5 random samples with 296 locations

set.seed(22343)
n <- 296
for ( i in 1:5) {
sample[[i]] <- runif(n, min = 0, max = 229354)
}

#Sorting the sample (because the palindrome locations in the DNA are
sorted)

for (i in 1:5) {
  sample[[i]] = sort(sample[[i]])
}

#Comparing DNA vs Random Sample distribution using histograms

par(mfrow = c(2,3))
```

```
for ( i in 1:5) {
hist(sample[[i]], breaks = 50, probability = TRUE, col = i+1,
main=paste("Random sample", i, sep=" "))
}
hist(data$location, breaks = 50, probability = TRUE, col = 4, main =
"CMV DNA palindrome locations")

#Calculating the gap between the 3 types of spacings

samplegap = c()
samplegap2 = c()
samplegap3 = c()

#Consecutive
gap = c()
for(i in 1:nrow(data))
{
  gap[i] = data$location[i+1]-data$location[i]
}
for(i in 1:length(sampleMean)-1)
{
samplegap[i] = sample[[1]][i+1]-sample[[1]][i]
}

#Consecutive w/ 1 in b/w
gap2 = c()
for(i in 1:nrow(data))
{
  gap2[i] = data$location[i+2]-data$location[i]
}
for(i in 1:length(sampleMean)-2)
{
samplegap2[i] = sample[[1]][i+2]-sample[[1]][i]
}

#Consecutive w/ 2 in b/w
gap3 = c()
for(i in 1:nrow(data))
{
  gap3[i] = data$location[i+3]-data$location[i]
}
for(i in 1:length(sampleMean)-3)
{
```

```
samplegap3[i] = sample[[1]][i+3]-sample[[1]][i]
}

#Creating intervals
tab <- table(cut(data$location, breaks = seq(0,230000, length.out =
k), include.lowest = TRUE))
count = as.vector(tab)
int = table(count)

#Setting up probabilities of the required length for chi-sq test using
dpois
Pois <- dpois(0:6, lambda = mean(count))
Pois[8] = 1-sum(Pois)
chisq.test(int,p=Pois)

#Zooming onto subsections
tab <- table(cut(data$location, breaks = seq(92400,92850, length.out
= 10), include.lowest = TRUE))
tab <- table(cut(data$location, breaks = seq(92500,92850, length.out
= 10), include.lowest = TRUE))
tab <- table(cut(data$location, breaks = seq(92500,92800, length.out
= 10), include.lowest = TRUE))
tab <- table(cut(data$location, breaks = seq(92600,93000, length.out
= 10), include.lowest = TRUE))
tab <- table(cut(data$location, breaks = seq(92600,93000, length.out
= 57), include.lowest = TRUE))
tab <- table(cut(data$location, breaks = seq(92600,93000, length.out
= 57), include.lowest = TRUE))
tab <- table(cut(data$location, breaks = seq(0,230000, length.out =
57), include.lowest = TRUE))
tab2 <- table(cut(data$location, breaks = seq(193000,196000,
length.out = 10), include.lowest = TRUE))
tab2 <- table(cut(data$location, breaks = seq(193000,196000,
length.out = 6), include.lowest = TRUE))
tab2 <- table(cut(data$location, breaks = seq(195000,195300,
length.out = 6), include.lowest = TRUE))
tab2 <- table(cut(data$location, breaks = seq(19500,195300,
length.out = 6), include.lowest = TRUE))
tab2 <- table(cut(data$location, breaks = seq(194800,195400,
length.out = 6), include.lowest = TRUE))
tab2 <- table(cut(data$location, breaks = seq(193000,196000,
length.out = 6), include.lowest = TRUE))
```

```
#Comparing different interval lengths to Poisson random variables of
that length
pois230 = rpois(230,lambda = mean(count230))
pois150 = rpois(150,lambda = mean(count150))
pois100 = rpois(100,lambda = mean(count100))
poisValues = rpois(57,lambda = mean(counts))
count230 = as.vector(tab230)
tab230 <- table(cut(data$location, breaks = seq(0, 230000, length.out
= k+1), include.lowest = TRUE))
count150 = as.vector(tab150)
tab150 <- table(cut(data$location, breaks = seq(0, 230000, length.out
= k+1), include.lowest = TRUE))
count100 = as.vector(tab100)
tab100 <- table(cut(data$location, breaks = seq(0, 230000, length.out
= k+1), include.lowest = TRUE))
```