Estimation and Testing on Patterns in DNA

by Edward Huynh

March 1, 2020

1. **Abstraction:** In this investigation report, student will determine whether the Poisson and Uniform distribution fit the estimation of palindrome locations in human cytomegalovirus (CMV) is potentially life-threatening disease. This investigation of palindromes can assit scientises to find a way to treat the diease. This statistical method help to narrow the research in both time and cost of examinate full DNA sequence of CMV. In this paper, student provide a mathematical basis for the assumption of Palindromes by showing randomly generated DNA sequences, the occurrences of palindromes can be approximated by a Poisson process.

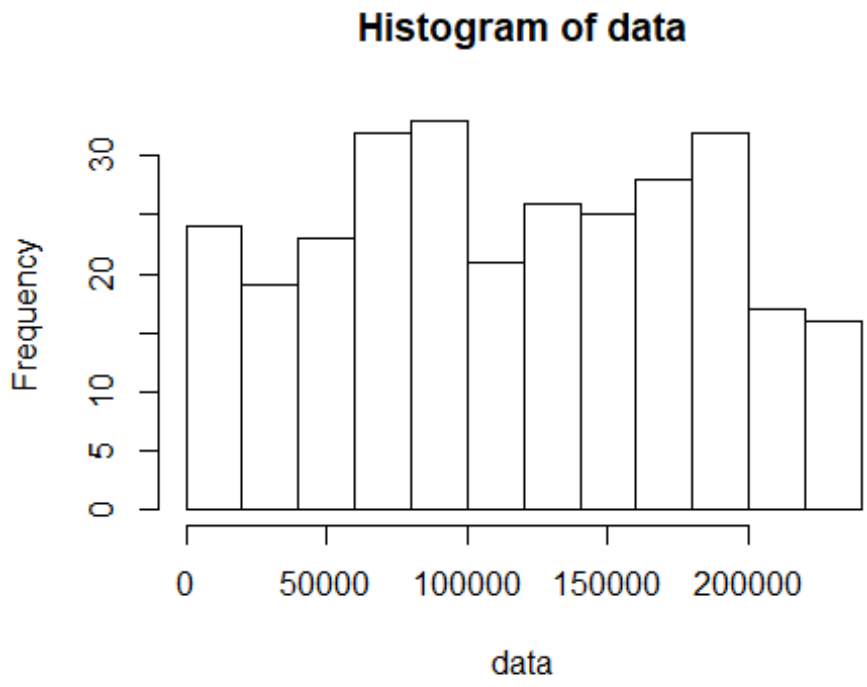
2. **Introduction:** From science poit of view, people with suppressed or deficient immune system, tne CMV can be life-threatening diease. Understanding how this virus replicates assists scientists to investigate the virus with better solution to treat it. The Palindromes refer to specfic patterns in DNA where sequence of letter reads the same in reverse direction as in the forward complementary sequence due to limited variety of bases. Identifying the palindrome locations is very helpful for examining stastically similar patterns of this virus. A computer can simulate 296 palindrome sites chosen at random along a DNA sequences of 229,354 bases by using a pseudorandom number generator. These sequences can be investigated via statistical analyses to determine amounts of palindromes in a chunk of a set length. A Poisson process is a model for randome phenomena. In this paper, student will investigate the rate of sequence at which points occur doesn't change with location. Student also figure out the number of points falling in seprate regions are independent. The distribution of frequency of palindrome occurrences in the CMV DNA sequence better fit either with Poisson distribution or uniform distribtion methods.

3. **Methods** In order to determine if the CMV DNA sequence better fits a Poisson or uniform distribution. Student will analyze the data set Palidromes in Cytomegalo Virus DNA from Stat.Berkeley labs. Using Rstudio to import the data and do basic statistical analysis for further investigate on histogram, computing summary, coding to get skewness and kurtosis values. Separate the data in different chunks to analyze. The number of palindromes in each chunk were counted and displayed as a dataframe to determine the frequency of palindrome count within a chunk. The frequency in observations are random variable for Poisson model. Then running a chi-squared test for dataset to see the goodness of fit tests.

4. Results:

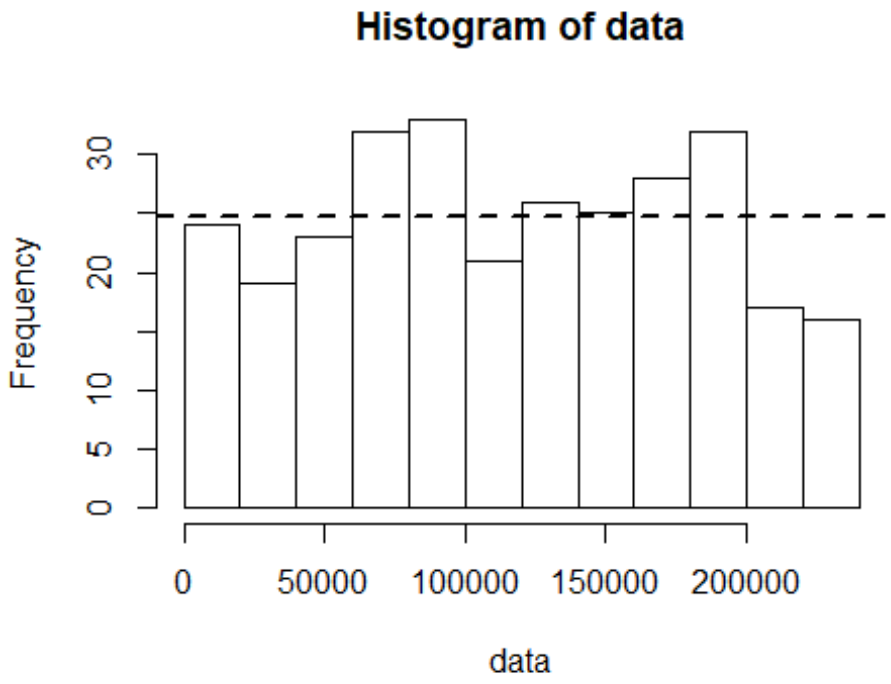
Input data from dataset in Stat.Berkeley labs. Data Frame then make Histogram of data

```
data<-read.table("C:/Users/Sang/Documents/hcmv.data", header=TRUE,sep="")
str(data)

## 'data.frame':    296 obs. of  1 variable:
##  $ location: int  177 1321 1433 1477 3248 3255 3286 7263 9023 9084 ...

data<-as.numeric(data$location)
h1<-hist(data)
```



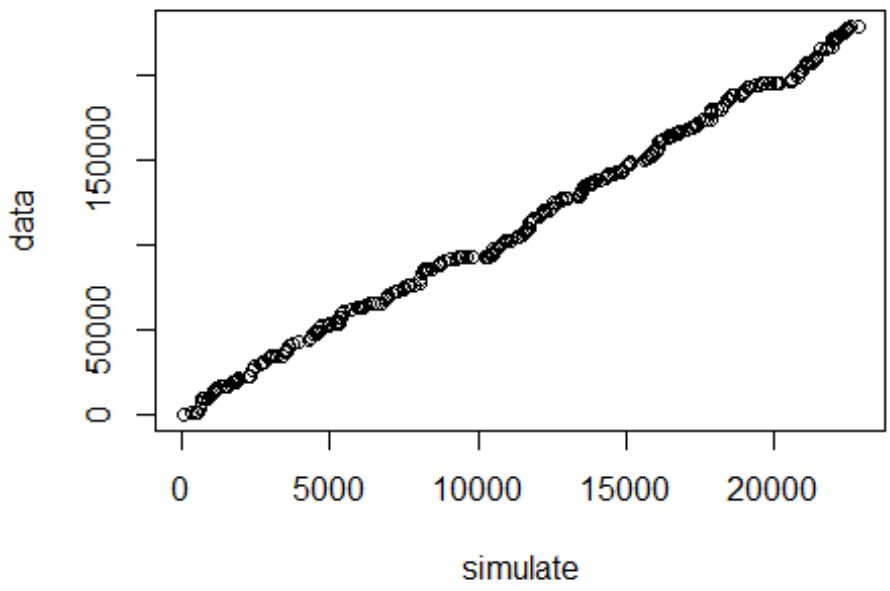
Mean valuable of dataset and make abline for histogram

```
mean(h1$counts)

## [1] 24.66667

plot(h1)
abline(h=24.66667, lwd=2,lty=2)
```



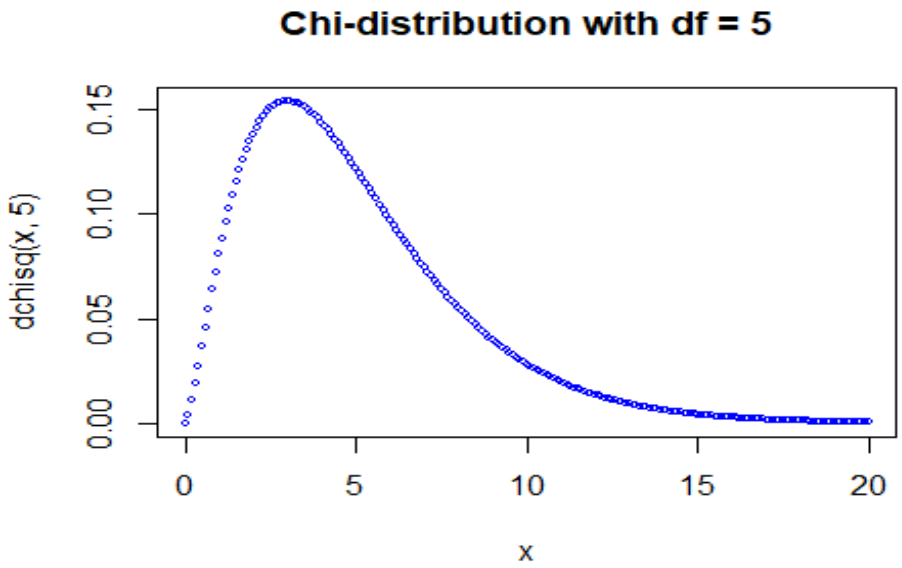
Summary values of data with q-q plot

```
summary(data)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     177   63714  117826  116960  171144  228953

sd(data)

## [1] 64732.03

library(moments)
skewness(data)

## [1] -0.02258943

kurtosis(data)

## [1] 1.864581

simulate<-sample(1:22900,296,replace=TRUE)
simulate<-as.numeric(simulate)
qqplot(simulate,data)
```



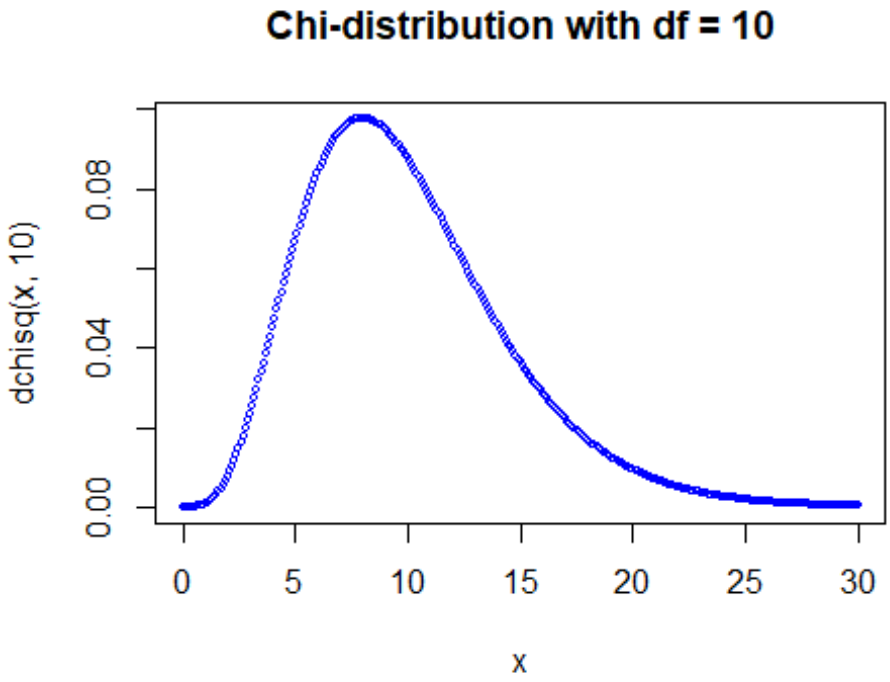
Making goodness of fit test with df 5

```
x=seq(0,20,0.1)
p1<-plot(x, cex = .5, dchisq(x,5),
main = "Chi-distribution with df = 5", col = "blue")
```



Making goodness of fit tests with df 10

```
x=seq(0,30,0.1)
p1<-plot(x, cex = .5, dchisq(x,10),
main = "Chi-distribution with df = 10", col = "blue")
```



Analyze Chi-square distribution test data

```
chisq.test(data)

##
##  Chi-squared test for given probabilities
##
## data:  data
## X-squared = 10568729, df = 295, p-value < 2.2e-16

f<-function(x) {dchisq(x,5)}
int<-integrate(f, lower=0, upper=1)
pPDF5<-1 -int$value
pPDF5

## [1] 0.9625658f<-function(x) {dchisq(x,10)}
int<-integrate(f, lower=0, upper=2)
pPDF10<-1 -int$value
pPDF10

## [1] 0.9963402
```

```r
#Uniform Distribution

observed = c(29, 21, 32,30,32, 31, 28, 32,34, 27)
expected = c(29.6/296, 29.6/296,29.6/296,29.6/296,
29.6/296,29.6/296,29.6/296,29.6/296,29.6/296,29.6/296)
chisq.test(x=observed, p=expected)

##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 4.1351, df = 9, p-value = 0.9023
```

#Poisson Distribution

```r
O3<-c()
O4<-c()
O5<-c()
for (i in 1:57)
  {O4[i]<-sum(data >= 4000*(i-1) & data <=4000*(i))
  O3[i]<-sum(data >= 3000*(i-1) & data <=3000*(i))
  O5[i]<-sum(data >= 5000*(i-1) & data <=5000*(i))}

O3
```

```
##  [1]  4  3  1  5  3  4  6  4  1  1  4  4  3  3  3  2  3  6  5  1  5  8  3
## [24]  3  5  7  1  1  6  3 13  6  3  2  5  2  3  1  4  5  3  2  5  4  2  8
## [47]  3  8  1  5  6  3  1  2  6  5  6
```

```r
O4
```

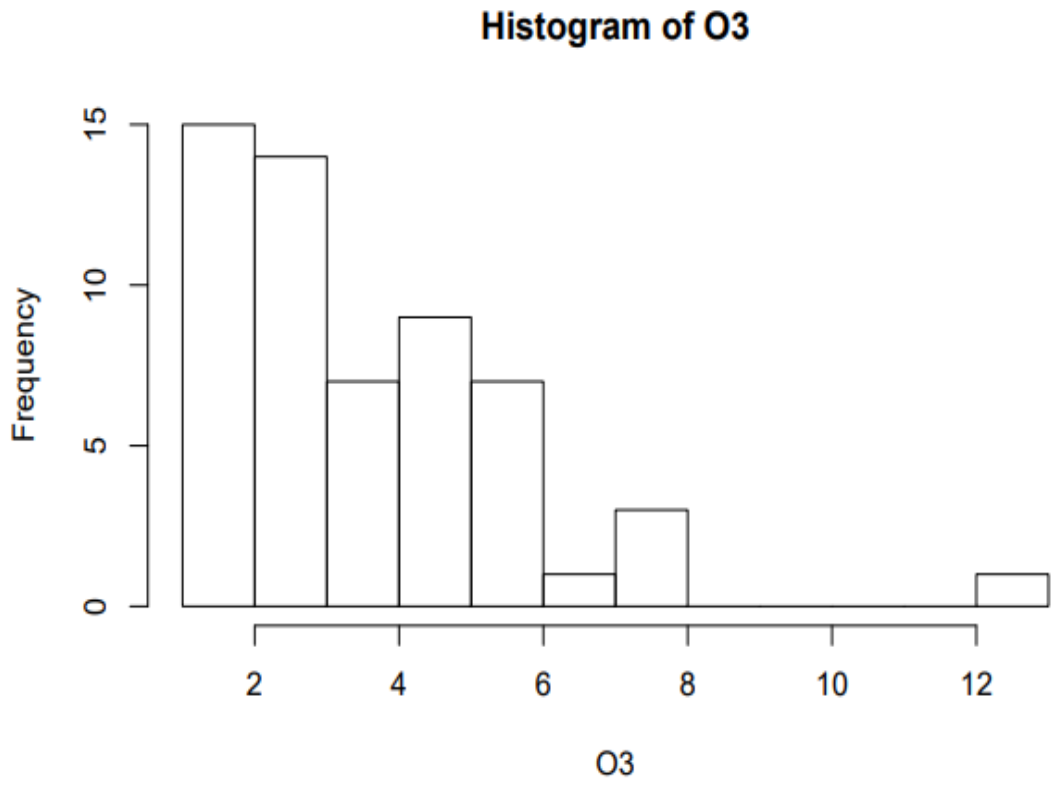
```
##  [1]  7  1  5  3  8  6  1  4  5  3  6  2  5  8  2  9  6  4  9  4  1  7  7
## [24] 14  4  4  4  3  5  5  3  6  5  3  9  9  4  5  6  1  7  6  7  5  3  4
## [47]  4  8 11  5  3  6  3  1  4  8  6
```

```r
O5
```

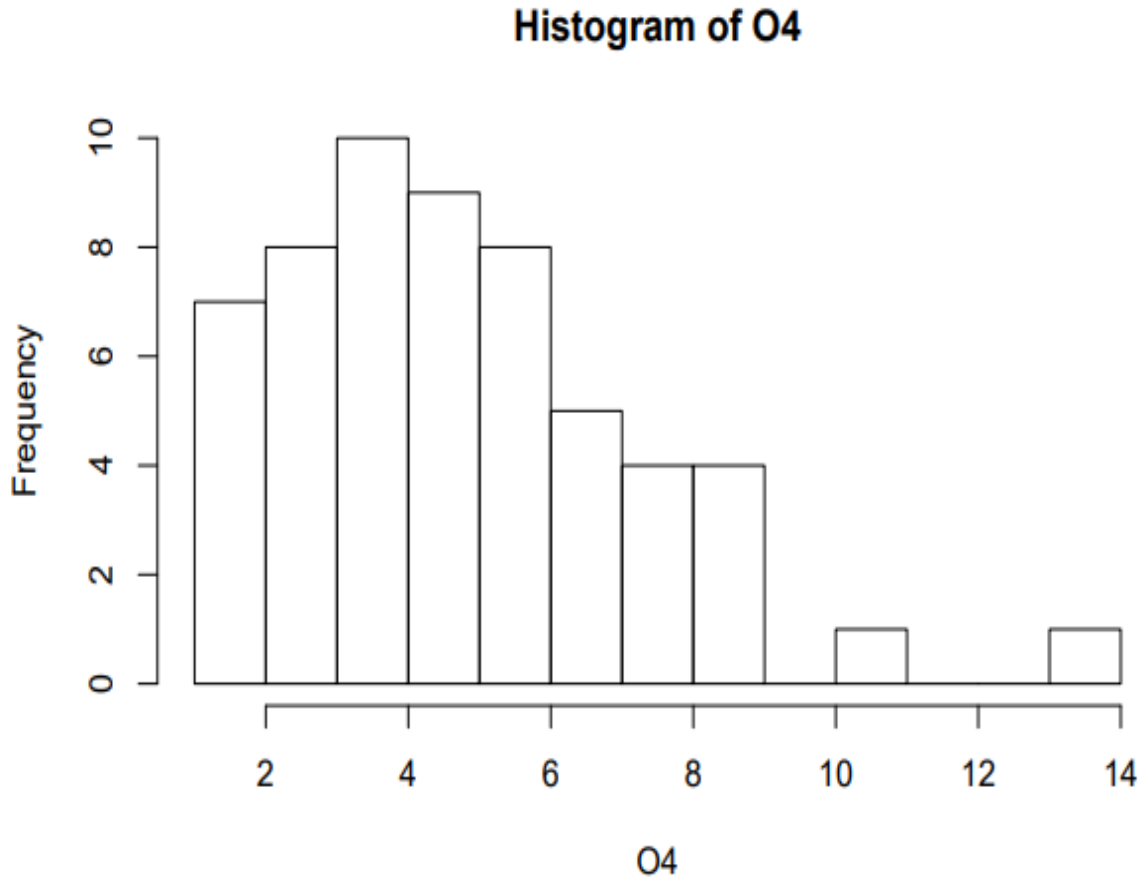
```
##  [1]  7  4  5  8  6  2  8  3  6  5  9  3 10  7  7  8  1  9 18  5  6  4  3
## [24]  8  4  8  4 10  9  6  7  3  8  8  9  3  4  8  8 12  4  7  1  5  9  7
## [47]  0  0  0  0  0  0  0  0  0  0  0
```

Number of Interval Observations

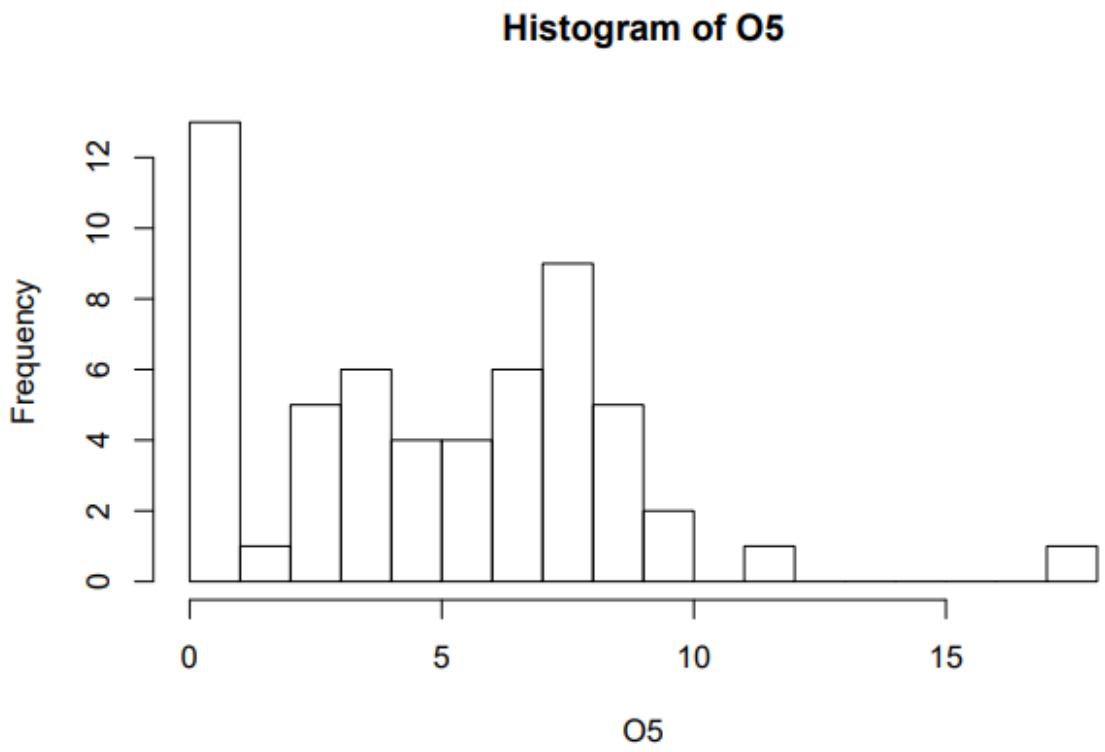
```
o3<-hist(O3,breaks=14)$counts
```



```r
o4<-hist(O4,breaks=14)$counts
```



```r
o5<-hist(O5,breaks=14)$counts
```




```r
o4[8]<-sum(o4[8:length(o4)])
o3[8]<-sum(o3[8:length(o3)])
o5[8]<-sum(o5[8:length(o5)])

o4<-o4[-c(9:length(o4))]
o3<-o3[-c(9:length(o3))]
o5<-o5[-c(9:length(o5))]

o4
```

```
## [1]  7  8 10  9  8  5  4  6
```

```r
o3
```

```
## [1] 15 14  7  9  7  1  3  1
```

```r
o5
```

```
## [1] 13  1  5  6  4  4  6 18
```

Number of Interval Expected

```
E4<- c()
```

```
E3<-c()
E5<-c()

l4<-round((sum(O4)/57),2)
l3<-round((sum(O3)/57),2)
l5<-round((sum(O5)/57),2)

E4[1]<-(1+l4+((l4^2)/2))*exp(-l4)*57
E3[1]<-(1+l3+((l3^2)/2))*exp(-l3)*57
E5[1]<-(1+l5+((l5^2)/2))*exp(-l5)*57

for (k in 3:13)
  {E4[k-1]<-(l4^k)*exp(-l4)*57/factorial(k)
  E3[k-1]<-(l3^k)*exp(-l3)*57/factorial(k)
  E5[k-1]<-(l5^k)*exp(-l5)*57/factorial(k)}

E4[8]<-sum(E4[8:12])
E3[8]<-sum(E3[8:12])
E5[8]<-sum(E5[8:12])

E4<-round(E4, 1 )
E3<-round(E3, 1 )
E5<-round(E5, 1 )

E4<-E4[-c(9:length(E4))]
E3<-E3[-c(9:length(E3))]
E5<-E5[-c(9:length(E5))]
```

```
E4
```

```
## [1]  6.4  7.5  9.7 10.0  8.6  6.3  4.1  4.5
```

```
E3
```

```
## [1] 14.5 11.4 11.1  8.7  5.6  3.1  1.5  1.0
```

```
E5
```

```
## [1]  6.2  7.4  9.6 10.0  8.6  6.4  4.1  4.6
```

Testing static of Expected interval

```
t3<-c()
t4<-c()
t5<-c()

for (i in 1:8)
  {t4[i]<-(o4[i]-E4[i])^(2)/E4[i]
  t3[i]<-(o3[i]- E3[i])^(2)/E3[i]
```

```
  t5[i]<-(o5[i]-E5[i])^(2)/E5[i]}
t4<-sum(t4[1:8])
t3<-sum(t3[1:8])
t5<-sum(t5[1:8])

t4<-round(t4,1)
t3<-round(t3,1)
t5<-round(t5,1)

t4
```

```
## [1] 1
```

```
t3
```

```
## [1] 5.4
```

```
t5
```

```
## [1] 60.1
```

Goodness of Fit Test

#Chi-square distribution with 6 degree of freedom for 4000 intervals

```
f <- function(x) {dchisq(x,6)}
  int<- integrate(f, lower = 0, upper = 1)
```

```
## [1] 0.9856123
```

#Chi-square distribution with 6 degree of freedom for 3000 intervals

```
f <- function(x) {dchisq(x,6)}
  int<- integrate(f, lower = 0, upper = 5.4)
```

```
## [1] 0.4936245
```

#Chi-square distribution with 6 degree of freedom for 5000 intervals

```
f <- function(x) {dchisq(x,6)}
  int<- integrate(f, lower = 0, upper = 60.1)
```

```
## [1] 4.295297e-11
```

#Summary Tests

```
tests<- matrix(c(t4,t3,t5,four, three,five), ncol=2,byrow=FALSE)

colnames(tests) <- c("Test Stat","chi-square")

rownames(tests) <- c("4000 int","3000 int","five int")

tests<- as.table(tests)
tests
```

```
##             Test Stat   chi-square
## 4000 int 1.000000e+00 9.856123e-01
## 3000 int 5.400000e+00 4.936245e-01
## five int 6.010000e+01 4.295297e-11
```

5. Discussion & Conclusions

The first histogram of the palindrome doesn't show any indication as to the uniform spread of the data because the values are not equal to the observed.

Since the value of theoretical discrete uniform distribution are difference with the investigate values above (skewness, and kurtosis), this distribution is not symmetric. Therefore, the palindrome locations are not uniform distributed due to not having the same frequency.

However, the Q-Q plot show that palindrome locations are uniformly distributed integers because it look like a straight line.

The Chi-square testing shows that the intervals different should not over 10 because the results of goodness if fit test is uniform distribution. Therefore, for larger number of intervals, it's better to use method of poisson distribution.

As follow the analysis mention in the textbook. In order to obtain the tables of Palindrome counts in first 57 non-overlapping intervals of 4000, 3000, and 5000 base pairs of CMV DNA, student analyzes for number of interval observed by constructing histograms at different intervals. Then find the number of intervals expected, and finally computed the test statistics to check whether the model is true. The result above show that the random scatter model is true at 4000 intervals by the test statistic which number showed less than 3, and the chi-square goodness of fit test is close to 1.