FATAL LAMB ANALYSIS

2024-06-05

```r
knitr::opts_chunk$set(echo = FALSE , warning=FALSE , comment = NULL)
```

INTRODUCTION

In this study, we are going to use the dataset Fetal Lamb; data on the counts of movements in five-second intervals of one fetal lamb(n= 240 intervals). Our goal is to to analyze this data and determine whether or not this data are well described by a probability distribution.Knowing the distribution that the data best fits will help us to make inferences and draw perfect conclusions. Probability distributions such as the Binomial , Poisson and Negative binomial distributions are used to model the data to find the one that best suit the data. Key features such as the variance - mean ratio, Method of Moments estimation, Maximum Likelihood estimation and other graphical displays such as histograms are also used to understand the distribution and relationships.

DATASET

The dataset used in this study is Fetal Lamb; data on the counts of movements in five-second intervals of one fetal lamb(n= 240 intervals).

OBJECTIVES

In this report, we seek to solve the following problems.

- Perform exploratory data analysis and determine the probability distribution that best fit the data.
- Find the method of moments estimators for the parameters.
- Find the maximum likelihood estimators for the parameters.
- Predict expected observations using the probability distribution selected.
- Perform chi-squared Goodness of fit (GOF) test
- Graph the fitted observation with the actual observations

PRELIMINARY ANALYSIS

Data Preprocessing

To begin, we first understand and preprocess the data.

```r
movements<-c(0,1,2,3,4,5,6,7)
counts<- c(182,41,12,2,2,0,0,1)
data<-cbind(movements,counts)
kable(data)
```

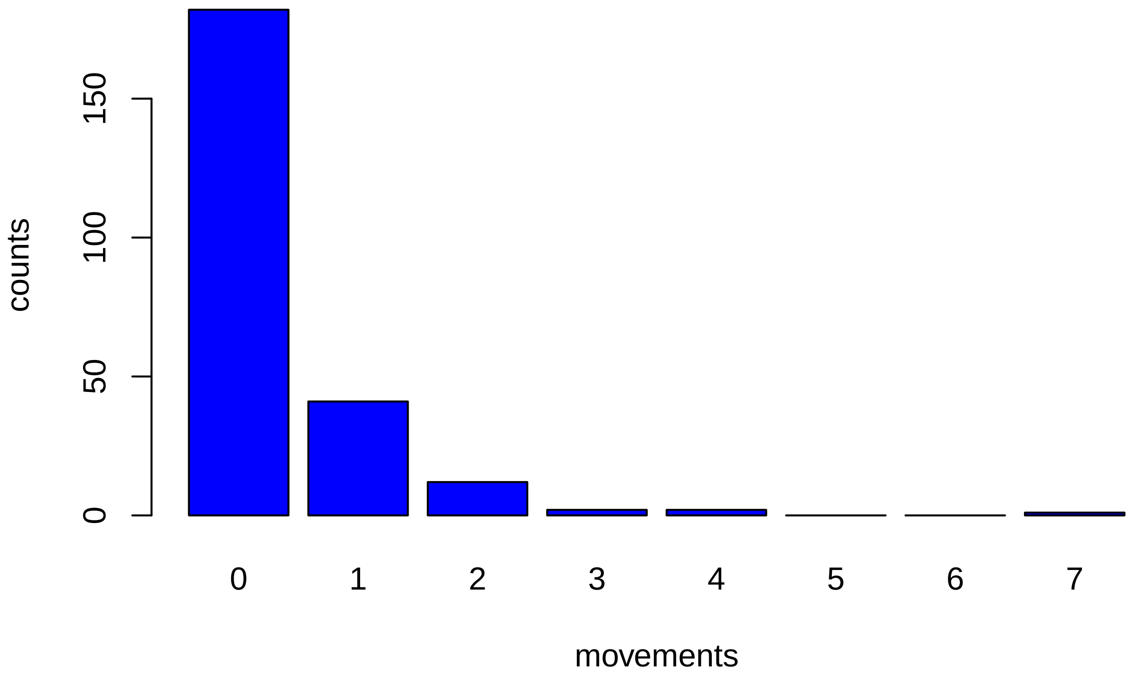
movements	counts
0	182
1	41
2	12
3	2
4	2
5	0

movements	counts
6	0
7	1

The above shows the fetal lamb movements and their frequencies. We can also see that the most of the fetal lambs (182) had zero movements.

Graphical Summaries

```r
## barplot and summary statistics
barplot(counts ~ movements, data = data, col = "blue")
```



From the barplot above we can see that the count for each movement decreases as the number of movements increases. The data is highly skewed to the right. In this case the mean will be greater than the median and the variance tends to be larger. We can also see that the most of the fetal lambs (182) had zero movements.

Checking for Variance - Mean Ratio

```r
# Calculate mean
mean <- sum(movements * counts) / sum(counts)
cat("The Mean  is ", mean, "\n")
```

```
The Mean  is  0.3583333
```

```
# Calculate variance
variance <- sum((movements - mean)^2 * counts) / (sum(counts)-1)
cat("The variance is ", variance, "\n")
```

```
The variance is  0.6576709
```

```
### variance mean ratio
var_mean_ratio <- variance / mean
cat("The variance-mean ratio  is ", var_mean_ratio , "\n")
```

```
The variance-mean ratio  is  1.835361
```

```
kable(cbind(mean,variance,var_mean_ratio))
```

mean	variance	var_mean_ratio
0.3583333	0.6576709	1.83536

Since we have each movement with the frequency we calculate the mean as

$$\bar{x} = \frac{\sum_{i=1}^{n}(x_i \cdot f_i)}{\sum_{i=1}^{n} f_i}$$

Where: $\bar{x}$: is the sample mean.

x_i : is each movement.

f_i : is the count in each movement.

n : is the total number of counts the movements.

then we caculate the variance also as ;

$$\sigma^2 = \frac{\sum_{i=1}^{n}(f_i \cdot (x_i - \bar{x})^2)}{\sum_{i=1}^{n} f_i - 1}$$

Where: σ^2: is the variance.

From the calculations above, we have that the mean is 0.3583333 , the variance is 0.6576709 and the variance - mean ratio is 1.835361. We can infer from these numerical summaries that the ideal distribution be a distribution with its variance greater than its mean and that the variance-mean ratio of that distribution should be greater than 1.

We Will now delve deeper into the probability distributions

FITTING DISTRIBUTIONS

In this part of our analysis, we will find the mean, variance and variance-mean ratio of some probability distributions to see if they fit our data. We will work on the Negative Binomial distribution , poisson distribution and the zero inflated poisson distribution.

Negative Binomial Distribution

$$P(X = k) = \binom{k + r - 1}{k} p^k (1 - p)^r$$

The formula for the expected value (mean) of a negative binomial distribution is given by:

$$E(X) = \frac{r \cdot (1 - p)}{p}$$

The formula for the variance of a negative binomial distribution is given by:

$$Var(X) = \frac{r(1 - p)}{p^2}$$

Now we will calculate the mean,variance and variance-mean ratio of the negative binomial distribution using the data given.

```
# Calculate mean of count data
mean_count <- sum(movements * counts) / sum(counts)

# Calculate variance of count data
var_count <- sum((movements - mean_count)^2 * counts) / (sum(counts) - 1)

# Estimate r (number of failures until the experiment is stopped) for negative binomial distribution
r_nb <- mean_count **2 / (var_count - mean_count)
cat("The shape parameter  is ", r_nb, "\n")
```

```
The shape parameter  is  0.4289565
```

```
p_nb<- r_nb/ (r_nb + mean_count)
cat("The probability   is ", p_nb, "\n")
```

```
The probability   is  0.5448521
```

```
## r <- (mean_count * p) / (1 - p) can also use this method
## pat<- mean_count/var_count

# Calculate mean of negative binomial distribution
mean_nb<- r_nb * (1 - p_nb) / p_nb
cat("The mean of the negative binomial distribution   is ", mean_nb, "\n")
```

```
The mean of the negative binomial distribution   is  0.3583333
```

```
# Calculate variance of negative binomial distribution
var_nb <- r_nb * (1 - p_nb) / p_nb**2
cat("The variance of the negative binomial distribution   is ", var_nb, "\n")
```

```
The variance of the negative binomial distribution   is  0.6576709
```

```
# Calculate variance-to-mean ratio
var_mean_ratio_nb <- var_nb / mean_nb
cat("The variance-mean ratio of the negative binomial distribution is ", var_mean_ratio_nb, "\n")
```

```
The variance-mean ratio of the negative binomial distribution is  1.835361
```

```
kable(cbind(mean_nb,var_nb,var_mean_ratio_nb))
```

mean_nb	var_nb	var_mean_ratio_nb
0.3583333	0.6576709	1.83536

We can see that the variance of the Negative Binomial distribution is greater than the mean of the Negative Binomial distribution. Also the variance- mean ratio of the Negative Binomial distribution is greater than one and equal to the variance- mean ratio of the actual dataset . This correspond to what we have in our original dataset, hence the Negative Binomial fits our data.

ESTIMATION

Now that we have concluded that our data follows a negative binomial distribution. We will like to estimate the parameters of our distribution and fit the probabilities using the estimated parameters. We will use the methods of of moments estimation and the maximum likelihood estimation in estimating our parameters.

We will start with the Method of Moment Estimation.

NEGATIVE BINOMIAL

Method Of Moment Estimation

The method of moments will be used to estimate the parameters r and p of a negative binomial distribution.

The moments estimate for $\hat{r}$ is :

$$\hat{r} = \frac{\bar{x}^2}{\mathrm{var}(x) - \bar{x}}$$

The moments estimate for $\hat{p}$ is:

$$\hat{p} = \frac{\bar{x}}{\mathrm{var}(x)}$$

```
x<-0:7;
freq<- c(182,41,12,2,2,0,0,1)
y= rep(x,freq)
n= length(y);n
```

```
[1] 240
```

```r
y.mean= mean(y); # sample mean
y.var= var(y) # sample variance
cat("The sample variance is ", y.var, "\n")
```

```
The sample variance is  0.6576709
```

```r
r_hat<- ((y.mean**2))/((y.var)-y.mean)
cat("The Method of Moment estimate (MME) of r ", r_hat, "\n")
```

```
The Method of Moment estimate (MME) of r  0.4289565
```

```r
p_hat<- y.mean/(y.var)
cat("The Method of Moment estimate (MME) of p ", p_hat, "\n")
```

```
The Method of Moment estimate (MME) of p  0.5448521
```

```r
# theta_hat.1<- r_hat/(y.mean + r_hat); theta_hat.1 ## same as theta hat

mu_hat<-y.mean; mu_hat
```

```
[1] 0.3583333
```

```r
cat("The sample mean is ", mu_hat, "\n")
```

```
The sample mean is  0.3583333
```

```r
kable(cbind(r_hat,p_hat,mu_hat))
```

r_hat	p_hat	mu_hat
0.4289565	0.5448521	0.3583333

Now that we have been able to calculate the Method of moments estimators for the parameters, we can find the fitted probabilities and frequencies

Estimated Expected frequencies (MME)

```r
## 1/31/2024 Estimated frequencies
px.1<- dnbinom(x,r_hat,p_hat); px.1
```

```
[1] 0.7706816076 0.1504668469 0.0489308090 0.0180315657 0.0070353869
[6] 0.0028364293 0.0011681258 0.0004882976
```

```r
y= rep(x,freq)
n = length(y) ##
nb = n*px.1  # Estimated expected frequencies
cat("The estimated expected frequencies is ", nb, "\n")
```

```
The estimated expected frequencies is  184.9636 36.11204 11.74339 4.327576 1.688493 0.680743 0.2803502
```

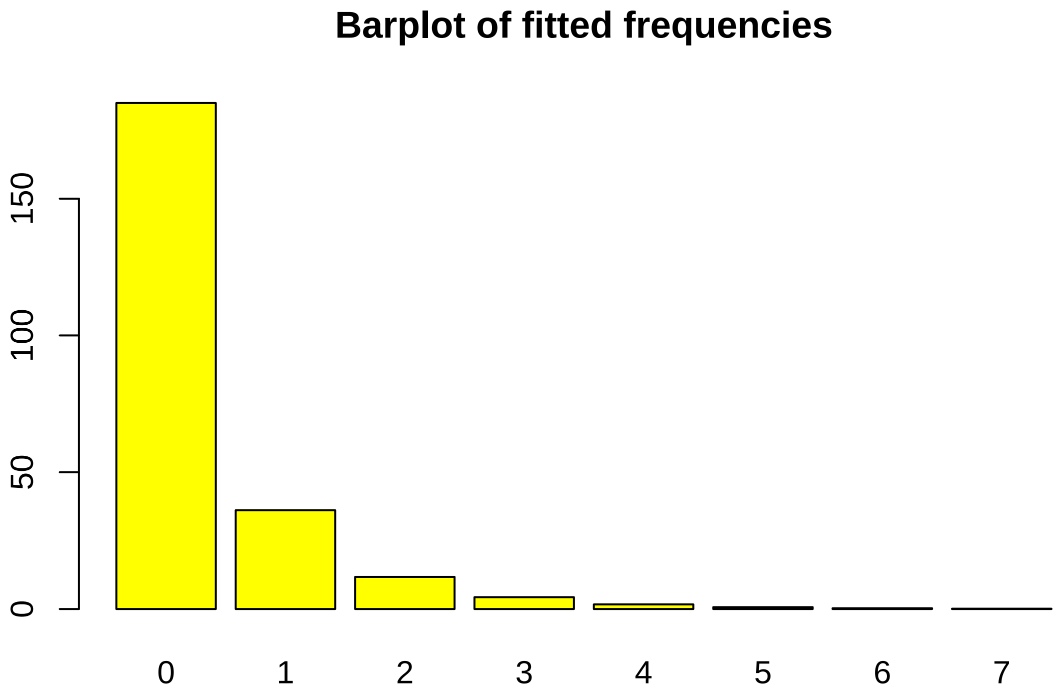
```
kable(cbind(x,freq,nb))## table with frequencies and estimated frequencies
```

x	freq	nb
0	182	184.9635858
1	41	36.1120433
2	12	11.7433941
3	2	4.3275758
4	2	1.6884929
5	0	0.6807430
6	0	0.2803502
7	1	0.1171914

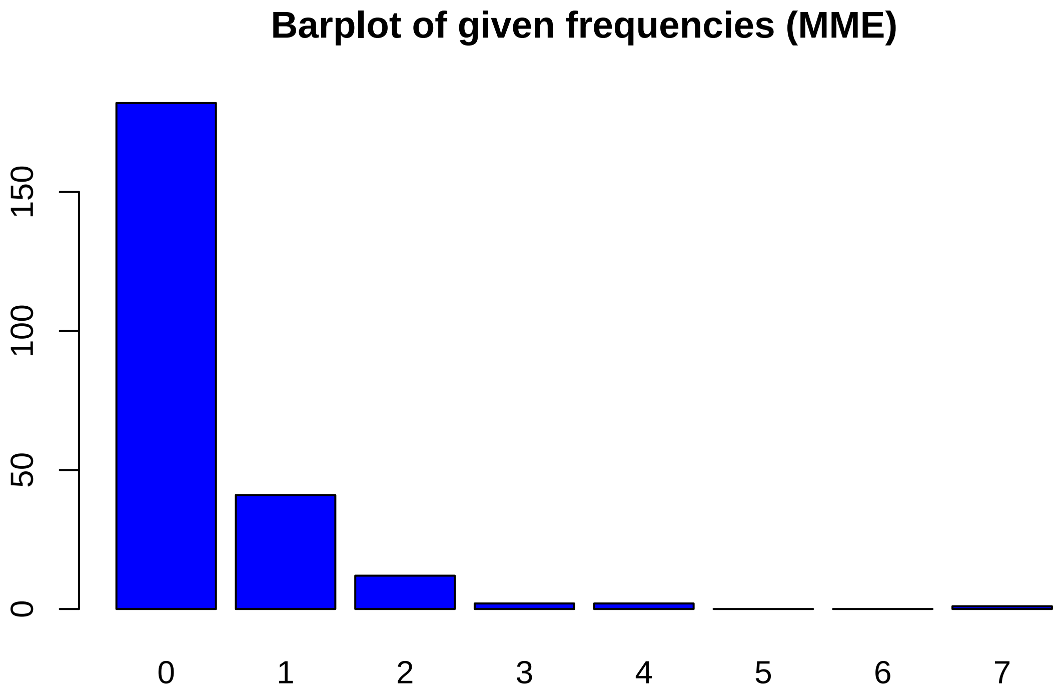
Now we compare the counts(actual frequencies from data) in each movements to the estimated counts using the negative binomial binomial distribution.

Barplot of Fitted Frequencies (MME) and Actual Frequencies

```
## fitted probabilities
x.0<-p_hat**r_hat
x.1<-choose(r_hat,1)*(p_hat**r_hat)*(1-p_hat)
x.2<-choose(r_hat+1,2)*(p_hat^r_hat)*((1-p_hat)^2)
x.3<-choose(r_hat+2,3)*(p_hat^r_hat)*((1-p_hat)^3)
x.4<-choose(r_hat+3,4)*(p_hat^r_hat)*((1-p_hat)^4)
x.5<-choose(r_hat+4,5)*(p_hat^r_hat)*((1-p_hat)^5)
x.6<-choose(r_hat+5,6)*(p_hat^r_hat)*((1-p_hat)^6)
x.7<-choose(r_hat+6,7)*(p_hat^r_hat)*((1-p_hat)^7)


fit.prob<-c(x.0,x.1,x.2,x.3,x.4,x.5,x.6,x.7)

## Fitted frequencies
fit.freq<- fit.prob*240
nw_fit.freq<-round(fit.freq, digits = 2)

## barplot of fitted frequencies
barplot(nw_fit.freq, names = x, col = "yellow", main = "Barplot of fitted frequencies")
```



```
## Barplot of given frequencies
barplot(freq, names = x, col = "blue", main = "Barplot of given frequencies (MME)")
```



The negative binomial seems to fit that data, however we will have to confirm this using the chi-squared goodness of fit test.

Chi-Squared goodness of fit test

The chi-squared goodness of fit test helps us to decide whether the estimated expected frequency is the same as the actual frequencies.

It can be calculated using :

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where: χ^2: is the chi-squared statistic.

O_i: is the observed frequency(counts of movements made).

E_i : is the expected frequency .

The degrees of freedom is calculated as $df = n - k - 1$, where n is the number of categories and k is the number of parameters estimated from the data.

Hypothesis

(H_0): The observed frequency distribution and the estimated expected frequency distribution are the same.

(H_1) : The observed frequency distribution and the expected frequency distribution are not equal.

```r
## 1/31/2024 Estimated frequencies
px.1<- dnbinom(x,r_hat,p_hat)
y= rep(x,freq)
n = length(y)
nb = n*px.1  # Estimated expected frequencies

k=sum(nb>5);
cat("The number  of categories in our data (observations greater 5 ) is ", k, "\n")
```

```
The number  of categories in our data (observations greater 5 ) is  3
```

```r
## chi-square value
x2<-sum(((freq-nb)**2/nb)[nb>5])## observations greater 5

cat("The chi-squared test statistics is ", x2, "\n")
```

```
The chi-squared test statistics is  0.7147022
```

```r
## p-value
p_value<-1-pchisq(0.7147022,0)
cat("The p-value is ", p_value, "\n")
```

```
The p-value is  0
```

The number of observations greater than five in our data is three , calculating the degrees of freedom gives us $df = 3 - 2 - 1 = 0$. And this does not make sense since the degrees of freedom cannot be zero.

We will try to add all observations less than 5 and treat as one observation and check the results of the chi squared also.

Adding all observations less than 5

```r
# Sum observations less than 5
sum_less_than_5 <- sum(nb[nb < 5])
k_bad=nb[nb>5];
freq
```

```
[1] 182  41  12   2   2   0   0   1
```

```r
new_expected_nb<- c(k_bad,sum_less_than_5)
x_new<-0:3
new_freq1<-sum(freq[freq < 5])
freq_bad=freq[freq>5];
new_freq<- c(freq_bad,new_freq1)
kable(cbind(x_new,new_freq,new_expected_nb))
```

x_new	new_freq	new_expected_nb
0	182	184.963586
1	41	36.112043
2	12	11.743394
3	5	7.094353

```r
k_new=sum(new_expected_nb>5);
cat("The number  of categories in our new data (observations greater 5 with sum of observations less tha
```

```
The number  of categories in our new data (observations greater 5 with sum of observations less than 5
```

The number of observations greater than five in our new data is four, calculating the degrees of freedom gives us $df = 4 - 2 - 1 = 1$. We will then use this degrees of freedom to calculate our chi-squared test statistic again.

Chi-Sqaured test statistic with New data

```r
## chi-square value
x2_new<-sum(((new_freq-new_expected_nb)**2/new_expected_nb))

cat("The chi-squared test statistics is ", x2_new, "\n")
```

```
The chi-squared test statistics is  1.332985
```

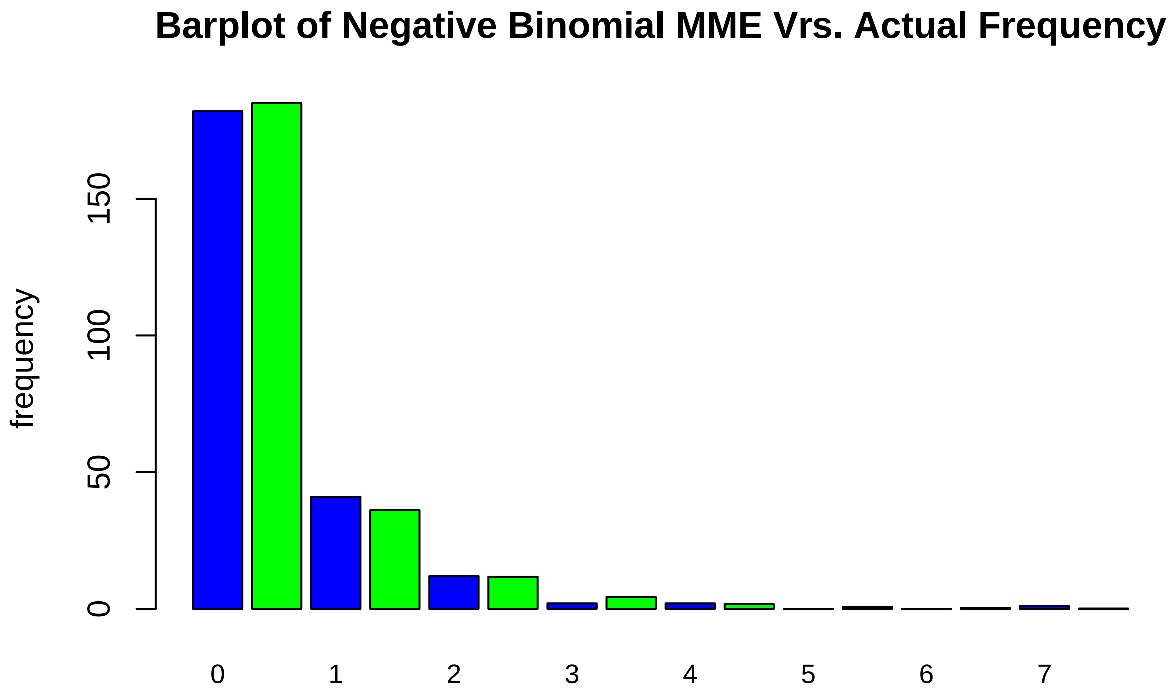
```r
## p-value
p_value<-1-pchisq(x2_new,1)
cat("The p-value is ", p_value, "\n")
```

```
The p-value is  0.2482749
```

The P-value is now 0.2482749. Since this p-value is greater than 0.05, we fail to reject our null hypothesis and conclude that the observed frequency distribution and the estimated expected frequency distribution are the same

combined barplots (actual vrs expected)

```r
Count = 1:16
Count[1:16%% 2!= 0]= freq
Count[1:16%%2 ==0]= nb
apor<-rep(0:7,2)
hawa<-rep(0:7,each=2)
X= as.character(rep(0:7, each = 2))

X[1:14%%2==0]= " "

barplot(Count, names = X, col= rep(c("blue","green"),8),ylab= "frequency", cex.names = 0.8, main = "Bar
```



Maximum Likelihood Estimation (MLE)

```
## 2/12/2024 ## MLE
hap= rep(x,freq)
fitdistr(hap, "negative binomial")$estimate
```

```
     size        mu
0.5289179 0.3583295
```

```
rhat<- as.numeric(fitdistr(hap, "negative binomial")$estimate[1])
rhat120<-rhat
cat("The maximum likelihood estimate for r is  ", rhat120, "\n")
```

```
The maximum likelihood estimate for r is   0.5289179
```

```
mu_hat<- as.numeric(fitdistr(hap, "negative binomial")$estimate[2])

p_hat2<- rhat/(rhat + mu_hat)
cat("The maximum likelihood estimate for p is  ", p_hat2, "\n")
```

```
The maximum likelihood estimate for p is   0.5961335
```

```
mu_hat2<- mean(hap)

kable(cbind(rhat,mu_hat2,p_hat2))
```

rhat	mu_hat2	p_hat2
0.5289179	0.3583333	0.5961335

Estimated probabilites and Estimated frequencies

```
n2=240
px2= dnbinom(x,rhat,p_hat2);px2
```

```
[1] 0.7606330139 0.1624805080 0.0501641272 0.0170782968 0.0060850471
[6] 0.0022260054 0.0008284245 0.0003120570
```

```
nb2 = n2*px2  # Estimated expected frequencies

kable(cbind(x,freq,px2,nb2))
```

x	freq	px2	nb2
0	182	0.7606330	182.5519233
1	41	0.1624805	38.9953219
2	12	0.0501641	12.0393905
3	2	0.0170783	4.0987912
4	2	0.0060850	1.4604113
5	0	0.0022260	0.5342413
6	0	0.0008284	0.1988219
7	1	0.0003121	0.0748937

Now we compare the counts(actual frequencies from data) in each movements to the estimated counts using the negative binomial binomial distribution.

Chi- Squared Goodness of Fit test

Hypothesis

(H_0): The observed frequency distribution and the expected frequency distribution are the same.

(H_1) : There is a difference between the observed frequency distribution and the expected frequency distribution.

```
## chi square
k2=sum(nb2>5);k2 ## counts where nb is greater is 5
```

```
[1] 3
```

```
cat("The number  of categories in our data (observations greater 5 ) is ", k2, "\n")
```

```
The number  of categories in our data (observations greater 5 ) is  3
```

```
df2= 3-2-1

## chi-square value
x3<-sum(((freq-nb2)**2/nb2)[nb2>5]);## observations greater 5
cat("The chi-squared test statistics is ", x3, "\n")
```

```
The chi-squared test statistics is  0.1048544
```

```
## p-value
p_value2<-1-pchisq(0.1048544,0)
cat("The p-value is ", p_value2, "\n")
```

```
The p-value is  0
```

Again, The number of the estimated expected observations greater than five in our data is three , calculating the degrees of freedom gives us $df = 3-2-1 = 0$. And this does not make sense since the degrees of freedom cannot be zero.

We will try to add all observations less than 5 and treat as one observation and check the results of the chi squared also.

Adding all observations less than 5

```
# Sum observations less than 5
sum_less_than_5_MLE <- sum(nb2[nb2 < 5])
k_bad2=nb2[nb2>5];
new_expected_nb2_MLE<- c(k_bad2,sum_less_than_5_MLE)
x_new_MLE<-0:3
new_freq1_MLE<-sum(freq[freq < 5])
freq_bad_MLE=freq[freq>5];
new_freq_MLE<- c(freq_bad_MLE,new_freq1_MLE)
kable(cbind(x_new_MLE,new_freq_MLE,new_expected_nb2_MLE))
```

x_new_MLE	new_freq_MLE	new_expected_nb2_MLE
0	182	182.551923
1	41	38.995322
2	12	12.039390
3	5	6.367159

```
k_new_MLE=sum(new_expected_nb2_MLE>5);
cat("The number  of categories in our new data (observations greater 5 with sum of observations less tha
```

```
The number  of categories in our new data (observations greater 5 with sum of observations less than 5
```

The number of observations greater than five in our new data is four, calculating the degrees of freedom gives us $df = 4-2-1 = 1$. We will then use this degrees of freedom to calculate our chi-squared test statistic again.

Chi-Sqaured test statistic with New data

```
## chi-square value
x2_new_MLE<-sum(((new_freq_MLE-new_expected_nb2_MLE)**2/new_expected_nb2_MLE))

cat("The chi-squared test statistics is ", x2_new_MLE, "\n")
```

```
The chi-squared test statistics is  0.3984115
```

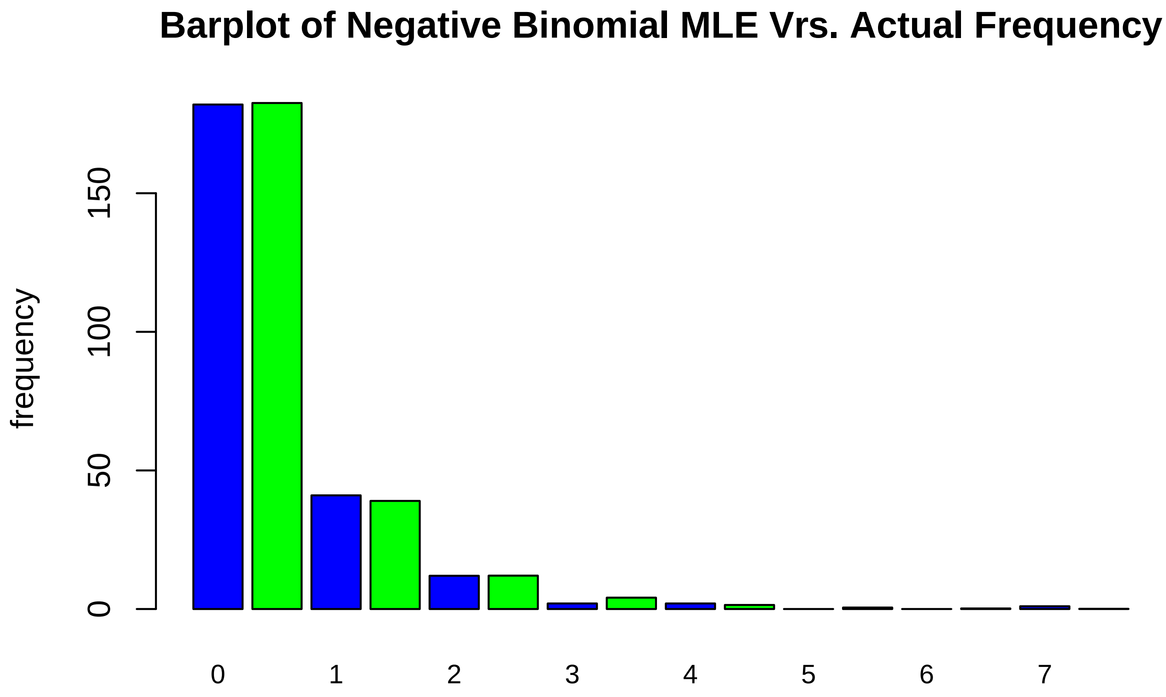
```
## p-value
p_value_MLE<-1-pchisq(x2_new_MLE,1)
cat("The p-value is ", p_value_MLE, "\n")
```

```
The p-value is  0.5279108
```

The P-value is now 0.5279108. Since this p-value is greater than 0.05, we fail to reject our null hypothesis and conclude that the observed frequency distribution and the estimated expected frequency distribution are the same

combined barplots (actual vrs expected)

```
Count = 1:16
Count[1:16%% 2!= 0]= freq
Count[1:16%%2 ==0]= nb2
apor<-rep(0:7,2)
hawa<-rep(0:7,each=2)
X= as.character(rep(0:7, each = 2))

X[1:14%%2==0]= " "

barplot(Count, names = X, col= rep(c("blue","green"),8),ylab= "frequency", cex.names = 0.8, main = "Barp
```



POISSON DISTRIBUTION

The probability mass function (PMF) of a Poisson distribution is given by:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Expectation (Mean)

The formula for the expected value (mean) of a Poisson distribution is given as:

$$E(X) = \lambda$$

Variance

The formula for the variance of a Poisson distribution is also given as:

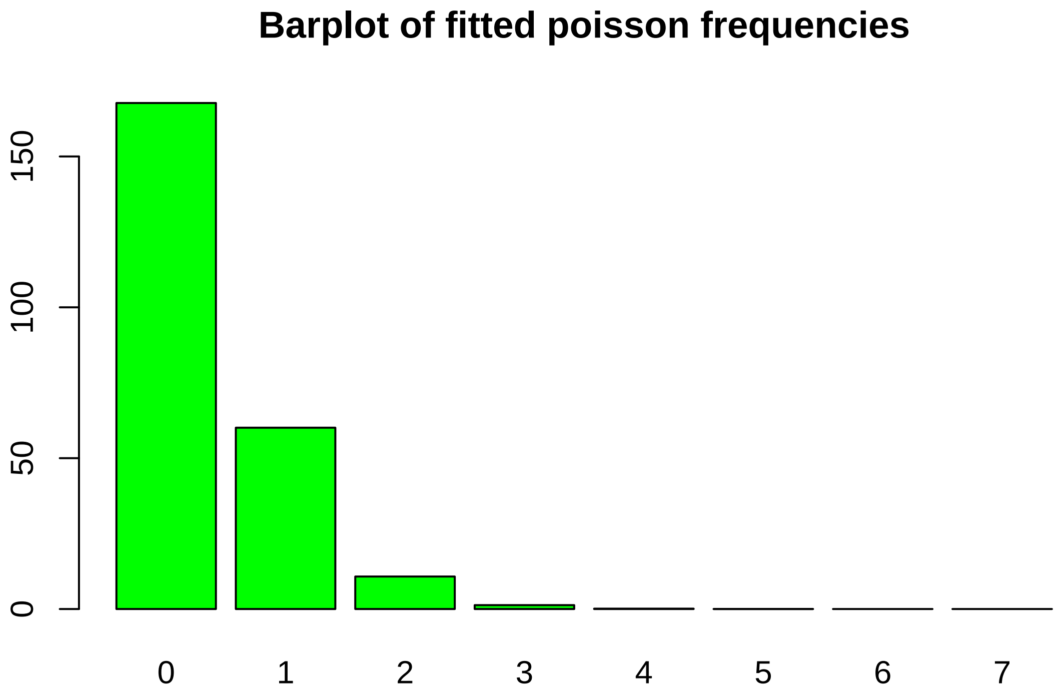
$$Var(X) = \lambda$$

Maximum Likelihood Estimation

```
## 2/12/2024 ## MLE
x<-0:7;
freq<- c(182,41,12,2,2,0,0,1)
n = 240

hap<-rep(movements,counts)
fit2<-fitdistr(hap,"poisson")
lambda.1<-fit2$estimate;
poisson_samples<-dpois(x,lambda= lambda.1)

poisson = n*poisson_samples
poisson_new= round(poisson, digits = 2)

kable(cbind(x,freq,poisson_new))
```

x	freq	poisson_new
0	182	167.72
1	41	60.10
2	12	10.77
3	2	1.29
4	2	0.12
5	0	0.01
6	0	0.00
7	1	0.00

```
barplot(poisson_new, names = x, col = "green", main = "Barplot of fitted poisson frequencies")
```



Chi- Squared Goodness of Fit test

Hypothesis

(H_0): The observed frequency distribution and the expected frequency distribution are the same.

(H_1) : There is a difference between the observed frequency distribution and the expected frequency distribution.

```
## chi square
k9=sum(poisson_new>5) ## counts where nb is greater is 5
cat("The number  of categories in our data (observations greater 5 ) is ", k9, "\n")
```

```
The number  of categories in our data (observations greater 5 ) is  3
```

```
df2= 3-2-1

## chi-square value
x.90<-sum(((freq-poisson_new)**2/poisson_new)[poisson_new>5]);## observations greater 5
cat("The chi-squared test statistics is ", x.90, "\n")
```

```
The chi-squared test statistics is  7.42635
```

```r
## p-value
p_value2<-1-pchisq(7.42635,0)
cat("The p-value is ", p_value2, "\n")
```

```
The p-value is  0
```

Again, The number of the estimated expected observations greater than five in our data is three , calculating the degrees of freedom gives us $df = 3 - 2 - 1 = 0$. And this does not make sense since the degrees of freedom cannot be zero.

We will try to add all observations less than 5 and treat as one observation and check the results of the chi squared also.

Adding all observations less than 5

```r
# Sum observations less than 5
sum_less_than_5_pp <- sum(n[n < 5])
k_badd12=poisson_new[poisson_new>5];
new_expected_nb2_pp<- c(k_badd12,sum_less_than_5_pp)
x_new_pp<-0:3
new_freq1_pp<-sum(freq[freq < 5])
freq_bad_pp=freq[freq>5];
new_freq_pp<- c(freq_bad_pp,new_freq1_pp)
kable(cbind(x_new_pp,new_freq_pp,new_expected_nb2_pp))
```

x_new_pp	new_freq_pp	new_expected_nb2_pp
0	182	167.72
1	41	60.10
2	12	10.77
3	5	0.00

```r
k_new_pp=4;
```

The number of observations in our new data is four, calculating the degrees of freedom gives us $df = 4 - 1 - 1 = 2$. We will then use this degrees of freedom to calculate our chi-squared test statistic again.

Chi-Sqaured test statistic with New data

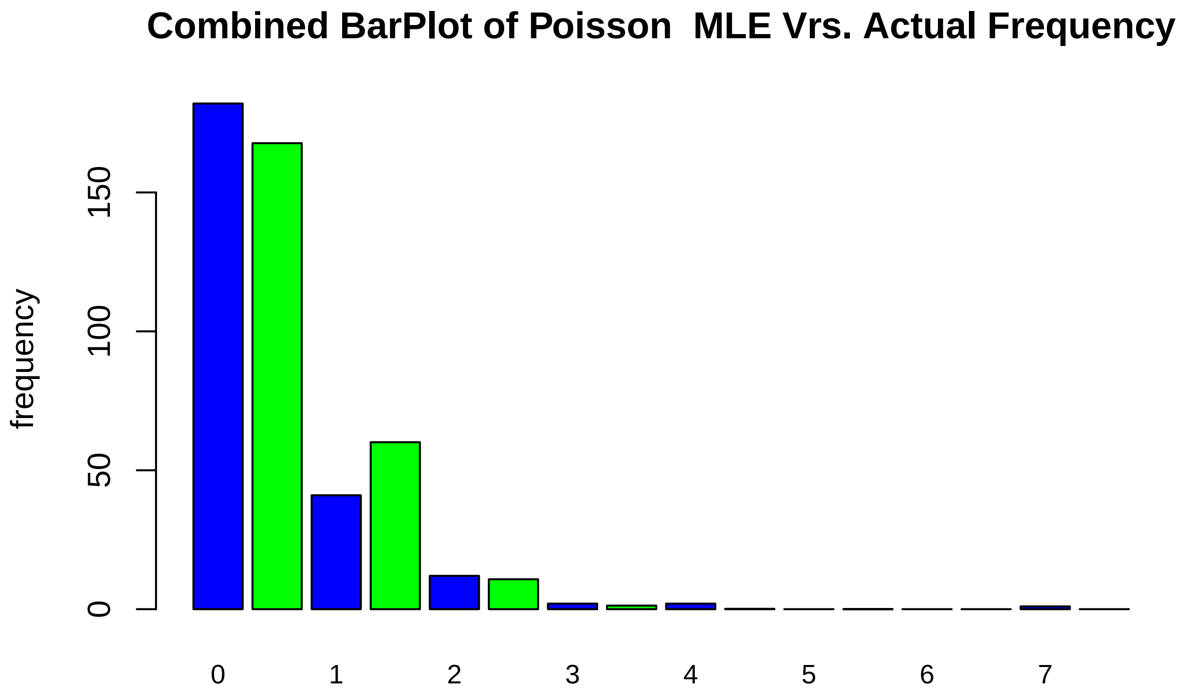
```r
## chi-square value

poisson_new5<- c(167.72,60.10,10.77,1.42)
x2_new_pp<-sum(((new_freq_pp-poisson_new5)**2/poisson_new5))

cat("The chi-squared test statistics is ", x2_new_pp, "\n")
```

```
The chi-squared test statistics is  16.45198
```

```
## p-value
p_value_new_pp<-1-pchisq(x2_new_pp,2)
cat("The p-value is ", p_value_new_pp, "\n")
```

```
The p-value is  0.0002676068
```

```
Count = 1:16
Count[1:16%% 2!= 0]= freq
Count[1:16%%2 ==0]= poisson_new
apor<-rep(0:7,2)
hawa<-rep(0:7,each=2)
X= as.character(rep(0:7, each = 2))

X[1:14%%2==0]= " "

barplot(Count, names = X, col= rep(c("blue","green"),8),ylab= "frequency", cex.names = 0.8, main = "Com
```



The P-value is now 0.0002676068 Since this p-value is less than 0.05, we reject our null hypothesis and conclude that the observed frequency distribution and the estimated expected frequency distribution are the different . That is the poisson distribution does not fit our data perfectly.

ZERO INFLATED POISSON DISTRIBUTION

The probability mass function (PMF) of the zero-inflated Poisson distribution is given by:

$$P(X = k) = \begin{cases} (1-\pi) \cdot e^{-\lambda} \cdot \frac{\lambda^k}{k!} & \text{if } k > 0 \\ \pi + (1-\pi) \cdot e^{-\lambda} & \text{if } k = 0 \end{cases}$$

Where: - X is the random variable following a zero-inflated Poisson distribution. - k is a non-negative integer representing the count value. - π is the probability of excess zeros. - λ is the mean parameter of the Poisson distribution.

Zero Inflated Poisson 4/3/2024

We write R function to find the MLE of the estimators

```
# Example count data with frequencies (replace this with your own data)

x<-0:7;
movements<-c(0,1,2,3,4,5,6,7)
freq<- c(182,41,12,2,2,0,0,1)
y= rep(x,freq)

## observations containing zeros
n0=sum(y==0)

### calculating the probability function for zip
dzip = function(y,theta,p)
  {
  Z= NULL
for (i in 1:length(y))
{
  if (y[i]==0) Z[i]= p + (1-p)*dpois(0,theta)
  else
    Z[i]= (1-p)*dpois(y[i],theta)
}
  Z
 }
dzip(0,0.846,0.566)
```

```
[1] 0.7522416
```

```
### 08/04/2024
ell = function(eta, y)
{
  -sum(log(dzip(y,eta[1],eta[2])))
}

ell(c(0.358,0), y=y)
```

```
[1] 201.0437
```

```
### calculate MLE
theta0 = mean(y)
n0= 180
```

```
n= 240
p0= n0/n

ell(c(theta0,p0) , y=y)
```

```
[1] 230.4781
```

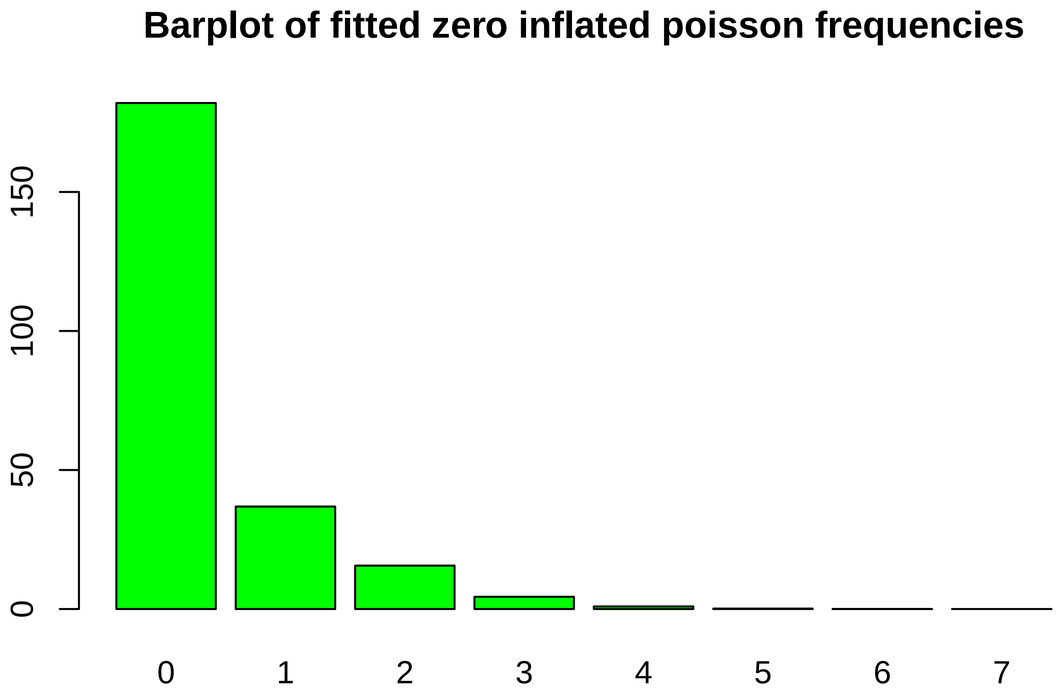
```
## maximum
mle=nlm(ell, c(theta0,p0), y=y)$estimate

cat("The maximum likelihood estimate for theta and p are  ", mle, "\n")
```

```
The maximum likelihood estimate for theta and p are   0.847278 0.577077
```

```
## estimation
zip_samples<-n*dzip(movements,mle[1], mle[2])
new.zip_samples<-round(zip_samples)


kable(cbind(x,freq,new.zip_samples))
```

x	freq	new.zip_samples
0	182	182
1	41	37
2	12	16
3	2	4
4	2	1
5	0	0
6	0	0
7	1	0

```
barplot(zip_samples, names = x, col = "green", main = "Barplot of fitted zero inflated poisson frequenc
```



Chi- Squared Goodness of Fit test

Hypothesis

(H_0): The observed frequency distribution and the expected frequency distribution are the same.

(H_1) : There is a difference between the observed frequency distribution and the expected frequency distribution.

```
## chi square
k10=sum(zip_samples>5);k10 ## counts where nb is greater is 5
```

```
[1] 3
```

```
cat("The number  of categories in our data (observations greater 5 ) is ", k10, "\n")
```

```
The number  of categories in our data (observations greater 5 ) is  3
```

```
df10= 3-2-1

## chi-square value
x10<-sum(((freq-zip_samples)**2/zip_samples)[zip_samples>5]);## observations greater 5
cat("The chi-squared test statistics is ", x10, "\n")
```

```
The chi-squared test statistics is  1.302166
```

```r
## p-value
p_value2<-1-pchisq(x10,0)
cat("The p-value is ", p_value2, "\n")
```

```
The p-value is  0
```

Again, The number of the estimated expected observations greater than five in our data is three , caculating the degrees of freedom gives us $df = 3-2-1 = 0$. And this does not make sense since the degrees of freedom cannot be zero.

We will try to add all observations less than 5 and treat as one observation and check the results of the chi squared also.

Adding all observations less than 5

```r
# Sum observations less than 5
sum_less_than_5_zip <- sum(zip_samples[zip_samples < 5])
k_bad20=zip_samples[zip_samples>5];
new_expected_nb2_zip<- c(k_bad20,sum_less_than_5_zip)
x_new_zip<-0:3
new_freq1_zip<-sum(freq[freq < 5])
freq_bad_zip=freq[freq>5];
new_freq_zip<- c(freq_bad_zip,new_freq1_zip)
kable(cbind(x_new_zip,new_freq_zip,new_expected_nb2_zip))
```

x_new_zip	new_freq_zip	new_expected_nb2_zip
0	182	181.999999
1	41	36.857874
2	12	15.614434
3	5	5.527378

```r
k_new_MLE_zip=sum(new_expected_nb2_zip>5);
cat("The number  of categories in our new data (observations greater 5 with sum of observations less tha
```

```
The number  of categories in our new data (observations greater 5 with sum of observations less than 5
```

The number of observations in our new data is four, calculating the degrees of freedom gives us $df = 4-2-1 = 1$. We will then use this degrees of freedom to calculate our chi-squared test statistic again.

Chi-Sqaured test statistic with New data

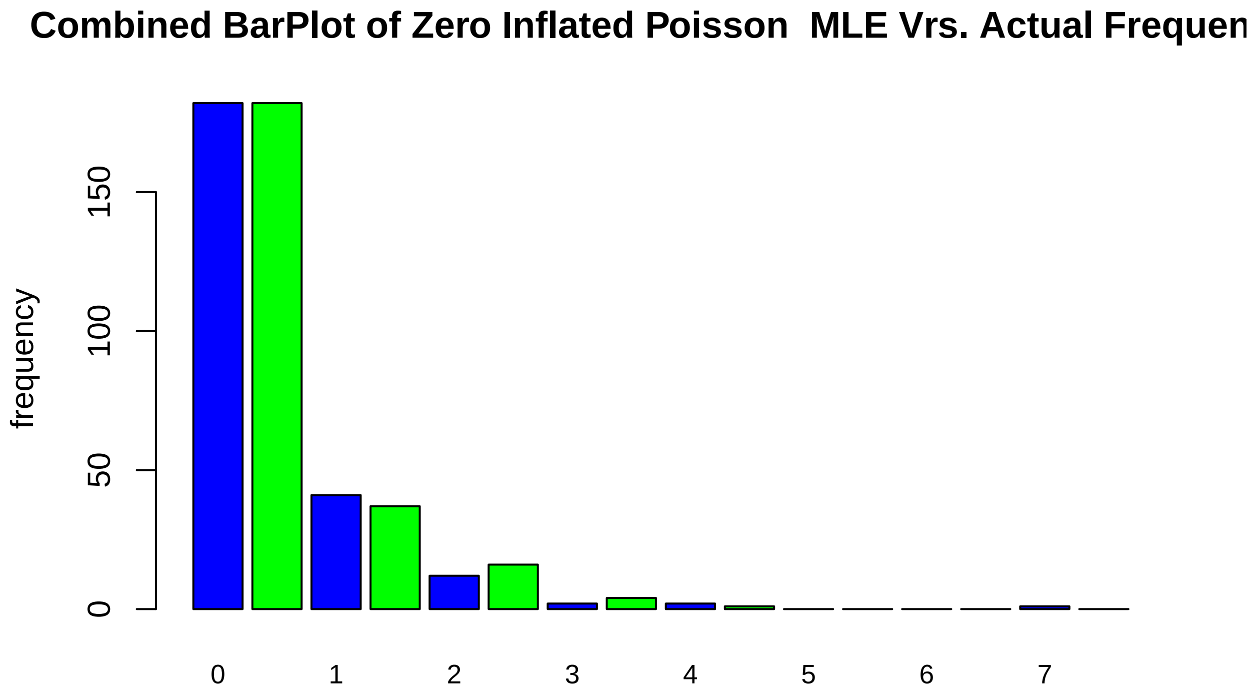
```r
## chi-square value
x2_new_zip<-sum(((new_freq_zip-new_expected_nb2_zip)**2/new_expected_nb2_zip))

cat("The chi-squared test statistics is ", x2_new_zip, "\n")
```

```
The chi-squared test statistics is  1.352485
```

```
## p-value
p_value_MLE<-1-pchisq(x2_new_zip,1)
cat("The p-value is ", p_value_MLE, "\n")
```

```
The p-value is  0.2448442
```

```
Count = 1:16
Count[1:16%% 2!= 0]= freq
Count[1:16%%2 ==0]= new.zip_samples
apor<-rep(0:7,2)
hawa<-rep(0:7,each=2)
X= as.character(rep(0:7, each = 2))

X[1:14%%2==0]= " "

barplot(Count, names = X, col= rep(c("blue","green"),8),ylab= "frequency", cex.names = 0.8, main = "Comb
```



The P-value is now 0.2484889 Since this p-value is greater than 0.05, we fail to reject our null hypothesis and conclude that the observed frequency distribution and the estimated expected frequency distribution are the same

CONCLUSION

We can see that the zero inflated poisson distribution fits our data well since we have many zeros in our data .