Data Analysis for the Life Sciences
Chapter 7: Statistical Models

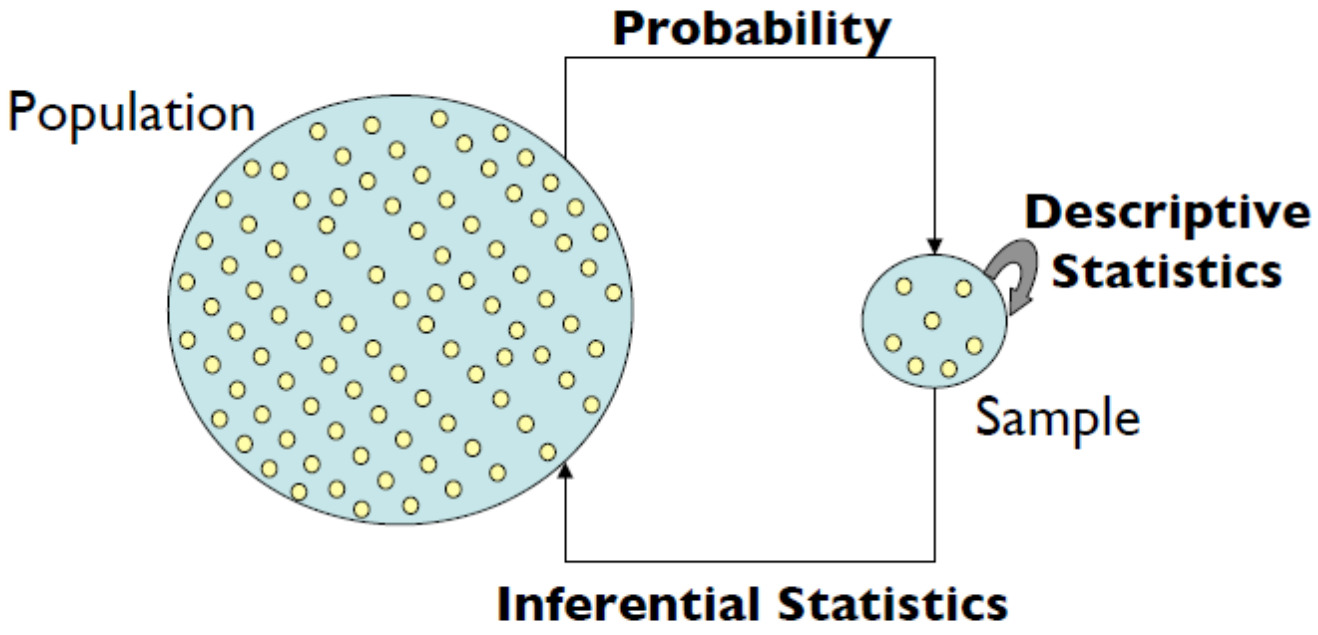
Jesica S. Rodriguez-Lopez

CUNY The City University of New York

BROAD INSTITUTE

Statistical Models: Estimation

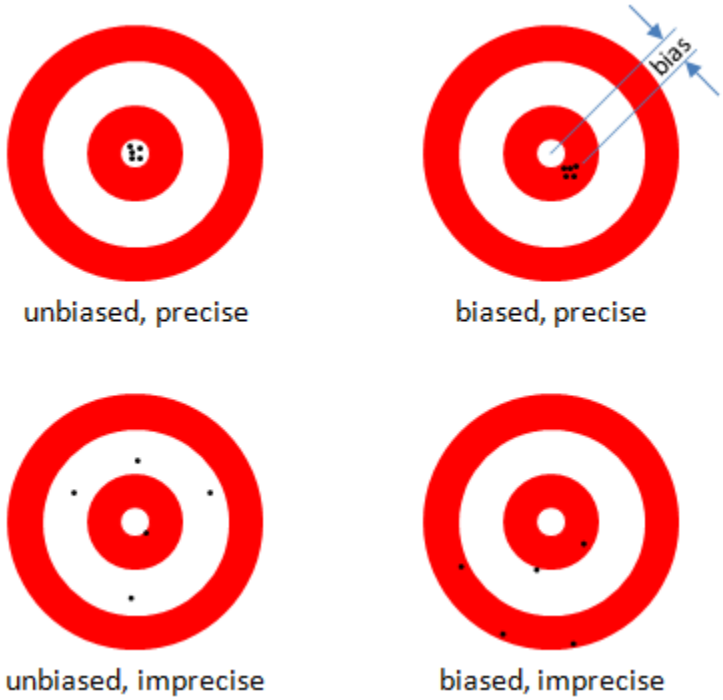
- Purpose of statistics:



Source: http://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture5.pdf

Statistical Models: Estimation

- Properties a good estimator:
  - Unbiased: Expected value of the estimator matches the real value of parameter
  - Precise: Small standard error
  - Consistent: As the sample size increases, the estimator gets closer to the true value of the parameter



Source: http://www.statisticalengineering.com/Weibull/precision-bias.html

Statistical Models: Estimation

- Methods:
  - Least Square Error: Already reviewed
    - Minimize

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

  - Maximum likelihood: Today's topic
  - Bayesian: Next topic

Maximum Likelihood

- Before an experiment is performed the outcome is **unknown**. Probability allows us to predict **unknown** outcomes based on **known** parameters:

$$Pr(Data|\theta)$$

- In the case of the binomial distribution:

$$Pr(S = k|N, p) = \binom{N}{k} p^k (1-p)^{N-k}$$

Source: http://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture5.pdf

Maximum Likelihood

- After an experiment is performed the outcome is ***known***. Now we talk about the ***likelihood*** that a parameter would generate the observed data:

$$L(\theta|Data) = Pr(Data|\theta)$$

- In the case of the binomial distribution:

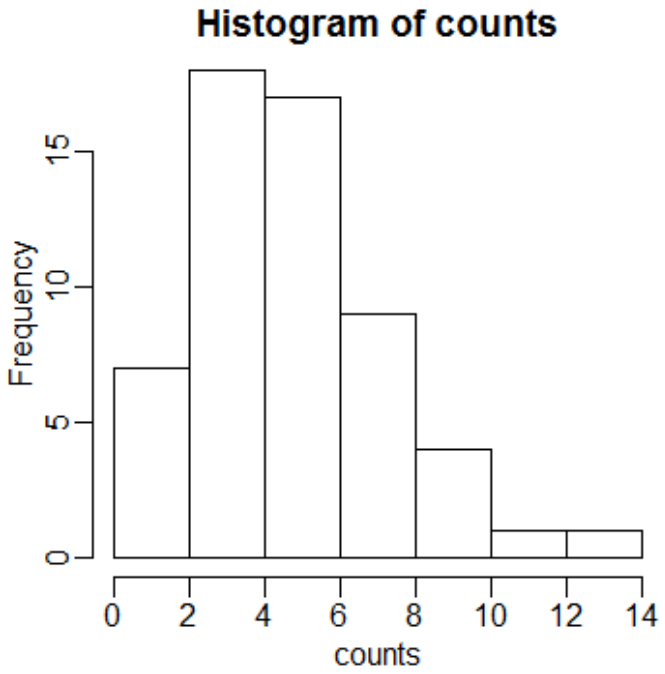
$$L(p|N,k) = \binom{N}{k} p^k (1-p)^{N-k}$$

- Estimation proceeds by finding the value of θ that makes the observed data most likely

Source: http://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture5.pdf

Maximum Likelihood

- Example: palindrome locations in the Human cytomegalovirus genome (HMCV) genome
  - Count the number of palindromes in each 4,000 basepair segments.



Palindrome count histogram

Maximum Likelihood

- Example: palindrome locations in the HMCV genome
  - Follows a Poisson distribution, with parameter λ :

$$Pr(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{k_i} exp(-\lambda)}{k_i!}$$

  - What is λ: The MLE is the value of λ that maximizes the likelihood:

$$L(\lambda|X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n) = exp\left(\sum_{i=1}^{n} ln\big(Pr(X_i = k_i|\lambda)\big)\right)$$

  - In practice, it is more convenient to maximize the log-likelihood which is the summation that is exponentiated in the expression above.

Maximum Likelihood

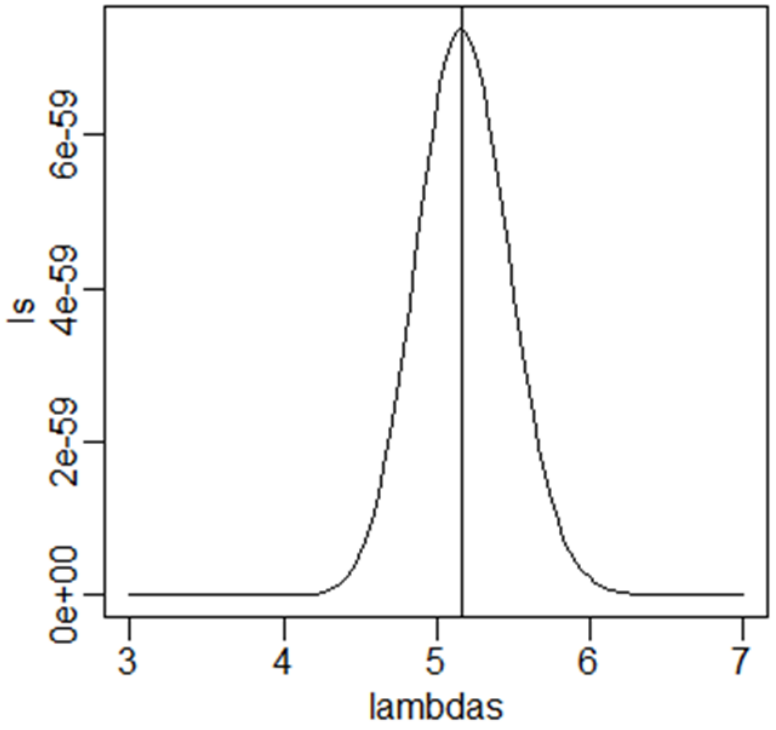
- Example: palindrome locations in the HMCV genome:

```
l<-function(lambda) sum(dpois(counts,lambda,log=TRUE))

lambdas<-seq(3,7,len=100)
ls <- exp(sapply(lambdas,l))

plot(lambdas,ls,type="l")

mle=optimize(l,c(0,10),maximum=TRUE)
abline(v=mle$maximum)
```

Maximum Likelihood

- Example: palindrome locations in the HMCV genome:
- If you work out the math and do a bit of calculus, you realize that this is a particularly simple example for which the MLE is the average:

```
print( c(mle$maximum, mean(counts) ) )
[1] 5.157894 5.157895
```



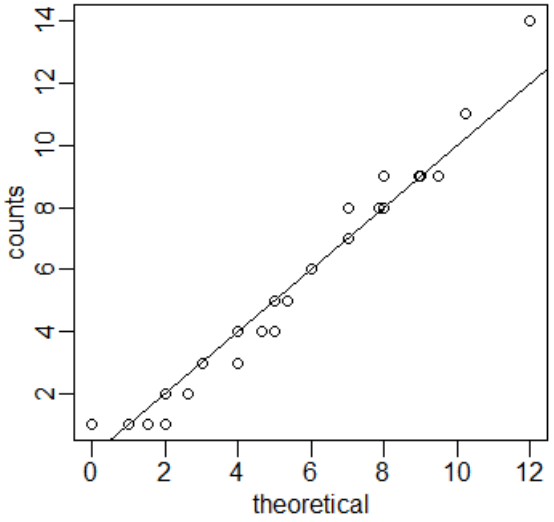
Likelihood versus lambda.

Maximum Likelihood

- Example: palindrome locations in the HMCV genome:
  - Note that a plot of observed counts versus counts predicted by the Poisson shows that the fit is quite good in this case:

```
theoretical<-
qpois((seq(0,99)+0.5)/100,mean(counts))
qqplot(theoretical,counts)
abline(0,1)
```

We therefore can model the palindrome count data with a Poisson with λ=5.16.

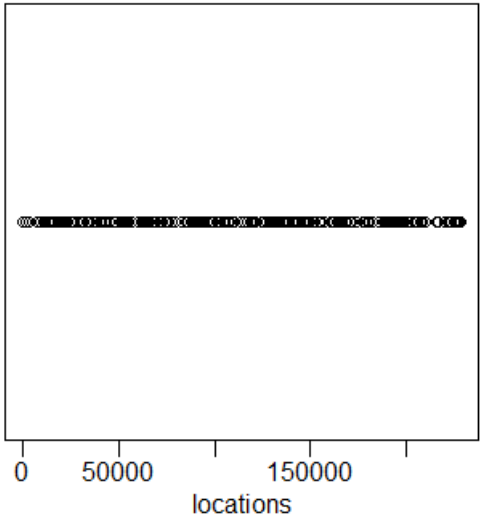


Observed counts versus theoretical Poisson counts.

Maximum Likelihood - Exercises

10. Now we are going to explore if palindromes are over-represented in some part of the HCMV genome. Make sure you have the latest version of the dagdata, load the palindrome data from the Human cytomegalovirus genome, and plot locations of palindromes on the genome for this virus:

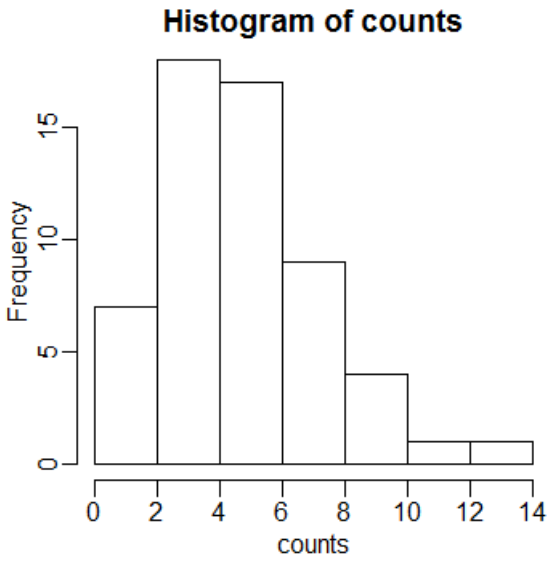
```
library(devtools)
install_github("genomicsclass/dagdata")
library(dagdata)
data(hcmv)
plot(locations,rep(1,length(locations)),ylab="",yaxt="n")
```



Maximum Likelihood - Exercises

10. (Cont.) These palindromes are quite rare, and therefore p is very small. If we break the genome into bins of 4000 basepairs, then we have Np not so small and we might be able to use Poisson to model the number of palindromes in each bin:

```
breaks=seq(0,4000*round(max(locations)/4000),4000)
tmp=cut(locations,breaks)
counts=as.numeric(table(tmp))
```



So if our model is correct, counts should follow a Poisson distribution. The distribution seems about right:

```
hist(counts)
```

Maximum Likelihood - Exercises

10. (Cont.) So let $X_1, X_2, \ldots, X_n$ be the random variables representing counts then

$$Pr(X_i = k|\lambda) = \frac{\lambda^k exp(-\lambda)}{k!}$$

and to fully describe this distribution, we need to know λ. For this we will use MLE. We can write the likelihood described in book in R. For example, for λ=4 we have:

```
probs <- dpois(counts,4)
likelihood <- prod(probs)
Likelihood
[1] 1.177527e-62
```

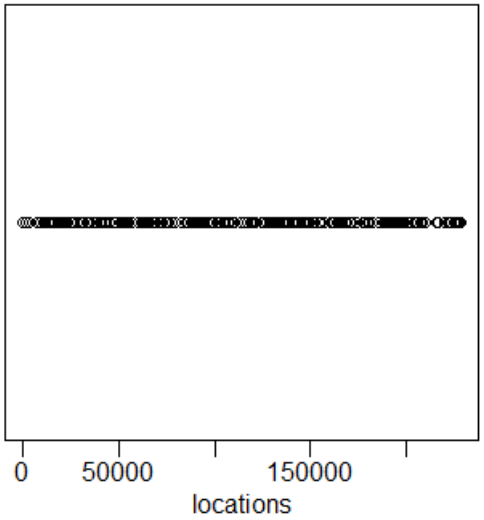
Notice that it's a tiny number. It is usually more convenient to compute log-likelihoods:

```
logprobs <- dpois(counts,4,log=TRUE)
loglikelihood <- sum(logprobs)
loglikelihood
[1] -142.5969
```

Maximum Likelihood - Exercises

10. (Cont.) Now write a function that takes λ and the vector of counts as input and returns the log-likelihood. Compute this log-likelihood for lambdas = seq(0,15,len=300) and make a plot. What value of lambdas maximizes the log-likelihood?

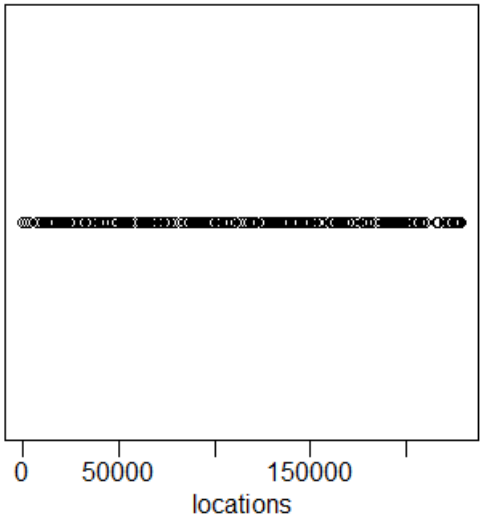
```
library(dagdata)
data(hcmv)

library(rafalib)
mypar()
plot(locations,rep(1,length(locations)),ylab="",yaxt="n")
```



Maximum Likelihood - Exercises

10. (Cont.) Now write a function that takes λ and the vector of counts as input and returns the log-likelihood. Compute this log-likelihood for lambdas = seq(0,15,len=300) and make a plot. What value of lambdas maximizes the log-likelihood?

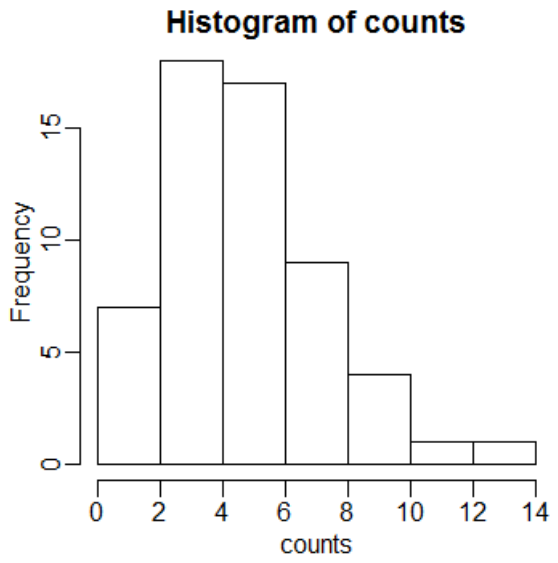
```
##Let's look at the data
library(dagdata)
data(hcmv)

library(rafalib)
mypar()
plot(locations,rep(1,length(locations)),ylab="",yaxt="n")
```



Maximum Likelihood - Exercises

10. (Cont.) Now write a function that takes λ and the vector of counts as input and returns the log-likelihood. Compute this log-likelihood for lambdas = seq(0,15,len=300) and make a plot. What value of lambdas maximizes the log-likelihood?

```
##break the genome into bins of 4000 basepairs and plot a histogram

breaks=seq(0,4000*round(max(locations)/4000),4000)
tmp=cut(locations,breaks)
counts=as.numeric(table(tmp))

hist(counts)
```



Maximum Likelihood - Exercises

10. (Cont.) Now write a function that takes λ and the vector of counts as input and returns the log-likelihood. Compute this log-likelihood for lambdas = seq(0,15,len=300) and make a plot. What value of lambdas maximizes the log-likelihood?

```
##Get Poisson Distribution for each of the values observed in the counts, with lambda=4. Define the likelihood as the product of those probabilities (this is assuming independence between the possible values the variable counts can take)

probs <- dpois(counts,4)
likelihood <- prod(probs)
likelihood
[1] 1.177527e-62
```

Maximum Likelihood - Exercises

10. (Cont.) Now write a function that takes λ and the vector of counts as input and returns the log-likelihood. Compute this log-likelihood for lambdas = seq(0,15,len=300) and make a plot. What value of lambdas maximizes the log-likelihood?

##Since the likelihood is a very tinny number, lets work with the log instead, as suggested

```
logprobs <- dpois(counts,4,log=TRUE)
loglikelihood <- sum(logprobs)
loglikelihood
```

[1] -142.5969

Maximum Likelihood - Exercises

10. (Cont.) Now write a function that takes λ and the vector of counts as input and returns the log-likelihood. Compute this log-likelihood for lambdas = seq(0,15,len=300) and make a plot. What value of lambdas maximizes the log-likelihood?

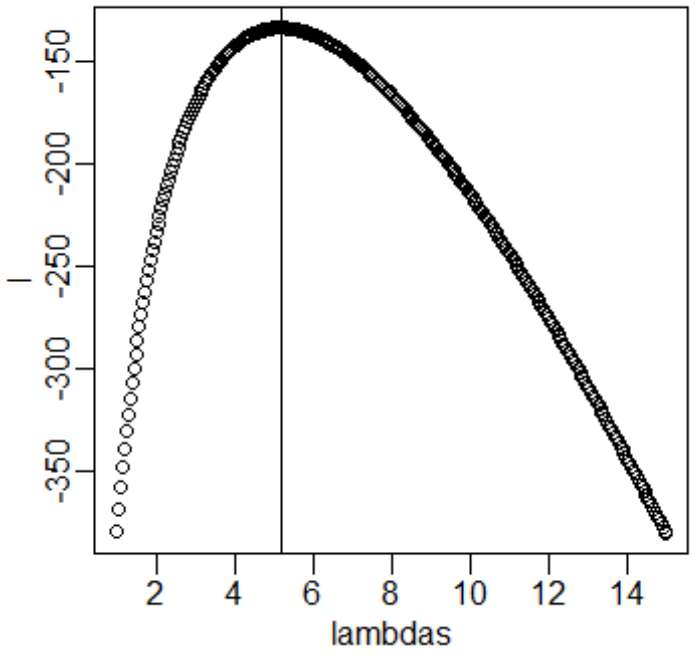
```
##Create the function that returns the log-likelihood based on lambda and the random variable
loglikelihood = function(lambda,x){
  sum(dpois(x,lambda,log=TRUE)) }
##Requested sequence of lambdas:
lambdas = seq(1,15,len=300)
##Create a data frame with the lambdas and the log-likelihood based on the lambdas and the counts,
l = sapply(lambdas,function(lambda) loglikelihood(lambda,counts))
```

Maximum Likelihood - Exercises

10. (Cont.) Now write a function that takes λ and the vector of counts as input and returns the log-likelihood. Compute this log-likelihood for lambdas = seq(0,15,len=300) and make a plot. What value c likelihood?

```
##Plot the lambdas against the log-likelihood
plot(lambdas,l)
##Get the MLE
mle=lambdas[which.max(l)]
abline(v=mle)
mle
```

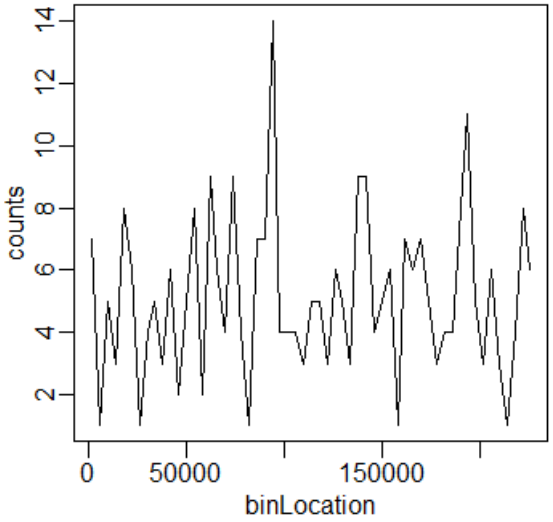
[1] 5.167224



Maximum Likelihood - Exercises

11. The point of collecting this dataset was to try to determine if there is a region of the genome that has a higher palindrome rate than expected. We can create a plot and see the counts per location:

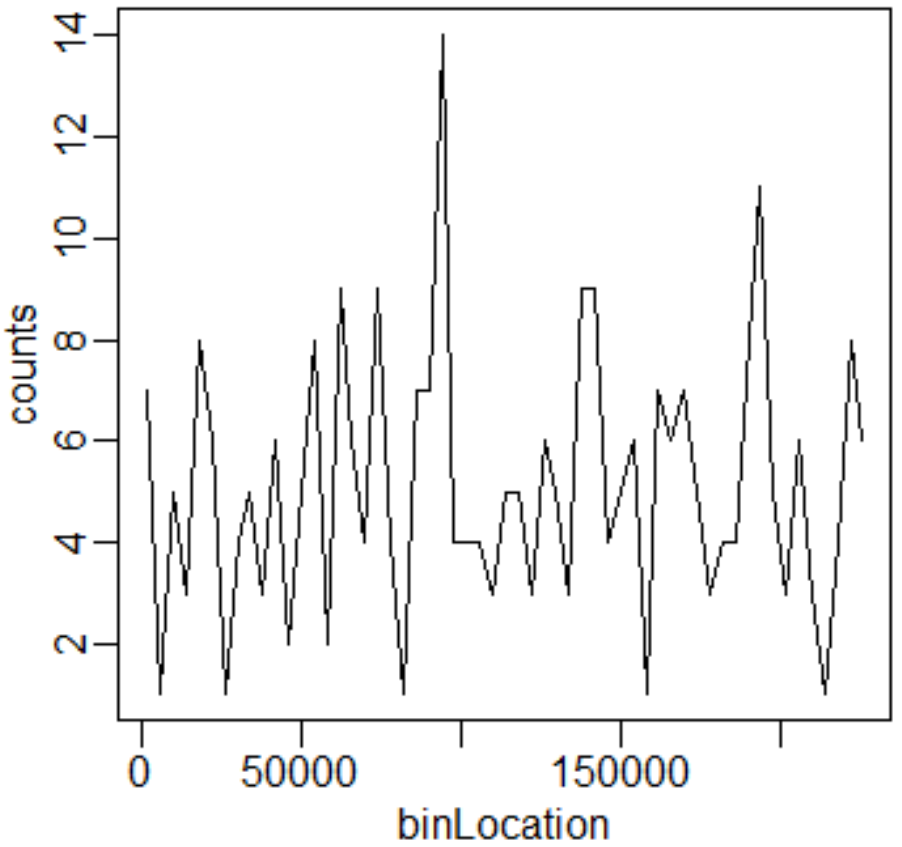
```
library(dagdata)
data(hcmv)
breaks=seq(0,4000*round(max(locations)/4000),4000)
tmp=cut(locations,breaks)
counts=as.numeric(table(tmp))
binLocation=(breaks[-1]+breaks[-length(breaks)])/2
plot(binLocation,counts,type="l",xlab=)
```



Maximum Likelihood - Exercises

11. (Cont.) What is the center of the bin with the highest count?

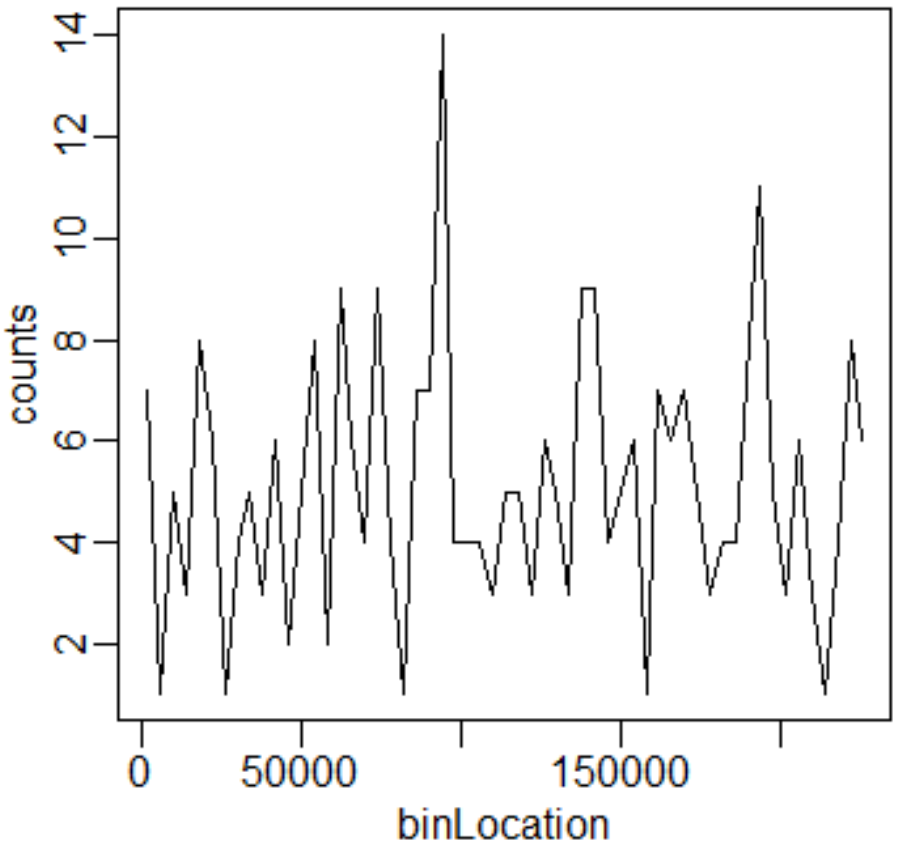
max(counts)
[1] 14



Maximum Likelihood - Exercises

12. For the question above, what is the maximum count?

```
binLocation[which.max(counts)
[1] 94000
```



Maximum Likelihood - Exercises

13. Once we have identified the location with the largest palindrome count, we want to know if we could see a value this big by chance. If X is a Poisson random variable with rate:
    lambda = mean(counts[- which.max(counts)])
    What is the probability of seeing a count of 14 or more?

##ppois give us the probability of getting a value lower than x, so we need to use the property of the complement of a cumulative probability to answer this questions

```
pval = 1 - ppois(13,lambda)
print(pval)
```
[1] 0.00069799

Maximum Likelihood - Exercises

14. So we obtain a p-value smaller than 0.001 for a count of 14. Why is it problematic to report this p-value as strong evidence of a location that is different?

A) Poisson in only an approximation.
B) We selected the highest region out of 57 and need to adjust for multiple testing.
C) λ is an estimate, a random variable, and we didn't take into account its variability.
D) We don't know the effect size.

Maximum Likelihood - Exercises

15. Use the Bonferonni correction to determine the p-value cut-off that guarantees a FWER (familywise error rate) of 0.05. What is this p-value cutoff?

Bonferonni correction is a multiple-comparison correction used when several dependent or independent statistical tests are being performed simultaneously
The FWER is the probability of making one Type I error (rejecting one true null hypothesis).
How many null hypothesis/p-values we can obtain from the data? As many as bin locations:

```
n=length(binLocation)
        [1] 57
```

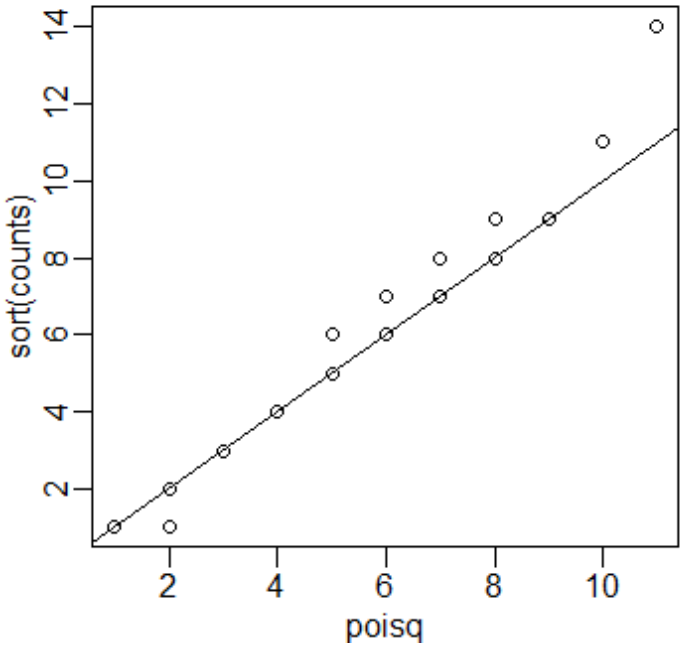
Bonferroni correction: FWER/n=0.05/57

```
alpha=0.05
bc=alpha/n
print(bc)
        [1] 0.000877193
```

Maximum Likelihood - Exercises

16. Create a qq-plot to see if our Poisson model is a good fit:

```
ps <- (seq(along=counts) - 0.5)/length(counts)
lambda <- mean( counts[ -which.max(counts)])
poisq <- qpois(ps,lambda)
plot(poisq,sort(counts))
abline(0,1)
```



Maximum Likelihood - Exercises

16. (Cont.) How would you characterize this qq-plot?

A) Poisson is a terrible approximation.
B) Poisson is a very good approximation except for one point that we actually think is a region of interest.
C) There are too many 1s in the data.
D) A normal distribution provides a better approximation.