DATA 581

Modeling and Simulation II

Lecture 2: Maximum Likelihood Estimation

What we discuss today

- **Likelihood Estimation**
- **Fitting Model using Maximum Likelihood Estimation**
- **Model checking using bootstrapping technique and MLE**

Motivation

- **Often in data analysis, we use a model to describe the relationship between two or more variables.**
  - **For example, you're going to discover the relationship between the revenue (y) of your company and the advertising budget (x).**
  - **Your guess is a linear model!**

$$y \approx ax + b$$

- **We have some data points $(x_i, y_i), i \in (1, 2, \ldots, n)$**
- **We want to find the parameters of a distribution (curve) that was most likely responsible for creating such data points.**

Motivation

- **The parameter values are found such that they maximize the likelihood that the process described by the model produced the data that were actually observed.**

- **What we want to calculate is the total probability of observing such a data (*Likelihood of observing such a data with the proposed model*);**

  - **if the probability of observing such a data following that models is high. the choice of the model is probably correct.**

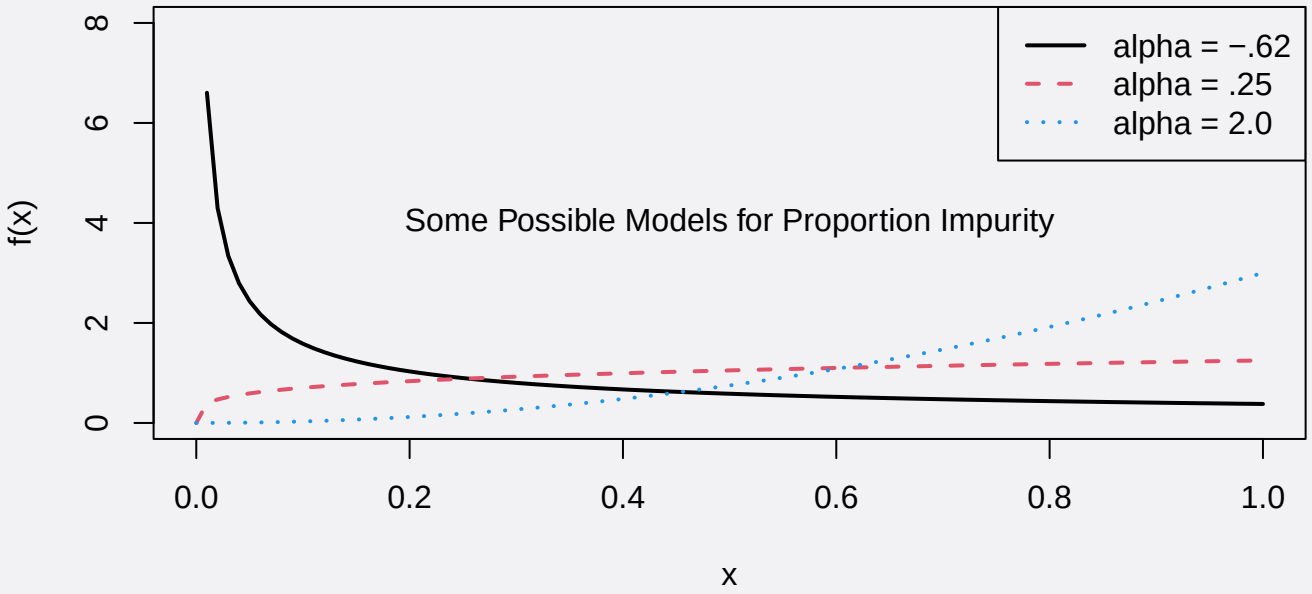
  - **note that we must have a parametric model to begin with, choice of the model comes from a background knowledge usually!**

Motivating Example

Suppose, from some background information, we know that the proportion of impurity x in an iron ore specimen can be modelled with the probability density function below:

$$f_X(x, \alpha) = (\alpha + 1)x^{\alpha}, \quad 0 \leq x \leq 1.$$

However, α is an unknown parameter.

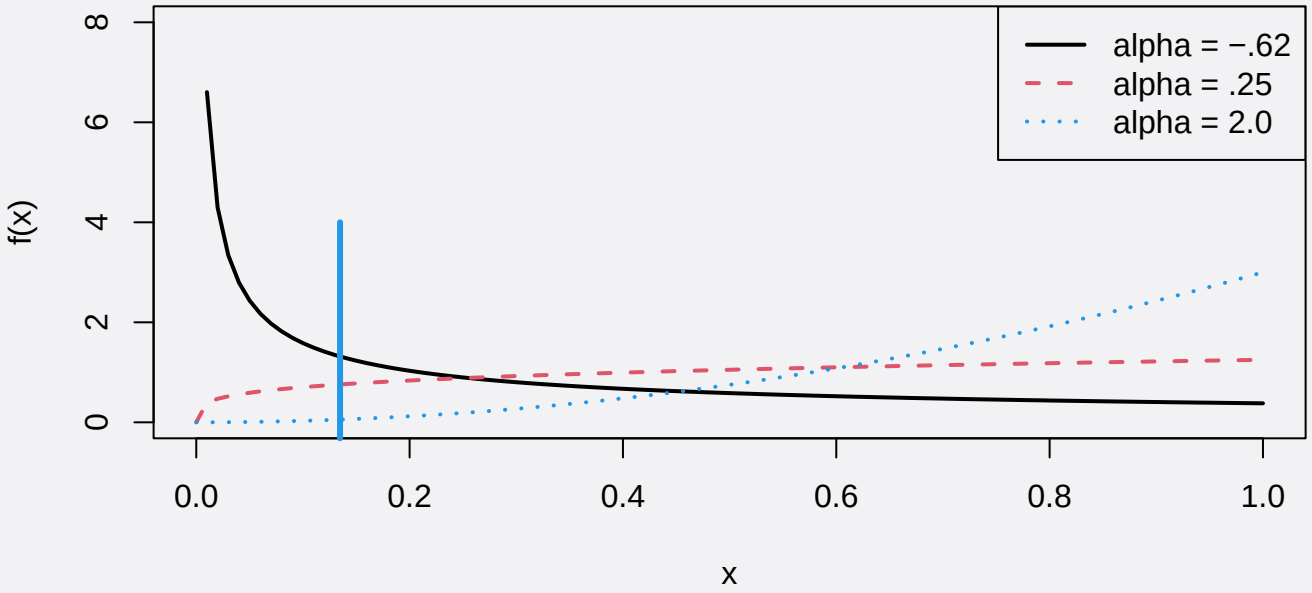


How can we find the right α?

Maximum likelihood estimation (MLE) is a method that determines values for the parameters (i.e α) of a model.

Taking one measurement

Suppose an impurity measurement is taken: $x = .1348$.



Unless we have taken an unusual measurement, the probability density function at our measurement should be high.

Which of the above curves appears to be the most likely?

Maximizing the Likelihood

If the parameter space consists only of the values $\{-.62, .25, 2.0\}$, then we would choose $\widehat{\alpha} = -.62$, since it seems to be the most likely value.

$\widehat{\alpha} = -.62$ is the MLE.

The maximum likelihood estimate is the value of the parameter which gives maximum probability density at the given data.

It is important to keep the possible set of parameter values in mind. Here there were 3 possibilities only.

In many cases there are an infinite number of possibilities.

Maximizing the Likelihood

Suppose the parameter space is the set where $\alpha > -1$.

The likelihood principle tells us to choose the value that gives highest density at the data point.

In other words, we need to see where

$$f_X(.1348, \alpha) = (\alpha + 1).1348^{\alpha}$$

is large.

Now, this is a function of α only,

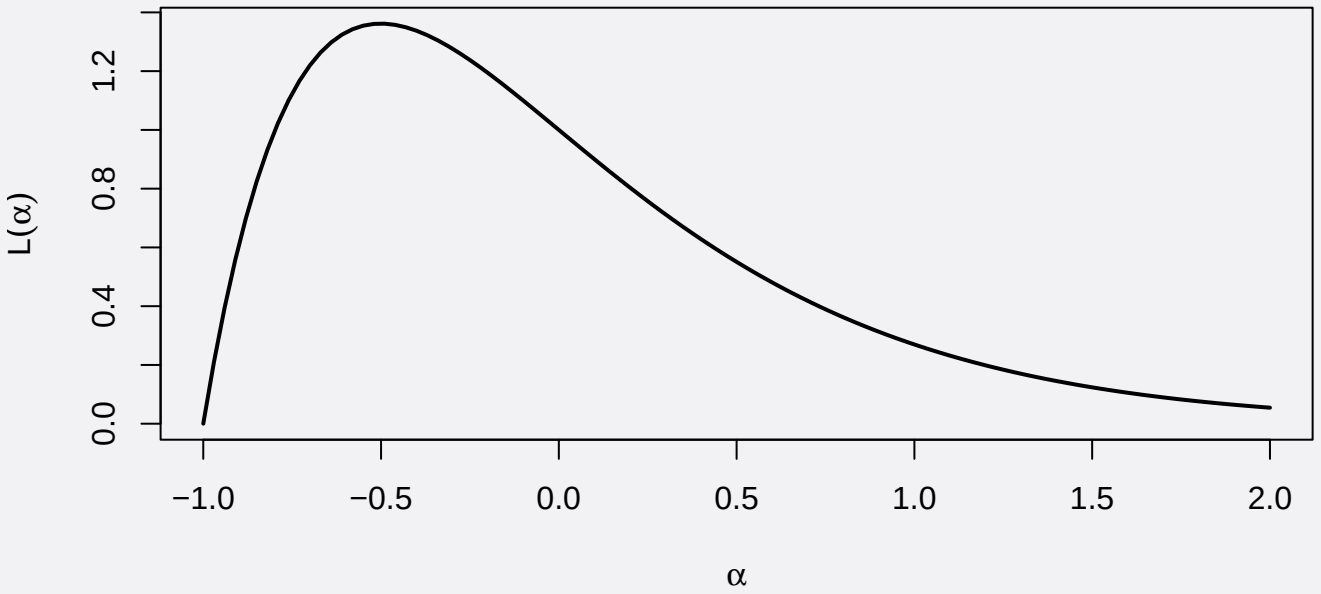
$$L(\alpha) = (\alpha + 1).1348^{\alpha}$$

$L(\alpha)$ is called the likelihood function.

Our goal is to find the value of α for which this function is maximized.

Choosing the Most Likely Density Function

When there is only one parameter, it is easy and a good idea to plot the graph of the likelihood function.

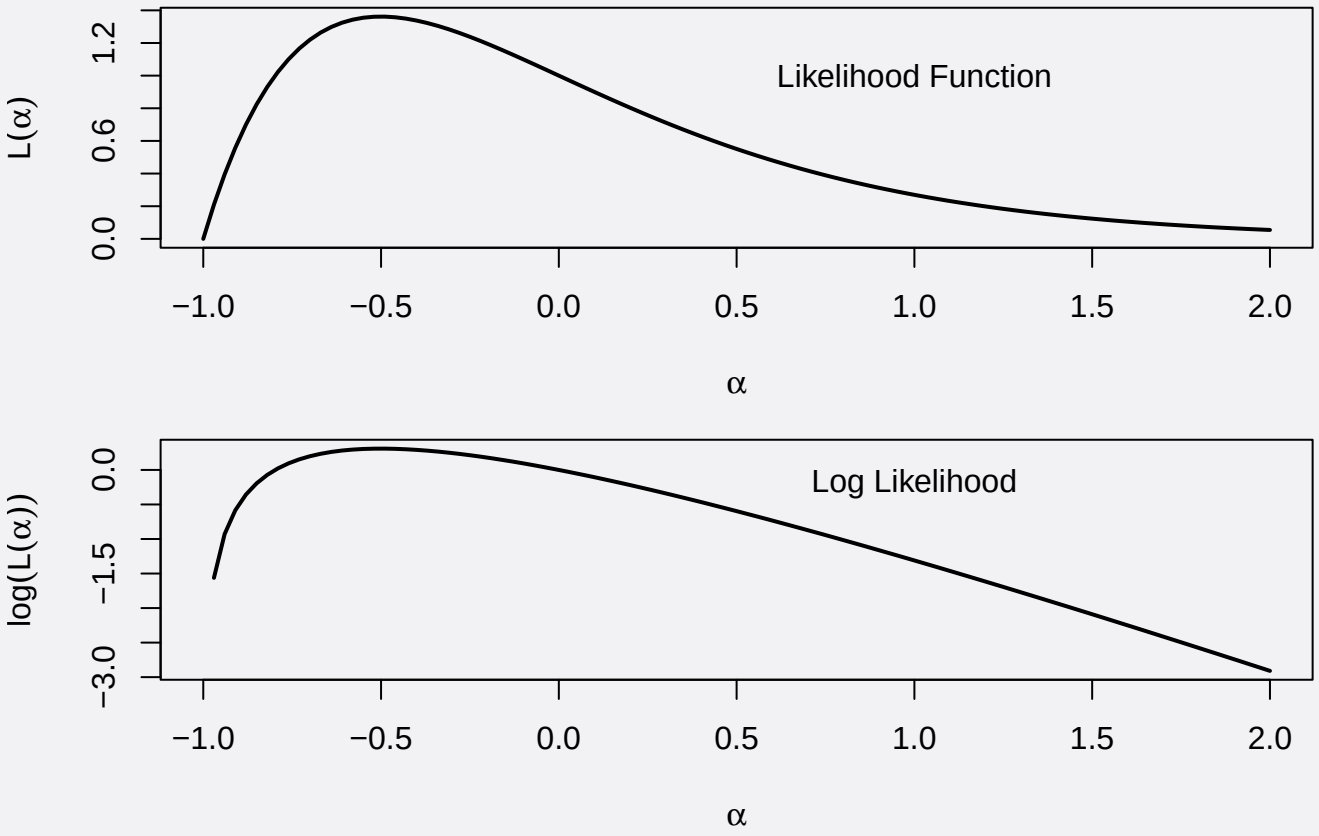


The function above measures the likelihood of our observed data as a function of α.

We want an α that maximizing this likelihood $\rightsquigarrow \alpha = -.5$.

Log Likelihood Function

Maximizing the log of the likelihood function gives the equivalent result, but is often computationally more convenient:



The maximum value occurs near $\alpha = -.5$ in both cases.

Review ; Some Calculus

The exact value that maximizes the likelihood function can *sometimes* be determined using calculus:

1. **Differentiate the likelihood $L(\alpha)$ or log-likelihood $\ell(\alpha) = \log L(\alpha)$ function with respect to α**

2. **Solve for α in $L'(\alpha) = 0$ to get the MLE.**

Example:

$$L(\alpha) = (\alpha + 1).1348^{\alpha}$$

$$\ell(\alpha) = \log(L(\alpha)) = \log(\alpha + 1) + \alpha \log(.1348)$$

Differentiate with respect to α:

$$\ell'(\alpha) = \frac{1}{\alpha + 1} + \log(.1348)$$

Solve $\ell'(\alpha) = 0$ for α:

$$\widehat{\alpha} = -1 - \frac{1}{\log(.1348)} = -0.501.$$

We should hope to get a better estimate of α if we have more than one measurement.

Estimating α with 2 Measurements

Assumption: **Data points are generated independently of each other.**

The joint density function for 2 independent impurity measurements, x_1 and x_2, is

$$f_X(x_1, x_2) = f_X(x_1) f_X(x_2)$$

$$(\alpha + 1)x_1^{\alpha}(\alpha + 1)x_2^{\alpha} = (\alpha + 1)^2(x_1 x_2)^{\alpha} = L(\alpha)$$

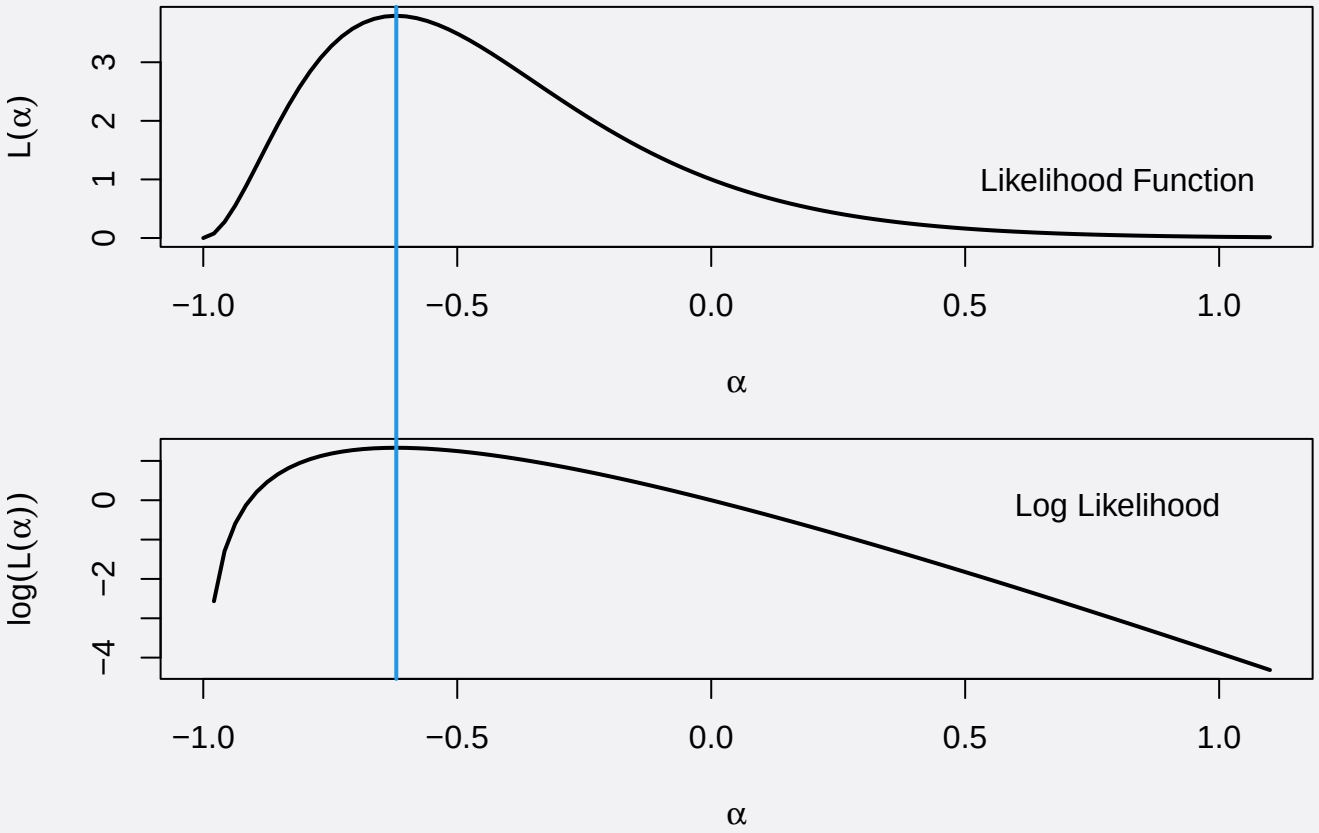
The log likelihood function is

$$\ell(\alpha) = \log L(\alpha) = 2\log(\alpha + 1) + \alpha \log(x_1 x_2).$$

Maximizing either of these functions with respect to α gives us the maximum likelihood estimate of α.

Estimating α with 2 Measurements

A second independent impurity measurement is .0381.



The maximum value occurs near $\alpha = -.6$ in both cases.

Estimating α with 2 Measurements

Evaluating the likelihood at the 2 measurements, we have

$$L(\alpha) = (\alpha + 1)^2 (.1348 \times .381)^{\alpha}$$

and the log likelihood, $\ell(\alpha)$ is

$$\ell(\alpha) = \log L(\alpha) = 2 \log(\alpha + 1) + \alpha \log(.00514)$$

By differentiating with respect to α, we can find the value of α that maximizes this. That is, solve

$$\ell'(\alpha) = \frac{2}{\alpha + 1} + \log(.00514) = 0.$$

The maximizer is $\widehat{\alpha} = -.62$.

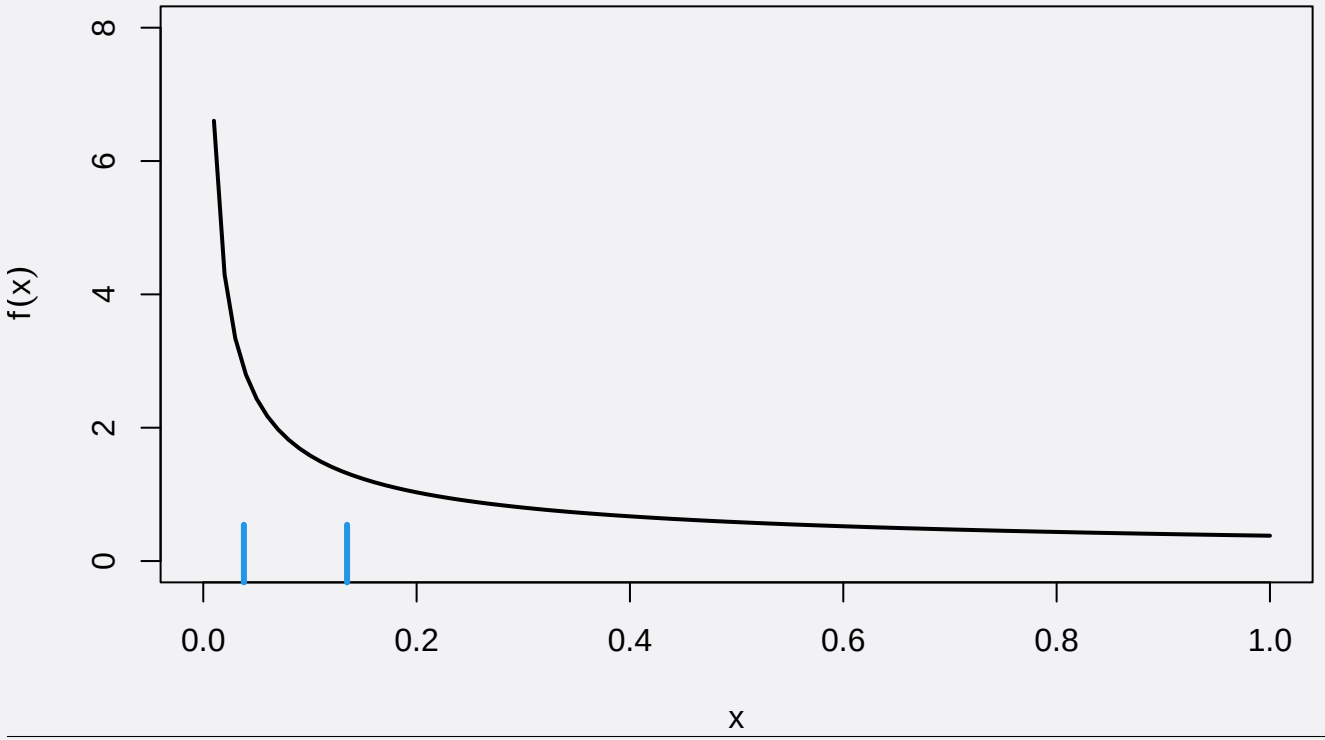
Estimating α with 2 Measurements

We can write

$$\widehat{f}_X(x) = (\widehat{\alpha} + 1)x^{\widehat{\alpha}} = .38x^{-.62}$$

as our estimate of the impurity pdf.

We can also plot it, together with the data, to see that it makes sense:



Estimating α with a Larger Sample

A better estimate can be obtained with a larger sample of impurity measurements.

Here is a sample of 10 independent measurements:

```
##  [1] 0.1348 0.0420 0.0003 0.0049 0.0002
##  [6] 0.0381 0.0018 0.0264 0.0366 0.0007
```

Because of independence, the joint density evaluated at the measurements is

$$f(x_1, x_2, \ldots, x_{10}; \alpha) = (\alpha + 1)^{10}(x_1 x_2 \ldots x_{10})^{\alpha}$$

Again, this is a function of the unknown parameter α. We can take the logarithm of this likelihood function:

$$\ell(\alpha) = \log L(\alpha) = 10 \log(\alpha + 1) + \alpha \sum_{j=1}^{10} \log(x_j)$$

Estimating α with 10 Measurements

The log likelihood function evaluates to

$$\ell(\alpha) = 10 \log(\alpha + 1) + \alpha(-50.9155).$$

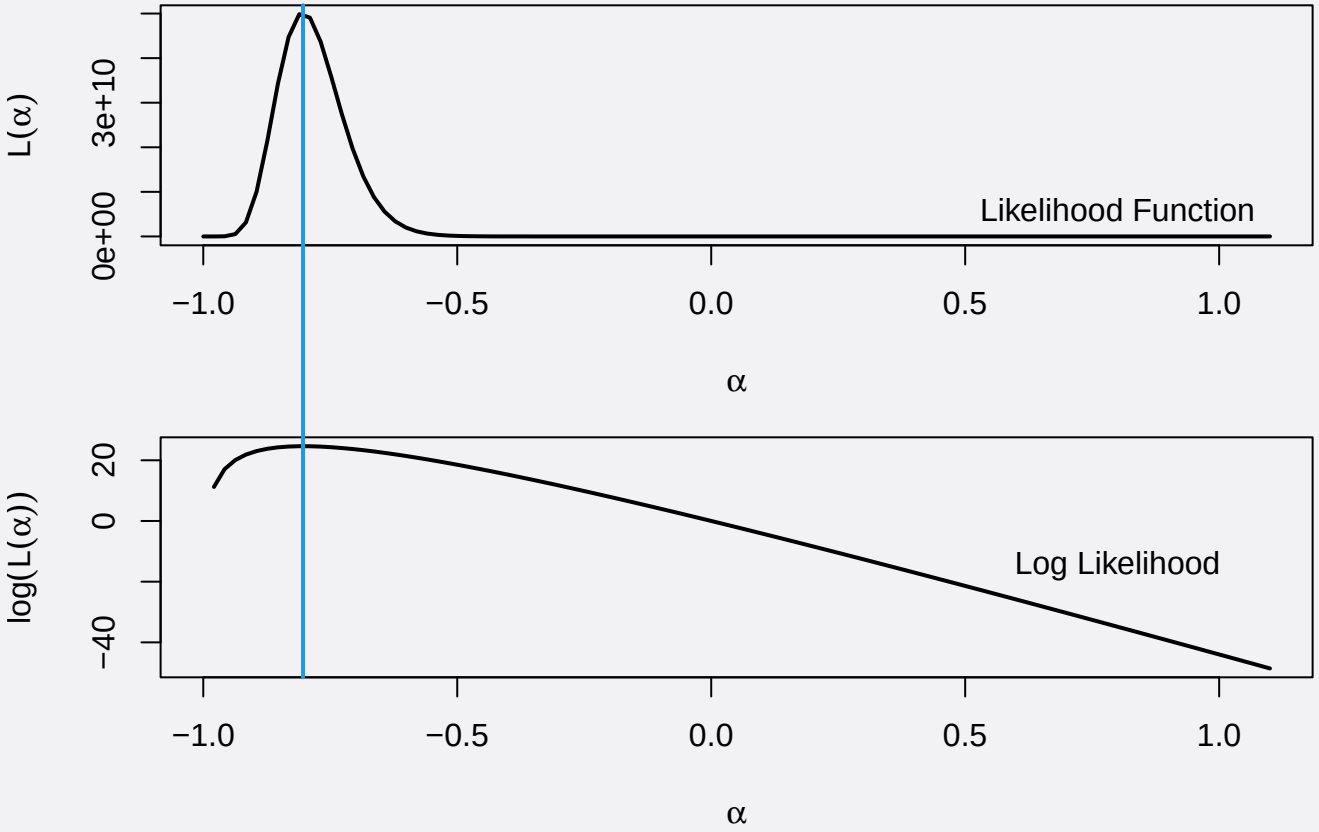
We maximize this by solving

$$\ell'(\alpha) = \frac{10}{\alpha + 1} - 50.9155 = 0.$$

$$\widehat{\alpha} = -0.8036.$$

Estimating α with 10 Measurements

The likelihood and log likelihood functions can be plotted:



The maximum value occurs at $\alpha = -0.8036$ in both cases.

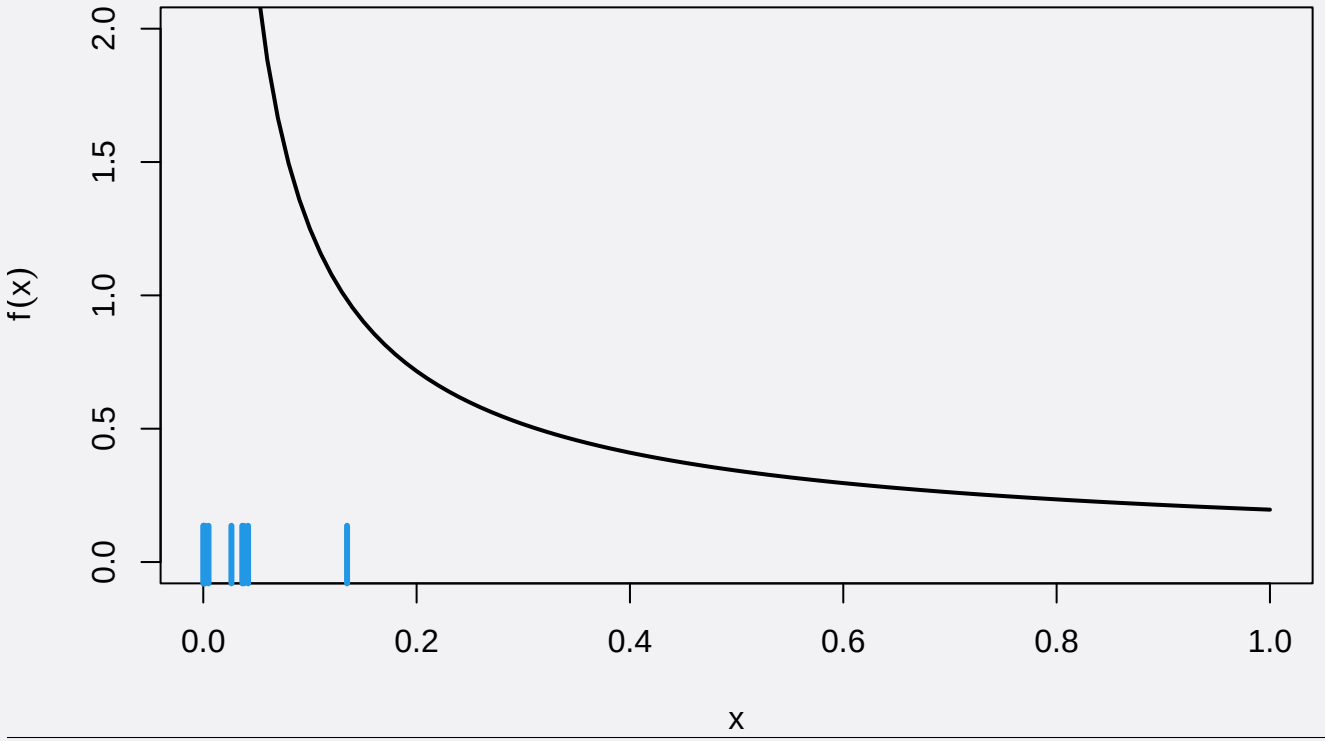
Note: The spread in the likelihood function is less than before $\rightsquigarrow$ the maximizer is more precise due to the increase in sample size.

Estimating α with 10 Measurements

The estimated pdf is now

$$\widehat{f}_X(x) = 0.1964 x^{-0.8036}.$$

Again, we plot the estimated pdf, together with the data, to see that it makes sense:



Summary of MLE; Poisson Distribution

Let's suppose we have some observations that we know are likely coming from a Poisson distribution.

Remember: Probability density function (PDF) of the Poisson distribution is :

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

We want to find the parameter λ.

1. **write the likelihood function which is simply the product of the PDF for the observed values $x_1, \ldots, x_n$.**

$$L(\lambda; x_1, \ldots, x_n) = \Pi_{j=1}^{n} \frac{\lambda_j^x e^{-\lambda}}{x_j!}$$

2. **to simplify the calculation, let's use log likelihood function;**

$$\ell(\lambda; x_1, \ldots, x_n) = log(\Pi_{j=1}^{n} \frac{\lambda_j^x e^{-\lambda}}{x_j!}) = \sum_{j=1}^{n} log(\frac{\lambda_j^x e^{-\lambda}}{x_j!}) = log(\lambda) \sum_{j=1}^{n} x_j - n\lambda - \sum_{j=1}^{n} log(x_j!)$$

3. **calculate the derivative of the log likelihood function with respect to λ.**

$$\frac{\partial \ell(\lambda; x_1, \ldots, x_n)}{\partial \lambda} = \frac{\partial}{\partial \lambda} \left(log(\lambda) \sum_{j=1}^{n} x_j - n\lambda - \sum_{j=1}^{n} log(x_j!) \right)$$

$$\frac{\partial \ell(\lambda; x_1, \ldots, x_n)}{\partial \lambda} = \frac{1}{\lambda} \sum_{j=1}^{n} x_j - n$$

4. **set the derivate to zero and solve for λ.**

$$\frac{1}{\lambda} \sum_{j=1}^{n} x_j - n = 0 \rightsquigarrow = \frac{\sum_{j=1}^{n} x_j}{n}$$

Conclusion **maximum likelihood estimates of the parameter λ in a Poisson distribution is equivalent to the sample mean.**

Using MLE and Bootstrapping for Model Checking

A Bootstrap Approach to Model-Checking

A dataset in MPV package contains numbers of cigarette butts at various distances from the smoking area on the university campus.

We assumed that the a poisson model well describe the relationship between the variables.

```
library(MPV)
y.glm <- glm(count~ distance, family = poisson,
    data = cigbutts)
a <- coef(y.glm)[1]
b <- coef(y.glm)[2]
coef(y.glm)

## (Intercept)    distance
##    3.553514   -0.001696
```

This means λ has the following form. $\rightsquigarrow \lambda = e^{(3.554 - 0.001 \times distance)}$

A Bootstrap Approach to Model-Checking

We can generate a Poisson probability density with the obtained parameters from our data.

The log likelihood contributions for each of the data points are:

```
count<-cigbutts$count
distance <- cigbutts$distance
log_likelihoodValues <- log( dpois(count, lambda = exp(a + b*
log_likelihoodValues

##  [1] -2.7749 -2.5004 -2.3738 -2.3333
##  [5] -2.4910 -1.6696 -2.0735 -1.4770
##  [9] -1.3414 -1.0139 -1.8848 -1.1129
## [13] -1.2804 -0.3029 -0.2158
```

These values don't seem too small, but how do we know what kind of value would be unacceptably small?

A Bootstrap Approach to Model-Checking

Idea: **if we could repeatedly simulate data from the true model, then we could calculate the likelihood contributions repeatedly, obtaining distributions of values that could be compared with our estimated likelihood contributions.**

We do not know the true model, but if we simulate from the fitted model, the fitted model will act as the true model in a "bootstrap world".

Thus, we apply the above idea to the fitted model, and compare the observed likelihood contributions from the observed data with distributions of likelihood contributions based on simulations from the fitted model.

A likelihood contribution beyond the lower end of a distribution range would be an indicator of a poor fit there.

- **Using the fitted model, N data sets are simulated and the log likelihood contributions for each of the x values are calculated.**

- **Side-by-side boxplots of the resulting log likelihood contributions are plotted and compared with the observed log likelihood contributions for each measurement.**

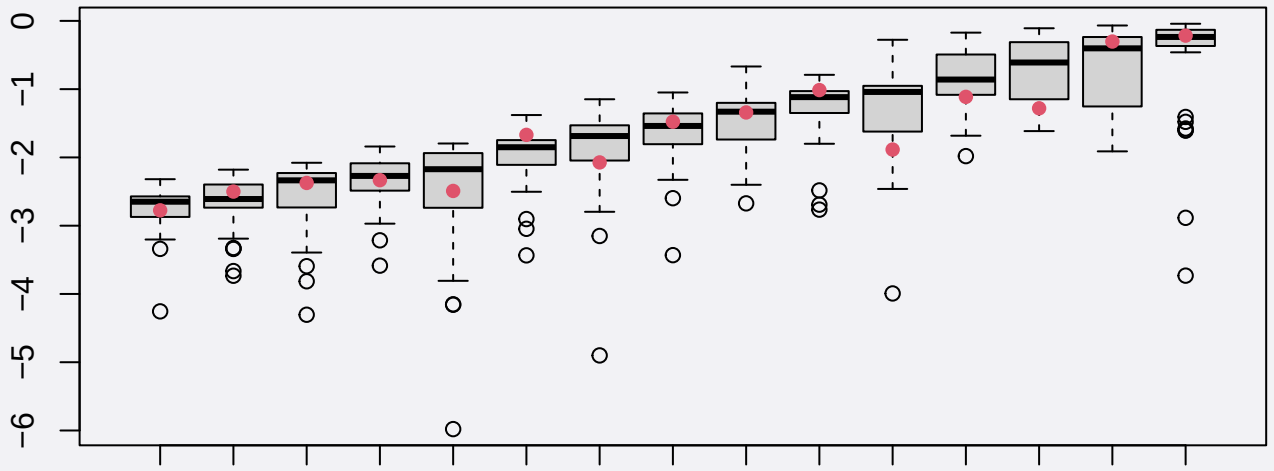
A Bootstrap Approach to Model-Checking - cigbutts Data

We can use the following function to simulate from the the fitted model and compute the log likelihood for every observation according to the simulated models.

```
N <- 40
# x = distance, y = cigbutts
simloglike <- function(x, y.glm, N=40) {
ll <- matrix(0, ncol=length(x), nrow=N)
for (j in 1:N) {
    y <- simulate(y.glm)$sim_1
    sim.glm <- glm(y ~ x, family = poisson) # family will change accordi
    a <- coef(sim.glm)[1]
    b <- coef(sim.glm)[2]
    log_likelihoodValues <- log(dpois(y, lambda = exp(a + b*x)))
    ll[j, ] <- log_likelihoodValues
        }
        ll <- data.frame(ll)
  names(ll) <- ""
  ll # this containes the MLE for each observation from all the simulated
}
```

A Bootstrap Approach to Model-Checking - cigbutts Data

```
ciglike <-  with(cigbutts, simloglike(distance, y.glm))
par(mar=c(1, 4, .1,.1))
boxplot(ciglike, ylim=range(c(log_likelihoodValues,ciglike)))
points(log_likelihoodValues, pch=16, col=2)
```



Distributions of observed likelihood contributions compared with resampled likelihood contributions. This is a check on the Poisson regression model for the cigbutts data.

The locations of the red dots within the range of the boxplots indicates a good fit.

- A similar simulation was run with an incorrect model

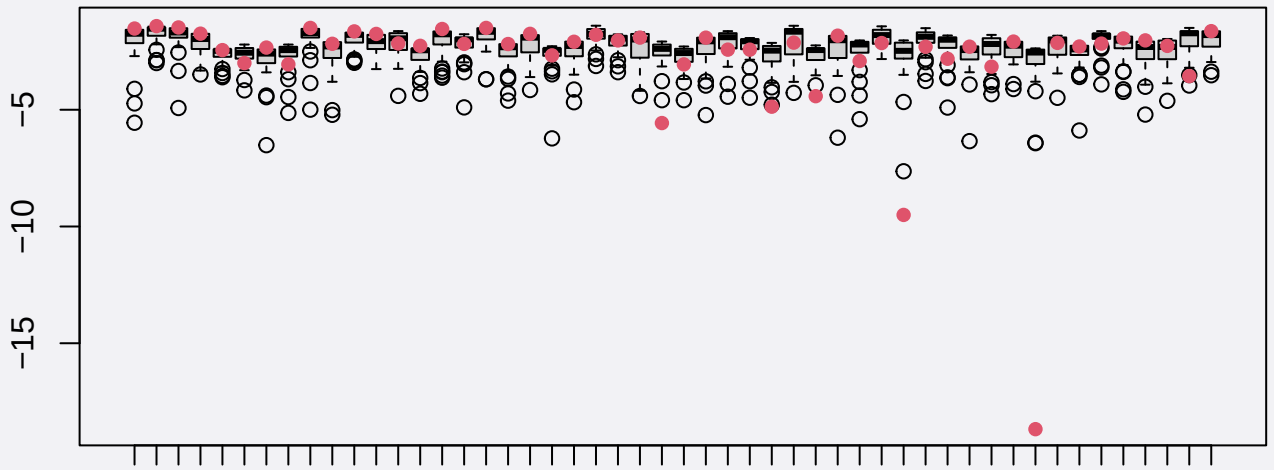
- the number of 0's has been inflated by multiplying the correct Poisson value by a random Bernoulli random variable.
  - When multiplied by 1, the value remains unchanged.
  - When multiplied by 0, the Poisson value may change to 0 from a nonzero value
  - 5% of the values are affected in this way.

Simulation of data where a Poisson regression model is inappropriately applied:

```
n <- 50
x <- runif(n, min = 0, max = 10)
theta <- 3 - .2*x
lambda <- exp(theta)
# manipulate y so doesn't fully follow a poisson model
y <- (rpois(n, lambda = lambda))*rbinom(n, 1, prob=.95)
#remember we know that poisson is not the good fit for this c
y.glm <- glm(y ~ x, family = poisson)
a <- coef(y.glm)[1]
b <- coef(y.glm)[2]
llikelihoodValues <- log(dpois(y, lambda = exp(a + b*x)))
```

A Bootstrap Approach to Model-Checking

```
simlike <- simloglike(x, y.glm)
par(mar=c(1, 4, .1,.1))
boxplot(simlike, ylim=range(c(llikelihoodValues,simlike)))
points(llikelihoodValues, pch=16, col=2)
```



Distributions of observed likelihood contributions compared with resampled likelihood contributions. The model is incorrect in this case, and we see log likelihood contributions that are far below the corresponding box plots.

A Bootstrap Approach to Model-Checking - Missile Success Data

We now try the technique out on a logistic regression example:

```
y.glm <- glm(y ~ x, family = binomial, data = p13.1)
a <- coef(y.glm)[1]
b <- coef(y.glm)[2]
llikelihoodValues <-  with(p13.1,
    log(dbinom(y, 1, prob = exp(a + b*x)/(1+exp(a + b*x)))))
```

A Bootstrap Approach to Model-Checking - Missile Success Data

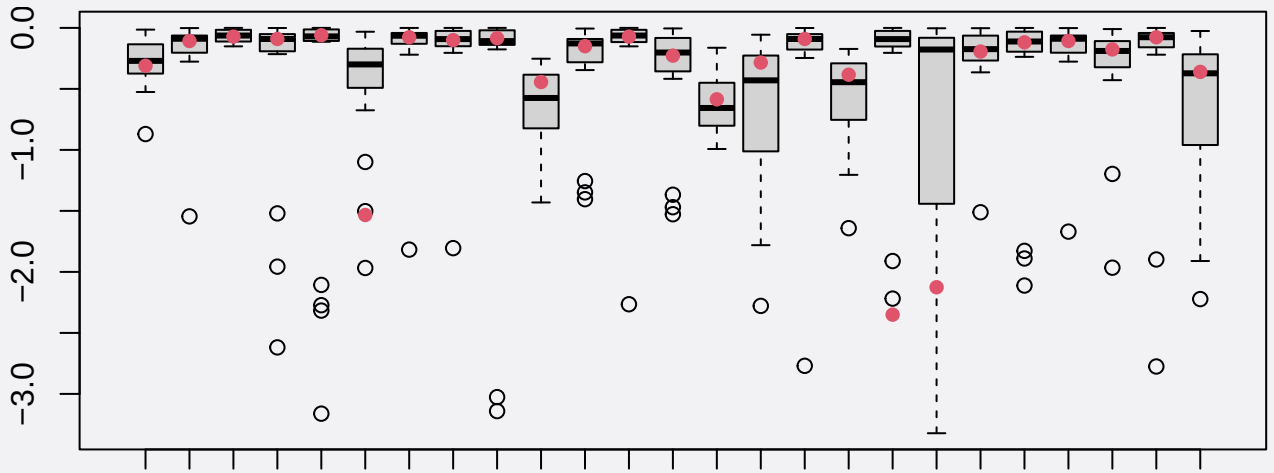
We need to re-write the `simloglike` function so that it takes the binomial model instead of Poisson:

```
simloglike <- function(x, y.glm, N=10) {
ll <- matrix(0, ncol=length(x), nrow=N)
for (j in 1:N) {
    y <- simulate(y.glm)$sim_1
    x <- p13.1$x
    sim.glm <- glm(y ~ x, family = binomial)
    a <- coef(sim.glm)[1]
    b <- coef(sim.glm)[2]
    loglikelihoodValues <- log(dbinom(y, 1,
        prob = exp(a + b*x)/(1+exp(a + b*x))))
    ll[j, ] <- loglikelihoodValues
}
ll <- data.frame(ll)   #names(ll) <- as.character(x)
names(ll) <- ""
ll
}
```

```
missilelike <-  with(p13.1, simloglike(x, y.glm, N=20))
```

Bootstrap Model-Checking - Missile Success Data

```
par(mar=c(1, 4, .1,.1))
boxplot(missilelike, ylim=range(c(llikelihoodValues,missilelike)))
points(llikelihoodValues, pch=16, col=2)
```



Distributions of observed likelihood contributions compared with resampled likelihood contributions. This is a check on the logistic regression model for the missile success data. The locations of the red dots within the range of the boxplots indicates a good fit. There are two problematic observations indicating that there are some difficulties with this model.

What to Take Away from this Lecture

- **The cpncept of Maximum Likelihood Estimation**
- **Maximizing the likelihood when there are a small number of possible parameter values**
- **Maximizing the likelihood when there is a continuous infinity of possible parameter values, using calculus**
- **Using bootstrapping and MLE to evaluate the fitness of a model to our data.**