POLSCI 9592

Lecture 1: Maximum Likelihood Estimation

Dave Armstrong

Goals for this class

1. Introductions
2. Discuss direction for the course.
3. Describe Maximum Likelihood Estimation (MLE)
4. Consider a couple of examples of MLE

Experiment

- Go to the following Google sheet
- In the column with your name, put the results of the following experiment:
    - Roll your die 4 times and count the number of *even* numbers you get.
    - Record the number of even rolls you get in Trial 1.
    - Repeat for Trials 2-5.

Questions About Experiment

1. What is the overall mean - how could we figure out how variable it is?
2. If we had a hypothesis about everyone's die being fair, how would we evaluate it?
3. What if we wanted to estimate $Pr(\text{Even})$ for each person?
4. What if we wanted to do this in a regression context?

Binomial Distribution

We could use the binomial distribution to figure this out. It assumes Bernoulli trials:

1. There are only two outcomes (success and failure, no judgment intended)
2. All trials have the same underlying probability p.
3. The trials are independent from each other.

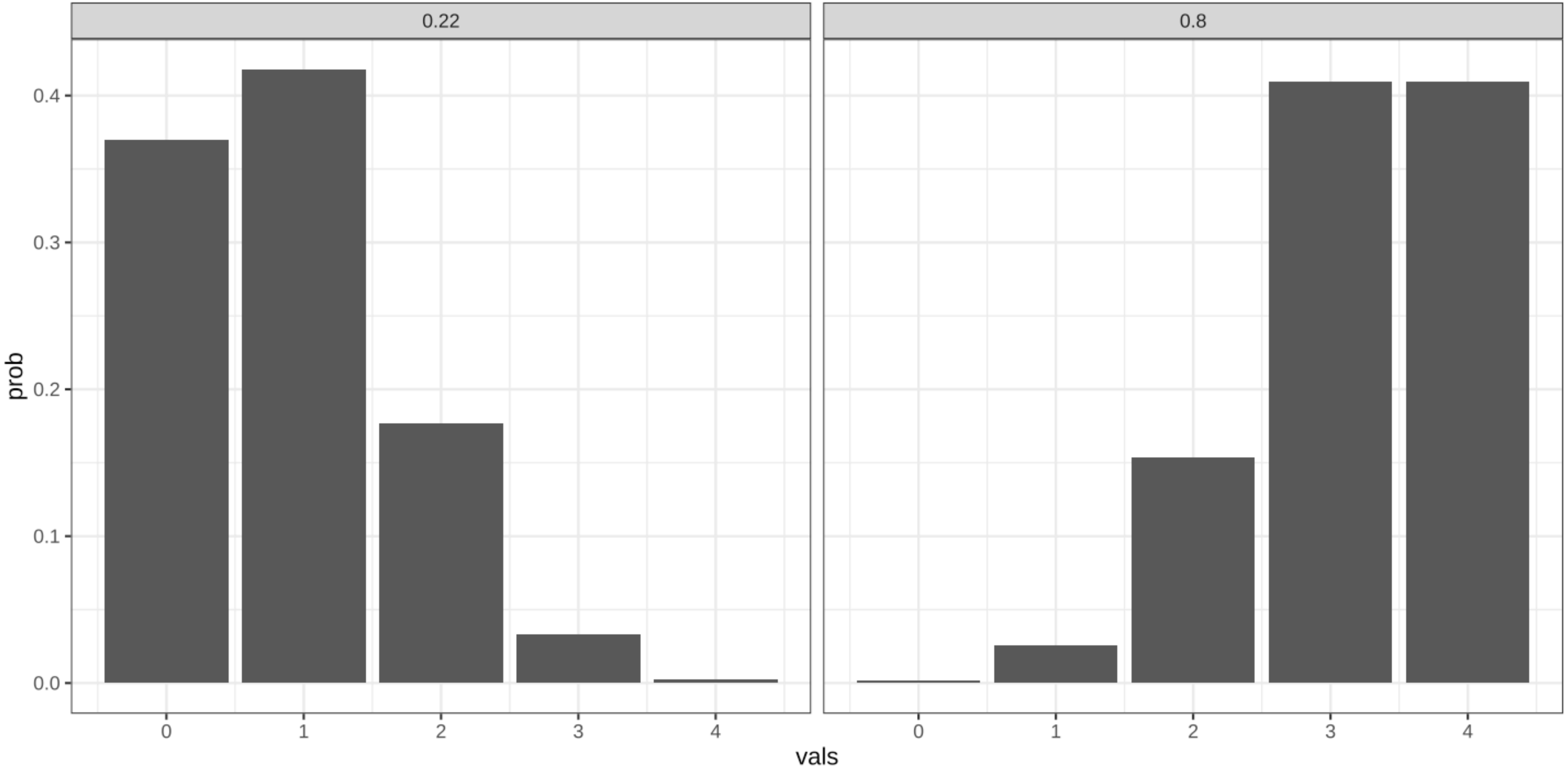
If y has a binomial distribution, then

$$f(y) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

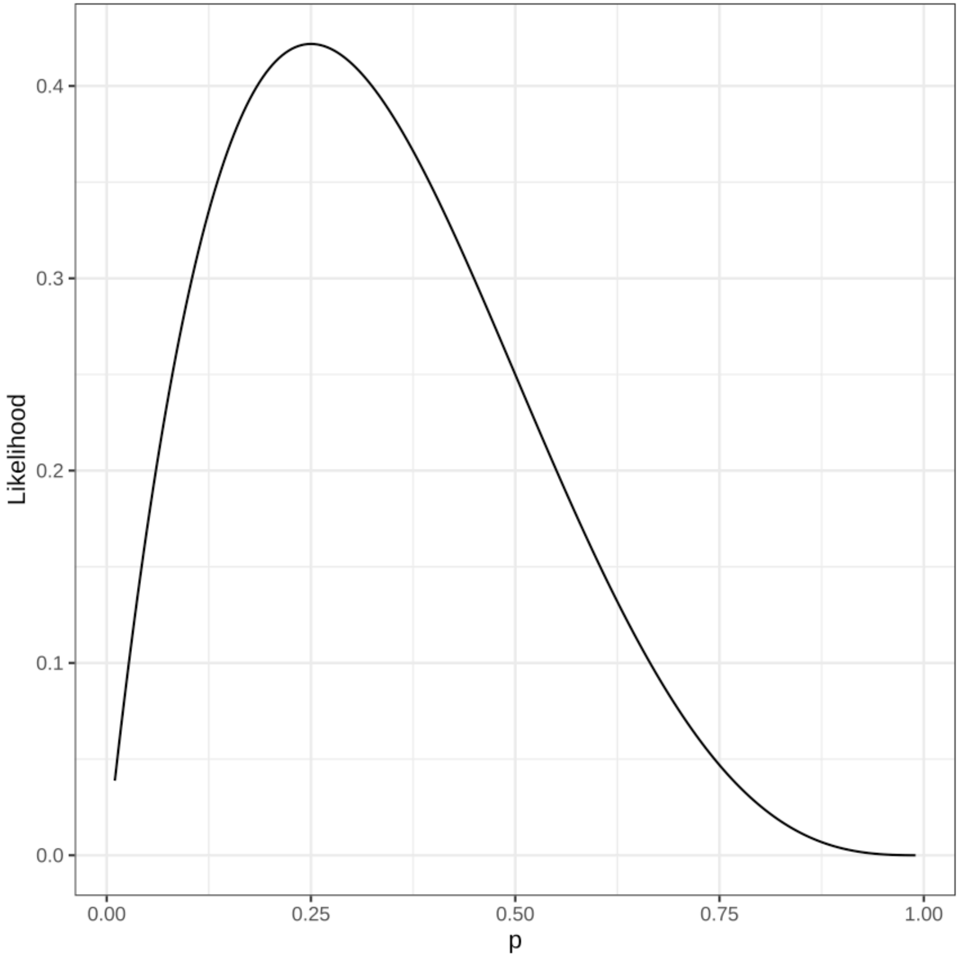
where $\binom{n}{k}$ is the binomial coefficient and is defined as $\frac{n!}{k!(n-k)!}$. Calculating this out tells us the number of possible possible outcomes of size n that have exactly k successes.

Binomial Example

Let's say I had one data point, 1 out of 4 rolls was even and I had to pick between two different values of p that produced these two distributions.

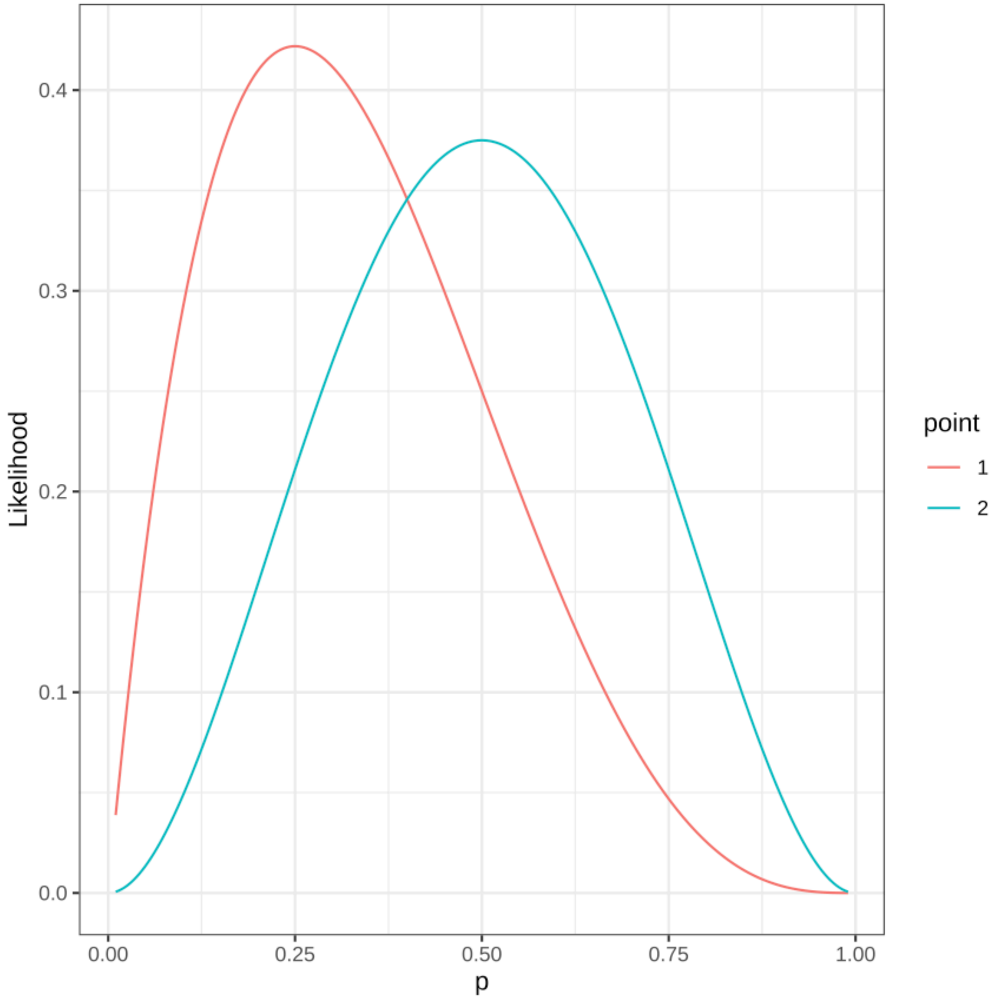


Look Over all Values of *p*



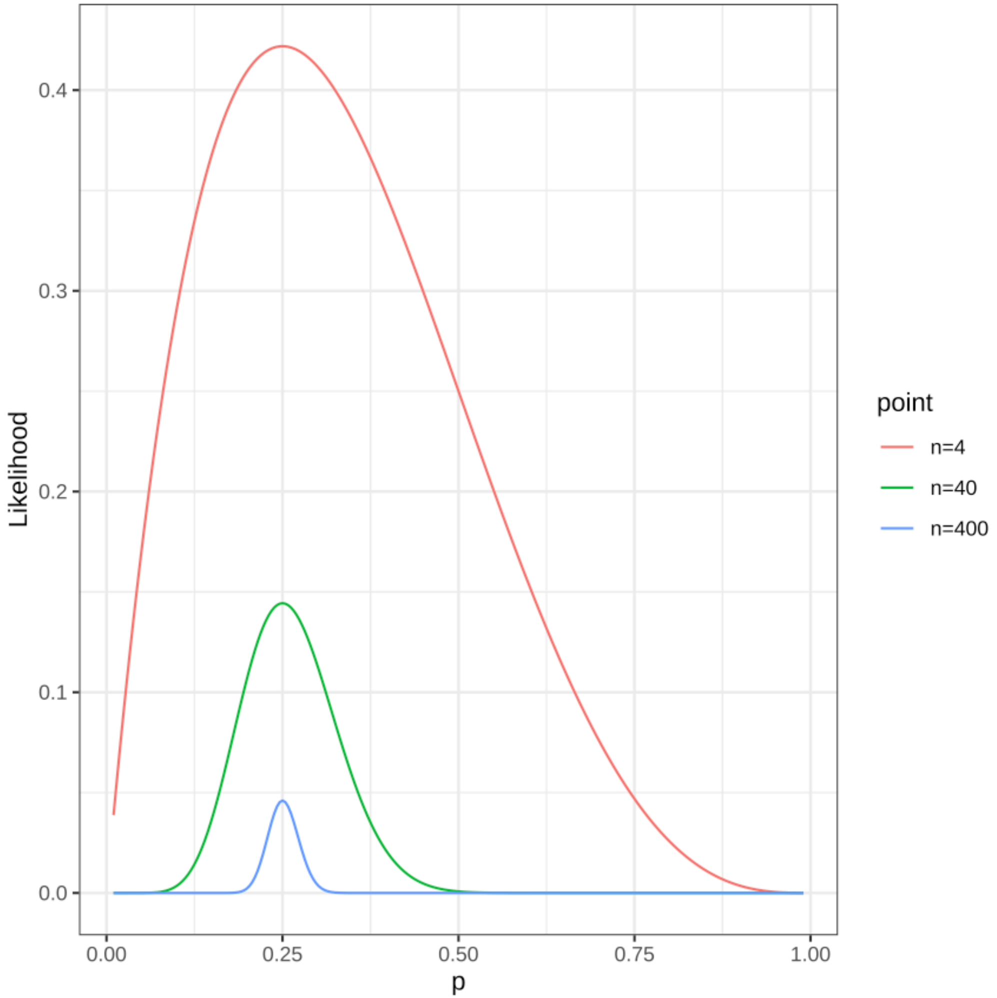
2 Data Points

What if we had 2 data points (1 and 2 even rolls out of 4)?



More Data!

What if, instead of 4 rolls and 1 even, I did 40 rolls with 10 events (or 400 rolls with 100 events)?



MLE

Maximum Likelihood Estimation is a different way of estimating statistical relationships.

- Least Squares is also a method and while robust to the violation of lots of its assumptions, it assumes a "continuous" dependent variable (or put differently, it assumes that the errors are normally distributed)
- Put another way, it assumes that each value of y, is normally distributed in repeated sampling.
- This assumption need not be made and as we will see, linear relationships can also be estimated with MLE, but so can lots of others.

Probability

Before we go into MLE, we need to refresh ourselves on the axioms of probability. Let's assume that the sample space is S (the set of all possible outcomes):

1. For any event A, $Pr(A \geq 0)$.
2. $Pr(S) = 1$
3. If events A, B and C are mutually exclusive, then $Pr(A \& B \& C) = Pr(A) \cdot Pr(B) \cdot Pr(C)$

Probability Density Function

One of the main elements of a likelihood estimation is a *probability density function* or *PDF* .

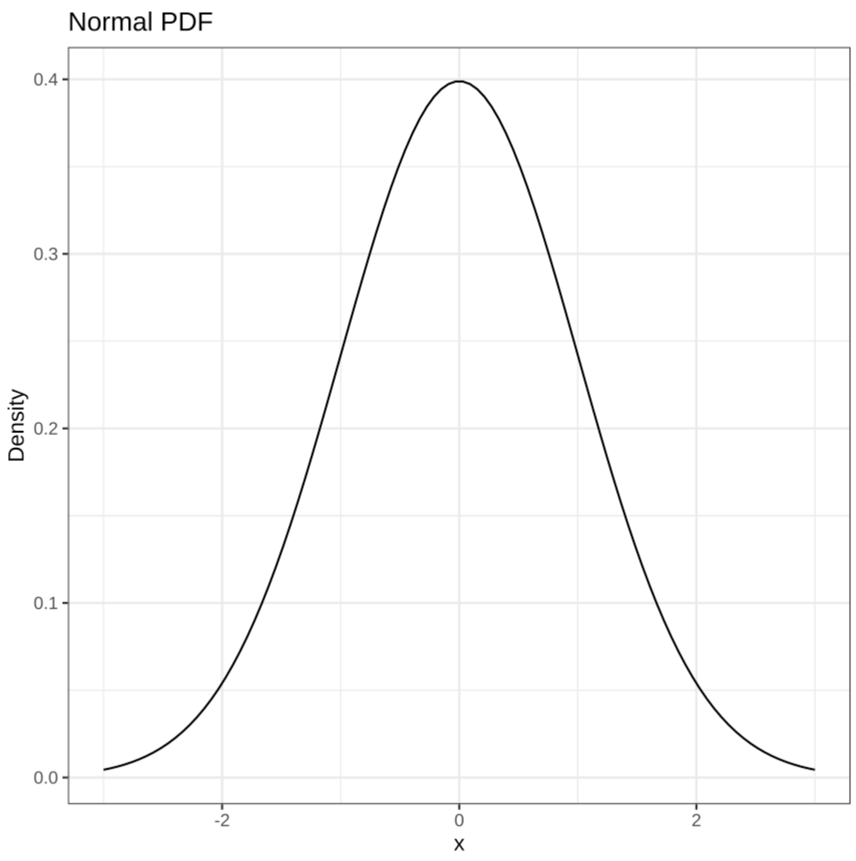
- The PDF gives the *relative likelihood* that an observation drawn randomly from a continuous random variable would take on a particular value.
- We can use these relative likelihoods to find the best parameters for our relationships.

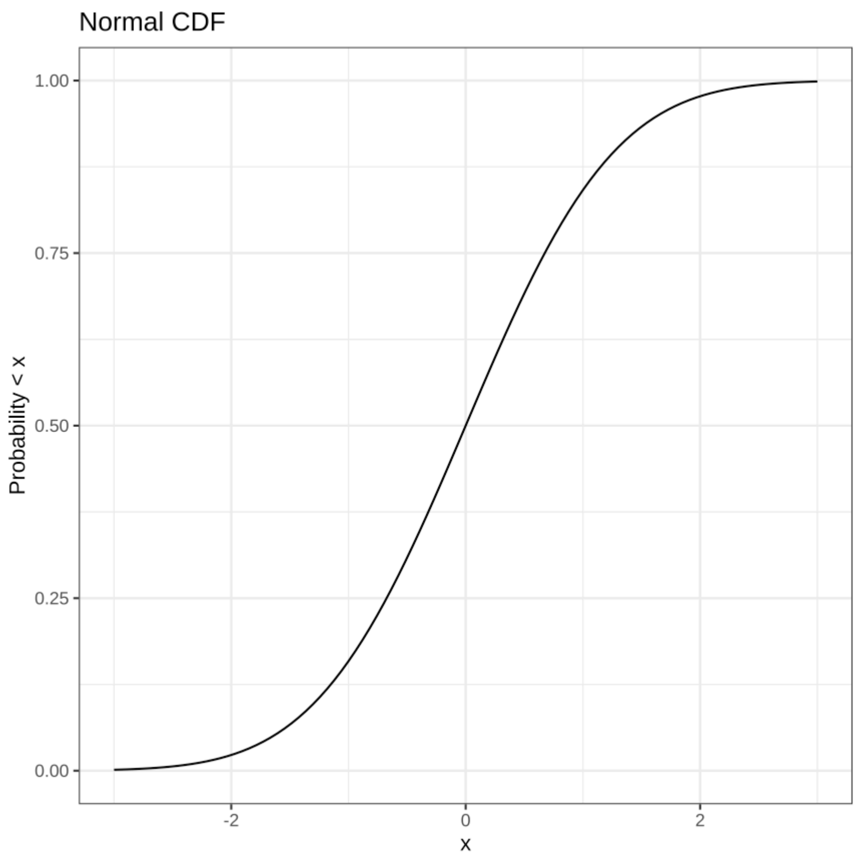
Cumulative Distribution Function

A counterpart to the PDF is the *Cumulative Distribution Function* or *CDF*.

- The CDF tells us the probability of being below (or alternatively above) a certain value of a random variable.
- The z- and t- tables in the back of your stats book from last semester give you the CDF for the normal and t distributions evaluated at lots of different points.

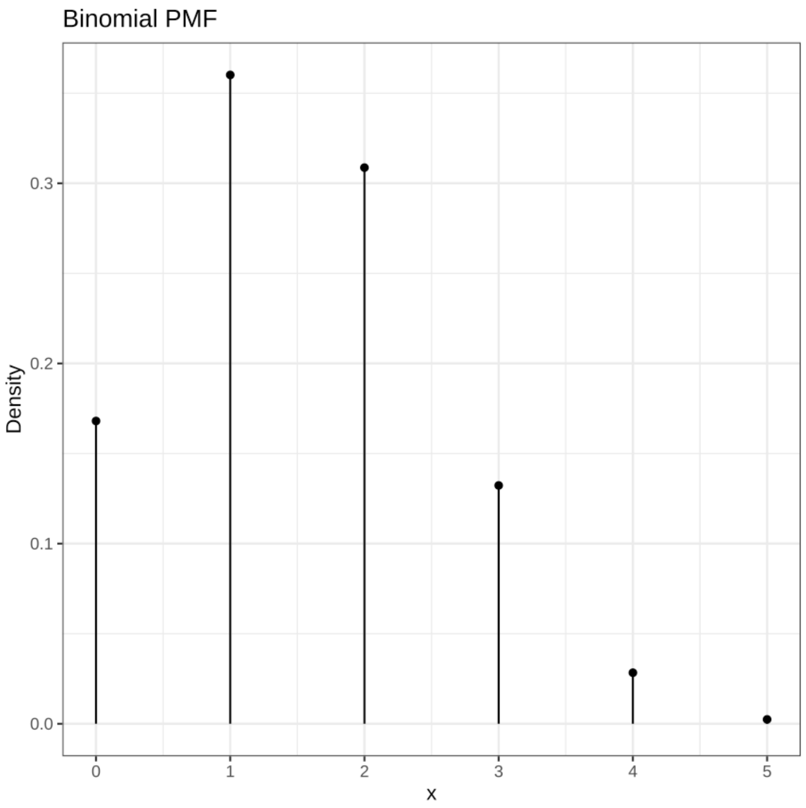
Normal PDF and CDF

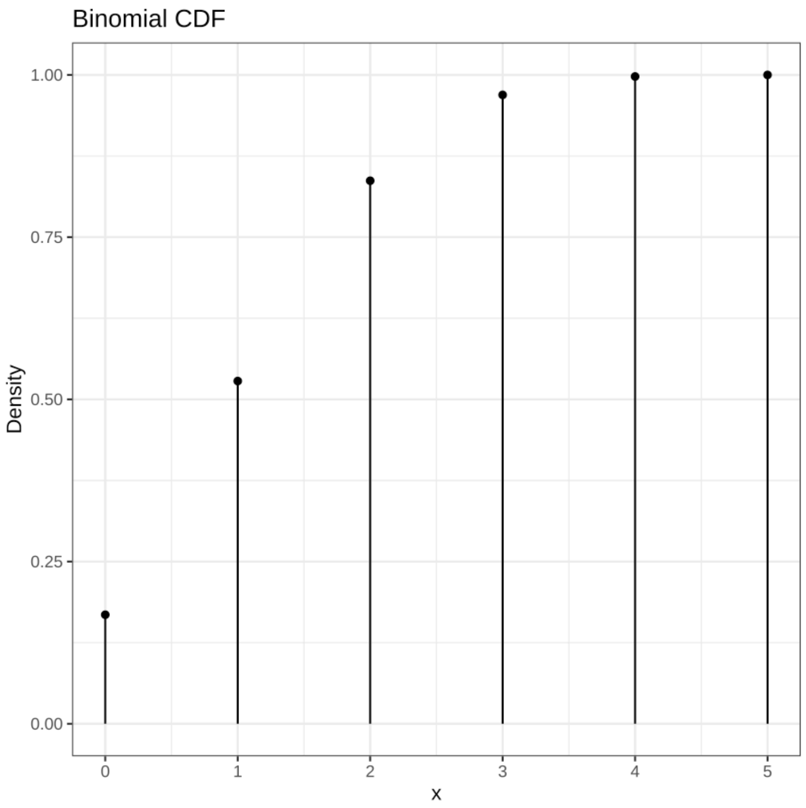




Discrete Distributions

With discrete distributions (those where some values are impossible, like counts), we have a *probability mass function* or *PMF* instead of a PDF.





What is Likelihood

The Likelihood Axiom is as follows:

$$\begin{aligned} L(\tilde{\theta}|y) &= k(y) f(y|\tilde{\theta}) \\ &\propto f(y|\tilde{\theta}) \end{aligned}$$

- $f(y|\tilde{\theta})$ is a probability density function of y given the hypothetical model parameters $\tilde{\theta}$
- $k(y)$ is an unknown function that depends only on the data, not the parameters.
- What Maximum Likelihood Estimation does is to pick the parameters $\hat{\theta}$ that make the data most likely to have been generated given the assumptions we make.

Evaluating the Likelihood function

Assuming that you've got lots of y values (i.e., y_i for $i = 1, \ldots, n$), then you would want to know the aggregate likelihood for all values (i.e., a single number).

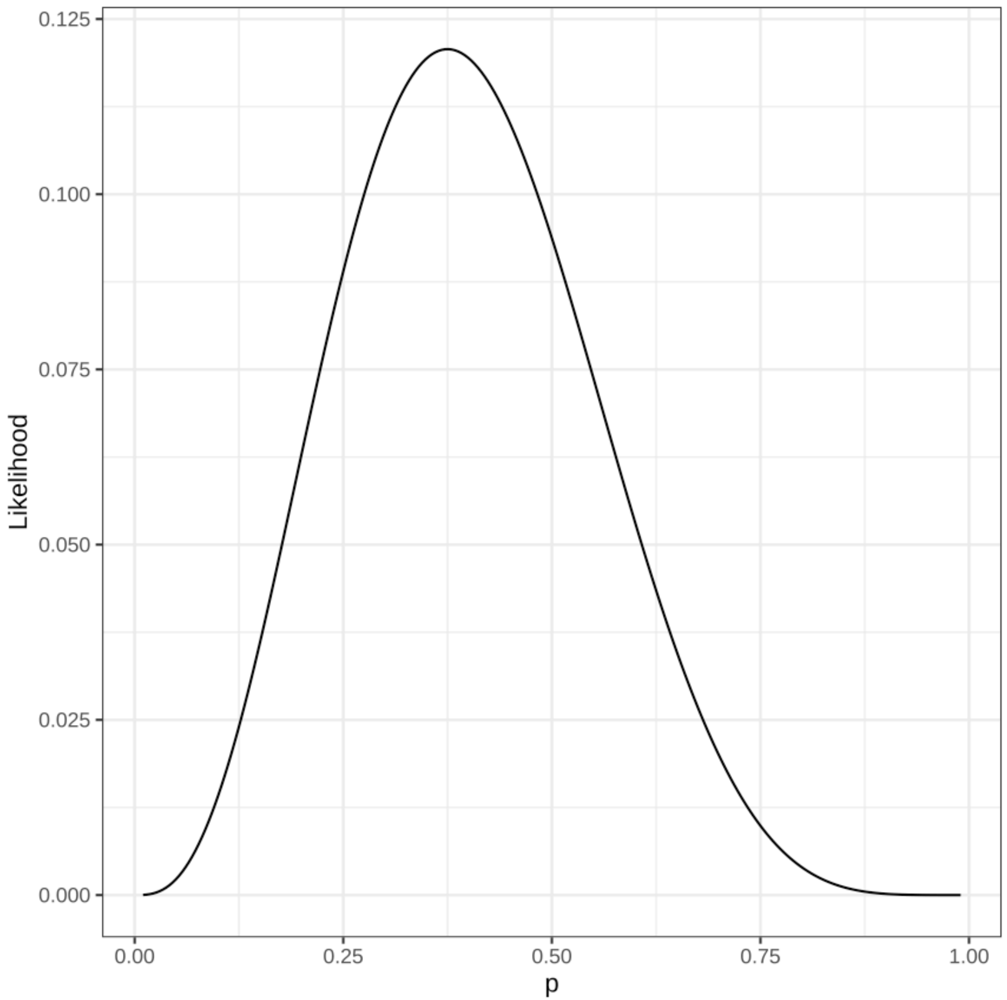
$$L(\theta|y) = \prod_{i=1}^{n} L(\theta, y_i) \propto \prod_{i=1}^{n} f(y_i|\theta)$$

While this would theoretically work fine, taking the product of a bunch of small numbers is going to generate something that the computer will have difficulty dealing with, thus we usually try to maximize the log-likelihood (LL).

$$LL(\theta|y) = \sum_{i=1}^{n} LL(\theta, y_i) \propto \sum_{i=1}^{n} log\left(f(y_i|\theta)\right)$$

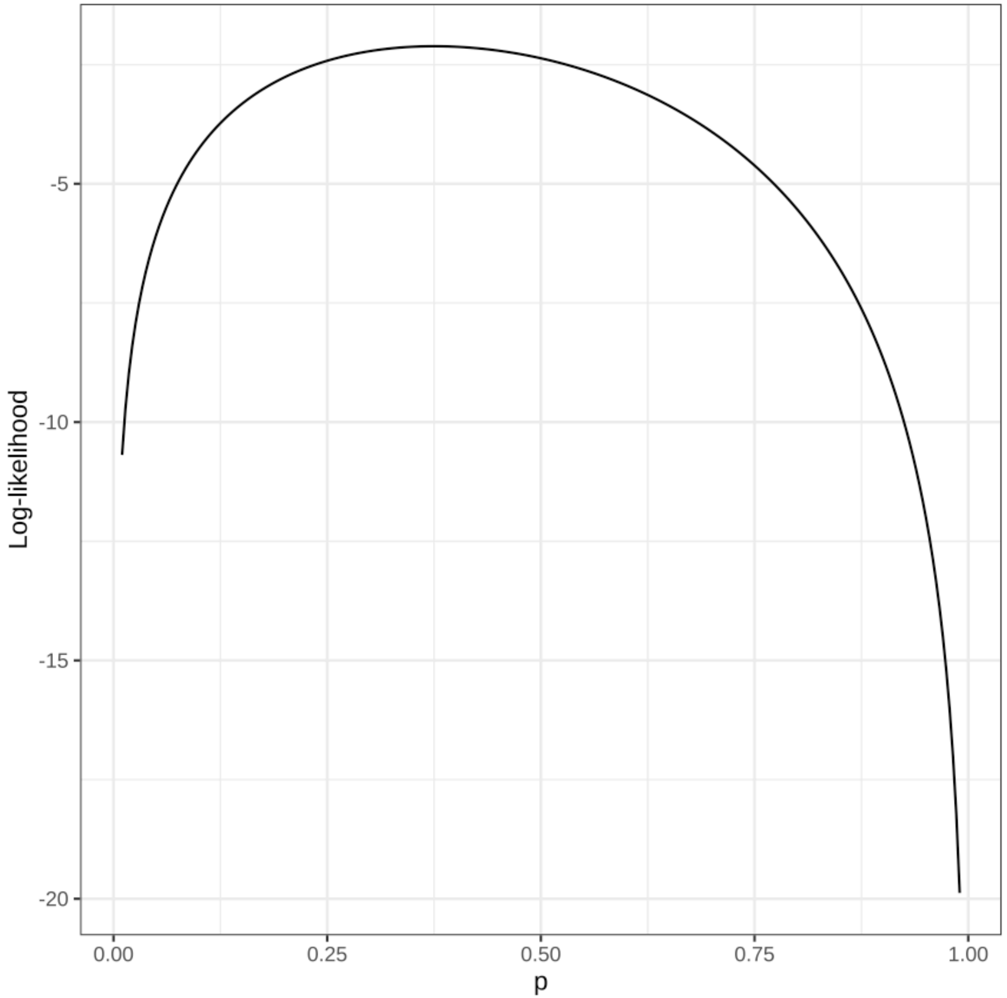
Back to the 2 data points

If we wanted to know what the combined probability was for data points 1 and 2 given a certain p, we would want to know $Pr(y_1 = 1|p) \times Pr(y_2 = 2|p)$.



2 data points: Log-likelihood

Using the product of the probabilities gives the likelihood, if we wanted the log-likelihood, we would take the sum of the log of the probabilities.

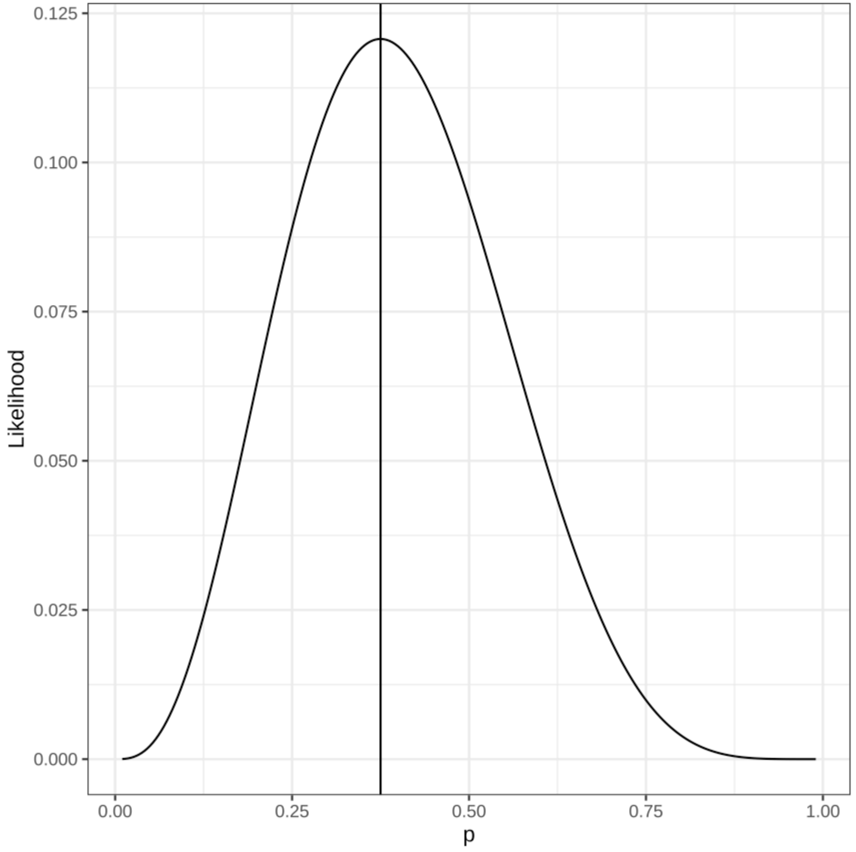


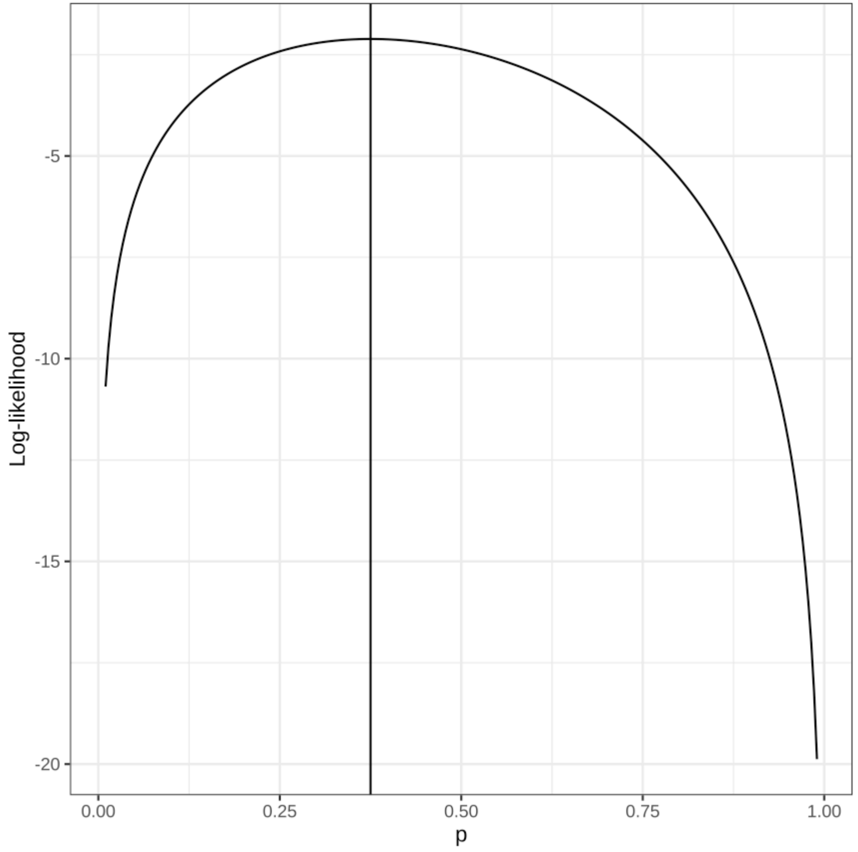
Functions in R

```r
library(maxLik)
llfun <- function(par, x){
    p <- dbinom(x, 4, par[1])
    sum(log(p))
}
out <- maxLik(llfun, start=.5, x=c(1,2))
summary(out)
```

```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 2 iterations
## Return code 1: gradient close to zero (gradtol)
## Log-Likelihood: -2.114452
## 1  free parameters
## Estimates:
##      Estimate Std. error t value Pr(> t)
## [1,]   0.3750     0.1712   2.191  0.0285 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

Likelihood and Log-Likelihood





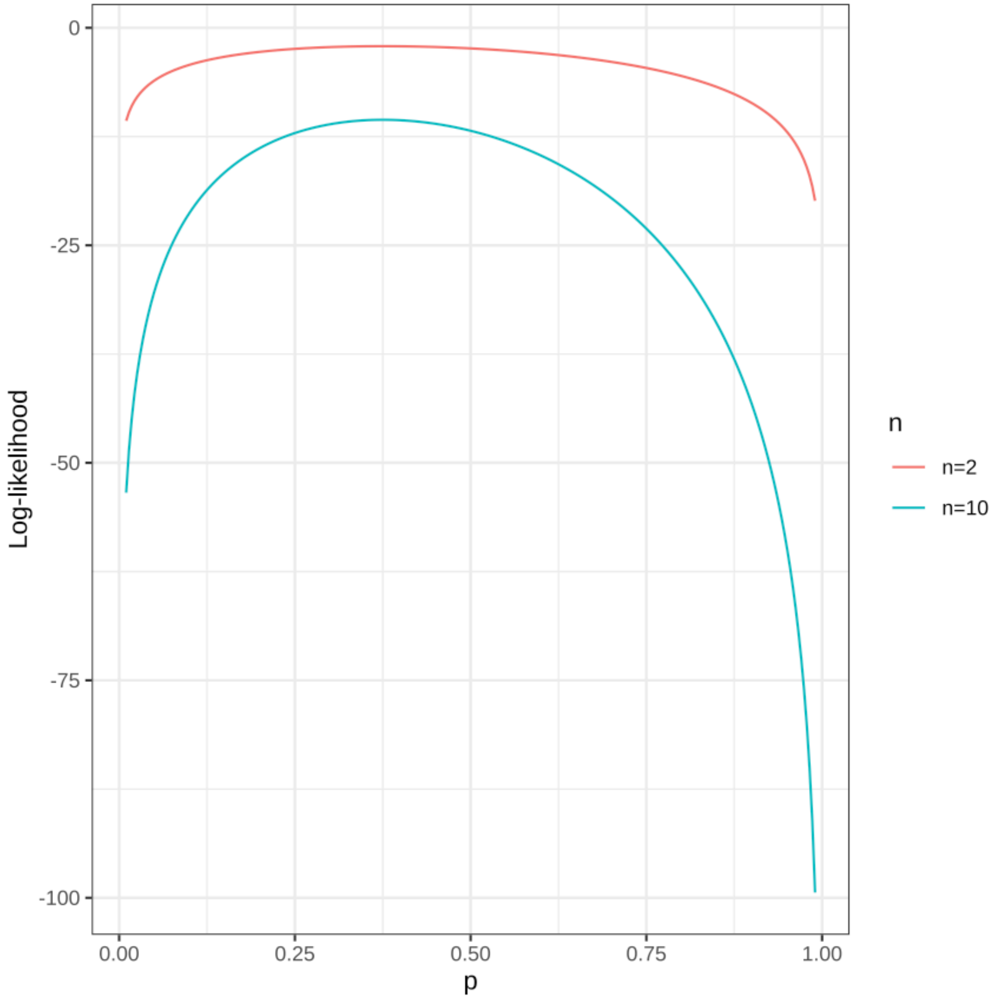
Functions in R

```r
library(maxLik)
llfun <- function(par, x){
    p <- dbinom(x, 4, par[1])
    sum(log(p))
}
out <- maxLik(llfun, start=.5, x=c(1,2,1,2,1,2,1,2,1,2))
summary(out)
```

```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 2 iterations
## Return code 1: gradient close to zero (gradtol)
## Log-Likelihood: -10.57226
## 1  free parameters
## Estimates:
##      Estimate Std. error t value  Pr(> t)
## [1,]  0.37500    0.07655   4.899 9.63e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

Log-Likelihood Functions

We could look at the two likelihood functions, one from two points and one from 10 points.



Back to the Experimental Data

```
library(googlesheets)
library(reshape)
gs <- gs_title("9591 Experiment")
g <- gs_read(gs)
g.long <- melt(g, id="Trial")
out1 <- maxLik(llfun, start=.5, x=g.long$value)
summary(out)
```

By Experimenter

```
llfun <- function(par, x, group){
    p <- dbinom(x, 4, par[group])
    sum(log(p))
}
g <- as.numeric(g.long$variable)

out <- maxLik(llfun, start=rep(.5, 10), x=g.long$value, group=g)
summary(out)
```

Testing the Two Models: Likelihood Ratio Test

If two models are nested, then we can use a likelihood ratio test to figure out whether the bigger one is "better".

$$s = -2\left(LL(M_{\text{small}}) - LL(M_{\text{big}})\right)$$

Under H_0 : Both Models Same, $s \sim \chi^2_{k_{\text{big}} - k_{\text{small}}}$.

```
lr <- -2*(logLik(out1) - logLik(out))
pchisq(lr, 9)
```

Properties of MLEs

- Consistent - As sample size increases the probability that the MLE differs from the true parameter by an arbitrarily small amount is zero
- Asymptotically efficient which means that the MLE's variance is the smallest among all possible consistent estimators.
- Asymptotically normally distributed

All of these are asymptotic properties, so they describe the properties of the estimators as n is close to ∞. They may behave differently in small samples. We will discuss this a bit later on

Linear Models: OLS

```r
data(Prestige, package="carData")
Prestige <- na.omit(Prestige)
summary(lm(prestige ~ education, data=Prestige))
```

```
##
## Call:
## lm(formula = prestige ~ education, data = Prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.605  -6.151   0.366   6.565  17.540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.8409     3.5285  -3.072  0.00276 **
## education     5.3884     0.3168  17.006  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.578 on 96 degrees of freedom
## Multiple R-squared:  0.7508,    Adjusted R-squared:  0.7482
## F-statistic: 289.2 on 1 and 96 DF,  p-value: < 2.2e-16
```

```r
X <- cbind(1, Prestige$education)
y <- matrix(Prestige$prestige, ncol=1)
llfun <- function(par, X, y, ...){
   n <- nrow(X)
   b <- par[1:2]
   yhat <- X %*% b
   sig2 <- par[3]
   sum(dnorm(y-yhat, 0, sqrt(exp(sig2)), log=TRUE))
}
lm_mle <- maxLik(llfun, X=X, y=y, start=c(0,0,1), tol=1E-15)
summary(lm_mle)
```

```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 19 iterations
## Return code 8: successive function values within relative tolerance lin
## Log-Likelihood: -348.6709
## 3  free parameters
## Estimates:
##      Estimate Std. error t value  Pr(> t)
## [1,] -10.8390     2.9643  -3.657 0.000256 ***
## [2,]   5.3882     0.2670  20.181  < 2e-16 ***
## [3,]   4.2777     0.1438  29.747  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

Recap

1. What is MLE?
    - $L(\theta|\text{Data}) \propto f(\text{Data}|\theta)$