**Maximum likelihood estimation**

The statistic that estimates μ (population mean) is

.

This was introduced by intuition. It made sense to use the “sample mean” to estimate the “population mean”. Also, we showed that E() = μ so that  was an unbiased estimator.

The statistic that estimates σ2 (population variance) is

.

This was introduced somewhat by intuition and justified by E(S2) = σ2.

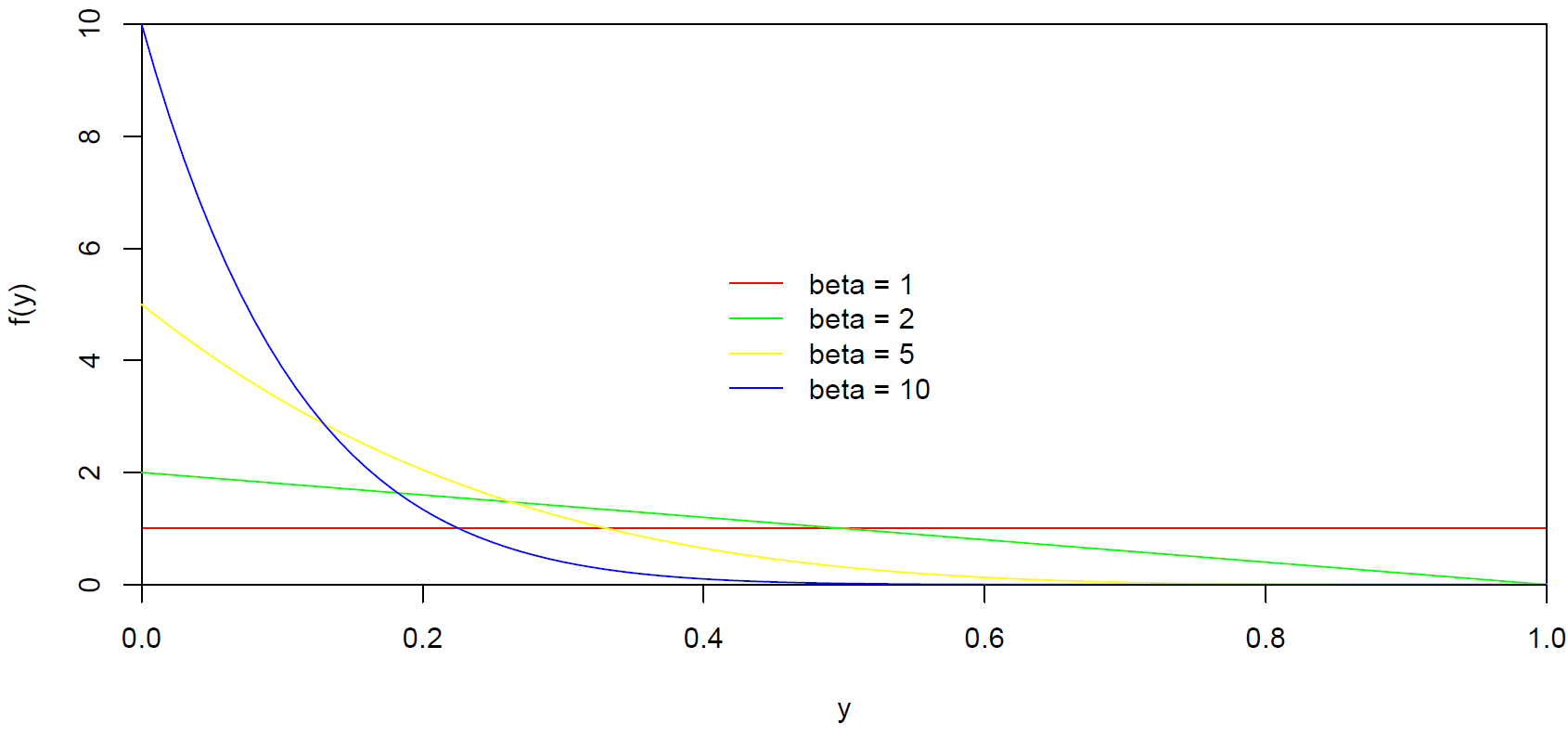
Suppose Y1, …, Yn are a random sample from a population characterized by the PDF



where β > 0.

This is a one-parameter beta PDF. Why is it important to have good estimate for β?

Let’s examine plots of the PDF (see beta.R).



Notice the plots change due to the particular value of β!

What is a statistic that estimates β? In other words, how should we use the random sample of Y1, …, Yn to estimate β?



In this case and many others like it, there are not necessarily any intuitive statistics to estimate the parameters. Thus, other methods need to be developed to determine what one should use as the parameter estimate!

The most common method used to find these estimators is maximum likelihood estimation. This estimation method will be described next in the context of an example.

Example: Field goal kicking (LikelihoodFunction.R, LikelihoodFunction.ipynb)

The success or failure of a field goal in football can be modeled with a Bernoulli PDF that has a probability of success parameter denoted by π. Define Yi = 0 for a failure and Yi = 1 as a success for the ith field goal. The PDF for Yi is:

 for yi = 0 or 1

Suppose we would like to estimate π for a 40 yard field goal. Let Y1, …, Yn denote a random sample of field goal results at 40 yards. The joint PDF for Y1, …, Yn is





Notice the use of product notation which is just like summation notation except items are multiplied. For example, .

Because our goal is to estimate π given our observed data, we re-write the joint PDF as the likelihood function:

.

Suppose = w = 4 successes are observed out of n = 10 trials. Given this observed information, we would like to find the corresponding parameter value for π that produces the largest probability of obtaining this particular sample. The following table can be formed to help find this parameter value:

> w <- 4 w is the sum of the y\_i’s

> n <- 10

> pi <- c(0.2, 0.3, 0.35, 0.39, 0.4, 0.41, 0.5)

> Lik <- pi^w\*(1-pi)^(n-w)

> data.frame(pi, Lik)

pi Lik

1 0.20 0.0004194304

2 0.30 0.0009529569

3 0.35 0.0011317547

4 0.39 0.0011918935

5 0.40 0.0011943936

6 0.41 0.0011919211

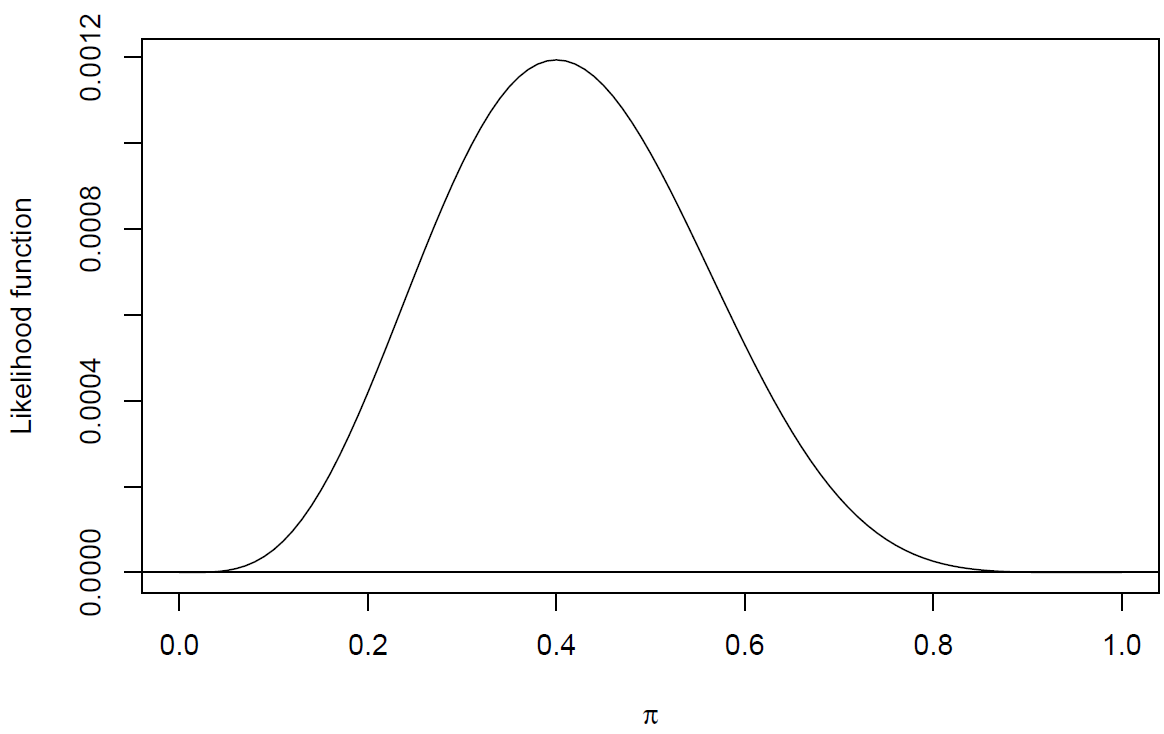
7 0.50 0.0009765625

> curve(expr = x^w\*(1-x)^(n-w), from = 0, to = 1,

xlab = expression(pi), ylab = "Likelihood

function")

> abline(h = 0)



Note that π = 0.4 is the “most plausible” value of π for the observed data because this maximizes the likelihood function. Therefore,  is the maximum likelihood estimate (MLE).

The MLE can be found using calculus:

1. Find the natural log of the likelihood function, 
2. Take the derivative of  with respect to π.
3. Set the derivative equal to 0 and solve for π to find the MLE. The solution is the maximum of  provided certain “regularity” conditions hold (see Mood et al. 1974).

The maximum for  occurs at the same location as for  because the log transformation is a monotone increasing transformation. The log transformation is used because it often results in mathematical simplifications.

For the field goal example:





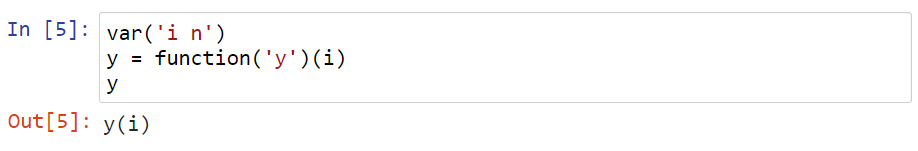


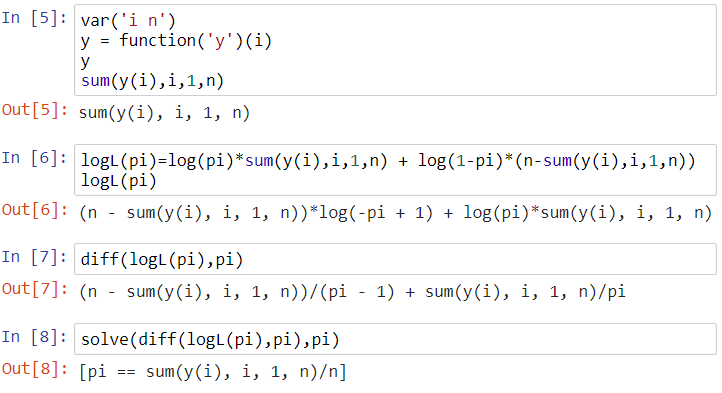


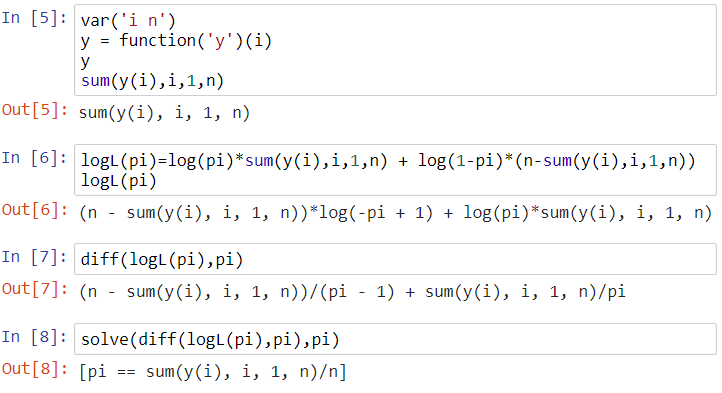


Therefore, the maximum likelihood estimate of π is the proportion of field goals made. To avoid confusion between a parameter and a statistic, one denotes the estimator as  = /n.

Sage:







This estimator can also be derived from a binomial PDF point of view. Suppose W is the number of successes out of n trials. Then

 for w = 0, …, n

Thus, we can view this as observing only one random variable w denoting the number of successes. The likelihood function then becomes



The MLE is found from



When one takes the derivation with respect to π, . This leads to the same steps as the previous case, so that  = w/n.

Important notes:

1. Due to properties of the PDFs we have not examined, usually we can just take the solution of



to be the MLE for a generic parameter θ.

1. What happens for two or more parameters? The MLEs becomes a little more difficult to find! In the case of two parameters, say θ1 and θ2, we will need to find the solutions to

 and



simultaneously. Thus, we have two equations and two unknowns. The upcoming normal PDF example gives one simple example.

1. There are often times when a solution cannot be written out as  = \_\_\_. Instead, iterative numerical methods (like a Newton-Raphson method) need to be used to find the MLE. For example, this is the case with finding the MLEs of α and β from a gamma PDF. A simpler type of iterative method is just to plug in multiple values of α and β until one finds the values that maximize the likelihood function.

Example: MLEs for μ and σ from a normal distribution

Let Y1, … ,Yn be a random sample from a normal PDF where each random variable has E(Yi) = μ and Var(Yi) = σ2. Find the MLE of μ and σ2. Remember that the normal distribution is:



Then the likelihood function is



Taking the log of the likelihood function produces



To find the maximum likelihood estimate of μ, take the derivative with respect to μ and set equal to 0.



Solving for μ produces:



To find the maximum likelihood estimate of σ, take the derivative with respect to σ and set equal to 0.



Solving for σ2 produces:



Notes:

1. Solving the two equations for two unknowns was a little easier here than in most other situations because the partial derivative with respect to μ was only a function of μ.
2. The MLE for σ2 is not the unbiased estimator ! Thus, the MLE here is referred to as a biased estimator. However, notice that as n increases, the difference between the MLE and s2 will become quite small.