**Maximum Likelihood Estimation**

Table of Contents

[Estimators and Estimates 4](#_Toc64283152)

[Likelihood Functions 5](#_Toc64283153)

[Log Likelihood 10](#_Toc64283154)

[Bayesian Estimates 17](#_Toc64283155)

In probability theory, we learnt how to take an experiment and create a probability model from it. Using this probability model, we were able to tell what the probabilities of generating certain pieces of data were.

In inferential statistics, we do the opposite. We take some data and try to fit that data onto known probability models. In some cases, the probability model may be given to us and in some cases, we might have to make a hypothesis about what the probability model should be based on things like the histogram of the data. It is even possible that we may never really know for certain which probability model is accurate.

Every probability model has some parameters. In probability theory, we knew these parameters, but in inferential statistics, our goal is to make an estimate about these parameters. Maximum Likelihood Estimation is one of the methods that helps us do this.

Generally, the parameter is denoted as , or if there are more than one parameter, and . For the case with more than one parameter, a vector may also be used, .

To further understand why we need to find the parameters, consider an example.

Say there is an election going on with two candidates, and . We can go around asking people which candidate they voted for. Every piece of data that we collect is a different random variable, . Each of the random variables can have one of two values, with the values denoted as .

The fact that we know that there are only two possible values means that each of the random variables is a Bernoulli distribution, and the population as a whole is also a Bernoulli distribution. Thus,

where is perhaps the probability that candidate wins. Thus, just by looking at the situation and the possible pieces of data we immediately found the probability model. That was not the difficult part. To fully describe the probability model though, we need to use the data we collected to find the parameters.

## Estimators and Estimates

With regards to this, there are two things that we will be needing, an estimator and an estimate.

An estimator is a function that works with the data given to us. Thus, an estimator is a statistic. For example, we saw a formula, . Here, can be considered an estimator.

An estimate on the other hand, is the actual value that we observe. If we collect data one time, we might get some set of values for the random variables , while on a second run, we might get a different set of values, say . The result of the estimator for one particular set of values is the estimate. Using the previous example again, would give us an estimate for the sample mean.

## Likelihood Functions

Now we will learn how to use the Maximum Likelihood Estimation (MLE) process to find the values of the probability model parameters. We will do this by using examples. As we will see, the process is pretty simple. In most cases, we will just end up using averages.

Example

Say we toss a coin times and observe heads. Let be the probability of getting a head in a single toss. Now, we must use MLE to estimate .

First, let’s define the problem formally.

Experiment: Tossing a coin and counting the number of heads.

Number of tosses .

Data: heads

Parameters: (unknown)

Likelihood/Likelihood Function:

We have seen the definition of likelihood before, which discussing the statistical approach to Bayes’ theorem. At the time, we saw that

We identified the term , the probability of getting a particular piece of data, given that a certain hypothesis is true, as the likelihood.

In this experiment, we can have a random variable, , which counts the number of heads. Thus, . is binomially distributed, since we are counting the number of ‘successes’ in a certain number of attempts, i.e. . Since is binomially distributed, we can have an equation for the probability of getting heads.

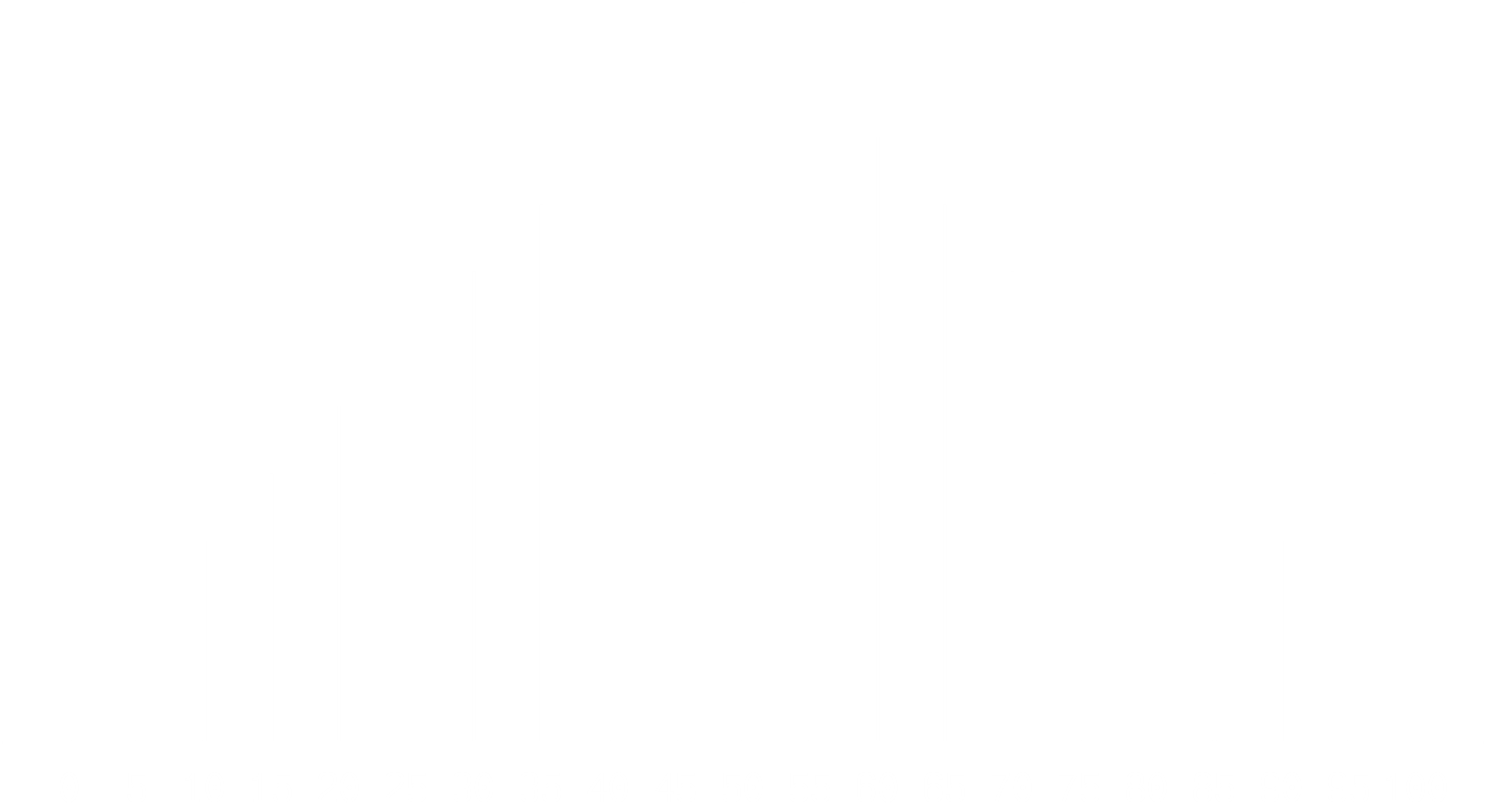
This equation tells us the probability of getting heads. It is not a fixed value, since it is conditioned on the value of . Thus, we can write this equation as

This is the likelihood function. The likelihood function tells us the probability of observing certain data given a parameter value or hypothesis.

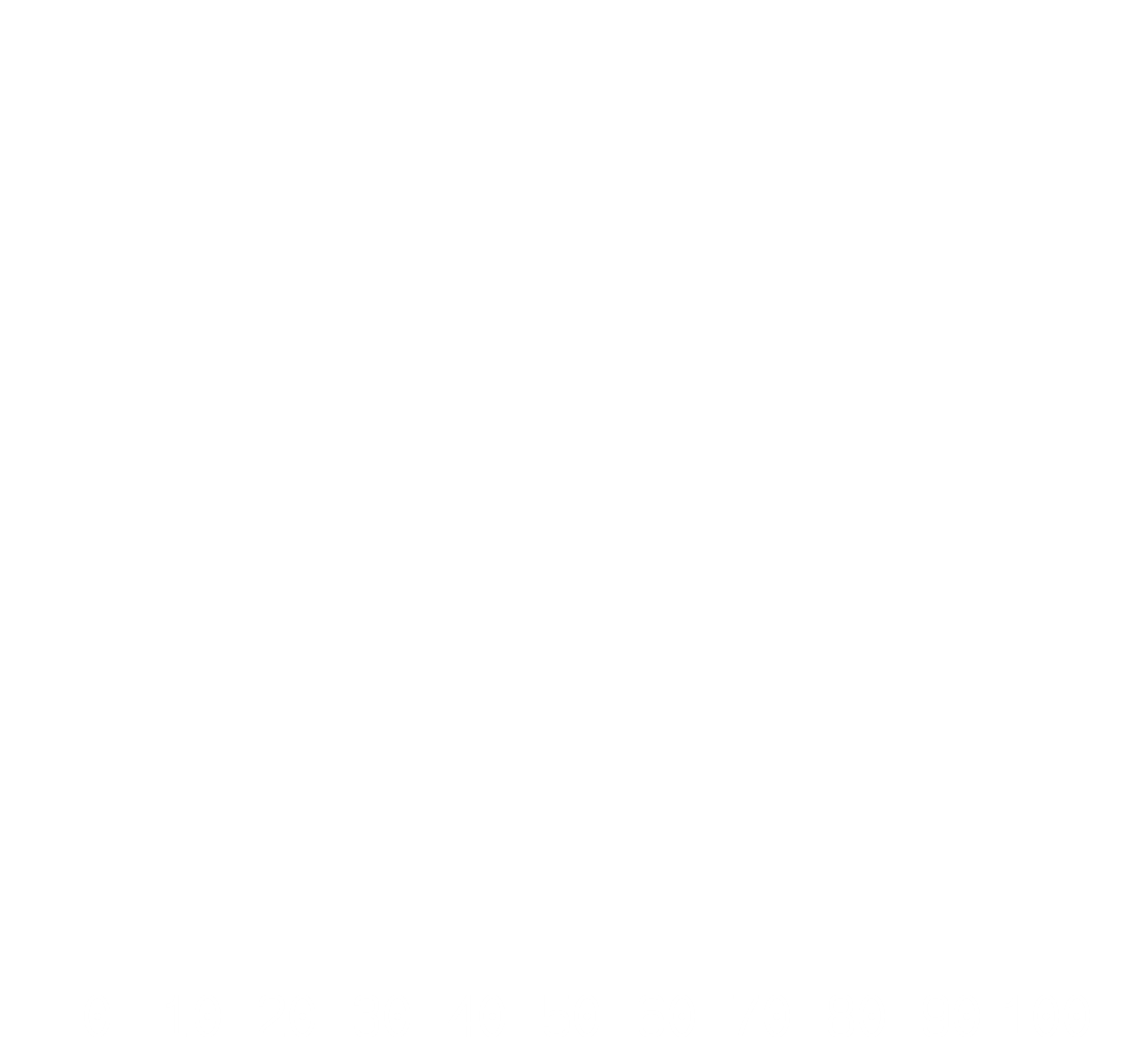
Now, what is the Maximum Likelihood Estimate. We already have a likelihood function which we just defined. For different values of , we will get different values for the function. The MLE gives us the value of that maximizes the result of the likelihood function.

When we define a probability model and its parameters for a set of data, we will judge whether or not we were correct based on how often our model keeps holding. We do not know for certain that we are right. We are making an estimation. So, the best thing for us to do to is to maximize the probability that we are right. That is what we are doing here.

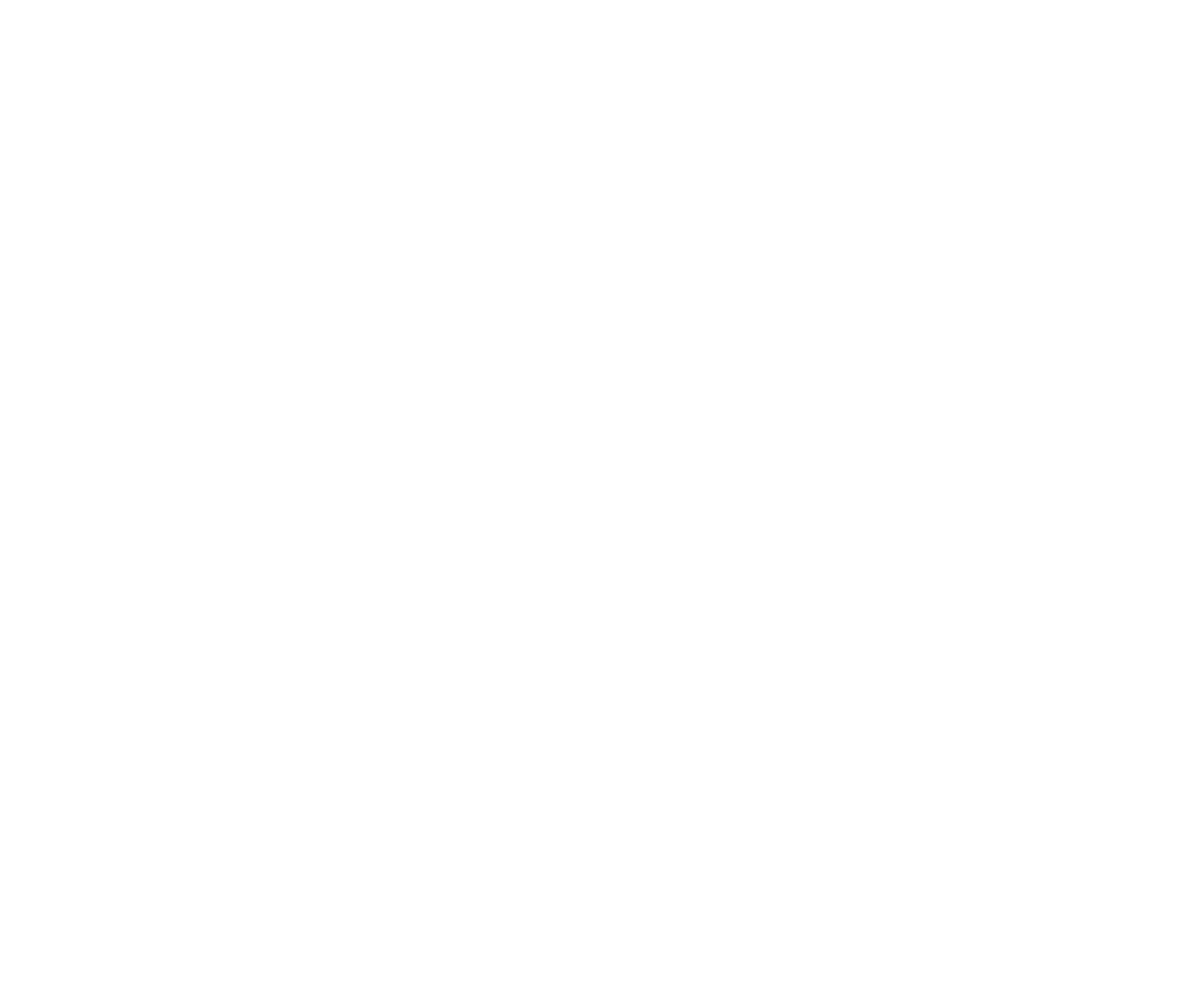
Say . If we plot the graph for this, it will look like this:



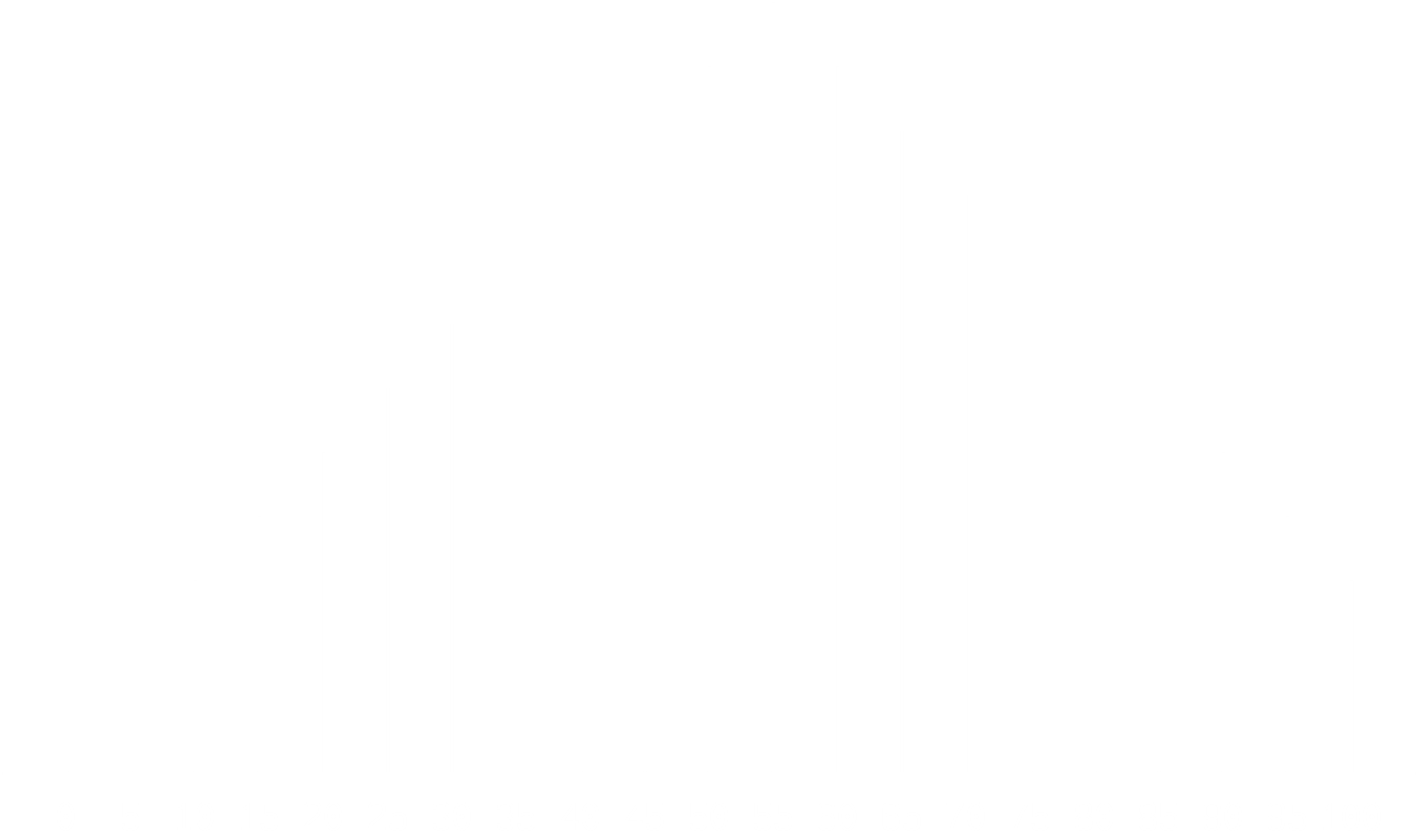
Now let’s change and make it . Then, the graph becomes:



Again, say .



And finally, say .



As we change the value of , the probability of getting heads is changing. Our actual data tells us that we got heads. The entire point of MLE is to change the value of and find a point where the probability of getting the given data is maximized.

Of course, this does not mean that this genuinely is the probability. It is entirely possible that in reality and we still got heads. However, it is most likely that, since we got heads, , since that is the value that maximizes the chances of getting heads again if we repeat the experiment.

What we did above was not the mathematical way of finding of course. This was just an example. We can hardly go around randomly selecting values, creating graphs and seeing what happens all the time.

To find the maximum value of the likelihood function, we find the point where the first derivative is .

At times, instead of writing , we write , to indicate that this is an estimate and not the actual value of .

Another observation we can make is that is the ratio of the number of ‘successes’ to the total number of attempts.

## Log Likelihood

The log likelihood allows us to calculate the parameter value for maximizing the likelihood function easily. This is because products become summations when taken in logarithmic form.

Exponential Distribution Example

Population: Lifetime of light bulbs

Common Distribution: Exponential

Parameter: , the average failure rate

Data: Lifetimes of lightbulbs, , , , and

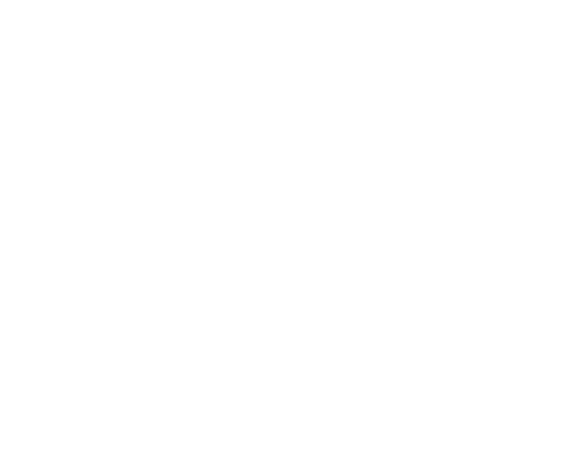
Sample Values: , , , ,

Random Sample: , , , ,

We want to use MLE to find . For this, we need the likelihood function, which we know can be found from the .

We know that for all . The value of represents the lifetime of the -th bulb. This must be a continuous random variable, since the bulb could fail at any given time.

Since all the random variables in our random sample are continuous random variables, we cannot find , since for continuous random variables. As such, we need to work with the PDF. We know that if we take , where is a very small range, we can say that this is the probability that has a value near .



As such, the likelihood function can be written as

The log likelihood function is given by

Digging a little deeper,

If we take the sum of the values from the random sample and find the average, we will also find that we get . Thus,

The point of using MLE is to find a value of such that the probability of getting the same set of values that we had in our random sample was the highest. For this value of , if we collect random samples of values repeatedly, we will see that this particular set occurs with the highest frequency.

This is useful. If we want to continue the experiment and we want to get data that is similar to the sample data we had, we can set the value of . We will have different types of data, but mostly, we will get data that is similar to our sample data.

Normal Population Example

Say we are considering a population of the delays of data packets, and we know that the common distribution is normal. Thus, the parameters are and .

We take a random sample, with the values . We want to find the MLE for and .

Thus, is just the arithmetic average.

Example

Say we are having a chess tournament with both male and female players. In a particular game, the probability of a male player playing is and the probability of a female player playing is .

Thus, in a given game, the probability of both players being male is , the probability of both players being female is and the probability of the players being of opposite genders is .

Say, in games, we find games where both players are male, games where the players are of opposite genders and games where both players are female.

We want to find the MLE of .

The population in this case is the games being player.

The common distribution is an empirical distribution, meaning the probabilities are just .

The data is the values of , and .

The likelihood function will be .

We are trying to find the value of that maximizes the probability of getting a particular set of values such that , and . This is like a multinomial distribution, which, for two variables, would have reduced to a binomial distribution.

## Bayesian Estimates

There are actually two schools of thought. The Frequentist school of thought works with the frequencies of data to estimate parameters, which is what we saw till now with MLE. The Bayesian school of thought on the other hand, relies on the Bayes theorem to estimate parameters.

The Bayes theorem tells us

We have previously defined as the prior probability, as the likelihood and as the posterior probability, and we found that

Using the Bayesian theorem, we want to maximize this value, i.e. we want to maximize , the product of . The which maximizes this value is the one we accept as the estimated value of .

Comparing this to the method used by the Frequentist school of thought, they want to maximize the likelihood, . This is because they assume that the prior probabilities for all the hypotheses are equal, meaning the value of is the same for all , so they ignore it.

Another debate between the two schools is regarding the parameter itself. Is the parameter actually a constant value that we are finding? That is what the Frequentist school of thought says. Or is the parameter a value that depends on the random sample we collect and thus varies depending on the sample. In the latter case, this would make the parameter itself a random variable. This is the notion proposed by the Bayesian school of thought. This difference contributes to the first difference we saw, where the probabilities of were considered only by the Bayesian school of thought.

You would imagine that these differences would give us different results, but if we have a large amount of data of good quality, we will actually get the same values.

The problems start with smaller values. Consider that we have a coin that we toss just times and somehow all tosses give us heads. By MLE, we can take the average and find that . Thus, MLE is telling us that this is a double headed coin. From the Bayesian method though, we would need to take , the prior probability, into account. Just from the concept that the parameter has a probability, we can tell that the value of cannot definitively be claimed to be . Before the first toss, the prior probability will have one value. This value will be the subjective probability as determined by some expert on tossing coins. Due to the results of the first toss, the prior probability will change before the second toss and so on.