Maximum Likelihood Estimation
by Addie Andromeda Evans
San Francisco State University
BIO 710 Advanced Biometry Spring 2008

Estimation Methods  Estimation of parameters is a fundamental problem in data analysis. This paper is about maximum likelihood estimation, which is a method that finds the most likely value for the parameter based on the data set collected. A handful of estimation methods existed before maximum likelihood, such as least squares, method of moments and bayesian estimation. This paper will discuss the development of maximum likelihood estimation, the mathematical theory and application of the method, as well as its relationship to other methods of estimation. A basic knowledge of statistics, probability theory and calculus is assumed.

Earlier Methods of Estimation  Estimation is the process of determining approximate values for parameters of different populations or events. How well the parameter is approximated can depend on the method, the type of data and other factors.

Gauss was the first to document the method of least squares, around 1794. This method tests different values of parameters in order to find the best fit model for the given data set. However, least squares is only as robust as the data points are close to the model and thus outliers can cause a least squares estimate to be outside the range of desired accuracy.

The method of moments is another way to estimate parameters. The 1st moment is defined to be the mean, and the 2nd moment the variance. The 3rd moment is the skewness and the 4th moment is the kurtosis. In complex models, with more than one parameter, it can be difficult to solve for these moments directly, and so moment generating functions were developed using sophisticated analysis. These moment generating functions can also be used to estimate their respective moments.

Bayesian estimation is based on Bayes’ Theorem for conditional probability. Bayesian analysis starts with little to no information about the parameter to be estimated. Any data collected can be used to adjust the function of the parameter, thereby improving the estimation of the parameter. This process of refinement can continue as new data is collected until a satisfactory estimate is found.
Evolution of Maximum Likelihood Estimation  It was none other than R. A. Fisher who developed maximum likelihood estimation. Fisher based his work on that of Karl Pearson, who promoted several estimation methods, in particular the method of moments. While Fisher agreed with Pearson that the method of moments is better than least squares, Fisher had an idea for an even better method. It took many years for him to fully conceptualize his method, which ended up with the name maximum likelihood estimation.

In 1912, when he was a third year undergraduate student, Fisher published a paper called ”Absolute criterion for fitting frequency curves.” The concepts in this paper were based on the principle of inverse probability, which Fisher later discarded. (If any method can be considered comparable to inverse probability it is Bayesian estimation.) Because Fisher was convinced that he had an idea for the superior method of estimation, criticism of his idea only fueled his pursuit of the precise definition. In the end, his debates with other statisticians resulted in the the creation of many statistical terms we use today, including the word ”estimation” itself and even ”statistics”. Finally, Fisher defined the difference between probability and likelihood and put his final touches on maximum likelihood estimation in 1922.

The distinction between probability and likelihood is indeed subtle. As this paper continues, the distinction will unfold and become clearer to the reader.

Mathematical Theory of Maximum Likelihood Estimation  Suppose we have flipped a coin three times and observed a sequence of events HHT. We know that flipping a coin is modeled by the binomial probability density function,

\[ P(k; n, p) = \binom{n}{k} p^k (1 - p)^{n-k}, \]

where we have \( k \) successes out of \( n \) Bernoulli trials and we define the random variable \( K \) as either ”heads” or ”not heads” on each toss. The parameter of this model is \( p \), the probability of flipping a coin and getting heads. So we define

\[ P(K = 1) = p \]
\[ P(K = 0) = (1 - p) \]

For our sequence HHT \( K_1 = 1, K_2 = 1, \) and \( K_3 = 0, \) and since these trials are independent, we get

\[ P(K_1 = 1 \cap K_2 = 1 \cap K_3 = 0) = P(K_1 = 1) \cdot P(K_2 = 1) \cdot P(K_3 = 0) \]

2
which means
\[ P(K_1 = 1 \cap K_2 = 1 \cap K_3 = 0) = p^2(1 - p) \]

Based on this data set, a good estimate for the mean of the binomial model is \( \frac{2}{3} \) since
\[ P(k) = \binom{n}{k} \left( \frac{2}{3} \right)^k \left( \frac{1}{3} \right)^{n-k} \]
is much more likely to predict HHT than
\[ P(k) = \binom{n}{k} \left( \frac{1}{2} \right)^k \left( \frac{1}{2} \right)^{n-k} \]
which is what we might have expected since most coins have a probability of getting heads of one half. But in this case, based on our known data, we expect to get heads two thirds of the time on future tosses with the same coin.

**Definition of a Likelihood Function**  We saw in the example above that we can construct a function based on the probabilities of the independent observations. We used this function to estimate the parameter. This function we created is called a likelihood function and the formal definition is:

**Definition 1** If \( y_1, y_2, \ldots, y_n \) is a random sample of size \( n \) from a discrete or continuous pdf, \( f_Y(y_i; \theta) \), where \( \theta \) is an unknown parameter then the likelihood function is written
\[ L(\theta) = \prod_{i=1}^{n} f_Y(y_i; \theta) \]

**The Definition of Maximum Likelihood Estimation**  For a likelihood function \( L(\theta) \), where \( \theta \) is an unknown parameter. Let \( \theta_e \) be a value of the parameter such that \( L(\theta_e) \geq L(\theta) \) for all possible values of \( \theta \). Then \( \theta_e \) is called a "maximum likelihood estimate" for \( \theta \). [?]

In calculus, the extreme value theorem states that if a real-valued function \( f \) is continuous on a closed, bounded interval \([a,b]\), then \( f \) has a maximum and minimum value at some point on the interval. [?] That is, there exist some numbers \( x_m \) and \( x_M \) such that:
\[ f(x_m) \leq f(x) \leq f(x_M) \text{ for all } x \in [a,b] \]
The extreme value theorem allows you to find the maximum and minimum by taking the derivative of the function and setting it equal to zero. This is an optimization technique that makes use of the fact that the slope of the tangent line of the maximum and minimum is 0. Since the function decreases in value after the maximum, then the function will have an negative second derivative at the maximum. Since the function increases after the minimum, then the function will have a positive second derivative at the minimum. This fact is used to determine whether the estimates obtained by setting the first derivative equal to 0 are maximums or minimums. Thus $\theta_{\text{max}}$ must have the properties such that $L'(\theta_{\text{max}}) = 0$ and $L''(\theta_{\text{max}}) < 0$.

**Estimating Population Size of Seals**  
This example is inspired by a study of the population size of harbour seals in the Dutch Wadden Sea. Suppose we want to estimate the size of a certain population of seals. We design a mark-recapture experiment where we radio tag two seals so we can detect them later. Then, one day we go out and count the number of seals on the haul-out. Also, we determine how many of the tagged seals are on the haul-out.

The proportion $p$ of the seal population on the haul out is the number of seals on the haul out divided by the population size $N$. Suppose we counted 500 seals on the beach. Then $p = \frac{500}{N}$. This is the probability that any given seal will be on the haul-out, and the probability that any given seal is not on the haul-out is $\left(1 - \frac{500}{N}\right)$.

Suppose we do not detect either tagged seal on the haul-out. Then the probability of these two observations is the product of their individual probabilities. Thus the likelihood function for seeing no tagged seals is

$$L_0(N) = \left(1 - \frac{500}{N}\right) \cdot \left(1 - \frac{500}{N}\right).$$

Suppose we detect the first tagged seal but not the second tagged seal or that we detect the second and not the first. Since there are two possibilities for detecting one tagged seal, then we sum the probabilities of each possibility. Thus the likelihood function for seeing one tagged seal is

$$L_1(N) = 2 \cdot \frac{500}{N} \cdot \left(1 - \frac{500}{N}\right).$$

Now suppose we detect both tagged seals on the haul-out. Then the likelihood function is

$$L_2(N) = \left(\frac{500}{N}\right) \cdot \left(\frac{500}{N}\right).$$

4
To get an idea about how the likelihood functions are different, we need to look at the graph of each.

![Graph of likelihood functions](image)

Notice that for the 0 seals, the likelihood function has a minimum at $N = 500$. Since we know there are at least 500 seals on the beach, we consider values to the right of 500. We can not see on the graph, but this function never decreases after $N = 500$ and so there is no maximum on the interval $[500, \infty)$. Thus we would interpret this as meaning that the number of seals is infinite. In reality, a study like this accounts for radio tag failure and this scenario would be interpreted as 100% radio tag failure, with no determinate population size estimate. If we take the derivative of the likelihood function

$$
\frac{dL_0}{dN} = \frac{1000N - 500,000}{N^3} \Rightarrow N_e = 500 \text{ seals}
$$
we see that we get an estimated 500 seals. If we were to take the second derivative, we would get a positive value at \( N = 500 \), since this is actually gives a minimum value of \( L(N) \), as we saw on the graph.

The likelihood function for 1 seal detected is quite nice since you can see that it’s maximum is at \( n = 1000 \). Thus we determine this to be our maximum likelihood estimate. Setting the derivative of \( L(N) \) equal to zero

\[
\frac{dL_1}{dN} = \frac{-1000N^2 + 1,000,000N}{N^4} \Rightarrow N_{\text{max}} = 1000 \text{ seals}
\]

we also get a maximum likelihood estimate of 1000 seals.

Finally, the likelihood function for detecting 2 seals has its own interesting twist. The function has a vertical asymptote at \( N = 0 \) and is strictly decreasing. Since \( N = 500 \) is our lower bound of the population size, this is our maximum likelihood estimate on our interval of possible values. Taking the derivative does not tell us any new information, but could be used without the graph to determine the same result:

\[
\frac{dL_2}{dN} = \frac{-500,000}{N^3} \Rightarrow N_{\text{max}} = 500 \text{ seals}
\]

Since we have three different scenarios, let us take this opportunity to compare their likelihood with their probability. Notice that the right most column is the sums of the probabilities of all the different scenarios. These all equal 1 as they should. Notice that the bottom row is the sums of the likelihood for a few example values of the parameter \( N \). These do not add up to 1.

<table>
<thead>
<tr>
<th>Likelihood versus Probability</th>
<th>N</th>
<th>0 tagged seals</th>
<th>1 tagged seal</th>
<th>2 tagged seals</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>500 seals</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>750 seals</td>
<td>0.11</td>
<td>0.44</td>
<td>0.44</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1000 seals</td>
<td>0.25</td>
<td>0.50</td>
<td>0.25</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1250 seals</td>
<td>0.36</td>
<td>0.48</td>
<td>0.16</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Likelihood</td>
<td>0.72</td>
<td>1.42</td>
<td>0.85</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

There are infinitely many estimates for any given parameter. The sum of the likelihood values for all the possible estimates need not be equal to 1 since likelihood is not the same as probability.
**Estimates and Estimators**  For a PDF in the geometric family, we have the likelihood function

\[ L(p) = \prod_{i=1}^{n} (1 - p)^{k_i - 1}p \]

Suppose we have no data to input to find \( p_e \). We can still solve for \( p_{\text{max}} \) to use later, once we collect some data, but until we collect some data, we call it \( \hat{p} \) and say that it is an estimator of the parameter \( p \).

**Finding the Maximum Likelihood Estimator**  To find \( \hat{p} \), we first must transform that product

\[ L(p) = \prod_{i=1}^{n} (1 - p)^{k_i - 1}p = (1 - p)^{\sum_{i=1}^{n} -np^n} \]

so it easier to take the derivative. And since most models are exponential, it is often easier to take the natural log before differentiating.

\[ \ln L(p) = \sum_{i=1}^{n} k_i \cdot \ln(1 - p) + n \ln p \]

Setting the derivative equal to 0 gives

\[ p \left( n - \sum_{i=1}^{n} k_i \right) + (1 - p)n = 0 \]

which tells us that

\[ p_{\text{max}} = \frac{n}{\sum_{i=1}^{n} k_i} \]

This is technically still called an estimate, where the \( k_i \)'s are considered to be unknown constants. However, we can then exchange the unknown constants for random variables \( K_1, K_2, ..., K_n \) and then we have

\[ \hat{p} = \frac{n}{\sum_{i=1}^{n} K_i} \]

the distinction between unknown constants and random variables is small, yet it important to keep track of what you are working with.
Properties of Estimators  The maximum likelihood estimator is just one of an infinite number of estimators. Perhaps, like Fisher we want to compare estimators to see if we can determine which one is best. Since we have made sure to define an estimator as a random variable, then they each have their own pdf, expected value, and variance which allow us to make comparisons.

While with a point estimate you have no way of knowing how precise it is, with estimators you can specify a confidence interval. Of course, the larger the sample size, the greater the precision of the estimator. The experimental design can incorporate the necessary sample size to provide the desired amount of precision. Additional properties are briefly described as follows:

Unbiasedness  If the distribution of the random variable is skewed by outliers, the estimates can be biased. Thus some \( \hat{\theta} \) will overestimate, and some will underestimate the true \( \theta \).

Efficiency  The estimator with the lowest variance is the most precise. For two unbiased estimators, the one with lower variance is preferable since it is the one that is more leptokurtic (peaked),

The Cramer-Rao Lower Bound  For any two unbiased estimators, it is impossible to compare their respective variance with the infinitely many other estimators. An important theorem says that any estimator can only have a variance that is so small, and this variance is the Cramer-Rao lower bound. If the variance for some \( \theta_e \) is equal to this lower bound, then there is no estimator that is more precise.

Sufficiency  Another triumph of Fisher, an estimator is determined to be sufficient if no other information from the data can be used to improve the estimator.

Consistency  An estimator is said to be consistent when the value of \( \hat{\theta} \) converges in probability to \( \theta \) as the sample size gets infinitely large.

Maximum Likelihood Estimation for More Than One Parameter  For families of probability models with more than one parameter, \( \theta_1, \theta_2, \ldots, \theta_k \), finding the maximum likelihood estimates for the \( \theta_i \)s can become quite complicated. If a closed form solution is possible, then the estimation requires solving a set of \( k \) simultaneous solutions. For \( k = 2 \) we have

\[
\frac{\partial \ln L(\theta_1, \theta_2)}{\partial \theta_1} = 0
\]
\[
\frac{\partial \ln L(\theta_1, \theta_2)}{\partial \theta_2} = 0
\]
We can take the partial derivative of function of two variables with respect to each variable individually, in order to solve for the variables. These kinds of partial differential equations can be difficult to solve, however not impossible, with two variables and simple models. The more parameters, the harder it is to solve.

Computational Applications of Maximum Likelihood Estimation  Maximum likelihood estimation is extremely useful with simple, normal data. But with more complicated models, maximum likelihood alone may not result in a closed form solution. Newton’s method can be used to find solutions when no closed form exists and it can converge quickly, especially if the initial value of the iteration is close to the actual solution. Here the importance of an efficient estimator is reinforced since the platykurtic nature of an inefficient estimator diminishes the ability of the algorithm to converge. However, with the rapid increase of computer speed, maximum likelihood estimation has become easier and has increased in popularity.

The Different Methods of Estimation  Since the method of moments uses a different function to individually estimate a given parameter, it is often easier and faster to calculate a method of moments estimate when more than one parameter is involved. These estimates can then be used as the initial value for the iterative process of finding the maximum likelihood estimates. Least squares corresponds to the maximum likelihood criterion if the experimental errors have a normal distribution. Bayesian estimates can be the same as maximum likelihood estimates, if the sample size is large enough. The different estimation methods seem to converge as the sample size increases, but with small sample sizes the properties of estimators are extremely useful for determining the best estimator.

Search for the Best Estimation Method  Its very name seems to imply that maximum likelihood is the best estimation method. Indeed, the theory of maximum likelihood estimation has a purity about it that is very appealing. Perhaps that is what made Fisher defend it so vehemently against other methods. Although, like all mathematical theories, maximum likelihood estimation is useful in some cases more than others. The fact that all methods of estimation are connected to each other in some way may explain why other statisticians couldn’t agree on the best method. In the end, the search for the best method at least produced a tool box of estimation methods for statisticians and other scientists to use as needed.
References


