Parameter Estimation

Introducing and Comparing Two Methods

Vishad Raj Onta

1 Introduction

Parameters are entities of great interest to statisticians, especially probabilists. For any given distribution, parameters control the behavior of a random variable, and define the shape, scale, and location of the probability density curve. However parameter values are not known and must be estimated from observed data. Only then can probability laws be fitted to the data. A scientist, for instance, would be directly interested in the parameters when a scientific theory suggests the form of a probability distribution. Further, a probability model may play a role in a complex and practical modeling situation such as that of pricing of utilities. Parameter estimation is also essential in the world of machine learning, where inferences are made from large amounts of data.

There are a few different methods of parameter estimation. In this document, I will introduce and develop two of them: the Method of Moments and the Method of Maximum Likelihood. I will go through steps and examples for each of the two methods. Given that there are a variety of ways to estimate parameters, an important question is that of optimal estimation: how do we decide which method is best? To answer this, I will discuss the methods of assessing uncertainty in estimates. In addition, I will apply the two aforementioned methods to a specific context and compare their performance. Lastly, I will look at how estimates reached through the maximum likelihood method behave asymptotically, that is, as the sample size approaches infinity.

The general approach I will be taking is by regarding observed data as realizations of random variables $X_1, X_2, \cdots, X_n$. Their joint distribution will depend on the unknown parameter θ.

2 Moments

I will start with the Method of Moments. Moments are mathematical quantities that provide insight into the properties of data, such as central tendency, spread, and shape. The kth moment of a probability law is defined as follows:

Definition 1. *[1]*

$$\mu_k = E(X^k)$$

The first moment, μ_1, is by definition the average or the expected value. The expected value measures the position of the center of the distribution. It is usually labeled by the letter μ (mu), and is defined as follows:

Definition 2. *[2] Given a random variable X whose outcomes have probabilities $p(i)$, and have the values $x_i, (i = 1, 2, \cdots, n)$, the expected value of the random variable X is defined to be*

$$E[X] = \sum_{i=1}^{n} x_i p(i)$$

While the expected value tells us about the center of the distribution, it does not reveal anything about its spread, which is how narrow or wide the distribution is. The second moment of a distribution, known as its variance and denoted by μ_2, does this. In particular, it tells you to what degree the elements deviate from the mean. A large variance means that the distribution is broad and the data is spread out, and a small variance means that the distribution is narrow and more data is clustered around the mean. Variance is defined as follows:

Definition 3. *[2] Given a random variable X whose outcomes have probabilities $p(i)$, and have the values $x_i, (i = 1, 2, \cdots, n)$, the variance of the random variable X is denoted by σ^2 and is defined to be*

$$V[X] = \sum_{i=1}^{n} (x_i - \mu)^2 p(i)$$

Note how in definition 1, the data points are taken as they are, while for definiton 2, each data point is normalized by subtracting from it the mean. While the expected value is known as a "raw" moment, or a moment about 0, the variance is known as a "central" moment, or a moment about the mean.

While moments of random variables may be found by using the above formulas, they can also be generated using the "moment-generating function"(mgf). Moment generating functions are defined as follows:

Definition 4. *[3] Let X be a random variable. The mgf of X is the real-valued function*

$$m(t) = E[e^{tX}]$$

defined for all t when this expectation exists.

To obtain moments of X from the mgf, we successively differentiate $m(t)$ and evaluate the function at $t = 0$. This is what that looks like for the Poisson distribution:

$$m(t) = E[e^{tX}] = \sum_{k=0}^{\infty} e^{tk} \frac{e^{-\lambda} \lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}$$

Differentiating gives

$$m'(t) = \lambda e^t e^{\lambda(e^t - 1)}$$

Evaluating at zero,

$$m'(0) = \lambda = E[X^1]$$

And the second derivative:

$$m''(t) = (\lambda e^t)(\lambda e^t + 1)(e^{\lambda(e^t - 1)})$$

With the second moment:

$$m''(0) = \lambda^2 + \lambda = E[X^2]$$

This gives $V[X] = E[X^2] - (E[X])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$. We see that the first two moments of the Poisson distribution–its expected value and variance–are both λ, as expected.

3 The Method of Moments

Moments can be used to estimate parameters through the Method of Moments. This method consists of three basic steps. First, the population moments are calculated and expressions for the moments are found in terms of the parameters. Next, these expressions are inverted to find formulas for the parameters in terms of the moments. Finally, the sample moments from the observed data are calculated and inserted into the expressions obtained in the previous step. This gives us equations that can be solved to get estimates of the unknown parameters.

Sample moments are defined as follows:

Definition 5. *[3] Let $X_1, \cdots, X_n$ be an independent and identically distributed sample from a probability distribution with moments that exist and are finite. The k^{th} sample moment is defined as*

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Let us apply the method of moments to the Poisson distribution. In section 2, we found using the mgf that the first moment for the Poisson distribution is $E(X) = \lambda$.

And, as the first sample moment is

$$\hat{\mu_1} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

the Method of Moments estimate of the parameter λ is simply $\hat{\lambda} = \overline{X}$, the sample mean.

As another example, we can use this method to estimate the parameters of the Normal Distribution. The first and second moments for the Normal distribution are [1]

$$\mu_1 = E[X^1] = \mu$$

$$\mu_2 = E[X^2] = \mu^2 + \sigma^2$$

Finding expressions for the parameters in terms of the moments,

$$\mu = \mu_1$$

$$\sigma^2 = \mu_2 - \mu_1^2$$

And finding the corresponding estimates of μ and σ^2 from the sample moments,

$$\hat{\mu} = \overline{X}$$

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

I will next demonstrate the efficacy of using the method of moments, by using simulations in R to consider the Negative Binomial distribution. The Negative Binomial distribution is a discrete probability distribution that is used to model situations where binary events occur with a certain probability of success. It is characterized by the parameters p, the probability of success for a single trial, and r, which can signify either a certain number of successes or a certain number of failures.

In the variation used by R, r signifies the number of failures before the rth success. For a random variable $X \sim \text{NegBin}(r, p)$, the first two theoretical moments are

$$E[X] = \frac{r(1-p)}{p}$$

and

$$E[X^2] = V[X] + E[X^2] = \frac{r(1-p)}{p^2} + \left(\frac{r(1-p)}{p}\right)^2$$

In R, I will first generate 100 instances of a random variable that follows the negative binomial distribution with $p = 0.3$ and $r = 4$

The first and second sample moments for these data are

$$\frac{1}{100}\sum_{i=1}^{100} X_i^1 = 9.45 \text{ and } \frac{1}{100}\sum_{i=1}^{100} X_i^2 = 121.57$$

Thus, the two method of moment equations are

$$\frac{r(1-p)}{p} = 9.45$$

And

$$\frac{r(1-p)}{p^2} + \left(\frac{r(1-p)}{p}\right)^2 = 121.57$$

Substituting the first equality into the second equation gives

$$121.57 = \frac{9.45}{p} + (9.45)^2$$

Solving and back substituting gives $p = 0.293$ and $r = 3.914$. We see that Method of Moment estimators are quite reasonable.

4 The Method of Maximum Likelihood

Another such method of parameter estimation based on observed data is the Method of Maximum Likelihood. The Method of Maximum Likelihood finds the value of the parameter that makes the observed data most likely. This is done by first writing down the likelihood function for the observed data, which is the probability of observing the data given the unknown parameter. Then, this function is maximized with respect to the to-be-estimated parameter, giving the Maximum Likelihood Estimator(MLE) of the parameter.

We start by examining the likelihood function, which is defined as follows:

Definition 6. *[1] Suppose that the random variables $X_1, \cdots, X_n$ have a joint density or frequency function $f(x_1, x_2, \cdots, x_n|\theta)$. Given observed values $X_i = x_i$, where $i = 1, \cdots, n$, the likelihood of θ as a function of $x_1, x_2, \cdots, x_n$ is defined as*

$$L(\theta) = f(x_1, x_2, \cdots, x_n|\theta)$$

Note that the joint density is being considered as a function of θ and not a function of the instances of X. The MLE of θ is the value of θ that maximizes the likelihood, which means that it makes the observed data most likely. If the random variables are independent and identically distributed, their joint density is the product of the marginal densities. That is,

$$L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

This function must be found and maximized. Often, the natural logarithm of this function is easier to maximize and thus used instead. We can do this without losing mathematical soundness because the logarithm is monotonic, that is, increasing on $(0, \infty)$. The log likelihood is as follows:

$$l(\theta) = \sum_{i=1}^{n} log[f(X_i|\theta)]$$

Let us look at an example of calculating the Maximum Likelihood Estimator for the Poisson distribution. If X follows a Poisson distribution, then its density function is

$$f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \ \ x = 0, 1, 2, \cdots$$

Assuming that the instances of the variable are independent and identically distributed, their joint frequency function is the product of the marginal frequency functions. That is,

$$L(\lambda) = f(x_1, x_2, ..., x_n|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}, \ \ \lambda > 0$$

And the log likelihood is

$$l(\lambda) = \log L(\lambda) = \sum_{i=1}^{n} \left(x_i \log \lambda - \lambda - \log(x_i!) \right), \;\; \lambda > 0$$

$$= \log \lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \log x_i!, \;\; \lambda > 0$$

Next, to maximize, we set its first derivative with respect to λ equal to zero and solve, finding

$$l'(\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n = 0$$

And hence the maximum likelihood estimate

$$\hat{\lambda} = \overline{X}$$

Note that in this case, the Method of Maximum Likelihood and the Method of Moments agree.

As another example, we can apply the Method of Maximum Likelihood to estimate the parameter λ in an Exponential distribution. We have the density function:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}, \lambda > 0$$

The likelihood function is given by

$$L(\lambda | x_1, x_2, ..., x_n) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} \;\; , \lambda > 0$$

And the log likelihood is given by

$$\ell(\lambda|_1, x_2, ..., x_n) = n \log(\lambda) - \lambda \sum_{i=1}^{n} x_i \;\; , \lambda > 0$$

Now, finding the first derivative

$$\frac{d\ell(\lambda | x_1, x_2, ..., x_n)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

and setting the result to zero

$$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

yields the MLE. for the exponential distribution

$$\hat{\lambda}_{MLE} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\overline{X}}$$

It is the reciprocal of the sample mean.

5 Assessing the Variability of Estimates

Having made estimates, we now turn to methods to assess their variability. The general idea is that estimates themselves, as functions of the observed data, are random variables as well and thus have a distribution. This distribution is referred to as the sampling distribution of the estimate. When presented with options of various parameters, it makes sense to choose the estimate who sampling distribution is most highly concentrated about the true parameter value.

Recall that in section 3, it was found that the Method of Moments estimate of λ for the Poisson distribution was equal to the arithmetic mean. This arithmetic mean is dependent on the observed data. So, the estimates will vary from sample to sample. As mentioned, the distribution of these estimates is referred to as its sampling distribution, and in the case of the Poisson distribution, can be derived theoretically.[2].

The model stipulated that the individual counts X_i of the random variable were independent and followed the Poisson distribution with true parameter λ_0. We found that the Method of Moments estimate, $\hat{\lambda}$, was equal to the arithmetic mean. Thus, if we denote $S = \sum X_i$, then $\hat{\lambda} = S/n$. We consider S/n to be a random variable with its distribution being the distribution of the estimate, that is the sampling distribution.

We can then use the fact that the sum of independent Poisson random variables is also Poisson to conclude that the distribution of S is Poisson$(n\lambda_0)$. We can write

$$P(\hat{\lambda} = v) = P(S = nv) = \frac{(n\lambda_0)^{nv} e^{-n\lambda_0}}{(nv!)}$$

And, using the expectation and variance of a Poisson distribution, we have

$$E[\hat{\lambda}] = \frac{1}{n} E[S] = \frac{1}{n} n\lambda_0 = \lambda_0$$

$$V[\hat{\lambda}] = \frac{1}{n^2} V[S] = \frac{1}{n^2} n\lambda_0 = \frac{\lambda_0}{n}$$

We note several things[1]. First, by the Central Limit Theorem, we know that if $n\lambda_0$ is large, then the distribution of S will be approximately normal. Thus, the distribution of $\hat{\lambda}$ is also approximately normal. Second, since the expectation of the estimate, $E(\hat{\theta})$, is the true parameter λ_0, we see that the sampling distribution is centered around λ_0. Third, since $V[\hat{\lambda}] = \frac{\lambda_0}{n}$, we see that as n increases, the sampling distribution gets more concentrated around λ_0.

If we take the square root of this value, we get the standard deviation, which is also called

the standard error of $\hat{\lambda}$. Of course, we do not know the true parameter, so the best we can do is substitute in the estimate, yielding the "estimate standard error" $s_{\hat{\lambda}} = \sqrt{\frac{\hat{\lambda}}{n}}$.

In summary, we see that the sampling distribution of the Method of Moments estimator for the Poisson distribution, $\hat{\theta}$, is approximately normal, centered around the true value of the parameter, and with a standard deviation of $\sqrt{\frac{\hat{\lambda}}{n}}$.

However, it is not always possible to derive the sampling distribution of an estimate theoretically as we have done here. In these cases, we use simulation: we generate many samples of size n of the random variable in question, we get the estimates and then look at histograms or analyze their statistical qualities.

5.1 Consistency of a parameter

A quality that was just described in the section above is the consistency of an estimator. An estimator is called consistent if it approaches the true value of the parameter, as the sample approaches infinity. More precisely,

Definition 7. *[1] Let $\hat{\theta_n}$ be an estimate of a parameter θ based on a sample of size n. $\hat{\theta_n}$ is said to be consistent in probability if $\hat{\theta_n}$ converges in probability to θ as n approaches ∞. That is, for any $\epsilon > 0$,*

$$P(|\hat{\theta_n} - \theta| > \epsilon) \to 0 \ \ as \ \ n \to \infty.$$

Recall that in finding the variance of the sampling distribution in the previous section, we substituted the estimate $\hat{\theta}$ for the true parameter value θ_0, which we don't know. The consistency of $\hat{\theta}$ justifies this substitution.

5.2 Bias

The bias of an estimator measures the difference in the expectation of the estimator and the true estimator. It is defined as follows:

Definition 8. *[4] Let $\hat{\theta}$ be an estimator for θ. The bias of $\hat{\theta}$ as an estimator for θ is*

$$Bias(\hat{\theta}, \theta) = E[\hat{\theta}] - \theta$$

In the case that $Bias(\hat{\theta}, \theta) = 0$, that is, $E[\hat{\theta}] = \theta$, we say that $\hat{\theta}$ is an unbiased estimator of θ.

As an example, let us go back to the Poisson distribution. Recall that the Method of Moments estimate for this distribution was the sample mean. That is,[4]

$$\hat{\theta}_{MoM} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

To see if this estimate is biased or unbiased, we take its expectation:

$$E[\hat{\theta}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E[x_i]$$

$$= \frac{1}{n}\sum_{i=1}^{n} \theta$$

$$= \frac{1}{n} n\theta$$

$$= \theta$$

With its expectation being the parameter, we see that $\hat{\theta}$ is indeed an unbiased estimator of θ.

5.3 Variance

The variance of an estimator is another way to assess its effectiveness. The variance in question is that of the variable's sampling distribution. The definition, which is simply the normal definition of variance applied to the estimate, is as follows:[4]

$$V[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2]$$

5.4 Mean Squared Error

As you can see in the formula above, the variance of the estimate is the expectation of the squared difference of the estimate and its expectation. In contrast, the Mean Squared Error measures the expectation of the squared difference of the estimate and the true parameter. It is defined as follows:

Definition 9. *[4] The mean squared error of an estimator $\hat{\theta}$ of θ is*

$$MSE(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2]$$

Note how if $\hat{\theta}$ is an unbiased estimator, that is, if its expectation is equal to the true value of the parameter, then the formulas for the MSE and Variance are equivalent. In fact, $\text{MSE}(\hat{\theta}, \theta) = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta}, \theta)^2$.

Let us look at a proof of this: [4]

$$MSE(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2]$$

r

$$= E[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)^2]$$
$$= E[(\hat{\theta} - E[\hat{\theta}])^2] + 2E[(\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta)] + E[(E[\hat{\theta}] - \theta)^2]$$
$$= Var[\hat{\theta}] + 0 + E[Bias(\hat{\theta}, \theta)^2]$$
$$= Var[\hat{\theta}] + Bias(\hat{\theta}, \theta)^2$$

5.5 Relative Efficiency

In the context of comparing two estimates to decide on which one to use, the ratio of their Mean Squared Errors are looked at. This is defined as follows: [1]

Given two estimates, $\hat{\theta}$ and $\tilde{\theta}$, the efficiency of $\hat{\theta}$ relative to $\tilde{\theta}$ is defined to be

$$\text{eff}(\hat{\theta}, \tilde{\theta}) = \frac{\text{Var}(\tilde{\theta})}{\text{Var}(\hat{\theta})}$$

Note that if both estimates being compared are unbiased, the comparison of their MSE's is the same as the comparison of their variances.

6 A Comparison of the two Methods on the Uniform$(0, \theta)$

Having seen two methods of parameter estimation as well as the ways that the variability of the estimates may be assessed, we can do a comparison. We will compare the Method of Moments and the Method of Maximum Likelihood for the Uniform$(0, \theta)$ distribution.

Consider a random variable W that follows the uniform distribution on the interval $[a, b]$.

If $W \sim U(a, b)$, we have the following density function:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

and the sample mean(and first moment)

$$\mu_1 = E[W^1] = \frac{1}{2}(a + b)$$

Now, if we consider a random variable X that follows the uniform distribution on the interval $(0, \theta)$, we have the density function

$$f_X(x) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

and the sample mean(and first moment)

$$\mu_1 = E[X^1] = \frac{1}{2}(0+\theta) = \frac{\theta}{2}$$

As per the method of moments, we set the first population moment equal to the sample mean, and get

$$\overline{X} = \frac{\hat{\theta}_{MOM}}{2} \rightarrow \hat{\theta}_{MOM} = 2\overline{X}$$

Hence, the estimate for θ by the Method of Moments is twice the sample mean. Also as a Method of Moments estimate, this estimator is unbiased.

Finding the Maximum Likelihood Estimator is more complex. First, we have the likelihood function

$$L(\theta) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} (\frac{1}{\theta}) = \frac{1}{\theta^n}$$

We want the value of θ that maximizes this value. Observe that this is a strictly decreasing function. Therefore, given that the observed data here is $x_1, x_2, \cdots, x_n$, the best bet we have for the M.L.E. is the maximum of these numbers. In other words, we know that

$$\max(x_1, x_2, \cdots, x_n) \leq \theta$$

And hence

$$\hat{\theta}_{\text{MLE}} = \max(x_1, x_2, \cdots, x_n)$$

More formally, given a sample $(x_1, x_2, \cdots, x_3)$ from the Uniform $(0, \theta]$ distribution, let us write the likelihood function where X_1 and X_n are minimum and maximum order statistics of the sample respectively.

We have,

$$\begin{aligned} L(\theta|X_1, \cdots, X_n) &= \prod_{i=1}^{n} f(x_i|\theta) \\ &= \frac{\prod_{i=1}^{n} I(0 \leq x_i \leq \theta)}{\theta^n} \\ &= \frac{I(0 \leq X_{(1)}) I(X_{(n)} \leq \theta)}{\theta^n} \end{aligned}$$

As a function of θ, we see that this function is monotonically decreasing. Also see that the likelihood becomes zero when $\hat{\theta} < X_{(n)}$. So, to maximize the likelihood, it is necessary that we should choose $\hat{\theta}_{\text{MLE}} = X_{(n)}$.

Next, we can compare the two methods by finding their relative efficiency. As the Method of Moments estimate is unbiased, its MSE is equal to its Variance, calculated as follows:

$$MSE(\hat{\theta}) = Var(\hat{\theta}_{MOM}) = Var(2\overline{X})$$

$$= Var(\frac{2\sum_{i=1}^{n} x_i}{n})$$

$$= \frac{4}{n^2}(\sum_{i=1}^{n} Var(x_i))$$

$$= \frac{4}{n^2} \cdot \frac{n\theta^2}{12} = \frac{\theta^2}{3n}$$

Hence

$$\text{MSE}(\hat{\theta}_{MOM}) = \frac{\theta^2}{3n}$$

Next, we calculate $\text{MSE}(\hat{\theta}_{MLE})$. For this estimate, we do need to consider the bias and consider the distribution of the maximum value in the sample, which is our estimate $\hat{\theta}_{MLE}$ in this context.

This is best done by looking at the cumulative distribution function(cdf) for $X_{(n)}$. As $X \sim U(0, \theta]$, $0 \leq X_i \leq \theta$. Let us take an arbitrary point x in this interval. Now, thinking about the distribution of $X_{(n)}$, we find the cdf at x:

$$F_{X_{(n)}}(x) = P(X_{(n)} \leq x)$$

$$= P(X_1 \leq x, X_2 \leq x, \cdots, X_n \leq x)$$

$$= P(X_1 \leq x)P(X_2 \leq x) \cdots P(X_n \leq x)$$

$$= [P(X \leq x)]^n$$

$$= [F(x)]^n$$

And now, to find the probability density function of X_n, we differentiate:

$$f(X_{(n)}) = \frac{d}{dx}[F(x)]^n$$

$$= n[F(x)]^{n-1} f(x)$$

$$= n\left(\frac{x}{\theta}\right)^{n-1}\left(\frac{1}{\theta}\right)$$

$$= \left(\frac{n}{\theta}\right)\left(\frac{x}{\theta}\right)^{n-1}$$

Next, to calculate the Variance, we use the formula

$$\begin{aligned}\mathrm{Var}[\hat{\theta}_{\mathrm{MLE}}] &= \mathrm{Var}[X_{(n)}] \\ &= E[X_{(n)}^2] - (E[X_{(n)}])^2 \\ &= \frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1}\right)^2 \\ &= \frac{n\theta^2}{(n+2)(n+1)^2}\end{aligned}$$

Then, we calculate the bias:

$$\begin{aligned}\mathrm{Bias} &= E(\hat{\theta}_{MLE}) - \theta \\ &= \frac{n}{n+1}\theta - \theta \\ &= \left(\frac{-\theta}{n+1}\right)\end{aligned}$$

Finally, we can find the $\mathrm{MSE}(\hat{\theta}_{MLE})$:

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}_{MLE}) &= (B(\hat{\theta}_{MLE}))^2 + \mathrm{Var}(\hat{\theta}_{MLE}) \\ &= \left(\frac{-\theta}{n+1}\right)^2 + \left(\frac{n\theta^2}{(n+2)(n+1)}\right) \\ &= \frac{\theta^2}{(n+1)^2}\left(1 + \frac{n}{n+2}\right) \\ &= \frac{\theta^2}{(n+1)^2}\left(\frac{2(n+1)}{n+2}\right) \\ &= \frac{2\theta^2}{(n+1)(n+2)}\end{aligned}$$

And now, we can look at the relative efficiency of the two estimates:

$$\begin{aligned}\mathrm{eff}(\hat{\theta}_{MOM}, \hat{\theta}_{MLE}) &= \frac{MSE(\hat{\theta}_{MLE})}{MSE(\hat{\theta}_{MOM})} \\ &= \frac{\frac{2\theta^2}{(n+1)(n+2)}}{\frac{\theta^2}{3n}} \\ &= \frac{6n}{n^2+3n+2}\end{aligned}$$

We see that the numerator has a linear term for n, while the denominator has a quadratic term. Therefore, as n increases to infinity, the efficiency is less than 1. Hence, for the Uniform$((0,\theta))$ distribution, $\hat{\theta}_{MLE}$ is a much better estimator of the true parameter than $\hat{\theta}_{MoM}$

7 The Asymptotic Normality of Maximum Likelihood Estimators

As we saw in the previous section, the Maximum Likelihood Estimator did a much better job estimating θ than the Method of Moments estimator did. It turns out that the Maximum Likelihood Estimator is asymptotically normal, that is,

$$\sqrt{n}(\hat{\theta} - \theta_0) \to N(0, \sigma^2_{MLE})$$

in distribution. In this section I will show this and compute σ^2_{MLE}, which is called the asymptotic variance and measures the quality of the MLE[5].

In doing so I will be using two facts from probability[5]:

- Law of Large Numbers: Given an independent and identically distributed sample $X_1, \cdots, X_n$ such that the expectation of X_1 is finite, the sample average

  $$\overline{X_n} = \frac{X_1 + \cdots + X_n}{n} \to E[X_1]$$

  converges to its expectation in probability.

- Central Limit Theorem: Given an independent and identically distributed sample $X_1, \cdots, X_n$ such that the expectation and variance of X_1 is finite,

  $$\sqrt{n}(\overline{X_n} - E[X_1]) \to N(0, \sigma^2)$$

  That is, $\sqrt{n}(\overline{X_n} - E[X_1])$ will behave like a random variable from the normal distribution as n gets large.

To show the asymptotic normality, we must first look at a quantity called Fisher Information. Recall that when finding the MLE, we often take the log of the likelihood function

$$l(X|\theta) = logf(X|\theta)$$

And if we denote the derivatives of $l(X|\theta)$ with respect to θ by $l'(X|\theta)$, $l''(X|\theta)$ etc., the Fisher Information of a random variable X with a probability distribution defined by the true parameter θ_0 is defined as

Definition 10.

$$I(\theta_0) = E[l'(X|\theta_0)]^2 = E[\frac{\partial}{\partial\theta} logf(X|\theta)|_{\theta=\theta_0}]^2$$

Note that the derivative

$$l'(X|\theta_0) = (logf(X|\theta_0))' = \frac{f'(X|\theta)}{f(X|\theta_0)}$$

tells us how quickly the distribution will change when we slightly change θ near θ_0. In the definition of Fisher Information, this value is squared and the expectation is taken, to give an averaged version of the measure. A large Fisher value denotes that the distribution changes quickly when we move the parameter, meaning that this particular distribution is distinct and the parameter should be easier to estimate based on the observed data. A smaller Fisher Information value denotes the opposite, that this distribution is similar to ones defined by parameters not close to θ_0. Hence the estimation of the parameter is worse.

We have an alternate formula for Fisher Information, as follows[5]:

Lemma 11.

$$E[l''(X|\theta_0)] = E\left[\frac{\partial^2}{\partial\theta^2} log f(X|\theta_0)\right] = -I(\theta_0)$$

This Lemma says that the expectation of the second derivative of the log likelihood is equal to the negative Fisher Information. A proof for this follows[5]:

Proof. We have

$$l'(X|\theta_0) = (log f(X|\theta_0))' = \frac{f'(X|\theta)}{f(X|\theta_0)} \tag{11.1}$$

And,

$$(log f(X|\theta))'' = \frac{f''(X|\theta)}{f(X|\theta)} - \frac{(f'(X|\theta))^2}{f^2(X|\theta)} \tag{11.2}$$

Next, we know that a p.d.f integrates to 1. That is,

$$\int f(X|\theta)dx = 1$$

If we take the first and second derivatives, assuming smoothness on the p.d.f. so that we can interchange integration and differentiation, we get

$$\int \frac{\partial}{\partial\theta} f(x|\theta)dx = 0 \text{ and } \int \frac{\partial^2}{\partial\theta^2} f(x|\theta)dx \int f''(x|\theta)dx = 0 \tag{11.3}$$

Now,

$$E[l''(X|\theta_0)] = E\left[\frac{\partial^2}{\partial\theta^2} log f(X|\theta)\right]$$

$$= \int (log f(x|\theta_0))'' f(x|\theta_0)dx$$

Using 11.2,

$$= \int \left(\frac{f''(x|\theta_0)}{f(x|\theta_0)} - (\frac{f'(x|\theta_0)}{f(x|\theta_0)})^2\right) f(x|\theta_0)dx$$

And now using 11.1,

$$= \int f''(x|\theta_0) - (l'(X|\theta))^2 dx$$

$$= \int f''(x|\theta_0)dx - E[l'(X|\theta_0)]^2$$

Finally, using 11.3 and definition 10,

$$= 0 - I(\theta_0) = -I(\theta_0)$$

□

Now, we can go onto prove the asymptotic normality of the estimator, that is, the theorem

Theorem 12.

$$\sqrt{n}(\hat{\theta} - \theta_0) \to N\left(0, \frac{1}{I(\theta_0)}\right)$$

As an aside, the variance of the asymptotic distribution is the reciprocal of the Fisher information. It makes sense that a large Fisher information, which would give more information about the random variable at hand, would result in a smaller variance and a more accurate estimate.

Proof. [5] We start by the fact that the MLE, $\hat{\theta}$, maximizes $L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} logf(X_i|\theta)$ by definition. Hence, we have $L_n'(\hat{\theta}) = 0$. Next, we make use of the Mean Value Theorem:

$$f(a) = f(b) + f'(c)(a - b)$$

with $f(\theta) = L'(\theta), a = \hat{\theta}$ and $b = \theta_0$. We have,

$$0 = L_n'(\hat{\theta}) = L_n'(\theta_0) + L_n''(\hat{\theta_1})(\hat{\theta} - \theta_0)$$

Or,

$$\hat{\theta} - \theta_0 = \frac{-L_n'(\theta_0)}{L_n''(\hat{\theta_1})}$$

Multiplying by $\sqrt{n}$ on both sides,

$$\sqrt{n}(\hat{\theta} - \theta_0)u = \frac{-\sqrt{n}L_n'(\theta_0)}{L_n''(\hat{\theta_1})} \tag{12.1}$$

Now, if we consider the numerator from 12.1, we have

$$\sqrt{n}L_n'(\theta_0)$$

$$= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} l'(X_i|\theta_0) - 0\right)$$

Using the fact that θ_0 maximizes $L(\theta)$, we can use $L'(\theta_0) = E[l'(X|\theta_0)] = 0$ and write

$$= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} l'(X_i|\theta_0) - E[l'(X_1|\theta_0)]\right)$$

And, by the Central Limit Theorem, we can say that this term will have the asymptotic distribution of $N(0, Var[l'(X_1|\theta_0)])$.

Next, we consider the denominator of 12.1. We use the fact that

$$L_n''(\theta) = \frac{1}{n}\sum l''(X_i|\theta) \to E[l''(X_1|\theta)]$$

by the law of large numbers. Since we know that $\hat{\theta_1}$ is in between $\hat{\theta}$ and θ_0, we have by consistency that $\hat{\theta_1}$ approaches θ_0. Using this fact and Lemma 11,

$$L_n''(\hat{\theta_1}) \to E[l''(X_1|\theta_0)] = -I(\theta_0)$$

Finally, combining the numerator and denominator and going back to 12.1,

$$\frac{-\sqrt{n}L_n'(\theta_0)}{L_n''(\hat{\theta_1})} \to N\left(0, \frac{Var(l'(X_1|\theta_0))}{(I(\theta_0))^2}\right)$$

And finding the variance, using the definition of Fisher Information and the fact that θ_0 maximizes,

$$\begin{aligned} Var(l'(X_1|\theta_0)) &= E[l'(X|\theta_0)]^2 - (E[l'(x|\theta_0))^2 \\ &= I(\theta_0) - 0 = I(\theta_0) \end{aligned}$$

Thus,

$$\sqrt{n}(\hat{\theta} - \theta_0) \to N\left(0, \frac{I(\theta_0)}{(I(\theta_0))^2}\right) = N\left(0, \frac{1}{I(\theta_0)}\right)$$

□

We have shown the asymptotic normality of the estimator.

8 Asymptotic Confidence Intervals of the Estimates

Having seen how to understand the variability of an estimate, and how estimates behave like normal random variables asymptotically, I will now discuss how to construct asymptotic confidence intervals for estimates. Note that besides asymptotically, there are two other methods to do this: using the exact sampling distribution of the estimator, or through simulation.

A confidence interval for a quantity is a random interval calculated from the data that contains that quantity with some specified probability. In parameter estimation, confidence intervals are constructed using estimates and their corresponding estimated standard errors[1]

We saw in the previous section that[1]

$$\sqrt{n}(\hat{\theta} - \theta_0) \to N\left(0, \frac{1}{I(\theta_0)}\right)$$

We may rewrite this as

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \to N(0, 1)$$

And substituting $I(\hat{\theta})$ for $I(\theta_0)$, we see that $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ approximately follows the normal distribution.

Since the standard normal is symmetric about zero, we can write

$$P\left(-z\left(\frac{\alpha}{2}\right) \leq \sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0) \leq z\left(\frac{\alpha}{2}\right)\right)$$

And rearranging gives us

$$\hat{\theta} \pm z\left(\frac{\alpha}{2}\right)\frac{1}{\sqrt{nI(\hat{\theta})}}$$

as an approximately $100(1 - \alpha)$ % confidence interval for the estimate θ.

9 Conclusion

This document started by motivating the study of parameter estimation: without it, known probability laws cannot be fit to observed data, and no conclusions or predictions can be made. Two methods of parameter estimation were examined, the Method of Moments and the Method of Maximum Likelihood. These were introduced theoretically, and examined through application and simulation. Next, various ways to assess the variability of these methods were discussed before letting them fight it out on the battlefield of Uniform$(0, \theta)$. The Maximum Likelihood Estimator came out on top on this occasion. Finally, the asymptotic characteristics of the MLE were considered, including its nature to approach the true value of the parameter, and follow a distribution with a variance of $(I(\theta_0)^{-1})$. Lastly, asymptotic confidence intervals for the estimator were shown.

References

[1] *Rice, John A. Mathematical statistics and data analysis. Cengage Learning*, 2006.

[2] *Hamming, Richard W. "The Art of Probability: for Scientists and Engineers"*, 1991.

[3] *Wagaman, Amy S. Probability: With Applications and R. John Wiley; Sons, Inc.* , 2021.

[4] https://web.stanford.edu/class/archive/cs/cs109/cs109.1218/files/studentdrive/7.6.pdf

[5] https://ocw.mit.edu/courses/18-443-statistics-for-applications-fall-2006/03b407da8a94b3fe22d987453807ca46lecture3.pdf