Point Estimation

Srikar Katta and Edoardo Airoldi

Contents

1 Point Estimation Strategies

Estimation strategies are techniques we can use to find the values of a model's parameters that maximize the probability of observing the empirical data. Point estimation strategies are techniques we employ to find the individual values (as opposed to the distribution) of the unknowns that are a "best fit" for our data. We present two point estimation techniques: the Methods of Moment Estimator and the Maximum Likelihood Estimator.

Before we dig into estimation strategies, let us establish some notation and terms. First, the model's parameters are called the estimand while the values that we believe are a reasonable characterization of the estimand are referred to as estimates; and the technique we use to find the estimand is known as the estimator. We typically denote out estimate with a hat; so if our estimand of interest is θ, then $\hat{\theta}$ would be the estimate of interest.

1.1 Method Moments Estimator

The Method of Moments Estimator (MOME) is a technique that takes advantage of moments in probability (i.e., quantities that can uniquely identify a distribution; e.g., mean and variance) and assumes that we can equate the empirical and theoretical moments to one another, as seen in Table 1.

The MOME has the following setup:

1. Represent the data as a probabilistic model f with unknown constants θ: $X_1, \ldots, X_n \sim f(\theta)$
2. If we have k estimands, then we set up k systems of equations where we equate each empirical and theoretical moment to one another

Moment	Theoretical	Empirical
Mean	$\mu = \mathbb{E}[X_i]$	$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$
Variance	$\mathbb{E}[(X_i - \mu)^2]$	$\frac{1}{N}\left(\sum_{i=1}^{N}(x_i - \bar{x})^2\right)$
$\vdots$	$\vdots$	$\vdots$
k^{th} centered moment	$\mathbb{E}[(X_i - \mu)^k]$	$\frac{1}{N}\left(\sum_{i=1}^{N}(x_i - \bar{x})^k\right)$

Table 1: Centered Empirical and Theoretical Moments for N samples of $X_i \sim f_X(\theta)$

3. Solve the system for the unknown constants; the solutions are labeled as $\hat{\theta}$

Example 1. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Normal(\mu, 1)$ with unknown constant μ. Let us find $\hat{\mu}_{MOME}$, the estimate of μ found via method of moments.

First, notice that we only have one unknown, μ. So, we will assume only the first theoretical and empirical moments are equivalent to one another. So, we assume $\mathbb{E}[X_i] = \bar{x}$. So,

$$\hat{\mu}_{MOME} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Example 2. We can now consider a slightly more complicated situation. Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} Normal(\mu, \sigma^2),$$

with unknown constants μ, σ^2. Let us find $\hat{\mu}_{MOME}, \hat{\sigma^2}_{MOME}$, the estimates of μ, σ^2 found via method of moments.

First, recognize that we only have two unknown constants, so we will only equate two moments to each other; we assume

$$\mathbb{E}[X_i] = \sum_{i=1}^{n} x_i \text{ and}$$

$$\mathbb{E}[(X_i - \mu)^2] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Notice that $\mu = \mathbb{E}[X_i]$, so the method of moments estimate for the true mean is simply the sample mean: $\hat{\mu}_{MOME} = \sum_{i=1}^{n} x_i$. Now, we can substitute $\hat{\mu}_{MOME}$ for μ in the calculation of

$$\mathbb{E}[(X_i - \mu)^2] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{MOME})^2.$$

Since $\sigma^2 = \mathbb{E}[(X_i - \mu)^2]$, the method of moments estimator for σ^2 is

$$\hat{\sigma}^2_{MOME} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \sum_{i=1}^{n} x_i\right)^2.$$

Example 3. One drawback of the method of moments estimator is that in a misspecified data generating

process, our method of moments estimator may yield nonsensical estimates. Consider the following problem. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Binomial(\theta, N)$ where θ is the probability of success and N is the number of independent Bernoulli trials. Suppose θ, N are both unknown but we observe the realizations $x_1, \ldots, x_n$. Let us estimate θ, N using the method of moments estimator.

We know that the first moment for a Bernoulli random variable is $N\theta$. So, our theoretical first moment is $\mathbb{E}[X] = N\theta$. And our empirical first moment is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Additionally, the second central theoretical moment is $\mathbb{E}[(X_i - \mathbb{E}[X])^2] = N\theta(1-\theta)$, and the second central empirical moment is $S_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i)^2$. Then, we can set up a system of equations by equating the theoretical and empirical moments together. So,

$$\begin{aligned} N\theta &= \bar{x} \\ N\theta(1-\theta) &= S_n. \end{aligned}$$

Then, we can solve for $\hat{\theta}$: $\hat{\theta} = \frac{\bar{x}}{N}$. And using our second moment, we can find N:

$$\begin{aligned} S_n &= N\theta(1-\theta) \\ &= N\left(\frac{\bar{x}}{N}\right)\left(1 - \frac{\bar{x}}{N}\right) \\ &= (\bar{x})\left(\frac{N - \bar{x}}{N}\right). \end{aligned}$$

We can then perform some algebra to find that $N = \frac{\bar{x}^2}{(\bar{x} - S_n)}$. So, $\hat{\theta}_{MOME} = \frac{\bar{x}}{N}$ and $\hat{N}_{MOME} = \frac{\bar{x}^2}{(\bar{x} - S_n)}$. So, we can plug N into

$$\begin{aligned} \hat{\theta}_{MOME} &= \frac{\bar{X}_n}{N} \\ &= \bar{X}_n \frac{(BarX_n - S_n)}{\bar{X}_n^2} \\ &= \frac{(BarX_n - S_n)}{\bar{X}_n}. \end{aligned}$$

So, the final solution reads as follows:

$$\begin{aligned} \hat{\theta}_{MOME} &= \frac{(BarX_n - S_n)}{\bar{X}_n} \\ \hat{N}_{MOME} &= \frac{\bar{X}_n^2}{(BarX_n - S_n)}. \end{aligned}$$

However, the MOME has one key issue: its estimators do not follow the same parametric space as the unknown constants, which could lead to wrong analyses. For example, under the Binomial distribution, $0 \leq \theta \leq 1$, so $0 \leq 1 - \theta \leq 1$, which implies $N\theta > N\theta(1-\theta)$. In other words, the theoretical variance must be greater than the theoretical mean in the case of the Binomial distribution. However, it is very possible that the empirical variance is less than the empirical mean, which will yield an estimate of $\hat{N}_{MOME} < 0$, which is

not possible. However, if there is a large sample size and the proposed model is not mis-specified, the method of moments estimator will likely give a good estimate.

This leads to a discussion of different samples. Often times, in estimation, people say there are empirical moments computed directly from the data and the theoretical moments. However, in practice, there are empirical moments that summarize the data, true theoretical moments from the true DGP that are accessible, and the model specified theoretical moments from the model/DGP approximation. In reality, the specified model may not be the true DGP but hopefully approximates the true DGP; if the specified model is not properly specified, there are many issues that may arise. So, careful precision is required in MOME.

1.2 Maximum Likelihood Estimator

The maximum likelihood estimation technique is another point estimation approach that identifies the set of parameter values that are most representative of the given data.

Let us motivate this section with an example.
Example 4. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Normal(\mu, 1)$ and represents the expression of gene called TPS1 in each of the n samples. Assume we observe the realizations $x_1, \ldots, x_n$ but not μ.

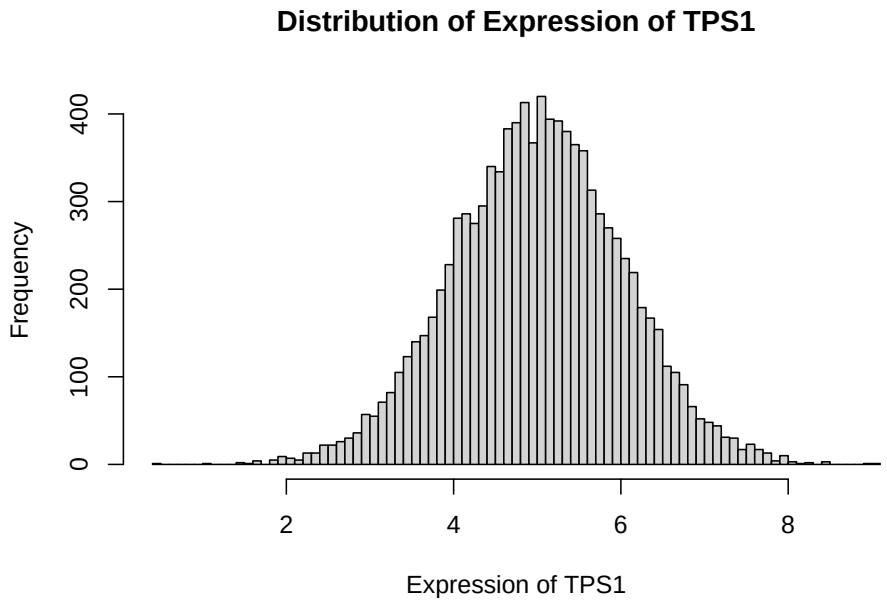
Then, the 2x2 table of this example would look as follows:

	Observed	**Unobserved**
Variable	$x_1, \ldots, x_n$	NA
Constant	NA	μ

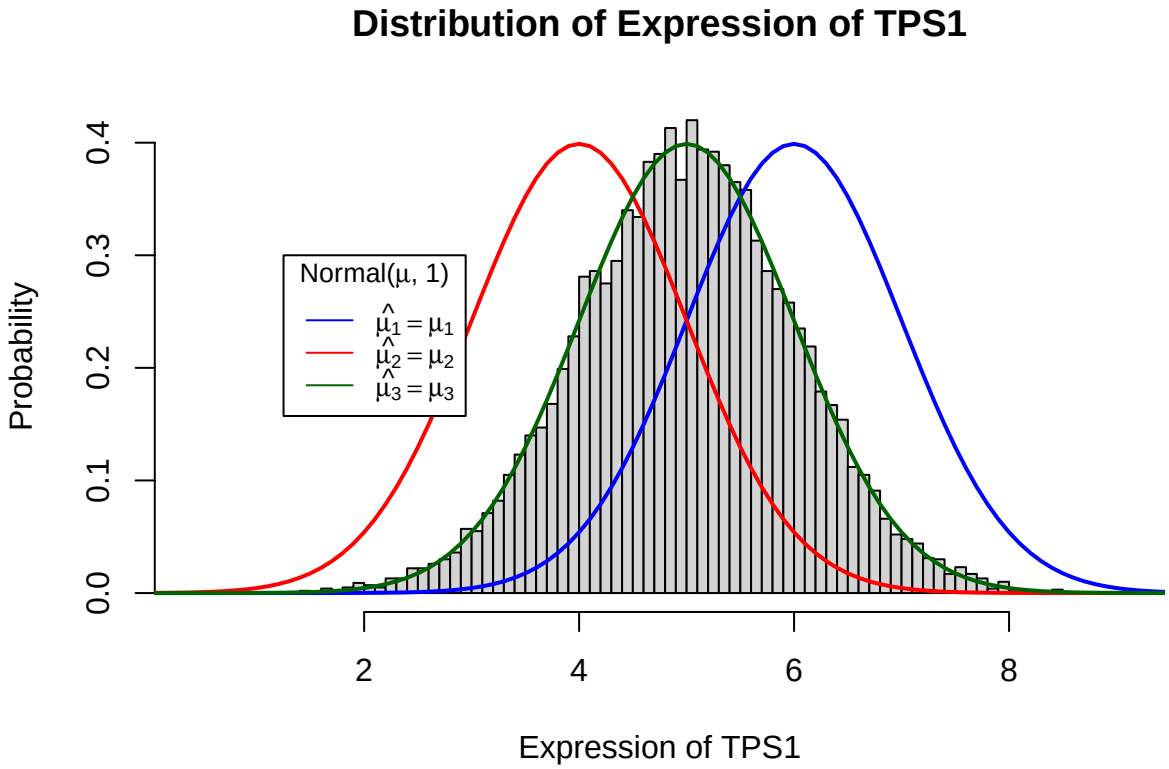
And the likelihood would look as follows:

$$\begin{aligned} likelihood &= \prod_{i=1}^{n} Normal(x_i|\mu, 1) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2}} \\ &= \frac{1}{\sqrt{2\pi}} \prod_{i=1}^{n} e^{\frac{-(x_i-\mu)^2}{2}} \\ &= \frac{1}{\sqrt{2\pi}} e^{\sum_{i=1}^{n} \frac{-(x_i-\mu)^2}{2}}. \end{aligned}$$

The likelihood is an expression of the unknown constant μ given all the constants. Simply, the likelihood is a function of $x_1, \ldots, x_n, \mu$, namely the probability of $x_1, \ldots, x_n$ given μ. Since μ is the only unknown, this is the same as saying the likelihood is a function of μ. Suppose the histogram of our data looks as follows:



We want to estimate μ, an unknown constant that describes the normal distribution for the data. Suppose we randomly select a μ_1, μ_2, μ_3. Each μ is a parameter for a distribution with their own forms, drawn below:



We have three random estimates for $\hat{\mu}$. Maximum likelihood estimation will allow us to find the best estimate for this distribution. Recall, these are ways of representing the likelihood:

$$likelihood = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n | \mu),$$

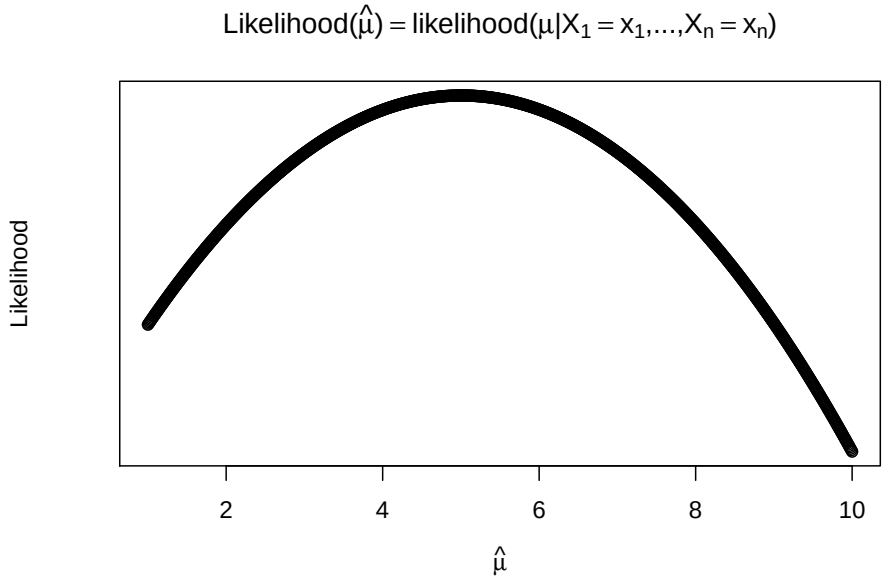
which is a function of μ.

The likelihood can be thought of as a score of the fit of a distribution for each of the unknown constants. And we want to maximize this score. Without μ, $\frac{1}{\sqrt{2\pi}} e^{\sum_{i=1}^{n} \frac{-(x_i-\mu)^2}{2}}$ is just a mathematical expression. However, if we consider specific $\hat{\mu}$, the $likelihood(\hat{\mu})$ is an actual value.

Now, the maximum likelihood estimate is the value of $\hat{\mu}$ that makes $x_1, \ldots, x_n$ most probable. In other words, the maximum likelihood estimator is a function of the observed random variables that maximizes the likelihood. Mathematically,

$$\hat{\mu}(x_1, \ldots, x_n) = \hat{\mu}_{MLE} = \arg\max_{\mu} likelihood(\mu).$$

Because the likelihood is essentially a function of μ, we can plot this function and find the maximum, as seen below:



So, to recap, the maximum likelihood estimator is the argument that maximizes the likelihood of the unknown constants given the data. So, in a more general case, the maximum likelihood estimator of the unknown constants θ is

$$\hat{\theta}_{MLE} = \arg\max_{\theta} likelihood(\theta).$$

1.2.1 Log Likelihood

One thing to recognize when maximizing likelihoods is that the argument that maximizes the likelihood is the same argument that maximizes the log of the likelihood because the logarithmic function is an always increasing function. Additionally, because the log of products is the sum of logs, the log likelihood is often times computationally easier to compute than the likelihood, which tends to have many products. The log likelihood is simpler and easier to maximize. We denote likelihood of θ as $\mathcal{L}(\theta)$ and the log likelihood of θ as $l(\theta)$. Please note that in general, we refer to log as being the natural log function.

Example 5. Let us find the log likelihood of

$$\mathcal{L}(\mu) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

Since the log of products is the product of logs,

$$\begin{aligned} l(\mu) &= \ln L(\mu) \\ &= \ln\left[\frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}\right] \\ &= \ln\frac{1}{\sqrt{2\pi}} + log\left[e^{\frac{-1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}\right]. \end{aligned}$$

Now, because $\ln(e^x) = x$,

$$l(\mu) = \ln\frac{1}{\sqrt{2\pi}} + \frac{-1}{2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Recall that $\log\left(\frac{a}{b}\right) = \log a - \log b$ and $\ln(1) = 0$. So,

$$\begin{aligned} l(\mu) &= \ln\frac{1}{\sqrt{2\pi}} + \frac{-1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 \\ &= \ln(1) - \ln(\sqrt{2\pi}) + \frac{-1}{2}\sum_{i=1}^{n}(x_i-\mu)^2. \end{aligned}$$

By applying the same log operations and expanding $(x_i-\mu)^2$,

$$\begin{aligned} &= 0 - \frac{1}{2}\ln(2\pi) + \frac{-1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 \\ &= -\frac{1}{2}[\ln(2)+\ln(\pi)] + \frac{-1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 \\ &= -\frac{1}{2}[\ln(2)+\ln(\pi)] + \frac{-1}{2}\sum_{i=1}^{n}(x_i^2+\mu^2-2x_i\mu) \\ &= -\frac{1}{2}[\ln(2)+\ln(\pi)] + \frac{-1}{2}\sum_{i=1}^{n}(x_i^2) + \sum_{i=1}^{n}(\mu^2) - \sum_{i=1}^{n}(2x_i\mu), \end{aligned}$$

which is easier to deal with analytically.

1.2.2 Invariance of Maximum Likelihood Estimator

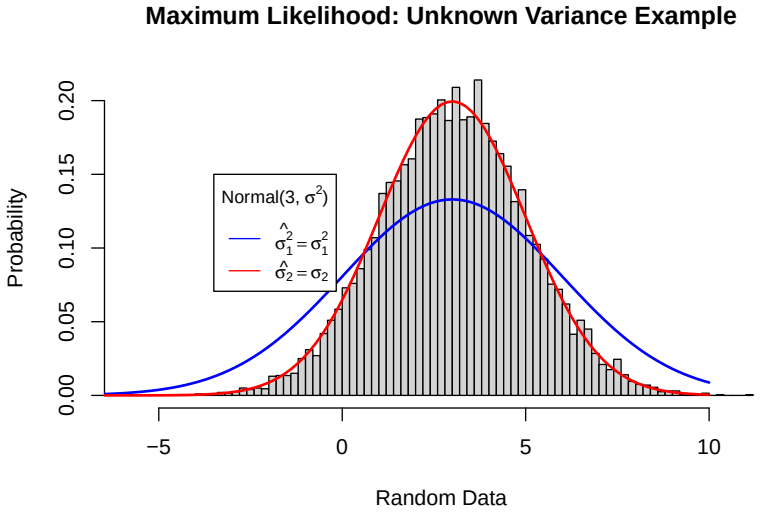
One key strength of maximum likelihood estimation is invariance. In short, if $\hat{\theta}_{MLE}$ is the maximum likelihood estimator for $L(\theta)$, then $g(\hat{\theta}_{MLE})$ is the maximum likelihood estimator for $L(g(\theta))$.

Definition 1.1: Invariance Property of MLE

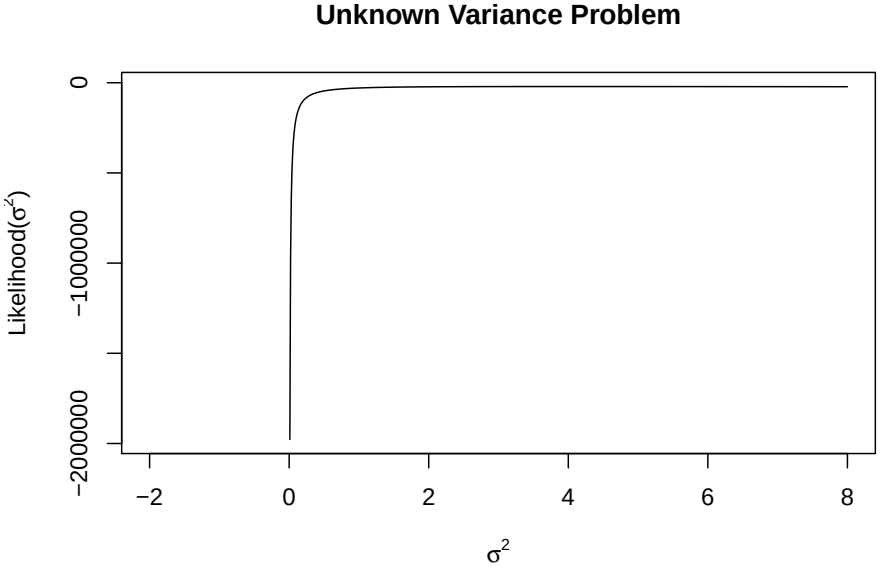
If the MLE estimate for θ is $\hat{\theta}$, then the MLE estimate for $f(\theta)$ is $f(\hat{\theta})$ for any function f.

Example 6. Consider the random variables $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Normal(3, \sigma^2)$. In this case, σ^2 is unknown. What is $\hat{\sigma^2}_{MLE}$.

Consider the following two distributions for our data; what is a better distribution to describe the data? Based on these two proposed distributions, the red distribution is a significantly better fit for the data because it follows it more closely.



Now, let us consider the likelihood function. Because we are looking for the σ^2 that maximizes the likelihood, which also maximizes the log likelihood, which ever one we consider really does not matter. While we can compute this by hand, consider the following plot of the likelihood/log likelihood over different values of σ^2:



So, $\hat{\sigma}^2_{MLE}$ is the value that maximizes the likelihood of the model in the first place. Notice, if the estimate if below 0, there is no likelihood associated with that value. That is because the likelihood estimator will follow

the same support as the constant we are hoping to estimate. Since variance is always greater than or equal to 0, the maximum likelihood estimator for the variance will always be greater than or equal to 0 as well. Because the likelihood is defined over the exact parametric space that defines the model (here, the model is $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Normal(3, \sigma^2)$), the possible for the maximum likelihood estimate will always like in the same parametric space. This property of maximum likelihood is one of the reasons it is so popular.

Example 7. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Normal(3, \sigma^2)$. Let us find σ^2_{MLE}.

So here, we want to estimate a function of σ^2, $\hat{\theta}_{MLE} = \frac{\sqrt{\sigma^2}}{\mu} = \frac{\sigma}{3}$. This computation is complicated. If we find the argument that maximizes $Likelihood(\sigma^2)$ known as $\hat{\sigma}^2_{MLE}$, then $\hat{\theta}_{MLE} = \frac{\sigma}{3} = \frac{\sqrt{\hat{\sigma}^2_{MLE}}}{3}$. In other words, if $\theta = \frac{\sqrt{\sigma^2}}{3}$, then $(3\theta)^2 = \sigma^2$. To find $\hat{\theta}_{MLE}$ without the invariance principle, we would have to maximize the likelihood with respect to θ: $\arg\max_\theta \mathcal{L}(\theta)$. But we no longer have to do that computation because we can utilize the relationship between θ and σ^2, which will be the same relationship as $\hat{\theta}_{MLE}$ and $\hat{\sigma}^2_{MLE}$.

Example 8. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Poisson(\lambda)$, where λ is unknown. Because X_i is from a Poisson distribution, the support of X_i is just the natural numbers (i.e., 0, 1, 2, 3,„. . . ,) and the support of $\lambda > 0$. Find $\hat{\lambda}_{MLE}$.

First, we need to find the likelihood of this model:

$$\begin{aligned}
Likelihood(\lambda) &= \mathbb{P}(x_1, \ldots, x_n|\lambda) \\
&= \mathbb{P}(x_1|\lambda)*, \ldots, *\mathbb{P}(x_n|\lambda) \\
&= \prod_{i=1}^{n} \mathbb{P}(x_i|\lambda) \\
&= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\
&= \frac{e^{-\sum_{i=1}^{n}\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} \\
&= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}.
\end{aligned}$$

To make the computation simpler, it makes sense to find the log likelihood:

$$\begin{aligned}
\ln \mathcal{L}(\lambda) &= \ln \left[\frac{e^{-\sum_{i=1}^{n} \lambda} \lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} \right] \\
&= \ln \left[e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i} \right] - \ln \prod_{i=1}^{n} x_i! \\
&= \ln \left[e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i} \right] - \ln \prod_{i=1}^{n} x_i! \\
&= \ln \left[e^{-n\lambda} \right] + \ln \left[\lambda^{\sum_{i=1}^{n} x_i} \right] - \ln \prod_{i=1}^{n} x_i! \\
&= -n\lambda + \sum_{i=1}^{n} x_i \ln(\lambda) - \sum_{i=1}^{n} \ln(x_i!) \\
&= -n\lambda + \ln(\lambda) \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln(x_i!).
\end{aligned}$$

Now, we can find the critical points of the log likelihood by taking the first derivative:

$$\begin{aligned}
\frac{d}{d\lambda} l(\lambda) &= \frac{d}{d\lambda} \left[-n\lambda + log(\lambda) \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} log(x_i!) \right] \\
&= \frac{d}{d\lambda} \left[-n\lambda + \frac{d}{d\lambda} log(\lambda) \sum_{i=1}^{n} x_i - \frac{d}{d\lambda} \sum_{i=1}^{n} log(x_i!) \right] \\
&= -n \frac{d}{d\lambda} \lambda + \sum_{i=1}^{n} x_i \frac{d}{d\lambda} log(\lambda) - \frac{d}{d\lambda} \sum_{i=1}^{n} log(x_i!) \\
&= -n(1) + \sum_{i=1}^{n} x_i \frac{1}{\lambda} - 0,
\end{aligned}$$

because $\frac{d}{d\lambda}\lambda = 1$, $\frac{d}{d\lambda} log(\lambda) = \frac{1}{\lambda}$, and the derivative with respect to λ of a term without λ is 0. Now, by setting the derivative equal to 0, we can find the critical points:

$$\begin{aligned}
0 = \frac{d}{d\lambda} l(\lambda) = -n + \sum_{i=1}^{n} x_i \frac{1}{\lambda}, &\iff n = \sum_{i=1}^{n} x_i \frac{1}{\lambda} \\
&\iff n\lambda = \sum_{i=1}^{n} x_i \\
&\iff \lambda^* = \frac{\sum_{i=1}^{n} x_i}{n}.
\end{aligned}$$

which implies that $\lambda^* = \frac{\sum_{i=1}^{n} x_i}{n}$, where λ^* is our proposed estimate. Now, we have to check if λ^* is a minimum or a maximum by finding the second derivative, which is just $-\sum_{i=1}^{n} x_i \frac{1}{\lambda^2}$. \end{align*}

Since $\lambda \geq 0$ and $x_i \geq 0$ always, the second derivative must be less than 0. So, that means λ^* is a maximum. So, $\hat{\lambda}_{MLE} = \lambda^* = \frac{\sum_{i=1}^{n} x_i}{n}$. If we were interested in the maximum of the likelihood, then we simply find $\mathcal{L}(\hat{\lambda}_{MLE})$.

We will now explore a very famous application of maximum likelihood, known as the "German Tank Problem."

In World War II, the Allied forces were destroying German and Italian tank. Each one that was destroyed had a serial number, and each serial number was sequential. A problem of key consideration was estimating how many tanks the German and Italians had. While this is an historic example, there are many current situations that this solution is still applicable to. For example, in finance, oftentimes traders assume they know log returns (a measure used to decide which stocks to buy or sell) are uniformly distributed. If one were to find the maximum negative log return (maximum loss), then this approach that will be laid out in this example will be of key importance.

Example 9. Assume $X_1...X_n \overset{\text{iid}}{\sim} Uniform[0, \theta]$, where θ is an unknown constant. Assume $x_1...x_n$ are observed. Let us find $\hat{\theta}_{MLE}$.

First, let us create the 2x2 table for this situation:

	Observed	**Unobserved**
Variable	$x_1...x_n$	NA
Constant	NA	θ

Because we are estimating with no latent variables and we want to ensure that our estimate is within its parameter space, maximum likelihood estimation is a perfect strategy for this problem. To reiterate, maximum likelihood estimation allows us to find the value of $\hat{\theta}$ that maximizes the chances (likelihood) of the observed data being seen, given some assumptions about the model's distributions.

For maximum likelihood estimation, we need to find the likelihood. Since we have no latent random variables, we can directly find the proper likelihood: $\mathcal{L}(\theta) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n|\theta)$. Since $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Uniform[0, \theta]$, the probability density function for the observed data is $\frac{1}{\theta}$. So,

$$\begin{aligned}\mathcal{L}(\theta) &= \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n|\theta) \\ &= \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(x_i)\end{aligned}$$

which is essentially saying that the probability of the observed data occurring given some θ is just $\frac{1}{\theta}$, as long as any observed data is not greater than θ. However, having this $x_i \in [0, \theta]$ term is a little cumbersome, which is why we use the indicator function, which is essentially a shorthanded form of a piece wise function. If the condition is met, then it returns 1, and it returns 0 if the condition is not met. In this example, we want $0 \leq x_i \leq \theta$, so we will the indicator function would work as follows:

$$\mathbb{1}_{[0,\theta]}(x_i) = \begin{cases} 1 & \text{if } 0 \leq x_i \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

If we refer back to the uniform distribution, θ represents the upper bound for the data, so if there exists some observed data point x_i that is greater than θ, then θ is no longer an upper bound. Since we take the product of probability of each individual observation occurring given θ (i.e., $\prod_{i=1}^{n} \frac{1}{\theta}$), if one of the observations is greater than θ, the entire likelihood evaluates to 0. For example, suppose $\theta = 5$ and some observation $x_k = 10$. Then, $\mathcal{L}(\theta = 10|x_k = 5) = 0$.

Now, notice that the likelihood is a function of $x_1, \ldots, x_n$ since it is the input for $\mathbb{1}_{[0,\theta]}(x_i)$. Because we want to maximize the likelihood only with respect to θ, we need to find some way to "swap" $[0, \theta]$ and x_i that is fair

for all x_i. So, for all x_i, $\theta \geq x_i \geq 0$. Since x_i belongs to a finite set of terms, we can find the maximum of all of the x_i: $\max(x_1, \ldots, x_n)$. By definition of the maximum, for all x_i, $\max(x_1, \ldots, x_n) \geq x_i$. So, that means if $\theta \geq \max(x_1, \ldots, x_n)$, then $\prod_{i=1}^{n} \mathbb{1}_{[0,\theta]}(x_i) = 1$. So, we can rewrite this as $\mathbb{1}_{[0,\theta]}(\max(x_1, \ldots, x_n)) = 1$, which says that as long as the maximum is between 0 and θ, then θ is valid. Now, we still need to swap the x_i and θ. Instead of saying $0 \leq \max(x_1, \ldots, x_n) \leq \theta$, we can say that $\max(x_1, \ldots, x_n) \leq \theta < \infty$. So, $\mathbb{1}_{[0,\theta]}(\max(x_1, \ldots, x_n)) = 1 = \mathbb{1}_{[\max(x_1,\ldots,x_n),\infty)}(\theta) = 1$. So,

$$\mathcal{L}(\theta) = \frac{1}{\theta^n} \mathbb{1}_{[\max(x_1,\ldots,x_n),\infty)}(\theta),$$

which is now truly a function of θ.

Now that we have a likelihood expressed in terms of θ, we can find $\hat{\theta}_{MLE}$. Let ϵ and a be some numbers greater than 0. Intuitively, we know that when $a < a + \epsilon$, so that means $\frac{1}{a} > \frac{1}{a+\epsilon}$. So, now, replacing that with θ^n, as θ^n increases, $\frac{1}{\theta^n}$ decreases. So, the $\hat{\theta}$ that maximizes the likelihood will be the $\hat{\theta}_{MLE}$ that is less than all other possible $\hat{\theta}$s. Since we have the restriction that $\theta \geq \max(x_1, \ldots, x_n)$, it is quite easy to see that the smallest $\hat{\theta}$ can be is $\max(x_1, \ldots, x_n)$, so $\hat{\theta}_{MLE} = \max(x_1, \ldots, x_n)$.

Analytically, we can take the derivative of the likelihood and find $\hat{\theta}_{MLE}$:

$$\frac{\partial \mathcal{L}}{\partial \theta} = -n \left(\frac{1}{\theta^{n-1}} \right),$$

which is less than 0 for all θ, suggesting a decreasing function. So the minimum θ maximizes the likelihood. Thus, $\hat{\theta}_{MLE} = \max(x_1, \ldots, x_n)$.

1.2.3 Sufficient Statistics

First, recall that a statistic is any function of the data that is not a function of unknown constants/parameters. A sufficient statistic is any statistic of the data that captures all of the relevant/necessary information for estimating the model's unknown parameters.

Definition 1.2: Sufficient statistic

Suppose we have a probabilistic model with unknown parameters θ and observed data $x_1, \ldots, x_n$. A statistic $T = r(X_1, \ldots, X_n)$ is sufficient if

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n | T = t, \theta) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n | T = t),$$

for all t.

Example 10. Let us look at sufficient statistics in action. Suppose X_i represents the number of words in document i. Assume $X_i \overset{\text{iid}}{\sim} Poisson(\lambda)$ where $\lambda > 0$. Suppose $x_1, \ldots, x_n$ are observed and λ is an unknown constant. Let us find $\hat{\lambda}_{MLE}$.

Since $X_1, \ldots, X_n$ have a theoretical distribution and λ is unknown without a theoretical distribution (note, $\lambda > 0$ is not a distribution), the 2x2 looks as follows:

	Observed	**Unobserved**
Variable	$X_1, \ldots, X_n$	NA
Constant	NA	λ

So, the likelihood is only a function of λ and the proper likelihood can be found directly:

$$\mathcal{L}(\lambda) = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}.$$

Because we would like to maximize the likelihood, we can instead maximize the log of the likelihood which will most likely be easier to take the derivative of:

$$\begin{aligned} l &= \ln \mathcal{L}(\lambda) \\ &= -n\lambda + \sum_{i=1}^{n} x_i \ln(\lambda) - \sum_{i=1}^{n} \ln(x_i!). \end{aligned}$$

To find $\hat{\lambda}$, we take the derivative of the log likelihood with respect to λ:

$$\begin{aligned} 0 = \frac{\partial l}{\partial \lambda} &= \frac{\partial}{\partial \lambda}\left[-n\lambda + \sum_{i=1}^{n} x_i \ln(\lambda) - \sum_{i=1}^{n} \ln(x_i!)\right] \\ &= -n\frac{\partial}{\partial \lambda}\lambda + \sum_{i=1}^{n} x_i \frac{\partial}{\partial \lambda}\ln(\lambda) - \frac{\partial}{\partial \lambda}\sum_{i=1}^{n} \ln(x_i!) \\ &= -n + \sum_{i=1}^{n} \frac{x_i}{\lambda} + 0. \end{aligned}$$

Notice, the derivative of the term $\sum_{i=1}^{n} \ln(x_i!)$, so it adds no value to estimating λ via MLE.

Then, the maximum likelihood estimator can be found from the non-zero terms of the first derivative, which in this case is

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

Using this information, we can propose $\lambda^* = \frac{\sum_{i=1}^{n} x_i}{n}$ as a potential MLE estimate. While we will not show the calculations here, by

$$\begin{aligned} 0 &= \frac{\partial}{\partial \lambda} l(\lambda) \\ \iff 0 &= 0 - n + \frac{\sum_{i=1}^{n} x_i}{\lambda} \\ \iff n &= \frac{\sum_{i=1}^{n} x_i}{\lambda} \\ \iff \lambda^* &= \frac{\sum_{i=1}^{n} x_i}{n}, \end{aligned}$$

so (as long as the second derivative of the log likelihood is less than 0, suggesting that $\lambda*$ is maximum), $\hat{\lambda}_{MLE} = \lambda^*$. Because $\sum_{i=1}^{n} x_i$ is a function of the data and is the minimum set of information necessary to compute the maximum likelihood, it is a sufficient statistic. \end{solution}

In summary, to compute the maximum likelihood estimator for θ, an unknown constant, we follow these steps:

- Get likelihood function: $\mathcal{L}(\theta)$
- Get log of the likelihood function: $l(\theta) = \ln \mathcal{L}(\theta)$
- Find the first derivative of the log likelihood: $\frac{d}{d\theta} l(\theta) = 0$, which proposes potential θ^*
- Find second derivative of the log likelihood and check that θ^* is in fact a maximum: $\frac{d^2}{d\theta^2} l < 0$
- If $\frac{d^2}{d\theta^2} l < 0$, then $\hat{\theta}_{MLE} = \theta^*$.