Statistical Inference: Maximum Likelihood Estimation *

Konstantin Kashin

Spring 2014

Contents

*This note was initially prepared to supplement course material for Gov 2001 for Spring 2013. It is in part based upon my notes from Stat 111 at Harvard University and also the resources cited at the end.

1 Overview: Maximum Likelihood vs. Bayesian Estimation

This note is about the mechanics of maximum likelihood estimation (MLE). However, before delving into the mechanics of finding the MLE, let's step back and lay out maximum likelihood as a theory of inference. Specifically, it will prove useful to compare maximum likelihood to Bayesian theory of inference.

In general, what is statistical inference? It's the process of making a statement about how data is generated in the world. We can think of the data that we observe in the world as a product of some data generation process (DGP) that is fundamentally unknown and likely highly convoluted. However, it is our goal as social scientists or applied statisticians to use observed data to learn something about the DGP. In parametric inference, we are going to assume that we can represent the DGP by a statistical model. Remember that whatever model we choose, it will essentially never be "right". Thus, the question is whether or not the model is useful. In the context of this note, we will limit ourselves to very common probability distributions as our models.

Once we have selected a probability distribution to represent the data generation process, we aren't done, for in fact we have not specified a unique distribution, but just a family of distributions. This is because we leave one (or more) parameters as unknown. The goal of statistical inference is then to use observed data to make a statement about parameters that govern our model.

To introduce some notation, we are going to call the model, or probability distribution, that we choose $f(\cdot)$. This probability distribution is going to depend on a parameter θ (or vector of parameters, $\boldsymbol{\theta} = \theta_1, \theta_2, ..., \theta_k$) that characterize the distribution. The set Ω of all possible values of a parameter or the vector of parameters is called the parameter space. We then observe some data, drawn from this distribution:

$$X \sim f(x|\theta)$$

The random variables $X_1, ... ,X_n$ are independent and identically distributed (iid) because they are drawn independently from the same DGP. The goal is to use the observed data $\mathbf{x}$ to learn about θ. Knowing this parameter specifies a particular distribution from the family of distributions we have selected to represent the data generation process. In the end, we hope that θ is a substantively meaningful quantity that teaches us something about the world (or at least can be used to derive a substantively interesting quantity of interest).

So far, we've just set up the general inferential goal. Now, we can introduce different theories of actually achieving said goal. Specifically, we focus on two general approaches to inference and estimation: **frequentist / maximum likelihood** and **Bayesian**. The two are distinguished by their sources of variability, the mathematical objects involved, and estimation and inference. It is important to keep track of the sources of randomness in each of these paradigms since different estimators are used for random variables as opposed to constants.

Let's formalize the notion of inference using Bayes' Rule. First, let's restate the goal of inference: it's to estimate the probability that the parameter governing our assumed distribution is θ conditional on the sample we observe, denoted as $\mathbf{x}$. We denote this probability as $\xi(\theta|\mathbf{x})$. Using Bayes' Rule, we can equate this probability to:

$$\xi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\xi(\theta)}{g_n(\mathbf{x})} = \frac{f_n(\mathbf{x}|\theta)\xi(\theta)}{\int_{\Omega} f_n(\mathbf{x}|\theta)\xi(\theta)}, \text{ for } \theta \in \Omega$$

Now, since the denominator in the expression above is a constant in the pdf of θ ($g_n(\mathbf{x})$ is simply a function of the observed

data), the expression can be rewritten as:

$$\underbrace{\xi(\theta|\mathbf{x})}_{\text{posterior}} \propto \underbrace{f_n(\mathbf{x}|\theta)}_{\text{likelihood}} \underbrace{\xi(\theta)}_{\text{prior}}$$

The two theories of inference – frequentist and Bayesian – diverge at this step. Under the frequentist paradigm, the parameter θ is a constant, albeit an unknown constant. Thus, the prior is meaningless and we can absorb $\xi(\theta)$ into the proportionality sign (along with the normalization constant / denominator from Bayes' Rule above). The result is what R.A. Fisher termed likelihood:

$$L(\theta|\mathbf{x}) \propto f_n(\mathbf{x}|\theta)$$

Sometimes, the likelihood is also written in terms of an unknown constant $k(x)$:

$$L(\theta|\mathbf{x}) = k(x) f_n(\mathbf{x}|\theta)$$

Since $k(x)$ is never known, likelihood is not a probability density. Instead, likelihood is some positive multiple of $f_n(\mathbf{x}|\theta)$.

To summarize, the parameters in the frequentist setting (likelihood theory of inference) are unknown constants. Therefore, we can ignore $\xi(\theta)$ and just focus on the likelihood since everything we know about the parameter based on the data is summarized in the likelihood function. The likelihood function is a function of θ: it conveys the relative likelihood of drawing the sample observations you observe given some value of θ.

In contrast to frequentist inference, in the Bayesian setting, the parameters are latent random variables, which means that there is some variability attached to the parameters. This variability is captured through one's prior beliefs about the value of θ and is incorporated through the prior, $\xi(\theta)$. The focus of Bayesian inference is estimating the posterior distribution of the parameter, $\xi(\theta|\mathbf{x})$.

The posterior distribution of θ, $\xi(\theta|\mathbf{x})$, is the distribution of the parameter conditional upon the observed data and provides some sense of (relative) uncertainty regarding our estimate for θ. Note that we cannot obtain an absolute measure of uncertainty since we do not truly know $\xi(\theta)$. However, even before the data is observed, the researcher may know where θ may lie in the parameter space Ω. This information can thus be incorporated through the prior, $\xi(\theta)$ in Bayesian inference. Finally, the data is conceptualized as a joint density function conditional on the parameters of the hypothesized model. That is, $f_n(x_1, x_2, \ldots, x_n|\theta) = f_n(\mathbf{x}|\theta)$. For an iid sample, we get $f(X_1|\theta) \cdot f(x_2|\theta) \cdots f(x_n|\theta)$. The term $f_n(\mathbf{x}|\theta)$ is known as the **likelihood**.

To recap, where does variability in the data we observe come from? In both frameworks, the sample is a source of variability. That is, $X_1, ..., X_n$ form a random sample drawn from some distribution. In the Bayesian mindset, however, there is some additional variability ascribed to the prior distribution on the parameter θ (or vector of parameters). This variability from the prior may or may not overlap with the variability from the sample. Frequentists, by contrast, treat the parameters as unknown constants.

As a result of the differences in philosophies, the estimation procedure and the approach to inference differ between frequentists and Bayesians. Specifically, under the frequentist framework, we use the likelihood theory of inference where the maximum likelihood estimator (MLE) is the single point summary of the likelihood curve. It is the point which maximizes the likelihood function. In contrast, the Bayesian approach tends to focus on the posterior distribution of θ and various estimators, such as

the posterior mean (PM) or maximum a posteriori estimator (MAP), which summarize the posterior distribution.

To summarize the distinction between the two approaches to inference, it helps to examine a typology of mathematical objects. It classifies objects based on whether they are random or not, and whether they are observed or not. When confronted with inference, one must always ask if is there a density on any given object. A presence of a density implies variability. Furthermore, one must ask if the quantity is observed or not observed.

	Observed	Not Observed
Variable (Var > 0)	**Random Variable** **X** (the data)	**Latent Random Variable** θ in Bayesian inference
Not Variable (Var = 0)	**Known Constant** α, β, which govern $\xi(\theta)$ in Bayesian inference	**Unknown Constant** θ in frequentist inference

2 Introduction to Maximum Likelihood Estimation

2.1 What is Likelihood and the MLE?

Say that one has a sample of n iid observations $X_1, X_2, ..., X_n$ that come from some probability density function characterized by an unknown parameter θ_o: $f_o = f(\cdot|\theta_o)$, where θ_o belongs to a parameter space Ω. We want to find $\hat{\theta}$ that is the best estimator of θ_o. Specifically, the main approach of maximum likelihood estimation (MLE) is to determine the value of θ that is most likely to have generated the vector of observed data, $\mathbf{x}$.

The **likelihood**, $L(\theta|\mathbf{x})$, is a function that assigns a value to each point in parameter space Ω which indicates how likely each value of the parameter is to have generated the data. This is proportional to the joint probability distribution of the data as a function of the unknown coefficients. According to the likelihood theory of inference, the likelihood function summarizes all the information we have about the parameters given the data we observe. The method of maximum likelihood obtains values of model parameters that define a distribution that is most likely to have resulted in the observed data. For many statistical models, the MLE estimator is just a function of the observed data. Furthermore, we often work with the log of the likelihood function, denoted as $\log L(\theta|\mathbf{x}) = \ell(\theta|\mathbf{x})$. Note that since the log is a monotonic function, this does not change any information we have about the parameter.

As has been alluded to, it is important to distinguish the likelihood of the parameter θ from the probability distribution of θ conditional upon the data, which is obtained via Bayes' Theorem. The likelihood is not a probability. Instead, the likelihood is a measure of **relative** uncertainty about the plausible values of θ, given by Ω. This relativity is exactly what allows us to work with the log of the likelihood and to scale the likelihood using monotonic transformations. As a result, we can only compare likelihoods within, not across, data sets.

Let us formally define the likelihood as proportional to the joint probability of the data conditional on the parameter:

$$L(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

The maximum likelihood estimate of θ, which we denote as $\hat{\theta}_{MLE}$, is the value of θ in parameter space Ω that maximizes the likelihood (or log-likelihood) function. It is the value of θ that is most likely to have generated the data.

Mathematically, we write the MLE as:

$$\hat{\theta}_{MLE} = \max_{\theta \in \Omega} L(\theta|\mathbf{x}) = \max_{\theta \in \Omega} \prod_{i=1}^{n} f(x_i|\theta)$$

Alternatively, we could work with the log-likelihood function because maximizing the logarithm of the likelihood is the same as maximizing the likelihood (due to monotonicity):

$$\hat{\theta}_{MLE} = \max_{\theta \in \Omega} \log L(\theta|\mathbf{x}) = \max_{\theta \in \Omega} \ell(\theta|\mathbf{x}) = \max_{\theta \in \Omega} \sum_{i=1}^{n} \log(f(x_i|\theta))$$

How do we actually find the MLE? There are two alternatives: analytic and numeric. We won't focus on numeric optimization

methods in this note, but analytically, finding the MLE involves taking the first derivative of the log-likelihood (or likelihood function), setting it to 0, and solving for the parameter θ. We then need to check that we have indeed obtained a maximum by calculating the second derivative at the critical value and checking that it is negative.

To further introduce some terminology, let us define the **score** as the first derivative of the log-likelihood function with respect to each of the parameters (gradient). For a single parameter:

$$S(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}$$

The first order condition thus involves setting the score to zero and solving for θ.

In the case of multiple parameters (a vector $\boldsymbol{\theta}$ of length k), the score is defined as:

$$S(\boldsymbol{\theta}) = \nabla \ell(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1} \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_2} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_k} \end{pmatrix}$$

We can visualize the log-likelihood curve quite easily using R (at least for the most common distributions). Let's visualize a log-likelihood curve for μ in a normal distribution with unknown μ and a known $\sigma = 1$. The observed data is $\mathbf{x} = \{7, 6, 5, 5, 7, 5, 6, 3, 4, 6\}$.

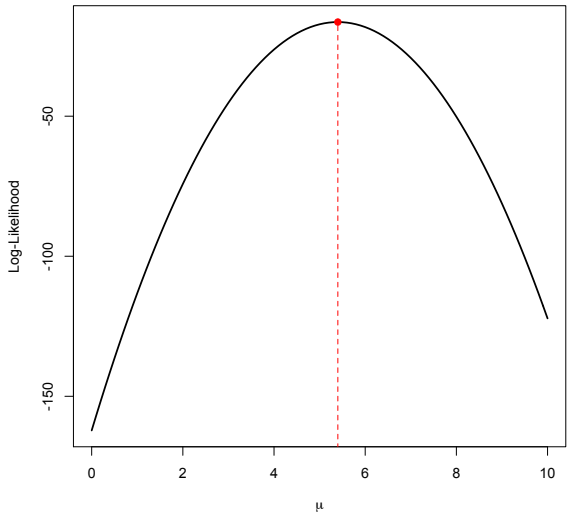
```
my.data <- c(7,6,5,5,7,5,6,3,4,6)
norm.ll<- function(x) return(sum(dnorm(my.data,mean=x,sd=1,log=TRUE)))
norm.ll <- Vectorize(norm.ll)
curve(norm.ll, from=0,to=10, lwd=2, xlab=expression(mu),ylab="Log-Likelihood")
```

Similarly, we can visualize a log-likelihood curve for the parameter λ in a Poisson distribution governed by that parameter. The observed data is $\mathbf{x} = \{2, 1, 1, 4, 4, 2, 1, 2, 1, 2\}$.

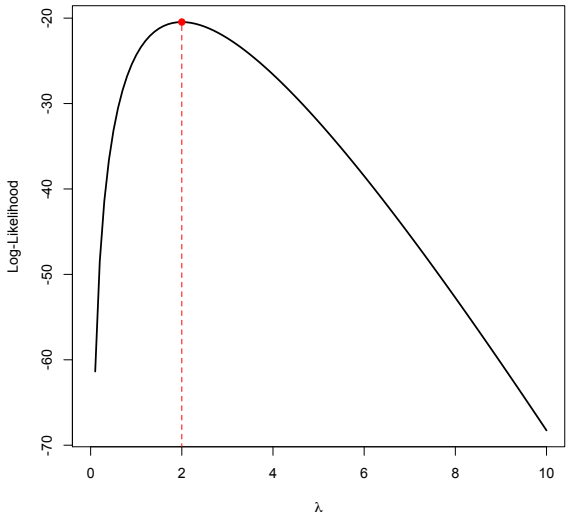
```
my.data <- c(2,1,1,4,4,2,1,2,1,2)
pois.ll<- function(x) return(sum(dpois(my.data,lambda=x,log=TRUE)))
pois.ll <- Vectorize(pois.ll)
curve(pois.ll, from=0,to=10, lwd=2, xlab=expression(lambda),ylab="Log-Likelihood")
```

The results of these two plots, along with the MLE estimates of the respective parameters, are presented in Figure 1. The log-likelihood surface for a 2-parameter example (a normal distribution with unknown mean and variance) is presented in Figure 2.

Figure 1: Examples of Log-Likelihood Functions

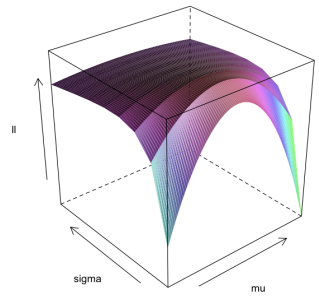


(a) Log-Likelihood curve for μ in a normal distribution based on the following data: $\{7, 6, 5, 5, 7, 5, 6, 3, 4, 6\}$. Note $\sigma = 1$.

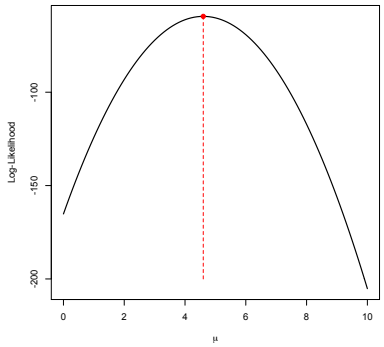


(b) Log-Likelihood curve for λ in a Poisson distribution based on the following data: $\{2, 1, 1, 4, 4, 2, 1, 2, 1, 2\}$.

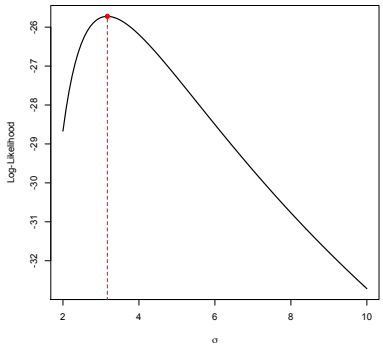
Figure 2: Example of Log-Likelihood Function for Two Parameters: μ and σ in a normal distribution



(a) Log-Likelihood surface for both μ and σ in a normal distribution.



(b) Marginal log-likelihood curve for μ.



(c) Marginal log-likelihood curve for σ.

2.2 Examples of Analytical MLE Derivations

2.2.1 MLE Estimation for Sampling from Bernoulli Distribution

$X_1 \ldots X_n$ form a random sample from a Bernoulli distribution with unknown parameter $0 \leq \theta \leq 1$. We need to find the MLE estimator for θ.

$$L(\theta \,|\, x_1, \ldots, x_n) = f_n(\mathbf{x}|\theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}$$

Taking the log of the function, we get:

$$\ell(\theta|\mathbf{x}) = (\sum_{i=1}^{n} x_i) \log \theta + (n = \sum_{i=1}^{n} x_i) \log(1-\theta)$$

Optimizing $\ell(\theta|\mathbf{x})$ by taking its derivative and finding its roots yields the estimator $\hat{\theta} = \bar{X}$.

2.2.2 MLE Estimation of Mean and Variance for Sampling from Normal Distribution

$X_1 \ldots X_n$ form a random sample from a Normal distribution with unknown parameters $\theta = (\mu, \sigma^2)$. We need to find the MLE estimator for θ.

Starting with the likelihood:

$$L(\theta \,|\, x_1, \ldots, x_n) = f_n(\mathbf{x}|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2]$$

Taking the log of the function, we get:

$$\ell(\theta \,|\, x_1, \ldots, x_n) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

The likelihood function needs to be optimized with respect to parameters μ and σ^2, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$.

First, treat σ^2 as known and find $\hat{\mu}(\sigma^2)$. Now, we can take the partial derivative of the log likelihood with respect to the mean parameter and set it equal to zero:

$$\frac{\partial \ell(\theta)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{X}_n$$

It is good practice to check that the obtained estimator is indeed the maximum using the second order condition:

$$\frac{\partial^2 \ell(\theta)}{\partial \mu^2} = \frac{-n}{\sigma^2} < 0 \;\forall\; n, \sigma^2 > 0$$

Now for the variance. Plugging $\hat{\mu} = \bar{x}_n$ in for μ in the log-likelihood function and taking the derivative with respect to σ^2:

$$\frac{\partial \ell(\theta)}{\partial \sigma^2} = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 = 0$$

Solving for σ^2, we get:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$$

The MLEs for μ and σ^2 are thus: $\boxed{\hat{\mu} = \bar{X} \text{ and } \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2}$

2.2.3 Gamma Distribution

For the gamma distribution, θ is the scale parameter and α is the shape parameter. We seek the conditions for the maximum likelihood estimates of (θ, α).

The likelihood function for a gamma distribution is the following:

$$L(\alpha, \theta|\mathbf{x}) = f_n(\mathbf{x}|\alpha, \theta) = \frac{1}{\Gamma^n(\alpha) \cdot \theta^{n\alpha}} \Big(\prod_{i=1}^{n} x_i \Big)^{\alpha - 1} \exp(-\sum_{i=1}^{n} \frac{x_i}{\alpha})$$

Taking the log, we obtain:

$$\ell(\alpha, \theta) = -n \cdot \log(\Gamma(\alpha)) - n\alpha \cdot \log(\theta) + (\alpha - 1) \sum_{i=1}^{n} \log(x_i) - \frac{1}{\theta} \sum_{i=1}^{n} x_i$$

Taking the derivative of the log likelihood with respect to θ and setting it equal to 0:

$$\frac{\partial \ell(\alpha, \theta)}{\partial \theta} = -\frac{n\alpha}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i = 0$$

Solving for θ, we obtain the following MLE for θ:

$$\hat{\theta}|\alpha = \frac{\sum_{i=1}^{n} x_i}{\alpha \cdot n} = \frac{1}{\alpha}\bar{x}_n$$

Plugging this back into the log likelihood function, taking its derivative with respect to α, and setting the result equal to 0:

$$\frac{dL(\alpha, \hat{\theta|\alpha})}{d\alpha} = -\frac{n \cdot \Gamma'(\alpha)}{\Gamma(\alpha)} - n \cdot \log(\frac{1}{\alpha}\bar{x}_n) + \sum_{i=1}^{n} \log(x_i) = 0$$

$$\frac{dL(\alpha, \hat{\theta|\alpha})}{d\alpha} = -\frac{n \cdot \Gamma'(\alpha)}{\Gamma(\alpha)} + n \cdot \log(\alpha) - n \cdot \log(\bar{x}_n) + \sum_{i=1}^{n} \log(x_i) = 0$$

Solving for α as far as we can (the answer remains in terms of the digamma function), we obtain the following condition for the MLE of α:

$$\log(\alpha) - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \log(\bar{x}_n) - \frac{\sum_{i=1}^{n} \log(x_i)}{n}$$

Therefore, the MLE values of α and θ must satisfy the following 2 equations (there is no unique solution) are:

$$\boxed{\log(\hat{\alpha}) - \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = \log(\bar{x}_n) - \frac{\sum_{i=1}^{n} \log(x_i)}{n}}$$

$$\boxed{\hat{\theta} = \frac{1}{\hat{\alpha}}\bar{x}_n}$$

2.2.4 Multinomial Distribution

The data follows a multinomial distribution. We begin by writing down the likelihood function for all the data:

$$L(\boldsymbol{\theta}|\mathbf{n}) = f_n(\mathbf{n}|\theta_1, \theta_2, ..., \theta_k) = \frac{n!}{n_1!n_2!\cdots n_k!}\theta_1^{n_1}\theta_2^{n_2}\cdots\theta_k^{n_k} \text{ for } \sum_{i=1}^{k} n_i = n$$

Taking the log of the likelihood function, we get:

$$\ell(\theta_1, ..., \theta_k) = \log(n!) - \sum_{i=1}^{k} \log(n_i!) + \sum_{i=1}^{k} n_i \cdot \log(\theta_i)$$

Before we can maximize the log of the likelihood function, we have to remember that we are maximizing it subject to the constraint that $\sum_{i=1}^{k} \theta_i = 1$ (a property of the multinomial distribution). Therefore, we will proceed by maximizing the following equation using Lagrange multipliers:

$$\Lambda(\theta_1, ..., \theta_k, \lambda) = L(\theta_1, ..., \theta_k) + \lambda \cdot (\sum_{i=1}^{k} \theta_i - 1)$$

Now, we solve $\nabla_{\theta_1,\theta_2,...,\theta_k,\lambda}\Lambda(\theta_1, ..., \theta_k, \lambda) = \nabla_{\theta_1,\theta_2,...,\theta_k,\lambda}\left(ln(n!) - \sum_{i=1}^{k} ln(n_i!) + \sum_{i=1}^{k} n_i \cdot ln(\theta_i) + \lambda \cdot (\sum_{i=1}^{k} \theta_i - 1)\right) = 0$ for θ_i and get n+1 first-order conditions:

$$\theta_i = \frac{n_i}{\lambda} \text{ for all } i = 1, 2, ..., k \text{ and } \sum_{i=1}^{k} \theta_i = 1$$

Since $\sum_{i=1}^{k} \theta_i = \sum_{i=1}^{k} \frac{n_i}{\lambda} = \frac{n}{\lambda} = 1, \lambda = n.$

$$\boxed{\therefore \hat{\theta}_{i,MLE} = \frac{n_i}{n} \text{ for all } i = 1, 2, ..., k}$$

Note: Alternatively, the same solution is obtained for $\hat{\theta}_{i,MLE}$ if we model being the ith type of individual as a success in a binomial distribution of n draws (where the failures are belonging to all the $k-1$ remaining types).

2.2.5 Uniform Distribution

We want to solve for the MLE of θ for a uniform distribution on the interval $[0, \theta]$. We begin by writing the likelihood function for θ, which is the joint density of the observed data conditional on θ:

$$f(\mathbf{x}|\theta) = \begin{cases} \frac{1}{\theta^n} & \text{for } 0 \leq x_i \leq \theta \ (i = 1, ..., n) \\ 0 & \text{otherwise} \end{cases}$$

Taking the log of the likelihood function:

$$L(\theta) = \begin{cases} -n \cdot \log(\theta) & \text{for } 0 \leq x_i \leq \theta \ (i = 1, ..., n) \\ 0 & \text{otherwise} \end{cases}$$

From the equation for $L(\theta)$ above, one can see that the MLE of θ must be a value of θ for which $0 \leq x_i \leq \theta$ for all $i = 1, ..., n$. Since $L(\theta)$ is a monotonically decreasing function of θ, we need the smallest value of θ such that $\theta \geq x_i$ for all $i = 1, ..., n$ in order to maximize the log likelihood. This value is $\theta = max(x_1, ..., x_n)$.

Therefore, the MLE is:
$\hat{\theta}_{MLE} = max(X_1, ..., X_n)$.

3 Properties of MLE: The Basics

Why do we use maximum likelihood estimation? It turns out that subject to **regularity conditions**, the following properties hold for the MLE (see proofs in Section 5):

1. **Consistency:** As sample size (n) increases, the MLE ($\hat{\theta}_{MLE}$) converges to the true parameter, θ_o:

$$\hat{\theta}_{MLE} \xrightarrow{p} \theta_o$$

2. **Normality:** As sample size (n) increases, the MLE is normally distributed with a mean equal to the true parameter (θ_o) and the variance equal to the inverse of the expected sample Fisher information at the true parameter (denoted as $\mathcal{I}_n(\theta_o)$):

$$\hat{\theta}_{MLE} \sim \mathcal{N}\left(\theta_o, \underbrace{\left(-\mathrm{E}\left[\frac{\partial^2 \ell(\theta|\mathbf{x})}{\partial \theta^2}\Big|_{\theta=\theta_o}\right]\right)^{-1}}_{\mathcal{I}_n(\theta_o)}\right)$$

   However, using the consistency property of the MLE and observed sample Fisher information, we can use the inverse of the observed sample Fisher information evaluated at the MLE, denoted as $\mathcal{J}_n(\hat{\theta}_{MLE})$ to approximate the variance. Note that the observed sample Fisher information, which will be defined in detail below, is the negation of the second derivative of the log-likelihood curve.

$$\hat{\theta}_{MLE} \sim \mathcal{N}\left(\theta_o, \underbrace{\left(-\left[\frac{\partial^2 \ell(\theta|\mathbf{x})}{\partial \theta^2}\Big|_{\theta=\hat{\theta}_{MLE}}\right]\right)^{-1}}_{\mathcal{J}_n(\hat{\theta}_{MLE})}\right)$$

3. **Efficiency:** As sample size (n) increases, MLE is the estimation procedure that generally provides the lowest variance.
4. In a finite sample, MLE gives the **minimum variance unbiased estimator**, or MVUE, if it exists.

3.1 Functional Invariance of MLE

A very useful property of the MLE is its **functional invariance**. The principle of invariance states that if we know that $\hat{\theta}$ is the MLE of θ, then the MLE of $g(\theta)$ is $g(\hat{\theta})$, where $g(\theta)$ is a function of θ.

Proof:
By definition of the MLE: $\hat{\theta}_{MLE} \in \Omega$ and $\ell(\hat{\theta}_{MLE}|\mathbf{x}) \geq \ell(\theta|\mathbf{x}) \forall \theta \in \Omega$. Thus, setting $\hat{\lambda} = g(\hat{\theta})$:

$$\ell(g^{-1}(\hat{\lambda})|\mathbf{x}) \geq \ell(g^{-1}(\lambda)|\mathbf{x}) \forall \lambda \in \Lambda$$

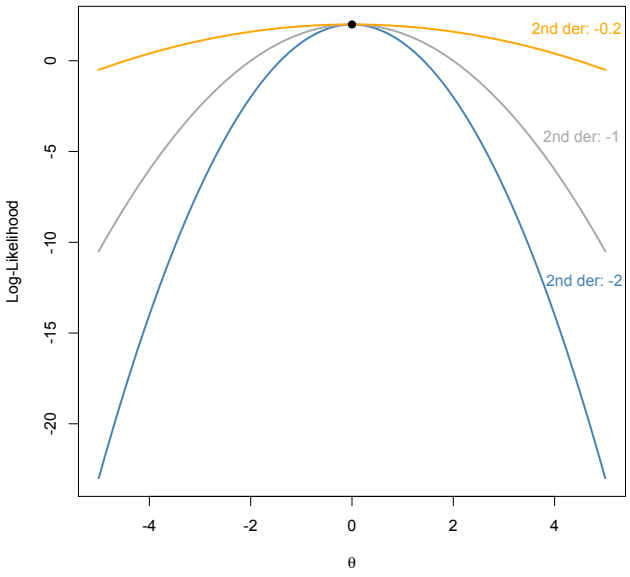
$\therefore \hat{\lambda} = g(\hat{\theta})$ is MLE of λ for a model with density $\ell(g^{-1}(\lambda)|\mathbf{x})$.

4 Fisher Information and Uncertainty of MLE

In addition to providing the MLE as a single point summary of the likelihood curve, we would like to quantify how certain we are in our estimate. We saw in the previous section that the variance of the MLE is asymptotically given by the inverse of the expected Fisher information. However, we usually approximate expected Fisher information with observed Fisher information. What are these two quantities? In this section, we'll define the observed and expected Fisher information, as well as state the intuition behind why these are useful quantities. The more theoretical derivations and proofs follow in subsequent sections.

First, let's develop some basic intuition regarding uncertainty of the MLE. Note that the curvature of the likelihood curve around the MLE contains information about how certain we are as to our estimate. Before we wade into the weeds of specific calculations, let's understand the intuition behind this. Looking at Figure 3, we can see that intuitively, we are more certain in our MLE if the curve has a steeper slope around the MLE than if it has a slope closer to 0. That is, we have more certainty in an MLE if the second derivative at the MLE is a more negative number. Recall that at the MLE, the slope of the log-likelihood function is 0, so the second derivative of the log-likelihood curve evaluated at the MLE captures how quickly the slope changes from 0 as you move away from the MLE in either direction. This comes from the interpretation of the second derivative as the rate of change of the slope.

Figure 3: Example of second derivatives.



How do we formalize this intuition? Let's define the **observed Fisher information**, termed $\mathcal{J}$, to be the negation of the second derivative of the log-likelihood function. Specifically, we will evaluate it at the MLE[1]:

$$\mathcal{J}(\hat{\theta}_{MLE}) = -\frac{\partial^2}{\partial\theta^2}\ell(\theta|\mathbf{x})\Big|_{\hat{\theta}_{MLE}} = -\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{x}|\theta)\Big|_{\hat{\theta}_{MLE}}$$

[1]Technically, we can calculate the observed Fisher information at any value of θ, but we will always talk about it as evaluated at the MLE.

Note that we negate the second derivative (which is always negative at the MLE) so that we always have a positive observed Fisher information. Now, the intuition we developed about the steepness of the curve follows through: the steeper the curve around the MLE, the larger the observed Fisher information.

Note that in the case of multiple parameters (we have a vector of k parameters $\boldsymbol{\theta}$), the observed Fisher information is the negation of the hessian, the matrix of second derivatives. Again, here we evaluate it at the MLE.

$$\mathcal{J}(\hat{\boldsymbol{\theta}}_{MLE}) = -\nabla\nabla^T \ell(\boldsymbol{\theta}|\boldsymbol{x})\big|_{\hat{\boldsymbol{\theta}}_{MLE}} = -\begin{pmatrix} \frac{\partial^2}{\partial\theta_1^2} & \frac{\partial^2}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2}{\partial\theta_1\partial\theta_k} \\ \frac{\partial^2}{\partial\theta_2\partial\theta_1} & \frac{\partial^2}{\partial\theta_2^2} & \cdots & \frac{\partial^2}{\partial\theta_2\partial\theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial\theta_k\partial\theta_1} & \frac{\partial^2}{\partial\theta_k\partial\theta_2} & \cdots & \frac{\partial^2}{\partial\theta_k^2} \end{pmatrix} \ell(\boldsymbol{\theta}|\boldsymbol{x})\big|_{\hat{\theta}_{MLE}}$$

What is the theoretical rationale for how the Fisher information is linked to the variance of the maximum likelihood estimate? It turns out that we can prove, using the Central Limit Theorem, that the MLE is asymptotically normal with a mean equal to the true parameter value and variance equal to the inverse of the expected Fisher information evaluated at the true parameter value (see proof). To understand this, we need to define the expected Fisher information. Having defined the observed Fisher information, we can define the **expected Fisher information** as the expectation of the observed Fisher information:

$$\mathcal{I}(\theta) = -\mathrm{E}\frac{\partial^2 \ell(\theta|\mathbf{x})}{\partial\theta^2}$$

The expected Fisher information is thus a function of θ – the parameter we are trying to estimate – that gives us the expected information across the samples we could draw from our distribution of interest. That is, imagine drawing 1,000,000 different samples (or even better, infinite samples!) from the distribution of interest (even though in reality we observe only one). Each sample will have a slightly different MLE and also a slightly different observed Fisher information (since observed Fisher information is a sample-specific quantity). The observed Fisher information we expect, on average, across all possible samples is the expected Fisher information. Moreover, note that since the MLEs are different from sample to sample, we also have a variance across MLEs – this is in fact the variance we are after (recall that in the frequentist inferential framework, all randomness comes from sampling and the parameters are fixed)! The inverse of the expected Fisher information captures this variance. For an example that will elucidate these concepts, see Figure 4.

We can also be more specific as to what kind of Fisher information we are dealing with. Specifically, we can define the **sample expected Fisher information** across all the random variables in our sample and the **unit expected Fisher information** for just one random variable.

The unit expected Fisher information is defined for one random variable from our distribution:

$$\mathcal{I}(\theta) = -\mathrm{E}\frac{\partial^2}{\partial\theta^2}\ell(\theta|x)$$

The sample expected Fisher information is defined as:

$$\mathcal{I}_n(\theta) = -\mathrm{E}\frac{\partial^2}{\partial\theta^2}\ell(\theta|\boldsymbol{x})$$

Note that for distinction, we have added an n subscript to explicitly differentiate the sample Fisher information from the unit Fisher information.

Since our sample is a set of n iid random variables, we can relate the sample Fisher information to the unit Fisher information using the linearity of expectation and the fact that we can bring the derivative operator within the summation sign :

$$\mathcal{I}_n(\theta) = -\mathrm{E}\frac{\partial^2}{\partial\theta^2}\ell(\theta|\boldsymbol{x}) = \mathrm{E}\frac{\partial^2}{\partial\theta^2}\Big[\sum_{i=1}^{n} f(x_i|\theta)\Big] = \sum_{i=1}^{n}\mathrm{E}\Big[\frac{\partial^2}{\partial\theta^2}f(\mathbf{x}|\theta)\Big] = n\cdot\mathrm{E}\Big[\frac{\partial^2}{\partial\theta^2}f(\mathbf{x}|\theta)\Big] = n\mathcal{I}(\theta)$$

$$\therefore \mathcal{I}_n(\theta) = n\mathcal{I}(\theta)$$

This tells us that for a sample of iid random variables, the expected sample information is just a sum of the individual expected informations across the n observations. This relationship becomes important in the proofs of the asymptotic distribution of the MLE.

Finally, you may ask how we can get away with using the observed Fisher information even though we have just stated that the MLE is asymptotically distributed normally with a variance equal to the inverse of the *expected* Fisher information? It in fact turns out that the observed Fisher information is consistent for the expected Fisher information. That is, we can prove using the law of large numbers that the observed information converges to the expected Fisher information as the sample size increases. Moreover, we can evaluate the observed Fisher information at the MLE instead of at the true (unknown) value of the parameter because of the consistency of the MLE (proven below). In most applications, we thus use the observed Fisher information.

4.0.1 Example of Calculating Fisher Information: Bernoulli Distribution

Suppose that we have a sample $\mathbf{X} = \{X_1, ..., X_n\}$ such that $X \sim Bern(p)$ (iid).

Let's write down the log-likelihood and solve for the MLE:

$$\ell(p) = \log\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$\ell(p) = \log p^{\sum x_i}(1-p)^{n-\sum x_i}$$

$$\ell(p) = \sum x_i \log p + (n - \sum x_i)\log(1-p)$$

Using the first order condition to solve for the MLE:

$$\frac{\partial\ell(p)}{\partial p} = \frac{\sum x_i}{p} - \frac{n-\sum x_i}{1-p} = 0$$

$$\hat{p}_{MLE} = \bar{X}$$

What is the second derivative of the log-likelihood?

$$\frac{\partial^2 \ell(p)}{\partial p^2} = -\frac{\sum x_i}{p^2} - \frac{n - \sum x_i}{(1-p)^2}$$

We now define expected Fisher information as the expected value of the negation of the second derivative

$$\mathcal{I}_n(p) = \mathrm{E}\left[\frac{\sum x_i}{p^2} + \frac{n - \sum x_i}{(1-p)^2}\right]$$

Since $\mathrm{E}[X] = p$:

$$\mathcal{I}_n(p) = \frac{np}{p^2} + \frac{n - np}{(1-p)^2}$$

Simplifying:

$$\mathcal{I}_n(p) = \frac{n}{p} + \frac{n}{1-p} = \frac{n}{p(1-p)}$$

We have found the *expected* sample fisher information for $X_1, ..., X_n$ for $\mathrm{Bern}(p)$.

The reason that this quantity is not particularly tractable for calculation of uncertainty is that it depends on p, the unknown parameter! Instead, we can use the observed Fisher information, evaluated at the MLE:

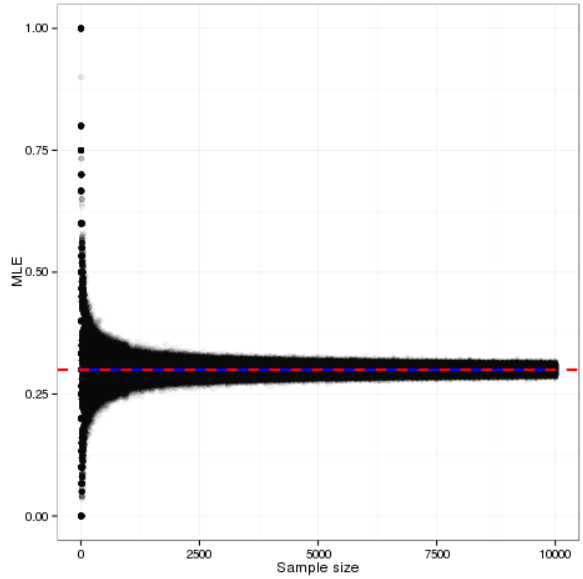
$$\mathcal{J}(\hat{p}_{MLE}) = \frac{n}{\hat{p}_{MLE}(1 - \hat{p}_{MLE})} = \frac{n}{\bar{X}(1-\bar{X})}$$

Asymptotically, we thus know that $\hat{p}_{MLE}$ has the following approximate distribution:

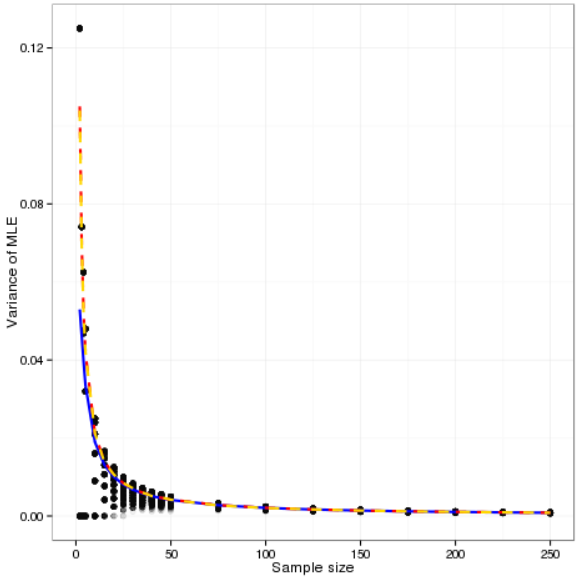
$$\hat{p}_{MLE} \sim \mathcal{N}\left(p, \mathcal{J}(\hat{p}_{MLE})^{-1}\right)$$

We can easily use this to calculate confidence intervals and test statistics.

Figure 4: Illustration of consistency properties of MLE and observed Fisher information. For this illustration, the sample size (n) was varied and at each n, 10,000 datasets were drawn from a Bern(0.3) distribution. For each dataset, the MLE and the observed Fisher information were calculated.



(a) Simulation of MLEs from Bern(0.3) model by sample size (n). For each n, 10,000 MLEs are plotted with black dots. The dotted red line denotes the true parameter (p = 0.3), while the dotted blue line represents the mean MLE across the 10,000 samples at each value of n.



(b) Simulation of observed Fisher information from Bern(0.3) model by sample size (n). For each n, 10,000 observed Fisher informations are plotted with black dots. The dotted red line denotes the expected Fisher information, derived in the example above. The gold line represents the simulated variance across the 10,000 MLEs at each sample size. Finally, the solid blue line represents the mean of the observed Fisher informations at each value of n.

4.1 The Theory of Fisher Information

4.1.1 Derivation of Unit Fisher Information

Let's begin by deriving the expected Fisher information for one random variable, X. To do this derivation, we need to impose some **regularity conditions** on the distribution of the random variable X:

- $f(x|\theta)$ is the pdf of X, where $X \in S$ (sample space) and $\theta \in \Omega$ (parameter space)
- Assume that $f(x|\theta) > 0$ for each value $x \in S$ and each value $\theta \in \Omega$
- Assume that the pdf of X is a twice differentiable function of θ

Recall that by definition of a pdf, the integral of a continuous density across the sample space S is 1:

$$\int_S f(x|\theta)\partial x = 1$$

Let's assume that we are able to distribute a derivative operator within the integration operator, such that:

$$\frac{\partial}{\partial x}\int_S f(x|\theta)\partial x = \int_S \frac{\partial}{\partial x} f(x|\theta)\partial x = \int_S f'(x|\theta)\partial x$$

and

$$\frac{\partial^2}{\partial x^2}\int_S f(x|\theta)\partial x = \int_S \frac{\partial^2}{\partial x^2} f(x|\theta)\partial x = \int_S f''(x|\theta)\partial x$$

Recall that the score of the log-likelihood is defined to be the first derivative of the log-likelihood:

$$\mathcal{S} = \frac{\partial}{\partial \theta}\log L(\theta|x) = \frac{\partial}{\partial \theta}\log f(x|\theta)$$

We also know the following about the score from the properties of derivatives, specifically the chain rule:

$$\ell'(\theta|x) = \frac{\partial}{\partial \theta}\log f(x|\theta) = \frac{1}{f(x|\theta)} \cdot f'(x|\theta)$$

$$\ell'(\theta|x) = \frac{f'(x|\theta)}{f(x|\theta)}$$

We can find the expected value of the score using the definition of expectation:

$$\mathrm{E}_\theta\left[\frac{\partial \ell(\theta|x)}{\partial \theta}\right] = \int_S \ell'(\theta|x) f(x|\theta)\partial x = \int_S \frac{f'(x|\theta)}{f(x|\theta)} f(x|\theta)\partial x = \int_S f'(x|\theta)\partial x$$

Using our ability to exchange the order of integration and taking a derivative:

$$\int_S f'(x|\theta)\partial x = \frac{\partial}{\partial \theta}\int_S f(x|\theta)\partial x = \frac{\partial}{\partial \theta} 1 = 0$$

$$\therefore \mathrm{E}_\theta\left[\ell'(\theta|x)\right] = 0$$

We have just shown that in expectation (across samples), the score will be 0.

Suppose that we define the expected unit Fisher information for random variable X as the expectation of the squared score (you'll see shortly how it relates to our previous expression for the expected information).

$$\mathcal{I}(\theta) = \mathrm{E}_\theta\left[\left(\frac{\partial \ell(\theta|x)}{\partial \theta}\right)^2\right] = \mathrm{E}_\theta[(\ell'(\theta|x))^2]$$

However, using the fact that $\mathrm{E}_\theta\left[\ell'(\theta|x)\right] = 0$ and the definition of variance ($\mathrm{Var}(Y) = \mathrm{E}[Y^2] - (\mathrm{E}[Y])^2$ for any random variable Y), the Fisher information can also be written as:

$$\mathcal{I}(\theta) = \mathrm{E}_\theta\left[(\ell'(\theta|x))^2\right] = \mathrm{E}_\theta\left[(\ell'(\theta|x))^2\right] - \underbrace{(\mathrm{E}_\theta\left[\ell'(\theta|x)\right])^2}_{\mathrm{E}_\theta[\ell'(\theta|x)]=0} = \mathrm{Var}[\ell'(\theta|x)]$$

We have just shown that the expected unit Fisher information is equivalent to the variance of the score:

$$\therefore \mathcal{I}(\theta) = \mathrm{Var}[\ell'(\theta|x)] = \mathrm{Var}(\mathcal{S}(\theta))$$

Since the first moment of the score (the expected value) is zero, the Fisher information is also the *second moment of the score*.

We can also derive the (more familiar) expression for the expected information in terms of the second derivative of the log-likelihood. First, let us define the second derivative of the log-likelihood (using the quotient rule) as:

$$\ell''(\theta|x) = \frac{f(x|\theta)f''(x|\theta) - [f'(x|\theta)]^2}{[f(x|\theta)]^2}$$

Separating the result into two fractions and simplifying:

$$\ell''(\theta|x) = \frac{f(x|\theta)f''(x|\theta)}{[f(x|\theta)]^2} - \frac{[f'(x|\theta)]^2}{[f(x|\theta)]^2} = \frac{f''(x|\theta)}{f(x|\theta)} - [\ell'(\theta|x)]^2$$

The final step above follows from the fact that $\ell'(\theta|x) = \frac{f'(x|\theta)}{f(x|\theta)}$.

Taking the expected value of both sides:

$$\mathrm{E}[\ell''(\theta|x)] = \underbrace{\mathrm{E}[\frac{f''(x|\theta)}{f(x|\theta)}]}_{\int_S f''(x|\theta)dx=0} - \underbrace{\mathrm{E}[[\ell'(\theta|x)]^2]}_{\mathcal{I}(\theta)}$$

The result is a more familiar version of the expected unit Fisher information:

$$\mathcal{I}(\theta) = -\mathrm{E}[\ell''(\theta|x)]$$

4.1.2 Derivation of Sample Fisher Information

Let $X_1, ..., X_n$ be an iid sample from the distribution $f(x|\theta)$. We can then expand the expression for the expected unit Fisher information to the case of a sample of random variables.

In an iid sample, we define the log-likelihood as:

$$\log L(\theta|\mathbf{x}) = \ell(\theta|\mathbf{x}) = \sum_{i=1}^{n} \ell(\theta|x_i)$$

Similarly, the score and second derivative of the log-likelihood for the sample can be expressed in terms of sums of unit scores and second-derivatives:

$$\mathcal{S} = \frac{\partial}{\partial\theta} \log L(\theta|\mathbf{x}) = \ell'(\theta|\mathbf{x}) = \sum_{i=1}^{n} \ell'(\theta|x_i)$$

$$\frac{\partial^2}{\partial\theta^2} \log L(\theta|\mathbf{x}) = \ell''(\theta|\mathbf{x}) = \sum_{i=1}^{n} \ell''(\theta|x_i)$$

Using these expressions and the linearity of expectation, the expected value of the negative of the second derivative of the log-likelihood function is:

$$\underbrace{\mathrm{E}[-\ell(\theta|\mathbf{x})]}_{\mathcal{I}_n(\theta)} = \sum_{i=1}^{n} \underbrace{\mathrm{E}[-\ell''(\theta|x_i)]}_{=\mathcal{I}(\theta)}$$

But recall that $\mathrm{E}[-\ell(\theta|\mathbf{x})]$ is just the definition of the expected sample Fisher information, denoted as $\mathcal{I}_n(\theta)$. We just proved that the following relation holds true for an iid sample of n random variables:

$$\mathcal{I}_n(\theta) = n \cdot \mathcal{I}(\theta)$$

4.1.3 Relationship between Expected and Observed Fisher Information

The observed Fisher information is defined as the negation of the second derivative of the log-likelihood:

$$\mathcal{J}_n(\theta) = -\frac{\partial^2}{\partial\theta^2} \ell(\theta|\mathbf{x}) = -\frac{\partial^2}{\partial\theta^2} \log f(\mathbf{x}|\theta)$$

We can rewrite the observed Fisher information as the sum of second derivatives:

$$\mathcal{J}_n(\theta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f(x_i|\theta)$$

$X_1, X_2, ..., X_n$ are iid hence the second derivatives on the right hand side of the expression above are also iid Thus, by the **law**

of large numbers, their average converges to the expectation of a single term:

$$\frac{1}{n}\mathcal{J}_n(\theta) \xrightarrow{p} \mathrm{E}\Big[\frac{\partial^2}{\partial\theta^2}\log f(x_i|\theta)\Big] = \mathcal{I}(\theta)$$

Therefore, when n is large:

$$\mathcal{J}_n(\theta) \approx n \cdot \mathcal{I}(\theta) = \mathcal{I}_n(\theta)$$

So by the consistency we have just shown, we can use $\mathcal{J}_n(\theta)$ instead of $\mathcal{I}_n(\theta)$. However, we still do not know θ_0, the true value of the parameter (remember that the variance of the MLE asymptotically is $\mathcal{I}_n(\theta_0)^{-1}$. But it turns out $\hat{\theta}_{MLE}$ is a consistent estimator for θ_0, and as a result we use $\mathcal{J}_n(\theta_0)^{-1}$ as an estimator for the variance of the MLE.

4.1.4 Extension to Multiple Parameters

All the proofs above are generalizable to the case of a vector of k parameters, $\boldsymbol{\theta}$. The multivariate extension of the score (first derivative of log-likelihood curve) is:

$$\mathcal{S}(\boldsymbol{\theta}) = \nabla\ell(\boldsymbol{\theta}|\boldsymbol{x}) = \begin{pmatrix} \frac{\partial\ell(\boldsymbol{\theta}|\boldsymbol{x})}{\theta_1} \\ \frac{\partial\ell(\boldsymbol{\theta}|\boldsymbol{x})}{\theta_2} \\ \vdots \\ \frac{\partial\ell(\boldsymbol{\theta}|\boldsymbol{x})}{\theta_k} \end{pmatrix}$$

In the multivariate case, we define expected sample Fisher information as:

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = -\mathrm{E}\Big[\nabla^2\ell(\boldsymbol{\theta}|\boldsymbol{x})\Big] = -\mathrm{E}\left[\begin{pmatrix} \frac{\partial^2}{\partial\theta_1^2} & \frac{\partial^2}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2}{\partial\theta_1\partial\theta_k} \\ \frac{\partial^2}{\partial\theta_2\partial\theta_1} & \frac{\partial^2}{\partial\theta_2^2} & \cdots & \frac{\partial^2}{\partial\theta_2\partial\theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial\theta_k\partial\theta_1} & \frac{\partial^2}{\partial\theta_k\partial\theta_2} & \cdots & \frac{\partial^2}{\partial\theta_k^2} \end{pmatrix}\ell(\boldsymbol{\theta}|\boldsymbol{x})\right]$$

We can show, analagously to the case with a single parameter, that the expected sample Fisher information is equal to the variance of the score:

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = -\mathrm{E}\Big[\nabla^2\ell(\boldsymbol{\theta}|\boldsymbol{x})\Big] = \mathrm{Var}\Big[\nabla\ell(\boldsymbol{\theta}|\boldsymbol{x})\Big] = \mathrm{Var}\Big[\mathcal{S}(\boldsymbol{\theta})\Big]$$

5 Proofs of Asymptotic Properties of MLE

In this section, we want to prove the asymptotic properties of the MLE: consistency, normality, and efficiency.

We first introduce some notation. Let $X_n = X_1, ..., X_n$ be our random sample, where the X_i's are iid. Let $\hat{\theta}_{MLE}$ be the maximum likelihood estimator for the parameter θ. θ_o is the true underlying value of the parameter. Note that $\hat{\theta}_{MLE}, \theta_o, \theta \in \Omega$ (the parameter space).

Furthermore, to prove the asymptotic properties of the maximum likelihood estimator, we will introduce several **regularity conditions**:

- The parameter space must be a bounded and closed set. That is, Ω must be a *compact subset*.
- The true value of the parameter, θ_o, must be an interior point of the parameter set: $\theta_o \in int(\Omega)$. Phrased differently, θ_o cannot be on the boundary of the set.
- The likelihood function is continuous in θ.
- The likelihood function is twice-differentiable in the neighborhood of θ_o.
- Integration and differentiation is interchangeable (as we defined above for the derivation of expected Fisher information).

5.1 Consistency of MLE

We want to show that the MLE ($\hat{\theta}_{MLE}$) converges in probability to the true value of the parameter (θ_o):

$$\hat{\theta}_{MLE} \xrightarrow{p} \theta_o$$

Since the observations in our sample are iid, we can write the log-likelihood as the sum of log-likelihoods for each observation x_i:

$$\ell(\theta|\mathbf{x}) = \sum_{i=1}^{n} \ell(\theta|x_i)$$

Let's divide by n, which we can do since it doesn't affect the maximization of the log-likelihood. Now, we have an expression that looks like the average of log-likelihoods across all the Xs. We can then show by the strong law of large numbers that that converges to the expected value of a log-likelihood of a single X:

$$\frac{1}{n}\sum_{i=1}^{n} \ell(\theta|x_i) \xrightarrow{a.s.} \mathrm{E}_{\theta_o}\ell(\theta|x) = \mathrm{E}_{\theta_o} \log f(x|\theta)$$

In this expression, E_{θ_o} represents the expectation of the density with respect to the true unknown parameter and thus we define a new function $\mathcal{L}(\theta)$, which is the expected log-likelihood function:

$$\mathcal{L}(\theta) = \mathrm{E}_{\theta_o} \log f(x|\theta) = \int_{-\infty}^{\infty} \log f(x|\theta) \cdot f(x|\theta_o)\partial x$$

As a result, the normalized log-likelihood converges to the expected log-likelihood function $\mathcal{L}(\theta)$ for any value of θ. This expression depends solely on θ and not on x since we integrate it out.

Now, let's look at the divergence between $\mathcal{L}(\theta)$ and $\mathcal{L}(\theta_0)$ (the expected log-likelihood function evaluated at an arbitrary parameter θ and the true parameter θ_0):

$$\mathcal{L}(\theta) - \mathcal{L}(\theta_0) = \mathrm{E}_{\theta_0}[\log f(x|\theta) - \log f(x|\theta_0)] = \mathrm{E}_{\theta_0}\Big[\log \frac{f(x|\theta)}{f(x|\theta_0)}\Big]$$

By Jensen's inequality:

$$\mathrm{E}_{\theta_0}\Big[\log \frac{f(x|\theta)}{f(x|\theta_0)}\Big] \leq \log \mathrm{E}_{\theta_0}\Big[\frac{f(x|\theta)}{f(x|\theta_0)}\Big] = \log \int_{-\infty}^{\infty} \frac{f(x|\theta)}{f(x|\theta_0)} \cdot f(x|\theta_0)dx = \log \underbrace{\int_{-\infty}^{\infty} f(x|\theta)dx}_{=1 \text{ by def. of pdf}} = 0$$

Thus:

$$\mathcal{L}(\theta) - \mathcal{L}(\theta_0) \leq 0$$

$$\mathcal{L}(\theta) \leq \mathcal{L}(\theta_0)$$

This inequality suggests that the expected log-likelihood when assuming that an arbitrary parameter θ is governing the data generation process is no greater than the expected log-likelihood when you correctly identify the true parameter θ_0 governing the data generation process. Note that this is closely related to the concept of the Kullback-Leibler divergence. In fact, we know from Gibbs' inequality that the Kullback-Leibler divergence between $f(x|\theta_0)$ and $f(x|\theta)$ must be non-negative:

$$D_{KL}(f(x|\theta_0)||f(x|\theta)) = \mathrm{E}_{\theta_0}\Big[\log \frac{f(x|\theta_0)}{f(x|\theta)}\Big] \geq 0$$

In more practical terms, this inequality suggests that no distribution describes the data as well as the true distribution that generated it. Therefore, on average, the greatest log-likelihood will be the one that is a function of the true parameter θ_0. Phrased differently, θ_0 is the maximizer of the expected log-likelihood, $\mathcal{L}(\theta)$.

Now, let's put the different pieces together. Recall that by the strong law of large numbers:

$$\frac{1}{n}\sum_{i=1}^{n} \ell(\theta|x_i) \overset{a.s.}{\to} \mathrm{E}_{\theta_0} \log f(x|\theta)$$

For a finite parameter space Ω, the following holds for the MLE (convergence is uniform from the uniform strong law of large numbers):

$$\hat{\theta}_{MLE} = \sup_{\theta \in \Omega} \frac{1}{n}\sum_{i=1}^{n} \ell(\theta|x_i) \overset{a.s.}{\to} \sup_{\theta \in \Omega} \mathrm{E}_{\theta_0} \log f(x|\theta) = \theta_0$$

$$\therefore \hat{\theta}_{MLE} \overset{a.s.}{\to} \theta_0$$

In sum, by the strong law of large numbers, the MLE is actually maximizing the expected log-likelihood as n increases asymptotically. But we showed that the expected log-likelihood is maximized at the true value of the parameter. Therefore, as $n \to \infty$,

the normalized log-likelihood of the data should approach the expected value of the log-likelihood of the random variable X. Another way of stating the consistency of the MLE is that it minimizes the Kullback-Leibler divergence between an arbitrary log-likelihood function and the true log-likelihood. If $\hat{\theta}_{MLE} = \theta_0$, the Kullback-Leibler divergence goes to 0 (by LLN).

Note also that the consistency of the MLE can also be proven for an infinite parameter space Ω as well as for non-compact parameter spaces.

5.2 Asymptotic Normality of MLE

We want to prove that the MLE, $\hat{\theta}_{MLE}$, is asymptotically normal:

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} \mathcal{N}(0, (\mathcal{I}_n)^{-1})$$

Ultimately, we want to rely on the fact that the score (first derivative of the log-likelihood) is a sum of iid terms, and as a result we know that its asymptotic distribution can be approximated using the law of large numbers and the central limit theorem. We can start the proof by noting that since the MLE maximizes the log-likelihood, the score is equal to zero at the MLE:

$$\mathcal{S}(\hat{\theta}_{MLE}) = \ell'(\hat{\theta}_{MLE}|\mathbf{x}) = 0$$

When the log-likelihood is twice differentiable, we can expand the score around the true parameter value θ_0 using a Taylor series approximation of the 1st order:

$$\mathcal{S}(\theta) = \ell'(\theta) = \ell'(\theta_0) + \ell''(\tilde{\theta})(\theta - \theta_0),$$

where $\tilde{\theta}$ is some point between θ and θ_0. More precisely:

$$\tilde{\theta} = \alpha\theta + (1 - \alpha)\theta_0, \text{ where } \alpha \in [0, 1]$$

Now we can plug the MLE in for θ and since we know that the score at evaluated at the MLE is 0, the expression simplifies to:

$$\ell'(\hat{\theta}_{MLE}) = 0 = \ell'(\theta_0) + \ell''(\tilde{\theta})(\hat{\theta}_{MLE} - \theta_0)$$

Note that in this case, $\tilde{\theta}$ is an arbitrary point located between the MLE ($\hat{\theta}_{MLE}$) and the true parameter (θ_0). One could also obtain this equation by utilizing the mean-value theorem which suggests that for a continuous function $f(x)$ there is a point c on the interval $[a, b]$ such that:

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

Applying the mean value theorem to the function $\ell'(\theta)$ and letting $a = \hat{\theta}_{MLE}$ and $b = \theta_0$:

$$\ell''(\tilde{\theta}) = \frac{\ell'(\hat{\theta}_{MLE}) - \ell'(\theta_0)}{\hat{\theta}_{MLE} - \theta_0}, \text{where } \tilde{\theta} \in [\hat{\theta}_{MLE}, \theta_0]$$

$$\ell''(\tilde{\theta})(\hat{\theta}_{MLE} - \theta_0) + \ell'(\theta_0) = \ell'(\hat{\theta}_{MLE}) = 0, \tilde{\theta} \in [\hat{\theta}_{MLE}, \theta_0]$$

In either case, we obtain the following relationship:

$$\hat{\theta}_{MLE} - \theta_0 = -\frac{\ell'(\theta_0)}{\ell''(\tilde{\theta})}$$

Multiplying both sides by $\sqrt{n}$:

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) = -\frac{\sqrt{n}\ell'(\theta_0)}{\ell''(\tilde{\theta})}$$

Let's consider the asymptotic distribution of the numerator and denominator in turn, starting with the asymptotic distribution of the numerator. We can first express the numerator as a sum of iid scores (first derivatives of the log-likelihood of iid random variables):

$$\ell'(\theta_0) = \sum_{i=1}^{n} \ell'(\theta_0|x_i)$$

Using the Central Limit Theorem, we will be able to make a statement regarding the asymptotic distribution of the average of the score, $\frac{1}{n}\ell'(\theta_0)$. First, though, note that in the section on Fisher information above, we proved that the first moment of the score is zero and the second moment of the score is the expected sample Fisher information:

$$\mathrm{E}_\theta[\ell'(\theta|\mathbf{x})] = 0$$

and

$$\mathrm{Var}_\theta[\ell'(\theta|\mathbf{x})] = -\mathrm{E}_\theta[\ell''(\theta|\mathbf{x})] = \mathcal{I}_n(\theta)$$

For a single observation, the first moment is 0 and the second moment is $\mathcal{I}_1(\theta)$. Invoking the Central Limit Theorem, we know that the mean of the score is distributed as a normal with a mean of 0 and a variance of $\mathcal{I}_1(\theta)/n$.

$$\frac{1}{n}\ell'(\theta) \xrightarrow{d} \mathcal{N}\Big(0, \frac{\mathcal{I}_1(\theta)}{n}\Big)$$

This implies that $\ell'(\theta)$ is distributed as a normal with a mean of 0 and a variance of $\mathcal{I}_n(\theta)$:

$$\ell'(\theta) \xrightarrow{d} \mathcal{N}\big(0, \mathcal{I}_n(\theta)\big)$$

Multiplying this by $\sqrt{n}$ to get the expression in the numerator:

$$\sqrt{n}\ell'(\theta) \xrightarrow{d} \mathcal{N}\big(0, n \cdot \mathcal{I}_n(\theta)\big)$$

Now, we can find the asymptotic behavior of the denominator – $\ell''(\tilde{\theta}|\mathbf{x})$ – in the Taylor series expansion of the score. We can use the law of large numbers to derive the following convergence in probability for any θ:

$$\frac{1}{n}\ell''(\tilde{\theta}|\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\ell''(\tilde{\theta}|x_i)}_{\mathrm{E}[\ell''(\tilde{\theta}|x_i)]=-\mathcal{I}_1(\tilde{\theta})} \xrightarrow{p} -\mathcal{I}_1(\tilde{\theta})$$

Moreover, since we know from the consistency of the MLE that $\hat{\theta}_{MLE} \xrightarrow{a.s} \theta_0$ and $\tilde{\theta} \in [\hat{\theta}_{MLE}, \theta_0]$, it follows that $\tilde{\theta} \xrightarrow{a.s} \theta_0$. Therefore:

$$\ell_n''(\tilde{\theta}) \xrightarrow{a.s} -\mathcal{I}_n(\theta_0)$$

Combining the asymptotic behavior of the numerator and denominator:

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) = -\frac{\sqrt{n}\ell'(\theta_0)}{\ell''(\tilde{\theta})} \xrightarrow{d} \mathcal{N}\big(0, \frac{n \cdot \mathcal{I}_n(\theta_0)}{(-\mathcal{I}_n(\theta_0))^2}\big) = \mathcal{N}\big(0, \frac{n}{\mathcal{I}_n(\theta_0)}\big)$$

In sum, we have shown:

$$\hat{\theta}_{MLE} \sim \mathcal{N}\big(\theta_0, \frac{1}{\mathcal{I}_n(\theta_0)}\big)$$

5.2.1 Efficiency of MLE

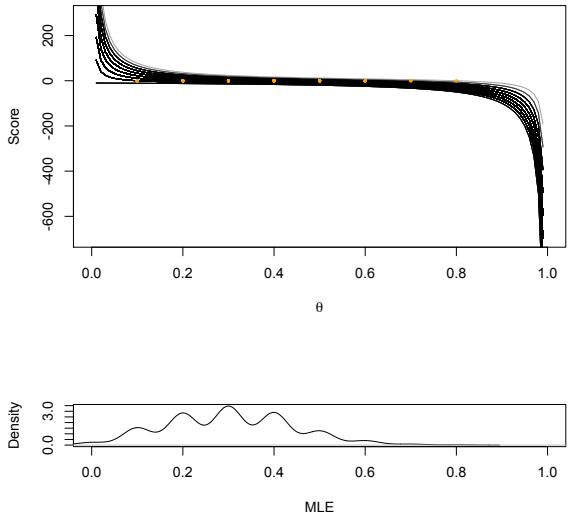
To show the efficiency of the MLE we need to show that the variance of the MLE is equal to the Cramer-Rao lower bound. However, we know that the Cramer-Rao lower bound is just the inverse of the Fisher information, $\mathcal{I}(\theta)$.

Since $\mathrm{Var}(\hat{\theta}_{MLE}) = \mathcal{I}(\theta)^{-1} = -\frac{1}{\mathrm{E}\big[\frac{\partial^2 \ell(\theta|\mathbf{x})}{\partial \theta^2}\big]}$, the efficiency of the MLE is:

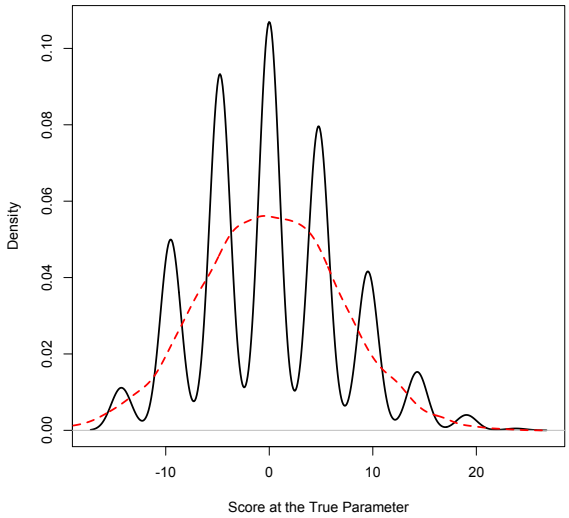
$$e(\hat{\theta}_{MLE}) = \frac{\mathcal{I}(\theta)^{-1}}{\mathrm{Var}(\hat{\theta}_{MLE})} = 1$$

$\therefore \hat{\theta}_{MLE}$ is an efficient estimator.

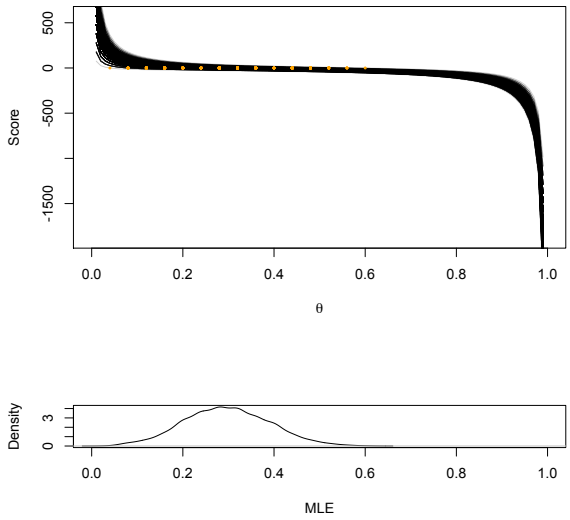
Figure 5: Illustration of Central Limit Theorem for the MLE and the score evaluated at the true parameter. For this illustration, we simulated 1000 datasets of sample size $n \in \{10, 25, 100\}$ from the Bern(0.3) distribution. For each dataset, we plotted the score function (as a function of θ) and the MLE. We also evaluated the score function at the true value of the parameter, $\theta_0 = 0.3$.



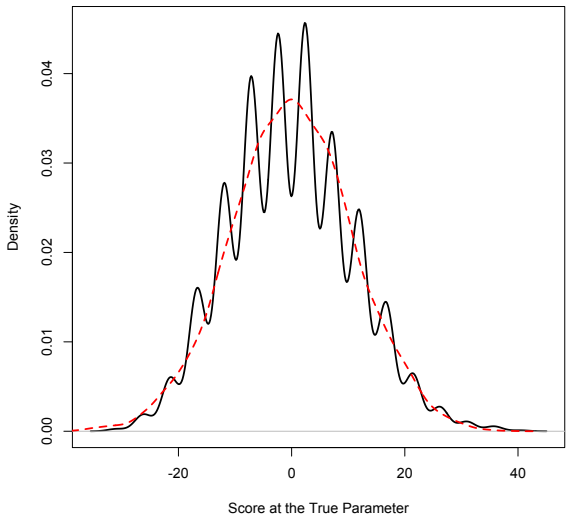
(a) The top plot portrays the 1000 score functions for simulated data with sample size $n = 10$. The MLEs are indicated with orange dots. The lower plot depicts the density function of the MLEs across the 1000 datasets. Note that the density of the MLEs does not yet look particularly normal.



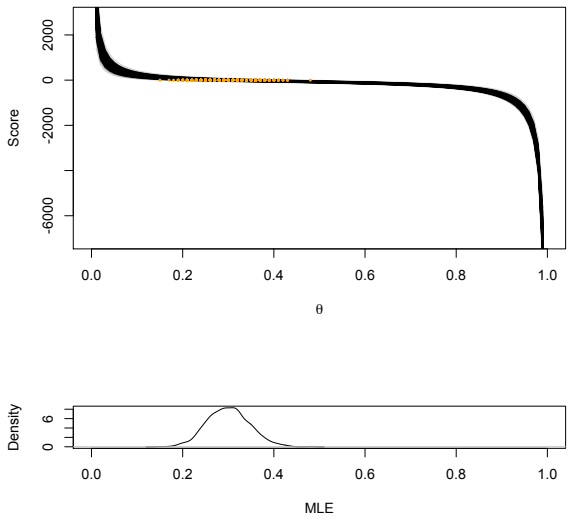
(b) A density plot of the score function evaluated at the true parameter, θ_0, across 1000 simulated datasets of sample size $n = 10$ (in black). For comparison in red, we have plotted the theoretical asymptotic density of the score, $\mathcal{N}(0, \mathcal{I}_n(\theta_0))$, derived using the CLT. Clearly the simulated density has not converged to the normal density.



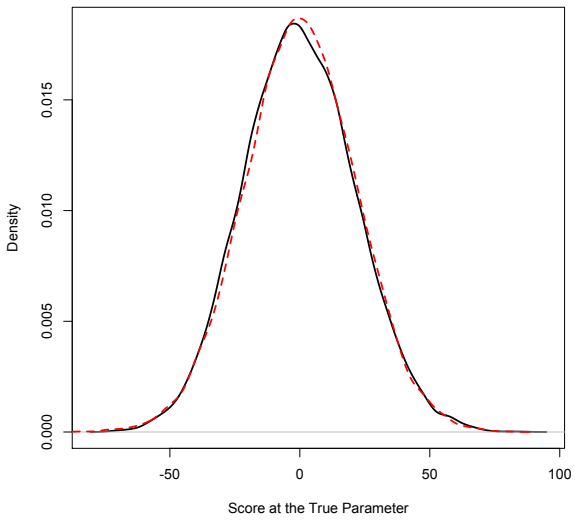
(c) The top plot portrays the 1000 score functions for simulated data with sample size $n = 25$. The MLEs are indicated with orange dots. The lower plot depicts the density function of the MLEs across the 1000 datasets.



(d) A density plot of the score function evaluated at the true parameter, θ_0, across 1000 simulated datasets of sample size $n = 25$ (in black). For comparison in red, we have plotted the theoretical asymptotic density of the score, $\mathcal{N}(0, \mathcal{I}_n(\theta_0))$, derived using the CLT. The simulated density is starting to converge to the expected normal density.



(a) The top plot portrays the 1000 score functions for simulated data with sample size $n = 100$. The MLEs are indicated with orange dots. The lower plot depicts the density function of the MLEs across the 1000 datasets. The MLE has essentially converged to the expected distribution.



(b) A density plot of the score function evaluated at the true parameter, θ_0, across 1000 simulated datasets of sample size $n = 100$ (in black). For comparison in red, we have plotted the theoretical asymptotic density of the score, $\mathcal{N}(0, \mathcal{I}_n(\theta_0))$. As predicted by the CLT, the distribution of the scores has essentially converged to the expected density.

6 References

Casella, George. and Roger L. Berger. *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury Press, 2002.

DeGroot, Morris H., and Mark J. Schervish. *Probability and Statistics*. 3rd ed. Boston, MA: Addison-Wesley, 2002.

Newey, Whitney K. and Daniel McFadden. Large sample estimation and hypothesis testing. In Engle, Robert F. and Daniel L. McFadden, editors, *Handbook of Econometrics*, vol. 4, 1994.

Convergence of Random Variables

What does it mean to say that a sequence converges? There are several notions of convergence for random variables. The two main ones are convergence in probability and convergence in distribution.

Convergence in Probability

Suppose we have a sequence of random variables denoted by $\{X_n\} = X_1, X_2, ..., X_n$. The sequence converges in probability to X if the probability distribution of the sequence $\{X_n\}$ is increasingly concentrated around X:

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1$$
or
$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$$

for every $\epsilon > 0$.

Convergence in probability is denoted as $X_n \xrightarrow{p} X$.

Note that convergence in probability implies convergence in distribution, but convergence in distribution implies convergence in probability only when the limiting variable X is a constant.

We can extend convergence in probability to the multivariate case. If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} X$, then $(X_n, Y_n) \xrightarrow{p} (X, Y)$.

Almost Sure Convergence

Suppose we have a sequence of random variables denoted by $\{X_n\} = X_1, X_2, ..., X_n$. The sequence converges almost surely to X if:

$$\lim_{n\to\infty} P\left(\sup_{m\geq n}|X_m - X| \geq \epsilon\right) = 0 \text{ for every } \epsilon > 0.$$

Almost sure convergence is denoted as $X_n \xrightarrow{a.s.} X$.

Note that almost sure convergence implies convergence in probability, but not vice versa.

Convergence in Distribution

Suppose we have a sequence of random variables denoted by $\{X_n\} = X_1, X_2, ..., X_n$. Let F_n denote the CDF of random variable X_n and F^* denote the CDF of random variable X^*.

The sequence converges in distribution to a random variable X if:

$$\boxed{\lim_{n\to\infty} F_n(x) = F^*(x) \text{ for every number } x \in \mathbf{R} \text{ where } F^* \text{ is continuous}}$$

Intuitively, convergence in distribution means that if n is sufficiently large, the probability for X_n to be in a given range is approximately equal to the probability that X^* is in the same range.

Convergence in probability is denoted as $X_n \xrightarrow{d} X^*$. X^* is the asymptotic distribution of X_n.

We can extend convergence in distribution to the multivariate case. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ (where c is a constant), then $(X_n, Y_n) \xrightarrow{p} (X, c)$.

Levy's Continuity Theorem

Levy's continuity theorem equates converges in distribution of a sequence of random variables to the pointwise converges of a sequence of corresponding characteristic functions.

Let $X_1, X_2, ..., X_n$ be a sequence of n random varaibles. Let F_n be the CDF of X_n and let ξ_n be the characteristic function of X_n, given by $\xi_n(t) = \mathrm{E}(e^{itX_n}) \ \forall t \in \mathbf{R}$. Let F^* denote the CDF and ξ^* denote the characteristic function of random variable X^*.

The sequence $X_1, X_2, ..., X_n$ converges in distribution to X^* if:

$$\lim_{n\to\infty} \xi_n(t) = \xi^*(t) \ \forall t \in \mathbf{R}$$

Convergence of Moment Generating Functions:
Levy's continuity theorem is a more general version of the following statement: a convergence of moment generating functions (MGFs) to one MGF is equivalent to the convergence in distribution of a sequence of random variables (and their CDFs) to one random variable (and its CDF).

Let $X_1, X_2, ..., X_n$ be a sequence of n random varaibles. Let F_n be the CDF of X_n and let ψ_n be the moment generating function (MGF) of X_n, given by $\psi_n(t) = \mathrm{E}(e^{tX_n}) \ \forall t \in \mathbf{R}$. As before, let F^* denote the CDF and ψ^* denote the MGF of random variable X^*. Assume that both MGFs exist.

Then, the sequence $X_1, X_2, ..., X_n$ converges in distribution to X^* if:

$$\lim_{n\to\infty} \psi_n(t) = \psi^*(t) \text{ for all values of } t \text{ in the neighborhood around } t = 0$$

Law of Large Numbers

This theorem describes the behavior of the sample mean of a large number of random variables. The **Law of Large Numbers** states that the mean of a sequence of iid random variables converges to the expected value of the random variables. Another way of stating this is that the sample mean converges to the population mean. This means that if the sample size is large, the probability that the sample mean is near the true population mean is large. Mathematically:

$$\bar{X}_n \xrightarrow{p} \mu$$

Proof of Weak Law of Large Numbers

Let $X_1, X_2, ..., X_n$ be a sequence of identically and independently distributed random variables where $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$.

Recall that Chebyshev's inequality indicates that for any random variable X:
$P(|X - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$

However, what if we want to know the behavior of $\frac{1}{n}\bar{X}$?

First, find the expected value and the variance of the mean of the sequence:
$\mathrm{E}[\bar{X}] = \mu$

$\mathrm{Var}[\bar{X}] = \frac{1}{n}\sigma^2$

Then, due to Chebyshev's inequality, for every number $\epsilon > 0$:

$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$

This is identical to:
$P(|\bar{X}_n - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}$

Note that as $n \to \infty$:
$P(|\bar{X}_n - \mu| < \epsilon) = 1$ or $P(|\bar{X}_n - \mu| > \epsilon) = 0$

$\therefore \bar{X}_n \xrightarrow{p} \mu$

Strong Law of Large Numbers

The strong law of large numbers states that the sample mean almost sure converges to the true population mean:

$\bar{X}_n \xrightarrow{a.s.} \mu$

Moreover, we can generalize the strong law of large numbers to any function of x. Let $X_1, X_2, ..., X_n$ be iid random variables. Then assume that $f(x, \theta)$ is a continuous function of x defined for all $\theta \in \Omega$. The strong law of large numbers states:

$\frac{1}{n}\sum_{i=1}^{n} f(X_i, \theta) \xrightarrow{a.s.} \mathrm{E}[f(X, \theta)]$

Uniform Strong Law of Large Numbers

The uniform strong law of large numbers states necessary conditions for almost sure convergence to be uniform across the parameter space Ω.

The **uniform strong law of large numbers** states:

$$\boxed{\sup_{\theta \in \Omega} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i, \theta) - \mathrm{E}[f(X, \theta)] \right| \xrightarrow{a.s.} 0}$$

The conditions that need to hold for the uniform strong law of large numbers to apply to a random variable X and a function $f(x, \theta)$ are:

- Ω must be a compact parameter space

- $f(x, \theta)$ should be semi-continuous in $\theta \in \Omega$ for all x
- There must be a function $K(x)$ where $\mathrm{E}[K(x)] \leq \infty$ and $|f(x, \theta)| \leq K(x)\ \forall x, \theta$

Central Limit Theorem

The **central limit theorem** (CLT) states gives the conditions under which a mean of independently and identically distributed random variables will converge to a normal distribution. Specifically, the theorem states that when a random sample of size n is taken from any distribution characterized by a mean of μ and a variance of $\sigma^2 < \infty$, then the sample mean $\bar{X}_n$ has a distribution that is approximately normal with mean μ and variance σ^2/n. An equivalent statement is that for a large random sample, the distribution of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ is approximated by the standard normal distribution. The approximation improves as n increases (this is a convergence in distribution). This approximation holds whether the original distribution is continuous or discrete.

More rigorously:
Let $X_1, \ldots, X_n$ form a random sample of size n from a distribution with mean μ and variance σ^2. Then for each fixed number x:

$$\lim_{n \to \infty} P\Big[\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x\Big] = \Phi(x)$$

In a sense, the CLT governs the shape of convergence of $\bar{X}_n$ to μ, which we proved using the law of large numbers. The CLT indicates that the limit as n goes to infinity of $\bar{X}_n - \mu$ is non-degenerate when the exponent on the n is $\frac{1}{2}$ and that the limiting distribution is normal.

Proof of Central Limit Theorem

Suppose $X_1, X_2, \ldots, X_n$ are iid random variables with mean μ and variance σ^2. We want to prove that as n increases, $\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$. We will assume that X has a a moment generating function (MGF). Thus, we will use MGFs to prove that the central limit theorem holds. Even if the MGF does not exist, characteristic functions (which do always exist) may be used to prove that the CLT still holds. Note that the MGF of the standard normal distribution is given by $\psi^*(t) = e^{\frac{1}{2}t^2}$.

We begin with the random variable $Z_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$, which converges in distribution to the standard normal.

Using some simple algebra:

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{\sqrt{n}(\sum_{i=1}^{n}(X_i/n) - \mu)}{\sigma} = \frac{\sqrt{n} \cdot \frac{1}{n}(\sum_{i=1}^{n}(X_i - \mu))}{\sigma} = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n}\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underbrace{\frac{(X_i - \mu)}{\sigma}}_{Y_i}$$

Let $\psi(t)$ denote the MGF of the random variable Y_i. Then, since the sum of independent random variables is the product of their MGFs, the MGF of $\sum_{i=1}^{n} Y_i$ is $(\psi(t))^n$. When we multiply the MGF by $\frac{1}{\sqrt{n}}$, we get that the MGF of the standardized sum of random variables, Z_n, is:

$$\psi_n(t) = \Big(\psi(\frac{t}{\sqrt{n}})\Big)^n$$

We can express the MGF of the standardized sum of random variables using a Taylor series expansion around the point $t = 0$. This enables us to incorporate the information that the first moment of the standardized RV is 0 and the second moment is 1:

First moment: $E(Y_i) = \psi'(0) = 0$

Second moment: $E(Y_i^2) = \psi''(0) = 1$ (since $\text{Var}(Y_i) = 1$)

The Taylor series expansion of $\psi(t)$ around $t = 0$ (hence Maclaurin series) is:

$$\psi(t) \approx \psi(0) + t\psi'(0) + \frac{t^2}{2!}\psi''(0) + \frac{t^3}{3!}\psi'''(0) + \ldots = \psi(0) + t\psi'(0) + \frac{t^2}{2!}\psi''(t^*) = 1 + \frac{t^2}{2}\psi''(t^*) \text{ for } 0 < t* < t$$

Note instead of writing an infinite series, I wrote the the first 2 terms of the series and then used a Lagrange remainder.

This means that the Taylor series expansion of $\psi_n(t)$ is simply:

$$\psi_n(t) \approx \left[1 + \frac{t^2}{2n}\psi''(t^*)\right]^n \text{ for } 0 < t* < \frac{t}{\sqrt{n}}$$

Now we can find the limit of $\psi_n(t)$ as $n \to \infty$. Note that as $n \to \infty$, $t^* \to 0$ because $\frac{t}{\sqrt{n}} \to 0$.

Replacing t^* with 0 yields $\psi"(0) = 1$. Taking the limit:
$\lim_{n\to\infty}\left[1 + \frac{t^2}{2n}\right]^n$

This now looks like a case of a notable limit from calculus, given in its general form by:

$$\lim_{x\to\infty}\left[1 + \frac{k}{x}\right]^x = e^k$$

$$\therefore \lim_{n\to\infty}\psi_n(t) = \lim_{n\to\infty}\left[1 + \frac{t^2/2}{n}\right]^n = e^{t^2/2} = \psi^*(t)$$

We have just proven that the MGF of the standardized sequence of any random variables converges to the MGF of the standard normal distribution. Therefore, the CDF of the standardized sequence of the random variables converges in distribution to the standard normal distribution.

$$\therefore Z_n \xrightarrow{d} Z \sim N(0,1)$$

Continuous Mapping Theorem

The continuous mapping theorem states that if one has a sequence of random variables that converges, a continuous function that maps the convergent sequence to another sequence will also yield a convergent sequence.

More technically:
Define X to be a random variable on metric space S. X_n represents a sequence of n random variables X. The continuous function $g(\cdot)$ then maps $S \to S'$.

The **continuous mapping theorem** indicates that the following hold (assuming $g(\cdot)$ is continuous at X):

- $X_n \xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X)$
- $X_n \xrightarrow{p} X \Rightarrow g(X_n) \xrightarrow{p} g(X)$

- $X_n \xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X)$

Furthermore, one can show that if $X_n \xrightarrow{p} a$ and $Y_n \xrightarrow{p} b$, then $g(X_n, Y_n) \xrightarrow{p} g(a, b)$ (assuming that $g(\cdot)$ is continuous at a, b).

Slutsky's Theorem

Suppose that we have two sets of sequences: $\{X_n\}$ and $\{Y_n\}$. Furthermore, suppose that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$.

Slutsky's theorem indicates that the following relationships hold between convergent sequences:

- $X_n + Y_n \xrightarrow{d} X + c$
- $Y_n \cdot X_n \xrightarrow{d} cX$
- $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$ $(c \neq 0)$

Proof: The proof of Slutsky's theorem is quite simple - in effect, the theorem is just a particular application of the continuous mapping theorem. We know that since $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, the joint vector $(X_n, Y_n) \xrightarrow{d} (X, c)$. Now, using the multivariate version of the continuous mapping theorem, we respectively let $g(x, y) = x + y$, $g(x, y) = xy$, and $g(x, y) = \frac{x}{y}$.