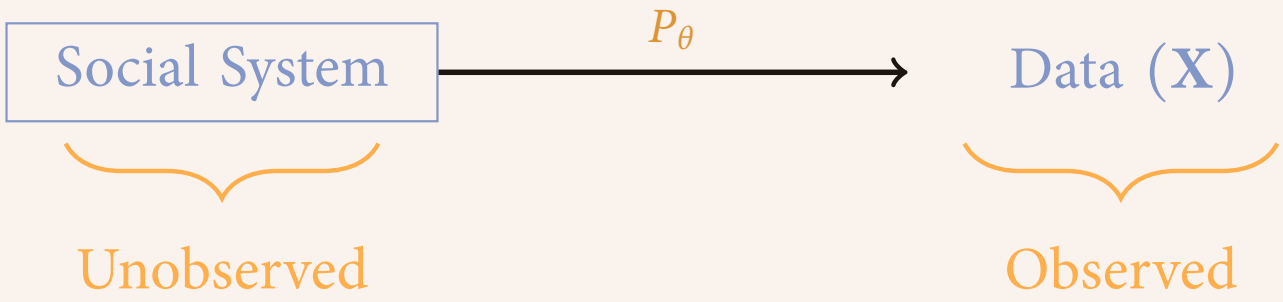
Gov 2001 Section 5: Theory of Maximum Likelihood Estimation

Konstantin Kashin

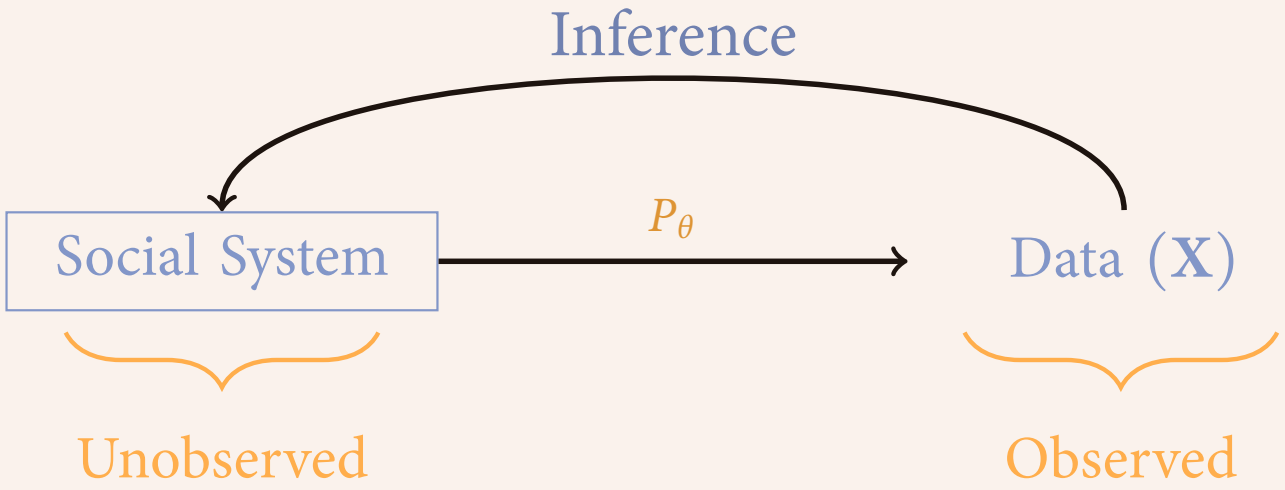
February 28, 2013

Outline

What is Statistical Inference?



What is Statistical Inference?



Model-Based (Parametric) Inference

Model-Based (Parametric) Inference

- We assume that the data we observe comes from a model / family of distributions:

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

Model-Based (Parametric) Inference

- We assume that the data we observe comes from a model / family of distributions:

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

- Not right or wrong, but useful, representation of data generating process (DGP)

Model-Based (Parametric) Inference

- We assume that the data we observe comes from a model / family of distributions:

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

- Not right or wrong, but useful, representation of data generating process (DGP)
- Goal of inference is to use the sample we observe $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ to say something about θ (the parameter that completely specifies the DGP) under our model assumptions

Model-Based (Parametric) Inference

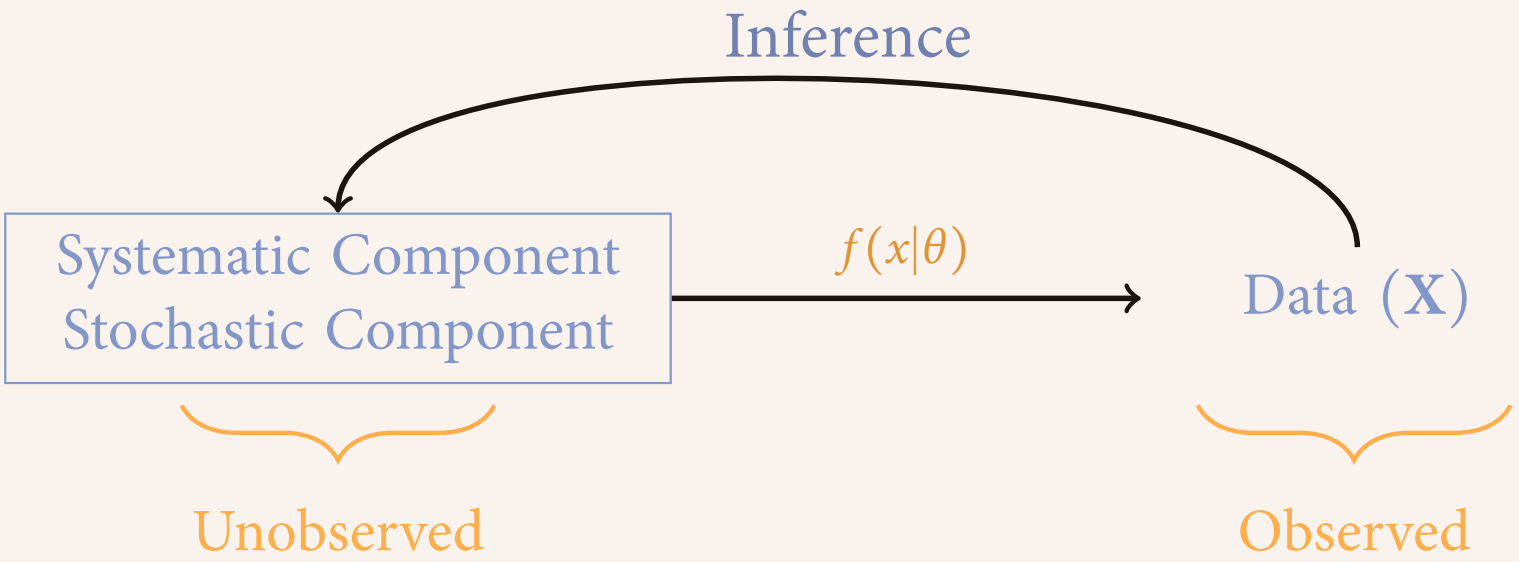
- We assume that the data we observe comes from a model / family of distributions:

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

- Not right or wrong, but useful, representation of data generating process (DGP)
- Goal of inference is to use the sample we observe $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ to say something about θ (the parameter that completely specifies the DGP) under our model assumptions

There are two main theories of doing this:
Frequentist (likelihood) and Bayesian

What is Statistical Inference?



Bayes' Rule

Intuitively, we would like to know the probability density over the unknown parameter θ conditional on the data we observe:

$$\xi(\theta|\mathbf{x})$$

Bayes' Rule

Intuitively, we would like to know the probability density over the unknown parameter θ conditional on the data we observe:

$$\xi(\theta|\mathbf{x})$$

By Bayes' Rule, we can write this probability as:

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

Frequentist / Likelihood Inference

$$\xi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\xi(\theta)}{g_n(\mathbf{x})}$$

Frequentist / Likelihood Inference

$$\xi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\xi(\theta)}{g_n(\mathbf{x})}$$

We absorb everything that is a constant of the data in $k(x)$:

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

Frequentist / Likelihood Inference

$$\xi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\xi(\theta)}{g_n(\mathbf{x})}$$

We absorb everything that is a constant of the data in $k(x)$:

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

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

Frequentist / Likelihood Inference

$$\xi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\xi(\theta)}{g_n(\mathbf{x})}$$

We absorb everything that is a constant of the data in $k(x)$:

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

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

- Parameter θ is an unknown constant

Frequentist / Likelihood Inference

$$\xi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\xi(\theta)}{g_n(\mathbf{x})}$$

We absorb everything that is a constant of the data in $k(x)$:

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

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

- Parameter θ is an unknown constant
- All fundamental variability (uncertainty) comes from sampling

Frequentist / Likelihood Inference

$$\xi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\xi(\theta)}{g_n(\mathbf{x})}$$

We absorb everything that is a constant of the data in $k(x)$:

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

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

- Parameter θ is an unknown constant
- All fundamental variability (uncertainty) comes from sampling
- Everything we know about the parameter based on the data is summarized in the likelihood function

Frequentist / Likelihood Inference

$$\xi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\xi(\theta)}{g_n(\mathbf{x})}$$

We absorb everything that is a constant of the data in $k(x)$:

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

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

- Parameter θ is an unknown constant
- All fundamental variability (uncertainty) comes from sampling
- Everything we know about the parameter based on the data is summarized in the likelihood function
- Focus of inference: characterize likelihood $L(\theta|\mathbf{x})$

Frequentist / Likelihood Inference

$$\xi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\xi(\theta)}{g_n(\mathbf{x})}$$

We absorb everything that is a constant of the data in $k(x)$:

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

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

- Parameter θ is an unknown constant
- All fundamental variability (uncertainty) comes from sampling
- Everything we know about the parameter based on the data is summarized in the likelihood function
- Focus of inference: characterize likelihood $L(\theta|\mathbf{x})$
- Point summary: maximum likelihood estimate

Bayesian Inference

Bayesian Inference

$$\xi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\xi(\theta)}{g_n(\mathbf{x})}$$

Bayesian Inference

$$\xi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\xi(\theta)}{g_n(\mathbf{x})}$$

We can drop the proportionality constant $g_n(\mathbf{x})$ since it's not a function of θ:

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

Bayesian Inference

$$\xi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\xi(\theta)}{g_n(\mathbf{x})}$$

We can drop the proportionality constant $g_n(\mathbf{x})$ since it's not a function of θ:

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

- Parameter θ is a latent (unobserved) random variable

Bayesian Inference

$$\xi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\xi(\theta)}{g_n(\mathbf{x})}$$

We can drop the proportionality constant $g_n(\mathbf{x})$ since it's not a function of θ:

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

- Parameter θ is a latent (unobserved) random variable
- All fundamental variability (uncertainty) comes from sampling & parameter (through the prior)

Bayesian Inference

$$\xi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\xi(\theta)}{g_n(\mathbf{x})}$$

We can drop the proportionality constant $g_n(\mathbf{x})$ since it's not a function of θ:

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

- Parameter θ is a latent (unobserved) random variable
- All fundamental variability (uncertainty) comes from sampling & parameter (through the prior)
- Probabilities still relative because we don't truly know $\xi(\theta)$

Bayesian Inference

$$\xi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\xi(\theta)}{g_n(\mathbf{x})}$$

We can drop the proportionality constant $g_n(\mathbf{x})$ since it's not a function of θ:

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

- Parameter θ is a latent (unobserved) random variable
- All fundamental variability (uncertainty) comes from sampling & parameter (through the prior)
- Probabilities still relative because we don't truly know $\xi(\theta)$
- Focus of inference: estimate posterior $\xi(\theta|\mathbf{x})$

Bayesian Inference

$$\xi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\xi(\theta)}{g_n(\mathbf{x})}$$

We can drop the proportionality constant $g_n(\mathbf{x})$ since it's not a function of θ:

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

- Parameter θ is a latent (unobserved) random variable
- All fundamental variability (uncertainty) comes from sampling & parameter (through the prior)
- Probabilities still relative because we don't truly know $\xi(\theta)$
- Focus of inference: estimate posterior $\xi(\theta|\mathbf{x})$
- Point summary: maximum a posteriori (MAP) or posterior mean (PM)

Outline

Likelihood

For an i.i.d. sample of $\mathbf{X} = X_1, \ldots, X_n$, we define the likelihood of parameter θ as:

Likelihood

For an i.i.d. sample of $\mathbf{X} = X_1, \ldots, X_n$, we define the likelihood of parameter θ as:

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

Likelihood

For an i.i.d. sample of $\mathbf{X} = X_1, \ldots, X_n$, we define the likelihood of parameter θ as:

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

Conceptually, $L(\theta|\mathbf{x})$, is a function that assigns a value to each point in parameter space Ω that indicates how likely each value of the parameter is to have generated the data.

Likelihood

For an i.i.d. sample of $\mathbf{X} = X_1, \ldots, X_n$, we define the likelihood of parameter θ as:

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

Conceptually, $L(\theta|\mathbf{x})$, is a function that assigns a value to each point in parameter space Ω that indicates how likely each value of the parameter is to have generated the data.

For a variety of reasons, we work with the log-likelihood:

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

MLE

The MLE is defined as:

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

MLE

The MLE is defined 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 define the MLE in terms of maximizing the log-likelihood:

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

Finding the MLE

Finding the MLE

- **Analytic**: solve first order condition for critical points, then check that second derivative at critical point is negative

Finding the MLE

- **Analytic**: solve first order condition for critical points, then check that second derivative at critical point is negative
  - Define the score as:

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

Finding the MLE

- **Analytic**: solve first order condition for critical points, then check that second derivative at critical point is negative
  - Define the score as:

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

  - Find critical values by setting score to 0 and solving for θ

Finding the MLE

- **Analytic**: solve first order condition for critical points, then check that second derivative at critical point is negative
  - Define the score as:

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

  - Find critical values by setting score to 0 and solving for θ
- **Numeric**: `optim()` in R

Other quantities we can calculate from the sample...

Other quantities we can calculate from the sample...

- Score evaluated at a given θ:

$$S(\theta) = \ell'(\theta|\mathbf{x})$$

Other quantities we can calculate from the sample...

- Score evaluated at a given θ:

$$S(\theta) = \ell'(\theta|\mathbf{x})$$

- Second derivative of log-likelihood (hessian if multiple parameters) at any θ:

$$\ell''(\theta|\mathbf{x})$$

Other quantities we can calculate from the sample...

- Score evaluated at a given θ:

$$S(\theta) = \ell'(\theta|\mathbf{x})$$

- Second derivative of log-likelihood (hessian if multiple parameters) at any θ:

$$\ell''(\theta|\mathbf{x})$$

- Observed Fisher information: negation of second derivative at any θ

$$J_n(\theta) = -\ell''(\theta|\mathbf{x}) = -\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{x}|\theta)$$

Other quantities we can calculate from the sample...

- Score evaluated at a given θ:

$$S(\theta) = \ell'(\theta|\mathbf{x})$$

- Second derivative of log-likelihood (hessian if multiple parameters) at any θ:

$$\ell''(\theta|\mathbf{x})$$

- Observed Fisher information: negation of second derivative at any θ

$$J_n(\theta) = -\ell''(\theta|\mathbf{x}) = -\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{x}|\theta)$$

- Inverse of observed Fisher information: inverse of negation of second derivative at any θ

In multiple dimensions...

If we have multiple parameters (a vector $\boldsymbol{\theta}$ of length k):

In multiple dimensions...

If we have multiple parameters (a vector $\boldsymbol{\theta}$ of length k):

- Log-likelihood:

$$\ell(\boldsymbol{\theta}|\mathbf{x}) = \log L(\boldsymbol{\theta}|\mathbf{x})$$

In multiple dimensions...

If we have multiple parameters (a vector $\boldsymbol{\theta}$ of length k):

- Log-likelihood:

$$\ell(\boldsymbol{\theta}|\mathbf{x}) = \log L(\boldsymbol{\theta}|\mathbf{x})$$

- Score:

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

In multiple dimensions...

If we have multiple parameters (a vector $\boldsymbol{\theta}$ of length k):

- Log-likelihood:

$$\ell(\boldsymbol{\theta}|\mathbf{x}) = \log L(\boldsymbol{\theta}|\mathbf{x})$$

- Score:

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

- Observed fisher information:

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

Outline

A Geometric MLE

$X_1 \ldots X_n$ form a random sample from a geometric distribution with unknown parameter $0 \leq p \leq 1$. A geometric distribution describes the number of failures before we observe a success in a series of independent Bernoulli trials, each with probability p of a success. We need to find the MLE estimator for p.

A Geometric MLE

$X_1 \ldots X_n$ form a random sample from a geometric distribution with unknown parameter $0 \leq p \leq 1$. A geometric distribution describes the number of failures before we observe a success in a series of independent Bernoulli trials, each with probability p of a success. We need to find the MLE estimator for p.

The pdf of a geometric distribution is:

$$f(x|p) = (1-p)^x p$$

A Geometric MLE

$X_1 \ldots X_n$ form a random sample from a geometric distribution with unknown parameter $0 \leq p \leq 1$. A geometric distribution describes the number of failures before we observe a success in a series of independent Bernoulli trials, each with probability p of a success. We need to find the MLE estimator for p.

The pdf of a geometric distribution is:

$$f(x|p) = (1-p)^x p$$

Thus, the likelihood for n i.i.d. draws is:

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

A Geometric MLE

$X_1 \ldots X_n$ form a random sample from a geometric distribution with unknown parameter $0 \leq p \leq 1$. A geometric distribution describes the number of failures before we observe a success in a series of independent Bernoulli trials, each with probability p of a success. We need to find the MLE estimator for p.

The pdf of a geometric distribution is:

$$f(x|p) = (1-p)^x p$$

Thus, the likelihood for n i.i.d. draws is:

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

Taking the log:

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

A Geometric MLE

The log likelihood function is:

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

A Geometric MLE

The log likelihood function is:

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

The score, which is the first derivative of the log-likelihood, is:

$$S(p) = \ell'(p|\mathbf{x}) = \frac{n}{p} - \frac{1}{1-p}\sum_{i=1}^{n} x_i$$

A Geometric MLE

To find the MLE, we set the score equal to 0 and solve for p:

$$S(p) = \frac{n}{p} - \frac{1}{1-p} \sum_{i=1}^{n} x_i = 0$$

A Geometric MLE

To find the MLE, we set the score equal to 0 and solve for p:

$$S(p) = \frac{n}{p} - \frac{1}{1-p}\sum_{i=1}^{n} x_i = 0$$

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

A Geometric MLE

To find the MLE, we set the score equal to 0 and solve for p:

$$S(p) = \frac{n}{p} - \frac{1}{1-p}\sum_{i=1}^{n} x_i = 0$$

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

A Geometric MLE

To find the MLE, we set the score equal to 0 and solve for p:

$$S(p) = \frac{n}{p} - \frac{1}{1-p}\sum_{i=1}^{n} x_i = 0$$

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

$$p\sum_{i=1}^{n} x_i = n - np$$

A Geometric MLE

To find the MLE, we set the score equal to 0 and solve for p:

$$S(p) = \frac{n}{p} - \frac{1}{1-p}\sum_{i=1}^{n} x_i = 0$$

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

$$p\sum_{i=1}^{n} x_i = n - np$$

$$p = \frac{n}{n + \sum_{i=1}^{n} x_i}$$

A Geometric MLE

To find the MLE, we set the score equal to 0 and solve for p:

$$S(p) = \frac{n}{p} - \frac{1}{1-p}\sum_{i=1}^{n} x_i = 0$$

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

$$p\sum_{i=1}^{n} x_i = n - np$$

$$p = \frac{n}{n + \sum_{i=1}^{n} x_i}$$

$$\hat{p}_{MLE} = \frac{1}{1+\bar{x}}$$

Observed Fisher Information

First, let's find the second derivative of the log-likelihood function:

$$\ell''(p|\mathbf{x}) = -\frac{n}{p^2} - \frac{1}{(1-p)^2}\sum_{i=1}^{n} x_i$$

Note that the second derivative captures the steepness of the curvature around the point p. A more negative second derivative implies that the function is more steeply concave down around the point p.
The observed Fisher information is just the negation of the second derivative:

$$J_n(p) = -\ell''(p|\mathbf{x}) = \frac{n}{p^2} + \frac{1}{(1-p)^2}\sum_{i=1}^{n} x_i$$

Outline

Variance of the MLE

We are interested in calculating a measure of uncertainty of our MLE. That is, we are after the following:

Variance of the MLE

We are interested in calculating a measure of uncertainty of our MLE. That is, we are after the following:

$$Var(\hat{\theta}_{MLE})$$

Variance of the MLE

We are interested in calculating a measure of uncertainty of our MLE. That is, we are after the following:

$$Var(\hat{\theta}_{MLE})$$

Conceptually, what is this quantity? We want to understand it conceptually before we can talk about calculating it!

How do we understand variance in frequentist framework?

How do we understand variance in frequentist framework?

In terms of drawing infinite samples with sample size n from the distribution of interest!

How do we understand variance in frequentist framework?

In terms of drawing infinite samples with sample size n from the distribution of interest!

Specifically, we can think of our sample as one of many possible samples we can draw from our population:

How do we understand variance in frequentist framework?

In terms of drawing infinite samples with sample size n from the distribution of interest!

Specifically, we can think of our sample as one of many possible samples we can draw from our population:

	Samples from Pop.				
	1	**2**	**3**	$\cdots$	∞
X_1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$\cdots$	$x_{1,\infty}$
X_2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$\cdots$	$x_{2,\infty}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
X_n	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$\cdots$	$x_{n,\infty}$
$\hat{\theta}_{MLE}$	$\hat{\theta}_{MLE,1}$	$\hat{\theta}_{MLE,2}$	$\hat{\theta}_{MLE,3}$	$\cdots$	$\hat{\theta}_{MLE,\infty}$
$J_n(\hat{\theta}_{MLE})$	$J_{n,1}(\hat{\theta}_{MLE})$	$J_{n,2}(\hat{\theta}_{MLE})$	$J_{n,3}(\hat{\theta}_{MLE})$	$\cdots$	$J_{n,\infty}(\hat{\theta}_{MLE})$

How do we understand variance in frequentist framework?

	Samples from Pop.				
	1	**2**	**3**	$\cdots$	∞
X_1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$\cdots$	$x_{1,\infty}$
X_2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$\cdots$	$x_{2,\infty}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
X_n	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$\cdots$	$x_{n,\infty}$
$\hat{\theta}_{MLE}$	$\hat{\theta}_{MLE,1}$	$\hat{\theta}_{MLE,2}$	$\hat{\theta}_{MLE,3}$	$\cdots$	$\hat{\theta}_{MLE,\infty}$
$J_n(\hat{\theta}_{MLE})$	$J_{n,1}(\hat{\theta}_{MLE})$	$J_{n,2}(\hat{\theta}_{MLE})$	$J_{n,3}(\hat{\theta}_{MLE})$	$\cdots$	$J_{n,\infty}(\hat{\theta}_{MLE})$

We see that $\hat{\theta}_{MLE}$ and $J_n(\hat{\theta}_{MLE})$ are random variables! Thus, they have some theoretical distributions.

$Var(\hat{\theta}_{MLE})$ is just the variance of this theoretical distribution!

Expectations of Random Variables

In fact, the following are all random variables (vary across samples):

Expectations of Random Variables

In fact, the following are all random variables (vary across samples):

- MLE: $\hat{\theta}_{MLE}$

Expectations of Random Variables

In fact, the following are all random variables (vary across samples):

- MLE: $\hat{\theta}_{MLE}$
- Log-likelihood evaluated at MLE (or any θ): $\ell(\hat{\theta}_{MLE}|\mathbf{x})$

Expectations of Random Variables

In fact, the following are all random variables (vary across samples):

- MLE: $\hat{\theta}_{MLE}$
- Log-likelihood evaluated at MLE (or any θ): $\ell(\hat{\theta}_{MLE}|\mathbf{x})$
- Score evaluated at the MLE (or any θ): $S(\hat{\theta}_{MLE}) = \ell'(\hat{\theta}_{MLE}|\mathbf{x})$

Expectations of Random Variables

In fact, the following are all random variables (vary across samples):

- MLE: $\hat{\theta}_{MLE}$
- Log-likelihood evaluated at MLE (or any θ): $\ell(\hat{\theta}_{MLE}|\mathbf{x})$
- Score evaluated at the MLE (or any θ): $S(\hat{\theta}_{MLE}) = \ell'(\hat{\theta}_{MLE}|\mathbf{x})$
- Observed Fisher information evaluated at MLE (or any θ): $J_n(\hat{\theta}_{MLE}) = -\ell''(\hat{\theta}_{MLE}|\mathbf{x})$

Expectations of Random Variables

In fact, the following are all random variables (vary across samples):

- MLE: $\hat{\theta}_{MLE}$
- Log-likelihood evaluated at MLE (or any θ): $\ell(\hat{\theta}_{MLE}|\mathbf{x})$
- Score evaluated at the MLE (or any θ): $S(\hat{\theta}_{MLE}) = \ell'(\hat{\theta}_{MLE}|\mathbf{x})$
- Observed Fisher information evaluated at MLE (or any θ): $J_n(\hat{\theta}_{MLE}) = -\ell''(\hat{\theta}_{MLE}|\mathbf{x})$
- Inverse observed Fisher information: $J_n(\hat{\theta}_{MLE})^{-1}$

Expectations of Random Variables

In fact, the following are all random variables (vary across samples):

- MLE: $\hat{\theta}_{MLE}$
- Log-likelihood evaluated at MLE (or any θ): $\ell(\hat{\theta}_{MLE}|\mathbf{x})$
- Score evaluated at the MLE (or any θ): $S(\hat{\theta}_{MLE}) = \ell'(\hat{\theta}_{MLE}|\mathbf{x})$
- Observed Fisher information evaluated at MLE (or any θ): $J_n(\hat{\theta}_{MLE}) = -\ell''(\hat{\theta}_{MLE}|\mathbf{x})$
- Inverse observed Fisher information: $J_n(\hat{\theta}_{MLE})^{-1}$

We thus often talk about the expectation of these random quantities across infinite samples. We can denote this, for the example of the MLE, as:

$$E[\hat{\theta}_{MLE}]$$

or

$$E_{\theta_0}[\hat{\theta}_{MLE}]$$

Variance of the MLE

Recall that the variance of the MLE is:

Variance of the MLE

Recall that the variance of the MLE is:

$$Var(\hat{\theta}_{MLE})$$

Variance of the MLE

Recall that the variance of the MLE is:

$$Var(\hat{\theta}_{MLE})$$

Now that we understand it conceptually, how do we estimate it?

Asymptotic Distribution of the MLEs

It can be shown that under certain regularity conditions, the MLE is distributed normally 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 $I_n(\theta_o)$):

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

Let's focus on understanding the variance for now...

Expected Fisher Information

Expected Fisher Information

- Asymptotically: $Var(\hat{\theta}_{MLE}) = I_n(\theta_o)$

Expected Fisher Information

- Asymptotically: $Var(\hat{\theta}_{MLE}) = I_n(\theta_o)$
- What is $I_n(\theta_o)$?

$$E[J_n(\theta_o)] = -E\left[\ \ell(\theta|\mathbf{x})\right]$$

Expected Fisher Information

- Asymptotically: $Var(\hat{\theta}_{MLE}) = I_n(\theta_o)$
- What is $I_n(\theta_o)$?

$$E[J_n(\theta_o)] = -E\left[\ \ell(\theta|\mathbf{x})\right]$$

- That is, it's the expectation of the observed Fisher information evaluated at the true parameter θ_o

Expected Fisher Information

- Asymptotically: $Var(\hat{\theta}_{MLE}) = I_n(\theta_o)$
- What is $I_n(\theta_o)$?

$$E[J_n(\theta_o)] = -E\left[\ \ell(\theta|\mathbf{x})\right]$$

- That is, it's the expectation of the observed Fisher information evaluated at the true parameter θ_o
- Conceptually, this is the *expected* curvature of the log-likelihood curve (or surface) across repeated samples at the point θ_o (the true parameter)

Expected Fisher Information

- Asymptotically: $Var(\hat{\theta}_{MLE}) = I_n(\theta_o)$
- What is $I_n(\theta_o)$?

$$E[J_n(\theta_o)] = -E\left[\ \ell(\theta|\mathbf{x})\right]$$

- That is, it's the expectation of the observed Fisher information evaluated at the true parameter θ_o
- Conceptually, this is the *expected* curvature of the log-likelihood curve (or surface) across repeated samples at the point θ_o (the true parameter)
- As $n \to \infty$, the observed Fisher information converges to the expected Fisher information and the $\hat{\theta}_{MLE}$ converges to θ_o

Expected Fisher Information

- Asymptotically: $Var(\hat{\theta}_{MLE}) = I_n(\theta_o)$
- What is $I_n(\theta_o)$?

$$E[J_n(\theta_o)] = -E\left[\ \ell(\theta|\mathbf{x})\right]$$

- That is, it's the expectation of the observed Fisher information evaluated at the true parameter θ_o
- Conceptually, this is the *expected* curvature of the log-likelihood curve (or surface) across repeated samples at the point θ_o (the true parameter)
- As $n \to \infty$, the observed Fisher information converges to the expected Fisher information and the $\hat{\theta}_{MLE}$ converges to θ_o

In practice, we use the inverse of the observed Fisher information evaluated at the MLE to approximate the true variance of the MLE!

Let's do an example...

Let's do an example...

- Suppose $X \sim Geom(0.5)$

Let's do an example...

- Suppose $X \sim Geom(0.5)$
- The true parameter is $p_0 = 0.5$

Let's do an example...

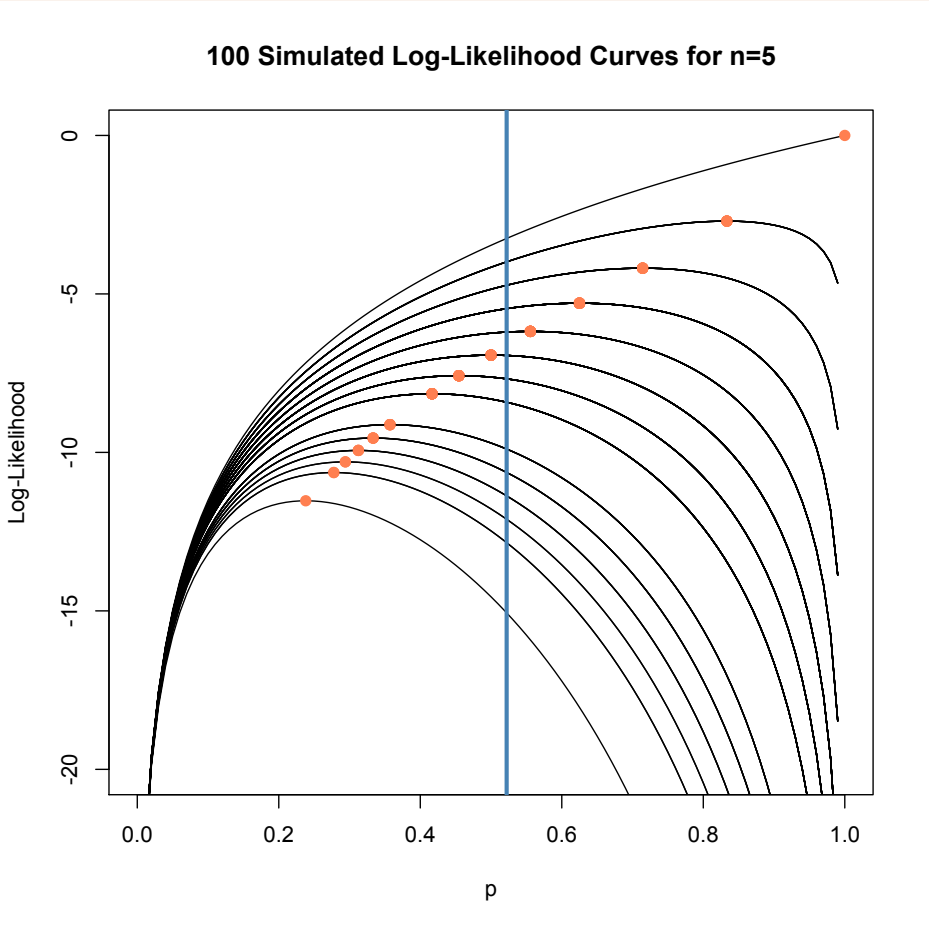
- Suppose $X \sim Geom(0.5)$
- The true parameter is $p_0 = 0.5$
- We want to estimate $\hat{p}_{MLE}$ and find the uncertainty around it

Let's do an example...

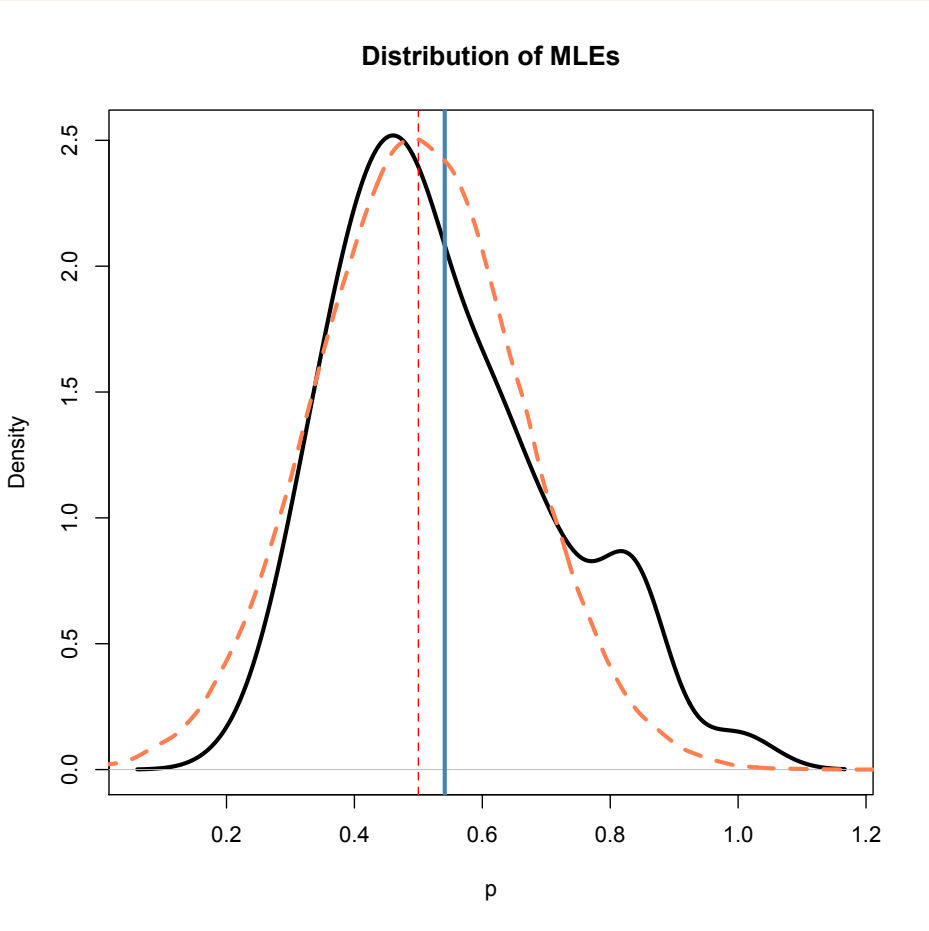
- Suppose $X \sim Geom(0.5)$
- The true parameter is $p_0 = 0.5$
- We want to estimate $\hat{p}_{MLE}$ and find the uncertainty around it

We shall see how as sample size (n) gets larger, $\hat{p}_{MLE} \approx p_0$ and $\left[-\ell''(\hat{p}_{MLE}|\mathbf{x})\right]^{-1} \approx Var(\hat{p}_{MLE})$

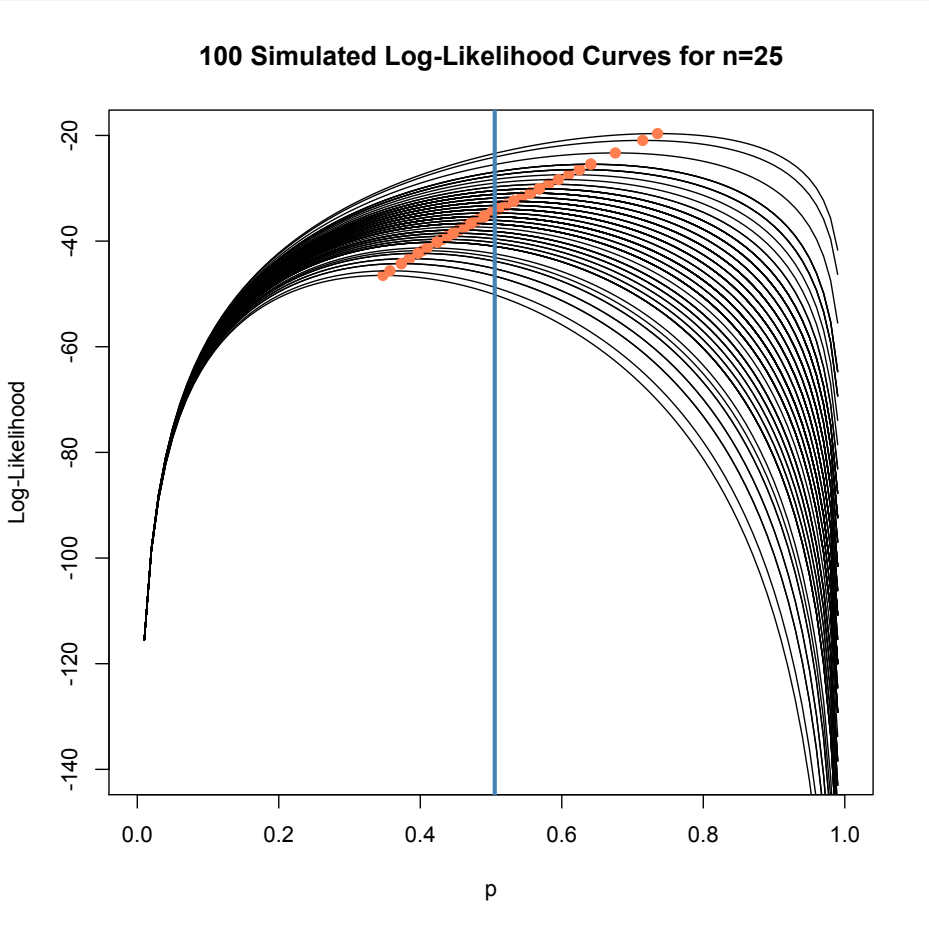
Log-Likelihoods for $n = 5$



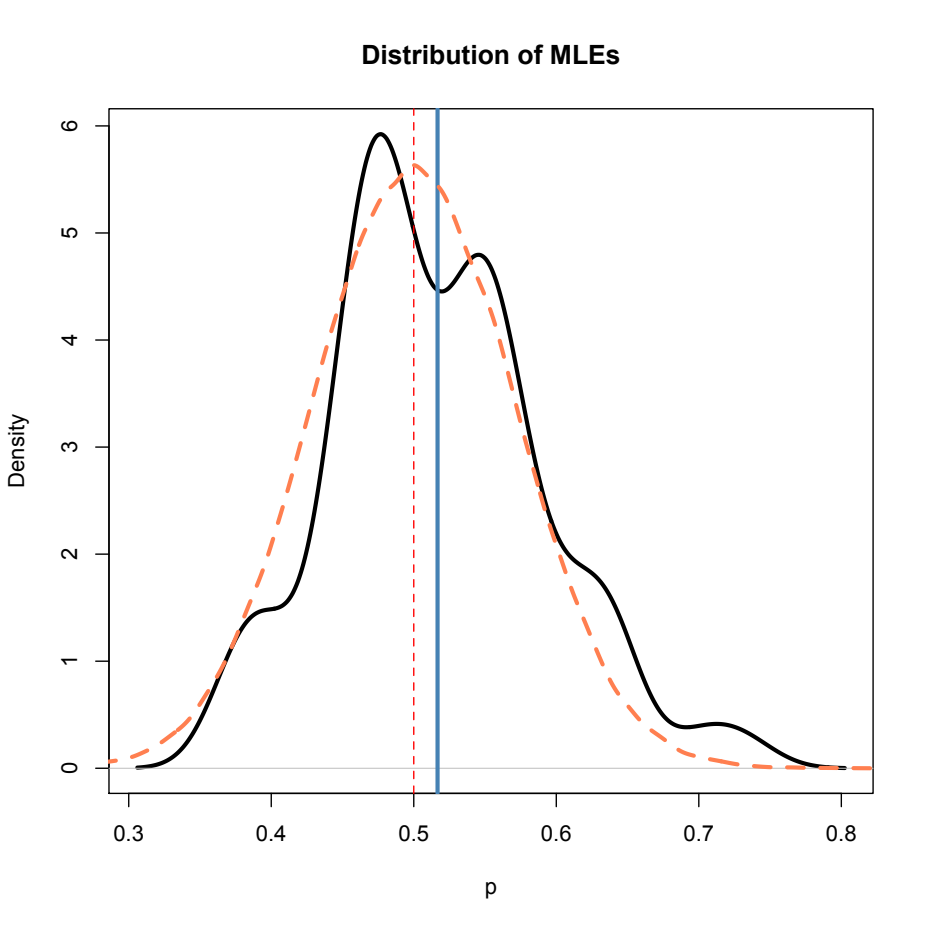
Distribution of MLEs for $n = 5$



Log-Likelihoods for $n = 25$



Distribution of MLEs for $n = 25$

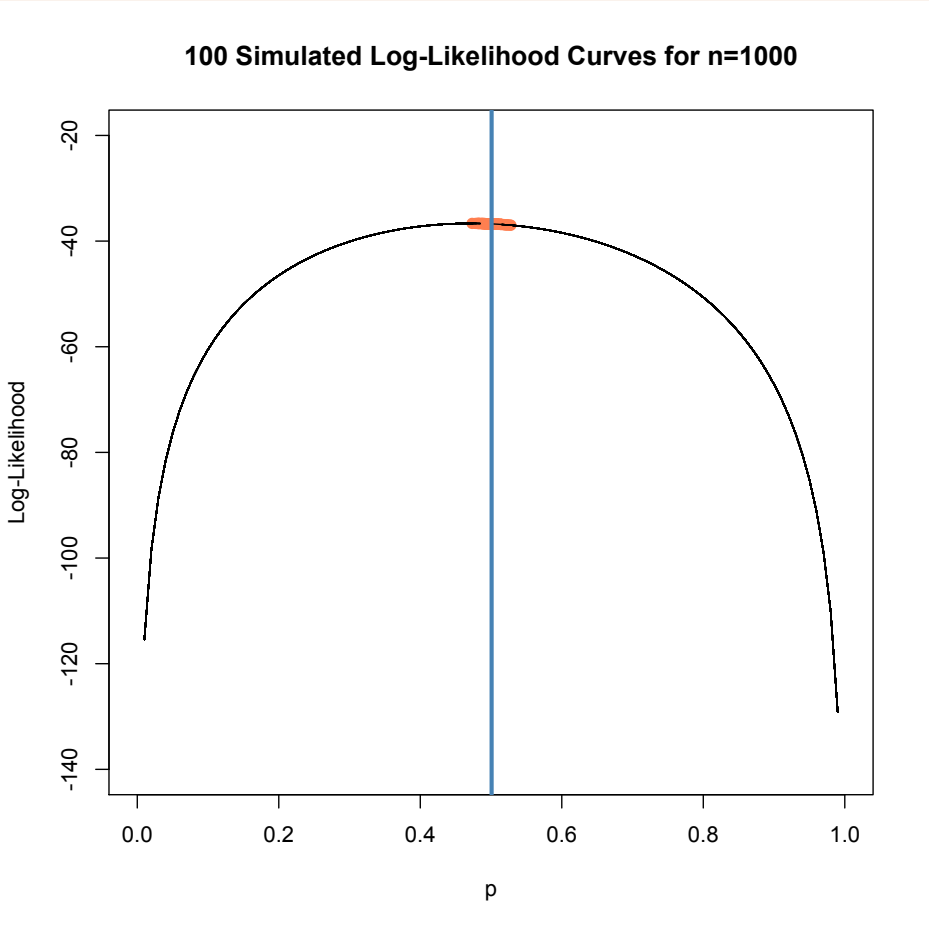


$$Var_{1000}(\hat{\theta}_{MLE}) = 0.005532$$

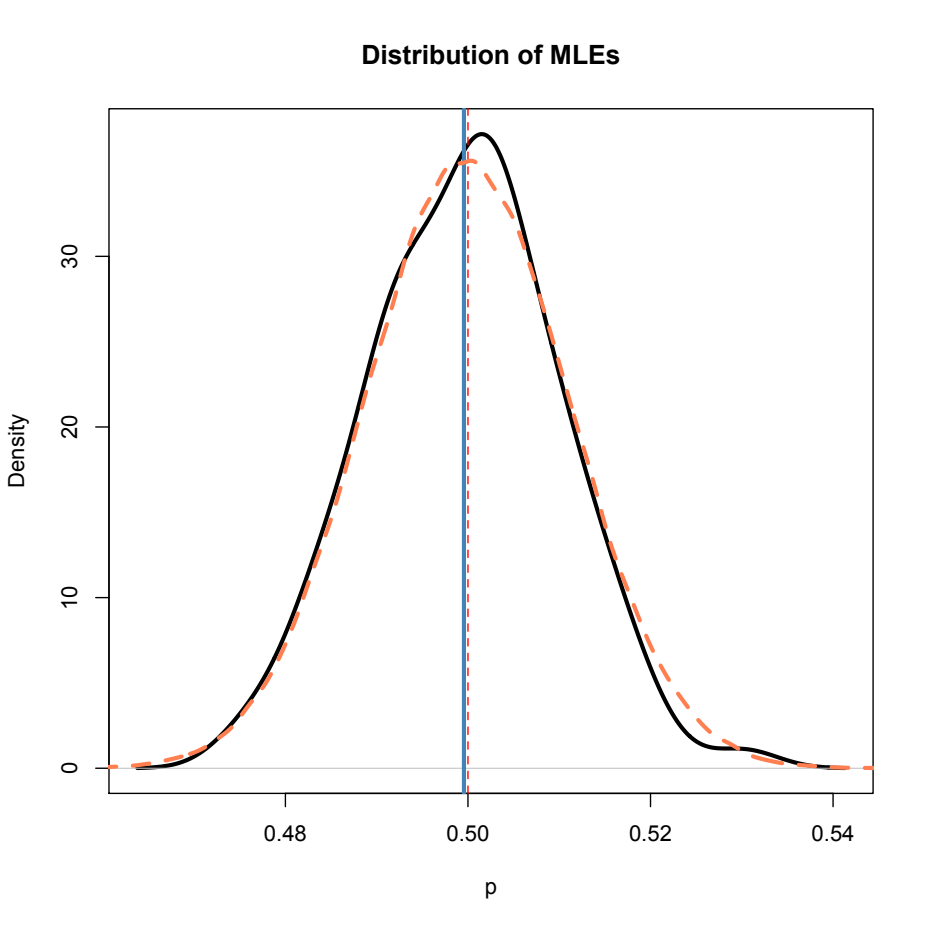
$$J(\hat{\theta}_{MLE})^{-1} = 0.00503$$

$$I(\theta_0)^{-1} = 0.005$$

Log-Likelihoods for $n = 1000$



Distribution of MLEs for $n = 1000$



$$Var_{1000}(\hat{\theta}_{MLE}) = 0.000106$$

$$\overline{J(\hat{\theta}_{MLE})^{-1}} = 0.000124$$

$$I(\theta_0)^{-1} = 0.000125$$

Outline

Summary of Asymptotic Properties

- **Consistency:**

$$\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$$

- **Normality:**

$$\hat{\theta}_{MLE} \sim N\left(\theta_0, \left(\underbrace{-E\left[\frac{\partial^2 \ell(\theta|\mathbf{x})}{\partial \theta^2}\Big|_{\theta=\theta_0}\right]}_{I_n(\theta_0)}\right)^{-1}\right)$$

- **Efficiency:** lowest mean square error amongst asymptotically unbiased estimators

Consistency

As sample size (n) increases, the MLE ($\hat{\theta}_{MLE}$) converges to the true parameter, θ_0:

$$\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$$

Proof relies upon the uniform law of large numbers.

Regularity Conditions for Consistency

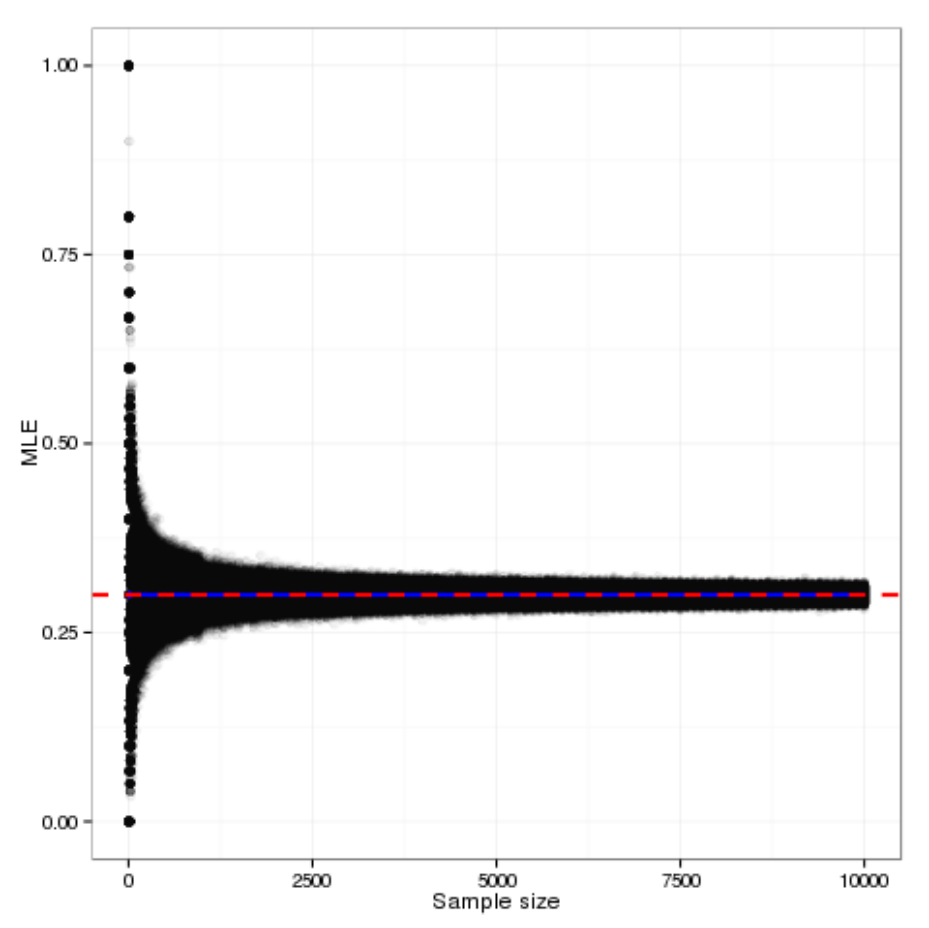
- Model identification: the true parameter θ_o is the unique global maximizer of $E_{\theta_o}[\ell(\theta|\mathbf{x})]$
- Compactness: the parameter space Ω must be a bounded and closed set. That is, Ω must be a *compact subset*.
- The log-likelihood function $\ell(\theta|\mathbf{x})$ is continuous in θ
- Note that compactness condition is sufficient but not necessary: it can be replaced with other conditions for non-compact and infinite parameter spaces

Visualizing Consistency of the MLE

Simple simulation study to look at properties of MLE:

- $X \sim Bern(0.3)$
- Vary sample size n
- For each n, simulate 10,000 datasets of size n and calculate MLE and observed Fisher information
- Take mean across MLEs for each n and compare to true value of the parameter

Visualizing Consistency of the MLE



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.

Normality

The standardized MLE is distributed normally with a mean of 0 and a variance equal to the expected unit Fisher information:

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \sim N\left(0, \left(\underbrace{-E\left[\frac{\partial^2 \ell(\theta|x)}{\partial \theta^2}\Big|_{\theta=\theta_0}\right]}_{I_1(\theta_0)} \right)^{-1}\right)$$

Proof relies upon the Central Limit Theorem.

Regularity Conditions for Normality

We need all conditions needed for consistency as well as:

- The true value of the parameter, θ_0, must be an interior point of the parameter set: $\theta_0 \in int(\Omega)$. Phrased differently, θ_0 cannot be on the boundary of the set.
  - This is equivalent to saying that Ω doesn't depend on θ
  - A violation of this is the uniform distribution $Unif[0, \theta]$ which has a biased MLE that is not asymptotically normally distributed (see Lehamnn and Casella 1998)
- The likelihood function (or probability distribution $f(x|\theta)$) is continuously twice-differentiable in the neighborhood of θ_0
- Fisher information matrix exists, is non-singular, and finitely bounded
-

$$\left| \frac{d^j \ln f(x|\theta)}{d\theta^j} \leq h_j(x) \right|, \text{ for } j = 1, 2, 3$$

$$E[h_j(y)] < \infty, \; j = 1, 2$$

$$E[h_3(x)] \text{ does not depend on } \theta$$

Normality

Another way to phrase asymptotic normality is:
As sample size (n) increases, the MLE is normally distributed with a mean equal to the true parameter (θ_0) and the variance equal to the inverse of the expected sample Fisher information at the true parameter (denoted as $I_n(\theta_0)$):

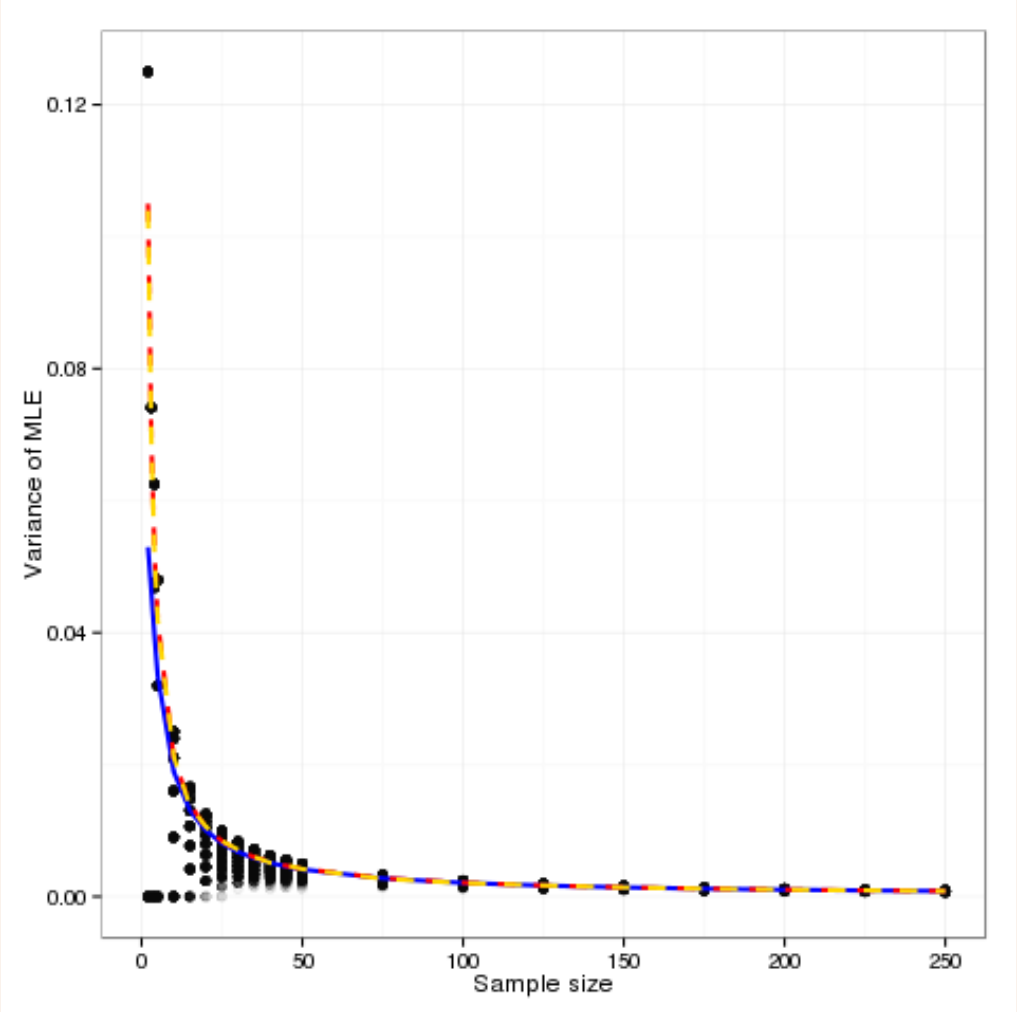
$$\hat{\theta}_{MLE} \sim N\left(\theta_0, \left(\underbrace{-E\left[\frac{\partial^2 \ell(\theta|\mathbf{x})}{\partial \theta^2}\Big|_{\theta=\theta_0}\right]}_{I_n(\theta_0)}\right)^{-1}\right)$$

Normality

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 $J_n(\hat{\theta}_{MLE})$ to approximate the variance:

$$\hat{\theta}_{MLE} \sim N\left(\theta_0, \left(\underbrace{-\left[\frac{\partial^2 \ell(\theta|\mathbf{x})}{\partial \theta^2}\Big|_{\theta=\hat{\theta}_{MLE}}\right]}_{J_n(\hat{\theta}_{MLE})}\right)^{-1}\right)$$

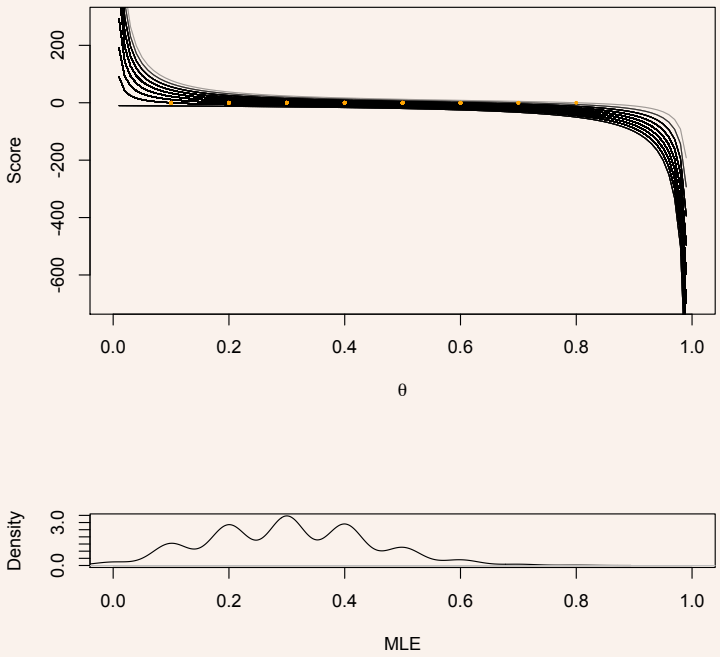
Consistency of the Observed Fisher Information

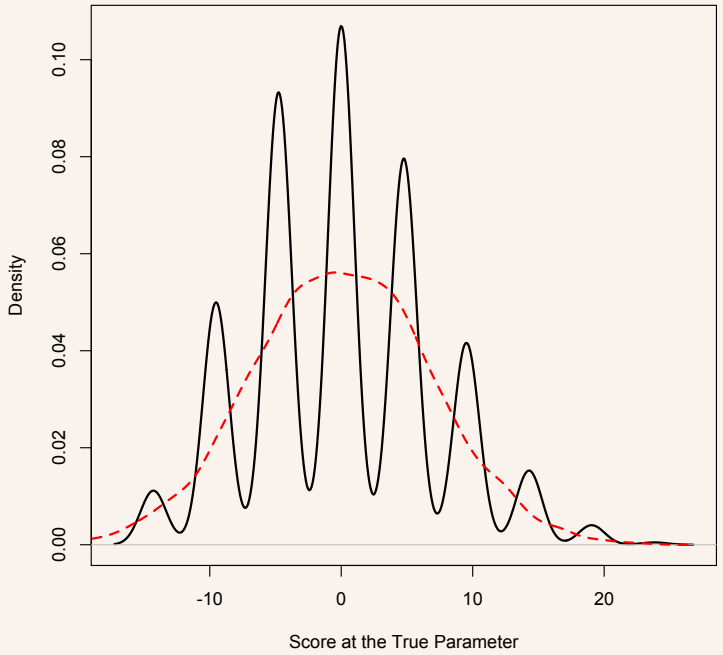


Average observed Fisher information in blue, expected Fisher information in red, and simulated variance across the 10,000 MLEs at each n in gold.

Visualizing Score Functions and Normality of MLE

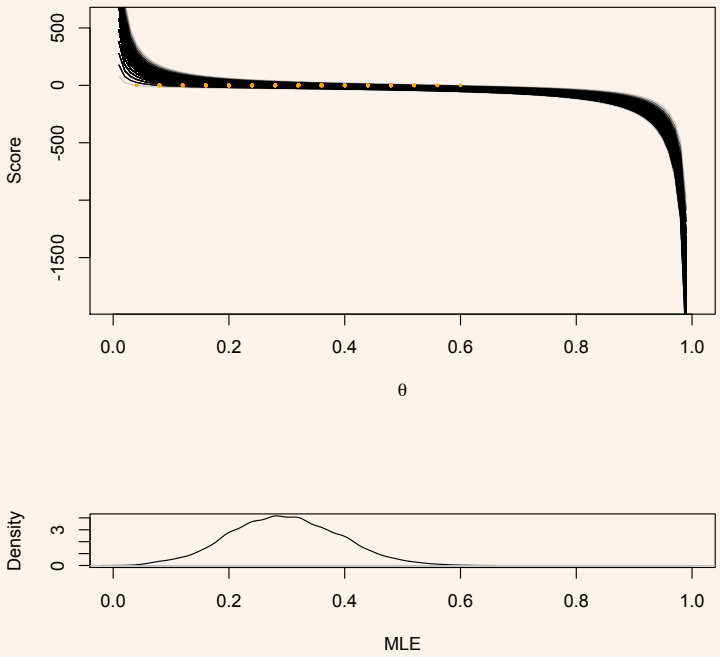
For a sample size of $n = 10$:

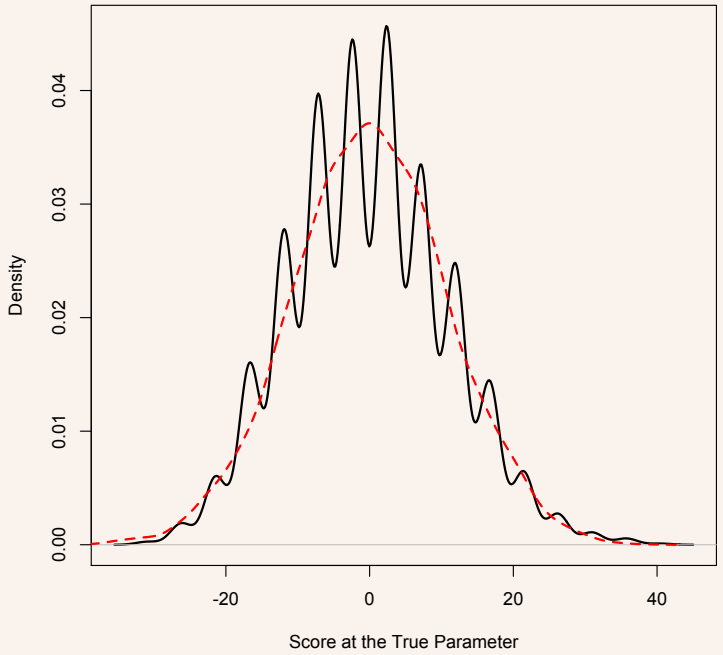




Visualizing Score Functions and Normality of MLE

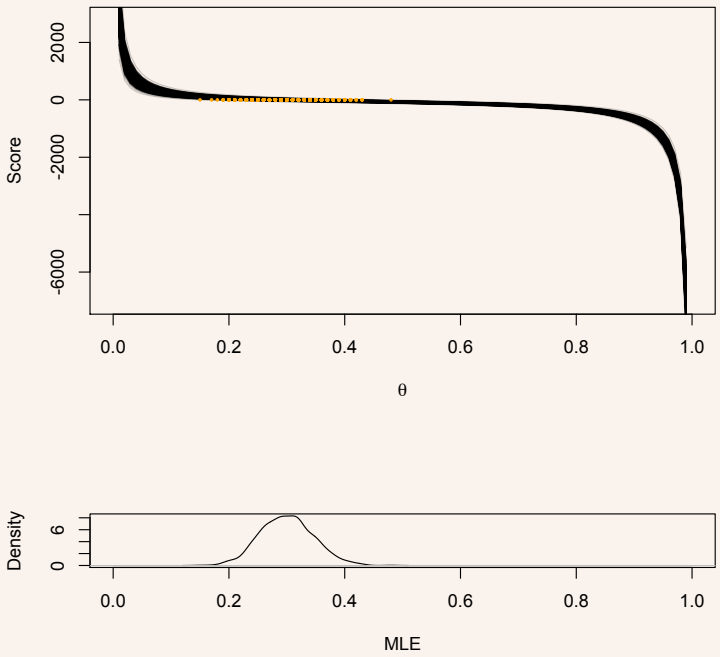
For a sample size of $n = 25$:

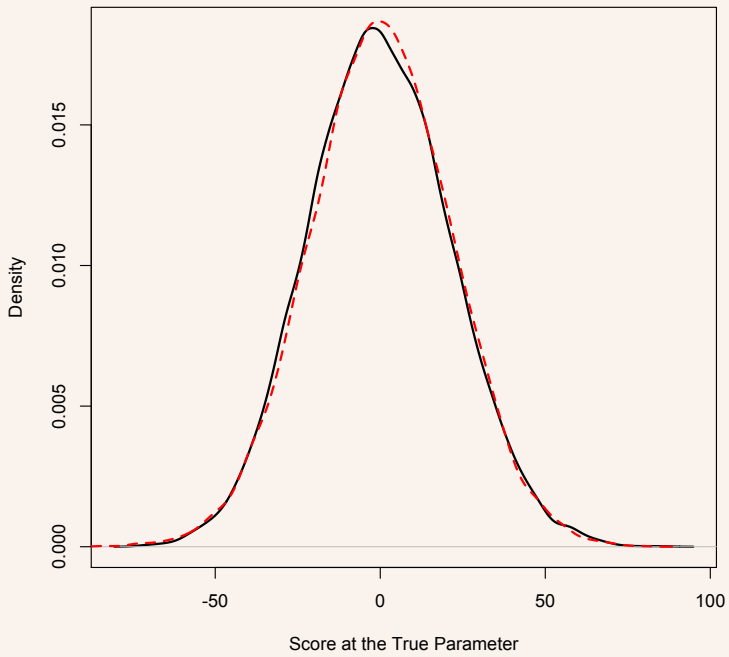




Visualizing Score Functions and Normality of MLE

For a sample size of $n = 100$:





Efficiency

- As sample size (n) increases, MLE is the estimation procedure that generally provides the lowest variance (in the class of other consistent and asymptotically normal estimators)
- ML estimator has a variance equal to Cramer-Rao lower bound