Political Methodology III: Model Based Inference

Justin Grimmer

Professor
Department of Political Science
Stanford University

April 4th, 2019

Statistical Inference

- Model based inference:
- Assume: data generated via distributional process
- Defines a likelihood function: parameter values $\rightarrow$ likelihood of parameters, given data
- Derive estimators that identify values that maximize likelihood

General Likelihood Theory $\rightarrow$ Example $\rightarrow$ General Theory $\rightarrow$ Example

Likelihood Inference, the Basics

Suppose that we collect a random sample of realizations from iid RVs

Likelihood Inference, the Basics

Suppose that we collect a random sample of realizations from iid RVs

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$$

Likelihood Inference, the Basics

Suppose that we collect a random sample of realizations from iid RVs

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$$

Suppose that the distributions have (unobserved) true parameter θ_0, Therefore,

Likelihood Inference, the Basics

Suppose that we collect a random sample of realizations from iid RVs

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$$

Suppose that the distributions have (unobserved) true parameter θ_0, Therefore,

$$f(\boldsymbol{y}|\theta_0) = \prod_{i=1}^{n} f(y_i|\theta_0)$$

Likelihood Inference, the Basics

Suppose that we collect a random sample of realizations from iid RVs

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$$

Suppose that the distributions have (unobserved) true parameter θ_0, Therefore,

$$f(\boldsymbol{y}|\theta_0) = \prod_{i=1}^{n} f(y_i|\theta_0)$$

We're going to be interested in making an inference about θ_0 using the observed data.

Likelihood Inference, the Basics

Assume we know the correct functional form for f–M^*

Likelihood Inference, the Basics

Assume we know the correct functional form for f–M^*
Define the likelihood function $L(\theta|\boldsymbol{y})$ as,

Likelihood Inference, the Basics

Assume we know the correct functional form for f–M^{*}
Define the likelihood function $L(\theta|\boldsymbol{y})$ as,

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

Likelihood Inference, the Basics

Assume we know the correct functional form for f–M^*

Define the likelihood function $L(\theta|\boldsymbol{y})$ as,

$$L(\theta|\boldsymbol{y}) \quad = \quad \prod_{i=1}^{n} f(y_i|\theta)$$

- $L(\theta|\boldsymbol{y}) : \underbrace{\Theta}_{\text{parameter space}} \rightarrow \Re$

Likelihood Inference, the Basics

Assume we know the correct functional form for f–M^*
Define the likelihood function $L(\theta|\boldsymbol{y})$ as,

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

- $L(\theta|\boldsymbol{y}) : \underbrace{\Theta}_{\text{parameter space}} \rightarrow \Re$
- Idea: values of $\theta \in \Theta$ will be more likely if they make the observed data a higher probability

Likelihood Inference, the Basic

If assumptions hold, then:

Likelihood Inference, the Basic

If assumptions hold, then:

- $L(\theta|\boldsymbol{y})$ encodes all information about likely values of θ, given data

Likelihood Inference, the Basic

If assumptions hold, then:

- $L(\theta|\boldsymbol{y})$ encodes all information about likely values of θ, given data
- Use $L(\theta|\boldsymbol{y})$ to find most likely value of θ and uncertainty about that estimate

Likelihood Inference, the Basic

If assumptions hold, then:

- $L(\theta|\boldsymbol{y})$ encodes all information about likely values of θ, given data
- Use $L(\theta|\boldsymbol{y})$ to find most likely value of θ and uncertainty about that estimate

Likelihood is a relative concept.

Likelihood Inference, the Basic

If assumptions hold, then:

- $L(\theta|\boldsymbol{y})$ encodes all information about likely values of θ, given data
- Use $L(\theta|\boldsymbol{y})$ to find most likely value of θ and uncertainty about that estimate

Likelihood is a relative concept. We can compare within models.

Likelihood Inference, the Basic

If assumptions hold, then:

- $L(\theta|\boldsymbol{y})$ encodes all information about likely values of θ, given data
- Use $L(\theta|\boldsymbol{y})$ to find most likely value of θ and uncertainty about that estimate

Likelihood is a relative concept. We can compare within models. But not across models

Likelihood Inference, the Basic

If assumptions hold, then:

- $L(\theta|\boldsymbol{y})$ encodes all information about likely values of θ, given data
- Use $L(\theta|\boldsymbol{y})$ to find most likely value of θ and uncertainty about that estimate

Likelihood is a relative concept. We can compare within models. But not across models

We also are only able to infer most likely value given modeling assumptions

Example: Bernoulli Trials

Suppose that we observe a sample of independentally and identically distributed observations that are Bernoulli distributed,

Example: Bernoulli Trials

Suppose that we observe a sample of independentally and identically distributed observations that are Bernoulli distributed,

$$Y_i \sim \text{Bernoulli}(\pi)$$

Example: Bernoulli Trials

Suppose that we observe a sample of independentally and identically distributed observations that are Bernoulli distributed,

$$Y_i \sim \text{Bernoulli}(\pi)$$

- Recall

Example: Bernoulli Trials

Suppose that we observe a sample of independentally and identically distributed observations that are Bernoulli distributed,

$$Y_i \sim \text{Bernoulli}(\pi)$$

- Recall

$$\text{Bernoulli}(\pi) \quad = \quad \pi^{y_i}(1-\pi)^{1-y_i}$$

Example: Bernoulli Trials

Suppose that we observe a sample of independentally and identically distributed observations that are Bernoulli distributed,

$$Y_i \sim \text{Bernoulli}(\pi)$$

- Recall

$$\text{Bernoulli}(\pi) = \pi^{y_i}(1-\pi)^{1-y_i}$$

- $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$

Example: Bernoulli Trials

Suppose that we observe a sample of independentally and identically distributed observations that are Bernoulli distributed,

$$Y_i \sim \text{Bernoulli}(\pi)$$

- Recall

$$\text{Bernoulli}(\pi) \quad = \quad \pi^{y_i}(1-\pi)^{1-y_i}$$

- $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$
- $y_i = 1$ or $y_i = 0$

Example: Bernoulli Trials

Example: Bernoulli Trials

$$L(\pi|\boldsymbol{y}) = f(\boldsymbol{y}|\pi)$$

Example: Bernoulli Trials

$$
\begin{aligned}
L(\pi|\boldsymbol{y}) &= f(\boldsymbol{y}|\pi) \\
&= \prod_{i=1}^{n} f(y_i|\pi)
\end{aligned}
$$

Example: Bernoulli Trials

$$
\begin{aligned}
L(\pi|\boldsymbol{y}) &= f(\boldsymbol{y}|\pi) \\
&= \prod_{i=1}^{n} f(y_i|\pi) \\
&= \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}
\end{aligned}
$$

Example: Bernoulli Trials

$$
\begin{aligned}
L(\pi|\boldsymbol{y}) &= f(\boldsymbol{y}|\pi) \\
&= \prod_{i=1}^{n} f(y_i|\pi) \\
&= \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i} \\
&= \pi^{\sum_{i=1}^{n} y_i}(1-\pi)^{n-\sum_{i=1}^{n} y_i}
\end{aligned}
$$

Example: Bernoulli Trials

$$\begin{aligned}
L(\pi|\boldsymbol{y}) &= f(\boldsymbol{y}|\pi) \\
&= \prod_{i=1}^{n} f(y_i|\pi) \\
&= \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i} \\
&= \pi^{\sum_{i=1}^{n} y_i}(1-\pi)^{n-\sum_{i=1}^{n} y_i}
\end{aligned}$$

We'll work with the natural logarithm of the likelihood,

Example: Bernoulli Trials

$$
\begin{aligned}
L(\pi|\boldsymbol{y}) &= f(\boldsymbol{y}|\pi) \\
&= \prod_{i=1}^{n} f(y_i|\pi) \\
&= \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i} \\
&= \pi^{\sum_{i=1}^{n} y_i}(1-\pi)^{n-\sum_{i=1}^{n} y_i}
\end{aligned}
$$

We'll work with the natural logarithm of the likelihood,

$$
\log L(\pi|\boldsymbol{y}) \equiv l(\pi|\boldsymbol{y}) = \sum_{i=1}^{n} y_i \log \pi + (n - \sum_{i=1}^{n} y_i)\log(1-\pi)
$$

Example: Bernoulli Trials

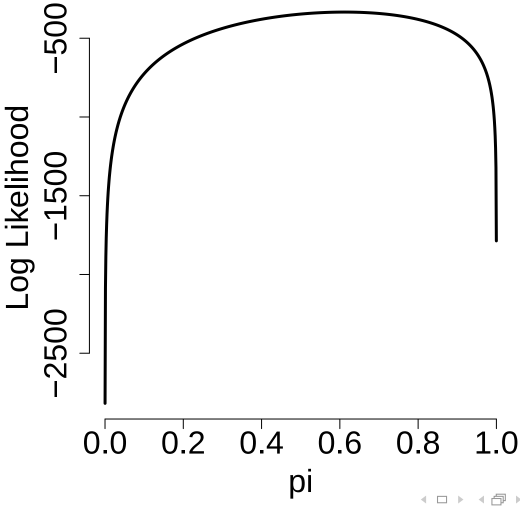
$$
\begin{aligned}
L(\pi|\boldsymbol{y}) &= f(\boldsymbol{y}|\pi) \\
&= \prod_{i=1}^{n} f(y_i|\pi) \\
&= \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i} \\
&= \pi^{\sum_{i=1}^{n} y_i}(1-\pi)^{n-\sum_{i=1}^{n} y_i}
\end{aligned}
$$

We'll work with the natural logarithm of the likelihood,

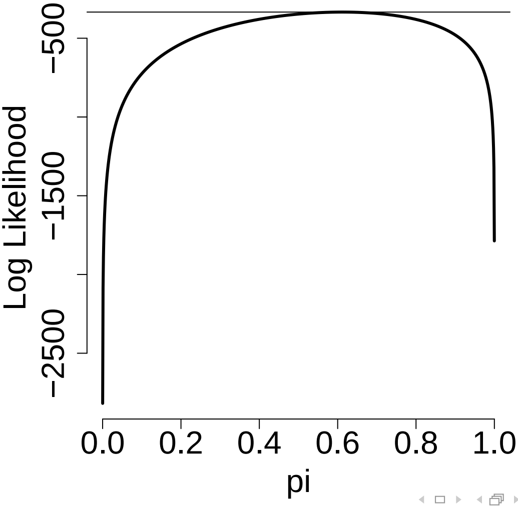
$$
\log L(\pi|\boldsymbol{y}) \equiv l(\pi|\boldsymbol{y}) = \sum_{i=1}^{n} y_i \log \pi + (n - \sum_{i=1}^{n} y_i) \log(1-\pi)
$$

For a fixed set of observations, what does this look like?

Example: Bernoulli Trials: Simulated Example with $\pi = 0.6$



Example: Bernoulli Trials: Simulated Example with $\pi = 0.6$



Example: Bernoulli Trials

Look for estimator that optimizes (log)-likelihood

Example: Bernoulli Trials

Look for estimator that optimizes (log)-likelihood

Previous plot: optimize by looking for stationary points

Example: Bernoulli Trials

Look for estimator that optimizes (log)-likelihood

Previous plot: optimize by looking for stationary points

$$l(\pi|\boldsymbol{y}) = \sum_{i=1}^{n} y_i \log \pi + (n - \sum_{i=1}^{n} y_i) \log(1-\pi)$$

Example: Bernoulli Trials

Look for estimator that optimizes (log)-likelihood

Previous plot: optimize by looking for stationary points

$$
\begin{aligned}
l(\pi|\boldsymbol{y}) &= \sum_{i=1}^{n} y_i \log \pi + (n - \sum_{i=1}^{n} y_i) \log(1-\pi) \\
\underbrace{\frac{\partial l(\pi|\boldsymbol{y})}{\partial \pi}}_{\text{Score Function}} &= \frac{\sum_{i=1}^{n} y_i}{\pi} - \frac{(n - \sum_{i=1}^{n} y_i)}{1-\pi}
\end{aligned}
$$

Example: Bernoulli Trials

Look for estimator that optimizes (log)-likelihood
Previous plot: optimize by looking for stationary points

$$\begin{aligned}
l(\pi|\boldsymbol{y}) &= \sum_{i=1}^{n} y_i \log \pi + (n - \sum_{i=1}^{n} y_i) \log(1-\pi) \\
\underbrace{\frac{\partial l(\pi|\boldsymbol{y})}{\partial \pi}}_{\text{Score Function}} &= \frac{\sum_{i=1}^{n} y_i}{\pi} - \frac{(n - \sum_{i=1}^{n} y_i)}{1-\pi} \\
0 &= \frac{\sum_{i=1}^{n} y_i}{\pi^*} - \frac{(n - \sum_{i=1}^{n} y_i)}{1-\pi^*}
\end{aligned}$$

Example: Bernoulli Trials

Look for estimator that optimizes (log)-likelihood

Previous plot: optimize by looking for stationary points

$$
\begin{aligned}
l(\pi|\boldsymbol{y}) &= \sum_{i=1}^{n} y_i \log \pi + (n - \sum_{i=1}^{n} y_i) \log(1-\pi) \\
\underbrace{\frac{\partial l(\pi|\boldsymbol{y})}{\partial \pi}}_{\text{Score Function}} &= \frac{\sum_{i=1}^{n} y_i}{\pi} - \frac{(n - \sum_{i=1}^{n} y_i)}{1-\pi} \\
0 &= \frac{\sum_{i=1}^{n} y_i}{\pi^*} - \frac{(n - \sum_{i=1}^{n} y_i)}{1-\pi^*} \\
\frac{(n - \sum_{i=1}^{n} y_i)}{1-\pi^*} &= \frac{\sum_{i=1}^{n} y_i}{\pi^*}
\end{aligned}
$$

Example: Bernoulli Trials

Look for estimator that optimizes (log)-likelihood
Previous plot: optimize by looking for stationary points

$$
\begin{aligned}
l(\pi|\boldsymbol{y}) &= \sum_{i=1}^{n} y_i \log \pi + (n - \sum_{i=1}^{n} y_i) \log(1-\pi) \\
\underbrace{\frac{\partial l(\pi|\boldsymbol{y})}{\partial \pi}}_{\text{Score Function}} &= \frac{\sum_{i=1}^{n} y_i}{\pi} - \frac{(n - \sum_{i=1}^{n} y_i)}{1-\pi} \\
0 &= \frac{\sum_{i=1}^{n} y_i}{\pi^*} - \frac{(n - \sum_{i=1}^{n} y_i)}{1-\pi^*} \\
\frac{(n - \sum_{i=1}^{n} y_i)}{1-\pi^*} &= \frac{\sum_{i=1}^{n} y_i}{\pi^*} \\
(n - \sum_{i=1}^{n} y_i)\pi^* &= (1-\pi^*) \sum_{i=1}^{n} y_i
\end{aligned}
$$

Example: Bernoulli Trials

Look for estimator that optimizes (log)-likelihood
Previous plot: optimize by looking for stationary points

$$
\begin{aligned}
l(\pi|\boldsymbol{y}) &= \sum_{i=1}^{n} y_i \log \pi + (n - \sum_{i=1}^{n} y_i) \log(1-\pi) \\
\underbrace{\frac{\partial l(\pi|\boldsymbol{y})}{\partial \pi}}_{\text{Score Function}} &= \frac{\sum_{i=1}^{n} y_i}{\pi} - \frac{(n - \sum_{i=1}^{n} y_i)}{1-\pi} \\
0 &= \frac{\sum_{i=1}^{n} y_i}{\pi^*} - \frac{(n - \sum_{i=1}^{n} y_i)}{1-\pi^*} \\
\frac{(n - \sum_{i=1}^{n} y_i)}{1-\pi^*} &= \frac{\sum_{i=1}^{n} y_i}{\pi^*} \\
(n - \sum_{i=1}^{n} y_i)\pi^* &= (1-\pi^*) \sum_{i=1}^{n} y_i \\
\pi^* n &= \sum_{i=1}^{n} y_i
\end{aligned}
$$

Example: Bernoulli Trials

Look for estimator that optimizes (log)-likelihood
Previous plot: optimize by looking for stationary points

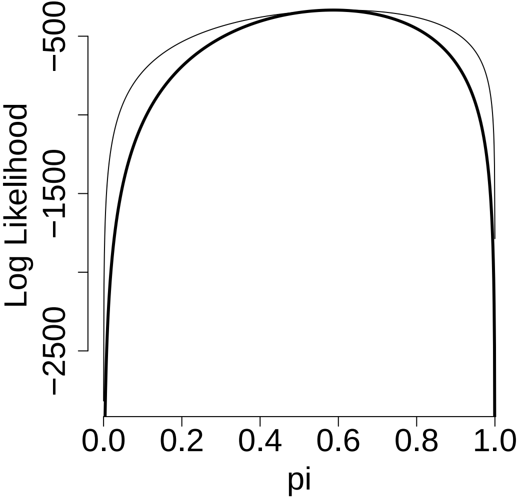
$$
\begin{aligned}
l(\pi|\boldsymbol{y}) &= \sum_{i=1}^{n} y_i \log \pi + (n - \sum_{i=1}^{n} y_i) \log(1-\pi) \\
\underbrace{\frac{\partial l(\pi|\boldsymbol{y})}{\partial \pi}}_{\text{Score Function}} &= \frac{\sum_{i=1}^{n} y_i}{\pi} - \frac{(n - \sum_{i=1}^{n} y_i)}{1-\pi} \\
0 &= \frac{\sum_{i=1}^{n} y_i}{\pi^*} - \frac{(n - \sum_{i=1}^{n} y_i)}{1-\pi^*} \\
\frac{(n - \sum_{i=1}^{n} y_i)}{1-\pi^*} &= \frac{\sum_{i=1}^{n} y_i}{\pi^*} \\
(n - \sum_{i=1}^{n} y_i)\pi^* &= (1-\pi^*) \sum_{i=1}^{n} y_i \\
\pi^* n &= \sum_{i=1}^{n} y_i \\
\pi^* &= \overline{y}
\end{aligned}
$$

Uncertainty About Mode

$\pi^* = \overline{y}$ maximizes $L(\pi|\boldsymbol{y})$.

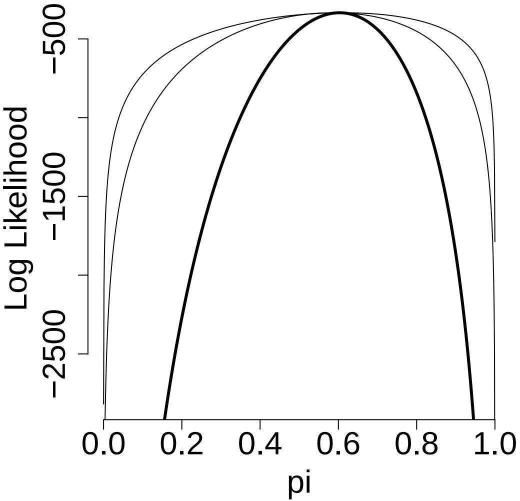
Uncertainty About Mode

$\pi^* = \overline{y}$ maximizes $L(\pi|\boldsymbol{y})$. How much uncertainty is there about this maximum?



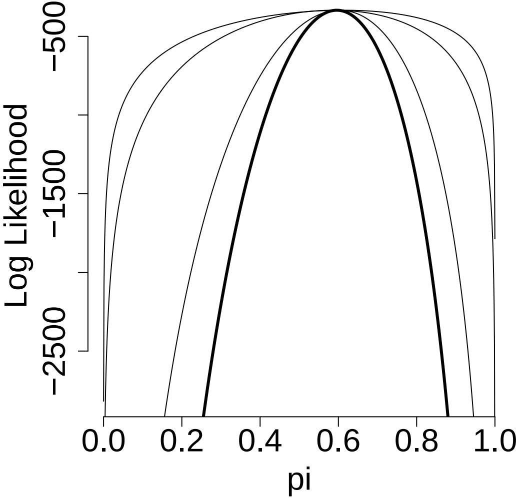
Uncertainty About Mode

$\pi^* = \bar{y}$ maximizes $L(\pi|\boldsymbol{y})$. How much uncertainty is there about this maximum?



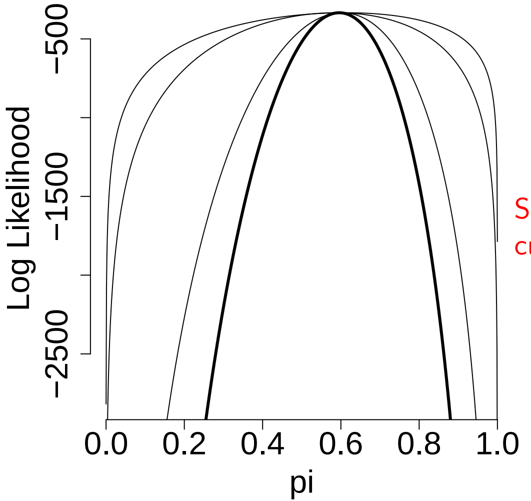
Uncertainty About Mode

$\pi^* = \overline{y}$ maximizes $L(\pi|\boldsymbol{y})$. How much uncertainty is there about this maximum?



Uncertainty About Mode

$\pi^* = \overline{y}$ maximizes $L(\pi|\boldsymbol{y})$. How much uncertainty is there about this maximum?



Second derivative captures this curvature

The Fisher Information measures the information that $\boldsymbol{y}$ conveys about the parameter $\boldsymbol{\theta}$. Define it using the two equivalent definitions:

Definition

The Fisher Information for a log-likelihood $l(\theta|\boldsymbol{Y})$ is

$$
\begin{aligned}
I(\theta) &= E\left[\left(\frac{\partial l(\theta|\boldsymbol{Y})}{\partial \theta}\right)^2 |\theta\right] \\
&= -E\left[\left(\frac{\partial^2 l(\theta|\boldsymbol{Y})}{\partial \theta \partial \theta}\right) |\theta\right]
\end{aligned}
$$

The observed Fisher information for a sample of n observations is given by

$$
I_n(\theta) = -\frac{\partial^2}{\partial \theta^2} l(\theta|\boldsymbol{y})
$$

Uncertainty About Mode

Information let's us know how much we learn about θ from our sample $\boldsymbol{y}$

Uncertainty About Mode

Information let's us know how much we learn about θ from our sample $\boldsymbol{y}$

- If $I_n(\theta)$ is big at MLE, more information. Or more concentrated around mode

Uncertainty About Mode

Information let's us know how much we learn about θ from our sample $\boldsymbol{y}$

- If $I_n(\theta)$ is big at MLE, more information. Or more concentrated around mode
- If $I_n(\theta)$ is small at MLE, less information. Or less concentrated around the mode

Uncertainty About Mode

Inverting the information provides the asymptotic variance for the maximum likelihood estimator (under some regulatory conditions we will discuss later)

$$
\begin{aligned}
\text{Variance}(\theta^*) &= \frac{1}{I_n(\theta^*)} \\
\text{Standard Error}(\theta^*) &= \sqrt{\frac{1}{I_n(\theta^*)}}
\end{aligned}
$$

Example: Bernoulli Trials

Calculating second derivate for Bernoulli example:

Example: Bernoulli Trials

Calculating second derivate for Bernoulli example:

$$\frac{\partial^2 l(\pi|\boldsymbol{y})}{\partial \pi^2} = -\frac{\partial^2}{\partial \pi^2} l(\pi|\boldsymbol{Y}) \quad = \quad \frac{\sum_{i=1}^{n} Y_i}{\pi^2} + \frac{n - \sum_{i=1}^{n} Y_i}{(1-\pi)^2}$$

Example: Bernoulli Trials

Calculating second derivate for Bernoulli example:

$$\frac{\partial^2 l(\pi|\boldsymbol{y})}{\partial \pi^2} = -\frac{\partial^2}{\partial \pi^2} l(\pi|\boldsymbol{Y}) = \frac{\sum_{i=1}^{n} Y_i}{\pi^2} + \frac{n - \sum_{i=1}^{n} Y_i}{(1-\pi)^2}$$

$$I_n(\pi^*) = \frac{n}{\overline{y}(1-\overline{y})}$$

Example: Bernoulli Trials

Calculating second derivate for Bernoulli example:

$$\frac{\partial^2 l(\pi|\boldsymbol{y})}{\partial \pi^2} = -\frac{\partial^2}{\partial \pi^2} l(\pi|\boldsymbol{Y}) = \frac{\sum_{i=1}^n Y_i}{\pi^2} + \frac{n - \sum_{i=1}^n Y_i}{(1-\pi)^2}$$

$$I_n(\pi^*) = \frac{n}{\overline{y}(1-\overline{y})}$$

- The bigger the n, the more curved at the mode.

Example: Bernoulli Trials

Calculating second derivate for Bernoulli example:

$$\frac{\partial^2 l(\pi|\boldsymbol{y})}{\partial \pi^2} = -\frac{\partial^2}{\partial \pi^2} l(\pi|\boldsymbol{Y}) = \frac{\sum_{i=1}^n Y_i}{\pi^2} + \frac{n - \sum_{i=1}^n Y_i}{(1-\pi)^2}$$
$$I_n(\pi^*) = \frac{n}{\overline{y}(1-\overline{y})}$$

- The bigger the n, the more curved at the mode.
- Calculating the variance

Example: Bernoulli Trials

Calculating second derivate for Bernoulli example:

$$\begin{aligned}
\frac{\partial^2 l(\pi|\boldsymbol{y})}{\partial \pi^2} = -\frac{\partial^2}{\partial \pi^2} l(\pi|\boldsymbol{Y}) &= \frac{\sum_{i=1}^{n} Y_i}{\pi^2} + \frac{n - \sum_{i=1}^{n} Y_i}{(1-\pi)^2} \\
I_n(\pi^*) &= \frac{n}{\overline{y}(1-\overline{y})}
\end{aligned}$$

- The bigger the n, the more curved at the mode.
- Calculating the variance

$$Var(\pi^*) = 1/I_n(\pi^*)$$

Example: Bernoulli Trials

Calculating second derivate for Bernoulli example:

$$\frac{\partial^2 l(\pi|\boldsymbol{y})}{\partial \pi^2} = -\frac{\partial^2}{\partial \pi^2} l(\pi|\boldsymbol{Y}) = \frac{\sum_{i=1}^n Y_i}{\pi^2} + \frac{n - \sum_{i=1}^n Y_i}{(1-\pi)^2}$$

$$I_n(\pi^*) = \frac{n}{\overline{y}(1-\overline{y})}$$

- The bigger the n, the more curved at the mode.
- Calculating the variance

$$\begin{aligned} Var(\pi^*) &= 1/I_n(\pi^*) \\ &= \frac{\overline{y}(1-\overline{y})}{n} \end{aligned}$$

Example: Bernoulli Trials

Calculating second derivate for Bernoulli example:

$$\begin{aligned}
\frac{\partial^2 l(\pi|\boldsymbol{y})}{\partial\pi^2} = -\frac{\partial^2}{\partial\pi^2} l(\pi|\boldsymbol{Y}) &= \frac{\sum_{i=1}^n Y_i}{\pi^2} + \frac{n - \sum_{i=1}^n Y_i}{(1-\pi)^2} \\
I_n(\pi^*) &= \frac{n}{\overline{y}(1-\overline{y})}
\end{aligned}$$

- The bigger the n, the more curved at the mode.
- Calculating the variance

$$\begin{aligned}
Var(\pi^*) &= 1/I_n(\pi^*) \\
&= \frac{\overline{y}(1-\overline{y})}{n}
\end{aligned}$$

- Curvature determines sampling distribution of maximum likelihood estimator

Properties of Maximum Likelihood Estimators

Likelihoods are summary estimators

Properties of Maximum Likelihood Estimators

Likelihoods are summary estimators: throw away data if you believe assumptions

Properties of Maximum Likelihood Estimators

Likelihoods are summary estimators: throw away data if you believe assumptions

Maximum likelihood estimator: point that maximizes likelihood

Properties of Maximum Likelihood Estimators

Likelihoods are summary estimators: throw away data if you believe assumptions

Maximum likelihood estimator: point that maximizes likelihood

Large sample properties given the correct distributional family is known

Properties of Maximum Likelihood Estimators

Likelihoods are summary estimators: throw away data if you believe assumptions

Maximum likelihood estimator: point that maximizes likelihood

Large sample properties given the correct distributional family is known

1) Consistent: As the sample size gets large, the mle converges on true value

Properties of Maximum Likelihood Estimators

Likelihoods are summary estimators: throw away data if you believe assumptions

Maximum likelihood estimator: point that maximizes likelihood

Large sample properties given the correct distributional family is known

1) Consistent: As the sample size gets large, the mle converges on true value
2) Convergence in Distribution As the sample size gets large, the mle becomes normally distributed, with mean at the true value and variance inverted information at maximum likelihood estimator

Properties of Maximum Likelihood Estimators

Likelihoods are summary estimators: throw away data if you believe assumptions

Maximum likelihood estimator: point that maximizes likelihood

Large sample properties given the correct distributional family is known

1) Consistent: As the sample size gets large, the mle converges on true value
2) Convergence in Distribution As the sample size gets large, the mle becomes normally distributed, with mean at the true value and variance inverted information at maximum likelihood estimator
3) Asymptotic Efficiency: MLE has the smallest variance for "well-behaved" estimator (Cramer-Rao Lower Bound obtained)

Properties of Maximum Likelihood Estimators

Likelihoods are summary estimators: throw away data if you believe assumptions

Maximum likelihood estimator: point that maximizes likelihood

Large sample properties given the correct distributional family is known

1) Consistent: As the sample size gets large, the mle converges on true value
2) Convergence in Distribution As the sample size gets large, the mle becomes normally distributed, with mean at the true value and variance inverted information at maximum likelihood estimator
3) Asymptotic Efficiency: MLE has the smallest variance for "well-behaved" estimator (Cramer-Rao Lower Bound obtained)

Small Sample Properties given correct distributional family is known

Small Sample Properties given correct distributional family is known

1) Equivariance: if $\widehat{\theta}$ is the MLE for θ and $g : \Theta \rightarrow \Re$ is a one-to-one function. Then $g(\widehat{\theta})$ is the MLE of $g(\theta)$.

Small Sample Properties given correct distributional family is known

1) Equivariance: if $\widehat{\theta}$ is the MLE for θ and $g : \Theta \rightarrow \Re$ is a one-to-one function. Then $g(\widehat{\theta})$ is the MLE of $g(\theta)$.
2) NO GUARANTEE OF UNBIASEDNESS!!!

Small Sample Properties given correct distributional family is known

1) Equivariance: if $\widehat{\theta}$ is the MLE for θ and $g : \Theta \rightarrow \Re$ is a one-to-one function. Then $g(\widehat{\theta})$ is the MLE of $g(\theta)$.
2) NO GUARANTEE OF UNBIASEDNESS!!!

Large sample properties given incorrect model

Small Sample Properties given correct distributional family is known

1) Equivariance: if $\widehat{\theta}$ is the MLE for θ and $g : \Theta \rightarrow \Re$ is a one-to-one function. Then $g(\widehat{\theta})$ is the MLE of $g(\theta)$.
2) NO GUARANTEE OF UNBIASEDNESS!!!

Large sample properties given incorrect model

1) The MLE gets as "close" as possible to the true answer

Properties of Maximum Likelihood Estimators

Properties of Maximum Likelihood Estimators

Definition

Consistent Let $\widehat{\theta}_n$ be an estimator for θ, with sample size n. Then $\widehat{\theta}_n$ converges in probability to θ if, for all $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\widehat{\theta}_n - \theta| > \epsilon) \quad \rightarrow \quad 0$$

Properties of Maximum Likelihood Estimators

Definition

Consistent Let $\widehat{\theta}_n$ be an estimator for θ, with sample size n. Then $\widehat{\theta}_n$ converges in probability to θ if, for all $\epsilon > 0$,

$$\lim_{n\rightarrow\infty} P(|\widehat{\theta}_n - \theta| > \epsilon) \quad \rightarrow \quad 0$$

Proposition

MLE is consistent: Assume $y_1, y_2, \ldots, y_n$ are simple random samples from $p(y|\theta_0)$. Define θ_n^* as the mle estimator with sample size n.
Then, as $n \rightarrow \infty$, $\theta_n^* \rightarrow \theta_0$ (in probability)

Properties of Maximum Likelihood Estimators

Simulated example with $\pi = 0.6$ and increasing n

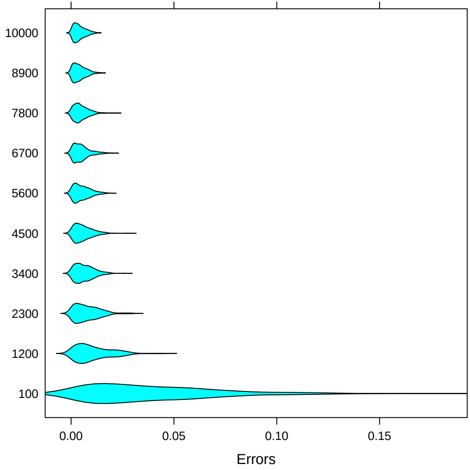
Properties of Maximum Likelihood Estimators

Simulated example with $\pi = 0.6$ and increasing n
Distribution of $|\pi^* - 0.6|$

Properties of Maximum Likelihood Estimators

Simulated example with $\pi = 0.6$ and increasing n

Distribution of $|\pi^* - 0.6|$



Definition

$\widehat{\theta}_n$, *with cdf* $F_n(x)$, *converges in distribution to random variable* Y *with cdf* $F(x)$ *if*

$$\lim_{n \rightarrow \infty} |F_n(x) - F(x)| = 0$$

For all $x \in \Re$ *where* $F(x)$ *is continuous.*

Definition

$\widehat{\theta}_n$, *with cdf* $F_n(x)$, *converges in distribution to random variable* Y *with cdf* $F(x)$ *if*

$$\lim_{n\to\infty} |F_n(x) - F(x)| = 0$$

For all $x \in \Re$ *where* $F(x)$ *is continuous.*

Properties of Maximum Likelihood Estimators

,

Properties of Maximum Likelihood Estimators

Proposition

Suppose that $y_1, y_2, \ldots, y_n$ *are simple random samples from* $p(y|\theta_0)$. *Define* θ^* *as the MLE. Then as* $n \rightarrow \infty$,

Properties of Maximum Likelihood Estimators

Proposition

Suppose that $y_1, y_2, \ldots, y_n$ *are simple random samples from* $p(y|\theta_0)$*.*
Define θ^* *as the MLE. Then as* $n \rightarrow \infty$ *,*

$$\sqrt{n}(\theta^* - \theta_0) \quad \rightarrow^d \quad \textit{Normal}(0, \frac{1}{I_n(\theta_0)})$$

Properties of Maximum Likelihood Estimators

Proposition

Suppose that $y_1, y_2, \ldots, y_n$ *are simple random samples from* $p(y|\theta_0)$*. Define* θ^* *as the MLE. Then as* $n \to \infty$ *,*

$$\sqrt{n}(\theta^* - \theta_0) \quad \to^d \quad \textit{Normal}(0, \frac{1}{I_n(\theta_0)})$$
$$\sqrt{n}(\theta^* - \theta_0) \quad \to^d \quad \textit{Normal}(0, \frac{1}{I_n(\theta^*)})$$

Properties of Maximum Likelihood Estimators

Proposition

Suppose that $y_1, y_2, \ldots, y_n$ are simple random samples from $p(y|\theta_0)$. Define θ^* as the MLE. Then as $n \to \infty$,

$$\sqrt{n}(\theta^* - \theta_0) \quad \rightarrow^d \quad \textit{Normal}(0, \frac{1}{I_n(\theta_0)})$$
$$\sqrt{n}(\theta^* - \theta_0) \quad \rightarrow^d \quad \textit{Normal}(0, \frac{1}{I_n(\theta^*)})$$

- MLE central limit theorem

Properties of Maximum Likelihood Estimators

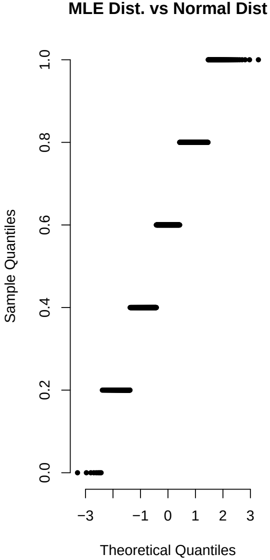
Proposition

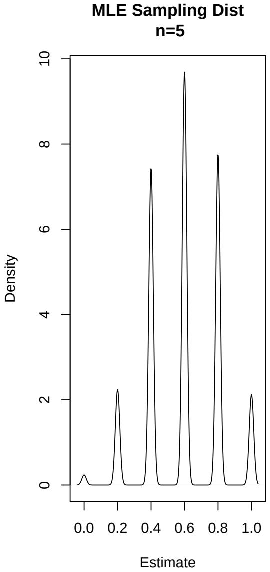
Suppose that $y_1, y_2, \ldots, y_n$ *are simple random samples from* $p(y|\theta_0)$. *Define* θ^* *as the MLE. Then as* $n \rightarrow \infty$,

$$\sqrt{n}(\theta^* - \theta_0) \quad \rightarrow^d \quad \textit{Normal}(0, \frac{1}{I_n(\theta_0)})$$
$$\sqrt{n}(\theta^* - \theta_0) \quad \rightarrow^d \quad \textit{Normal}(0, \frac{1}{I_n(\theta^*)})$$

- MLE central limit theorem
- As we have more observations, the MLE converges, in distribution to a normal distribution

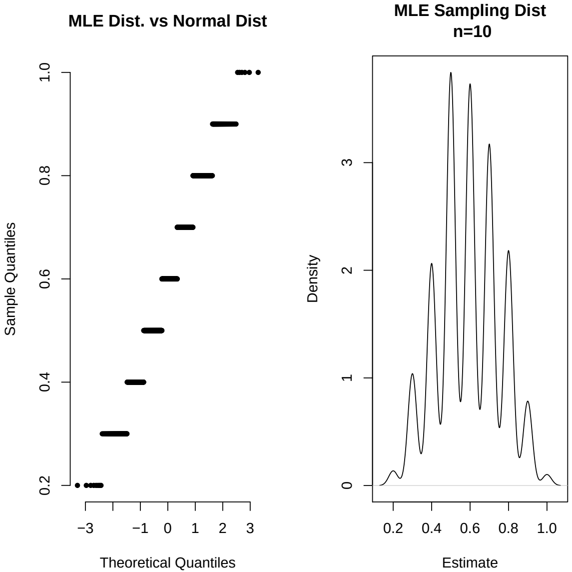
Central Limit Theorem for Maximum Likelihood Estimators





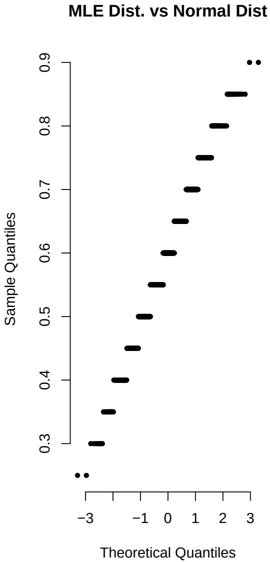
Example:
$\pi = 0.6$
Increasing n

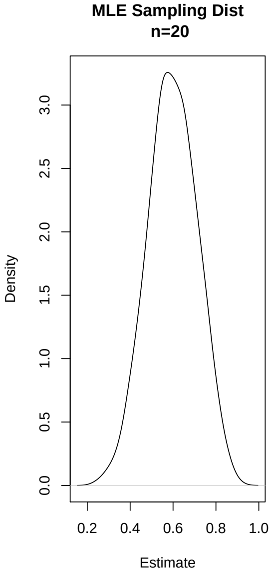
Central Limit Theorem for Maximum Likelihood Estimators



Example:
$\pi = 0.6$
Increasing n

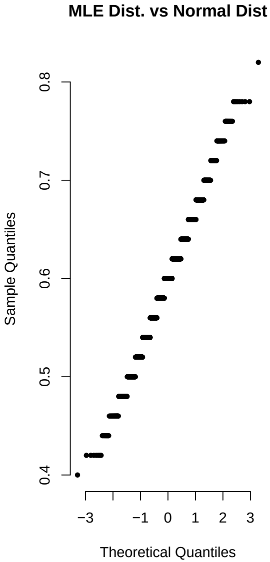
Central Limit Theorem for Maximum Likelihood Estimators

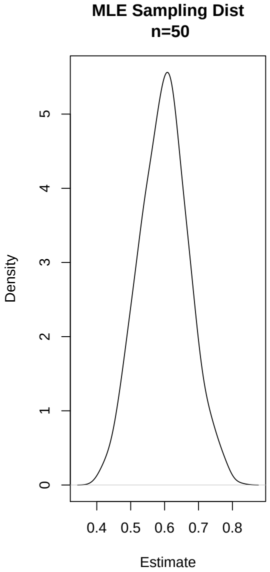




Example:
$\pi = 0.6$
Increasing n

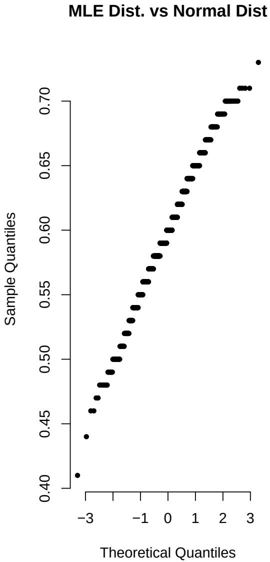
Central Limit Theorem for Maximum Likelihood Estimators

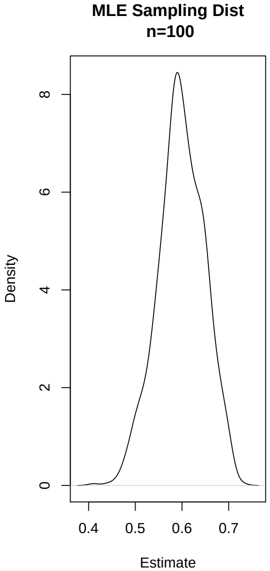




Example:
$\pi = 0.6$
Increasing n

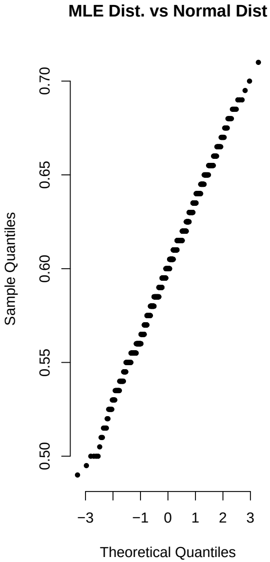
Central Limit Theorem for Maximum Likelihood Estimators

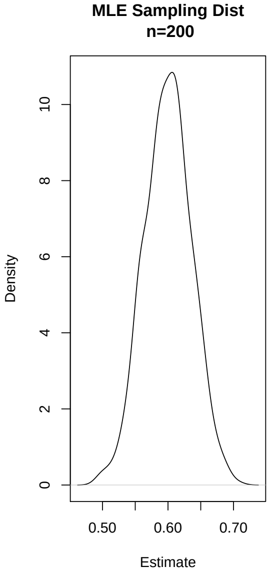




Example:
$\pi = 0.6$
Increasing n

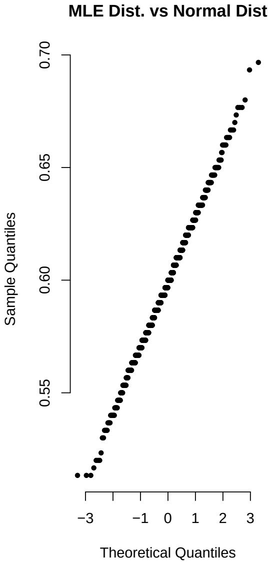
Central Limit Theorem for Maximum Likelihood Estimators

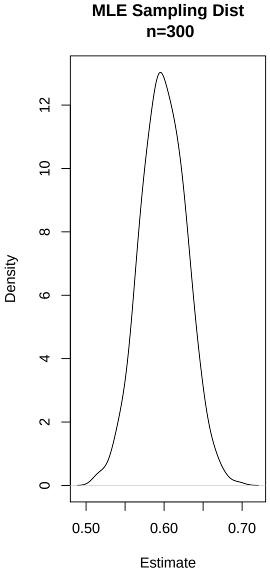




Example:
$\pi = 0.6$
Increasing n

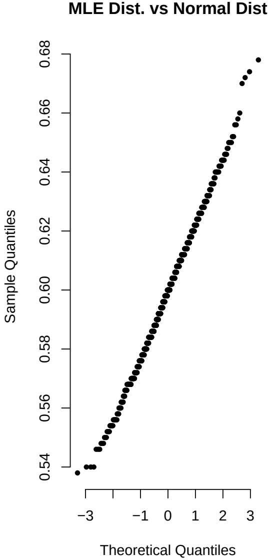
Central Limit Theorem for Maximum Likelihood Estimators

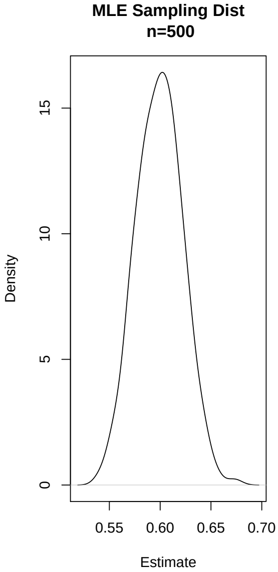




Example:
$\pi = 0.6$
Increasing n

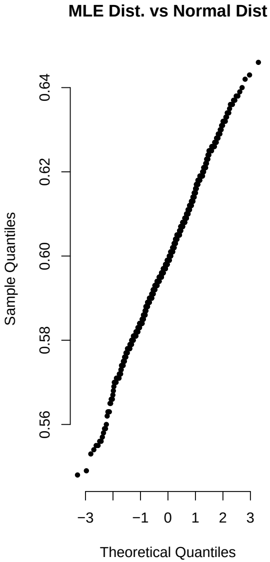
Central Limit Theorem for Maximum Likelihood Estimators

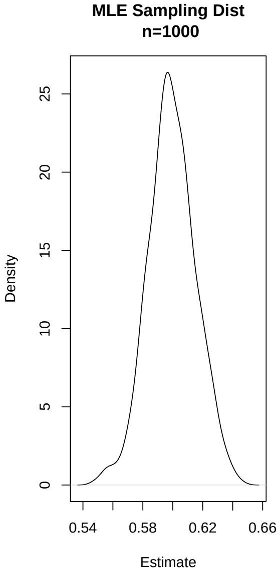




Example:
$\pi = 0.6$
Increasing n

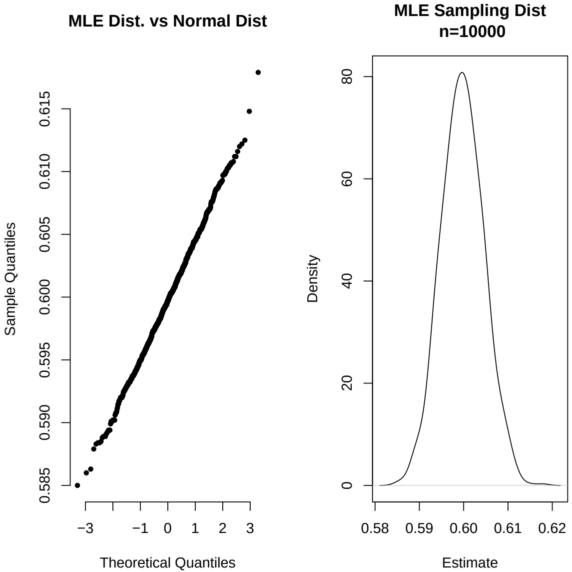
Central Limit Theorem for Maximum Likelihood Estimators





Example:
$\pi = 0.6$
Increasing n

Central Limit Theorem for Maximum Likelihood Estimators



Example:
$\pi = 0.6$
Increasing n

Returning to Our Examples

Summary: θ^* is an MLE and θ_0 is the true value and we know the right distributional family

1) As $n \rightarrow \infty$, $\theta_n^* \rightarrow \theta_0$
2) As $n \rightarrow \infty$, $\sqrt{n}(\theta_n^* - \theta_0) \rightarrow \text{Normal}(0, \frac{1}{I(\theta_0)})$

where $I(\theta_0) = -\frac{\partial^2}{\partial \pi^2} l(\theta_0 | y)$ or curvature of log-likelihood at true value of θ_0

Two-parameter MLE

Multivariate Normal Distribution

Suppose that we have a vector of random variables,

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$$

Then we'll say that $X \sim$ Multivariate Normal$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where,

$$\begin{aligned}
\boldsymbol{\mu} &= (\mu_1, \mu_2, \ldots, \mu_k) \\
\boldsymbol{\Sigma} &= \begin{pmatrix} \sigma_1^2 & Cov(X_1, X_2) & Cov(X_1, X_3) & \ldots & Cov(X_1, X_n) \\ Cov(X_1, X_2) & \sigma_2^2 & Cov(X_2, X_3) & \ldots & Cov(X_2, X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov(X_1, X_k) & Cov(X_2, X_k) & Cov(X_3, X_k) & \ldots & \sigma_k^2 \end{pmatrix}
\end{aligned}$$

Multivariate Normal Distribution

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^{'}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right)$$

Multivariate Version

Suppose that we simple random samples $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ from multivariate distribution $p(\boldsymbol{y}|\boldsymbol{\theta}_0)$.

Multivariate Version

Suppose that we simple random samples $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ from multivariate distribution $p(\boldsymbol{y}|\boldsymbol{\theta}_0)$. Define $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, \theta_k^*)$ as vector valued MLE.

Multivariate Version

Suppose that we simple random samples $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ from multivariate distribution $p(\boldsymbol{y}|\boldsymbol{\theta}_0)$. Define $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, \theta_k^*)$ as vector valued MLE. Then as $n \rightarrow \infty$,

Multivariate Version

Suppose that we simple random samples $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ from multivariate distribution $p(\boldsymbol{y}|\boldsymbol{\theta}_0)$. Define $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, \theta_k^*)$ as vector valued MLE. Then as $n \rightarrow \infty$,

$$p(\boldsymbol{\theta}^*) \quad \rightarrow^d \quad \text{Multivariate Normal}(\boldsymbol{\theta}_0, I_n(\boldsymbol{\theta}_0)^{-1})$$

Multivariate Version

Suppose that we simple random samples $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ from multivariate distribution $p(\boldsymbol{y}|\boldsymbol{\theta}_0)$. Define $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, \theta_k^*)$ as vector valued MLE. Then as $n \rightarrow \infty$,

$$p(\boldsymbol{\theta}^*) \quad \rightarrow^d \quad \text{Multivariate Normal}(\boldsymbol{\theta}_0, I_n(\boldsymbol{\theta}_0)^{-1})$$

Where $I(\boldsymbol{\theta}_0)$ is the observed Fisher Information Matrix, (Negative Hessian) or

Multivariate Version

Suppose that we simple random samples $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ from multivariate distribution $p(\boldsymbol{y}|\boldsymbol{\theta}_0)$. Define $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, \theta_k^*)$ as vector valued MLE. Then as $n \rightarrow \infty$,

$$p(\boldsymbol{\theta}^*) \quad \rightarrow^d \quad \text{Multivariate Normal}(\boldsymbol{\theta}_0, I_n(\boldsymbol{\theta}_0)^{-1})$$

Where $I(\boldsymbol{\theta}_0)$ is the observed Fisher Information Matrix, (Negative Hessian) or

$$I_n(\boldsymbol{\theta}_0) \quad = \quad - \begin{pmatrix} \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \theta_1^2} & \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \theta_2^2} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \theta_k^2} \end{pmatrix}$$

Multivariate Version

Suppose that we simple random samples $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ from multivariate distribution $p(\boldsymbol{y}|\boldsymbol{\theta}_0)$. Define $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, \theta_k^*)$ as vector valued MLE. Then as $n \rightarrow \infty$,

$$p(\boldsymbol{\theta}^*) \quad \rightarrow^d \quad \text{Multivariate Normal}(\boldsymbol{\theta}_0, I_n(\boldsymbol{\theta}_0)^{-1})$$

Where $I(\boldsymbol{\theta}_0)$ is the observed Fisher Information Matrix, (Negative Hessian) or

$$I_n(\boldsymbol{\theta}_0) \quad = \quad - \begin{pmatrix} \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \theta_1^2} & \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \theta_2^2} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta}_0)}{\partial \theta_k^2} \end{pmatrix}$$

Inverting the Fisher-information matrix provides Variance-Covariance Matrix

Maximum Likelihood Estimation, Normal Distribution

Example 2:

Maximum Likelihood Estimation, Normal Distribution

Example 2:
Suppose that we draw an independent and identically distributed random sample of n observations from a normal distribution,

Maximum Likelihood Estimation, Normal Distribution

Example 2:
Suppose that we draw an independent and identically distributed random sample of n observations from a normal distribution,

$$Y_i \quad \sim \quad \text{Normal}(\mu, \sigma^2)$$

Maximum Likelihood Estimation, Normal Distribution

Example 2:
Suppose that we draw an independent and identically distributed random sample of n observations from a normal distribution,

$$\begin{aligned} Y_i &\sim \textsf{Normal}(\mu, \sigma^2) \\ \boldsymbol{y} &= (y_1, y_2, \ldots, y_n) \end{aligned}$$

Maximum Likelihood Estimation, Normal Distribution

Example 2:
Suppose that we draw an independent and identically distributed random sample of n observations from a normal distribution,

$$\begin{aligned} Y_i &\sim \textsf{Normal}(\mu, \sigma^2) \\ \boldsymbol{y} &= (y_1, y_2, \ldots, y_n) \end{aligned}$$

Our task:

Maximum Likelihood Estimation, Normal Distribution

Example 2:
Suppose that we draw an independent and identically distributed random sample of n observations from a normal distribution,

$$\begin{aligned} Y_i &\sim \text{Normal}(\mu, \sigma^2) \\ \boldsymbol{y} &= (y_1, y_2, \ldots, y_n) \end{aligned}$$

Our task:

- Obtain likelihood (summary estimator)

Maximum Likelihood Estimation, Normal Distribution

Example 2:
Suppose that we draw an independent and identically distributed random sample of n observations from a normal distribution,

$$\begin{aligned} Y_i &\sim \text{Normal}(\mu, \sigma^2) \\ \boldsymbol{y} &= (y_1, y_2, \ldots, y_n) \end{aligned}$$

Our task:

- Obtain likelihood (summary estimator)
- Derive maximum likelihood estimators for μ and σ^2

Maximum Likelihood Estimation, Normal Distribution

Example 2:
Suppose that we draw an independent and identically distributed random sample of n observations from a normal distribution,

$$\begin{aligned} Y_i &\sim \text{Normal}(\mu, \sigma^2) \\ \boldsymbol{y} &= (y_1, y_2, \ldots, y_n) \end{aligned}$$

Our task:

- Obtain likelihood (summary estimator)
- Derive maximum likelihood estimators for μ and σ^2
- Characterize sampling distribution

Maximum Likelihood Estimation, Normal Distribution

Maximum Likelihood Estimation, Normal Distribution

$$L(\mu, \sigma^2 | \boldsymbol{y}) = \prod_{i=1}^{n} f(y_i | \mu, \sigma^2)$$

Maximum Likelihood Estimation, Normal Distribution

$$
\begin{aligned}
L(\mu, \sigma^2 | \boldsymbol{y}) &= \prod_{i=1}^{n} f(y_i | \mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{\exp[-\frac{(y_i - \mu)^2}{2\sigma^2}]}{\sqrt{2\pi\sigma^2}}
\end{aligned}
$$

Maximum Likelihood Estimation, Normal Distribution

$$
\begin{aligned}
L(\mu, \sigma^2 | \boldsymbol{y}) &= \prod_{i=1}^{n} f(y_i | \mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{\exp[-\frac{(y_i - \mu)^2}{2\sigma^2}]}{\sqrt{2\pi\sigma^2}} \\
&= \frac{\exp[-\sum_{i=1}^{n} \frac{(y_i-\mu)^2}{2\sigma^2}]}{(2\pi)^{n/2}\sigma^{2n/2}}
\end{aligned}
$$

Maximum Likelihood Estimation, Normal Distribution

$$
\begin{aligned}
L(\mu, \sigma^2 | \boldsymbol{y}) &= \prod_{i=1}^{n} f(y_i | \mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{\exp[-\frac{(y_i - \mu)^2}{2\sigma^2}]}{\sqrt{2\pi\sigma^2}} \\
&= \frac{\exp[-\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2}]}{(2\pi)^{n/2}\sigma^{2n/2}}
\end{aligned}
$$

Taking the logarithm, we have

Maximum Likelihood Estimation, Normal Distribution

$$
\begin{aligned}
L(\mu, \sigma^2|\boldsymbol{y}) &= \prod_{i=1}^{n} f(y_i|\mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{\exp[-\frac{(y_i-\mu)^2}{2\sigma^2}]}{\sqrt{2\pi\sigma^2}} \\
&= \frac{\exp[-\sum_{i=1}^{n} \frac{(y_i-\mu)^2}{2\sigma^2}]}{(2\pi)^{n/2}\sigma^{2n/2}}
\end{aligned}
$$

Taking the logarithm, we have

$$
l(\mu, \sigma^2|\boldsymbol{y}) = -\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2} - \frac{n}{2} log(2\pi) - \frac{n}{2}\log(\sigma^2)
$$

Maximum Likelihood Estimation, Normal Distribution

$$
\begin{aligned}
L(\mu, \sigma^2|\boldsymbol{y}) &= \prod_{i=1}^{n} f(y_i|\mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{\exp[-\frac{(y_i-\mu)^2}{2\sigma^2}]}{\sqrt{2\pi\sigma^2}} \\
&= \frac{\exp[-\sum_{i=1}^{n} \frac{(y_i-\mu)^2}{2\sigma^2}]}{(2\pi)^{n/2}\sigma^{2n/2}}
\end{aligned}
$$

Taking the logarithm, we have

$$
\begin{aligned}
l(\mu, \sigma^2|\boldsymbol{y}) &= -\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2} - \frac{n}{2} log(2\pi) - \frac{n}{2}\log(\sigma^2) \\
&= -\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2} - \frac{n}{2}\log(\sigma^2) + c
\end{aligned}
$$

Maximum Likelihood Estimation, Normal Distribution

Let's find μ^* and $(\sigma^2)^*$ that maximizes log-likelihood.

$$
\begin{aligned}
l(\mu, \sigma^2 | \boldsymbol{y}) &= -\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2} - \frac{n}{2}\log(\sigma^2) + c \\
\frac{\partial l(\mu, \sigma^2)|\boldsymbol{y})}{\partial \mu} &= -\sum_{i=1}^{n} \frac{2(y_i - \mu)}{2\sigma^2} \\
\frac{\partial l(\mu, \sigma^2)|\boldsymbol{y})}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i - \mu)^2
\end{aligned}
$$

Maximum Likelihood Estimation, Normal Distribution

$$
\begin{aligned}
0 &= -\sum_{i=1}^{n} \frac{2(y_i - \mu^*)}{2\sigma^2} \\
0 &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \mu^*)^2
\end{aligned}
$$

Solving for μ and σ^2 yields,

$$
\begin{aligned}
\mu^* &= \frac{\sum_{i=1}^{n} y_i}{n} \\
\sigma^2 &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2
\end{aligned}
$$

Maximum Likelihood Estimation, Normal Distribution

Multivariate analogy: observed Fisher information matrix (Negative Hessian) (Negative matrix of second derivatives)

Maximum Likelihood Estimation, Normal Distribution

Multivariate analogy: observed Fisher information matrix (Negative Hessian) (Negative matrix of second derivatives)

$$I_n(\mu^*, \widehat{\sigma}^2) = -\begin{pmatrix} \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial \mu^2} & \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial \sigma^2 \partial \mu} \\ \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial^2 \sigma^2} \end{pmatrix}$$

Maximum Likelihood Estimation, Normal Distribution

Multivariate analogy: observed Fisher information matrix (Negative Hessian) (Negative matrix of second derivatives)

$$I_n(\mu^*, \widehat{\sigma}^2) = - \begin{pmatrix} \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial \mu^2} & \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial \sigma^2 \partial \mu} \\ \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial^2 \sigma^2} \end{pmatrix}$$

Taking derivatives and evaluating at MLE's yields

Maximum Likelihood Estimation, Normal Distribution

Multivariate analogy: observed Fisher information matrix (Negative Hessian) (Negative matrix of second derivatives)

$$I_n(\mu^*, \widehat{\sigma}^2) = - \begin{pmatrix} \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial \mu^2} & \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial \sigma^2 \partial \mu} \\ \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial^2 \sigma^2} \end{pmatrix}$$

Taking derivatives and evaluating at MLE's yields

$$I_n(\mu^*, \widehat{\sigma}^2) = \begin{pmatrix} \frac{n}{\widehat{\sigma}^2} & 0 \\ 0 & \frac{n}{(\widehat{\sigma}^2)^2} \end{pmatrix}$$

Maximum Likelihood Estimation, Normal Distribution

Multivariate analogy: observed Fisher information matrix (Negative Hessian) (Negative matrix of second derivatives)

$$I_n(\mu^*, \widehat{\sigma}^2) = -\begin{pmatrix} \frac{\partial^2 l(\mu, \sigma^2|\boldsymbol{y})}{\partial \mu^2} & \frac{\partial^2 l(\mu, \sigma^2|\boldsymbol{y})}{\partial \sigma^2 \partial \mu} \\ \frac{\partial^2 l(\mu, \sigma^2|\boldsymbol{y})}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 l(\mu, \sigma^2|\boldsymbol{y})}{\partial^2 \sigma^2} \end{pmatrix}$$

Taking derivatives and evaluating at MLE's yields

$$I_n(\mu^*, \widehat{\sigma}^2) = \begin{pmatrix} \frac{n}{\widehat{\sigma}^2} & 0 \\ 0 & \frac{n}{(\widehat{\sigma}^2)^2} \end{pmatrix}$$

Therefore, as $n \rightarrow \infty$, we have that

Maximum Likelihood Estimation, Normal Distribution

Multivariate analogy: observed Fisher information matrix (Negative Hessian) (Negative matrix of second derivatives)

$$I_n(\mu^*, \widehat{\sigma}^2) = -\begin{pmatrix} \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial \mu^2} & \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial \sigma^2 \partial \mu} \\ \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial^2 \sigma^2} \end{pmatrix}$$

Taking derivatives and evaluating at MLE's yields

$$I_n(\mu^*, \widehat{\sigma}^2) \quad = \quad \begin{pmatrix} \frac{n}{\widehat{\sigma}^2} & 0 \\ 0 & \frac{n}{(\widehat{\sigma}^2)^2} \end{pmatrix}$$

Therefore, as $n \rightarrow \infty$, we have that

$$p(\mu, \sigma^2) \quad \rightarrow^d \quad \text{Multivariate Normal}\left(\left(\overline{y}, \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 \right), \begin{pmatrix} \frac{\widehat{\sigma}^2}{n} & 0 \\ 0 & \frac{(\widehat{\sigma}^2)^2}{n} \end{pmatrix} \right)$$

Maximum Likelihood Estimation, Normal Distribution

Multivariate analogy: observed Fisher information matrix (Negative Hessian) (Negative matrix of second derivatives)

$$I_n(\mu^*, \widehat{\sigma}^2) = -\begin{pmatrix} \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial \mu^2} & \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial \sigma^2 \partial \mu} \\ \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial^2 \sigma^2} \end{pmatrix}$$

Taking derivatives and evaluating at MLE's yields

$$I_n(\mu^*, \widehat{\sigma}^2) = \begin{pmatrix} \frac{n}{\widehat{\sigma}^2} & 0 \\ 0 & \frac{n}{(\widehat{\sigma}^2)^2} \end{pmatrix}$$

Therefore, as $n \rightarrow \infty$, we have that

$$p(\mu, \sigma^2) \rightarrow^d \text{ Multivariate Normal} \left(\left(\overline{y}, \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 \right), \begin{pmatrix} \frac{\widehat{\sigma}^2}{n} & 0 \\ 0 & \frac{(\widehat{\sigma}^2)^2}{n} \end{pmatrix} \right)$$

Because normal distribution $\Rightarrow$ that mle of μ and σ^2 are independent!

Maximum Likelihood Estimation, Normal Distribution

Multivariate analogy: observed Fisher information matrix (Negative Hessian) (Negative matrix of second derivatives)

$$I_n(\mu^*, \widehat{\sigma}^2) = -\begin{pmatrix} \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial \mu^2} & \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial \sigma^2 \partial \mu} \\ \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 l(\mu,\sigma^2|\boldsymbol{y})}{\partial^2 \sigma^2} \end{pmatrix}$$

Taking derivatives and evaluating at MLE's yields

$$I_n(\mu^*, \widehat{\sigma}^2) = \begin{pmatrix} \frac{n}{\widehat{\sigma}^2} & 0 \\ 0 & \frac{n}{(\widehat{\sigma}^2)^2} \end{pmatrix}$$

Therefore, as $n \rightarrow \infty$, we have that

$$p(\mu, \sigma^2) \rightarrow^d \text{Multivariate Normal}\left(\left(\overline{y}, \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2\right), \begin{pmatrix} \frac{\widehat{\sigma}^2}{n} & 0 \\ 0 & \frac{(\widehat{\sigma}^2)^2}{n} \end{pmatrix}\right)$$

Because normal distribution $\Rightarrow$ that mle of μ and σ^2 are independent!
This is an asymptotic result: results will vary with small sample sizes

Up next:

1) Linear regression in maximum likelihood
2) Logit/Probit
3) Numerical optimization