Soc504: Inference

Brandon Stewart[1]

Princeton

February 13, 2017

[1] Much of the material in section 2-7 is edited from Gary King's slides for Gov2000 at Harvard. Section 8 is heavily influenced by Patrick Lam.

Where We've Been and Where We're Going...

- Last Week
  - Intro and Class Overview
- This Week+
  - Theories of inference
  - Likelihood Estimation
  - Simulation
- Next Week+
  - Generalized Linear Models
- Long Run
  - likelihood $\rightarrow$ GLMs $\rightarrow$ advanced methods

Followup

Followup

- Questions?

Followup

- Questions?
- Replication Stories?

Followup

- Questions?
- Replication Stories?
- How is Perusall working for everyone?

Followup

- Questions?
- Replication Stories?
- How is Perusall working for everyone?
- Problem Set Plan

Followup

- Questions?
- Replication Stories?
- How is Perusall working for everyone?
- Problem Set Plan
- Note that today we will be discussing things that border on philosophy. Thus it is particularly important that you ask questions. Often the simplest questions are the most profound!

1. History
2. Likelihood Inference
3. Bayesian Inference
4. Neyman-Pearson
5. Likelihood Example
6. Properties and Tests
7. Simulation
8. Fun With Bayes

1 History

2 Likelihood Inference

3 Bayesian Inference

4 Neyman-Pearson

5 Likelihood Example

6 Properties and Tests

7 Simulation

8 Fun With Bayes

Historical Backdrop

Historical Backdrop

- We start with the fundamental question of how to learn from experience

Historical Backdrop

- We start with the fundamental question of how to learn from experience
- This isn't a question with easy answers despite the fact that humans do it all the time.

Historical Backdrop

- We start with the fundamental question of how to learn from experience
- This isn't a question with easy answers despite the fact that humans do it all the time.
- We've been using algorithms of various sorts for a long time, least squares dates back to Legendre and Gauss 1795-1805.

Historical Backdrop

- We start with the fundamental question of how to learn from experience
- This isn't a question with easy answers despite the fact that humans do it all the time.
- We've been using algorithms of various sorts for a long time, least squares dates back to Legendre and Gauss 1795-1805.
- While an algorithm tells us what to compute and provides a summary of the data, inference answers the question of why we are doing something (i.e. what properties it has).

Historical Backdrop

- We start with the fundamental question of how to learn from experience
- This isn't a question with easy answers despite the fact that humans do it all the time.
- We've been using algorithms of various sorts for a long time, least squares dates back to Legendre and Gauss 1795-1805.
- While an algorithm tells us what to compute and provides a summary of the data, inference answers the question of why we are doing something (i.e. what properties it has).
- For our purposes the central question of inference will be, how do we assess the accuracy of an estimate?

Historical Backdrop: Frequentist Thought

- Starting around 1900, a group of statisticians including Fisher, Hotelling, Neyman and Pearson provide an answer to the question of how we think about estimator accuracy: frequentism

Historical Backdrop: Frequentist Thought

- Starting around 1900, a group of statisticians including Fisher, Hotelling, Neyman and Pearson provide an answer to the question of how we think about estimator accuracy: frequentism
- At the times you had very few data points which were typically collected in laborious experiments. Thus we want a maximally efficient analysis method.

Historical Backdrop: Frequentist Thought

- Starting around 1900, a group of statisticians including Fisher, Hotelling, Neyman and Pearson provide an answer to the question of how we think about estimator accuracy: frequentism
- At the times you had very few data points which were typically collected in laborious experiments. Thus we want a maximally efficient analysis method.
- Frequentism is based on a clever intellectual pivot: we treat the probabilistic accuracy of the *estimator* as the accuracy of the *estimate*.

Historical Backdrop: Frequentist Thought

- Starting around 1900, a group of statisticians including Fisher, Hotelling, Neyman and Pearson provide an answer to the question of how we think about estimator accuracy: frequentism
- At the times you had very few data points which were typically collected in laborious experiments. Thus we want a maximally efficient analysis method.
- Frequentism is based on a clever intellectual pivot: we treat the probabilistic accuracy of the *estimator* as the accuracy of the *estimate*.
- Thus we attribute to a single number, the probabilistic properties of the estimator. (maybe it should have been called behaviorism!)

Historical Backdrop: Frequentist Thought

- Starting around 1900, a group of statisticians including Fisher, Hotelling, Neyman and Pearson provide an answer to the question of how we think about estimator accuracy: frequentism
- At the times you had very few data points which were typically collected in laborious experiments. Thus we want a maximally efficient analysis method.
- Frequentism is based on a clever intellectual pivot: we treat the probabilistic accuracy of the *estimator* as the accuracy of the *estimate*.
- Thus we attribute to a single number, the probabilistic properties of the estimator. (maybe it should have been called behaviorism!)
- We often talk about this as frequentists posing the question: 'what would happen if we reran the same situation over and over again?'

Historical Backdrop: Frequentist Thought

- Starting around 1900, a group of statisticians including Fisher, Hotelling, Neyman and Pearson provide an answer to the question of how we think about estimator accuracy: frequentism
- At the times you had very few data points which were typically collected in laborious experiments. Thus we want a maximally efficient analysis method.
- Frequentism is based on a clever intellectual pivot: we treat the probabilistic accuracy of the *estimator* as the accuracy of the *estimate*.
- Thus we attribute to a single number, the probabilistic properties of the estimator. (maybe it should have been called behaviorism!)
- We often talk about this as frequentists posing the question: 'what would happen if we reran the same situation over and over again?'
- Why is this hard? Well we need to calculate properties of an estimator obtained from a true distribution F even though F is unknown.

Historical Backdrop: Bayesian Approaches

- In Soc500 we implicitly worked in the frequentist domain: we talked about bias and variance and considered repeated trials.

Historical Backdrop: Bayesian Approaches

- In Soc500 we implicitly worked in the frequentist domain: we talked about bias and variance and considered repeated trials.
- An alternative view is Bayesian where we treat the data as fixed and the parameter as varying.

Historical Backdrop: Bayesian Approaches

- In Soc500 we implicitly worked in the frequentist domain: we talked about bias and variance and considered repeated trials.
- An alternative view is Bayesian where we treat the data as fixed and the parameter as varying.
- Efron and Hastie (2016) describe frequentism and Bayesianism as orthogonal because they both start with a family of probability distributions but then proceed to reason over different dimensions

Historical Backdrop: Bayesian Approaches

- In Soc500 we implicitly worked in the frequentist domain: we talked about bias and variance and considered repeated trials.
- An alternative view is Bayesian where we treat the data as fixed and the parameter as varying.
- Efron and Hastie (2016) describe frequentism and Bayesianism as orthogonal because they both start with a family of probability distributions but then proceed to reason over different dimensions

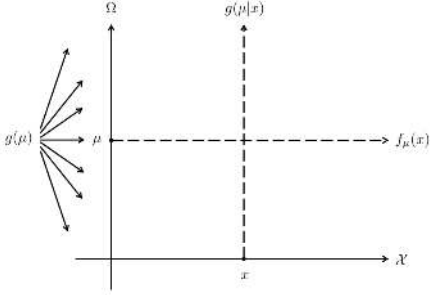


Figure 3.5 Bayesian inference proceeds vertically, given x; frequentist inference proceeds horizontally, given μ.

Likelihood

- We will spend the majority of our time on Fisher's Maximum Likelihood Theory.

Likelihood

- We will spend the majority of our time on Fisher's Maximum Likelihood Theory.
- ML dominated the twentieth century for a few reasons:

Likelihood

- We will spend the majority of our time on Fisher's Maximum Likelihood Theory.
- ML dominated the twentieth century for a few reasons:
  - It easily generates estimators: one theory provides us an estimator for almost every situation which is generally not true of other frequentist approaches.

Likelihood

- We will spend the majority of our time on Fisher's Maximum Likelihood Theory.
- ML dominated the twentieth century for a few reasons:
  - It easily generates estimators: one theory provides us an estimator for almost every situation which is generally not true of other frequentist approaches.
  - these approaches have excellent frequentist properties: they tend to be nearly unbiased and be reasonably efficient.

Likelihood

- We will spend the majority of our time on Fisher's Maximum Likelihood Theory.
- ML dominated the twentieth century for a few reasons:
  - It easily generates estimators: one theory provides us an estimator for almost every situation which is generally not true of other frequentist approaches.
  - these approaches have excellent frequentist properties: they tend to be nearly unbiased and be reasonably efficient.
  - the estimators have a bayesian interpretation.

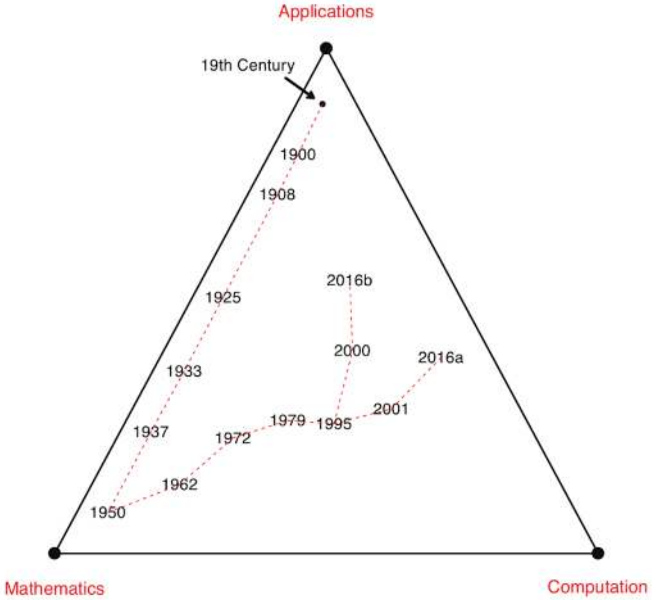
Likelihood

- We will spend the majority of our time on Fisher's Maximum Likelihood Theory.
- ML dominated the twentieth century for a few reasons:
  - It easily generates estimators: one theory provides us an estimator for almost every situation which is generally not true of other frequentist approaches.
  - these approaches have excellent frequentist properties: they tend to be nearly unbiased and be reasonably efficient.
  - the estimators have a bayesian interpretation.
- We will also see that likelihood lends itself nicely to situations where we care a lot about the outcome rather than the coefficients themselves.

Likelihood

- We will spend the majority of our time on Fisher's Maximum Likelihood Theory.
- ML dominated the twentieth century for a few reasons:
  - It easily generates estimators: one theory provides us an estimator for almost every situation which is generally not true of other frequentist approaches.
  - these approaches have excellent frequentist properties: they tend to be nearly unbiased and be reasonably efficient.
  - the estimators have a bayesian interpretation.
- We will also see that likelihood lends itself nicely to situations where we care a lot about the outcome rather than the coefficients themselves.
- For those interested Stigler's "The epic story of maximum likelihood" is a fantastic account of the history of the idea.

A Perspective on a Historical Arc (Efron and Hastie 2016)



The Problem of Inference

The Problem of Inference

1. Probability:

$$\mathbb{P}(y|M) = \mathbb{P}(\text{known}|\text{unknown})$$

The Problem of Inference

1. Probability:

$$\mathbb{P}(y|M) = \mathbb{P}(\text{known}|\text{unknown})$$

2. The goal of inverse probability:

$$\mathbb{P}(M|y) = \mathbb{P}(\text{unknown}|\text{known})$$

The Problem of Inference

1. Probability:

$$\mathbb{P}(y|M) = \mathbb{P}(\text{known}|\text{unknown})$$

2. The goal of inverse probability:

$$\mathbb{P}(M|y) = \mathbb{P}(\text{unknown}|\text{known})$$

3. A more reasonable, limited goal. Let $M = \{M^*, \theta\}$, where M^* is assumed & θ is to be estimated:

$$\mathbb{P}(\theta|y, M^*) \equiv \mathbb{P}(\theta|y)$$

The Problem of Inference

4. Bayes Theorem:

The Problem of Inference

4. Bayes Theorem:

$$\mathbb{P}(\theta|y) = \frac{\mathbb{P}(\theta, y)}{\mathbb{P}(y)} \qquad \text{[Defn. of conditional probability]}$$

The Problem of Inference

4. Bayes Theorem:

$$\mathbb{P}(\theta|y) = \frac{\mathbb{P}(\theta, y)}{\mathbb{P}(y)} \qquad \text{[Defn. of conditional probability]}$$

$$= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\mathbb{P}(y)} \qquad [\mathbb{P}(AB) = \mathbb{P}(B)\mathbb{P}(A|B)]$$

The Problem of Inference

4. Bayes Theorem:

$$\mathbb{P}(\theta|y) = \frac{\mathbb{P}(\theta, y)}{\mathbb{P}(y)} \qquad \text{[Defn. of conditional probability]}$$

$$= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\mathbb{P}(y)} \qquad [\mathbb{P}(AB) = \mathbb{P}(B)\mathbb{P}(A|B)]$$

$$= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} \qquad [\mathbb{P}(A) = \int \mathbb{P}(AB)dB]$$

The Problem of Inference

4. Bayes Theorem:

$$\mathbb{P}(\theta|y) = \frac{\mathbb{P}(\theta, y)}{\mathbb{P}(y)} \qquad \text{[Defn. of conditional probability]}$$

$$= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\mathbb{P}(y)} \qquad [\mathbb{P}(AB) = \mathbb{P}(B)\mathbb{P}(A|B)]$$

$$= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} \qquad [\mathbb{P}(A) = \int \mathbb{P}(AB)dB]$$

5. If we knew the right side, we could compute the inverse probability.

The Problem of Inference

4. Bayes Theorem:

$$\mathbb{P}(\theta|y) = \frac{\mathbb{P}(\theta, y)}{\mathbb{P}(y)} \qquad \text{[Defn. of conditional probability]}$$

$$= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\mathbb{P}(y)} \qquad [\mathbb{P}(AB) = \mathbb{P}(B)\mathbb{P}(A|B)]$$

$$= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} \qquad [\mathbb{P}(A) = \int \mathbb{P}(AB)dB]$$

5. If we knew the right side, we could compute the inverse probability.
6. We will discuss two alternative interpretations of this theorem. Likelihood and Bayesian

The Problem of Inference

4. Bayes Theorem:

$$\mathbb{P}(\theta|y) = \frac{\mathbb{P}(\theta, y)}{\mathbb{P}(y)} \qquad \text{[Defn. of conditional probability]}$$

$$= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\mathbb{P}(y)} \qquad [\mathbb{P}(AB) = \mathbb{P}(B)\mathbb{P}(A|B)]$$

$$= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} \qquad [\mathbb{P}(A) = \int \mathbb{P}(AB)dB]$$

5. If we knew the right side, we could compute the inverse probability.
6. We will discuss two alternative interpretations of this theorem. Likelihood and Bayesian
7. In both, $\mathbb{P}(y|\theta)$ is a traditional probability density

The Problem of Inference

4. Bayes Theorem:

$$\mathbb{P}(\theta|y) = \frac{\mathbb{P}(\theta, y)}{\mathbb{P}(y)} \qquad \text{[Defn. of conditional probability]}$$

$$= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\mathbb{P}(y)} \qquad [\mathbb{P}(AB) = \mathbb{P}(B)\mathbb{P}(A|B)]$$

$$= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} \qquad [\mathbb{P}(A) = \int \mathbb{P}(AB)dB]$$

5. If we knew the right side, we could compute the inverse probability.
6. We will discuss two alternative interpretations of this theorem. Likelihood and Bayesian
7. In both, $\mathbb{P}(y|\theta)$ is a traditional probability density
8. The two differ on what is fixed and what is random

1. History
2. Likelihood Inference
3. Bayesian Inference
4. Neyman-Pearson
5. Likelihood Example
6. Properties and Tests
7. Simulation
8. Fun With Bayes

1 History

2 Likelihood Inference

3 Bayesian Inference

4 Neyman-Pearson

5 Likelihood Example

6 Properties and Tests

7 Simulation

8 Fun With Bayes

Interpretation 1: The Likelihood Theory of Inference

Interpretation 1: The Likelihood Theory of Inference

1. R.A. Fisher's idea

Interpretation 1: The Likelihood Theory of Inference

1. R.A. Fisher's idea
2. θ is fixed and y is random

Interpretation 1: The Likelihood Theory of Inference

1. R.A. Fisher's idea
2. θ is fixed and y is random
3. Let:

$$k(y) \equiv \frac{\mathbb{P}(\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} \implies P(\theta|y) = \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} = k(y)\mathbb{P}(y|\theta)$$

Interpretation 1: The Likelihood Theory of Inference

1. R.A. Fisher's idea
2. θ is fixed and y is random
3. Let:

$$k(y) \equiv \frac{\mathbb{P}(\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} \implies P(\theta|y) = \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} = k(y)\mathbb{P}(y|\theta)$$

4. Define $K(y)$ as an unknown function of y with θ fixed at its true value

Interpretation 1: The Likelihood Theory of Inference

1. R.A. Fisher's idea
2. θ is fixed and y is random
3. Let:

$$k(y) \equiv \frac{\mathbb{P}(\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} \implies P(\theta|y) = \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} = k(y)\mathbb{P}(y|\theta)$$

4. <u>Define</u> $K(y)$ as an <u>unknown</u> function of y with θ fixed at its true value
5. $\leadsto$ the likelihood theory of inference has four axioms: the 3 probability axioms plus the likelihood axiom:

Interpretation 1: The Likelihood Theory of Inference

1. R.A. Fisher's idea
2. θ is fixed and y is random
3. Let:

$$k(y) \equiv \frac{\mathbb{P}(\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} \implies P(\theta|y) = \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} = k(y)\mathbb{P}(y|\theta)$$

4. <u>Define</u> $K(y)$ as an <u>unknown</u> function of y with θ fixed at its true value
5. $\leadsto$ the likelihood theory of inference has four axioms: the 3 probability axioms plus the likelihood axiom:

$$L(\theta|y) \equiv k(y)\mathbb{P}(y|\theta)$$

Interpretation 1: The Likelihood Theory of Inference

1. R.A. Fisher's idea
2. θ is fixed and y is random
3. Let:

$$k(y) \equiv \frac{\mathbb{P}(\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} \implies P(\theta|y) = \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} = k(y)\mathbb{P}(y|\theta)$$

4. Define $K(y)$ as an unknown function of y with θ fixed at its true value
5. $\leadsto$ the likelihood theory of inference has four axioms: the 3 probability axioms plus the likelihood axiom:

$$\begin{aligned} L(\theta|y) &\equiv k(y)\mathbb{P}(y|\theta) \\ &\propto \mathbb{P}(y|\theta) \end{aligned}$$

Interpretation 1: The Likelihood Theory of Inference

1. R.A. Fisher's idea
2. θ is fixed and y is random
3. Let:

$$k(y) \equiv \frac{\mathbb{P}(\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} \implies P(\theta|y) = \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} = k(y)\mathbb{P}(y|\theta)$$

4. Define $K(y)$ as an unknown function of y with θ fixed at its true value
5. $\rightsquigarrow$ the likelihood theory of inference has four axioms: the 3 probability axioms plus the likelihood axiom:

$$\begin{aligned} L(\theta|y) &\equiv k(y)\mathbb{P}(y|\theta) \\ &\propto \mathbb{P}(y|\theta) \end{aligned}$$

6. $L(\theta|y)$ is a function: for y fixed at the observed values, it gives the "likelihood" of any value of θ.

Interpretation 1: The Likelihood Theory of Inference

7. Typically we assume independence in the observations to get $L(\theta|y) \propto \prod_{i=1}^{N} \mathbb{P}(y_i|\theta)$

Interpretation 1: The Likelihood Theory of Inference

7. Typically we assume independence in the observations to get $L(\theta|y) \propto \prod_{i=1}^{N} \mathbb{P}(y_i|\theta)$
8. Likelihood: a relative measure of uncertainty, changing with the data

Interpretation 1: The Likelihood Theory of Inference

7. Typically we assume independence in the observations to get $L(\theta|y) \propto \prod_{i=1}^{N} \mathbb{P}(y_i|\theta)$
8. Likelihood: a relative measure of uncertainty, changing with the data
9. Comparing the value of $L(\theta|y)$ for different θ values in one data set y is meaningful.

Interpretation 1: The Likelihood Theory of Inference

7. Typically we assume independence in the observations to get $L(\theta|y) \propto \prod_{i=1}^{N} \mathbb{P}(y_i|\theta)$
8. Likelihood: a relative measure of uncertainty, changing with the data
9. Comparing the value of $L(\theta|y)$ for different θ values in one data set y is meaningful.
10. Comparing values of $L(\theta|y)$ across data sets is meaningless. (just as you can't compare R^2 values across equations with different dependent variables.)

Interpretation 1: The Likelihood Theory of Inference

7. Typically we assume independence in the observations to get $L(\theta|y) \propto \prod_{i=1}^{N} \mathbb{P}(y_i|\theta)$
8. Likelihood: a relative measure of uncertainty, changing with the data
9. Comparing the value of $L(\theta|y)$ for different θ values in one data set y is meaningful.
10. Comparing values of $L(\theta|y)$ across data sets is meaningless. (just as you can't compare R^2 values across equations with different dependent variables.)
11. The likelihood principle: the data only affect inferences through the likelihood function

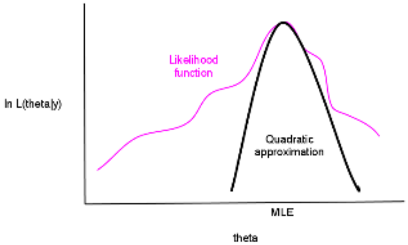
Visualizing the Likelihood

Visualizing the Likelihood

- For algebraic simplicity and numerical stability, we use the log-likelihood (the shape changes, but the max is in the same place)

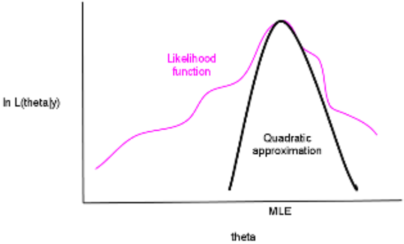
Visualizing the Likelihood

- For algebraic simplicity and numerical stability, we use the log-likelihood (the shape changes, but the max is in the same place)
- If θ has one element, we can plot:



Visualizing the Likelihood

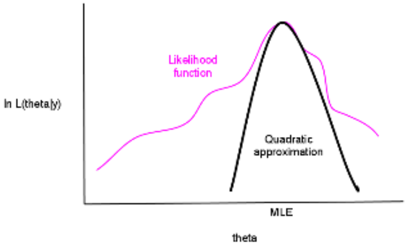
- For algebraic simplicity and numerical stability, we use the log-likelihood (the shape changes, but the max is in the same place)
- If θ has one element, we can plot:



- The full likelihood curve is a Summary Estimator. The likelihood principle means that once this is plotted, we can discard the data (if the model is correct!).

Visualizing the Likelihood

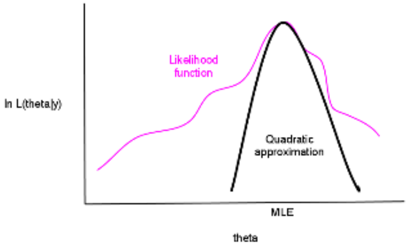
- For algebraic simplicity and numerical stability, we use the log-likelihood (the shape changes, but the max is in the same place)
- If θ has one element, we can plot:



- The full likelihood curve is a Summary Estimator. The likelihood principle means that once this is plotted, we can discard the data (if the model is correct!).
- A one-point summary at the maximum is the MLE

Visualizing the Likelihood

- For algebraic simplicity and numerical stability, we use the log-likelihood (the shape changes, but the max is in the same place)
- If θ has one element, we can plot:



- The full likelihood curve is a Summary Estimator. The likelihood principle means that once this is plotted, we can discard the data (if the model is correct!).
- A one-point summary at the maximum is the MLE
- Uncertainty of point estimate: curvature at the maximum

Logarithm review!

- Logs turn exponentiation into multiplication and multiplication into summation.

Logarithm review!

- Logs turn exponentiation into multiplication and multiplication into summation.
  - $\log(A \times B) = \log(A) + \log(B)$

Logarithm review!

- Logs turn exponentiation into multiplication and multiplication into summation.
  - $\log(A \times B) = \log(A) + \log(B)$
  - $\log(A/B) = \log(A) - \log(B)$

Logarithm review!

- Logs turn exponentiation into multiplication and multiplication into summation.
  - $\log(A \times B) = \log(A) + \log(B)$
  - $\log(A/B) = \log(A) - \log(B)$
  - $\log(A^b) = b \times \log(A)$

Logarithm review!

- Logs turn exponentiation into multiplication and multiplication into summation.
  - $\log(A \times B) = \log(A) + \log(B)$
  - $\log(A/B) = \log(A) - \log(B)$
  - $\log(A^b) = b \times \log(A)$
  - $\log(e) = \ln(e) = 1$

Logarithm review!

- Logs turn exponentiation into multiplication and multiplication into summation.
  - $\log(A \times B) = \log(A) + \log(B)$
  - $\log(A/B) = \log(A) - \log(B)$
  - $\log(A^b) = b \times \log(A)$
  - $\log(e) = \ln(e) = 1$
  - $\log(1) = 0$

Logarithm review!

- Logs turn exponentiation into multiplication and multiplication into summation.
  - $\log(A \times B) = \log(A) + \log(B)$
  - $\log(A/B) = \log(A) - \log(B)$
  - $\log(A^b) = b \times \log(A)$
  - $\log(e) = \ln(e) = 1$
  - $\log(1) = 0$
- Notational note: log in math is almost always used as short-hand for the natural log (ln) as opposed to the base-10 log.

Example 1: Bernoulli Trials

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

Example 1: 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 1: 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 1: 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 1: 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

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

- $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)$

Example 1: 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 1: Bernoulli Trials

Example 1: Bernoulli Trials

$$L(\pi|\boldsymbol{Y}) \quad \propto \quad f(\boldsymbol{Y}|\pi)$$

Example 1: Bernoulli Trials

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

Example 1: Bernoulli Trials

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

Example 1: Bernoulli Trials

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

Example 1: Bernoulli Trials

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

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

Example 1: Bernoulli Trials

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

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

$$\log L(\pi|\boldsymbol{Y}) \equiv \ell(\pi|\boldsymbol{Y}) = \sum_{i=1}^{n} Y_i \log \pi + (n - \sum_{i=1}^{n} Y_i) \log(1-\pi) + c$$

Example 1: Bernoulli Trials

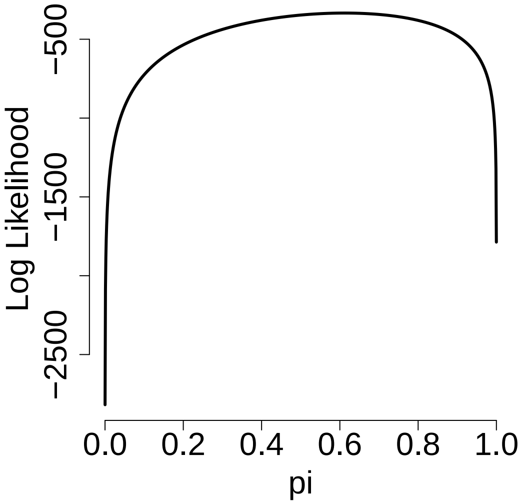
$$
\begin{aligned}
L(\pi|\boldsymbol{Y}) &\propto f(\boldsymbol{Y}|\pi) \\
&= \prod_{i=1}^{n} f(Y_i|\pi) \\
&= \prod_{i=1}^{n} \pi^{Y_i}(1-\pi)^{1-Y_i} \\
&= \pi^{\sum_{i=1}^{n} Y_i}(1-\pi)^{n-\sum_{i=1}^{n} Y_i}
\end{aligned}
$$

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

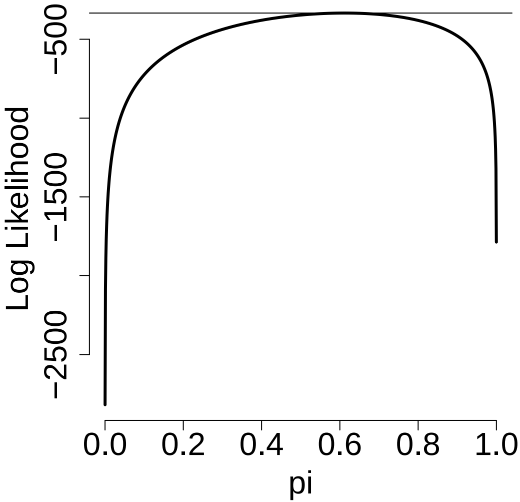
$$
\log L(\pi|\boldsymbol{Y}) \equiv \ell(\pi|\boldsymbol{Y}) = \sum_{i=1}^{n} Y_i \log \pi + (n - \sum_{i=1}^{n} Y_i) \log(1-\pi) + c
$$

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

Example 1: Bernoulli Trials: Simulated Example



Example 1: Bernoulli Trials: Simulated Example



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

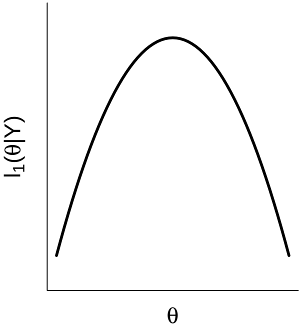
$\pi^* = \overline{Y}$ maximizes $L(\pi|\boldsymbol{Y})$. How much uncertainty is there about this maximum?

Q. Which log-likelihood function contains more information?:

Uncertainty About Mode

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

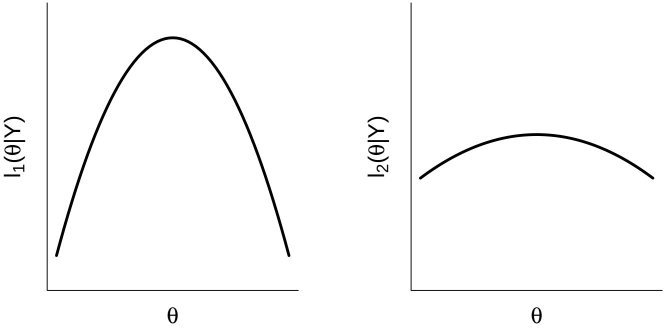
Q. Which log-likelihood function contains more information?:



Uncertainty About Mode

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

Q. Which log-likelihood function contains more information?:



Second derivative captures this curvature

1. History
2. Likelihood Inference
3. Bayesian Inference
4. Neyman-Pearson
5. Likelihood Example
6. Properties and Tests
7. Simulation
8. Fun With Bayes

1. History
2. Likelihood Inference
3. Bayesian Inference
4. Neyman-Pearson
5. Likelihood Example
6. Properties and Tests
7. Simulation
8. Fun With Bayes

Interpretation 2: The Bayesian Theory of Inference

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' idea published after his death by Richard Price as part of a proof of the existence of God

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' idea published after his death by Richard Price as part of a proof of the existence of God
- Recall:

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' idea published after his death by Richard Price as part of a proof of the existence of God
- Recall:

$$\mathbb{P}(\theta|y) = \frac{\mathbb{P}(\theta, y)}{\mathbb{P}(y)} \qquad \text{[Defn. of conditional probability]}$$

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' idea published after his death by Richard Price as part of a proof of the existence of God
- Recall:

$$\mathbb{P}(\theta|y) = \frac{\mathbb{P}(\theta, y)}{\mathbb{P}(y)} \qquad \text{[Defn. of conditional probability]}$$

$$= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\mathbb{P}(y)} \qquad [\mathbb{P}(AB) = \mathbb{P}(B)\mathbb{P}(A|B)]$$

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' idea published after his death by Richard Price as part of a proof of the existence of God
- Recall:

$$
\begin{aligned}
\mathbb{P}(\theta|y) &= \frac{\mathbb{P}(\theta, y)}{\mathbb{P}(y)} && \text{[Defn. of conditional probability]} \\
&= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\mathbb{P}(y)} && [\mathbb{P}(AB) = \mathbb{P}(B)\mathbb{P}(A|B)] \\
&= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} && [\mathbb{P}(A) = \int \mathbb{P}(AB)dB]
\end{aligned}
$$

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' idea published after his death by Richard Price as part of a proof of the existence of God
- Recall:

$$\begin{aligned}
\mathbb{P}(\theta|y) &= \frac{\mathbb{P}(\theta, y)}{\mathbb{P}(y)} && \text{[Defn. of conditional probability]} \\
&= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\mathbb{P}(y)} && [\mathbb{P}(AB) = \mathbb{P}(B)\mathbb{P}(A|B)] \\
&= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} && [\mathbb{P}(A) = \int \mathbb{P}(AB)dB] \\
&\propto \mathbb{P}(\theta)\mathbb{P}(y|\theta)
\end{aligned}$$

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' idea published after his death by Richard Price as part of a proof of the existence of God
- Recall:

$$\begin{aligned}
\mathbb{P}(\theta|y) &= \frac{\mathbb{P}(\theta, y)}{\mathbb{P}(y)} && \text{[Defn. of conditional probability]} \\
&= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\mathbb{P}(y)} && [\mathbb{P}(AB) = \mathbb{P}(B)\mathbb{P}(A|B)] \\
&= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} && [\mathbb{P}(A) = \int \mathbb{P}(AB)dB] \\
&\propto \mathbb{P}(\theta)\mathbb{P}(y|\theta)
\end{aligned}$$

- $\mathbb{P}(\theta|y)$ the posterior density

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' idea published after his death by Richard Price as part of a proof of the existence of God
- Recall:

$$
\begin{aligned}
\mathbb{P}(\theta|y) &= \frac{\mathbb{P}(\theta, y)}{\mathbb{P}(y)} && \text{[Defn. of conditional probability]} \\
&= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\mathbb{P}(y)} && [\mathbb{P}(AB) = \mathbb{P}(B)\mathbb{P}(A|B)] \\
&= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} && [\mathbb{P}(A) = \int \mathbb{P}(AB)dB] \\
&\propto \mathbb{P}(\theta)\mathbb{P}(y|\theta)
\end{aligned}
$$

- $\mathbb{P}(\theta|y)$ the posterior density
- $\mathbb{P}(y|\theta)$ the traditional probability ($\propto$ likelihood)

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' idea published after his death by Richard Price as part of a proof of the existence of God
- Recall:

$$
\begin{aligned}
\mathbb{P}(\theta|y) &= \frac{\mathbb{P}(\theta, y)}{\mathbb{P}(y)} && \text{[Defn. of conditional probability]} \\
&= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\mathbb{P}(y)} && [\mathbb{P}(AB) = \mathbb{P}(B)\mathbb{P}(A|B)] \\
&= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} && [\mathbb{P}(A) = \int \mathbb{P}(AB)dB] \\
&\propto \mathbb{P}(\theta)\mathbb{P}(y|\theta)
\end{aligned}
$$

- $\mathbb{P}(\theta|y)$ the posterior density
- $\mathbb{P}(y|\theta)$ the traditional probability ($\propto$ likelihood)
- $\mathbb{P}(y)$ a constant, computable

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' idea published after his death by Richard Price as part of a proof of the existence of God
- Recall:

$$\begin{aligned}
\mathbb{P}(\theta|y) &= \frac{\mathbb{P}(\theta, y)}{\mathbb{P}(y)} && \text{[Defn. of conditional probability]} \\
&= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\mathbb{P}(y)} && [\mathbb{P}(AB) = \mathbb{P}(B)\mathbb{P}(A|B)] \\
&= \frac{\mathbb{P}(\theta)\mathbb{P}(y|\theta)}{\int \mathbb{P}(\theta)\mathbb{P}(y|\theta)d\theta} && [\mathbb{P}(A) = \int \mathbb{P}(AB)dB] \\
&\propto \mathbb{P}(\theta)\mathbb{P}(y|\theta)
\end{aligned}$$

- $\mathbb{P}(\theta|y)$ the posterior density
- $\mathbb{P}(y|\theta)$ the traditional probability ($\propto$ likelihood)
- $\mathbb{P}(y)$ a constant, computable
- $\mathbb{P}(\theta)$, the prior density — the way Bayes differs from likelihood

What is the prior density, $P(\theta)$?

What is the prior density, $P(\theta)$?

1. A probability density that represents all prior evidence about θ.

What is the prior density, $P(\theta)$?

1. A probability density that represents all prior evidence about θ.
2. An opportunity: a way of getting other information outside the data set into the model

What is the prior density, $P(\theta)$?

1. A probability density that represents all prior evidence about θ.
2. An opportunity: a way of getting other information outside the data set into the model
3. An annoyance: the "other information" is required

What is the prior density, $P(\theta)$?

1. A probability density that represents all prior evidence about θ.
2. An opportunity: a way of getting other information outside the data set into the model
3. An annoyance: the "other information" is required
4. A philosophical assumption that nonsample information should matter (as it always does) and be formalized and included in all inferences.

Principles of Bayesian analysis

Principles of Bayesian analysis

1. All unknown quantities (θ, Y) are treated as random variables and have a joint probability distribution.

Principles of Bayesian analysis

1. All unknown quantities (θ, Y) are treated as random variables and have a joint probability distribution.
2. All known quantities (y) are treated as fixed.

Principles of Bayesian analysis

1. All unknown quantities (θ, Y) are treated as random variables and have a joint probability distribution.
2. All known quantities (y) are treated as fixed.
3. If we have observed variable B and unobserved variable A, then we are usually interested in the conditional distribution of A, given B: $P(A|B) = P(A, B)/P(B)$

Principles of Bayesian analysis

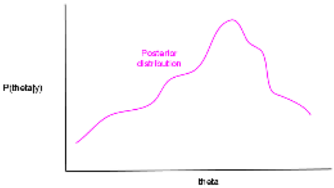
1. All unknown quantities (θ, Y) are treated as random variables and have a joint probability distribution.
2. All known quantities (y) are treated as fixed.
3. If we have observed variable B and unobserved variable A, then we are usually interested in the conditional distribution of A, given B: $P(A|B) = P(A, B)/P(B)$
4. If variables A and B are both unknown, then the distribution of A alone is $P(A) = \int \mathbb{P}(A, B)dB = \int P(A|B)P(B)dB$.

The posterior density, $\mathbb{P}(\theta|y)$

The posterior density, $\mathbb{P}(\theta|y)$

- Like L, it's a summary estimator

The posterior density, $\mathbb{P}(\theta|y)$



- Like L, it's a summary estimator
- Unlike L, it's a real probability density, from which we can derive probabilistic statements (via integration)

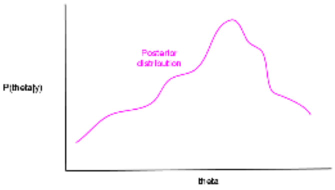
The posterior density, $\mathbb{P}(\theta|y)$

- Like L, it's a summary estimator
- Unlike L, it's a real probability density, from which we can derive probabilistic statements (via integration)
- To compare across applications or data sets, you may need different priors. So, the posterior is also relative, just like likelihood.

The posterior density, $\mathbb{P}(\theta|y)$

- Like L, it's a summary estimator
- Unlike L, it's a real probability density, from which we can derive probabilistic statements (via integration)
- To compare across applications or data sets, you may need different priors. So, the posterior is also relative, just like likelihood.
- Bayesian inference obeys the likelihood principle: the data set only affects inferences through the likelihood function

The posterior density, $\mathbb{P}(\theta|y)$



- Like L, it's a summary estimator
- Unlike L, it's a real probability density, from which we can derive probabilistic statements (via integration)
- To compare across applications or data sets, you may need different priors. So, the posterior is also relative, just like likelihood.
- Bayesian inference obeys the likelihood principle: the data set only affects inferences through the likelihood function
- If $\mathbb{P}(\theta) = 1$, i.e., is uniform in the relevant region, then $L(\theta|y) = \mathbb{P}(\theta|y)$.

Being a Bayesian

Being a Bayesian

- If $\mathbb{P}(\theta)$ is <u>diffuse</u>, differences from likelihood are minor, but numerical stability (and "identification") is improved

Being a Bayesian

- If $\mathbb{P}(\theta)$ is <u>diffuse</u>, differences from likelihood are minor, but numerical stability (and "identification") is improved
- Philosophical differences from likelihood: Huge

Being a Bayesian

- If $\mathbb{P}(\theta)$ is diffuse, differences from likelihood are minor, but numerical stability (and "identification") is improved
- Philosophical differences from likelihood: Huge
- Practical differences when we can compute both: often Minor (unless the prior matters)

Being a Bayesian

- If $\mathbb{P}(\theta)$ is diffuse, differences from likelihood are minor, but numerical stability (and "identification") is improved
- Philosophical differences from likelihood: Huge
- Practical differences when we can compute both: often Minor (unless the prior matters)
- Advantages: more information produces more efficiency;

Being a Bayesian

- If $\mathbb{P}(\theta)$ is diffuse, differences from likelihood are minor, but numerical stability (and "identification") is improved
- Philosophical differences from likelihood: Huge
- Practical differences when we can compute both: often Minor (unless the prior matters)
- Advantages: more information produces more efficiency;
- Few fights now between Bayesians and likelihoodists

Being a Bayesian

- If $\mathbb{P}(\theta)$ is <u>diffuse</u>, differences from likelihood are minor, but numerical stability (and "identification") is improved
- Philosophical differences from likelihood: Huge
- Practical differences when we can compute both: often Minor (unless the prior matters)
- Advantages: more information produces more efficiency;
- Few fights now between Bayesians and likelihoodists
- In general simple computation is easier under MLE, complex computation is *dramatically* easier under Bayes (the more parameters- the more you should think about Bayes).

Being a Bayesian

- If $\mathbb{P}(\theta)$ is <u>diffuse</u>, differences from likelihood are minor, but numerical stability (and "identification") is improved
- Philosophical differences from likelihood: Huge
- Practical differences when we can compute both: often Minor (unless the prior matters)
- Advantages: more information produces more efficiency;
- Few fights now between Bayesians and likelihoodists
- In general simple computation is easier under MLE, complex computation is *dramatically* easier under Bayes (the more parameters- the more you should think about Bayes).
- A perspective of growing importance is empirical Bayes which we will discuss later in the semester.

1. History
2. Likelihood Inference
3. Bayesian Inference
4. Neyman-Pearson
5. Likelihood Example
6. Properties and Tests
7. Simulation
8. Fun With Bayes

1 History

2 Likelihood Inference

3 Bayesian Inference

4 Neyman-Pearson

5 Likelihood Example

6 Properties and Tests

7 Simulation

8 Fun With Bayes

A 3rd Theory: Neyman-Pearson Hypothesis Testing

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}
2. Strict but arbitrary distinction: null H_0 vs alternative H_1 hypotheses

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}
2. Strict but arbitrary distinction: null H_0 vs alternative H_1 hypotheses
3. All tests are "under" (i.e., assuming) H_0

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}
2. Strict but arbitrary distinction: null H_0 vs alternative H_1 hypotheses
3. All tests are "under" (i.e., assuming) H_0

For example, is $\beta = 0$ in $E(Y) = \beta_0 + \beta X$?

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}
2. Strict but arbitrary distinction: null H_0 vs alternative H_1 hypotheses
3. All tests are "under" (i.e., assuming) H_0

For example, is $\beta = 0$ in $E(Y) = \beta_0 + \beta X$?

- H_0: $\beta = 0$ vs. H_1: $\beta > 0$

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}
2. Strict but arbitrary distinction: null H_0 vs alternative H_1 hypotheses
3. All tests are "under" (i.e., assuming) H_0

For example, is $\beta = 0$ in $E(Y) = \beta_0 + \beta X$?

- H_0: $\beta = 0$ vs. H_1: $\beta > 0$
- Choose Type I error, probability of deciding H_1 is right when H_0 is really true: say $\alpha = 0.05$

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}
2. Strict but arbitrary distinction: null H_0 vs alternative H_1 hypotheses
3. All tests are "under" (i.e., assuming) H_0

For example, is $\beta = 0$ in $E(Y) = \beta_0 + \beta X$?

- H_0: $\beta = 0$ vs. H_1: $\beta > 0$
- Choose Type I error, probability of deciding H_1 is right when H_0 is really true: say $\alpha = 0.05$
- (Type II error, the power to detect H_1 if it is true, is a consequence of choosing an estimator, not an ex ante decision like choosing α.)

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}
2. Strict but arbitrary distinction: null H_0 vs alternative H_1 hypotheses
3. All tests are "under" (i.e., assuming) H_0

For example, is $\beta = 0$ in $E(Y) = \beta_0 + \beta X$?

- H_0: $\beta = 0$ vs. H_1: $\beta > 0$
- Choose Type I error, probability of deciding H_1 is right when H_0 is really true: say $\alpha = 0.05$
- (Type II error, the power to detect H_1 if it is true, is a consequence of choosing an estimator, not an ex ante decision like choosing α.)
- Assume n is large enough for the CLT to kick in

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}
2. Strict but arbitrary distinction: null H_0 vs alternative H_1 hypotheses
3. All tests are "under" (i.e., assuming) H_0

For example, is $\beta = 0$ in $E(Y) = \beta_0 + \beta X$?

- H_0: $\beta = 0$ vs. H_1: $\beta > 0$
- Choose Type I error, probability of deciding H_1 is right when H_0 is really true: say $\alpha = 0.05$
- (Type II error, the power to detect H_1 if it is true, is a consequence of choosing an estimator, not an ex ante decision like choosing α.)
- Assume n is large enough for the CLT to kick in
- Then $b|(\beta = 0) \sim N(0, \sigma_b^2)$

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}
2. Strict but arbitrary distinction: null H_0 vs alternative H_1 hypotheses
3. All tests are "under" (i.e., assuming) H_0

For example, is $\beta = 0$ in $E(Y) = \beta_0 + \beta X$?

- H_0: $\beta = 0$ vs. H_1: $\beta > 0$
- Choose Type I error, probability of deciding H_1 is right when H_0 is really true: say $\alpha = 0.05$
- (Type II error, the power to detect H_1 if it is true, is a consequence of choosing an estimator, not an ex ante decision like choosing α.)
- Assume n is large enough for the CLT to kick in
- Then $b|(\beta = 0) \sim N(0, \sigma_b^2)$
- or

$$(TS)_\beta|(\beta = 0) \equiv \frac{b - \beta}{\hat{\sigma}_b} \equiv \frac{b}{\hat{\sigma}_b} \sim N(0, 1).$$

Neyman-Pearson Hypothesis Testing

Neyman-Pearson Hypothesis Testing

- Derive critical value, CV, e.g., the right tail:

$$\int_{(CV)}^{\infty} N(b|0, \sigma_b^2)db = \alpha$$

Neyman-Pearson Hypothesis Testing

- Derive critical value, CV, e.g., the right tail:

$$\int_{(CV)}^{\infty} N(b|0, \sigma_b^2)db = \alpha$$

- This means in principle: write your prospectus, plan your experiment, report the CV, and write your concluding chapter (loosely as follows):

Neyman-Pearson Hypothesis Testing

- Derive critical value, CV, e.g., the right tail:

$$\int_{(CV)}^{\infty} N(b|0, \sigma_b^2) db = \alpha$$

- This means in principle: write your prospectus, plan your experiment, report the CV, and write your concluding chapter (loosely as follows):

$$\text{Decision} =$$

Neyman-Pearson Hypothesis Testing

- Derive critical value, CV, e.g., the right tail:

$$\int_{(CV)}^{\infty} N(b|0, \sigma_b^2)db = \alpha$$

- This means in principle: write your prospectus, plan your experiment, report the CV, and write your concluding chapter (loosely as follows):

$$\text{Decision} = \begin{cases} \beta > 0 \text{ (I was right)} & \text{if } (TS) > (CV) \\ \beta = 0 \text{ (I was wrong)} & \text{if } (TS) \leq (CV) \end{cases}$$

Neyman-Pearson Hypothesis Testing

- Derive critical value, CV, e.g., the right tail:

$$\int_{(CV)}^{\infty} N(b|0, \sigma_b^2)db = \alpha$$

- This means in principle: write your prospectus, plan your experiment, report the CV, and write your concluding chapter (loosely as follows):

$$\text{Decision} = \begin{cases} \beta > 0 \text{ (I was right)} & \text{if } (TS) > (CV) \\ \beta = 0 \text{ (I was wrong)} & \text{if } (TS) \leq (CV) \end{cases}$$

  And then collect your data. You may not revise your hypothesis or your theory.

Neyman-Pearson Hypothesis Testing

- Derive critical value, CV, e.g., the right tail:

$$\int_{(CV)}^{\infty} N(b|0, \sigma_b^2) db = \alpha$$

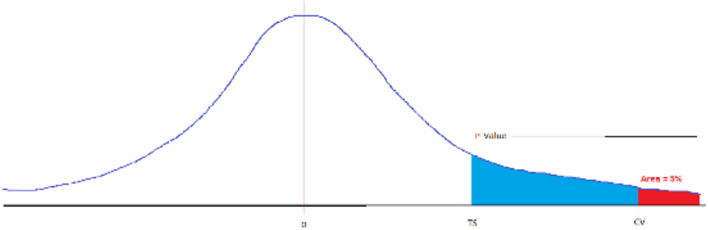
- This means in principle: write your prospectus, plan your experiment, report the CV, and write your concluding chapter (loosely as follows):

$$\text{Decision} = \begin{cases} \beta > 0 \text{ (I was right)} & \text{if } (TS) > (CV) \\ \beta = 0 \text{ (I was wrong)} & \text{if } (TS) \leq (CV) \end{cases}$$

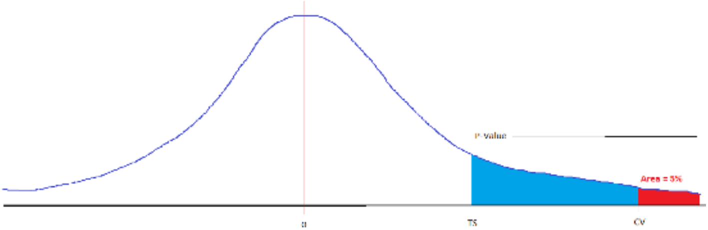
And then collect your data. You may not revise your hypothesis or your theory. When is this good?

Neyman-Pearson Hypothesis Testing

Neyman-Pearson Hypothesis Testing

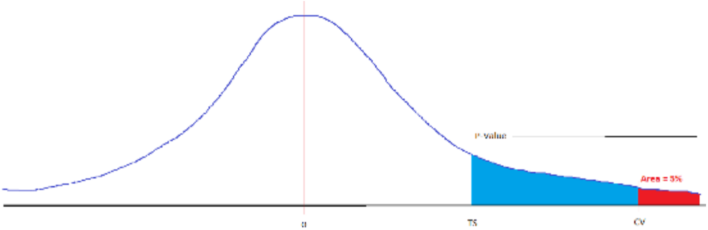


Neyman-Pearson Hypothesis Testing



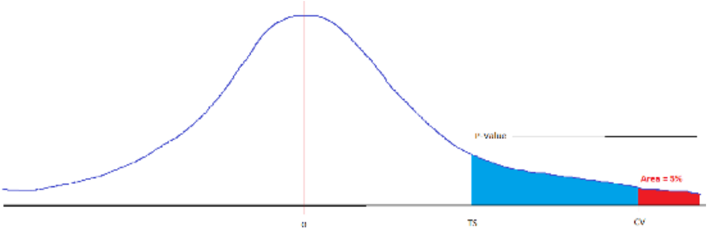
- In this example, $(TS) < (CV)$ and so we conclude that we can't reject $\beta = 0$.

Neyman-Pearson Hypothesis Testing



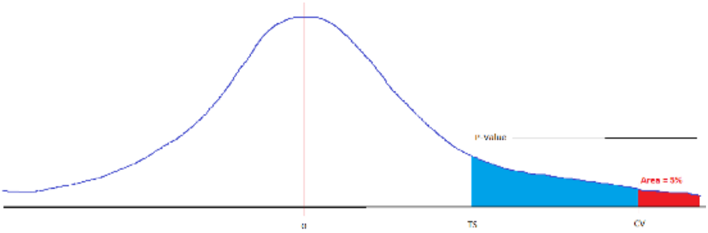
- In this example, $(TS) < (CV)$ and so we conclude that we can't reject $\beta = 0$.
- What's our best guess? We don't have one- it is a decision.

Neyman-Pearson Hypothesis Testing



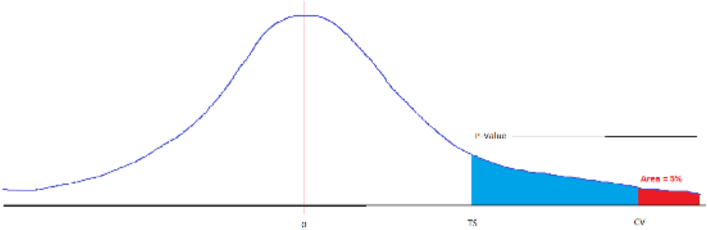
- In this example, $(TS) < (CV)$ and so we conclude that we can't reject $\beta = 0$.
- What's our best guess? We don't have one- it is a decision.
- Decision will be wrong 5% of the time; what about this time?

Neyman-Pearson Hypothesis Testing



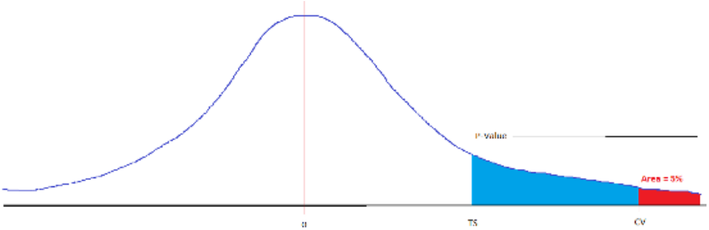
- In this example, $(TS) < (CV)$ and so we conclude that we can't reject $\beta = 0$.
- What's our best guess? We don't have one- it is a decision.
- Decision will be wrong 5% of the time; what about this time?
- What about when n is large or under control of the investigator?

Neyman-Pearson Hypothesis Testing



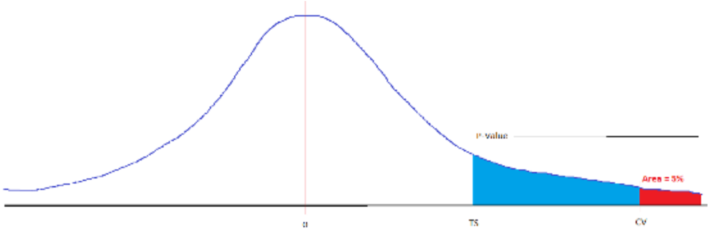
- In this example, $(TS) < (CV)$ and so we conclude that we can't reject $\beta = 0$.
- What's our best guess? We don't have one- it is a decision.
- Decision will be wrong 5% of the time; what about this time?
- What about when n is large or under control of the investigator?
- In practice, hypothesis testing is used with p-values:

Neyman-Pearson Hypothesis Testing



- In this example, $(TS) < (CV)$ and so we conclude that we can't reject $\beta = 0$.
- What's our best guess? We don't have one- it is a decision.
- Decision will be wrong 5% of the time; what about this time?
- What about when n is large or under control of the investigator?
- In practice, hypothesis testing is used with p-values: The probability under the null of getting a value as weird or weirder than the value we got — the area to the right of the realized value of (TS).

Neyman-Pearson Hypothesis Testing



- In this example, $(TS) < (CV)$ and so we conclude that we can't reject $\beta = 0$.
- What's our best guess? We don't have one- it is a decision.
- Decision will be wrong 5% of the time; what about this time?
- What about when n is large or under control of the investigator?
- In practice, hypothesis testing is used with p-values: The probability under the null of getting a value as weird or weirder than the value we got — the area to the right of the realized value of (TS).
- Is this really our quantity of interest?

What is the right theory of inference?

What is the right theory of inference?

1. Likelihood?

What is the right theory of inference?

1. Likelihood? Bayes?

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson?

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson?
   Criteria estimators?

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson?
   Criteria estimators? Finite or asymptotic based theory?

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson?
   Criteria estimators? Finite or asymptotic based theory? Decision theory?

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson?
   Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics?

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson?
   Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics?

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson?
   Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference?

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson?
   Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference? Superpopulation-based inference?

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson?
   Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference? Superpopulation-based inference? etc.

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson?
   Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference? Superpopulation-based inference? etc.
2. No

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson?
   Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference? Superpopulation-based inference? etc.
2. None

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson?
   Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference? Superpopulation-based inference? etc.
2. None of

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson?
   Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference? Superpopulation-based inference? etc.
2. None of these.

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson?
   Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference? Superpopulation-based inference? etc.
2. None of these.
3. The right theory of inference:

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson?
   Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference? Superpopulation-based inference? etc.
2. None of these.
3. The right theory of inference: pragmatism

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson?
   Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference? Superpopulation-based inference? etc.
2. None of these.
3. The right theory of inference: pragmatism
4. Methods for applied researchers: either useful or irrelevant $\rightarrow$ learn something then validate it.

Unification of Theories of Inference

Unification of Theories of Inference

- Can't bank on agreement on normative issues!

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't
- Alternative convergence is occurring: different methods giving the same result.

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't
- Alternative convergence is occurring: different methods giving the same result.
  - Likelihood or Bayes with careful goodness of fit checks

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't
- Alternative convergence is occurring: different methods giving the same result.
  - Likelihood or Bayes with careful goodness of fit checks
  - Various types of robust or semi-parametric methods

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't
- Alternative convergence is occurring: different methods giving the same result.
  - Likelihood or Bayes with careful goodness of fit checks
  - Various types of robust or semi-parametric methods
  - Matching for use as preprocessing for parametric analysis

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't
- Alternative convergence is occurring: different methods giving the same result.
  - Likelihood or Bayes with careful goodness of fit checks
  - Various types of robust or semi-parametric methods
  - Matching for use as preprocessing for parametric analysis
  - Some models with highly flexible functional forms

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't
- Alternative convergence is occurring: different methods giving the same result.
  - Likelihood or Bayes with careful goodness of fit checks
  - Various types of robust or semi-parametric methods
  - Matching for use as preprocessing for parametric analysis
  - Some models with highly flexible functional forms
- The key: No assumptions can always be trusted;

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't
- Alternative convergence is occurring: different methods giving the same result.
  - Likelihood or Bayes with careful goodness of fit checks
  - Various types of robust or semi-parametric methods
  - Matching for use as preprocessing for parametric analysis
  - Some models with highly flexible functional forms
- The key: No assumptions can always be trusted;

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't
- Alternative convergence is occurring: different methods giving the same result.
  - Likelihood or Bayes with careful goodness of fit checks
  - Various types of robust or semi-parametric methods
  - Matching for use as preprocessing for parametric analysis
  - Some models with highly flexible functional forms
- The key: No assumptions can always be trusted; all theories of inference condition on assumptions and so data analysts always struggle trying to understand and check them

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't
- Alternative convergence is occurring: different methods giving the same result.
  - Likelihood or Bayes with careful goodness of fit checks
  - Various types of robust or semi-parametric methods
  - Matching for use as preprocessing for parametric analysis
  - Some models with highly flexible functional forms
- The key: No assumptions can always be trusted; all theories of inference condition on assumptions and so data analysts always struggle trying to understand and check them
- This motivates different views of the core material such as agnostic and robust statistics.

1. History
2. Likelihood Inference
3. Bayesian Inference
4. Neyman-Pearson
5. Likelihood Example
6. Properties and Tests
7. Simulation
8. Fun With Bayes

1 History

2 Likelihood Inference

3 Bayesian Inference

4 Neyman-Pearson

5 Likelihood Example

6 Properties and Tests

7 Simulation

8 Fun With Bayes

A Simple Likelihood Model: Stylized Normal, no X

A Simple Likelihood Model: Stylized Normal, no X

The model:

A Simple Likelihood Model: Stylized Normal, no X

The model:

1. $Y_i \sim f_{\text{stn}}(y_i|\mu_i)$, normal stochastic component

A Simple Likelihood Model: Stylized Normal, no X

The model:

1. $Y_i \sim f_{\text{stn}}(y_i|\mu_i)$, normal stochastic component
2. $\mu_i = \beta$, a constant systematic component (no covariates)

A Simple Likelihood Model: Stylized Normal, no X

The model:

1. $Y_i \sim f_{\text{stn}}(y_i|\mu_i)$, normal stochastic component
2. $\mu_i = \beta$, a constant systematic component (no covariates)
3. Y_i and Y_j are independent $\forall\ i \neq j$.

A Simple Likelihood Model: Stylized Normal, no X

The model:

1. $Y_i \sim f_{\text{stn}}(y_i|\mu_i)$, normal stochastic component
2. $\mu_i = \beta$, a constant systematic component (no covariates)
3. Y_i and Y_j are independent $\forall\ i \neq j$.

Derive the full probability density of all observations, Pr(data|model) (Recall: if A and B are independent, $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$):

A Simple Likelihood Model: Stylized Normal, no X

The model:

1. $Y_i \sim f_{\text{stn}}(y_i|\mu_i)$, normal stochastic component
2. $\mu_i = \beta$, a constant systematic component (no covariates)
3. Y_i and Y_j are independent $\forall\ i \neq j$.

Derive the full probability density of all observations, Pr(data|model) (Recall: if A and B are independent, $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$):

$$\mathbb{P}(y|\mu) \equiv \mathbb{P}(y_1, \ldots, y_n|\mu_1, \ldots, \mu_n) = \prod_{i=1}^{n} f_{\text{stn}}(y_i|\mu_i)$$

A Simple Likelihood Model: Stylized Normal, no X

The model:

1. $Y_i \sim f_{\text{stn}}(y_i|\mu_i)$, normal stochastic component
2. $\mu_i = \beta$, a constant systematic component (no covariates)
3. Y_i and Y_j are independent $\forall\ i \neq j$.

Derive the full probability density of all observations, Pr(data|model)
(Recall: if A and B are independent, $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$):

$$\begin{aligned}
\mathbb{P}(y|\mu) \equiv \mathbb{P}(y_1, \ldots, y_n|\mu_1, \ldots, \mu_n) &= \prod_{i=1}^{n} f_{\text{stn}}(y_i|\mu_i) \\
&= \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2}\right)
\end{aligned}$$

A Simple Likelihood Model: Stylized Normal, no X

The model:

1. $Y_i \sim f_{\text{stn}}(y_i|\mu_i)$, normal stochastic component
2. $\mu_i = \beta$, a constant systematic component (no covariates)
3. Y_i and Y_j are independent $\forall\ i \neq j$.

Derive the full probability density of all observations, Pr(data|model) (Recall: if A and B are independent, $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$):

$$
\begin{aligned}
\mathbb{P}(y|\mu) \equiv \mathbb{P}(y_1, \ldots, y_n|\mu_1, \ldots, \mu_n) &= \prod_{i=1}^{n} f_{\text{stn}}(y_i|\mu_i) \\
&= \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2}\right)
\end{aligned}
$$

reparameterizing with $\mu_i = \beta$:

A Simple Likelihood Model: Stylized Normal, no X

The model:

1. $Y_i \sim f_{\text{stn}}(y_i|\mu_i)$, normal stochastic component
2. $\mu_i = \beta$, a constant systematic component (no covariates)
3. Y_i and Y_j are independent $\forall\ i \neq j$.

Derive the full probability density of all observations, Pr(data|model) (Recall: if A and B are independent, $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$):

$$\mathbb{P}(y|\mu) \equiv \mathbb{P}(y_1, \ldots, y_n|\mu_1, \ldots, \mu_n) = \prod_{i=1}^{n} f_{\text{stn}}(y_i|\mu_i)$$
$$= \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2}\right)$$

reparameterizing with $\mu_i = \beta$:

$$\mathbb{P}(y|\beta) \equiv \mathbb{P}(y_1, \ldots, y_n|\beta) = \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left(\frac{-(y_i - \beta)^2}{2}\right)$$

A Simple Likelihood Model: Stylized Normal, no X

The model:

1. $Y_i \sim f_{\text{stn}}(y_i|\mu_i)$, normal stochastic component
2. $\mu_i = \beta$, a constant systematic component (no covariates)
3. Y_i and Y_j are independent $\forall\ i \neq j$.

Derive the full probability density of all observations, Pr(data|model) (Recall: if A and B are independent, $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$):

$$
\begin{aligned}
\mathbb{P}(y|\mu) \equiv \mathbb{P}(y_1, \dots, y_n|\mu_1, \dots, \mu_n) &= \prod_{i=1}^{n} f_{\text{stn}}(y_i|\mu_i) \\
&= \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2}\right)
\end{aligned}
$$

reparameterizing with $\mu_i = \beta$:

$$
\mathbb{P}(y|\beta) \equiv \mathbb{P}(y_1, \dots, y_n|\beta) = \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left(\frac{-(y_i - \beta)^2}{2}\right)
$$

- What can you do with this probability density?

Stylized Normal Likelihood Function

Stylized Normal Likelihood Function

The likelihood of β (conditional on the model) having generated the data we observe.

Stylized Normal Likelihood Function

The likelihood of β (conditional on the model) having generated the data we observe.

$$L(\beta|y) = k(y)\prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta)$$

Stylized Normal Likelihood Function

The likelihood of β (conditional on the model) having generated the data we observe.

$$L(\beta|y) = k(y)\prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \propto \prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta)$$

Stylized Normal Likelihood Function

The likelihood of β (conditional on the model) having generated the data we observe.

$$
\begin{aligned}
L(\beta|y) &= k(y)\prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \propto \prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \\
&= \prod_{i=1}^{n}(2\pi)^{-1/2}\exp\left(\frac{-(y_i-\beta)^2}{2}\right)
\end{aligned}
$$

Stylized Normal Likelihood Function

The likelihood of β (conditional on the model) having generated the data we observe.

$$L(\beta|y) = k(y)\prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \propto \prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta)$$
$$= \prod_{i=1}^{n}(2\pi)^{-1/2}\exp\left(\frac{-(y_i-\beta)^2}{2}\right)$$

The log-likelihood (Recall: $\ln(ab) = \ln(a) + \ln(b)$):

Stylized Normal Likelihood Function

The likelihood of β (conditional on the model) having generated the data we observe.

$$L(\beta|y) = k(y)\prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \propto \prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta)$$
$$= \prod_{i=1}^{n}(2\pi)^{-1/2}\exp\left(\frac{-(y_i-\beta)^2}{2}\right)$$

The log-likelihood (Recall: $\ln(ab) = \ln(a) + \ln(b)$):

$$\ln L(\beta|y) = \ln[k(y)] + \sum_{i=1}^{n} \ln f_{\text{stn}}(y_i|\beta)$$

Stylized Normal Likelihood Function

The likelihood of β (conditional on the model) having generated the data we observe.

$$L(\beta|y) = k(y)\prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \propto \prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta)$$
$$= \prod_{i=1}^{n}(2\pi)^{-1/2}\exp\left(\frac{-(y_i-\beta)^2}{2}\right)$$

The log-likelihood (Recall: $\ln(ab) = \ln(a) + \ln(b)$):

$$\ln L(\beta|y) = \ln[k(y)] + \sum_{i=1}^{n} \ln f_{\text{stn}}(y_i|\beta)$$
$$= \ln[k(y)] + \sum_{i=1}^{n} \ln[(2\pi)^{-1/2}] - \sum_{i=1}^{n} \frac{1}{2}(y_i-\beta)^2$$

Stylized Normal Likelihood Function

The likelihood of β (conditional on the model) having generated the data we observe.

$$L(\beta|y) = k(y)\prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \propto \prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta)$$
$$= \prod_{i=1}^{n}(2\pi)^{-1/2}\exp\left(\frac{-(y_i-\beta)^2}{2}\right)$$

The log-likelihood (Recall: $\ln(ab) = \ln(a) + \ln(b)$):

$$\ln L(\beta|y) = \ln[k(y)] + \sum_{i=1}^{n}\ln f_{\text{stn}}(y_i|\beta)$$
$$= \ln[k(y)] + \sum_{i=1}^{n}\ln[(2\pi)^{-1/2}] - \sum_{i=1}^{n}\frac{1}{2}(y_i-\beta)^2$$
$$\doteq \sum_{i=1}^{n} -\frac{1}{2}(y_i-\beta)^2$$

Stylized Normal Likelihood Function

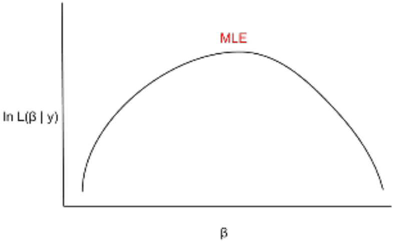
The likelihood of β (conditional on the model) having generated the data we observe.

$$L(\beta|y) = k(y)\prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \propto \prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta)$$
$$= \prod_{i=1}^{n}(2\pi)^{-1/2}\exp\left(\frac{-(y_i-\beta)^2}{2}\right)$$

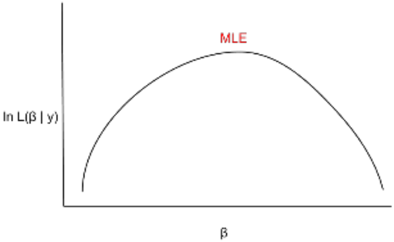
The log-likelihood (Recall: $\ln(ab) = \ln(a) + \ln(b)$):

$$\ln L(\beta|y) = \ln[k(y)] + \sum_{i=1}^{n} \ln f_{\text{stn}}(y_i|\beta)$$
$$= \ln[k(y)] + \sum_{i=1}^{n} \ln[(2\pi)^{-1/2}] - \sum_{i=1}^{n}\frac{1}{2}(y_i-\beta)^2$$
$$\doteq \sum_{i=1}^{n} -\frac{1}{2}(y_i-\beta)^2 = -\frac{1}{2}\sum_{i=1}^{n}(y_i-\beta)^2$$

Log-likelihood interpretation

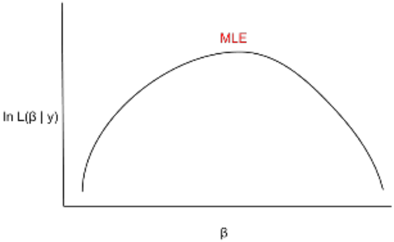


Log-likelihood interpretation



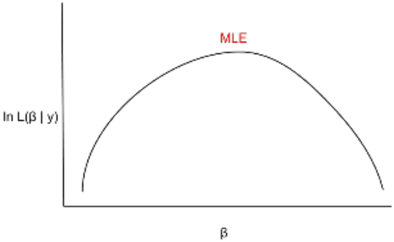
1. The log-likelihood is quadratic

Log-likelihood interpretation



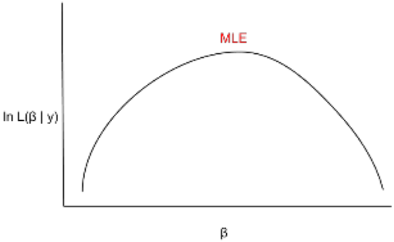
1. The log-likelihood is quadratic
2. This curve summarizes all information the data gives about β, assuming the model.

Log-likelihood interpretation



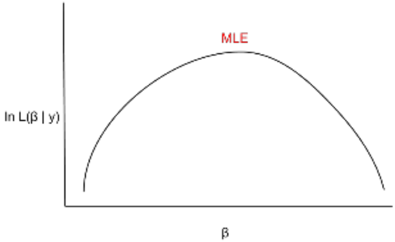
1. The log-likelihood is quadratic
2. This curve summarizes all information the data gives about β, assuming the model.
3. The MLE is at the same point as the MVLUE (minimum variance linear unbiased estimator)

Log-likelihood interpretation



1. The log-likelihood is quadratic
2. This curve summarizes all information the data gives about β, assuming the model.
3. The MLE is at the same point as the MVLUE (minimum variance linear unbiased estimator)
4. The maximum is at the same point as the least squares point

Log-likelihood interpretation



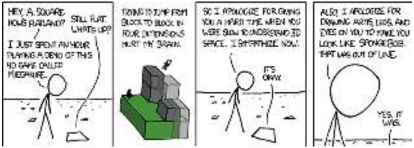
1. The log-likelihood is quadratic
2. This curve summarizes all information the data gives about β, assuming the model.
3. The MLE is at the same point as the MVLUE (minimum variance linear unbiased estimator)
4. The maximum is at the same point as the least squares point
5. No reason to summarize this curve with only the MLE

Summarizing k-dimensional space

- The problem of Flatland

Summarizing k-dimensional space

- The problem of Flatland



Summarizing k-dimensional space

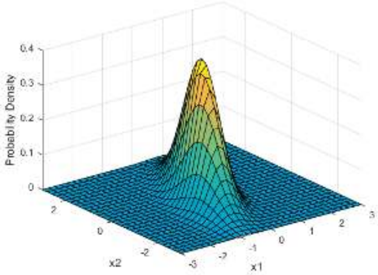
- The problem of Flatland
- Graphs

Summarizing k-dimensional space

- The problem of Flatland
- Graphs
- The curse of dimensionality

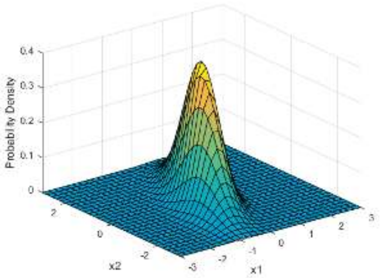
Summarizing k-dimensional space

- The problem of Flatland
- Graphs
- The curse of dimensionality
- Maximum



Summarizing k-dimensional space

- The problem of Flatland
- Graphs
- The curse of dimensionality
- Maximum
- The curvature at the maximum (standard errors, about which more shortly)



How to find the maximum?

How to find the maximum?

Goal: Find the value of $\theta \equiv \{\theta_1, \ldots, \theta_k\}$ that maximizes $L(\theta|y)$

How to find the maximum?

Goal: Find the value of $\theta \equiv \{\theta_1, \ldots, \theta_k\}$ that maximizes $L(\theta|y)$

1. Analytically — often impossible or too hard

How to find the maximum?

Goal: Find the value of $\theta \equiv \{\theta_1, \ldots, \theta_k\}$ that maximizes $L(\theta|y)$

1. Analytically — often impossible or too hard
   - Take the derivative of $\ln L(\theta|y)$ w.r.t. θ

How to find the maximum?

Goal: Find the value of $\theta \equiv \{\theta_1, \ldots, \theta_k\}$ that maximizes $L(\theta|y)$

1. Analytically — often impossible or too hard
   - Take the derivative of $\ln L(\theta|y)$ w.r.t. θ
   - Set to 0, substituting $\hat{\theta}$ for θ

How to find the maximum?

Goal: Find the value of $\theta \equiv \{\theta_1, \ldots, \theta_k\}$ that maximizes $L(\theta|y)$

1. Analytically — often impossible or too hard
   - Take the derivative of $\ln L(\theta|y)$ w.r.t. θ
   - Set to 0, substituting $\hat{\theta}$ for θ

$$\left| \frac{\partial \ln L(\theta|y)}{\partial \theta} \right|_{\theta=\hat{\theta}} = 0$$

How to find the maximum?

Goal: Find the value of $\theta \equiv \{\theta_1, \ldots, \theta_k\}$ that maximizes $L(\theta|y)$

1. Analytically — often impossible or too hard
   - Take the derivative of $\ln L(\theta|y)$ w.r.t. θ
   - Set to 0, substituting $\hat{\theta}$ for θ

$$\left| \frac{\partial \ln L(\theta|y)}{\partial \theta} \right|_{\theta=\hat{\theta}} = 0$$

   - If possible, solve for θ, and label it $\hat{\theta}$

How to find the maximum?

Goal: Find the value of $\theta \equiv \{\theta_1, \ldots, \theta_k\}$ that maximizes $L(\theta|y)$

1. Analytically — often impossible or too hard
   - Take the derivative of $\ln L(\theta|y)$ w.r.t. θ
   - Set to 0, substituting $\hat{\theta}$ for θ

$$\left| \frac{\partial \ln L(\theta|y)}{\partial \theta} \right|_{\theta=\hat{\theta}} = 0$$

   - If possible, solve for θ, and label it $\hat{\theta}$
   - Check the second order condition: see if the second derivative w.r.t. θ is negative (so its a maximum rather than a minimum)

How to find the maximum?

Goal: Find the value of $\theta \equiv \{\theta_1, \ldots, \theta_k\}$ that maximizes $L(\theta|y)$

1. Analytically — often impossible or too hard
   - Take the derivative of $\ln L(\theta|y)$ w.r.t. θ
   - Set to 0, substituting $\hat{\theta}$ for θ

$$\left| \frac{\partial \ln L(\theta|y)}{\partial \theta} \right|_{\theta = \hat{\theta}} = 0$$

   - If possible, solve for θ, and label it $\hat{\theta}$
   - Check the second order condition: see if the second derivative w.r.t. θ is negative (so its a maximum rather than a minimum)
2. Numerically — let the computer do the work for you

How to find the maximum?

Goal: Find the value of $\theta \equiv \{\theta_1, \ldots, \theta_k\}$ that maximizes $L(\theta|y)$

1. Analytically — often impossible or too hard
   - Take the derivative of $\ln L(\theta|y)$ w.r.t. θ
   - Set to 0, substituting $\hat{\theta}$ for θ

$$\left| \frac{\partial \ln L(\theta|y)}{\partial \theta} \right|_{\theta=\hat{\theta}} = 0$$

   - If possible, solve for θ, and label it $\hat{\theta}$
   - Check the second order condition: see if the second derivative w.r.t. θ is negative (so its a maximum rather than a minimum)
2. Numerically — let the computer do the work for you
   - We'll show you how in precept

How to find the maximum?

Goal: Find the value of $\theta \equiv \{\theta_1, \ldots, \theta_k\}$ that maximizes $L(\theta|y)$

1. Analytically — often impossible or too hard
   - Take the derivative of $\ln L(\theta|y)$ w.r.t. θ
   - Set to 0, substituting $\hat{\theta}$ for θ

$$\left|\frac{\partial \ln L(\theta|y)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0$$

   - If possible, solve for θ, and label it $\hat{\theta}$
   - Check the second order condition: see if the second derivative w.r.t. θ is negative (so its a maximum rather than a minimum)
2. Numerically — let the computer do the work for you
   - We'll show you how in precept
   - Most commonly gradient descent

How to find the maximum?

Goal: Find the value of $\theta \equiv \{\theta_1, \ldots, \theta_k\}$ that maximizes $L(\theta|y)$

1. Analytically — often impossible or too hard
   - Take the derivative of $\ln L(\theta|y)$ w.r.t. θ
   - Set to 0, substituting $\hat{\theta}$ for θ

$$\left|\frac{\partial \ln L(\theta|y)}{\partial\theta}\right|_{\theta=\hat{\theta}} = 0$$

   - If possible, solve for θ, and label it $\hat{\theta}$
   - Check the second order condition: see if the second derivative w.r.t. θ is negative (so its a maximum rather than a minimum)
2. Numerically — let the computer do the work for you
   - We'll show you how in precept
   - Most commonly gradient descent
   - Not a sharp divide- some analytic work helps numerical optimization

Example 2: Age distribution of ER visits due to wall punching

- We have a dataset from the U.S. Consumer Product Safety Commission's National Electronic Injury Surveillance System (NEISS) containing data on ER visits in 2014.

Example 2: Age distribution of ER visits due to wall punching

- We have a dataset from the U.S. Consumer Product Safety Commission's National Electronic Injury Surveillance System (NEISS) containing data on ER visits in 2014.
- Let's take a look at one injury category – wall punching.

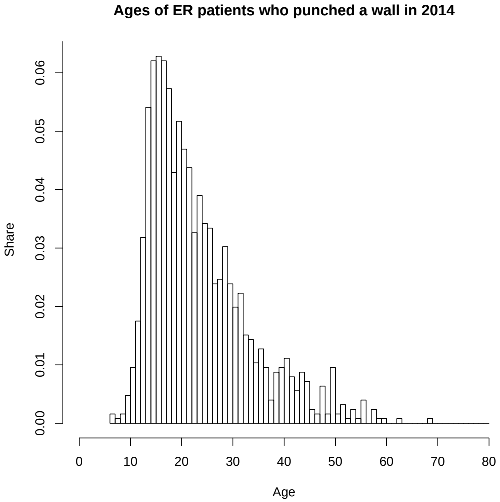
Example 2: Age distribution of ER visits due to wall punching

- We have a dataset from the U.S. Consumer Product Safety Commission's National Electronic Injury Surveillance System (NEISS) containing data on ER visits in 2014.
- Let's take a look at one injury category – wall punching. We're interested in modelling the distribution of the ages of individuals who visit the ER having punched a wall.

Example 2: Age distribution of ER visits due to wall punching

- We have a dataset from the U.S. Consumer Product Safety Commission's National Electronic Injury Surveillance System (NEISS) containing data on ER visits in 2014.
- Let's take a look at one injury category – wall punching. We're interested in modelling the distribution of the ages of individuals who visit the ER having punched a wall.
- To do this, we write down a probability model for the data.

Empirical distribution of wall-punching ages



A Model for the Data – Log-Normal distribution

- We observe n observations of ages, $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$.

A Model for the Data – Log-Normal distribution

- We observe n observations of ages, $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$.
- A normal distribution doesn't seem like a reasonable model since age is strictly positive and the distribution is somewhat right-skewed.

A Model for the Data – Log-Normal distribution

- We observe n observations of ages, $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$.
- A normal distribution doesn't seem like a reasonable model since age is strictly positive and the distribution is somewhat right-skewed.
- But a log-normal might be reasonable!

A Model for the Data – Log-Normal distribution

- We observe n observations of ages, $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$.
- A normal distribution doesn't seem like a reasonable model since age is strictly positive and the distribution is somewhat right-skewed.
- But a log-normal might be reasonable!
- We assume that each $Y_i \sim \text{Log-Normal}(\mu, \sigma^2)$, and that each Y_i is independently and identically distributed.

A Model for the Data – Log-Normal distribution

- We observe n observations of ages, $\mathbf{Y} = \{Y_1, \dots, Y_n\}$.
- A normal distribution doesn't seem like a reasonable model since age is strictly positive and the distribution is somewhat right-skewed.
- But a log-normal might be reasonable!
- We assume that each $Y_i \sim \text{Log-Normal}(\mu, \sigma^2)$, and that each Y_i is independently and identically distributed. (Later we could extend this model by adding covariates (e.g. $\mu_i = X_i\beta$)).

Example: Age distribution of ER visits due to wall punching

The density of the log-normal distribution is given by

$$f(Y_i|\mu,\sigma^2) = \frac{1}{Y_i\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln(Y_i)-\mu)^2}{2\sigma^2}\right)$$

Example: Age distribution of ER visits due to wall punching

The density of the log-normal distribution is given by

$$f(Y_i|\mu, \sigma^2) = \frac{1}{Y_i \sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln(Y_i) - \mu)^2}{2\sigma^2}\right)$$

Basically the same as saying $\ln(Y_i)$ is normally distributed!

Writing a likelihood

- After writing a probability model for the data, we can write the likelihood of the parameters given the data

Writing a likelihood

- After writing a probability model for the data, we can write the likelihood of the parameters given the data
- By definition of likelihood

$$\mathcal{L}(\mu, \sigma^2|\mathbf{Y}) \propto f(\mathbf{Y}|\mu, \sigma^2)$$

Writing a likelihood

- After writing a probability model for the data, we can write the likelihood of the parameters given the data
- By definition of likelihood

$$\mathcal{L}(\mu, \sigma^2 | \mathbf{Y}) \propto f(\mathbf{Y} | \mu, \sigma^2)$$

- Unfortunately, $f(\mathbf{Y}|\mu, \sigma^2)$ is an *n*-dimensional density, and *n* is huge!

Writing a likelihood

- After writing a probability model for the data, we can write the likelihood of the parameters given the data
- By definition of likelihood

$$\mathcal{L}(\mu, \sigma^2 | \mathbf{Y}) \propto f(\mathbf{Y} | \mu, \sigma^2)$$

- Unfortunately, $f(\mathbf{Y}|\mu, \sigma^2)$ is an *n*-dimensional density, and *n* is huge! How do we simplify this?

Writing a likelihood

- After writing a probability model for the data, we can write the likelihood of the parameters given the data
- By definition of likelihood

$$\mathcal{L}(\mu, \sigma^2 | \mathbf{Y}) \propto f(\mathbf{Y} | \mu, \sigma^2)$$

- Unfortunately, $f(\mathbf{Y}|\mu, \sigma^2)$ is an *n*-dimensional density, and *n* is huge! How do we simplify this? The *i.i.d.* assumption lets us factor the density!

$$\mathcal{L}(\mu, \sigma^2 | \mathbf{Y}) \propto \prod_{i=1}^{N} f(Y_i | \mu, \sigma^2)$$

Writing a likelihood

- Now we can plug in our assumed density for Y.

$$\mathcal{L}(\mu, \sigma^2 | \mathbf{Y}) \propto \prod_{i=1}^{N} \frac{1}{Y_i \sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln(Y_i) - \mu)^2}{2\sigma^2} \right)$$

Writing a likelihood

- Now we can plug in our assumed density for Y.

$$\mathcal{L}(\mu, \sigma^2|\mathbf{Y}) \propto \prod_{i=1}^{N} \frac{1}{Y_i \sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln(Y_i) - \mu)^2}{2\sigma^2}\right)$$

- However, if we tried to calculate this in `R`, the value would be incredibly small!

Writing a likelihood

- Now we can plug in our assumed density for Y.

$$\mathcal{L}(\mu, \sigma^2|\mathbf{Y}) \propto \prod_{i=1}^{N} \frac{1}{Y_i \sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(Y_i) - \mu)^2}{2\sigma^2}\right)$$

- However, if we tried to calculate this in `R`, the value would be incredibly small! It's the product of a bunch of probabilities which are between 0 and 1.

Writing a likelihood

- Now we can plug in our assumed density for Y.

$$\mathcal{L}(\mu, \sigma^2|\mathbf{Y}) \propto \prod_{i=1}^{N} \frac{1}{Y_i \sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln(Y_i) - \mu)^2}{2\sigma^2}\right)$$

- However, if we tried to calculate this in R, the value would be incredibly small! It's the product of a bunch of probabilities which are between 0 and 1. Computers have problems with numbers that small and round them to 0.

Writing a likelihood

- Now we can plug in our assumed density for Y.

$$\mathcal{L}(\mu, \sigma^2|\mathbf{Y}) \propto \prod_{i=1}^{N} \frac{1}{Y_i\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(Y_i) - \mu)^2}{2\sigma^2}\right)$$

- However, if we tried to calculate this in `R`, the value would be incredibly small! It's the product of a bunch of probabilities which are between 0 and 1. Computers have problems with numbers that small and round them to 0.
- It's also often analytically easier to work with sums over products.

Writing a likelihood

- Now we can plug in our assumed density for Y.

$$\mathcal{L}(\mu, \sigma^2|\mathbf{Y}) \propto \prod_{i=1}^{N} \frac{1}{Y_i \sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(Y_i) - \mu)^2}{2\sigma^2}\right)$$

- However, if we tried to calculate this in R, the value would be incredibly small! It's the product of a bunch of probabilities which are between 0 and 1. Computers have problems with numbers that small and round them to 0.
- It's also often analytically easier to work with sums over products.
- This is why we typically work with the log-likelihood (often denoted ℓ).

Writing a likelihood

- Now we can plug in our assumed density for Y.

$$\mathcal{L}(\mu, \sigma^2|\mathbf{Y}) \propto \prod_{i=1}^{N} \frac{1}{Y_i \sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln(Y_i) - \mu)^2}{2\sigma^2}\right)$$

- However, if we tried to calculate this in `R`, the value would be incredibly small! It's the product of a bunch of probabilities which are between 0 and 1. Computers have problems with numbers that small and round them to 0.
- It's also often analytically easier to work with sums over products.
- This is why we typically work with the log-likelihood (often denoted ℓ). Because taking the log is a monotonic transformation, it retains the proportionality!

Deriving the log-likelihood

$$\ell(\mu, \sigma^2|\mathbf{Y}) = \ln\left[\prod_{i=1}^{N} f(Y_i|\mu, \sigma^2)\right]$$

Deriving the log-likelihood

$$\ell(\mu, \sigma^2|\mathbf{Y}) = \ln\left[\prod_{i=1}^{N} f(Y_i|\mu, \sigma^2)\right]$$

$$= \ln\left[\prod_{i=1}^{N} \frac{1}{Y_i\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(Y_i) - \mu)^2}{2\sigma^2}\right)\right]$$

Deriving the log-likelihood

$$\ell(\mu, \sigma^2 | \mathbf{Y}) = \ln \left[\prod_{i=1}^{N} f(Y_i | \mu, \sigma^2) \right]$$

$$= \ln \left[\prod_{i=1}^{N} \frac{1}{Y_i \sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln(Y_i) - \mu)^2}{2\sigma^2} \right) \right]$$

$$= \sum_{i=1}^{N} \ln \left[\frac{1}{Y_i \sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln(Y_i) - \mu)^2}{2\sigma^2} \right) \right]$$

Deriving the log-likelihood

$$\ell(\mu, \sigma^2|\mathbf{Y}) = \ln\left[\prod_{i=1}^{N} f(Y_i|\mu, \sigma^2)\right]$$

$$= \ln\left[\prod_{i=1}^{N} \frac{1}{Y_i\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln(Y_i)-\mu)^2}{2\sigma^2}\right)\right]$$

$$= \sum_{i=1}^{N} \ln\left[\frac{1}{Y_i\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln(Y_i)-\mu)^2}{2\sigma^2}\right)\right]$$

$$= \sum_{i=1}^{N} -\ln(Y_i) - \ln(\sigma) - \ln(\sqrt{2\pi}) + \ln\left[\exp\left(-\frac{(\ln(Y_i)-\mu)^2}{2\sigma^2}\right)\right]$$

Deriving the log-likelihood

$$\ell(\mu, \sigma^2|\mathbf{Y}) = \ln\left[\prod_{i=1}^{N} f(Y_i|\mu,\sigma^2)\right]$$

$$= \ln\left[\prod_{i=1}^{N} \frac{1}{Y_i\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln(Y_i)-\mu)^2}{2\sigma^2}\right)\right]$$

$$= \sum_{i=1}^{N} \ln\left[\frac{1}{Y_i\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln(Y_i)-\mu)^2}{2\sigma^2}\right)\right]$$

$$= \sum_{i=1}^{N} -\ln(Y_i) - \ln(\sigma) - \ln(\sqrt{2\pi}) + \ln\left[\exp\left(-\frac{(\ln(Y_i)-\mu)^2}{2\sigma^2}\right)\right]$$

$$= \sum_{i=1}^{N} -\ln(Y_i) - \ln(\sigma) - \ln(\sqrt{2\pi}) - \frac{(\ln(Y_i)-\mu)^2}{2\sigma^2}$$

Deriving the log-likelihood

$$
\begin{aligned}
\ell(\mu, \sigma^2|\mathbf{Y}) &= \ln\left[\prod_{i=1}^{N} f(Y_i|\mu, \sigma^2)\right] \\
&= \ln\left[\prod_{i=1}^{N} \frac{1}{Y_i\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(Y_i)-\mu)^2}{2\sigma^2}\right)\right] \\
&= \sum_{i=1}^{N} \ln\left[\frac{1}{Y_i\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(Y_i)-\mu)^2}{2\sigma^2}\right)\right] \\
&= \sum_{i=1}^{N} -\ln(Y_i) - \ln(\sigma) - \ln(\sqrt{2\pi}) + \ln\left[\exp\left(-\frac{(\ln(Y_i)-\mu)^2}{2\sigma^2}\right)\right] \\
&= \sum_{i=1}^{N} -\ln(Y_i) - \ln(\sigma) - \ln(\sqrt{2\pi}) - \frac{(\ln(Y_i)-\mu)^2}{2\sigma^2}
\end{aligned}
$$

Deriving the log-likelihood

- To simplify further, we can drop multiplicative constants in the likelihood (additive on the log scale) that are not functions of the the parameters since that retains proportionality.

Deriving the log-likelihood

- To simplify further, we can drop multiplicative constants in the likelihood (additive on the log scale) that are not functions of the the parameters since that retains proportionality.

$$= \sum_{i=1}^{N} - \ln(Y_i) - \ln(\sigma) - \ln(\sqrt{2\pi}) - \frac{(\ln(Y_i) - \mu)^2}{2\sigma^2}$$

Deriving the log-likelihood

- To simplify further, we can drop multiplicative constants in the likelihood (additive on the log scale) that are not functions of the the parameters since that retains proportionality.

$$= \sum_{i=1}^{N} -\ln(Y_i) - \ln(\sigma) - \ln(\sqrt{2\pi}) - \frac{(\ln(Y_i) - \mu)^2}{2\sigma^2}$$

$$\doteq \sum_{i=1}^{N} -\ln(\sigma) - \frac{(\ln(Y_i) - \mu)^2}{2\sigma^2}$$

Deriving the log-likelihood

- To simplify further, we can drop multiplicative constants in the likelihood (additive on the log scale) that are not functions of the the parameters since that retains proportionality.

$$= \sum_{i=1}^{N} -\ln(Y_i) - \ln(\sigma) - \ln(\sqrt{2\pi}) - \frac{(\ln(Y_i) - \mu)^2}{2\sigma^2}$$

$$\doteq \sum_{i=1}^{N} -\ln(\sigma) - \frac{(\ln(Y_i) - \mu)^2}{2\sigma^2}$$

Plotting the log-likelihood

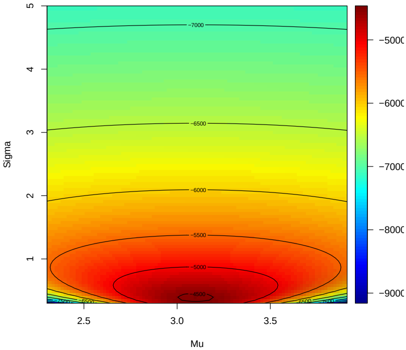


Figure: Contour plot of the log-likelihood for different values of μ and σ

Plotting the likelihood

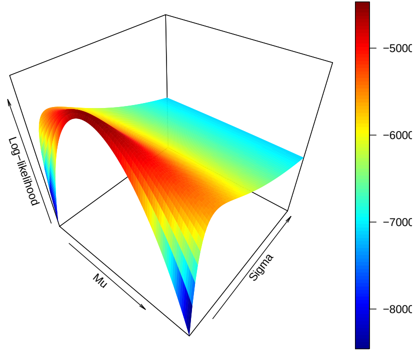


Figure: Plot of the log-likelihood for different values of μ and σ

Plotting the likelihood

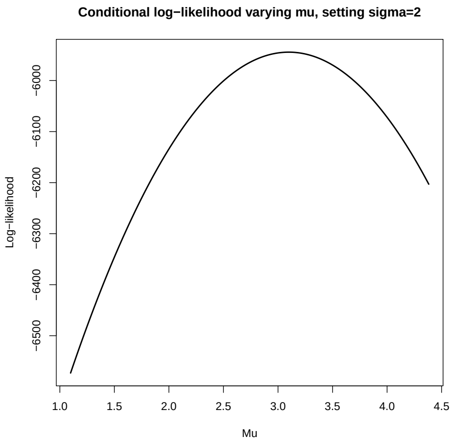


Figure: Plot of the conditional log-likelihood of μ given $\sigma = 2$

Comparing models using likelihood

- Example 1: $\mu = 4$, $\sigma = .2$: Log-likelihood $= -18048.79$

Comparing models using likelihood

- Example 1: $\mu = 4$, $\sigma = .2$: Log-likelihood $= -18048.79$
- Example 2: $\mu = 3.099$, $\sigma = 0.379$: Log-likelihood $= -4461.054$

Comparing models using likelihood

- Example 1: $\mu = 4$, $\sigma = .2$: Log-likelihood $= -18048.79$
- Example 2: $\mu = 3.099$, $\sigma = 0.379$: Log-likelihood $= -4461.054$ (actually the MLE)!

Comparing models using likelihood

- Example 1: $\mu = 4$, $\sigma = .2$: Log-likelihood $= -18048.79$
- Example 2: $\mu = 3.099$, $\sigma = 0.379$: Log-likelihood $= -4461.054$ (actually the MLE)!
- Let's plot the implied distribution of Y_i for each parameter set over the empirical histogram!

Comparing models using likelihood

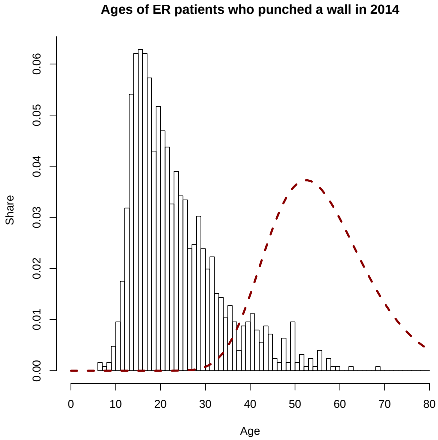


Figure: Empirical distribution of ages vs. log-normal with $\mu = 4$ and $\sigma = .2$

Comparing models using likelihood

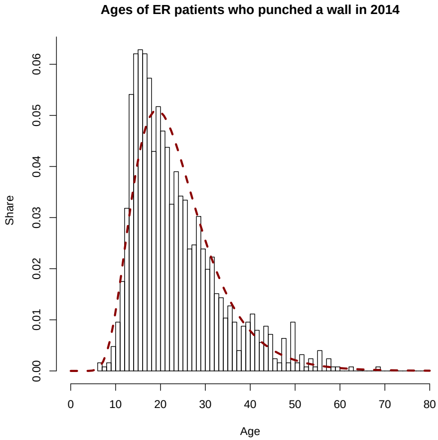


Figure: Empirical distribution of ages vs. log-normal using MLEs of parameters

1. History
2. Likelihood Inference
3. Bayesian Inference
4. Neyman-Pearson
5. Likelihood Example
6. Properties and Tests
7. Simulation
8. Fun With Bayes

1 History

2 Likelihood Inference

3 Bayesian Inference

4 Neyman-Pearson

5 Likelihood Example

6 Properties and Tests

7 Simulation

8 Fun With Bayes

Finite Sample Properties of the MLE

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:
     - Definition: $E(\hat{\theta}) = \theta$

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:
     - ★ Definition: $E(\hat{\theta}) = \theta$
     - ★ Example: $E\left(\bar{Y}\right) =$

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:
     - Definition: $E(\hat{\theta}) = \theta$
     - Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) =$

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:
     - ★ Definition: $E(\hat{\theta}) = \theta$
     - ★ Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) =$

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:
     - ★ Definition: $E(\hat{\theta}) = \theta$
     - ★ Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}n\mu$

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:
     - Definition: $E(\hat{\theta}) = \theta$
     - Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}n\mu = \mu$

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:
     - ★ Definition: $E(\hat{\theta}) = \theta$
     - ★ Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}n\mu = \mu$
   - Minimum variance ("efficiency")

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:
     - ⋆ Definition: $E(\hat{\theta}) = \theta$
     - ⋆ Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}n\mu = \mu$
   - Minimum variance ("efficiency")
     - ⋆ Variance to be minimized: $V(\hat{\theta})$

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
    - Unbiasedness:
        - ★ Definition: $E(\hat{\theta}) = \theta$
        - ★ Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}n\mu = \mu$
    - Minimum variance ("efficiency")
        - ★ Variance to be minimized: $V(\hat{\theta})$
        - ★ Example: $V\left(\bar{Y}\right) =$

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:
     - Definition: $E(\hat{\theta}) = \theta$
     - Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}n\mu = \mu$
   - Minimum variance ("efficiency")
     - Variance to be minimized: $V(\hat{\theta})$
     - Example: $V\left(\bar{Y}\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) =$

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:
     - ★ Definition: $E(\hat{\theta}) = \theta$
     - ★ Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}n\mu = \mu$
   - Minimum variance ("efficiency")
     - ★ Variance to be minimized: $V(\hat{\theta})$
     - ★ Example: $V\left(\bar{Y}\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) =$

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:
     - ★ Definition: $E(\hat{\theta}) = \theta$
     - ★ Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n} n\mu = \mu$
   - Minimum variance ("efficiency")
     - ★ Variance to be minimized: $V(\hat{\theta})$
     - ★ Example: $V\left(\bar{Y}\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2} n\sigma^2 =$

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:
     - Definition: $E(\hat{\theta}) = \theta$
     - Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}n\mu = \mu$
   - Minimum variance ("efficiency")
     - Variance to be minimized: $V(\hat{\theta})$
     - Example: $V\left(\bar{Y}\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2}n\sigma^2 = \sigma^2/n$

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:
     - Definition: $E(\hat{\theta}) = \theta$
     - Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}n\mu = \mu$
   - Minimum variance ("efficiency")
     - Variance to be minimized: $V(\hat{\theta})$
     - Example: $V\left(\bar{Y}\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2}n\sigma^2 = \sigma^2/n$
     - There is a lower bound on the variance of consistent estimators: the Cramer-Rao Lower Bound (CRLB). An MVUE meets that variance.

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:
     - Definition: $E(\hat{\theta}) = \theta$
     - Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n} n\mu = \mu$
   - Minimum variance ("efficiency")
     - Variance to be minimized: $V(\hat{\theta})$
     - Example: $V\left(\bar{Y}\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2} n\sigma^2 = \sigma^2/n$
     - There is a lower bound on the variance of consistent estimators: the Cramer-Rao Lower Bound (CRLB). An MVUE meets that variance.
   - If there is a MVUE, ML will find it (although there may be no unbiased estimator that meets CRLB.)

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
    - Unbiasedness:
        - Definition: $E(\hat{\theta}) = \theta$
        - Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n} n\mu = \mu$
    - Minimum variance ("efficiency")
        - Variance to be minimized: $V(\hat{\theta})$
        - Example: $V\left(\bar{Y}\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2} n\sigma^2 = \sigma^2/n$
        - There is a lower bound on the variance of consistent estimators: the Cramer-Rao Lower Bound (CRLB). An MVUE meets that variance.
    - If there is a MVUE, ML will find it (although there may be no unbiased estimator that meets CRLB.)
    - If there isn't one, ML will still usually find a good estimator

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
    - Unbiasedness:
        - Definition: $E(\hat{\theta}) = \theta$
        - Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}n\mu = \mu$
    - Minimum variance ("efficiency")
        - Variance to be minimized: $V(\hat{\theta})$
        - Example: $V\left(\bar{Y}\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2}n\sigma^2 = \sigma^2/n$
        - There is a lower bound on the variance of consistent estimators: the Cramer-Rao Lower Bound (CRLB). An MVUE meets that variance.
    - If there is a MVUE, ML will find it (although there may be no unbiased estimator that meets CRLB.)
    - If there isn't one, ML will still usually find a good estimator
2. Invariance to Reparameterization

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:
     - Definition: $E(\hat{\theta}) = \theta$
     - Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n} n\mu = \mu$
   - Minimum variance ("efficiency")
     - Variance to be minimized: $V(\hat{\theta})$
     - Example: $V\left(\bar{Y}\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2} n\sigma^2 = \sigma^2/n$
     - There is a lower bound on the variance of consistent estimators: the Cramer-Rao Lower Bound (CRLB). An MVUE meets that variance.
   - If there is a MVUE, ML will find it (although there may be no unbiased estimator that meets CRLB.)
   - If there isn't one, ML will still usually find a good estimator
2. Invariance to Reparameterization
   - Estimate σ with $\hat{\sigma}$ and calculate $\hat{\sigma}^2$ or estimate $\hat{\sigma}^2$: both are MLEs

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:
     - Definition: $E(\hat{\theta}) = \theta$
     - Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}n\mu = \mu$
   - Minimum variance ("efficiency")
     - Variance to be minimized: $V(\hat{\theta})$
     - Example: $V\left(\bar{Y}\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2}n\sigma^2 = \sigma^2/n$
     - There is a lower bound on the variance of consistent estimators: the Cramer-Rao Lower Bound (CRLB). An MVUE meets that variance.
   - If there is a MVUE, ML will find it (although there may be no unbiased estimator that meets CRLB.)
   - If there isn't one, ML will still usually find a good estimator
2. Invariance to Reparameterization
   - Estimate σ with $\hat{\sigma}$ and calculate $\hat{\sigma}^2$ or estimate $\hat{\sigma}^2$: both are MLEs
   - Not true for other methods of inference: e.g. $\bar{y}$ is an unbiased estimate of μ. What is an unbiased estimate of $1/\mu$? $E(1/\bar{y}) \neq 1/E(\bar{y})$.

Finite Sample Properties of the MLE

1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:
     - Definition: $E(\hat{\theta}) = \theta$
     - Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}n\mu = \mu$
   - Minimum variance ("efficiency")
     - Variance to be minimized: $V(\hat{\theta})$
     - Example: $V\left(\bar{Y}\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2}n\sigma^2 = \sigma^2/n$
     - There is a lower bound on the variance of consistent estimators: the Cramer-Rao Lower Bound (CRLB). An MVUE meets that variance.
   - If there is a MVUE, ML will find it (although there may be no unbiased estimator that meets CRLB.)
   - If there isn't one, ML will still usually find a good estimator
2. Invariance to Reparameterization
   - Estimate σ with $\hat{\sigma}$ and calculate $\hat{\sigma}^2$ or estimate $\hat{\sigma}^2$: both are MLEs
   - Not true for other methods of inference: e.g. $\bar{y}$ is an unbiased estimate of μ. What is an unbiased estimate of $1/\mu$? $E(1/\bar{y}) \neq 1/E(\bar{y})$.
3. Invariance to sampling plans

Asymptotic Properties of the MLE

Asymptotic Properties of the MLE

1. Consistency (from the Law of Large Numbers). As $n \to \infty$, the sampling distribution of the MLE collapses to a spike over the parameter value

Asymptotic Properties of the MLE

1. Consistency (from the Law of Large Numbers). As $n \to \infty$, the sampling distribution of the MLE collapses to a spike over the parameter value
2. Asymptotic normality (from the central limit theorem):

Asymptotic Properties of the MLE

1. Consistency (from the Law of Large Numbers). As $n \to \infty$, the sampling distribution of the MLE collapses to a spike over the parameter value
2. Asymptotic normality (from the central limit theorem):
   - As $n \to \infty$, the distribution of MLE/se(MLE) converges to a Normal.

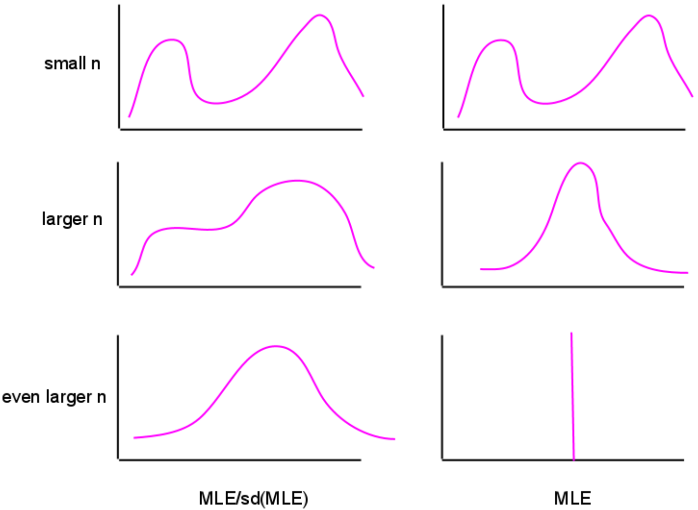
Asymptotic Properties of the MLE

1. Consistency (from the Law of Large Numbers). As $n \to \infty$, the sampling distribution of the MLE collapses to a spike over the parameter value
2. Asymptotic normality (from the central limit theorem):
   - As $n \to \infty$, the distribution of MLE/se(MLE) converges to a Normal.
   - Why do we care? If N is large enough, the asymptotic distribution is a good approximation in finite samples

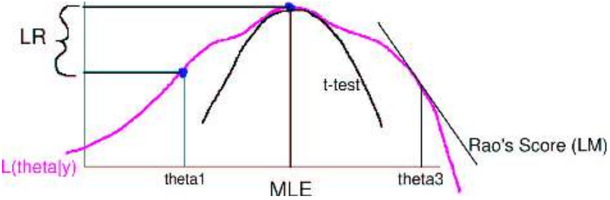
Asymptotic Properties of the MLE

1. Consistency (from the Law of Large Numbers). As $n \to \infty$, the sampling distribution of the MLE collapses to a spike over the parameter value
2. Asymptotic normality (from the central limit theorem):
   - As $n \to \infty$, the distribution of MLE/se(MLE) converges to a Normal.
   - Why do we care? If N is large enough, the asymptotic distribution is a good approximation in finite samples
3. Asymptotic efficiency. The MLE contains as much information as can be packed into a point estimator.

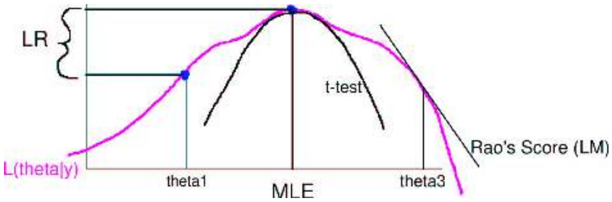
Sampling distributions of the MLE: CLT vs LLN



Uncertainty: Likelihood Ratios for nested models

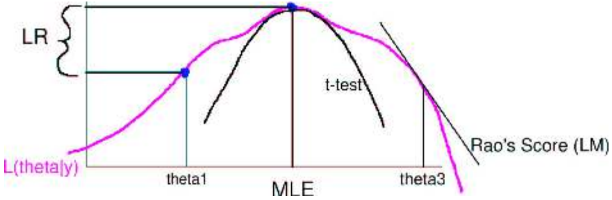


Uncertainty: Likelihood Ratios for nested models



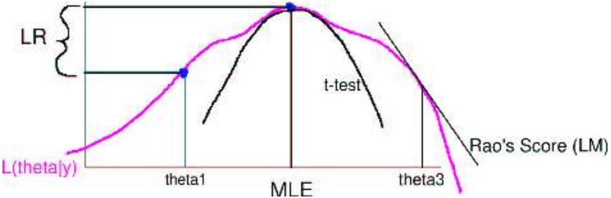
- L^* is the likelihood value for the unrestricted model

Uncertainty: Likelihood Ratios for nested models



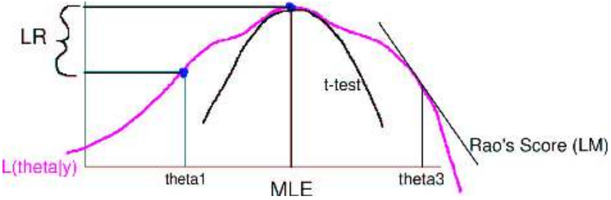
- L^* is the likelihood value for the unrestricted model
- L^*_R is the likelihood value for the (nested) restricted model

Uncertainty: Likelihood Ratios for nested models



- L^* is the likelihood value for the unrestricted model
- L^*_R is the likelihood value for the (nested) restricted model
- $\implies L^* \geq L^*_R \implies \frac{L^*_R}{L^*} \leq 1$

Uncertainty: Likelihood Ratios for nested models



- L^* is the likelihood value for the unrestricted model
- L^*_R is the likelihood value for the (nested) restricted model
- $\implies L^* \geq L^*_R \implies \frac{L^*_R}{L^*} \leq 1$
- This is a direct generalization of F-tests that we learned about in regression.

Meaning of the likelihood ratio

Meaning of the likelihood ratio

- Substantively, its the ratio of 2 traditional probabilities:

Meaning of the likelihood ratio

- Substantively, its the ratio of 2 traditional probabilities:

$$L(\theta_1|y) \propto k(y)\mathbb{P}(y|\theta_1)$$

Meaning of the likelihood ratio

- Substantively, its the ratio of 2 traditional probabilities:

$$L(\theta_1|y) \propto k(y)\mathbb{P}(y|\theta_1)$$
$$L(\theta_2|y) \propto k(y)\mathbb{P}(y|\theta_2)$$

Meaning of the likelihood ratio

- Substantively, its the ratio of 2 traditional probabilities:

$$L(\theta_1|y) \propto k(y)\mathbb{P}(y|\theta_1)$$
$$L(\theta_2|y) \propto k(y)\mathbb{P}(y|\theta_2)$$
$$\frac{L(\theta_1|y)}{L(\theta_2|y)} = \frac{k(y)}{k(y)}\frac{\mathbb{P}(y|\theta_1)}{\mathbb{P}(y|\theta_2)}$$

Meaning of the likelihood ratio

- Substantively, its the ratio of 2 traditional probabilities:

$$
\begin{aligned}
L(\theta_1|y) &\propto k(y)\mathbb{P}(y|\theta_1) \\
L(\theta_2|y) &\propto k(y)\mathbb{P}(y|\theta_2) \\
\frac{L(\theta_1|y)}{L(\theta_2|y)} &= \frac{k(y)}{k(y)}\frac{\mathbb{P}(y|\theta_1)}{\mathbb{P}(y|\theta_2)} \\
&= \frac{\mathbb{P}(y|\theta_1)}{\mathbb{P}(y|\theta_2)}
\end{aligned}
$$

Meaning of the likelihood ratio

- Substantively, its the ratio of 2 traditional probabilities:

$$\begin{aligned} L(\theta_1|y) &\propto k(y)\mathbb{P}(y|\theta_1) \\ L(\theta_2|y) &\propto k(y)\mathbb{P}(y|\theta_2) \\ \frac{L(\theta_1|y)}{L(\theta_2|y)} &= \frac{k(y)}{k(y)}\frac{\mathbb{P}(y|\theta_1)}{\mathbb{P}(y|\theta_2)} \\ &= \frac{\mathbb{P}(y|\theta_1)}{\mathbb{P}(y|\theta_2)} \end{aligned}$$

- Statistically (from the Neyman-Pearson Hypothesis Testing viewpoint), let

Meaning of the likelihood ratio

- Substantively, its the ratio of 2 traditional probabilities:

$$\begin{aligned} L(\theta_1|y) &\propto k(y)\mathbb{P}(y|\theta_1) \\ L(\theta_2|y) &\propto k(y)\mathbb{P}(y|\theta_2) \\ \frac{L(\theta_1|y)}{L(\theta_2|y)} &= \frac{k(y)}{k(y)}\frac{\mathbb{P}(y|\theta_1)}{\mathbb{P}(y|\theta_2)} \\ &= \frac{\mathbb{P}(y|\theta_1)}{\mathbb{P}(y|\theta_2)} \end{aligned}$$

- Statistically (from the Neyman-Pearson Hypothesis Testing viewpoint), let

$$R = -2\ln\left(\frac{L_R^*}{L^*}\right) = 2(\ln L^* - \ln L_R^*)$$

Meaning of the likelihood ratio

- Substantively, its the ratio of 2 traditional probabilities:

$$
\begin{aligned}
L(\theta_1|y) &\propto k(y)\mathbb{P}(y|\theta_1)\\
L(\theta_2|y) &\propto k(y)\mathbb{P}(y|\theta_2)\\
\frac{L(\theta_1|y)}{L(\theta_2|y)} &= \frac{k(y)}{k(y)}\frac{\mathbb{P}(y|\theta_1)}{\mathbb{P}(y|\theta_2)}\\
&= \frac{\mathbb{P}(y|\theta_1)}{\mathbb{P}(y|\theta_2)}
\end{aligned}
$$

- Statistically (from the Neyman-Pearson Hypothesis Testing viewpoint), let

$$
R = -2\ln\left(\frac{L_R^*}{L^*}\right) = 2(\ln L^* - \ln L_R^*)
$$

Then, under the null of no difference between the 2 models,

Meaning of the likelihood ratio

- Substantively, its the ratio of 2 traditional probabilities:

$$
\begin{aligned}
L(\theta_1|y) &\propto k(y)\mathbb{P}(y|\theta_1) \\
L(\theta_2|y) &\propto k(y)\mathbb{P}(y|\theta_2) \\
\frac{L(\theta_1|y)}{L(\theta_2|y)} &= \frac{k(y)}{k(y)}\frac{\mathbb{P}(y|\theta_1)}{\mathbb{P}(y|\theta_2)} \\
&= \frac{\mathbb{P}(y|\theta_1)}{\mathbb{P}(y|\theta_2)}
\end{aligned}
$$

- Statistically (from the Neyman-Pearson Hypothesis Testing viewpoint), let

$$R = -2\ln\left(\frac{L_R^*}{L^*}\right) = 2(\ln L^* - \ln L_R^*)$$

Then, under the null of no difference between the 2 models,

$$R \sim f_{\chi^2}(r|m)$$

Meaning of the likelihood ratio

- Substantively, its the ratio of 2 traditional probabilities:

$$\begin{aligned}
L(\theta_1|y) &\propto k(y)\mathbb{P}(y|\theta_1) \\
L(\theta_2|y) &\propto k(y)\mathbb{P}(y|\theta_2) \\
\frac{L(\theta_1|y)}{L(\theta_2|y)} &= \frac{k(y)}{k(y)}\frac{\mathbb{P}(y|\theta_1)}{\mathbb{P}(y|\theta_2)} \\
&= \frac{\mathbb{P}(y|\theta_1)}{\mathbb{P}(y|\theta_2)}
\end{aligned}$$

- Statistically (from the Neyman-Pearson Hypothesis Testing viewpoint), let

$$R = -2\ln\left(\frac{L_R^*}{L^*}\right) = 2(\ln L^* - \ln L_R^*)$$

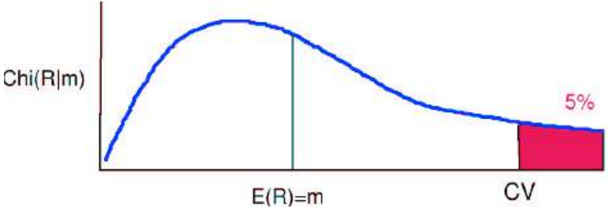
Then, under the null of no difference between the 2 models,

$$R \sim f_{\chi^2}(r|m)$$

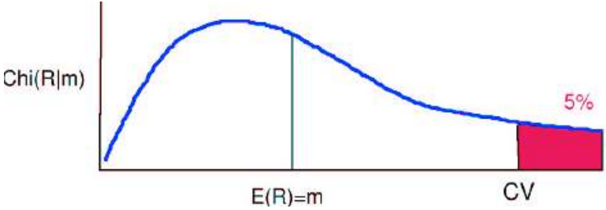
where r is the observed value of R and m is the number of restricted parameters.

Meaning of the likelihood ratio

Meaning of the likelihood ratio

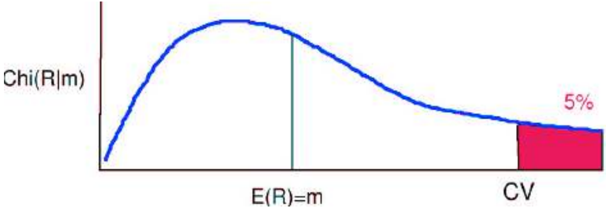


Meaning of the likelihood ratio



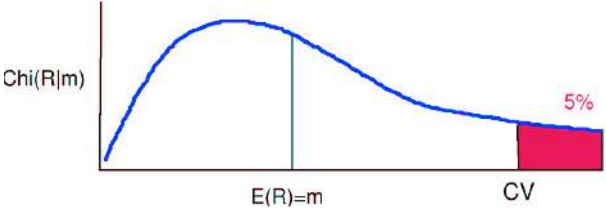
- If restrictions have no effect, $E(R) = m$.

Meaning of the likelihood ratio



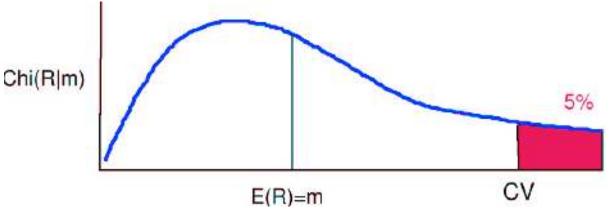
- If restrictions have no effect, $E(R) = m$.
- So only if $r >> m$ will the test parameters be clearly different from zero.

Meaning of the likelihood ratio



- If restrictions have no effect, $E(R) = m$.
- So only if $r >> m$ will the test parameters be clearly different from zero.
- Disadvantage: Too many likelihood ratio tests may be required to test all points of interest

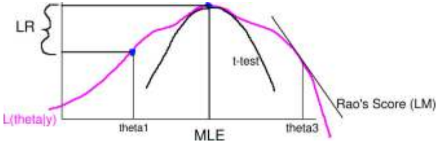
Meaning of the likelihood ratio



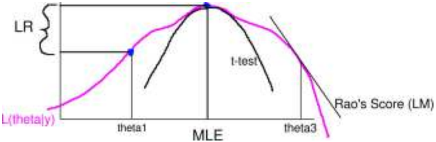
- If restrictions have no effect, $E(R) = m$.
- So only if $r >> m$ will the test parameters be clearly different from zero.
- Disadvantage: Too many likelihood ratio tests may be required to test all points of interest
- Thus, it might be nice to have a summary of uncertainty for every parameter separately $\rightsquigarrow$ standard errors

Standard Errors

Standard Errors

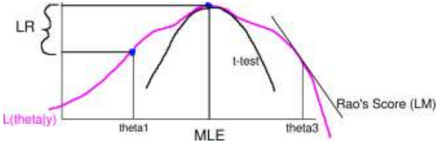


Standard Errors



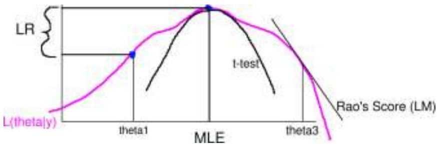
1. Instead of (a) plotting the entire likelihood hyper-surface or (b) computing numerous likelihood ratio tests, we can summarize the all info about the curvature near the maximum with one number

Standard Errors



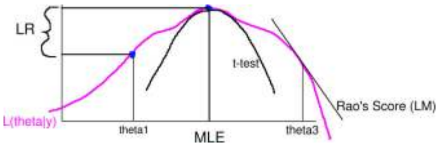
1. Instead of (a) plotting the entire likelihood hyper-surface or (b) computing numerous likelihood ratio tests, we can summarize the all info about the curvature near the maximum with one number
2. We will use the normal likelihood to approximate all likelihoods

Standard Errors



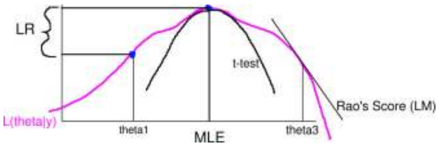
1. Instead of (a) plotting the entire likelihood hyper-surface or (b) computing numerous likelihood ratio tests, we can summarize the all info about the curvature near the maximum with one number
2. We will use the normal likelihood to approximate all likelihoods
3. (one justification) as $n \to \infty$, likelihoods become normal.

Standard Errors



1. Instead of (a) plotting the entire likelihood hyper-surface or (b) computing numerous likelihood ratio tests, we can summarize the all info about the curvature near the maximum with one number
2. We will use the normal likelihood to approximate all likelihoods
3. (one justification) as $n \to \infty$, likelihoods become normal.
4. Reformulate the normal (not stylized) likelihood with $E(Y) = \mu_i = \beta$:

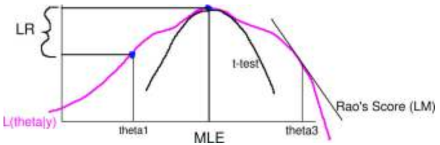
Standard Errors



1. Instead of (a) plotting the entire likelihood hyper-surface or (b) computing numerous likelihood ratio tests, we can summarize the all info about the curvature near the maximum with one number
2. We will use the normal likelihood to approximate all likelihoods
3. (one justification) as $n \to \infty$, likelihoods become normal.
4. Reformulate the normal (not stylized) likelihood with $E(Y) = \mu_i = \beta$:

$$L(\beta|y) \propto \prod_{i=1}^{n} N(y_i|\mu_i, \sigma^2)$$

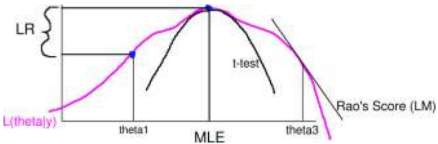
Standard Errors



1. Instead of (a) plotting the entire likelihood hyper-surface or (b) computing numerous likelihood ratio tests, we can summarize the all info about the curvature near the maximum with one number
2. We will use the normal likelihood to approximate all likelihoods
3. (one justification) as $n \to \infty$, likelihoods become normal.
4. Reformulate the normal (not stylized) likelihood with $E(Y) = \mu_i = \beta$:

$$L(\beta|y) \propto \prod_{i=1}^{n} N(y_i|\mu_i, \sigma^2)$$
$$= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right)$$

Standard Errors



1. Instead of (a) plotting the entire likelihood hyper-surface or (b) computing numerous likelihood ratio tests, we can summarize the all info about the curvature near the maximum with one number
2. We will use the normal likelihood to approximate all likelihoods
3. (one justification) as $n \to \infty$, likelihoods become normal.
4. Reformulate the normal (not stylized) likelihood with $E(Y) = \mu_i = \beta$:

$$\begin{aligned} L(\beta|y) &\propto \prod_{i=1}^{n} N(y_i|\mu_i, \sigma^2) \\ &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right) \\ &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(\frac{-(y_i - \beta)^2}{2\sigma^2}\right) \end{aligned}$$

Justifying Standard Errors

$$\ln L(\beta|y) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta)^2$$

Justifying Standard Errors

$$\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\beta + \beta^2)
\end{aligned}$$

Justifying Standard Errors

$$
\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\beta + \beta^2) \\
&= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2
\end{aligned}
$$

Justifying Standard Errors

$$
\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\beta + \beta^2) \\
&= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2 \\
&= a + b\beta + c\beta^2, \qquad \text{A quadratic equation}
\end{aligned}
$$

Justifying Standard Errors

$$
\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2-2y_i\beta+\beta^2) \\
&= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2 \\
&= a + b\beta + c\beta^2, \qquad \text{A quadratic equation}
\end{aligned}
$$

5. $\left(\frac{-n}{2\sigma^2}\right)$ is the degree of curvature. Curvature is larger when:

Justifying Standard Errors

$$
\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\beta + \beta^2) \\
&= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2 \\
&= a + b\beta + c\beta^2, \qquad \text{A quadratic equation}
\end{aligned}
$$

5. $\left(\frac{-n}{2\sigma^2}\right)$ is the degree of curvature. Curvature is larger when:
   - n is large

Justifying Standard Errors

$$
\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\beta + \beta^2) \\
&= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2 \\
&= a + b\beta + c\beta^2, \qquad \text{A quadratic equation}
\end{aligned}
$$

5. $\left(\frac{-n}{2\sigma^2}\right)$ is the degree of curvature. Curvature is larger when:
   - n is large
   - σ^2 is small

Justifying Standard Errors

$$
\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\beta + \beta^2) \\
&= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2 \\
&= a + b\beta + c\beta^2, \qquad \text{A quadratic equation}
\end{aligned}
$$

5. $\left(\frac{-n}{2\sigma^2}\right)$ is the degree of curvature. Curvature is larger when:
   - n is large
   - σ^2 is small
6. For normal likelihood, $\left(\frac{-n}{2\sigma^2}\right)$ is a summary. The bigger the (negative) number...

Justifying Standard Errors

$$\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\beta + \beta^2) \\
&= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2 \\
&= a + b\beta + c\beta^2, \qquad \text{A quadratic equation}
\end{aligned}$$

5. $\left(\frac{-n}{2\sigma^2}\right)$ is the degree of curvature. Curvature is larger when:
   - n is large
   - σ^2 is small
6. For normal likelihood, $\left(\frac{-n}{2\sigma^2}\right)$ is a summary. The bigger the (negative) number...
   - the better

Justifying Standard Errors

$$
\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\beta + \beta^2) \\
&= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2 \\
&= a + b\beta + c\beta^2, \qquad \text{A quadratic equation}
\end{aligned}
$$

5. $\left(\frac{-n}{2\sigma^2}\right)$ is the degree of curvature. Curvature is larger when:
   - n is large
   - σ^2 is small
6. For normal likelihood, $\left(\frac{-n}{2\sigma^2}\right)$ is a summary. The bigger the (negative) number...
   - the better
   - the more information exists in the MLE

Justifying Standard Errors

$$\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\beta + \beta^2) \\
&= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2 \\
&= a + b\beta + c\beta^2, \qquad \text{A quadratic equation}
\end{aligned}$$

5. $\left(\frac{-n}{2\sigma^2}\right)$ is the degree of curvature. Curvature is larger when:
   - n is large
   - σ^2 is small
6. For normal likelihood, $\left(\frac{-n}{2\sigma^2}\right)$ is a summary. The bigger the (negative) number...
   - the better
   - the more information exists in the MLE
   - the larger the likelihood ratio would be in comparing the MLE with any other parameter value.

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

$$\frac{\partial^2 \ln L(\beta|y)}{\partial\beta\partial\beta'} = \frac{-n}{\sigma^2}$$

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

$$\frac{\partial^2 \ln L(\beta|y)}{\partial\beta\partial\beta'} = \frac{-n}{\sigma^2}$$

More generally, this second derivative will give us a way to compute the coefficient on the squared term.

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

$$\frac{\partial^2 \ln L(\beta|y)}{\partial\beta\partial\beta'} = \frac{-n}{\sigma^2}$$

   More generally, this second derivative will give us a way to compute the coefficient on the squared term.
8. We invert the curvature to provide a statistical interpretation:

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

$$\frac{\partial^2 \ln L(\beta|y)}{\partial\beta\partial\beta'} = \frac{-n}{\sigma^2}$$

   More generally, this second derivative will give us a way to compute the coefficient on the squared term.
8. We invert the curvature to provide a statistical interpretation:

$$\hat{V}(\hat{\theta}) = \left[-\frac{\partial^2 \ln L(\theta|y)}{\partial\theta\partial\theta'}\right]^{-1}_{\theta=\hat{\theta}} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \dots \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

$$\frac{\partial^2 \ln L(\beta|y)}{\partial\beta\partial\beta'} = \frac{-n}{\sigma^2}$$

   More generally, this second derivative will give us a way to compute the coefficient on the squared term.

8. We invert the curvature to provide a statistical interpretation:

$$\hat{V}(\hat{\theta}) = \left[-\frac{\partial^2 \ln L(\theta|y)}{\partial\theta\partial\theta'}\right]^{-1}_{\theta=\hat{\theta}} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \cdots \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

   - Statistical interpretation: variance and covariance across repeated samples

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

$$\frac{\partial^2 \ln L(\beta|y)}{\partial\beta\partial\beta'} = \frac{-n}{\sigma^2}$$

   More generally, this second derivative will give us a way to compute the coefficient on the squared term.

8. We invert the curvature to provide a statistical interpretation:

$$\hat{V}(\hat{\theta}) = \left[-\frac{\partial^2 \ln L(\theta|y)}{\partial\theta\partial\theta'}\right]^{-1}_{\theta=\hat{\theta}} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \cdots \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

   - Statistical interpretation: variance and covariance across repeated samples
   - Works in general for a k-dimensional θ vector

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

$$\frac{\partial^2 \ln L(\beta|y)}{\partial\beta\partial\beta'} = \frac{-n}{\sigma^2}$$

   More generally, this second derivative will give us a way to compute the coefficient on the squared term.

8. We invert the curvature to provide a statistical interpretation:

$$\hat{V}(\hat{\theta}) = \left[-\frac{\partial^2 \ln L(\theta|y)}{\partial\theta\partial\theta'}\right]^{-1}_{\theta=\hat{\theta}} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \cdots \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

   - Statistical interpretation: variance and covariance across repeated samples
   - Works in general for a k-dimensional θ vector
   - Can be computed numerically

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

$$\frac{\partial^2 \ln L(\beta|y)}{\partial\beta\partial\beta'} = \frac{-n}{\sigma^2}$$

   More generally, this second derivative will give us a way to compute the coefficient on the squared term.

8. We invert the curvature to provide a statistical interpretation:

$$\hat{V}(\hat{\theta}) = \left[-\frac{\partial^2 \ln L(\theta|y)}{\partial\theta\partial\theta'}\right]^{-1}_{\theta=\hat{\theta}} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \dots \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

   - Statistical interpretation: variance and covariance across repeated samples
   - Works in general for a k-dimensional θ vector
   - Can be computed numerically
   - Known as the variance matrix, or variance-covariance matrix, or covariance matrix

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

$$\frac{\partial^2 \ln L(\beta|y)}{\partial\beta\partial\beta'} = \frac{-n}{\sigma^2}$$

   More generally, this second derivative will give us a way to compute the coefficient on the squared term.
8. We invert the curvature to provide a statistical interpretation:

$$\hat{V}(\hat{\theta}) = \left[-\frac{\partial^2 \ln L(\theta|y)}{\partial\theta\partial\theta'}\right]^{-1}_{\theta=\hat{\theta}} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \cdots \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

   - Statistical interpretation: variance and covariance across repeated samples
   - Works in general for a k-dimensional θ vector
   - Can be computed numerically
   - Known as the variance matrix, or variance-covariance matrix, or covariance matrix
9. This is an estimate of a quadratic approximation to the log-likelihood.

MLE Under Misspecification

MLE Under Misspecification

- When the model is correct, MLE is asymptotically the best estimator (asymptotically: consistent, unbiased, efficient)

MLE Under Misspecification

- When the model is correct, MLE is asymptotically the best estimator (asymptotically: consistent, unbiased, efficient)
- Can we say what happens when the model is wrong? i.e. what happens if we estimate $f(Y|\theta)$ but the true DGP is $g(Y|\theta)$

MLE Under Misspecification

- When the model is correct, MLE is asymptotically the best estimator (asymptotically: consistent, unbiased, efficient)
- Can we say what happens when the model is wrong? i.e. what happens if we estimate $f(Y|\theta)$ but the true DGP is $g(Y|\theta)$
- Our MLE is inconsistent $\text{plim}_{n\rightarrow\infty}\hat{\theta} = \theta^* \neq \theta$

MLE Under Misspecification

- When the model is correct, MLE is asymptotically the best estimator (asymptotically: consistent, unbiased, efficient)
- Can we say what happens when the model is wrong? i.e. what happens if we estimate $f(Y|\theta)$ but the true DGP is $g(Y|\theta)$
- Our MLE is inconsistent $\text{plim}_{n\to\infty}\hat{\theta} = \theta^* \neq \theta$
- θ^* minimizes the Kullback-Leibler (KL) divergence between f and g defined as:

$$E[\log g(Y|\theta) - \log f(Y|\theta)]$$

MLE Under Misspecification

- When the model is correct, MLE is asymptotically the best estimator (asymptotically: consistent, unbiased, efficient)
- Can we say what happens when the model is wrong? i.e. what happens if we estimate $f(Y|\theta)$ but the true DGP is $g(Y|\theta)$
- Our MLE is inconsistent $\text{plim}_{n\to\infty}\hat{\theta} = \theta^* \neq \theta$
- θ^* minimizes the Kullback-Leibler (KL) divergence between f and g defined as:

$$E[\log g(Y|\theta) - \log f(Y|\theta)]$$

- We call this the quasi-maximum likelihood estimator.

MLE Under Misspecification

- When the model is correct, MLE is asymptotically the best estimator (asymptotically: consistent, unbiased, efficient)
- Can we say what happens when the model is wrong? i.e. what happens if we estimate $f(Y|\theta)$ but the true DGP is $g(Y|\theta)$
- Our MLE is inconsistent $\text{plim}_{n\to\infty}\hat{\theta} = \theta^* \neq \theta$
- θ^* minimizes the Kullback-Leibler (KL) divergence between f and g defined as:

$$E[\log g(Y|\theta) - \log f(Y|\theta)]$$

- We call this the quasi-maximum likelihood estimator.
- In certain settings we can still prove the point estimate is consistent and derive consistent estimators of the sampling variance (heteroskedasticity and serial correlation in normal model, clustering in logit and probit models, overdispersion in GLMs)

1. History
2. Likelihood Inference
3. Bayesian Inference
4. Neyman-Pearson
5. Likelihood Example
6. Properties and Tests
7. Simulation
8. Fun With Bayes

1 History

2 Likelihood Inference

3 Bayesian Inference

4 Neyman-Pearson

5 Likelihood Example

6 Properties and Tests

7 Simulation

8 Fun With Bayes

Simulation for any ML Model

Simulation for any ML Model

- If the model is correct, a consistent point estimate of θ is the MLE, $\hat{\theta}$.

Simulation for <u>any</u> ML Model

- If the model is correct, a consistent point estimate of θ is the MLE, $\hat{\theta}$.
- True variance of the sampling distribution of $\hat{\theta}$: $V(\hat{\theta})$

Simulation for any ML Model

- If the model is correct, a consistent point estimate of θ is the MLE, $\hat{\theta}$.
- True variance of the sampling distribution of $\hat{\theta}$: $V(\hat{\theta})$
- Estimate of $V(\hat{\theta})$: $\hat{V}(\hat{\theta})$, the inverse of the negative of the matrix of second derivatives of $\ln L(\theta|y)$, evaluated at $\hat{\theta}$.

Simulation for any ML Model

- If the model is correct, a consistent point estimate of θ is the MLE, $\hat{\theta}$.
- True variance of the sampling distribution of $\hat{\theta}$: $V(\hat{\theta})$
- Estimate of $V(\hat{\theta})$: $\hat{V}(\hat{\theta})$, the inverse of the negative of the matrix of second derivatives of $\ln L(\theta|y)$, evaluated at $\hat{\theta}$.
- As n gets large,

Simulation for <u>any</u> ML Model

- If the model is correct, a consistent point estimate of θ is the MLE, $\hat{\theta}$.
- True variance of the sampling distribution of $\hat{\theta}$: $V(\hat{\theta})$
- Estimate of $V(\hat{\theta})$: $\hat{V}(\hat{\theta})$, the inverse of the negative of the matrix of second derivatives of $\ln L(\theta|y)$, evaluated at $\hat{\theta}$.
- As n gets large,
  - The standardized sampling distribution of $\hat{\theta}$ becomes normal.

Simulation for <u>any</u> ML Model

- If the model is correct, a consistent point estimate of θ is the MLE, $\hat{\theta}$.
- True variance of the sampling distribution of $\hat{\theta}$: $V(\hat{\theta})$
- Estimate of $V(\hat{\theta})$: $\hat{V}(\hat{\theta})$, the inverse of the negative of the matrix of second derivatives of $\ln L(\theta|y)$, evaluated at $\hat{\theta}$.
- As n gets large,
  - The standardized sampling distribution of $\hat{\theta}$ becomes normal.
  - the quadratic approximation implied (from the second derivative of the log-likelihood) improves

Simulation for any ML Model

- If the model is correct, a consistent point estimate of θ is the MLE, $\hat{\theta}$.
- True variance of the sampling distribution of $\hat{\theta}$: $V(\hat{\theta})$
- Estimate of $V(\hat{\theta})$: $\hat{V}(\hat{\theta})$, the inverse of the negative of the matrix of second derivatives of $\ln L(\theta|y)$, evaluated at $\hat{\theta}$.
- As n gets large,
  - The standardized sampling distribution of $\hat{\theta}$ becomes normal.
  - the quadratic approximation implied (from the second derivative of the log-likelihood) improves
- To simulate θ,

Simulation for any ML Model

- If the model is correct, a consistent point estimate of θ is the MLE, $\hat{\theta}$.
- True variance of the sampling distribution of $\hat{\theta}$: $V(\hat{\theta})$
- Estimate of $V(\hat{\theta})$: $\hat{V}(\hat{\theta})$, the inverse of the negative of the matrix of second derivatives of $\ln L(\theta|y)$, evaluated at $\hat{\theta}$.
- As n gets large,
  - The standardized sampling distribution of $\hat{\theta}$ becomes normal.
  - the quadratic approximation implied (from the second derivative of the log-likelihood) improves
- To simulate θ,
  - we'll draw from the multivariate normal: $\tilde{\theta} \sim N\left(\hat{\theta}, \hat{V}(\hat{\theta})\right)$

Simulation for any ML Model

- If the model is correct, a consistent point estimate of θ is the MLE, $\hat{\theta}$.
- True variance of the sampling distribution of $\hat{\theta}$: $V(\hat{\theta})$
- Estimate of $V(\hat{\theta})$: $\hat{V}(\hat{\theta})$, the inverse of the negative of the matrix of second derivatives of $\ln L(\theta|y)$, evaluated at $\hat{\theta}$.
- As n gets large,
  - The standardized sampling distribution of $\hat{\theta}$ becomes normal.
  - the quadratic approximation implied (from the second derivative of the log-likelihood) improves
- To simulate θ,
  - we'll draw from the multivariate normal: $\tilde{\theta} \sim N\left(\hat{\theta}, \hat{V}(\hat{\theta})\right)$
  - This is an asymptotic approximation and can be wrong sometimes.

Simulation for <u>any</u> ML Model

- If the model is correct, a consistent point estimate of θ is the MLE, $\hat{\theta}$.
- True variance of the sampling distribution of $\hat{\theta}$: $V(\hat{\theta})$
- Estimate of $V(\hat{\theta})$: $\hat{V}(\hat{\theta})$, the inverse of the negative of the matrix of second derivatives of $\ln L(\theta|y)$, evaluated at $\hat{\theta}$.
- As n gets large,
  - The standardized sampling distribution of $\hat{\theta}$ becomes normal.
  - the quadratic approximation implied (from the second derivative of the log-likelihood) improves
- To simulate θ,
  - we'll draw from the multivariate normal: $\tilde{\theta} \sim N\left(\hat{\theta}, \hat{V}(\hat{\theta})\right)$
  - This is an asymptotic approximation and can be wrong sometimes.
  - We'll discuss later how how to improve the approximation.

ML Example: k Parameters, including an Ancillary Parameter, with Simulation to Interpret.

ML Example: k Parameters, including an Ancillary Parameter, with Simulation to Interpret.

Forecasting Presidential Elections.

ML Example: k Parameters, including an Ancillary Parameter, with Simulation to Interpret.

Forecasting Presidential Elections.

The Data

ML Example: k Parameters, including an Ancillary Parameter, with Simulation to Interpret.

Forecasting Presidential Elections.

The Data

i	U.S. state, for $i = 1, \dots, 50$

ML Example: k Parameters, including an Ancillary Parameter, with Simulation to Interpret.

Forecasting Presidential Elections.

The Data

i	U.S. state, for $i = 1, \ldots, 50$
t	election year, for $t = 1948, 1952, \ldots, 2012$

ML Example: k Parameters, including an Ancillary Parameter, with Simulation to Interpret.

Forecasting Presidential Elections.

The Data

i	U.S. state, for $i = 1, \ldots, 50$
t	election year, for $t = 1948, 1952, \ldots, 2012$
y_{it}	Democratic fraction of the two-party vote

ML Example: k Parameters, including an Ancillary Parameter, with Simulation to Interpret.

Forecasting Presidential Elections.

The Data

i	U.S. state, for $i = 1, \ldots, 50$
t	election year, for $t = 1948, 1952, \ldots, 2012$
y_{it}	Democratic fraction of the two-party vote
X_{it}	a list of covariates (economic conditions, polls, home state, etc)

ML Example: k Parameters, including an Ancillary Parameter, with Simulation to Interpret.

Forecasting Presidential Elections.

The Data

i	U.S. state, for $i = 1, \ldots, 50$
t	election year, for $t = 1948, 1952, \ldots, 2012$
y_{it}	Democratic fraction of the two-party vote
X_{it}	a list of covariates (economic conditions, polls, home state, etc)
$X_{i,2016}$	the same covariates as X_{it} but measured in 2016

ML Example: k Parameters, including an Ancillary Parameter, with Simulation to Interpret.

Forecasting Presidential Elections.

The Data

i	U.S. state, for $i = 1, \dots, 50$
t	election year, for $t = 1948, 1952, \dots, 2012$
y_{it}	Democratic fraction of the two-party vote
X_{it}	a list of covariates (economic conditions, polls, home state, etc)
$X_{i,2016}$	the same covariates as X_{it} but measured in 2016
E_i	The number of electoral college votes for each state in 2016

The Model

The Model

1. $Y_{it} \sim N(\mu_{it}, \sigma^2)$.

The Model

1. $Y_{it} \sim N(\mu_{it}, \sigma^2)$.
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant

The Model

1. $Y_{it} \sim N(\mu_{it}, \sigma^2)$.
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$, conditional on X.

The Model

1. $Y_{it} \sim N(\mu_{it}, \sigma^2)$.
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$, conditional on X.

The Likelihood Model for the *i*th observation

The Model

1. $Y_{it} \sim N(\mu_{it}, \sigma^2)$.
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$, conditional on X.

The Likelihood Model for the *i*th observation

$$L(\mu_{it}, \sigma | y_{it}) \propto N(y_{it} | \mu_{it}, \sigma^2)$$

The Model

1. $Y_{it} \sim N(\mu_{it}, \sigma^2)$.
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$, conditional on X.

The Likelihood Model for the *i*th observation

$$\begin{aligned} L(\mu_{it}, \sigma | y_{it}) &\propto N(y_{it} | \mu_{it}, \sigma^2) \\ &= (2\pi\sigma^2)^{-1/2} e^{\frac{-(y_{it}-\mu_{it})^2}{2\sigma^2}} \end{aligned}$$

Likelihood model for all n observations

Likelihood model for all n observations

$$L(\beta, \sigma^2 | y) \propto \prod_{i=1}^{n} \prod_{t=1}^{T} f_N(y_{it} | \mu_{it}, \sigma^2)$$

Likelihood model for all n observations

$$L(\beta, \sigma^2|y) \propto \prod_{i=1}^{n}\prod_{t=1}^{T} f_N(y_{it}|\mu_{it}, \sigma^2)$$

$$\ln L(\beta, \sigma^2|y) \doteq \sum_{i=1}^{n}\sum_{t=1}^{T} \ln f_N(y_{it}|\mu_{it}, \sigma^2)$$

Likelihood model for all n observations

$$\begin{aligned}
L(\beta, \sigma^2 | y) &\propto \prod_{i=1}^{n} \prod_{t=1}^{T} f_N(y_{it} | \mu_{it}, \sigma^2) \\
\ln L(\beta, \sigma^2 | y) &\doteq \sum_{i=1}^{n} \sum_{t=1}^{T} \ln f_N(y_{it} | \mu_{it}, \sigma^2) \\
&= \sum_{i=1}^{n} \sum_{t=1}^{T} \left\{ -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(y_{it} - \mu_{it})^2}{2\sigma^2} \right\}
\end{aligned}$$

Likelihood model for all n observations

$$
\begin{aligned}
L(\beta, \sigma^2 | y) &\propto \prod_{i=1}^{n} \prod_{t=1}^{T} f_N(y_{it} | \mu_{it}, \sigma^2) \\
\ln L(\beta, \sigma^2 | y) &\doteq \sum_{i=1}^{n} \sum_{t=1}^{T} \ln f_N(y_{it} | \mu_{it}, \sigma^2) \\
&= \sum_{i=1}^{n} \sum_{t=1}^{T} \left\{ -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(y_{it} - \mu_{it})^2}{2\sigma^2} \right\} \\
&= \color{red}{\sum_{i=1}^{n} \sum_{t=1}^{T} -\frac{1}{2} \ln(2\pi)} + \sum_{i=1}^{n} \sum_{t=1}^{T} -\frac{1}{2} \left[\ln \sigma^2 + \frac{(y_{it} - \mu_{it})^2}{\sigma^2} \right]
\end{aligned}
$$

Likelihood model for all n observations

$$
\begin{aligned}
L(\beta, \sigma^2 | y) &\propto \prod_{i=1}^{n} \prod_{t=1}^{T} f_N(y_{it} | \mu_{it}, \sigma^2) \\
\ln L(\beta, \sigma^2 | y) &\doteq \sum_{i=1}^{n} \sum_{t=1}^{T} \ln f_N(y_{it} | \mu_{it}, \sigma^2) \\
&= \sum_{i=1}^{n} \sum_{t=1}^{T} \left\{ -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(y_{it} - \mu_{it})^2}{2\sigma^2} \right\} \\
&= \sum_{i=1}^{n} \sum_{t=1}^{T} -\frac{1}{2} \ln(2\pi) + \sum_{i=1}^{n} \sum_{t=1}^{T} -\frac{1}{2} \left[\ln \sigma^2 + \frac{(y_{it} - \mu_{it})^2}{\sigma^2} \right] \\
&\doteq \sum_{i=1}^{n} \sum_{t=1}^{T} -\frac{1}{2} \left[\ln \sigma^2 + \frac{(y_{it} - \mu_{it})^2}{\sigma^2} \right]
\end{aligned}
$$

Likelihood model for all n observations

$$
\begin{aligned}
L(\beta, \sigma^2 | y) &\propto \prod_{i=1}^{n} \prod_{t=1}^{T} f_N(y_{it} | \mu_{it}, \sigma^2) \\
\ln L(\beta, \sigma^2 | y) &\doteq \sum_{i=1}^{n} \sum_{t=1}^{T} \ln f_N(y_{it} | \mu_{it}, \sigma^2) \\
&= \sum_{i=1}^{n} \sum_{t=1}^{T} \left\{ -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(y_{it} - \mu_{it})^2}{2\sigma^2} \right\} \\
&= \sum_{i=1}^{n} \sum_{t=1}^{T} -\frac{1}{2} \ln(2\pi) + \sum_{i=1}^{n} \sum_{t=1}^{T} -\frac{1}{2} \left[\ln \sigma^2 + \frac{(y_{it} - \mu_{it})^2}{\sigma^2} \right] \\
&\doteq \sum_{i=1}^{n} \sum_{t=1}^{T} -\frac{1}{2} \left[\ln \sigma^2 + \frac{(y_{it} - \mu_{it})^2}{\sigma^2} \right] \\
&= \sum_{i=1}^{n} \sum_{t=1}^{T} -\frac{1}{2} \left[\ln \sigma^2 + \frac{(y_{it} - X_{it}\beta)^2}{\sigma^2} \right]
\end{aligned}
$$

Estimation

Estimation

- k: number of explanatory variables

Estimation

- k: number of explanatory variables
- Reparameterize on the unbounded scale; use: $\sigma^2 = e^\gamma$

Estimation

- k: number of explanatory variables
- Reparameterize on the unbounded scale; use: $\sigma^2 = e^\gamma$
- Let $\theta = \{\beta, \gamma\}$, a $k + 2 \times 1$ vector.

Estimation

- k: number of explanatory variables
- Reparameterize on the unbounded scale; use: $\sigma^2 = e^\gamma$
- Let $\theta = \{\beta, \gamma\}$, a $k + 2 \times 1$ vector.
- Maximize the likelihood; save $\hat{\theta} = \{\hat{\beta}, \hat{\gamma}\}$.

Estimation

- k: number of explanatory variables
- Reparameterize on the unbounded scale; use: $\sigma^2 = e^\gamma$
- Let $\theta = \{\beta, \gamma\}$, a $k + 2 \times 1$ vector.
- Maximize the likelihood; save $\hat{\theta} = \{\hat{\beta}, \hat{\gamma}\}$.
- Compute and save $\hat{V}(\hat{\theta})$, which is $k + 2 \times k + 2$

R Code for the Log-Likelihood

R Code for the Log-Likelihood

- Mathematical Form:

$$\ln L(\beta, \sigma^2 | y) = \sum_{i=1}^{n} \sum_{t=1}^{T} -\frac{1}{2} \left[\ln \sigma^2 + \frac{(y_{it} - X_{it}\beta)^2}{\sigma^2} \right]$$

R Code for the Log-Likelihood

- Mathematical Form:

$$\ln L(\beta, \sigma^2 | y) = \sum_{i=1}^{n} \sum_{t=1}^{T} -\frac{1}{2} \left[\ln \sigma^2 + \frac{(y_{it} - X_{it}\beta)^2}{\sigma^2} \right]$$

- An R function:

```
ll.normal <- function(par, X, Y) {
X <- as.matrix(cbind(1, X))
beta <- par[1:ncol(X)]
sigma2 <- exp(par[ncol(X) + 1])
-1/2 * sum( log(sigma2) + ((Y - X %*% beta)^2)/sigma2 )
}
```

R Code for the Log-Likelihood

- Mathematical Form:

$$\ln L(\beta, \sigma^2 | y) = \sum_{i=1}^{n} \sum_{t=1}^{T} -\frac{1}{2} \left[\ln \sigma^2 + \frac{(y_{it} - X_{it}\beta)^2}{\sigma^2} \right]$$

- An R function:

```
ll.normal <- function(par, X, Y) {
X <- as.matrix(cbind(1, X))
beta <- par[1:ncol(X)]
sigma2 <- exp(par[ncol(X) + 1])
-1/2 * sum( log(sigma2) + ((Y - X %*% beta)^2)/sigma2 )
}
```

- Calling it:

```
ll.normal(c(2,1,2,1,33,4,3.2),x,y)
ll.normal(c(2,1,2,1,33,4,3.7),x,y)
ll.normal(c(2,1,2,1,33,4,3.5),x,y)
```

Quantities of Interest

Quantities of Interest

- (Reasons we care about the regression coefficients:)

Quantities of Interest

- (Reasons we care about the regression coefficients: N)

Quantities of Interest

- (Reasons we care about the regression coefficients: No)

Quantities of Interest

- (Reasons we care about the regression coefficients: Non)

Quantities of Interest

- (Reasons we care about the regression coefficients: None)

Quantities of Interest

- (Reasons we care about the regression coefficients: None)
- The posterior distribution of electoral college delegates for the Democrat.

Quantities of Interest

- (Reasons we care about the regression coefficients: None)
- The posterior distribution of electoral college delegates for the Democrat.
- Expected number of electoral college delegates for the Democrat.

Quantities of Interest

- (Reasons we care about the regression coefficients: None)
- The posterior distribution of electoral college delegates for the Democrat.
- Expected number of electoral college delegates for the Democrat.
- Probability that the Democratic candidate gets more than $\sum_{i=1}^{n} E_i/n > 0.5$ proportion of electoral college delegates.

Predictive distribution of electoral college delegates in 2016

Predictive distribution of electoral college delegates in 2016

- Goal: Simulations of E_i in each state

Predictive distribution of electoral college delegates in 2016

- Goal: Simulations of E_i in each state
- Should we allocate E_i using the point estimate $\hat{y}_{i,2016}$ winner in each state?

Predictive distribution of electoral college delegates in 2016

Predictive distribution of electoral college delegates in 2016

- Goal: Simulations of E_i in each state

Predictive distribution of electoral college delegates in 2016

- Goal: Simulations of E_i in each state
- Draw many simulations of $y_{i,2016}$ ($\tilde{y}_{i,2016}$) from its approximate posterior distribution for U.S. state i, $\mathbb{P}(y_{i,2016} | y_{it}, t < 2016; X_{it'}, t' \leq 2016)$, i.e. $\mathbb{P}(\texttt{unknown}|\texttt{data})$. (Details shortly.)

Predictive distribution of electoral college delegates in 2016

- Goal: Simulations of E_i in each state
- Draw many simulations of $y_{i,2016}$ ($\tilde{y}_{i,2016}$) from its approximate posterior distribution for U.S. state i, $\mathbb{P}(y_{i,2016}|y_{it}, t < 2016; X_{it'}, t' \leq 2016)$, i.e. $\mathbb{P}(\texttt{unknown}|\texttt{data})$. (Details shortly.)
- For each simulation of state i, if $y_{i,2016} > 0.5$ the Democrat "wins" $\tilde{E}_i$ electoral college delegates; otherwise, the Democrat gets 0.

Predictive distribution of electoral college delegates in 2016

- Goal: Simulations of E_i in each state
- Draw many simulations of $y_{i,2016}$ ($\tilde{y}_{i,2016}$) from its approximate posterior distribution for U.S. state i, $\mathbb{P}(y_{i,2016}|y_{it}, t < 2016; X_{it'}, t' \leq 2016)$, i.e. $\mathbb{P}(\texttt{unknown}|\texttt{data})$. (Details shortly.)
- For each simulation of state i, if $y_{i,2016} > 0.5$ the Democrat "wins" $\tilde{E}_i$ electoral college delegates; otherwise, the Democrat gets 0.
- Add the number of electoral college delegates the Democrat wins in the entire country by adding simulated winnings from each state.

Predictive distribution of electoral college delegates in 2016

- Goal: Simulations of E_i in each state
- Draw many simulations of $y_{i,2016}$ ($\tilde{y}_{i,2016}$) from its approximate posterior distribution for U.S. state i, $\mathbb{P}(y_{i,2016}|y_{it}, t < 2016; X_{it'}, t' \leq 2016)$, i.e. $\mathbb{P}(\texttt{unknown}|\texttt{data})$. (Details shortly.)
- For each simulation of state i, if $y_{i,2016} > 0.5$ the Democrat "wins" $\tilde{E}_i$ electoral college delegates; otherwise, the Democrat gets 0.
- Add the number of electoral college delegates the Democrat wins in the entire country by adding simulated winnings from each state.
- Repeat Steps 1–3 $M = 1,000$ times, and plot a histogram of the results.

How to draw simulations of $y_{i,2016}$?

How to draw simulations of $y_{i,2016}$?

1. Choose values of explanatory variables. In this case, $X_{i,2016}$

How to draw simulations of $y_{i,2016}$?

1. Choose values of explanatory variables. In this case, $X_{i,2016}$
2. Simulate estimation uncertainty:

How to draw simulations of $y_{i,2016}$?

1. Choose values of explanatory variables. In this case, $X_{i,2016}$
2. Simulate estimation uncertainty:
   - Draw θ from its sampling distribution, $N(\hat{\theta}, \hat{V}(\hat{\theta}))$. Label the random draw $\tilde{\theta} = \{\tilde{\beta}, \tilde{\gamma}\}$.

How to draw simulations of $y_{i,2016}$?

1. Choose values of explanatory variables. In this case, $X_{i,2016}$
2. Simulate estimation uncertainty:
   - Draw θ from its sampling distribution, $N(\hat{\theta}, \hat{V}(\hat{\theta}))$. Label the random draw $\tilde{\theta} = \{\tilde{\beta}, \tilde{\gamma}\}$.
   - Pull out $\tilde{\beta}$ and save.

How to draw simulations of $y_{i,2016}$?

1. Choose values of explanatory variables. In this case, $X_{i,2016}$
2. Simulate estimation uncertainty:
   - Draw θ from its sampling distribution, $N(\hat{\theta}, \hat{V}(\hat{\theta}))$. Label the random draw $\tilde{\theta} = \{\tilde{\beta}, \tilde{\gamma}\}$.
   - Pull out $\tilde{\beta}$ and save.
   - Pull out $\tilde{\gamma}$, "un-reparameterize", and save $\tilde{\sigma}^2 = e^{\tilde{\gamma}}$

How to draw simulations of $y_{i,2016}$?

1. Choose values of explanatory variables. In this case, $X_{i,2016}$
2. Simulate estimation uncertainty:
   - Draw θ from its sampling distribution, $N(\hat{\theta}, \hat{V}(\hat{\theta}))$. Label the random draw $\tilde{\theta} = \{\tilde{\beta}, \tilde{\gamma}\}$.
   - Pull out $\tilde{\beta}$ and save.
   - Pull out $\tilde{\gamma}$, "un-reparameterize", and save $\tilde{\sigma}^2 = e^{\tilde{\gamma}}$
3. Compute the simulated systematic component:

How to draw simulations of $y_{i,2016}$?

1. Choose values of explanatory variables. In this case, $X_{i,2016}$
2. Simulate estimation uncertainty:
   - Draw θ from its sampling distribution, $N(\hat{\theta}, \hat{V}(\hat{\theta}))$. Label the random draw $\tilde{\theta} = \{\tilde{\beta}, \tilde{\gamma}\}$.
   - Pull out $\tilde{\beta}$ and save.
   - Pull out $\tilde{\gamma}$, "un-reparameterize", and save $\tilde{\sigma}^2 = e^{\tilde{\gamma}}$
3. Compute the simulated systematic component: $\tilde{\mu}_{it} = X_{i,2016}\tilde{\beta}$

How to draw simulations of $y_{i,2016}$?

1. Choose values of explanatory variables. In this case, $X_{i,2016}$
2. Simulate estimation uncertainty:
   - Draw θ from its sampling distribution, $N(\hat{\theta}, \hat{V}(\hat{\theta}))$. Label the random draw $\tilde{\theta} = \{\tilde{\beta}, \tilde{\gamma}\}$.
   - Pull out $\tilde{\beta}$ and save.
   - Pull out $\tilde{\gamma}$, "un-reparameterize", and save $\tilde{\sigma}^2 = e^{\tilde{\gamma}}$
3. Compute the simulated systematic component: $\tilde{\mu}_{it} = X_{i,2016}\tilde{\beta}$
4. Add fundamental uncertainty: draw $\tilde{y}_{i,2016} \sim N(\tilde{\mu}_{i,2016}, \tilde{\sigma}^2)$

How to do it with a LS Regression Program

How to do it with a LS Regression Program

1. Run `lm` of y_{it} on X_{it} and get $\hat{\beta}$ and $V(\hat{\beta})$

How to do it with a LS Regression Program

1. Run `lm` of y_{it} on X_{it} and get $\hat{\beta}$ and $V(\hat{\beta})$
2. Draw β randomly from its sampling distribution, $N(\beta|\hat{\beta}, V(\hat{\beta}))$. Label the random draw $\tilde{\beta}$.

How to do it with a LS Regression Program

1. Run `lm` of y_{it} on X_{it} and get $\hat{\beta}$ and $V(\hat{\beta})$
2. Draw β randomly from its sampling distribution, $N(\beta|\hat{\beta}, V(\hat{\beta}))$. Label the random draw $\tilde{\beta}$.
3. Draw σ^2 from its sampling distribution, $1/\chi^2(\hat{\sigma}^2, N-k)$, labeling it $\tilde{\sigma}^2$

How to do it with a LS Regression Program

1. Run `lm` of y_{it} on X_{it} and get $\hat{\beta}$ and $V(\hat{\beta})$
2. Draw β randomly from its sampling distribution, $N(\beta|\hat{\beta}, V(\hat{\beta}))$. Label the random draw $\tilde{\beta}$.
3. Draw σ^2 from its sampling distribution, $1/\chi^2(\hat{\sigma}^2, N-k)$, labeling it $\tilde{\sigma}^2$
4. Either:

How to do it with a LS Regression Program

1. Run `lm` of y_{it} on X_{it} and get $\hat{\beta}$ and $V(\hat{\beta})$
2. Draw β randomly from its sampling distribution, $N(\beta|\hat{\beta}, V(\hat{\beta}))$. Label the random draw $\tilde{\beta}$.
3. Draw σ^2 from its sampling distribution, $1/\chi^2(\hat{\sigma}^2, N-k)$, labeling it $\tilde{\sigma}^2$
4. Either:
   - Draw ϵ_{it} from $N(0, \tilde{\sigma}^2)$, label it $\tilde{\epsilon}_{it}$ and compute: $\tilde{y}_{i,2016} = \tilde{X}_{i,2016}\tilde{\beta} + \tilde{\epsilon}_{it}$

How to do it with a LS Regression Program

1. Run `lm` of y_{it} on X_{it} and get $\hat{\beta}$ and $V(\hat{\beta})$
2. Draw β randomly from its sampling distribution, $N(\beta|\hat{\beta}, V(\hat{\beta}))$. Label the random draw $\tilde{\beta}$.
3. Draw σ^2 from its sampling distribution, $1/\chi^2(\hat{\sigma}^2, N-k)$, labeling it $\tilde{\sigma}^2$
4. Either:
   - Draw ϵ_{it} from $N(0, \tilde{\sigma}^2)$, label it $\tilde{\epsilon}_{it}$ and compute: $\tilde{y}_{i,2016} = \tilde{X}_{i,2016}\tilde{\beta} + \tilde{\epsilon}_{it}$
   - Or, in our preferred notation, draw $\tilde{y}_{i,2016}$ from $N(X_{i,2016}\tilde{\beta}, \tilde{\sigma}^2)$

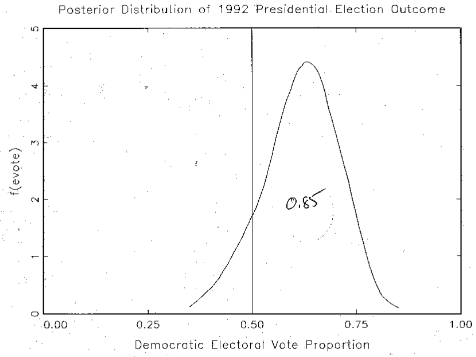
Actual Results for 1992

(calculated before the election by Gelman and King)

(actual results: 69%)

Actual Results for 1992

(calculated before the election by Gelman and King)



(actual results: 69%)

Variance Function Models

Variance Function Models

1. $Y_{it} \sim N(y_{it}|\mu_{it}, \sigma^2_{it})$

Variance Function Models

1. $Y_{it} \sim N(y_{it}|\mu_{it}, \sigma^2_{it})$
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant

Variance Function Models

1. $Y_{it} \sim N(y_{it}|\mu_{it}, \sigma^2_{it})$
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. $\sigma^2_{it} = \exp(z_{it}\gamma)$, where z_{it} is a vector of explanatory variables possibly overlapping x_{it}

Variance Function Models

1. $Y_{it} \sim N(y_{it}|\mu_{it}, \sigma^2_{it})$
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. $\sigma^2_{it} = \exp(z_{it}\gamma)$, where z_{it} is a vector of explanatory variables possibly overlapping x_{it}
4. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$, conditional on X and Z.

Variance Function Models

1. $Y_{it} \sim N(y_{it}|\mu_{it}, \sigma^2_{it})$
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. $\sigma^2_{it} = \exp(z_{it}\gamma)$, where z_{it} is a vector of explanatory variables possibly overlapping x_{it}
4. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$, conditional on X and Z.

The log-likelihood:

Variance Function Models

1. $Y_{it} \sim N(y_{it}|\mu_{it}, \sigma^2_{it})$
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. $\sigma^2_{it} = \exp(z_{it}\gamma)$, where z_{it} is a vector of explanatory variables possibly overlapping x_{it}
4. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$, conditional on X and Z.

The log-likelihood:

$$\ln L(\beta, \sigma^2|y) = \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\left[\ln \sigma^2 + \frac{(y_{it} - X_{it}\beta)^2}{\sigma^2}\right]$$

Variance Function Models

1. $Y_{it} \sim N(y_{it}|\mu_{it}, \sigma^2_{it})$
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. $\sigma^2_{it} = \exp(z_{it}\gamma)$, where z_{it} is a vector of explanatory variables possibly overlapping x_{it}
4. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$, conditional on X and Z.

The log-likelihood:

$$\begin{aligned}
\ln L(\beta, \sigma^2|y) &= \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\left[\ln \sigma^2 + \frac{(y_{it} - X_{it}\beta)^2}{\sigma^2}\right] \\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\left[z_{it}\gamma + \frac{(y_{it} - X_{it}\beta)^2}{\exp(z_{it}\gamma)}\right]
\end{aligned}$$

Variance Function Models

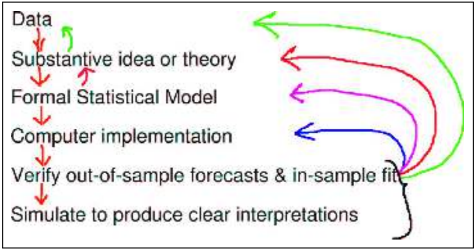
1. $Y_{it} \sim N(y_{it}|\mu_{it}, \sigma^2_{it})$
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. $\sigma^2_{it} = \exp(z_{it}\gamma)$, where z_{it} is a vector of explanatory variables possibly overlapping x_{it}
4. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$, conditional on X and Z.

The log-likelihood:

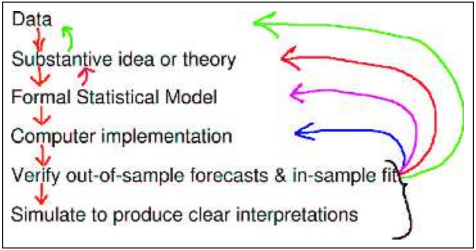
$$\begin{aligned}
\ln L(\beta, \sigma^2|y) &= \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\left[\ln \sigma^2 + \frac{(y_{it} - X_{it}\beta)^2}{\sigma^2}\right] \\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\left[z_{it}\gamma + \frac{(y_{it} - X_{it}\beta)^2}{\exp(z_{it}\gamma)}\right]
\end{aligned}$$

- For what applications would this model be informative?

An Outline of the Research Process

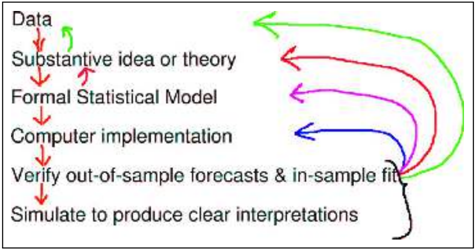


An Outline of the Research Process



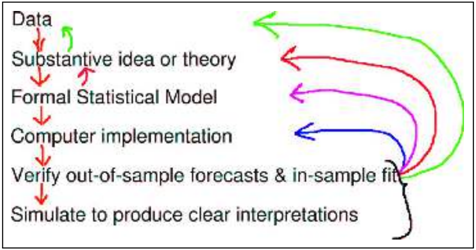
1. These figures are always wild simplifications.

An Outline of the Research Process



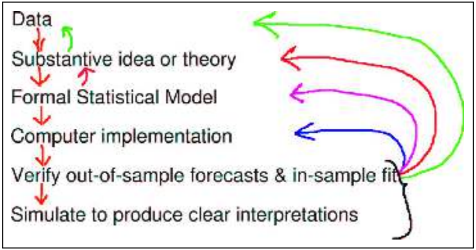
1. These figures are always wild simplifications.
2. Items are roughly in order.

An Outline of the Research Process



1. These figures are always wild simplifications.
2. Items are roughly in order.
3. You can start at any point.

An Outline of the Research Process



1. These figures are always wild simplifications.
2. Items are roughly in order.
3. You can start at any point.
4. Don't miss any parts.

Speculation Time!

- Likelihood dominated the 20th century. If I had to prognosticate, I would guess that empirical Bayes will dominate the 21st century.

Speculation Time!

- Likelihood dominated the 20th century. If I had to prognosticate, I would guess that empirical Bayes will dominate the 21st century.

Speculation Time!

- Likelihood dominated the 20th century. If I had to prognosticate, I would guess that empirical Bayes will dominate the 21st century.
- Empirical Bayes provides us with ways to share information from similar cases.

Speculation Time!

- Likelihood dominated the 20th century. If I had to prognosticate, I would guess that empirical Bayes will dominate the 21st century.
- Empirical Bayes provides us with ways to share information from similar cases.
- The analyst's job becomes to specify what cases are similar.

Speculation Time!

- Likelihood dominated the 20th century. If I had to prognosticate, I would guess that empirical Bayes will dominate the 21st century.
- Empirical Bayes provides us with ways to share information from similar cases.
- The analyst's job becomes to specify what cases are similar.
- Thus empirical bayes utilizes indirect evidence which is often what we have available in an era of bigger and bigger datasets.

Speculation Time!

- Likelihood dominated the 20th century. If I had to prognosticate, I would guess that empirical Bayes will dominate the 21st century.
- Empirical Bayes provides us with ways to share information from similar cases.
- The analyst's job becomes to specify what cases are similar.
- Thus empirical bayes utilizes indirect evidence which is often what we have available in an era of bigger and bigger datasets.
- We will talk about this more in the last couple of weeks.

1. History
2. Likelihood Inference
3. Bayesian Inference
4. Neyman-Pearson
5. Likelihood Example
6. Properties and Tests
7. Simulation
8. Fun With Bayes

1. History
2. Likelihood Inference
3. Bayesian Inference
4. Neyman-Pearson
5. Likelihood Example
6. Properties and Tests
7. Simulation
8. Fun With Bayes

Probability

Probability

Everyone agrees on the *axioms* of probability...

Probability

Everyone agrees on the *axioms* of probability. . .

1. Pr(an event in the event space) is greater than, or equal to, zero.
   - $\Pr(E) \geq 0$

Probability

Everyone agrees on the *axioms* of probability...

1. Pr(an event in the event space) is greater than, or equal to, zero.
   - $\Pr(E) \geq 0$

e.g. Pr(you are pregnant) is at least zero.

Probability

Everyone agrees on the *axioms* of probability...

1. Pr(an event in the event space) is greater than, or equal to, zero.
   - $\Pr(E) \geq 0$

e.g. Pr(you are pregnant) is at least zero.

2. Pr(*an* event in sample space) is 1.

Probability

Everyone agrees on the *axioms* of probability...

1. Pr(an event in the event space) is greater than, or equal to, zero.
   - $\Pr(E) \geq 0$

e.g. Pr(you are pregnant) is at least zero.

2. Pr(*an* event in sample space) is 1.
   - $\Pr(\omega) = 1$

Probability

Everyone agrees on the *axioms* of probability...

1. Pr(an event in the event space) is greater than, or equal to, zero.
   - $\Pr(E) \geq 0$

e.g. Pr(you are pregnant) is at least zero.

2. Pr(*an* event in sample space) is 1.
   - $\Pr(\omega) = 1$

e.g. a coin will come down 'heads' or 'tails'... not 'sausages'

Probability

Everyone agrees on the *axioms* of probability...

1. Pr(an event in the event space) is greater than, or equal to, zero.
   - $\Pr(E) \geq 0$

e.g. Pr(you are pregnant) is at least zero.

2. Pr(*an* event in sample space) is 1.
   - $\Pr(\omega) = 1$

e.g. a coin will come down 'heads' or 'tails'... not 'sausages'

3. sum of the probability of *mutually exclusive* events is equal to the union of the probability of those events:

Probability

Everyone agrees on the *axioms* of probability...

1. Pr(an event in the event space) is greater than, or equal to, zero.
   - $\Pr(E) \geq 0$

e.g. Pr(you are pregnant) is at least zero.

2. Pr(*an* event in sample space) is 1.
   - $\Pr(\omega) = 1$

e.g. a coin will come down 'heads' or 'tails'... not 'sausages'

3. sum of the probability of *mutually exclusive* events is equal to the union of the probability of those events:
   - $\Pr(E_1 \bigcup E_2 \bigcup \ldots) = \sum \Pr(E_i)$

Probability

Everyone agrees on the *axioms* of probability...

1. Pr(an event in the event space) is greater than, or equal to, zero.
   - $\Pr(E) \geq 0$

e.g. Pr(you are pregnant) is at least zero.

2. Pr(*an* event in sample space) is 1.
   - $\Pr(\omega) = 1$

e.g. a coin will come down 'heads' or 'tails'... not 'sausages'

3. sum of the probability of *mutually exclusive* events is equal to the union of the probability of those events:
   - $\Pr(E_1 \bigcup E_2 \bigcup \ldots) = \sum \Pr(E_i)$

e.g. when rolling a die, the probability of a 3 *or* 5 is just $\frac{1}{6} + \frac{1}{6}$

They don't agree on the *nature* of probability.

They don't agree on the *nature* of probability.

1. for frequentists, probability refers to a long run, limiting frequency:

They don't agree on the *nature* of probability.

1. for frequentists, probability refers to a long run, limiting frequency:
   - relative frequency: probability is the number of successes (heads) out of the number of trials (coin flips).

They don't agree on the *nature* of probability.

1. for frequentists, probability refers to a long run, limiting frequency:
    - relative frequency: probability is the number of successes (heads) out of the number of trials (coin flips).
    - that is, $\Pr(X = \text{heads}) \approx \frac{n_x}{n} = \Pr(x)$, where n_x number of trials in which x occurs, n is number of trials.

They don't agree on the *nature* of probability.

1. for frequentists, probability refers to a long run, limiting frequency:
    - relative frequency: probability is the number of successes (heads) out of the number of trials (coin flips).
    - that is, $\Pr(X = \text{heads}) \approx \frac{n_x}{n} = \Pr(x)$, where n_x number of trials in which x occurs, n is number of trials.
    - more controversially: for infinite number of trials, the relative frequency converges to the probability itself

They don't agree on the *nature* of probability.

1. for frequentists, probability refers to a long run, limiting frequency:

    - relative frequency: probability is the number of successes (heads) out of the number of trials (coin flips).
    - that is, $\Pr(X = \text{heads}) \approx \frac{n_x}{n} = \Pr(x)$, where n_x number of trials in which x occurs, n is number of trials.

    - more controversially: for infinite number of trials, the relative frequency converges to the probability itself
    - that is, $\Pr(x) = \lim_{n\to\infty} \frac{n_x}{n}$. (appeal to ∞ is not uncontroversial)

They don't agree on the *nature* of probability.

1. for frequentists, probability refers to a long run, limiting frequency:
   - relative frequency: probability is the number of successes (heads) out of the number of trials (coin flips).
   - that is, $\Pr(X = \text{heads}) \approx \frac{n_x}{n} = \Pr(x)$, where n_x number of trials in which x occurs, n is number of trials.

   - more controversially: for infinite number of trials, the relative frequency converges to the probability itself
   - that is, $\Pr(x) = \lim_{n \to \infty} \frac{n_x}{n}$. (appeal to ∞ is not uncontroversial)
   - we say heads or tails are equally likely because equal proportions are what we observe in (very) large number of trials.

They don't agree on the *nature* of probability.

1. for frequentists, probability refers to a long run, limiting frequency:

   - relative frequency: probability is the number of successes (heads) out of the number of trials (coin flips).
   - that is, $\Pr(X = \text{heads}) \approx \frac{n_x}{n} = \Pr(x)$, where n_x number of trials in which x occurs, n is number of trials.

   - more controversially: for infinite number of trials, the relative frequency converges to the probability itself
   - that is, $\Pr(x) = \lim_{n\to\infty} \frac{n_x}{n}$. (appeal to ∞ is not uncontroversial)
   - we say heads or tails are equally likely because equal proportions are what we observe in (very) large number of trials.

   - "objective", dominant paradigm in statistics, and cheerleaders incl Fischer.

But. . .

But...

1. we could view probability as subjective ...

But. . .

1. we could view probability as subjective . . .

   - probability as a personal belief: it need not be constant across all people at all times (cf. frequentist)

But...

1. we could view probability as subjective ...

    - probability as a personal belief: it need not be constant across all people at all times (cf. frequentist)
    - connected to the idea of a wager:

But...

1. we could view probability as subjective ...
   - probability as a personal belief: it need not be constant across all people at all times (cf. frequentist)
   - connected to the idea of a wager: your willingness to bet (possibly your own money!) on an outcome.

But. . .

1. we could view probability as subjective . . .

    - probability as a personal belief: it need not be constant across all people at all times (cf. frequentist)
    - connected to the idea of a wager: your willingness to bet (possibly your own money!) on an outcome.
    - more objective/axiomatic approaches require that the various beliefs are not contradictory (e.g. transitivity).

But...

1. we could view probability as subjective ...

   - probability as a personal belief: it need not be constant across all people at all times (cf. frequentist)
   - connected to the idea of a wager: your willingness to bet (possibly your own money!) on an outcome.
   - more objective/axiomatic approaches require that the various beliefs are not contradictory (e.g. transitivity).

   - formally capture the belief(s) via a prior: a distribution of probabilities over the possible events.

But...

1. we could view probability as subjective ...

   - probability as a personal belief: it need not be constant across all people at all times (cf. frequentist)
   - connected to the idea of a wager: your willingness to bet (possibly your own money!) on an outcome.
   - more objective/axiomatic approaches require that the various beliefs are not contradictory (e.g. transitivity).

   - formally capture the belief(s) via a prior: a distribution of probabilities over the possible events.
   - idea will be to update beliefs (about parameter values) on observing the data

But...

1. we could view probability as subjective ...

    - probability as a personal belief: it need not be constant across all people at all times (cf. frequentist)
    - connected to the idea of a wager: your willingness to bet (possibly your own money!) on an outcome.
    - more objective/axiomatic approaches require that the various beliefs are not contradictory (e.g. transitivity).

    - formally capture the belief(s) via a prior: a distribution of probabilities over the possible events.
    - idea will be to update beliefs (about parameter values) on observing the data
    - example: our prior over a coin's outcomes (Bernoulli process) might be $p = \frac{1}{2}$ or $p = \frac{1}{3}$ or $p = 1$ ('degenerate')

But...

1. we could view probability as subjective ...

    - probability as a personal belief: it need not be constant across all people at all times (cf. frequentist)
    - connected to the idea of a wager: your willingness to bet (possibly your own money!) on an outcome.
    - more objective/axiomatic approaches require that the various beliefs are not contradictory (e.g. transitivity).

    - formally capture the belief(s) via a prior: a distribution of probabilities over the possible events.
    - idea will be to update beliefs (about parameter values) on observing the data
    - example: our prior over a coin's outcomes (Bernoulli process) might be $p = \frac{1}{2}$ or $p = \frac{1}{3}$ or $p = 1$ ('degenerate')— we can then conduct our trials (the tosses themselves).

But...

1. we could view probability as subjective ...

   - probability as a personal belief: it need not be constant across all people at all times (cf. frequentist)
   - connected to the idea of a wager: your willingness to bet (possibly your own money!) on an outcome.
   - more objective/axiomatic approaches require that the various beliefs are not contradictory (e.g. transitivity).

   - formally capture the belief(s) via a prior: a distribution of probabilities over the possible events.
   - idea will be to update beliefs (about parameter values) on observing the data
   - example: our prior over a coin's outcomes (Bernoulli process) might be $p = \frac{1}{2}$ or $p = \frac{1}{3}$ or $p = 1$ ('degenerate')— we can then conduct our trials (the tosses themselves). Alternatively, we might have a prior on the value of some $\hat{\beta}$ in a regression.

Maximum Likelihood

Maximum Likelihood

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\
&= p(y|\theta)k(y)
\end{aligned}
$$

Maximum Likelihood

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\
&= p(y|\theta)k(y) \\
&\propto p(y|\theta)
\end{aligned}
$$

Maximum Likelihood

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\
&= p(y|\theta)k(y) \\
&\propto p(y|\theta)
\end{aligned}
$$

$$
L(\theta|y) = p(y|\theta)
$$

Maximum Likelihood

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ &= p(y|\theta)k(y) \\ &\propto p(y|\theta) \end{aligned}$$

$$L(\theta|y) = p(y|\theta)$$

There is a fixed, true value of θ,

Maximum Likelihood

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ &= p(y|\theta)k(y) \\ &\propto p(y|\theta) \end{aligned}$$

$$L(\theta|y) = p(y|\theta)$$

There is a fixed, true value of θ, and we maximize the likelihood to estimate θ and make assumptions to generate uncertainty about our estimate of θ.

Bayesian

Bayesian

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

Bayesian

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\
&\propto p(y|\theta)p(\theta)
\end{aligned}
$$

Bayesian

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\
&\propto p(y|\theta)p(\theta)
\end{aligned}
$$

- θ is a random variable.

Bayesian

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\
&\propto p(y|\theta)p(\theta)
\end{aligned}
$$

- θ is a random variable.
  - θ is stochastic and changes from time to time.

Bayesian

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ &\propto p(y|\theta)p(\theta) \end{aligned}$$

- θ is a random variable.
  - θ is stochastic and changes from time to time.
  - θ is truly fixed, but we want to reflect our uncertainty about it.

Bayesian

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$
$$\propto p(y|\theta)p(\theta)$$

- θ is a random variable.
  - θ is stochastic and changes from time to time.
  - θ is truly fixed, but we want to reflect our uncertainty about it.
- We have a prior subjective belief about θ, which we update with the data to form posterior beliefs about θ.

Bayesian

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\
&\propto p(y|\theta)p(\theta)
\end{aligned}
$$

- θ is a random variable.
  - θ is stochastic and changes from time to time.
  - θ is truly fixed, but we want to reflect our uncertainty about it.
- We have a prior subjective belief about θ, which we update with the data to form posterior beliefs about θ.
- The posterior is a probability distribution that must integrate to 1.

Bayesian

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ &\propto p(y|\theta)p(\theta) \end{aligned}$$

- θ is a random variable.
  - θ is stochastic and changes from time to time.
  - θ is truly fixed, but we want to reflect our uncertainty about it.
- We have a prior subjective belief about θ, which we update with the data to form posterior beliefs about θ.
- The posterior is a probability distribution that must integrate to 1.
- The prior is usually a probability distribution that integrates to 1 (proper prior).

θ as Fixed versus as a Random Variable

θ as Fixed versus as a Random Variable

Non-Bayesian approach (θ fixed):

θ as Fixed versus as a Random Variable

Non-Bayesian approach (θ fixed):

- Estimate θ with measures of uncertainty (SE, CIs)

θ as Fixed versus as a Random Variable

Non-Bayesian approach (θ fixed):

- Estimate θ with measures of uncertainty (SE, CIs)
- 95% Confidence Interval: 95% of the time, θ is in the 95% interval that is estimated each time.

θ as Fixed versus as a Random Variable

Non-Bayesian approach (θ fixed):

- Estimate θ with measures of uncertainty (SE, CIs)
- 95% Confidence Interval: 95% of the time, θ is in the 95% interval that is estimated each time.
  - $P(\theta \in 95\% \text{ CI}) =$

θ as Fixed versus as a Random Variable

Non-Bayesian approach (θ fixed):

- Estimate θ with measures of uncertainty (SE, CIs)
- 95% Confidence Interval: 95% of the time, θ is in the 95% interval that is estimated each time.
  - P($\theta \in$ 95% CI) = 0 or 1

θ as Fixed versus as a Random Variable

Non-Bayesian approach (θ fixed):

- Estimate θ with measures of uncertainty (SE, CIs)
- 95% Confidence Interval: 95% of the time, θ is in the 95% interval that is estimated each time.
  - $P(\theta \in 95\% \text{ CI}) =$ 0 or 1
- $P(\theta > 2) =$

θ as Fixed versus as a Random Variable

Non-Bayesian approach (θ fixed):

- Estimate θ with measures of uncertainty (SE, CIs)
- 95% Confidence Interval: 95% of the time, θ is in the 95% interval that is estimated each time.
  - $\text{P}(\theta \in 95\% \text{ CI}) =$ 0 or 1
- $\text{P}(\theta > 2) =$ 0 or 1

θ as Fixed versus as a Random Variable

Non-Bayesian approach (θ fixed):

- Estimate θ with measures of uncertainty (SE, CIs)
- 95% Confidence Interval: 95% of the time, θ is in the 95% interval that is estimated each time.
  - $P(\theta \in 95\%\ CI) =$ 0 or 1
- $P(\theta > 2) =$ 0 or 1

Bayesian approach (θ random):

θ as Fixed versus as a Random Variable

Non-Bayesian approach (θ fixed):

- Estimate θ with measures of uncertainty (SE, CIs)
- 95% Confidence Interval: 95% of the time, θ is in the 95% interval that is estimated each time.
  - P($\theta \in$ 95% CI) = 0 or 1
- P($\theta > 2$) = 0 or 1

Bayesian approach (θ random):

- Find the posterior distribution of θ.

θ as Fixed versus as a Random Variable

Non-Bayesian approach (θ fixed):

- Estimate θ with measures of uncertainty (SE, CIs)
- 95% Confidence Interval: 95% of the time, θ is in the 95% interval that is estimated each time.
  - $P(\theta \in 95\%\ CI) =$ 0 or 1
- $P(\theta > 2) =$ 0 or 1

Bayesian approach (θ random):

- Find the posterior distribution of θ.
- Take quantities of interest from the distribution (posterior mean, posterior SD, posterior credible intervals)

θ as Fixed versus as a Random Variable

Non-Bayesian approach (θ fixed):

- Estimate θ with measures of uncertainty (SE, CIs)
- 95% Confidence Interval: 95% of the time, θ is in the 95% interval that is estimated each time.
  - $\text{P}(\theta \in 95\% \text{ CI}) =$ 0 or 1
- $\text{P}(\theta > 2) =$ 0 or 1

Bayesian approach (θ random):

- Find the posterior distribution of θ.
- Take quantities of interest from the distribution (posterior mean, posterior SD, posterior credible intervals)
- We can make probability statements regarding θ.

θ as Fixed versus as a Random Variable

Non-Bayesian approach (θ fixed):

- Estimate θ with measures of uncertainty (SE, CIs)
- 95% Confidence Interval: 95% of the time, θ is in the 95% interval that is estimated each time.
  - P($\theta \in$ 95% CI) $=$ 0 or 1
- P($\theta > 2$) $=$ 0 or 1

Bayesian approach (θ random):

- Find the posterior distribution of θ.
- Take quantities of interest from the distribution (posterior mean, posterior SD, posterior credible intervals)
- We can make probability statements regarding θ.
  - 95% Credible Interval:

θ as Fixed versus as a Random Variable

Non-Bayesian approach (θ fixed):

- Estimate θ with measures of uncertainty (SE, CIs)
- 95% Confidence Interval: 95% of the time, θ is in the 95% interval that is estimated each time.
  - $\text{P}(\theta \in 95\% \text{ CI}) =$ 0 or 1
- $\text{P}(\theta > 2) =$ 0 or 1

Bayesian approach (θ random):

- Find the posterior distribution of θ.
- Take quantities of interest from the distribution (posterior mean, posterior SD, posterior credible intervals)
- We can make probability statements regarding θ.
  - 95% Credible Interval: $\text{P}(\theta \in 95\% \text{ CI}) = 0.95$

θ as Fixed versus as a Random Variable

Non-Bayesian approach (θ fixed):

- Estimate θ with measures of uncertainty (SE, CIs)
- 95% Confidence Interval: 95% of the time, θ is in the 95% interval that is estimated each time.
  - $\text{P}(\theta \in 95\% \text{ CI}) =$ 0 or 1
- $\text{P}(\theta > 2) =$ 0 or 1

Bayesian approach (θ random):

- Find the posterior distribution of θ.
- Take quantities of interest from the distribution (posterior mean, posterior SD, posterior credible intervals)
- We can make probability statements regarding θ.
  - 95% Credible Interval: $\text{P}(\theta \in 95\% \text{ CI}) = 0.95$
  - $\text{P}(\theta > 2) = (0, 1)$

Critiques

Critiques

$$\text{Posterior} \quad = \quad \text{Evidence} \times \text{Prior}$$

Critiques

$$\text{Posterior} = \text{Evidence} \times \text{Prior}$$

NB: *Bayesians introduce priors that are not justifiable.*

Critiques

$$\text{Posterior} \quad = \quad \text{Evidence} \times \text{Prior}$$

NB: *Bayesians introduce priors that are not justifiable.*

B: Non-Bayesians are just doing Bayesian statistics with uninformative priors, which may be equally unjustifiable.

Critiques

$$\text{Posterior} \quad = \quad \text{Evidence} \times \text{Prior}$$

NB: *Bayesians introduce priors that are not justifiable.*
B: Non-Bayesians are just doing Bayesian statistics with uninformative priors, which may be equally unjustifiable.

NB: *Unjustified Bayesian priors are driving the results.*

Critiques

$$\textcolor{blue}{\text{Posterior}} = \text{Evidence} \times \textcolor{red}{\text{Prior}}$$

NB: *Bayesians introduce priors that are not justifiable.*
B: Non-Bayesians are just doing Bayesian statistics with uninformative priors, which may be equally unjustifiable.

NB: *Unjustified Bayesian priors are driving the results.*
B: Bayesian results $\approx$ non-Bayesian results as n gets larger (the data overwhelm the prior).

Critiques

$$\text{Posterior} = \text{Evidence} \times \text{Prior}$$

NB: *Bayesians introduce priors that are not justifiable.*
B: Non-Bayesians are just doing Bayesian statistics with uninformative priors, which may be equally unjustifiable.

NB: *Unjustified Bayesian priors are driving the results.*
B: Bayesian results $\approx$ non-Bayesian results as n gets larger (the data overwhelm the prior).

NB: *Bayesian is too hard. Why use it?*

Critiques

$$\textcolor{blue}{\text{Posterior}} = \text{Evidence} \times \textcolor{red}{\text{Prior}}$$

NB: *Bayesians introduce priors that are not justifiable.*
B: Non-Bayesians are just doing Bayesian statistics with uninformative priors, which may be equally unjustifiable.

NB: *Unjustified Bayesian priors are driving the results.*
B: Bayesian results $\approx$ non-Bayesian results as n gets larger (the data overwhelm the prior).

NB: *Bayesian is too hard. Why use it?*
B: Bayesian methods allow us to easily estimate models that are too hard to estimate (cannot computationally find the MLE) or unidentified (no unique MLE exists) with non-Bayesian methods.

Critiques

$$\text{Posterior} = \text{Evidence} \times \text{Prior}$$

NB: *Bayesians introduce priors that are not justifiable.*
B: Non-Bayesians are just doing Bayesian statistics with uninformative priors, which may be equally unjustifiable.

NB: *Unjustified Bayesian priors are driving the results.*
B: Bayesian results $\approx$ non-Bayesian results as n gets larger (the data overwhelm the prior).

NB: *Bayesian is too hard. Why use it?*
B: Bayesian methods allow us to easily estimate models that are too hard to estimate (cannot computationally find the MLE) or unidentified (no unique MLE exists) with non-Bayesian methods. Bayesian methods also allow us to incorporate prior/qualitative information into the model.

Running a Model

Running a Model

Non-Bayesian:

Running a Model

Non-Bayesian:

1. Specify a probability model (distribution for Y).

Running a Model

Non-Bayesian:

1. Specify a probability model (distribution for Y).
2. Find MLE $\hat{\theta}$ and measures of uncertainty (SE, CI). Assume $\hat{\theta}$ follows a (multivariate) normal distribution.

Running a Model

Non-Bayesian:

1. Specify a probability model (distribution for Y).
2. Find MLE $\hat{\theta}$ and measures of uncertainty (SE, CI). Assume $\hat{\theta}$ follows a (multivariate) normal distribution.
3. Estimate quantities of interest analytically or via simulation.

Running a Model

Non-Bayesian:

1. Specify a probability model (distribution for Y).
2. Find MLE $\hat{\theta}$ and measures of uncertainty (SE, CI). Assume $\hat{\theta}$ follows a (multivariate) normal distribution.
3. Estimate quantities of interest analytically or via simulation.

Bayesian:

1. Specify a probability model (distribution for Y and priors on θ).

Running a Model

Non-Bayesian:

1. Specify a probability model (distribution for Y).
2. Find MLE $\hat{\theta}$ and measures of uncertainty (SE, CI). Assume $\hat{\theta}$ follows a (multivariate) normal distribution.
3. Estimate quantities of interest analytically or via simulation.

Bayesian:

1. Specify a probability model (distribution for Y and priors on θ).
2. Solve for posterior and summarize it (mean, SD, credible interval, etc.). We can do both analytically or via simulation.

Running a Model

Non-Bayesian:

1. Specify a probability model (distribution for Y).
2. Find MLE $\hat{\theta}$ and measures of uncertainty (SE, CI). Assume $\hat{\theta}$ follows a (multivariate) normal distribution.
3. Estimate quantities of interest analytically or via simulation.

Bayesian:

1. Specify a probability model (distribution for Y and priors on θ).
2. Solve for posterior and summarize it (mean, SD, credible interval, etc.). We can do both analytically or via simulation.
3. Estimate quantities of interest analytically or via simulation.

Running a Model

Non-Bayesian:

1. Specify a probability model (distribution for Y).
2. Find MLE $\hat{\theta}$ and measures of uncertainty (SE, CI). Assume $\hat{\theta}$ follows a (multivariate) normal distribution.
3. Estimate quantities of interest analytically or via simulation.

Bayesian:

1. Specify a probability model (distribution for Y and priors on θ).
2. Solve for posterior and summarize it (mean, SD, credible interval, etc.). We can do both analytically or via simulation.
3. Estimate quantities of interest analytically or via simulation.

There is a Bayesian way to do any non-Bayesian parametric model.

A Simple (Beta-Binomial) Model: The Canonical Example

A Simple (Beta-Binomial) Model: The Canonical Example

You flip a coin 82 times

A Simple (Beta-Binomial) Model: The Canonical Example

You flip a coin 82 times 65 are heads.

A Simple (Beta-Binomial) Model: The Canonical Example

You flip a coin 82 times 65 are heads. Suppose the coin is heads with probability π.

A Simple (Beta-Binomial) Model: The Canonical Example

You flip a coin 82 times 65 are heads. Suppose the coin is heads with probability π. Estimate π.

A Simple (Beta-Binomial) Model: The Canonical Example

You flip a coin 82 times 65 are heads. Suppose the coin is heads with probability π. Estimate π.

We have 82 Bernoulli observations or one observation Y, where

$$Y \sim \text{Binomial}(n, \pi)$$

with $n = 82$.

A Simple (Beta-Binomial) Model: The Canonical Example

You flip a coin 82 times 65 are heads. Suppose the coin is heads with probability π. Estimate π.

We have 82 Bernoulli observations or one observation Y, where

$$Y \sim \text{Binomial}(n, \pi)$$

with $n = 82$.

Assumptions:

A Simple (Beta-Binomial) Model: The Canonical Example

You flip a coin 82 times 65 are heads. Suppose the coin is heads with probability π. Estimate π.

We have 82 Bernoulli observations or one observation Y, where

$$Y \sim \text{Binomial}(n, \pi)$$

with $n = 82$.

Assumptions:

- Each flip is a Bernoulli trial.

A Simple (Beta-Binomial) Model: The Canonical Example

You flip a coin 82 times 65 are heads. Suppose the coin is heads with probability π. Estimate π.

We have 82 Bernoulli observations or one observation Y, where

$$Y \sim \mathrm{Binomial}(n, \pi)$$

with $n = 82$.

Assumptions:

- Each flip is a Bernoulli trial.
- The coin has the same probability of landing heads each flip .

A Simple (Beta-Binomial) Model: The Canonical Example

You flip a coin 82 times 65 are heads. Suppose the coin is heads with probability π. Estimate π.

We have 82 Bernoulli observations or one observation Y, where

$$Y \sim \mathrm{Binomial}(n, \pi)$$

with $n = 82$.

Assumptions:

- Each flip is a Bernoulli trial.
- The coin has the same probability of landing heads each flip .
- The outcomes are independent.

We can use the beta distribution as a prior for π since it has support over [0,1].

We can use the beta distribution as a prior for π since it has support over [0,1].

$$p(\pi|y) \quad \propto \quad p(y|\pi)p(\pi)$$

We can use the beta distribution as a prior for π since it has support over [0,1].

$$\begin{aligned} p(\pi|y) &\propto p(y|\pi)p(\pi) \\ &= \text{Binomial}(n,\pi) \times \text{Beta}(\alpha,\beta) \end{aligned}$$

We can use the beta distribution as a prior for π since it has support over [0,1].

$$\begin{aligned} p(\pi|y) &\propto p(y|\pi)p(\pi) \\ &= \text{Binomial}(n,\pi) \times \text{Beta}(\alpha,\beta) \\ &= \binom{n}{y}\pi^{y}(1-\pi)^{(n-y)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{(\alpha-1)}(1-\pi)^{(\beta-1)} \end{aligned}$$

We can use the beta distribution as a prior for π since it has support over [0,1].

$$\begin{aligned} p(\pi|y) &\propto p(y|\pi)p(\pi) \\ &= \text{Binomial}(n,\pi) \times \text{Beta}(\alpha,\beta) \\ &= \binom{n}{y}\pi^{y}(1-\pi)^{(n-y)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{(\alpha-1)}(1-\pi)^{(\beta-1)} \\ &\propto \pi^{y}(1-\pi)^{(n-y)}\pi^{(\alpha-1)}(1-\pi)^{(\beta-1)} \end{aligned}$$

We can use the beta distribution as a prior for π since it has support over [0,1].

$$\begin{aligned}
p(\pi|y) &\propto p(y|\pi)p(\pi) \\
&= \text{Binomial}(n,\pi) \times \text{Beta}(\alpha,\beta) \\
&= \binom{n}{y}\pi^{y}(1-\pi)^{(n-y)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{(\alpha-1)}(1-\pi)^{(\beta-1)} \\
&\propto \pi^{y}(1-\pi)^{(n-y)}\pi^{(\alpha-1)}(1-\pi)^{(\beta-1)}
\end{aligned}$$

$$p(\pi|y) \propto \pi^{y+\alpha-1}(1-\pi)^{n-y+\beta-1}$$

We can use the beta distribution as a prior for π since it has support over [0,1].

$$\begin{aligned} p(\pi|y) &\propto p(y|\pi)p(\pi) \\ &= \text{Binomial}(n,\pi) \times \text{Beta}(\alpha,\beta) \\ &= \binom{n}{y}\pi^{y}(1-\pi)^{(n-y)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{(\alpha-1)}(1-\pi)^{(\beta-1)} \\ &\propto \pi^{y}(1-\pi)^{(n-y)}\pi^{(\alpha-1)}(1-\pi)^{(\beta-1)} \end{aligned}$$

$$p(\pi|y) \propto \pi^{y+\alpha-1}(1-\pi)^{n-y+\beta-1}$$

The posterior distribution is simply a Beta$(y+\alpha, n-y+\beta)$ distribution.

We can use the beta distribution as a prior for π since it has support over [0,1].

$$
\begin{aligned}
p(\pi|y) &\propto p(y|\pi)p(\pi) \\
&= \text{Binomial}(n, \pi) \times \text{Beta}(\alpha, \beta) \\
&= \binom{n}{y} \pi^y (1-\pi)^{(n-y)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{(\alpha-1)} (1-\pi)^{(\beta-1)} \\
&\propto \pi^y (1-\pi)^{(n-y)} \pi^{(\alpha-1)} (1-\pi)^{(\beta-1)}
\end{aligned}
$$

$$
p(\pi|y) \propto \pi^{y+\alpha-1}(1-\pi)^{n-y+\beta-1}
$$

The posterior distribution is simply a $\text{Beta}(y+\alpha, n-y+\beta)$ distribution. Effectively, our prior is just adding $\alpha-1$ successes and $\beta-1$ failures to the dataset.

We can use the beta distribution as a prior for π since it has support over [0,1].

$$
\begin{aligned}
p(\pi|y) &\propto p(y|\pi)p(\pi) \\
&= \text{Binomial}(n,\pi) \times \text{Beta}(\alpha,\beta) \\
&= \binom{n}{y}\pi^{y}(1-\pi)^{(n-y)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{(\alpha-1)}(1-\pi)^{(\beta-1)} \\
&\propto \pi^{y}(1-\pi)^{(n-y)}\pi^{(\alpha-1)}(1-\pi)^{(\beta-1)}
\end{aligned}
$$

$$
p(\pi|y) \propto \pi^{y+\alpha-1}(1-\pi)^{n-y+\beta-1}
$$

The posterior distribution is simply a Beta$(y+\alpha, n-y+\beta)$ distribution. Effectively, our prior is just adding $\alpha-1$ successes and $\beta-1$ failures to the dataset.

Bayesian priors are just adding pseudo observations to the data.

Since we know the posterior is a Beta$(y + \alpha, n - y + \beta)$ distribution, we can summarize it analytically or via simulation with the following quantities:

Since we know the posterior is a $\text{Beta}(y + \alpha, n - y + \beta)$ distribution, we can summarize it analytically or via simulation with the following quantities:

- posterior mean

Since we know the posterior is a Beta$(y + \alpha, n - y + \beta)$ distribution, we can summarize it analytically or via simulation with the following quantities:

- posterior mean
- posterior standard deviation

Since we know the posterior is a Beta$(y + \alpha, n - y + \beta)$ distribution, we can summarize it analytically or via simulation with the following quantities:

- posterior mean
- posterior standard deviation
- posterior credible intervals (credible sets)

Since we know the posterior is a Beta$(y + \alpha, n - y + \beta)$ distribution, we can summarize it analytically or via simulation with the following quantities:

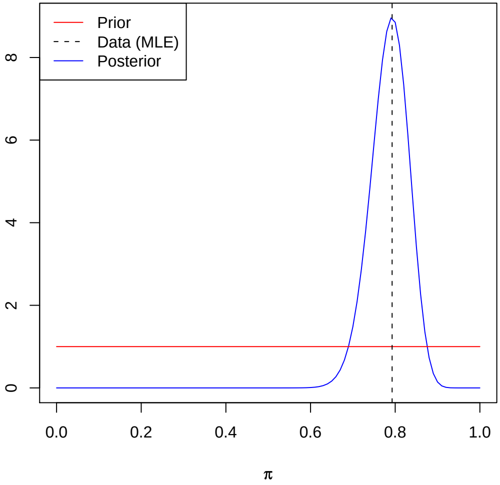
- posterior mean
- posterior standard deviation
- posterior credible intervals (credible sets)
- highest posterior density region

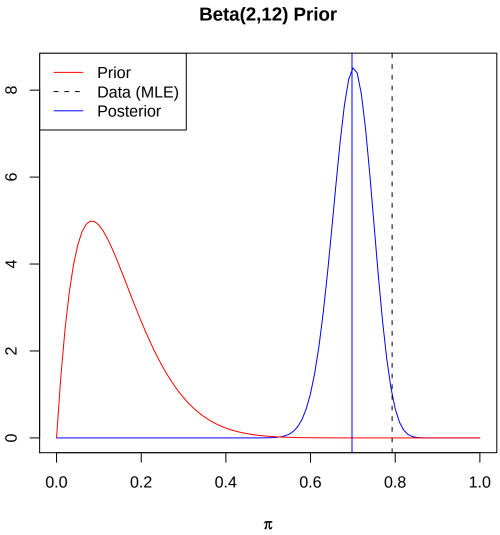
Since we know the posterior is a $\text{Beta}(y + \alpha, n - y + \beta)$ distribution, we can summarize it analytically or via simulation with the following quantities:

- posterior mean
- posterior standard deviation
- posterior credible intervals (credible sets)
- highest posterior density region

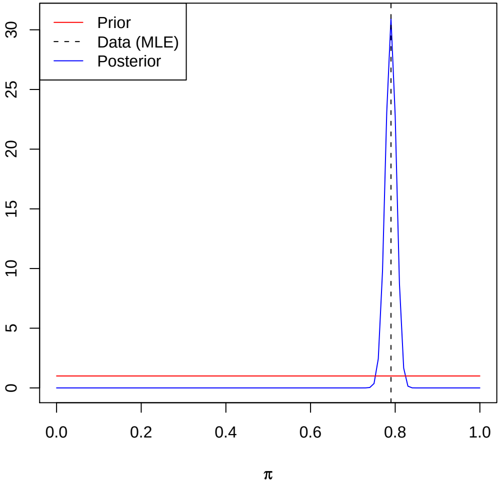
Big Point: Bayesian inference necessitates the estimation of distributions rather than parameters

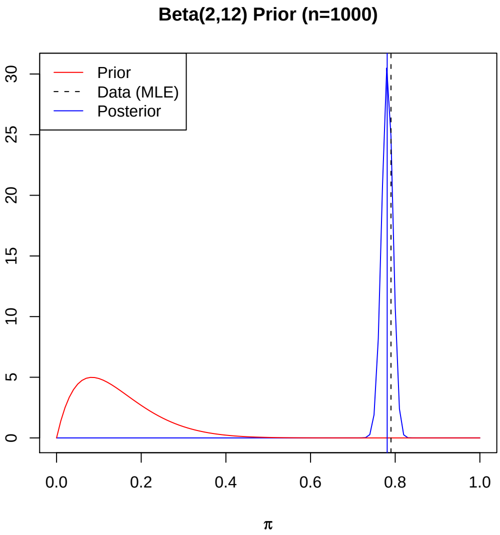
Uninformative Beta(1,1) Prior




Beta(2,12) Prior
Prior
Data (MLE)
Posterior
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
π

Uninformative Beta(1,1) Prior (n=1000)




Beta(2,12) Prior (n=1000)
Prior
Data (MLE)
Posterior
0
5
10
15
20
25
30
0.0
0.2
0.4
0.6
0.8
1.0
π

In the previous model, we had

$$\text{Beta}(y+\alpha, n-y+\beta) \quad = \quad \frac{\text{Binomial}(n,\pi) \times \text{Beta}(\alpha,\beta)}{p(y)}$$

In the previous model, we had

$$\mathrm{Beta}(y+\alpha, n-y+\beta) \quad = \quad \frac{\mathrm{Binomial}(n,\pi) \times \mathrm{Beta}(\alpha,\beta)}{p(y)}$$

We knew that the likelihood $\times$ prior produced something that looked like a Beta distribution up to a constant of proportionality.

In the previous model, we had

$$\mathrm{Beta}(y+\alpha, n-y+\beta) \quad = \quad \frac{\mathrm{Binomial}(n,\pi) \times \mathrm{Beta}(\alpha,\beta)}{p(y)}$$

We knew that the likelihood $\times$ prior produced something that looked like a Beta distribution up to a constant of proportionality.

Since the posterior must be a probability distribution, we know that it is a Beta distribution and we can easily solve for the normalizing constant

In the previous model, we had

$$\mathrm{Beta}(y+\alpha, n-y+\beta) \quad = \quad \frac{\mathrm{Binomial}(n,\pi) \times \mathrm{Beta}(\alpha,\beta)}{p(y)}$$

We knew that the likelihood $\times$ prior produced something that looked like a Beta distribution up to a constant of proportionality.

Since the posterior must be a probability distribution, we know that it is a Beta distribution and we can easily solve for the normalizing constant (although we don't need to since we already have the posterior).

In the previous model, we had

$$\text{Beta}(y+\alpha, n-y+\beta) = \frac{\text{Binomial}(n,\pi)\times \text{Beta}(\alpha,\beta)}{p(y)}$$

We knew that the likelihood $\times$ prior produced something that looked like a Beta distribution up to a constant of proportionality.

Since the posterior must be a probability distribution, we know that it is a Beta distribution and we can easily solve for the normalizing constant (although we don't need to since we already have the posterior).

When the posterior is the same distribution family as the prior, we have **conjugacy**.

In the previous model, we had

$$\mathrm{Beta}(y + \alpha, n - y + \beta) \quad = \quad \frac{\mathrm{Binomial}(n, \pi) \times \mathrm{Beta}(\alpha, \beta)}{p(y)}$$

We knew that the likelihood $\times$ prior produced something that looked like a Beta distribution up to a constant of proportionality.

Since the posterior must be a probability distribution, we know that it is a Beta distribution and we can easily solve for the normalizing constant (although we don't need to since we already have the posterior).

When the posterior is the same distribution family as the prior, we have **conjugacy**.

Conjugate models are great because we can find the exact posterior, but...

The Problem

many real posteriors look like this:

The Problem

many real posteriors look like this:

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\pi}, \boldsymbol{\tau} | \boldsymbol{Y}) \quad \propto \quad \prod_{k=1}^{K} \prod_{s=1}^{S} \frac{\exp(-\frac{\alpha_{ks}}{1/4})}{1/4} \times \frac{\Gamma(\sum_{w=1}^{W} \lambda_w)}{\prod_{w=1}^{W} \Gamma(\lambda_w)} \prod_{w=1}^{W} \theta_{k,w}^{\lambda_w - 1} \times$$

$$\prod_{i=1}^{n} \prod_{t=2005}^{2007} \prod_{s=1}^{S} \left[\beta_s \frac{\Gamma(\sum_{k=1}^{K} \alpha_{ks})}{\prod_{k=1}^{K} \Gamma(\alpha_{ks})} \prod_{k=1}^{K} \pi_{itk}^{\alpha_{ks} - 1} \prod_{j=1}^{D_{it}} \prod_{k=1}^{K} \left[\pi_{itk} \prod_{w=1}^{W} \theta_{kw}^{y_{ijtw}} \right]^{\tau_{ijtk}} \right]^{\sigma_{its}}$$

Example of Useful Prior Information: Girosi and King (2008)

One nice property of Bayesian analysis is that we can incorporate prior information about almost anything given a little work.

Example of Useful Prior Information: Girosi and King (2008)

One nice property of Bayesian analysis is that we can incorporate prior information about almost anything given a little work.

Generally we put prior on the parameter. But what if we don't know about the parameter?

Example of Useful Prior Information: Girosi and King (2008)

One nice property of Bayesian analysis is that we can incorporate prior information about almost anything given a little work.

Generally we put prior on the parameter. But what if we don't know about the parameter?

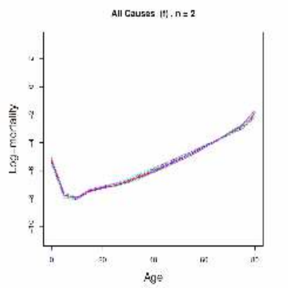
In mortality forecasting we know two key things: (1) mortality looks a lot alike for different causes and (2) it has a distinctive check shape.

Example of Useful Prior Information: Girosi and King (2008)

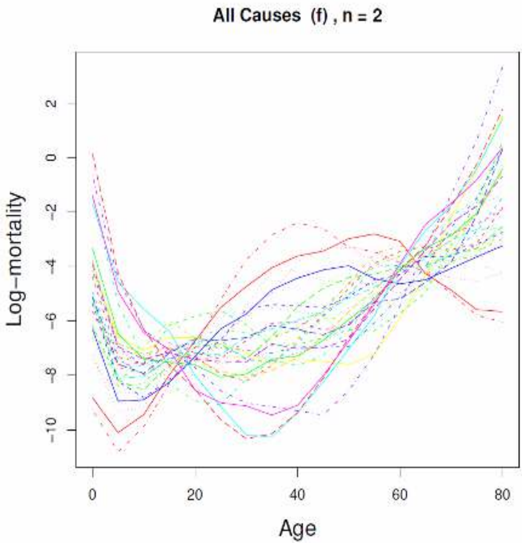
One nice property of Bayesian analysis is that we can incorporate prior information about almost anything given a little work.

Generally we put prior on the parameter. But what if we don't know about the parameter?

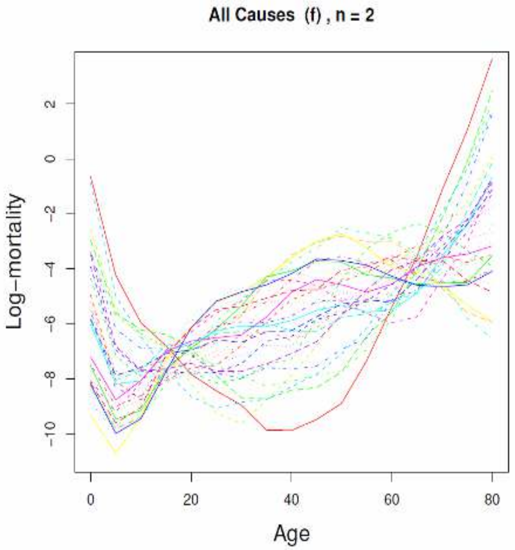
In mortality forecasting we know two key things: (1) mortality looks a lot alike for different causes and (2) it has a distinctive check shape.



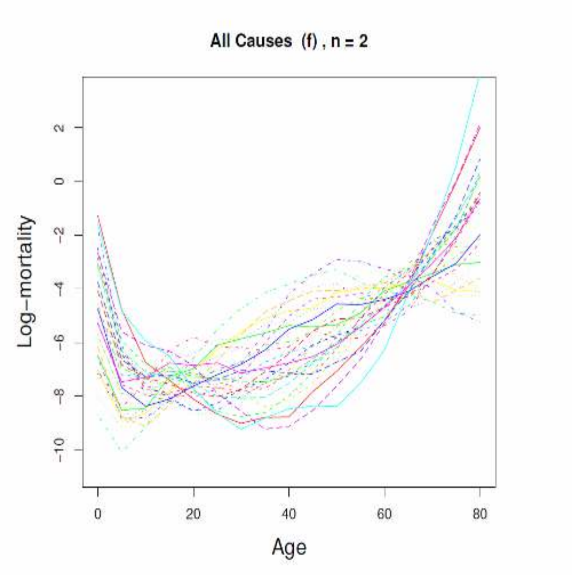
Example: Girosi and King (2008)



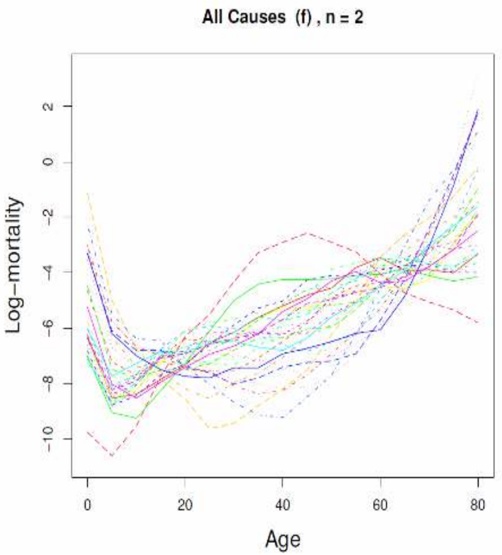
Example: Girosi and King (2008)



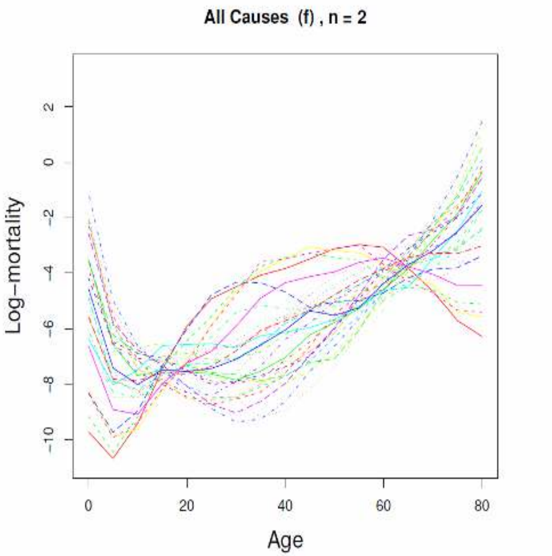
Example: Girosi and King (2008)



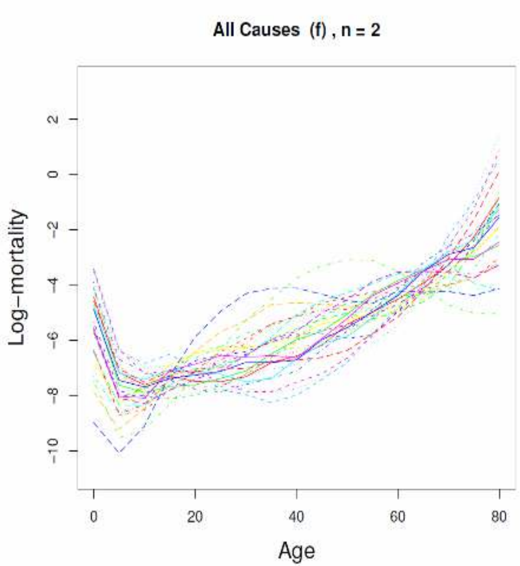
Example: Girosi and King (2008)



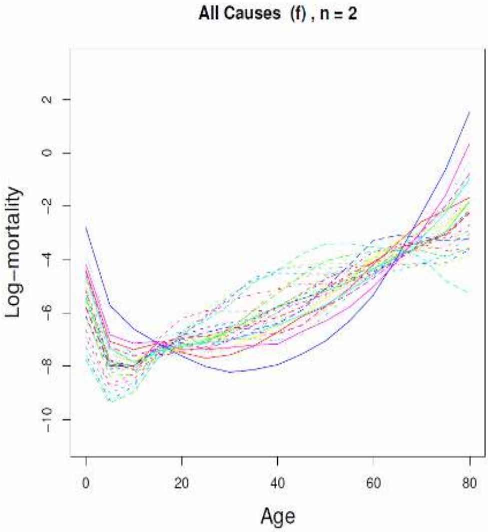
Example: Girosi and King (2008)



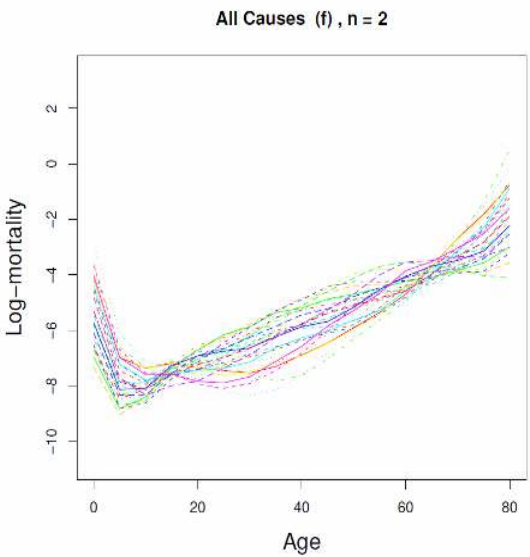
Example: Girosi and King (2008)



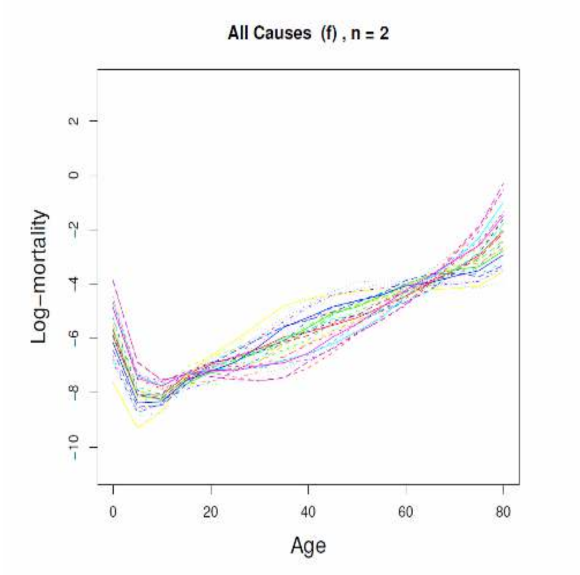
Example: Girosi and King (2008)



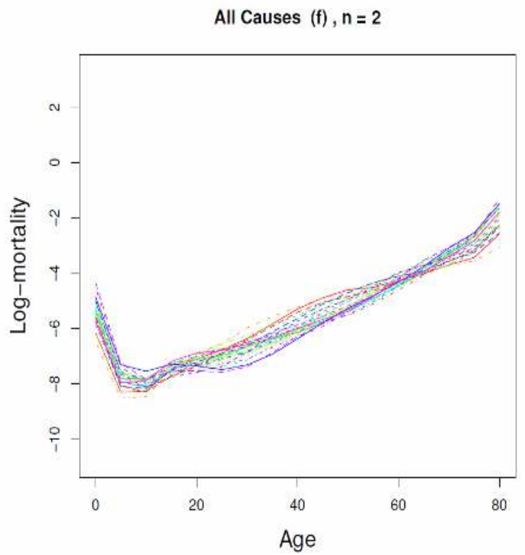
Example: Girosi and King (2008)



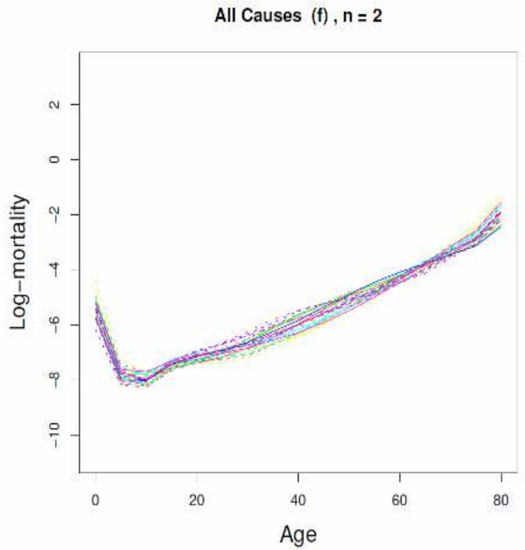
Example: Girosi and King (2008)



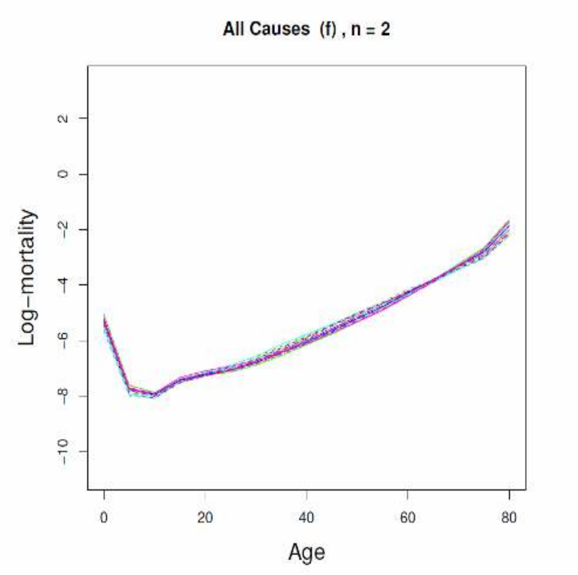
Example: Girosi and King (2008)



Example: Girosi and King (2008)



Example: Girosi and King (2008)



Example: Girosi and King (2008)

This is a tough problem:

Example: Girosi and King (2008)

This is a tough problem:

- Multidimensional Data Structures: 24 causes of death, 17 age groups, 2 sexes, 191 countries, all for 50 annual observations.

Example: Girosi and King (2008)

This is a tough problem:

- Multidimensional Data Structures: 24 causes of death, 17 age groups, 2 sexes, 191 countries, all for 50 annual observations.
- One time series analysis for each of 155,856 cross-sections: with 1 minute to analyze each, one run takes 108 days

Example: Girosi and King (2008)

This is a tough problem:

- Multidimensional Data Structures: 24 causes of death, 17 age groups, 2 sexes, 191 countries, all for 50 annual observations.
- One time series analysis for each of 155,856 cross-sections: with 1 minute to analyze each, one run takes 108 days
- Explanatory variables:

Example: Girosi and King (2008)

This is a tough problem:

- Multidimensional Data Structures: 24 causes of death, 17 age groups, 2 sexes, 191 countries, all for 50 annual observations.
- One time series analysis for each of 155,856 cross-sections: with 1 minute to analyze each, one run takes 108 days
- Explanatory variables:
  - Available in many countries: tobacco consumption, GDP, human capital, trends, fat consumption, total fertility rates, etc.

Example: Girosi and King (2008)

This is a tough problem:

- Multidimensional Data Structures: 24 causes of death, 17 age groups, 2 sexes, 191 countries, all for 50 annual observations.
- One time series analysis for each of 155,856 cross-sections: with 1 minute to analyze each, one run takes 108 days
- Explanatory variables:
  - Available in many countries: tobacco consumption, GDP, human capital, trends, fat consumption, total fertility rates, etc.
  - Numerous variables specific to a cause, age group, sex, and country

Example: Girosi and King (2008)

This is a tough problem:

- Multidimensional Data Structures: 24 causes of death, 17 age groups, 2 sexes, 191 countries, all for 50 annual observations.
- One time series analysis for each of 155,856 cross-sections: with 1 minute to analyze each, one run takes 108 days
- Explanatory variables:
  - Available in many countries: tobacco consumption, GDP, human capital, trends, fat consumption, total fertility rates, etc.
  - Numerous variables specific to a cause, age group, sex, and country
  - Most time series are very short. A majority of countries have only a few isolated annual observations; only 54 countries have at least 20 observations;

Example: Girosi and King (2008)

This is a tough problem:

- Multidimensional Data Structures: 24 causes of death, 17 age groups, 2 sexes, 191 countries, all for 50 annual observations.
- One time series analysis for each of 155,856 cross-sections: with 1 minute to analyze each, one run takes 108 days
- Explanatory variables:
  - Available in many countries: tobacco consumption, GDP, human capital, trends, fat consumption, total fertility rates, etc.
  - Numerous variables specific to a cause, age group, sex, and country
  - Most time series are very short. A majority of countries have only a few isolated annual observations; only 54 countries have at least 20 observations;

All solved using Bayesian Hierarchical Models! (See *Demographic Forecasting* and the *YourCast* package.