# An Introduction to Bayesian Statistics

David Carlson

March 13, 2019

# The Scientific Method

- 1) Ask a question or pose a problem

# The Scientific Method

- 1) Ask a question or pose a problem
- 2) Collect the existing relevant information

# The Scientific Method

- 1) Ask a question or pose a problem
- 2) Collect the existing relevant information
- 3) Based on information, design an investigation or experiment to address the question

# The Scientific Method

- 1) Ask a question or pose a problem
- 2) Collect the existing relevant information
- 3) Based on information, design an investigation or experiment to address the question
- 4) Carry out the investigation or experiment

# The Scientific Method

- 1) Ask a question or pose a problem
- 2) Collect the existing relevant information
- 3) Based on information, design an investigation or experiment to address the question
- 4) Carry out the investigation or experiment
- 5) Use the evidence to update previously known information, then draw conclusions

# The Scientific Method

- 1) Ask a question or pose a problem
- 2) Collect the existing relevant information
- 3) Based on information, design an investigation or experiment to address the question
- 4) Carry out the investigation or experiment
- 5) Use the evidence to update previously known information, then draw conclusions
- 6) Repeat 3—5 as necessary

# Where Does Statistics Fit In?

- Central to steps 2, 3, and 5

# Where Does Statistics Fit In?

- Central to steps 2, 3, and 5
  - Experimental design, survey sampling

# Where Does Statistics Fit In?

- Central to steps 2, 3, and 5
  - Experimental design, survey sampling
  - Statistical inference

# Where Does Statistics Fit In?

- Central to steps 2, 3, and 5
  - Experimental design, survey sampling
  - Statistical inference
- Bayesian statistics is particularly well-suited for step 2 and 5

# Where Does Statistics Fit In?

- Central to steps 2, 3, and 5
  - Experimental design, survey sampling
  - Statistical inference
- Bayesian statistics is particularly well-suited for step 2 and 5
  - Step 2: Prior information

# Where Does Statistics Fit In?

- Central to steps 2, 3, and 5
  - Experimental design, survey sampling
  - Statistical inference
- Bayesian statistics is particularly well-suited for step 2 and 5
  - Step 2: Prior information
  - Step 5: Prior information $\rightarrow$ posterior information

# Economic Applications of Bayesian Statistics

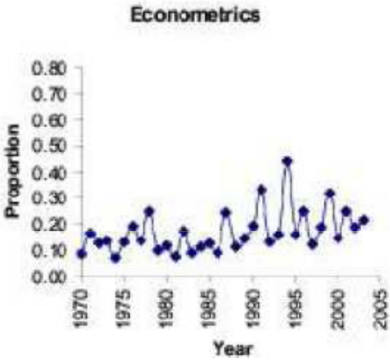



Figure 5: Econometrica Containing "Bayes" or "Bayesian"

# Common Applications of Bayesian Statistics

- Economics

# Common Applications of Bayesian Statistics

- Economics
- Marketing

# Common Applications of Bayesian Statistics

- Economics
- Marketing
- Education

# Common Applications of Bayesian Statistics

- Economics
- Marketing
- Education
- Medical research

# Common Applications of Bayesian Statistics

- Economics
- Marketing
- Education
- Medical research
- Genomics

# Common Applications of Bayesian Statistics

- Economics
- Marketing
- Education
- Medical research
- Genomics
- Weather

# Common Applications of Bayesian Statistics

- Economics
- Marketing
- Education
- Medical research
- Genomics
- Weather
- The list goes on...

# Theoretical Differences to Frequentist Approaches

- Probability as the subjective experience of uncertainty

# Theoretical Differences to Frequentist Approaches

- Probability as the subjective experience of uncertainty
- No notion of infinitely repeating an event of interest

# Theoretical Differences to Frequentist Approaches

- Probability as the subjective experience of uncertainty
- No notion of infinitely repeating an event of interest
- Using as much prior information as possible as well as personal judgment (placing a bet)

# Theoretical Differences to Frequentist Approaches

- Probability as the subjective experience of uncertainty
- No notion of infinitely repeating an event of interest
- Using as much prior information as possible as well as personal judgment (placing a bet)
- Once an outcome is revealed, prior information is updated

# Table of Frequentist vs. Bayesian Interpretations

| | Frequentist statistics | Bayesian statistics |
|---|---|---|
| Definition of the $p$ value | The probability of observing the same or more extreme data assuming that the null hypothesis is true in the population | The probability of the (null) hypothesis |
| Large samples needed? | Usually, when normal theory-based methods are used | Not necessarily |
| Inclusion of prior knowledge possible? | No | Yes |
| Nature of the parameters in the model | Unknown but fixed | Unknown and therefore random |
| Population parameter | One true value | A distribution of values reflecting uncertainty |
| Uncertainty is defined by | The sampling distribution based on the idea of infinite repeated sampling | Probability distribution for the population parameter |
| Estimated intervals | Confidence interval: Over an infinity of samples taken from the population, 95% of these contain the true population value | Credibility interval: A 95% probability that the population value is within the limits of the interval |

# Bayesian vs. Frequentist

- Frequentist: Parameter is unknown, but fixed

# Bayesian vs. Frequentist

- Frequentist: Parameter is unknown, but fixed
- Bayesian: Unknown parameters are treated as uncertain and described by probability distribution

# Bayesian vs. Frequentist

- Frequentist: Parameter is unknown, but fixed
- Bayesian: Unknown parameters are treated as uncertain and described by probability distribution
- Frequentist: 95 of 100 replications of same experiment capture the fixed but unknown parameter

# Bayesian vs. Frequentist

- Frequentist: Parameter is unknown, but fixed
- Bayesian: Unknown parameters are treated as uncertain and described by probability distribution
- Frequentist: 95 of 100 replications of same experiment capture the fixed but unknown parameter
- Bayesian: Probability that a parameter lies in the credible interval

# Priors

- Understanding of parameters of interest given our current data depends on our prior knowledge about the parameters of interest weighted by the current evidence given those parameters of interest

# Priors

- Understanding of parameters of interest given our current data depends on our prior knowledge about the parameters of interest weighted by the current evidence given those parameters of interest
- Priors: Equally important to quantify ignorance as to quantify cumulative understanding of a problem

# Priors

- Understanding of parameters of interest given our current data depends on our prior knowledge about the parameters of interest weighted by the current evidence given those parameters of interest
- Priors: Equally important to quantify ignorance as to quantify cumulative understanding of a problem
- Objective prior: Pure ignorance

# Priors

- Understanding of parameters of interest given our current data depends on our prior knowledge about the parameters of interest weighted by the current evidence given those parameters of interest
- Priors: Equally important to quantify ignorance as to quantify cumulative understanding of a problem
- Objective prior: Pure ignorance
- Prior reflects knowledge about parameters before observing current data

# Priors

- Understanding of parameters of interest given our current data depends on our prior knowledge about the parameters of interest weighted by the current evidence given those parameters of interest
- Priors: Equally important to quantify ignorance as to quantify cumulative understanding of a problem
- Objective prior: Pure ignorance
- Prior reflects knowledge about parameters before observing current data
- Science can be accumulative!

# Bayes' Theorem

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A) \times \mathrm{P}(A)}{\mathrm{P}(B)}$$

# An Example of Bayes' Theorem

- Enzyme-Linked Immuno Sobert Assay (ELISA)

# An Example of Bayes' Theorem

- Enzyme-Linked Immuno Sobert Assay (ELISA)
  - Diagnosis tool for HIV

# An Example of Bayes' Theorem

- Enzyme-Linked Immuno Sobert Assay (ELISA)
  - Diagnosis tool for HIV
  - Low cost

# An Example of Bayes' Theorem

- Enzyme-Linked Immuno Sobert Assay (ELISA)
  - Diagnosis tool for HIV
  - Low cost
  - Less accurate compared to more expensive tools

# An Example of Bayes' Theorem

- Enzyme-Linked Immuno Sobert Assay (ELISA)
  - Diagnosis tool for HIV
  - Low cost
  - Less accurate compared to more expensive tools
- Let $A = \{\text{the patient is positive}\}$. Is it proper to use $\Pr(A)$?

# An Example of Bayes' Theorem

- Enzyme-Linked Immuno Sobert Assay (ELISA)
  - Diagnosis tool for HIV
  - Low cost
  - Less accurate compared to more expensive tools
- Let $\mathrm{A} = \{\text{the patient is positive}\}$. Is it proper to use $\Pr(\mathrm{A})$?
- How should we update the uncertainty after a test?

# An Example of Bayes' Theorem (cont.)

- Data: Test result (positive/negative)

# An Example of Bayes' Theorem (cont.)

- Data: Test result (positive/negative)
- Update the uncertainty using Bayes' theorem

$$\mathrm{P}(A|+) = \frac{\mathrm{P}(+|A)\mathrm{P}(A)}{\mathrm{P}(+|A)\mathrm{P}(A) + \mathrm{P}(+|\neg A)\mathrm{P}(\neg A)}$$

# An Example of Bayes' Theorem (cont.)

- Data: Test result (positive/negative)
- Update the uncertainty using Bayes' theorem

$$\mathrm{P}(A|+) = \frac{\mathrm{P}(+|A)\mathrm{P}(A)}{\mathrm{P}(+|A)\mathrm{P}(A) + \mathrm{P}(+|\neg A)\mathrm{P}(\neg A)}$$

- Sensitivity: $\mathrm{P}(+|A) = 0.95$

# An Example of Bayes' Theorem (cont.)

- Data: Test result (positive/negative)
- Update the uncertainty using Bayes' theorem

$$\mathrm{P}(A|+) = \frac{\mathrm{P}(+|A)\mathrm{P}(A)}{\mathrm{P}(+|A)\mathrm{P}(A) + \mathrm{P}(+|\neg A)\mathrm{P}(\neg A)}$$

- Sensitivity: $\mathrm{P}(+|A) = 0.95$
- Specificity: $\mathrm{P}(-|\neg A) = 1 - \mathrm{P}(+|\neg A) = 0.98$

# An Example of Bayes' Theorem (cont.)

- Data: Test result (positive/negative)
- Update the uncertainty using Bayes' theorem

$$\mathrm{P}(A|+) = \frac{\mathrm{P}(+|A)\mathrm{P}(A)}{\mathrm{P}(+|A)\mathrm{P}(A) + \mathrm{P}(+|\neg A)\mathrm{P}(\neg A)}$$

- Sensitivity: $\mathrm{P}(+|A) = 0.95$
- Specificity: $\mathrm{P}(-|\neg A) = 1 - \mathrm{P}(+|\neg A) = 0.98$
- Population prevalance: $\mathrm{P}(A) = 0.01$

# An Example of Bayes' Theorem (cont.)

- Data: Test result (positive/negative)
- Update the uncertainty using Bayes' theorem

$$\mathrm{P}(A|+) = \frac{\mathrm{P}(+|A)\mathrm{P}(A)}{\mathrm{P}(+|A)\mathrm{P}(A) + \mathrm{P}(+|\neg A)\mathrm{P}(\neg A)}$$

- Sensitivity: $\mathrm{P}(+|A) = 0.95$
- Specificity: $\mathrm{P}(-|\neg A) = 1 - \mathrm{P}(+|\neg A) = 0.98$
- Population prevalance: $\mathrm{P}(A) = 0.01$
- $\mathrm{P}(A|+) = 0.32$

# An Example of Bayes' Theorem (cont.)

- Data: Test result (positive/negative)
- Update the uncertainty using Bayes' theorem

$$P(A|+) = \frac{P(+|A)P(A)}{P(+|A)P(A) + P(+|\neg A)P(\neg A)}$$

- Sensitivity: $P(+|A) = 0.95$
- Specificity: $P(-|\neg A) = 1 - P(+|\neg A) = 0.98$
- Population prevalance: $P(A) = 0.01$
- $P(A|+) = 0.32$
- False positive rate: $P(\neg A|+) = 1 - P(A|+) = 0.68$

# An Example of Bayes' Theorem (cont.)

- Data: Test result (positive/negative)
- Update the uncertainty using Bayes' theorem

$$\mathrm{P}(A|+) = \frac{\mathrm{P}(+|A)\mathrm{P}(A)}{\mathrm{P}(+|A)\mathrm{P}(A) + \mathrm{P}(+|\neg A)\mathrm{P}(\neg A)}$$

- Sensitivity: $\mathrm{P}(+|A) = 0.95$
- Specificity: $\mathrm{P}(-|\neg A) = 1 - \mathrm{P}(+|\neg A) = 0.98$
- Population prevalance: $\mathrm{P}(A) = 0.01$
- $\mathrm{P}(A|+) = 0.32$
- False positive rate: $\mathrm{P}(\neg A|+) = 1 - \mathrm{P}(A|+) = 0.68$
- False negative rate: $\mathrm{P}(A|-) = 0.00052$

# General Form of Bayes' Theorem

- Model: $y|\theta \sim f(y|\theta)$

# General Form of Bayes' Theorem

- Model: $y|\theta \sim f(y|\theta)$
  - $y$: data

# General Form of Bayes' Theorem

- Model: $y|\theta \sim f(y|\theta)$
  - $y$: data
  - $\theta$: parameter

# General Form of Bayes' Theorem

- Model: $y|\theta \sim f(y|\theta)$
  - $y$: data
  - $\theta$: parameter
- Bayes' theorem

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

# General Form of Bayes' Theorem

- Model: $y|\theta \sim f(y|\theta)$
  - $y$: data
  - $\theta$: parameter
- Bayes' theorem

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

- Posterior $\propto$ likelihood $\times$ prior

# Key Issues in Bayesian Statistics

- How to specify prior distributions?

# Key Issues in Bayesian Statistics

- How to specify prior distributions?
- How to compute the posterior $f(\theta|y)$?

# Key Issues in Bayesian Statistics

- How to specify prior distributions?
- How to compute the posterior $f(\theta|y)$?
  - In general we need to know the marginal distribution $f(y) = \int f(y|\theta)f(\theta)d\theta$

# Key Issues in Bayesian Statistics

- How to specify prior distributions?
- How to compute the posterior $f(\theta|y)$?
  - In general we need to know the marginal distribution $f(y) = \int f(y|\theta)f(\theta)d\theta$
  - Direct computation of this integral is difficult especially when $\theta$ is high-dimensional

# Key Issues in Bayesian Statistics

- How to specify prior distributions?
- How to compute the posterior $f(\theta|y)$?
  - In general we need to know the marginal distribution $f(y) = \int f(y|\theta)f(\theta)d\theta$
  - Direct computation of this integral is difficult especially when $\theta$ is high-dimensional
  - Simulation is the general solution

# Key Issues in Bayesian Statistics

- How to specify prior distributions?
- How to compute the posterior $f(\theta|y)$?
  - In general we need to know the marginal distribution $f(y) = \int f(y|\theta)f(\theta)d\theta$
  - Direct computation of this integral is difficult especially when $\theta$ is high-dimensional
  - Simulation is the general solution
- How to summarize the posterior?

# Key Issues in Bayesian Statistics

- How to specify prior distributions?
- How to compute the posterior $f(\theta|y)$?
  - In general we need to know the marginal distribution $f(y) = \int f(y|\theta)f(\theta)d\theta$
  - Direct computation of this integral is difficult especially when $\theta$ is high-dimensional
  - Simulation is the general solution
- How to summarize the posterior?
  - Posterior mean, median, or mode

# Key Issues in Bayesian Statistics

- How to specify prior distributions?
- How to compute the posterior $f(\theta|y)$?
  - In general we need to know the marginal distribution $f(y) = \int f(y|\theta)f(\theta)d\theta$
  - Direct computation of this integral is difficult especially when $\theta$ is high-dimensional
  - Simulation is the general solution
- How to summarize the posterior?
  - Posterior mean, median, or mode
  - Credible interval

# Comparison of Bayesian and Frequentist Methods

- Interval estimation

# Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confindence interval

# Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confindence interval
    - The experiment should be repeatable

# Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confindence interval
    - The experiment should be repeatable
    - Not conditional on the data

# Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confindence interval
    - ⋆ The experiment should be repeatable
    - ⋆ Not conditional on the data
  - Bayesian: Credible interval

# Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confindence interval
    - The experiment should be repeatable
    - Not conditional on the data
  - Bayesian: Credible interval
- Hypothesis testing

# Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confindence interval
    - The experiment should be repeatable
    - Not conditional on the data
  - Bayesian: Credible interval
- Hypothesis testing
  - Consider testing $H_0 : \theta = 0.5$ vs. $H_1 : \theta > 0.5$

# Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confindence interval
    - The experiment should be repeatable
    - Not conditional on the data
  - Bayesian: Credible interval
- Hypothesis testing
  - Consider testing $H_0 : \theta = 0.5$ vs. $H_1 : \theta > 0.5$
  - Frequentist:

# Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confindence interval
    - The experiment should be repeatable
    - Not conditional on the data
  - Bayesian: Credible interval
- Hypothesis testing
  - Consider testing $H_0 : \theta = 0.5$ vs. $H_1 : \theta > 0.5$
  - Frequentist:
    - $p$-value can only be computed when the paramter value is given under the null

# Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confindence interval
    - The experiment should be repeatable
    - Not conditional on the data
  - Bayesian: Credible interval
- Hypothesis testing
  - Consider testing $H_0 : \theta = 0.5$ vs. $H_1 : \theta > 0.5$
  - Frequentist:
    - *p*-value can only be computed when the paramter value is given under the null
    - If we switch the role of two hypotheses, frequentists cannot compute *p*-value

# Comparison of Bayesian and Frequentist Methods

- Interval estimation
  - Frequentist: Confindence interval
    - The experiment should be repeatable
    - Not conditional on the data
  - Bayesian: Credible interval
- Hypothesis testing
  - Consider testing $H_0 : \theta = 0.5$ vs. $H_1 : \theta > 0.5$
  - Frequentist:
    - $p$-value can only be computed when the paramter value is given under the null
    - If we switch the role of two hypotheses, frequentists cannot compute $p$-value
  - Bayesian: No trouble

# Advantages and Disadvantages of Bayesian

- Disadvantages of Bayesian

# Advantages and Disadvantages of Bayesian

- Disadvantages of Bayesian
  - Subjective choice of prior

# Advantages and Disadvantages of Bayesian

- Disadvantages of Bayesian
  - Subjective choice of prior
  - Computation of posterior can be difficult

# Advantages and Disadvantages of Bayesian

- Disadvantages of Bayesian
  - Subjective choice of prior
  - Computation of posterior can be difficult
- Advantages of Bayesian

# Advantages and Disadvantages of Bayesian

- Disadvantages of Bayesian
  - Subjective choice of prior
  - Computation of posterior can be difficult
- Advantages of Bayesian
  - Can easily incorporate prior information

# Advantages and Disadvantages of Bayesian

- Disadvantages of Bayesian
  - Subjective choice of prior
  - Computation of posterior can be difficult
- Advantages of Bayesian
  - Can easily incorporate prior information
  - Inferences are conditional on the actual data

# Advantages and Disadvantages of Bayesian

- Disadvantages of Bayesian
  - Subjective choice of prior
  - Computation of posterior can be difficult
- Advantages of Bayesian
  - Can easily incorporate prior information
  - Inferences are conditional on the actual data
  - Results are more easily interpretable

# Advantages and Disadvantages of Bayesian

- Disadvantages of Bayesian
  - Subjective choice of prior
  - Computation of posterior can be difficult
- Advantages of Bayesian
  - Can easily incorporate prior information
  - Inferences are conditional on the actual data
  - Results are more easily interpretable
  - Inferences are based on posterior, which is conceptually simple

# Advantages and Disadvantages of Bayesian

- Disadvantages of Bayesian
  - Subjective choice of prior
  - Computation of posterior can be difficult
- Advantages of Bayesian
  - Can easily incorporate prior information
  - Inferences are conditional on the actual data
  - Results are more easily interpretable
  - Inferences are based on posterior, which is conceptually simple
  - In principle, all problems can be solved

# Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior

# Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form

# Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form
  - Computationally efficient

# Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form
  - Computationally efficient
  - Analytically tractible

# Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form
  - Computationally efficient
  - Analytically tractible
- Elicited (informative) prior

# Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form
  - Computationally efficient
  - Analytically tractible
- Elicited (informative) prior
  - Histogram

# Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form
  - Computationally efficient
  - Analytically tractible
- Elicited (informative) prior
  - Histogram
  - Parameters in priors using moments or quantiles

# Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form
  - Computationally efficient
  - Analytically tractible
- Elicited (informative) prior
  - Histogram
  - Parameters in priors using moments or quantiles
  - Interactive computer programs

# Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form
  - Computationally efficient
  - Analytically tractible
- Elicited (informative) prior
  - Histogram
  - Parameters in priors using moments or quantiles
  - Interactive computer programs
  - Using a hierarchical structure is often helpful

# Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form
  - Computationally efficient
  - Analytically tractible
- Elicited (informative) prior
  - Histogram
  - Parameters in priors using moments or quantiles
  - Interactive computer programs
  - Using a hierarchical structure is often helpful
- Noninformative prior

# Priors

- Mixture of conjugate prior $\rightarrow$ mixture of conjugate posterior
  - Prior $\times$ posterior takes known form
  - Computationally efficient
  - Analytically tractible
- Elicited (informative) prior
  - Histogram
  - Parameters in priors using moments or quantiles
  - Interactive computer programs
  - Using a hierarchical structure is often helpful
- Noninformative prior
- Empirical Bayes

# Noninformative Prior

- Flat (uniform)

# Noninformative Prior

- Flat (uniform)
  - Improper when the support is not finite

# Noninformative Prior

- Flat (uniform)
  - Improper when the support is not finite
  - Improper prior can lead to proper posterior

# Noninformative Prior

- Flat (uniform)
  - Improper when the support is not finite
  - Improper prior can lead to proper posterior
  - Usually not flat after transformation

# Noninformative Prior

- Flat (uniform)
  - Improper when the support is not finite
  - Improper prior can lead to proper posterior
  - Usually not flat after transformation
- Jeffrey's prior: Invariant under one-to-one transformation

# Noninformative Prior

- Flat (uniform)
  - Improper when the support is not finite
  - Improper prior can lead to proper posterior
  - Usually not flat after transformation
- Jeffrey's prior: Invariant under one-to-one transformation
- Location invariant prior for location family

# Noninformative Prior

- Flat (uniform)
  - Improper when the support is not finite
  - Improper prior can lead to proper posterior
  - Usually not flat after transformation
- Jeffrey's prior: Invariant under one-to-one transformation
- Location invariant prior for location family
- Scale invariant prior for scale family

# Some Basic Models

- Beta-binomial model

# Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$

# Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$
  - $\theta \sim \text{beta}(a, b)$

# Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$
  - $\theta \sim \text{beta}(a, b)$
  - Posterior: beta

# Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$
  - $\theta \sim \text{beta}(a, b)$
  - Posterior: beta
- Poisson model

# Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$
  - $\theta \sim \text{beta}(a, b)$
  - Posterior: beta
- Poisson model
  - $y_1, \ldots, y_n|\theta \sim \text{Poisson}$

# Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$
  - $\theta \sim \text{beta}(a, b)$
  - Posterior: beta
- Poisson model
  - $y_1, \ldots, y_n|\theta \sim \text{Poisson}$
  - Prior: gamma

# Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$
  - $\theta \sim \text{beta}(a, b)$
  - Posterior: beta
- Poisson model
  - $y_1, \ldots, y_n|\theta \sim \text{Poisson}$
  - Prior: gamma
  - Marginal model: Negative-binomial

# Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$
  - $\theta \sim \text{beta}(a, b)$
  - Posterior: beta
- Poisson model
  - $y_1, \ldots, y_n|\theta \sim \text{Poisson}$
  - Prior: gamma
  - Marginal model: Negative-binomial
- Exponential model

# Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$
  - $\theta \sim \text{beta}(a, b)$
  - Posterior: beta
- Poisson model
  - $y_1, \dots, y_n|\theta \sim \text{Poisson}$
  - Prior: gamma
  - Marginal model: Negative-binomial
- Exponential model
  - $y_1, \dots, y_n|\theta \sim \text{Exp}(\theta)$

# Some Basic Models

- Beta-binomial model
  - $y|\theta \sim \text{binomial}(n, \theta)$
  - $\theta \sim \text{beta}(a, b)$
  - Posterior: beta
- Poisson model
  - $y_1, \ldots, y_n|\theta \sim \text{Poisson}$
  - Prior: gamma
  - Marginal model: Negative-binomial
- Exponential model
  - $y_1, \ldots, y_n|\theta \sim \text{Exp}(\theta)$
  - Prior: gamma

# Some Common Conjugate Priors

| Likelihood | Prior | Posterior |
|---|---|---|
| $X\|\theta \sim \mathcal{N}(\theta, \sigma^2)$ | $\theta \sim \mathcal{N}(\mu, \tau^2)$ | $\theta\|X \sim \mathcal{N}(\frac{\tau^2}{\sigma^2+\tau^2}X + \frac{\sigma^2}{\sigma^2+\tau^2}\mu, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2})$ |
| $X\|\theta \sim \mathcal{B}(n, \theta)$ | $\theta \sim \mathcal{B}e(\alpha, \beta)$ | $\theta\|X \sim \mathcal{B}e(\alpha + x, n - x + \beta)$ |
| $X_1, \ldots, X_n\|\theta \sim \mathcal{P}(\theta)$ | $\theta \sim \mathcal{G}a(\alpha, \beta)$ | $\theta\|X_1, \ldots, X_n \sim \mathcal{G}a(\sum_i X_i + \alpha, n + \beta)$. |
| $X_1, \ldots, X_n\|\theta \sim \mathcal{NB}(m, \theta)$ | $\theta \sim \mathcal{B}e(\alpha, \beta)$ | $\theta\|X_1, \ldots, X_n \sim \mathcal{B}e(\alpha + mn, \beta + \sum_{i=1}^n x_i)$ |
| $X \sim \mathcal{G}(n/2, 2\theta)$ | $\theta \sim \mathcal{IG}(\alpha, \beta)$ | $\theta\|X \sim \mathcal{IG}(n/2 + \alpha, (x/2 + \beta^{-1})^{-1})$ |
| $X_1, \ldots, X_n\|\theta \sim \mathcal{U}(0, \theta)$ | $\theta \sim \mathcal{P}a(\theta_0, \alpha)$ | $\theta\|X_1, \ldots, X_n \sim \mathcal{P}a(\max\{\theta_0, x_1, \ldots, x_n\}\alpha + n)$ |
| $X\|\theta \sim \mathcal{N}(\mu, \theta)$ | $\theta \sim \mathcal{IG}(\alpha, \beta)$ | $\theta\|X \sim \mathcal{IG}(\alpha + 1/2, \beta + (\mu - X)^2/2)$ |
| $X\|\theta \sim \mathcal{G}a(\nu, \theta)$ | $\theta \sim \mathcal{G}a(\alpha, \beta)$ | $\theta\|X \sim \mathcal{G}a(\alpha + \nu, \beta + x)$ |

# Likelihood Principle

- Likelihood principle: In the inference about $\theta$, after $y$ is observed, all relevant experimental information is contained in the likelihood function for the observed $y$. Furthermore, two likelihood functions contain the same information about $\theta$ if they are proportional to each other

# Likelihood Principle

- Likelihood principle: In the inference about $\theta$, after $y$ is observed, all relevant experimental information is contained in the likelihood function for the observed $y$. Furthermore, two likelihood functions contain the same information about $\theta$ if they are proportional to each other
- Consider testing the fairness of a coin:

$$H_0 : \theta = \frac{1}{2} \text{ vs. } H_1 : \theta > \frac{1}{2}$$

# Likelihood Principle

- Likelihood principle: In the inference about $\theta$, after $y$ is observed, all relevant experimental information is contained in the likelihood function for the observed $y$. Furthermore, two likelihood functions contain the same information about $\theta$ if they are proportional to each other
- Consider testing the fairness of a coin:

$$H_0 : \theta = \frac{1}{2} \text{ vs. } H_1 : \theta > \frac{1}{2}$$

- Data: An experiment is conducted and 9 heads and 3 tails are observed

# Two Possible Experiments

- Binomial: 12 toss in total

# Two Possible Experiments

- Binomial: 12 toss in total
- Negative binomial: keep tossing until get three tails

# Two Possible Experiments

- Binomial: 12 toss in total
- Negative binomial: keep tossing until get three tails
- Likelihoods are proportional

# Two Possible Experiments

- Binomial: 12 toss in total
- Negative binomial: keep tossing until get three tails
- Likelihoods are proportional
- Conclusions based on $p$-values are contradictive $\rightarrow$ violation of likelihood principle

# Two Possible Experiments

- Binomial: 12 toss in total
- Negative binomial: keep tossing until get three tails
- Likelihoods are proportional
- Conclusions based on $p$-values are contradictive $\rightarrow$ violation of likelihood principle
- Bayesian method has no difficulty $\rightarrow$ the same conclusion under both scenarios

# Some General Facts About Bayesian Inference

- On average, posterior distribution is less variable than the prior distribution

# Some General Facts About Bayesian Inference

- On average, posterior distribution is less variable than the prior distribution
  - Prior variance: $var(\theta)$

# Some General Facts About Bayesian Inference

- On average, posterior distribution is less variable than the prior distribution
  - Prior variance: $var(\theta)$
  - Posterior variance: $var(\theta|y)$

# Some General Facts About Bayesian Inference

- On average, posterior distribution is less variable than the prior distribution
  - Prior variance: $var(\theta)$
  - Posterior variance: $var(\theta|y)$
  - $var(\theta) = E(var(\theta|y)) + var(E(\theta|y)) \geq E(var(\theta|y))$

# Some General Facts About Bayesian Inference

- On average, posterior distribution is less variable than the prior distribution
  - Prior variance: $var(\theta)$
  - Posterior variance: $var(\theta|y)$
  - $var(\theta) = E(var(\theta|y)) + var(E(\theta|y)) \geq E(var(\theta|y))$
- Sequential updates in Bayesian inference

# Some General Facts About Bayesian Inference

- On average, posterior distribution is less variable than the prior distribution
  - Prior variance: $var(\theta)$
  - Posterior variance: $var(\theta|y)$
  - $var(\theta) = E(var(\theta|y)) + var(E(\theta|y)) \geq E(var(\theta|y))$
- Sequential updates in Bayesian inference
  - Prior: $p(\theta)$

# Some General Facts About Bayesian Inference

- On average, posterior distribution is less variable than the prior distribution
  - Prior variance: $var(\theta)$
  - Posterior variance: $var(\theta|y)$
  - $var(\theta) = E(var(\theta|y)) + var(E(\theta|y)) \geq E(var(\theta|y))$
- Sequential updates in Bayesian inference
  - Prior: $p(\theta)$
  - After first batch of data $y_1 \rightarrow p(\theta|y_1) \propto p(y_1|\theta)p(\theta)$

# Some General Facts About Bayesian Inference

- On average, posterior distribution is less variable than the prior distribution
  - Prior variance: $var(\theta)$
  - Posterior variance: $var(\theta|y)$
  - $var(\theta) = E(var(\theta|y)) + var(E(\theta|y)) \geq E(var(\theta|y))$
- Sequential updates in Bayesian inference
  - Prior: $p(\theta)$
  - After first batch of data $y_1 \rightarrow p(\theta|y_1) \propto p(y_1|\theta)p(\theta)$
  - After second batch $y_2$ (assume it is conditionally independent from $y_1$) $\rightarrow p(\theta|y_1, y_2) \propto p(y_2|\theta)p(\theta|y_1) \propto p(y_2|\theta)p(y_1\theta)p(\theta) = p(y_1, y_2|\theta)p(\theta)$

# Some General Facts About Bayesian Inference

- On average, posterior distribution is less variable than the prior distribution
  - Prior variance: $var(\theta)$
  - Posterior variance: $var(\theta|y)$
  - $var(\theta) = E(var(\theta|y)) + var(E(\theta|y)) \geq E(var(\theta|y))$
- Sequential updates in Bayesian inference
  - Prior: $p(\theta)$
  - After first batch of data $y_1 \rightarrow p(\theta|y_1) \propto p(y_1|\theta)p(\theta)$
  - After second batch $y_2$ (assume it is conditionally independent from $y_1$) $\rightarrow p(\theta|y_1, y_2) \propto p(y_2|\theta)p(\theta|y_1) \propto p(y_2|\theta)p(y_1\theta)p(\theta) = p(y_1, y_2|\theta)p(\theta)$
  - This is the same as if we observed both batches together

# Simulate Normal Random Variables: Box–Muller Transformation

- We require two random variables, $U$ and $V$, uniformly distributed on $[0, 1]$. Set

$$\begin{aligned} R &= \sqrt{-2 \log V}, \\ \theta &= 2\pi U, \end{aligned}$$

and

$$\begin{aligned} Z_1 &= R \cos\theta, \\ Z_2 &= R \sin\theta. \end{aligned}$$

Then they are independent standard normal variables. To obtain two standard normal variables with correlation $\rho$, take

$$\begin{aligned} X &= Z_1 \\ Y &= \rho Z_1 + \sqrt{1-\rho^2} Z_2 \end{aligned}$$

# Multiple Linear Regression

- Response: $y$

# Multiple Linear Regression

- Response: $y$
- Explanatory variables: $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$

# Multiple Linear Regression

- Response: $y$
- Explanatory variables: $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$
- Goal: Find $y = f(\boldsymbol{X})$

# Multiple Linear Regression

- Response: $y$
- Explanatory variables: $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$
- Goal: Find $y = f(\boldsymbol{X})$
- Data: $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$

# Multiple Linear Regression

- Response: $y$
- Explanatory variables: $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$
- Goal: Find $y = f(\boldsymbol{X})$
- Data: $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$
- Model: $y_i = f(\boldsymbol{x}_i) + \epsilon_i,$

# Multiple Linear Regression

- Response: $y$
- Explanatory variables: $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$
- Goal: Find $y = f(\boldsymbol{X})$
- Data: $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$
- Model: $y_i = f(\boldsymbol{x}_i) + \epsilon_i$,
- Typical assumption for normal linear model:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) iid$$

# Multiple Linear Regression

- Response: $y$
- Explanatory variables: $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$
- Goal: Find $y = f(\boldsymbol{X})$
- Data: $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$
- Model: $y_i = f(\boldsymbol{x}_i) + \epsilon_i,$
- Typical assumption for normal linear model:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) iid$$

- $E(\boldsymbol{y}|\boldsymbol{X}) = f(\boldsymbol{X})$

# Multiple Linear Regression

- Response: $y$
- Explanatory variables: $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$
- Goal: Find $y = f(\boldsymbol{X})$
- Data: $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$
- Model: $y_i = f(\boldsymbol{x}_i) + \epsilon_i$,
- Typical assumption for normal linear model:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) iid$$

- $E(\boldsymbol{y}|\boldsymbol{X}) = f(\boldsymbol{X})$
- $\boldsymbol{y}|\boldsymbol{X} \sim \mathcal{N}_n(\boldsymbol{X}\beta, \sigma^2 I)$

# Frequentist Inference

- Ordinary least squares

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$
$$\hat{\sigma}^2 = s^2 = \frac{1}{n-k}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})'(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})$$

# Bayesian Inference

- Noninformative prior:

$$f(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{X}) \propto \sigma^{-2}$$

# Bayesian Inference

- Noninformative prior:

$$f(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{X}) \propto \sigma^{-2}$$

- Posterior:

$$\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y} \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, V_\beta \sigma^2)$$

$$\frac{(n-k)s^2}{\sigma^2} | \boldsymbol{y} \sim \chi^2_{n-k}$$

$$\sigma^2 | \boldsymbol{y} \sim Inv - \chi^2(n-k, s^2)$$

# Bayesian Inference

- Noninformative prior:

$$f(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{X}) \propto \sigma^{-2}$$

- Posterior:

$$\begin{aligned} \boldsymbol{\beta}, \sigma^2 | \boldsymbol{y} &\sim \mathcal{N}(\hat{\boldsymbol{\beta}}, V_\beta \sigma^2) \\ \frac{(n-k)s^2}{\sigma^2} | \boldsymbol{y} &\sim \chi^2_{n-k} \\ \sigma^2 | \boldsymbol{y} &\sim Inv - \chi^2(n-k, s^2) \end{aligned}$$

- Marginal posterior of $\boldsymbol{\beta} | \boldsymbol{y}$ is the multivariate $t$-distribution with $n-k$ degrees of freedom

# Hierarchical Model

- Powerful technique for describing complex models. Idea is to break the model down into smaller easier understood pieces, which when put together describes the joint distribution of all data and parameters

# Hierarchical Model

- Powerful technique for describing complex models. Idea is to break the model down into smaller easier understood pieces, which when put together describes the joint distribution of all data and parameters
- Why go hierarchical?

# Hierarchical Model

- Powerful technique for describing complex models. Idea is to break the model down into smaller easier understood pieces, which when put together describes the joint distribution of all data and parameters
- Why go hierarchical?
  - Non-hierarchical models with few parameters generally don't fit the data well

# Hierarchical Model

- Powerful technique for describing complex models. Idea is to break the model down into smaller easier understood pieces, which when put together describes the joint distribution of all data and parameters
- Why go hierarchical?
  - Non-hierarchical models with few parameters generally don't fit the data well
  - Non-hierarchical models with many parameters then to fit the data well, but have poor predictive ability (overfitting)

# Hierarchical Model

- Powerful technique for describing complex models. Idea is to break the model down into smaller easier understood pieces, which when put together describes the joint distribution of all data and parameters
- Why go hierarchical?
  - Non-hierarchical models with few parameters generally don't fit the data well
  - Non-hierarchical models with many parameters then to fit the data well, but have poor predictive ability (overfitting)
  - Hierarchical models can often fit data with a small number of parameters but can also do well in prediction

# A Simple Example

- Observed datum is $\boldsymbol{X}$

$$\boldsymbol{X}|\theta \sim \mathcal{N}(\theta, 1)$$

# A Simple Example

- Observed datum is $\boldsymbol{X}$

$$\boldsymbol{X}|\theta \sim \mathcal{N}(\theta, 1)$$

- First-stage prior: $\theta|\tau^2 \sim \mathcal{N}(0, \tau^2)$

# A Simple Example

- Observed datum is $\boldsymbol{X}$

$$\boldsymbol{X}|\theta \sim \mathcal{N}(\theta, 1)$$

- First-stage prior: $\theta|\tau^2 \sim \mathcal{N}(0, \tau^2)$
- Second-stage prior: $\tau^2 \sim \pi$ where $\pi$ is completely specified

# A Simple Example

- Observed datum is $\boldsymbol{X}$

$$\boldsymbol{X}|\theta \sim \mathcal{N}(\theta, 1)$$

- First-stage prior: $\theta|\tau^2 \sim \mathcal{N}(0, \tau^2)$
- Second-stage prior: $\tau^2 \sim \pi$ where $\pi$ is completely specified
- Possibly more levels

# A Simple Example

- Observed datum is $\boldsymbol{X}$

$$\boldsymbol{X}|\theta \sim \mathcal{N}(\theta, 1)$$

- First-stage prior: $\theta|\tau^2 \sim \mathcal{N}(0, \tau^2)$
- Second-stage prior: $\tau^2 \sim \pi$ where $\pi$ is completely specified
- Possibly more levels
  - Second-stage prior: $\tau^2|\alpha \sim gamma(\alpha, 1)$

# A Simple Example

- Observed datum is $\boldsymbol{X}$

$$\boldsymbol{X}|\theta \sim \mathcal{N}(\theta, 1)$$

- First-stage prior: $\theta|\tau^2 \sim \mathcal{N}(0, \tau^2)$
- Second-stage prior: $\tau^2 \sim \pi$ where $\pi$ is completely specified
- Possibly more levels
  - Second-stage prior: $\tau^2|\alpha \sim \mathit{gamma}(\alpha, 1)$
  - Third-stage prior: $\alpha \sim \mathit{Exp}(1)$

# Hierarchical Linear Model

$$
\begin{aligned}
Y|X, \beta, \Sigma &\sim \mathcal{N}(X\beta, \Sigma) \\
\beta|X_\beta, \alpha, \Sigma_\beta &\sim \mathcal{N}(X_\beta \alpha, \Sigma_\beta) \\
\alpha|\alpha_0, \Sigma_\alpha &\sim \mathcal{N}(\alpha_0, \Sigma_\alpha)
\end{aligned}
$$

# Simple Random Effects Model

- $J$ groups

# Simple Random Effects Model

- $J$ groups
- Data in group $j$: $Y_{1j}, \ldots, Y_{n_j j}$

# Simple Random Effects Model

- $J$ groups
- Data in group $j$: $Y_{1j}, \ldots, Y_{n_j j}$
- $Y_{ij} | \beta_j, \sigma^2 \sim \mathcal{N}(\beta_j, \sigma^2)$ independent, $j = 1, \ldots, J$, $i = 1, \ldots, n_j$

# Simple Random Effects Model

- $J$ groups
- Data in group $j$: $Y_{1j}, \ldots, Y_{n_j j}$
- $Y_{ij} | \beta_j, \sigma^2 \sim \mathcal{N}(\beta_j, \sigma^2)$ independent, $j = 1, \ldots, J$, $i = 1, \ldots, n_j$
- Random effects: $\beta_j | \alpha, \sigma_\beta^2 \sim \mathcal{N}(\alpha, \sigma_\beta^2)$ iid

# Simple Random Effects Model

- $J$ groups
- Data in group $j$: $Y_{1j}, \ldots, Y_{n_j j}$
- $Y_{ij} | \beta_j, \sigma^2 \sim \mathcal{N}(\beta_j, \sigma^2)$ independent, $j = 1, \ldots, J$, $i = 1, \ldots, n_j$
- Random effects: $\beta_j | \alpha, \sigma_\beta^2 \sim \mathcal{N}(\alpha, \sigma_\beta^2)$ iid
- $\alpha \sim \mathcal{N}(\alpha_0, \sigma_\alpha^2)$

# Posterior Inference by Simulation

- Approach 1: Independence sampling

# Posterior Inference by Simulation

- Approach 1: Independence sampling
  - Simulate independent samples from the posterior distributions

# Posterior Inference by Simulation

- Approach 1: Independence sampling
  - Simulate independent samples from the posterior distributions
  - Draw $\theta^{(1)}, \ldots, \theta^{(M)}$ iid from the posterior distribution $f(\theta|y)$

# Posterior Inference by Simulation

- Approach 1: Independence sampling
  - Simulate independent samples from the posterior distributions
  - Draw $\theta^{(1)}, \ldots, \theta^{(M)}$ iid from the posterior distribution $f(\theta|y)$
  - This is all we have discussed so far

# Posterior Inference by Simulation

- Approach 1: Independence sampling
  - Simulate independent samples from the posterior distributions
  - Draw $\theta^{(1)}, \ldots, \theta^{(M)}$ iid from the posterior distribution $f(\theta|y)$
  - This is all we have discussed so far
- Approach 2: Markov Chain Monte Carlo (MCMC)

# Posterior Inference by Simulation

- Approach 1: Independence sampling
  - Simulate independent samples from the posterior distributions
  - Draw $\theta^{(1)}, \ldots, \theta^{(M)}$ iid from the posterior distribution $f(\theta|y)$
  - This is all we have discussed so far
- Approach 2: Markov Chain Monte Carlo (MCMC)
  - Draw $\theta^{(i+1)}$ from $g(\theta^{(i+1)}|\theta^{(i)}$ such that

$$f(\theta^{(i+1)}) \to f(\theta|y)$$

# Posterior Inference by Simulation

- Approach 1: Independence sampling
  - Simulate independent samples from the posterior distributions
  - Draw $\theta^{(1)}, \ldots, \theta^{(M)}$ iid from the posterior distribution $f(\theta|y)$
  - This is all we have discussed so far
- Approach 2: Markov Chain Monte Carlo (MCMC)
  - Draw $\theta^{(i+1)}$ from $g(\theta^{(i+1)}|\theta^{(i)}$ such that

$$f(\theta^{(i+1)}) \rightarrow f(\theta|y)$$

  - Gibbs sampling

# Posterior Inference by Simulation

- Approach 1: Independence sampling
  - Simulate independent samples from the posterior distributions
  - Draw $\theta^{(1)}, \ldots, \theta^{(M)}$ iid from the posterior distribution $f(\theta|y)$
  - This is all we have discussed so far
- Approach 2: Markov Chain Monte Carlo (MCMC)
  - Draw $\theta^{(i+1)}$ from $g(\theta^{(i+1)}|\theta^{(i)}$ such that

$$f(\theta^{(i+1)}) \to f(\theta|y)$$

  - Gibbs sampling
  - Metropolis-Hastings

# Posterior Inference by Simulation

- Approach 1: Independence sampling
  - Simulate independent samples from the posterior distributions
  - Draw $\theta^{(1)}, \dots, \theta^{(M)}$ iid from the posterior distribution $f(\theta|y)$
  - This is all we have discussed so far
- Approach 2: Markov Chain Monte Carlo (MCMC)
  - Draw $\theta^{(i+1)}$ from $g(\theta^{(i+1)}|\theta^{(i)}$ such that

$$f(\theta^{(i+1)}) \rightarrow f(\theta|y)$$

  - Gibbs sampling
  - Metropolis-Hastings
  - Hamiltonian dynamics

# Gibbs Sampler

- Used for multiparameter models

# Gibbs Sampler

- Used for multiparameter models
- Parameter: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$

# Gibbs Sampler

- Used for multiparameter models
- Parameter: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$
- An iterative algorithm

# Gibbs Sampler

- Used for multiparameter models
- Parameter: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$
- An iterative algorithm
  - Draw $\theta_1$ from $p(\theta_1|\theta_2, \theta_3, \ldots, \theta_k, y)$

# Gibbs Sampler

- Used for multiparameter models
- Parameter: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$
- An iterative algorithm
  - Draw $\theta_1$ from $p(\theta_1 | \theta_2, \theta_3, \ldots, \theta_k, y)$
  - Draw $\theta_2$ from $p(\theta_2 | \theta_1, \theta_3, \ldots, \theta_k, y)$

# Gibbs Sampler

- Used for multiparameter models
- Parameter: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$
- An iterative algorithm
  - Draw $\theta_1$ from $p(\theta_1 | \theta_2, \theta_3, \ldots, \theta_k, y)$
  - Draw $\theta_2$ from $p(\theta_2 | \theta_1, \theta_3, \ldots, \theta_k, y)$
  - $\ldots$

# Gibbs Sampler

- Used for multiparameter models
- Parameter: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$
- An iterative algorithm
  - Draw $\theta_1$ from $p(\theta_1|\theta_2, \theta_3, \ldots, \theta_k, y)$
  - Draw $\theta_2$ from $p(\theta_2|\theta_1, \theta_3, \ldots, \theta_k, y)$
  - $\ldots$
  - Draw $\theta_k$ from $p(\theta_k|\theta_1, \theta_2, \ldots, \theta_{k-1}, y)$

# Gibbs Sampler (cont.)

- Full conditional distribution $p(\theta_j|\theta_{-j}, y)$, where $\theta_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k)$

# Gibbs Sampler (cont.)

- Full conditional distribution $p(\theta_j|\theta_{-j}, y)$, where $\theta_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k)$
- In iteration $t$, draw $\theta_j^t = p(\theta_j|\theta_{-j}^t, y)$, where $(\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_k^{t-1})$

# Gibbs Sampler (cont.)

- Full conditional distribution $p(\theta_j|\theta_{-j}, y)$, where $\theta_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k)$
- In iteration $t$, draw $\theta_j^t = p(\theta_j|\theta_{-j}^t, y)$, where $(\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_k^{t-1})$
- Each $\theta_j$ is updated conditional on the latest values of $\theta$

# Example: Simulate from a Bivariate Normal Distribution

S

- Joint distribution

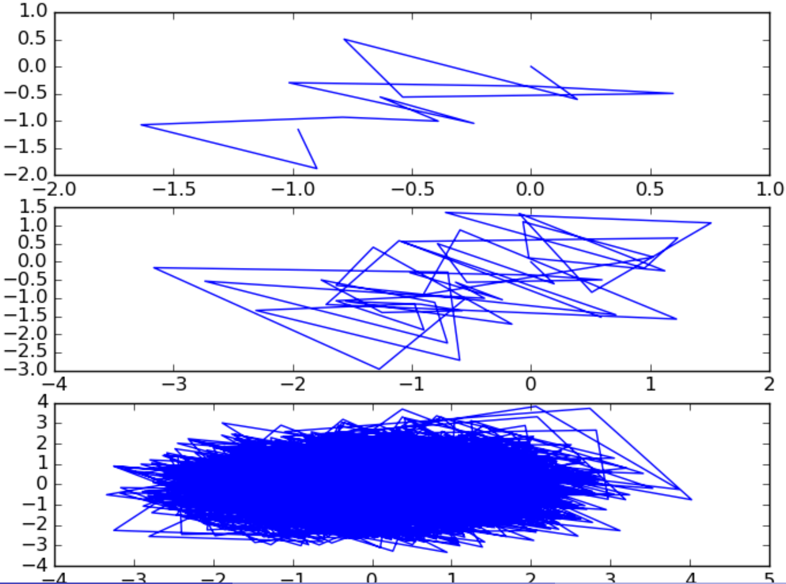
$$\boldsymbol{Z} = (X, Y)' \sim \mathcal{N}(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix})$$

# Python Code

```
import numpy as np
x = [0]
y = [0]
rho = .9
c = np.sqrt(1 - rho**2)
for i in range (6000):
    x.append(rho*y[i - 1] + c*np.random.normal(0, 1, 1))
    y.append(rho*x[i] + c*np.random.normal(0, 1, 1))

%matplotlib qt5
import matplotlib as mpl
import matplotlib.pyplot as plt
plt.style.use('classic')
plt.figure()
plt.subplot(3, 1, 1)
plt.plot(x[0:14], y[0:14], '-')
plt.subplot(3, 1, 2)
```

# The Resulting Posterior



## General Property of Gibbs Sampler

- Output: A dependent sequence

$$\begin{aligned}
\theta^{(1)} &= \{\theta_1^{(1)}, \ldots, \theta_p^{(1)}\} \\
\theta^{(2)} &= \{\theta_1^{(2)}, \ldots, \theta_p^{(2)}\} \\
&\vdots \\
\theta^{(S)} &= \{\theta_1^{(S)}, \ldots, \theta_p^{(S)}\}
\end{aligned}$$

# General Property of Gibbs Sampler

- Output: A dependent sequence

$$
\begin{aligned}
\theta^{(1)} &= \{\theta_1^{(1)}, \ldots, \theta_p^{(1)}\} \\
\theta^{(2)} &= \{\theta_1^{(2)}, \ldots, \theta_p^{(2)}\} \\
&\vdots \\
\theta^{(S)} &= \{\theta_1^{(S)}, \ldots, \theta_p^{(S)}\}
\end{aligned}
$$

- $\theta^{(S)}$ depends on $\theta^{(0)}, \ldots, \theta^{(S-1)}$ only through $\theta^{(S-1)}$

# General Property of Gibbs Sampler

- Output: A dependent sequence

$$\theta^{(1)} = \{\theta_1^{(1)}, \ldots, \theta_p^{(1)}\}$$
$$\theta^{(2)} = \{\theta_1^{(2)}, \ldots, \theta_p^{(2)}\}$$
$$\vdots$$
$$\theta^{(S)} = \{\theta_1^{(S)}, \ldots, \theta_p^{(S)}\}$$

- $\theta^{(S)}$ depends on $\theta^{(0)}, \ldots, \theta^{(S-1)}$ only through $\theta^{(S-1)}$
- $\theta^{(S)}$ is conditionally independent of $\theta^{(0)}, \ldots, \theta^{(S-2)}$ given $\theta^{(S-1)}$

# General Property of Gibbs Sampler

- Output: A dependent sequence

$$\begin{aligned}
\theta^{(1)} &= \{\theta_1^{(1)}, \ldots, \theta_p^{(1)}\} \\
\theta^{(2)} &= \{\theta_1^{(2)}, \ldots, \theta_p^{(2)}\} \\
&\vdots \\
\theta^{(S)} &= \{\theta_1^{(S)}, \ldots, \theta_p^{(S)}\}
\end{aligned}$$

- $\theta^{(S)}$ depends on $\theta^{(0)}, \ldots, \theta^{(S-1)}$ only through $\theta^{(S-1)}$
- $\theta^{(S)}$ is conditionally independent of $\theta^{(0)}, \ldots, \theta^{(S-2)}$ given $\theta^{(S-1)}$
- This is called a Markov property, and the sequence a Markov chain

# General Property of Gibbs Sampler

- Output: A dependent sequence

$$\theta^{(1)} = \{\theta_1^{(1)}, \ldots, \theta_p^{(1)}\}$$
$$\theta^{(2)} = \{\theta_1^{(2)}, \ldots, \theta_p^{(2)}\}$$
$$\vdots$$
$$\theta^{(S)} = \{\theta_1^{(S)}, \ldots, \theta_p^{(S)}\}$$

- $\theta^{(S)}$ depends on $\theta^{(0)}, \ldots, \theta^{(S-1)}$ only through $\theta^{(S-1)}$
- $\theta^{(S)}$ is conditionally independent of $\theta^{(0)}, \ldots, \theta^{(S-2)}$ given $\theta^{(S-1)}$
- This is called a Markov property, and the sequence a Markov chain
- For the models in this class, the sampling distribution of $\theta^{(S)}$ approaches the target distribution as $S \to \infty$, regardless of starting value

$$Pr(\theta^{(S)} \in A \to \int_A p(\theta) d\theta \text{ as } S \to \infty$$

# General Property of Gibbs Sampler (cont.)

- More importantly, for most functions $g$ of interest,

$$\frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}) \rightarrow E[g(\theta)] = \int g(\theta)p(\theta)d\theta \text{ as } S \rightarrow \infty$$

# General Property of Gibbs Sampler (cont.)

- More importantly, for most functions $g$ of interest,

$$\frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}) \rightarrow E[g(\theta)] = \int g(\theta)p(\theta)d\theta \text{ as } S \rightarrow \infty$$

- One can approximate $E[g(\theta)]$ with sample average of $\{g(\theta^{(1)}), \ldots, g(\theta^{(S)})\}$. This is the Monte Carlo part

# General Property of Gibbs Sampler (cont.)

- More importantly, for most functions $g$ of interest,

$$\frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}) \to E[g(\theta)] = \int g(\theta)p(\theta)d\theta \text{ as } S \to \infty$$

- One can approximate $E[g(\theta)]$ with sample average of $\{g(\theta^{(1)}), \ldots, g(\theta^{(S)})\}$. This is the Monte Carlo part
- Hence, we call this method Markov chain Monte Carlo (MCMC)

# Distinguishing Parameter Estimation from Posterior Approximation

- Bayesian data analysis using Monte Carlo methods

# Distinguishing Parameter Estimation from Posterior Approximation

- Bayesian data analysis using Monte Carlo methods
  - Data analysis: The statistical part

# Distinguishing Parameter Estimation from Posterior Approximation

- Bayesian data analysis using Monte Carlo methods
  - Data analysis: The statistical part
  - Numerical approximation: The Monte Carlo part

# Distinguishing Parameter Estimation from Posterior Approximation

- Bayesian data analysis using Monte Carlo methods
  - Data analysis: The statistical part
  - Numerical approximation: The Monte Carlo part
- Ingredients of Bayesian data analysis

# Distinguishing Parameter Estimation from Posterior Approximation

- Bayesian data analysis using Monte Carlo methods
  - Data analysis: The statistical part
  - Numerical approximation: The Monte Carlo part
- Ingredients of Bayesian data analysis
  - Model specification

# Distinguishing Parameter Estimation from Posterior Approximation

- Bayesian data analysis using Monte Carlo methods
  - Data analysis: The statistical part
  - Numerical approximation: The Monte Carlo part
- Ingredients of Bayesian data analysis
  - Model specification
  - Prior specification

# Distinguishing Parameter Estimation from Posterior Approximation

- Bayesian data analysis using Monte Carlo methods
  - Data analysis: The statistical part
  - Numerical approximation: The Monte Carlo part
- Ingredients of Bayesian data analysis
  - Model specification
  - Prior specification
  - Posterior summary

# Distinguishing Parameter Estimation from Posterior Approximation

- Bayesian data analysis using Monte Carlo methods
  - Data analysis: The statistical part
  - Numerical approximation: The Monte Carlo part
- Ingredients of Bayesian data analysis
  - Model specification
  - Prior specification
  - Posterior summary
- When the posterior distribution is complicated, we can "look at" the posterior by studying Monte Carlo samples from the posterior

# Distinguishing Parameter Estimation from Posterior Approximation

- Monte Carlo and MCMC sampling algorithms

# Distinguishing Parameter Estimation from Posterior Approximation

- Monte Carlo and MCMC sampling algorithms
  - Are not models

# Distinguishing Parameter Estimation from Posterior Approximation

- Monte Carlo and MCMC sampling algorithms
  - Are not models
  - They do not generate more information than is in $y$ and $p(\theta)$

# Distinguishing Parameter Estimation from Posterior Approximation

- Monte Carlo and MCMC sampling algorithms
  - Are not models
  - They do not generate more information than is in $y$ and $p(\theta)$
  - They are simply ways of looking at $p(\theta|y)$

# Distinguishing Parameter Estimation from Posterior Approximation

- Monte Carlo and MCMC sampling algorithms
  - Are not models
  - They do not generate more information than is in $y$ and $p(\theta)$
  - They are simply ways of looking at $p(\theta|y)$
- Estimation: How we use $p(\theta|y)$ to make inferences about $\theta$

# Distinguishing Parameter Estimation from Posterior Approximation

- Monte Carlo and MCMC sampling algorithms
  - Are not models
  - They do not generate more information than is in $y$ and $p(\theta)$
  - They are simply ways of looking at $p(\theta|y)$
- Estimation: How we use $p(\theta|y)$ to make inferences about $\theta$
- Approximation: The use of Monte Carlo procedures to approximate integrals

# Additional Issues

- Length of a chain

# Additional Issues

- Length of a chain
- Multiple chains (in parallel)

# Additional Issues

- Length of a chain
- Multiple chains (in parallel)
- Burn-in

# Additional Issues

- Length of a chain
- Multiple chains (in parallel)
- Burn-in
- Reduce autocorrelation by thinning

# Additional Issues

- Length of a chain
- Multiple chains (in parallel)
- Burn-in
- Reduce autocorrelation by thinning
- Convergence assessment

# Additional Issues

- Length of a chain
- Multiple chains (in parallel)
- Burn-in
- Reduce autocorrelation by thinning
- Convergence assessment
- We will deal with these later