Lecture 4: Intro to Bayesian Analysis

UNIVERSITY OF
SAN FRANCISCO

James D. Wilson

MSAN 628

Computational Statistics

Statistical Analysis

Aim: develop practical methods for making inferences from data using probability models for (a) quantities we observe and (b) quantities we wish to learn.

We develop a statistical model to describe population quantities using observed samples or data.

A statistical model relies on the following components

- θ: parameter(s) that describe features of a population
- y: observed data / sample for which we make inference

Statistical Analysis

- In traditional *frequentist* analysis, we posit that the observed data y arises from a family of possible models $\{f(y \mid \theta) : \theta \in \Theta\}$.
  - θ is treated as an unknown constant
  - We use the likelihood (or distribution) $L(\theta \mid y) = f(y \mid \theta)$ to make inference on θ.

Bayesian Analysis

- In *Bayesian* analysis, we treat θ as a random quantity that is observed *jointly* with the data y. In particular, we suppose that

$$p(\theta \mid y) = \frac{p(\theta, y)}{p(y)} = \frac{\pi(\theta) f(y \mid \theta)}{p(y)}$$

  - $p(\theta \mid y)$ = the posterior distribution of θ given y
  - $f(y \mid \theta)$ = the likelihood of the data y given θ
  - $\pi(\theta)$ = the prior distribution of θ
  - $p(y) = \int \pi(\theta) f(y \mid \theta) d\theta$ = the marginal distribution of y.

The Three Steps of Bayesian Analysis

There are three main steps to Bayesian Analysis:

1) *Set up a full probability model* - a joint probability distribution for all observable and unobservable quantities in a problem, namely for (y, θ). The model should be consistent with
   - the underlying problem
   - the data collection process

   Note: Bayesian models requires that observed data $y_1, \ldots, y_n$ are exchangeable, namely $p(y_1, \ldots, y_n)$ is invariant to permutations of the indices.

The Three Steps of Bayesian Analysis

2) *Condition on observed data* - calculate and interpret the *posterior distribution* - the conditional distribution of the unobserved quantities of interest given the observed data.
   - In other words, define a conditional density $f(y \mid \theta)$ and calculate $p(\theta \mid y)$ using Bayes rule.

The Three Steps of Bayesian Analysis

3) *Validation* - evaluate the fit of the model and implications of the resulting posterior distribution:
   - how well does the model fit the data?
   - are the conclusions reasonable?
   - how sensitive are the results to the modeling assumptions?

A Brief Digression

- In many cases, the marginalizing constant (or denominator) $p(y)$ is difficult or impossible to calculate.
- However, when we are trying to calculate the posterior density of θ, it is often enough to calculate the numerator of Bayes rule:

$$p(\theta \mid y) \propto \pi(\theta) f(y \mid \theta)$$

- $\pi(\theta) f(y \mid \theta)$ is known as the kernel of the distribution. It contains all information about θ, but does *not* integrate to 1, and thus is not a distribution itself.

Estimating the Probability of a Female Birth

- The currently accepted proportion of female births in European-race populations is 0.485
- We would like to devise a statistical model for θ = the probability of a female birth in a European country
- *Prior*: we have no idea, so suppose θ is uniform on the unit interval
- *Data*: observe y girls of n recorded births in the past year
- *Question*: Using these bits of information, what is a reliable model for $\theta \mid y$?

Estimating the Probability of a Female Birth

- Data suggests a binomial model for y:

$$P(Y = y \mid \theta) = \text{Bin}(n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

- Applying Bayes rule, we get the kernel of the posterior density:

$$p(\theta) \propto \pi(\theta) f(y \mid \theta) = 1 \binom{n}{y} \theta^y (1-\theta)^{n-y} \propto \theta^y (1-\theta)^{n-y}$$

- The above is exactly the kernel of the Beta$(y+1, n-y+1)$ distribution. So, our posterior model is

$$\theta \mid y \sim \text{Beta}(y+1, n-y+1)$$

Posterior Predictions

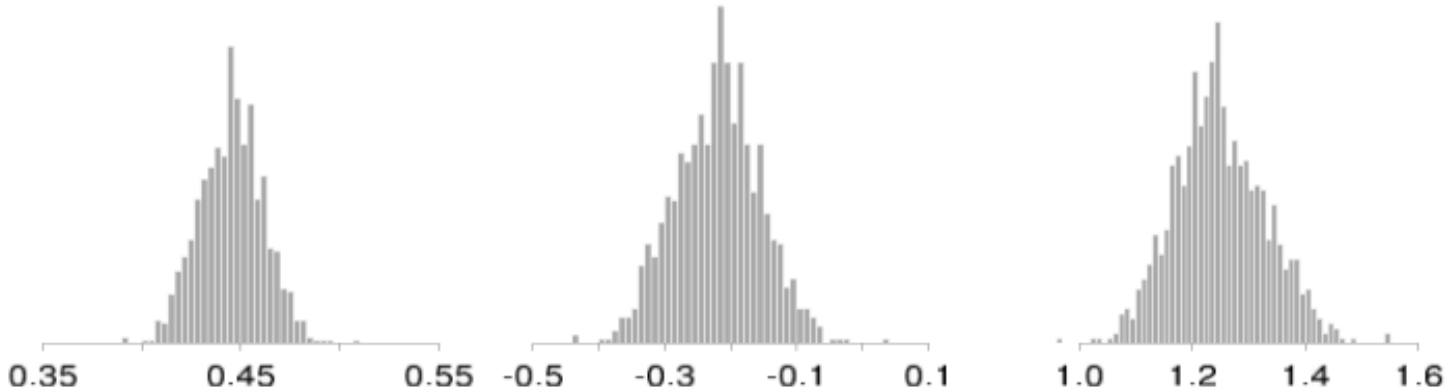


Figure 2.3 *Draws from the posterior distribution of (a) the probability of female birth, θ; (b) the logit transform, logit(θ); (c) the male-to-female sex ratio, $\phi = (1-\theta)/\theta$.*

Compromise between the data and the prior

Recall the laws of total expectation, which give:

$$\mathbb{E}[\theta] = \mathbb{E}[\mathbb{E}[\theta \mid y]] \tag{1}$$

$$\text{Var}(\theta) = \mathbb{E}[\text{Var}(\theta \mid y)] + \text{Var}(\mathbb{E}[\theta \mid y]) \tag{2}$$

These laws imply two important things about the posterior distribution:

- (1) $\rightarrow$ *the prior mean of* θ *is the average of all possible posterior means over the distribution of possible data*
- (2) $\rightarrow$ *the posterior variance is on average smaller than the prior variance*

Notes about Posterior Inference

- The posterior distribution $p(\theta \mid y)$ contains **all** information about the parameter θ
- Thus, all inference can be done using $p(\theta \mid y)$:
  - point estimation: the posterior mode, median, or mean
  - intervals: posterior intervals
  - hypothesis testing: directly from $p(\theta \mid y)$
- Compare the above to the usual frequentist approach for inference
- We will revisit these modes of inference throughout the class

Posterior Predictions

- Suppose that y arises from a likelihood $f(y \mid \theta)$ and θ has posterior distribution $p(\theta \mid y)$. Then, the posterior predictive distribution of a future observation $\tilde{y}$ is given by:

$$p(\tilde{y} \mid y) = \int f(\tilde{y} \mid \theta) p(\theta \mid y) d\theta$$

- A α% posterior interval for θ is an interval $[a, b]$ such that

$$\int_a^b p(\theta \mid y) d\theta = \alpha$$

- A α% posterior interval for a new observation $\tilde{y}$ is an interval $[a, b]$ such that

$$\int_a^b p(\tilde{y} \mid y) d\tilde{y} = \alpha$$

Posterior Predictions

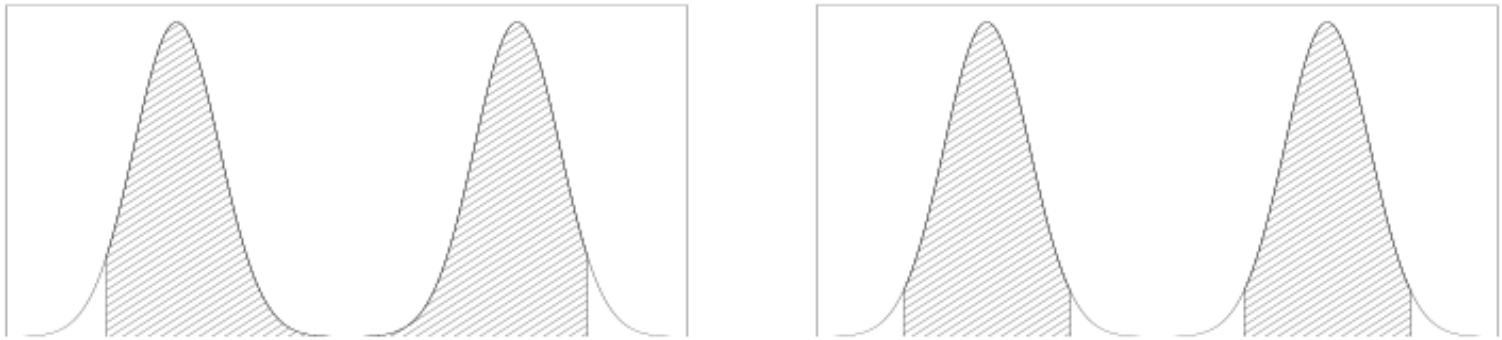


Figure 2.2 *Hypothetical density for which the 95% central interval and 95% highest posterior density region dramatically differ: (a) central posterior interval, (b) highest posterior density region.*

Informative Prior Distributions

The choice of a prior distribution for θ makes a big impact on its posterior inference, as we have seen. Thus, constructing a prior is an important step in Bayesian analysis.

In general, there are two interpretations for prior distributions:

1. *Population Interpretation*: the prior distribution represents a population of possible parameter values from which θ has been drawn
2. *State of knowledge Interpretation*: the prior distribution expresses our knowledge (and uncertainty) about θ as if its value could be thought of as a random realization from the prior distribution

The Binomial Model Revisited

- In general, choosing an appropriate prior distribution is not an easy task.
- Given a data generative model $f(y \mid \theta)$, one common way to choose a prior is to do it in such a way that the *posterior has the same functional form as the prior*.
- Consider the Binomial model as an example

$$f(y \mid \theta) \propto \theta^{y}(1-\theta)^{n-y}$$

Goal: Choose $\pi(\theta)$ so that $p(\theta \mid y)$ has the same form.

The Binomial Model Revisited

Let's keep it really simple. How about a Beta(α, β) distribution, where

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Then,

$$\begin{aligned} p(\theta \mid y) &\propto \theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}\theta^{\beta-1} \\ &= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \\ &= \text{Beta}(\alpha+y, \beta+n-y) \end{aligned}$$

Conclusion: Both $\pi(\theta)$ and $p(\theta \mid y)$ are Beta distributions. The model $y \sim \text{Bin}(n, \theta), \theta \sim \text{Beta}(\alpha, \beta)$ is known as the Beta-Binomial model with hyperparameters α and β.

Conjugate Priors

The Beta-Binomial model is an example of a conjugate family, because the prior and posterior distribution of θ had the same functional form. This idea can be formalized.

Conjugate priors

If $\mathcal{F}$ is a class of sampling distributions $f(y \mid \theta)$ and $\mathcal{P}$ is a class of prior distributions for θ, then the class $\mathcal{P}$ is conjugate for $\mathcal{F}$ if

$$p(\theta \mid y) \in \mathcal{P} \text{ for all } f(\cdot \mid \theta) \in \mathcal{F} \text{ and } \pi(\cdot) \in \mathcal{P}$$

Key Advantages: (1) computationally tractable posteriors, (2) interpretable as additional data does not change the model family of θ

Conjugate Priors for Exponential Family Models

Exponential Family

The class $\mathcal{F}$ is an exponential family if all of its members $f(\cdot \mid \theta) \in \mathcal{F}$ have the form

$$f(y \mid \theta) = g(y)h(\theta)\exp(\phi(\theta)^T u(y))$$

Common Examples:

- $N(\mu, \sigma^2)$
- $Exp(\lambda)$
- $Bin(n, p)$
- $Poisson(\lambda)$

Conjugate Priors for Exponential Family Models

Suppose that $y_1, \ldots, y_n$ have density $f(\cdot \mid \theta)$, then the likelihood of $\mathbf{y} = (y_1, \ldots, y_n)$ is

$$f(\mathbf{y} \mid \theta) = \left(\prod_{i=1}^{n} g(y_i) \right) h(\theta)^n \exp\left(\phi(\theta)^T \sum_{i=1}^{n} u(y_i) \right)$$

which suggests that *as a function of* θ:

$$f(\mathbf{y} \mid \theta) \propto h(\theta)^n \exp\left(\phi(\theta)^T \sum_{i=1}^{n} u(y_i) \right),$$

where $t(\mathbf{y}) = \sum_{i=1}^{n} u(y_i)$ is a sufficient statistic for θ because the likelihood for θ depends on the data $\mathbf{y}$ only through the value of $t(\mathbf{y})$.

Conjugate Priors for Exponential Family Models

If $y_1, \ldots, y_n$ have an exponential family so that

$$f(\mathbf{y} \mid \theta) \propto h(\theta)^n \exp\left(\phi(\theta)^T \sum_{i=1}^{n} u(y_i)\right)$$

then any prior density

$$\pi(\theta) \propto h(\theta)^\eta \exp\left(\phi(\theta)^T \nu\right)$$

for some hyperparameters η and ν is a conjugate prior since

$$p(\theta \mid y) \propto h(\theta)^{\eta+n} \exp\left(\phi(\theta)^T (\nu + t(\mathbf{y}))\right)$$

Conclusion: Exponential families have natural conjugate priors!

Common Conjugate Families

- **Beta-Binomial**:

$$y \mid \theta \sim Bin(n, \theta); \theta \sim Beta(\alpha, \beta)$$

- **Normal - Normal**:

$$y \mid \theta \sim N(\theta, \sigma^2); \theta \sim N(\mu_0, \tau_0^2)$$

- **Poisson - Gamma**:

$$y \mid \theta \sim Poisson(\theta); \theta \sim Gamma(\alpha, \beta)$$

- **Exponential - Gamma**:

$$y \mid \theta \sim Exp(1/\theta); \theta \sim Gamma(\alpha, \beta)$$

Informative vs. Noninformative Priors

- One natural way to construct a prior is through the use of conjugate priors
  - intuitive reasoning
  - analytical form of the posterior and marginal density of y
- But... these are not always available! We can use noninformative priors:
  - "flat" prior (e.g., Uniform prior) which gives no stronger weight to any value of θ
  - lets the "data speak for itself"
  - closely related to frequentist approach

Non-conjugate Priors

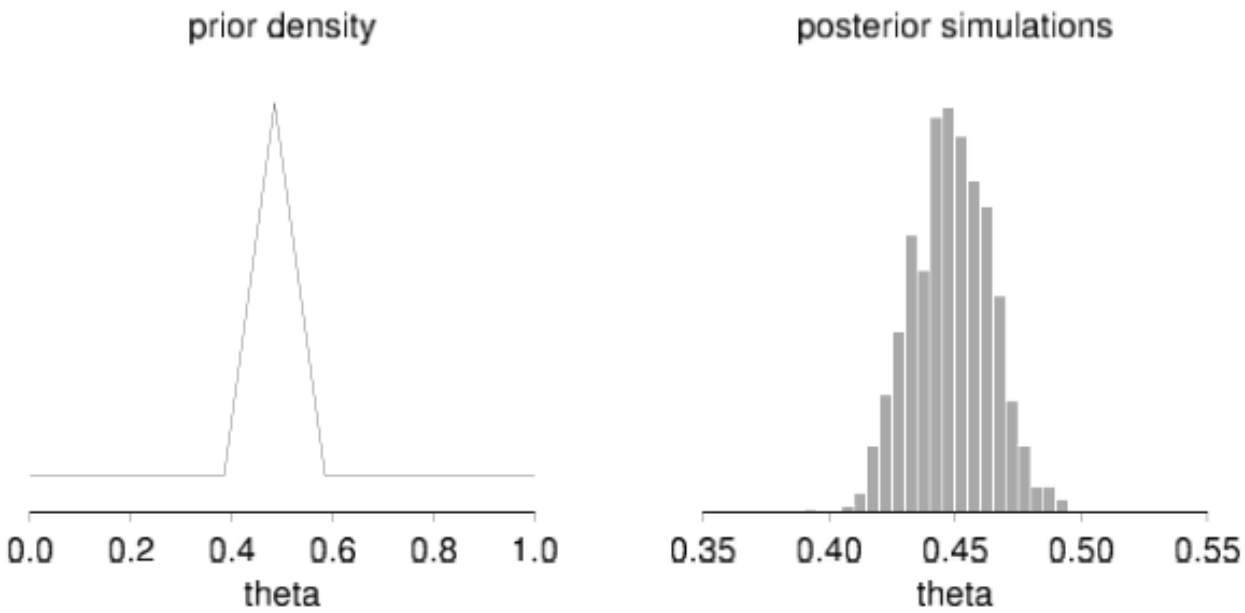


Figure 2.4 *(a) Prior density for θ in an example nonconjugate analysis of birth ratio example; (b) histogram of 1000 draws from a discrete approximation to the posterior density. Figures are plotted on different scales.*

Improper Priors

- Typically, we assume that the prior $\pi(\theta)$ is proper, namely that $\pi(\cdot)$ satisfies the axioms.
- $\pi(\cdot)$ that do not satisfy the axioms (i.e., it integrates to some value other than 1) are known as an improper prior.
- **Examples:**
  - $Beta(0, 0)$
  - $Uniform(0, \infty)$
- Improper $\pi(\theta)$ can still lead to a valid posterior distribution for θ!

Weakly Informative Priors

Weakly Informative Priors

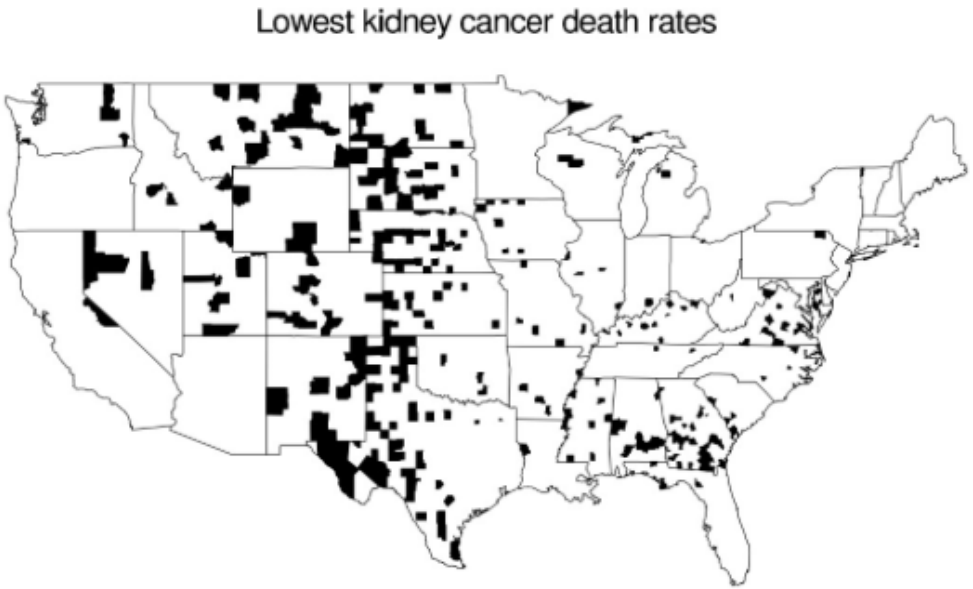
A prior $\pi(\cdot)$ is said to be a weakly informative prior if it is proper and provides information that is weaker than whatever actual prior knowledge is available.

Example

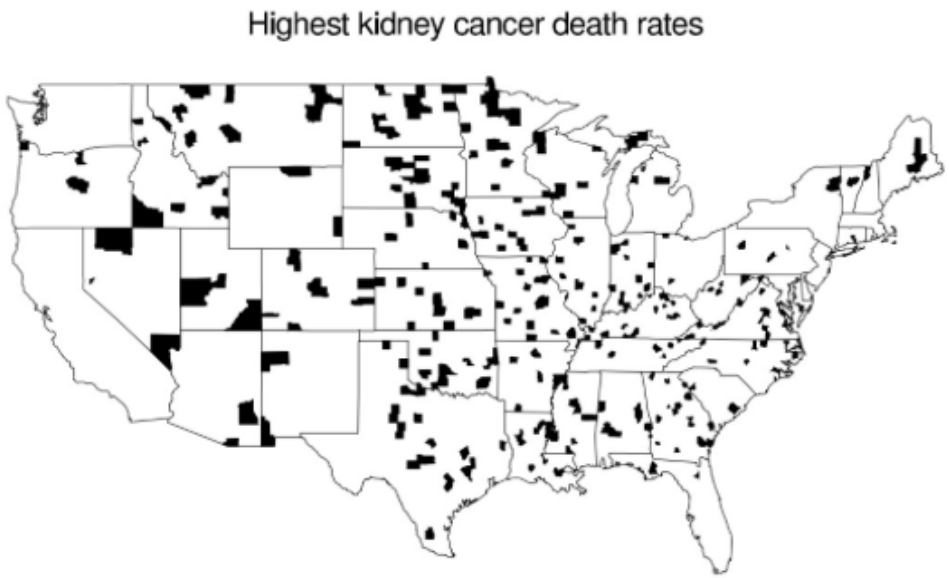
In the previous example about estimating the proportion of female births in Europe. One could specify a $N(0.5, 0.1^2)$ prior, or for the sake of conjugacy, a $Beta(20, 20)$ prior.

Case Study for U.S. Kidney Cancer Death Rates

The following two images show the counties in the United States with the highest kidney cancer death rates and the lowest kidney cancer death rates during the 1980s.



Case Study for U.S. Kidney Cancer Death Rates



Case Study for U.S. Kidney Cancer Death Rates

- Interesting trend: many of the counties in the Great Plains and in the middle of the country tend to have extreme numbers of kidney cancer death rates, but relatively few counties near the coasts are shaded.
- Why?
  - air and water pollution?
  - diet?
  - rate of seeking medical attention?
  - quality of health care?

Case Study for U.S. Kidney Cancer Death Rates

Actually none of these are needed! The real reason is sample size!

- Kidney cancer is a relatively rare disease
- If even one person has kidney cancer in a small county (of which there are a lot in the Great Plains), it is enough to put them in the top 10% of occurrences.
- Since there are so many low-population counties in the middle of the country, these counties are overrepresented in both maps (high variability)
- Any model for these cancer rates should account for the population size in a county

Case Study for U.S. Kidney Cancer Death Rates

Reasonable model:

- y_j = number of kidney cancer deaths in county j; n_j = population of the county; θ_j = underlying rate in units of deaths per person per year
- **Data**: $y_j \sim Poisson(10 n_j \theta_j)$
- **Prior**: $\theta_j \sim Gamma(\alpha, \beta)$ with hyperparameters $\alpha = 20$ and $\beta = 430,000$ chosen to match the mean and standard deviation rate of the observed data
- **Posterior**: $\theta_j \mid y_j \sim Gamma(20 + y_j, 430,000 + 10 n_j)$

Next Time

- Multiparameter models
- Hierarchical models
- Ties between Bayesian and Frequentist inference