Bayesian Inference

STAT 251, Unit 3

Overview

Interpretation of Bayes Theorem

Bayesian Approach to Statistics

Useful R code for Overlaying Densities in a Plot

Interpretation of Bayes Theorem

Recall Definition of Probability

For every event A, we would like to associate it with a number that describes *how likely the event is to occur*. This number is called **probability**.

Recall Definition of Probability

For every event A, we would like to associate it with a number that describes *how likely the event is to occur*. This number is called **probability**.

- The expression $\mathrm{P}(A)$ denotes the probability that the event A occurs.

Interpretations of Probability

Not all statisticians interpret *probability* in the same manner. We will distinguish between two prevalent notions of what $\mathrm{P}(A)$ represents.

Interpretations of Probability

Not all statisticians interpret *probability* in the same manner. We will distinguish between two prevalent notions of what $\mathrm{P}(A)$ represents.

- $\mathrm{P}(A)$ is the proportion of times that event A would occur in the long run, if the experiment were to be repeated over and over again. This long-run-frequency interpretation is a **frequentist** interpretation.
- $\mathrm{P}(A)$ is the degree of belief we have that A will occur. This degree-of-belief interpretation is characteristic of the **Bayesian** interpretation.

Why the Bayesian interpretation?

Sometimes the frequentist interpretation of what probability represents is hard to justify because only one realization of the experiment is obtainable.

For example, suppose event A was that Trump wins the 2016 presidential election.

However, the Bayesian interpretation is still logical in this one-realization scenario.

Discussion Questions:

In each of the following events, determine what the frequentist interpretation of the probability would be, and assess whether or not this is meaningful. Then, determine what the Bayesian interpretation of the probability would be, and assess whether or not this is meaningful.

- $A =$ I roll a five when I toss a standard, six-sided die.
- $A =$ I roll a nine when I toss a standard, six-sided die.
- $A =$ BYU wins the 2018 football game against Utah.
- $A =$ It snows this Saturday.
- $A =$ Trump is re-elected as US President in the 2020 election.

Bayes Theorem

Recall what Bayes Theorem says (From Unit 2A):

Bayes Rule: Let $A_1, \ldots, A_n$ be mutually exclusive and exhaustive events, and let B be any event, then

$$\mathrm{P}(A_k|B) = \frac{\mathrm{P}(B|A_k)\mathrm{P}(A_k)}{\sum_{i=1}^{n} \mathrm{P}(B|A_i)P(A_i)}$$

One way to view this theorem is that we can get "inverse" probabilities. That is, given the $P(B|A_k)$'s (and the $P(A_k)$'s), we can get the $P(A_k|B)$'s—the reverse (or "inverse") conditioning.

Remember: $P(B|A) \neq P(A|B)$ in general

My favorite example to help remember that $P(A|B)$ does not represent the same thing as $P(B|A)$ is from an example provided by Louis Lyons:

What is $P(\textit{pregnant}|\textit{female})$?

`http://blogs.sch.gr/ggrammat/files/2014/11/lyons_colloquium_Stat2007.pdf`, slides 20–21, accessed 28 January 2017

Remember: $P(B|A) \neq P(A|B)$ in general

My favorite example to help remember that $P(A|B)$ does not represent the same thing as $P(B|A)$ is from an example provided by Louis Lyons:

What is $P(\textit{pregnant}|\textit{female})$?

What is $P(\textit{female}|\textit{pregnant})$?

`http://blogs.sch.gr/ggrammat/files/2014/11/lyons_colloquium_Stat2007.pdf`, slides 20–21, accessed 28 January 2017

Alternate Interpretation of Bayes' Theorem

Recall what Bayes Theorem says (From Unit 2A):

Bayes Rule: Let $A_1, \ldots, A_n$ be mutually exclusive and exhaustive events, and let B be any event, then

$$\mathrm{P}(A_k|B) = \frac{\mathrm{P}(B|A_k)\mathrm{P}(A_k)}{\sum_{i=1}^{n} \mathrm{P}(B|A_i)P(A_i)}$$

Bayes Theorem provides the basis on how to **update** probabilities in a coherent fashion. After witnessing some event B, we can update $P(A_k)$ to get $P(A_k|B)$.

Disease Example, revisited

The proportion of people in a given community who have a certain disease is 0.005. A test is available to diagnose the disease. If a person has the disease, the probability that the test will produce a positive signal is 0.99. If a person does not have the disease, the probability that the test will produce a positive signal is 0.02.

Q1 What is the probability that a random person has the disease?

Q2 If a random person tests positive, what is the probability that the person actually has the disease?

Disease Example, revisited

The proportion of people in a given community who have a certain disease is 0.005. A test is available to diagnose the disease. If a person has the disease, the probability that the test will produce a positive signal is 0.99. If a person does not have the disease, the probability that the test will produce a positive signal is 0.02.

Q1 What is the probability that a random person has the disease?

Q2 If a random person tests positive, what is the probability that the person actually has the disease?

A1 $P(Disease) = 0.005$

A2 Using Bayes' theorem, we see that $P(Disease|+) = 0.1992$

The *prior* probability (before getting additional information) was 0.005, but the *posterior* probability (after observing the "data") is updated to 0.1992.

Airline Example (revisited)

Suppose that two airlines operate at a small airport. 55% of the flights at the airport are operated by Airline A and 45% of the flights are operated by airline B. Of the flights operated by Airline A, 90% are on-time. Of the flights operated by Airline B, 75% are on time.

Q1 What is the probability that a random flight was operated by airline B?

Q2 If a random flight is on-time, what is the probability the flight was operated by airline B?

Airline Example (revisited)

Suppose that two airlines operate at a small airport. 55% of the flights at the airport are operated by Airline A and 45% of the flights are operated by airline B. Of the flights operated by Airline A, 90% are on-time. Of the flights operated by Airline B, 75% are on time.

Q1 What is the probability that a random flight was operated by airline B?

Q2 If a random flight is on-time, what is the probability the flight was operated by airline B?

A1 *prior probability*: $P(\text{airline B}) = 0.45$

A2 *posterior probability*: $P(\text{airline B} \mid \text{on-time}) = 0.4054$

The *prior* probability (before getting additional information) was 0.45, but the *posterior* probability (after observing the "data") is updated to 0.4054.

Let's recap ...

1. We had the probability of a random flight being airline B. (prior probability)
2. We had conditional probabilities of a flight being late, given the airline. (likelihood)
3. We observe the punctuality of the flight. (data)
4. We used these to come up with the conditional probability of the flight being airline B given the punctuality. (posterior probability)

Bayesian Approach to Statistics

Bayesian Approach to Statistics

Concept of inverse probability applies more generally.

- Have an assumed parametric model: $f(data|parameters)$.
  - The **Likelihood** of data (given specific values of the parameters).
- **Prior** information about the parameters expressed via a distribution reflecting belief; i.e., $\pi(parameters)$
- Use observed data to determine **posterior** information about parameters. Often described as updating:

$$\pi(parameters|data) = \frac{f(data|parameters)\pi(parameters)}{f(data)}$$

Compare with $Pr(B|A) = \frac{Pr(A|B)Pr(B)}{Pr(A)}$.

We start with some beliefs, the data then affect our beliefs, and we iteratively update beliefs as new data becomes available.

Bayesian Updating

The cycle of learning from data will be drawn below:

To be a Bayesian, we ...

- Identify a question of interest
- Consider the form of the data to be collected
- Thoughtfully select a likelihood (model for the data)
- Elicit the prior distribution for the likelihood's parameters.
- Collect/observe the data
- Update the prior by blending the prior and likelihood via Bayes' rule to form the posterior distribution
- Make inferences on parameters and/or predictions on future observations.

Which steps are unique to Bayesians?

Key Points

- Stronger prior beliefs $\Rightarrow$ less influence of data on posterior.
- More data $\Rightarrow$ less influence of prior.
- Prior influences posterior for better or worse.

Bayesian vs. Frequentist

Common to use "frequentist" or classical to describe statistical techniques that do not adhere to the Bayesian paradigm.

- Bayesians explicitly include prior beliefs
- Bayesians make inference based on observed data, not all data that could have been observed
- Frequentists compare observed data with behavior of statistics in repeated sampling
- Bayesians use probability as "fundamental measure or yardstick of uncertainty" (Gelman, Carlin, Stern, and Rubin (2004) p. 11)

See also Louis Lyons' 2007 SLAC presentation slides (especially last three) at `http://blogs.sch.gr/ggrammat/files/2014/11/lyons_colloquium_Stat2007.pdf`.

Statistical Ethics—Soapbox time

Common critique against Bayesians: prior is not objective ("My answer is different from yours!")

Rebuttal 1: Methods exist to obtain "default" priors

Rebuttal 2: If you already have some idea, why ignore that information?

Rebuttal 3: Not everyone necessarily has the same information.

Rebuttal 4: Bayesians necessarily must be explicit in likelihood and prior (C.S. Reese)

Rebuttal 5: When using hypothesis tests, penalty for overoptimistic prior (V.E. Johnson)

Big picture outline for process in remainder of course

- Select likelihood (model)
- Select prior
- Determine posterior distribution
- Make inference with posterior distribution.

Notation: We will use

- $\mathbf{y}$ to generically represent the *data*,
- $\boldsymbol{\theta}$ to generically represent the *parameters*,
- $f(\mathbf{y}|\boldsymbol{\theta})$ to represent the *likelihood* of the data given the parameters
- $\pi(\boldsymbol{\theta})$ to represent the *prior distribution* of the parameters
- $\pi(\boldsymbol{\theta}|\mathbf{y})$ to represent the *posterior distribution* of the parameters given the observed data.

Selecting Likelihood, $f(\mathbf{y}|\theta)$

- Virtually always necessary (Bayesian or not)
- Binomial distribution for success/failure data
- Multinomial distribution for categorical data
- Poisson distribution for counts across time/space.
- Gamma distribution for certain types of positive data
- Normal distribution for bell-curve data

Selecting Prior Distribution, $\pi(\theta)$

Strategies for choosing prior

- Objective (by using default prior)
- Subjective (by using substantive beliefs)

Form of prior distribution:

- Convenience (by using conjugate prior)
- Can be VERY flexible

Caution: Some priors lead to improper posteriors, meaning posterior inference not possible.

Posterior Distribution, $\pi(\boldsymbol{\theta}|\mathbf{y})$

Likelihood, data, and prior will determine posterior.

$$Posterior = \frac{Likelihood \times prior}{marginal\ likelihood}$$

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\mathbf{y})}$$

Note: the denominator (the unconditional, or *marginal*, likelihood of the data) is

$$f(\mathbf{y}) = \begin{cases} \int_{\boldsymbol{\Theta}} f(\mathbf{y}, \boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\boldsymbol{\Theta}} f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}, & \text{if } \theta \text{ continuous,} \\ \sum_{\boldsymbol{\Theta}} f(\mathbf{y}, \boldsymbol{\theta}) = \sum_{\boldsymbol{\Theta}} f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), & \text{if } \theta \text{ discrete.} \end{cases}$$

Posterior Distribution (cont.)

To determine, can use:

- Exact derivation,
- Mathematical Approximation, or
- Monte Carlo (simulation).

Summarizing the Posterior Distribution

Once we have the posterior distribution, what do we do with it?

The entire distribution has value.

Summarizing the Posterior Distribution

Once we have the posterior distribution, what do we do with it?

The entire distribution has value.

But if we insist on a summary measure, a natural choice is the expected value (i.e., mean) of the posterior distribution.

Other choices are the distribution's mode, median, standard deviation, and variance, as well as the probability that θ is in a particular region of interest.

Summarizing the Posterior Distribution

Once we have the posterior distribution, what do we do with it?

The entire distribution has value.

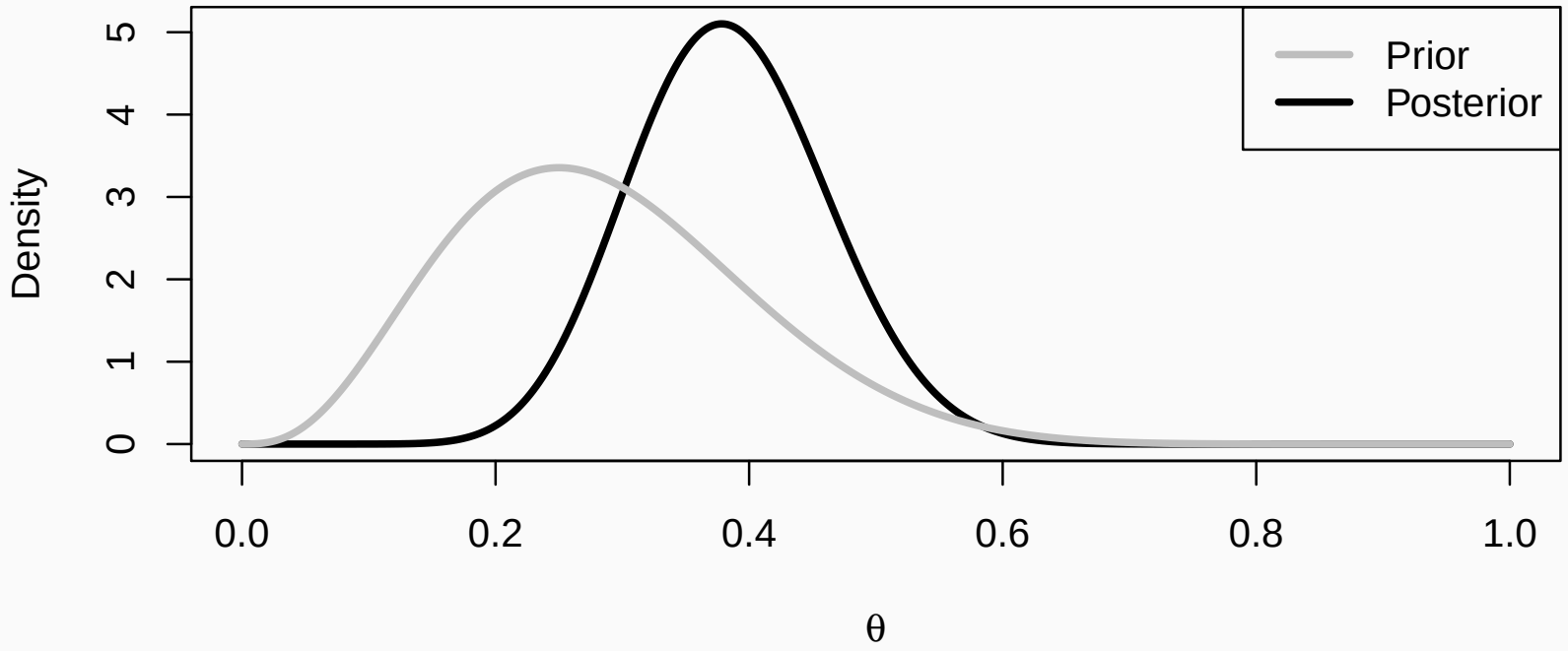
But if we insist on a summary measure, a natural choice is the expected value (i.e., mean) of the posterior distribution.

Other choices are the distribution's mode, median, standard deviation, and variance, as well as the probability that θ is in a particular region of interest.

Popular to also create a **credible interval**: an interval with a specified probability of containing θ.

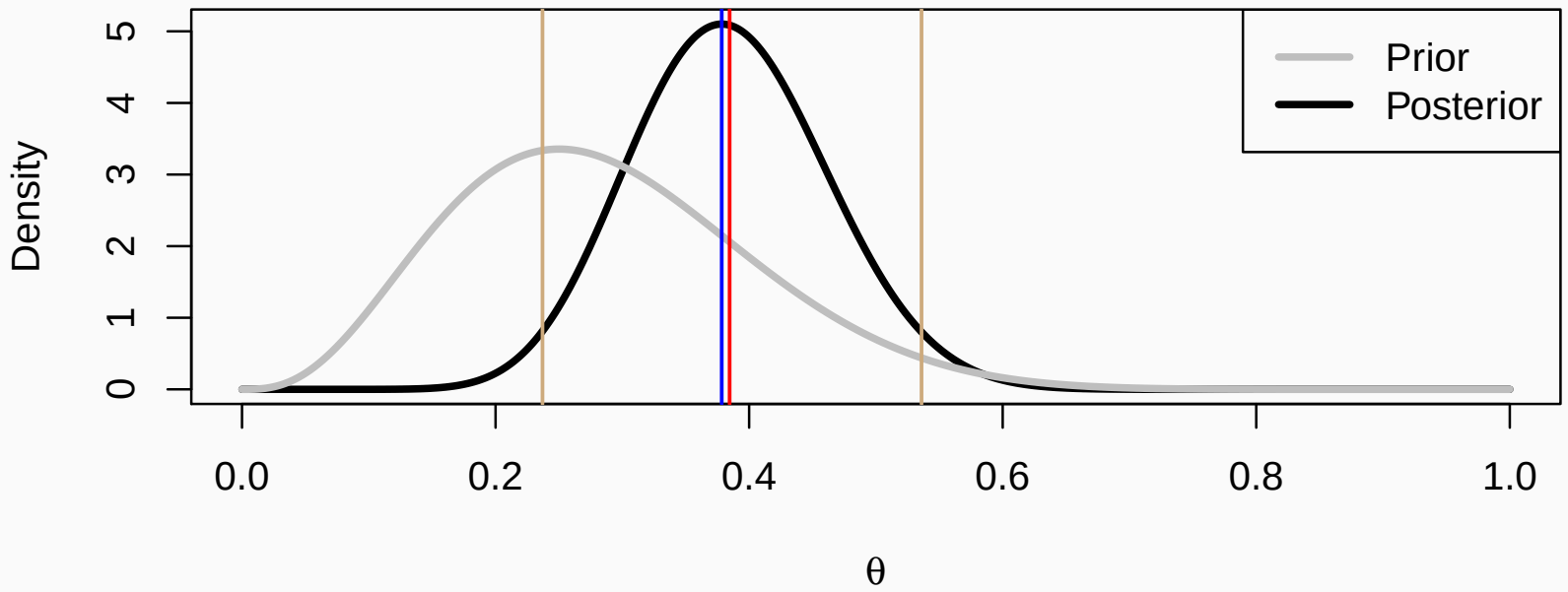
Example

In a particular setting**, we would move from the prior distribution depicted below to the posterior distribution below.



** Having a Beta(4,10) prior on $\theta \equiv$ probability of success, and then observing 11 successes in 25 trials, produces a Beta(15, 24) posterior for $\theta|\mathbf{y}$. More on this in Unit 4.

Posterior Summaries



It can be shown that for this posterior,

posterior mean $= E(\theta|\mathbf{y}) = \int_\theta \pi(\theta|\mathbf{y})d\theta = \frac{15}{39} = 0.3846$

posterior mode of $\pi(\theta|\mathbf{y}) = \frac{14}{37} = 0.3783$

95% posterior credible interval: (0.237, 0.536)

Useful R code for Overlaying Densities in a Plot

R Code to create previous graphic

```
curve(dbeta(x,15,24), from=0, to=1, n=1001,
      type="l", xlab=expression(theta),
      ylab="Density", lwd=3)

lines(curve(dbeta(x,4,10), from=0, to=1,
            n=1001, add=TRUE), col="gray", lwd=3)

legend("topright", c("Prior", "Posterior"),
       col=c("gray", "black"), lwd=3)

abline(v=(15/39), col="red", lwd=1.5)
abline(v=14/37, col="blue", lwd=1.5)
abline(v=c(.237, .536), col="burlywood3", lwd=1.5)
```

Alternate R Code to create previous graphic

```
 xx <- seq(0, 1, length.out=1001)
 plot(xx, dbeta(xx, 15, 24),
     type="l", xlab=expression(theta),
     ylab="Density", lwd=3)

lines(xx, dbeta(xx, 4, 10), col="gray", lwd=3)

legend("topright", c("Prior", "Posterior"),
       col=c("gray", "black"), lwd=3)

abline(v=(15/39), col="red", lwd=1.5)
abline(v=14/37, col="blue", lwd=1.5)
abline(v=c(.237, .536), col="burlywood3", lwd=1.5)
```

Useful R functions

plot(, , type="l") # see Unit 2C for practice

lines(, , col="color1") # Add connect-the-dots lines to plot

abline(v=num) # add vertical line to existing plot at x-axis=num

abline(v=c(num1, num2)) # add two vertical lines, one at num1, other at num2

legend("topright", c("text1", "text2"), col=c("color1", "color2"), lwd=linewidth)

add legend at top-right of plot, with two lines of text (text1, then text2)

and with two lines that are *linewidth* wide

having *color1*, then *color2*.

expression(theta) # To display in plot as θ

Remember: Big-Picture Process for Remainder of Course

- We have a desire to learn more

  to better understand a data-generating process, or to make better predictions, or to assess relationships between variables
- We will use data to help us learn
- We assume data follow a statistical model (i.e., distribution) that depends on parameter(s); we call this distribution the **likelihood**.
- We express our current state of belief/knowledge about parameter(s) with a separate statistical distribution; the **prior distribution**
- We use a generalization of Bayes's rule to update our current state of belief/knowledge about parameter(s) in light of the data; we call this the **posterior distribution**.