Introduction to Bayesian Data Analysis

Graduate Workshop on Environmental Data Analytics

Alix I. Gitelman

Statistics Department
Oregon State University
gitelman@science.oregonstate.edu

July 25, 2016

Overview

- Let's Play Cards
- Statistical Paradigms
  - The Frequentist Paradigm
  - The Bayesian Paradigm
  - Frequentist vs Bayesian
- A Simple Example
  - The Bayesian Approach
  - Bayes Theorem (Preview)
- Notation
- Bayes Theorem
  - Discrete Probabilities
  - A More General Rule
- The Binomial Problem

Paradigm (definitions)

> *"a philosophical and theoretical framework of a scientific school or discipline within which theories, laws, and generalizations and the experiments performed in support of them are formulated"*

–Merriam-Webster online

> *"the set of common beliefs and agreements shared between scientists about how problems should be understood and addressed"*

–Thomas Kuhn

What's the Frequentist Paradigm?

Specifically, what are the "beliefs and agreements" about how statistical problems should be "understood and addressed?"

Let's back up even further: what are "statistical problems?" also called inferential problems

- **estimation**
- **hypothesis testing**

In general, frequentists do estimation and testing

- **using sampling distributions**

The Frequentist Paradigm

But what is a sampling distribution?

...somewhere along the way the word "frequency" ought to come up...

- The sampling distribution of a particular statistic, T_n, is the relative frequency distribution (think histogram) of T_n constructed from repeated samples of size n from the population of interest.

The frequentist paradigm relies on the notion of relative frequency probability for dealing with statistical problems.

Relative Frequency Probability

Suppose that A is one possible outcome among a finite set of possible outcomes of some experiment E. Then $Pr(A)$ is defined as the relative frequency of A's occurrence in an infinite sequence of repeated experiments, E.

The hypotheticals in that definition:

1. We need to assume that it's possible to run E more than once, let alone an infinite number of times...
2. We need to wait around for an infinity of trials...

How Frequentist Modeling Works

In the frequentist approach we:

1. Sample data from an unknown population distribution
2. Condition on a true, fixed, *unknown* quantity or set of quantities (called parameters)
3. Determine how unusual a sample we have relative to all other possible samples from a null hypothesized population distribution

The Bayesian Paradigm

Like the frequentist paradigm, the Bayesian paradigm has at its heart a notion of probability.

- Subjective probability is defined in terms of bets: A probability p attached to an event E is defined as the fraction $p \in [0, 1]$ at which you would bet p cents for a return of \$1 if E occurs.

Notice that this means that I might assign a different probability to a certain event than you might.

How Much Would You Bet?

Here's the setup:

- I will draw cards, one at a time from one of the two decks.
- Before I draw a card, you must place a bet on whether the card will be red or black.
- Your bet consists of standing on one foot for 0 to 10 seconds—if you stand for 0 seconds, you're certain that a red card will NOT appear, if you stand for 10 seconds, you're certain that a red card WILL appear.
- The return on your bet if you win is a bag of M&M®
- Volunteers?

The Bayesian Paradigm

In the Bayesian approach, we:

1. Specify a prior model for the parameter(s) of interest
2. Obtain data from a relevant population
3. Condition on the observed data to update our prior model to a **posterior model** that we use to make inference.

Notice the similarity to the Scientific Method

Frequentist vs Bayesian

In terms of making inferences there are two essential differences between classical (frequentist) and Bayesian statistics:

1. How we think about parameters (fixed versus random)
2. How we think about probability (long-run frequency versus “subjective”)

To a frequentist, parameters are “true, fixed, unknown” quantities.

By contrast, a Bayesian models uncertainty about parameters. In that regard, parameters are thought of as random.

Frequentist vs Bayesian

To Frequentists:

- Parameters are fixed
- Data are random

To Bayesians:

- Parameters are random
- Data are obtained through some random mechanism, but in an analysis, they are treated as fixed

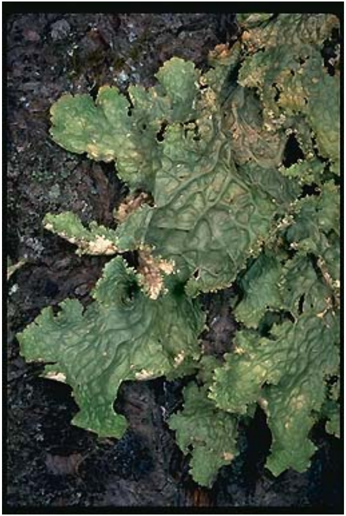
Lichen Presence/Absence

Surveying the diversity and abundance of lichens in forests can be useful for several reasons:

1. To assist in classification of stands as “old-growth”
2. To evaluate climate conditions and effects
3. To evaluate stand health vis-a-vis airborne pollution.

In an oversimplification of a lichen survey, we’ll look at the presence/absence of one species, *lobaria oregana* or “lettuce lichen,” which is relatively common in the Oregon Cascades.

Lettuce Lichen



Lichen Presence/Absence

Suppose in a given forest stand in the Oregon Cascades, we obtain a sample of $n = 57$ spatially distinct trees.

1. Each tree is evaluated for the presence/absence of a particular lichen species, *lobaria oregana*, or “lettuce lichen.”
2. We’ll assume that being “spatially distinct” is enough to ensure that the observations are statistically independent.
3. In all, $X = 22$ of the trees have the lichen present.

The Binomial Probability Mass Function

A reasonable probability model for these data is the Binomial probability mass function (pmf):

$$Pr(X = x|\pi) = \binom{57}{x} \pi^{x}(1-\pi)^{57-x}$$

for $x = 0, 1, \ldots, 57$ and $0 \leq \pi \leq 1$.

Here, π is the population probability of presence.

In general, the Binomial pmf is written,

$$P(X = x|\pi) = \binom{n}{x} \pi^{x}(1-\pi)^{n-x}$$

for $x = 0, 1, \ldots, n$ and $0 \leq \pi \leq 1$.

Frequentist Approach

How would a frequentist analyze these data?

- estimate π with $\hat{\pi} = 22/57 = 0.3860$
- find $SE(\hat{\pi}) = \sqrt{\hat{\pi}(1-\hat{\pi})/57} = 0.0645$
- identify the sampling distribution for $\hat{\pi}$.

Some questions:

1. Why is $\hat{\pi} = x/n$ a good estimate for π?
2. What do we use as a sampling distribution?

Frequentist Approach

In the frequentist approach, we often use **maximum likelihood estimates** or MLE.

- In this case the MLE for π is $\hat{\pi} = 22/57$, the number of trees with the lichen divided by the total number of trees in the sample.

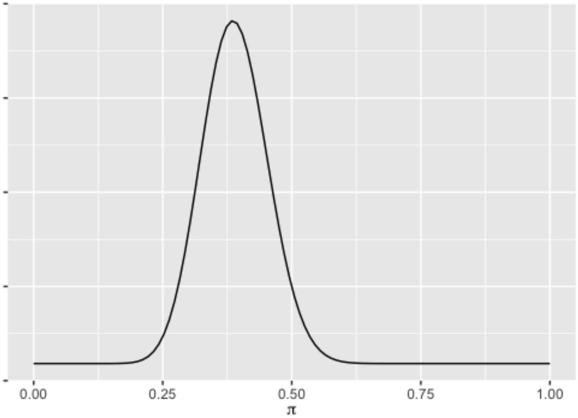
- We appeal to the Central Limit Theorem to use the Normal distribution as the sampling distribution for $\hat{\pi}$.

- In R:

```
> prop.test(22,57)

data:  22 out of 57, null probability 0.5
X-squared = 2.5263, df = 1, p-value = 0.1120
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.2629116 0.5243842
```

The Likelihood Function



Bayesian Approach

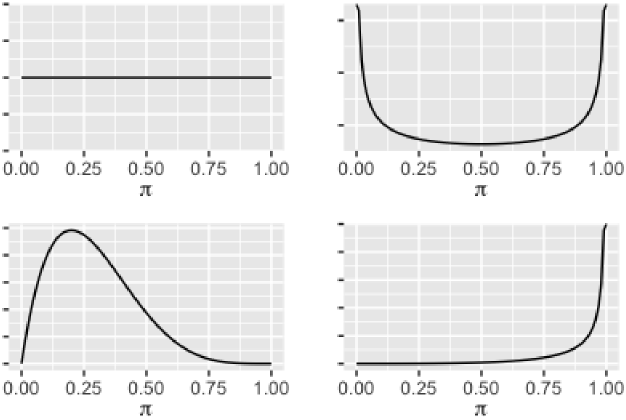
Here, we'll again use the Binomial pmf for X:

$$P(X = x|\pi) = \binom{n}{x}\pi^x(1-\pi)^{57-x}$$

And now, we have to specify a prior probability model for π.

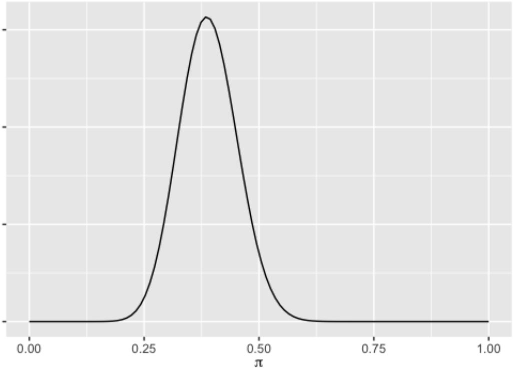
How to do this? Can we draw a picture? What if we had some prior information?

Some Possible Prior Distributions



Bayesian Approach

Using Bayes Theorem (details later), I now combine the likelihood model and uniform (or non-informative or reference) prior model to obtain the **posterior distribution** for π; namely, $f(\pi|X)$.



Bayesian Approach

Using this **posterior distribution**, I can report the **posterior mean** of π and a **95% posterior interval** for π (this is the Bayesian analog to the frequentist confidence interval):

- The posterior mean, mean$(\pi|X) = \tilde{\pi} = 0.40$.
- A 95% posterior interval is (0.28,0.52).

Compare this to the frequentist estimate, $\hat{\pi} = 0.38$ and confidence interval: (0.26,0.52).

Interpretations

For the frequentist confidence interval:

> *"In 95% of repeated samples, the 95% confidence interval for π will cover π."*

Notice a couple of things:

1. We don't say anything about **this particular** interval, we just have to make a general statement about hypothetical intervals like this one.

2. The language "will cover π" reflects the fact that the confidence interval is a probability statement about the sample proportion, $\hat{\pi}$; it's not a probability statement about π.

Interpretations

For the Bayesian posterior interval:

> *"Under a uniform prior for π, and given the observed data, the probability that π is between 0.28 and 0.52 is 95%."*

Some things to notice:

1. This is a statement about **this** interval, not some hypothetical collection of intervals.
2. The probability statement is about π, the parameter we actually want to make inference about.

A Comment on Notation

Suppose that $(X_1, \ldots, X_n) \equiv \mathbf{X}$ is a sample of data, and let $f(\mathbf{X}; \theta)$ denote the joint probability distribution (pdf or pmf) of $\mathbf{X}$.

- This joint pdf is also called the **likelihood function**, and it's used in both frequentist and Bayesian statistics (you've probably encountered maximum likelihood estimates).
- In Bayesian statistics, we think of the likelihood function as the joint probability distribution of the data, *conditioned* on θ, and so we typically write the likelihood as $L(\mathbf{X}|\theta) \equiv f(\mathbf{X}; \theta)$.
- I'll adopt the convention of using square brackets to denote probability distributions, so instead of $L(\mathbf{X}|\theta)$ I'll write simply $[\mathbf{X}|\theta]$.

A Comment on Notation

The “pipe” $(\cdot|\cdot)$ is a notation to denote conditioning.

- $Pr(A|B = b)$ is read “the probability of A given that $B = b$”
- That is, in $Pr(A|B = b)$, we take B to be a fixed, known value, b
- To write $X \sim N(0, 1)$ is to say that

$$f(X|\mu = 0, \sigma = 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

To write $X \sim N(10, 2)$ is to say that

$$f(X|\mu = 10, \sigma = 2) = \frac{1}{\sqrt{8\pi}} e^{-\frac{(x-10)^2}{8}}$$

Bayes Theorem: Discrete Probabilities

Suppose that A and B are discrete events where each can take a finite number of possible values; that is, $A = a$ for $a \in \{a_1, a_2, \ldots, a_k\}$ and $B = b$ for $b \in \{b_1, b_2, \ldots, b_\ell\}$.

Bayes Theorem:

$$\begin{aligned} P(A = a|B = b) &= \frac{P(B = b|A = a)P(A = a)}{\sum_{j=1}^{k} P(B = b|A = a_j)P(A = a_j)} \\ &= \frac{P(B = b|A = a)P(A = a)}{P(B = b)} \end{aligned}$$

provided that $P(B = b) \neq 0$.

Bayes Theorem: Discrete Probabilities

Bayes Theorem:

$$P(A = a|B = b) = \frac{P(B = b|A = a)P(A = a)}{P(B = b)}$$

provided that $P(B = b) \neq 0$.

Notice that the **order of conditioning** changes from the right hand side $[P(B = b|A = a)]$ to the left hand side $[P(A = a|B = b)]$

Bayes Theorem: Generalization

Bayes rule also applies to probability distribution functions.

In what follows:

- $[\mathbf{X}|\theta]$ denotes the **likelihood function** of a sample of data, $\mathbf{X}$, conditional on some parameter(s), θ.
- $[\theta]$ is the **prior distribution** of θ
- $[\theta|\mathbf{X}]$ is the **posterior distribution** of θ given the sample of data, $\mathbf{X}$

Bayes Theorem: Generalization

Bayes Theorem applied to probability distribution (mass) functions:

General Version of Bayes Theorem:

$$[\theta|\mathbf{X}] = \frac{[\mathbf{X}|\theta][\theta]}{\int[\mathbf{X}|\theta][\theta]d\theta}$$

provided some regularity conditions are met.

The regularity conditions essentially ensure that the denominator is finite (this is analogous to the condition that $P(B = b) \neq 0$ in the discrete case).

Bayesian Modeling

In a Bayesian model:

1. We specify a probability model for the observed data—this is the likelihood function;
2. we specify a prior probability model for the parameters of interest (i.e., the parameters of the likelihood model);
3. and, we use Bayes theorem to combine the likelihood and the prior to obtain the posterior distribution.

The general idea is that the prior distribution models our current view of the parameters, and the posterior distribution is an update to that prior distribution using the available data.

Back To Bayes Theorem

Bayes Theorem:

$$[\theta|\mathbf{X}] = \frac{[\mathbf{X}|\theta][\theta]}{\int[\mathbf{X}|\theta][\theta]d\theta}$$

provided some regularity conditions are met.

- We have to deal with the denominator term, which can be daunting, especially if θ is multi-dimensional.
- Notice, however, that $\int[\mathbf{X}|\theta][\theta]d\theta$ does not depend on θ—that's the whole point of the integration, to get rid of (average over) θ.

Normalizing Constant

Therefore, in Bayes Theorem, since the left hand side, $[\theta|\mathbf{X}]$, is a function of θ, and the denominator on the right hand side is a constant with respect to θ, we can think of the denominator as a **scaling factor** or **normalizing constant** and write:

$$\begin{aligned}[\theta|\mathbf{X}] &= \frac{[\mathbf{X}|\theta][\theta]}{[\mathbf{X}]} \\ &\propto [\mathbf{X}|\theta][\theta]\end{aligned}$$

That is: **posterior** $\propto$ **likelihood** $\times$ **prior**.

(the symbol $\propto$ is read: "is proportional to")

Back to the Binomial Problem

For our Binomial problem with $X = 22$ and $n = 57$ we have

$$[X = 22|\pi] \equiv L(X = 22|\pi) = \binom{57}{22}\pi^{22}(1-\pi)^{35}.$$

And, using a Uniform(0,1) prior for π, we have

$$[\pi] = 1 \text{ for } 0 < \theta < 1.$$

So that

$$\begin{aligned} [\pi|X = 22] &= \frac{\binom{57}{22}\pi^{22}(1-\pi)^{35}}{\int_0^1 \binom{57}{22}\pi^{22}(1-\pi)^{35} d\pi} \\ &\propto \pi^{22}(1-\pi)^{35} \end{aligned}$$

The Binomial Problem

It turns out that:

$$f(\pi|X=22) \quad \propto \quad \underbrace{\pi^{22}(1-\pi)^{35}}_{\text{kernel of a Beta distribution}}$$

That is, the posterior distribution has a known, closed form—it's a Beta distribution with parameters $\alpha_n = 23$ and $\beta_n = 36$ (More on the Beta distribution in the next session).

If there's time: show LichenDemo of posterior distribution for π in R

Next Steps

In our next session, you'll do some activities in R:

- Run through LichenDemo.R of the posterior distribution for π
- Beta prior distributions (Shiny demo, Brian Reich)

And, I'll give an introduction to MCMC:

- What's it all about?
- Heuristics of how it works
- Writing Bayesian models in BUGS code