Idea of Statistical Modeling

ISTA 410 / INFO 510 - Bayesian Modeling and Inference

University of Arizona School of Information

August 25, 2021

Outline

Outline for today:

- Goals of Bayesian analysis
- Example: kidney cancer death rates
- Types of uncertainty
- More involved example: vaccine effectiveness

Motivation: Bayesian analysis

Goal: analyze and quantify uncertainty

- uncertain quantities get a probability distribution
- probability distribution is updated based on new observations

Bayesian approach:

- Named for Thomas Bayes – English minister in the 18th century
- Considered the problem of *inverse probability*
- Didn't invent the whole theory, but was one of the earliest to solve a problem with it (along with Laplace)

Generative probabilistic models

Core component: generative models

- given values of model parameters, can generate outcomes
    - given a value for the probability a coin comes up heads, we can simulate a sequence of flips
    - given a mean and standard deviation, we can simulate normally distributed values

Generative probabilistic models

Core component: generative models

- given values of model parameters, can generate outcomes
    - given a value for the probability a coin comes up heads, we can simulate a sequence of flips
    - given a mean and standard deviation, we can simulate normally distributed values

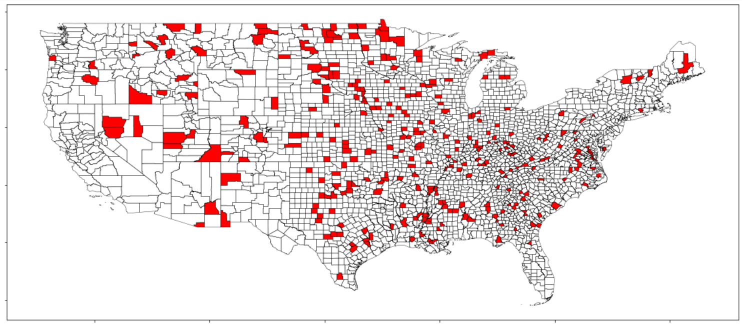
The inverse problem is: given the outcomes, infer a probability distribution for the parameters

- Both Bayes's and Laplace's early work deal with a binomial model (like the coin flip)

Case study: kidney cancers

Where is kidney cancer highest?

The following map shows the counties with the highest 10% death rates due to kidney cancer (1980-89).



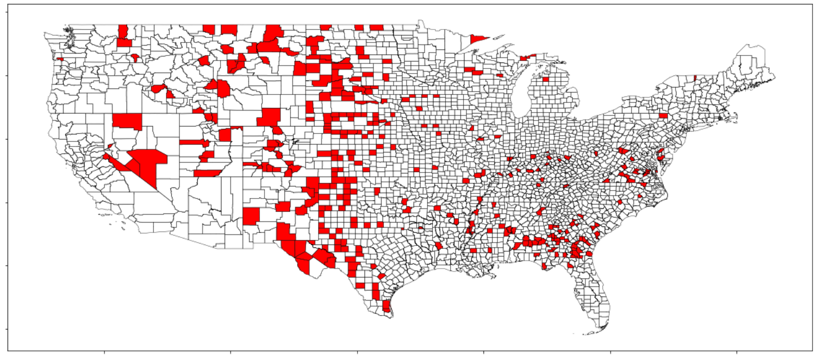
What do we notice?

What we can notice: most counties in the middle of the country, not coasts.

Why?

Where is kidney cancer lowest?

The following map shows the counties with the *lowest* 10% death rates due to kidney cancer (1980-89).



Explaining both of these

It would be nice if we could explain both of these features at once:

Explaining both of these

It would be nice if we could explain both of these features at once:

Possible explanation: sample size

Explaining both of these

It would be nice if we could explain both of these features at once:

Possible explanation: sample size

- Rates in small samples are more variable than larger samples
- Kidney cancer is a rare cause of death, and USA has a lot of very low population counties
- A county with 1000 people is likely to record zero deaths
- A county with 1000 people that records 1 death jumps to a rate of 100 deaths per 100,000 people, easily enough to jump to the top 10%

A simple model

To see if the sample size effect is enough to explain this, we can try a generative model.

- Model is a procedure for generate simulated versions of our observed outcomes
- What are our outcomes? Deaths due to kidney cancer.
- Idea: establish a minimal model, see if it replicates the qualitative behavior of the real data
- We need to pick a probability distribution for our outcomes

Common choice for this sort of count: Poisson distribution

Poisson distribution

- Defined on natural numbers $\{0, 1, 2, \ldots\}$
- Models a count of events occurring independently at a fixed rate
- Depends on a rate parameter most often written λ

Probability mass function:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Our model

We'll make the simplest possible assumption: there is no effect of geography on kidney cancer, and the only relevant feature of a county is population.

So, our model has one parameter, θ (the underlying death rate).

Then, the death count in each county, y_j, follows a Poisson distribution:

$$y_j \sim \mathrm{Poisson}(\theta n_j)$$

where n_j is the population of the county. (In Poisson models this scaling factor is sometimes called an *exposure*.)

Let's try simulating...

Case study for inference

Inference with a generative model

Previous example:

- Generative procedure allowed us to explain one of the qualitative features of the data set
- However, we didn't set the model up to do any *inference*, e.g. estimating the actual death rate
  - we used a fixed value of θ
  - we used the same θ for every county
- In practice, we often want to estimate un-observed parameters

Using a Bayesian model

Steps:

- Set up a probabilistic model for the observed data, dependent on un-observed parameters
- Apply a *prior distribution* to the parameters, representing our knowledge before observing data
- Apply Bayes' theorem to update the distribution of the parameters, resulting in a *posterior distribution*
- Summarize relevant results

Bayes' theorem

Recall Bayes' theorem:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Terminology:

- $P(H|E)$ – *posterior probability*
- $P(H)$ – *prior probability*
- $P(E|H)$ – *likelihood*
- $P(E)$ – *normalizing constant*

Cookies

Variant on example from last time: cookie problem. Bowl 1 has 1/2 chocolate, 1/2 vanilla cookies; Bowl 2 is the same; Bowl 3 has 3/4 vanilla, 1/4 chocolate. We draw a chocolate cookie; what's the probability we are drawing from bowl 3?

Cookies

Variant on example from last time: cookie problem. Bowl 1 has 1/2 chocolate, 1/2 vanilla cookies; Bowl 2 is the same; Bowl 3 has 3/4 vanilla, 1/4 chocolate. We draw a chocolate cookie; what's the probability we are drawing from bowl 3?

$$P(\text{Bowl 3}|\text{chocolate}) = \frac{P(\text{chocolate}|\text{Bowl 3})P(\text{Bowl 3})}{P(\text{chocolate})}$$

Cookies

Variant on example from last time: cookie problem. Bowl 1 has 1/2 chocolate, 1/2 vanilla cookies; Bowl 2 is the same; Bowl 3 has 3/4 vanilla, 1/4 chocolate. We draw a chocolate cookie; what's the probability we are drawing from bowl 3?

$$P(\text{Bowl } 3|\text{chocolate}) = \frac{P(\text{chocolate}|\text{Bowl } 3)P(\text{Bowl } 3)}{P(\text{chocolate})}$$

- $P(\text{Bowl } 3) =$
- $P(\text{chocolate}|\text{Bowl } 3) =$
- $P(\text{chocolate}) =$

Cookies

We can plug the numbers in:

$$P(\text{Bowl 3}|\text{chocolate}) = \frac{(1/4) \times (1/3)}{5/12} = 1/5$$

Cookies

We can plug the numbers in:

$$P(\text{Bowl 3}|\text{chocolate}) = \frac{(1/4) \times (1/3)}{5/12} = 1/5$$

More commonly, our parameters are not discrete (bowl 1 vs. 2 vs. 3) but continuous.

Differences:

- instead of finitely many "hypotheses", any allowed value of
- we have to work with probability density functions

Bayes' theorem with densities

Most commonly we have probability density functions that depend on unknown parameters:

- y – data
- θ – parameters

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

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

The normalizing constant $p(y)$ is gotten by marginalizing over θ by computing $\int p(y|\theta)p(\theta)d\theta$; this integral may be intractable, so we work with the proportionality statement.

Binomial model

If we are observing binary categorical outcomes, a binomial likelihood makes sense. $\mathrm{Binomial}(n, \theta)$ is the distribution of the count of "successes" in n independent trials with a fixed probability θ of success.

$$p(y \text{ successes}|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

A continuous prior

A common choice of prior for a binomial likelihood is a beta distribution:

$$\theta \sim \mathrm{Beta}(\alpha, \beta)$$

where $\alpha, \beta > 0$ are chosen ahead of time.

Beta distribution: defined on $[0, 1]$ by the PDF

$$p(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

$B(\alpha, \beta)$ is the normalizing constant, called a *Beta function*. There are formulas for it but not important for us right now.

What is the data-generating process?

The generative procedure now:

1. Draw a value of θ from $\mathrm{Beta}(\alpha, \beta)$
2. Draw a value of y from $\mathrm{Binomial}(n, \theta)$

What is the data-generating process?

The generative procedure now:

1. Draw a value of θ from $\mathrm{Beta}(\alpha, \beta)$
2. Draw a value of y from $\mathrm{Binomial}(n, \theta)$

- The cookie problem: y's distribution is a finite mixture of binomials, with equal weight
- Now: y's distribution is an infinite mixture of binomials, weighted by the PDF of θ

Conjugate prior

One reason for the choice of beta prior: *conjugacy*

A distribution $p(\theta)$ is conjugate to a likelihood $p(y|\theta)$ if the posterior distribution $p(\theta|y)$ is a member of the same family as $p(\theta)$:

In the beta-binomial model,

$$p(\theta|y) = \frac{1}{p(y)} \frac{1}{B(\alpha,\beta)} p(k|\theta) \begin{pmatrix} n \\ k \end{pmatrix} \theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

The leading three factors don't depend on θ, so we absorb them into a single constant.

Conjugate prior

Now:

$$p(\theta|y) = \frac{1}{Z}\theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$
$$= \frac{1}{Z}\theta^{\alpha+y-1}(1-\theta)^{\beta+(n-y)-1}$$

Since the dependence of the density on θ is that of a beta distribution with parameters $(\alpha + y, \beta + (n-y))$, the constant Z must be the corresponding beta function, and

$$\theta|y \sim \mathrm{Beta}(\alpha + y, \beta + (n-y))$$

Computationally very convenient! Convenience less important these days than it used to be, though.

Posterior distribution

So, in a beta-binomial model:

$$
\begin{aligned}
y &\sim \text{Binomial}(n, \theta) \\
\theta &\sim \text{Beta}(\alpha, \beta)
\end{aligned}
$$

if we observe y successes and $n - y$ failures, the posterior distribution of θ is

$$\theta | y \sim \text{Beta}(\alpha + y, \beta + (n - y))$$

Posterior distribution

So, in a beta-binomial model:

$$y \sim \mathrm{Binomial}(n, \theta)$$
$$\theta \sim \mathrm{Beta}(\alpha, \beta)$$

if we observe y successes and $n - y$ failures, the posterior distribution of θ is

$$\theta | y \sim \mathrm{Beta}(\alpha + y, \beta + (n - y))$$

Interpretation: we may think of the prior parameters $\alpha - 1, \beta - 1$ as *pseudocounts*

Summarizing inferences; example

Inferences from the posterior

The posterior distribution is the primary product of inference; it contains all that we know about the parameter after incorporating prior and data.

In practice, often want to distill out some summary statistics:

- posterior mean – expected value of θ under the posterior distribution
- posterior intervals – 95% common, but arbitrary. Note difference between highest density and central intervals
- maximum a posteriori estimate – often not a good choice, especially if the model has many parameters

Example: Pfizer's vaccine trial

Prominent recent example: beta-binomial model in analysis of Pfizer's COVID-19 vaccine

Trial procedure:

- Study participants divided randomly into two "arms": control/placebo and vaccine
- Control arm given placebo, vaccine arm given vaccine
- Watch both groups and count cases, running the analysis when a predetermined number of cases is observed

Beta-binomial model

Defining parameters:

- π_c: probability that a control subject becomes ill
- π_v: probability that a vaccinated subject becomes ill
- Derived quantity: Vaccine efficacy:

$$VE = 1 - \frac{\pi_v}{\pi_c}$$

Parameter for the model:

$$\theta = \frac{1 - VE}{2 - VE} = \frac{\pi_v}{\pi_v + \pi_c}$$

Measures the probability that a case came from the vaccine arm

Pfizer's prior

Let y be the number of cases that come from the vaccinated group.

The model:

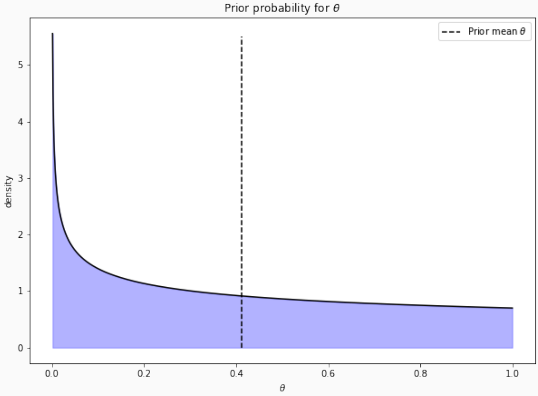
$$y \sim \mathrm{Binomial}(\theta, n)$$
$$\theta \sim \mathrm{Beta}(0.700102, 1)$$

Prior was stated in Pfizer's press release. No specific reason given for these parameters, but:

- VE at prior mean θ is 30%
- fairly uninformative: 95% interval is about $(-26.2, 0.995)$.

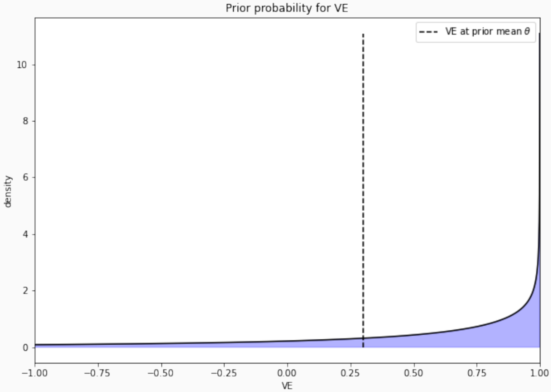
Pfizer's prior

On the θ scale:



Pfizer's prior

On the VE scale



What's the data?

The result of the study submitted to the FDA to obtain an emergency use authorization had a total of 170 observed cases, 8 of which were in the vaccine arm. So:

$$\theta | y \sim \mathrm{Beta}(0.700102 + 8, 1 + 162)$$

Let's examine this graphically...

Next week

Next week:

- More models
- Going beyond conjugate priors with various approximations