Anatomy of Bayesian inference

ISTA 410 / INFO 510 - Bayesian Modeling and Inference

University of Arizona School of Information

August 26, 2020

Outline for today

- Review basic components of Bayesian inference
- Simplest possible example
- Summarizing posterior inferences
- Predictive distributions
- Return to the kidney cancer example

HW 1 on D2L, due 9/04 11:59 PM

The basic components

Priors, likelihoods, posteriors

A Bayesian model provides a way to express the joint probability distribution for observed data y and an unobserved parameter θ as a product of two densities:

$$p(y, \theta) = p(\theta) p(y|\theta)$$

- $p(\theta)$ – *prior density* for θ
- $p(y|\theta)$ – *sampling distribution* or *data distribution*

Passing from prior to posterior

We do inference by conditioning on some known data to get:

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

(Bayes' rule)

Un-normalized version:

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

We think of the right hand side as a function of θ, not y, in which case $p(y|\theta)$ is called the *likelihood function*.

A simple binomial model

The unfair coin

We have a coin that, when spun on its edge, falls on the “heads” side with some unknown probability θ. In our experiment, we spin it n times and observe the number of heads, y.

For the moment we will assume a uniform prior distribution for θ.

The unfair coin

We have a coin that, when spun on its edge, falls on the "heads" side with some unknown probability θ. In our experiment, we spin it n times and observe the number of heads, y.

For the moment we will assume a uniform prior distribution for θ.

Our model, written explicitly, is

$$\begin{aligned} y &\sim \mathrm{Binomial}(n, \theta) \\ \theta &\sim \mathrm{Uniform}(0, 1) \end{aligned}$$

so

$$\begin{aligned} p(y|n, \theta) &= \left(\begin{array}{c} n \\ y \end{array} \right) \theta^{y} (1-\theta)^{n-y} \\ p(\theta) &= 1 \qquad 0 \leq \theta \leq 1 \end{aligned}$$

Calculating a posterior

Suppose we spin the coin 10 times and get 7 heads.

Starting from a uniform prior, $p(\theta = 1)$, we get that

$$p(\theta|y_{\rm obs}) = \frac{\left(\begin{array}{c} 10 \\ 7 \end{array}\right)\theta^7(1-\theta)^3}{P(y_{\rm obs})}$$

Calculating a posterior

Suppose we spin the coin 10 times and get 7 heads.

Starting from a uniform prior, $p(\theta = 1)$, we get that

$$p(\theta | y_{\mathrm{obs}}) = \frac{\left(\begin{array}{c} 10 \\ 7 \end{array} \right) \theta^7 (1 - \theta)^3}{P(y_{\mathrm{obs}})}$$

The normalizing constant is

$$P(y_{\mathrm{obs}}) = \left(\begin{array}{c} 10 \\ 7 \end{array} \right) \int_0^1 \theta^7 (1 - \theta)^3 d\theta$$

representing the probability of getting the data in all the possible "universes" (different values of θ)

Calculating a posterior

Often we don't need to calculate normalizing constants explicitly:

- we can recognize the part of the posterior that depends on θ as a well-known distribution (we're in this situation here)
- we are going to simulate the distribution instead of trying to calculate analytically (when we do more complex models, we'll usually be here)

Beta posterior

Beta(α, β) distribution:

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

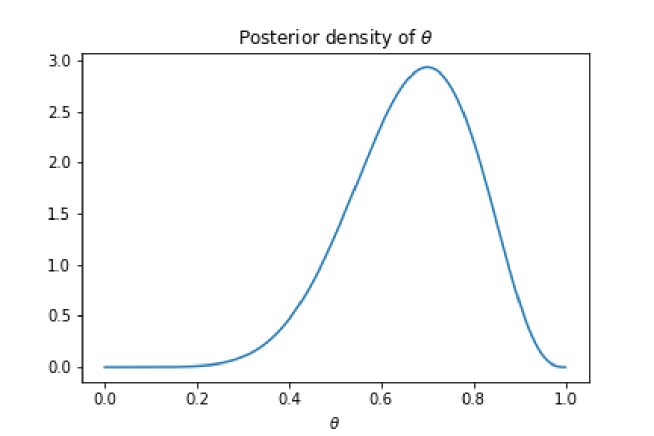
where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

So our posterior is Beta(8, 4).

(Note also that the uniform distribution on $[0, 1]$ is the same as Beta(1, 1).)

Beta posterior

Graphing the posterior:



Conjugate prior

If we look carefully, we can notice that if the prior is Beta(α, β), then the posterior will always be:

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

which is also Beta, but with different parameters.

The Beta distribution is said to be a *conjugate prior* for the binomial likelihood; this means the posterior remains a Beta distribution.

Conjugate priors

Many common likelihoods have conjugate priors.

Conjugate priors are useful for convenience:

- if you want analytical expressions, conjugate priors are often the only way
- in the case of simulation, it's often faster to sample from conventional distributions

That said, modern MCMC methods are efficient enough that conjugate priors often carry little computational advantage, so use what's best for your model

Summarizing inference

Summarizing the posterior

The posterior distribution $p(\theta|y_{\mathrm{obs}})$ is the end result of inference – it contains all we know about θ.

But it's often convenient to summarize this in some way: graphically, or by summary statistics.

Common summaries of location

- Posterior mean: expected value of the posterior distribution

  $$E(\theta|y)$$

  When an analytic form of the posterior is available (usually when using a conjugate prior) this is often easy to compute – also efficiently computed by MCMC simulations
- Posterior mode (a.k.a. maximum a posteriori (MAP) estimate). Can be interpreted as "most likely" value but *beware* using this for models with many parameters. MAP estimate can be far away from almost 100% of the probability mass in a joint distribution over many parameters.

For the coin example

For our coin example:

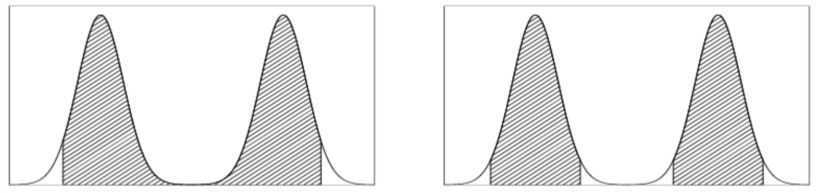
- Posterior mean: $\frac{\alpha}{\alpha+\beta} = 8/12$
- Posterior mode: $\frac{\alpha-1}{\alpha+\beta-2} = 7/10$

Quantiles and intervals

Posterior median: 50th %ile of posterior distribution

Posterior intervals: capture where “most” of the probability mass lives

- Central posterior interval: cut off $\alpha/2$ of probability at top / bottom
- Highest posterior density: find the region with highest posterior density where $(100 - \alpha)\%$ mass is located. Not guaranteed to be an interval!



Posterior intervals

In many practical cases the central interval and highest posterior density give roughly the same result.

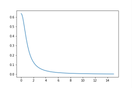
If you have a bimodal distribution, probably your parameter is a mixture of two densities, so model it like one – explicitly include the categorical variable that distinguishes those modes.

Another case where these two may differ: skewed densities

Posterior intervals

Example: half-Cauchy distribution

$$p(\theta) = \begin{cases} \frac{2}{\pi}\frac{1}{\theta^2+1} & \theta \geq 0 \\ 0 & \theta < 0 \end{cases}$$



90% highest density: $[0, 6.31]$

90% central interval: $[0.08, 12.71]$

Predictive distributions

Prior and posterior predictive distributions

Recall: every Bayesian model is *generative*, meaning it tells us how to generate data/make predictions

- Prior predictive distribution $p(y)$ – distribution of observations under prior
- Posterior predictive distribution $p(y|y_{\mathrm{obs}})$ – distribution of future observations given current ones. Generally assume conditional independence of y, y_{obs} given θ.

These are obtained by integrating over all values of θ.

Prior and posterior predictive distributions

Prior predictive distribution:

$$p(y) = \int p(y, \theta) d\theta = \int p(y|\theta) p(\theta) d\theta$$

Posterior predictive distribution:

$$\begin{aligned} p(y|y_{\mathrm{obs}}) &= \int p(y, \theta | y_{\mathrm{obs}}) d\theta \\ &= \int p(y|\theta) p(\theta | y_{\mathrm{obs}}) d\theta \end{aligned}$$

Prior predictive distribution for the coin example

Let's consider the prior predictive distribution for the coin:

Question: before observing any data, what is $P(y = \mathrm{heads})$?

Prior predictive distribution for the coin example

Let's consider the prior predictive distribution for the coin:

Question: before observing any data, what is $P(y = \mathrm{heads})$? 1/2

Prior predictive distribution for the coin example

Let's consider the prior predictive distribution for the coin:

Question: before observing any data, what is $P(y = \mathrm{heads})$? 1/2

Question: how is this calculated?

Prior predictive distribution for the coin example

Let's consider the prior predictive distribution for the coin:

Question: before observing any data, what is $P(y = \mathrm{heads})$? 1/2

Question: how is this calculated? By integrating out θ

Prior predictive distribution

The predictive distribution is really

$$y \sim \mathrm{Bernoulli}(\theta)$$

$$p(y = \mathrm{heads}|\theta) = \theta$$

To extract $p(y)$ from this we have to *marginalize*:

$$p(y) = \int_0^1 \theta p(\theta) d\theta$$

Prior predictive distributions

It happens that for a uniform $p(\theta)$ (or any other prior symmetric around $1/2$),

$$\int_0^1 \theta p(\theta) d\theta = 1/2$$

so our prior prediction is $p(y = \mathrm{heads}) = 1/2$.

But this isn't the same as saying $p(y = \mathrm{heads}) = 1/2$ because we assume an unknown coin is fair.

Predictive distributions are always based on *all* values of the parameter, not one.

The problem with the binomial model

The problem with the binomial model is that it is too simple to really show this distinction:

$$\begin{aligned} p(y|y_{\text{obs}}) &= \int_0^1 p(y|\theta)p(\theta|y_{\text{obs}})d\theta \\ &= \int_0^1 \theta p(\theta|y_{\text{obs}})d\theta \end{aligned}$$

so the posterior predictive probability of heads is always the same as the posterior mean.

This is a feature of the binomial likelihood, though – not a feature of posterior means in general.

Two types of uncertainty

Predictions of observations come with two types of uncertainty:

- *aleatoric* (chance) uncertainty: comes from variability of outcomes
- *epistemic* (knowledge) uncertainty: comes from uncertainty in model parameters

Integrating over θ is what incorporates the epistemic uncertainty into our predictions

Return to the kidney cancer example

Inference for the simple model

Last time we proposed a simple model:

- One underlying rate θ
- Every county is an observation of this rate:

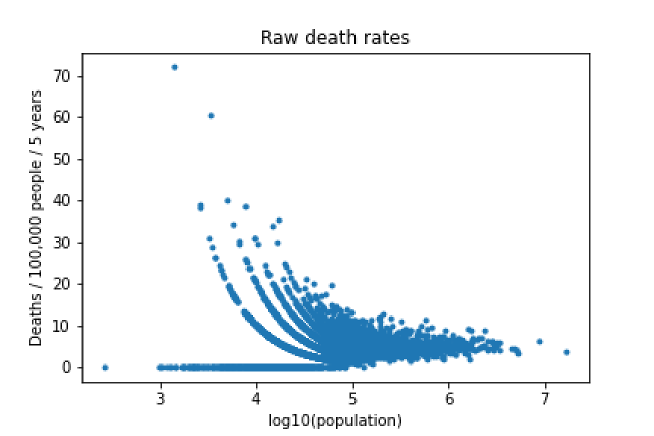
$$y_j \sim \mathrm{Poisson}(\theta n_j)$$

We ran a simulation (really a prior predictive simulation) showing that the qualitative behavior seen in the data set is mostly captured by this model.

But we didn't do any inference, so let's try that now.

First, a plot

Plot of raw death rates vs. log of population



The simple model

For convenience we take a Gamma prior on θ :

$$\begin{aligned} y_j|\theta &\sim \mathrm{Poisson}(\theta n_j) \\ \theta &\sim \mathrm{Gamma}(10, 10^5) \end{aligned}$$

A $\mathrm{Gamma}(\alpha, \beta)$ prior is chosen for convenience (it's the conjugate prior to our Poisson likelihood).

The parameters α, β are chosen to give a prior mean of 10^{-4} for θ (which is what we plugged into our simulation last time).

Inference for the simple model

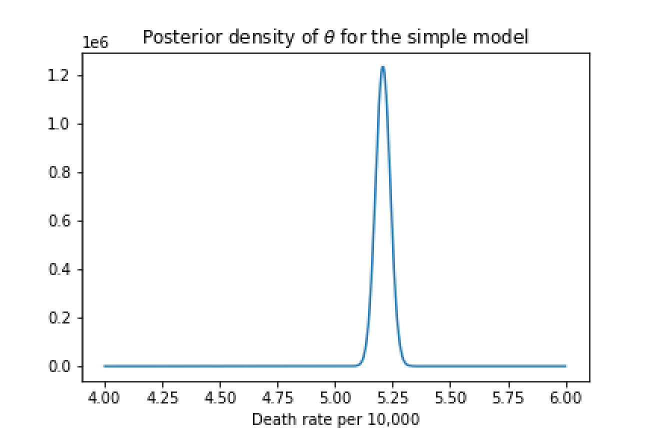
Since we are using a conjugate prior, we can just pool the deaths and populations, and get:

$$\begin{aligned} Y &= \sum y_j \\ N &= \sum n_j \\ \theta | Y, N &\sim \mathrm{Gamma}(10 + Y, 100,000 + N) \end{aligned}$$

Inference for the simple model

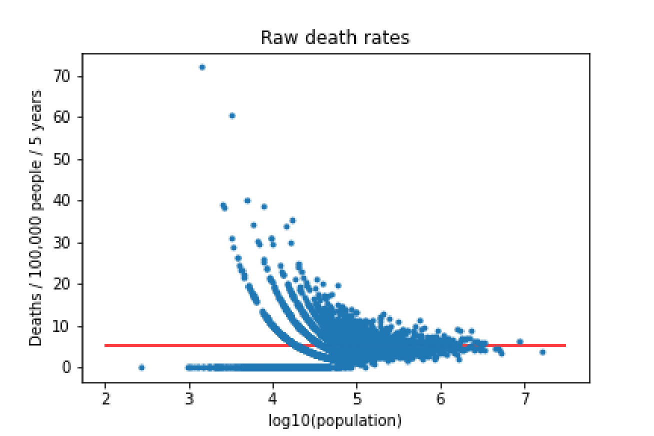
Plugging in values from our data set, we get

$$\theta | Y, N \sim \mathrm{Gamma}(10 + 25962, 100000 + 498650740)$$



Inference for the simple model

Essentially this just takes the pooled sample mean as an estimate of the death rate.



The more complex model

The second attempt at modeling: allow different θ_j for each county.

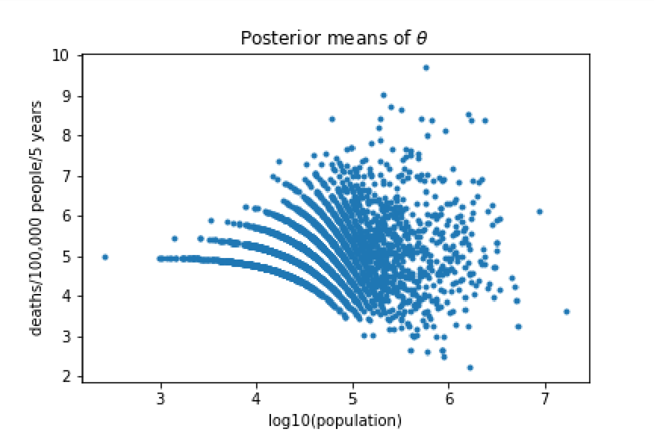
$$\begin{aligned} y_j|\theta_j &\sim \mathrm{Poisson}(\theta n_j) \\ \theta_j &\sim \mathrm{Gamma}(10, 2 \times 10^5) \end{aligned}$$

Why 2×10^5? Ad-hoc matching of the prior to the posterior mean from the simple model[1]. (More sophisticated version of this in BDA3 section 2.8)

[1] This is kind of cheating!

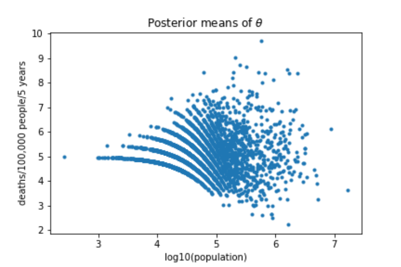
The posterior means

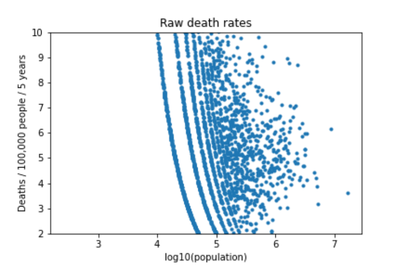
Posterior means of all of the counties' death rates:



It's informative to look at the raw death rates on the same *y*-axis scale

The posterior means





Effect of the new model: "shrink" estimates toward nationwide average, with shrinkage dependent on population

We cheated to get this effect by setting the prior with our knowledge of the nationwide average. More Bayesian approach: allow the counties to share information. More on that later in the course!

Summary

Summary

- Basic modeling approach
- Summaries of posteriors: point estimates, intervals
- Predictive distributions: integrate over parameters
- Kidney cancer example: population-dependent shrinkage of estimates

Next week

- More one- and multi-parameter models
- Regression models from a Bayesian perspective
- Choice of priors

HW1 due 9/04.