More Single Parameter Models

ISTA 410 / INFO 510 - Bayesian Modeling and Inference

University of Arizona School of Information

February 11, 2021

Outline

Last time:

- One parameter models:
    - Beta-binomial models and conjugate priors
- Summarizing posterior inferences

Today:

- Normal model with known variance
- A little bit about priors

Beta-binomial model

Defining parameters:

- π_c: probability that a control subject becomes ill
- π_v: probability that a vaccinated subject becomes ill

never deal with these directly.

- Derived quantity: Vaccine efficacy:

☆ given

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

Parameter for the model:

☆ given

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

Observe n cases.

y = # from vaccine side

$y \sim \text{Binomial}(n, \theta)$

Measures the probability that a case came from the vaccine arm

Alternative: y_v = # from vaccine arm $\quad y_v \sim \text{Binomial}(n_v, \pi_v)$

y_c = # from control $\quad y_c \sim \text{Binomial}(n_c, \pi_c)$

vary with length of study.

Normal model, known variance

Normal likelihood

We all know the normal distribution:

mean

$$p(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2}\frac{(y-\mu)^2}{\sigma^2}\right)$$

std. dev.

For now, to keep this a one-parameter model, we'll treat σ as a known constant. (We could also fix μ and use σ as the unknown parameter, in principle)

The normal model, known variance

We'll start with a normal model, and as an example case we'll use a data set for basketball scores: final scores y_i from all NCAA men's tournament games from about 1939-1995. We're interested in inferring what an "average" total score is in a game.

combined score of both teams.

- Often normal models get used out of convenience or out of tradition
- When justified, usually justified by the central limit theorem: sum or average of many IID components gives rise to normal distribution

independent, identically distributed.

A visual inspection of the data distribution shows a normal distribution really does fit here, but it's reasonably well justified from first principles

The normal model, known variance

As usual, our starting point is specifying a model and priors for our parameters:

unknown mean

$$y_i \sim \text{Normal}(\theta, \sigma)$$
$$\theta \sim \text{Normal}(\mu_0, \tau_0) \leftarrow \text{prior distribution for } \theta.$$

Take $\sigma = 14$. Here, we are choosing a normal prior for convenience (it's conjugate to the normal likelihood)

How can we choose values for μ_0, τ_0

Say low estimate for θ: ≈ 100
high estimate for θ: ≈ 300

Choose μ_0, τ_0 so that most of prob. for θ falls between 100 and 300.

Checking our prior: prior predictive simulations

Prior predictive simulations: draw observations (i.e. values of y_i) using the prior distribution

This can be used to check the reasonableness of a prior, by making sure it doesn't produce impossible results.

- We're not looking for the prior predictions to be a perfect model for the data
- But, if our predictive draws have games with negative score, or teams scoring 500 points, maybe something is off

Assume we start with one observation y. Since we are using a conjugate prior, the posterior is analytically expressible:

$$
\begin{aligned}
p(\theta|y) &\propto \overbrace{\exp\left(-\frac{1}{2}\frac{(y-\theta)^2}{\sigma^2}\right)}^{\text{likelihood}} \overbrace{\exp\left(-\frac{1}{2}\frac{(\theta-\mu_0)^2}{\tau_0^2}\right)}^{\text{prior density}} \\
&= \exp\left(-\frac{1}{2}\left(\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\tau_0^2}\right)\theta^2 - \left(\frac{2y}{\sigma^2}+\frac{2\mu_0}{\tau_0^2}\right)\theta \underbrace{+ \frac{y^2}{\sigma^2} + \frac{\mu_0^2}{\tau_0^2}}_{\text{constant / independent of } \theta}\right)
\end{aligned}
$$

Then some magic happens...

Calculating the posterior

one observation

$$\theta|y \sim \text{Normal}(\mu_1, \tau_1)$$

where

$$\left.\begin{aligned}\mu_1 &= \frac{\frac{1}{\sigma^2}y + \frac{1}{\tau_0^2}\mu_0}{\frac{1}{\sigma^2} + \frac{1}{\tau_0^2}} \\ \frac{1}{\tau_1} &= \frac{1}{\sigma_2} + \frac{1}{\tau_0^2}\end{aligned}\right\} \text{updated parameters.}$$

The inverse variances $1/\sigma^2, 1/\tau^2$ are called the *precisions* of these distributions

(Where's the magic? Complete the square (exercise 2.14(a)) in the book)

The posterior as a compromise

Three ways of writing the posterior mean of θ:

$$\mu_1 = \frac{\frac{1}{\sigma^2} y + \frac{1}{\tau_0^2}\mu_0}{\frac{1}{\sigma^2} + \frac{1}{\tau_0^2}}$$

variance of prior distribution.

$$\mu_1 = \mu_0 + (y - \mu_0)\frac{\boxed{\tau_0^2}}{\sigma^2 + \tau_0^2}$$

$$\mu_1 = y - (y - \mu_0)\frac{\sigma^2}{\boxed{\sigma^2} + \tau_0^2}$$

variance of observations

- Weighted average of μ_0 and y
- Prior mean μ_0 adjusted toward the data
- Data "shrunk" toward the prior mean

Generalizing to many observations

We don't have to iterate this process a thousand times to incorporate our thousand games (although the ability to incorporate observations one by one can be considered a feature of the Bayesian approach); the posterior depends on $y_1, y_2, \ldots$ only through the sample mean $\bar{y}$[1]

$$\theta | (y_1, y_2, \ldots y_n) \sim \mathrm{Normal}(\mu_n, \tau_n)$$

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

[1] $\bar{y}$ is called a *sufficient statistic* in this model

Posterior inferences

Let's make some posterior inferences:

- posterior mean?
- 92% interval?

Posterior predictive distribution

One feature of the posterior we haven't met yet: the *posterior predictive distribution*

- probability distribution of future observations
- will the next case observed in the vaccine trial come from the vaccine or placebo group?
- what will the combined score be of the next game?

In some cases, we can calculate the posterior predictive distribution explicitly; but if not, we can sample in stages (like in the dice problem)

Posterior predictions

The posterior predictive distribution is (unsurprisingly) also normal (details in section 2.5 of BDA) with

posterior mean

$$E(y|y_{\text{obs}}) = \boxed{\mu_n}$$

$$\text{var}(y|y_{\text{obs}}) = \sigma^2 + \tau_n^2$$

posterior variance

likelihood variance.

Intuitively:

- mean prediction is posterior mean of θ
- uncertainty of prediction is the uncertainty in θ (epistemic uncertainty, τ_n^2) plus the uncertainty of individual observations (aleatoric uncertainty, σ^2)

chance uncertainty.

Informative vs. uninformative priors

Most often, priors are categorized as *informative* or *uninformative priors* depending on whether they incorporate outside scientific information

- informative priors: bring in knowledge about the application domain, or results of previous study, as a starting point for estimation and inference
- uninformative priors: avoid using external knowledge, "let the data speak for itself"

Proper and improper prior distributions

The prior precision $1/\tau_0^2$ is the weight given to the prior mean in the posterior distribution; if τ_0^2 is very large, then this weight is very small and the posterior is dominated by the data.

Specifically, if $n/\sigma^2 \gg 1/\tau_0^2$, then the posterior distribution is approximately

$$p(\theta|y) \sim \mathrm{Normal}(\bar{y}, \sigma/\sqrt{n})$$

and so the prior has essentially no input in the posterior.

To eliminate the influence of the prior, we could assign a completely flat/uniform prior on θ. Problem: has an infinite integral, so it's not really a probability distribution.

Improper prior distributions can produce proper posteriors

This example shows that even with an improper uniform prior on θ, the posterior distribution is proper – i.e. $p(\theta|y)$ has a finite integral for any possible data y (as long as there is at least one observation).

- This must be checked any time you use an improper prior
- Most reasonable interpretation of the posterior: as an approximation, valid when the likelihood dominates the prior density
- This is generally dependent on both sufficient amount of data and sufficiently localized likelihood

Flat priors

Some issues about uninformative priors:

- uninformative doesn't always mean "flat" / uniform
  - a prior that is flat in one parameterization may be non-flat if you change variables
  - flat priors can be improper
  - flat priors can be practically nonsensical

Weakly informative priors

A compromise between the informative and uninformative priors is so-called "weakly informative" priors, which generally attempt to include enough outside knowledge to ensure that the prior is proper and sensible, but the information in the prior is intentionally weaker than the availabele outside informations.

- Our basketball prior example: I asked you to come up with rough bounds, but we used them loosely
- Example: in the coin spinning problem, take $\mathrm{Beta}(3, 3)$ in place of uniform or $\mathrm{Beta}(1, 1)$.
- Example (from the book): in estimating the proportion of female births, choose a prior with the probability mass concentrated between, say, 0.4 and 0.6 (e.g. $\mathrm{Normal}(0.5, 0.1)$)

What to do?

We'll return to problems of prior choice frequently, but for now:

- Weakly informative often a good choice
- Flat priors can be problematic
- Avoid assigning probability 0 to anything unless you are really sure
- Prior predictive checks to avoid truly nonsensical values

Summary

Today:

- Normal model, known variance
- Informative vs. uninformative priors

Next week:

- Some multi-parameter models
- More on priors