Several Simple Models

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

August 31, 2020

Outline

Previously:

- Binomial model (spinning coin)
- Poisson model (kidney cancer)

Now:

- Normal models (known or unknown variance)
- Some commentary on priors
- Linear regression as a normal model

Normal model with known variance

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.

- 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

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:

$$y_i \sim \text{Normal}(\theta, \sigma)$$
$$\theta \sim \text{Normal}(\mu_0, \tau_0)$$

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

Can we choose a value for μ_0?

Checking our prior: prior predictive simulations

We've set a prior – we should check that it's at least slightly reasonable.

Simple thing to do: draw some samples, make sure they're not off-the-wall ridiculous

- 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

Calculating the posterior

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

$$
\begin{aligned}
p(\theta|y) &\propto \exp\left(-\frac{1}{2}\frac{(y-\theta)^2}{\sigma^2}\right)\exp\left(-\frac{1}{2}\frac{(\theta-\mu_0)^2}{\tau_0^2}\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{(y-\theta)^2}{\sigma^2}+\frac{(\theta-\mu_0)^2}{\tau_0^2}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\tau_0^2}\right)\theta^2-\left(\frac{2y}{\sigma^2}+\frac{2\mu_0}{\tau_0^2}\right)\theta+\frac{y^2}{\sigma^2}+\frac{\mu_0^2}{\tau_0^2}\right)
\end{aligned}
$$

Then some magic happens...

Calculating the posterior

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

where

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

$$\frac{1}{\tau_1} = \frac{1}{\sigma_2} + \frac{1}{\tau_0^2}$$

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:

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

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

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

- 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]

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

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

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

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

Posterior predictions

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

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

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

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)

Posterior predictions

Let's do some posterior predictions...

Priors

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"

In reality, informative vs. uninformative is not always a sharp binary distinction.

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 this precision is very small, it is as if $\tau^2 = \infty$.

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

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

We could imagine that we had assigned a prior that is constant/uniform. Problem: has an infinite integral!

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

Uninformative 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
  - from Aki Vehtari – how about a flat prior on "the amount of money in my wallet"?

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.

- 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. $\mathrm{Normal}(0.5, 0.1)$)

Normal model, unknown variance

Introduction to multi-parameter models

The known-variance assumption isn't necessarily particularly realistic. So instead, we can allow σ^2 to be an unknown parameter in our model.

New model (simple, improper priors):

$$y_i \sim \text{Normal}(\mu, \sigma)$$
$$\mu \sim ???$$
$$\sigma \sim ???$$

Next time:

- Priors for μ, σ
- Considerations for multi-parameter models
- Failure of a normal model