More on Priors

Bayesian Data Analysis

Steve Buyske

Revisiting priors

- There are three ways that Bayesian statistics differs from frequentist statistics
  - The language of probability is used start to finish. Everyone likes that.
  - Computational aspect
    - used to be almost intractable outside of special cases
    - last 30 years workable by specialists
    - last 5 years workable by others.
  - And …

- Priors
  - Lots of people don’t like this.
  - Doesn’t seem objective. It’s not, but neither is most of statistics, which is full of choices.
  - $\text{Prob}(\theta|\text{data}) \propto \text{Prob}(\text{data}|\theta)\text{Prob}(\theta)$
    - The choice of prior is external to doing a correct analysis
  - There are different approaches to choosing priors.

Cromwell Rule

- Before we get into priors, it is worth mentioning the Cromwell Rule:
- Avoid priors that assign 0 probability to logically possible events.
- Named for Oliver Cromwell, who said in a speech in 1650,

> I beseech you, in the bowels of Christ, think it possible that you may be mistaken.

- Because 0 times anything is 0, if the prior has a region of zero probability then the posterior will also have zero probability in that region, no matter how strong the evidence is.
- It’s fine, however, to have very, very small probability.

Sequential approach to choosing priors

- The prior for a second experiment is the posterior from the first experiment.

$$\text{Prob}(\theta|\text{data2}) \propto \text{Prob}(\text{data2}|\theta)\text{Prob}(\theta)$$

$$\propto \text{Prob}(\text{data2}|\theta)\text{Prob}(\theta|\text{data1})$$

- We end up with $\text{Prob}(\theta|\text{data2}, \text{data1})$

Sequential approach to choosing priors cont.

- We update the prior as we get new data.
- The initial prior has less and less effect as we get more and more data.
- You might infer $\text{Prob}(\theta|\text{data1})$ from published work
  - You don’t need a published posterior—you could work backward from a mean, a standard error, and a sample size.
- People often start with a prior from published work and then weaken it by increasing the scale—that is, a larger standard deviation.

Conjugate priors

- We have seen that the beta distribution is a conjugate prior for a Bernoulli (or Binomial) distribution.
- The idea is that the posterior distribution and the prior distribution are in the same family.
- Huge advantage that you don’t need to use an MCMC process.
- Disadvantage in that conjugate priors are very restrictive.

Conjugate priors cont.

- Some common examples, just for the record:

Distribution used in likelihood	Conjugate prior
Bernoulli	beta
Binomial	beta
Poisson	gamma
Normal with fixed variance	normal
Normal	normal and inverse gamma

Subjective priors

- Subjective priors cause the most controversy.
- One approach is to boil down expert opinion, known as “eliciting a prior.”
- For example, people will sometimes give experts 20 stickers, each representing 5% probability, and ask them to predict the effect of a drug.
    - The result is essentially one expert’s prior; you might average over several such priors to get a final one.

Subjective priors cont.

- A variant sometimes used for clinical trials is to elicit a “skeptics’ prior” and an “enthusiasts’ prior.”
  - If the resulting posteriors are nearly the same, then there is a sort of consensus.
  - One way to get a skeptics’ prior would be to use only the most skeptical priors of the ones you elicited
    - “most skeptical” here meaning the most probability of little to no effect
  - The enthusiasts’ prior would be the similar, but using priors that have most of the probability away from no effect.
  - A related approach would be to have the skeptics’ prior be centered at no effect, with, say, 5% of an effect larger than the minimally clinically relevant effect size.

Non-informative priors

- The idea of non-informative or uninformative priors is to keep the Bayesian framework of probability, but to use a prior with a minimal role in affacting the posterior.
- This is known as the Bayes-Laplace *Principle of indifference*, and was the reason for Laplace picking a uniform prior for his study of the female proportion of births.

Flat priors

- An important subset of non-informative priors are *flat priors*.
  - The prior is simply constant; for a bounded region (like $(0, 1)$ for Laplace), the prior is a uniform distribution.
  - Can even do a version of this for parameters defined on the entire real line, by just saying that $p(\theta) = 1$ for all θ.
    - This prior is not an actual distribution, since it won't integrate to 1, and is therefore called an *improper prior.*
- With a flat prior, the posterior is proportional to the likelihood.
  - If you've taken Theory of Statistics: that means the mode of the posterior will be the maximum likelihood estimator.
- The `brms` package uses flat priors as defaults.

Flat priors cont.

- Flat priors seem appealing, but don't really reflect a state of ignorance.
  - Suppose β is a linear regression coefficient.
    - With a flat prior, there's very little probability near 0.
    - In fact, for any interval, with an improper flat prior there's more prior probability outside the interval than inside it.
    - Consequently, flat priors put much more probability on extreme values than more realistic ones.
- Important to note that a flat prior will not remain flat after a transformation:
  - Suppose $\text{Prob}(1 < \sigma < 2) = \text{Prob}(2 < \sigma < 3)$. Then $\text{Prob}(1 < \sigma^2 < 4) = \text{Prob}(4 < \sigma^2 < 9)$, which means that the prior is not flat for σ^2.

Jeffreys prior

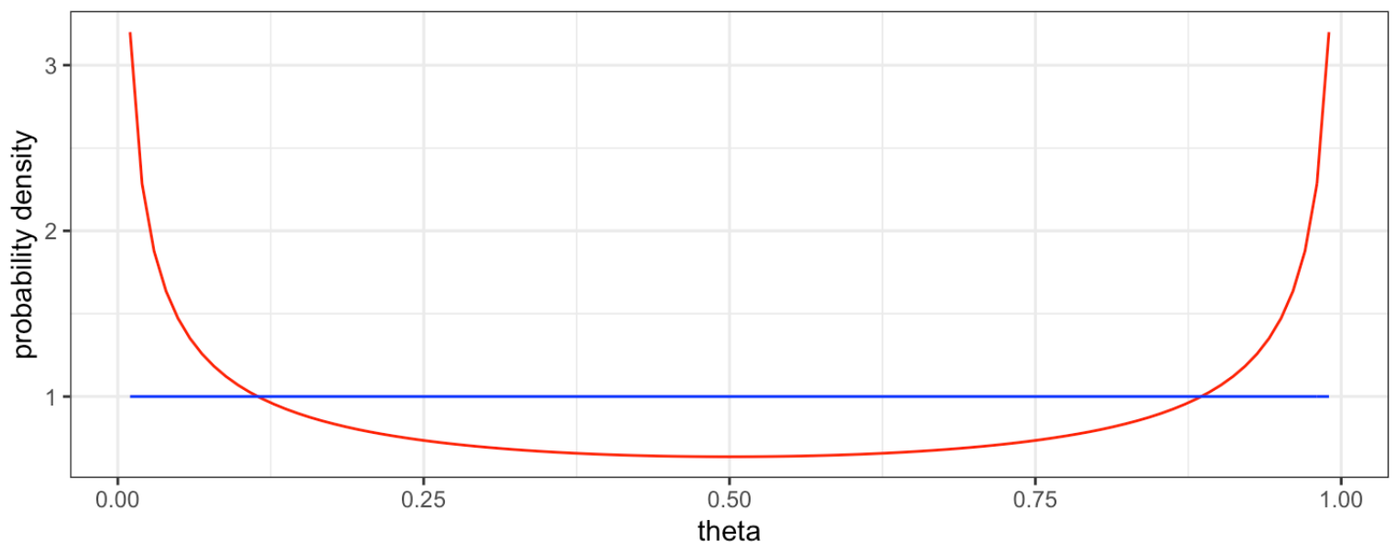
- Another non-informative prior is due to Jeffreys
- The idea is to use priors that provide the same conclusion regardless of transformation
  - It turns out that can be done with $p(\theta) \propto \sqrt{\text{Fisher Information}}$.
    - The Fisher Information for the simple example of the mean of normal distribution turns out to be $1/\sigma^2$
    - We are not going to use the concept, but Fisher information is a measure of how much information there is for an unknown parameter—it has to do with how curved the likelihood is.
    - In case you're really interested, one definition is

$$I(\theta) = -E_\theta(l''(x|\theta),$$

where $l(x|\theta)$ is the log likelihood.

Jeffreys prior cont.

- As an example, for the Bernoulli likelihood, the Jeffreys prior is $\text{beta}(1/2, 1/2)$
  - In the plot, the blue line is the uniform distribution, while the red curve is the Jeffreys prior



Weakly informative priors

- The defaults in the `rstanarm` package are what are known as *weakly informative priors*.
- The idea is that there is that they should not have a strong effect on the prior, but there is enough information to “regularize” the posterior.
- Such a prior should (one hopes) give little probability to unreasonable values, while not ruling out values that might make sense.

Weakly informative priors cont.

- It is intentionally weaker—more spread out—than whatever prior knowledge we might have.
- (As a general principle, it's better to have too weak a prior than too strong.)
- **Do not** do something like, for a coefficient of a variable with standard deviation 1, say something like $\beta \sim N(0, 500)$
  - That puts too much probability on absurd values
- `rstanarm`'s default is $N(0, 2.5)$ when the variables are standardized to mean 0 and sd 1.
- Another popular choice is a t-distribution with mean 0, sd 2.5, and 3 degrees of freedom.

How informative are weakly informative priors?

- Clearly there's some subjective judgment in how informative to make a weakly informative prior.
- Here's one rule of thumb:
  - If the standard deviation of the posterior is more than 1/10 the standard deviation of the prior, then let your readers/listeners/clients know that the prior is affecting the posterior
    - (Of course, that's always literally true, but here we mean in an important way.)
- In practice, for a critical analysis, people may rerun the analysis with different priors (a *sensitivity analysis*) to make sure that the effect of the prior is minor.

How to incorporate choice of priors into R?

- Next week!

Closing notes

- If you are *really* interested and have a lot of background, you might find the essay at https://mc-stan.org/users/documentation/case-studies/weakly_informative_shapes.html of interest.

- At the start, I mentioned that priors are the most controversial aspect of Bayesian statistics.

    - Those who are critical will say that it introduces too much subjectivity into the analysis. - Bayesians might reply that all statistical analysis is subjective, but priors, at least, make that subjectivity explicit.

        - “If you don’t like *my* prior, you can redo the analysis with *your* prior.”