An Example of the Analytic Approach

Bayesian Data Analysis

Steve Buyske

An Analytic Example of a Bayesian Analysis

- I tossed a coin and got **7** heads in **10** tosses.
- Let's write θ for the probability of getting heads.
- Let's also write $Y = 1$ for 1 heads in 1 toss, and $Y = 0$ for 0 heads in 1 toss.
- Then $\text{Prob}(Y = 1 \mid \theta) = \theta$ and $\text{Prob}(Y = 0 \mid \theta) = 1 - \theta$.
- We can combine these into a single equation with

$$\text{Prob}(Y \mid \theta) = \theta^{Y}(1 - \theta)^{1-Y} \qquad \text{for } Y \in \{0, 1\}.$$

- This is the likelihood for a single coin toss. The distribution is known as the **Bernoulli** distribution.

An Analytic Example cont.

- My sequence of tosses were $\{Y_i\} = \{1, 1, 0, 1, 1, 0, 0, 1, 1, 1\}$.
- With independent events, we can multiply the probabilities to get, in general,
- $$\begin{aligned}\text{Prob}(\{Y_i\} \mid \theta) &= \prod \text{Prob}(Y_i \mid \theta) \\ &= \prod \theta^{Y_i}(1-\theta)^{1-Y_i} \\ &= \prod \theta^{Y_i} \prod (1-\theta)^{1-Y_i} \\ &= \theta^{\sum Y_i}(1-\theta)^{\sum(1-Y_i)} \\ &= \theta^{\text{number of heads}}(1-\theta)^{\text{number of tails}}.\end{aligned}$$
- Write Z for the number of heads and N for the number of tosses (*not the number of tails*), and we have
- $$\text{Prob}(\{Y_i\} \mid \theta) = \theta^Z(1-\theta)^{N-Z}.$$

- We have

  $$\text{Prob}(\{Y_i\} \mid \theta) = \theta^Z (1 - \theta)^{N-Z},$$

  which is going to be our likelihood.
- You may recognize that this looks a lot like the binomial distribution (given the *number* of heads, say, in N tosses), but without the constant in front.
- The difference is that this distribution is for the particular sequence of heads and tails.
- From the right hand side, though, you can see that the probability depends only on the number of heads and the number of tosses, and not on the particular sequence.
- In our framework, where we standardize at the end to get the posterior, the difference between the Bernoulli and the binomial distributions doesn’t really matter.

- Note that when we write $\text{Prob}(\{Y_i\} \mid \theta)$,
- if we think of $\{Y_i\}$ as random and θ as fixed, then $\text{Prob}(\{Y_i\} \mid \theta)$ is a distribution.
- If we think of $\{Y_i\}$ as fixed and θ as random, then $\text{Prob}(\{Y_i\} \mid \theta)$ is not actually a distribution, since it will not generally add up to 1.0.
- In this context, people often write this as $L(\theta \mid \{Y_i\})$ to emphasize the idea that $\{Y_i\}$ is fixed.
- Not adding up to 1.0 is not too much of a concern for us, since you can see from

$$\text{Prob}(\theta \mid \{Y_i\}) = \frac{\text{Prob}(\{Y_i\} \mid \theta)\,\text{Prob}(\theta)}{\text{Prob}(\{Y_i\})}$$

  that regardless we will have to standardize the right side so that we get an actual distribution.

An Analytic Example cont

- Now we’ve settled on a likelihood, namely $\mathrm{Prob}(\{Y_i\} \mid \theta) = \theta^Z (1-\theta)^{N-Z}$.
- Next, we need a prior. In particular, we want a prior that will be nice to work with.
- We would like $\mathrm{Prob}(\{Y_i\} \mid \theta)\mathrm{Prob}(\theta)$ to have the same form as $\mathrm{Prob}(\theta)$.
- If it does, then the posterior can be come a new prior *in the same form*. Such a prior is called a **conjugate prior**.
- We would like $\mathrm{Prob}(\{Y_i\}) = \int \mathrm{Prob}(\{Y_i\} \mid \theta)p(\theta)\, d\theta$ to have a nice form so that we can work with the integral.

- In our example, the likelihood has the form $\theta^Z(1-\theta)^{N-Z}$.
- If the prior was proportional to something similar, say,
  - $\theta^u(1-\theta)^v$,
  - then the product of the likelihood and the prior would be proportional to
  - $\theta^{Z+u}(1-\theta)^{N-Z+v}$
  - *which has the very same form as the prior*.

The Beta Distribution

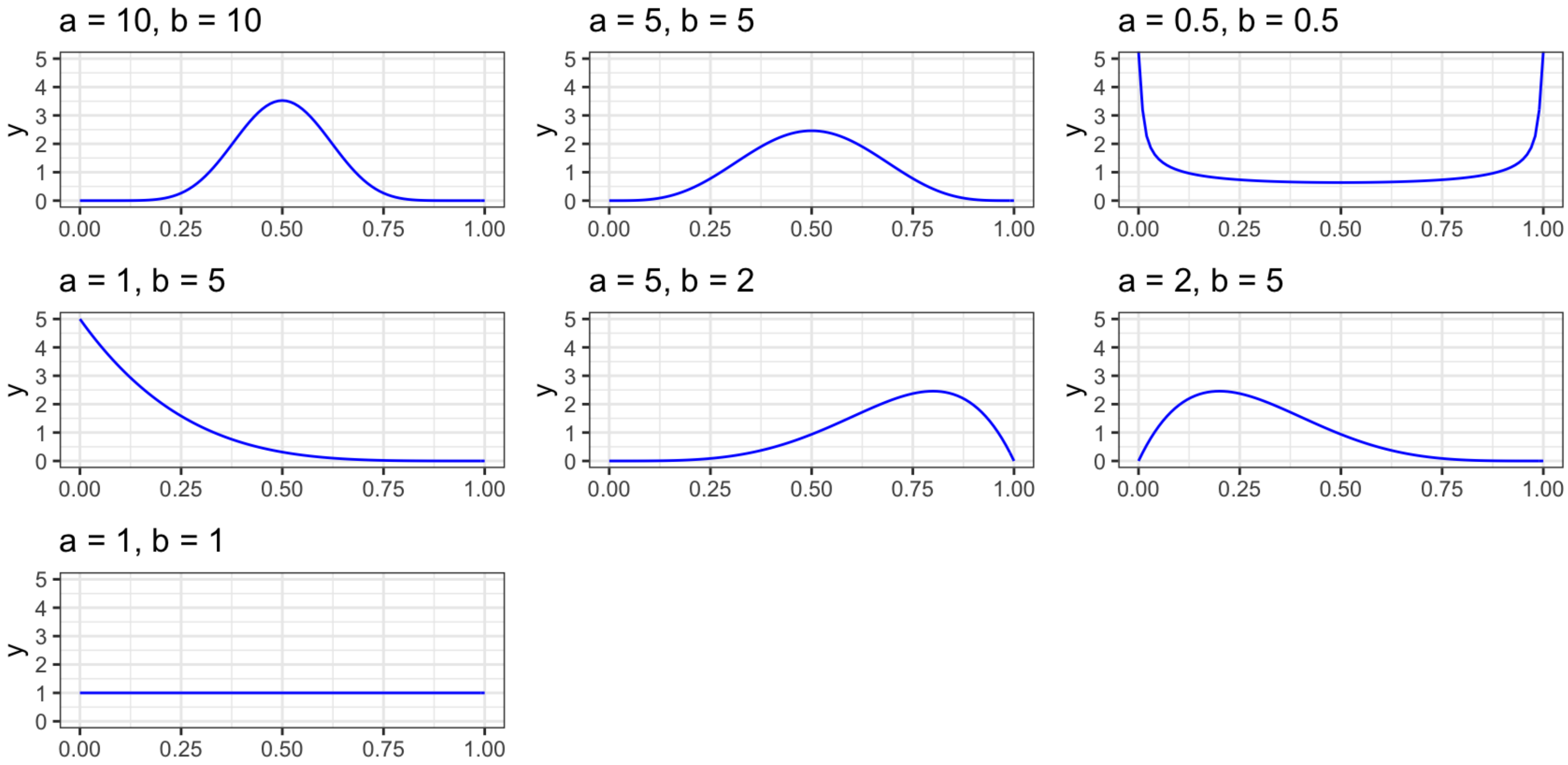
- The **beta distribution** is the distribution that we need (see pp 173–174 of the text).
- It is defined by

$$\text{beta}(\theta \mid a, b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)},$$

  where $a > 0$, $b > 0$, $0 < \theta < 1$, and $B(a, b)$, called the beta function, is defined by

$$B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}\, d\theta.$$

- If you know Gamma functions, then $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$.



Various beta distributions

Notice that the larger a is, the lower the density goes on the *left*, while the larger b is, the lower the density goes on the *right*.

- The **mean** of a beta distribution if

$$\frac{a}{a+b}$$

(which will be 1/2 if $a = b$, making the distribution symmetric).

- The **mode**, when $a > 1$ and $b > 1$, is

$$\frac{a-1}{a+b-2}.$$

- The **variance** is

$$\frac{ab}{(a+b)^2(a+b+1)}.$$

The posterior of a beta prior and a Bernoulli likelihood

With a prior for θ of $\mathrm{beta}(a, b)$, our posterior will be

$$\begin{aligned}
\mathrm{Prob}(\theta \mid Z, N) &= \frac{P(Z \mid \theta, N)\mathrm{beta}(a, b)}{\mathrm{Prob}(Z \mid N)} \\
&= \frac{\theta^{Z}(1-\theta)^{N-Z}\theta^{a-1}(1-\theta)^{b-1}/B(a, b)}{\mathrm{Prob}(Z \mid N)} \\
&= \frac{\theta^{Z+a-1}(1-\theta)^{N-Z+b-1}}{B(a, b)\mathrm{Prob}(Z \mid N)} \\
&= \frac{\theta^{Z+a-1}(1-\theta)^{N-Z+b-1}}{B(Z+a, N-Z+b)},
\end{aligned}$$

so Z heads in N tosses changes our prior from a $\mathrm{beta}(a, b)$ distribution to a $\mathrm{beta}(Z + a, N - Z + b)$ distribution!

- The mean of the posterior, $\mathrm{beta}(Z+a, N-Z+b)$, is

$$\frac{Z+a}{Z+a+N-Z+b} = \frac{Z+a}{N+a+b}.$$

- We can rewrite this as

$$\frac{Z}{N}\,\frac{N}{N+a+b} + \frac{a}{a+b}\,\frac{a+b}{N+a+b}.$$

- We can think of this as
- (mean from the data)(weight of the data) + (mean of the prior)(weight of the prior).

- Knowing that the mean of the posterior is $\frac{Z}{N}\frac{N}{N+a+b} + \frac{a}{a+b}\frac{a+b}{N+a+b}$ helps us interpret the prior:
  - The ratio $\frac{a}{a+b}$ is the mean of the prior.
  - The ratio $\frac{a+b}{N+a+b}$ is how much weight the prior gets.
- Since a plays exactly the same role as the number of heads, and b the same role as the number of tails,
- this means that if we use a prior of $\mathrm{beta}(5, 5)$ we are saying that our prior belief is the same as if we had tossed the coin and gotten 5 heads and 5 tails.
- If we were much more confident that the probability of heads is 0.5, then maybe we'd use a prior of $\mathrm{beta}(100, 100)$, saying that our prior belief is the same as if we had tossed the coin and gotten 100 heads and 100 tails.
- In the other direction, if we want to say that we know so little about the probability of heads that we want to say that all values are equally likely, then we might use a prior of $\mathrm{beta}(1, 1)$—the uniform distribution.

Laplace and the ratio of female births

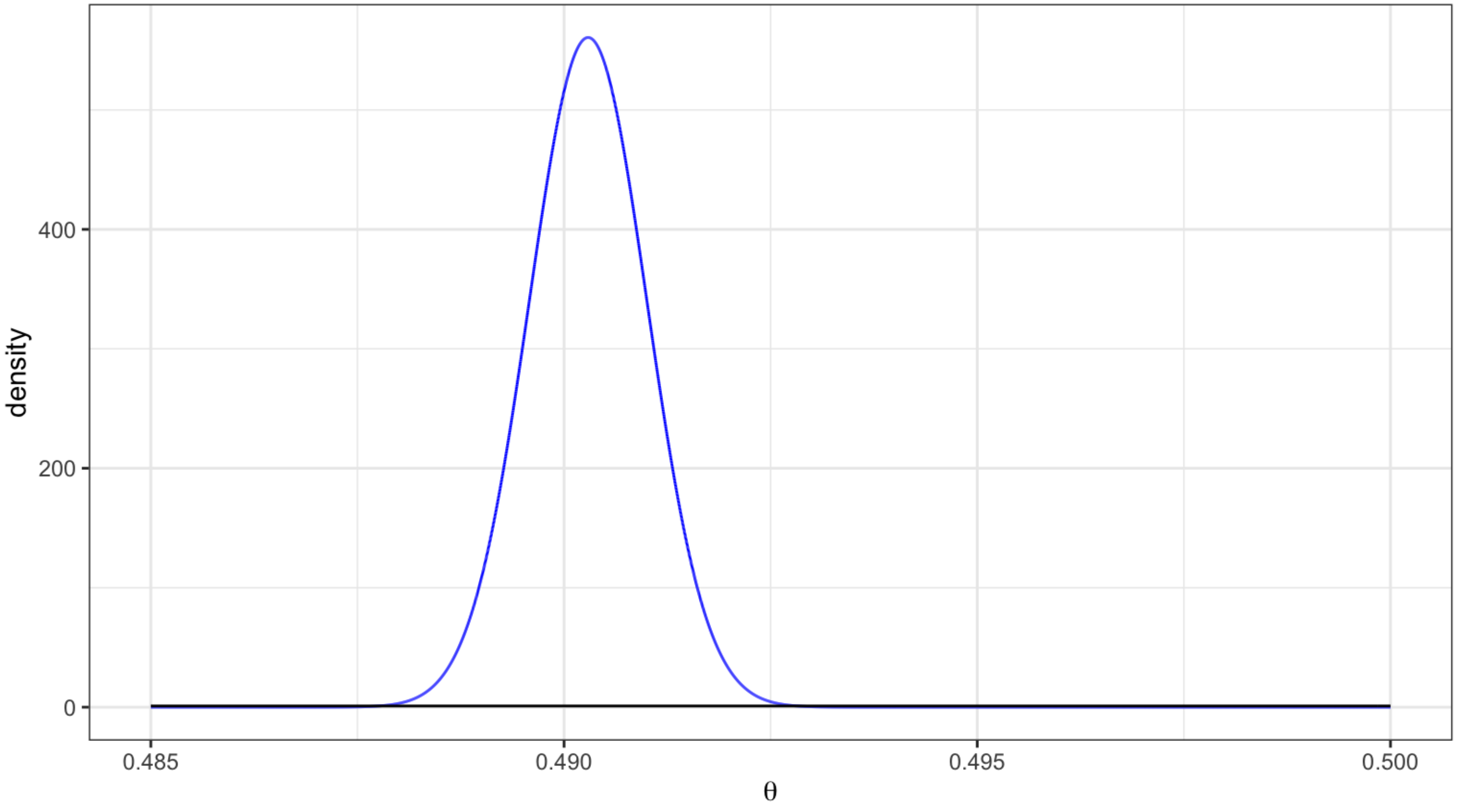
- Pierre-Simon Laplace did the first Bayesian data analysis.

- He was interested in the proportion of births that were girls; let's call it θ.
- For a prior, he decided that all values of θ were equally likely.
- In other words, he decided to use $\theta \sim \text{beta}(1,1)$.
- For data, he was able to find that in Paris between 1745 to 1770, there were 241,945 girls born out of 493,472 births.
- He calculated that the posterior for θ was

$$\theta \mid \text{data} \sim \text{beta}(241945 + 1, 493472 - 241945 + 1) = \text{beta}(241946, 251528).$$

- Knowing that, Laplace could calculate the probability that θ was less than 0.5 —it's $1 - 1.146 \times 10^{-42}$—and concluded that he was "morally certain" that θ was less than 0.5.

Corrected from the original version of the slides



The posterior for the proportion of female births

- In terms of the expression

$$\frac{Z}{N}\,\frac{N}{N+a+b}+\frac{a}{a+b}\,\frac{a+b}{N+a+b}.$$

for the mean of the posterior, Laplace's calculation looks

$$\frac{241945}{493472}\,\frac{493472}{493472+1+1}+\frac{1}{1+1}\,\frac{1+1}{493472+1+1}.$$

$$= 0.4902912 \times 0.9999959 + 0.5 \times 0.000041$$

$$= 0.4903097.$$

- You can see how the weight of the evidence overwhelms the weight of the prior.

Appendix: Two final notes not in the video

1. As mentioned, because the likelihood using the Bernoulli model depends only on the number of heads and the number of tosses, the posterior will be exactly the same whether we use a Bernoulli model or a binomial model. The combination of the beta prior and the Bernoulli or binomial model is commonly called the *beta-binomial* model.
2. In addition to the nice properties already mentioned for the beta distribution as a conjugate prior for the Bernoulli model, it also has the useful property that it does not allow nonsensical values for θ like -5 or 1.1.