STA 365: Applied Bayesian Statistics

Boris Babic
Assistant Professor, University of Toronto

Week 3A: Beta/Binomial

Improper Priors

A proper distribution whose density integrates to 1.

$$\sum_{y=0}^{n} \pi(Y = y \mid \theta) = 1, \quad \int \pi(\theta \mid \alpha, \beta) d\theta = 1.$$

An improper distribution whose "density" does not integrate to one (or even a finite number for that matter) over the support of its argument.

$$h(\theta) = \lim_{\alpha \to 0, \beta \to 0} \pi(\theta \mid \alpha, \beta), \qquad \int h(\theta) d\theta = \infty.$$

When the posterior distribution is proper?

$$\int \pi(y \mid \theta) \pi(\theta) d\theta < \infty.$$

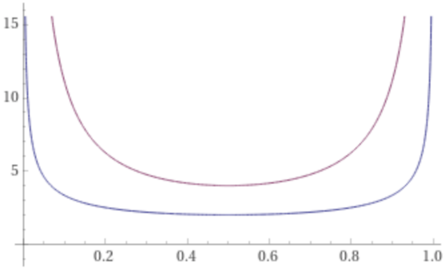
Improper Priors

- If $\alpha = \beta = 1/2$, then

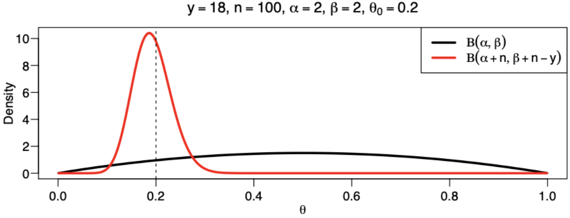
$$\pi(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}}$$

- If $\alpha = \beta = 0$, then

$$\pi(\theta) \propto \frac{1}{\theta(1-\theta)}$$



Prior/Post Expectation and Variance



Prior expectation is the average of the posterior expectation by *the law of iteration expectation:*

$$E(\theta) = E\{E(\theta \mid y)\}.$$

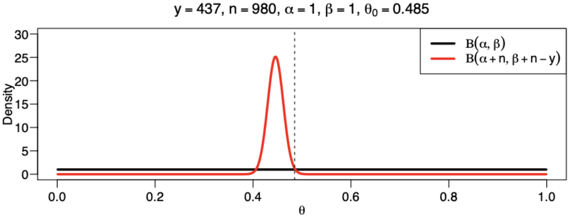
Posterior variance is (on average) smaller than the prior variance by *the law of total variation:*

$$\mathrm{Var}(\theta) = E\{\mathrm{Var}(\theta \mid y)\} + \mathrm{Var}\{E(\theta \mid y)\}.$$

Example: Placenta Previa

- A study was conducted in Germany of 980 births from women with placenta previa. Out of the 980 births, $y = 437$ were baby girls.
- Note: $X \sim \text{Bernoulli}(\theta)$ and $Y = \sum X_i \sim \text{Binomial}(n, \theta)$
- The established proportion of female births in the genral population is 0.485.
- The scientific question of interest is whether the proportion of female births in this subpopulation is less than that in the general population.
- Let θ denote the proportion of female births.
- Assume $\theta \sim \text{Beta}(1, 1) = \text{Uniform}(0, 1)$.
- Find $\theta | y$
- Find $\text{E}(\theta | y)$
- Find 95% credible interval of $\theta | y$
- To run R code, you can log in to UofT Jupyter Hub: `http://jupyter.utoronto.ca`
- You can also execute simple commands here: `https://rdrr.io/snippets/`
- Repeat the above with $\alpha = 9.7, \beta = 10.3$
- Repeat the above with $\alpha = 97, \beta = 103$
- Create a table with the prior mean, posterior mean, and credible interval

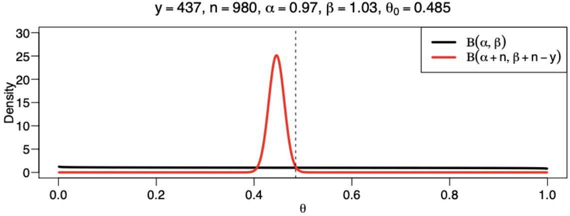
Example: Placenta Previa



$E(\theta)$	$\alpha + \beta$	$E(\theta \mid y)$	95% Credible Interval
0.5	2	0.446	(0.415, 0.477)
0.485	2	0.446	(0.415, 0.477)
0.485	20	0.447	(0.416, 0.478)
0.485	200	0.453	(0.424, 0.481)

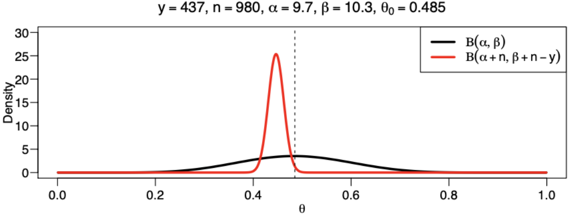
Note that $y/n = 0.445$.

Example: Placenta Previa



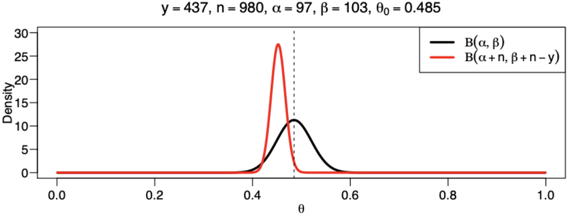
$E(\theta)$	$\alpha + \beta$	$E(\theta \mid y)$	95% Credible Interval
0.5	2	0.446	(0.415, 0.477)
0.485	2	0.446	(0.415, 0.477)
0.485	20	0.447	(0.416, 0.478)
0.485	200	0.453	(0.424, 0.481)

Example: Placenta Previa



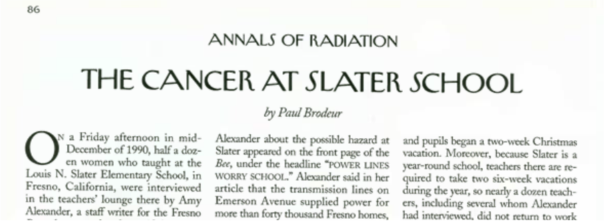
$E(\theta)$	$\alpha+\beta$	$E(\theta \mid y)$	95% Credible Interval
0.5	2	0.446	(0.415, 0.477)
0.485	2	0.446	(0.415, 0.477)
0.485	20	0.447	(0.416, 0.478)
0.485	200	0.453	(0.424, 0.481)

Example: Placenta Previa



$E(\theta)$	$\alpha + \beta$	$E(\theta \mid y)$	95% Credible Interval
0.5	2	0.446	(0.415, 0.477)
0.485	2	0.446	(0.415, 0.477)
0.485	20	0.447	(0.416, 0.478)
0.485	200	0.453	(0.424, 0.481)

The New Yorker (December 7, 1992)



86

ANNALS OF RADIATION

THE CANCER AT SLATER SCHOOL

by Paul Brodeur

On a Friday afternoon in mid-December of 1990, half a dozen women who taught at the Louis N. Slater Elementary School, in Fresno, California, were interviewed in the teachers' lounge there by Amy Alexander, a staff writer for the Fresno

Alexander about the possible hazard at Slater appeared on the front page of the *Bee*, under the headline "POWER LINES WORRY SCHOOL." Alexander said in her article that the transmission lines on Emerson Avenue supplied power for more than forty thousand Fresno homes,

and pupils began a two-week Christmas vacation. Moreover, because Slater is a year-round school, teachers there are required to take two six-week vacations during the year, so nearly a dozen teachers, including several whom Alexander had interviewed, did not return to work

- The Slater school is an elementary school in Fresno, California
- Teachers and staff were "concerned about the presence of two high-voltage transmission lines that ran past the school ..."
- Their concern centered on the "high incidence of cancer at Slater ..."
- To address their concern, Dr. Raymond Neutra of the California Department of Health Services's Special Epidemiological Studies Program conducted a statistical analysis
- He found that there were 8 instances of cancer out of 145 total staff.
- He also calculated that given national rate, the average ages, the gender distribution, the expected number, or prevalance, would have been 4.2.

Likelihood

- Let θ be the chance of cancer (same for each employee)
- Let Y be the number of cancers out of 145 employees
- Then $Y \sim \text{Binomial}(n, \theta)$:

$$\Pr(Y = y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y},\ y = 0, ...n$$

- In particular, $Y \sim \text{Binomial}(145, \theta)$.
- We observed: The event $\{Y = 8\}$
- According to Dr. Neutra, the expected number of cancers is 4.2. We formulate a theory (Theory A): $\theta = 4.2/145 = 0.03$
- An alternative theory (Theory B): $\theta = 0.06$.

Likelihood

- To compare the theories we see how well each one explains the data. That is, for each value of θ, we calculate

$$\Pr(Y = 8|\theta) = \binom{145}{8}\theta^8(1-\theta)^{137}$$

- $\Pr(Y = 9|\theta = 0.03) \approx 0.036$
-
- $\Pr(Y = 9|\theta = 0.06) \approx 0.136$
- Hence, the alternative theory explains the data about four times as well.

The Likelihood Principle

- $\Pr(Y = y|\theta)$ is a function of two variables, y and θ. Once $Y = 8$ has been observed, then $\Pr(Y = 8|\theta)$ describes how well each theory, or value of θ, explains the data.
- It is a function only of θ; no value of Y other than 8 is relevant.
- Is $\Pr(Y = 9|\theta = 0.03)$ relevant? Does it describe how well theory explains the observed data?
- The Likelihood Principle: Once Y has been observed, say $Y = y_0$, then no other value of Y matters and we should treat $\Pr(Y = y_0|\theta)$ as a function only of θ.
- This principle is central to Bayesian thinking.

Prior and Likelihood

- Suppose Pr(Theory A) = Pr(Theory B) = 1/2
- Then,
$$\Pr(A|Y=8) = \frac{\Pr(Y=8|A)\Pr(A)}{\Pr(Y=8|A)\Pr(A) + \Pr(Y=8|B)\Pr(B)}$$
- This is ≈ 0.21
- What about Theory B?
- This is ≈ 0.79
- Hence, theory B is almost four times as likely after observing $Y = 8$.

A Frequentist Approach

Consider the hypothesis testing problem

$$H_0 : \theta = 0.03, \quad \text{versus} \quad H_1 : \theta > 0.03.$$

What is the p-value?

- The probability under H_0 of observing an outcome at least as extreme as the outcome actually observed.
- In the Slater problem,

$$\text{p-value} = \Pr(Y = 8 \mid \theta = 0.03) + \ldots + \Pr(Y = 145 \mid \theta = 0.03) \approx 0.07$$

Why the p-value is not appropriate here?

- Hypotheses should be compared by how well they explain the data,
- the p-value does not account for how well the alternative hypotheses explain the data, and
- the summands of $\Pr(Y = 9 \mid \theta = 0.03), \ldots, \Pr(Y = 145 \mid \theta = 0.03)$ are irrelevant because they do not describe how well any hypothesis explains any observed data

The p-value does not obey the Likelihood Principle!

Back to the Coin Example

$$X_1, X_2, ..., X_n \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta) \text{ where } n = 12.$$

$$H_0 : \theta = 0.5$$
$$H_1 : \theta > 0.5$$

Recall that from our experiment,

$$H, T, H, H, H, H, H, T, H, H, H, T$$

$$9H, 3T$$

Need to compute the probability of observing our result, or a more extreme result, under $H_0 : \theta = 0.5$.

P-value in the Coin Example

Let $Y = \sum X_i$. Then

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

$$\begin{aligned} P(Y \geq 9|\theta = 0.5, n = 12) &= \sum_{y=9}^{12} \binom{12}{y_i} \left(\frac{1}{2}\right)^{y_i} \left(\frac{1}{2}\right)^{12-y_i} \\ &= \left[\binom{12}{3} + \binom{12}{2} + \binom{12}{1} + \binom{12}{0}\right] \left(\frac{1}{2}\right)^{12} \\ &= \frac{299}{4096} \approx 0.07 \end{aligned}$$

The result is not statistically significant.

An Unexpected Mixup

- After you report the results back to your RA, you learn there was a mix up!
- You thought you told the RA to toss the coin 12 times.
- But your RA actually tossed it until observing 3 tails.
- As it happens, it took 12 tosses to get 3 tails.
- Should this matter? The evidence is what it is isn't it?
- But now n is random and Y is fixed.
- Thus, a result more extreme than $(9,3)$ is no longer $(10,2)$, $(11,1)$, and $(12,0)$. Rather, it is $(10,3)$, $(11,3)$, $(12,3)$, and so on.
- We have to re-calculate the p-value. Thoughts on how to do this?

$$\Pr(N = n|\theta, r) = \binom{n-1}{r-1}\theta^r(1-\theta)^{n-r}$$

where r is the number of tails.

One More Try

$$
\begin{aligned}
P(N \geq 12|\theta = 0.5, r = 3) &= \sum_{n=12}^{\infty} \binom{n_i - 1}{r - 1} .5^r .5^{n_i - r} \\
&= \sum_{n=12}^{\infty} \binom{n_i - 1}{2} .5^{n_i} \\
&= 1 - \sum_{n=1}^{11} \binom{n_i - 1}{2} .5^{n_i} \\
&\approx 0.03
\end{aligned}
$$

Now the result *is* statistically significant!

What if the RA stopped tossing the coin so that they can get a coffee?

Or to watch Narcos on Netflix?

Sometimes there are ethical reasons to stop collecting data
(HIV antiretroviral drugs, Covid-19 antiviral drugs)

Prediction

- An important feature of Bayesian inference is the existence of a predictive distribution for new observations.
- Let $y_1, ...y_n$ be the outcomes from a $Y_1, ..., Y_n \sim \text{Bernoulli}(\theta)$ sample. And let $\widetilde{Y} \in \{0, 1\}$ be an additional outcoem from the same population that has yet to be observed.
- The predictive distribution of $\widetilde{Y}$ is the conditional distribution of $\widetilde{Y}$ given $\{Y_1 = y_1, ..., Y_n = y_n\}$. For conditionally iid Bernoulli RVs this distribution can be dervived from the distribution of $\widetilde{Y}$ given θ and the posterior distribution of θ.

Prediction

$$
\begin{aligned}
\Pr(\widetilde{Y} = 1 | y_1, ... y_n) &= \int_0^1 \Pr(\widetilde{Y} = 1, \theta | y_1, ... y_n) d\theta \\
&= \int_0^1 \Pr(\widetilde{Y} = 1 | \theta, y_1, ... y_n) \pi(\theta | y_1, ... y_n) d\theta \\
&= \int_0^1 \theta \pi(\theta | y_1, ... y_n) d\theta \\
&= \mathrm{E}[\theta | y_1, ... y_n] = \frac{\alpha + \sum_{i=1}^n y_i}{\alpha + \beta + n} \\
\rightarrow \Pr(\widetilde{Y} = 0 | y_1, ... y_n) &= 1 - \mathrm{E}[\theta | y_1, ... y_n] \\
&= \frac{\beta + \sum_{i=1}^n (1 - y_i)}{\alpha + \beta + n}
\end{aligned}
$$

Prediction

- The predictive distribution does not depend on any unknown quantities. If it did, we would not be able to use it to make predictions.
- The predictive distribution depends on observed data. Otherwise, we could never infer anything about the unsampled population from the sampled cases.
- Laplace's Rule of Succession: $(y + 1)/(n + 2)$. The predictive distribution under a uniform prior.

Normal Models

- Important in many statistical modeling problems
- Often useful as approximation or a component in more complicated models
- We will treat separately cases with known variance and known mean.

Natural Conjugate Priors

If $\mathcal{F}$ is a class of sampling distributions $\pi(y \mid \theta)$, and $\mathcal{P}$ is a class of prior distributions for θ, then the class $\mathcal{P}$ is conjugate for $\mathcal{F}$ is

$$\pi(\theta \mid y) \in \mathcal{P}, \text{ for all } \pi(\cdot \mid \theta) \in \mathcal{F} \text{ and } \pi(\cdot) \in \mathcal{P}.$$

Is beta distribution conjugate for binomial distribution?

What distributions are conjugate for normal distribution?

The natural conjugate prior families: $\mathcal{P}$ is the set of all densities having the same functional form as the likelihood.

Normal Model: Unknown Mean, Known Variance

- Suppose $x_i|\mu \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$ with σ^2 known. Let $x = (x_1, ..., x_n)$.
- What is the likelihood?

$$\begin{aligned}\pi(x|\mu, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\ &\propto \exp\left[-(2\sigma^2)^{-1}\sum_{i=1}^{n}(x_i-\mu)^2\right]\end{aligned}$$

- What is the natural conjugate prior?
- Goal: pick a prior that has the same functional form as the likelihood, then derive the posterior and evaluate whether it too will have the same functional form.
- Expanding the quadratic term in the exponent, we see that $\pi(x_1, ..., x_n|\mu, \sigma^2)$ depends on $x_1, ..., x_n$ through:

$$\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}x_i^2 - 2\frac{\mu}{\sigma^2}\sum_{i=1}^{n}x_i + n\frac{\mu^2}{\sigma^2}$$

- We simplify this and isolate μ on the next slide.

Unknown mean, known variance

- Write the likelihood as follows:

$$\pi(x|\mu) \propto \exp\left[A\mu^2 + B\mu + C\right]$$

- Note that:

  $A = -n(2\sigma^2)^{-1}$

  $B = \sigma^{-2}\sum_{i=1}^n x_i$

  $C = -(2\sigma^2)^{-1}\sum_{i=1}^n x_i^2$

- To specify natural conjugate prior, set

$$\pi(\mu) \propto \exp\left[a^*\mu^2 + b^*\mu + c^*\right] = \exp\left[-\frac{1}{2\tau_0^2}(\mu - \mu_0)^2\right]$$

- This implies that $\mu \sim \mathrm{N}(\mu_0, \tau_0^2)$ with hyperparameters μ_0 and τ_0^2.
- Now: if $\pi(\mu|\sigma^2) \sim \mathrm{N}$ and $x_1, \ldots x_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ is $\pi(\mu|x_1, \ldots x_n, \sigma^2)$ also normal?