Intro to Bayesian Statistics: Inference

CSDE Workshop

Jessica Godwin

March 2, 2023

Resources and Support

Resources and Support

Texts

- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013). Bayesian Data Analysis, 3rd ed. Chapman and Hall/CRC.
- McElreath, R. (2020). Statistical Rethinking: A Bayesian Course with Examples in R and Stan, 2nd ed. Chapman and Hall/CRC.
- Casella, G., & Berger, R. L. (2002). Statistical Inference, 2nd ed. Cengage Learning.

What is Bayesian inference?

3 Steps of Bayesian data analysis

According to Gelman et al. (2013), there are three steps to Bayesian data analysis:

1. Setting up a **full probability model**.
   - Specify joint probability distribution for all observable $(y)$and unobservable quantities (θ).
2. Conditioning on observed data, then calculating & interpreting the **posterior distribution**.
3. Evaluating the fit of the model.
   - How well does the model fit the data?
   - Are the substantive conclusions reasonable?
   - How sensitive are the results to modeling assumptions in Step 1?

Step 1: Specifying a full probability model.

Suppose we have observations y_i, $i = 1, \ldots, n$ and we assume:

- they come from some probability distribution with parameters θ, i.e. specify the **likelihood** $p(\mathbf{y}|\theta)$, and
- assume a priori what values of θ might be plausible, i.e. specify the **prior** $p(\theta)$.

Then, we can either perform:

- **Bayesian inference**, i.e. learn something about θ from our observed data, or
- **Bayesian prediction**, i.e. learn something about unobserved (but potentially observable) data, $\tilde{y}$, from our observed data.

Step 2: Inference

After specifying our full probability model, i.e. the likelihood and the prior, we can calculate the **posterior distribution** of θ, $p(\theta|y)$:

$$\underbrace{p(\theta|y)}_{\text{posterior distribution}} = \frac{\overbrace{p(\theta, y)}^{\text{sampling distribution}}}{\underbrace{p(y)}_{\text{prior predictive distribution}}} = \frac{\overbrace{p(y|\theta)}^{\text{likelihood}}\ \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{\int_\theta p(y|\theta)p(\theta)d\theta}_{\text{marginal distribution}}}.$$

- **Point estimates:** posterior mean, median or mode
- **Uncertainty:** posterior standard deviation or interquartile range, posterior intervals, or highest density posterior intervals
- **Both:** full posterior distribution (histograms, densities, contour plots)

Step 2: Prediction

After specifying our full probability model, i.e. the likelihood and the prior, we can calculate the **posterior predictive distribution distribution** of $\tilde{y}$, $p(\tilde{y}|y)$ with some tricks from conditional probability:

$$\underbrace{p(\theta|y)}_{\text{posterior predictive distribution}} = \int_\theta p(\tilde{y},\theta|y)d\theta = \int_\theta p(\tilde{y}|\theta,y)\,\overbrace{p(\theta|y)}^{\text{posterior}}\,d\theta = \int_\theta p(\tilde{y}|\theta)\,\overbrace{p(\theta|y)}^{\text{posterior}}\,d\theta.$$

- **Point estimates:** posterior predictive mean, median or mode
- **Uncertainty:** posterior predictive standard deviation or interquartile range, posterior predictive intervals, or highest density posterior predictive intervals
- **Both:** full posterior predictive distribution (histograms, densities, contour plots)

Step 2: Computing the marginal distribution

Calculating $p(y) = \int_\theta p(y|\theta)p(\theta)d\theta$ can be difficult.

Possibile Methods:

- Calculate it analytically (often not easy or possible)
  - Choosing conjugate likelihood-prior pairs leads to a known, closed-form posterior.
- Approximate the posterior distribution
  - Examples: grid approximation, quadratic or Normal approximation, Laplace approximation (INLA, TMB)
- Sampling from the posterior distribution
  - Markov Chain Monte Carlo (WinBUGS, JAGS)– Gibbs sampling & Metropolis-Hastings, Hamiltonian Monte Carlo (Stan)

Inference: Conjugate Priors

Conjugacy

- The property that the posterior distribution follows the same parametric form as the prior distribution is called **conjugacy**.
- The prior and posterior distributions that have this property with a particular likelihood are called a **conjugate family** to the likelihood.

Examples of conjugate families:

Likelihood	Conjugate family
Binomial	Beta
Multinomial	Dirichlet
Poisson	Gamma
Exponential	Gamma
Normal (mean)	Normal
Normal (mean, variance)	Normal, Inverse Gamma

The Beta-Binomial Model

- **Likelihood:** Let $X_1, \ldots, X_n$ be iid Bernoulli(p), so that $Y = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p)$.
- **Prior:** If we assume $p \sim \text{Beta}(\alpha, \beta)$,
- **Posterior:** what is the distribution of $p|y, n$?

$$
\begin{aligned}
p|y, n &\sim \text{Beta}(\alpha + y, \beta + n - y) \\
p(p|y, n) &= \frac{\Gamma(\alpha + \beta + n)}{\Gamma(\alpha + y)\Gamma(\beta + n - y)} p^{\alpha + y - 1}(1 - p)^{\beta + n - y - 1} \\
E[p|y, n] &= \frac{\alpha + y}{\alpha + y + \beta + n - y} = \frac{\alpha + y}{\alpha + \beta + n} \\
Var(p|y, n) &= \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}
\end{aligned}
$$

The Beta-Binomial Model: An Example

Suppose we sample $n = 100$ individuals from a population in an attempt to estimate the **support ratio**, or ratio of individuals who are 15-64 to those who are 65+. Let y be a binary outcome indicating an individual is 65+.

Age	N	y
0-14	13	
15-64	72	0
65+	15	1

Step 1: Specify **binomial likelihood** for y,

$$p(y = 15|p, n = 87) = \begin{pmatrix} 87 \\ 15 \end{pmatrix} p^{15}(1 - p)^{87},$$

and specify a **Beta(α =2,β =2) prior** for p,

$$p(p) = \frac{\Gamma(4)}{\Gamma(2)\Gamma(2)} p^{2-1}(1 - p)^{2-1}.$$

The Beta-Binomial Model: An Example

Step 2: Calculate the **Beta($\alpha + y$, $\beta + n - y$) posterior distribution** for p,

$$p(p|y,n) = \frac{\Gamma(4+87)}{\Gamma(2+15)\Gamma(2+72)} p^{2+15-1}(1-p)^{2-1}.$$

Step 2, cont'd: Make inference about p.

$$E[p|y=15, n=87, \alpha=2, \beta=2] = \frac{\alpha+y}{\alpha+\beta+n} = \frac{17}{2+2+87} = 0.187$$

$$\sqrt{Var(p|y,n)} = \sqrt{\frac{(\alpha+y)(\beta+n-y)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}} = \sqrt{\frac{17\times 74}{91^2\times 92}} = 0.041.$$

The Beta-Binomial Model: An Example

Step 2, cont'd: Make inference about p.

$$p|y = 15, n = 87, \alpha = 2, \beta = 2 \sim \text{Beta}(17, 74)$$

$$E[p|y = 15, n = 87, \alpha = 2, \beta = 2] = \frac{17}{91} = 0.187 \; \sqrt{Var(p|y, n)} = \sqrt{\frac{17 \times 74}{91^2 \times 92}} = 0.041.$$

The posterior mean proportion of individuals aged 15 or older who are 65+ is 0.187 (0.041).

```
qbeta(c(0.025, 0.975), shape1 = 17, shape2 = 74)
```

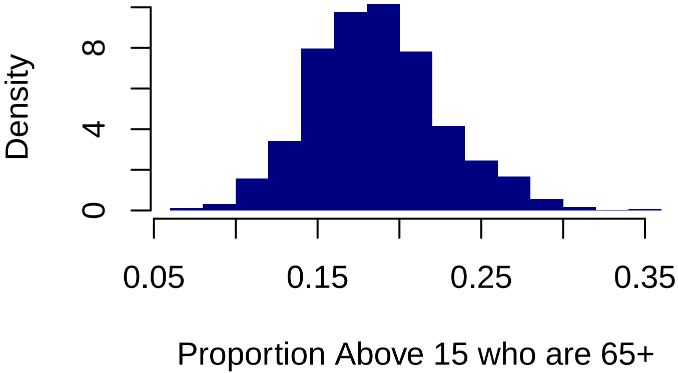
```
## [1] 0.1140597 0.2725848
```

The 95% posterior (or credible) interval for the proportion of individuals age 15 or older who are 65+ is (.114, .273).

The Beta-Binomial Model: An Example

Step 2, cont'd: Make inference about p using $p|y \sim \text{Beta}(17, 74)$

```
post_prob <- rbeta(n = 1000, shape1 = 17, shape2 = 74)
hist(post_prob, main = "", xlab = "Proportion Above 15 who are 65+",
     border = FALSE, col = "navy", freq = FALSE)
```



The Beta-Binomial Model: An Example

Step 2, cont'd: Make inference about **the support ratio** using $p|y \sim \text{Beta}(17, 74)$

```
support_ratio <- (1 - post_prob)/post_prob
c(mean(support_ratio), sd(support_ratio))
```

```
## [1] 4.643837 1.308973
```

```
quantile(support_ratio, probs = c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 2.714640 7.929339
```

The posterior mean of the support ratio is 4.64 (1.31) persons 15-64 for every person 65+. The 95% posterior interval for the support ratio is (2.71, 7.93).

Inference: Grid Approximation

Inference: Grid Approximation

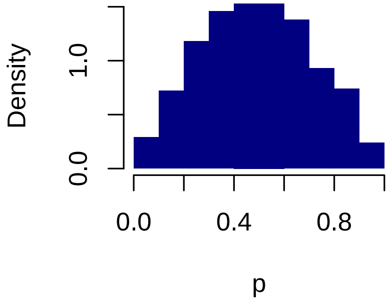
One option for approximating the posterior distribution is **grid approximation**:

1. Specify the likelihood ($p(y|\theta)$) and prior distributions ($p(\theta)$).
2. Pick S values of θ that span the support of the prior $p(\theta)$.
3. Evaluate $p(\theta_s)$ and $p(y|\theta_s)$ for all $s = 1, \ldots, S$.
4. Calculate $p(y) = \sum_{s=1}^{S} p(y|\theta_s)p(\theta_s)$.
5. Evaluate the posterior $\dfrac{p(y|\theta_s)p(\theta_s)}{p(y)}$ for all $s = 1, \ldots, S$.
6. Use the S values of the posterior to produce point estimates of θ, quantify uncertainty about those estimates, or to approximate the posterior distribution as a whole.

Grid Approximation: An Example

Let's return to our previous example estimating the proportion of individuals above 15 who are 65+ and using that to estimate the support ratio.

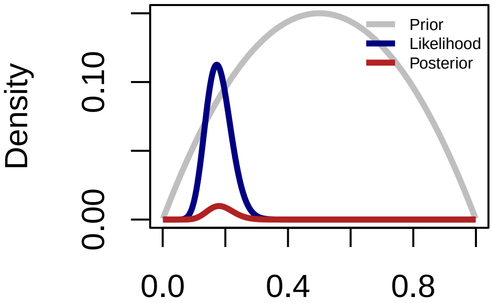
$$y|p, n = 87 \sim \text{Bin}(n = 87, p)$$
$$p \sim \text{Beta}(2,2)$$



Grid Approximation: An Example

The support for the Beta(2, 2) distribution is (0,1).

```
p_grid <- seq(0.001, 0.999, .001)
prior_eval <- dbeta(p_grid, shape1 = 2, shape2 = 2)
likelihood_eval <- dbinom(15, size = 87, prob = p_grid)
marg_calc <- sum(likelihood_eval*prior_eval)
post_eval <- (1/marg_calc)*likelihood_eval*prior_eval
```



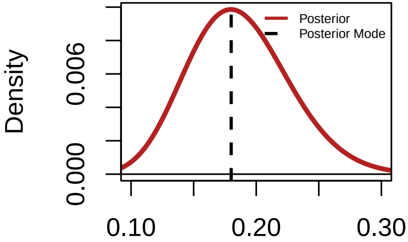
Grid Approximation: An Example

At what value of p does the posterior distribution attain its maximum?

```
max_val_idx <- which.max(post_eval)
p_grid[max_val_idx]
```

```
## [1] 0.18
```

The posterior mode proportion of individuals aged 15 or older who are 65+ is 0.18.



The Normal-Normal Model

- **Likelihood:** Let $Y_i \sim \text{Normal}(\mu, \sigma^2)$, for $i = 1, \ldots, n$ where σ^2 is known.
- **Prior:** If we assume $\mu \sim \text{Normal}(\theta, \tau^2)$, where θ and τ^2 are known values,
- **Posterior:** what is the distribution of $\mu|\mathbf{y}, \sigma^2, \theta, \tau^2$?

$$\mu|\mathbf{y}, \sigma^2, \theta, \tau^2 \sim \text{Normal}\left(\bar{y} \times \frac{\tau^2}{\sigma^2/n + \tau^2} + \theta \times \frac{\sigma^2/n}{\sigma^2/n + \tau^2}, \frac{\tau^2\sigma^2/n}{\tau^2 + \sigma^2/n}\right)$$

$$E[\mu|\mathbf{y}, \sigma^2, \theta, \tau^2] = \bar{y} \times \frac{\tau^2}{\sigma^2/n + \tau^2} + \theta \times \frac{\sigma^2/n}{\sigma^2/n + \tau^2}$$

$$Var(\mu|\mathbf{y}, \sigma^2, \theta, \tau^2) = \frac{\tau^2\sigma^2/n}{\tau^2 + \sigma^2/n}$$

Question: What happens when $n \to \infty$?

The Normal-Normal Model: Regression

Let $Y_1, \ldots, Y_n \stackrel{iid}{\sim} \text{Normal}(\beta_0 + \beta_1 X_i, \sigma^2)$, where σ^2 is known. If we assume $\beta \sim \text{Normal}(\theta, \Sigma_\theta)$, where $\theta = \begin{bmatrix} \theta_0 & \theta_1 \end{bmatrix}$ and Σ_θ are known values, what is the distribution of $\beta | y, \sigma^2, \theta, \Sigma_\theta$ where $\beta = \begin{bmatrix} \beta_0 & \beta_1 \end{bmatrix}$?

$$P(\beta|\mathbf{y}, \mathbf{x}, \sigma^2, \theta, \Sigma) \sim \text{Normal}\left(\left[\Sigma_\theta + \frac{\sum_{i=1}^n x_i^2}{\sigma^2}\right]^{-1} \frac{\sum_{i=1}^n x_i y_i}{\sigma^2}, \left[\Sigma_\theta + \frac{\sum_{i=1}^n x_i^2}{\sigma^2}\right]^{-1}\right)$$

$$P(\beta|\mathbf{y}, \mathbf{x}, \sigma^2, \theta, \Sigma) \sim \text{Normal}\left(\left[\Sigma_\theta + \Sigma^{-1}\mathbf{x}^T\mathbf{x}\right]^{-1} \Sigma^{-1}\mathbf{x}^T\mathbf{y}, \left[\Sigma_\theta + \Sigma^{-1}\mathbf{x}^T\mathbf{x}\right]^{-1}\right)$$

$$E[\beta|\mathbf{y}, \mathbf{x}, \sigma^2, \theta, \Sigma] = \left[\Sigma_\theta + \Sigma^{-1}\mathbf{x}^T\mathbf{x}\right]^{-1} \Sigma^{-1}\mathbf{x}^T\mathbf{y}$$

$$Var(\beta|\mathbf{y}, \mathbf{x}, \sigma^2, \theta, \Sigma) = \left[\Sigma_\theta + \Sigma^{-1}\mathbf{x}^T\mathbf{x}\right]^{-1}$$