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.

Review

Reviewing Bayes' (and Laplace's) Rule

- **Rules of conditional probability**

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

- **Bayes' Rule**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Probability: Bayes' Rule - Testing Example

A disease has a prevalence of 1% in the population. A blood test for the disease has high sensitivity (the probability of a positive test if someone is sick) and specificity (the probability of a negative test if someone is not sick).

- If someone has the disease, there is a 98% chance they will test positive.
- If someone does not have the disease, there is a 95% chance they will test negative

Suppose you test positive for the disease and you want to figure out the chance that you have the disease. That is, given someone has tested positive for the disease what is the probability that they have the disease?

Probability: Bayes' Rule - Testing Example

A disease has a prevalence of 1% in the population. A blood test for the disease has high sensitivity (the probability of a positive test if someone is sick) and specificity (the probability of a negative test if someone is not sick).

- If someone has the disease, there is a 98% chance they will test positive.
- If someone does not have the disease, there is a 95% chance they will test negative

Suppose you test positive for the disease and you want to figure out the chance that you have the disease. That is, given someone has tested positive for the disease what is the probability that they have the disease?

What do we know?

Probability: Bayes' Rule - Testing Example

A disease has a prevalence of 1% in the population. A blood test for the disease has high sensitivity (the probability of a positive test if someone is sick) and specificity (the probability of a negative test if someone is not sick).

- If someone has the disease, there is a 98% chance they will test positive.
- If someone does not have the disease, there is a 95% chance they will test negative

Suppose you test positive for the disease and you want to figure out the chance that you have the disease. That is, given someone has tested positive for the disease what is the probability that they have the disease?

What do we know?

- $P(+|D^+) = 0.98$, $P(-|D^+) = 0.02$

Probability: Bayes' Rule - Testing Example

A disease has a prevalence of 1% in the population. A blood test for the disease has high sensitivity (the probability of a positive test if someone is sick) and specificity (the probability of a negative test if someone is not sick).

- If someone has the disease, there is a 98% chance they will test positive.
- If someone does not have the disease, there is a 95% chance they will test negative

Suppose you test positive for the disease and you want to figure out the chance that you have the disease. That is, given someone has tested positive for the disease what is the probability that they have the disease?

What do we know?

- $P(+|D^+) = 0.98,\ P(-|D^+) = 0.02$
- $P(-|D^-) = 0.95,\ P(+|D^-) = 0.05$

Probability: Bayes' Rule - Testing Example

A disease has a prevalence of 1% in the population. A blood test for the disease has high sensitivity (the probability of a positive test if someone is sick) and specificity (the probability of a negative test if someone is not sick).

- If someone has the disease, there is a 98% chance they will test positive.
- If someone does not have the disease, there is a 95% chance they will test negative

Suppose you test positive for the disease and you want to figure out the chance that you have the disease. That is, given someone has tested positive for the disease what is the probability that they have the disease?

What do we know?

- $P(+|D^+) = 0.98$, $P(-|D^+) = 0.02$
- $P(-|D^-) = 0.95$, $P(+|D^-) = 0.05$
- $P(D^+) = 0.01$, $P(D^-) = 0.99$

Probabilty: Bayes' Rule - Testing Example

- $P(+|D^+) = 0.98,\ P(-|D^+) = 0.02$
- $P(-|D^-) = 0.95,\ P(+|D^-) = 0.05$
- $P(D^+) = 0.01,\ P(D^-) = 0.99$

$$
\begin{aligned}
P(D^+|+) = \frac{P(+|D^+)P(D^+)}{P(+)} &= \frac{P(+|D^+)P(D^+)}{P(+ \cap D^+) + P(+ \cap D^-)} \\
&= \frac{P(+|D^+)P(D^+)}{P(+|D^+)P(D^+) + P(+|D^-)P(D^-)} \\
&= \frac{0.98 \cdot 0.01}{0.98 \cdot 0.01 + 0.05 \cdot 0.99} \\
&= 0.165
\end{aligned}
$$

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:

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

- **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 $p|y, n$?

$$p|y, n \sim \text{Beta}(\alpha + y, \beta + n - y)$$

$$p(p|y, n) = \frac{\Gamma(\alpha + \beta + n)}{\Gamma(\alpha + y)\Gamma(\beta + n - y)} p^{\alpha+y-1}(1-p)^{\beta+n-y-1}$$

$$E[p|y, n] = \frac{\alpha + y}{\alpha + y + \beta + n - y} = \frac{\alpha + y}{\alpha + \beta + n}$$

$$Var(p|y, n) = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}$$

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).

On Frequentist vs. Bayesian intervals

- Frequentist confidence intervals have the interpretation that, if one were to repeat the experiment that led to their observed data over and over again $(1-\alpha)\%$ of the time the confidence interval

$$\hat{\theta} \pm z_{1-\alpha/2} \times SE(\hat{\theta})$$

would contain the true parameter of interest, θ.

  - Any single frequentist CI is NOT a probability statement, the true parameter θ is either in that interval or not.
  - However, people often WANT to interpret frequentist CIs the way a Bayesian interval CAN BE interpretted.

- Bayesian intervals ARE probability statements with the interpretation that based on the model (prior and likelihood choice), there is a $(1-\alpha)\%$ chance θ lies in the interval.

On Frequentist vs. Bayesian intervals

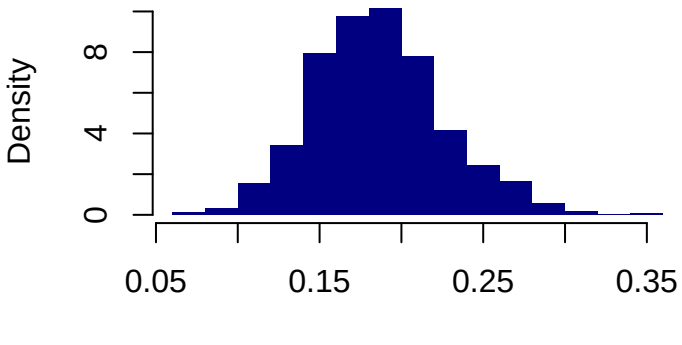
- Bayesian intervals ARE probability statements with the interpretation that based on the model (prior and likelihood choice), there is a $(1-\alpha)\%$ chance θ lies in the interval.

The 95% posterior (or credible) interval for the proportion of individuals age 15 or older who are 65+ is (.114, .273), i.e. given our prior choice and observed data, there is a 95% chance the true proportion of individuals 15 or older who are 65+ lies between (.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 p using $p|y \sim \text{Beta}(17, 74)$

```
sum(post_prob > 15/87)/length(post_prob)
```

```
## [1] 0.605
```

```
sum(post_prob > 0.25)/length(post_prob)
```

```
## [1] 0.071
```

Based on our observed data and prior choice, there is a 60.5% chance that the true proportion of individuals aged 15 or older who are 65+ is higher than what we observed.

Based on our observed data and prior choice, there is a 7.1% chance a quarter or more of the individuals 15+ are 65+.

These are called **exceedance probabilities**.

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).

The Beta-Binomial Model: An Example

Step 2, cont'd: Simulate predicted values from $p(\tilde{y}|y) = \int_p p(\tilde{y}|p, n)p(p|y, n)dp$:

```
post_pred <- rbinom(1000, size = 87, prob = post_prob)
c(mean(post_pred), sd(post_pred))
```

```
## [1] 16.13900  4.97035
```

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

```
##  2.5% 97.5%
##     7    27
```

The posterior predictive mean is is 16.14 (4.97) persons 65+ for every sample of $n = 87$ individuals 15 and above. The 95% **posterior predictive interval** (7, 27).

Inference: Grid Approximation

Inference: Grid Approximation

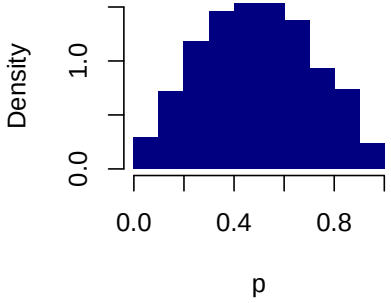
One option for approximating the posterior distribution is **grid approximation**:

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.

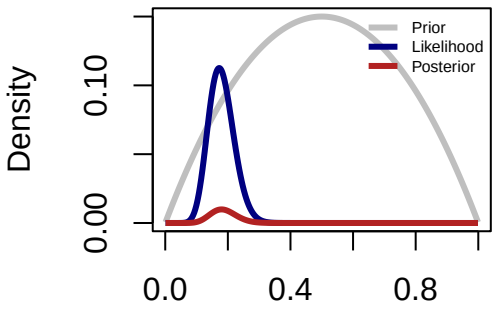
$$y|p, n = 87 \sim \text{Bin}(n = 87, p)$$
$$p \sim \text{Beta}(2,2)$$



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
```



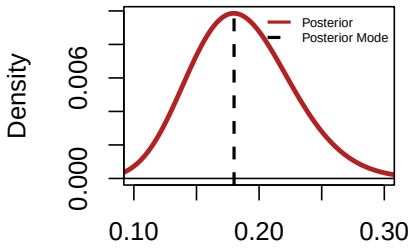
Grid Approximation: An Example

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.



Grid Approximation: An Example

Posterior medians and intervals by sampling from the approximated posterior:

```
post_samp_grid <- sample(p_grid, size = 1000, replace = TRUE,
                         prob = post_eval)
quantile(post_samp_grid, c(0.025, .5, 0.975))
```

```
##      2.5%       50%     97.5%
## 0.115000 0.185000 0.267025
```

The posterior median proportion of individuals 15+ who are 65+ is 0.184, and the 95% posterior interval is (0.11, 0.27).

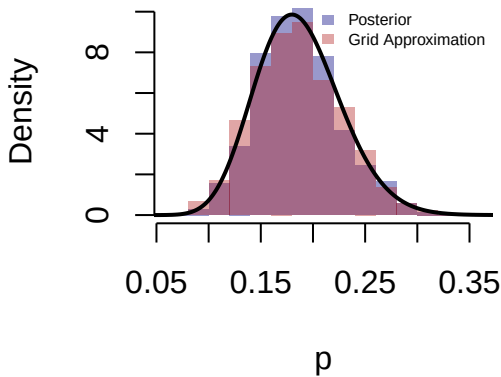
Comparison to the analytical posterior

Let's compare our inference based on this grid approximation of the posterior to the inference based on the true posterior distribution we know via conjugacy.

Measure	True Value	Samples	Grid Approx
Mean	0.187	0.1858	0.1863
Median	0.1845	0.1841	0.1840
SD	0.0406	0.0398	0.0421
95% CI	(0.1140, 0.2726)	(0.1120, 0.2692)	(0.1110, 0.2721)
$P(p \geq 0.25)$	0.0683	0.0710	0.075

Comparison to the analytical posterior

Let's compare our inference based on this grid approximation of the posterior to the inference based on the true posterior distribution we know via conjugacy.



Why quadratic approximation or MCMC?

- As we have seen, grid approximation can provide a very good approximation to the posterior. So, why do we need to talk about other methods?
- Grid approximation becomes unwieldy as the number of parameters grow, we have only looked at a single parameter model so far.
- **Quadratic approximation** involves two steps:
  1. Find posterior mode via optimization
  2. Calculate curvature (second derivative at peak)
     - Quality of approximation increases as $n \to \infty$
     - Often equal to the maximum likelihood estimate
     - Breaks if the Hessian (matrix of 2nd derivatives) can't be calculated
- **MCMC** is useful when quadratic approximation doesn't work or lots and lots of parameters