GENERAL ASSEMBLY

BAYESIAN INFERENCE

Matt Brems

DSI+

LEARNING OBJECTIVES

- Describe the relationships among parameter, statistic, sample, and population.
- Understand how Bayes' Theorem connects to Bayesian inference.
- Describe the posterior distribution.
- Identify methods for choosing a prior and a likelihood.
- Define improper prior, uninformative prior, informative prior, hierarchical modeling, and hyperparameter.
- Understand conjugacy and describe its benefits.
- Understand how simulations play such a large role in Bayesian inference.

REVIEW OF INFERENCE

FREQUENTIST VS. BAYESIAN INFERENCE OF PARAMETERS

- Frequentist inference and Bayesian inference have different interpretations of probability, and these interpretations give rise to different methods of analysis.

FREQUENTIST VS. BAYESIAN INFERENCE OF PARAMETERS

- Frequentist inference and Bayesian inference have different interpretations of probability, and these interpretations give rise to different methods of analysis.

  - Example: The average height of women at Ohio State, denoted μ.
    - Frequentists treat μ as fixed: $\mu = 64 \text{ inches}$
    - Bayesians treat μ as a parameter with a distribution: $\mu \sim N(64, 2)$

FREQUENTIST VS. BAYESIAN INFERENCE OF PARAMETERS

- Example: The average height of women at Ohio State, denoted μ.
  - Frequentists treat μ as fixed: $\mu = 64 \text{ inches}$
  - Bayesians treat μ as a parameter with a distribution: $\mu \sim N(64, 2)$
- Frequentist:
- Bayesian:

RECALL BAYES' RULE

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

- $P(A)$ is the probability that A occurs given no supplemental information.
- $P(B|A)$ is the likelihood of seeing evidence (data) B assuming that A is true.
- $P(B)$ is what we scale $P(B|A)P(A)$ by to ensure we are only looking at A within the context of B occurring.

BAYES' RULE: DIACHRONIC INTERPRETATION

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \Rightarrow P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

BAYES' RULE: DIACHRONIC INTERPRETATION

Suppose we flip a coin and want to see if it's likely that the probability of flipping heads is 50%.

$$P(p = 0.5|data) = \frac{P(data|p = 0.5)P(p = 0.5)}{P(data)}$$

WHAT IF WE LOOK AT ALL POSSIBLE HYPOTHESES?

$$P(p = 0|D) = \frac{P(D|p = 0)P(p = 0)}{P(D)}$$

$$P(p = 0.001|D) = \frac{P(D|p = 0.001)P(p = 0.001)}{P(D)}$$

$$\vdots$$

$$P(p = 1|D) = \frac{P(D|p = 1)P(p = 1)}{P(D)}$$

WHAT IF WE LOOK AT ALL POSSIBLE HYPOTHESES?

- Instead of manually writing out every possible hypothesis (time-consuming, impossible every time we want to learn about a continuous parameter), what if we combined each of these individual probabilities into one distribution?

$$P(p = 0|D) = \frac{P(D|p = 0)P(p = 0)}{P(D)} \Rightarrow f(p|D) = \frac{f(D|p)f(p)}{f(D)}$$

BAYES' RULE: PARAMETER INFERENCE

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \Rightarrow f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)}$$

- $f(\theta)$ is the distribution of θ given no supplemental information.
  - "Prior Distribution of θ"
- $f(y|\theta)$ is the likelihood function relating y and θ.
  - "Likelihood"
- $f(y)$ is the normalizing constant to ensure $f(\theta|y)$ is a valid probability distribution.
  - "Marginal Likelihood of y"

“PROPORTIONAL TO”

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)} \propto f(y|\theta)f(\theta)$$

- We often ignore the $f(y)$ component in the denominator and simply say that the posterior $f(\theta|y)$ is **proportional to** $f(y|\theta)f(\theta)$.

- Why?

"PROPORTIONAL TO"

- Autocorrect Example: I type the word "radom" into my phone. My phone has to decide to leave the word as "radom," change to "radon," or change to "random."

“PROPORTIONAL TO”

- Autocorrect Example: I type the word “radom” into my phone. My phone has to decide to leave the word as “radom,” change to “radon,” or change to “random.”

- If we have three values of θ and we calculate:

$$P(\theta = radom|y) \propto P(y|\theta = radom)P(\theta = radom) = 5$$
$$P(\theta = radon|y) \propto P(y|\theta = radon)P(\theta = radon) = 5$$
$$P(\theta = random|y) \propto P(y|\theta = random)P(\theta = random) = 10$$

…it’s very easy for us to convert $f(\theta|y)$ into a valid probability distribution.

POSTERIOR DISTRIBUTION

- The posterior distribution $f(\theta|y)$ represents all possible values of θ and how frequently we observe each of these values, given the data we've observed.
  - The posterior distribution is a **complete summary of our parameter of interest $\boldsymbol{\theta}$ that takes into account our data $\boldsymbol{y}$.**

POSTERIOR DISTRIBUTION

- The posterior distribution $f(\theta|y)$ represents all possible values of θ and how frequently we observe each of these values, given the data we've observed.
  - The posterior distribution is a **complete summary of our parameter of interest $\boldsymbol{\theta}$ that takes into account our data $\boldsymbol{y}$.**

- In order to construct this posterior distribution $f(\theta|y)$, we need two things:
  - $f(\theta)$, the prior distribution of θ.
  - $f(y|\theta)$, the likelihood of observing the data y under some model.

POSTERIOR DISTRIBUTION

- The posterior distribution $f(\theta|y)$ represents all possible values of θ and how frequently we observe each of these values, given the data we've observed.
  - The posterior distribution is a **complete summary of our parameter of interest $\boldsymbol{\theta}$ that takes into account our data $\boldsymbol{y}$.**

- In order to construct this posterior distribution $f(\theta|y)$, we need two things:
  - $f(\theta)$, the prior distribution of θ.
  - $f(y|\theta)$, the likelihood of observing the data y under some model.

- We can think of our posterior distribution $f(\theta|y)$ as a combination of our data and our prior.
  - $f(\theta|y) \propto f(y|\theta) \times f(\theta) = likelihood \times prior$

ESTIMATING A PRIOR DISTRIBUTION

PRIOR INFLUENCE ON THE POSTERIOR

- We can think of our posterior distribution $f(\theta|y)$ as a combination of our data and our prior.
  - $f(\theta|y) \propto f(y|\theta) \times f(\theta) = likelihood \times prior$

- If our prior is too specific, then our posterior will be "dominated by" the prior.

- If our prior is too vague, then our posterior will be "dominated by" the data through the likelihood.

PRIOR INFLUENCE ON THE POSTERIOR

- If our prior is too specific, then our posterior will be "dominated by" the prior.

- If our prior is too vague, then our posterior will be "dominated by" the data through the likelihood.

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

- If $P(A) = 0, P(A|B) = 0$.

- If $P(A) = 1, P(B|A) = P(B) \Rightarrow P(A|B) = 1$.

TERMS

- Improper Priors
  - Priors that are not valid probability functions.

- Uninformative Priors
  - Includes minimal information about θ (i.e. physical limitations)

- Informative Priors
  - Includes prior knowledge about θ by taking past data and information into account. (i.e. scientific research)

BAYESIAN & FREQUENTIST STATISTICS

- Say we want to conduct inference on μ, the mean height of American adults.
  - Recall: A prior summarizes our beliefs about μ before observing any data.
  - What is an example of an **improper prior**?
  - What is an example of an **uninformative prior**?
  - What is an example of an **informative prior**?

BAYESIAN & FREQUENTIST STATISTICS

- Frequentist analysis makes no assumptions about the prior distribution of the parameter.

- You can think of a completely flat Uniform, improper prior distribution - this is equivalent to frequentism!

SPECIFYING THE LIKELIHOOD

DEFINITIONS

- We can think of our posterior distribution $f(\theta|y)$ as a combination of our data and our prior.
  - $f(\theta|y) \propto f(y|\theta) \times f(\theta) = likelihood \times prior$

- We want our likelihood to reflect the model that allows us to observe the data we observe.
  - If my data was observing k heads out of n coin flips, the Binomial distribution is probably a good model for how many heads I observe.
  - If my data was observing the number of people who visit my website in a fixed amount of time, the Poisson or Negative Binomial distribution might be a good model.

LIKELIHOOD PRINCIPLE

- The likelihood principle tells us that the data influences our posterior distribution **only** through the likelihood function.
  - The data should not influence our posterior distribution through the prior!

LIKELIHOOD PRINCIPLE

- The likelihood principle tells us that the data influences our posterior distribution **only** through the likelihood function.
  - The data should not influence our posterior distribution through the prior!
    - We may estimate a prior distribution from a pilot study or previous knowledge, but the data for our experiment/analysis should only affect our posterior through the likelihood!

CONJUGACY

- Certain likelihood functions give rise to particularly nice posterior distributions.
  - Normal prior, Normal likelihood ⇒ Normal posterior.
  - Beta prior, Binomial likelihood ⇒ Beta posterior.
  - Gamma prior, Poisson likelihood ⇒ Gamma posterior.

- This is called **conjugacy**.
  - Prior and posterior follow the same parametric distribution.

CONJUGACY

- Conjugacy used to be a very important concept in statistics. Why?

CONJUGACY

- This requires a working knowledge of common statistical distributions, your data-generating process, and your subject area.
    - "Think Bayes!" walks through these well.

WHAT HAPPENS WITHOUT CONJUGACY?

- Suppose I want to conduct inference on some parameter θ.
  - Before observing data, I **really** believe that θ follows a Wishart distribution.
  - I **really** believe that my data generating process $y|\theta$ follows a Cauchy distribution.

WHAT HAPPENS WITHOUT CONJUGACY?

- Suppose I want to conduct inference on some parameter θ.
  - Before observing data, I **really** believe that θ follows a Wishart distribution.
  - I **really** believe that my data generating process $y|\theta$ follows a Cauchy distribution.

- Strategy 1: Instead of picking Wishart/Cauchy distributions, I pick distributions that might reflect the real world less in order for my prior and likelihood to "play nicely" together.

WHAT HAPPENS WITHOUT CONJUGACY?

- Suppose I want to conduct inference on some parameter θ.
  - Before observing data, I **really** believe that θ follows a Wishart distribution.
  - I **really** believe that my data generating process $y|\theta$ follows a Cauchy distribution.

- Strategy 2: Monte Carlo simulations!

CALCULATING THE POSTERIOR

SIMULATING THE POSTERIOR

$$f(\theta|y) \propto f(y|\theta) \times f(\theta)$$

1. Specify $f(y|\theta)$ and $f(\theta)$.
2. Simulate one value from $f(\theta)$, called θ'.
3. Using the value θ', find and plot the height of $f(y|\theta')$.
4. Repeat this large number of times.

SIMULATING THE POSTERIOR

$$f(\theta|y) \propto f(y|\theta) \times f(\theta)$$

- Once we've simulated the posterior distribution, we can do whatever we want to do with it.
  - Estimate the average value of θ.
  - Estimate the median value of θ.
  - Estimate the range of the middle 95% values of θ.

BONUS SECTION

REFERENCE: STATISTICAL DISTRIBUTIONS

Distribution	Support	Continuous vs. Discrete	Common Use Case
Normal			
Exponential			
Gamma			
Beta			
Binomial			
Poisson			
Negative Binomial			

UPDATING INFORMATION

- Prior: $f(\theta) \Rightarrow$ Posterior: $f(\theta|y_1)$
- Prior: $f(\theta|y_1) \Rightarrow$ Posterior: $f(\theta|y_1, y_2)$
- Prior: $f(\theta|y_1, y_2) \Rightarrow$ Posterior: $f(\theta|y_1, y_2, y_3)$

EXAMPLE

- "Disentangling Bias and Variance in Election Polls"

$$y_i \sim N\left(\nu_{r[i]} + \alpha_{r[i]} + t_i\beta_{r[i]}, \sqrt{\frac{\nu_{r[i]}\left(1-\nu_{r[i]}\right)}{n_i} + \tau_{r[i]}}\right)$$

- $y_i =$ outcome of poll i
- $\nu_{r[i]} =$ final two-party vote share for Republican candidate
- $\alpha_{r[i]} + t_i\beta_{r[i]} =$ bias of ith poll with t in months
- $\sqrt{\frac{\nu_{r[i]}\left(1-\nu_{r[i]}\right)}{n_i}} =$ standard error of $\nu_{r[i]}$ under SRS
- $\tau_{r[i]} =$ election-specific variance

EXAMPLE

- "Disentangling Bias and Variance in Election Polls"

$$y_i \sim N\left(\nu_{r[i]} + \alpha_{r[i]} + t_i\beta_{r[i]}, \sqrt{\frac{\nu_{r[i]}\left(1-\nu_{r[i]}\right)}{n_i}} + \tau_{r[i]}\right)$$

- $\alpha_r \sim N(\mu_\alpha, \sigma_\alpha); \mu_\alpha \sim N(0,0.05); \sigma_\alpha \sim N_+(0,0.05)$
- $\beta_r \sim N\left(\mu_\beta, \sigma_\beta\right); \mu_\beta \sim N(0,0.05); \sigma_\beta \sim N_+(0,0.05)$
- $\tau_r \sim N_+(0, \sigma_\tau); \sigma_\tau \sim N_+(0, 0.02)$

- Think of these **hyperparameters** as **tuning parameters**.

SO ***HOW*** DO WE DO BAYESIAN STATISTICS?

- Goal: Find posterior distribution of parameter θ given our evidence y.
  - This is written as $f(\theta|y)$.

- Needed:
  - Prior distribution for parameter θ.
  - Likelihood of data y given parameter θ.
  - Marginal likelihood of data y with no knowledge of parameter.*