Practical Applications of Bayesian Statistics

for Data Scientists in Business

Matt DiNauta
Principal Applied Scientist
Zillow Group

About Me

- Principal Applied Scientist at Zillow Group
- Work on models and data products supporting our internal teams: finance, business operations, product, etc.
- Run a Bayesian statistics interest group for data practitioners across Zillow Group

Intended Audience

- Data scientists new to Bayesian Statistics
- Data scientists of the “analyst” variety: those whose work involves producing insights from data

“I am a busy data scientist. I’m aware of Bayesian statistics but it seems very theoretical and it’s not clear to be why should I add this to my toolbox over other the many other things I could spend my time learning.”

Goals

- Introduce common terms and notation so you are well prepared for further reading
- Recognize the next time you run into a data science problem that is particularly well suited for applying Bayesian methods

Core Concepts

Bayes' Theorem

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

A, B = events

$P(A|B)$ = probability of A given B is true

$P(B|A)$ = probability of B given A is true

$P(A), P(B)$ = the independent probabilities of A and B

Bayes Rule Annotated

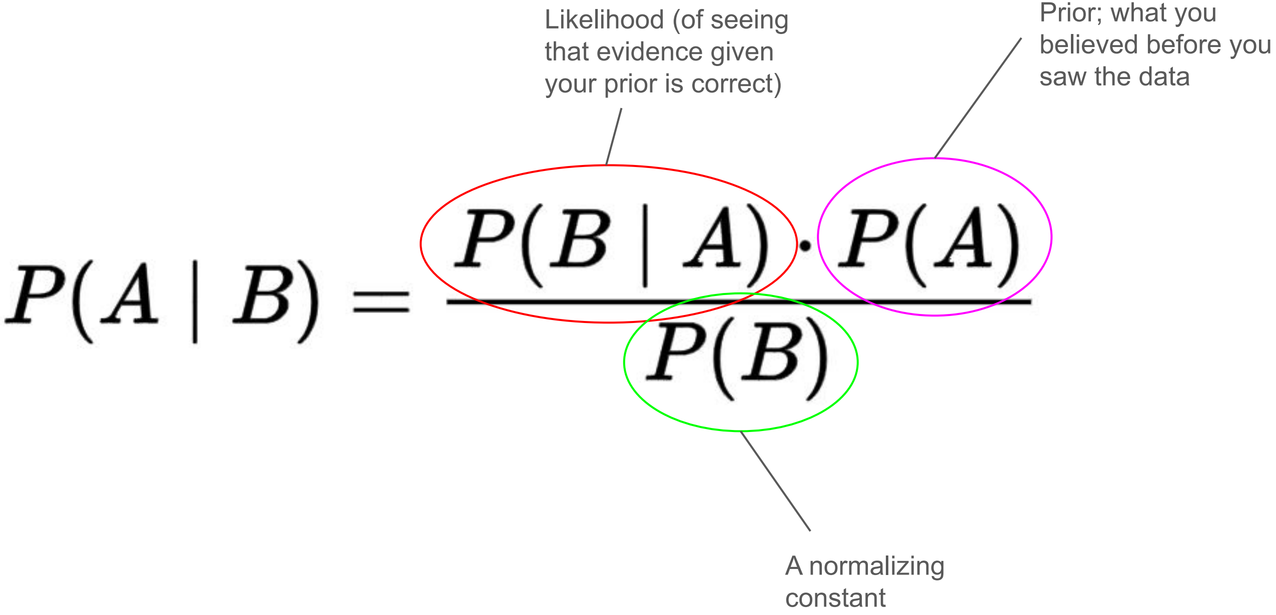
Prior; what you
believed before you
saw the data

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Bayes Rule Annotated

$$P(A \mid B) = \frac{\overbrace{P(B \mid A)}^{\text{Likelihood (of seeing that evidence given your prior is correct)}} \cdot \overbrace{P(A)}^{\text{Prior; what you believed before you saw the data}}}{P(B)}$$

Bayes Rule Annotated



$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Likelihood (of seeing that evidence given your prior is correct)

Prior; what you believed before you saw the data

A normalizing constant

Bayes Rule Annotated

Likelihood (of seeing
that evidence given
your prior is correct)

Prior; what you
believed before you
saw the data

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Posterior

A normalizing
constant

Canonical example of applying Bayes rule

An individual tests positive for a rare disease. What is the probability they actually have the disease?

- Disease affects 1 in every 10,000 people in the population (P(D))
- Test has a 1% false positive rate (P(T|~D)); 1% false negative rate (P(~T|D))
- P(D|T)?

Canonical example of applying Bayes rule

An individual tests positive for a rare disease. What is the probability they actually have the disease?

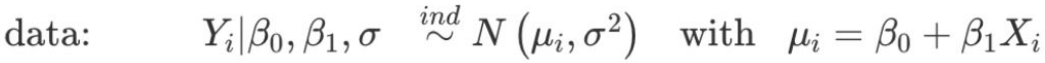
- Disease affects 1 in every 10,000 people in the population ($P(D)$)
- Test has a 1% false positive rate ($P(T|{\sim}D)$); 1% false negative rate ($P({\sim}T|D)$)
- $P(D|T)$?

- Bayes rule: $P(D|T) = [P(T|D) * P(D)] / [P(T|D) * P(D) + P(T|{\sim}D) * P({\sim}D)]$
  - $P(D|T) = [0.99 * 0.0001] / [0.99 * 0.0001 + 0.01 * 0.9999]$
- $P(D|T) = {\sim}1\%$

Introducing some notation

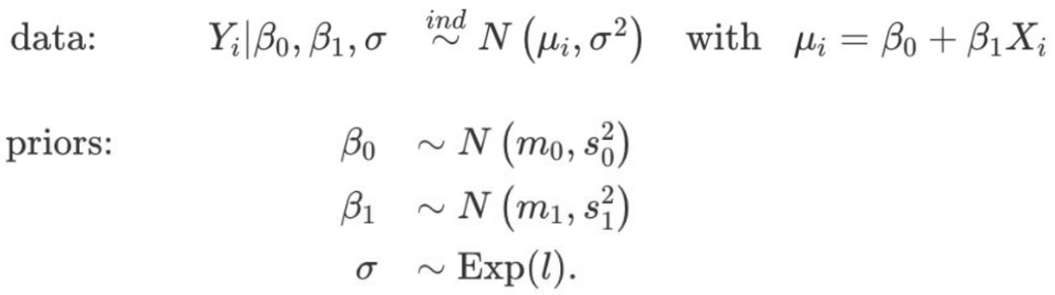
- Regression equation we are all familiar with: $Y = \beta_0 + X_1\beta_1 + \epsilon$
- Bayesian approach specifies priors for β
- These priors are *probability distributions*

Introducing some notation



data: $$Y_i|\beta_0,\beta_1,\sigma \overset{ind}{\sim} N\left(\mu_i,\sigma^2\right) \quad \text{with} \quad \mu_i = \beta_0 + \beta_1 X_i$$

Introducing some notation

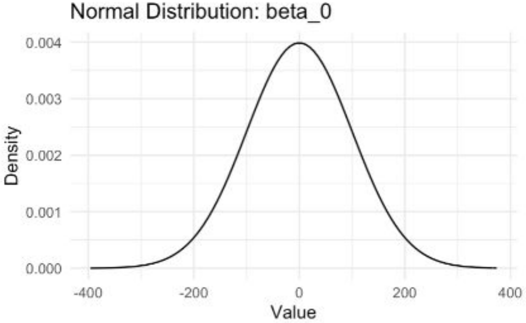


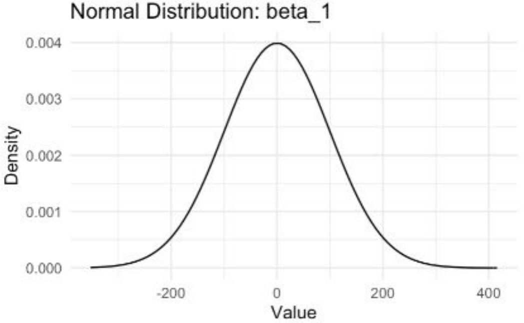
$$
\begin{array}{lll}
\text{data:} & Y_i|\beta_0,\beta_1,\sigma & \overset{ind}{\sim} N\left(\mu_i,\sigma^2\right) \quad \text{with} \quad \mu_i = \beta_0 + \beta_1 X_i \\
\text{priors:} & \beta_0 & \sim N\left(m_0, s_0^2\right) \\
 & \beta_1 & \sim N\left(m_1, s_1^2\right) \\
 & \sigma & \sim \text{Exp}(l).
\end{array}
$$

Introducing some notation

$$\text{data:} \qquad Y_i|\beta_0,\beta_1,\sigma \quad \stackrel{ind}{\sim} N\left(\mu_i,\sigma^2\right) \quad \text{with} \quad \mu_i = \beta_0 + \beta_1 X_i$$

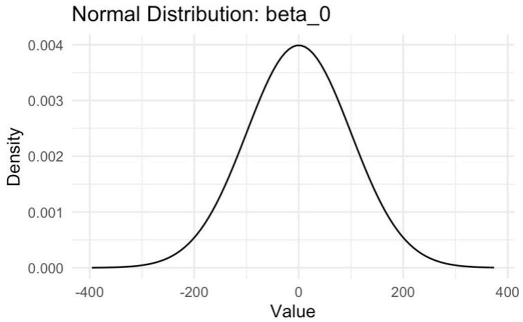
$$\begin{aligned} \text{priors:} \qquad \beta_0 & \sim N\left(m_0, s_0^2\right) \\ \beta_1 & \sim N\left(m_1, s_1^2\right) \\ \sigma & \sim \text{Exp}(l). \end{aligned}$$

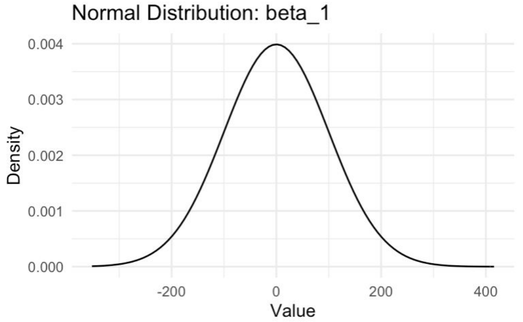


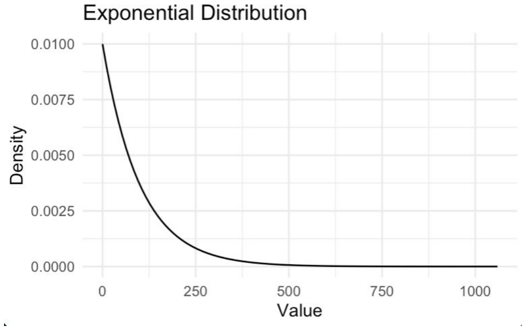


Introducing some notation

$$\begin{array}{lrl} \text{data:} & Y_i|\beta_0,\beta_1,\sigma & \overset{ind}{\sim} N\left(\mu_i,\sigma^2\right) \quad \text{with} \quad \mu_i = \beta_0 + \beta_1 X_i \\ \text{priors:} & \beta_0 & \sim N\left(m_0, s_0^2\right) \\ & \beta_1 & \sim N\left(m_1, s_1^2\right) \\ & \sigma & \sim \text{Exp}(l). \end{array}$$


Normal Distribution: beta_0


Normal Distribution: beta_1


Exponential Distribution

Practical Examples

Comparing many proportions

Imagine we are analyzing…

- Sports statistics: baseball batting averages, basketball free-throw percentages
- Real estate agents: conversion rate (deals closed / leads)

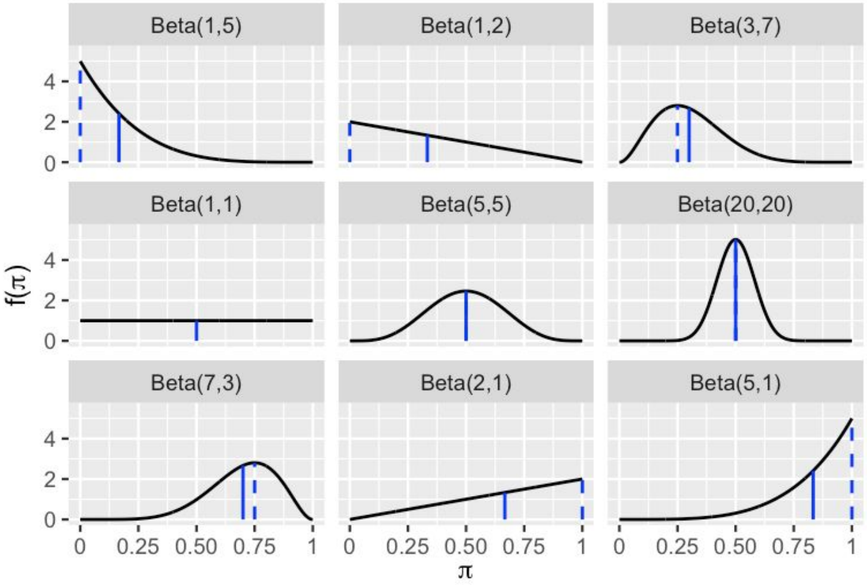
We need to somehow account for the varying number of observations in each dimension.

About the Beta distribution

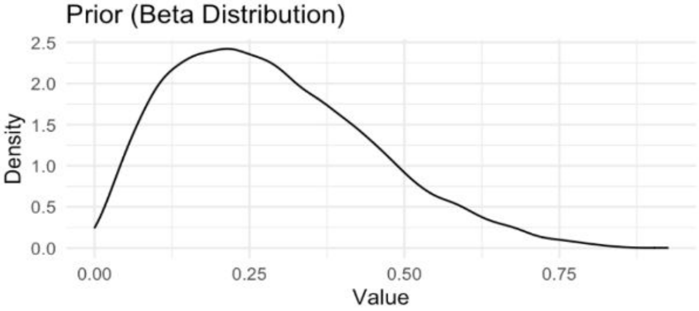
- Often used for proportions
- Described by two parameters, α and β

$$E(\theta) = \frac{\alpha}{\alpha+\beta}$$

- α is the number of successes
- α + β is the number of trials



The beta-binomial model

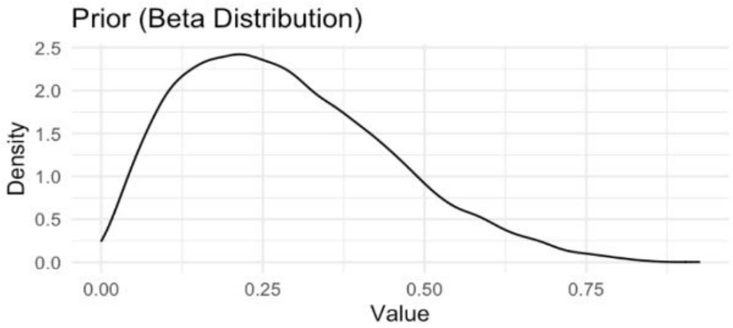


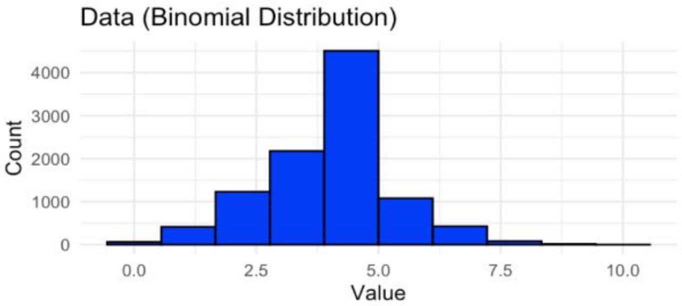
$$\pi \sim Beta(\alpha, \beta)$$

The beta-binomial model

$$\pi \sim Beta(\alpha, \beta)$$

$$Y \sim Binomial(n, \pi)$$


Prior (Beta Distribution)

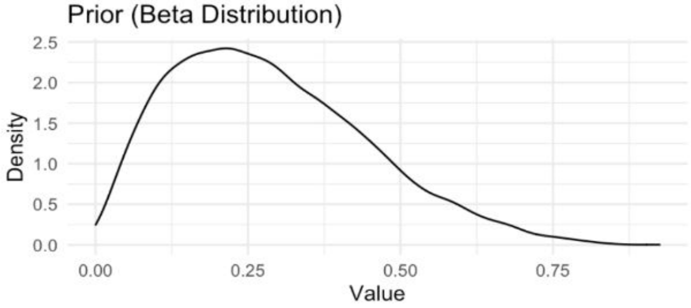

Data (Binomial Distribution)

The beta-binomial model

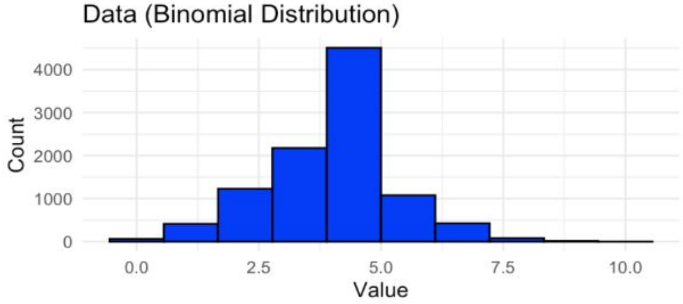
$$\pi \sim Beta(\alpha, \beta)$$

$$Y \sim Binomial(n, \pi)$$

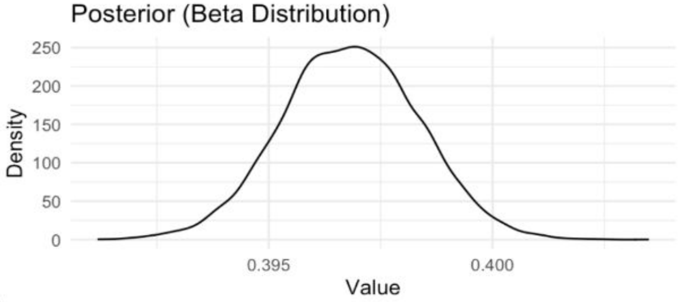
$$\pi | Y \sim Beta(\alpha + Y, \beta + n - Y)$$



Prior (Beta Distribution)



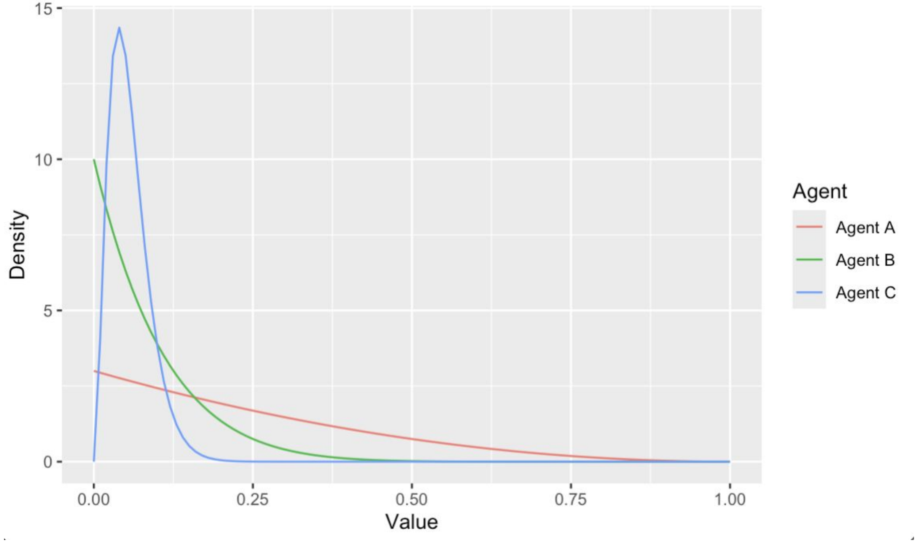
Data (Binomial Distribution)



Posterior (Beta Distribution)

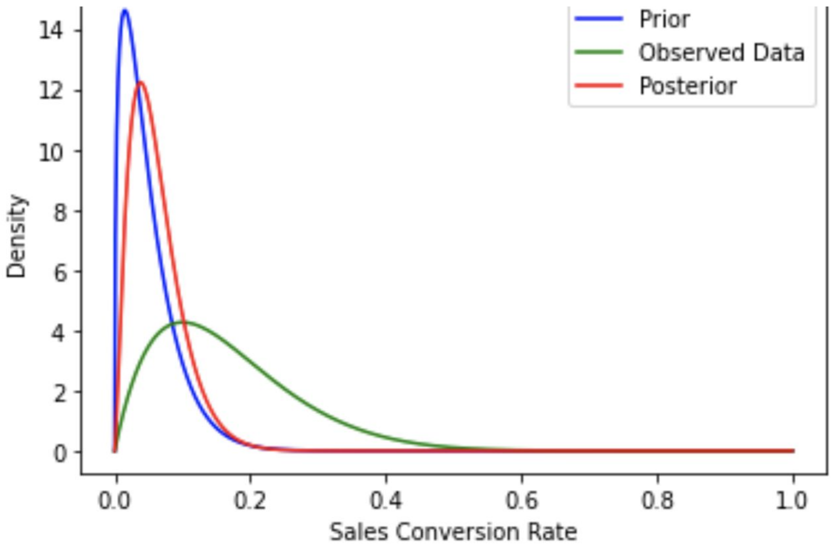
Ranking performance

- Agent A: 1 sale from 3 leads (30%)
- Agent B: 1 sales from 10 leads (10%)
- Agent C: 3 sales from 50 leads (6%)



Add a prior

- Prior = mean 5% and sd 4%.
- Beta(α ≅1.4, β ≅27.3)



Rank our sales people again

- Agent 1: (1 + 1.4) / (3 + 1.4 + 27.3) = 7.6%
- Agent 2: (1 + 1.4) / (10 + 1.4 + 27.3) = 6.2%
- Agent 3: (3 + 1.4) / (50 + 1.4 + 27.3) = 5.6%

Intuition:

1. “We are ‘starting off’ all agents with 1.4 sales and 28.7 attempts, to reflect that the average agent has a 5% CVR.”
2. “As we collect more data on their performance, the influence of the prior gets smaller.”

Extending to hierarchical models

- Rather than a single prior, real estate agents in a particular geographic area probably look more other agents in that geographic area than they do agents from other geographic regions
- We want a different prior for each geographic region. The geographic regions themselves may be considered as belonging to another, global distribution of regions.

This type of situation is the motivation for hierarchical models

The Key Idea: Partial Pooling

- Complete pooling - ignore groups
- No pooling - analyze each group separately
- Partial pooling - middle ground. All groups are connected and thus might contain valuable information about one another.

A Hierarchical Model Example

- Example from Bayes Rules! An Introduction to Applied Bayesian Modeling
- Spotify: modeling the popularity of songs
  - What’s the typical popularity of a Spotify song?
  - To what extent does popularity vary **from artist to artist**?
  - **For any single artist**, how much might popularity vary from song to song?

A Hierarchical Model Example

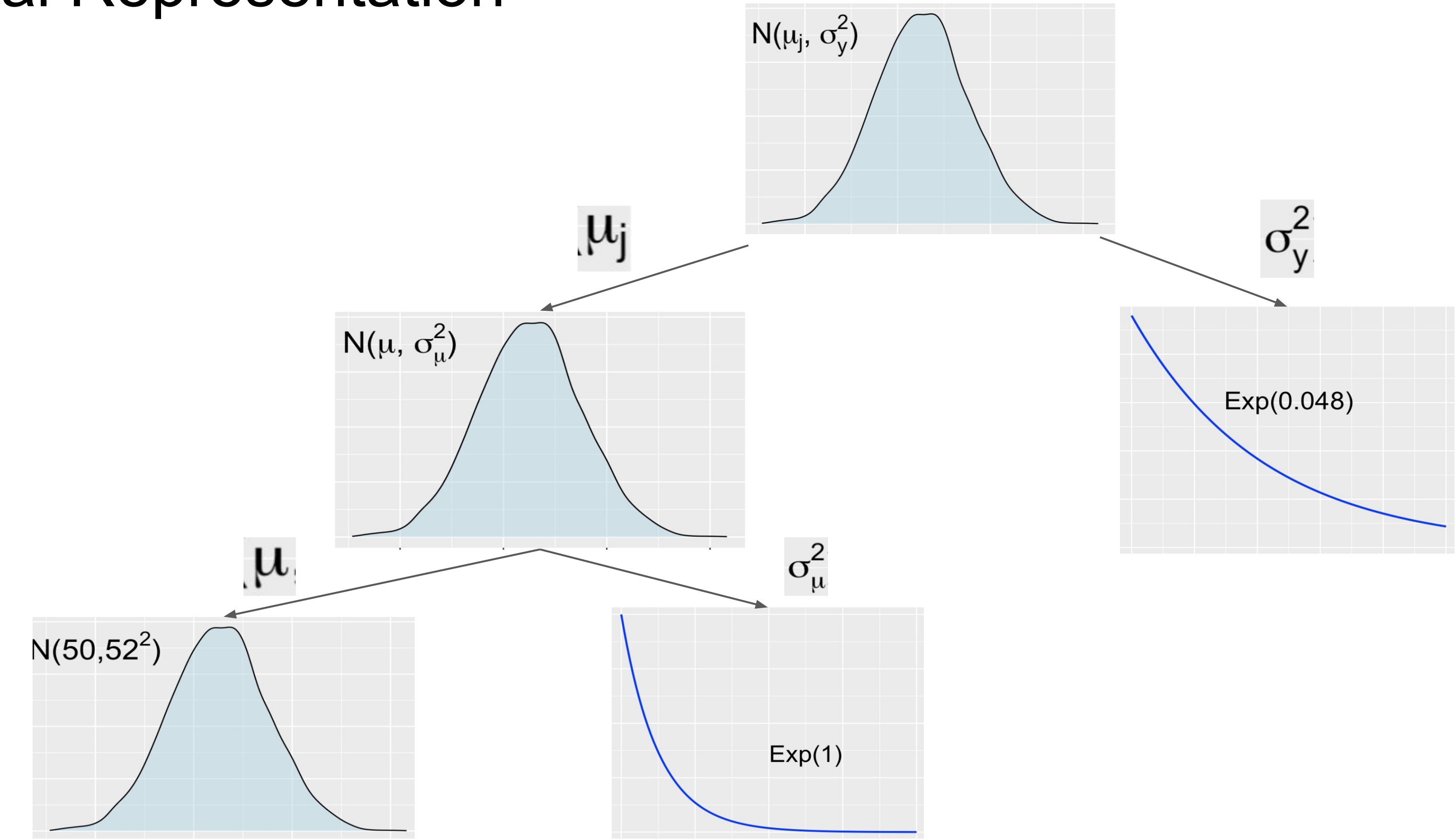
- Example from Bayes Rules! An Introduction to Applied Bayesian Modeling
- Spotify: modeling the popularity of songs
  - What's the typical popularity of a Spotify song?
  - To what extent does popularity vary from artist to artist?
  - For any single artist, how much might popularity vary from song to song?

- Complete pooling: ignore artists and lump all songs together
- No pooling: separately analyze each artist
- Partial pooling: even though artists differ in popularity, they might share information about each other

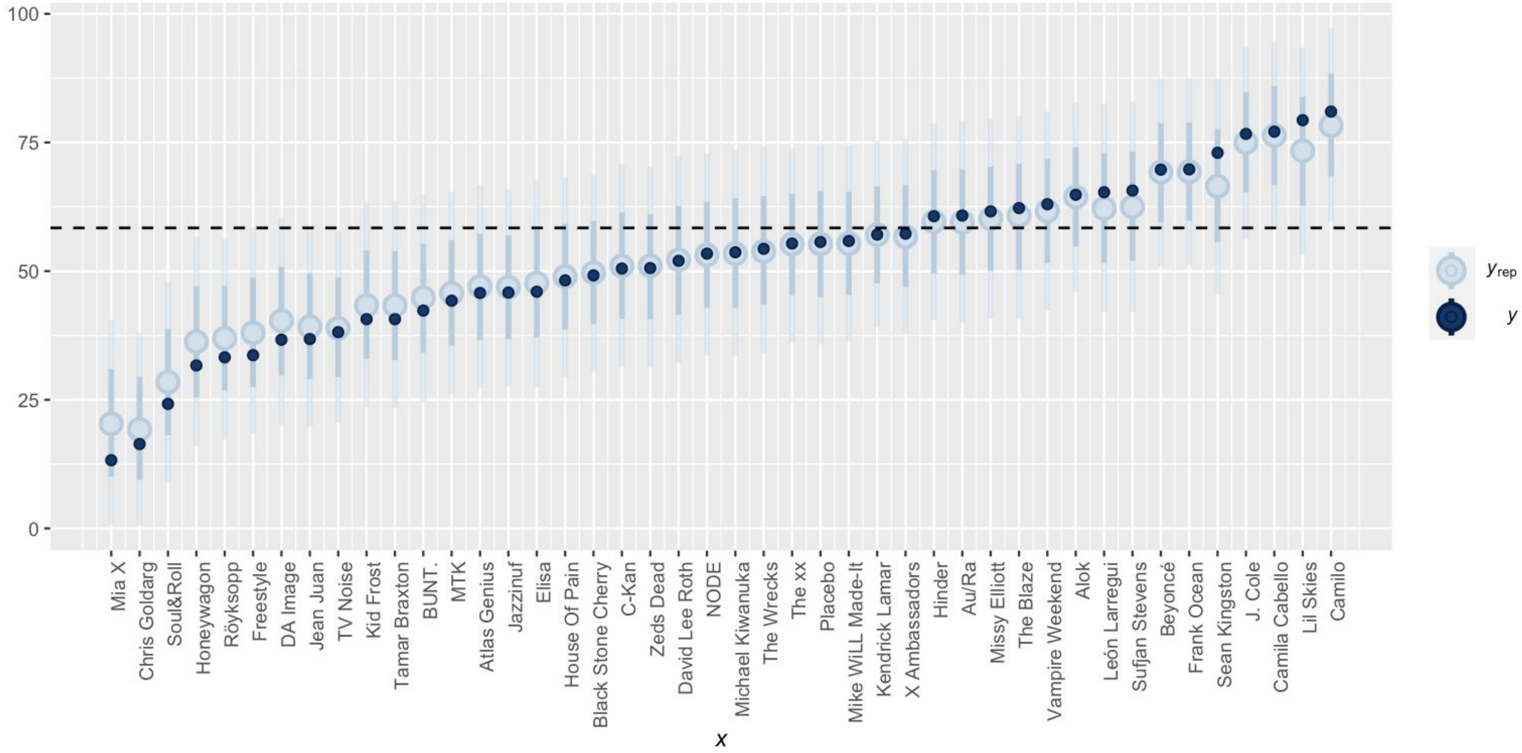
Hierarchical Model Specification

$$\begin{array}{lrll}
\text{Layer 1:} & Y_{ij}|\mu_j,\sigma_y & \sim N(\mu_j,\sigma_y^2) & \text{model of individual songs within artist } j \\
\text{Layer 2:} & \mu_j|\mu,\sigma_\mu & \overset{ind}{\sim} N(\mu,\sigma_\mu^2) & \text{model of variability between artists} \\
\text{Layer 3:} & \left\{\begin{array}{r} \mu \\ \sigma_y \\ \sigma_\mu \end{array}\right. & \begin{array}{l} \sim N(50,52^2) \\ \sim \text{Exp}(0.048) \\ \sim \text{Exp}(1) \end{array} & \begin{array}{l} \text{prior models on global parameters} \\ \\ \\ \end{array}
\end{array}$$

Visual Representation



Estimates “shrunk” to global mean



Recap

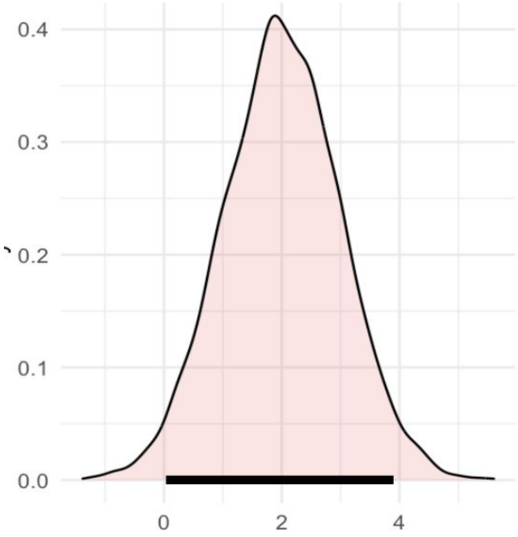
- Priors are useful when we have small amounts of data
- A situation where they can be very useful is when we have a lot of information globally, but little information in the dimensions we are interested in
- Take this idea further by exploiting hierarchical structure

Last Step: Making Comparisons

- We have a bunch of posterior distributions. How do we compare them?
- Frequentist approach will use p-values, confidence intervals
- Bayesian approach is arguably more intuitive

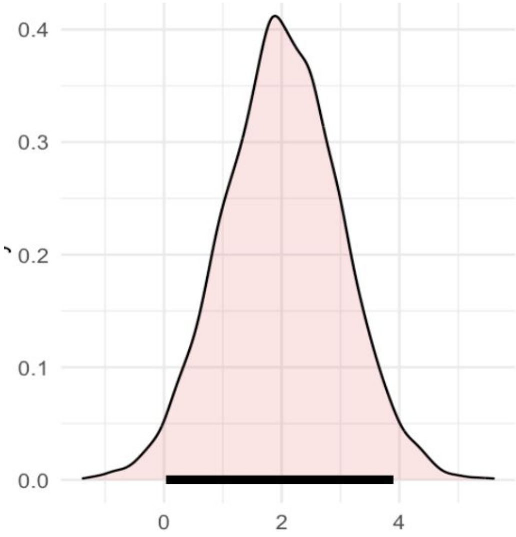
Last Step: Making Comparisons

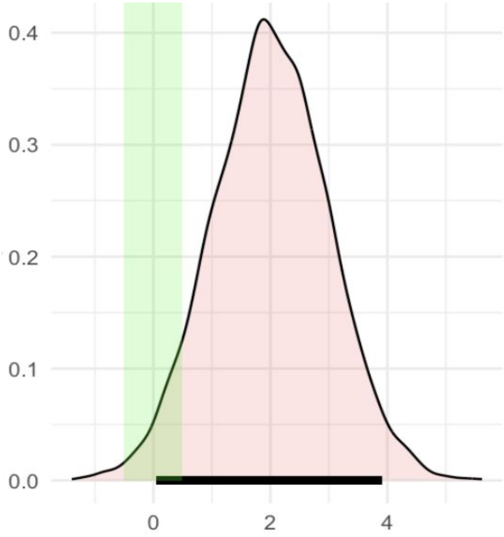
- HDI → highest density interval. Smallest possible interval that covers n% (e.g. 95%) of the probability density



Last Step: Making Comparisons

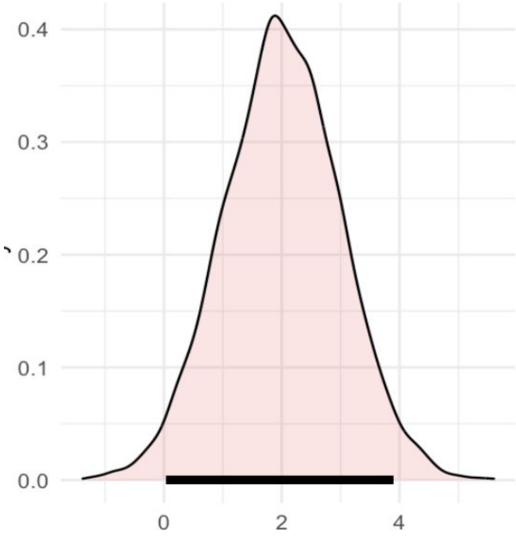
- HDI → highest density interval. Smallest possible interval that covers n% (e.g. 95%) of the probability density
- ROPE → “region of practical equivalence”

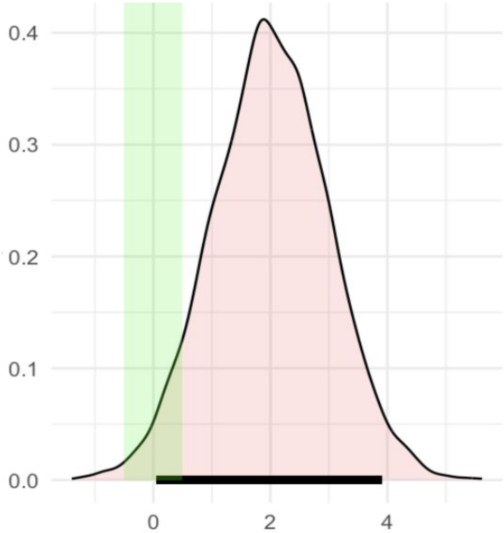


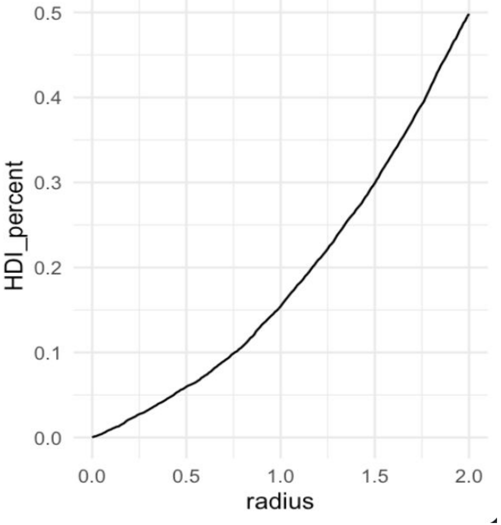


Last Step: Making Comparisons

- HDI → highest density interval. Smallest possible interval that covers n% (e.g. 95%) of the probability density
- ROPE → “region of practical equivalence”







Goals: Revisited

- Introduce common terms and notation so you are well prepared for further reading
  - Bayes' theorem: prior, likelihood, and posterior
  - The beta-binomial model
  - Hierarchical models: partial pooling

- Recognize the next time you run into a problem that is particular well suited for applying Bayesian methods
  - You have small data, but some prior beliefs
  - Hierarchical structure to data
    - Real estate agents within geographic regions
    - Songs within artists (within genres?)
  - Quantifying cost/benefit is important - the *direct probabilistic* interpretations we get from Bayesian approach lends itself well to this

Appendix

Recommended Further Reading

- Bayes Rules!
  - Freely available online
  - Succinct, lots of practical examples

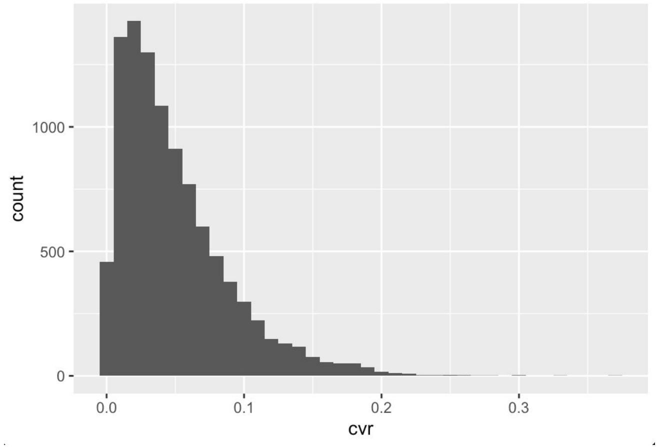
- Doing Bayesian Statistics
  - Lots of direct comparisons between the frequentist way of doing things vs. Bayesian
  - E.g. “what is the Bayesian analogue for a t test”?

- Rethinking Statistics
  - More of a complete introduction
  - Ties together Bayesian statistics and causal inference

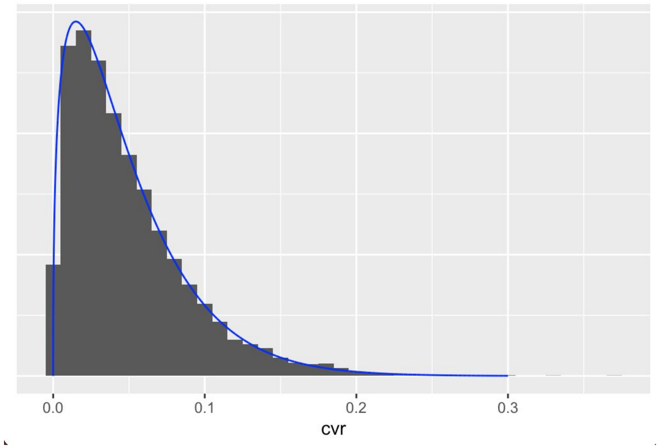
How to determine a prior?

- Priors can come from previous research, domain experts, or the data itself (empirical Bayes)
  - I find that the empirical Bayes approach is particularly useful for practical data science applications
- Consider an empirical distribution of salesperson-level CVRs (a)
- Fit a Beta distribution to the data - we get the two parameters of the Beta distribution that best fits (b)
- This Beta distribution is used as our prior for estimating the CVR of individual sales people

(a)



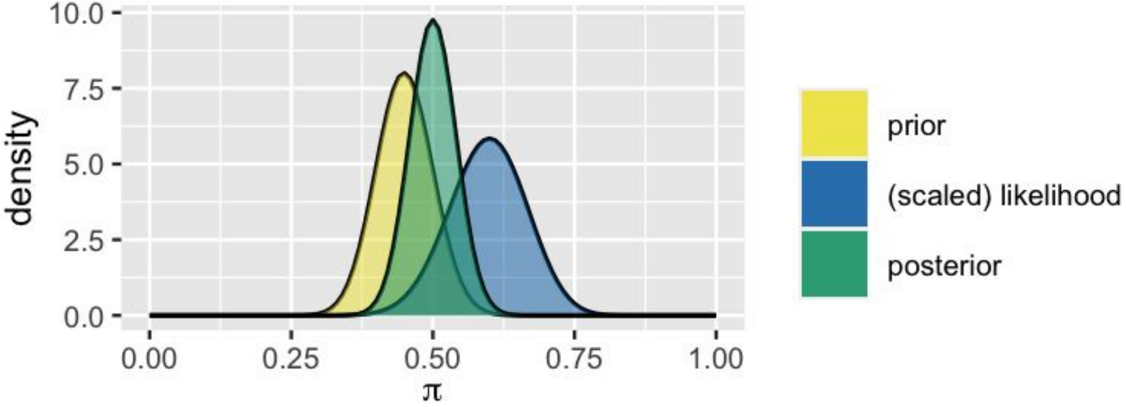
(b)



How do we actually fit these models

- Terms to know: Probabilistic programming, MCMC
- Probabilistic programming
    - Python: pymc
    - R: Stan, JAGS
- Declarative: you describe the model; the libraries fit
- Practical benefit: do not need to find, or write, an implementation for different methods. E.g. a library for time-series modeling, a library for hierarchical/mixed models. If you can describe it, pymc will (try to) fit it

“Shrinkage” Intuition


density
π
prior
(scaled) likelihood
posterior