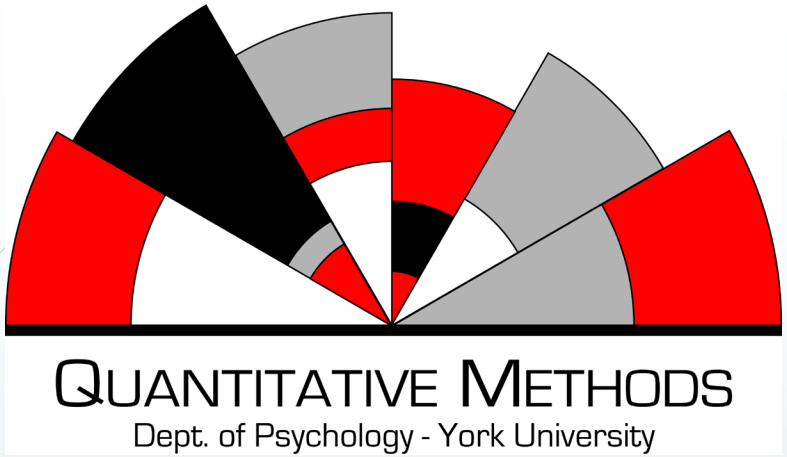
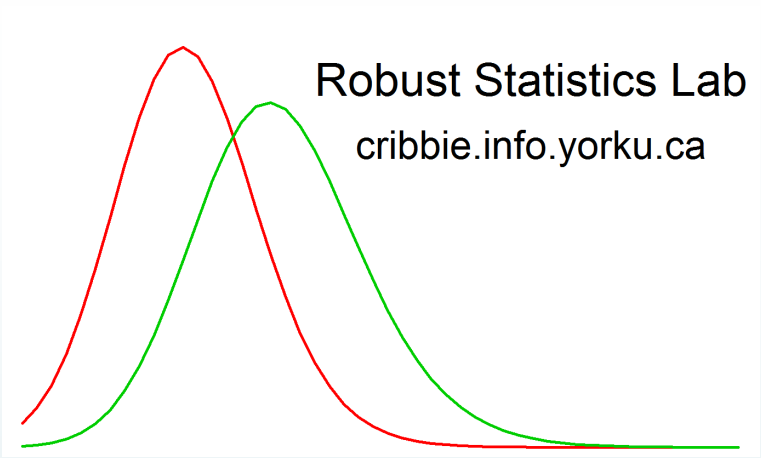




Bayesian Regression: Getting Started

Rob Cribbie (feat. SCS)

Why an *Intro to Bayesian Regression* at the QM Forum?

- I have been spewing about the advantages of Bayesian methods for years, but have had little experience applying the methods
- As a statistical consultant … I would like to feel confident recommending and conducting Bayesian analyses
- Grad students have limited exposure to Bayesian methods in classes
- Regression is the building block for many advanced methods (e.g., multilevel modeling, SEM)
- I would like to get feedback regarding my approach to the Bayesian regression analysis
- I had nothing better to talk about ☺

Outline

- Why Use Bayesian Statistics?
- Brief Introduction to Bayesian Inference and Model Fitting
- Analyzing a Simple Regression in R via Bayesian Methods

Inference/Estimation from a Traditional Frequentist Perspective

- *p*-Values
  - The probability of observing a test statistic as extreme, or more extreme, than that found, assuming the null hypothesis is true
  - Common interpretational mistake
    - The *p*-value is the probability that the null hypothesis is true
- 95% Confidence Intervals (CIs)
  - If we were to sample repeatedly from the population, and calculate a CI for each sample, 95% of the intervals would contain the population parameter
  - Common interpretational mistake
    - There is a 95% chance that the calculated CI contains the population parameter

Sample Results: Regression, Frequentist Perspective

with work stability being independently associated with a higher resilience attitude (9.3 point increase in resilience scores; 95% confidence interval: −17.-62–0.95; p = 0.039).

$$H_0\colon\ b^*_{ws} = 0$$

Embracing resilience in multiple sclerosis: a new perspective from COVID-19 pandemic

Elvira Sbragia, Eleonora Colombo, Chiara Pollio, Maria Cellerino, Caterina Lapucci, Matilde Inglese, Gianluigi Mancardi & Giacomo Boffa

What can the researcher conclude?

- *p*-Value
  - There is a 3.9% chance of observing a test statistic for the regression coefficient as extreme, or more extreme, than that found, assuming the partial relationship between work stability and resilience attitude is null in the population
  - *p*(data | hypothesis)
- Confidence Interval
  - If we were to sample repeatedly from the population, and calculate a CI for each sample, 95% of the intervals would contain the population partial regression coefficient
  - Says nothing about the probability of the parameter falling in a certain interval
- Wouldn't it be nice to be able to say something about the p(Null Hypothesis) or the probability that the regression coefficient falls within a given confidence interval?

Frequentist Model Fitting

- There is nothing more frustrating in model fitting than:

** ERROR ** model did NOT converge

Frequentist Model Fitting

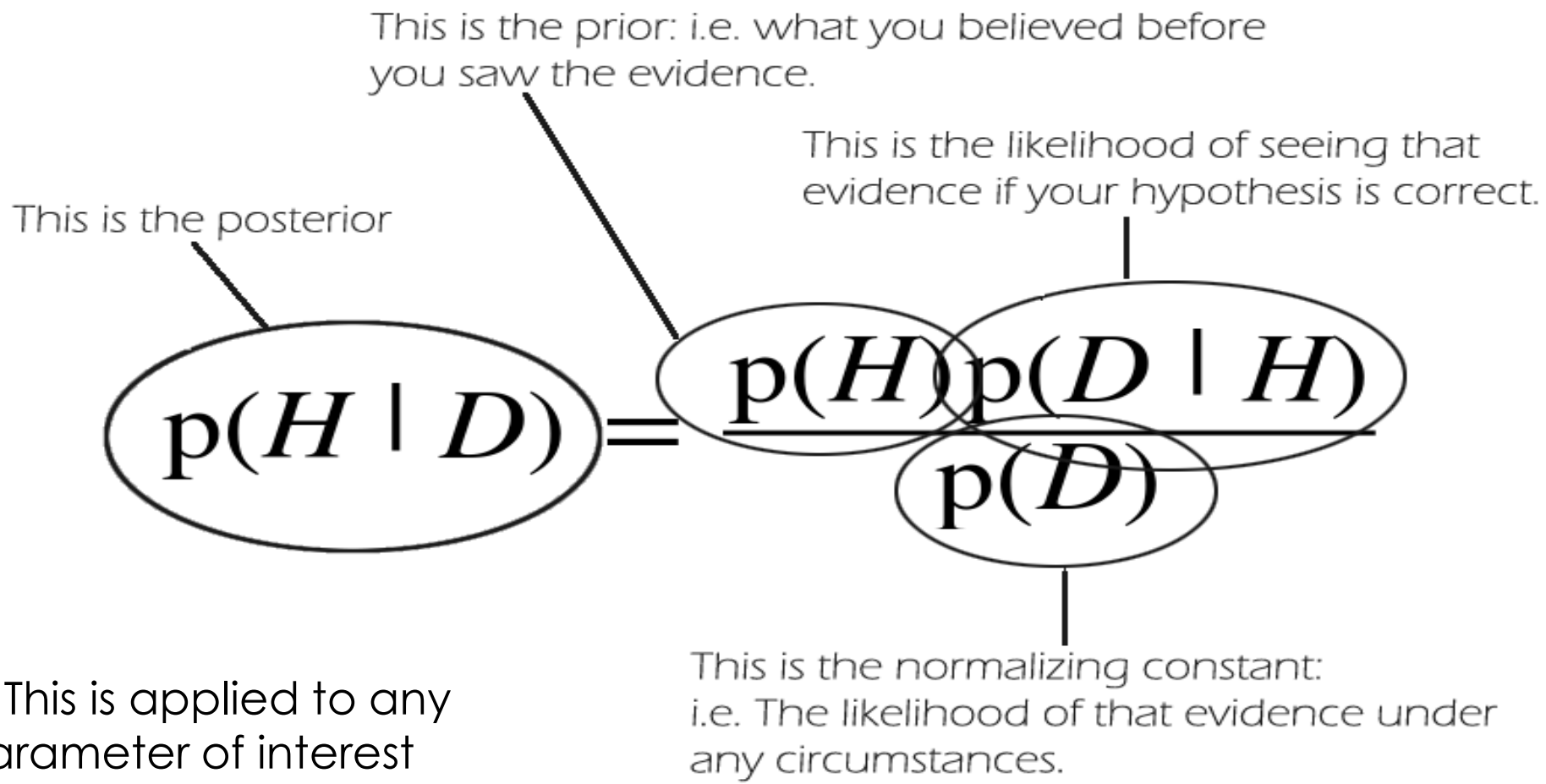
- Bayesian models are typically estimated using Markov Chain Monte Carlo (MCMC)
- Models that lead to convergence issues with ML, REML, etc. (e.g., many parameters, complex hierarchical structure) converge using MCMC.

Bayes Theorem

- $p(H|D) = \frac{p(D|H)p(H)}{p(D)}$
- $p(H|D) = \frac{p(D|H)p(H)}{\sum_1^N p(D|H)p(H)}$
  - N = number of possible hypotheses
- $p(H|D) = \frac{Likelihood * Prior}{Normalizing\ Constant}$

Main Point: We are solving for p(H|D), not P(D|H)!

Bayes Theorem



** This is applied to any parameter of interest

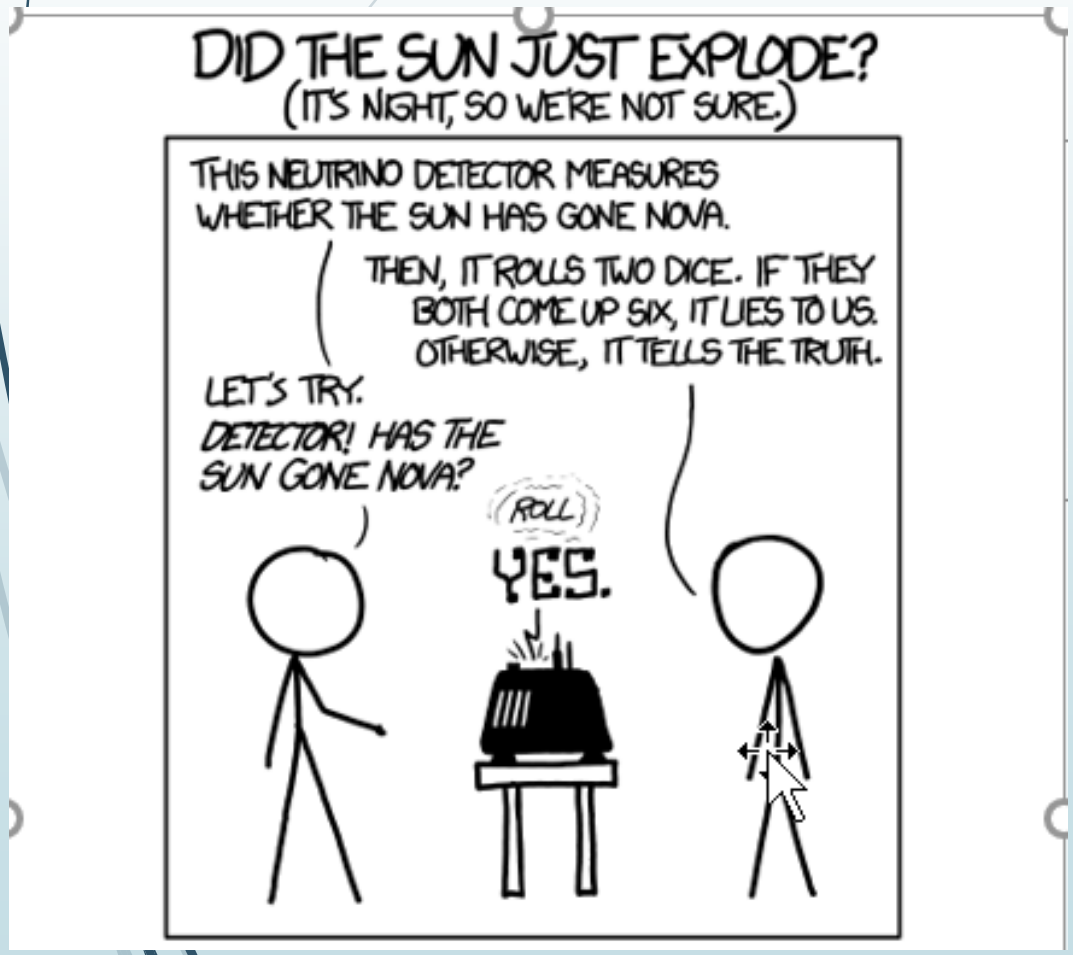
Definitions of Probability

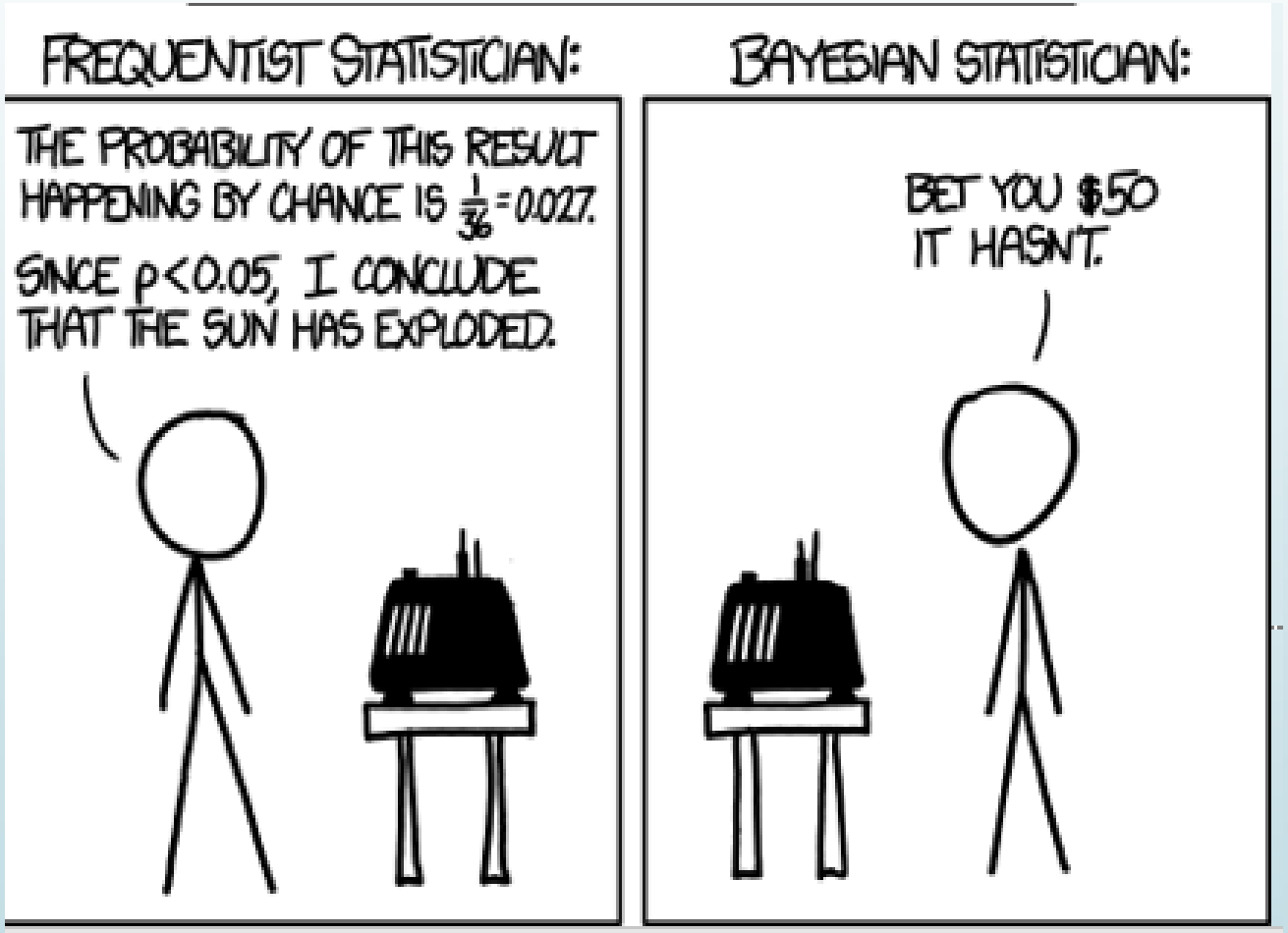
- Frequentist
  - Long run probability
  - E.g., probability of success in therapy (over many clients)
    - $p(S) = \frac{\#\ successes}{\#\ clients}$
  - The parameter [$p(S)$] is fixed, and we are using the data to try to detect it

Definitions of Probability

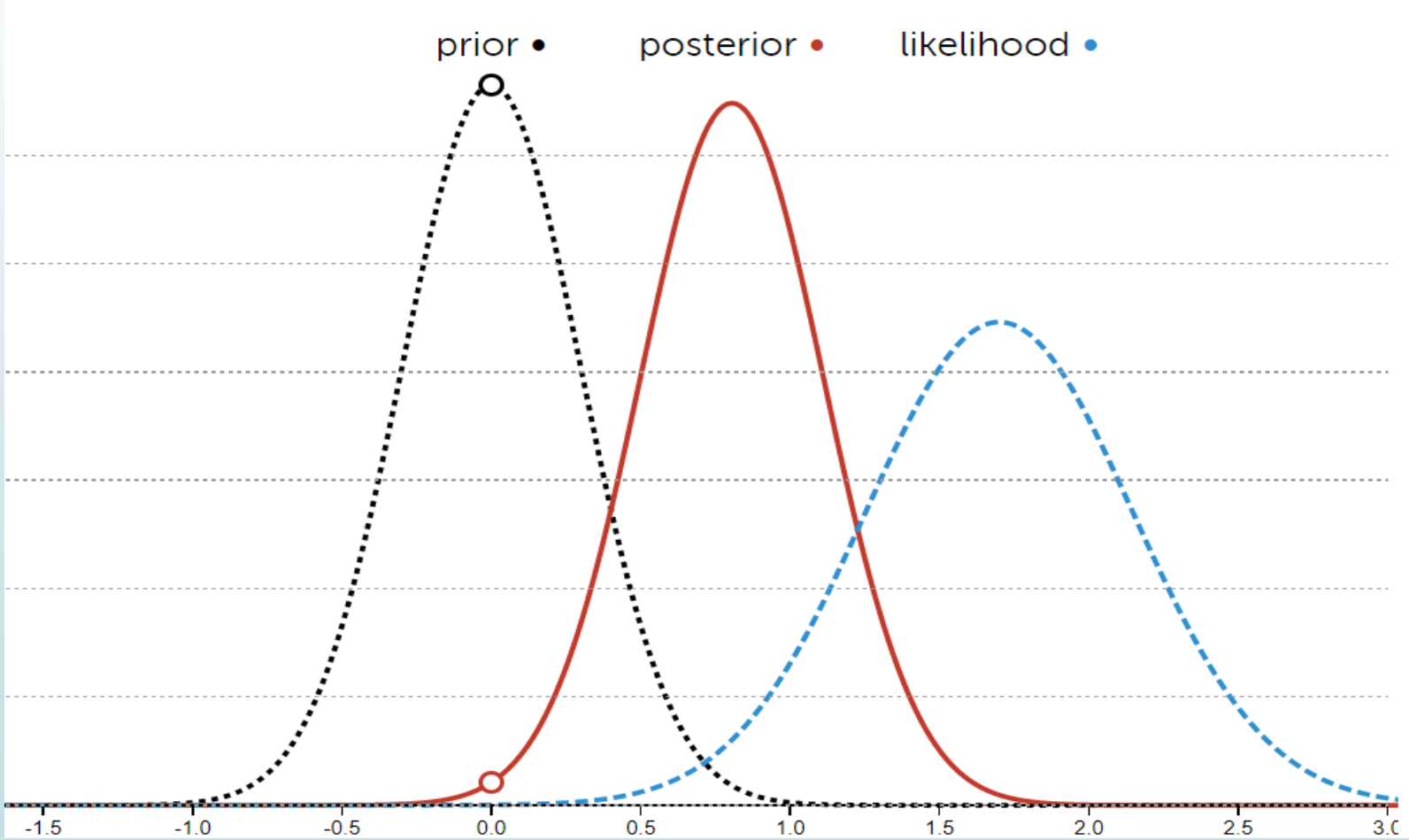
- Bayesian
  - Parameters are not fixed
    - E.g., there is no TRUE *p(S)*
  - Instead, parameters can vary
    - Our job is to use prior beliefs and the data to determine the relative probability of observing the different possibilities for p(S)
  - The data and prior are fixed, but the parameter(s) can take on different values; probability is a degree of belief

Frequentist vs Bayesian Example

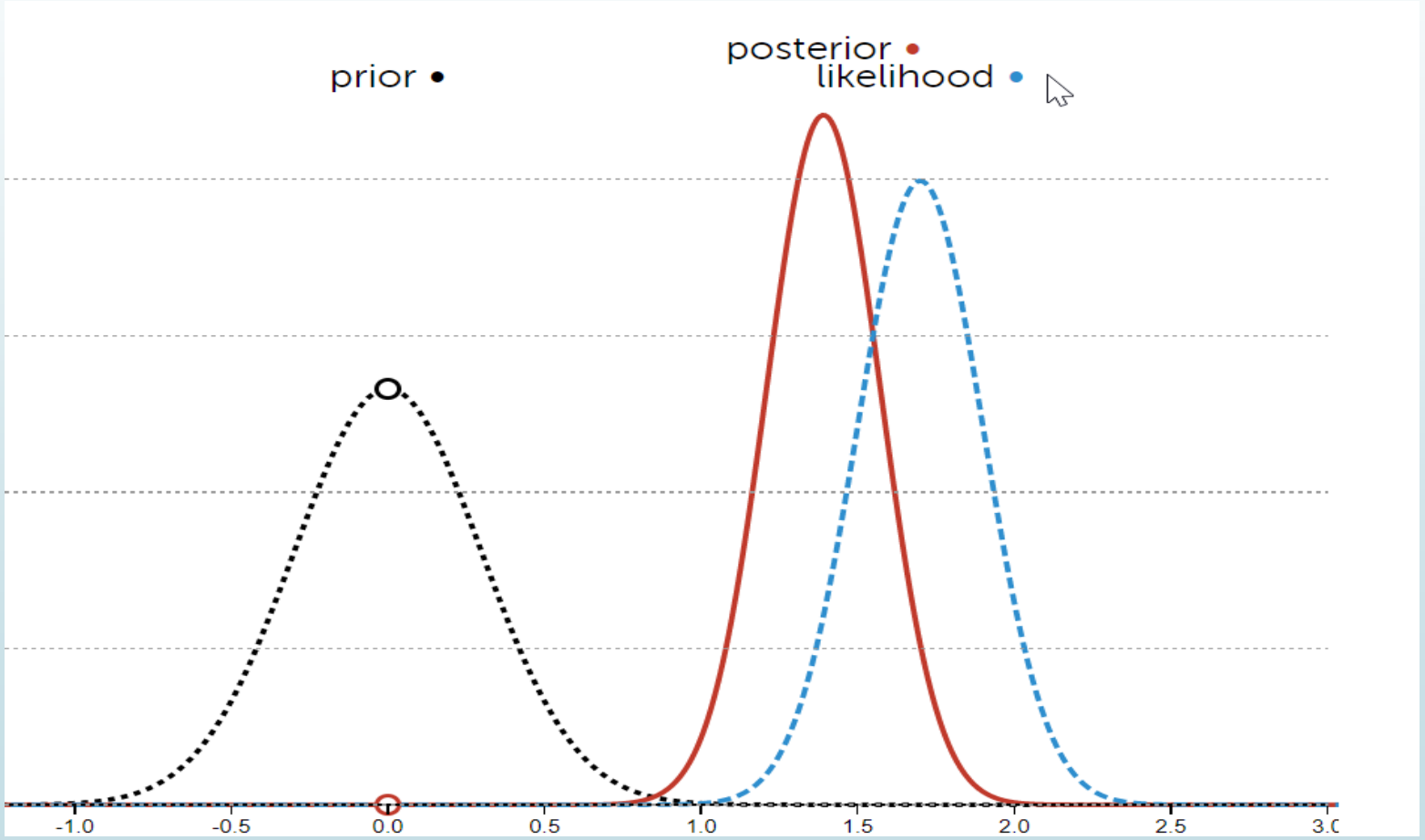




Prior, Likelihood and Posterior Distributions (small N)



Prior, Likelihood and Posterior Distributions (larger N)



Understanding the Posterior Distribution

- The posterior distribution is the outcome of interest in Bayesian analysis
  - The posterior distribution represents the probability of an event (e.g., *b*), after all evidence (data) and background information have been taken into account
  - We can think of the posterior distribution as an updating of the prior distribution
- The posterior can provide information regarding the probability of certain value for the parameter of interest
  - E.g., probability of a regression coefficient being greater than 0
  - E.g., probability of a regression coefficient > .4

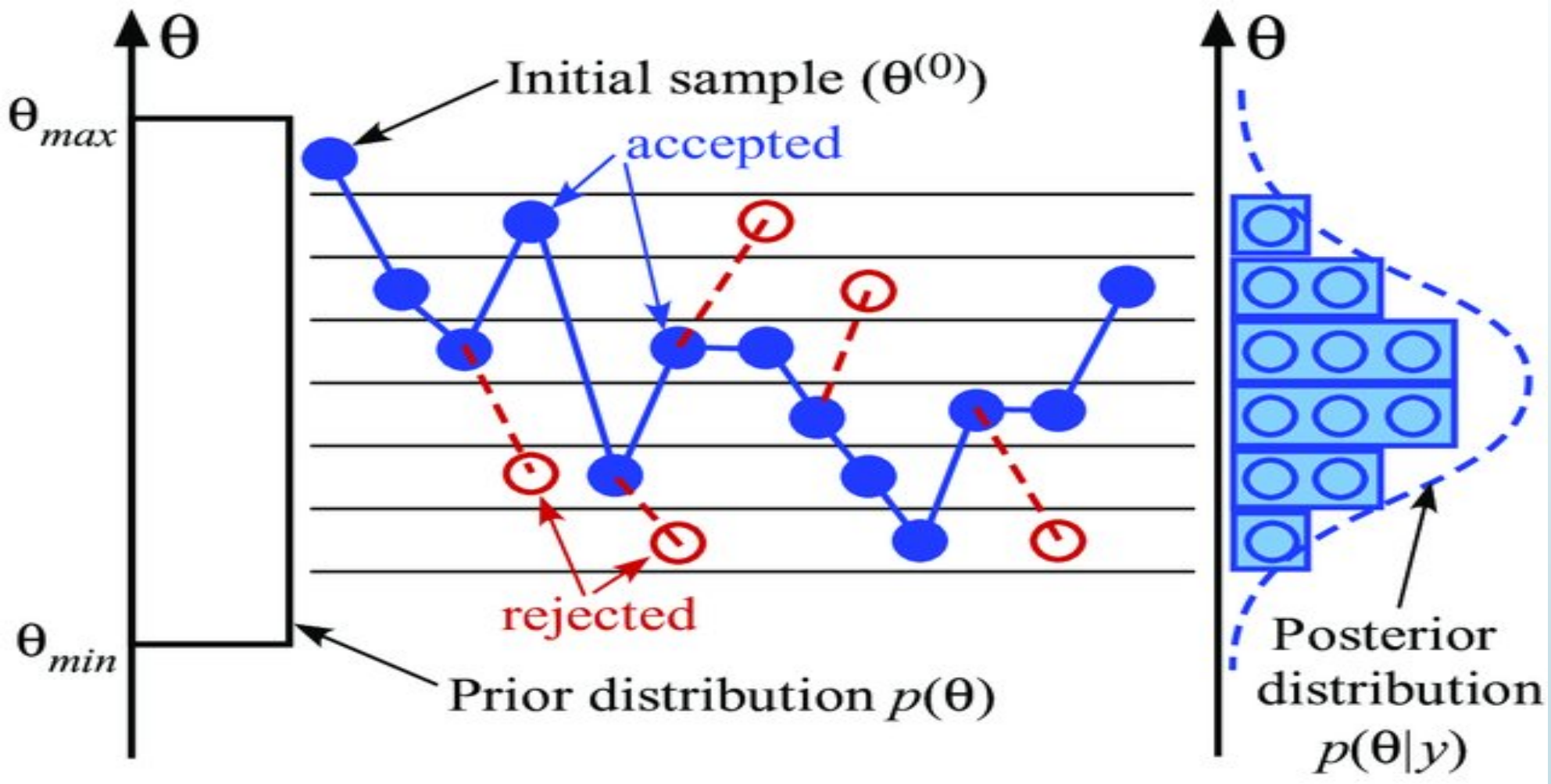
Markov Chain Monte Carlo (MCMC)

- An algorithm for sampling from probability distributions (*Monte Carlo*)
- Samples are drawn such that the (k + 1)th sample is dependent on the kth sample
  - This process is called a *Markov Chain*
- This allows the algorithm to narrow in on the quantity that is being approximated from the distribution (e.g., b), even with a large number of random variables

MCMC – Some Important Points

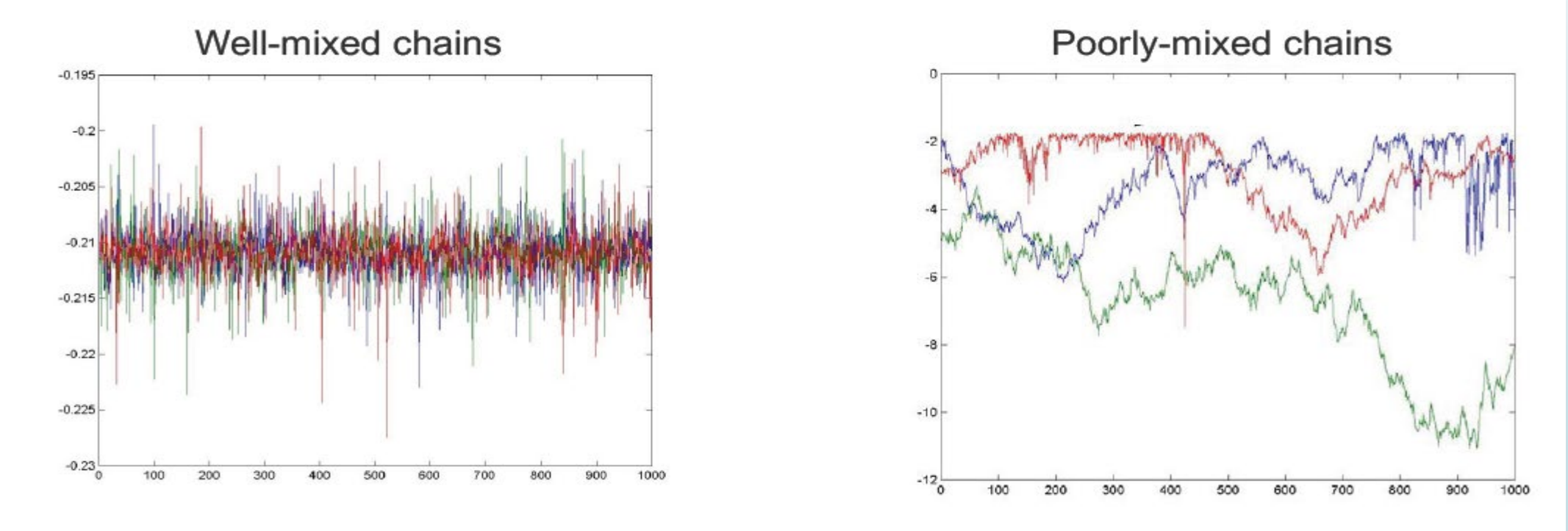
- The first step is drawing a 'potential parameter', after considering the prior and likelihood (proposal distribution)
  - This is our first "guess" of the parameter
- We then draw another 'potential parameter', related to the first, and we compare the parameters in terms of which explains the data better
- Whether we "accept" a new parameter depends on how it explains the data relative to the previous parameter
  - If it explains the data better, it is definitely kept
  - if it does not explain the data as well, the probability of keeping the proposed parameter depends on how much worse it explains the data
- Preliminary "guesses" are thrown out as they are unlikely to be reliable (burn-in)

MCMC Example



MCMC Diagnostics

- Mixing
  - Multiple chains are usually implemented (e.g., 3-5 chains)
  - It is expected that all chains will sample the same parameter space (i.e., mix)



MCMC Diagnostics

- Potential Scale Reduction Factor
  - Also known as the Gelman-Rubin Statistic or $\hat{R}$
    - $\hat{R} = \frac{Var_{Between\ Chain}}{Var_{Within\ Chain}}$
      - If convergence has been met, the between and within chain variance should be similar
  - $\hat{R}$ < 1.05 indicates convergence, whereas $\hat{R}$ > 1.05 indicates that more samples may be necessary for convergence
    - $\hat{R}$ > 1.25 can be indicative of convergence failure, such as chains heading towards local maxima

MCMC Diagnostics

- Effective Sample Size
  - Due to autocorrelation in the samples, the *effective* number of samples will be less than the total number of samples
    - The higher the better (e.g., >25% of the total number of samples)
- Monte Carlo Standard Error (per parameter)
  - The standard error of the mean of the posterior draws
  - MCSE should be small relative to the posterior standard deviation

MCMC Diagnostics

- Posterior Predictive Check
  - If a model is a good fit, then we should be able to use it to generate values for the outcome that are very similar to what we observed
    - In other words, we can use our observed values on the predictor(s) and the final model to generate a posterior predictive distribution
- Posterior Predictive Distribution Mean
  - If the mean of the posterior predictive distribution is not similar to the simple mean of the outcome variable, there may be convergence issues

Simple Regression

- $Y_i = b_0 + b_1 X_i + e_i$
  - b_0 is the intercept (value of Y when X = 0)
  - b_1 is the expected change in Y for a 1 unit increase in X
  - e_i = residual = $Y_i - (b_0 + b_1 X_i)$

- Parameters of Interest for Bayesian Analysis
  - b_0
  - b_1
  - σ_e (standard deviation of the e_i)

Conjugate Priors on the Parameters

- The normal distribution is a conjugate prior for the intercept and regression coefficient and the inverse gamma distribution is a conjugate prior for the residual variance (σ_e^2)
  - Conjugate priors result in posteriors with the same distribution as the prior

- $b_0 \sim N(\mu_{0\_b_0}, \tau_{0\_b_0})$
- $b_1 \sim N(\mu_{0\_b_1}, \tau_{0\_b_1})$
- $\sigma_e^2 \sim IG(\alpha, \beta)$

Default Priors for `stan_glm`

- The normal distribution is the default prior for the intercept and regression coefficient and the exponential distribution is the default prior for the residual standard deviation (σ_e)

- More specifically, the default priors are
  - $b_0 \sim N(\mu = \bar{Y},\ \sigma = 2.5s_Y)$
  - $b_1 \sim N(\mu = 0, \sigma = 2.5\,{}^{s_Y}/_{s_X})$
  - $\sigma_e \sim Exp(\lambda = 1/s_Y)$

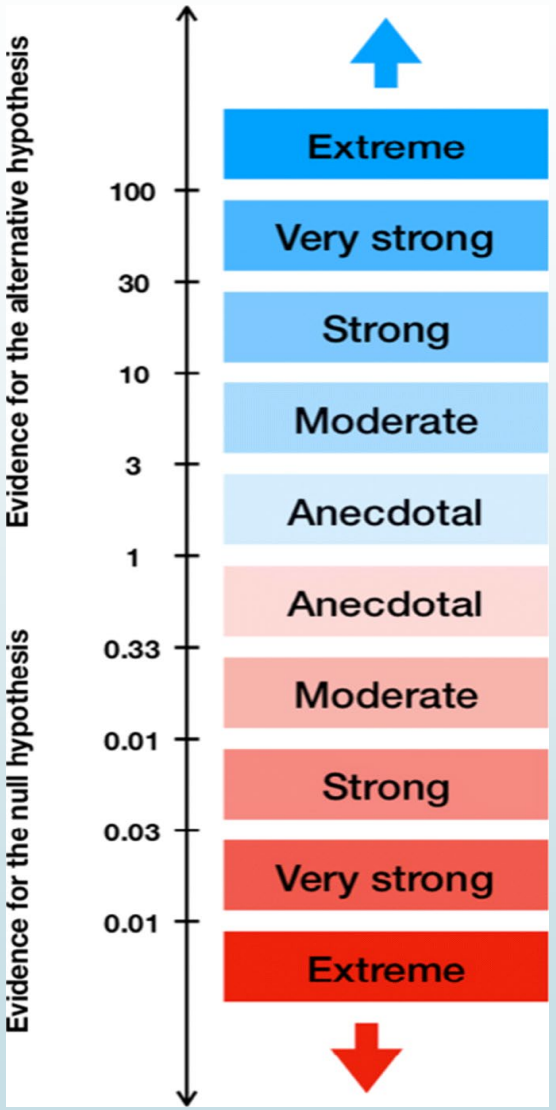
Informative Priors

- In many cases we have better "guesses" about the priors than the default priors
  - For example, previous (similar) studies exploring the relationship among the variables of interest
  - E.g., imagine past results found *b* = .5, *s* = 1.5
    - We could set the prior for the regression coefficient accordingly
      - $b_1 \sim N(.5, 1.5)$
    - But we might increase the scale of the prior to be conservative
      - E.g., $b_1 \sim N(.5, 3)$
- This is in the spirit of "updating" prior information via the data

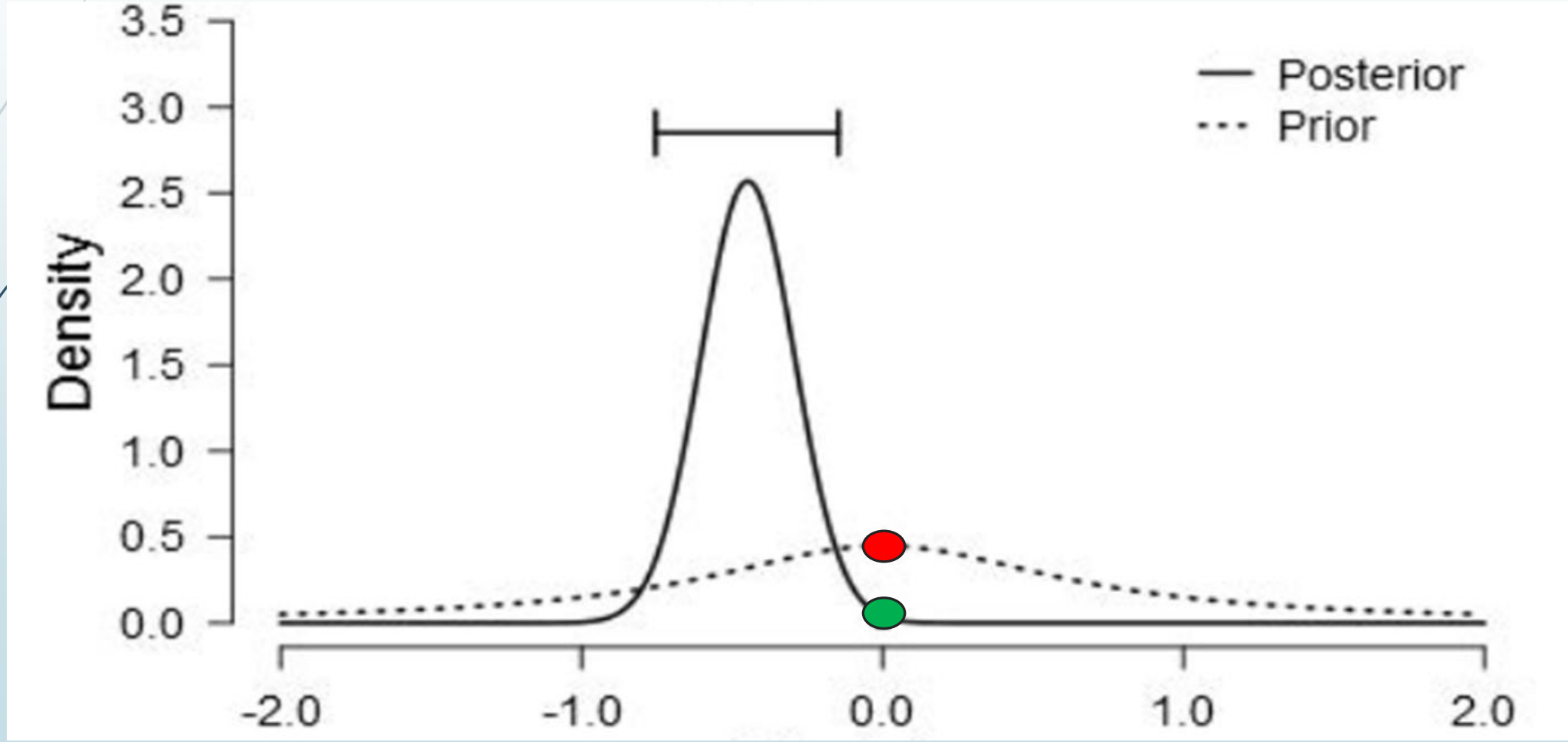
Bayes Factor

- The ratio of the likelihood of one particular hypothesis (e.g., alternative) to the likelihood of another (e.g., null)
  - Savage-Dickey Ratio
  - I.e., weight of the evidence in favor of a given hypothesis

- $BF = \frac{P(D|M_1)}{P(D|M_2)}$
- $BF = \frac{P(D|Alternative)}{P(D|Null)}$



Bayes Factor



Time to try things out ….