PSY 504: Advanced Statistics

Introduction to Bayesian Data Analysis

Jason Geller, Ph.D. (he/him/his)

Princeton University

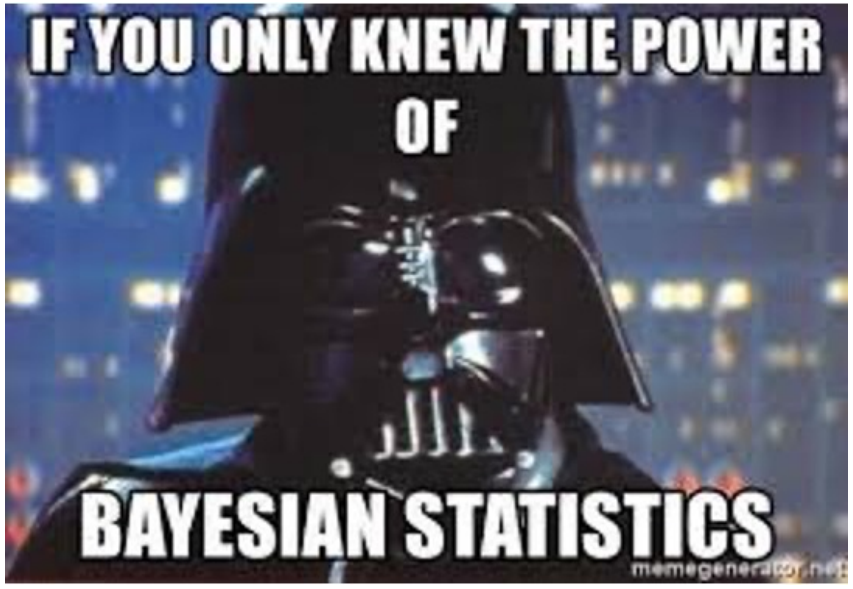
Updated:2023-03-18

Today

- Understand basic concepts of Bayesian statistics
- Learn how to conduct and interpret a simple Bayesian regression using **`brms`**

Bayesian statistics as a tool

- A lot of discussion on philosophical issues:
  - Subjective vs. objective probabilities
  - Frequentist vs. Bayesian (why I am better than you are)
  - p-values vs. subjective probabilities



What is Bayesian data analysis?

- It is when you use probability to represent uncertainty in all parts of a statistical model
- A flexible extension of maximum likelihood *(hey I know what that is!)*
- Can be computationally intensive

What is Bayesian data analysis?

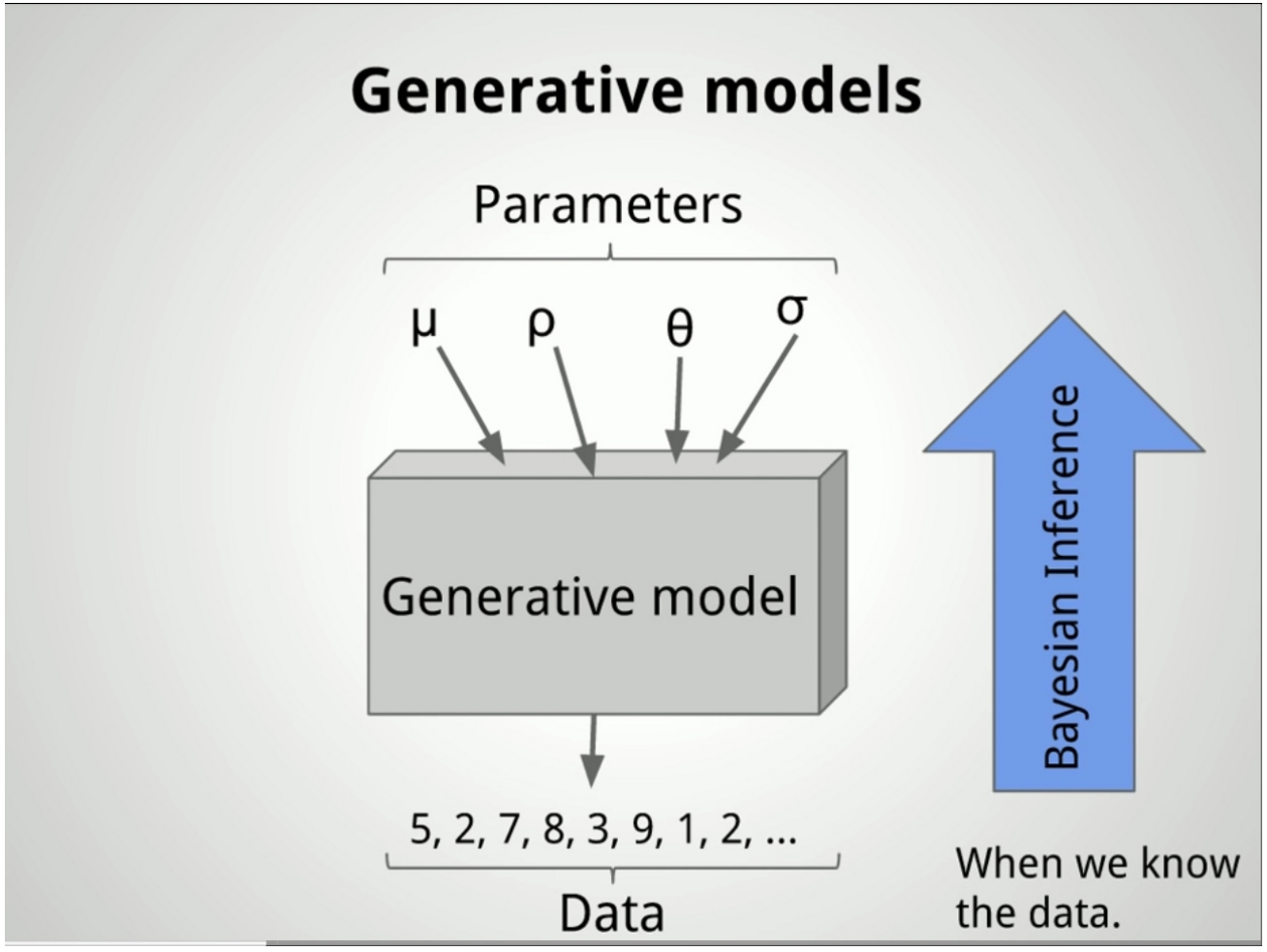
- A method for figuring out unknowns that requires three things:

1. Prior (what we know before data is collected)

2. Data

3. Generative models

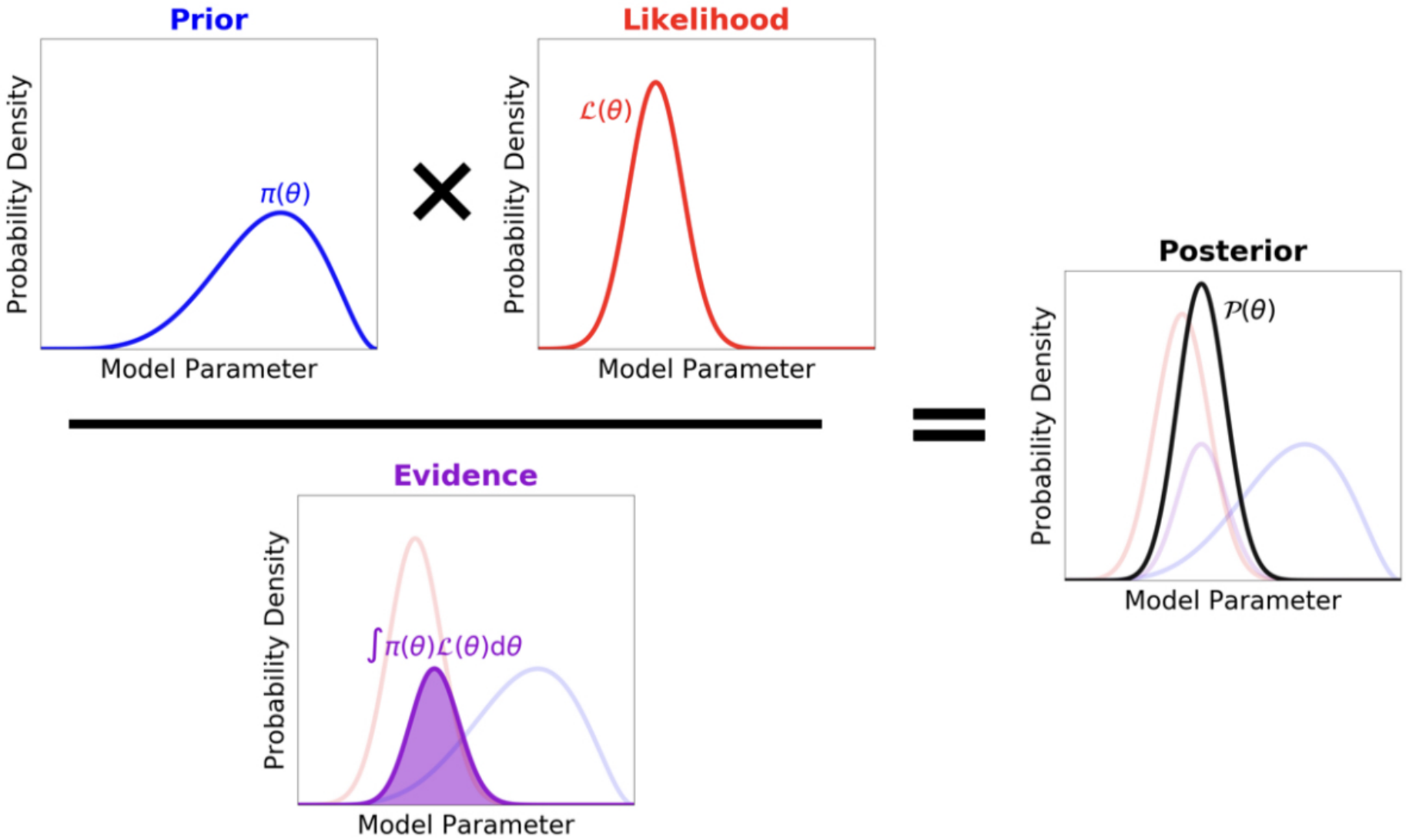
Generative model



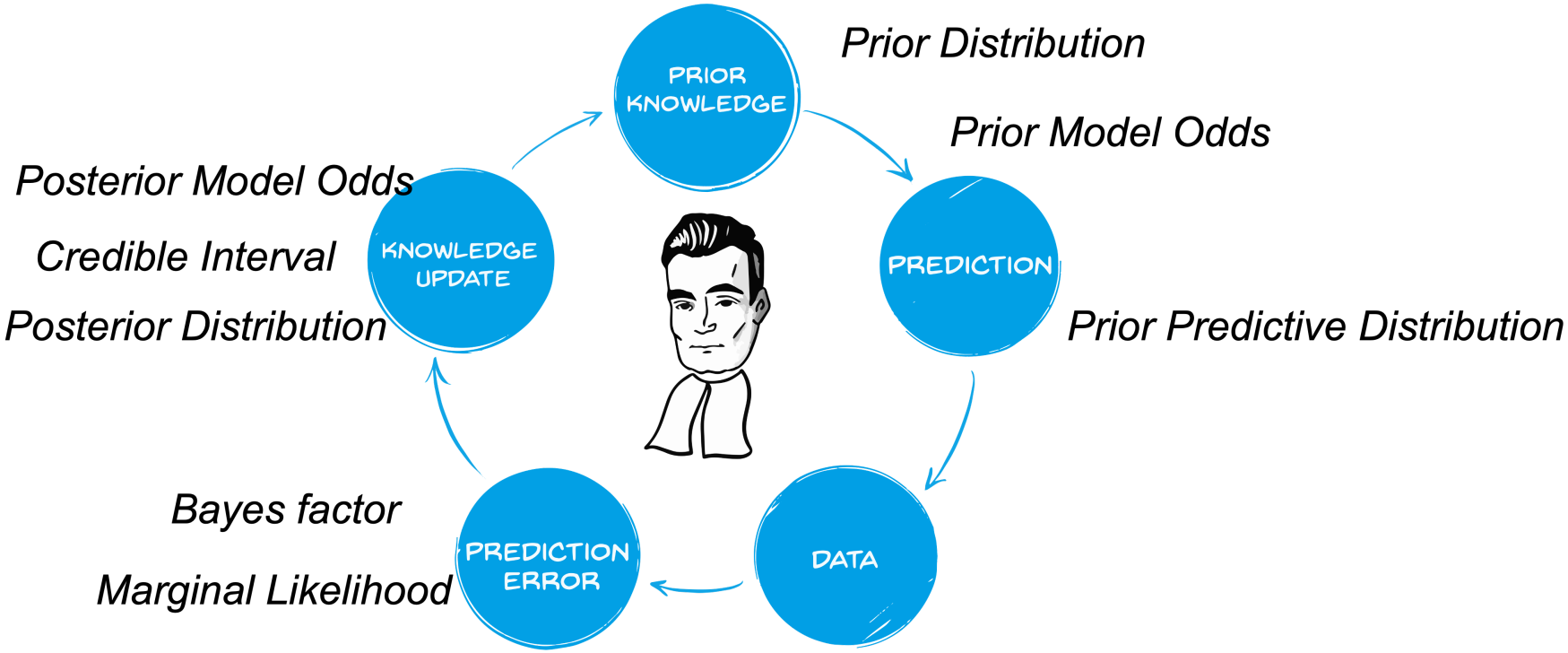
Bayes therom

$$\underbrace{p(\theta \mid \text{data})}_{\text{Posterior beliefs}} = \underbrace{p(\theta)}_{\text{Prior beliefs}} \times \underbrace{\overbrace{\frac{p(\text{data} \mid \theta)}{p(\text{data})}}^{\text{Prediction for specific } \theta}}_{\substack{\text{Average prediction} \\ \text{across all } \theta's}} .$$

- $p(\theta \mid \text{data})$ - The question you always wanted to test (posterior)
- $p(\theta)$ - expectation/prior belief (priors)
- $p(data \mid \theta)$ - How well data fits given estimated parameter value (likelihood)
- $p(data \mid M_k)$ - Marginal likelihood or evidence


Prior
Probability Density
$\pi(\theta)$
Model Parameter
Likelihood
Probability Density
$\mathcal{L}(\theta)$
Model Parameter
Evidence
Probability Density
$\int \pi(\theta)\mathcal{L}(\theta)\mathrm{d}\theta$
Model Parameter
Posterior
Probability Density
$\mathcal{P}(\theta)$
Model Parameter

Bayesian belief updating



Example

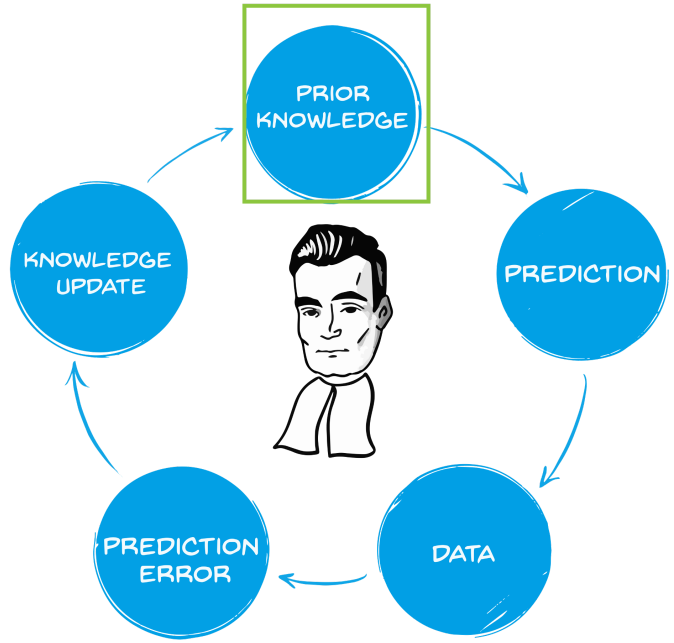
- We are interested in the percentage of dog people in the US
- People can be classified at dog people or cat people
- Data:
    - 0 = Cat person
    - 1 = Dog person

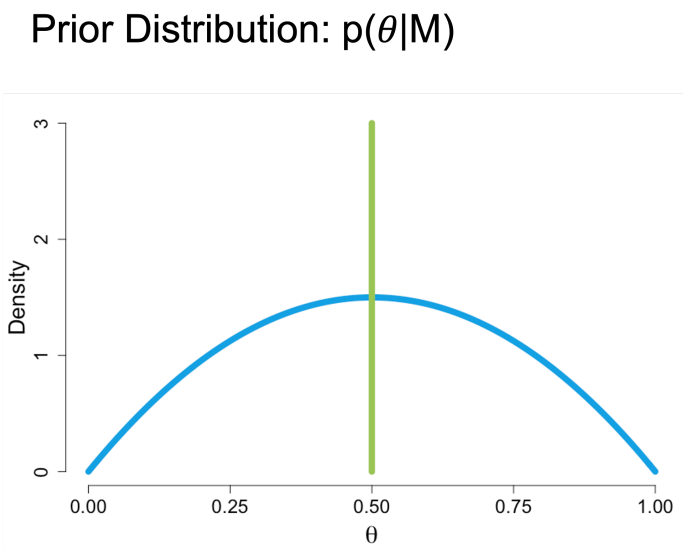
Parameters:

- θ = Proportion of dog people

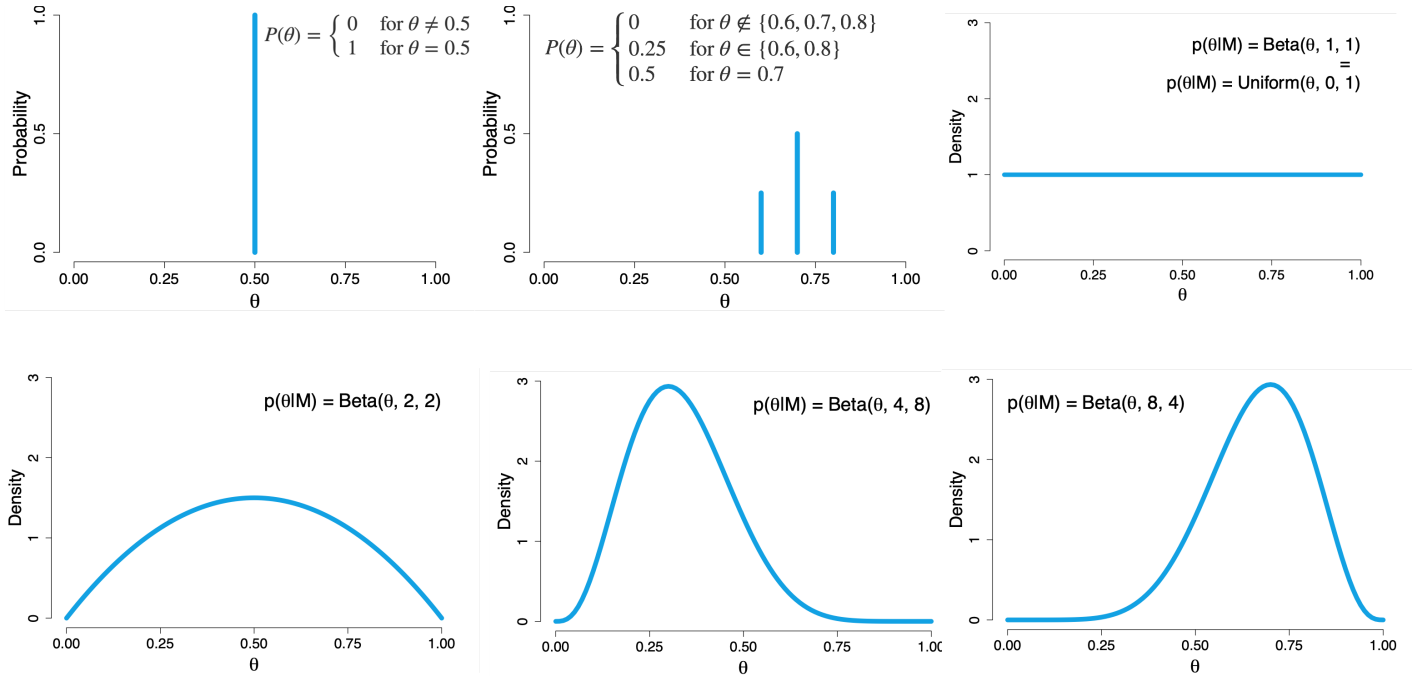
Bayesian belief updating

- Prior Probability
  - An unconditional probability distribution representing belief about a parameter BEFORE DATA COLLECTION





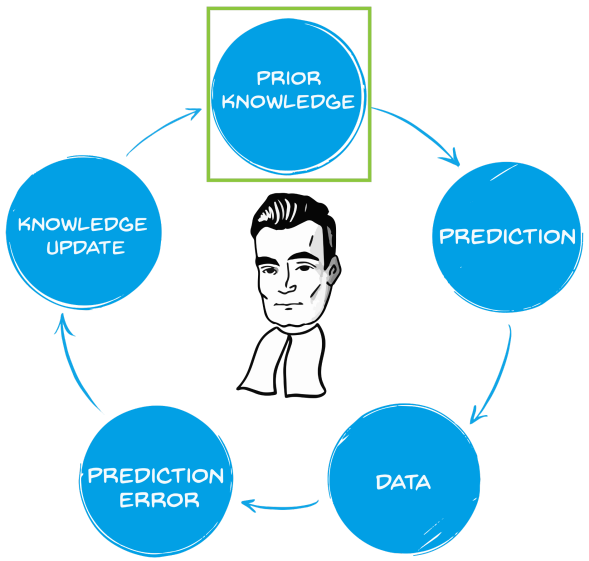
$$p(\theta \mid M)$$

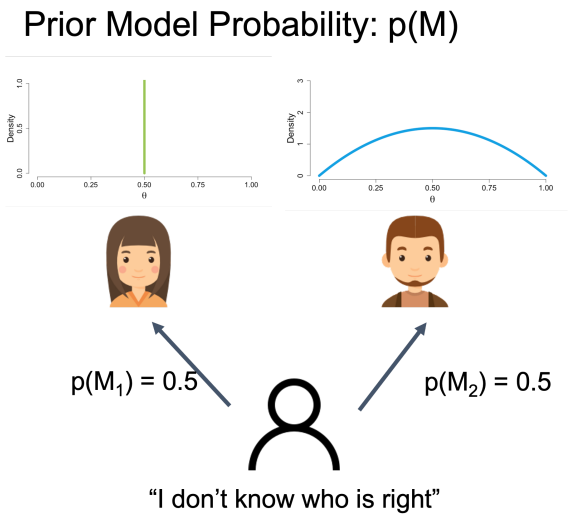

$P(\theta) = \begin{cases} 0 & \text{for } \theta \neq 0.5 \\ 1 & \text{for } \theta = 0.5 \end{cases}$
$P(\theta) = \begin{cases} 0 & \text{for } \theta \notin \{0.6, 0.7, 0.8\} \\ 0.25 & \text{for } \theta \in \{0.6, 0.8\} \\ 0.5 & \text{for } \theta = 0.7 \end{cases}$
Probability
Density
θ
0.00
0.25
0.50
0.75
1.00
p(θ|M) = Beta(θ, 1, 1)
=
p(θ|M) = Uniform(θ, 0, 1)
p(θ|M) = Beta(θ, 2, 2)
p(θ|M) = Beta(θ, 4, 8)
p(θ|M) = Beta(θ, 8, 4)

Bayesian belief updating

- Prior odds
  - Compares the relative plausibility of two models before data collection

$$\frac{p(M1)}{p(M2)}$$





Declaring Priors

You have to identify:

- Distribution of every statistic you want to estimate, including the dependent variable and each parameter of its distribution
  - (e.g., DV ~ N(μ , σ))
- Expected values for the location and spread of the distributions

How to choose

- People argue about priors
  - Priors differ in how informative they are
  - Priors differ in how proper they are
- Creates two camps:
  - “Subjective Bayesians” vs. “Objective Bayesians"

Informativeness of priors

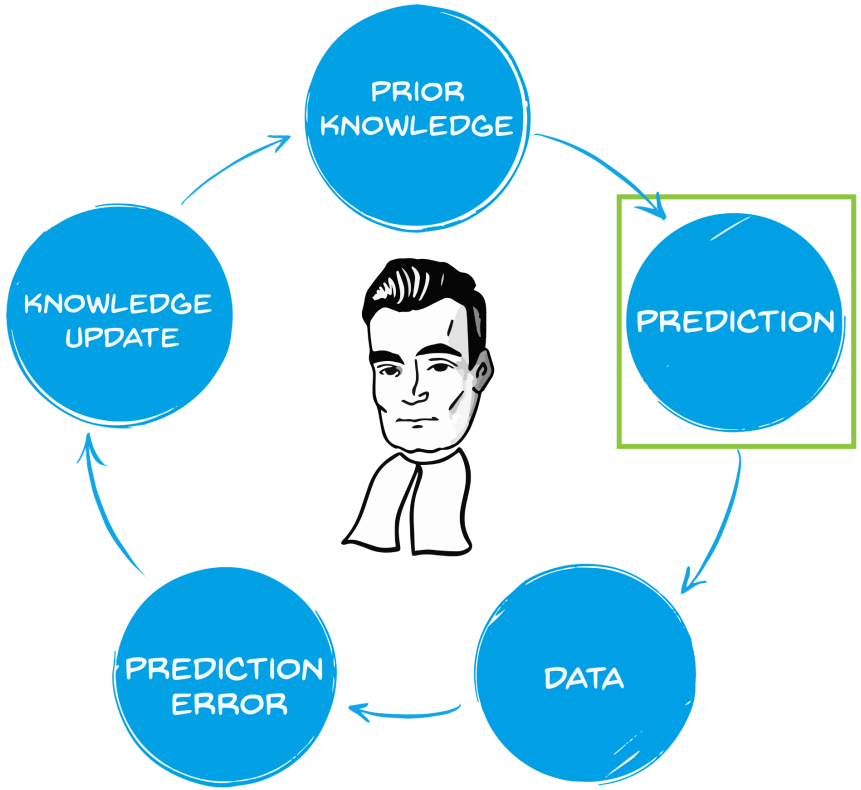
- Informative Priors (“Subjective Bayesians”)
  - Prior distributions that are specific about the values of model parameters (e.g., true correlation ≈ N(μ = -0.5)
- Non-informative Priors (“Objective Bayesians”)
  - Usually, uniform distributions that includes all values of a parameter (e.g., -1 ≤ true correlation ≤ +1, with every value having equal probability)
- Weakly-informative priors (“WIP”; Most Bayesians)
  - Specifying the distribution (e.g., Normal), with starting values known to bias estimates the least

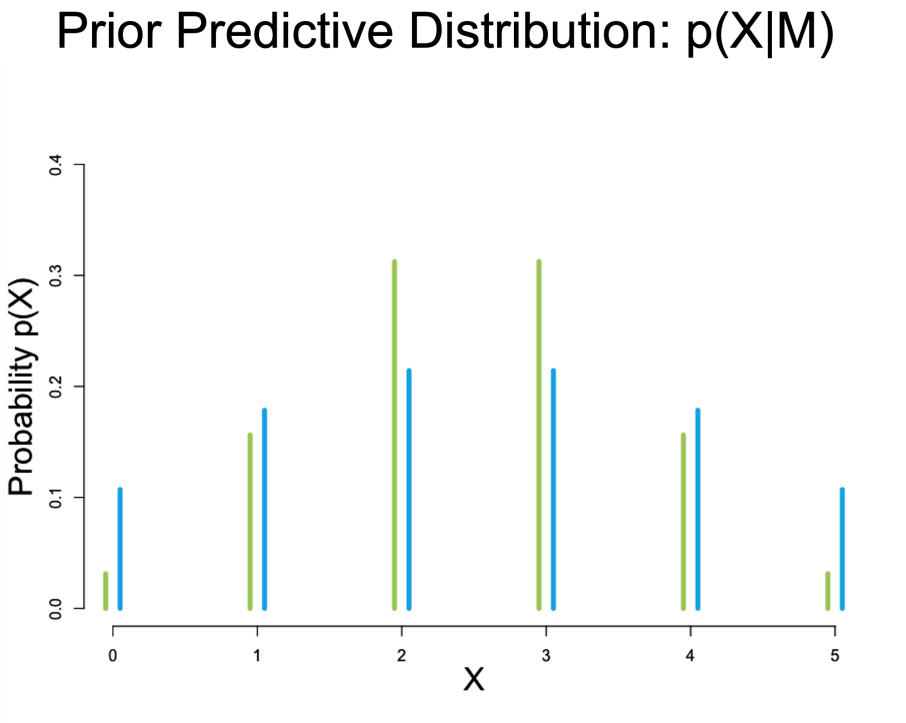
Informativeness of priors

- People vary in how strongly they state their prior beliefs
- If you state your belief strongly
  - E.g., the true correlation is ~N(0.3, 0.06)
  - Pitfall: Your beliefs have greater influence over the shape of the posterior distribution
- If you state your belief weakly
  - E.g., true correlation is equally likely at any real value between -1 and 1
  - Pitfall: You run the risk of overestimating the relative densities of the posterior distribution to the prior distribution

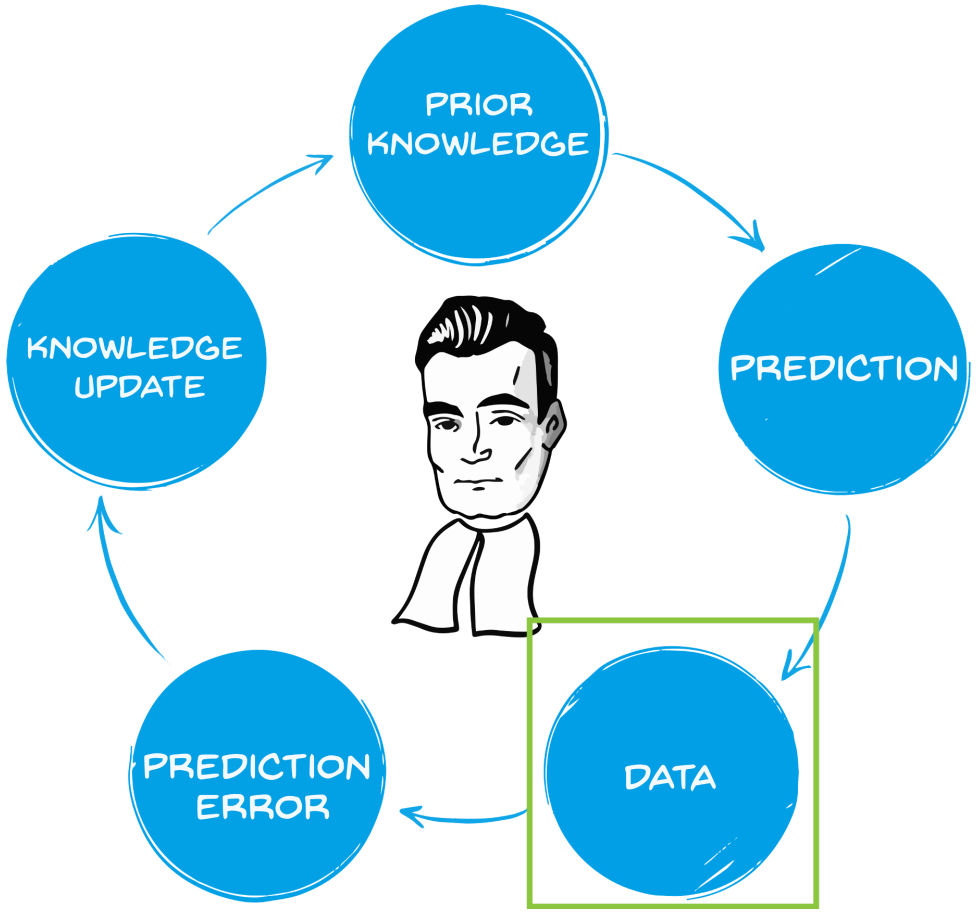
Bayesian belief updating

- What is the probability to observe 0, 1, 2, … dog people in a random sample of 5 people given our model?





Bayesian belief updating



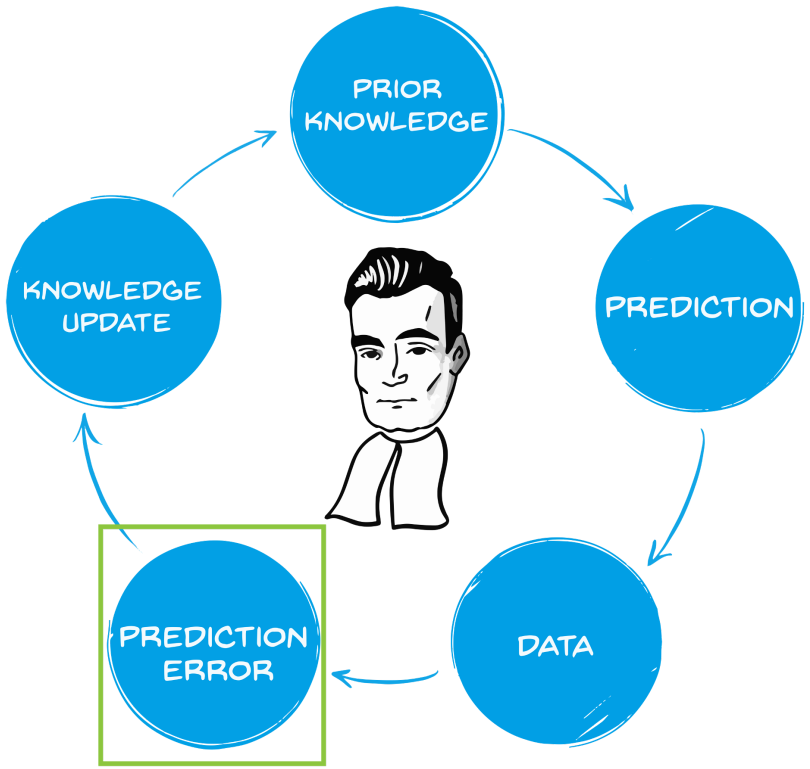
For example:

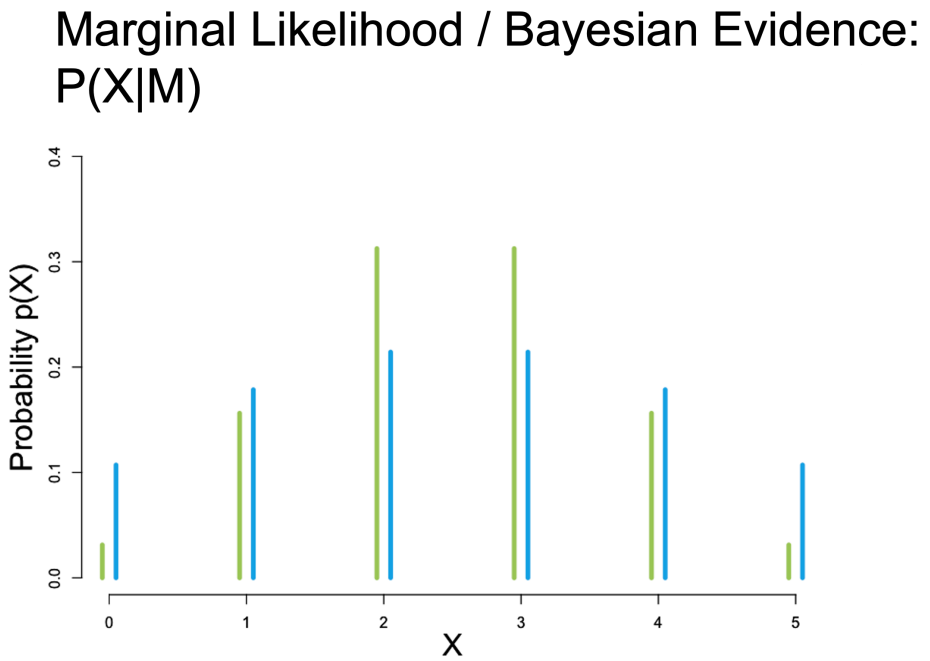
In a random sample of N = 5 we observe:

x = 3 Dog people
5-x = 2 Cat people

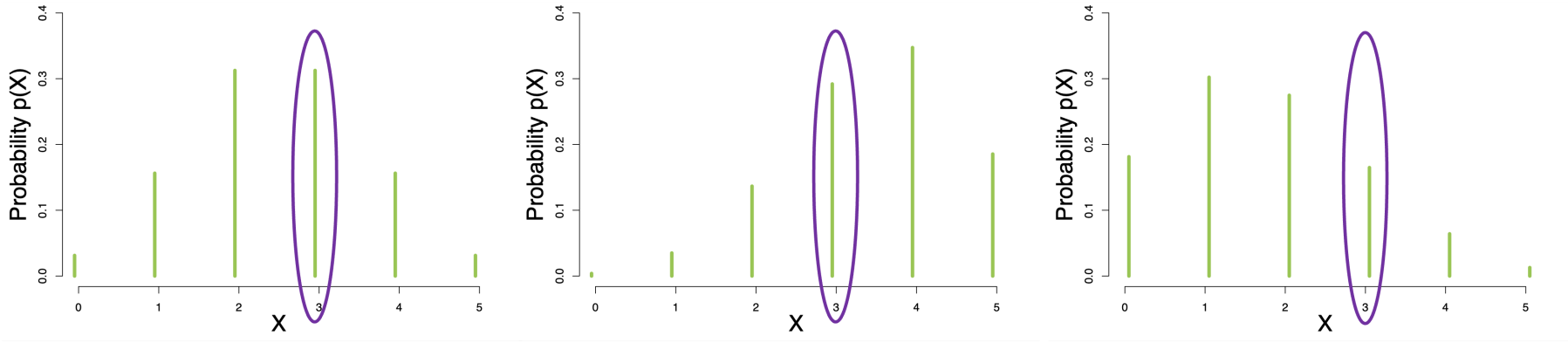
Bayesian belief updating

- How plausible are the observed data under the model?
  - Evaluation of the prior predictive distribution at the observed data

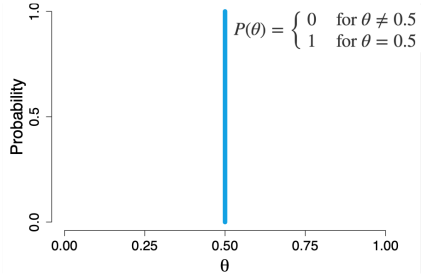


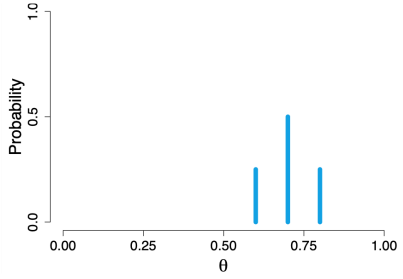


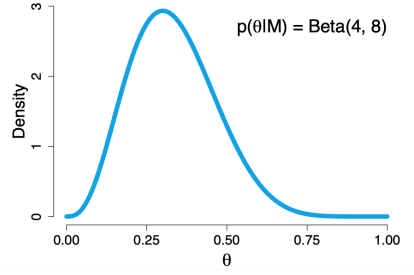
Bayesian belief updating



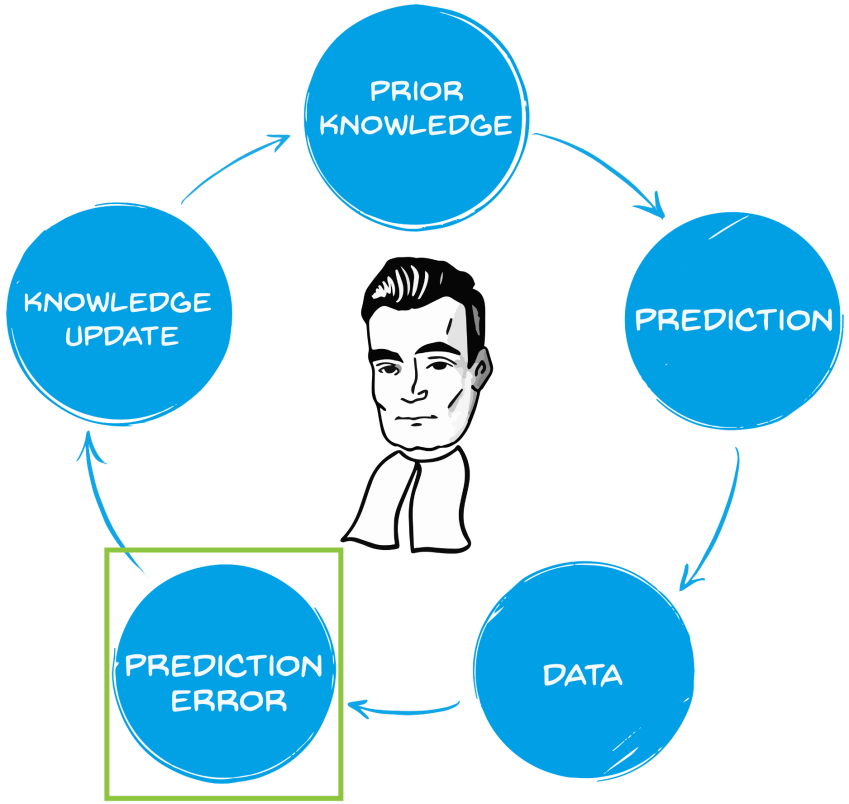
Prior predictive distributions based on the binomial likelihood and the following prior distributions



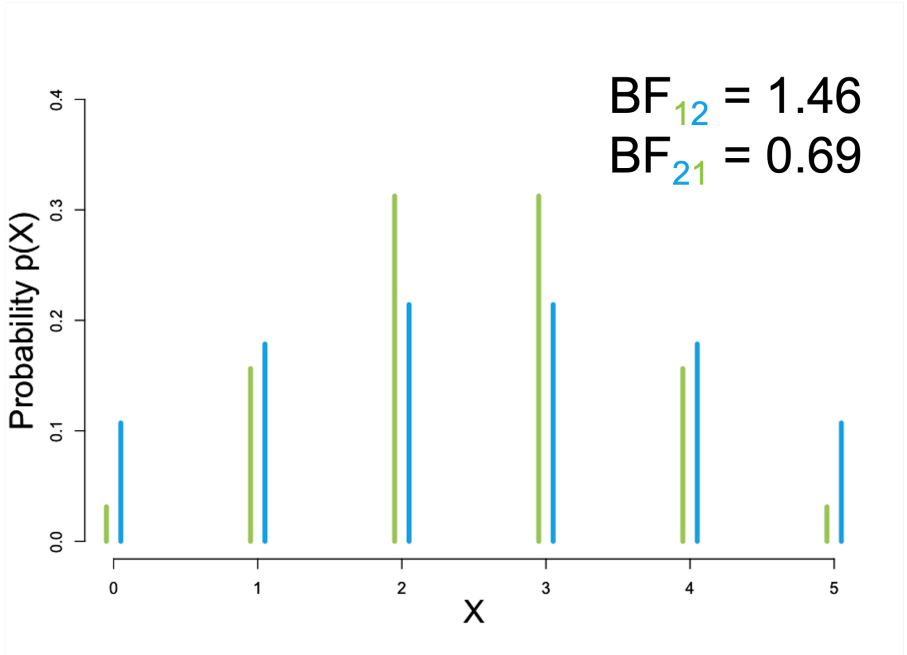




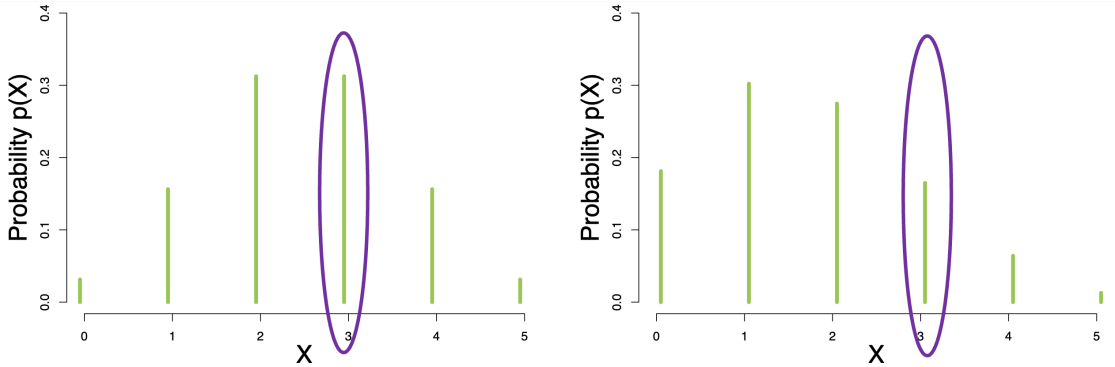
Bayesian belief updating



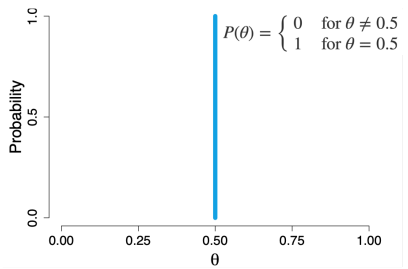
Bayes factor (BF): $P(X|M_1) / P(X|M_2)$

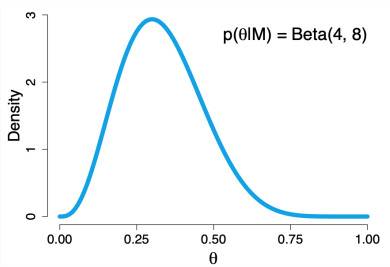


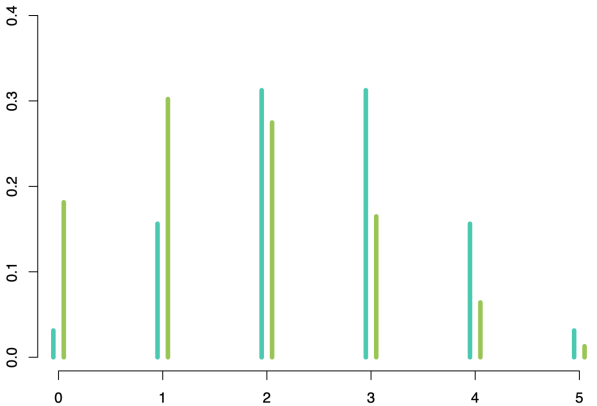
$p(x|M_1)/p(x|M_2)$



Prior predictive distributions for the binomial models with the following prior distributions on θ







$p(x=3|M_1) = 0.3125$
$p(x=3|M_2) = 0.1648$

$BF_{12} = 1.896$

Bayes factors

- Frequentists have *p* values
- Bayesians have Bayes factors (BF)
  - Tells you how much more likely the observed data are under one model than under another model
  - Can be interpreted as degree of relative evidence for a model
  - Typically: Spike prior under the null model, distribution under the alternative model

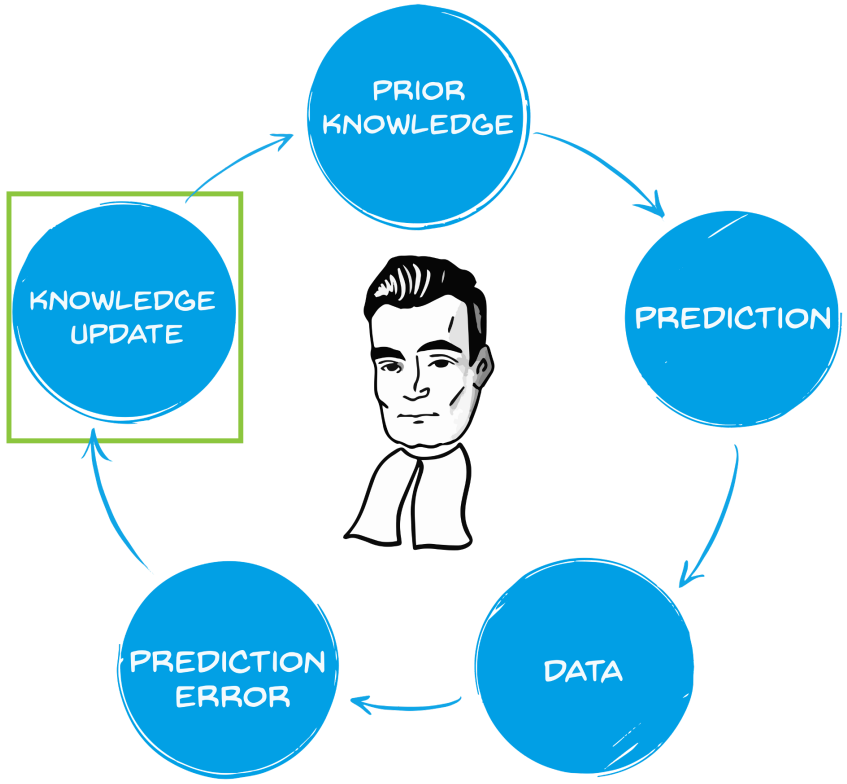
$$\text{Bayes factor}(BF) = \frac{P(\mathcal{D}|H_1)}{P(\mathcal{D}|H_0)}$$

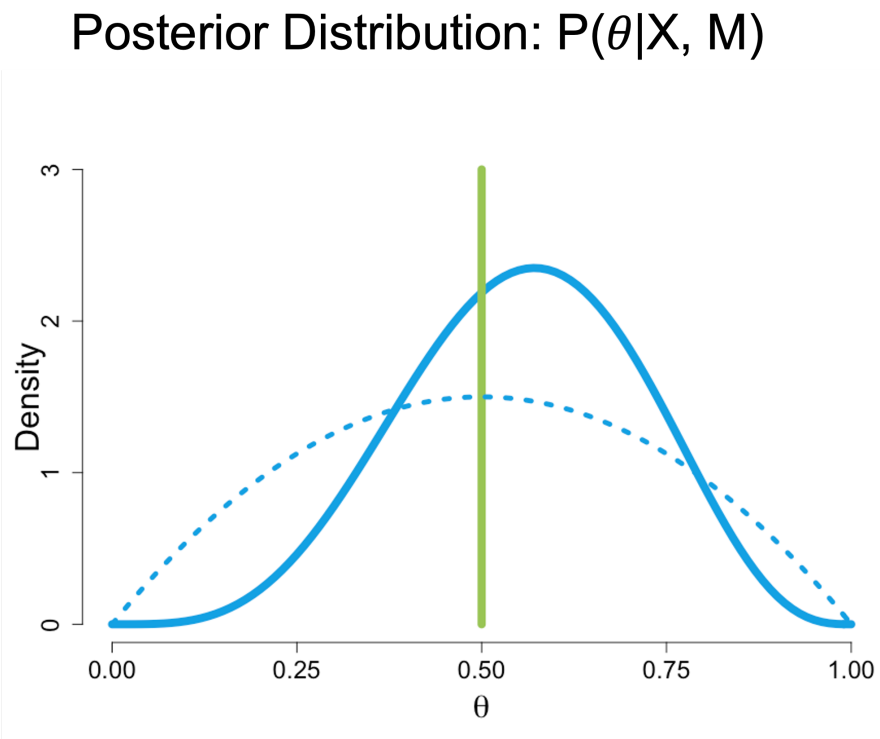
Bayes factors

Bayes factor	Evidence category
> 100	Extreme evidence for $\mathcal{H}_1$
30 - 100	Very strong evidence for $\mathcal{H}_1$
10 - 30	Strong evidence for $\mathcal{H}_1$
3 - 10	Moderate evidence for $\mathcal{H}_1$
1 - 3	Anecdotal evidence for $\mathcal{H}_1$
1	No evidence
1/3 - 1	Anecdotal evidence for $\mathcal{H}_0$
1/10 - 1/3	Moderate evidence for $\mathcal{H}_0$
1/30 - 1/10	Strong evidence for $\mathcal{H}_0$
1/100 - 1/30	Very strong evidence for $\mathcal{H}_0$

Bayesian belief updating

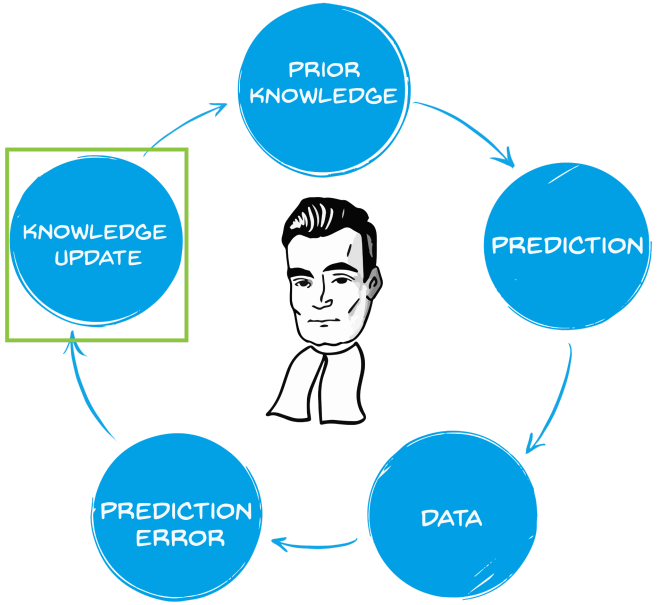
- A posterior distribution is a conditional probability distribution that represents belief about a parameter, taking the evidence into account

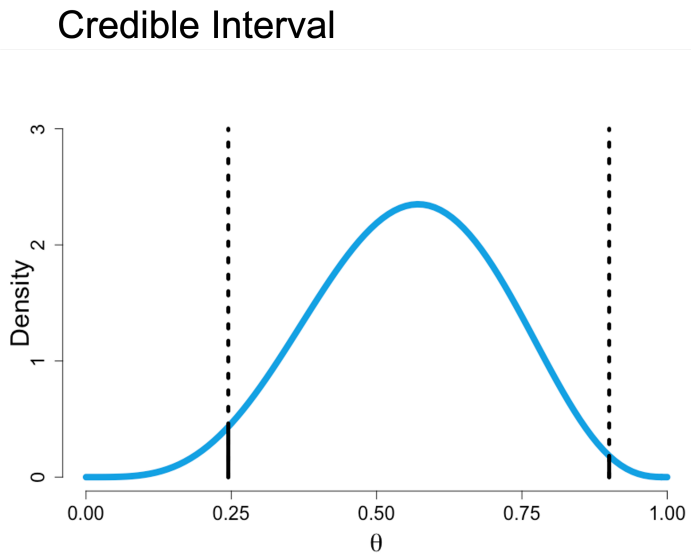




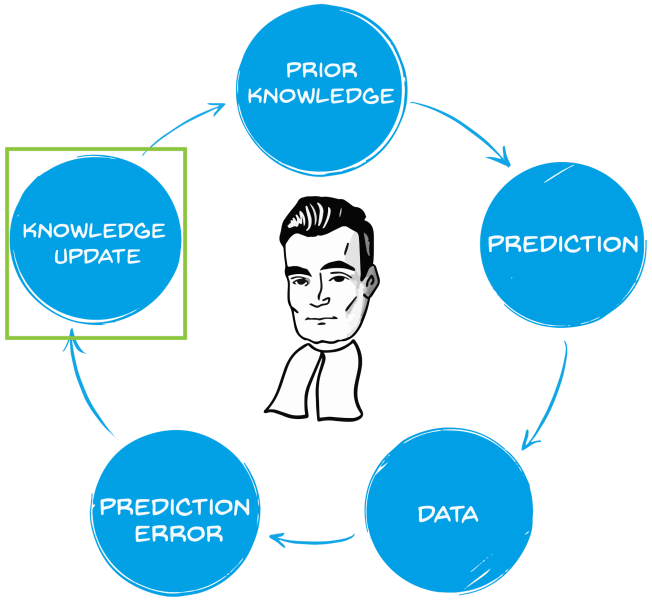
Bayesian belief updating

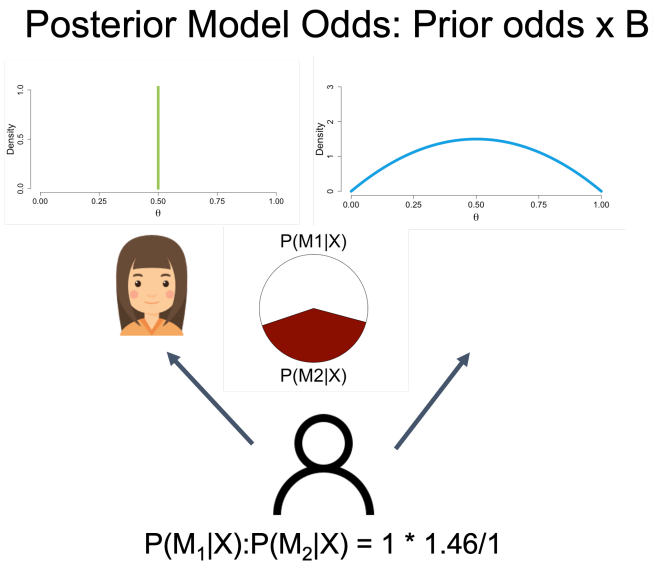
- Credible Intervals (Highest Density Intervals)
  - With a probability of x%, the parameter lies within this interval
  - Defined by the posterior distribution



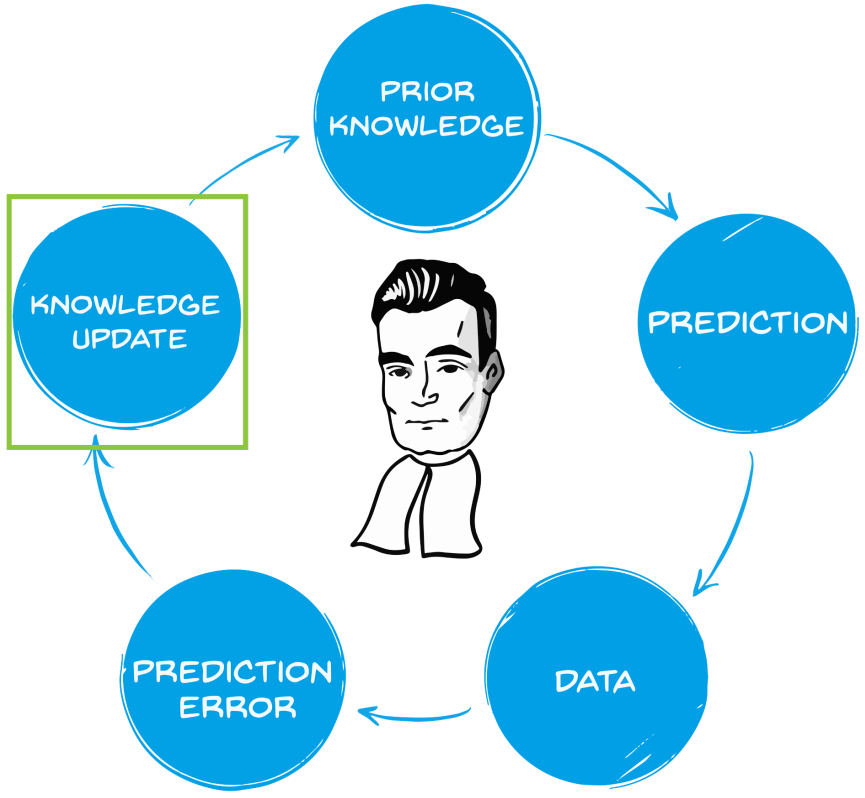


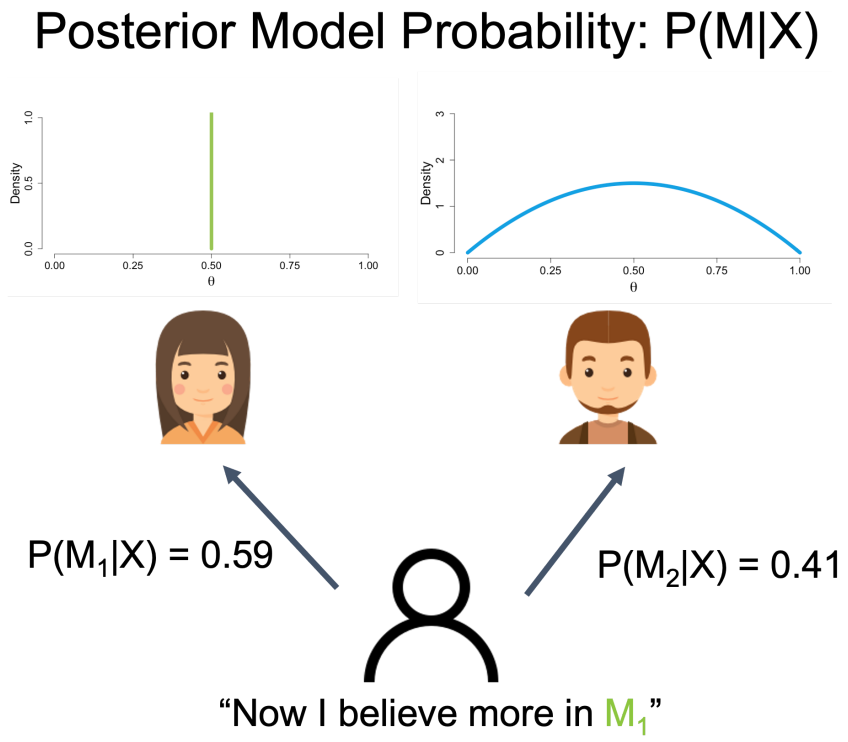
Bayesian belief updating





Bayesian belief updating





Today

- ✅ Understand basic concepts of Bayesian statistics
- Learn how to conduct and interpret a simple Bayesian regression using **`brms`**

Bayesain regression example

- Does synchronous attendance matter in hybrid courses?
  - 33 students in Fall 2020 statistics course
  - Looked at:
    - **`Final course grade`**: Max 100
    - **`Mode of attendance`**: (0=asychronous; 1=synchronous)
    - **`Average standardized viewing time for recorded lectures`**: in minutes

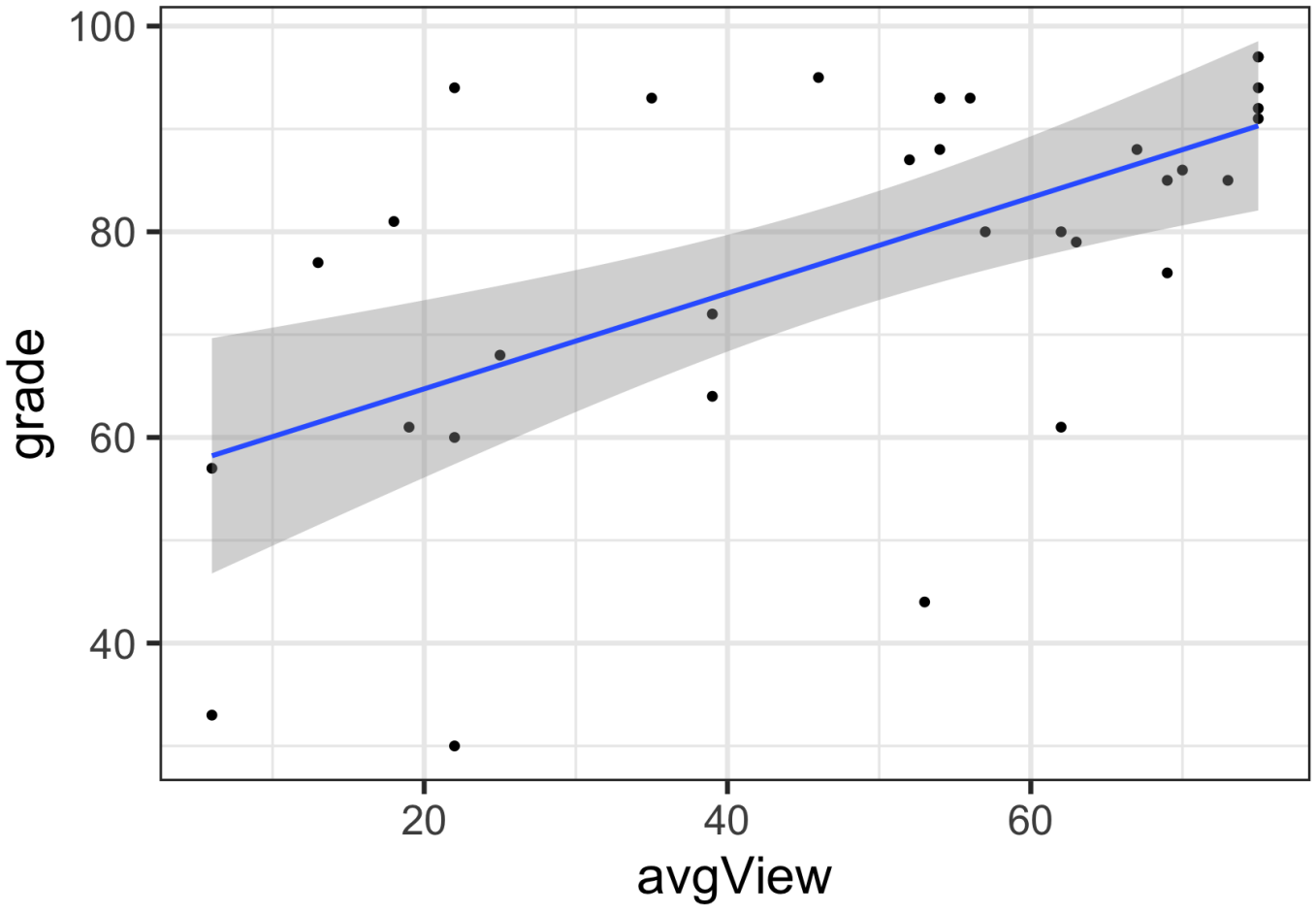
Data

```
library(httr)
#can read directly from osf
data<-read_csv("https://osf.io/sxk2a/download")
```

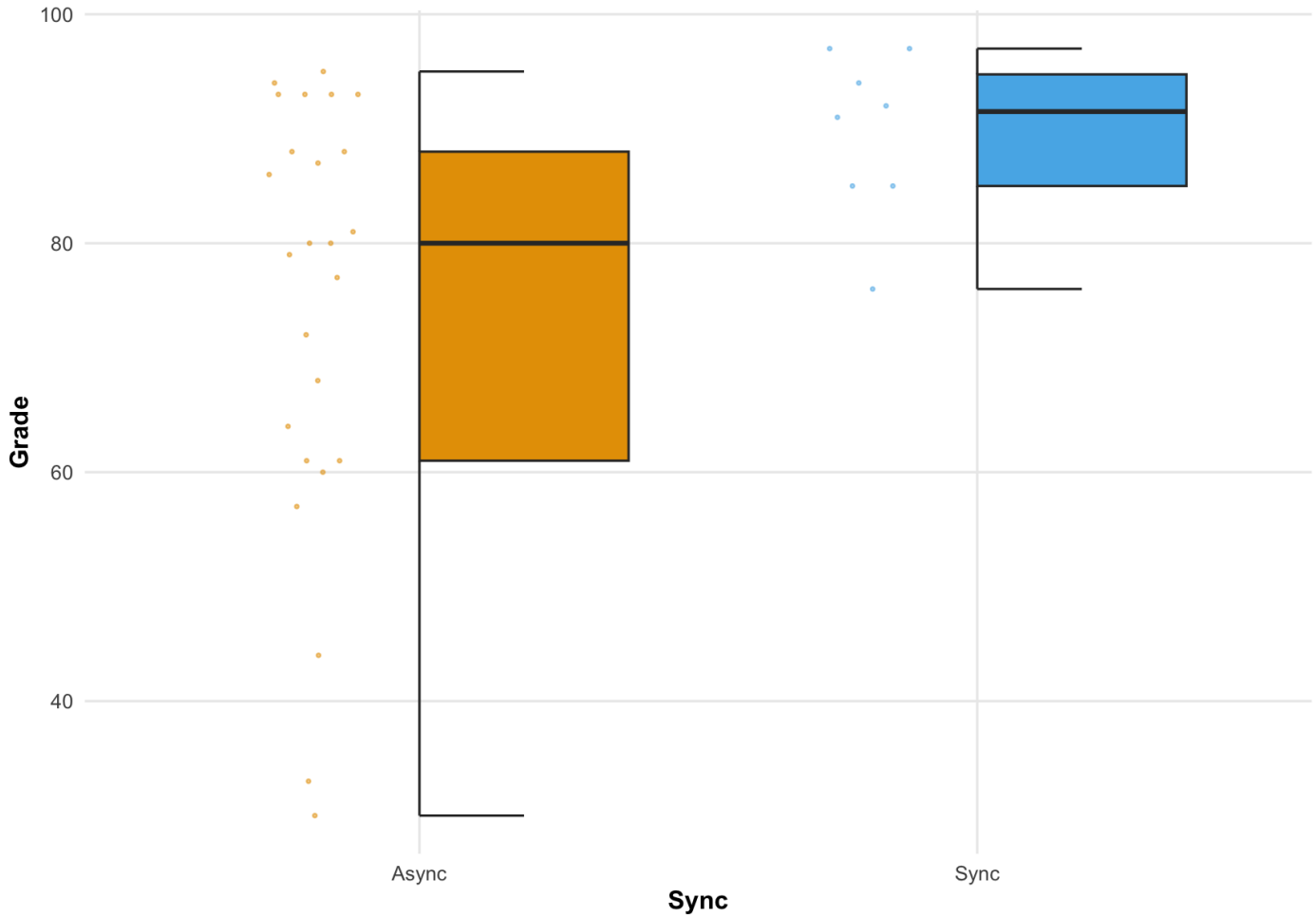
Packages

```
library(brms) # run bayes lm
library(emmeans) # get posteriors
library(ggeffects) # graph
library(easystats) # easystats packages # bayesttestR
library(bayesplot) # graph trace plots
```

avgView plot



sync plot



Simple regression

```
lm_class <- lm(grade~avgView+sync_cont, data=data)

kable(tidy(lm_class), digits=3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	55.676	9.230	6.032	0.000
avgView	0.461	0.153	3.020	0.005
sync_cont	0.297	7.920	0.038	0.970

brms

- Bayesian regression models in Stan (**brms**)

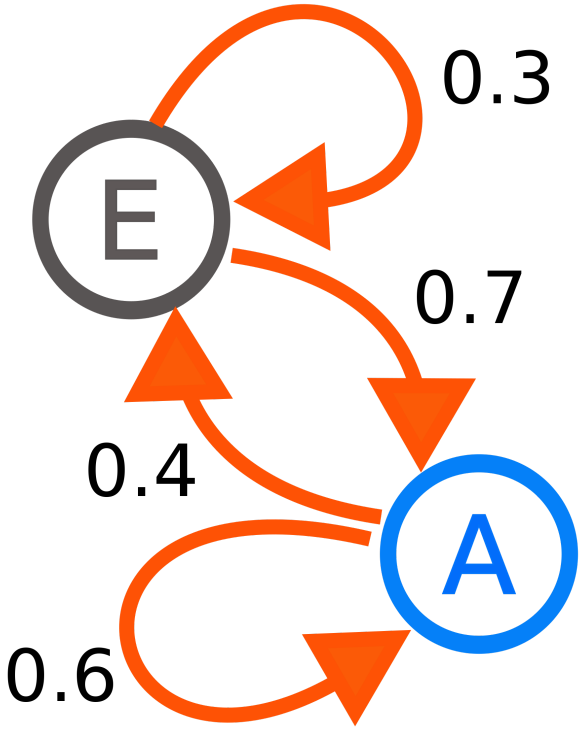
```
library(brms)
brm_class1 <- brm(grade~avgView, data=data,
family= gaussian(),#distribution
prior=NULL,
chains=4, # how many chains are run
core=4, #computer cores to use
warmup = 2000, # warm-up for MCMC
iter = 5000) # number of MCMC samples
```

- Let's go to R to run this

Computing the posterior

- Markov chain Monte Carlo (MCMC) sampler!
- Given possible priors and your data, a computer uses a Monte Carlo sampling technique to build stochastic Markov Chains, a process referred to as MCMC
- We run multiple chains (e.g., 4 chains in **`brms`**) with equal numbers of iterations (e.g., 5000 iterations) in each chain to estimate convergence/stability
- MCMC chains contain samples from the posterior distribution of the theory given the data

Markov chains



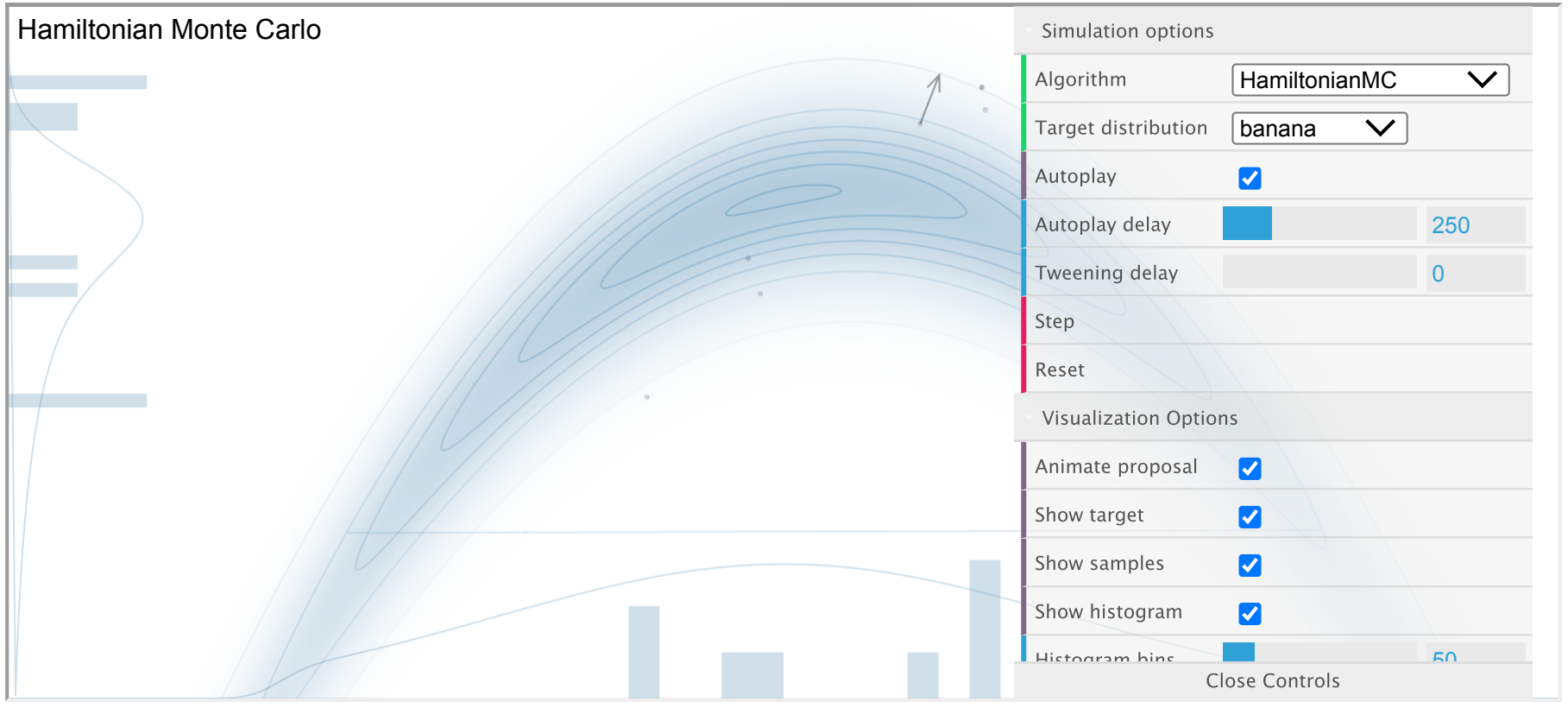
- Chain of discrete events, moving forward in time
  - Probability of each event is a conditional probability, given the last event
- Each event is wholly predicted by the immediately preceding event
- They can stay the same/loop
- Future events can be predicted by knowing only the current event

Differences Between MCMC and Bootstrapping

- An entire set of bootstrapping resamples would be practically equivalent to one MCMC “chain” in the analysis
- Bootstrap resamples are independent of each other, MCMC iterations are dependent on each other
- MCMC iterations can get stuck
- The first n iterations or samples in an MCMC chain are generated as “burn in” samples that will be the priors of the recorded MCMC samples

MCMC in action

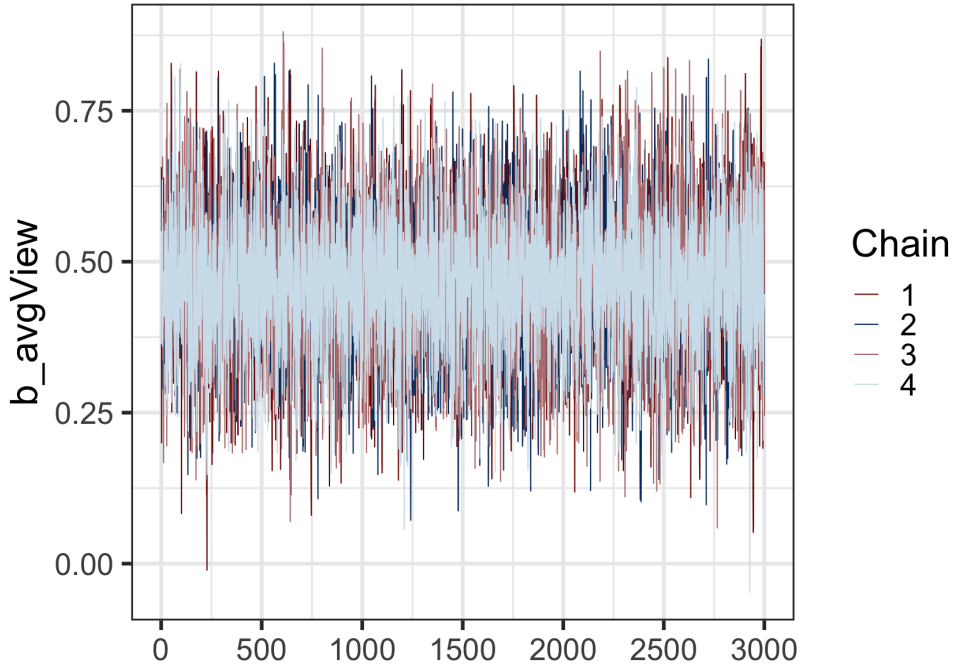
```
knitr::include_url("https://chi-feng.github.io/mcmc-demo/app.html")
```



MCMC Diagnostics

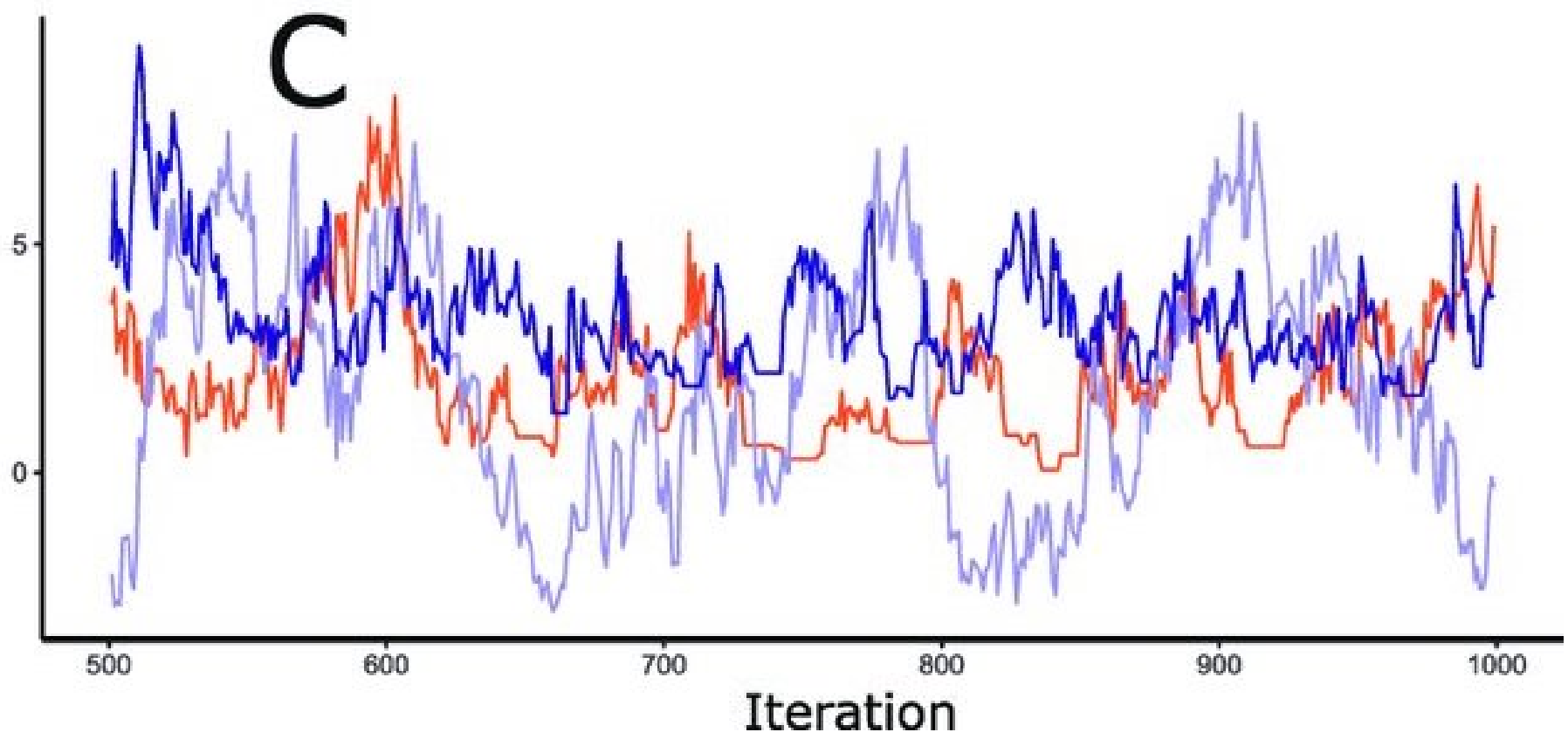
- Look for the fuzzy catapliers

```
bayesplot::color_scheme_set("mix-blue-red")
bayesplot::mcmc_trace(brm_class1, pars = c("b_avgView"),
           facet_args = list(ncol = 1, strip.position = "left"))
```



MCMC Diagnostics

- Bad plots



Other diagnostics

```
kable(diagnostic_posterior(brm_class1), digits=3)
```

Parameter	Rhat	ESS	MCSE
b_avgView	1	10528.96	0.001
b_Intercept	1	10952.91	0.060

- $\hat{R}$
  - Measure of consistency of Markov chains
  - Should be close to 1 (not larger than 1.01)
  - Ratio of variance (like *F* test)

Other Diagnostics

```
kable(diagnostic_posterior(brm_class1), digits=3)
```

Parameter	Rhat	ESS	MCSE
b_avgView	1	10528.96	0.001
b_Intercept	1	10952.91	0.060

- Effective sample size
  - MCMC chains are autocorrelated
  - Number of independent pieces there is in autocorrelated chains (Krushke, 2015, p182-3)
  - Should be > 1000

Priors

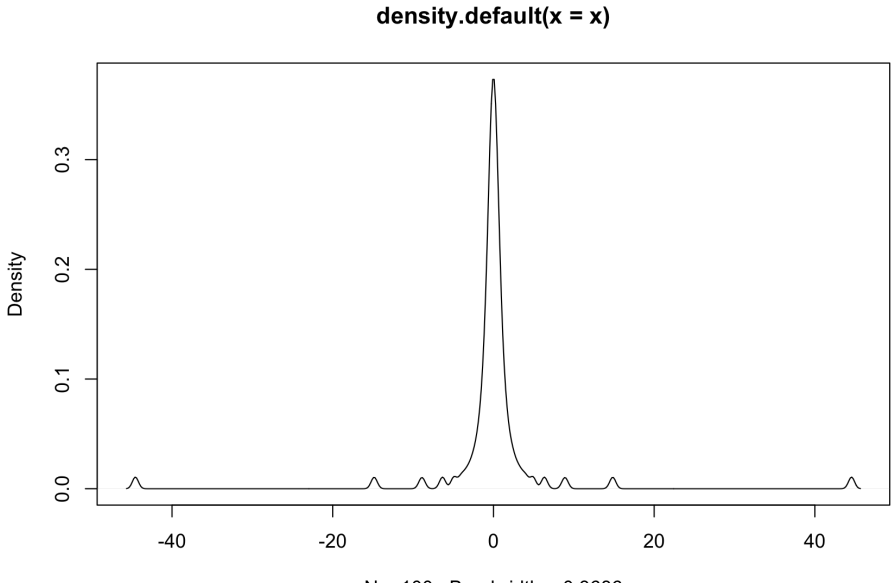
```
prior_summary(brm_class1)
```

```
##                  prior     class    coef group resp dpar nlpar lb ub
##                 (flat)         b                                    
##                 (flat)         b avgView                            
##  student_t(3, 85, 11.9) Intercept                                    
##   student_t(3, 0, 11.9)     sigma                                  0
##         source
##        default
##  (vectorized)
##        default
##        default
```

Weakly informative priors

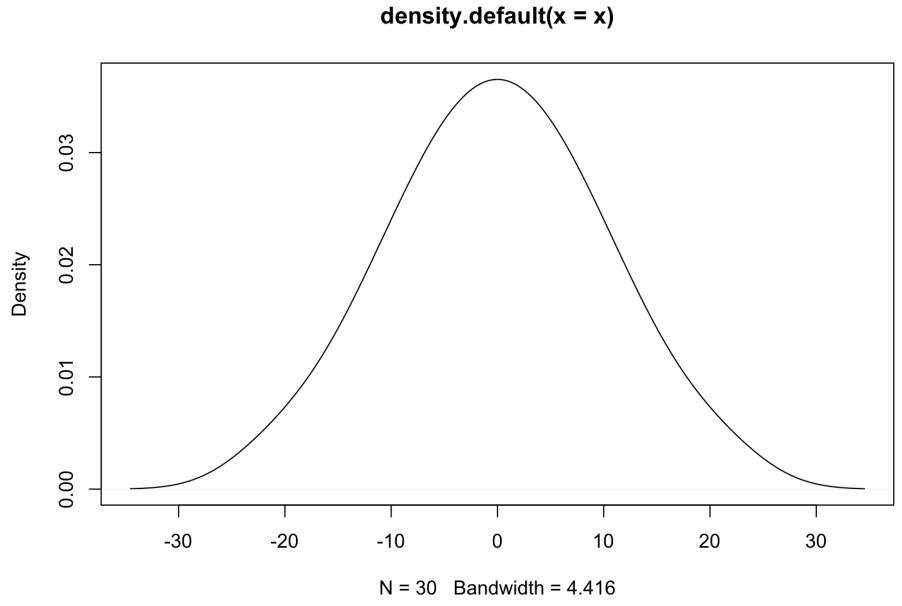
- cauchy(0, .7)
- cauchy(0, 2)

```
x=distribution_cauchy(100, location=0,

plot(density(x))
```



- normal(0, 10); normal(0, empirical rule)

```
x=distribution_normal(30, 0, 10)
plot(density(x))
```

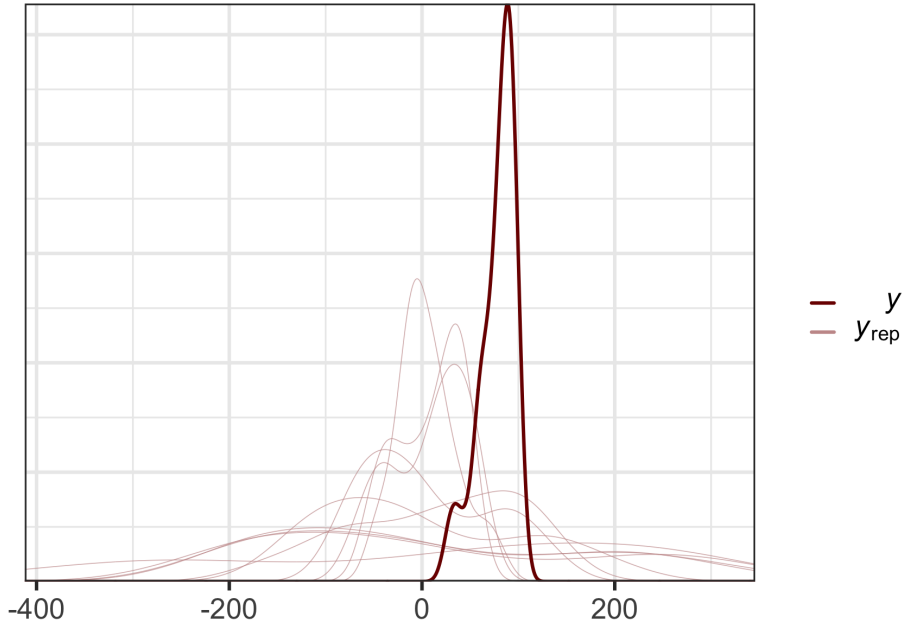


Visualize prior predictive distribution

- Make sure prior distribution makes sensible predictions

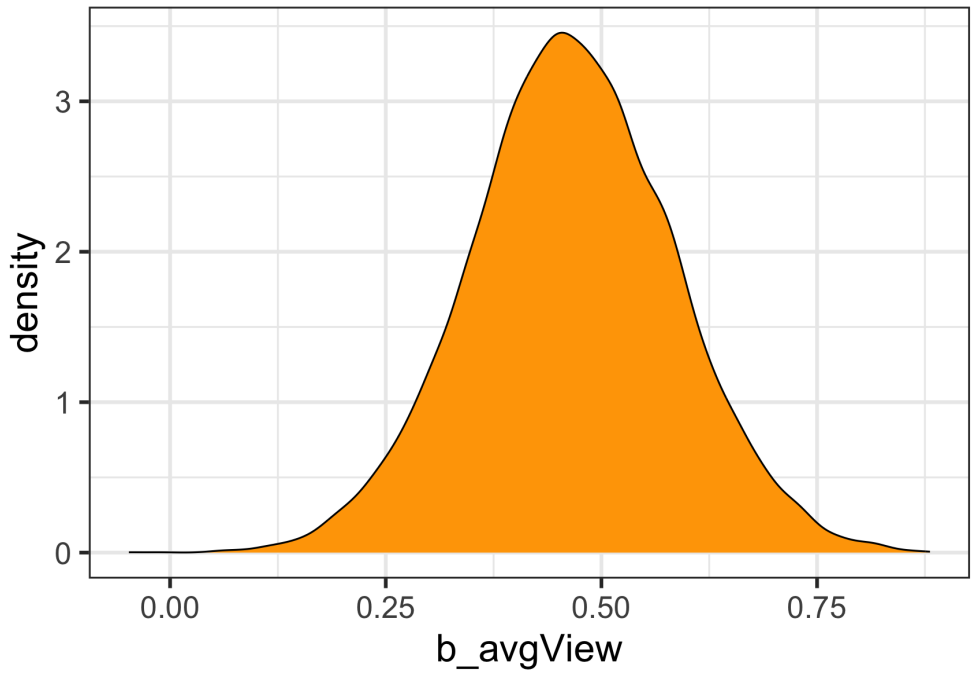
```
# set prior
bprior <- c(prior("normal(0,10)", class
            prior("normal(0,10)", class

prior1 <- prior(cauchy(0, .707), class
prior2 <- prior(normal(0, 10), class =
prior3 <- prior(normal(0, .51), class =

#include prior
# only sample from prior so we can plo
brm_class_prior <- brm(grade~avgView, d
prior=bprior, # add in prior informati
family= gaussian(),
warmup = 2000,
iter = 5000)

# check prior
```

```
pp_check(brm_class_prior)
```



Visualing the posterior distribution

```
posteriors <- get_parameters(brm_class1)

ggplot(posteriors, aes(x = b_avgView)) +
  geom_density(fill = "orange")
```



Describing the Posterior

1. A point-estimate which is a one-value summary (similar to the β in frequentist regressions)

2. A credible interval representing the associated uncertainty

3. Some indices of significance, giving information about the relative importance of this effect (e.g., Bayes Factors)

Point-estimate

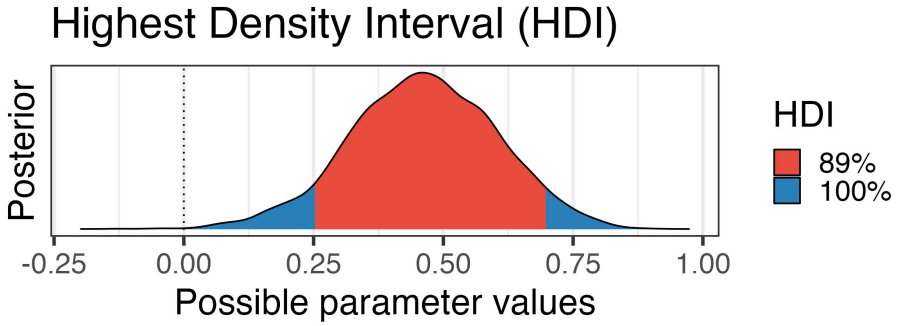
```
describe_posterior(
  brm_class1,
  effects = "all",
  component = "all",
  centrality = "all"
)
```

```
## Summary of Posterior Distribution
##
## Parameter   | Median |  Mean |   MAP |         95% CI |     pd |          ROPE | % in ROP
## ------------------------------------------------------------------------------------
## (Intercept) |  55.77 | 55.81 | 55.16 | [43.18, 68.09] |   100% | [-1.80, 1.80] |        0
## avgView     |   0.46 |  0.47 |  0.46 | [ 0.23,  0.70] | 99.98% | [-1.80, 1.80] |      100
##
## # Fixed effects sigma
##
## Parameter | Median |  Mean |   MAP |         95% CI |   pd |          ROPE | % in ROPE |
## ------------------------------------------------------------------------------------
## sigma     |  14.96 | 15.15 | 14.59 | [11.95, 19.46] | 100% | [-1.80, 1.80] |        0% |
```

Uncertainty: Credible intervals

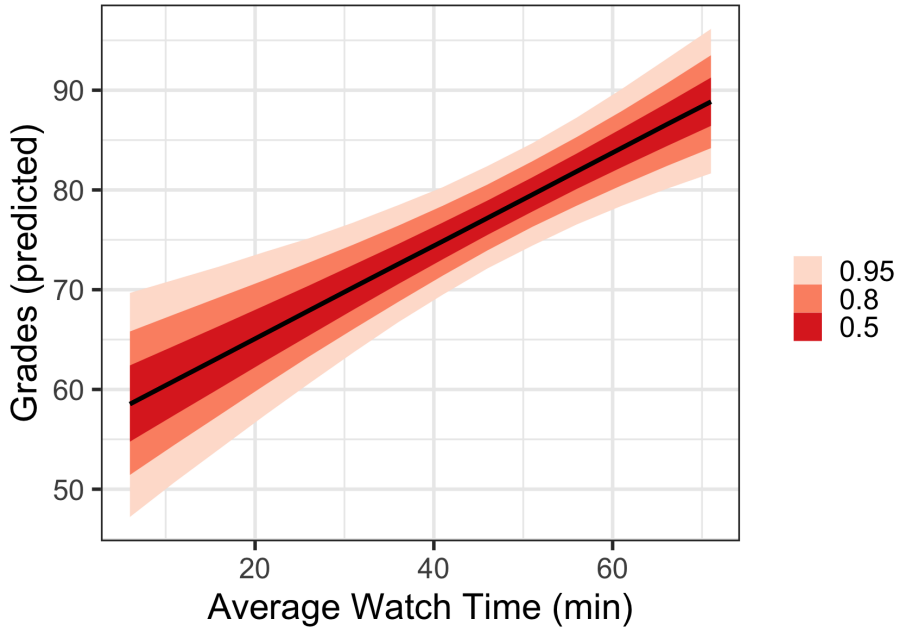
- Credible intervals (high density intervals)
  - Common to use 89% HDIs (why?)
  - Provides more stable estimates
  - If 95%, need to increase number of iterations

```
resuls=hdi(posteriors$b_avgView, ci=0.
```



Visualizing uncertainty

```
pred <- predictions(brm_class1,
                    newdata = datagrid
        avgView = seq(6, 75, by = 5)))
        posterior_draws()

pred_fig <- ggplot(pred, aes(x = avgVi
    stat_lineribbon() +
    scale_fill_brewer(palette = "Reds"
    labs(x = "Average Watch Time (min)
         y = "Grades (predicted)",
         fill = "")
```



Significance

- Does the credible interval contain 0?
  - If yes, "not significant"
  - If no, "significant"

Significance

- Test if effect is greater than 0, or equal to 0

```
brm_class_pr <- brm(grade~avgView, data=data,#use this to check prior pulls
prior=bprior,
sample_prior = TRUE,
family= gaussian(),
warmup = 2000,
iter = 5000)
```

```
BF <- bayestestR::bayesfactor_parameters(brm_class_pr, null=0)

BF
```

```
## Bayes Factor (Savage-Dickey density ratio)
##
## Parameter   |       BF
## ----------------------
## (Intercept) | 2.91e+03
## avgView     |     4.14
##
```

Adding categorical predictor

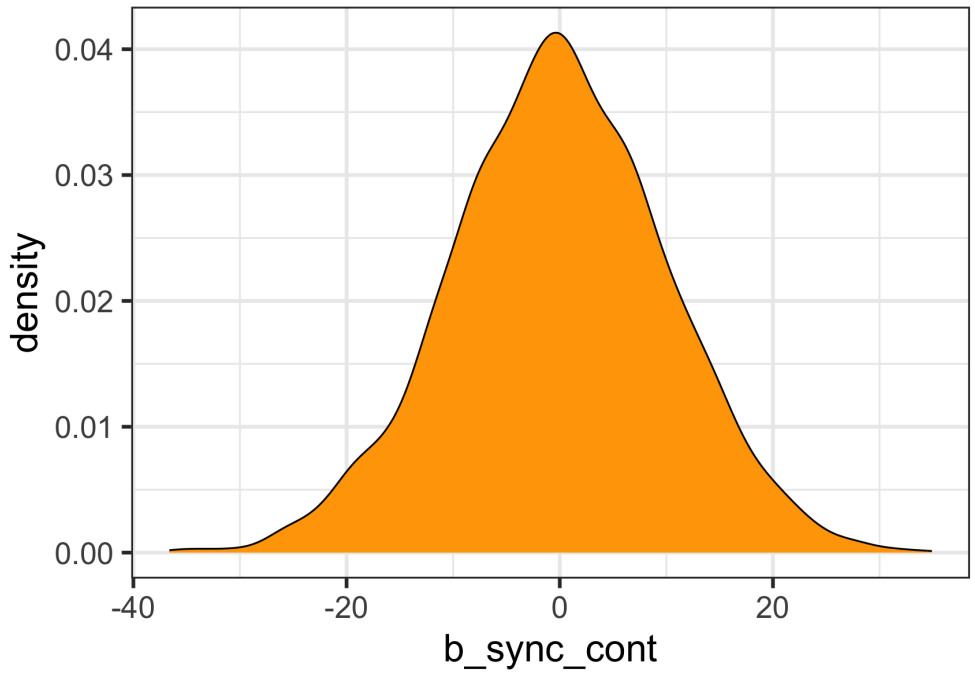
- **avgView** seems to have an effect on grades
- Let's add **sync** to our model

```
data$sync_cont<-ifelse(data$sync==0, -.5, .5) # contrast code var

brm_class_cat <- brm(grade~avgView + sync_cont, data=data, prior=bprior, silent = T

brm_class <- brm(grade~avgView + sync, data=data, prior=bprior, silent = T)

posteriors <- get_parameters(brm_class_cat)
```

effect	component	group	term	estimate	std.error	conf.low	conf.high
fixed	cond	NA	(Intercept)	49.528	8.644	32.260	65.776
fixed	cond	NA	avgView	0.465	0.129	0.207	0.725
fixed	cond	NA	sync_cont	-0.089	10.096	-20.170	19.671
ran_pars	cond	Residual	sd__Observation	16.380	2.431	12.519	21.954

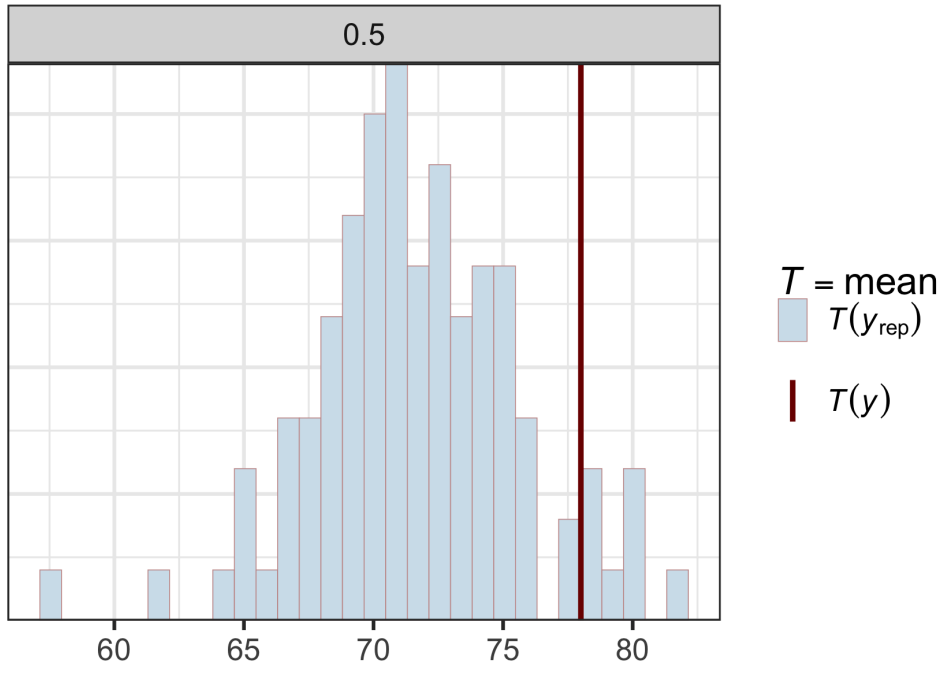
Visualizing the posterior distribution

```
ggplot(posteriors, aes(x = b_sync_cont)) +
  geom_density(fill = "orange")
```



Posterior distribution plot

```
pp_check(brm_class_cat, type="stat_grouped", group="sync_cont", ndraws = 100)
```



Point-estimate

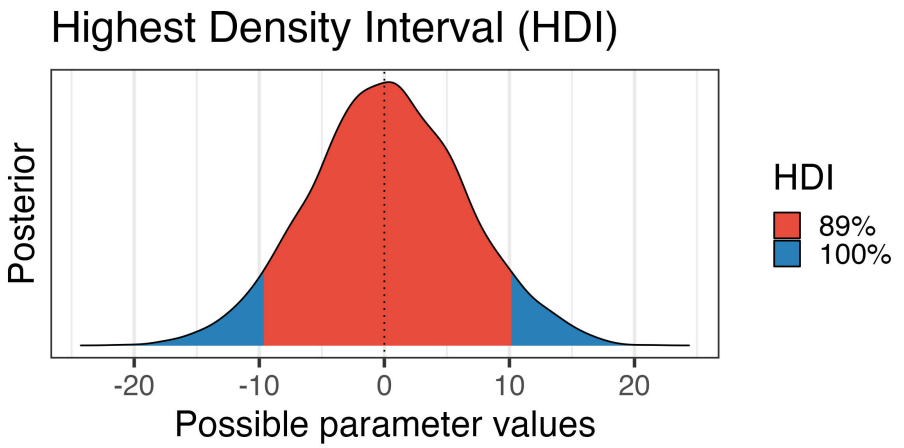
```
describe_posterior(
  brm_class_cat,
  effects = "fixed",
  component = "all",
  centrality = "all"
)
```

```
## Summary of Posterior Distribution
##
## Parameter   | Median |  Mean |   MAP |          95% CI |     pd |          ROPE | % in RO
## ------------------------------------------------------------------------------------
## (Intercept) |  49.65 | 49.53 | 49.13 | [ 32.26, 65.78] |   100% | [-1.80, 1.80] |
## avgView     |   0.46 |  0.47 |  0.46 | [  0.21,  0.72] | 99.98% | [-1.80, 1.80] |      10
## sync_cont   |  -0.15 | -0.09 | -0.40 | [-20.17, 19.67] | 50.52% | [-1.80, 1.80] |    16.2
##
## # Fixed effects sigma
##
## Parameter | Median |  Mean |   MAP |          95% CI |   pd |          ROPE | % in ROPE |
## ------------------------------------------------------------------------------------
## sigma     |  16.10 | 16.38 | 15.81 | [ 12.52, 21.95] | 100% | [-1.80, 1.80] |        0% |
```

Uncertainty: Credible intervals

- Credible intervals (HDI)

```
library(see)

results=hdi(posteriors$b_sync_cont, ci
```



Significance

- Does the credible interval contain 0?
    - If yes, "not significant"
    - If no, "significant"

Significant differences

- Use **emmeans** to get mean differences between variables

```
em_syn <- emmeans(brm_class, ~sync) %>%
  pairs() %>%
  kable("html")

em_syn
```

contrast	estimate	lower.HPD	upper.HPD
Async - Sync	-0.3179415	-14.04305	11.93909

Significant differences

- Is the effect 0?

```
library(bayestestR) # bayes functions easystats
# contrast
#BF only if you use weakly-strong priors
BF <- bayestestR::bayesfactor_parameters(brm_class_cat, null = 0)

BF
```

```
## Bayes Factor (Savage-Dickey density ratio)
##
## Parameter   |     BF
## --------------------
## (Intercept) | 438.99
## avgView     |   5.94
## sync_cont   |  1.000
##
## * Evidence Against The Null: 0
```

Model comparisons

- Use **`bayestestR::bayesfactor_models`** to get a BF for model selection

```
# Model 1: grade ~ sync + avgView
#save_pars for bayes factors
brm_class1 <- brm(grade~avgView + sync, data=data , family = gaussian(),prior=bpri
#grade ~ avgView
brm_class2 <- brm(grade~avgView, data=data, prior=bprior, family = gaussian(), sam
    warmup = 2000, iter = 5000)
```

```
# testing models
# compared to intercept-only or null model
bayesfactor_models(brm_class1, brm_class2)
```

```
## Bayes Factors for Model Comparison
##
##     Model     BF
## [2] avgView 1.55
##
## * Against Denominator: [1] avgView + sync
## *   Bayes Factor Type: marginal likelihoods (bridgesampling)
```

Original question

- Do my students' course grades depend on whether they attend lectures synchronously or asynchronously?
  - Maybe?
  - BF for model comparison suggests equivocal evidence ($BF_{avgtime vs.syn}$ = 1.56)
  - Average viewing time does matter
  - Moderate evidence that effect not zero (BF = 3.47)
- What do we get from Bayesian analysis that we don't get from regular linear regression?

Reporting bayesian analysis

```
report_bayes=report::report(brm_class1)
```

We fitted a Bayesian linear model (estimated using MCMC sampling with 4 chains of 5000 iterations and a warmup of 2000) to predict grade with avgView and sync (formula: grade ~ avgView + sync). Priors over parameters were set as normal (mean = 0.00, SD = 10.00) distributions. The model's explanatory power is substantial (R2 = 0.34, 95% CI [0.10, 0.53], adj. R2 = 0.10). Within this model:

- The effect of b Intercept (Median = 49.58, 95% CI [33.65, 63.64]) has a 100.00% probability of being positive (> 0), 100.00% of being significant (> 0.90), and 100.00% of being large (> 5.41). The estimation successfully converged (Rhat = 1.000) and the indices are reliable (ESS = 8658)
- The effect of b avgView (Median = 0.46, 95% CI [0.15, 0.77]) has a 99.64% probability of being positive (> 0), 0.39% of being significant (> 0.90), and 0.00% of being large (> 5.41). The estimation successfully converged (Rhat = 1.000) and the indices are reliable (ESS = 9063)

Bayesian pros

- Evidence can be gathered in favor of a hypothesis (the null)
- Quantify the amount of support for one hypothesis relative to another
- Parsimony is rewarded
- Sample size does not affect estimates as much as it does the likelihood
- Optional stopping is okay

Bayesian cons:

- Priors
- Computationaly intensive

Caveat: What can Bayes not do?

- Ban questionable research practices (e.g., HARKing)
- Provide a remedy for:
  - Small sample sizes
  - Unrepresentative samples
  - Poor experimental design