Probabilistic Modeling

Bayesian Statistics, Maximum likelihood, and all that

Diogo Melo

What does probability mean?

The many faces of probability

Frequentist probabilities

- The probability of an event as the limit of its relative frequency in a large number of trials
- Probabilities describe objective properties of the physical world
- Probabilities only apply to repeated events

Cox probabilities

- Probability is an extension of formal logic, and allows us to discuss events about which we lack complete information
- Probabilities describe states of knowledge, not real properties of the world
- Probabilities apply to any statement for which we lack complete information

Probabilities as extensions of logic

A implies B

What does B say about A?

A: it is raining

B: There are clouds

If A is true, then B is true

If B is true, my assessment about the plausibility of A should change?

How much should it change?

Probabilities as extensions of logic

A implies B

What does B say about A?

A: it is raining

B: There are clouds

If A is true, then B is true

If B is true, my assessment about the plausibility of A should change?

How much should it change?

Cox showed that the correct tool for this type of reasoning was probabilities!

Likelihood of an observation

- The Likelihood asks us to imagine a possible world, and in that world, figure out what could happen
- Think of the likelihood in a story. It is not how plausible it is that it really happened in the real world, but how consistent the events of the story are with the world in which it happens

$$P(\text{Events} \mid \text{Assumed World})$$

- Mathematically, the Likelihood is just a count of how many ways something could happen under some hypothesis about the world
- The more ways to produce a data set under a hypothesis, the higher the likelihood attributed to a data set under that hypothesis

$$P(y \mid \theta)$$

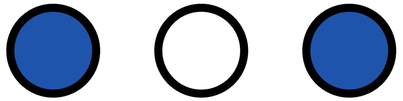
Likelihood as counting

- An evil statistician created an urn that contain 4 white or blue balls

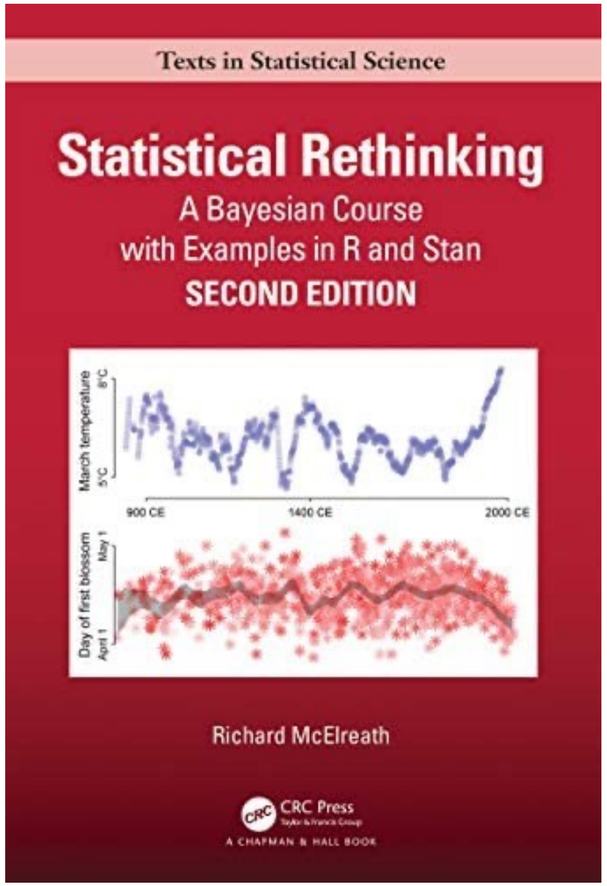
Five possible configurations (worlds)

(1) [○○○○], (2) [●○○○], (3) [●●○○], (4) [●●●○], (5) [●●●●]

- We observe 3 observations with replacement:



Example from:



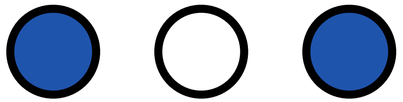
Likelihood as counting

- An evil statistician created an urn that contain 4 white or blue balls

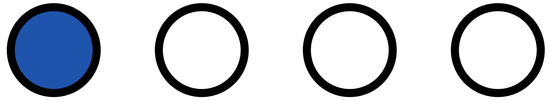
Five possible configurations (worlds)

(1) [○○○○], (2) [●○○○], (3) [●●○○], (4) [●●●○], (5) [●●●●]

- We observe 3 observations with replacement:



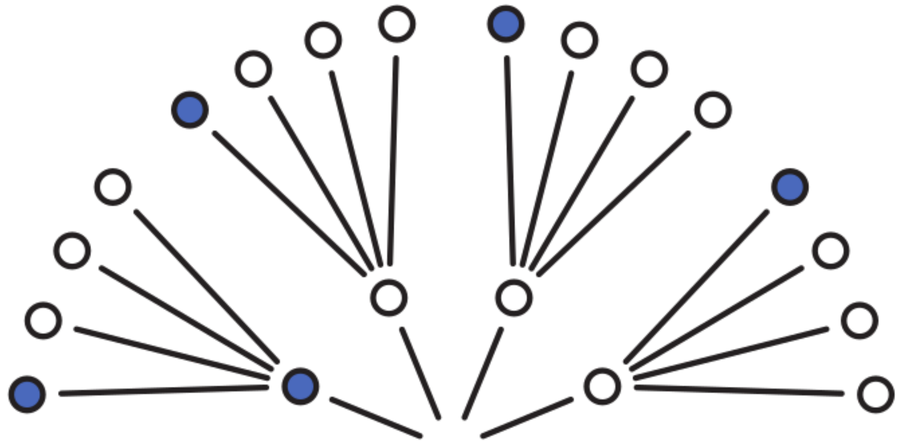
Assume the urn has the composition:



These are the 4 possible outcome of a draw:

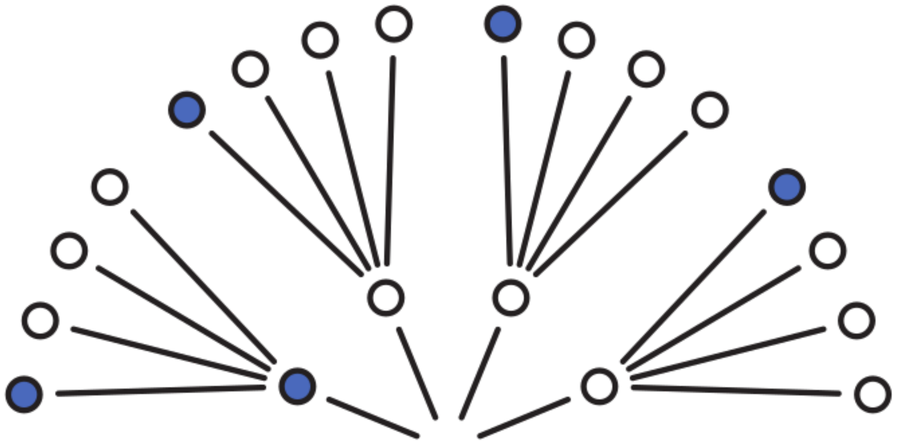
Likelihood as counting

These are the 4^2 possible outcomes of 2 draws:

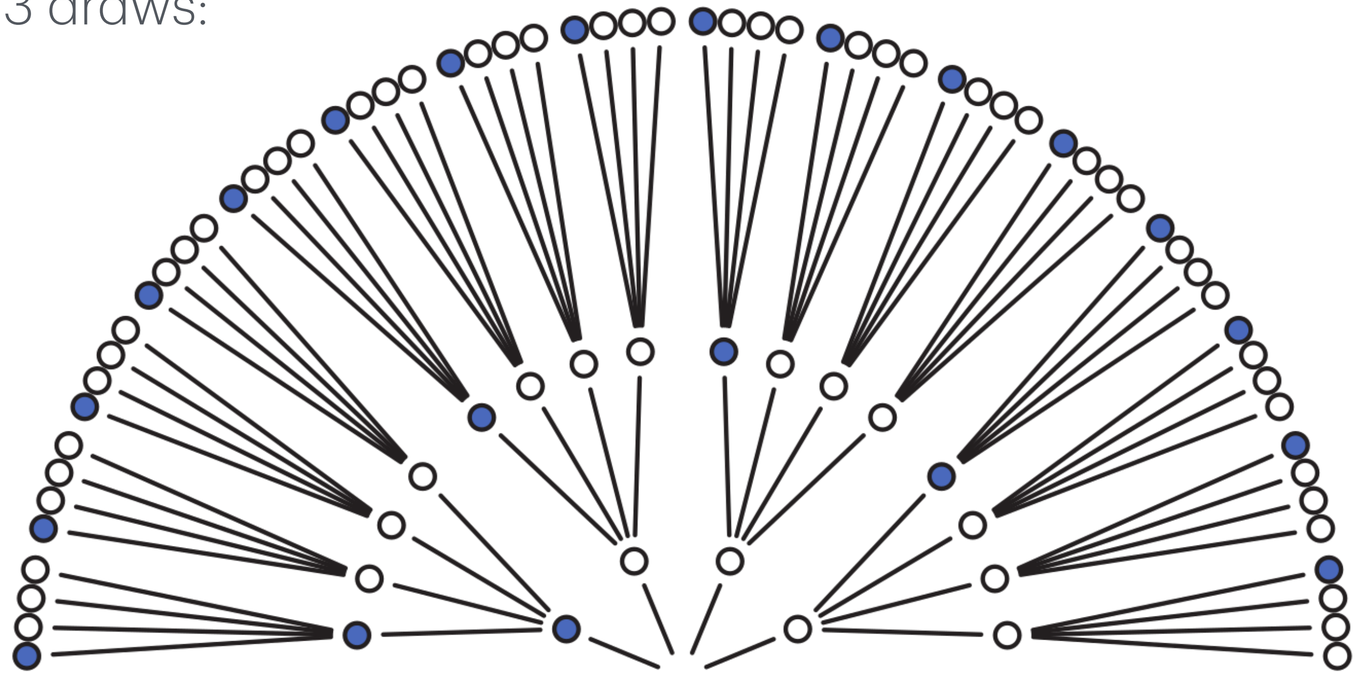


Likelihood as counting

These are the 4^2 possible outcomes of 2 draws:

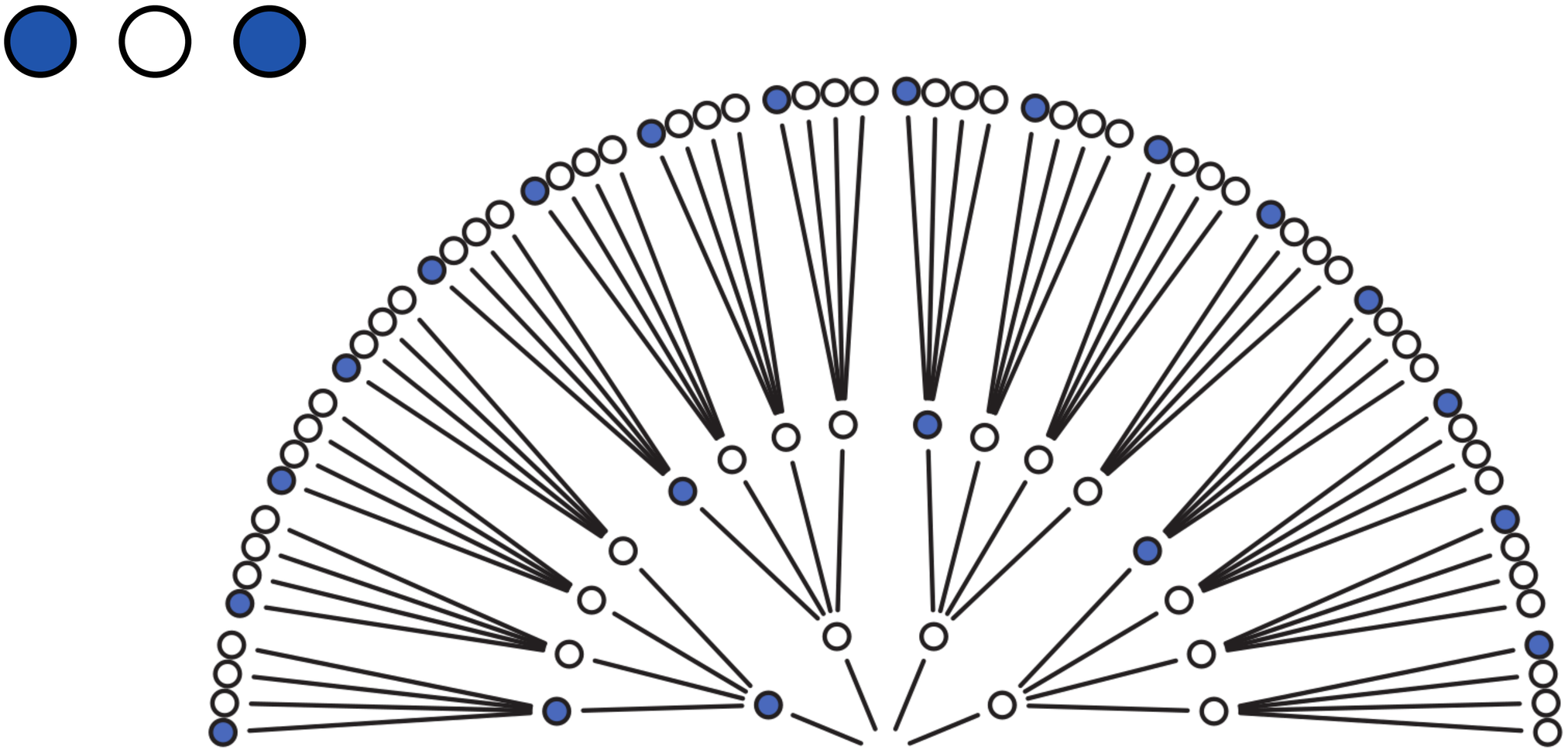


These are the 4^3 possible outcomes of 3 draws:



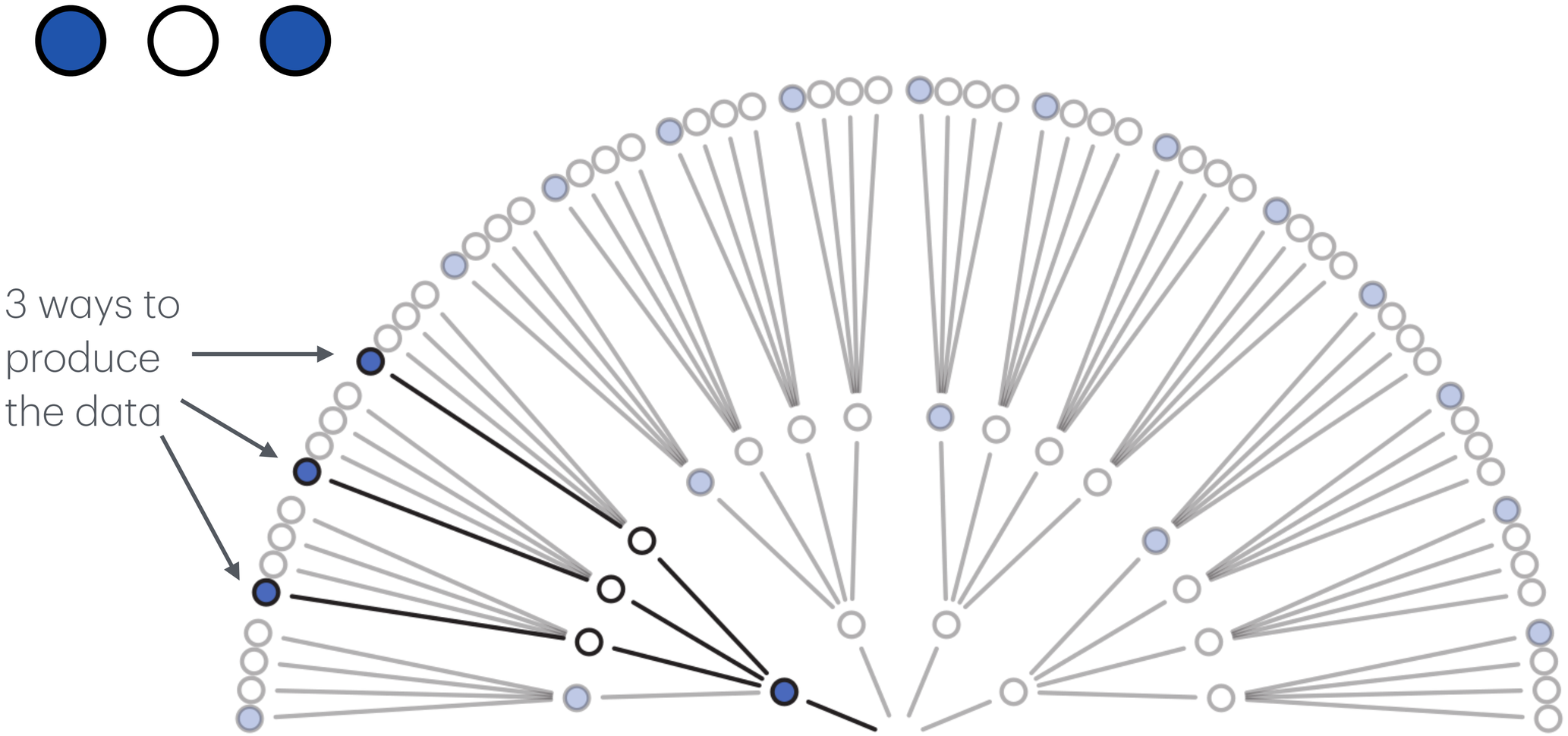
Likelihood as counting

How many of these possible draws are compatible with the data?



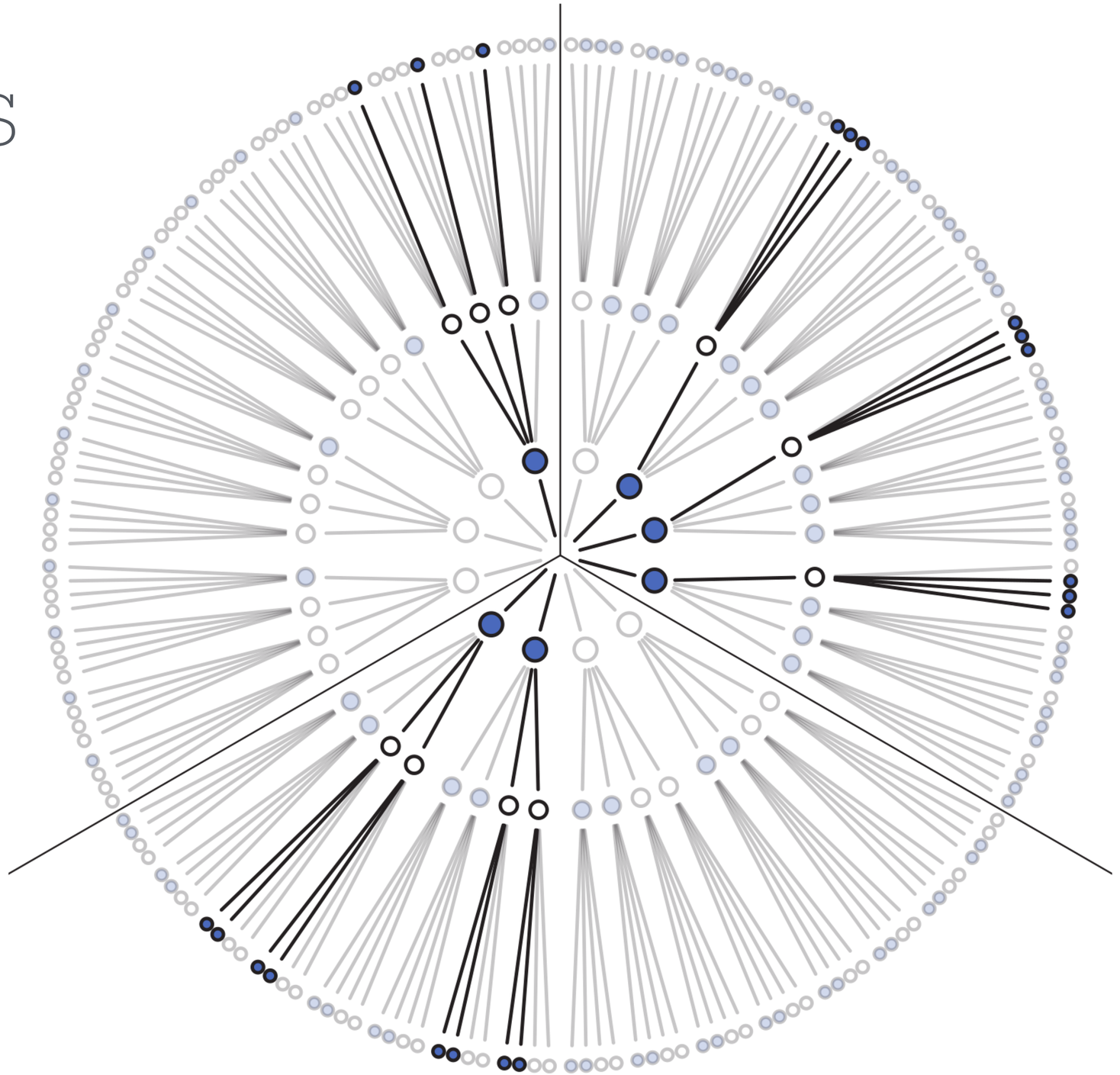
Likelihood as counting

How many of these possible draws are compatible with the data?

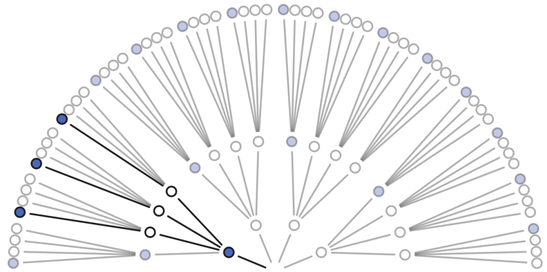


All the possibilities

Conjecture	Ways to produce ●○●
[○○○○]	$0 \times 4 \times 0 = 0$
[●○○○]	$1 \times 3 \times 1 = 3$
[●●○○]	$2 \times 2 \times 2 = 8$
[●●●○]	$3 \times 1 \times 3 = 9$
[●●●●]	$4 \times 0 \times 4 = 0$



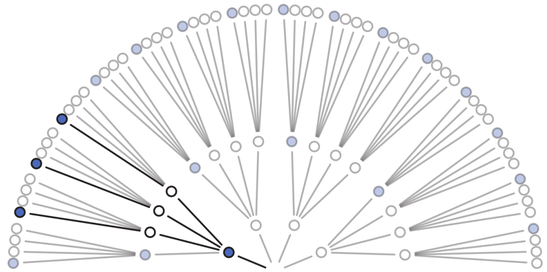
Likelihood is just a scaled measure of these counts



Conjecture	Ways to produce ●○●
[○○○○]	$0 \times 4 \times 0 = 0$
[●○○○]	$1 \times 3 \times 1 = 3$
[●●○○]	$2 \times 2 \times 2 = 8$
[●●●○]	$3 \times 1 \times 3 = 9$
[●●●●]	$4 \times 0 \times 4 = 0$

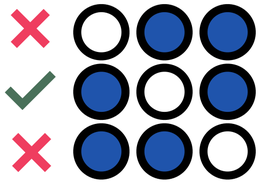
```
> 3/(4^3)
[1] 0.046875
> dbinom(2, 3, 0.25)/3
[1] 0.046875
```

Likelihood is just a scaled measure of these counts



Conjecture	Ways to produce ●○●
[○○○○]	$0 \times 4 \times 0 = 0$
[●○○○]	$1 \times 3 \times 1 = 3$
[●●○○]	$2 \times 2 \times 2 = 8$
[●●●○]	$3 \times 1 \times 3 = 9$
[●●●●]	$4 \times 0 \times 4 = 0$

```
> 3/(4^3)
[1] 0.046875
> dbinom(2, 3, 0.25)/3
[1] 0.046875
```



Likelihood is just a scaled measure of these counts

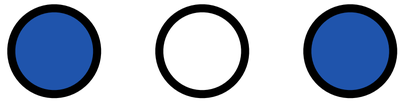
We use parameters to index the conjectures

Possible composition	p	Ways to produce data	Plausibility
[○○○○]	0	0	0
[●○○○]	0.25	3	0.15
[●●○○]	0.5	8	0.40
[●●●○]	0.75	9	0.45
[●●●●]	1	0	0

And scale the plausibilities so we don't work with huge numbers of counts

What if we draw another blue ball?

Prior data:



New observation:

Conjecture	Ways to produce ●	Previous counts	New count
[○○○○]	0	0	$0 \times 0 = 0$
[●○○○]	1	3	$3 \times 1 = 3$
[●●○○]	2	8	$8 \times 2 = 16$
[●●●○]	3	9	$9 \times 3 = 27$
[●●●●]	4	0	$0 \times 4 = 0$

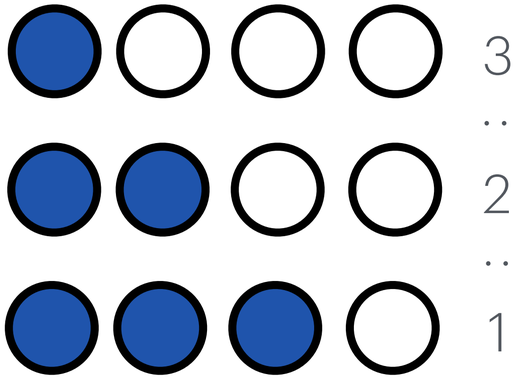
New plausibilities is the product of the number of ways to produce the data under the conjectures

What if we get some information from the urn factory?

Our urn

Data: ● ○ ●

Proportions of urns from the factory:



Conjecture	Prior count	Factory count	New count
[○○○○]	0	0	$0 \times 0 = 0$
[●○○○]	3	3	$3 \times 3 = 9$
[●●○○]	16	2	$16 \times 2 = 32$
[●●●○]	27	1	$27 \times 1 = 27$
[●●●●]	0	0	$0 \times 0 = 0$

New plausibilities is the product of the number of ways to produce the data under the conjectures by the prior information from the factory

Bringing it all together

$$P(\theta = 0.25 \,|\, [B, W, B]) \propto P([B, W, B] \,|\, \theta = 0.25) \times P(\theta = 0.25)$$

Bringing it all together

$$P(\theta = 0.25 \mid [B, W, B]) \propto P([B, W, B] \mid \theta = 0.25) \times P(\theta = 0.25)$$

Likelihood, how many ways to produce the data under the assumed conjecture

Bringing it all together

$$P(\theta = 0.25 \mid [B, W, B]) \propto P([B, W, B] \mid \theta = 0.25) \times P(\theta = 0.25)$$

Likelihood, how many ways to produce the data under the assumed conjecture

Prior, other information about the conjecture not related to the data

Bringing it all together

Posterior probability of the conjeture, considering all the information in the data and the prior

Likelihood, how many ways to produce the data under the assumed conjecture

Prior, other information about the conjecture not related to the data

What is Bayesian Statistics?

ML: What is the parameter value that maximizes the probability of having generated the data:

$$\underset{\theta}{\operatorname{argmax}} \left[P(y \mid \theta) \right]$$

Bayesian: What is the probability distribution of parameter values given the data:

$$P(\theta \mid y) \propto P(\theta) P(y \mid \theta)$$

What is Bayesian Statistics?

ML: What is the parameter value that maximizes the probability of having generated the data:

$$\underset{\theta}{\operatorname{argmax}} \left[P(y \mid \theta) \right]$$

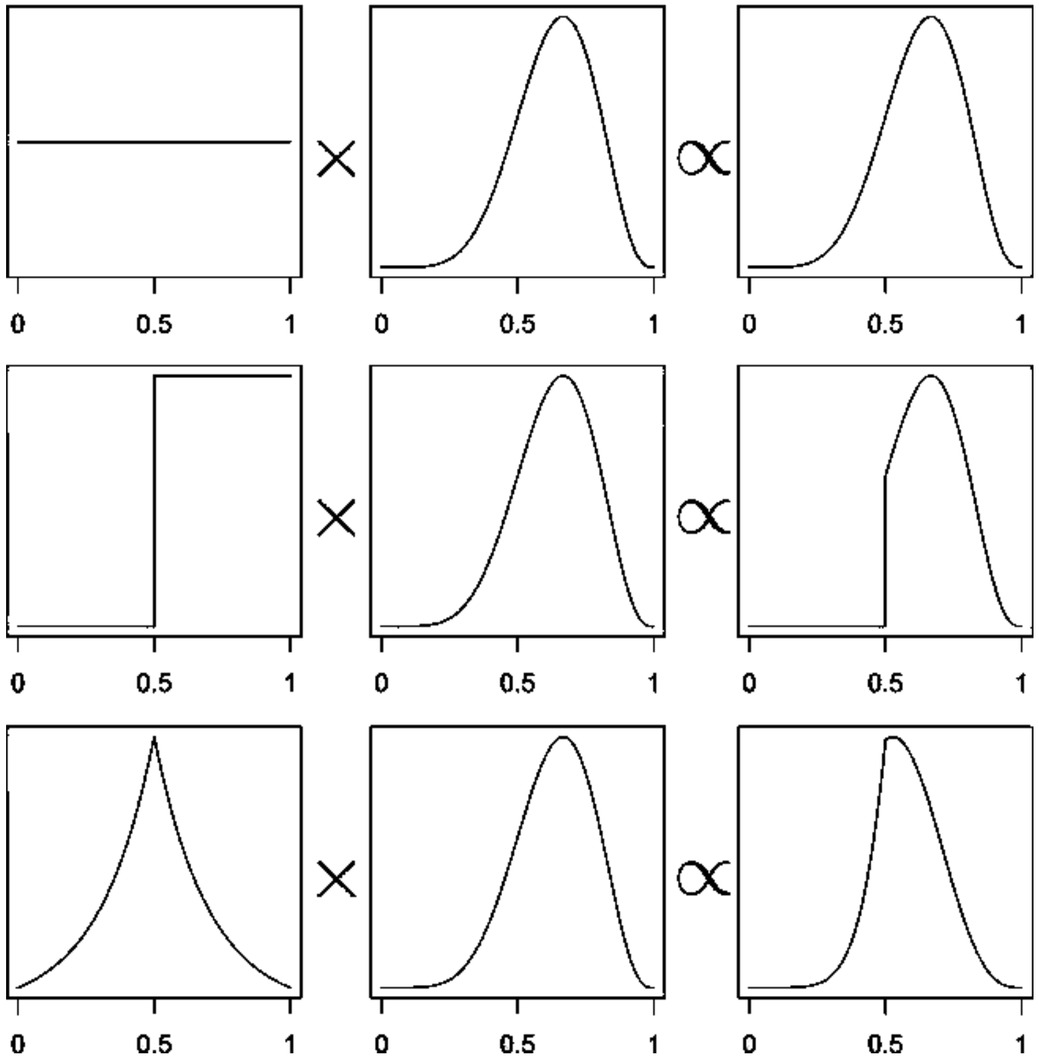
Bayesian: What is the probability distribution of parameter values given the data:

$$P(\theta \mid y) \propto P(\theta) P(y \mid \theta)$$

Attention! No maximization!

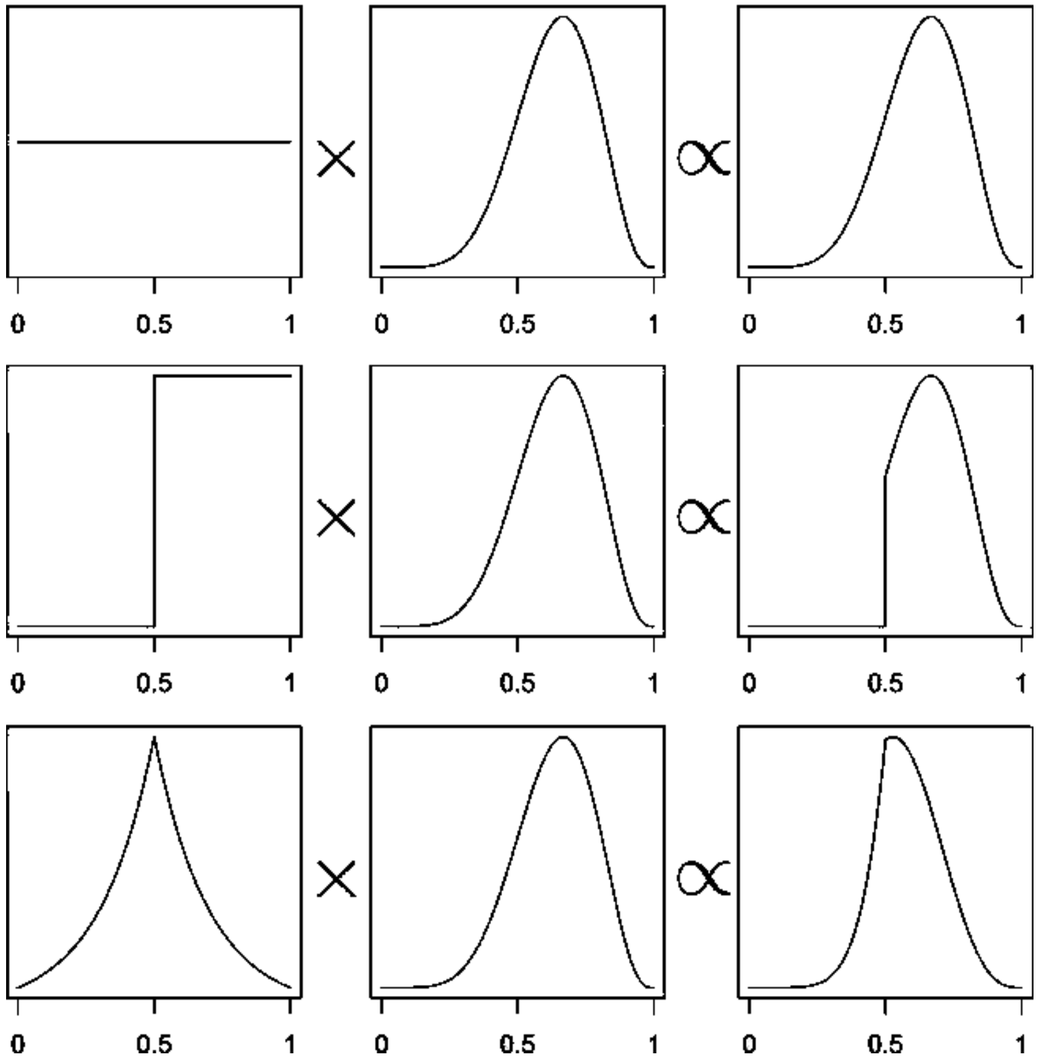
Prior × Likelihood ∝ Posterior

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



Prior × Likelihood ∝ Posterior

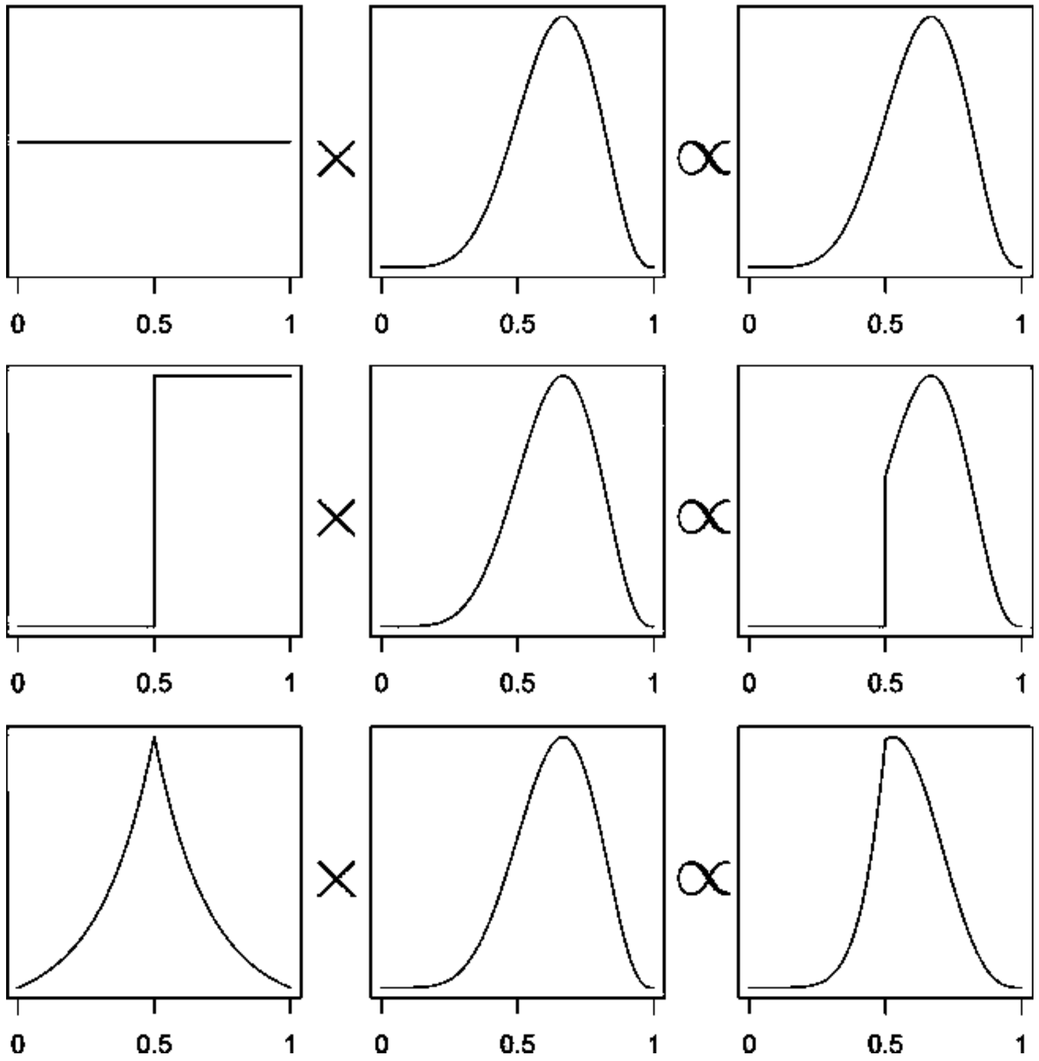
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$



- $P(\theta \mid y)$ represents our knowledge of parameters using probability.

Prior × Likelihood ∝ Posterior

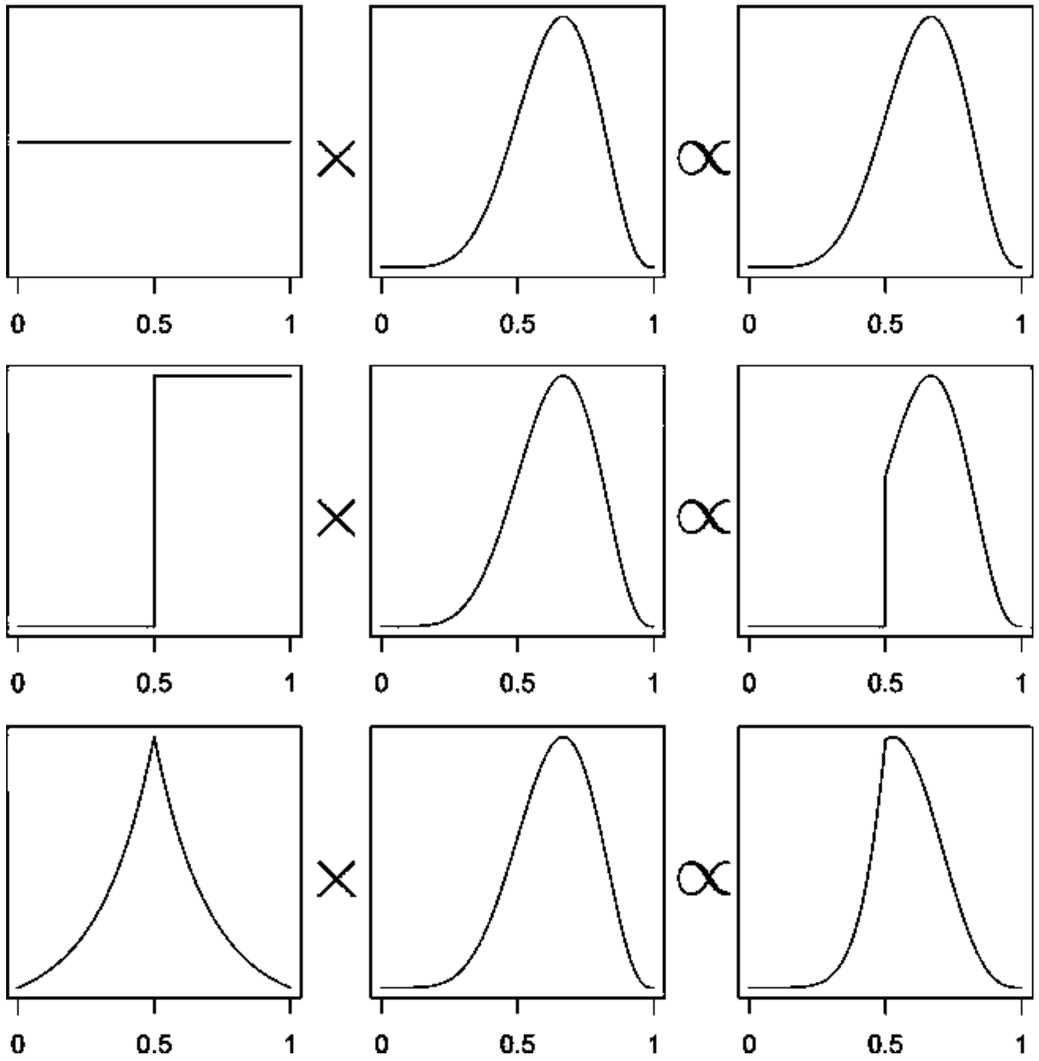
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$



- $P(\theta \mid y)$ represents our knowledge of parameters using probability.
    - this representation fully encapsulates our beliefs.

Prior × Likelihood ∝ Posterior

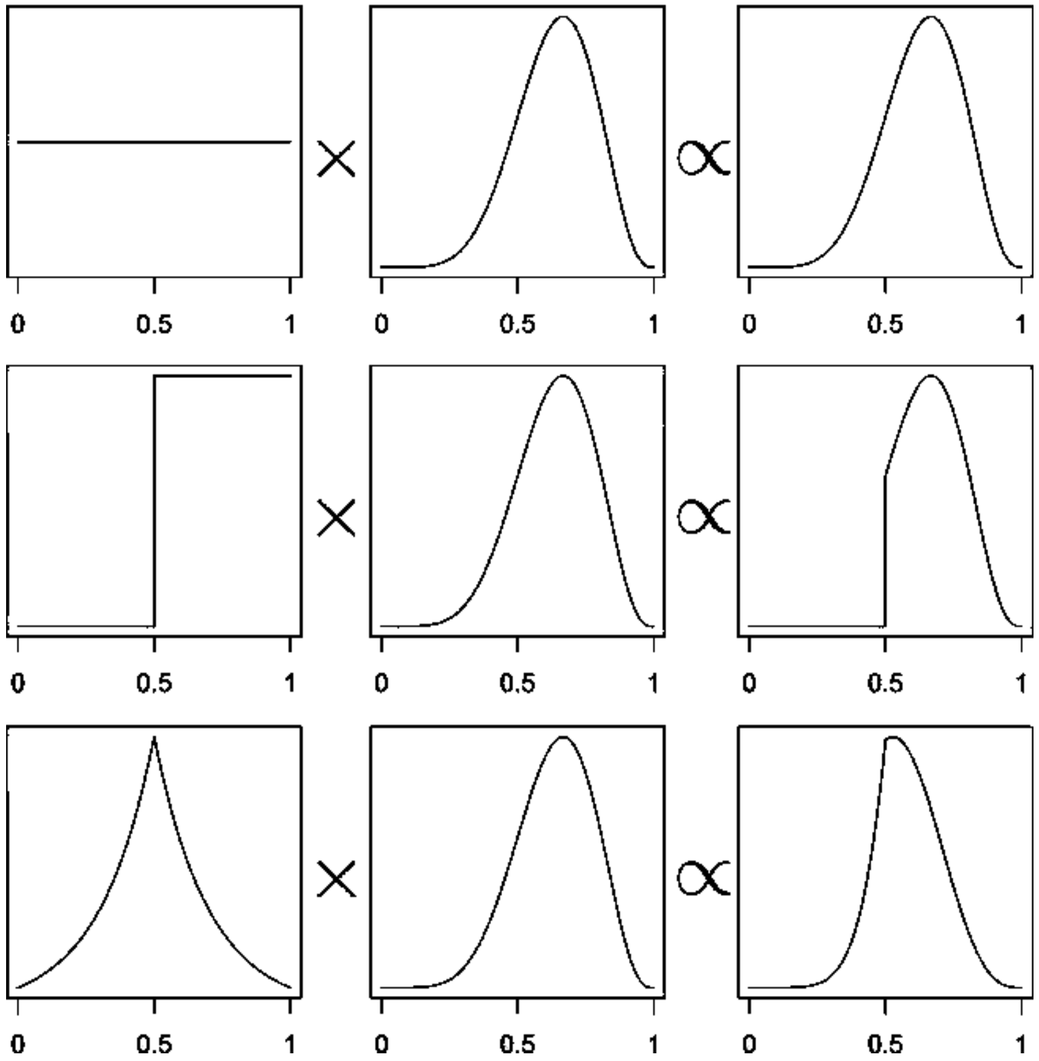
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$



- $P(\theta \mid y)$ represents our knowledge of parameters using probability.
  - this representation fully encapsulates our beliefs.
  - Includes all the uncertainty

Prior × Likelihood ∝ Posterior

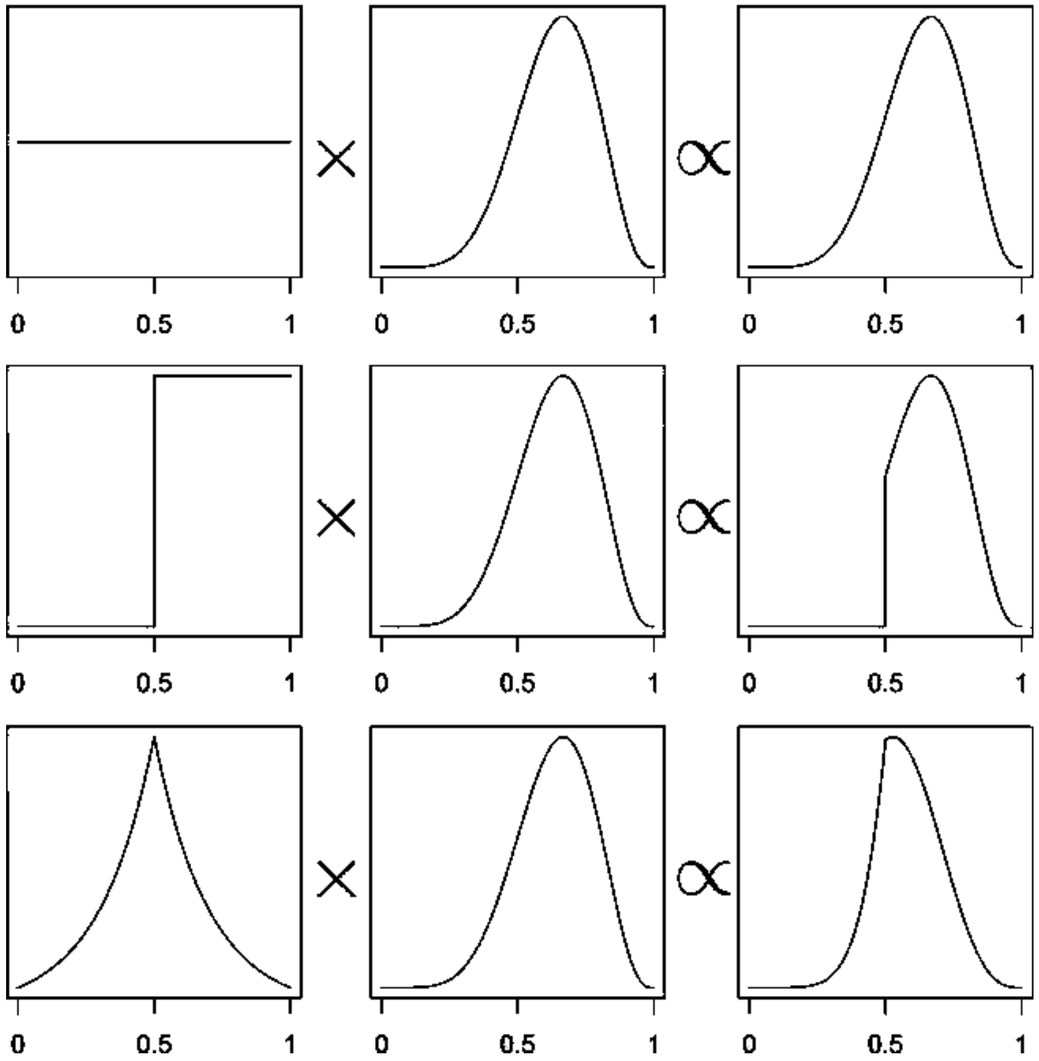
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$



- $P(\theta \mid y)$ represents our knowledge of parameters using probability.
  - this representation fully encapsulates our beliefs.
  - Includes all the uncertainty
- $P(\theta)$, the prior, can encode useful information:

Prior × Likelihood ∝ Posterior

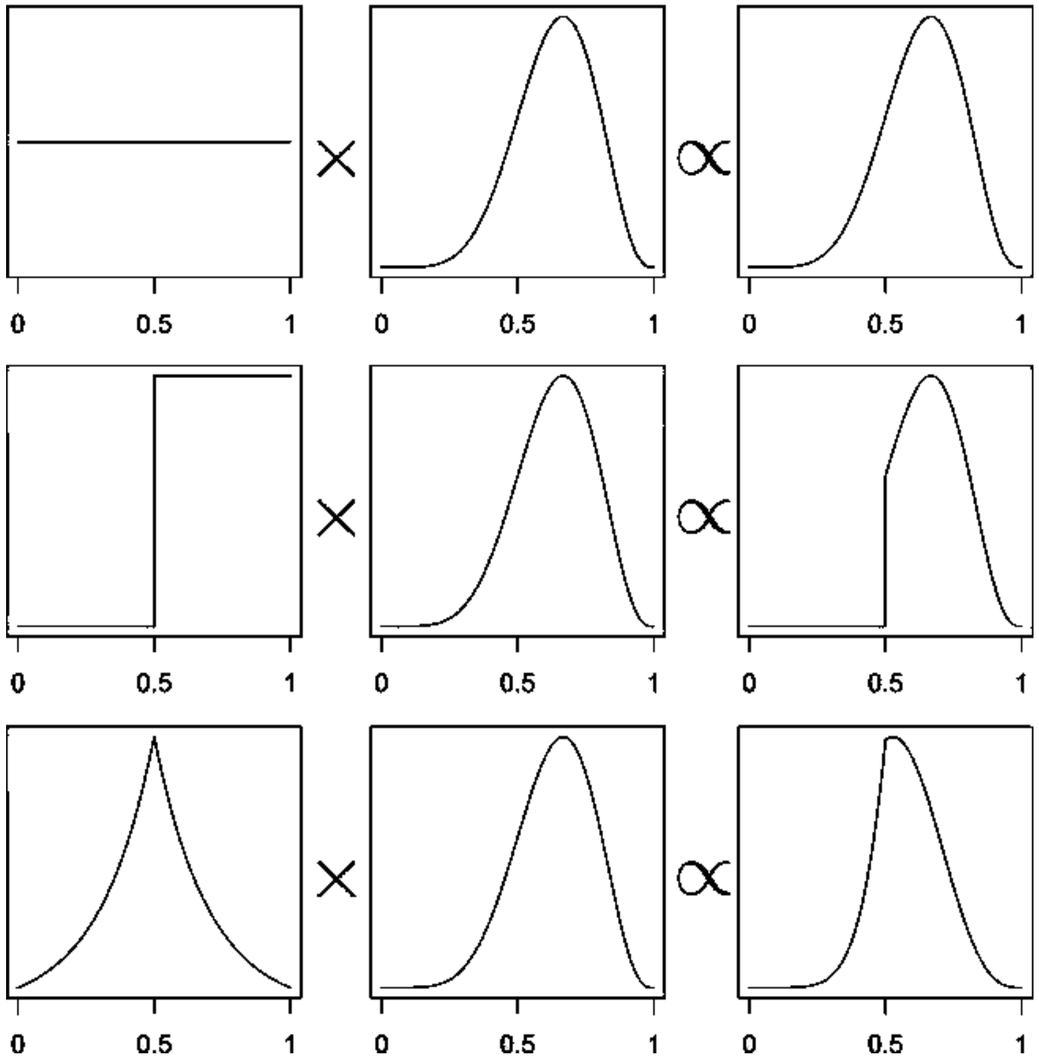
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$



- $P(\theta \mid y)$ represents our knowledge of parameters using probability.
  - this representation fully encapsulates our beliefs.
  - Includes all the uncertainty
- $P(\theta)$, the prior, can encode useful information:
  - parameter scale, shared structure, permitted values…

Prior × Likelihood ∝ Posterior

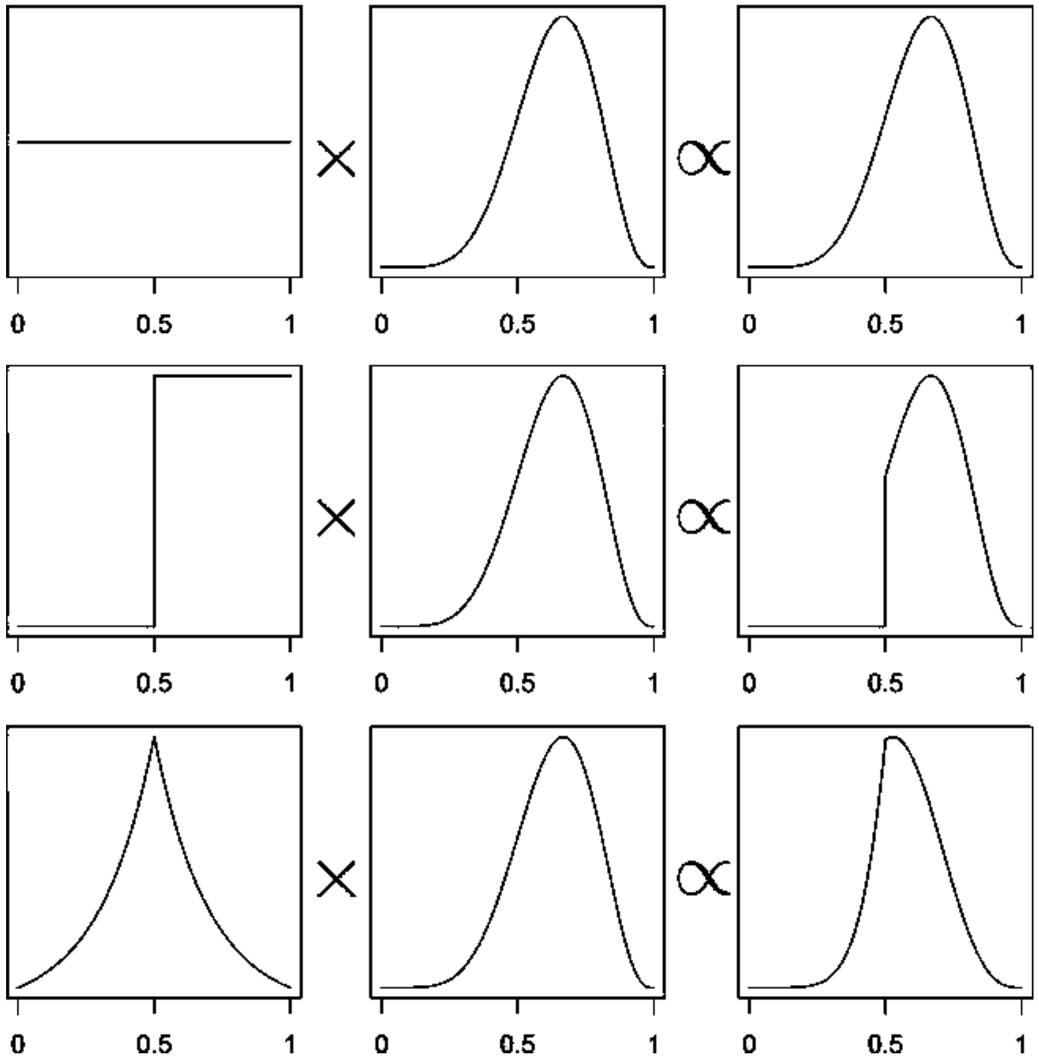
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$



- $P(\theta \mid y)$ represents our knowledge of parameters using probability.
  - this representation fully encapsulates our beliefs.
  - Includes all the uncertainty
- $P(\theta)$, the prior, can encode useful information:
  - parameter scale, shared structure, permitted values...
- Isn't the MLE the best estimator? (depends on the criteria...)

Prior × Likelihood ∝ Posterior

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



- $P(\theta \mid y)$ represents our knowledge of parameters using probability.
    - this representation fully encapsulates our beliefs.
    - Includes all the uncertainty
- $P(\theta)$, the prior, can encode useful information:
    - parameter scale, shared structure, permitted values...
- Isn't the MLE the best estimator? (depends on the criteria...)
    - Sometimes... but not $\rho = f(\hat{\theta})$

Using the posterior

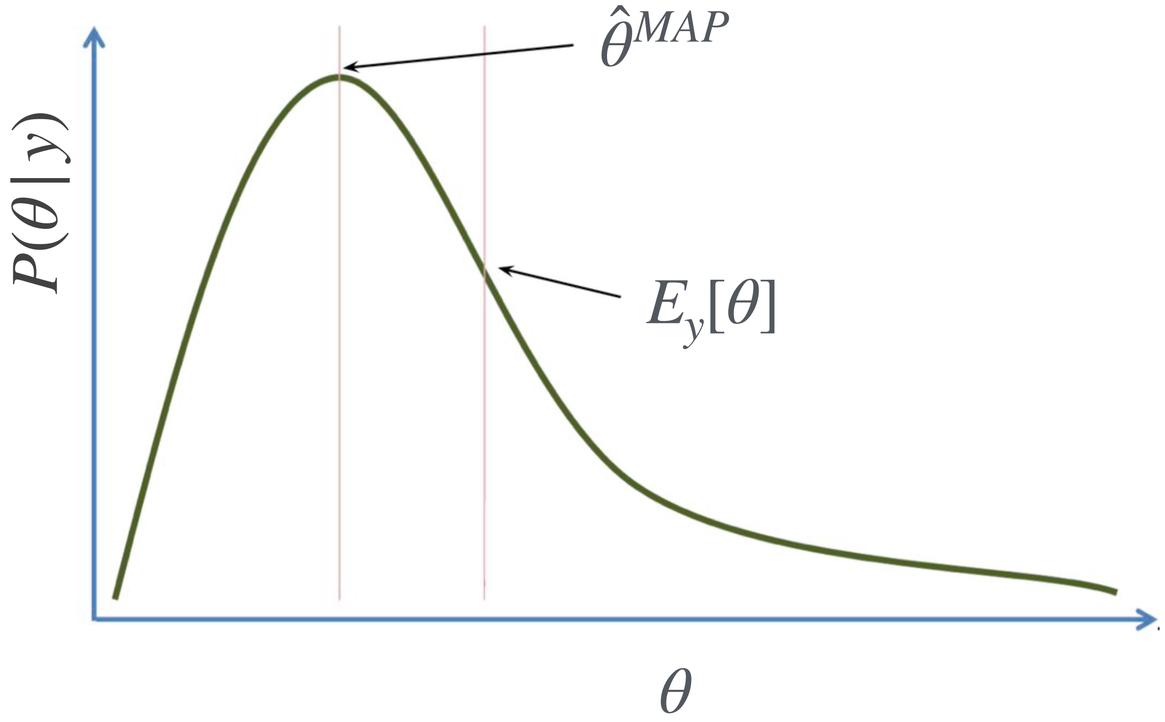
Posterior estimators

- Bayesian equivalent to MLE is the **M**aximum **A** **P**osteriori (MAP):

$$\hat{\theta}^{MAP} = \underset{\theta \in \Omega}{\operatorname{argmax}} \left[P(\theta \mid y) \right]$$

- The posterior mean is more common:

$$E_y[\theta] = \sum_{\theta \in \Omega} \theta P(\theta \mid y)$$

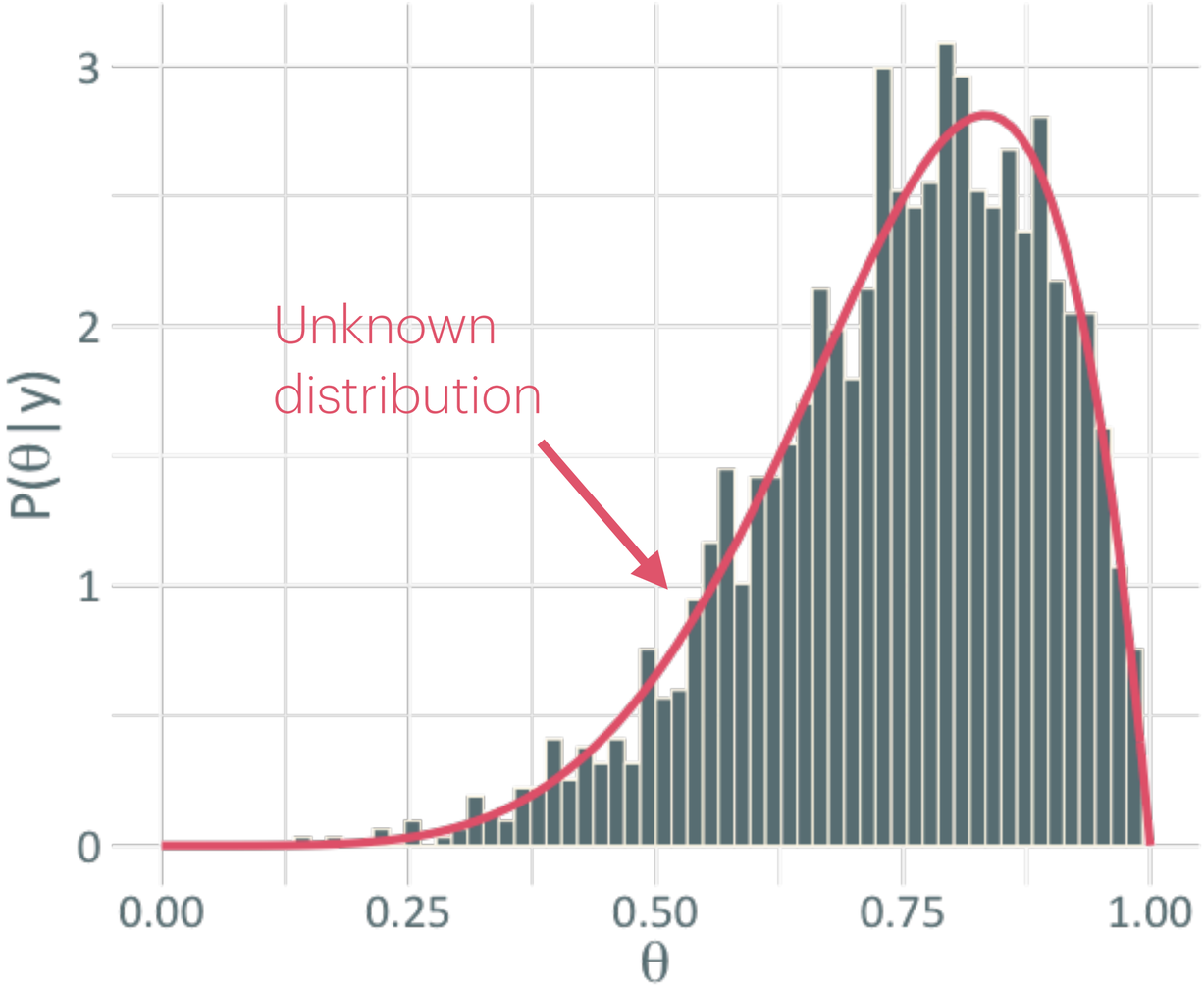


(Posterior median is also used occasionally.)

Posterior Approximations

- For a small number of models we can write the posterior distribution directly (really small, don't bother).
- For most models, we use posterior samples to approximate the posterior.

$$\{\theta_1, \cdots, \theta_N\} \sim P(\theta \,|\, y)$$



Posterior derived quantities

- This sample can be used to calculate any quantity of interest:

$$\{\theta_1, \cdots, \theta_N\} \sim P(\theta \mid y)$$

For example, the posterior mean is just:

$$\frac{\theta_1 + \theta_2 + \cdots + \theta_N}{N} \approx \sum_{\theta \in \Omega} \theta P(\theta \mid y)$$

Posterior derived quantities

- This sample can be used to calculate any quantity of interest:

$$\{\theta_1, \cdots, \theta_N\} \sim P(\theta \mid y)$$

For example, the posterior mean is just:

$$\frac{\theta_1 + \theta_2 + \cdots + \theta_N}{N} \approx \sum_{\theta \in \Omega} \theta P(\theta \mid y)$$

Other quantities

- Any other functions of the parameters can be estimated from the samples.
- A common use is to calculate contrast between categorical levels, estimating the difference between groups.
- Quantiles, values above a value, confidence intervals...

Building a model

Out usual regression model

- Given the matched pairs:

$$(x, y) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$$

- Define a likelihood:

$$\left.\begin{aligned} y_i &\sim Normal(\mu_i, \sigma) \\ \mu_i &= \alpha + \beta x_i \end{aligned}\right\} P(y \mid \theta)$$

Building a model

Out usual regression model

- Given the matched pairs:

$$(x, y) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$$

- Define a likelihood:

$$\left.\begin{aligned} y_i &\sim Normal(\mu_i, \sigma) \\ \mu_i &= \alpha + \beta x_i \end{aligned}\right\} P(y \mid \theta)$$

- And a set of priors on the parameters:

$$\begin{aligned} \alpha &\sim P(\alpha) \\ \beta &\sim P(\beta) \\ \sigma &\sim P(\sigma) \end{aligned}$$

Building a model

Out usual regression model

- Given the matched pairs:

$$(x, y) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$$

- Define a likelihood:

$$\left.\begin{aligned} y_i &\sim Normal(\mu_i, \sigma) \\ \mu_i &= \alpha + \beta x_i \end{aligned}\right\} P(y \mid \theta)$$

- And a set of priors on the parameters:

$$\left.\begin{aligned} \alpha &\sim P(\alpha) \\ \beta &\sim P(\beta) \\ \sigma &\sim P(\sigma) \end{aligned}\right\} P(\theta)$$

How do we choose the priors?!

$$\alpha \sim P(\alpha)$$
$$\beta \sim P(\beta)$$
$$\sigma \sim P(\sigma)$$

How do we choose the priors?!

$$\alpha \sim P(\alpha)$$
$$\beta \sim P(\beta)$$
$$\sigma \sim P(\sigma)$$

- Agnostic choices

How do we choose the priors?!

$$\alpha \sim P(\alpha)$$
$$\beta \sim P(\beta)$$
$$\sigma \sim P(\sigma)$$

- Agnostic choices
  - Laplace and the Principle of indifference

How do we choose the priors?!

$$\alpha \sim P(\alpha)$$
$$\beta \sim P(\beta)$$
$$\sigma \sim P(\sigma)$$

- Agnostic choices
  - Laplace and the Principle of indifference
  - "Uninformative" priors

How do we choose the priors?!

$$\alpha \sim P(\alpha)$$
$$\beta \sim P(\beta)$$
$$\sigma \sim P(\sigma)$$

- Agnostic choices
  - Laplace and the Principle of indifference
  - "Uninformative" priors
- Maximum entropy priors

How do we choose the priors?!

$$\alpha \sim P(\alpha)$$
$$\beta \sim P(\beta)$$
$$\sigma \sim P(\sigma)$$

- Agnostic choices
  - Laplace and the Principle of indifference
  - "Uninformative" priors
- Maximum entropy priors
  - priors that encode the least amount of information given constraints

How do we choose the priors?!

$$\alpha \sim P(\alpha)$$
$$\beta \sim P(\beta)$$
$$\sigma \sim P(\sigma)$$

- Agnostic choices
  - Laplace and the Principle of indifference
  - "Uninformative" priors
- Maximum entropy priors
  - priors that encode the least amount of information given constraints
- Jeffreys priors

How do we choose the priors?!

$$\alpha \sim P(\alpha)$$
$$\beta \sim P(\beta)$$
$$\sigma \sim P(\sigma)$$

- Agnostic choices
  - Laplace and the Principle of indifference
  - "Uninformative" priors
- Maximum entropy priors
  - priors that encode the least amount of information given constraints
- Jeffreys priors
  - invariant under a change of coordinates

How do we choose the priors?!

$$\alpha \sim P(\alpha)$$
$$\beta \sim P(\beta)$$
$$\sigma \sim P(\sigma)$$

- Agnostic choices
  - Laplace and the Principle of indifference
  - "Uninformative" priors
- Maximum entropy priors
  - priors that encode the least amount of information given constraints
- Jeffreys priors
  - invariant under a change of coordinates
- Hard constraints

How do we choose the priors?!

$$\alpha \sim P(\alpha)$$
$$\beta \sim P(\beta)$$
$$\sigma \sim P(\sigma)$$

- Agnostic choices
  - Laplace and the Principle of indifference
  - "Uninformative" priors
- Maximum entropy priors
  - priors that encode the least amount of information given constraints
- Jeffreys priors
  - invariant under a change of coordinates
- Hard constraints
  - restricted domains (e.g. variance must be positive)

How do we choose the priors?!

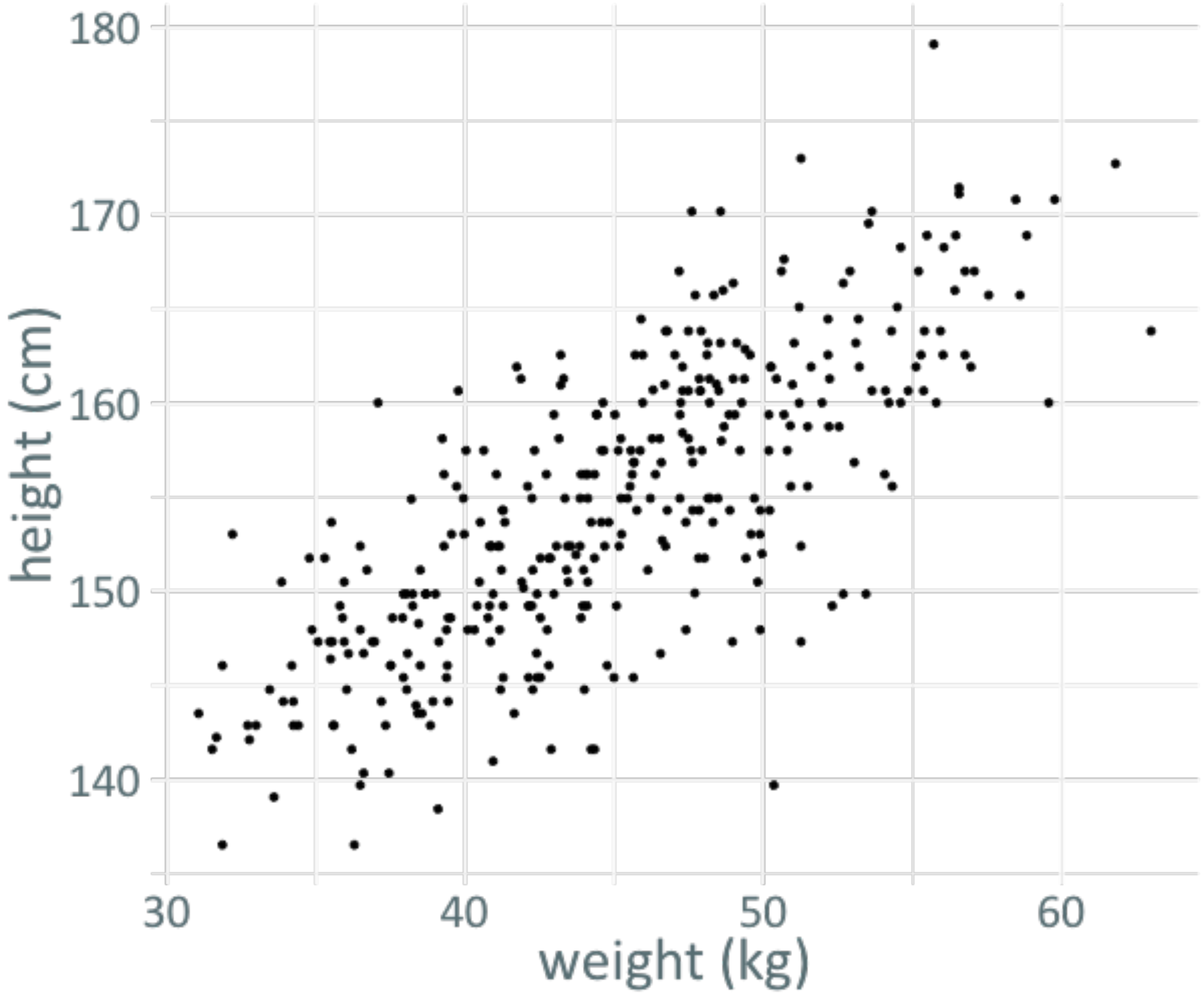
$$\alpha \sim P(\alpha)$$
$$\beta \sim P(\beta)$$
$$\sigma \sim P(\sigma)$$

- Agnostic choices
  - Laplace and the Principle of indifference
  - "Uninformative" priors
- Maximum entropy priors
  - priors that encode the least amount of information given constraints
- Jeffreys priors
  - invariant under a change of coordinates
- Hard constraints
  - restricted domains (e.g. variance must be positive)

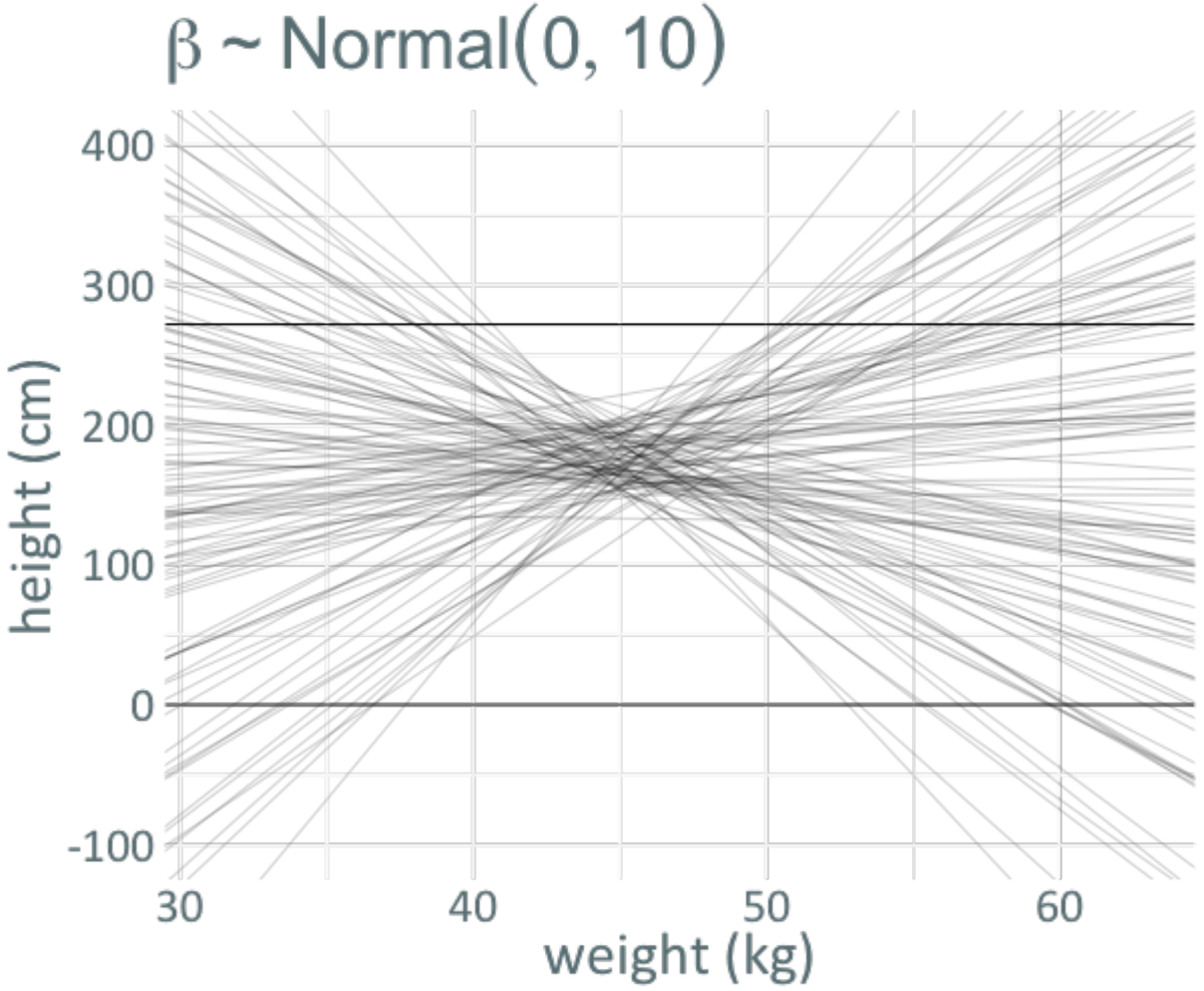
Good prior choices:

- Use domain expertise!
- Knowledge of scale (height by weight example)
- Experimental design (more in the hierarchical models class)
- Using simulations to understand the implications of priors

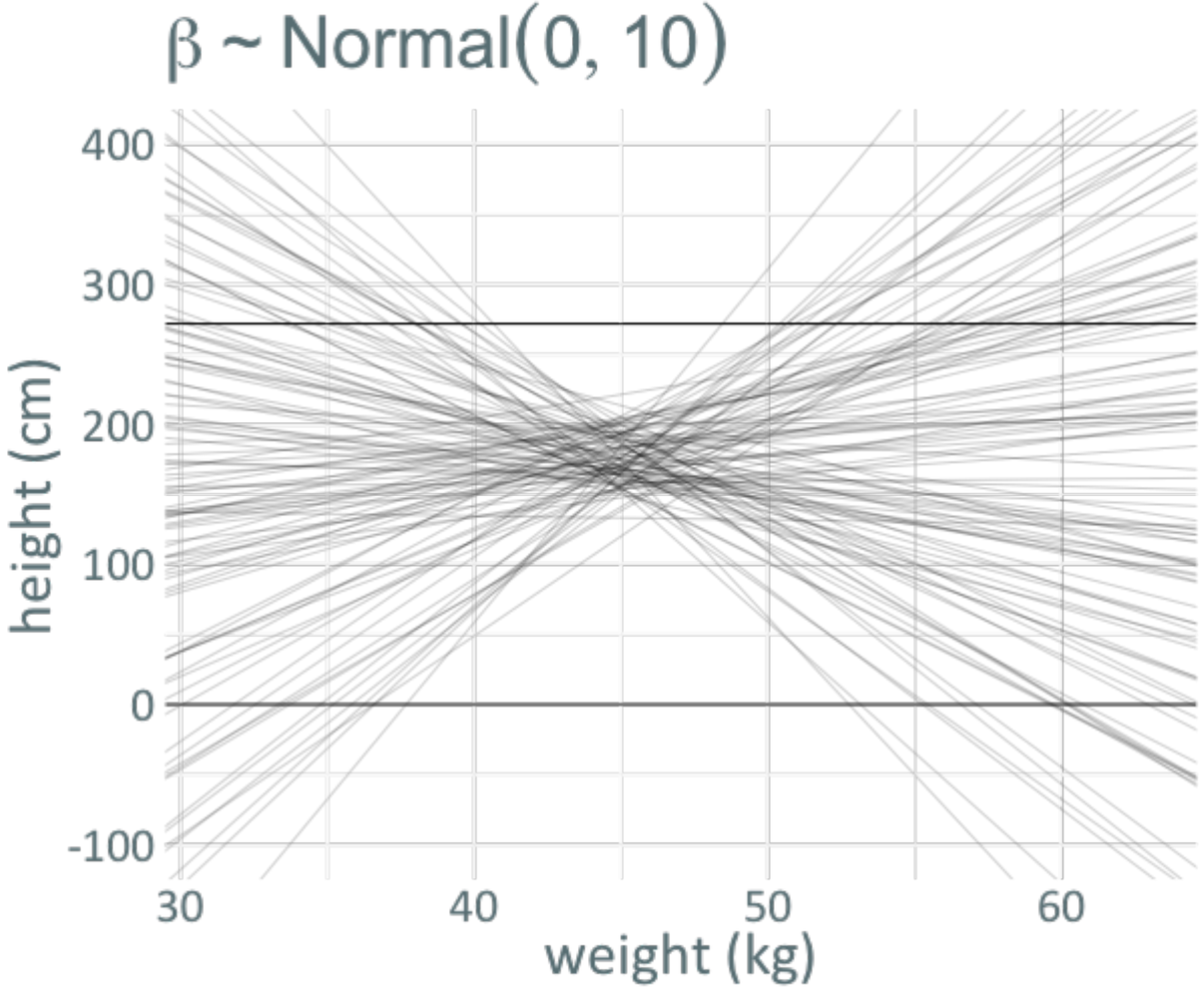
Priors can be used to encode scale information



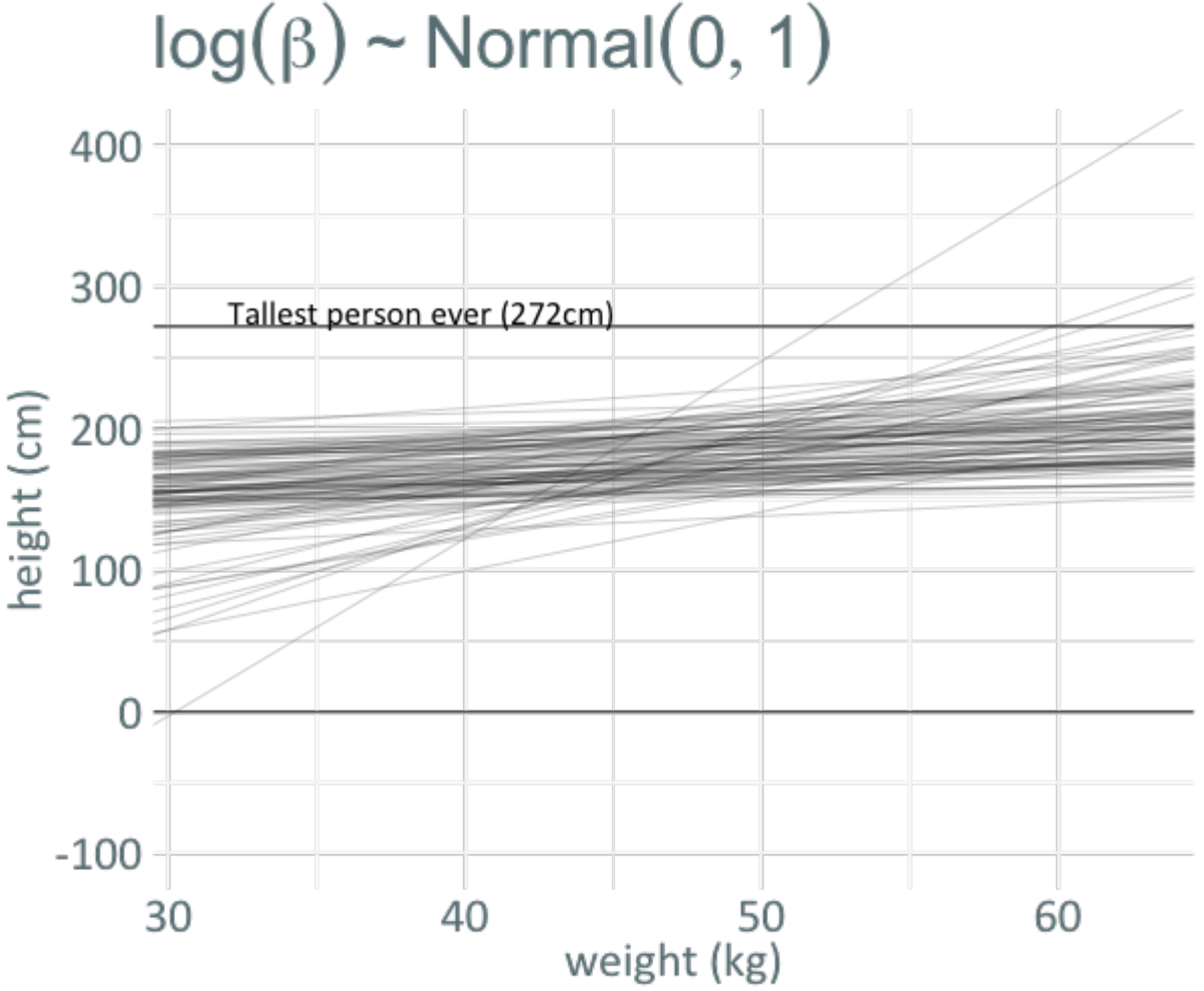
Adapted from Statistical Rethinking



This is sometimes called a non-informative prior



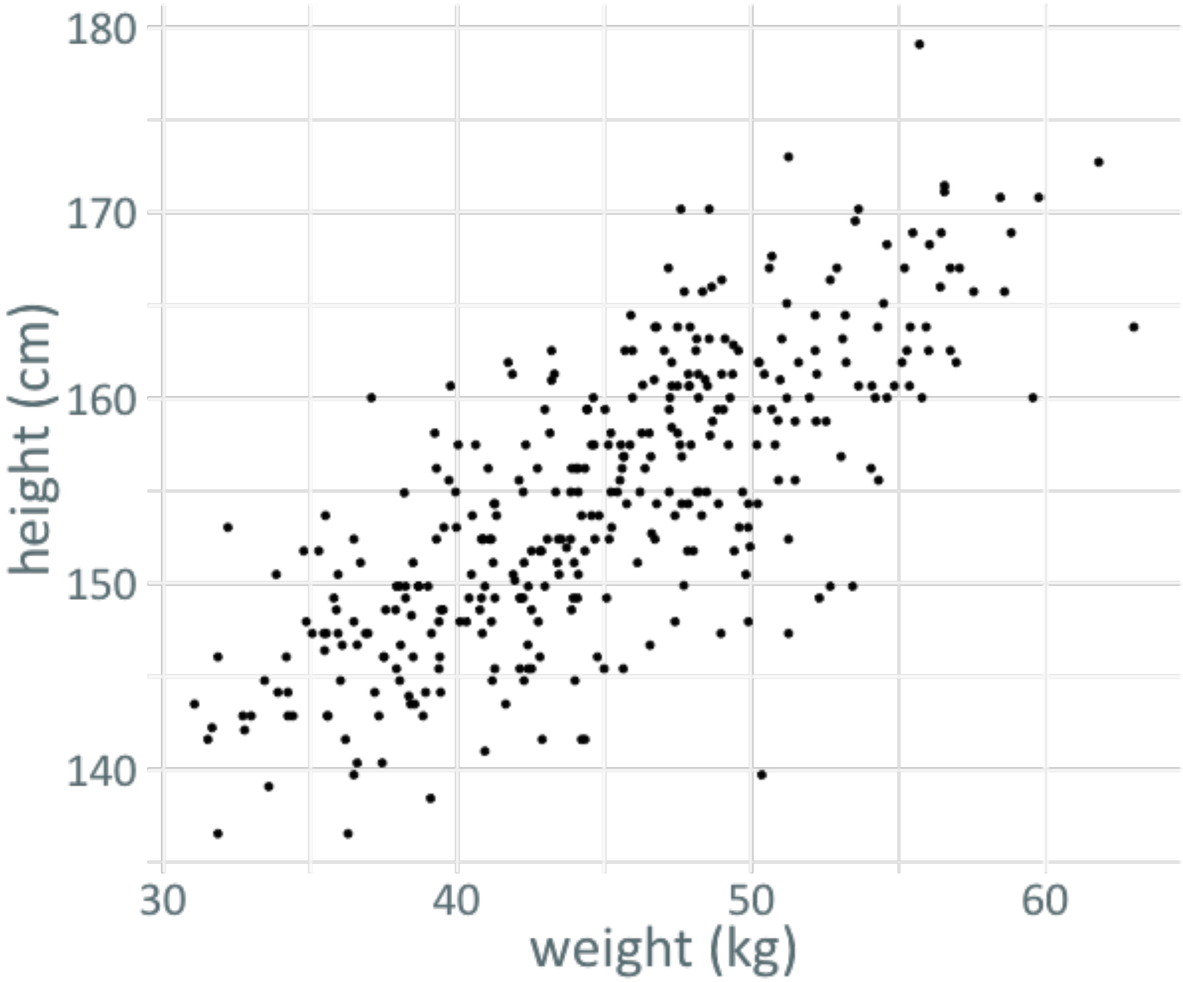
This is sometimes called a non-informative prior



This prior is informative, but in a good way!

Our model for the height data

$$y_i \sim Normal(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta x_i$$
$$\alpha \sim Normal(0, 20)$$
$$\beta \sim lognormal(0, 1)$$
$$\sigma \sim Exponential(1)$$

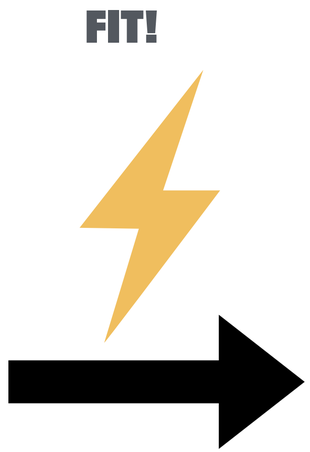


Posterior samples

$$
\begin{aligned}
y_i &\sim Normal(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta x_i \\
\alpha &\sim Normal(0, 20) \\
\beta &\sim lognormal(0, 1) \\
\sigma &\sim Exponential(1)
\end{aligned}
$$

Posterior samples

$$y_i \sim Normal(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta x_i$$
$$\alpha \sim Normal(0, 20)$$
$$\beta \sim lognormal(0, 1)$$
$$\sigma \sim Exponential(1)$$



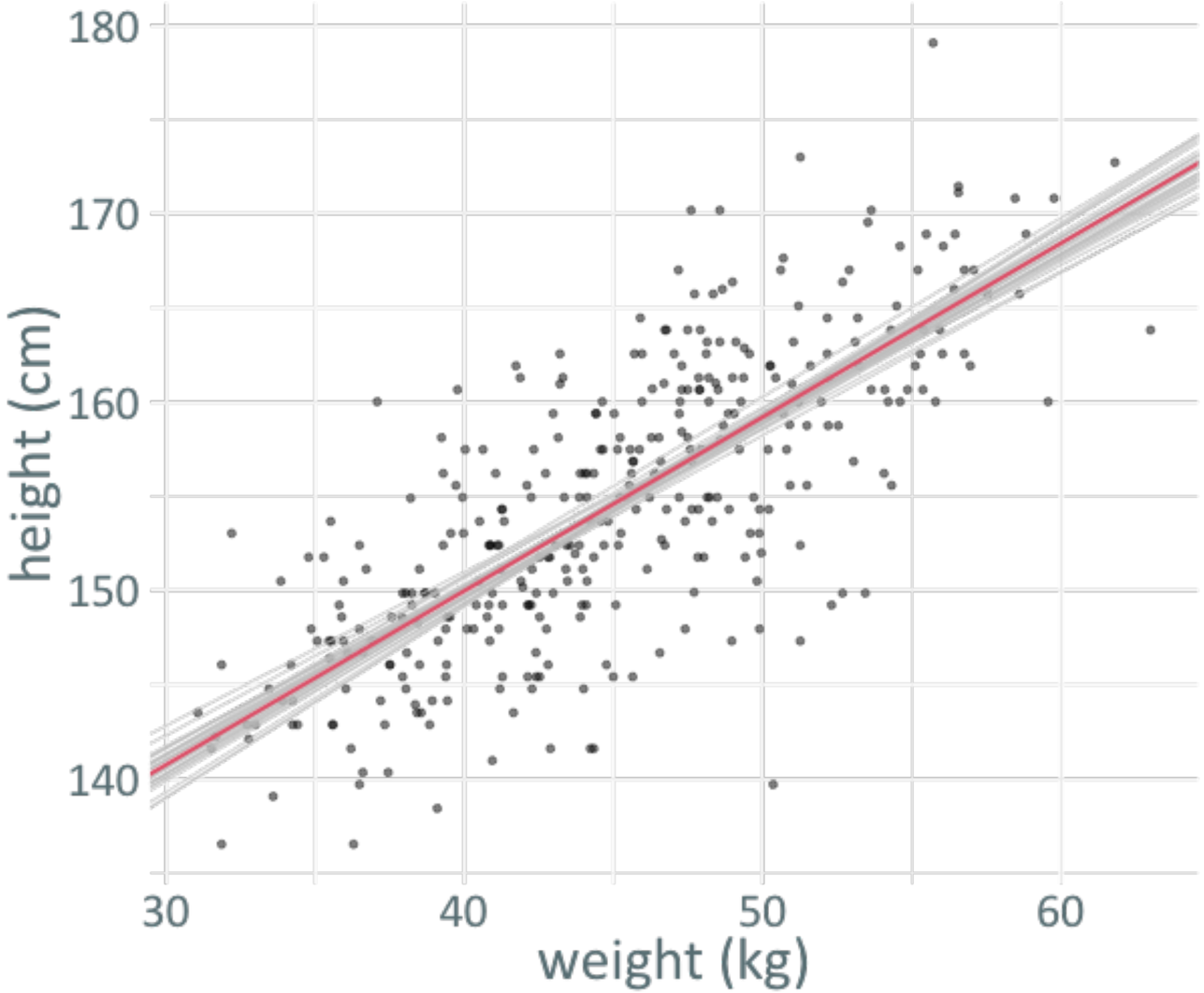

```
> samples
# A tibble: 2,000 × 3
             a          b     sigma
   <dbl[1d]> <dbl[1d]> <dbl[1d]>
 1      115.     0.889       4.78
 2      109.     1.02        5.30
 3      112.     0.928       5.07
 4      111.     0.949       5.30
 5      111.     0.955       5.04
 6      115.     0.872       5.19
 7      109.     1.01        5.13
 8      117.     0.844       5.00
 9      115.     0.882       4.94
10      112.     0.939       4.95
# … with 1,990 more rows
```

Posterior mean estimates

$$E_y[\theta] =$$

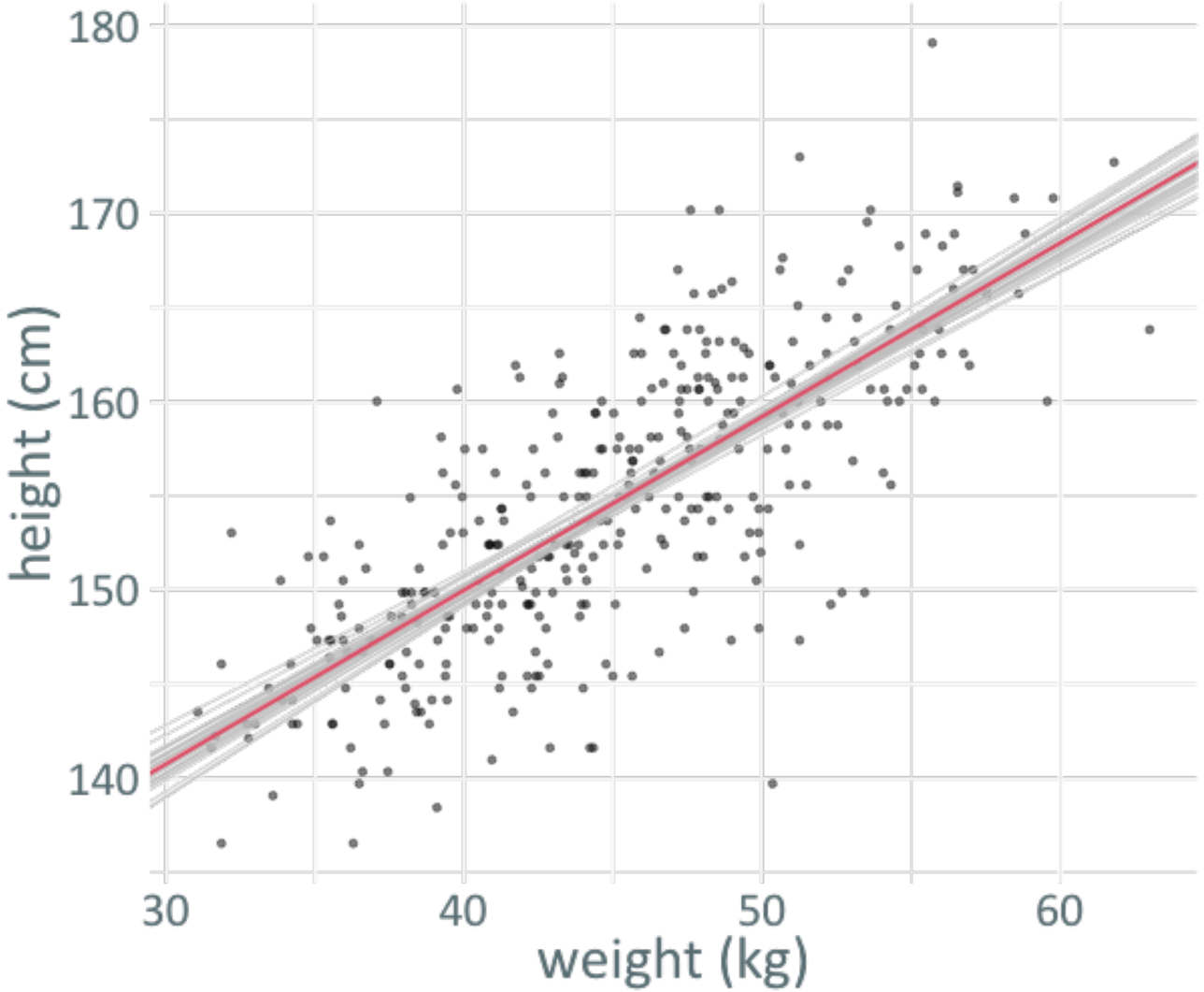
```
> colMeans(samples)
          a           b       sigma
112.9296580   0.9253803   5.0453651
```

Model fit




```
> colMeans(samples)
          a           b       sigma
112.9296580   0.9253803   5.0453651
```

Model fit

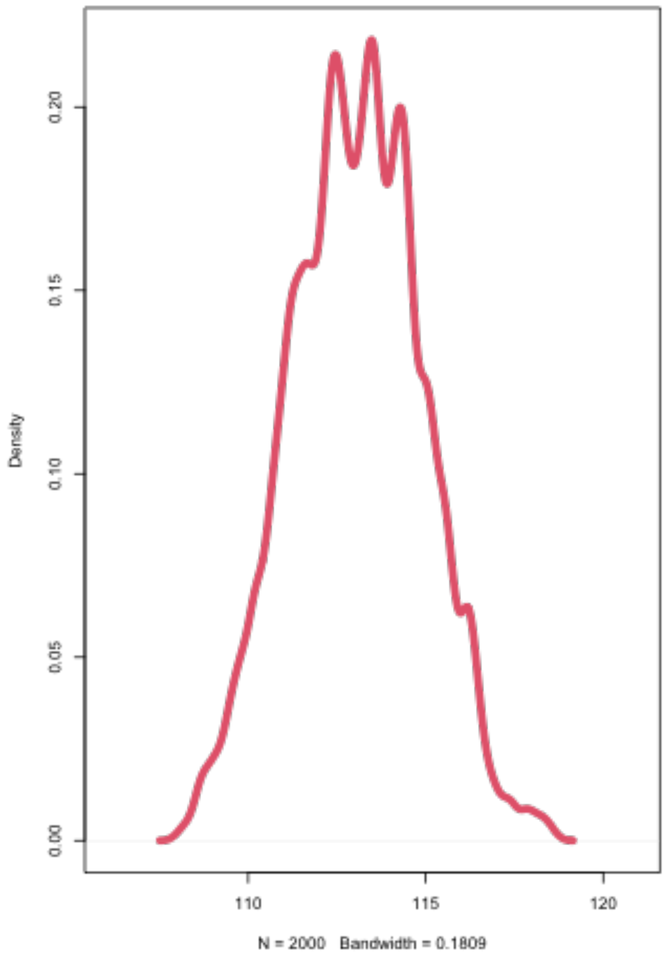



```
> colMeans(samples)
          a           b       sigma
112.9296580   0.9253803   5.0453651
```

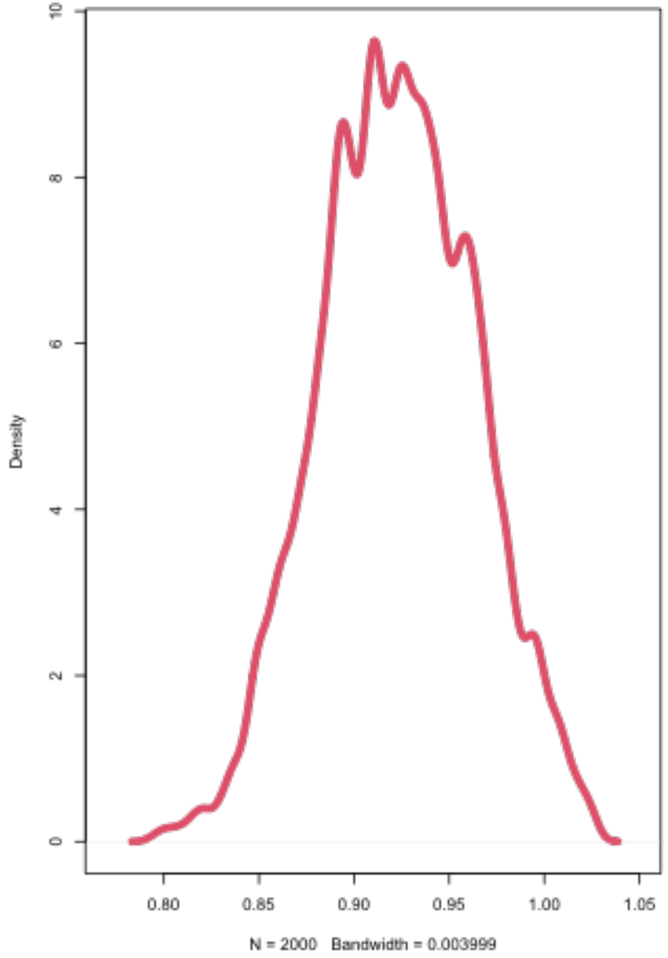
```
> samples
# A tibble: 2,000 × 3
          a           b       sigma
   <dbl[1d]> <dbl[1d]> <dbl[1d]>
 1      115.      0.889      4.78
 2      109.      1.02       5.30
 3      112.      0.928      5.07
 4      111.      0.949      5.30
 5      111.      0.955      5.04
 6      115.      0.872      5.19
 7      109.      1.01       5.13
 8      117.      0.844      5.00
 9      115.      0.882      4.94
10      112.      0.939      4.95
# … with 1,990 more rows
```

The Posterior parameter distribution

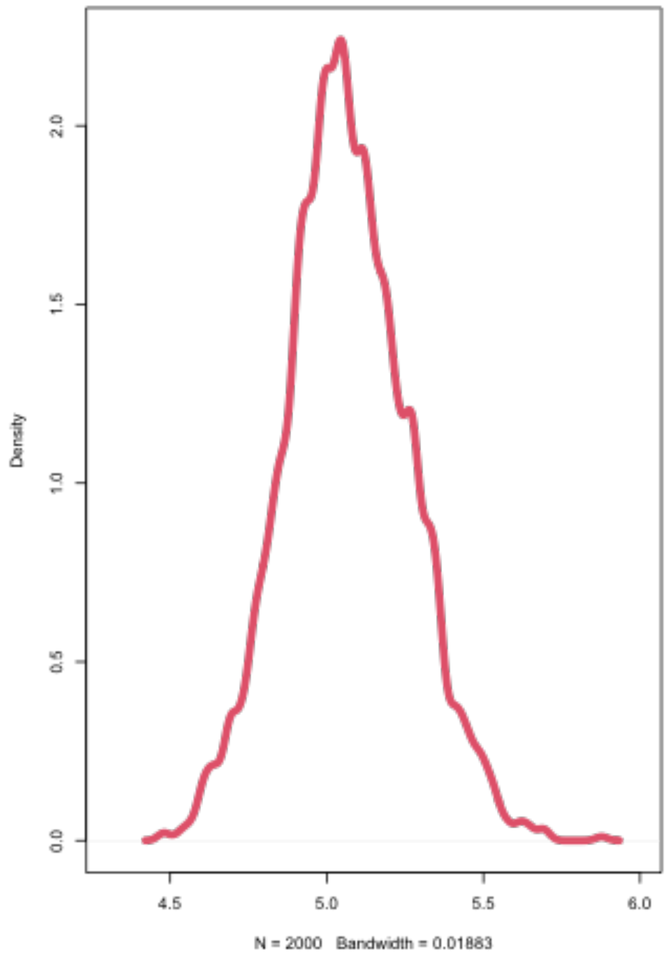
a



b

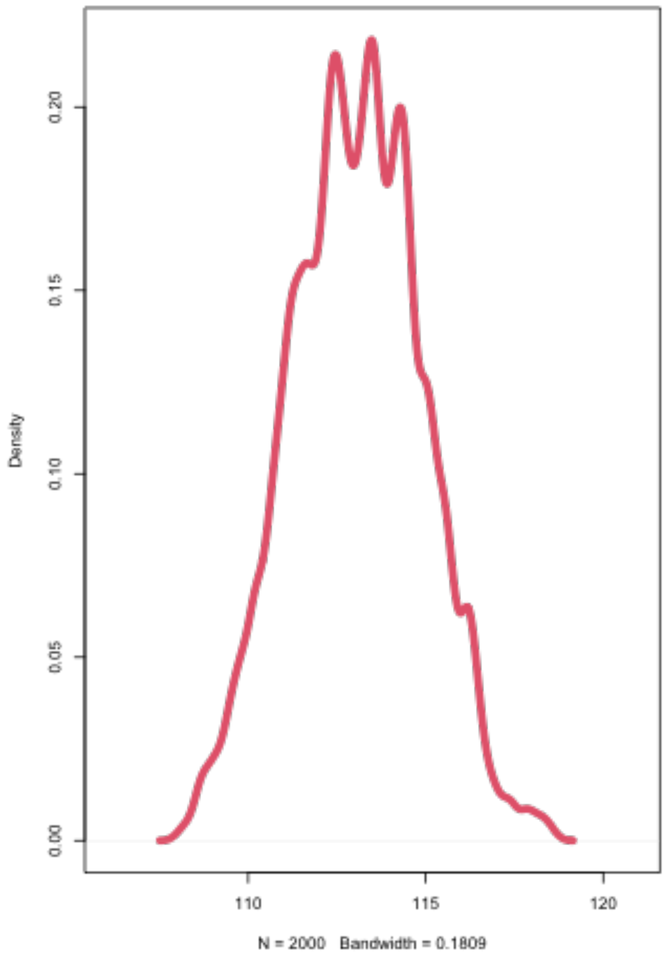


sigma

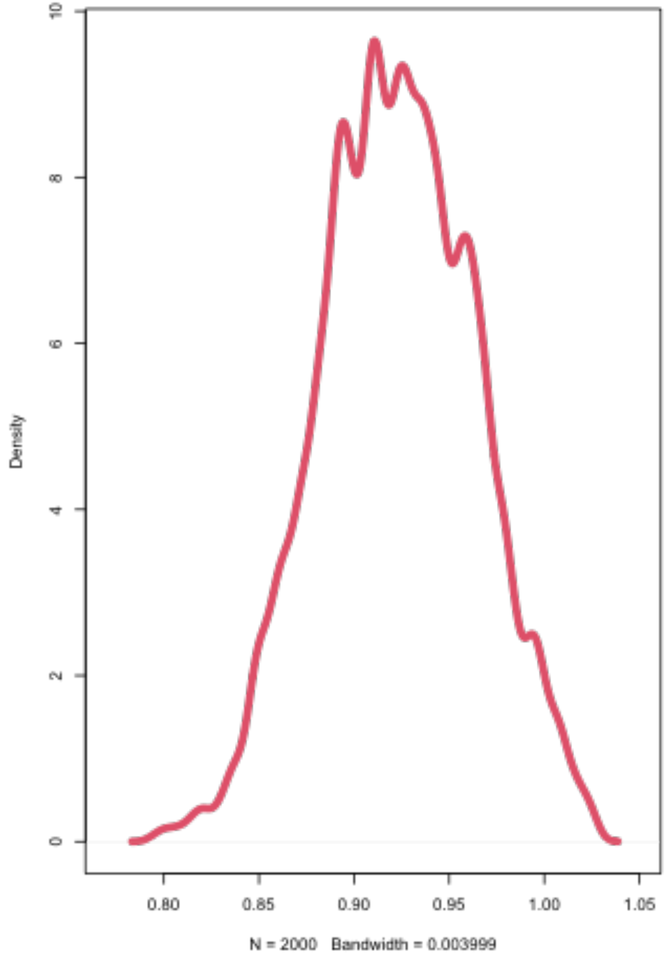


The Posterior parameter distribution

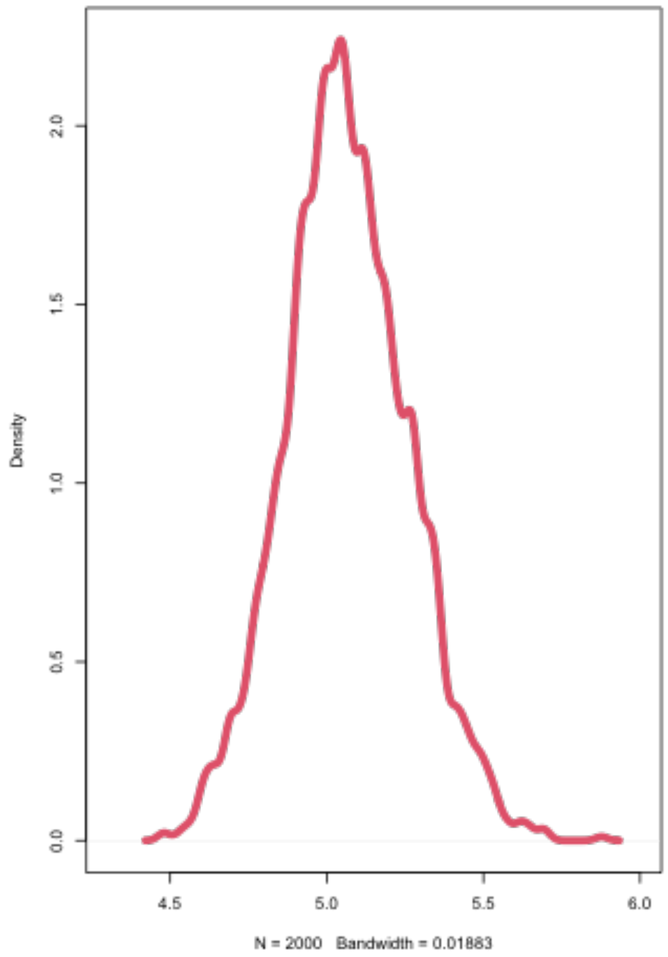
a



b



sigma



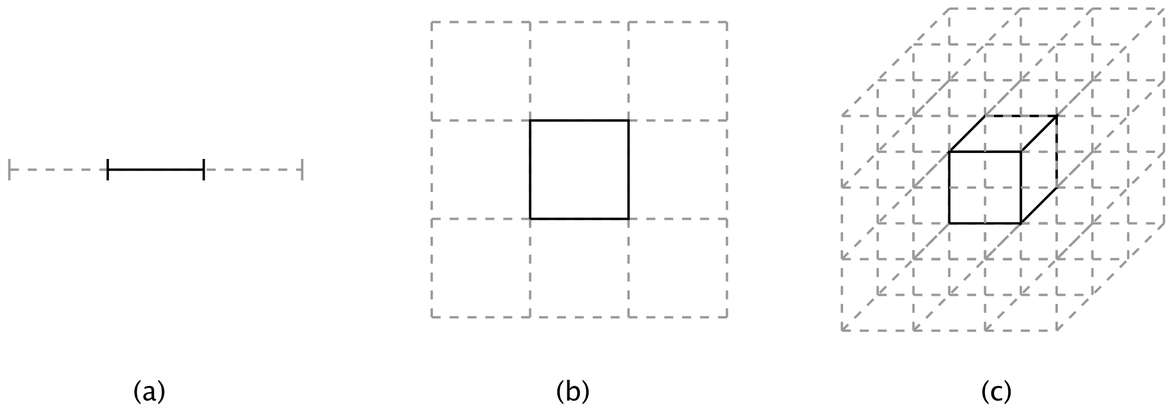

```
> samples
# A tibble: 2,000 × 3
       a     b sigma
   <dbl> <dbl> <dbl>
 1  115.  0.889  4.78
 2  109.  1.02   5.30
 3  112.  0.928  5.07
 4  111.  0.949  5.30
 5  111.  0.955  5.04
 6  115.  0.872  5.19
 7  109.  1.01   5.13
 8  117.  0.844  5.00
 9  115.  0.882  4.94
10  112.  0.939  4.95
# … with 1,990 more rows
```

Intermission

How to get posterior samples?

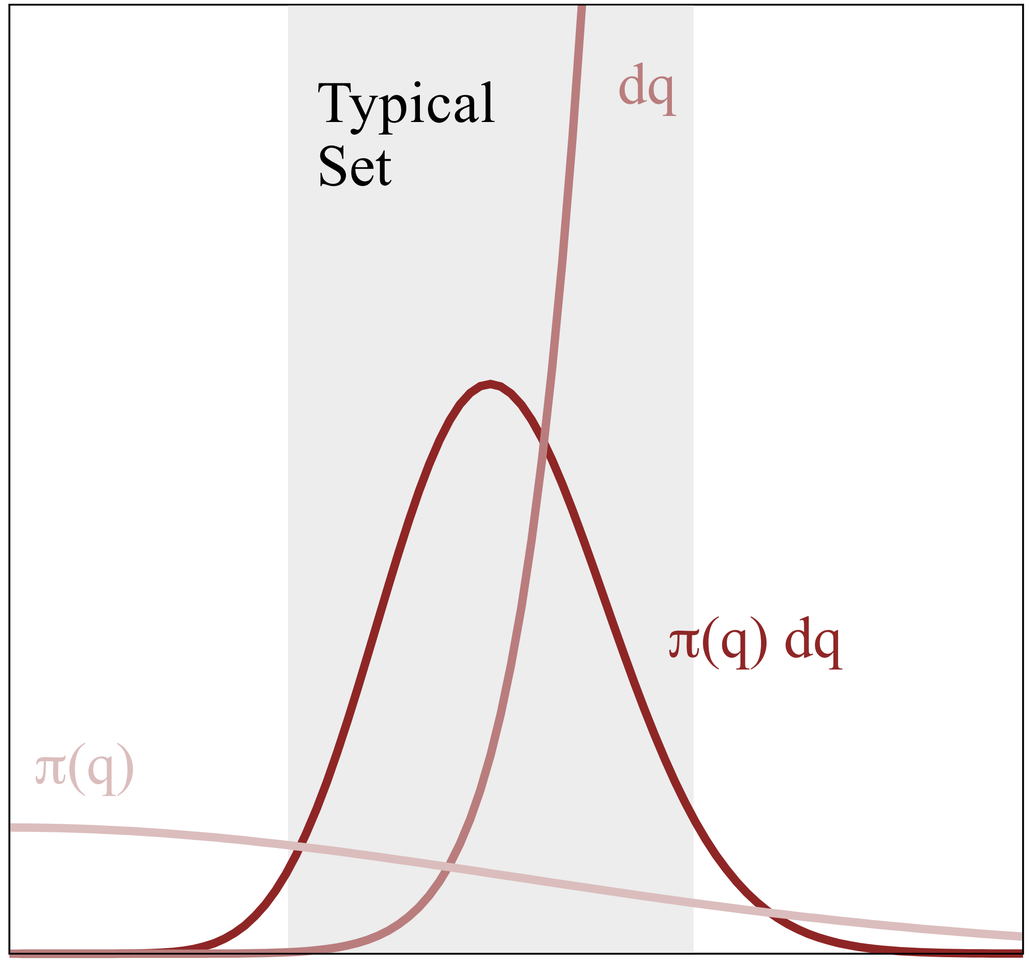
- There is no general method to find high probability regions in arbitrary probability distributions.
- This mean most models are fit using purely computational methods.
- For simple parameters spaces, we can do grid search or some brute force method to find high probability regions

- This breaks down quickly as the number of parameters increases



Betancourt, M. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv [stat.ME]* (2017)

Typical set



Betancourt, M. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv [stat.ME]* (2017)

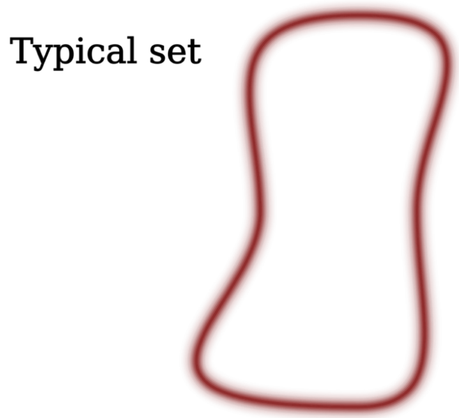
Finding the typical set

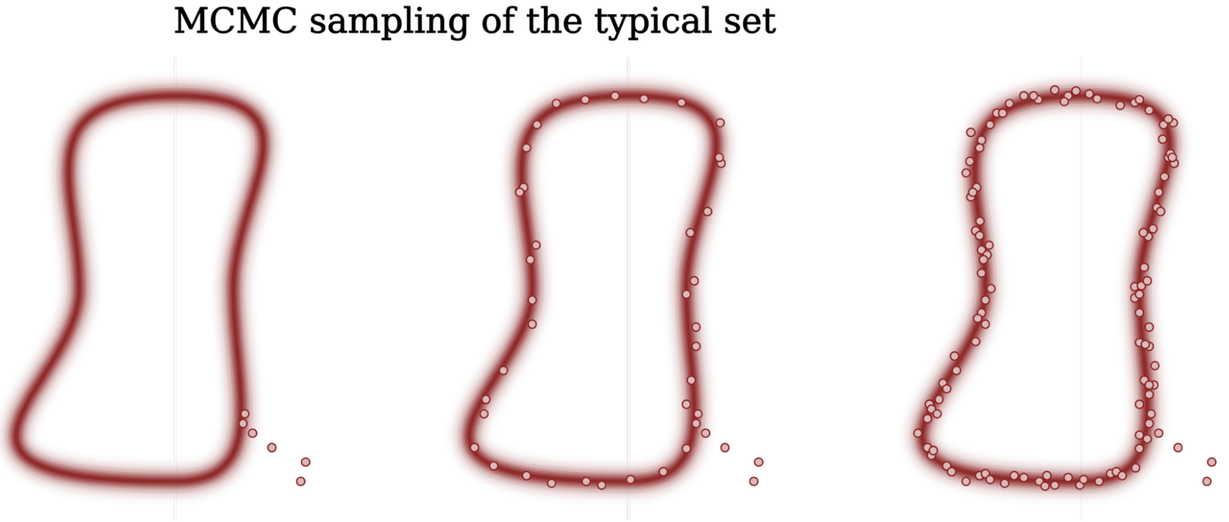
Find a sequence of points in the parameter space that converge to the typical set:

$$\theta_1 \rightarrow \theta_2 \rightarrow \theta_3 \rightarrow \theta_4 \rightarrow \cdots$$

Such that:

$$\{\theta_1, \cdots, \theta_n\} \sim P(\theta \mid y)$$





MCMC samplers

- Metropolis–Hastings algorithms (broad class of samplers, very general).
  - Most methods in the wild are some flavor of this.
- Reversible Jump MCMC (used in many phylogenetic packages).
  - Allows for posterior distributions with variable dimensionality.
- Usable non-mcmc methods: R-INLA - integrated nested Laplace approximation.
  - Great for structural equation modeling, much faster for some classes of models.

MCMC samplers

- Metropolis–Hastings algorithms (broad class of samplers, very general).
  - Most methods in the wild are some flavor of this.
- Reversible Jump MCMC (used in many phylogenetic packages).
  - Allows for posterior distributions with variable dimensionality.
- Usable non-mcmc methods: R-INLA - integrated nested Laplace approximation.
  - Great for structural equation modeling, much faster for some classes of models.
- Gibbs samplers.
  - Mostly surpassed, but still in wide use.
  - Can sample discrete parameters.
  - Requires particular types of priors.
  - Software: WinBugs, Bugs, Jags...
- Hamiltonian Monte Carlo samplers
  - Discrete parameters must be integrated.
  - Can fit dynamic models using differential equations.
  - Software: PyMC3, Edward, Stan (rethinking engine)...

What makes these samplers different?

Basically the transition proposal distribution

$$P(\theta_{i+1} \mid \theta_i)$$

$$\theta_1 \rightarrow \theta_2 \rightarrow \theta_3 \rightarrow \theta_4 \rightarrow \cdots$$

We can visualize what is going on with different samplers:

https://chi-feng.github.io/mcmc-demo/app.html

Our standard model

All the ingredients for a computational fit

$$
\begin{aligned}
y_i &\sim Normal(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta x_i \\
\alpha &\sim Normal(0, 20) \\
\beta &\sim lognormal(0, 1) \\
\sigma &\sim Exponential(1)
\end{aligned}
$$

```
# Data
library(rethinking)
d2 <- Howell1[ Howell1$age >= 18 , ]

# Model
ulam(alist(
  y ~ normal(mu, sigma),
  mu <- a + b * x,
  a ~ normal(0, 20),
  b ~ lognormal(0, 1),
  sigma ~ exponential(1)),
  data = list(y = d2$height,
              x = d2$weight),
  iter = 1000, chains = 4, cores = 4)
```

Our standard model

All the ingredients for a computational fit

$$y_i \sim Normal(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta x_i$$
$$\alpha \sim Normal(0, 20)$$
$$\beta \sim lognormal(0, 1)$$
$$\sigma \sim Exponential(1)$$

```
# Data
library(rethinking)
d2 <- Howell1[ Howell1$age >= 18 , ]

# Model
ulam(alist(
  y ~ normal(mu, sigma),
  mu <- a + b * x,
  a ~ normal(0, 20),
  b ~ lognormal(0, 1),
  sigma ~ exponential(1)),
  data = list(y = d2$height,
              x = d2$weight),
  iter = 1000, chains = 4, cores = 4)
```

rethinking generates Stan code

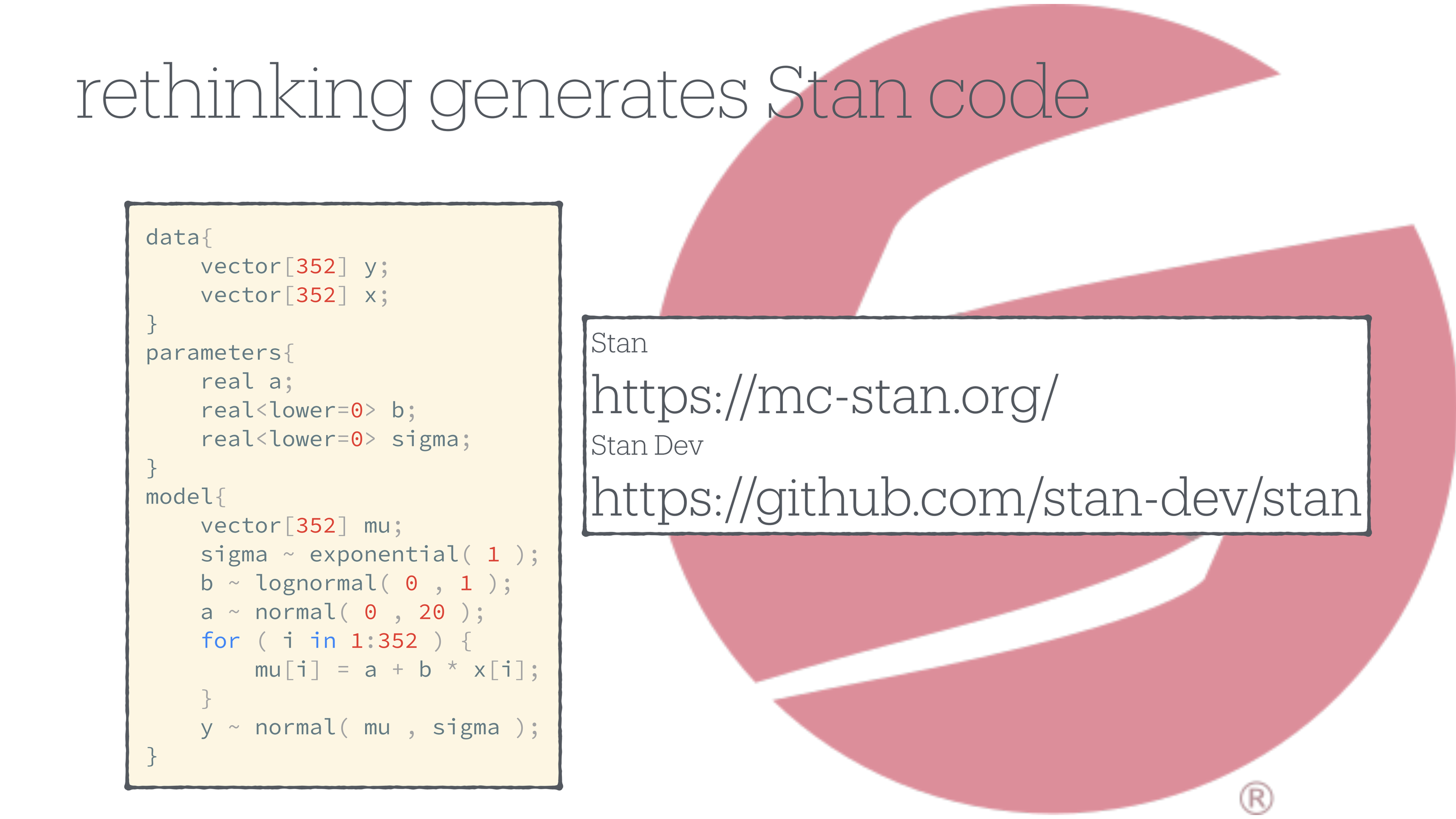
```
data{
    vector[352] y;
    vector[352] x;
}
parameters{
    real a;
    real<lower=0> b;
    real<lower=0> sigma;
}
model{
    vector[352] mu;
    sigma ~ exponential( 1 );
    b ~ lognormal( 0 , 1 );
    a ~ normal( 0 , 20 );
    for ( i in 1:352 ) {
        mu[i] = a + b * x[i];
    }
    y ~ normal( mu , sigma );
}
```

Stan

https://mc-stan.org/

Stan Dev

https://github.com/stan-dev/stan



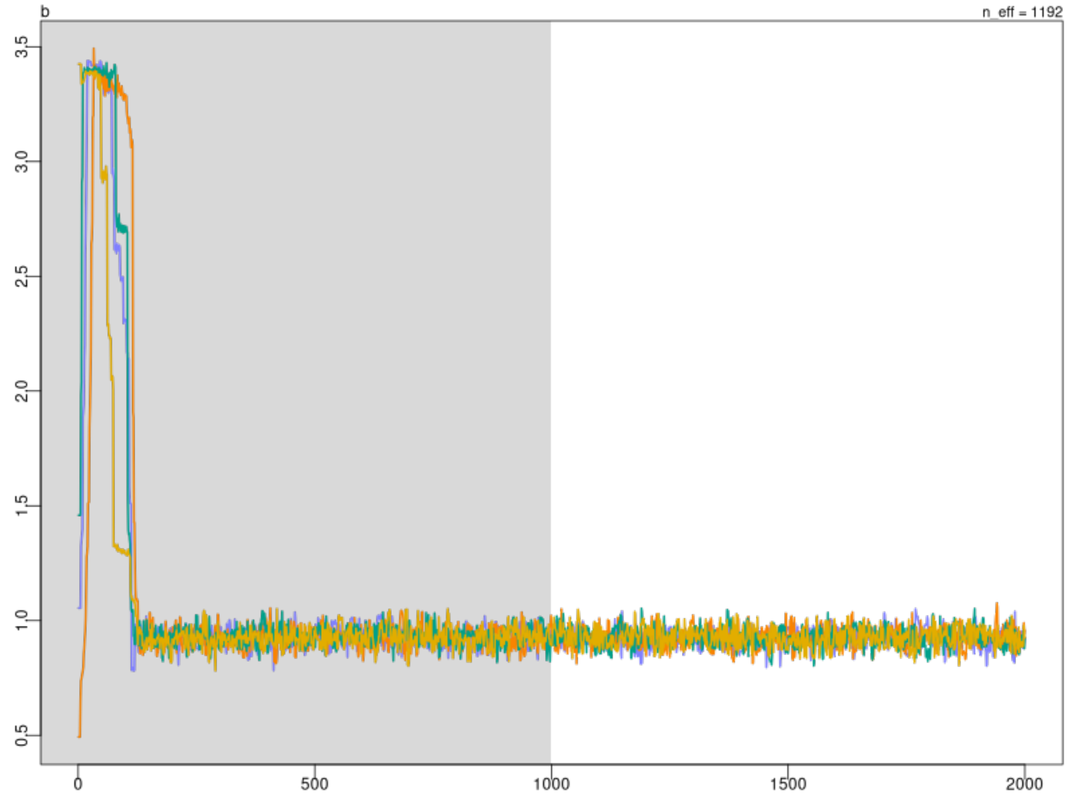
Sampler arguments

- Chains: fit the model several times
- Cores: do the fits in parallel
- Iterations: how many total samples
- Warm-up (or burn-in): starting samples that get discarded

Model summary

```
> precis(fit)
        mean   sd   5.5%  94.5% n_eff Rhat4
a     112.97 2.00 109.83 116.19  1013  1.01
b       0.92 0.04   0.86   0.99  1004  1.01
sigma   5.08 0.20   4.78   5.40  1543  1.00
```

Chains and convergence



Model checking

After fitting the model, we can use the posterior to simulate synthetic data and compare to the data used to fit the model. Discrepancies can suggest paths to improve the model.

$$y_{sim} \sim P(y_{sim}|y) = \sum_{\theta} P(y_{sim}|\theta)P(\theta|y)$$

For each value of the parameters ($\theta_i = \{a_i, b_i, \sigma_i\}$) we can simulate a synthetic dataset y_{sim} and compare to the observed data y.

$$y_{sim} = Normal(a_i + b_i x, \sigma_i)$$

Step by step for posterior simulations

1. Extract the **posterior samples** for the parameters a, b, σ from the fitted model.

2. For each set of parameter values (a_i, b_i, σ_i):

  - Compute the predicted outcome: $y_{pred} = a + bx$.

  - Add random noise to y_{pred}, where the noise is drawn from a normal distribution with mean 0 and standard deviation σ_i.

This gives the synthetic data y_{sim}.

3. Compare the synthetic data y_{sim} to the observed data y.

  - Compute summary statistics (e.g., mean, variance, **quantiles**) for both y_{sim} and y.

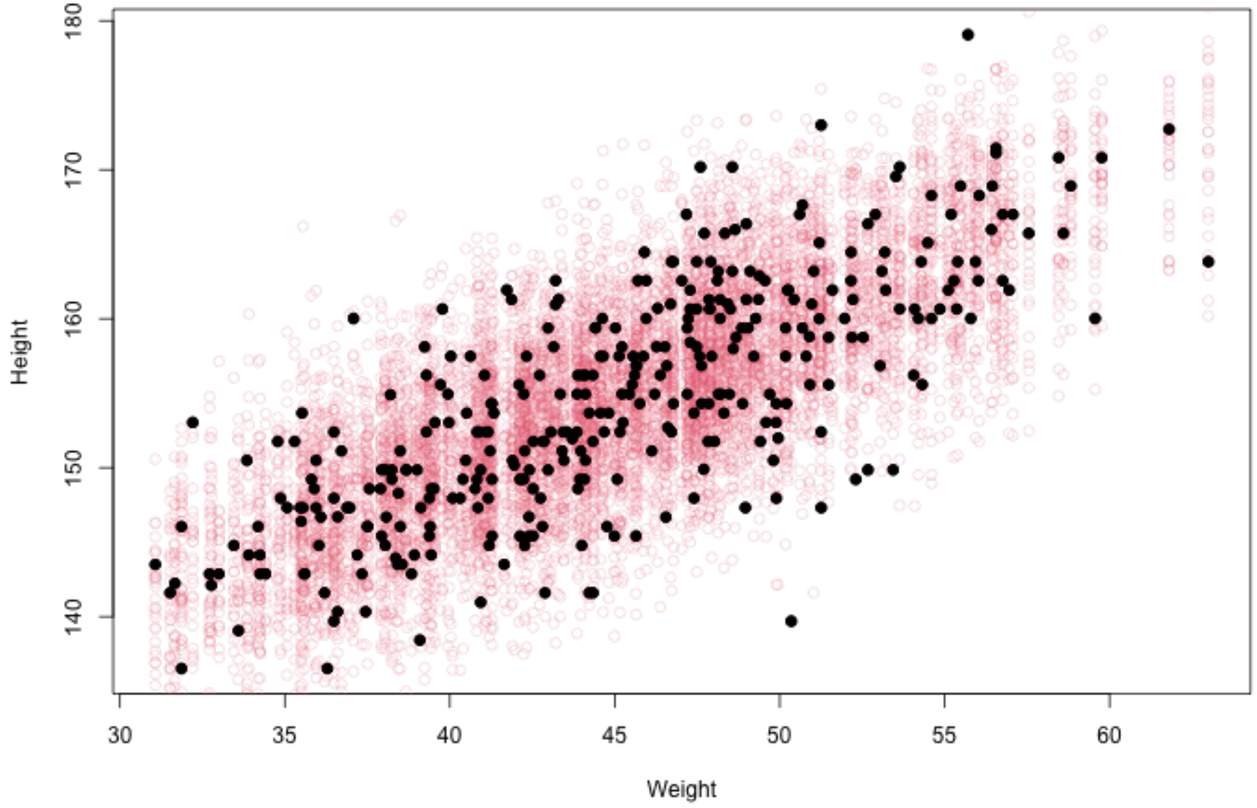
  - If the summary statistics are similar for y_{sim} and y, this suggests that the model is a good fit to the data.

4. Repeat steps 2-3 for all sets of parameter values to get a distribution of summary statistics for the synthetic data.

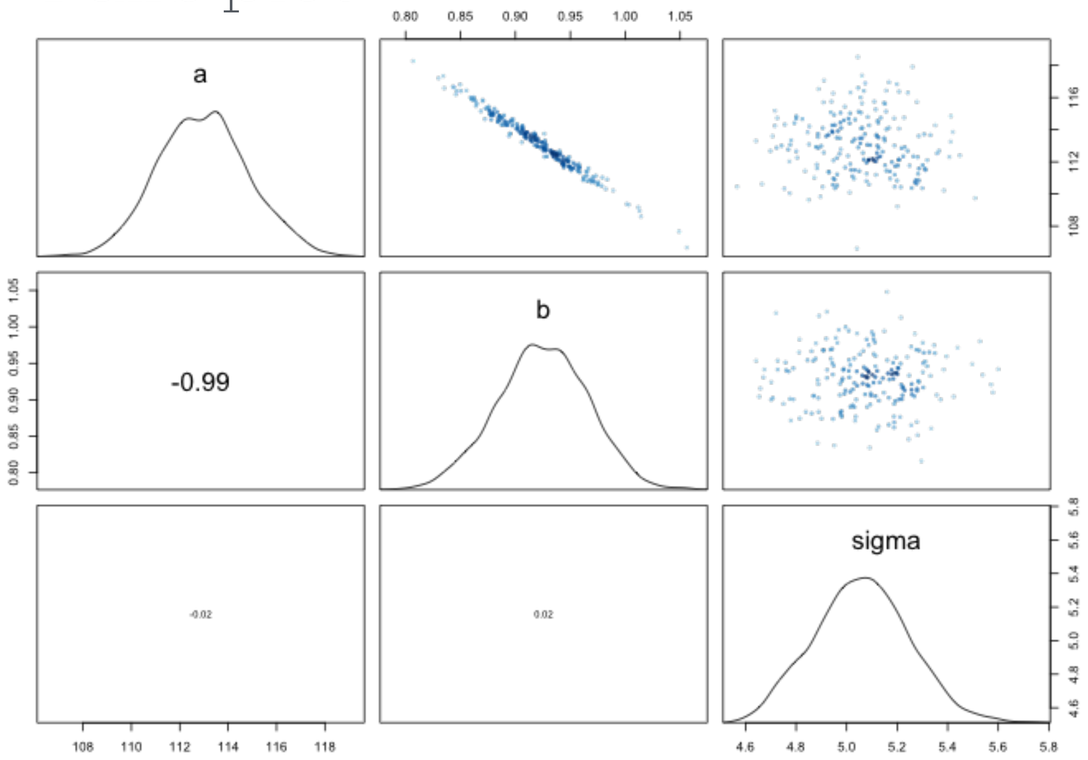
5. Compare the distribution of summary statistics for the synthetic data to the corresponding summary statistics for the observed data. If they are similar, this suggests that the model is a good fit to the data. If they are not similar, this suggests that the model may need to be improved.

Model Check

Posterior simulations



Pairs plot



Summary

- The posterior distribution contains a lot of useful information not accessible by other methods, all including uncertainty.
- We need to understand the computation methods we use to probe the posterior.
  - Convergence checks are fundamental and can help us diagnose bad models.
  - Stan provides many (many!) tools for model checking.
- There are no true unique residuals in Bayesian models, but can use posterior simulations to make and understand our predictions.