Statistical Techniques for Data Science & Robotics

Week 9

Quiz

Objectives for today

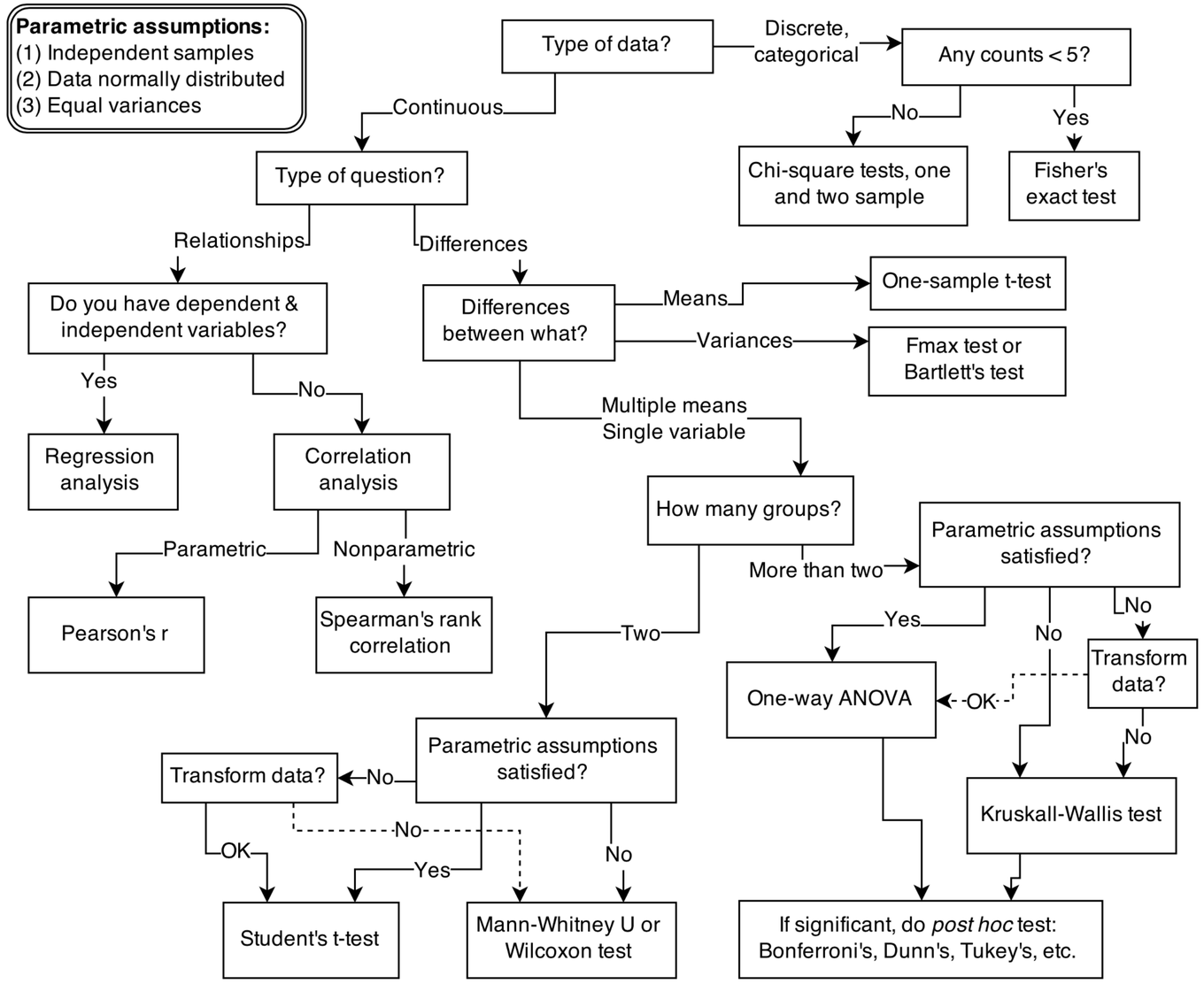
- to recap and understand
  - Bayesian approach to statistics

- to understand
  - Bayesian statistics. Examples.
  - Applications of Bayesian statistics

What we did in the first part of the course

- Estimation
  - collect data
  - choose estimator
  - analyze variability
  - Bootstrap
  - Histogram,
  - KDE

- Inference
  - Hypothesis Testing
  - Significance Tests
  - Confidence intervals

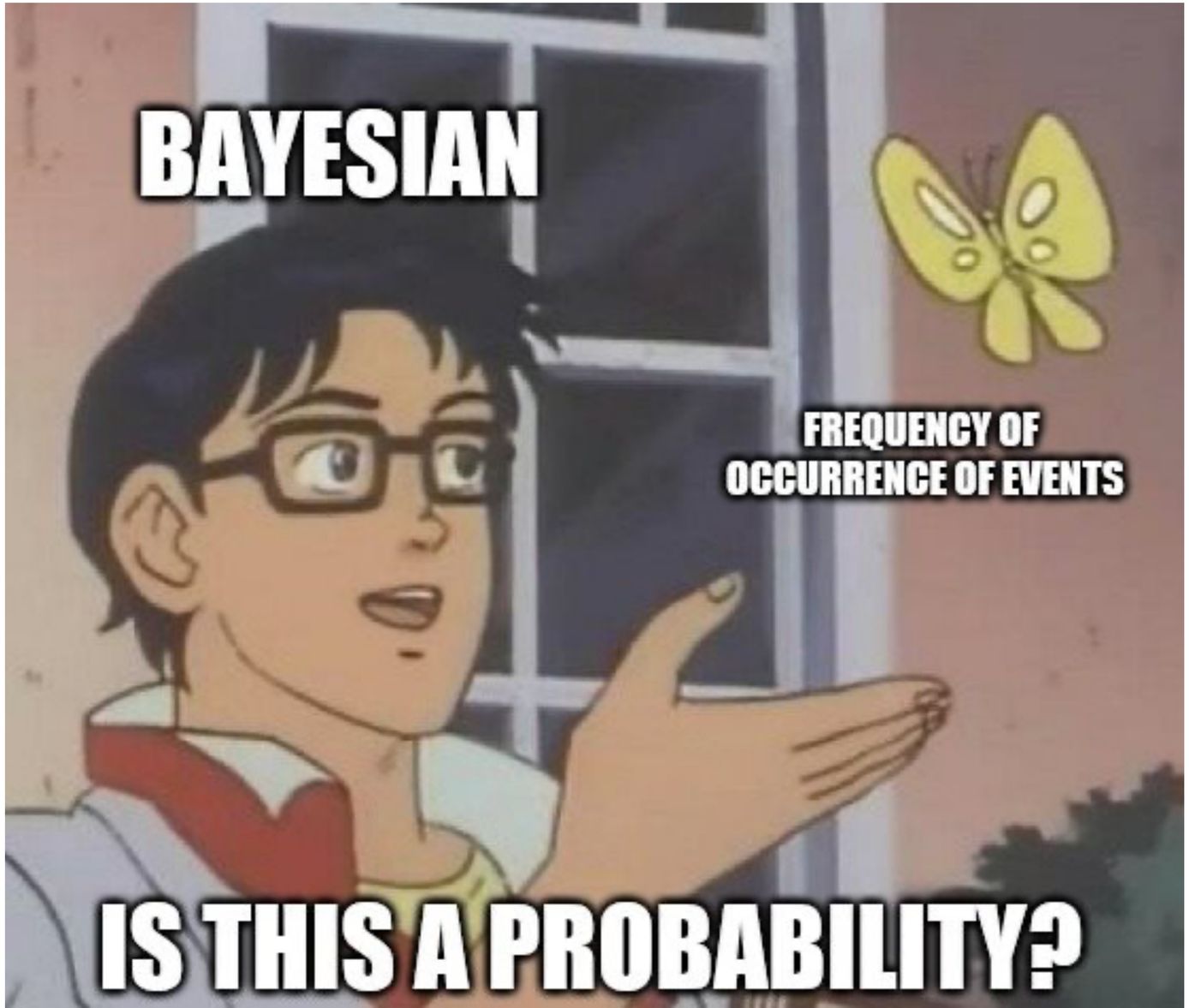


Source: Richard McElreath “Statistical Rethinking”, 2nd Edition, 2019

Overview of the remaining part of the course

- Bayesian statistics
- Bandit Algorithms
- Sampling methods
- MCMC

Bayesian Approach to Statistics



Intuition behind interval estimate

- Confidence interval (U,V)
  - We construct two statistics, say U and V, $U < V$,
  - such that we have a probability p that the random interval (U, V) covers a **fixed but unknown** point (parameter, say μ).

Intuition behind interval estimate

- We adopted this principle:
  - Use the experimental results to compute the values of U and V, say u and v; then call the interval (u, v) a **100p** % confidence interval for the parameter μ.
  - parameter μ is **fixed but unknown**
  - interval (u, v) **is random** (depends on a sample, error depends on sample size…)

…there could be other principles

Bayesian statistics

- **Bayesian statistics** takes into account any prior knowledge of the experiment that the statistician has
- Example:
  - We want to get an observation from $X \sim \text{Poisson}(\lambda)$,
  - to estimate the parameter λ
  - p.d.f. $f(x, \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$, where $x \in 0,1,2,\ldots$

…there could be other principles

Bayesian statistics

- We want to get an observation from $X \sim \text{Poisson}(\lambda)$,
- to estimate the parameter λ
- p.d.f. $f(x, \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$, where $x \in 0,1,2,\ldots$

- We **believe** that λ can have either of two values (2 or 4)
  - 2 with prior probability $= 0.8$
  - 4 with prior probability $= 0.2$
  - this destribution is **prior**

Example cont.

$$f(x,\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}$$

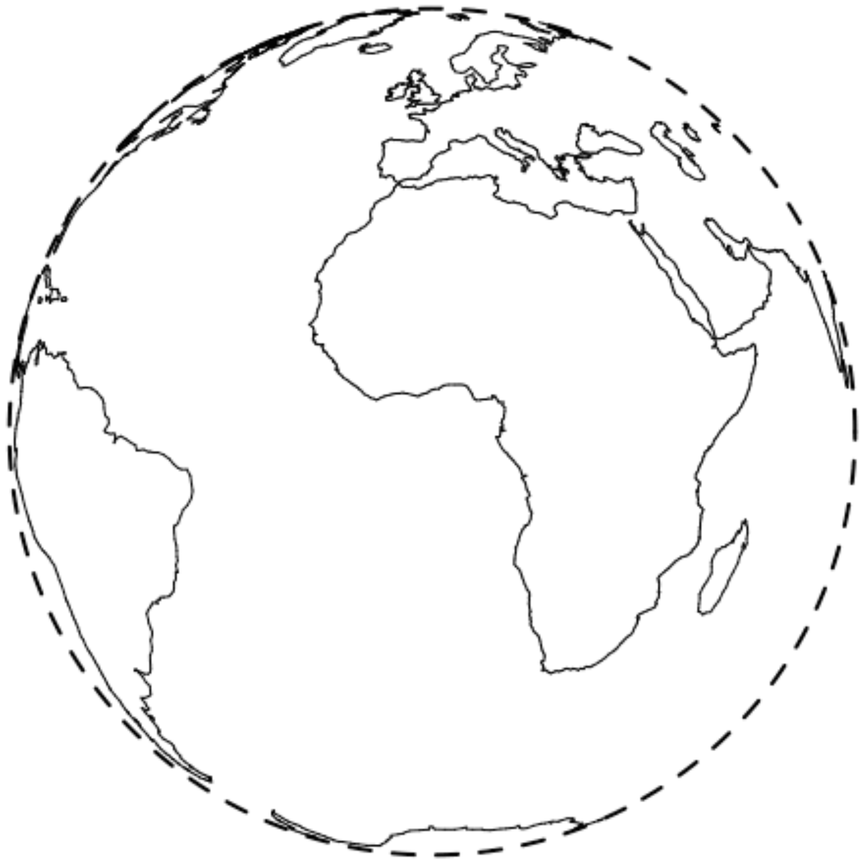
- So, we have run the experiment and got x=6
  - P(X=6 | λ=2)=0.012 $f(X=6\,|\,\lambda=2) = \frac{2^6}{6!}e^{-2}$
  - P(X=6 | λ=4)=0.104 $f(X=6\,|\,\lambda=4) = \frac{4^6}{6!}e^{-4}$
- Those are **likelihoods**
- Now, we can compute the conditional probability:
  - $P(\lambda=2\,|\,X=6) = \frac{P(\lambda=2, X=6)}{P(X=6)} = \frac{P(\lambda=2)P(X=6\,|\,\lambda=2)}{P(\lambda=2)P(X=6\,|\,\lambda=2)+P(\lambda=4)P(X=6\,|\,\lambda=4)} = 0.316$
  - $P(\lambda=4\,|\,X=6) = \frac{P(\lambda=4, X=6)}{P(X=6)} = \frac{P(\lambda=4)P(X=6\,|\,\lambda=4)}{P(\lambda=2)P(X=6\,|\,\lambda=2)+P(\lambda=4)P(X=6\,|\,\lambda=4)} = 0.648$
- these two are **posterior probability distribution**


I HAVE A PRIOR
I HAVE A MEASUREMENT MODEL
BOOM!
POSTERIOR
meme-arsenal.ru

Water on Earth: a Case study

Slides taken from the “Statistical rethinking” course

What proportion of the surface is covered with water?



How should we use the sample?

How to produce a summary?

How to represent uncertainty?

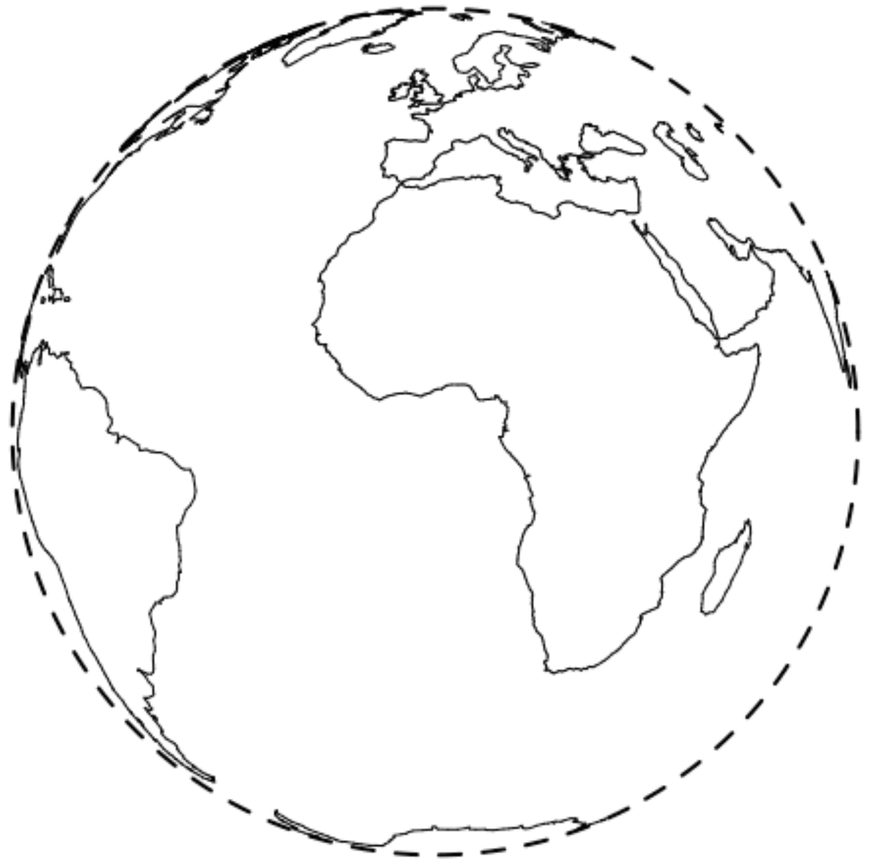
Globe of Forking Water

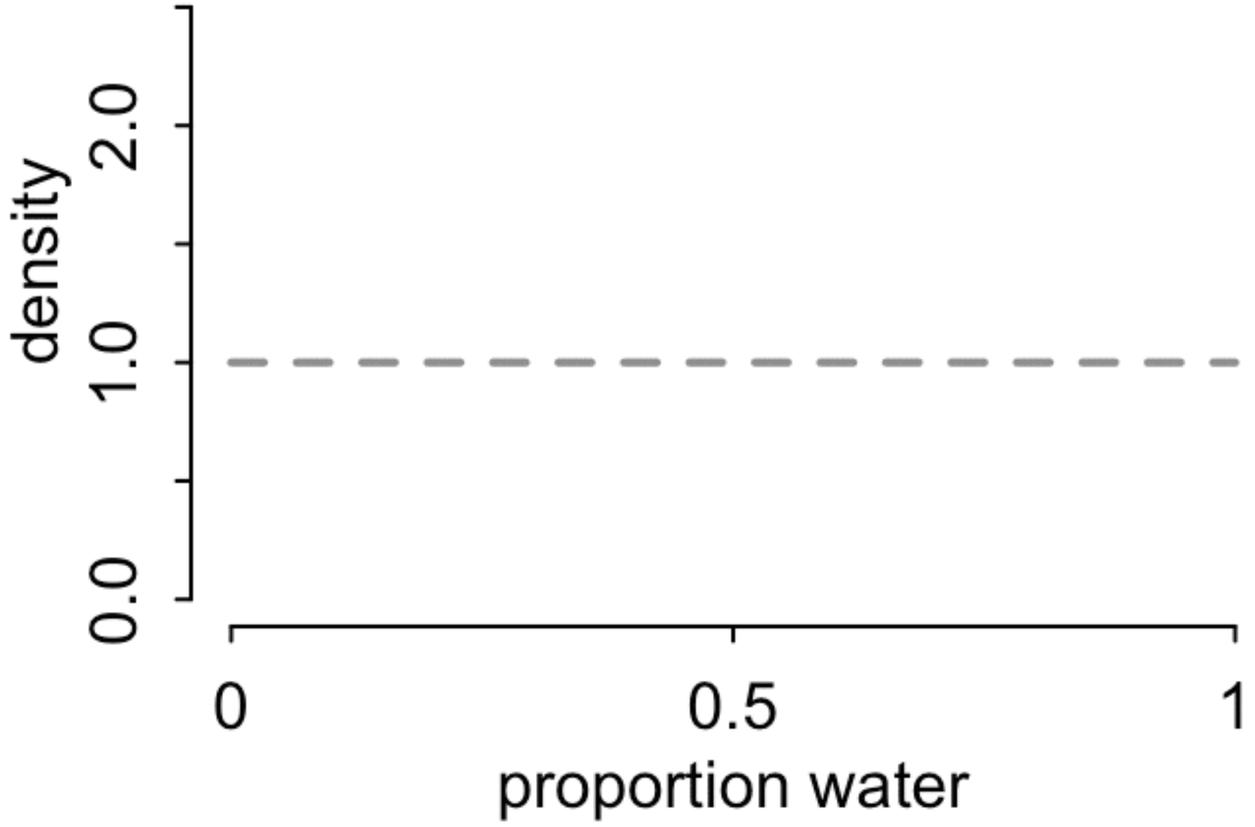
For each possible proportion of water,

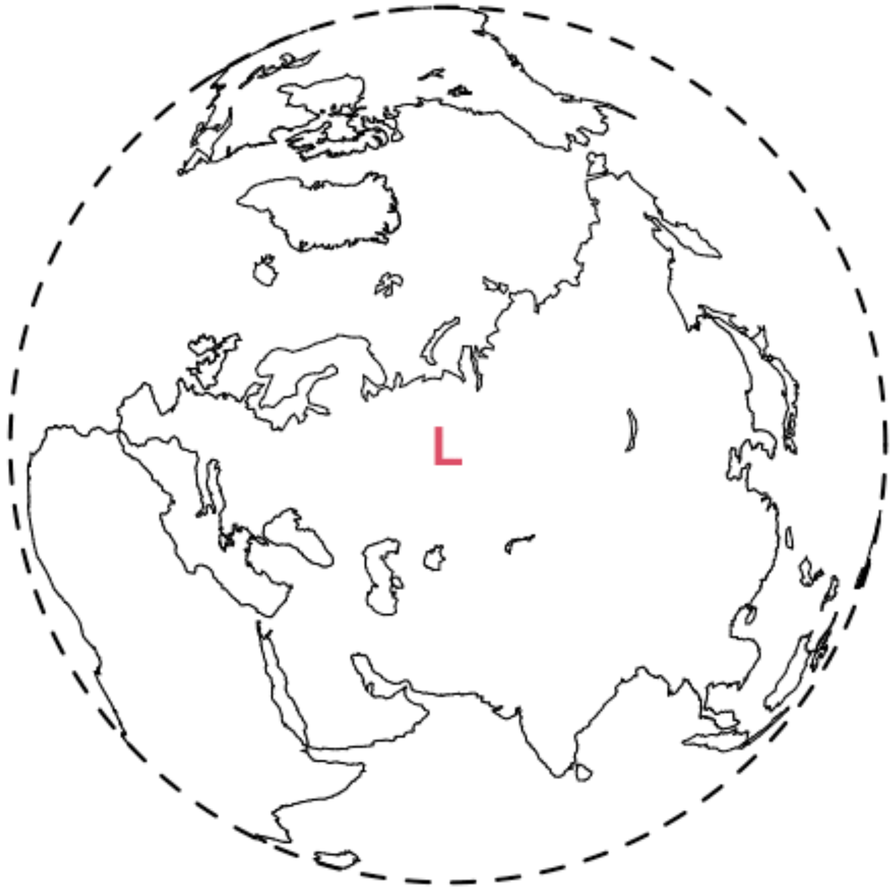
Count number of ways data could happen.

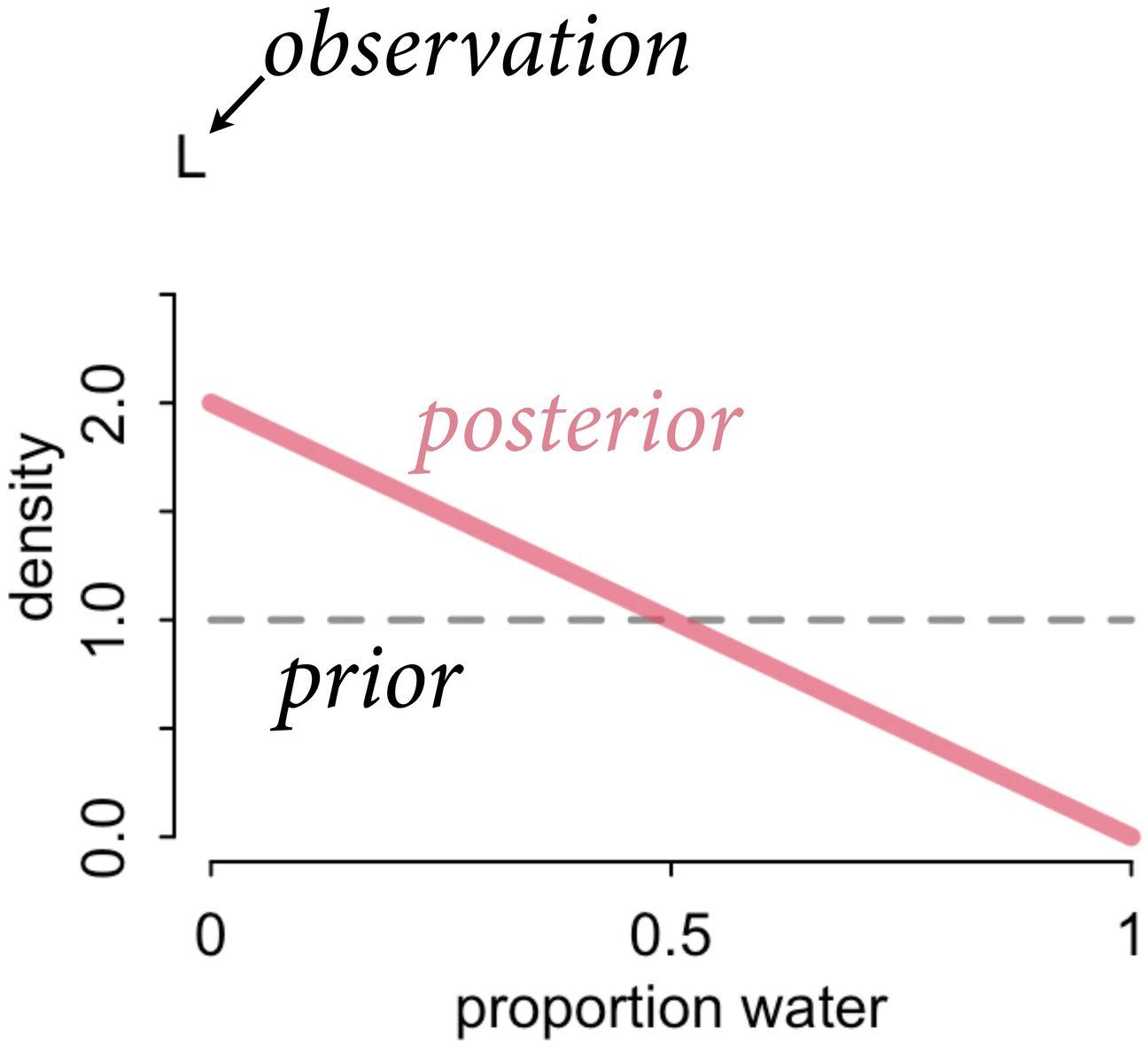
Must state how observations are generated

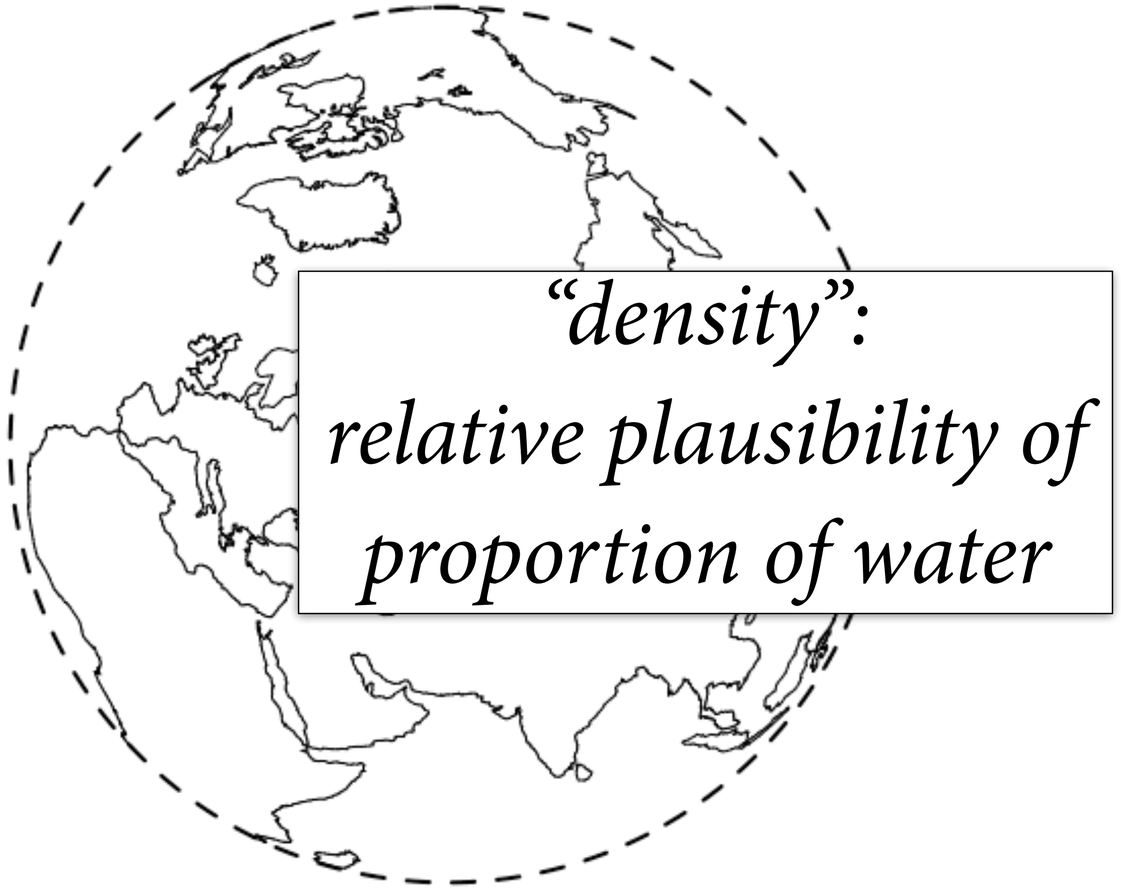
Toss The First



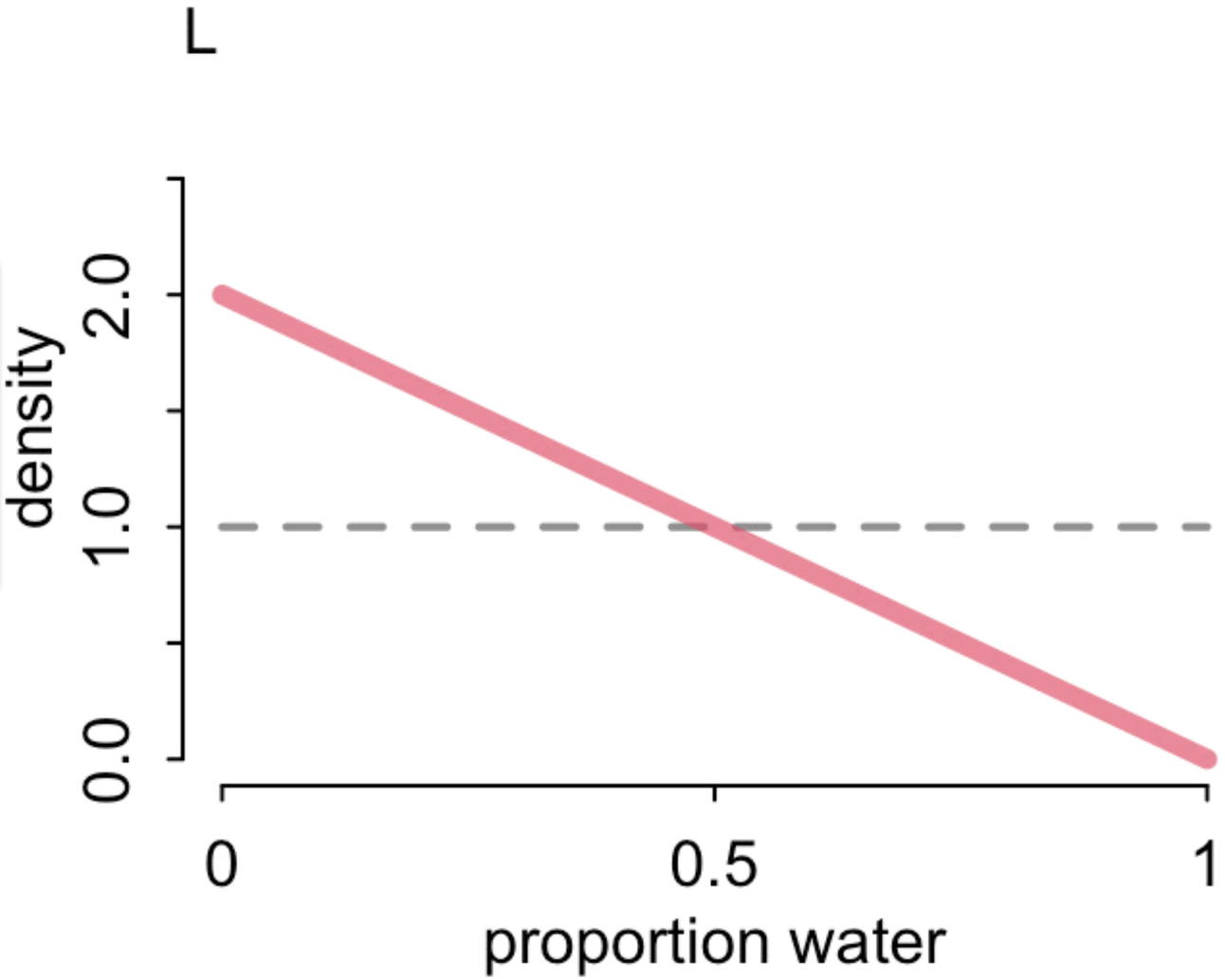



L

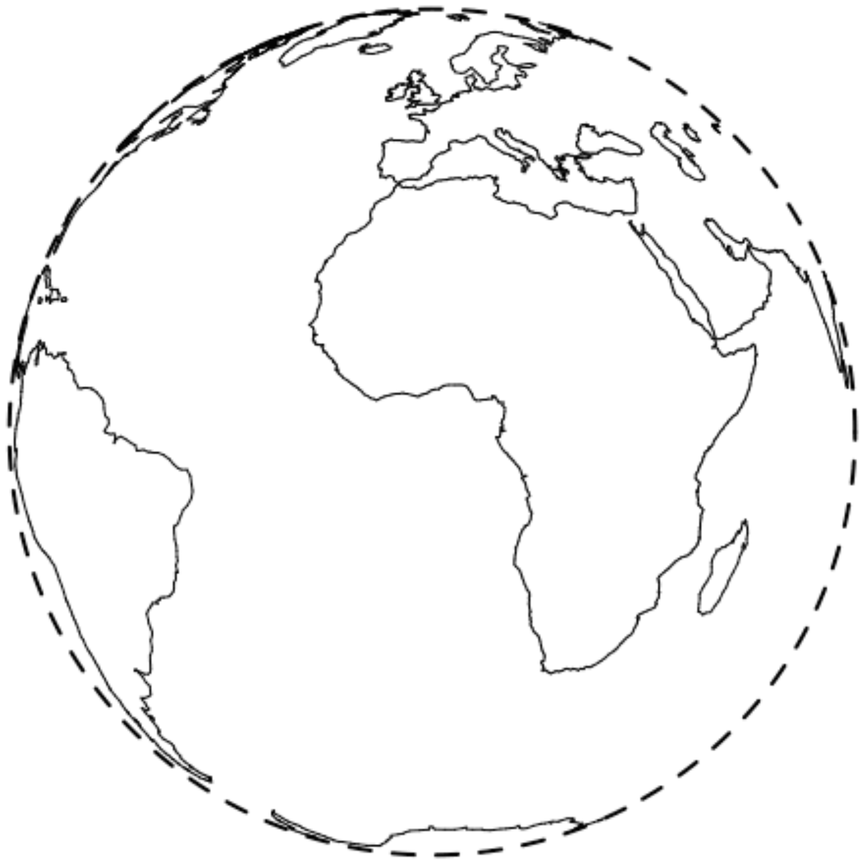

observation
posterior
prior
density
0.0
1.0
2.0
0
0.5
1
proportion water

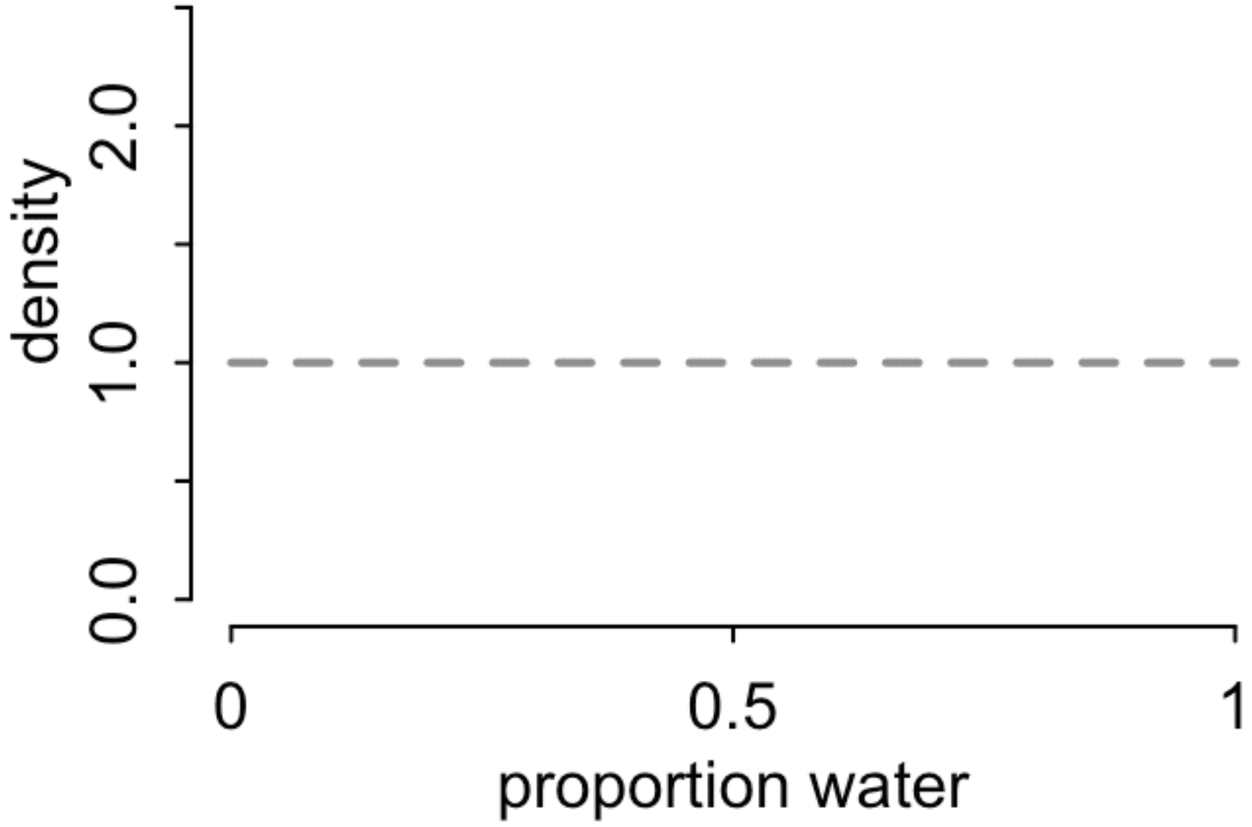


"density":
relative plausibility of
proportion of water

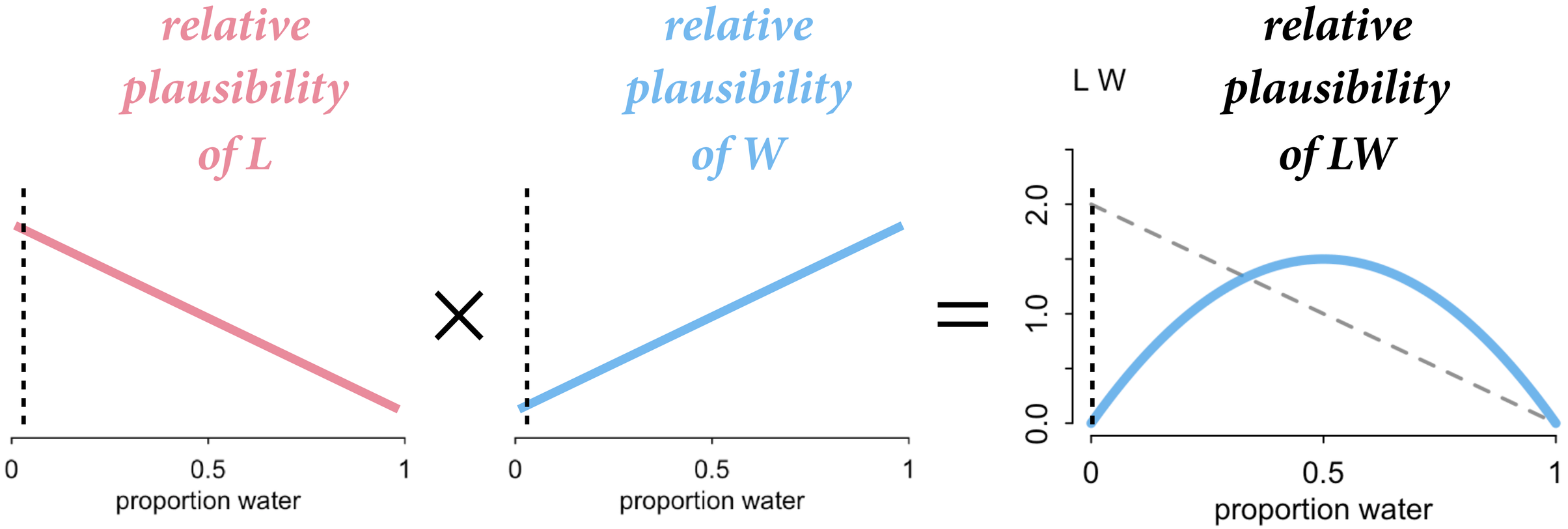


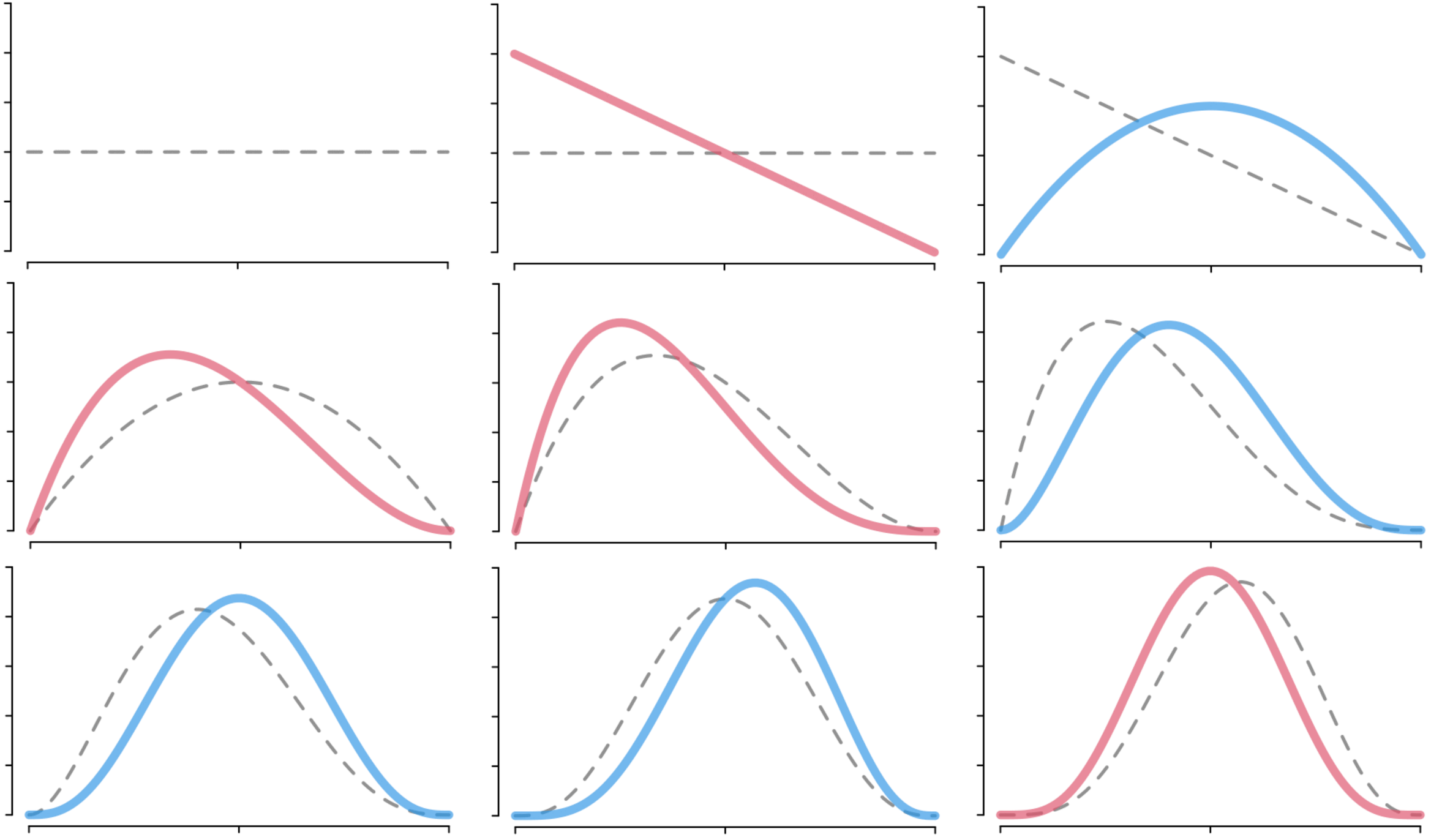
Toss The Second



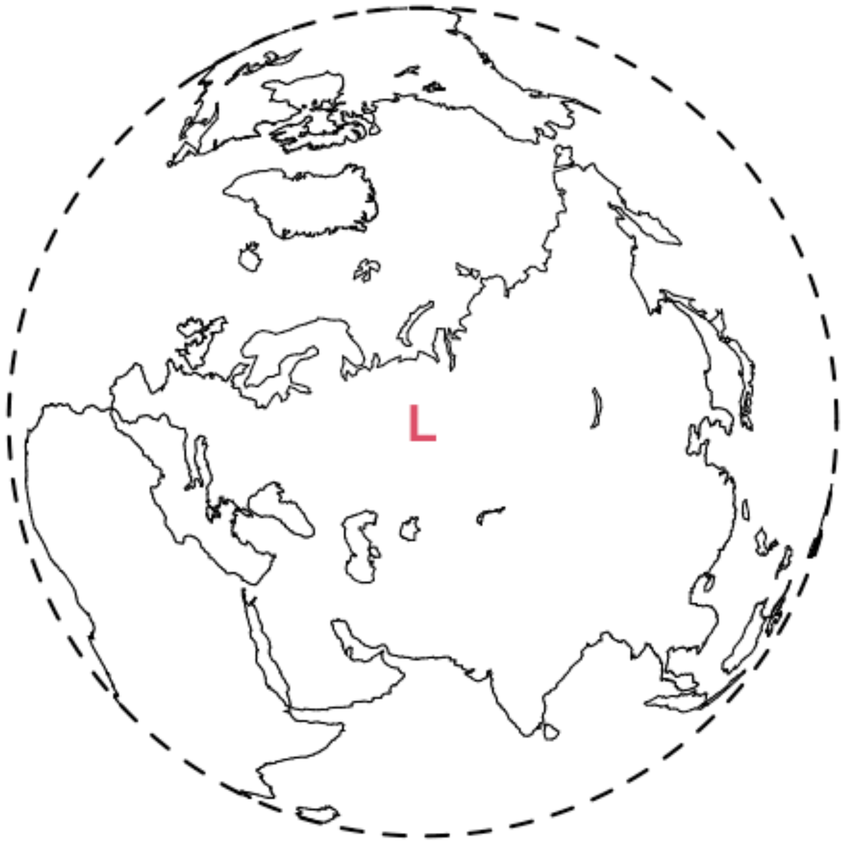


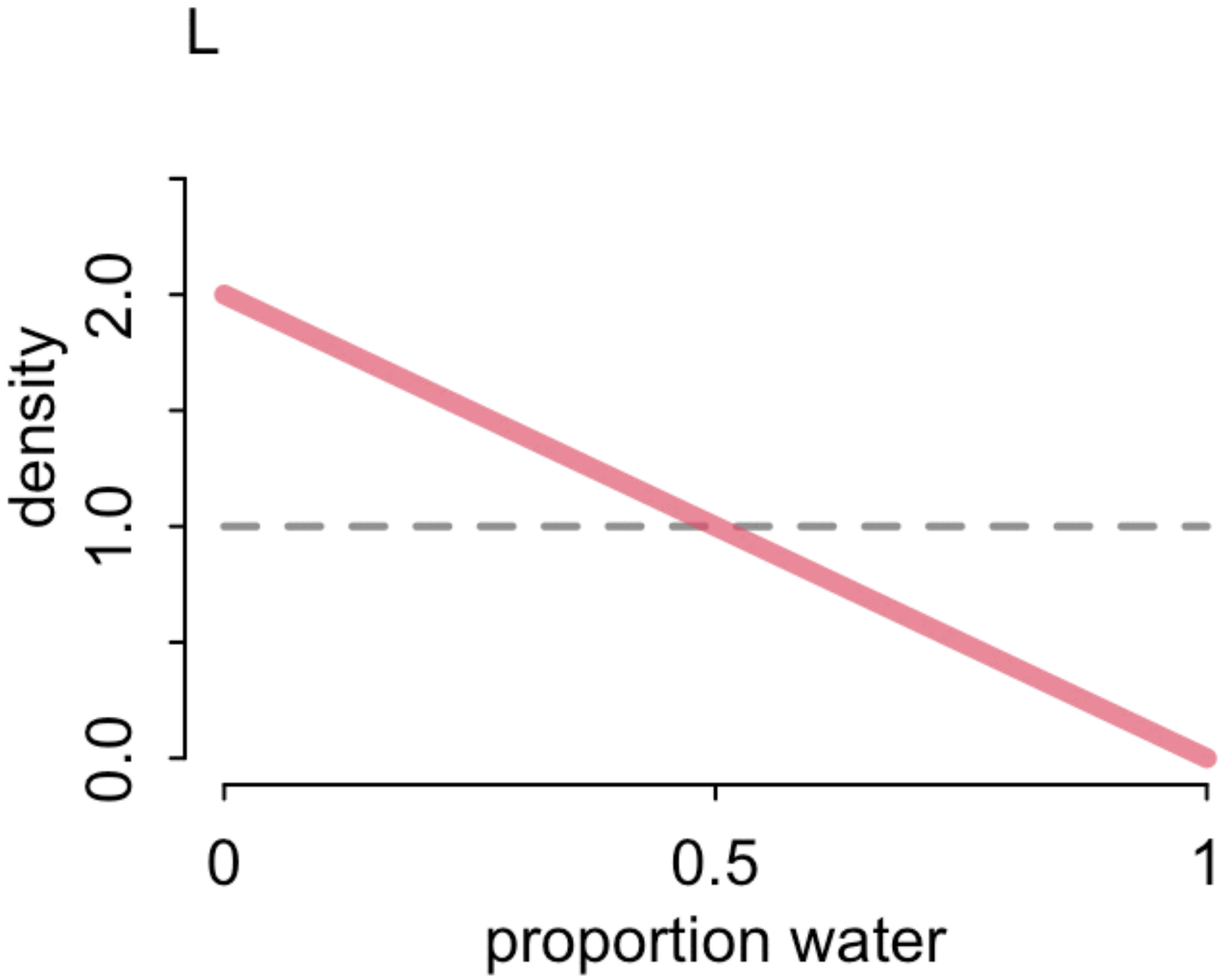
Toss The Second



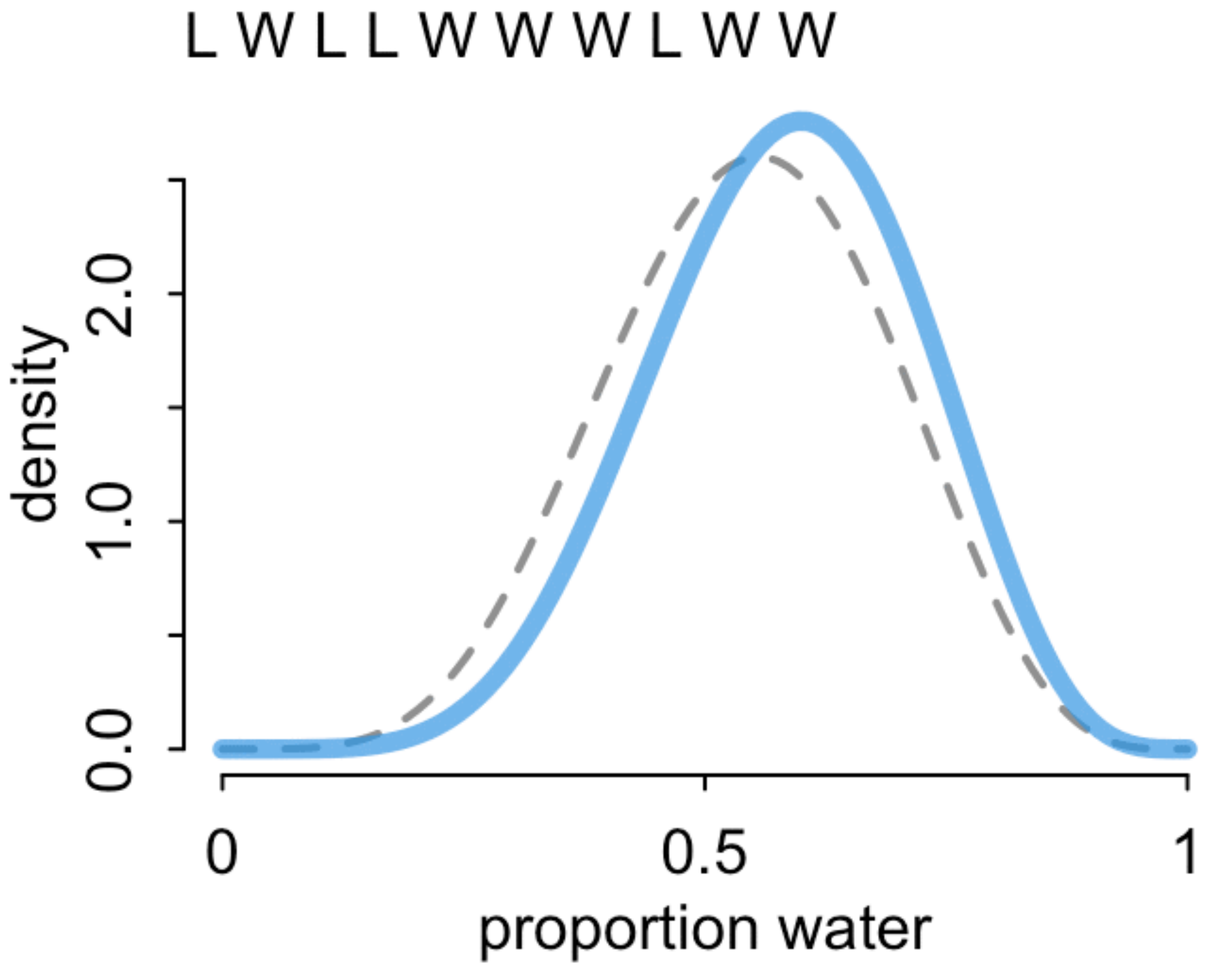


(1) No minimum sample size

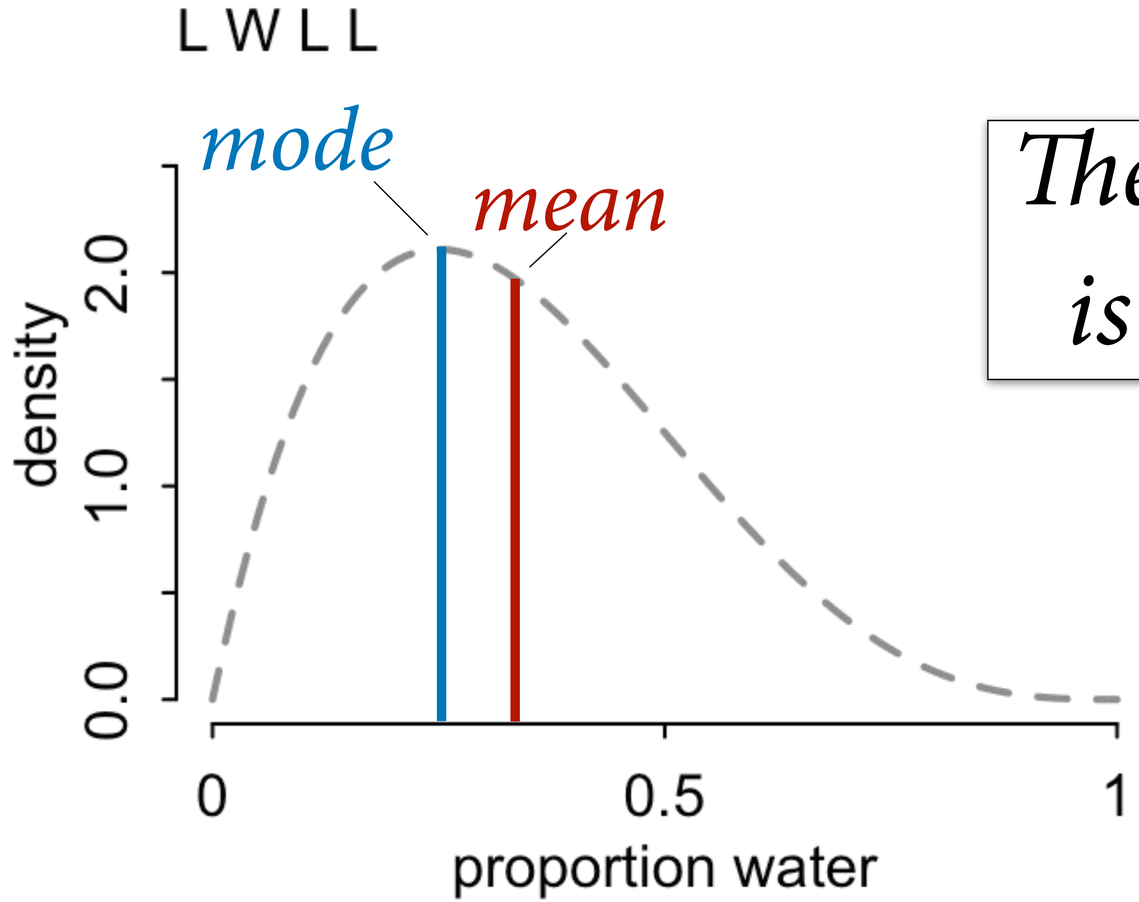




(2) Shape embodies sample size



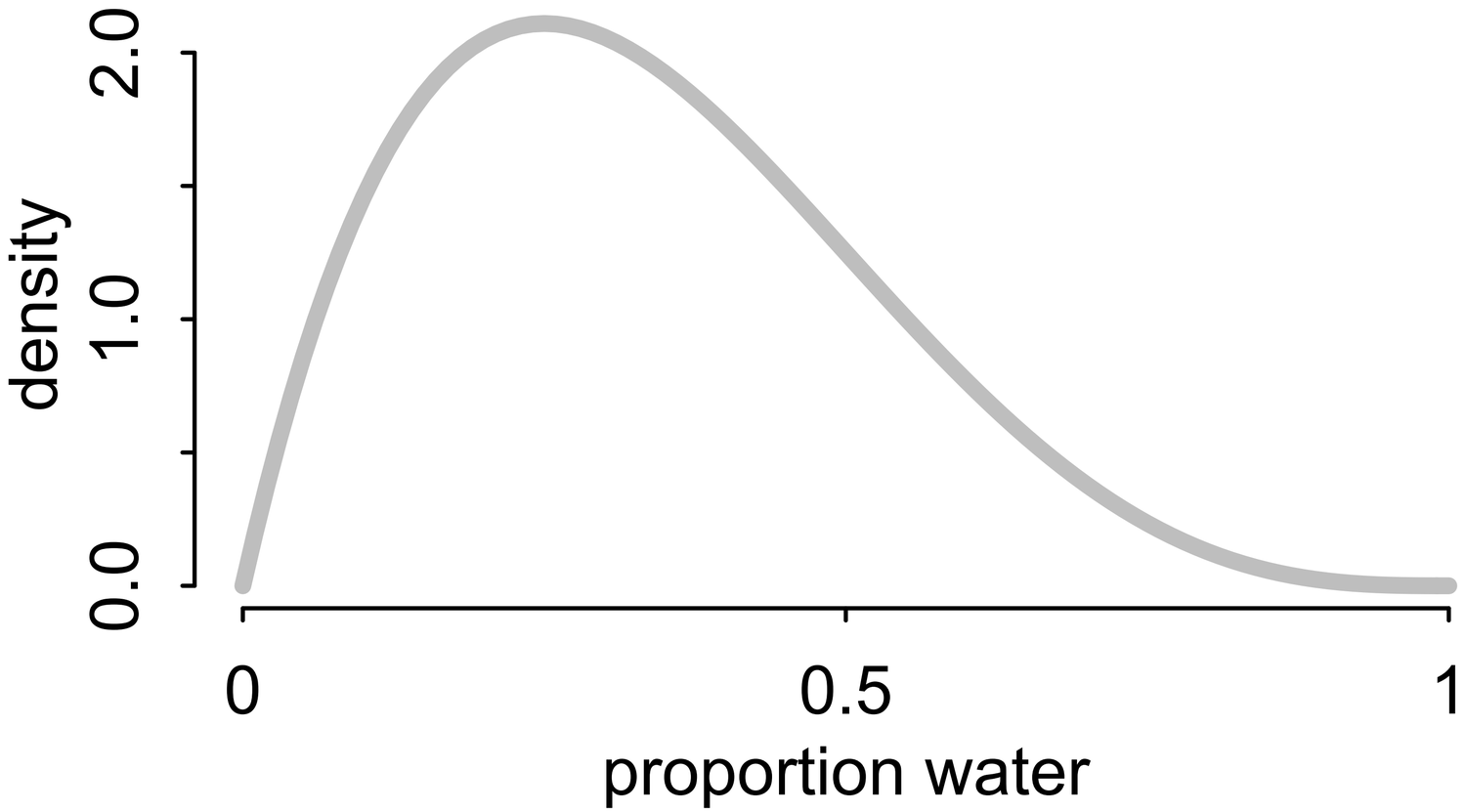
(3) No point estimate



*The **distribution** is the estimate*

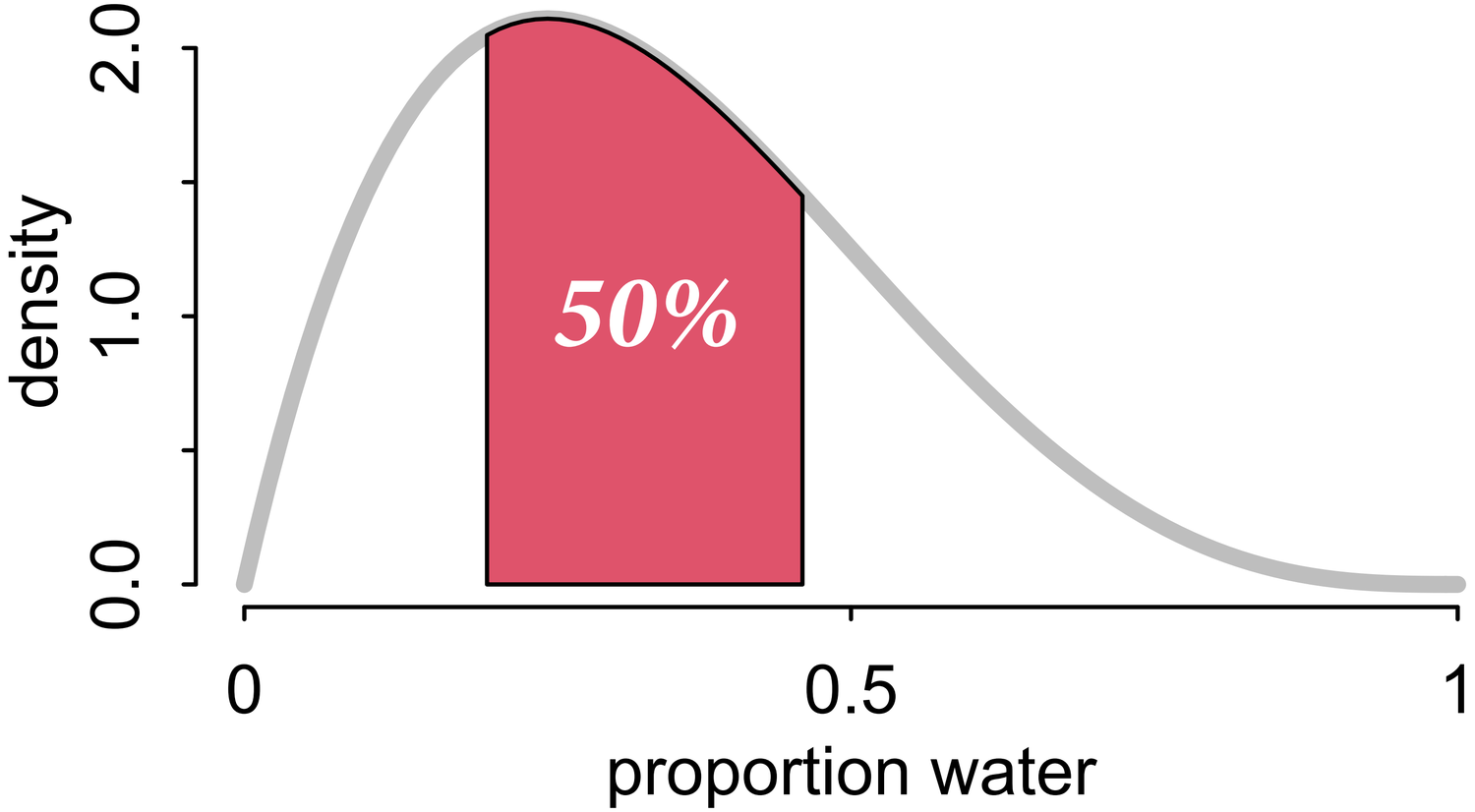
Always use the entire distribution

(4) No one true interval



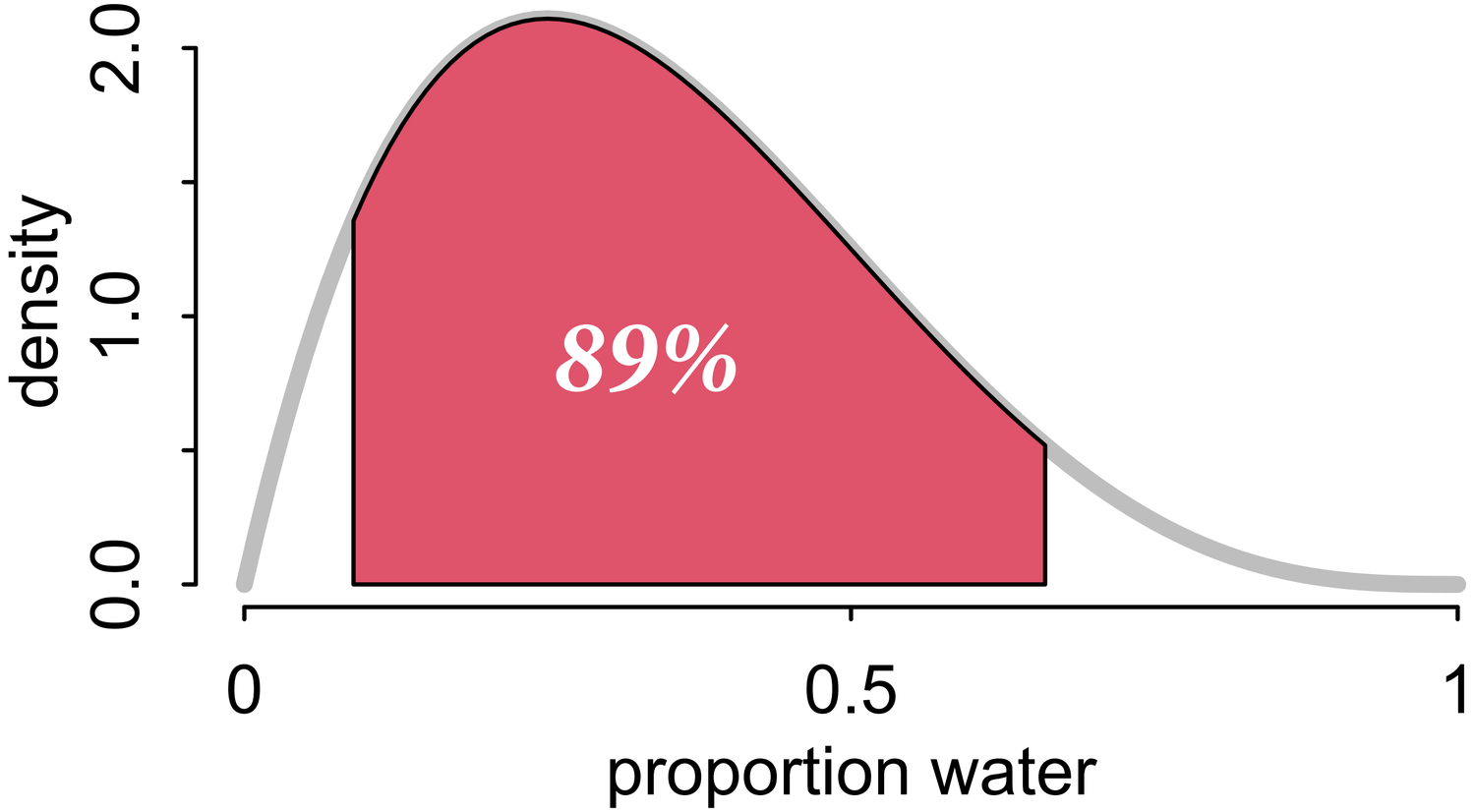
Intervals communicate shape of posterior

(4) No one true interval



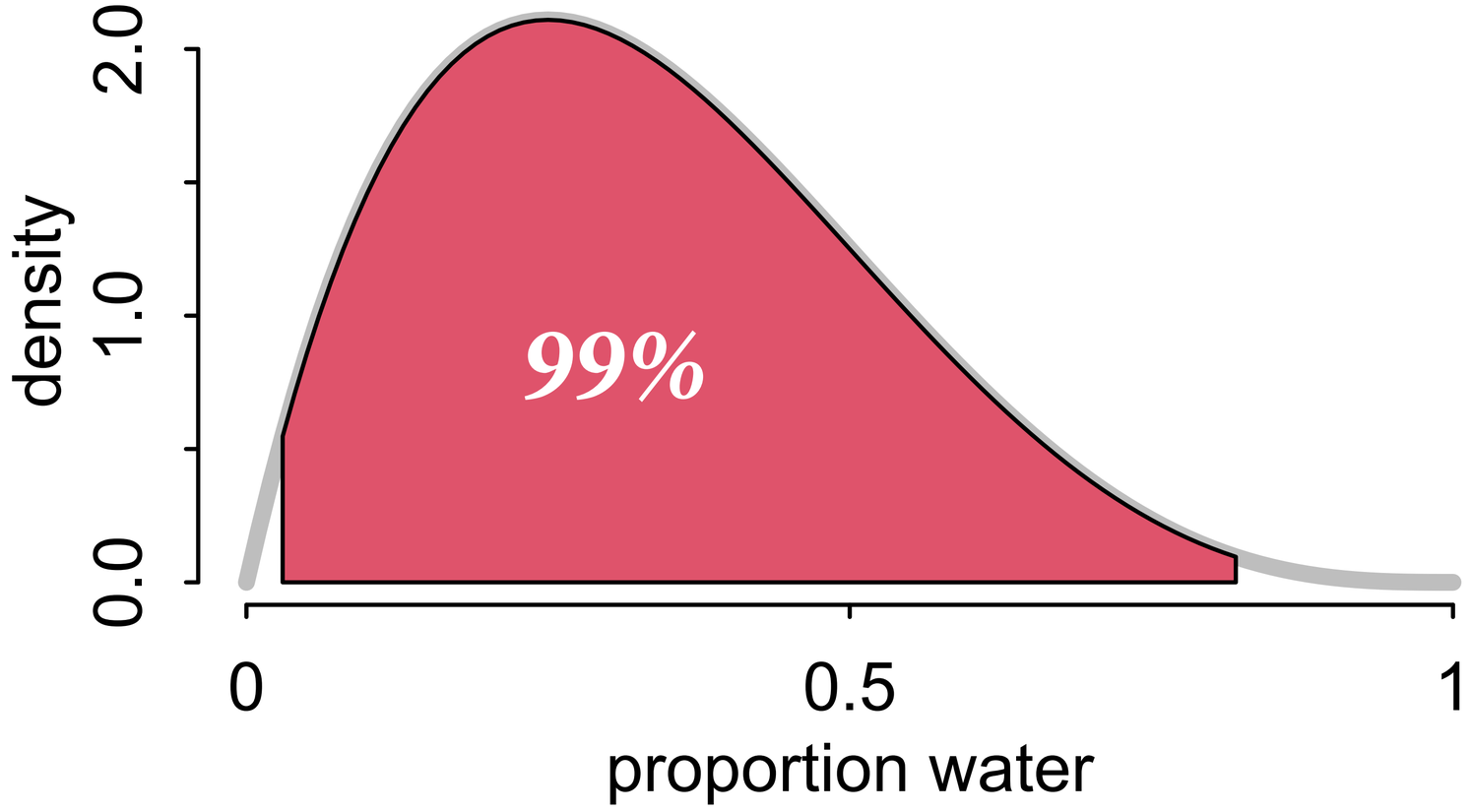
Intervals communicate shape of posterior

(4) No one true interval



Intervals communicate shape of posterior

(4) No one true interval

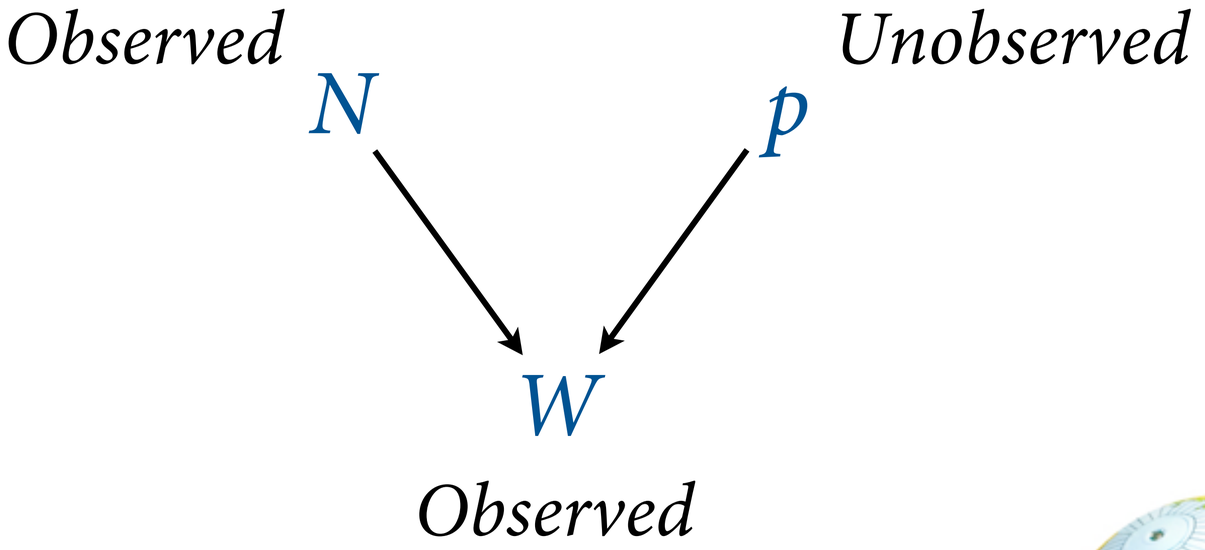


Intervals communicate shape of posterior

95% is obvious superstition. Nothing magical happens at the boundary.

Bayesian Modeling

Slides taken from the “Statistical rethinking” course


Observed
Unobserved
N
p
W
Observed

The Joint Model

$$W \sim \text{Binomial}(N, p)$$
$$p \sim \text{Uniform}(0, 1)$$

Components of the model

- Assume:
  (1) Likelihood
  (2) Parameters
  (3) Prior
- Deduce: Posterior

Likelihood

- Pr(data|assumptions)
  - Defines probability of each observation, conditional "|" on assumptions
  - i.e. relative count of number of ways of seeing data, given a particular conjecture
- In this case, binomial probability:

$$\Pr(n_{\mathrm{W}}|n,p) = \frac{n!}{n_{\mathrm{W}}!(n-n_{\mathrm{W}})!}p^{n_{\mathrm{W}}}(1-p)^{n-n_{\mathrm{W}}}$$

Parameters

- Likelihood contains symbols: n_W, n, p
- Some are data (n_W, n)
- Others parameters (p)
  - Define targets of inference, what is updated
  - These were the *conjectures* in the bag example
- Which are data and which parameters depend upon your context and question
  - e.g. mark-recapture: know n_W, must infer n, p

The Formalities

In practice, we write the model in a way that communicates all of the probability assumptions.

The observations (data) and explanations (parameters) are variables

For each variable, must say how it is generated

The Formalities

Data: W and L, the number of water and land observations

$$\Pr(W, L|p) = \frac{(W+L)!}{W!L!} p^W (1-p)^L$$

The number of ways to realize W,L given p

The Formalities

Data: W and L, the number of water and land observations

$$\Pr(W, L|p) = \frac{(W+L)!}{W!L!} p^W (1-p)^L$$

The number of ways to realize W,L given p

Parameters: p, the proportion of water on the globe

$$\Pr(p) = \frac{1}{1-0} = 1.$$

Relative plausibility of each possible p

The Formalities

$$\Pr(W, L|p) = \frac{(W+L)!}{W!L!} p^W (1-p)^L \qquad \Pr(p) = \frac{1}{1-0} = 1.$$

Posterior is (normalized) product:

$$\Pr(p|W, L) = \frac{\Pr(W, L|p)\Pr(p)}{\Pr(W, L)}$$

Relative plausibility of each possible p, after learning W,L

With Numbers

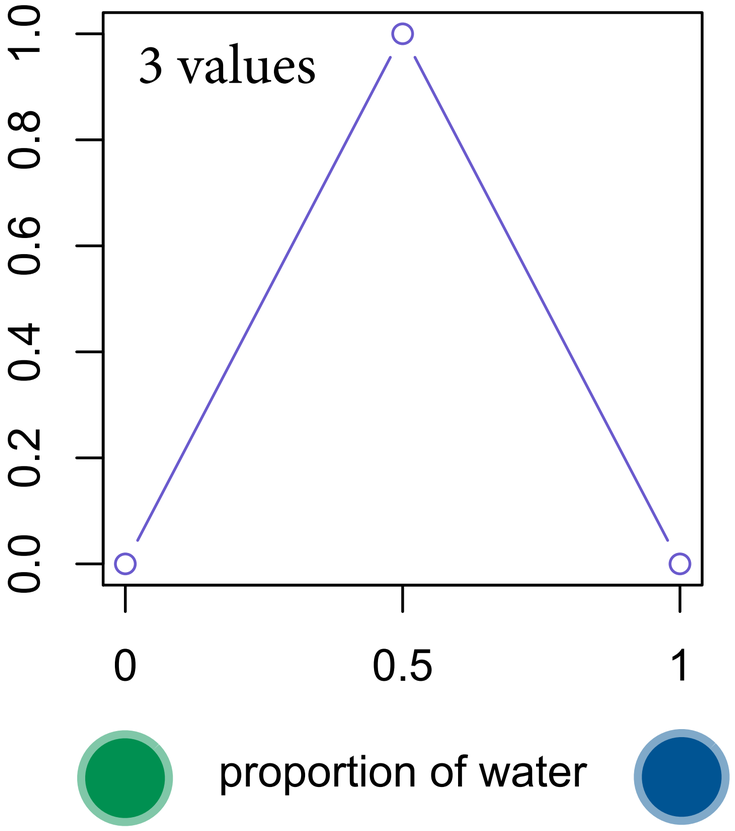
Ignore the mathematics for the moment and just draw the owl with numbers

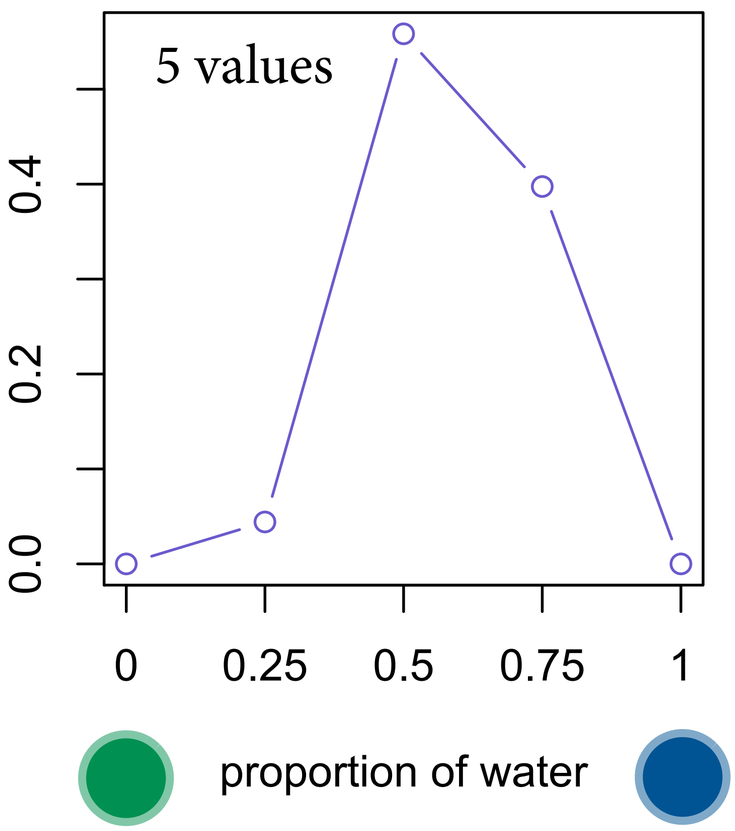
1. For each possible value of p

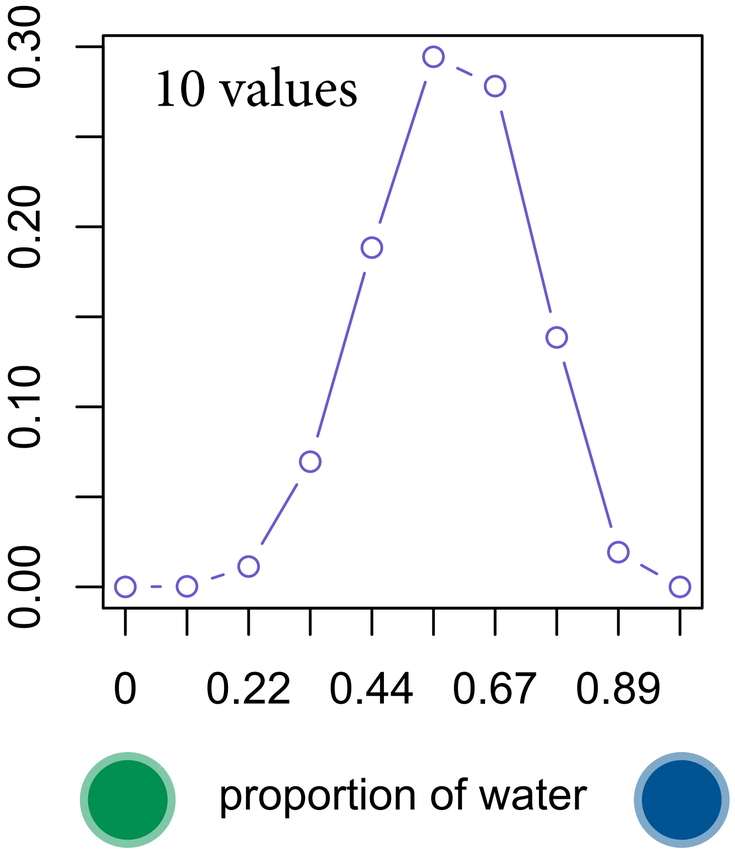
2. Compute product $\Pr(W,L|p)\Pr(p)$

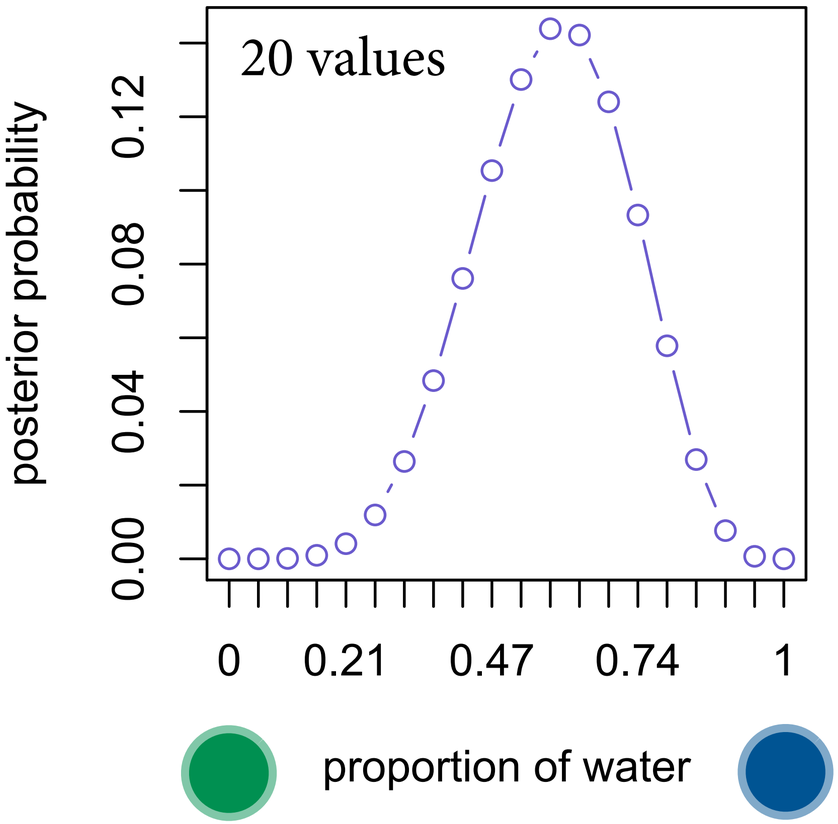
3. Relative sizes of products in (2) are posterior probabilities

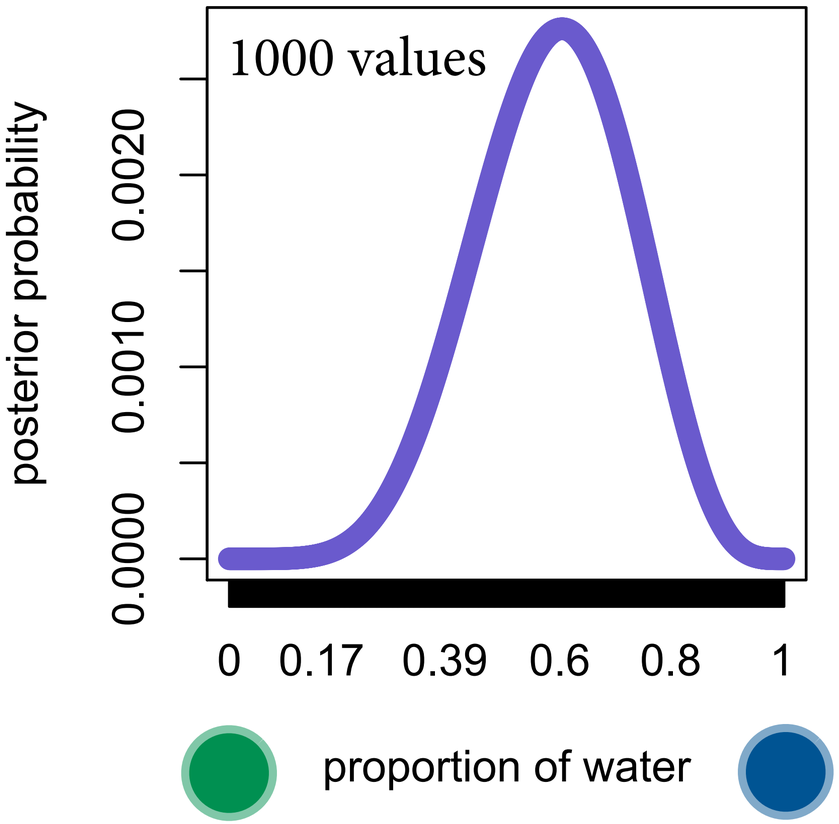
Bayesian owl


3 values
0.0
0.2
0.4
0.6
0.8
1.0
0
0.5
1
proportion of water


5 values
0.0
0.2
0.4
0
0.25
0.5
0.75
1
proportion of water


10 values
0.00
0.10
0.20
0.30
0
0.22
0.44
0.67
0.89
proportion of water


20 values
posterior probability
0.00
0.04
0.08
0.12
0
0.21
0.47
0.74
1
proportion of water


1000 values
posterior probability
0.0000
0.0010
0.0020
0
0.17
0.39
0.6
0.8
1
proportion of water

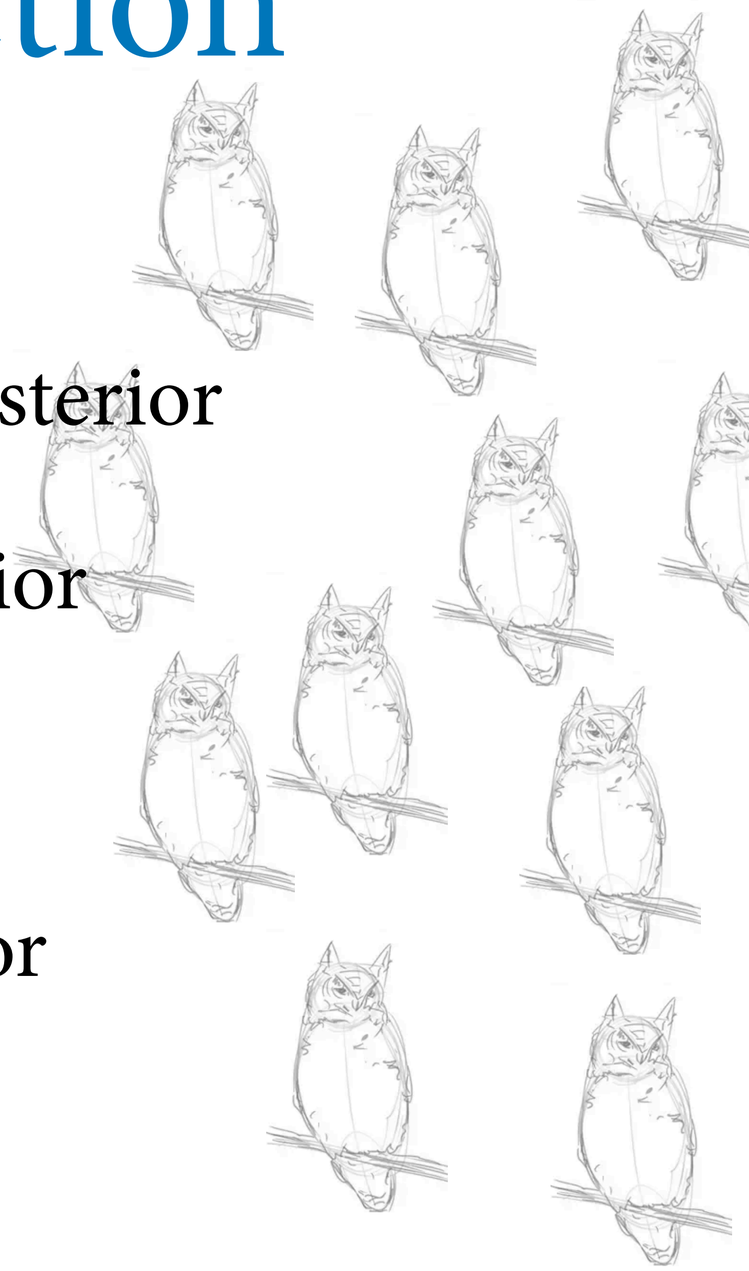
From Posterior to Prediction

Implications of model depend upon **entire** posterior

Must average any inference over entire posterior

This usually requires integral calculus

OR we can just take samples from the posterior



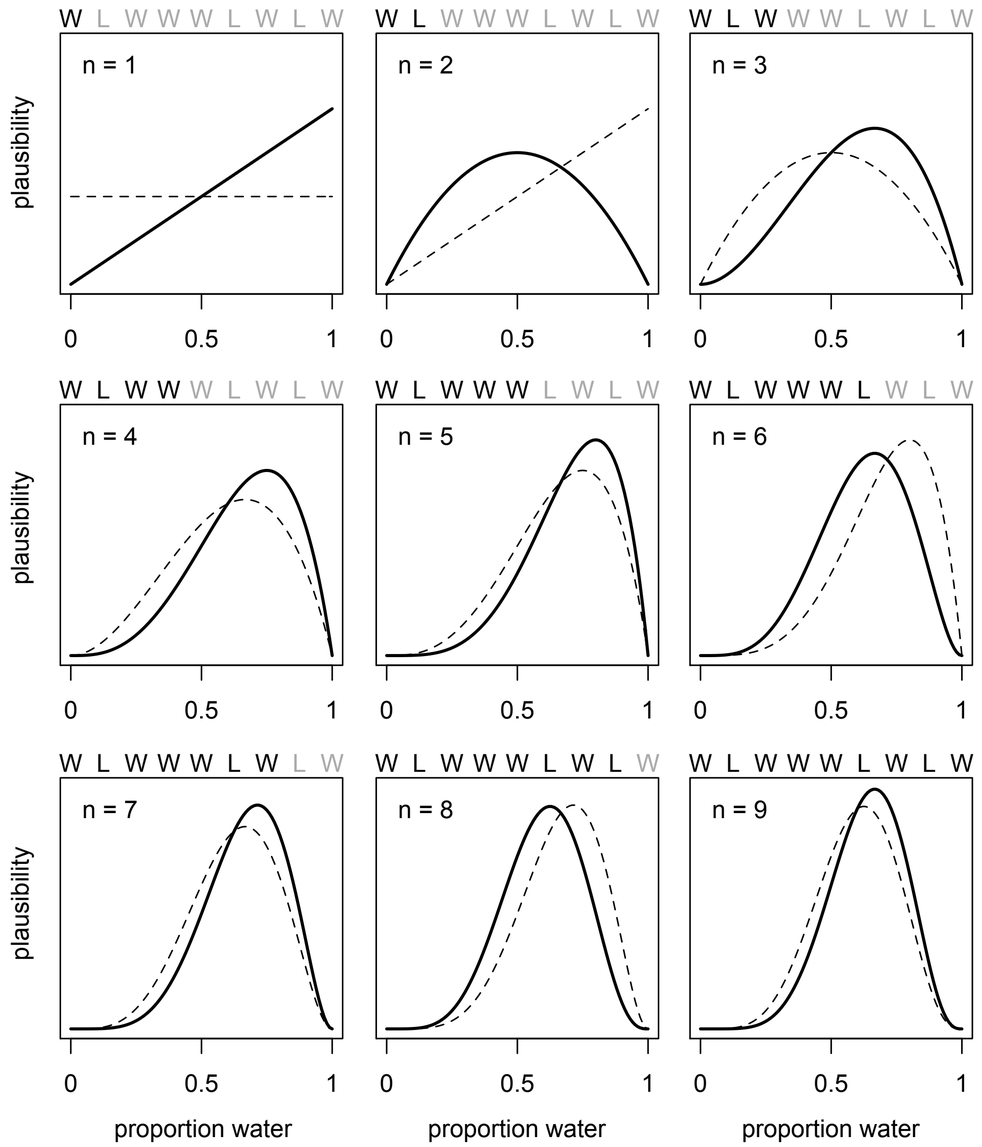
Bayesian data analysis

For each possible explanation of the data,

Count all the ways data can happen.

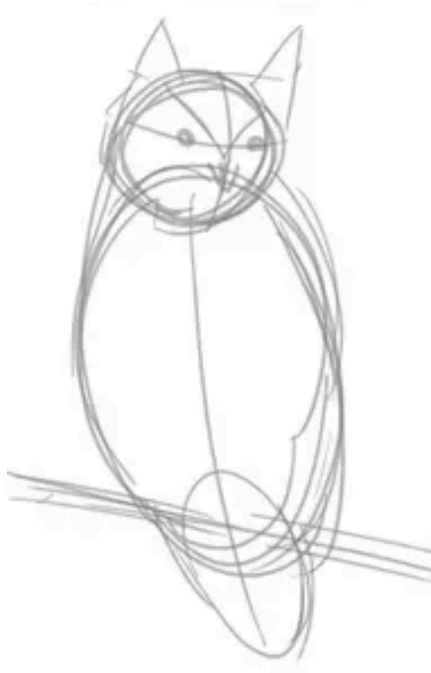
Explanations with more ways to produce the data are more plausible.

"How plausible is each proportion of water, given these data?"



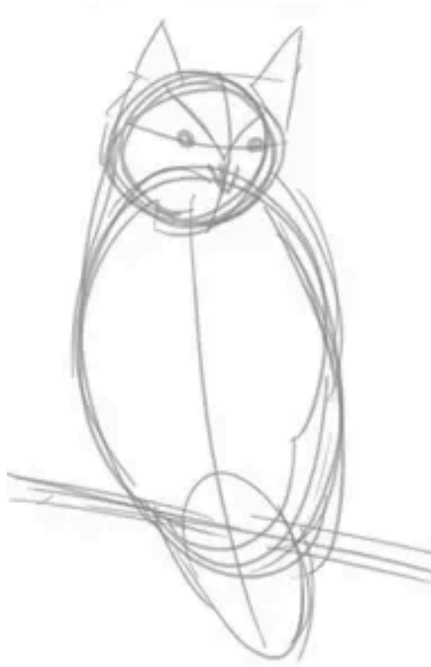
Now make it compute — arrange as probability statements

$$\Pr(W|N, p) = \text{Binomial}(W|N, p)$$
$$\Pr(p) = \text{Uniform}(p|0, 1)$$



Now make it compute — arrange as probability statements

$$\Pr(W|N, p) = \text{Binomial}(W|N, p)$$
$$\Pr(p) = \text{Uniform}(p|0, 1)$$



Posterior distribution

$$\Pr(p|W, N) \propto \text{Binomial}(W|N, p)\ \text{Uniform}(p|0,1)$$

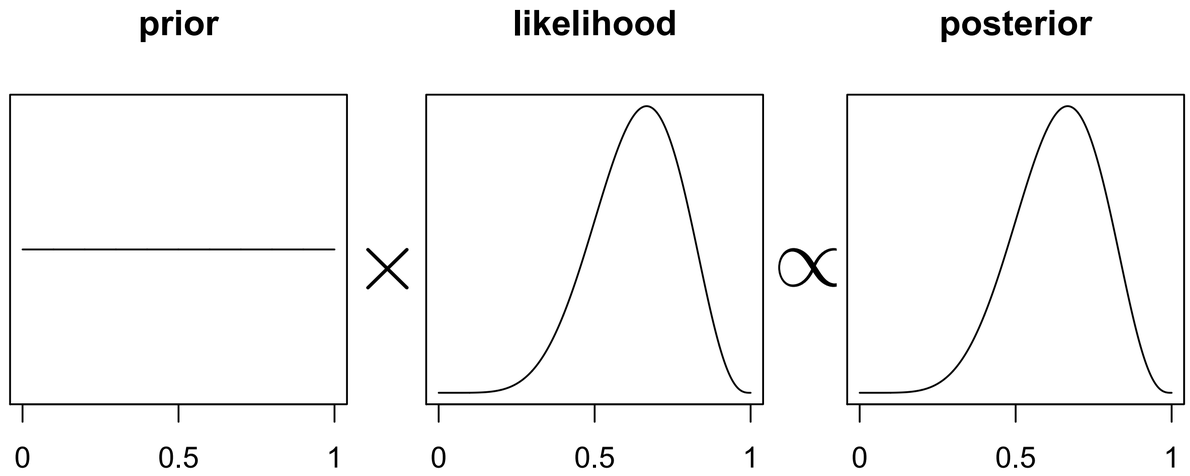
"proportional to"

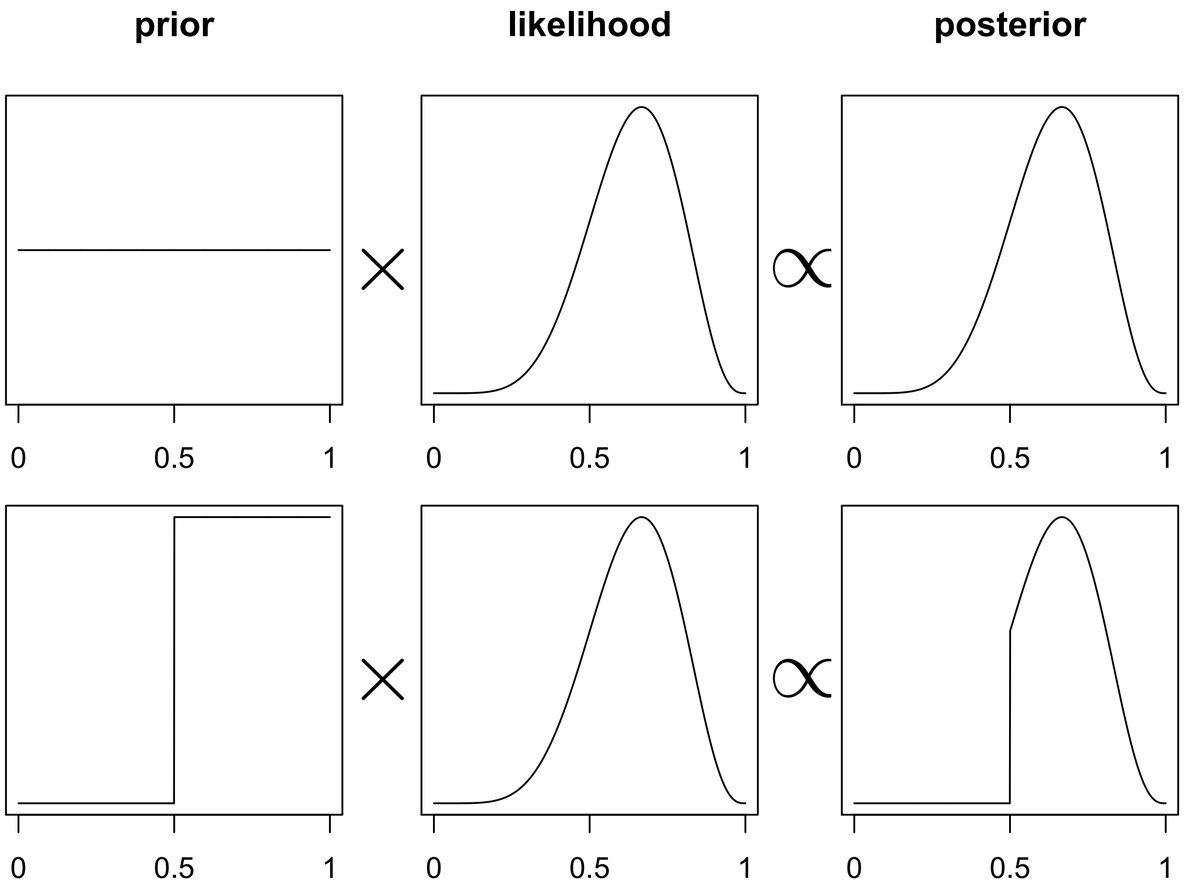
Posterior probability

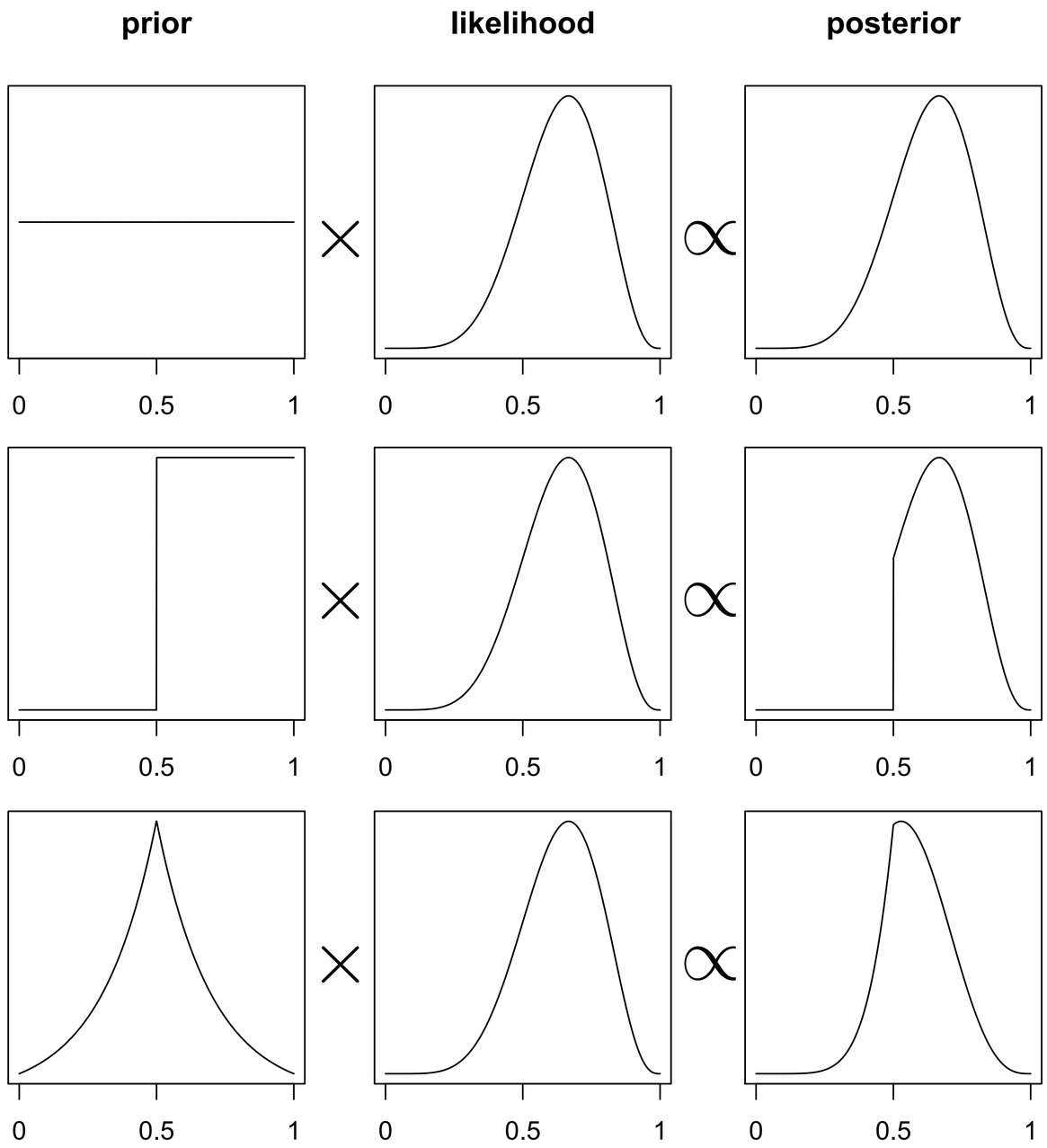
- Bayesian "estimate" is always *posterior distribution over parameters*, Pr(parameters|data)
- Here: Pr($p|W,N$)
- Compute using *Bayes' theorem*:

$$\Pr(p|W,N) = \frac{\Pr(W|N,p)\Pr(p)}{\sum \Pr(W|N,p)\Pr(p) \text{ for all } p}$$

$$\text{Posterior} = \frac{(\text{Prob observed variables}) \times (\text{Prior})}{\text{Normalizing constant}}$$

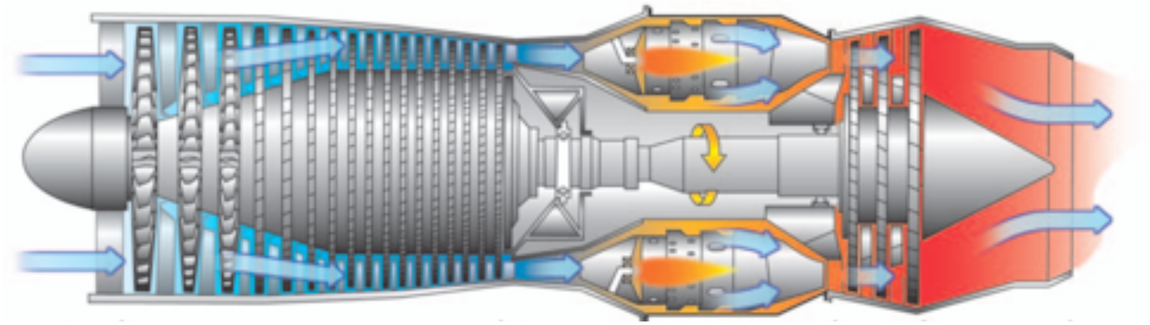

prior
likelihood
posterior
0
0.5
1
0
0.5
1
0
0.5
1


prior
likelihood
posterior
0
0.5
1
0
0.5
1
0
0.5
1
0
0.5
1
0
0.5
1
0
0.5
1


prior
likelihood
posterior
0
0.5
1
×
∝

Computing the posterior

1. Analytical approach (often impossible)
2. Grid approximation (very intensive)
3. Quadratic approximation (limited)
4. Markov chain Monte Carlo (intensive)



Sampling from the posterior

- Incredibly useful to *sample randomly* from the posterior
  - Visualize uncertainty
  - Compute confidence intervals
  - Simulate observations
- MCMC produces only samples
- Above all, *easier to think with samples*
- Transforms a hard calculus problem into an easy data summary problem

Once you have posterior, new questions arise

- How much posterior probability lies below some parameter value?
- How much posterior probability lies between two parameter values?
- Which parameter value marks the lower 5% of the posterior probability?
- Which range of parameter values contains 90% of the posterior probability?
- Which parameter value has highest posterior probability?

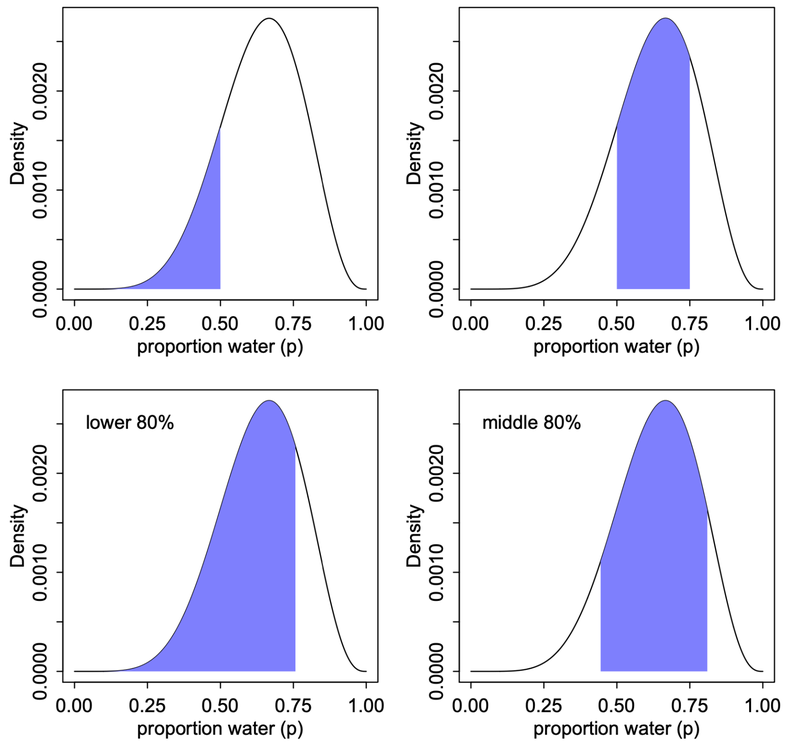
Groups of questions

- Questions about:
  - (1) intervals of defined boundaries,
  - (2) intervals of defined probability mass, and
  - (3) point estimates.
- Sampling from the posterior helps to answer those questions

Intervals

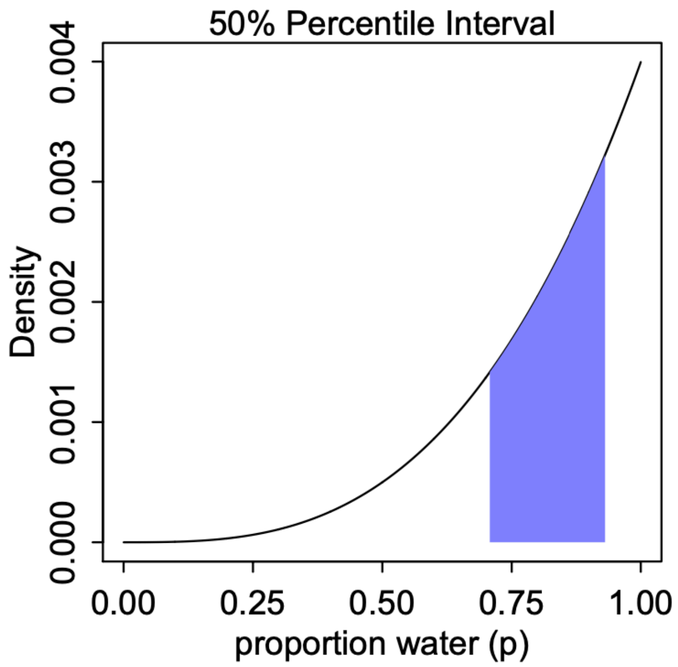
- defined boundaries: say, what is probability that proportion of water is less than 50%?
  - Sample from grid approximation and add up posterior probability where $p < 0.5$

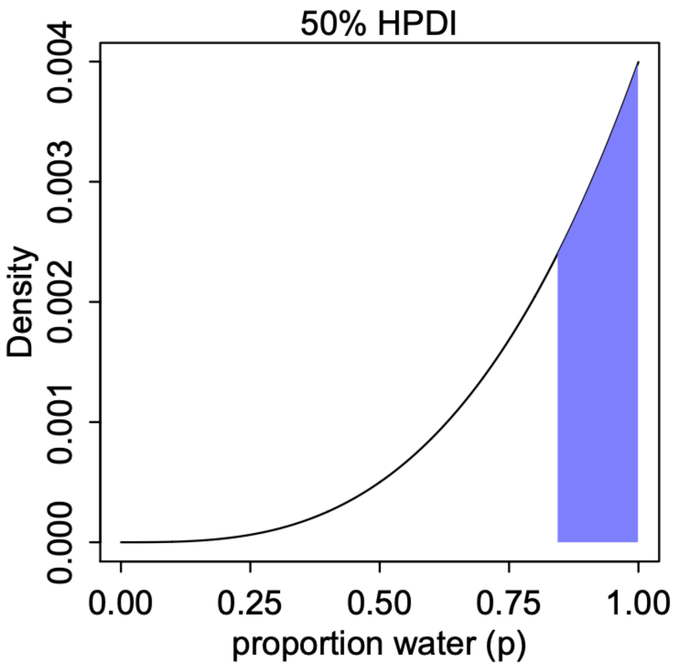
- defined mass:
  - here you have options



Percentile interval (PI) and HPDI

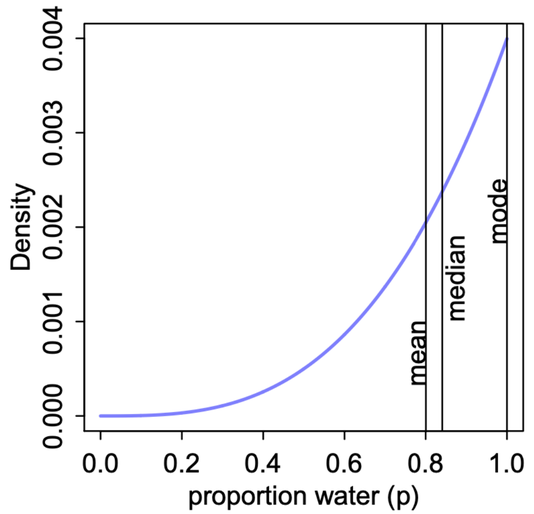
- 50% PI assigns equal mass (25%) to both the left and right tail
- 50% HDPI is the highest posterior density interval, HPD

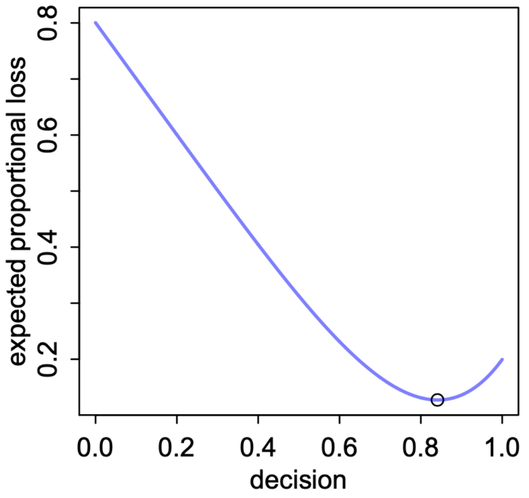




Point estimates of a parameter

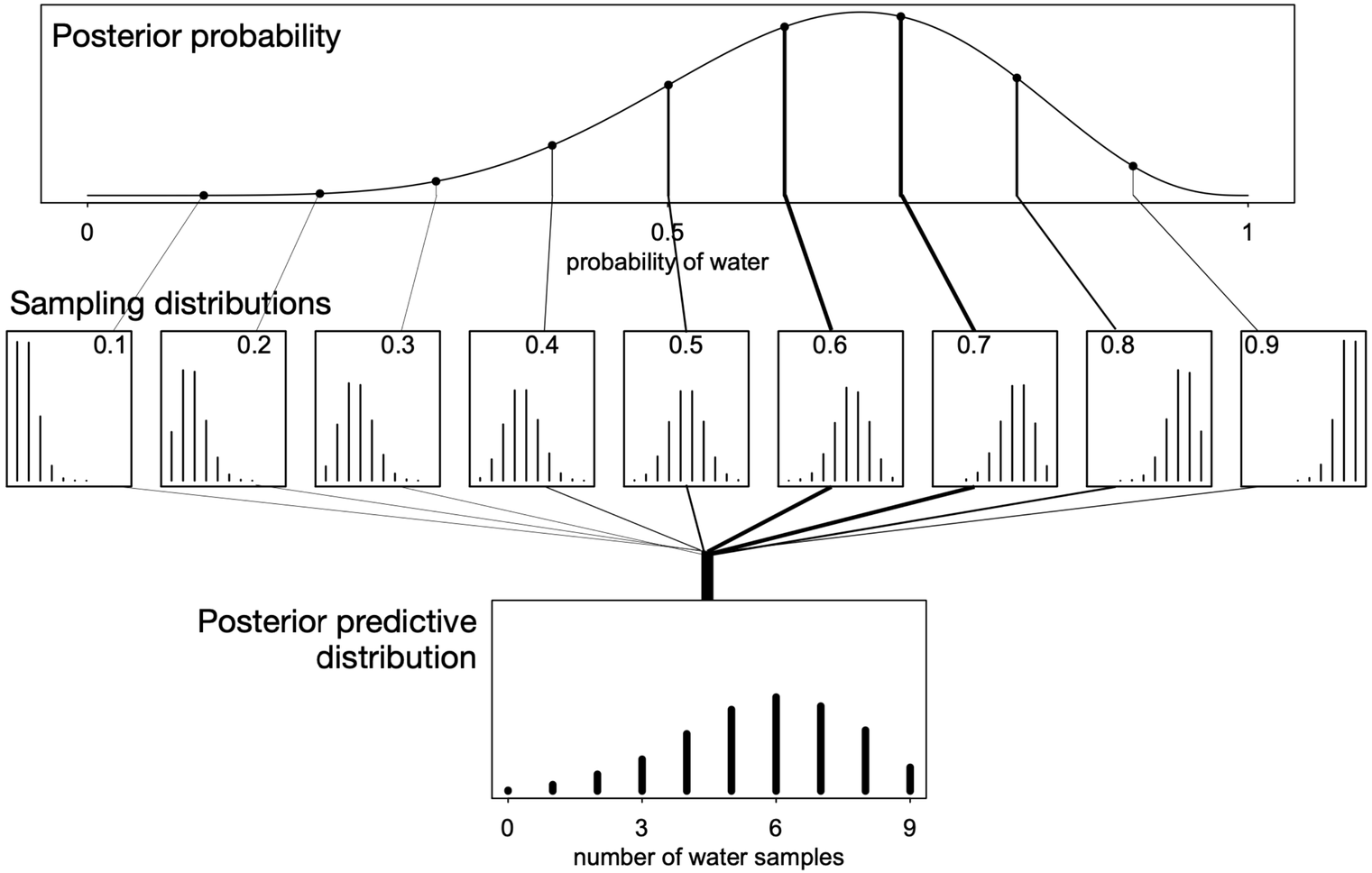
- if you must produce a single point to summarize the posterior
  - then you have many options
- parameter value with highest posterior probability,
  - a maximum a posteriori (MAP) estimate.
- there are also mean, median..
- depending on the loss function:
  - absolute error —> medain
  - mse —> mean





Simulating predictions from the total posterior

- This distribution propagates uncertainty about parameter to uncertainty about prediction



Linear Regression (bayesian approach)

Slides taken from the “Statistical rethinking” course
https://speakerdeck.com/rmcelreath/

Linear Regression

Model of **mean** and **variance** of variable

Mean as **weighted sum** of other variables

Why Normal?

Two arguments

(1) Generative: Summed fluctuations tend towards normal distribution

(2) Statistical: For estimating mean and variance, normal distribution is least informative distribution (maxent)

Variable does not have to be normally distributed for normal model to be useful

Making Normal Models

Goals:

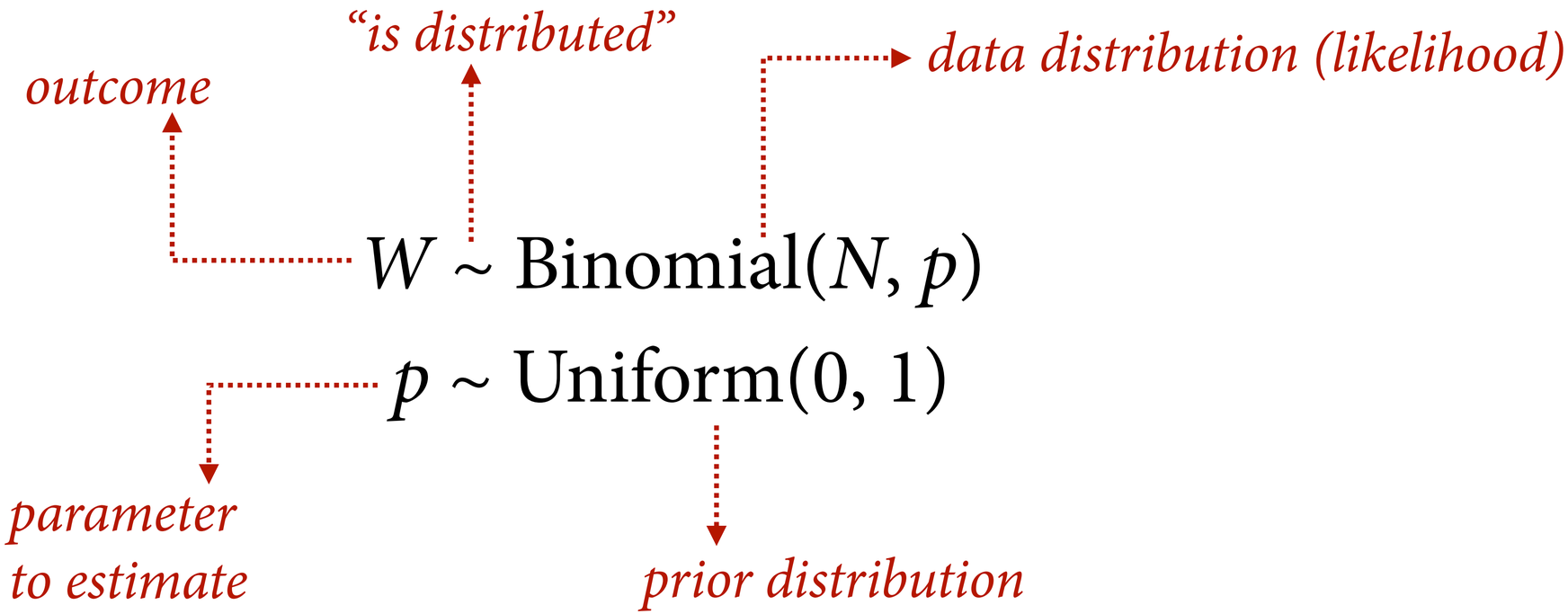
(1) Language for representing models

(2) How to calculate bigger posterior distributions

(3) Constructing & understanding linear models

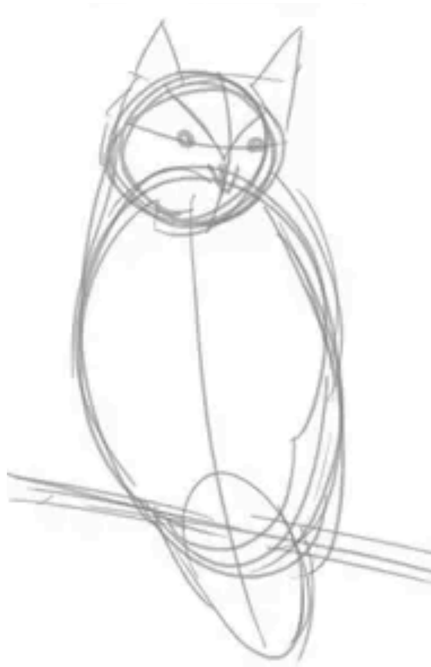
Language for modeling

Revisit globe tossing model:



Now make it compute — arrange as probability statements

$$\Pr(W|N, p) = \text{Binomial}(W|N, p)$$
$$\Pr(p) = \text{Uniform}(p|0, 1)$$



Posterior distribution

$$\Pr(p|W, N) \propto \text{Binomial}(W|N, p)\ \text{Uniform}(p|0,1)$$

"proportional to"

Linear Regression

```
library(rethinking)
data(Howell1)
```

Drawing the Owl

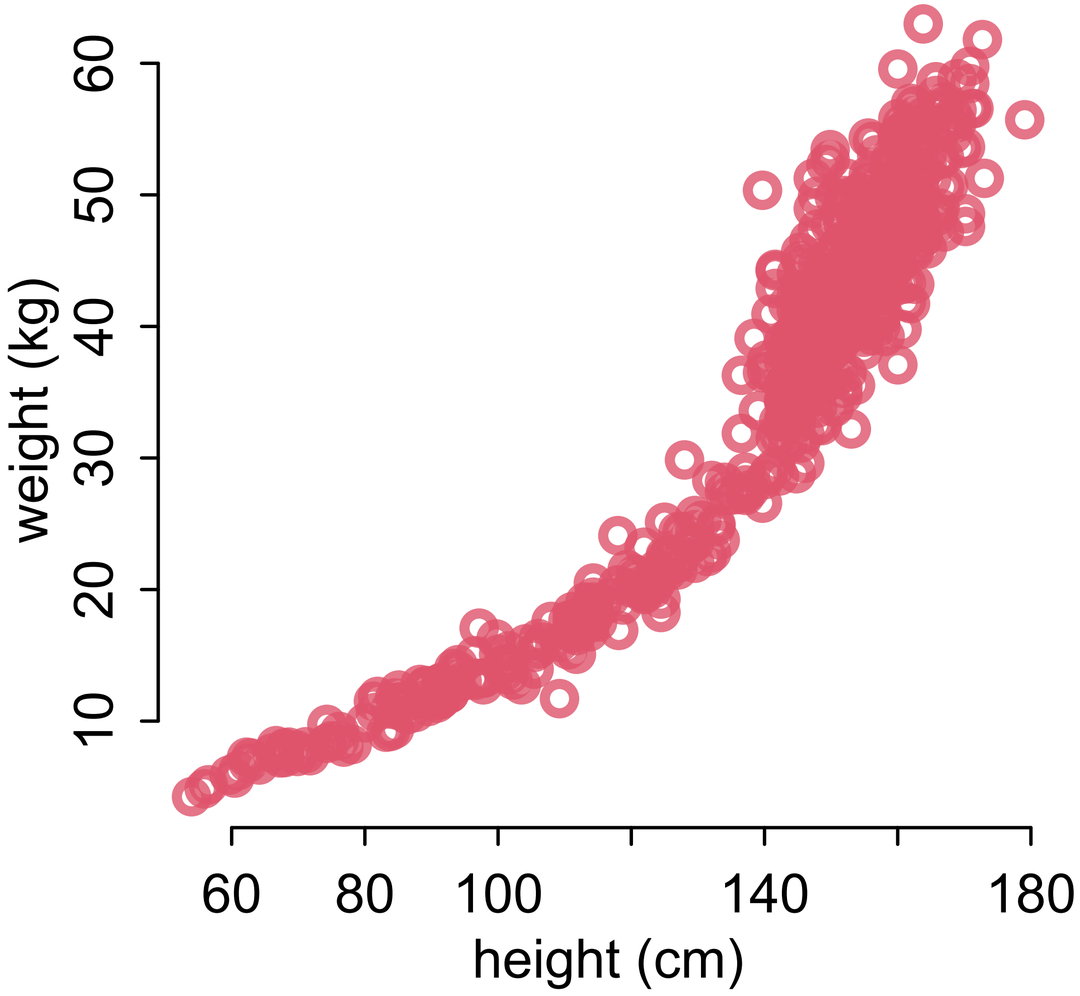
(1) Question/goal/estimand

(2) Scientific model

(3) Statistical model(s)

(4) Validate model

(5) Analyze data



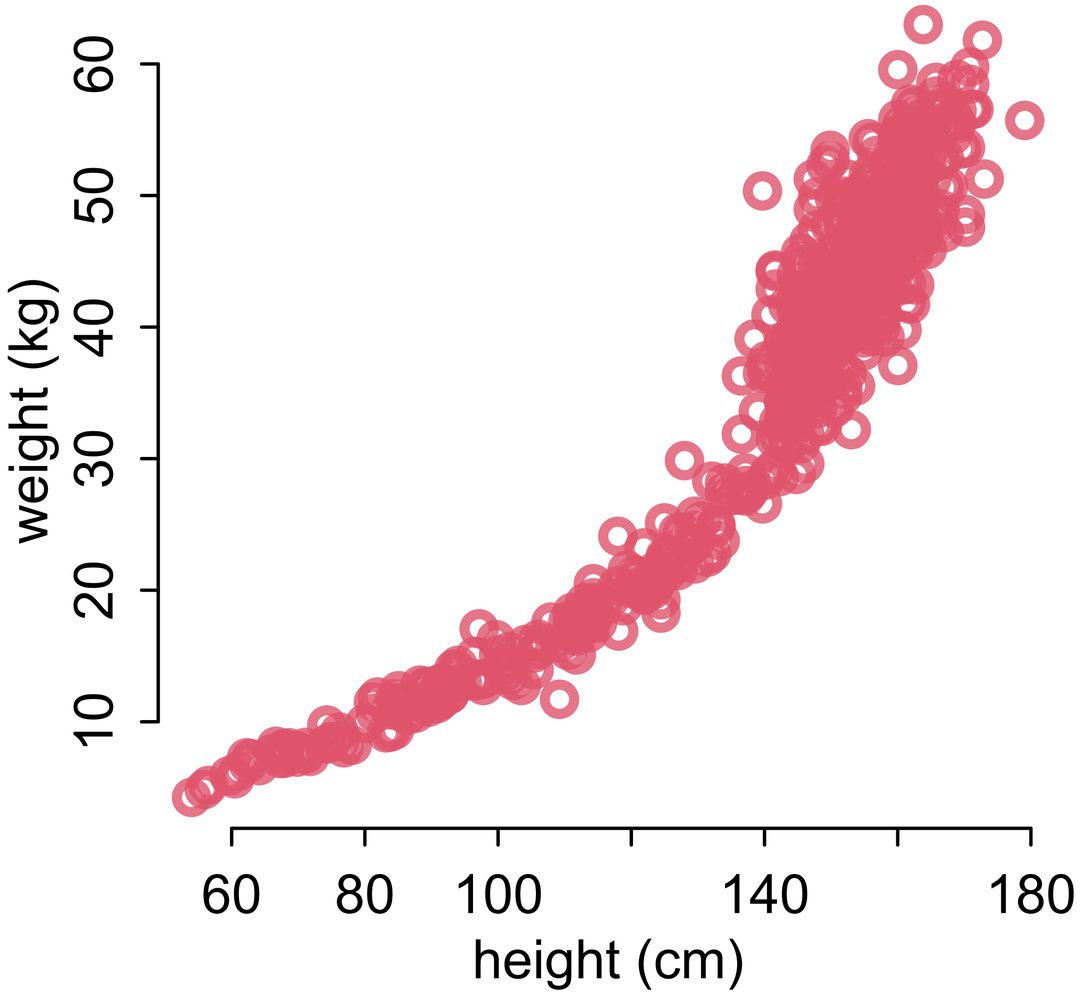
Linear Regression

```
library(rethinking)
data(Howell1)
```

Drawing the Owl

(1) Question/goal/estimand

Describe association between **weight** and **height**



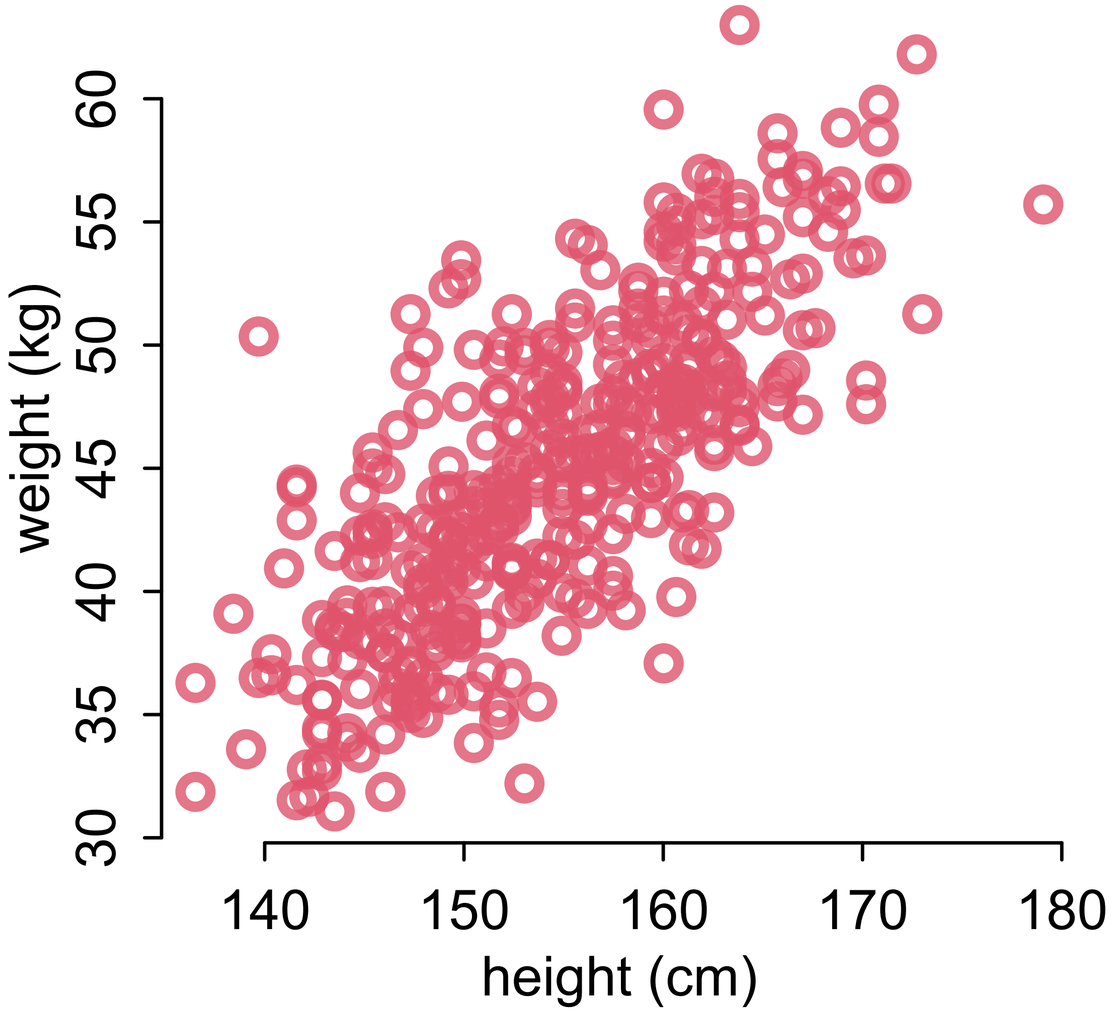
Linear Regression

```
data(Howell1)
d <- Howell1[Howell1$age>=18,]
```

Drawing the Owl

(1) Question/goal/estimand

Describe association between ADULT **weight** and **height**



Linear Regression

```
data(Howell1)
d <- Howell1[Howell1$age>=18,]
```

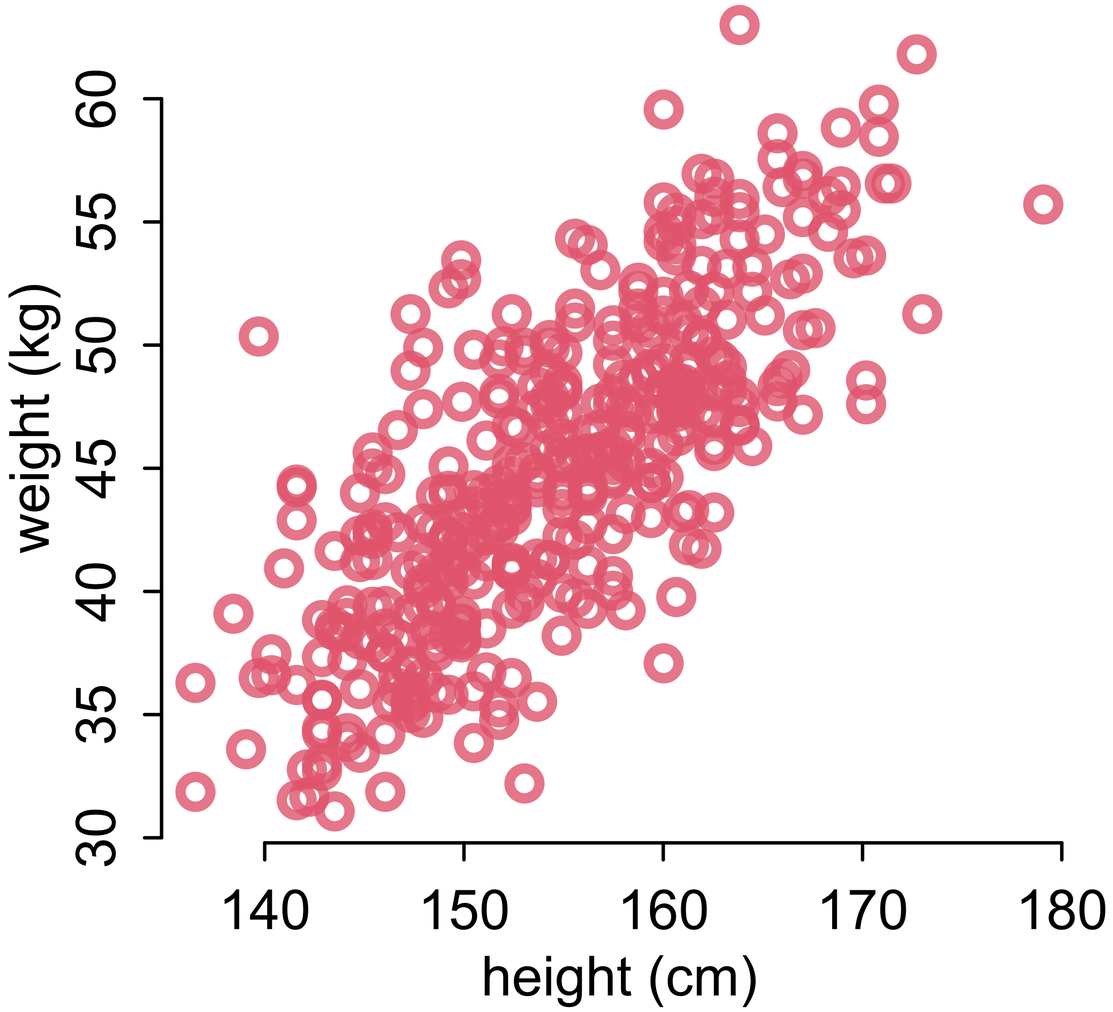
(2) Scientific model

How does **height** influence **weight**?

$$H \longrightarrow W$$

$$W = f(H)$$

"Weight is some function of height"



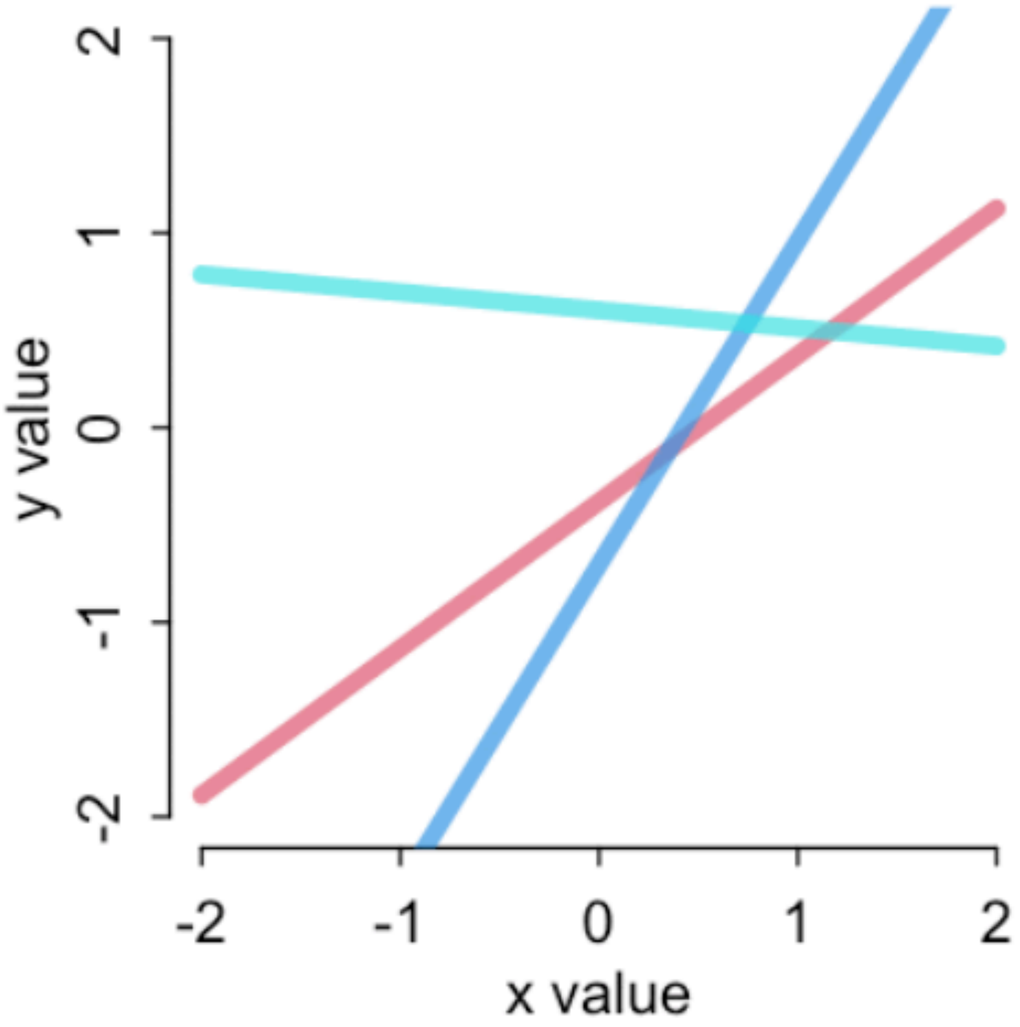
Anatomy of a linear model

$$y_i = \alpha + \beta x_i$$

index

intercept

slope



Anatomy of a linear model

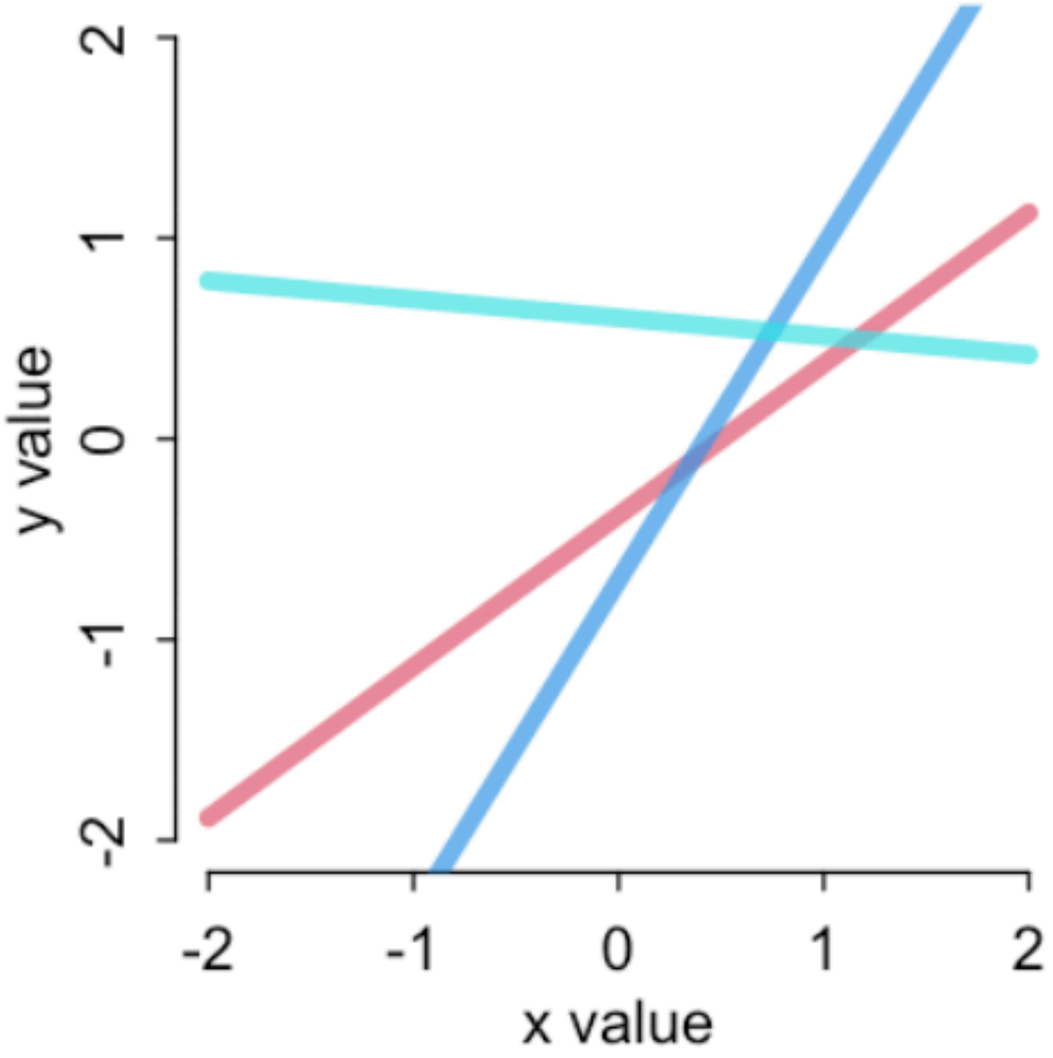
standard deviation

expectation

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta x_i$$

"Each x value has a different expectation, $E(y|x) = \mu$"



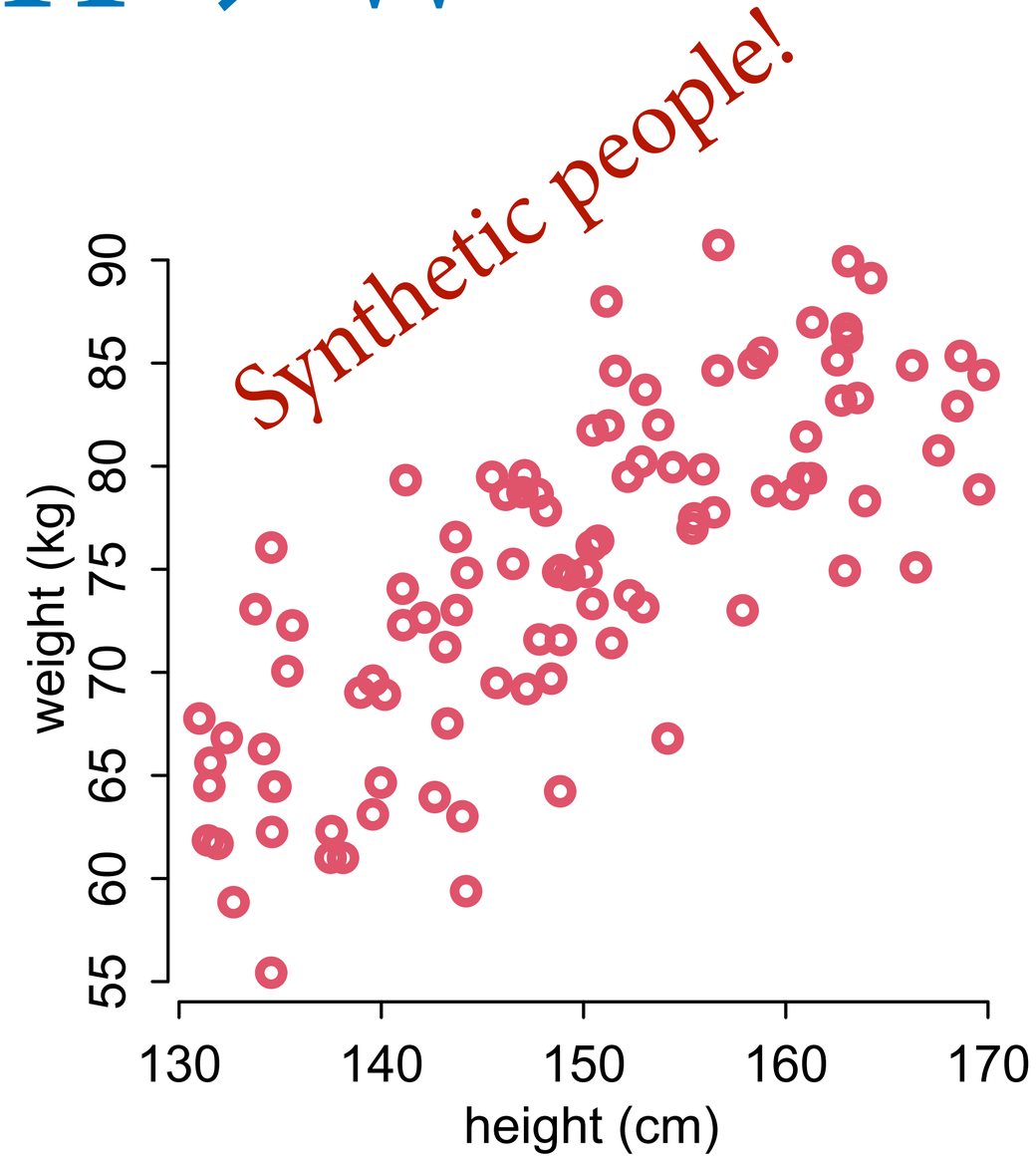
Generative model: H → W

$$W_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta H_i$$

```
alpha <- 0
beta <- 0.5
sigma <- 5
n_individuals <- 100

H <- runif(n_individuals,130,170)

mu <- alpha + beta*H
W <- rnorm(n_individuals,mu,sigma)
```



Linear Regression

```
data(Howell1)
d <- Howell1[Howell1$age>=18,]
```

Drawing the Owl

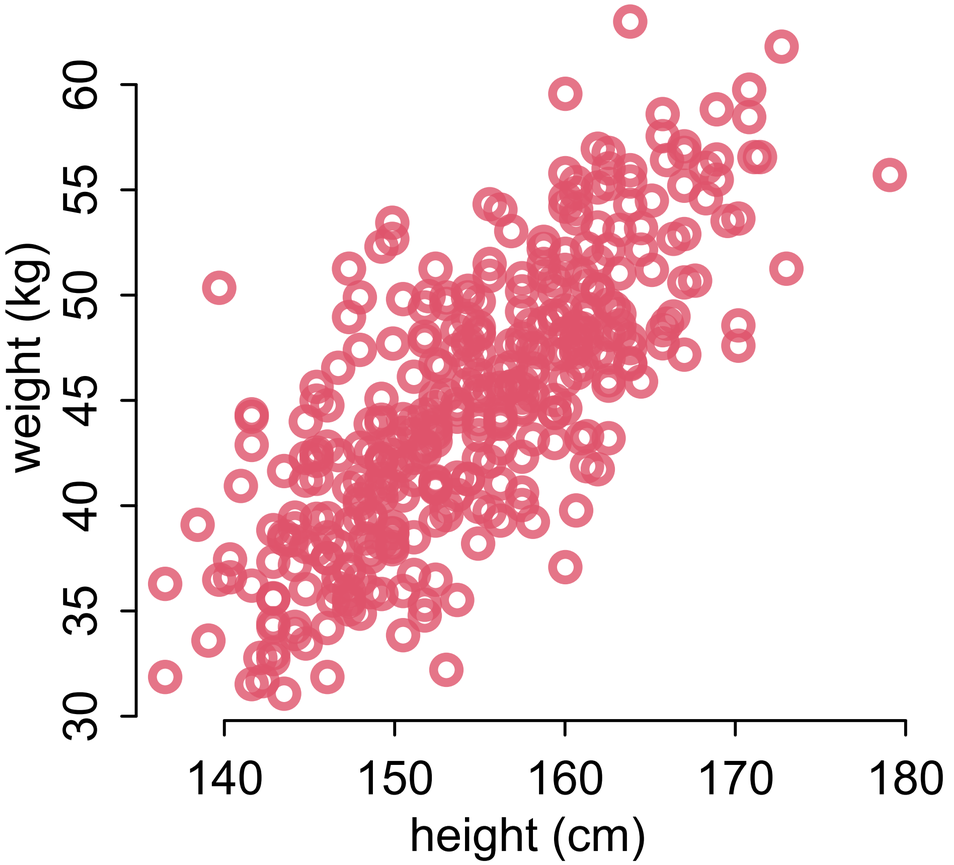
(1) Question/goal/estimand

(2) Scientific model

(3) Statistical model(s)

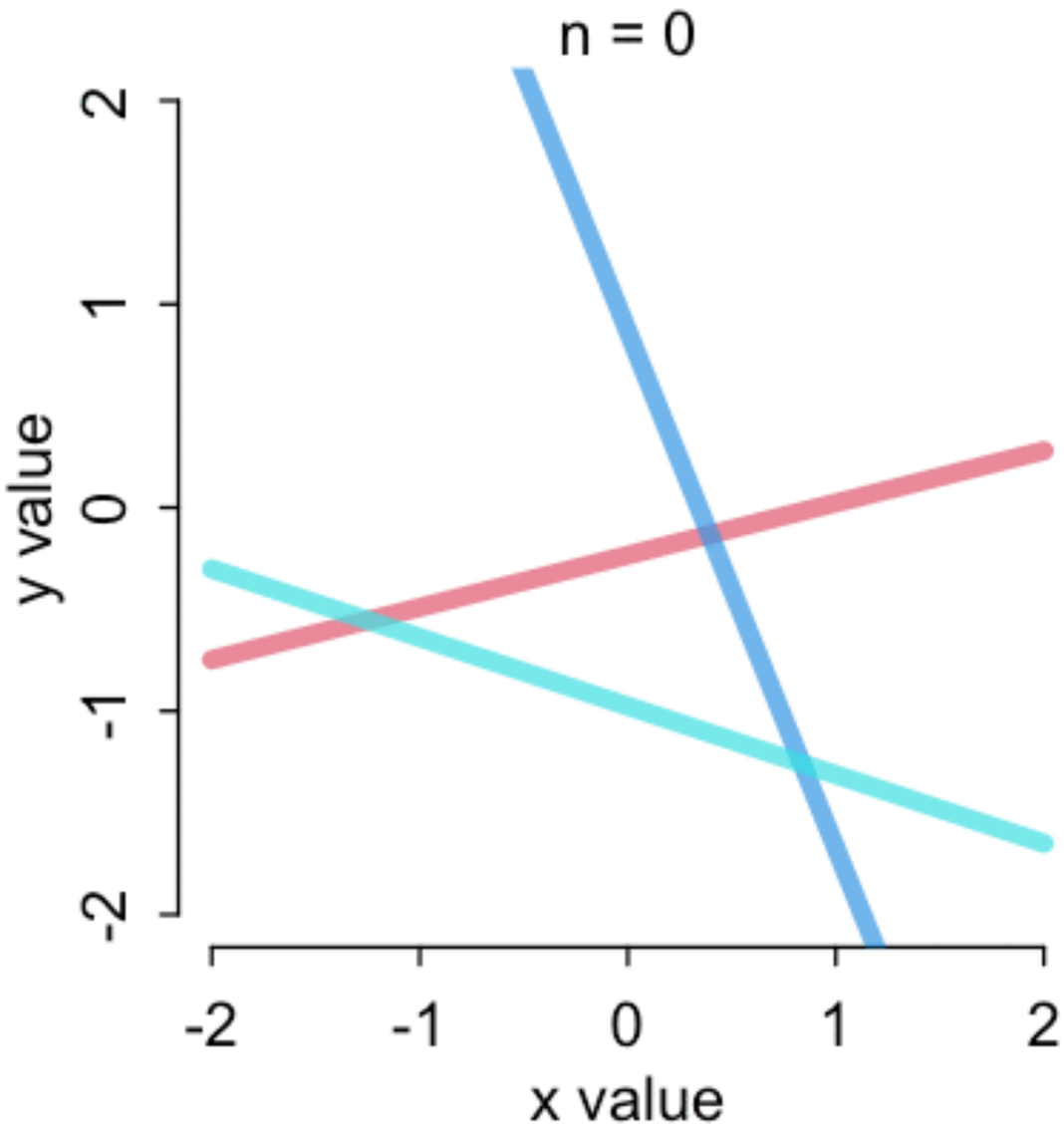
(4) Validate model

(5) Analyze data

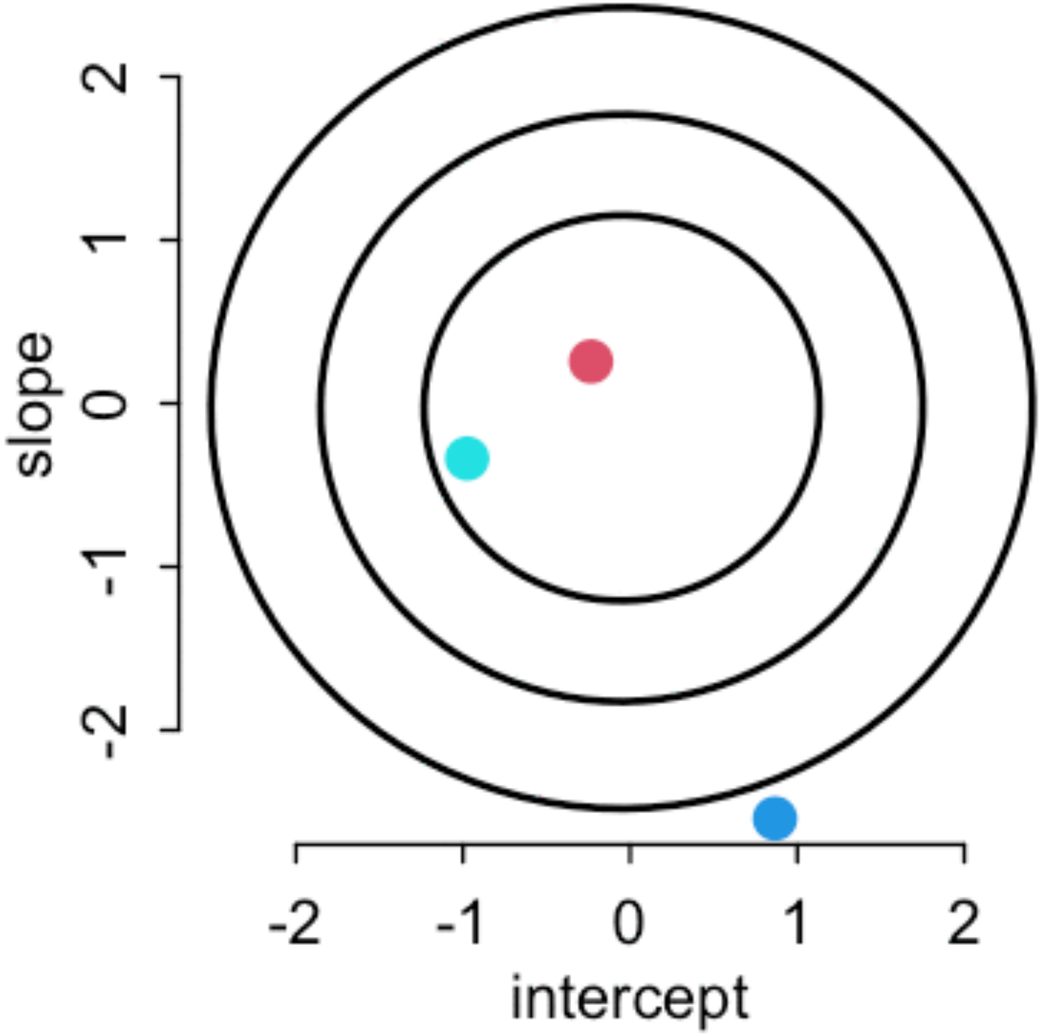


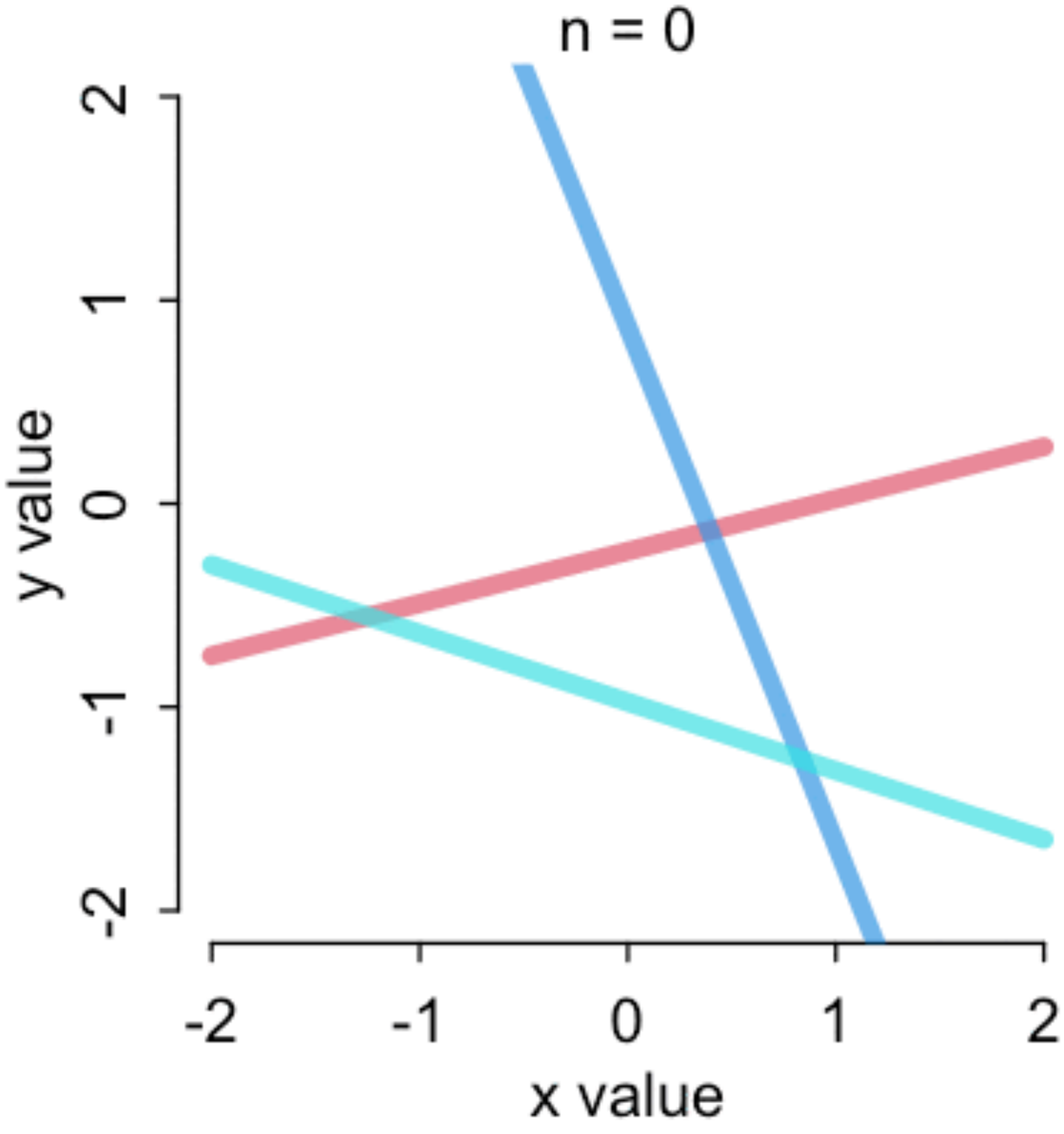
Anatomy of a linear model

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta x_i$$
$$\alpha \sim \text{Normal}(0,1)$$
$$\beta \sim \text{Normal}(0,1)$$
$$\sigma \sim \text{Uniform}(0,1)$$

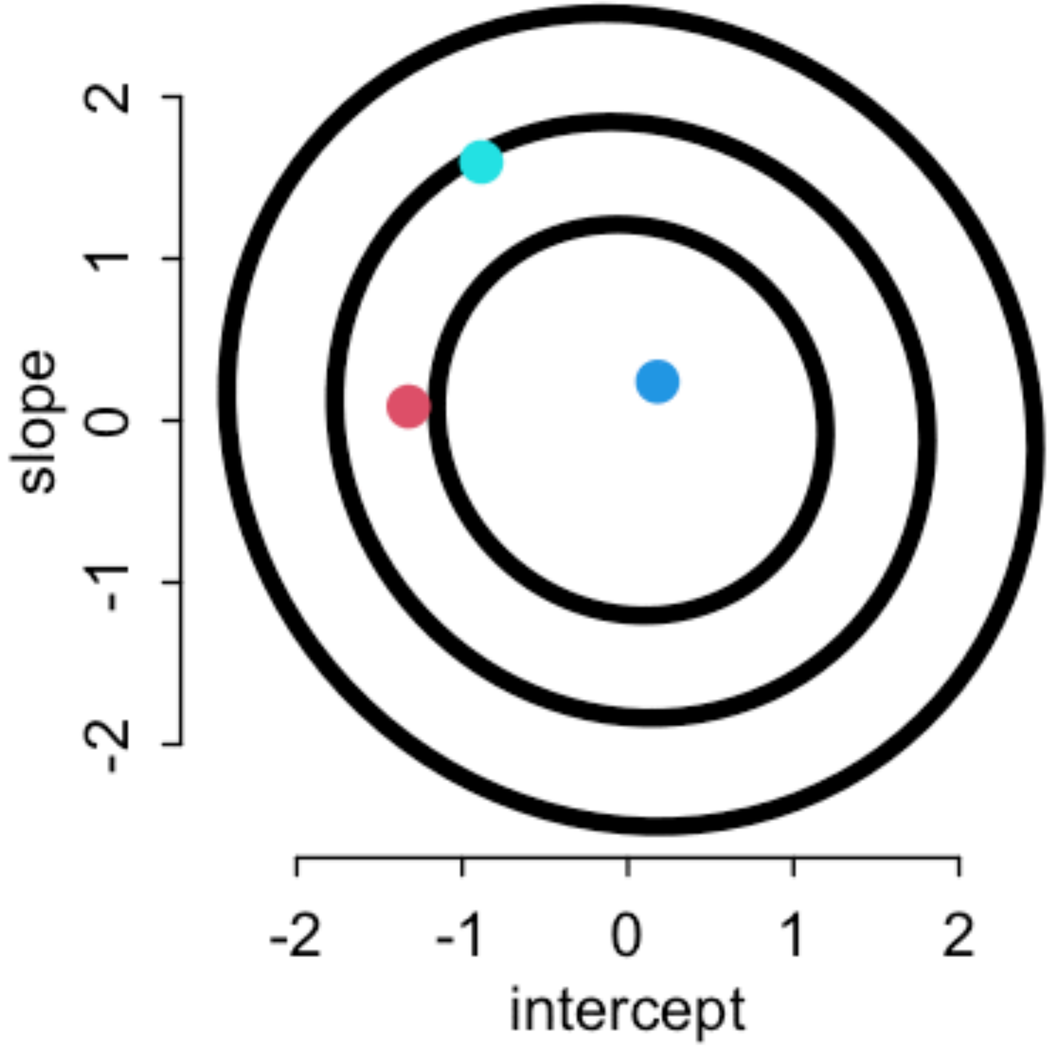


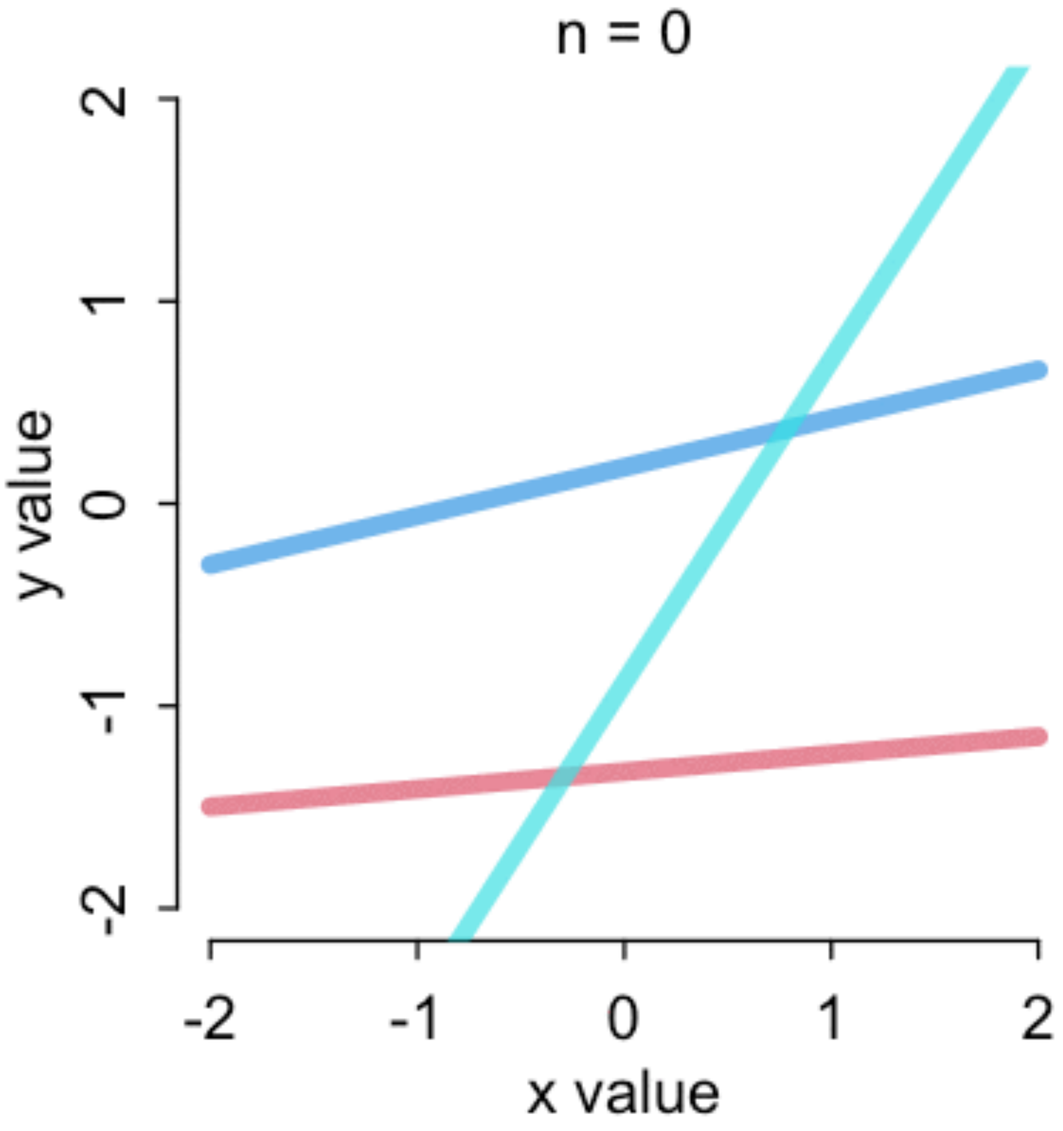
Sampling the prior distribution





Updating the posterior





Statistical model for H → W

Structure of statistical model similar to generative model, BUT

(1) Useful to re-scale variables

(2) Must think about priors

These two things go together

$$W_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta(H_i - \bar{H})$$
$$\alpha \sim \text{Normal}(?, ?)$$
$$\beta \sim \text{Normal}(?, ?)$$
$$\sigma \sim \text{Uniform}(0, ?)$$

Statistical model for H → W

Re-scaling **height** so that the **intercept** makes sense

$$W_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta(H_i - \bar{H})$$

value of μ when $H_i - \bar{H} = 0$

mean value of H_i

Statistical model for H → W

Now what are scientifically reasonable priors?

α: average adult weight

β: kilograms per centimeter

Region	Adult population (millions)	Average weight
Africa	535	60.7 kg (133.8 lb)
Asia	2,815	57.7 kg (127.2 lb)

$$W_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta(H_i - \bar{H})$$

$$\alpha \sim \text{Normal}(60,10)$$

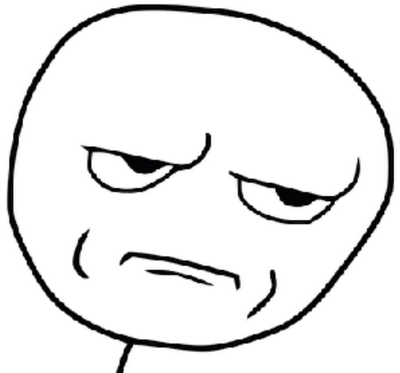
$$\beta \sim \text{Normal}(0,10)$$

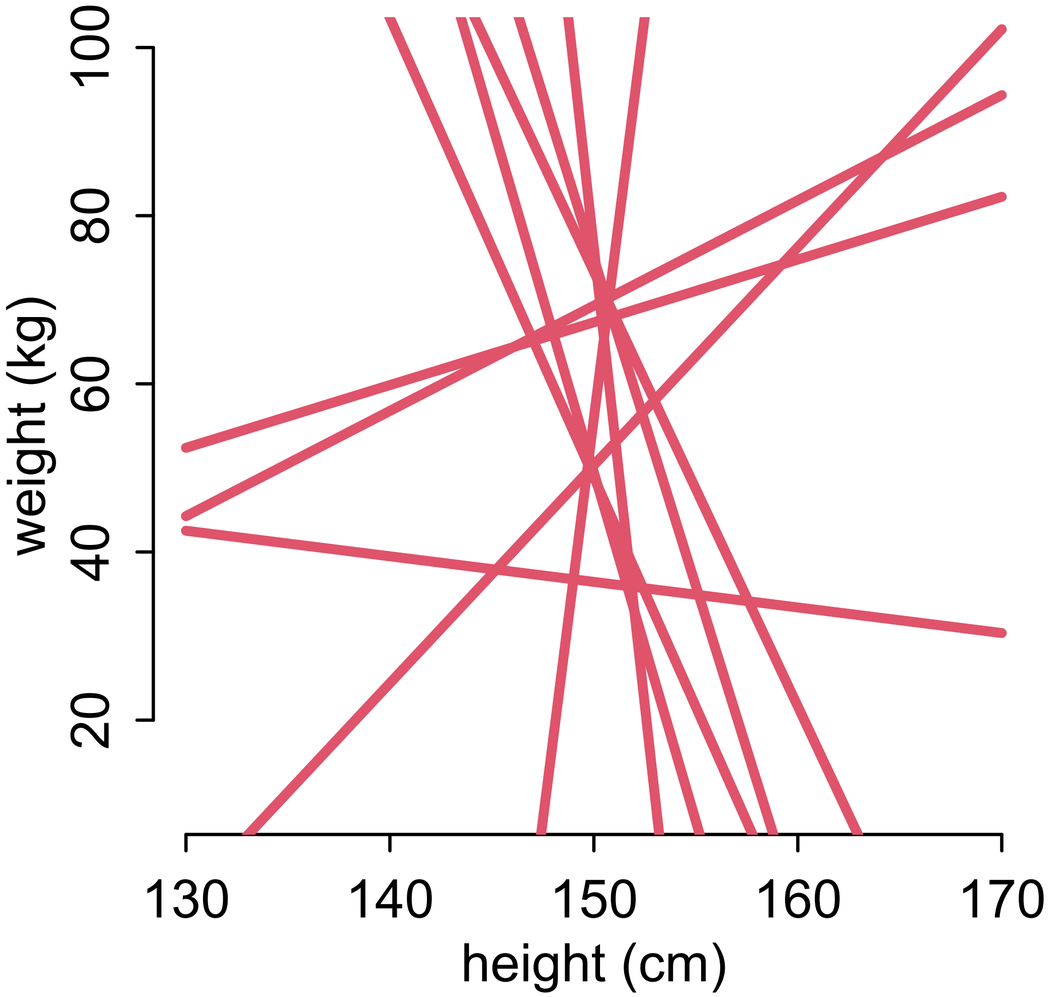
$$\sigma \sim \text{Uniform}(0,10)$$

Sampled regression lines

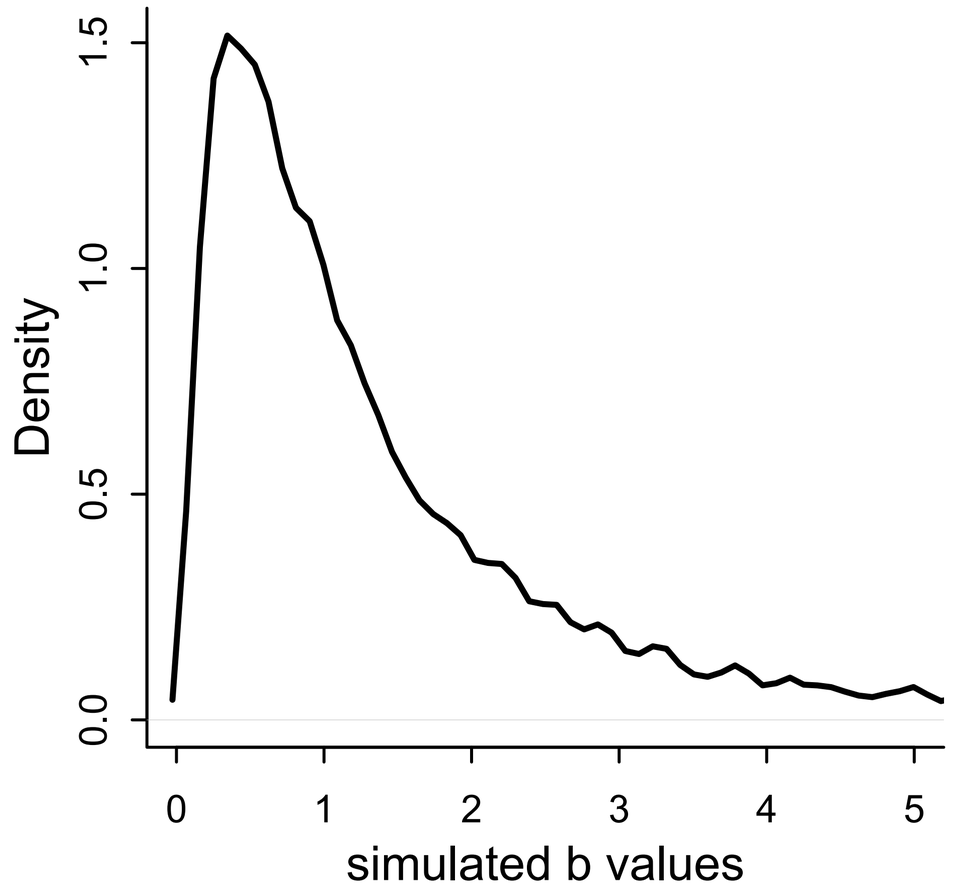
```
n <- 10
alpha <- rnorm(n,60,10)
beta <- rnorm(n,0,10)

Hbar <- 150
Hseq <- seq(from=130,to=170,len=30)
plot(NULL,xlim=c(130,170),ylim=c(10,100),
    xlab="height (cm)",ylab="weight (kg)")
for ( i in 1:n )
    lines( Hseq , alpha[i] + beta[i]*(Hseq-Hbar) ,
    lwd=3 , col=2 )
```





Statistical model for H → W



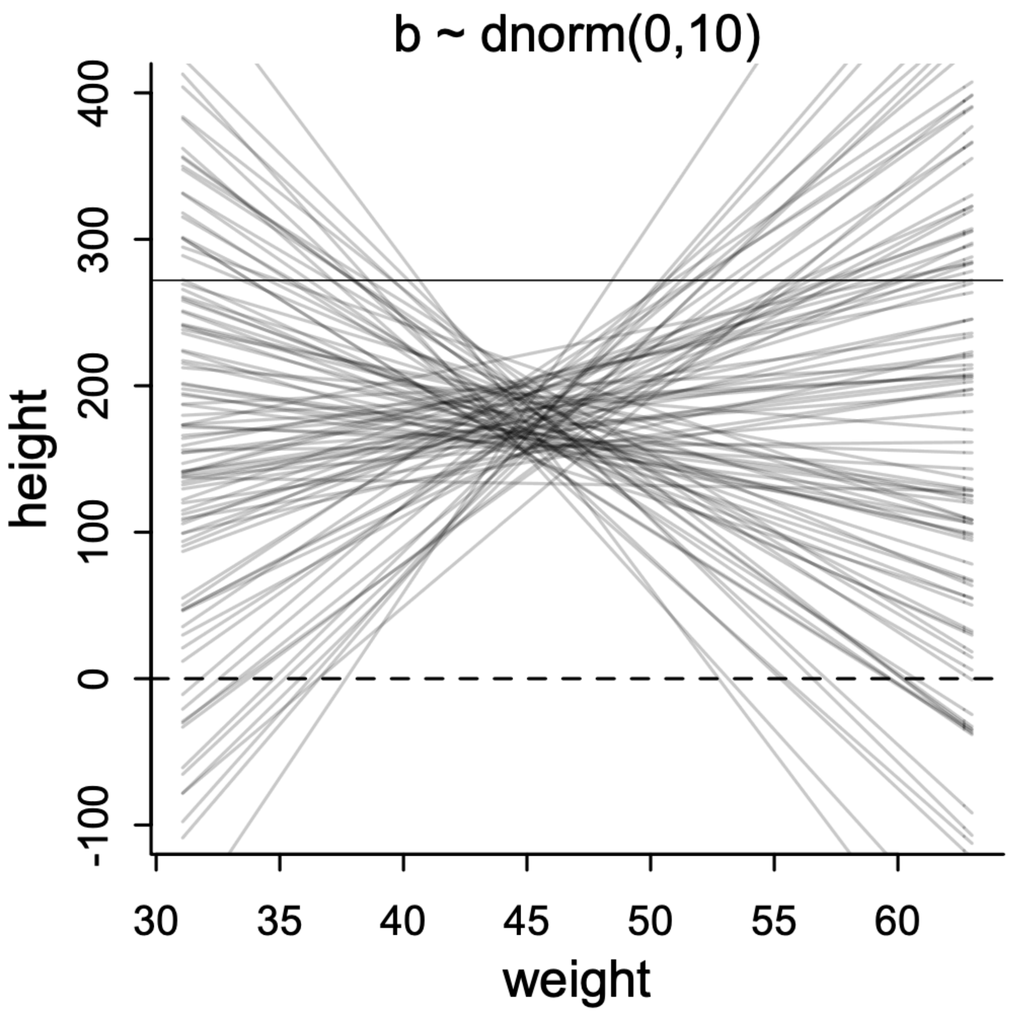
$$W_i \sim \text{Normal}(\mu_i, \sigma)$$

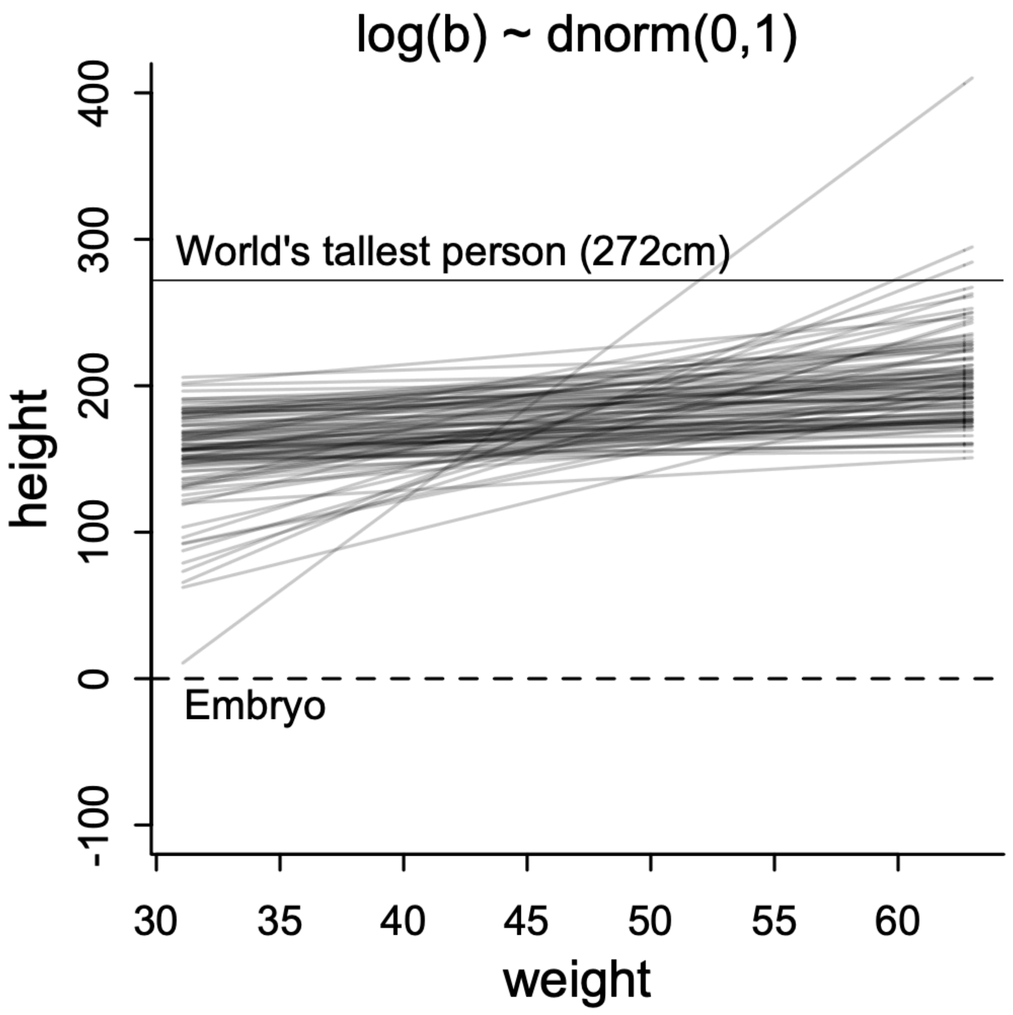
$$\mu_i = \alpha + \beta(H_i - \bar{H})$$

$$\alpha \sim \text{Normal}(60,10)$$

$$\beta \sim \text{LogNormal}(0,1)$$

$$\sigma \sim \text{Uniform}(0,10)$$

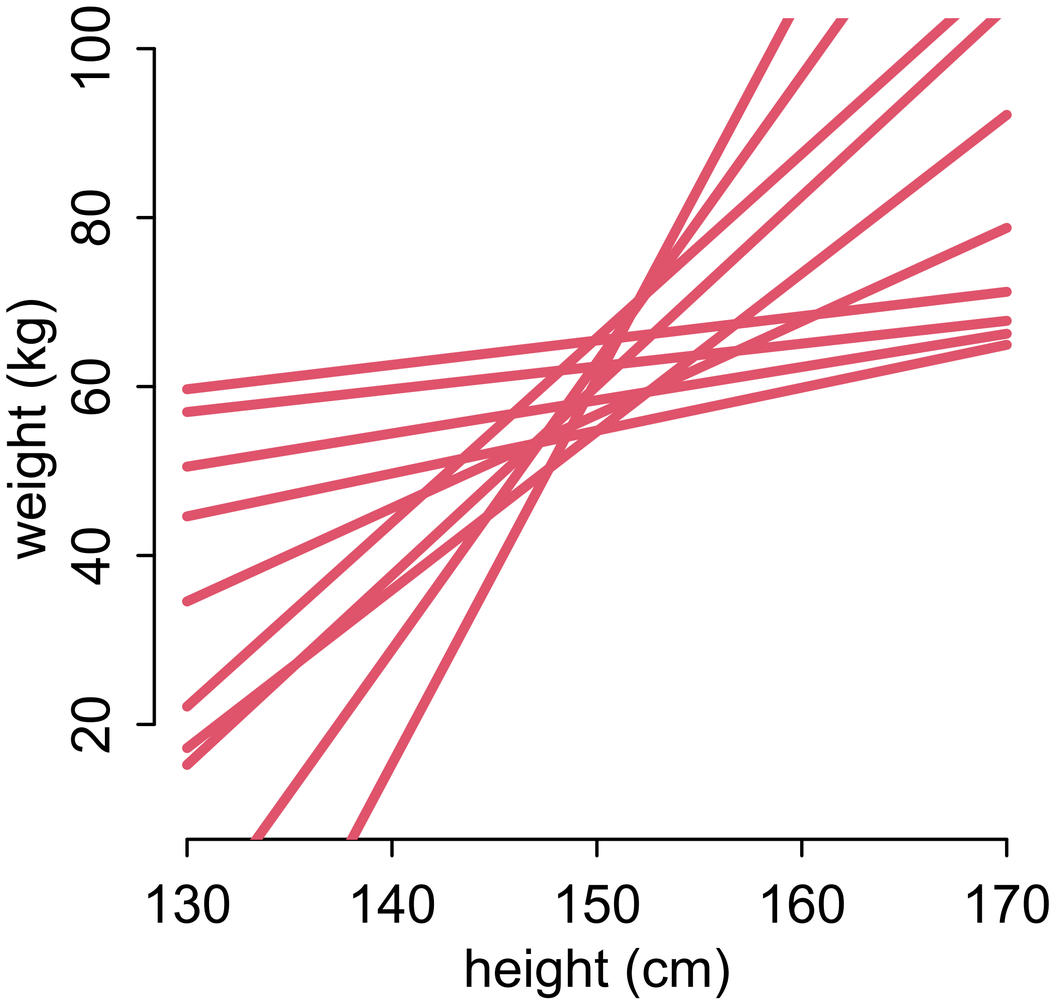

b ~ dnorm(0,10)
height
weight
400
300
200
100
0
-100
30
35
40
45
50
55
60


log(b) ~ dnorm(0,1)
World's tallest person (272cm)
Embryo
height
weight
400
300
200
100
0
-100
30
35
40
45
50
55
60

Sampled regression lines

```
n <- 10
alpha <- rnorm(n,60,10)
beta <- rlnorm(n,0,1)

Hbar <- 150
Hseq <- seq(from=130,to=170,len=30)
plot(NULL,xlim=c(130,170),ylim=c(10,100),
    xlab="height (cm)",ylab="weight (kg)")
for ( i in 1:n )
    lines( Hseq , alpha[i] + beta[i]*(Hseq-Hbar) ,
    lwd=3 , col=2 )
```



Fitting the model

$$\Pr(\alpha, \beta, \sigma \mid W, H) \propto \text{Normal}(W \mid \mu, \sigma) \times \text{Normal}(\alpha \mid 60, 10) \times \text{LogNormal}(\beta \mid 0, 1) \times \text{Uniform}(\sigma \mid 0, 10)$$

Grid approximation expensive:
100 values of each parameter => 1 million calculations

Linear Regression

```
data(Howell1)
d <- Howell1[Howell1$age>=18,]
```

Drawing the Owl

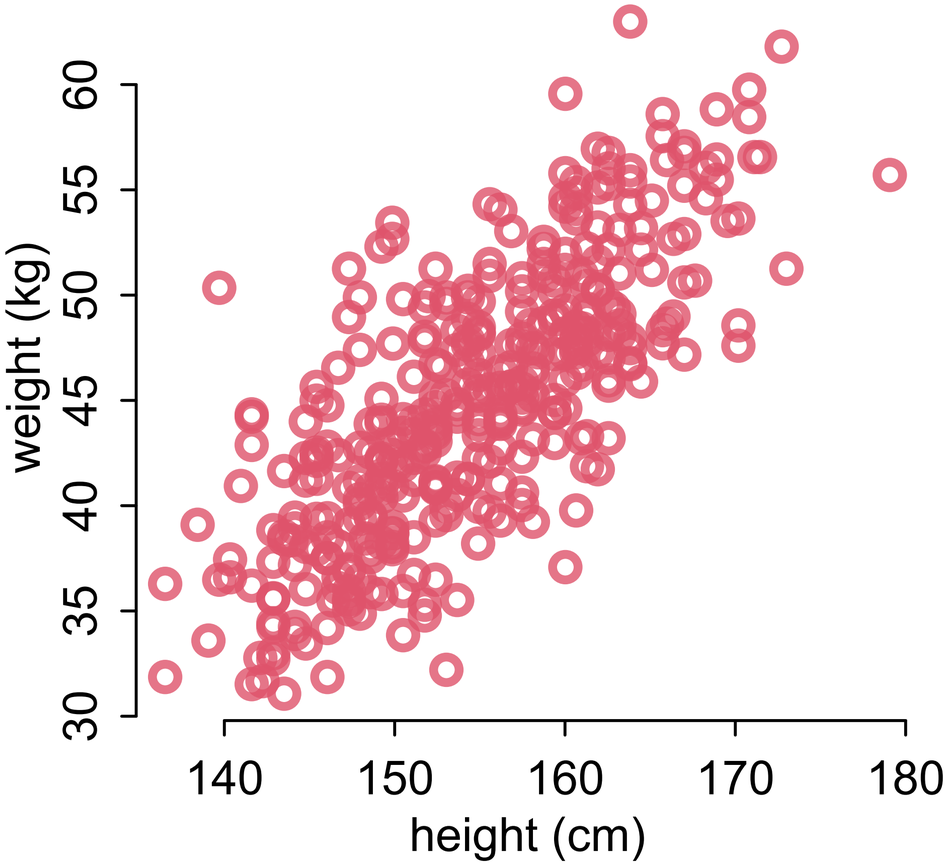
(1) Question/goal/estimand

(2) Scientific model

(3) Statistical model(s)

(4) Validate model

(5) Analyze data



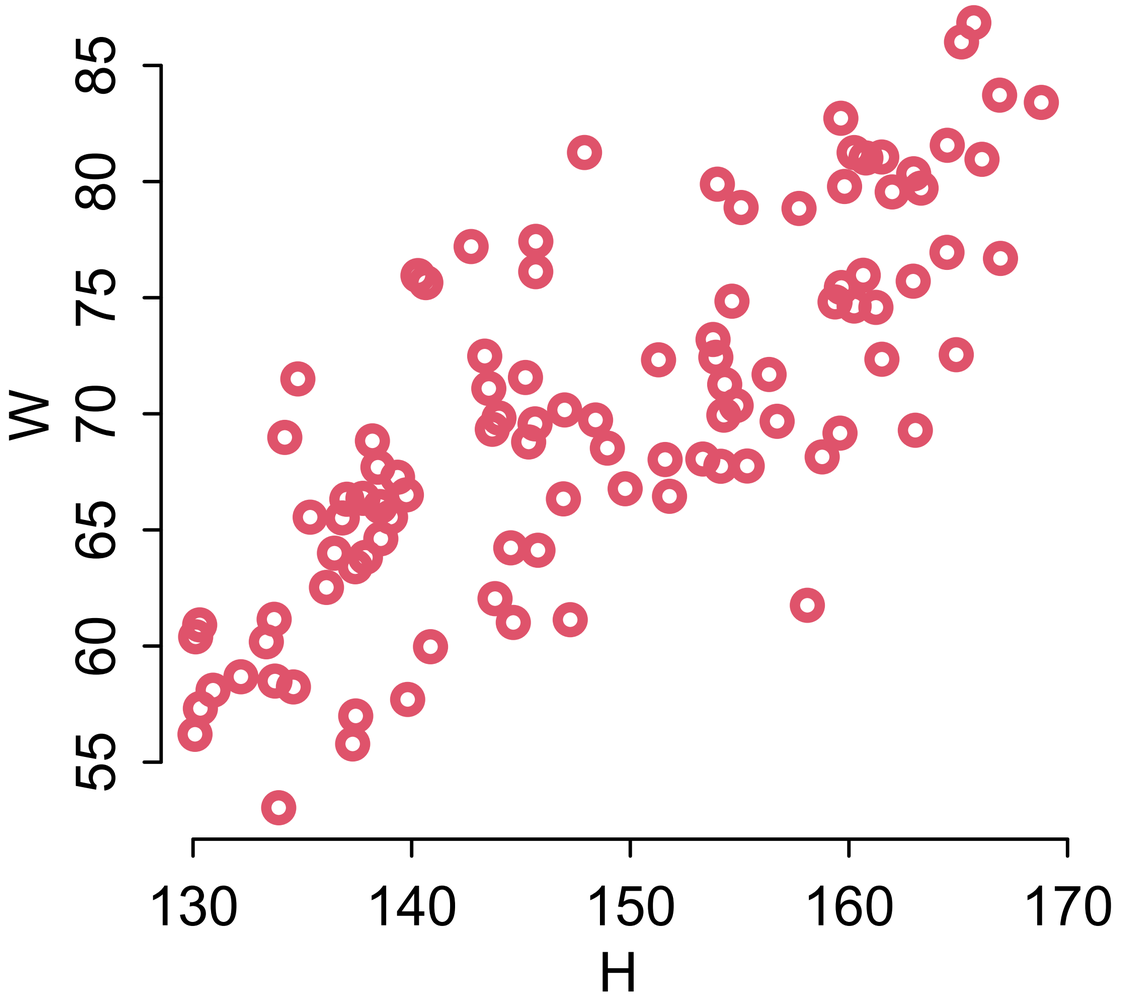
First validate with simulation

```
alpha <- 70
beta <- 0.5
sigma <- 5
n_individuals <
H <- runif(n_individuals,130,170)
mu <- alpha + beta*(H-mean(H))
W <- rnorm(n_individuals,mu,sigma)

dat <- list( H=H , W=W , Hbar=mean(H) )

m_validate <- quap(
    alist(
        W ~ dnorm(mu,sigma),
        mu <- a + b*(H-Hbar),
        a ~ dnorm(60,10),
        b ~ dlnorm(0,1),
        sigma ~ dunif(0,10)
    ), data=dat )
```

```
> precis(m_validate)
        mean   sd  5.5% 94.5%
a      69.57 0.45 68.85 70.28
b       0.49 0.04  0.43  0.56
sigma   4.48 0.32  3.97  4.98
>
```



Now with the real data

```
data(Howell1)
d <- Howell1
d <- d[ d$age>=18 , ]

dat <- list(
    W = d$weight,
    H = d$height,
    Hbar = mean(d$height) )

m_adults <- quap(
    alist(
        W ~ dnorm(mu,sigma),
        mu <- a + b*(H-Hbar),
        a ~ dnorm(60,10),
        b ~ dlnorm(0,1),
        sigma ~ dunif(0,10)
    ), data=dat )
```

$$W_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta(H_i - \bar{H})$$
$$\alpha \sim \text{Normal}(60,10)$$
$$\beta \sim \text{LogNormal}(0,1)$$
$$\sigma \sim \text{Uniform}(0,10)$$

Obey The Law

First Law of Statistical Interpretation:

The **parameters are not independent** of one another and cannot always be independently interpreted

Instead:

Push out **posterior predictions** and describe/interpret those

```
> precis(m_adults)
       mean   sd  5.5% 94.5%
a     45.00 0.23 44.64 45.36
b      0.63 0.03  0.58  0.68
sigma  4.23 0.16  3.97  4.48
>
```

```
> post <- extract.samples(m_adults)
> head(post)
         a         b    sigma
1 45.14733 0.7045790 4.380254
2 44.97759 0.6461353 4.372925
3 44.94856 0.6537192 4.111149
4 44.85016 0.6597310 4.379347
5 44.75898 0.6532690 4.200026
6 44.91711 0.6090434 4.105432
>
```

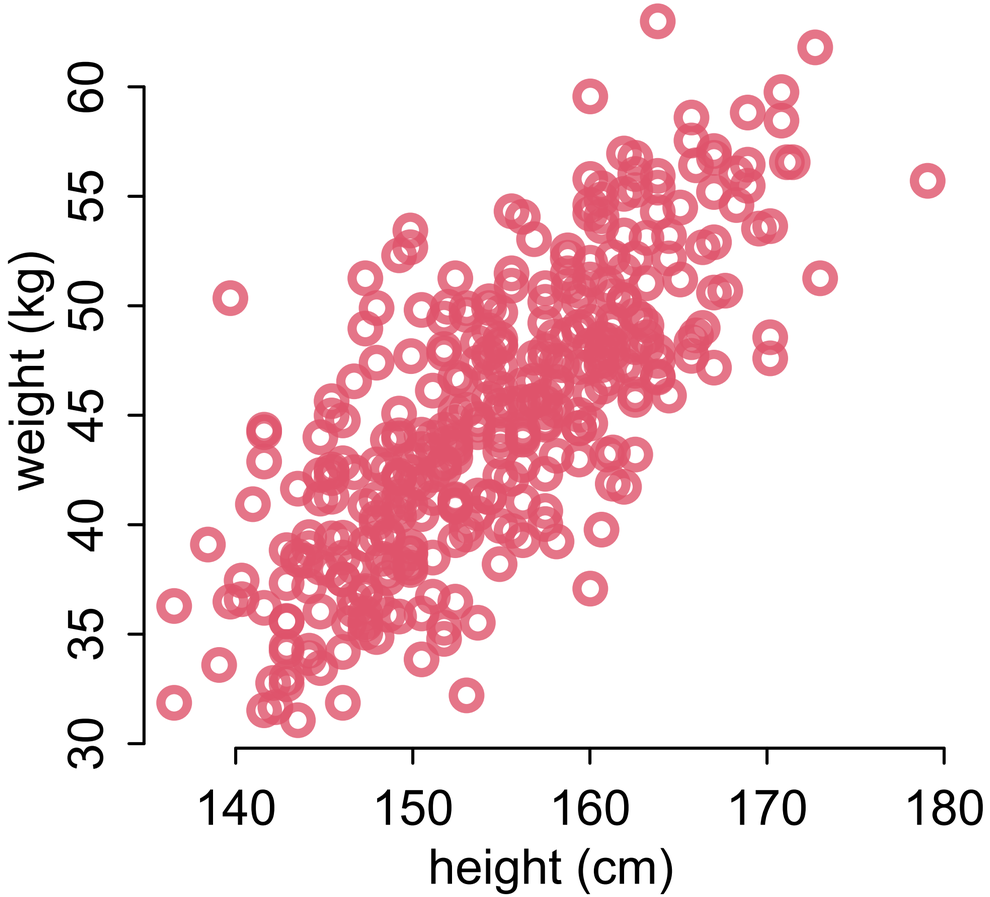
Posterior predictive distribution

(1) Plot the sample

(2) Plot the posterior mean

(3) Plot uncertainty of the mean

(4) Plot uncertainty of predictions



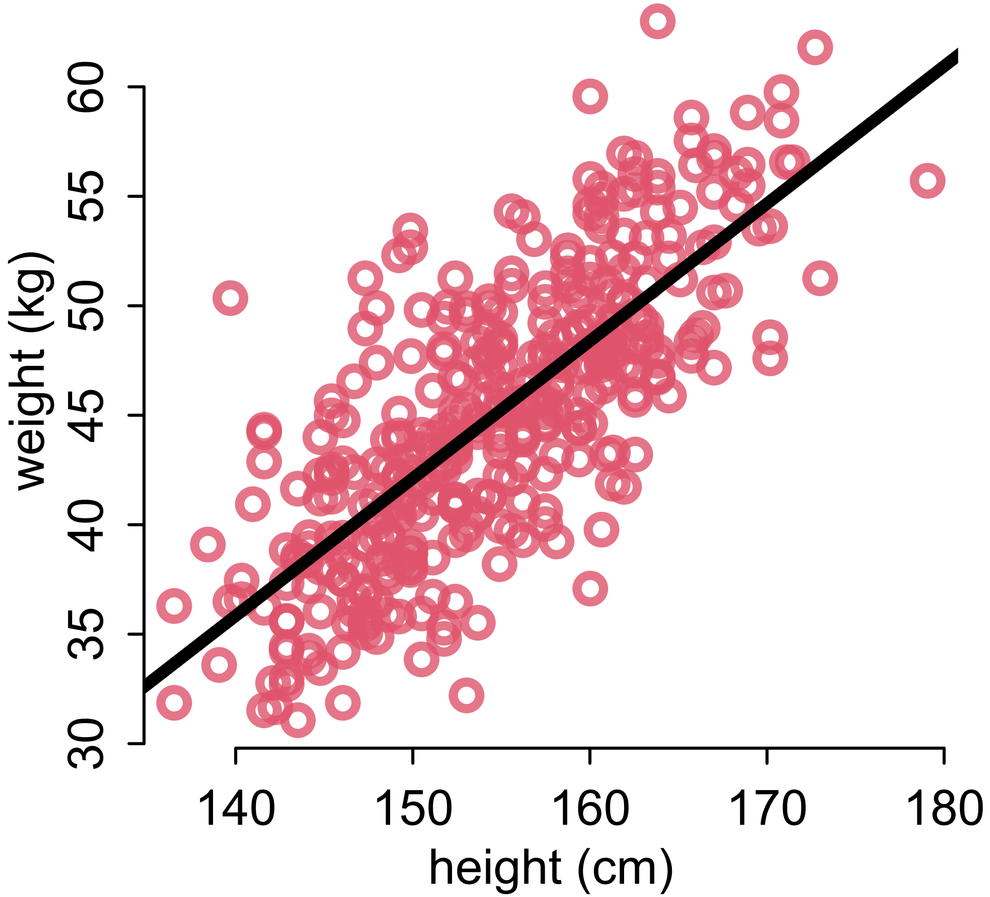
Posterior predictive distribution

(1) Plot the sample

(2) Plot the posterior mean

(3) Plot uncertainty of the mean

(4) Plot uncertainty of predictions



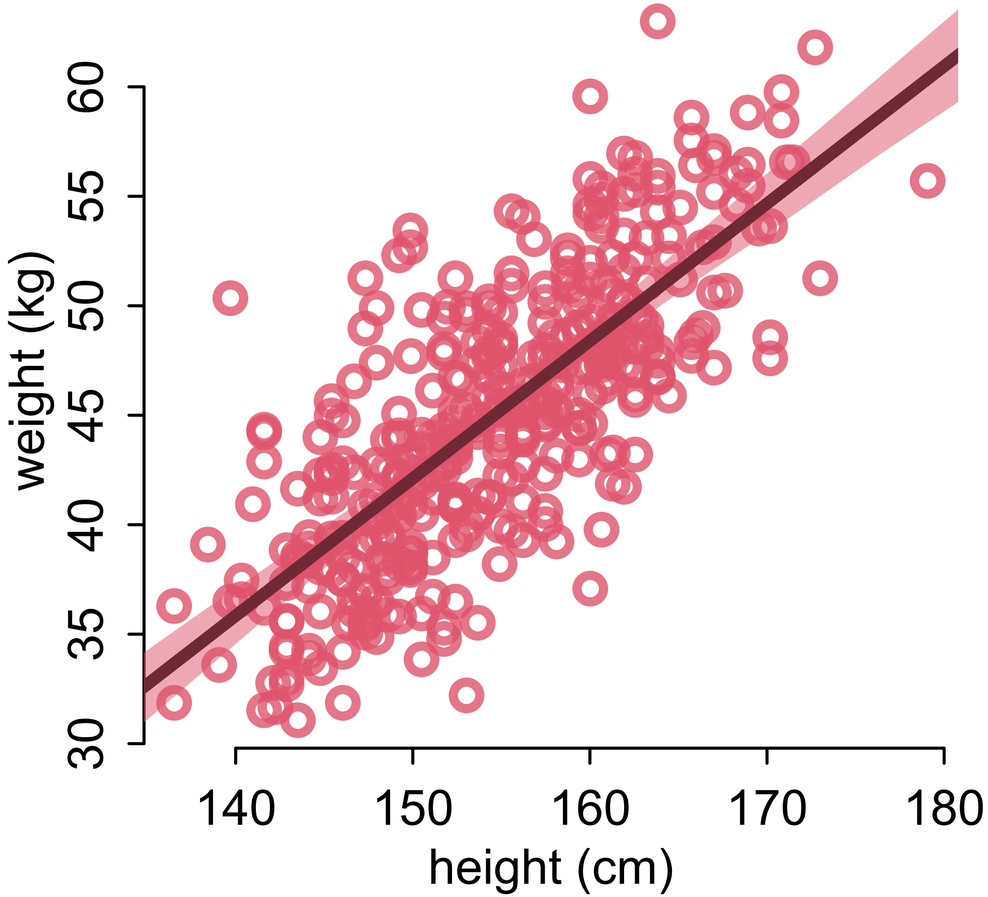
Posterior predictive distribution

(1) Plot the sample

(2) Plot the posterior mean

(3) Plot uncertainty of the mean

(4) Plot uncertainty of predictions



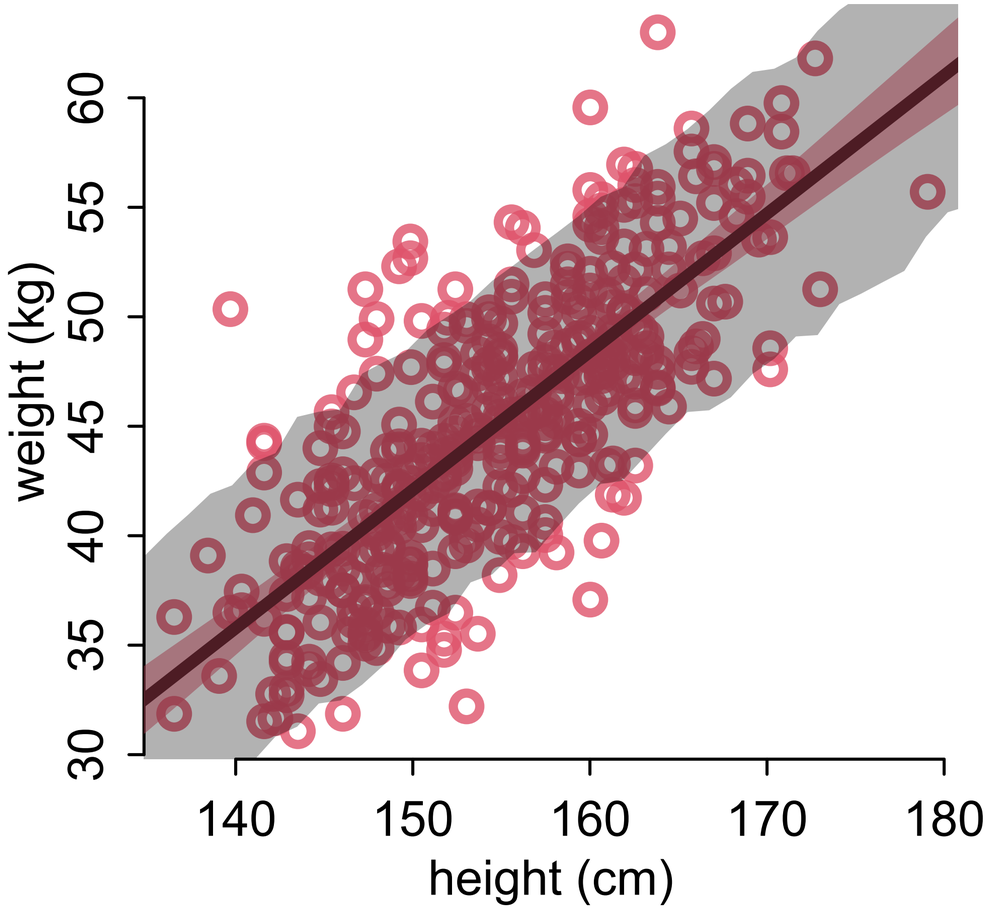
Posterior predictive distribution

(1) Plot the sample

(2) Plot the posterior mean

(3) Plot uncertainty of the mean

(4) Plot uncertainty of predictions

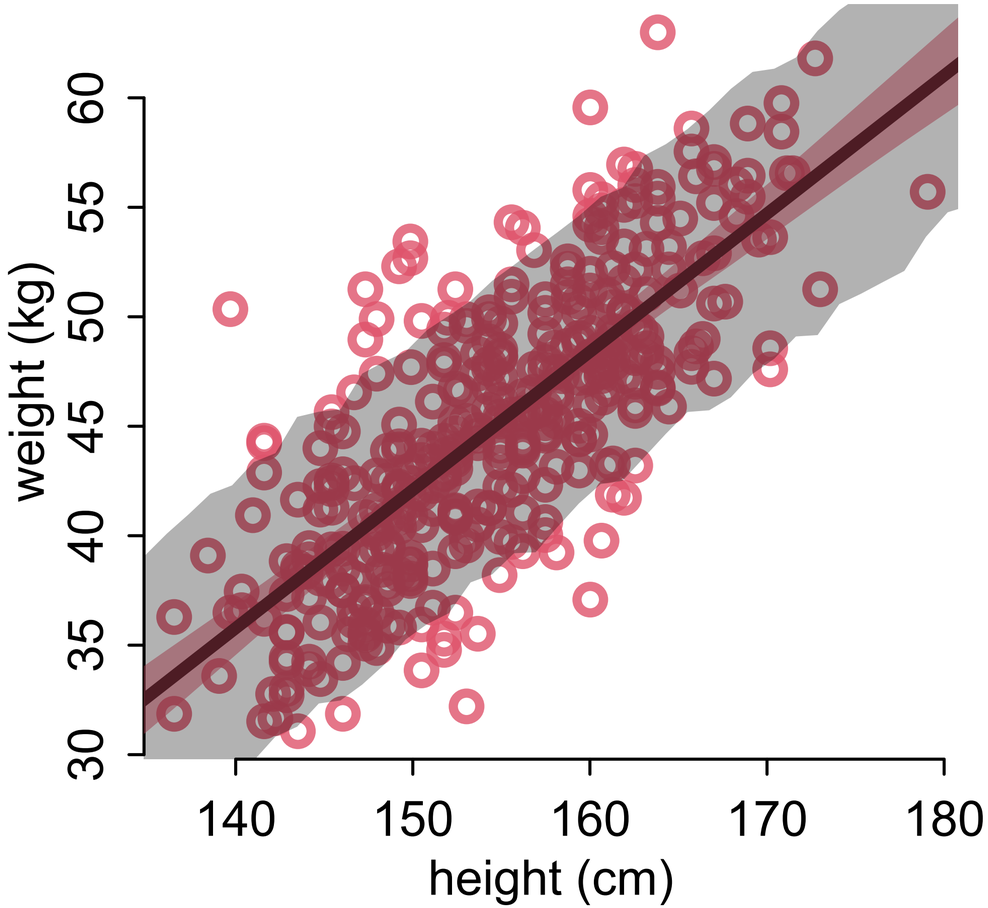


Posterior predictive distribution

```
# plot sample
col2 <- col.alpha(2,0.8)
plot( d$height , d$weight , col=col2 , lwd=3 ,
   cex=1.2 , xlab="height (cm)" , ylab="weight (kg)" )

# expectation with 99% compatibility interval
xseq <- seq(from=130,to=190,len=50)
mu <- link(m0,data=list(H=xseq,Hbar=mean(d$height)))
lines( xseq , apply(mu,2,mean) , lwd=4 )
shade( apply(mu,2,PI,prob=0.99) , xseq ,
   col=col.alpha(2,0.5) )

# 89% prediction interval
W_sim <- sim(m0,data=list(H=xseq,Hbar=mean(d$height)))
shade( apply(W_sim,2,PI,prob=0.89) , xseq ,
   col=col.alpha(1,0.3) )
```



Summary

- Write a 2-3 sentences
  - Difference betweeen frequentists and bayesian statistics