Introduction to the course and Bayesian data analysis

Frank Edwards

1/24/2020

Getting Started

Before we get started: R, Rstudio, and packages

1. Install / update R if needed (`https://cran.r-project.org/`)
2. Install / update RStudio if needed (`https://rstudio.com/`)
3. Install rstan (`https://github.com/stan-dev/rstan/wiki/RStan-Getting-Started`)
4. Install rethinking package (`https://github.com/rmcelreath/rethinking/tree/Experimental`)

Before we get started: Git

- Install Git (`https://git-scm.com/`)

1. Open your terminal [Mac, Linux] or Git Bash [Windows]

```
cd Dropbox [Arbitrary! choose any location you like]
git clone https://github.com/f-edwards/intermediate_stats.git
```

2. Then, to update with the latest slides / data / etc

```
cd Dropbox/intermediate_stats
git fetch
```

If you feel lost

- This course closely follows McElreath's *Statistical Rethinking*
- His lectures and slides for the course are available for further review (`https://github.com/rmcelreath/statrethinking_winter2019`)
- My slides borrows liberally from the book and his materials

All models are wrong...

So why do statistics?

- Learn about something we can't see (parameters) from something we can (data)

So why do statistics?

- Learn about something we can't see (parameters) from something we can (data)
- Models (**golems** for McElreath) are very powerful, but very dumb

So why do statistics?

- Learn about something we can't see (parameters) from something we can (data)
- Models (**golems** for McElreath) are very powerful, but very dumb
- Statistical models $\neq$ scientific models

All models are wrong, but some are useful

- We create stylized abstractions of reality in our models (McElreath's **small world**)

All models are wrong, but some are useful

- We create stylized abstractions of reality in our models (McElreath's **small world**)
- They are always incomplete representations of reality (**large world**)

All models are wrong, but some are useful

- We create stylized abstractions of reality in our models (McElreath's **small world**)
- They are always incomplete representations of reality (**large world**)
- Models answer questions about the **small world**. It's up to us to carefully translate them to the large world.

All models are wrong, but some are useful

- We create stylized abstractions of reality in our models (McElreath's **small world**)
- They are always incomplete representations of reality (**large world**)
- Models answer questions about the **small world**. It's up to us to carefully translate them to the large world.
- Also, Columbus was a slaver and initiated a catastrophic genocide

Interpret these results

```r
library(broom)
data(mtcars)
## mpg is miles per gallon, wt is weight in 1000 lbs
m1 <- lm(mpg ~ wt, data = mtcars)
tidy(m1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    37.3      1.88      19.9  8.24e-19
## 2 wt             -5.34     0.559     -9.56 1.29e-10
```

```r
confint(m1)
```

```
##                 2.5 %    97.5 %
## (Intercept) 33.450500 41.119753
## wt          -6.486308 -4.202635
```

Motivating two common approaches

- Frequentist: The truth is fixed, we can estimate it using repeated sampling and large number theorems, assuming our measurement is one of many that could have resulted

- Frequentist: The truth is fixed, we can estimate it using repeated sampling and large number theorems, assuming our measurement is one of many that could have resulted
- Bayesian: Given our assumptions and the data, which probability distribution is the most plausible answer?

- A more intuitive approach to interpreting statistical models (no more sampling distributions!)

Why Bayesian?

- A more intuitive approach to interpreting statistical models (no more sampling distributions!)
- Computational costs have decreased rapidly

Why Bayesian?

- A more intuitive approach to interpreting statistical models (no more sampling distributions!)
- Computational costs have decreased rapidly
- Priors are a useful way to incorporate what we already know

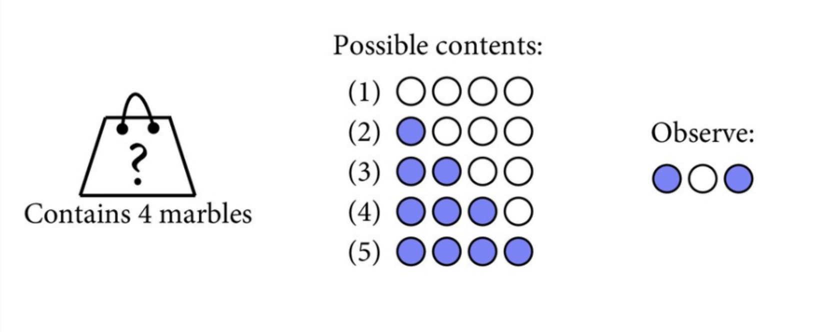
Why Bayesian?

- A more intuitive approach to interpreting statistical models (no more sampling distributions!)
- Computational costs have decreased rapidly
- Priors are a useful way to incorporate what we already know
- Avoids overfitting by not trusting the data too much

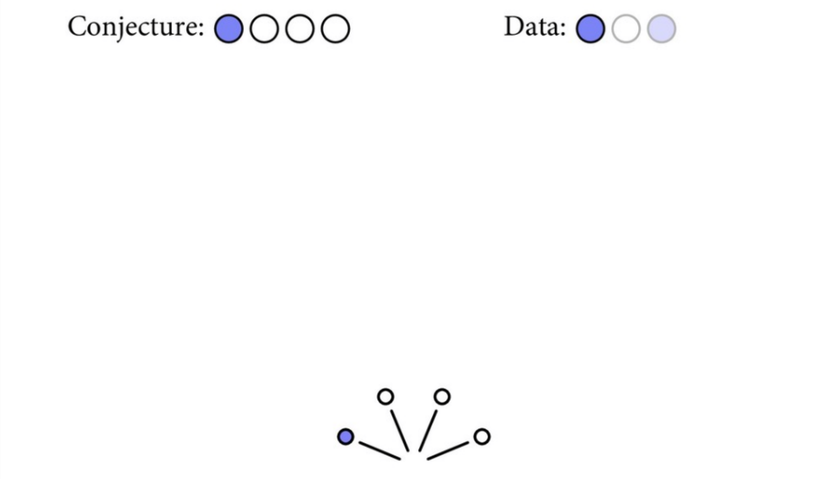
Why Bayesian?

- A more intuitive approach to interpreting statistical models (no more sampling distributions!)
- Computational costs have decreased rapidly
- Priors are a useful way to incorporate what we already know
- Avoids overfitting by not trusting the data too much
- Applications in scientific inference, prediction, machine learning

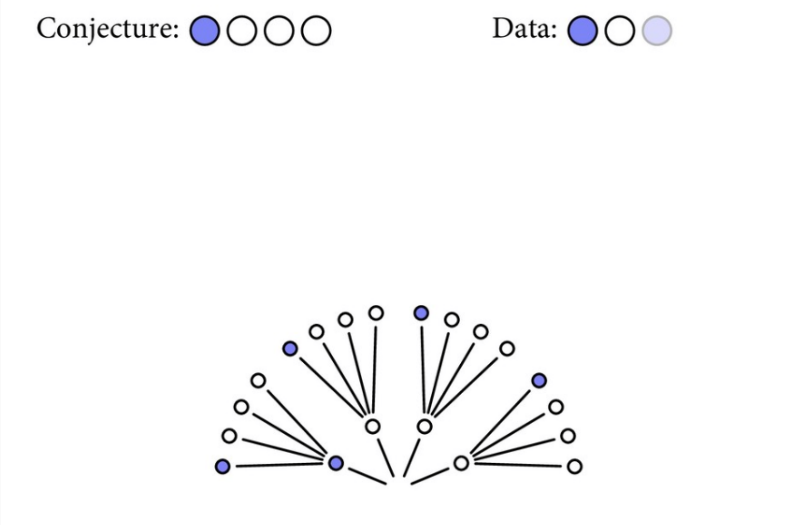
Probability: how many ways could what happened have happened?

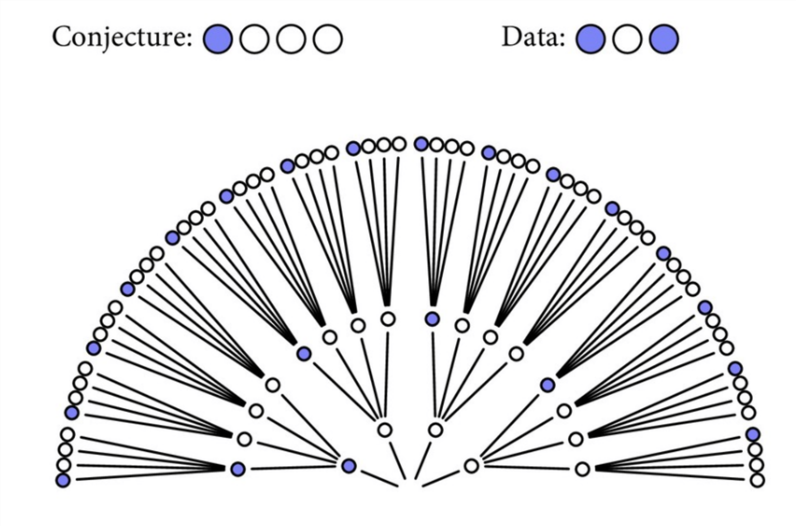


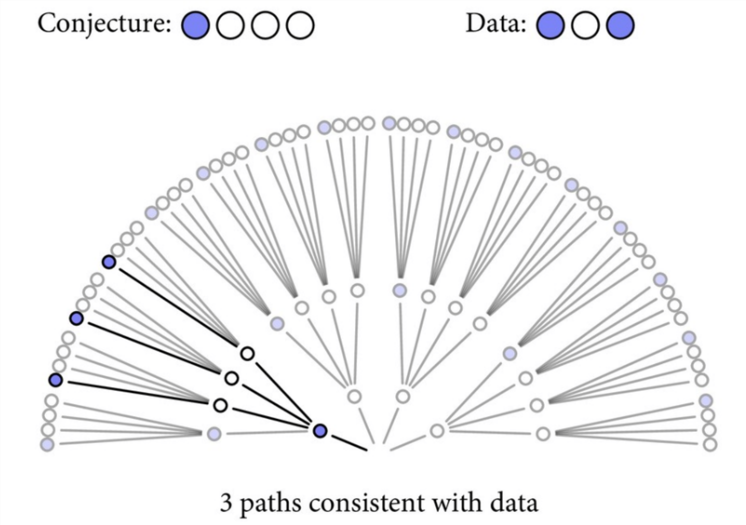
Stolen from McElreath's slides

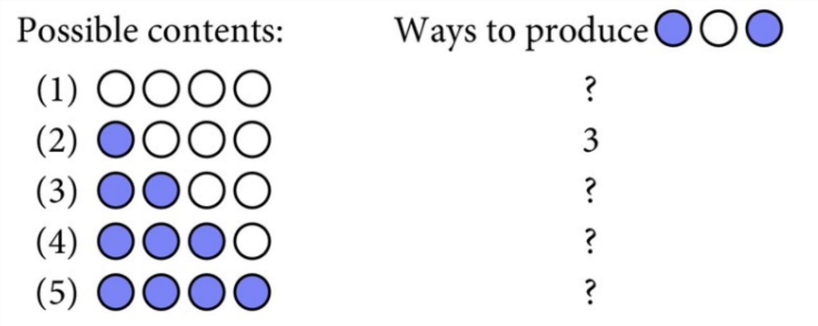


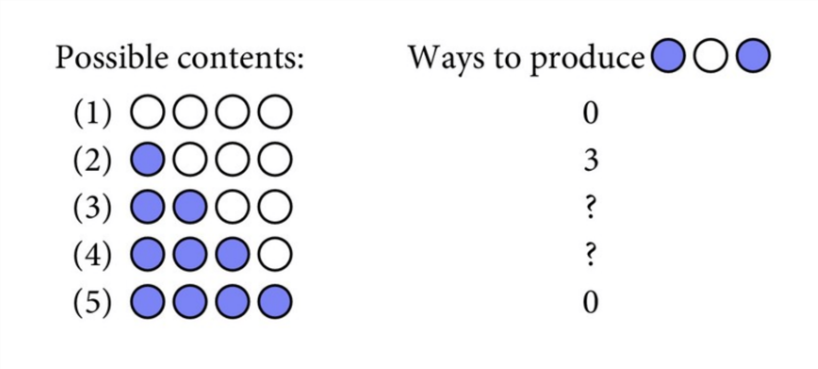
Conjecture:
Data:

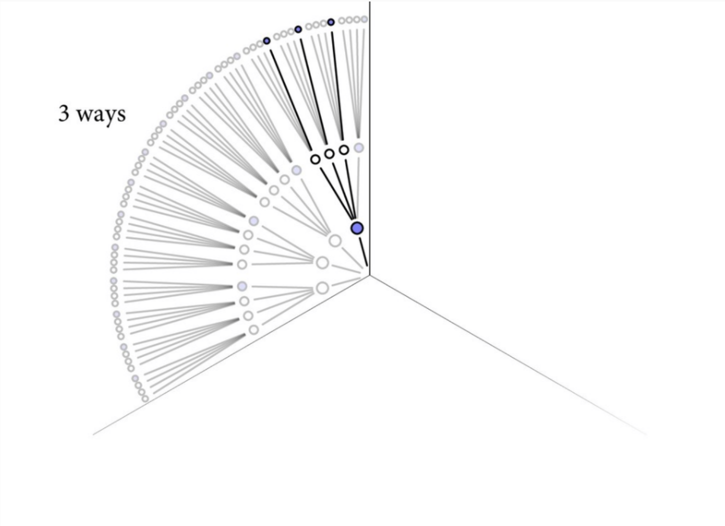

Conjecture:
Data:

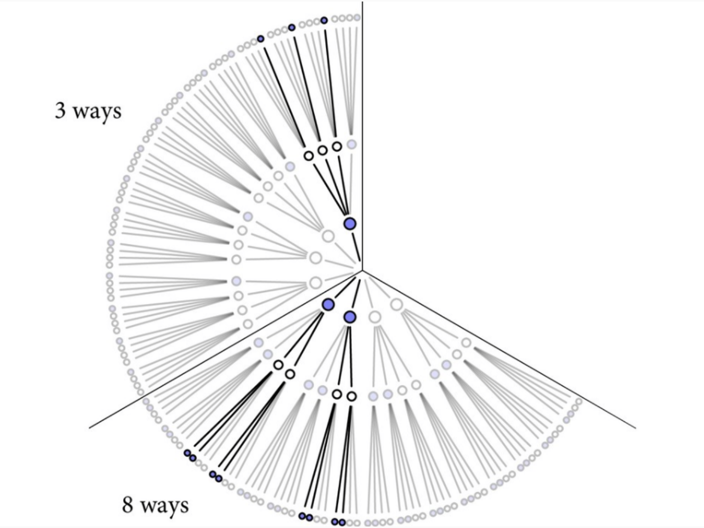

Conjecture:
Data:

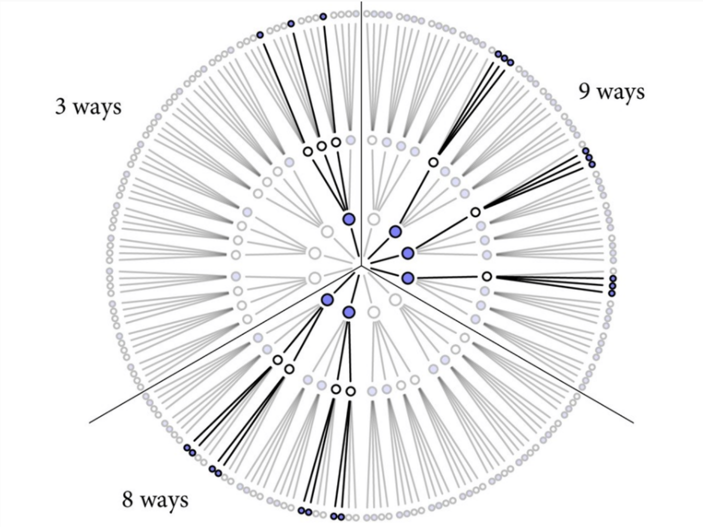

Conjecture:
Data:
3 paths consistent with data


Possible contents:
Ways to produce
(1)
?
(2)
3
(3)
?
(4)
?
(5)
?

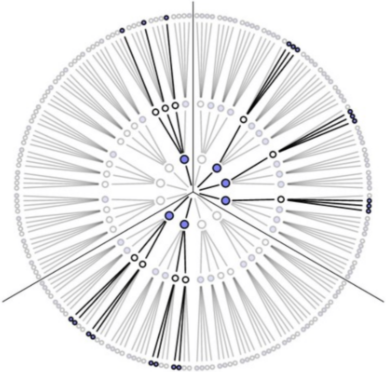

Possible contents:
Ways to produce
(1)
0
(2)
3
(3)
?
(4)
?
(5)
0


3 ways


3 ways
8 ways


3 ways
9 ways
8 ways

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



Draw one new marble: it is Blue

Conjecture	Ways to produce B	Prior counts	New count
WWWW	0	0	$0 \times 0 = 0$
BWWW	1	3	$1 \times 3 = 3$
BBWW	2	8	$2 \times 8 = 16$
BBBW	3	9	$3 \times 9 = 27$
BBBB	4	0	$4 \times 0 = 0$

Moving to probability

plausability of [BWWW] after seeing [B] $\propto$

ways [BWWW] can produce B $\times$

prior plausbility of [BWWW] based on draw [BWB]

Bayes' rule

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

$$p(\text{parameter}|\text{data}) = \frac{p(\text{data}|\text{parameter})p(\text{parameter})}{p(\text{data})}$$

Or in Bayesian vernacular:

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Average probability of the data}}$$

Updating our estimate of plausability of each parameter

Let's indicate the possible bag compositions with the parameter θ

θ	$p(B\|\theta)$	Prior $p(\theta)$	Posterior $pl(\theta\|B)$
WWWW	$\frac{0}{4}$	$\frac{0}{20}$	$\frac{0}{4} \times \frac{0}{20}$
BWWW	$\frac{1}{4}$	$\frac{3}{20}$	$\frac{1}{4} \times \frac{3}{20}$
BBWW	$\frac{2}{4}$	$\frac{8}{20}$	$\frac{2}{4} \times \frac{8}{20}$
BBBW	$\frac{3}{4}$	$\frac{9}{20}$	$\frac{3}{4} \times \frac{9}{20}$
BBBB	$\frac{4}{4}$	$\frac{0}{20}$	$\frac{4}{4} \times \frac{0}{20}$

Nomenclature

- A hypothetical composition of the bag of marbles θ is a **parameter**, and is unknown
- The number of ways that a parameter could produce the data is a **likelihood**, an enumeration of all sequences thta could have happened, then eliminating those that are logically inconsistent with the data
- The prior plausability of any value of θ is a **prior probability**
- The new, updated plausability of any value of θ is a **posterior probability**

Model building

- New experiment: Toss a globe, catch it, and note whether your right index finger has landed on water or land
- Suppose the first nine attempts (samples) result in the data:
- Our observed data for 9 trials: Water, Land, Water, Water, Water, Land, Water, Land, Water: WLWWWLWLW

The model building design sequence

1. Design a model (a story about how the data might arise)
2. Update: Educate the model by conditioning on the data
3. Evaluate: compare, critique, and revise the model
4. Repeat!

Design a model (tell a story)

How do we obtain the data we observed?

1. The true proportion of water on the globe is p
2. A single toss of the glove has probability p of producing a water (W) observation, and $1 - p$ of producing a land (L) observation
3. Each toss is independent of all other tosses

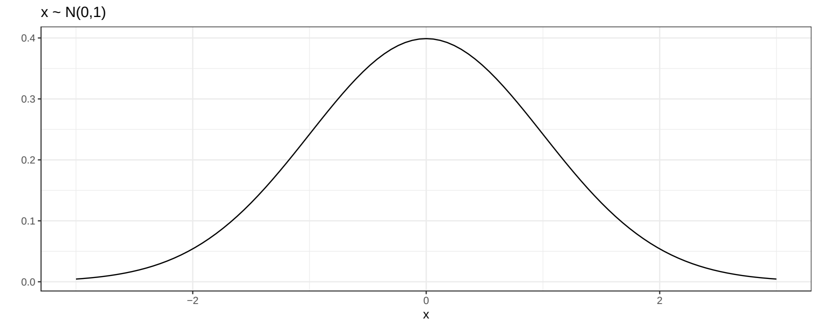
How do we use the evidence to evaluate which proportion of water on the globe is true?

- Begin with a set of plausabilities for each possible value of the parameter p (prior)
- Update these plausabilities after collecting the data (posterior)

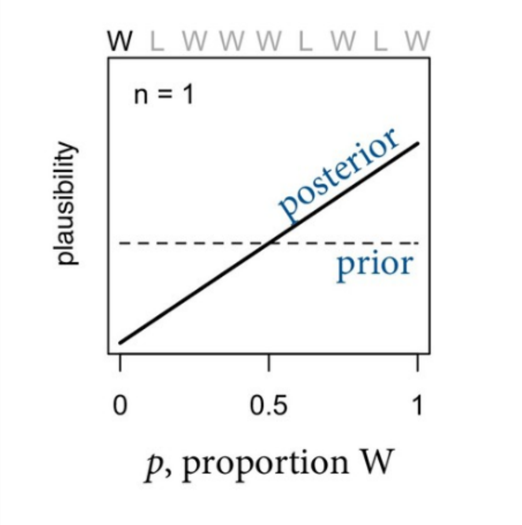
To begin, let's assume a prior where each value of p is equally likely (a uniform distribution)

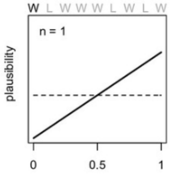
Probability densities

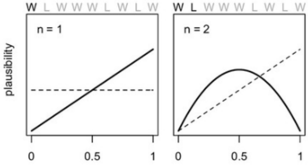
Recall that we can use a **probability density function** to describe the likelihood that a parameter takes on any particular value.



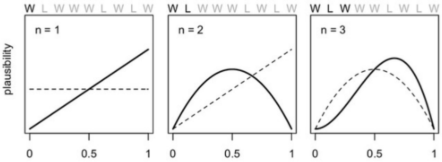
The prior and posterior



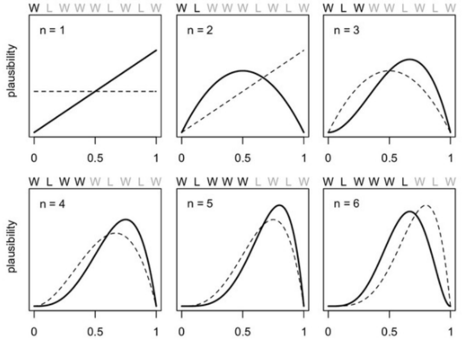

W L W W W L W L W
n = 1
plausibility
0
0.5
1



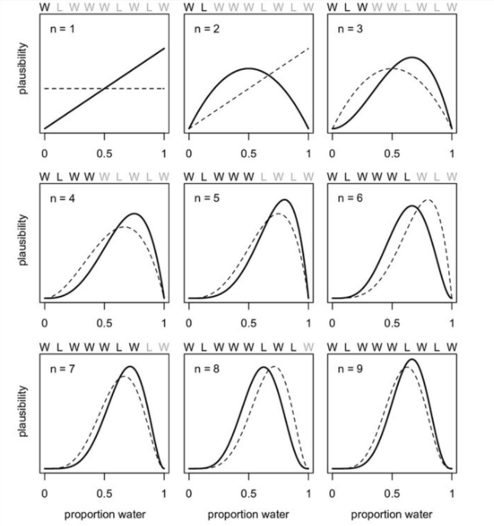
W L W W W L W L W
n = 1
plausibility
0
0.5
1
W L W W W L W L W
n = 2
0
0.5
1



W L W W W L W L W
n = 1
plausibility
0
0.5
1
W L W W W L W L W
n = 2
0
0.5
1
W L W W W L W L W
n = 3
0
0.5
1



W L W W W L W L W
n = 1
n = 2
n = 3
n = 4
n = 5
n = 6
plausibility
0
0.5
1



W L W W W L W L W
n = 1
n = 2
n = 3
n = 4
n = 5
n = 6
n = 7
n = 8
n = 9
plausibility
0
0.5
1
proportion water

Evaluate

Our approach gives us a logical answer to this question:

"How plausible is each proportion of water, given these data and our model".

1. Model certainty is no guarantee that your model accurately captures the real world
2. Check how the answer changes based on changes in your assumptions (priors, model)

Components of a model

We have two kinds of variables: *parameters* (unobserved), and observed variables

- For each *parameter*, we must specify a **prior** distribution that tells us the plausability of each possible value of the parameter

We have two kinds of variables: *parameters* (unobserved), and observed variables

- For each *parameter*, we must specify a **prior** distribution that tells us the plausability of each possible value of the parameter
- For observed variables, we define how likely each combination of observed variables is for a specific value of p, called a **likelihood**. Instead of counting outcomes, we'll use a probability distribution function

- What is the parameter of interest?

For the globe tossing model

- What is the parameter of interest?
- What are the observed variables?

For the globe tossing model

- What is the parameter of interest? p: *the proportion of the globe covered in water*
- What are the observed variables? W, L: *the counts of water and land results*

Likelihood for the observed variables

With two possible outcomes $[W, L]$, and the assumptions that

1. Each toss is independent
2. The probability of W is the same on every toss

We can estimate the probability of a set number of W values for N tosses for each possible value of p using the binomial distribution as our likelihood.

$$W \sim Binomial(p, N)$$

Or: The count of W's is distributed binomially, with a probability of a Water result p on each toss, and N total tosses

Using the binomial distribution in R

If we want to know the probability of obtaining $W = 5$ when $N = 7$ and $p = 0.5$

```r
## binomial probability density
dbinom(x = 5, size = 7, prob = 0.5)
```

```
## [1] 0.1640625
```

A prior distribution for each parameter

- Each unobseved variable, or parameter *p*, must be assigned a distribution of *prior* plausability.
- The prior is an initial assignment of how likely each possible value of *p* is.
- Priors may reflect information from other sources, or beliefs
- The prior choice is arbitrary, but consequential!
- Priors are assumptions, and can be modified and critiqued

For the globe tossing model

Let's assume that all values of p are equally likely: that the globe could have any proportion of water between 0 and 1

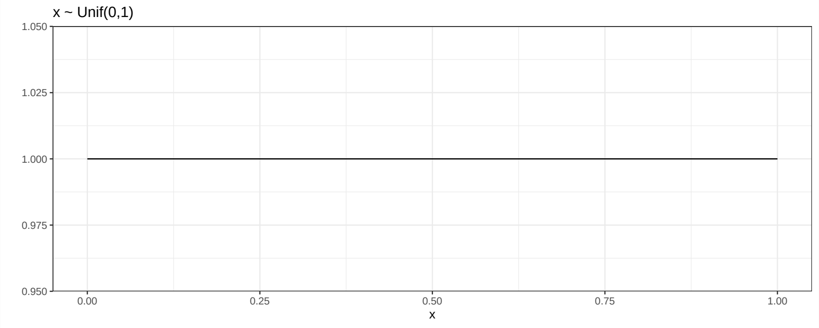
This prior is described by a *Uniform* distribution

$$p \sim Uniform(0, 1)$$

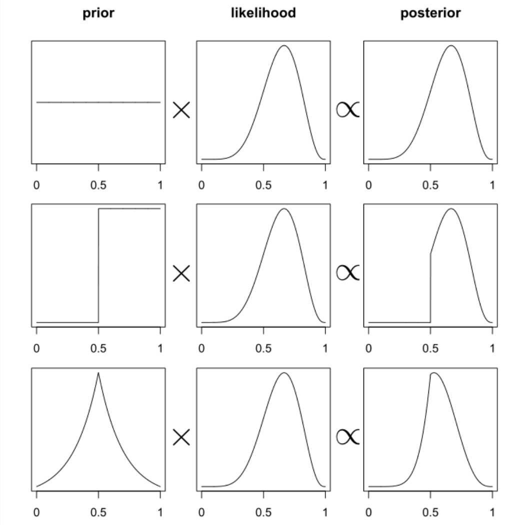
$$Pr(p) = \frac{1}{1 - 0}$$

Probability densities: prior distribution

We are assuming that the globe could have any proportion of water between 0 and 1, and that each proportion is equally likely - a uniform distribution (flat prior)



How priors influence our inferences



Estimation strategies

Because we are computing the product of probability distributions there sometimes aren't exact solutions. We'll rely on 3 algorithms to *approximate* posterior distributions to condition the prior on the likelihood of the data.

- Grid approximation (today)
- Quadratic approximation (weeks 2 on)
- Markov chain Monte Carlo (MCMC) (week 7 or 8 on)

Grid approximation algorithm

1. Define the grid
2. Compute the prior for each parameter value on the grid
3. Compute the likelihood for each parameter value on the grid
4. Multiply the prior by the likelihood
5. Divide by the sum of all values

Grid approximation in R

```
length <- 7
### make our grid
grid <- seq(from = 0, to = 1, length.out = length)
grid
```

```
## [1] 0.0000000 0.1666667 0.3333333 0.5000000 0.6666667 0.8333333 1.0000000
```

```
### make our prior and likelihood remember uniform distributions are p(x) = 1/b-a,
### and b=1, a =0 and we observe 6 Waters in 9 Trials
prior <- rep(1, length)
prior
```

```
## [1] 1 1 1 1 1 1 1
```

```
likelihood <- dbinom(6, size = 9, prob = grid)
likelihood
```

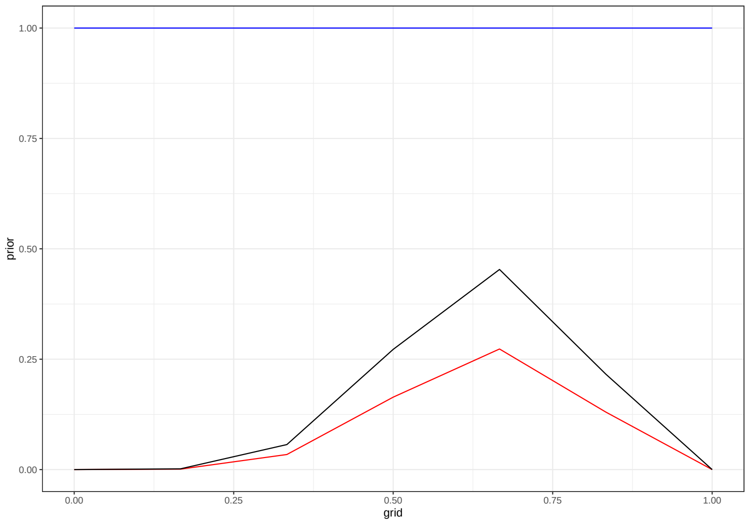
```
## [1] 0.000000000 0.001041905 0.034141137 0.164062500 0.273129096 0.130238102
## [7] 0.000000000
```

```
posterior <- prior * likelihood/sum(prior * likelihood)
posterior
```

```
## [1] 0.000000000 0.001728979 0.056655186 0.272251961 0.453241490 0.216122384
## [7] 0.000000000
```

Plot it, grid size 7

```
plot_dat <- tibble(grid, prior, likelihood, posterior)
ggplot(plot_dat, aes(x = grid)) + geom_line(aes(y = prior), color = "blue") + geom_line(aes(y = likelihood),
    color = "red") + geom_line(aes(y = posterior), color = "black")
```



Plot it, grid size 20