BEGINNING BAYES IN R

Comparing two proportions

Learning about many parameters

- **Chapters 2-3**: single parameter (one proportion or one mean)
- **Chapter 4**: multiple parameters
  - Two proportions from independent samples
  - Normal sampling where both M, S are unknown
  - Simple regression models

Types of inferences

- Making comparisons between groups:
  - Is one proportion larger than another?
- Regression effects (e.g. comparing two means):
  - Does Rafael Nadal take longer than Roger Federer to serve?

Exercise among college students

What proportion of students exercise 10 hours a week?

Does this proportion vary between men and women?

Inferential problem

- Let p_W and p_M represent the proportions of college women and men who exercise 10 hours a week, respectively
- Various hypotheses:
  - $p_W > p_M$ (women exercise more)
  - $p_W = p_M$ (women and men exercise about the same)

Models: A discrete approach

- A model is a pair: (p_W, p_M)
- Suppose each could be one of nine values 0.1, 0.2, 0.3, ..., 0.9
- Have 9 x 9 = 81 possible models

Here are the 81 models

Row is p_W, column is p_M, each X corresponds to model:

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	x	x	x	x	x	x	x	x	x
0.2	x	x	x	x	x	x	x	x	x
0.3	x	x	x	x	x	x	x	x	x
0.4	x	x	x	x	x	x	x	x	x
0.5	x	x	x	x	x	x	x	x	x
0.6	x	x	x	x	x	x	x	x	x
0.7	x	x	x	x	x	x	x	x	x
0.8	x	x	x	x	x	x	x	x	x
0.9	x	x	x	x	x	x	x	x	x

A prior

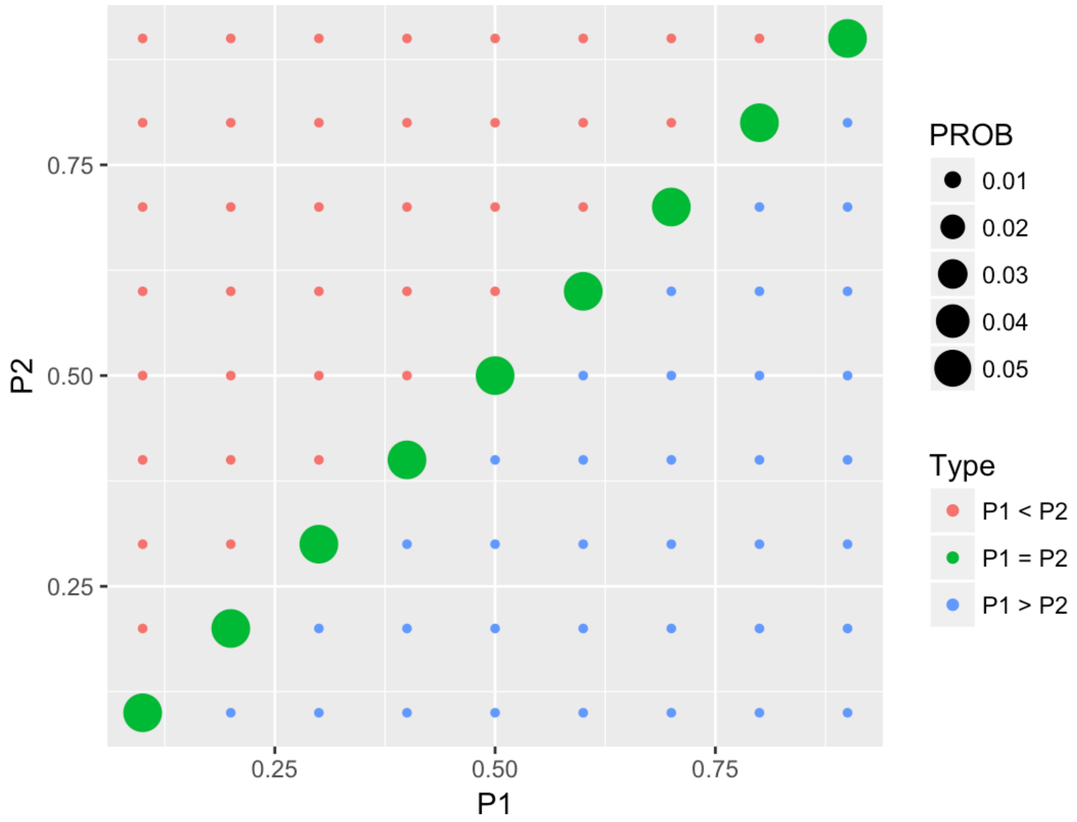
- Difficult to construct
- Describes a relationship between the proportions:
  - There is a 50% chance that $p_W = p_M$
  - Otherwise, you don't know about relative likelihoods

Testing prior

```
> # Construct a prior for Prob(p1 = p2)
> library(TeachBayes)
> prior <- testing_prior(lo = 0.1, hi = 0.9,
                         np = 9, pequal = 0.5)
> round(prior, 3)
      0.1   0.2   0.3   0.4   0.5   0.6   0.7   0.8   0.9
0.1 0.056 0.007 0.007 0.007 0.007 0.007 0.007 0.007 0.007
0.2 0.007 0.056 0.007 0.007 0.007 0.007 0.007 0.007 0.007
0.3 0.007 0.007 0.056 0.007 0.007 0.007 0.007 0.007 0.007
0.4 0.007 0.007 0.007 0.056 0.007 0.007 0.007 0.007 0.007
0.5 0.007 0.007 0.007 0.007 0.056 0.007 0.007 0.007 0.007
0.6 0.007 0.007 0.007 0.007 0.007 0.056 0.007 0.007 0.007
0.7 0.007 0.007 0.007 0.007 0.007 0.007 0.056 0.007 0.007
0.8 0.007 0.007 0.007 0.007 0.007 0.007 0.007 0.056 0.007
0.9 0.007 0.007 0.007 0.007 0.007 0.007 0.007 0.007 0.056
```

Plot of testing prior

```
> library(TeachBayes)
> draw_two_p(prior)
```



Likelihood

- We survey 40 students on their exercise habits
- 10 out of 20 women exercise; 14 out of 20 men exercise
- Assuming independent samples, likelihood is a product of binomial densities

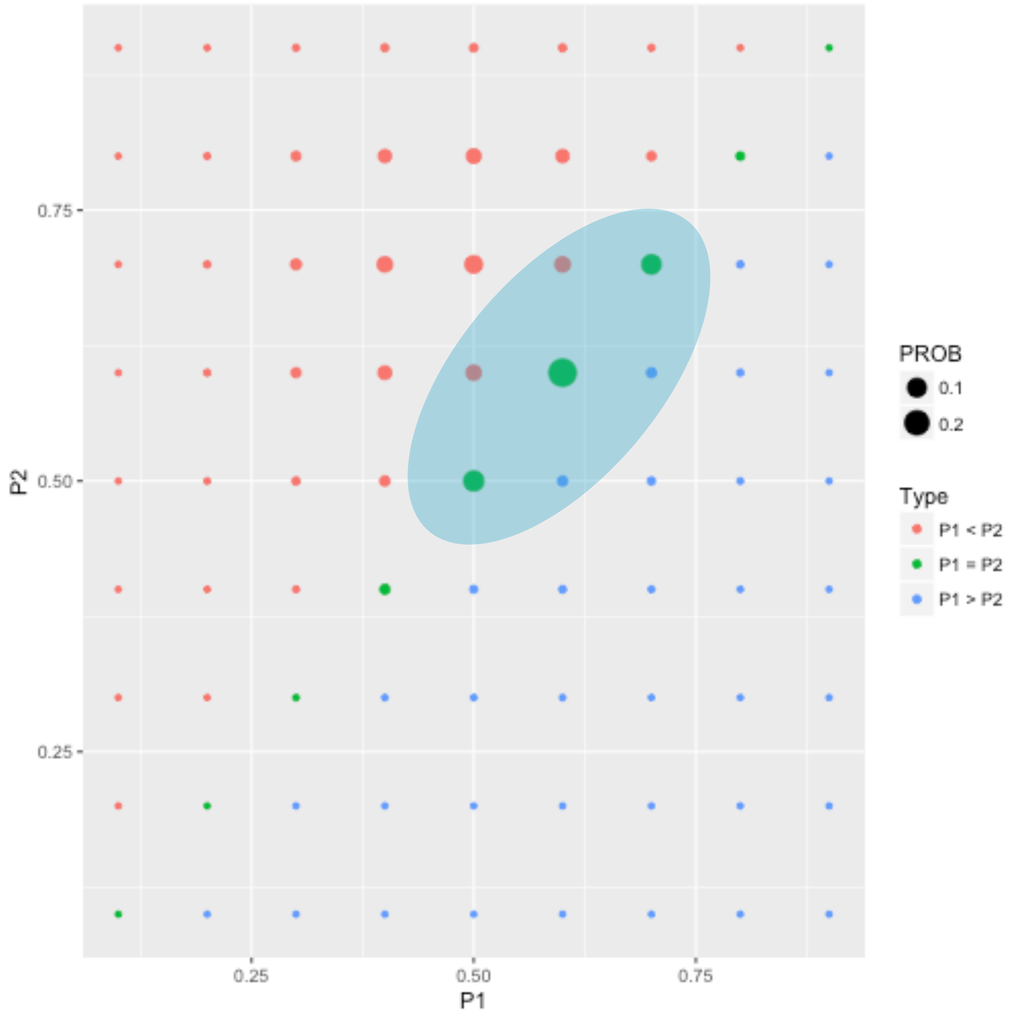
```
> Likelihood <- dbinom(10, size = 20, prob = pW) *
                dbinom(14, size = 20, prob = pM)
```

Posterior probabilities

```
> # Recall prior:
> library(TeachBayes)
> prior <- testing_prior(lo = 0.1, hi = 0.9,
                         np = 9, pequal = 0.5)

> # Multiply prior by likelihood, then normalize products
> post <- two_p_update(prior, c(10, 10), c(14, 6))
```

Plot of posterior

```
> library(TeachBayes)
> draw_two_p(post)
```



Summarize the posterior

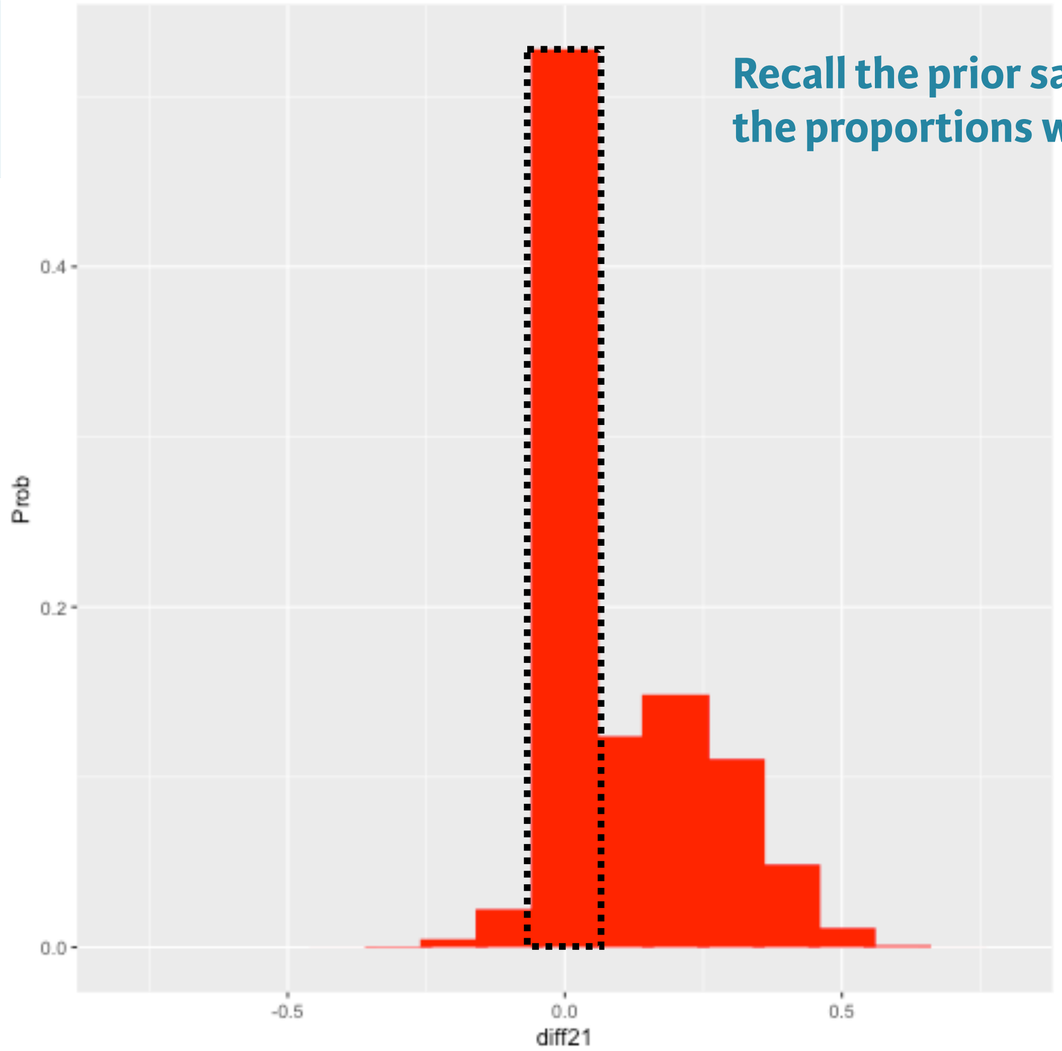
- Interested in proportions of men and women who exercise
- Posterior probabilities of the difference: $d = p_M - p_W$
- `two_p_summarize()`: finds posterior probabilities of d

Compute posterior of d

```
> library(TeachBayes)
> d <- two_p_summarize(post)
> head(d)
# A tibble: 6 × 2
  diff21          Prob
   <dbl>         <dbl>
1   -0.8 3.150309e-15
2   -0.7 2.645338e-11
3   -0.6 1.137921e-08
4   -0.5 1.247954e-06
5   -0.4 4.039640e-05
6   -0.3 5.966738e-04
```

Graph of probabilities of d

```
> library(TeachBayes)
> prob_plot(d)
```



Recall the prior said there was 50% chance the proportions were equal (i.e. d = 0)

Interpret

	$P(p_W < p_M)$	$P(p_W = p_M)$	$P(p_W > p_M)$
Prior	0.25	0.50	0.25
Posterior	0.444	0.528	0.028

There is little evidence to say that the two proportions are different

BEGINNING BAYES IN R

Let’s practice!

BEGINNING BAYES IN R

Proportions with continuous priors

Exercise among college students

- Interested in proportions of women and men who exercise
- Let p_W and p_M represent the proportions of college women and men who exercise at least 10 hours a week
- Does this proportion vary between men and women?

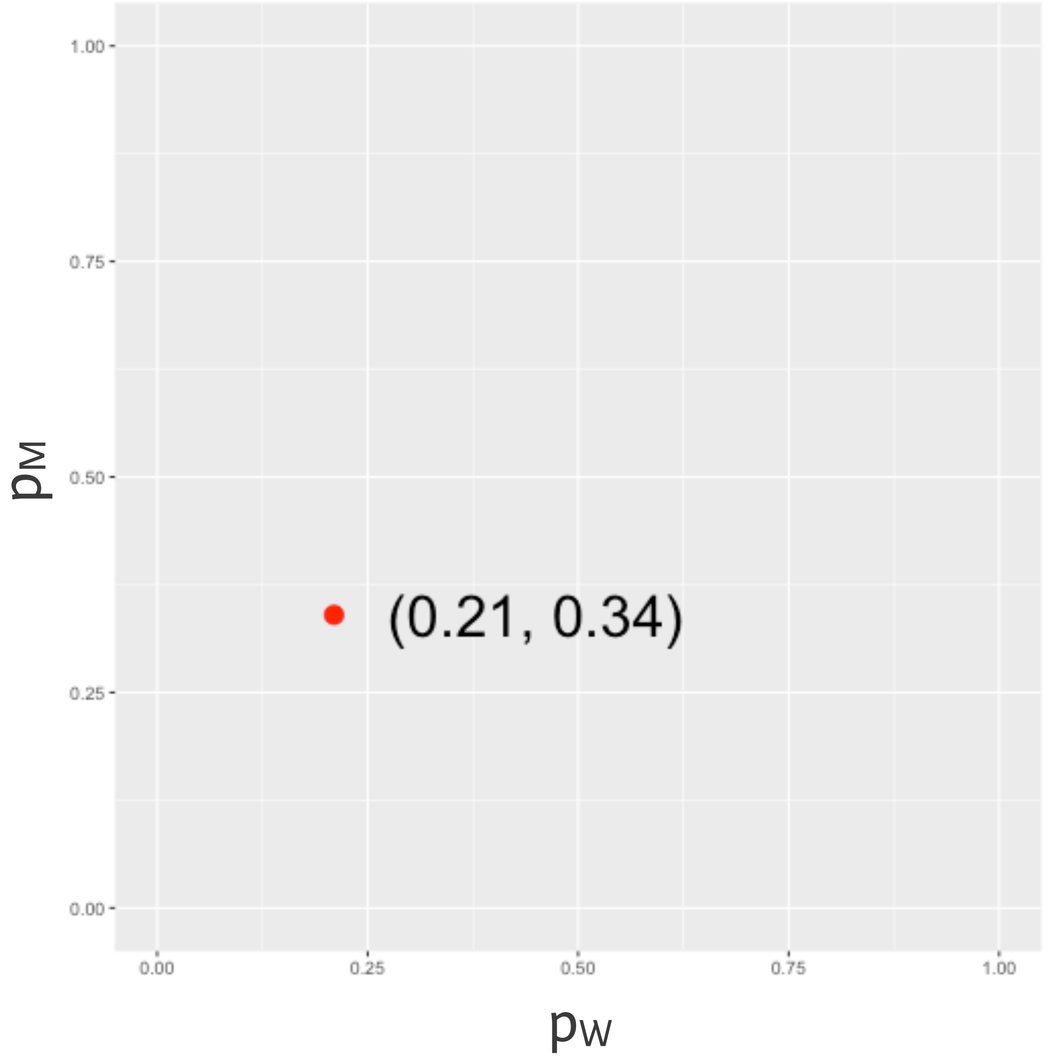
Inferential problem

- Various hypotheses:
  - $p_W = p_M$ (women and men exercise about the same)
  - $p_W > p_M$ (women exercise more)

Continuous models

- Previously, considered discrete prior models for two proportions
- View each proportion as continuous from 0 to 1

One model



Prior?

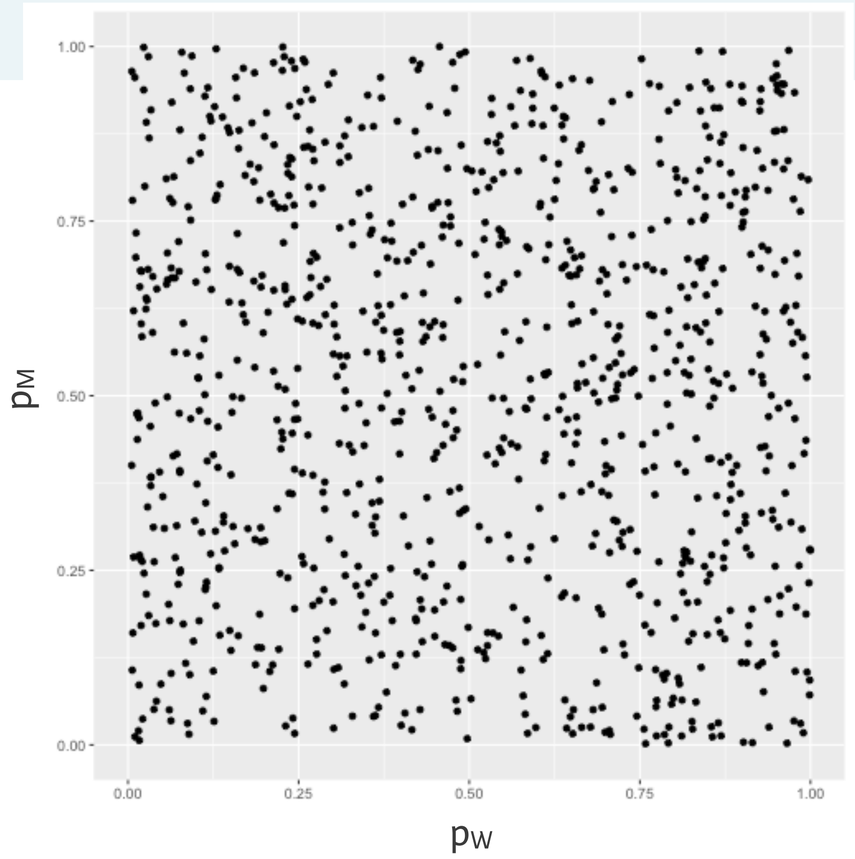
- Unit square represents all possible pairs of proportions
- Probabilities represented by smooth surface over unit square
- Difficult to construct priors that reflect dependence between two proportions p_W and p_M

Prior using beta densities

- Assume beliefs about p_W are independent of beliefs about p_M
- Use one beta curve to represent beliefs about p_W, another to represent beliefs about p_M
- Here we illustrate uniform priors:
  - p_W is beta(1, 1)
  - p_M is beta(1, 1)

1000 simulations from prior

```
> df <- data.frame(pW = rbeta(1000, 1, 1),
                   pM = rbeta(1000, 1, 1))
> ggplot(df, aes(pW, pM)) + geom_point() +
    xlim(0, 1) + ylim(0, 1)
```



Updating ...

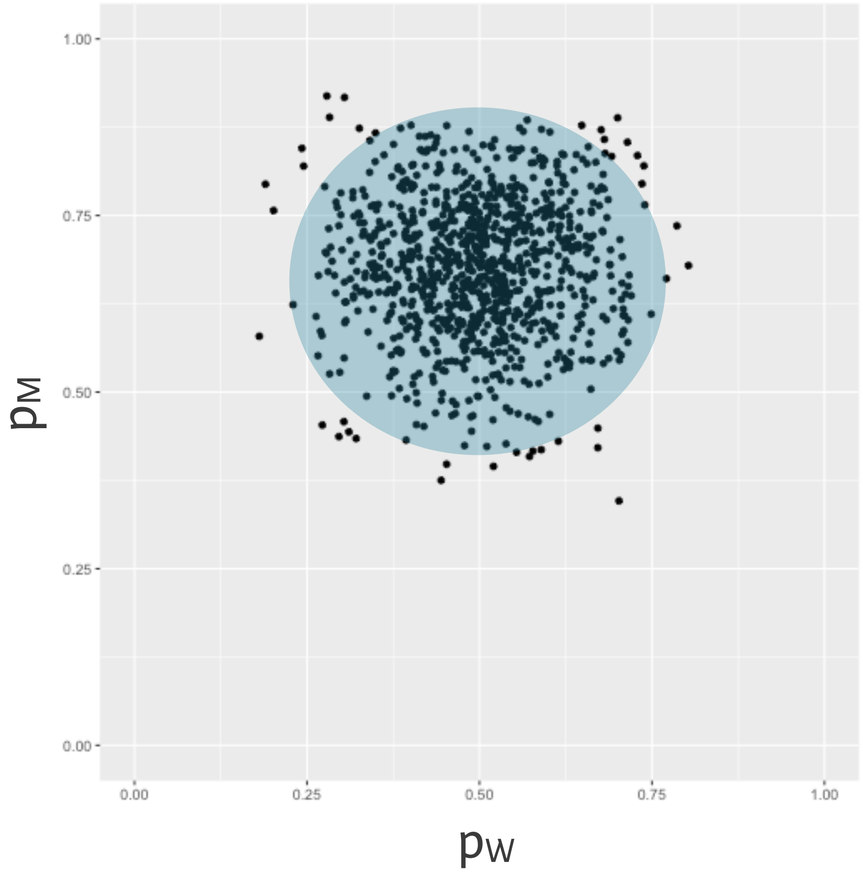
- We surveyed 40 students on their exercise habits
- 10 out of 20 women exercise; 14 out of 20 men exercise
- Prior assumed two independent beta(1, 1) curves
- Posterior of (p_W, p_M) is also a beta curve:
  - p_W is beta(10 + 1, 10 + 1)
  - p_M is beta(14 + 1, 6 + 1)

Simulation to summarize posterior

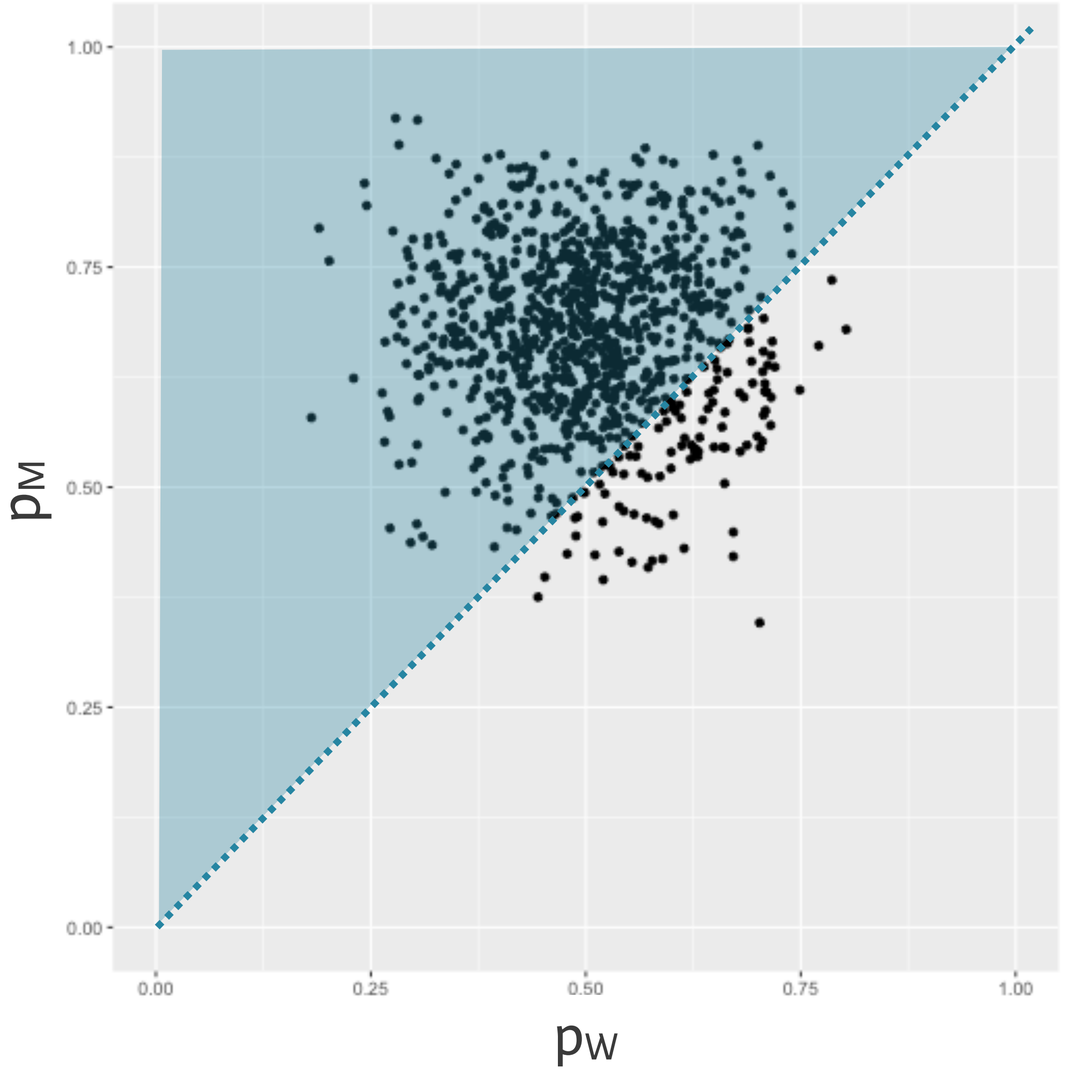
```
> # Simulate pW from beta(11, 11) curve
> pW <- rbeta(1000, 11, 11)

> # Simulate pM from beta(15, 7) curve
> pM <- rbeta(1000, 15, 7)
```

Graph of posterior of (p_W, p_M)

```
> df <- data.frame(pW, pM)
> ggplot(df, aes(pW, pM)) + geom_point() + xlim(0, 1) + ylim(0, 1)
```



Prob($p_W < p_M$)?



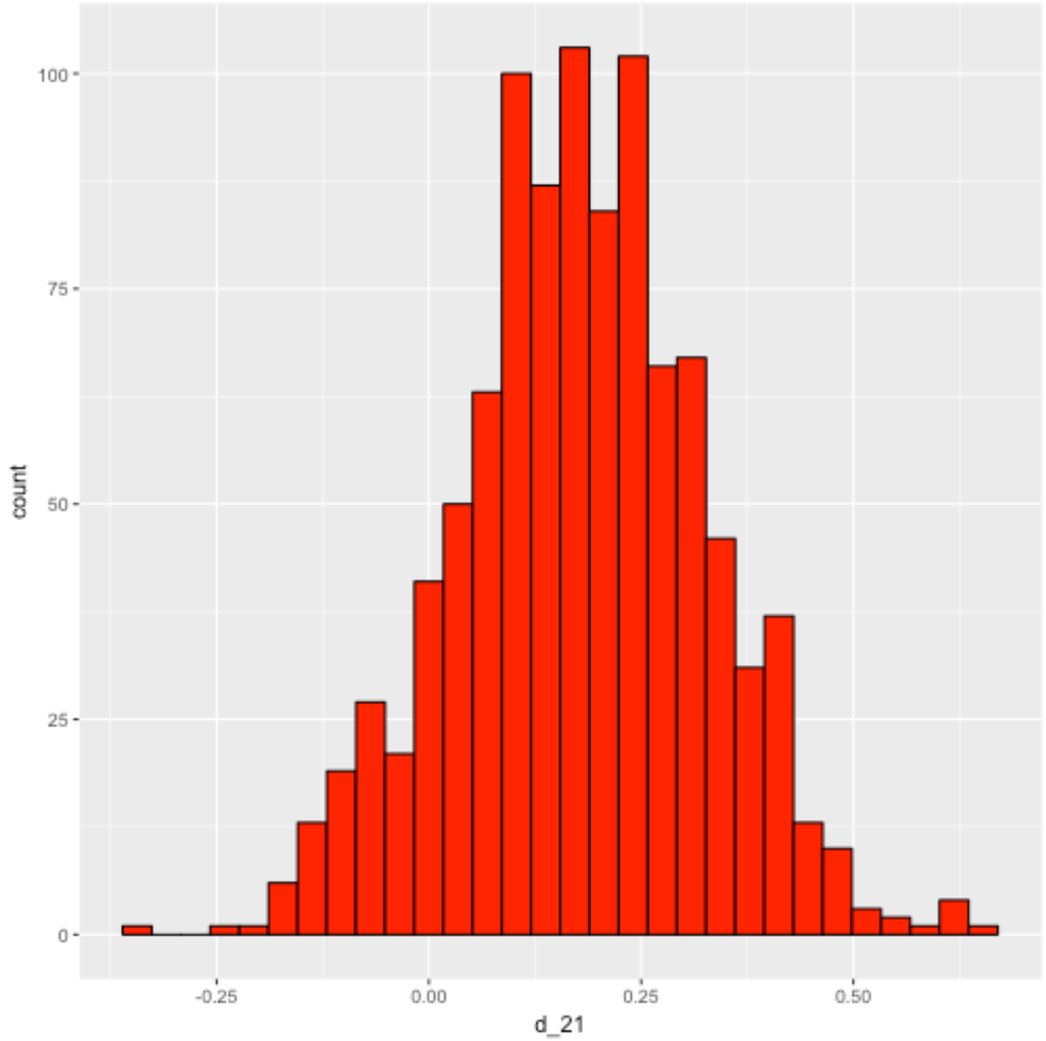
Prob($p_W < p_M$)?

```
> # Probability that pW < pM
> with(df, sum(pW < pM) / 1000)
[1] 0.891
```

Posterior of difference p_M - p_W

```
> # For each simulated (pW, pM), compute d = pM - pW
> df$d_21 <- with(df, pM - pW)

> # Plot histogram
> ggplot(df, aes(d_21)) +
    geom_histogram(color = "black", fill = "red")
```

Posterior of difference p_M - p_W



Probability interval for d

```
> # Compute 90% interval
> (Q <- quantile(df$d_21, c(0.05, 0.95)))
         5%          95%
-0.07442153   0.41724768
```

P(-0.07 < pM - pW < 0.42) = 0.9

Since the interval contains zero, there's no significant evidence to say the proportions are different

BEGINNING BAYES IN R

Let’s practice!

BEGINNING BAYES IN R

Normal model inference

Learning about a normal model

- **Chapter 3**: inference on mean M of a normal sampling model, assumed standard deviation S
- **Chapter 4**: mean M and standard deviation S are both unknown
- Revisit Roger Federer's time-to-serve data

Prior?

- Both M and S are continuous
- Not easy to think about beliefs about pairs (M, S)
- So we focus on the use of a standard "non-informative" prior

Non-informative prior

- Standard non-informative prior for mean M and standard deviation S looks like:

$$g(M, S) = \frac{1}{S}$$

- How to understand this prior?
  - Assign M a normal prior with large standard deviation
  - Assign S a normal prior with large standard deviation
  - These beliefs approximate non-informative prior

The data

```
> # Input observed times-to-serve
> Fed <- data.frame(Player = "Federer",
                    Time_to_Serve = c(20.9, 17.8, 14.9, 12.0, 14.1,
                                      22.8, 14.6, 15.3, 21.2, 20.7,
                                      12.2, 16.2, 15.6, 19.4, 22.3,
                                      14.1, 18.1, 23.6, 11.0, 17.3))
```

Posterior?

- Likelihood of this data is given:

```
> Likelihood <- prod(dnorm(Time_to_Serve, mean = M, sd = S))
```

- Posterior density of (M, S):

$$Posterior = Likelihood \times \frac{1}{S}$$

Non-informative prior

Posterior calculation

- Simulate (M, S) from the 2-parameter posterior
- Summarize posterior sample to perform inference
- Simulate using the `sim()` method from the `arm` package

Using the arm package

```
> # Regression model with only an intercept
> fit <- lm(Time_to_Serve ~ 1, data = Fed)
> summary(fit)

Call:
lm(formula = Time_to_Serve ~ 1, data = Fed)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   17.205      0.851   20.22 2.62e-14 ***
---
Residual standard error: 3.806 on 19 degrees of freedom
```

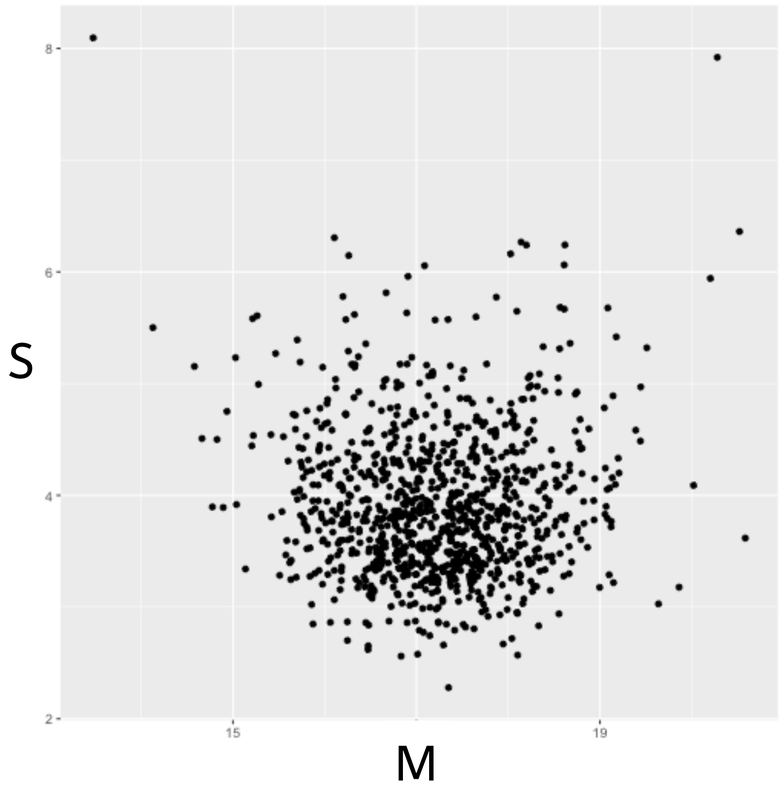
Simulate from lm()

```
> library(arm)
> sim_fit <- sim(fit, n.sims = 1000)
> sim_M <- coef(sim_fit)
> sim_S <- sigma.hat(sim_fit)
```

- Simulates from posterior of (M, S) using non-informative prior
- Extract the simulated values of M and S, respectively

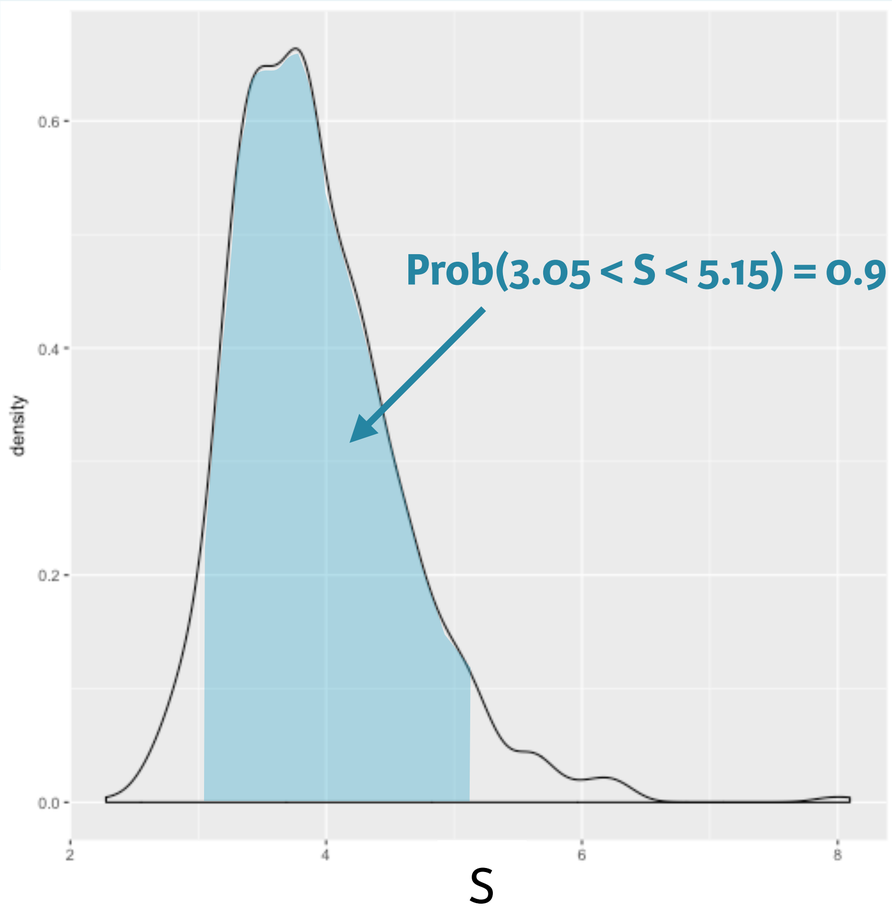
Plot the posterior sample of (M, S)

```
> library(ggplot2)
> ggplot(data.frame(sim_M, sim_S), aes(sim_M, sim_S)) +
    geom_point()
```



Learn about standard deviation S?

```
> ggplot(data.frame(sim_S), aes(sim_S)) +
    geom_density()

> quantile(sim_S, c(0.05, 0.95))
      5%      95%
3.049684 5.154036
```



BEGINNING BAYES IN R

Let’s practice!

BEGINNING BAYES IN R

Bayesian regression

Comparing time-to-serve

- We were exploring time-to-serve data for Roger Federer
- Let's compare him with a slow player like Rafael Nadal
- How much slower is Rafa than Roger?

Sampling model

- Regression framework:
  - Response variable: `Time_to_serve`
  - Single covariate: `Player` (Federer or Nadal)

```
> # Fit regression line using lm()
> lm(Time_to_serve ~ Player, data = Tennis)
```

Bayesian model

- Modeling time-to-serve (in seconds)
- Sampling level:

$$y \sim N(\beta_0 + \beta_1 I(Player = Nadal), S)$$

- Prior:

$$(\beta_0, \beta_1, S) \sim \frac{1}{S}$$

Likelihood and posterior

```
> Likelihood <- prod(dnorm(Time_to_Serve,
                           mean = beta0 + beta1 * I(Nadal), sd = S))
```

$$Posterior \propto Likelihood \times \frac{1}{S}$$

Using the arm package

- `lm()` - Fit regression model
- `sim()` - Simulate from posterior density
- `coef()` - Extract draws of regression parameters β_0, β_1
- `sigma.hat()` - Extract simulated draws of S

The data

```
> Fed <- data.frame(Player = "Federer",
                    Time_to_Serve = c(20.9, 17.8, 14.9, 12.0, 14.1,
                                      22.8, 14.6, 15.3, 21.2, 20.7,
                                      12.2, 16.2, 15.6, 19.4, 22.3,
                                      14.1, 18.1, 23.6, 11.0, 17.3))

> Rafa <- data.frame(Player = "Nadal",
                     Time_to_Serve = c(20.5, 25.1, 21.4, 25.6, 41.2,
                                       23.9, 22.6, 19.0, 29.7, 36.4,
                                       18.4, 20.3, 26.9, 28.9, 22.9,
                                       31.5, 39.6, 29.4, 26.9, 24.5))

> Tennis <- rbind(Fed, Rafa)
```

Fit the regression model

```
> fit <- lm(Time_to_Serve ~ Player, data = Tennis)
> fit

Call:
lm(formula = Time_to_Serve ~ Player, data = Tennis)

Coefficients:
(Intercept)  PlayerNadal
      17.20         9.53
```

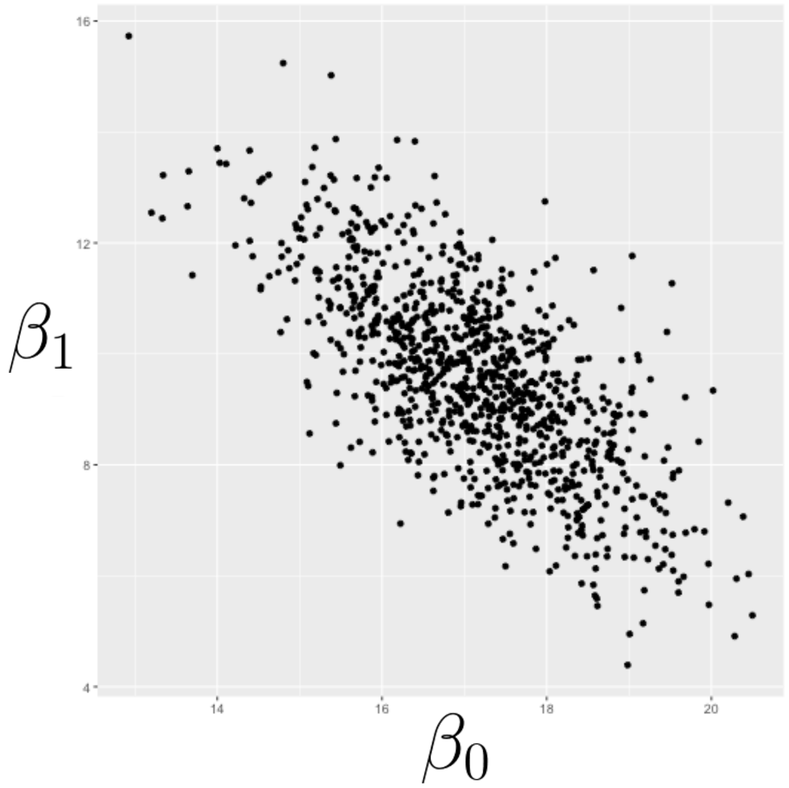
	Federer	Nadal
Time-to-serve	17.2	17.20 + 9.53

The arm package

```
> library(arm)
> sim_fit <- sim(fit, n.sims = 1000)
> sim_beta <- coef(sim_fit)
> sim_S <- sigma.hat(sim_fit)
```

Graph of posterior of regression

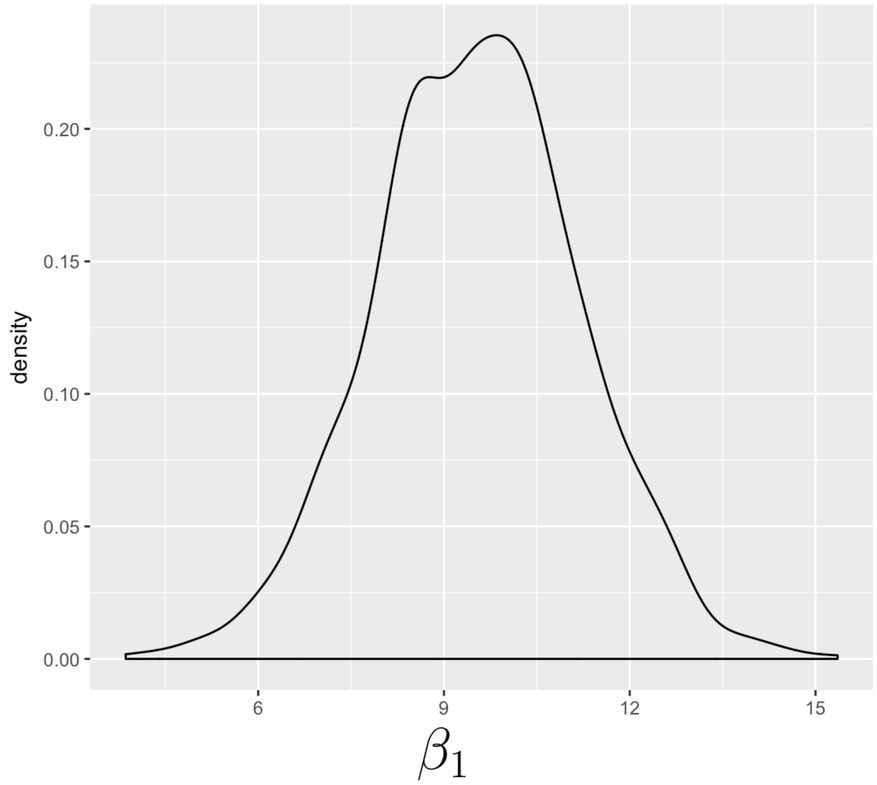
```
> sim_beta <- data.frame(sim_beta)
> names(sim_beta)[1] <- "Intercept"
> ggplot(sim_beta, aes(Intercept, PlayerNadal)) + geom_point()
```



How much slower is Rafa?

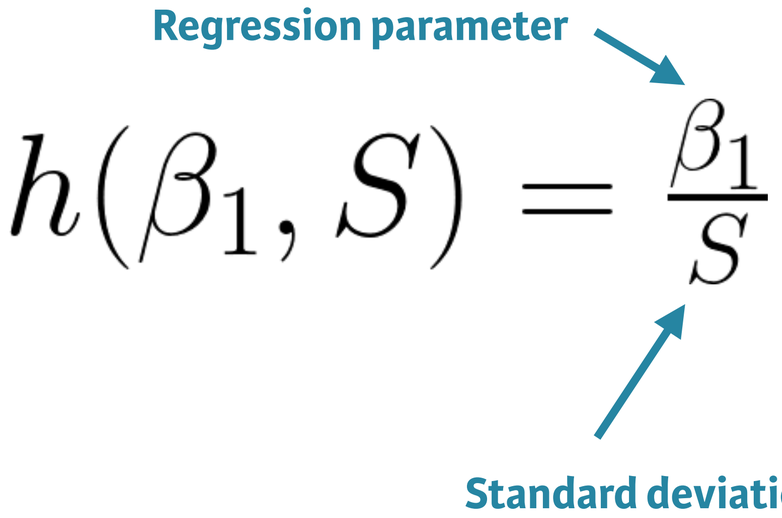
Look at the posterior of β_1, the difference in means

```
> ggplot(sim_beta, aes(PlayerNadal)) + geom_density()
```



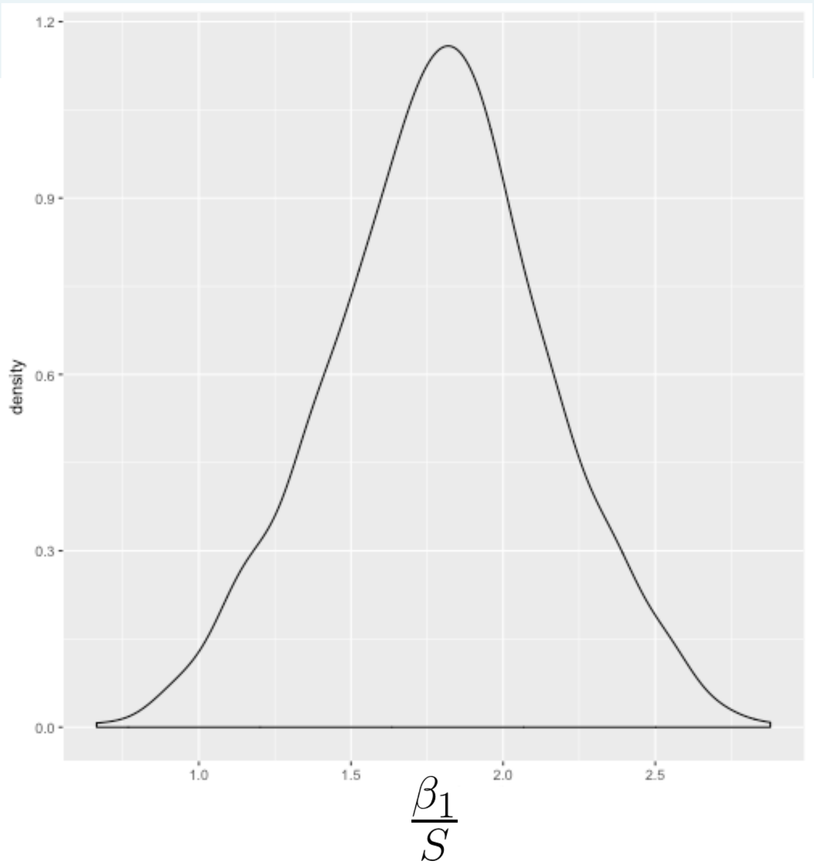
Interested in standardized effect

Standardized effect: average time that Rafa is slower than Federer, measured in standard deviation units



Posterior of standardized effect

```
> posterior <- data.frame(sim_beta, sim_S)
> standardized_effect <- with(posterior, PlayerNadal / sim_S)
> ggplot(posterior, aes(standardized_effect)) +
    geom_density()
```



90% probability interval for standardized effect

```
> sim_beta <- data.frame(sim_beta)
> quantile(sim_beta$PlayerNadal / sim_S, c(0.05, 0.95))
      5%      95%
1.147181 2.411296
```

90% probability interval

BEGINNING BAYES IN R

Let’s practice!

BEGINNING BAYES IN R

Wrap-up and review

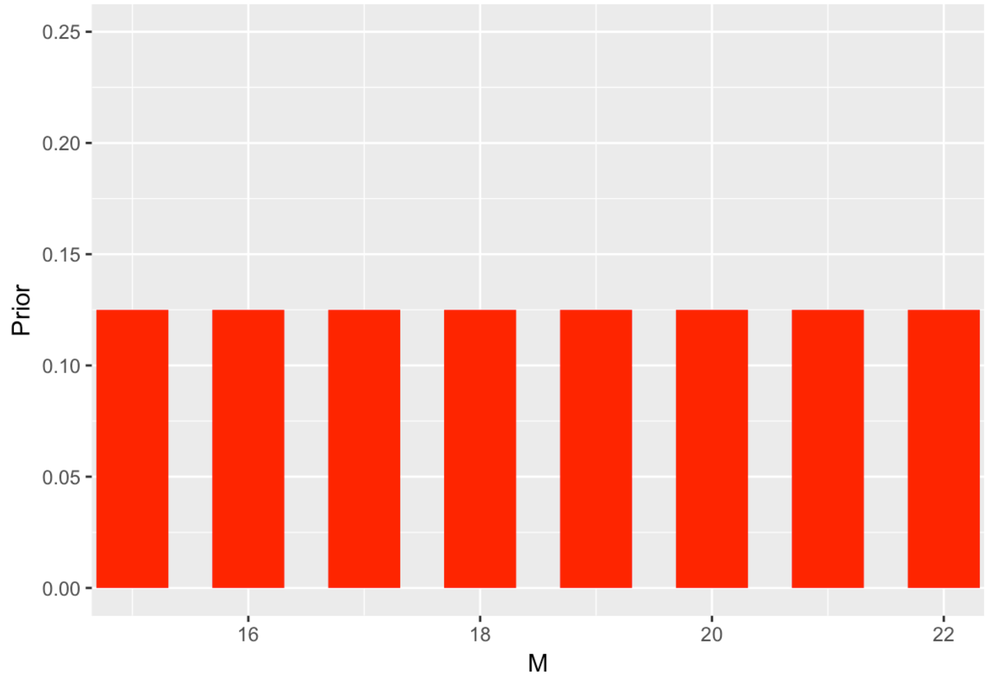
Update probabilities using Bayes' rule

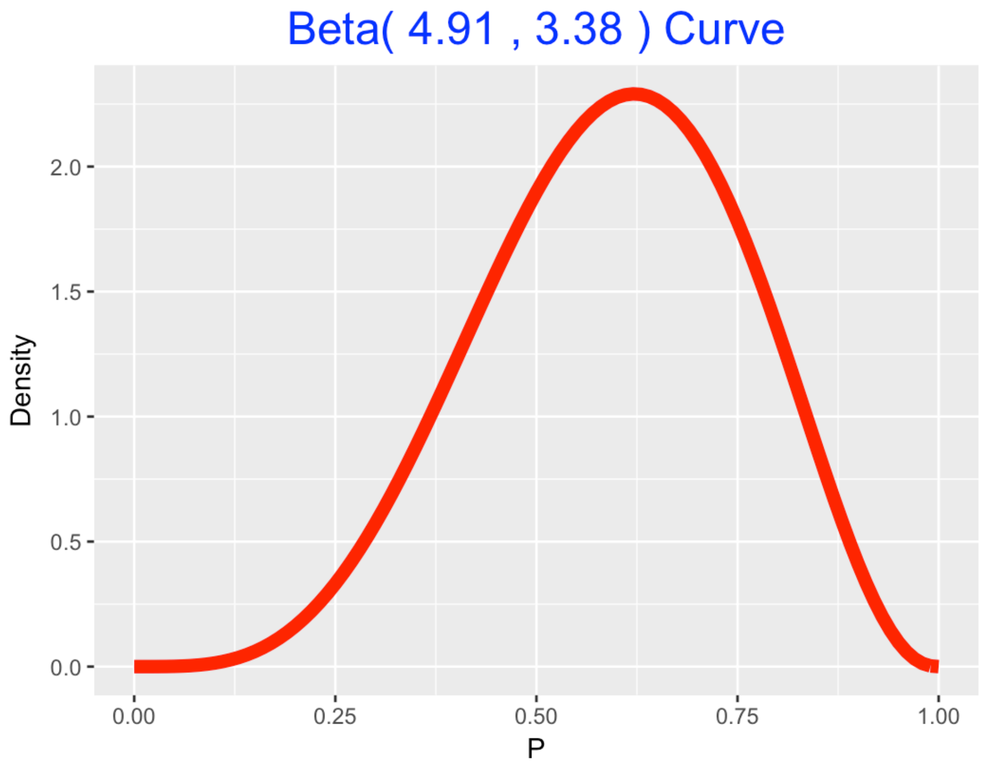
```
> library(TeachBayes)
> bayesian_crank(bayes_df)
      Model Prior Likelihood Product Posterior
1 Spinner A  0.25       0.33  0.0825     0.264
2 Spinner B  0.25       0.50  0.1250     0.400
3 Spinner C  0.25       0.25  0.0625     0.200
4 Spinner D  0.25       0.17  0.0425     0.136
```

Prior x Likelihood = Product

Product / sum(Product) = Posterior

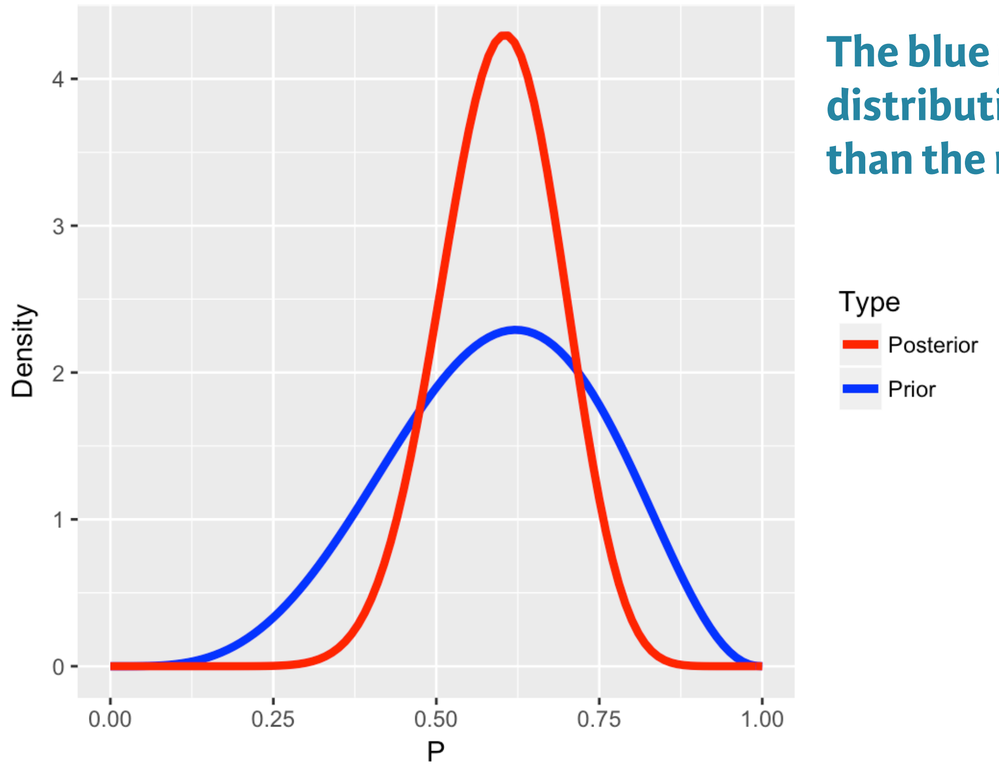
Construct priors





Obtain a posterior

```
> # Overlay posterior on prior curve
> library(TeachBayes)
> beta_prior_post(prior_par, post_par)
```



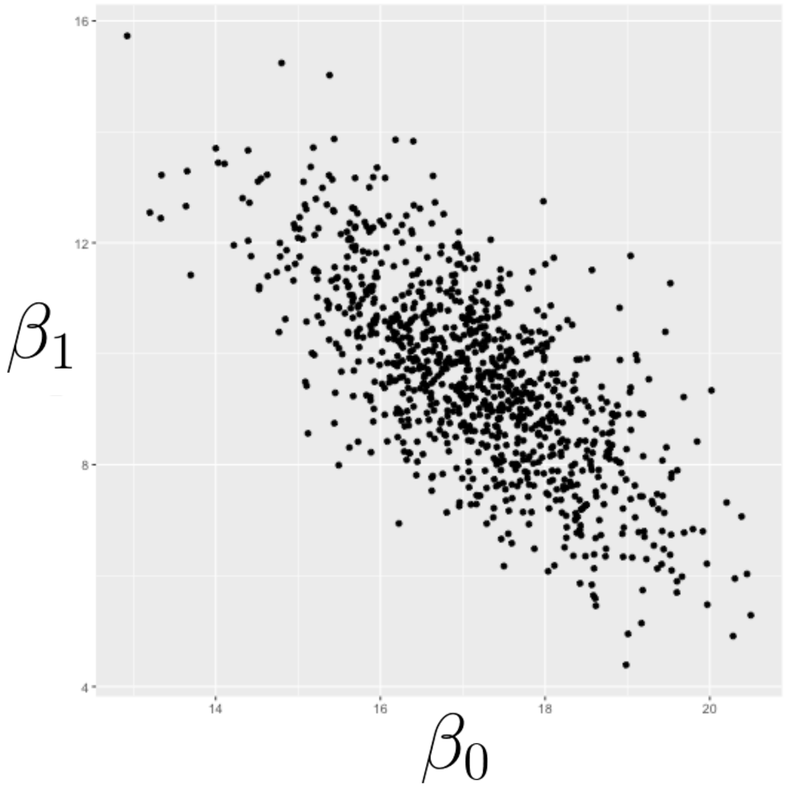
The blue prior curve shows a wider distribution, showing more uncertainty than the red posterior curve

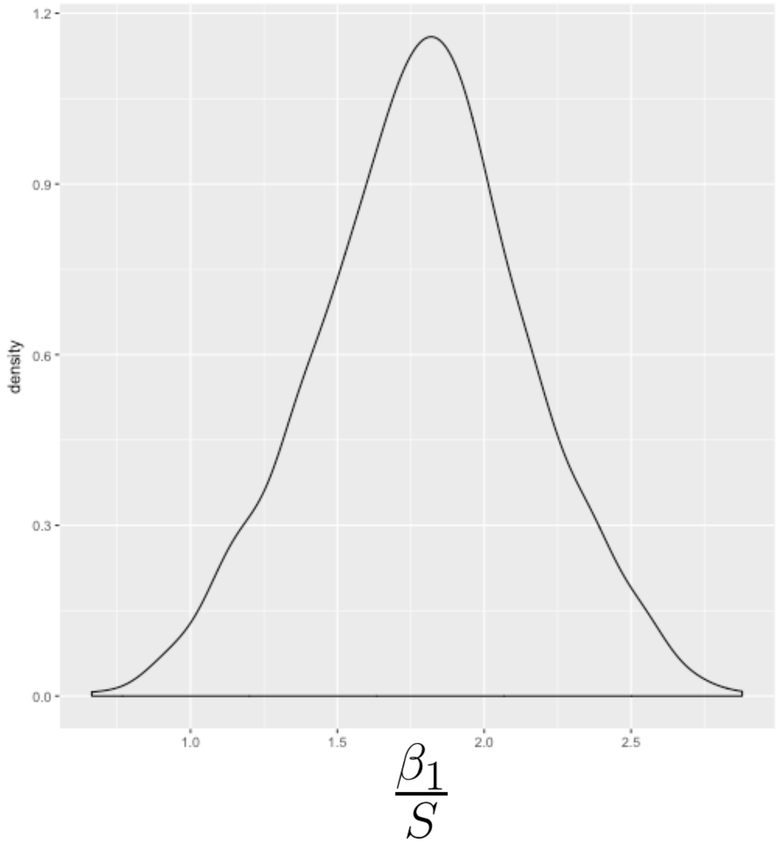
Summarize using simulation

```
> library(arm)
> sim_fit <- sim(fit, n.sims = 1000)
> sim_M <- coef(sim_fit)
> sim_S <- sigma.hat(sim_fit)
```

- `sim()` simulates from posterior of (M, S) using a non-informative prior
- `coef()` and `sigma.hat()` extract the simulated values of M and S, respectively

Inference about multiple parameters





Thanks!