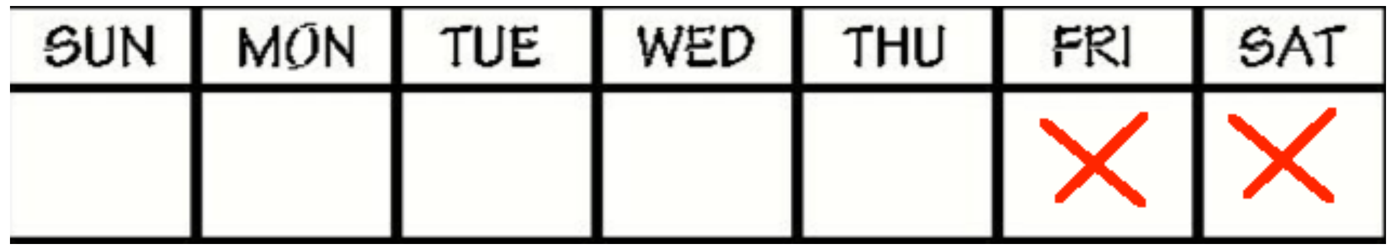
Bayes with discrete models

Survey on eating out

What is your favorite day for eating out?

Construct a prior for p

- **Define p**: proportion of all students who answer "Friday or Saturday"



- **Form opinion about p**: 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 are plausible values
    - 0.5 and 0.6 are most likely
    - Each of these is twice as likely as other values

Construct a prior for p

```
> bayes_df <- data.frame(P = seq(0.3, 0.8, by = 0.1),
                         Weight = c(1, 1, 2, 2, 1, 1),
                         Prior = c(1, 1, 2, 2, 1, 1) / 8)
> bayes_df
    P Weight Prior
1 0.3      1 0.125
2 0.4      1 0.125
3 0.5      2 0.250
4 0.6      2 0.250
5 0.7      1 0.125
6 0.8      1 0.125
```

Collect data

- Survey random sample of 20 students
- 12 out of 20 say “Friday or Saturday”
- This is a **binomial experiment**

Likelihood

Chance of 12 "successes" out of 20 where p is probability of success:

$$LIKELIHOOD = \binom{20}{12} p^{12} (1-p)^{8}$$

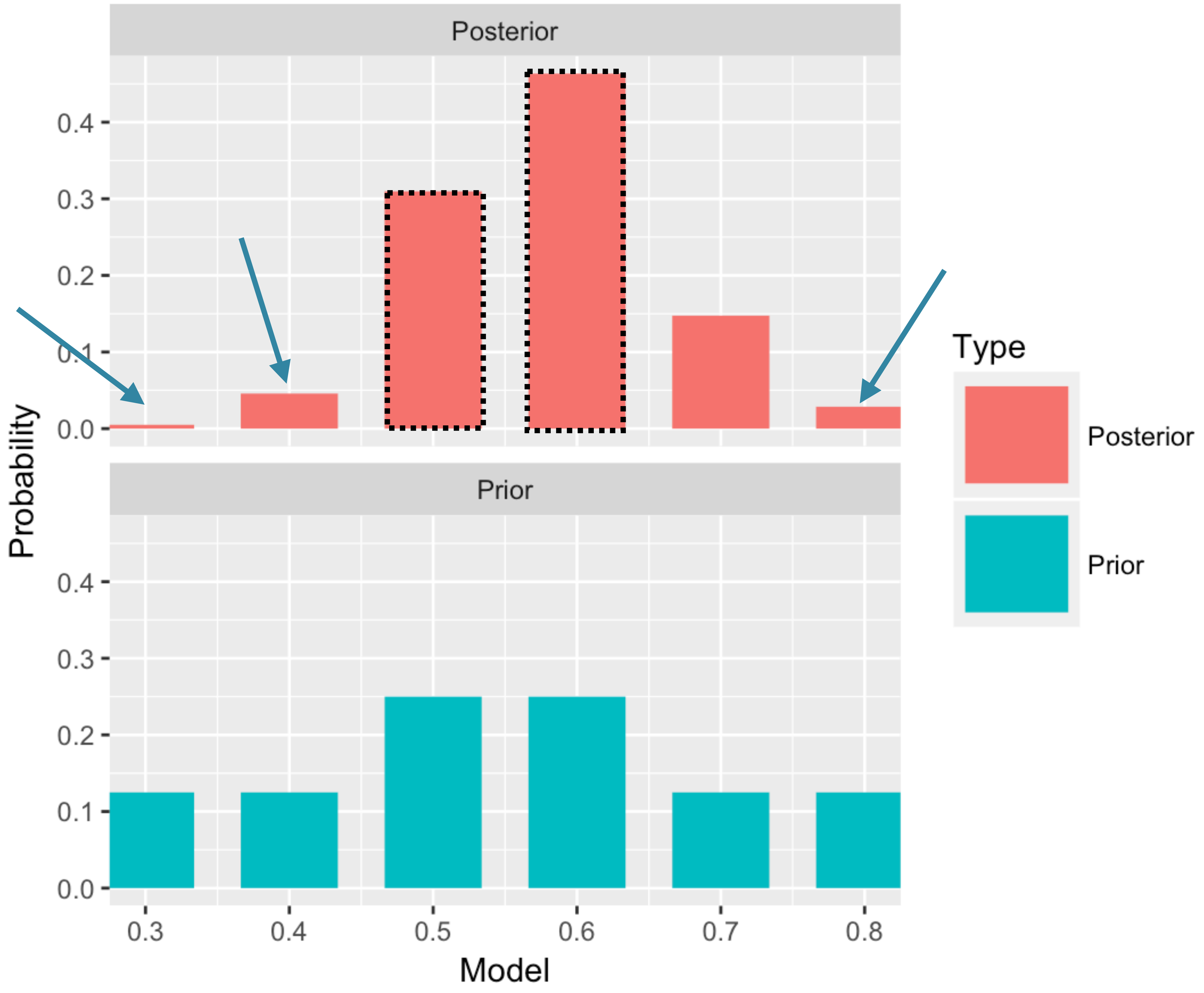
```
> # Add likelihood to the table:
> bayes_df$Likelihood <- dbinom(12, size = 20, prob = bayes_df$P)

> bayes_df <- bayes_df[, c(1, 3, 4)]
> round(bayes_df, 3)

    P Prior Likelihood
1 0.3 0.125      0.004
2 0.4 0.125      0.035
3 0.5 0.250      0.120
4 0.6 0.250      0.180
5 0.7 0.125      0.114
6 0.8 0.125      0.022
```

Turn the Bayesian crank

```
> library(TeachBayes)
> bayes_df <- bayesian_crank(bayes_df)
> round(bayes_df, 3)

    P Prior Likelihood Product Posterior
1 0.3 0.125      0.004   0.000     0.005
2 0.4 0.125      0.035   0.004     0.046
3 0.5 0.250      0.120   0.030     0.310
4 0.6 0.250      0.180   0.045     0.463
5 0.7 0.125      0.114   0.014     0.147
6 0.8 0.125      0.022   0.003     0.029
```

Compare prior and posterior

```
> library(TeachBayes)
> prior_post_plot(bayes_df)
```



Statistical inference

What is the probability that p is larger than 0.5?

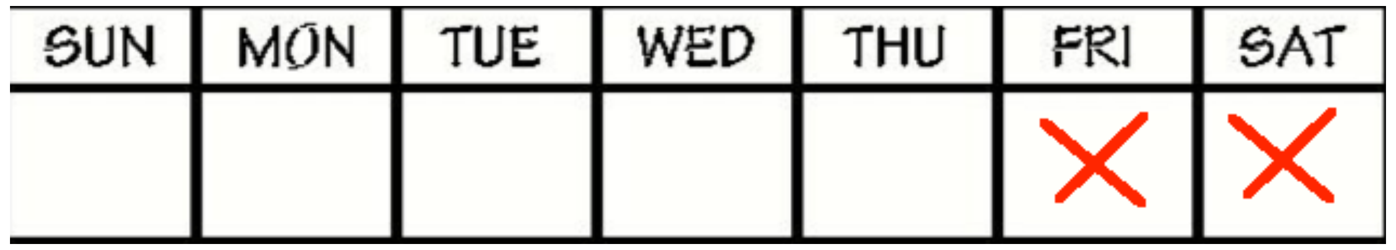
```
> round(bayes_df[, c("P", "Posterior")], 3)

    P  Posterior
1 0.3      0.005
2 0.4      0.046
3 0.5      0.310
4 0.6      0.463
5 0.7      0.147
6 0.8      0.029
```

$$Prob(p > 0.5) = 0.463 + 0.147 + 0.029 = 0.639$$

Let's practice!

Continuous prior

Dining survey example



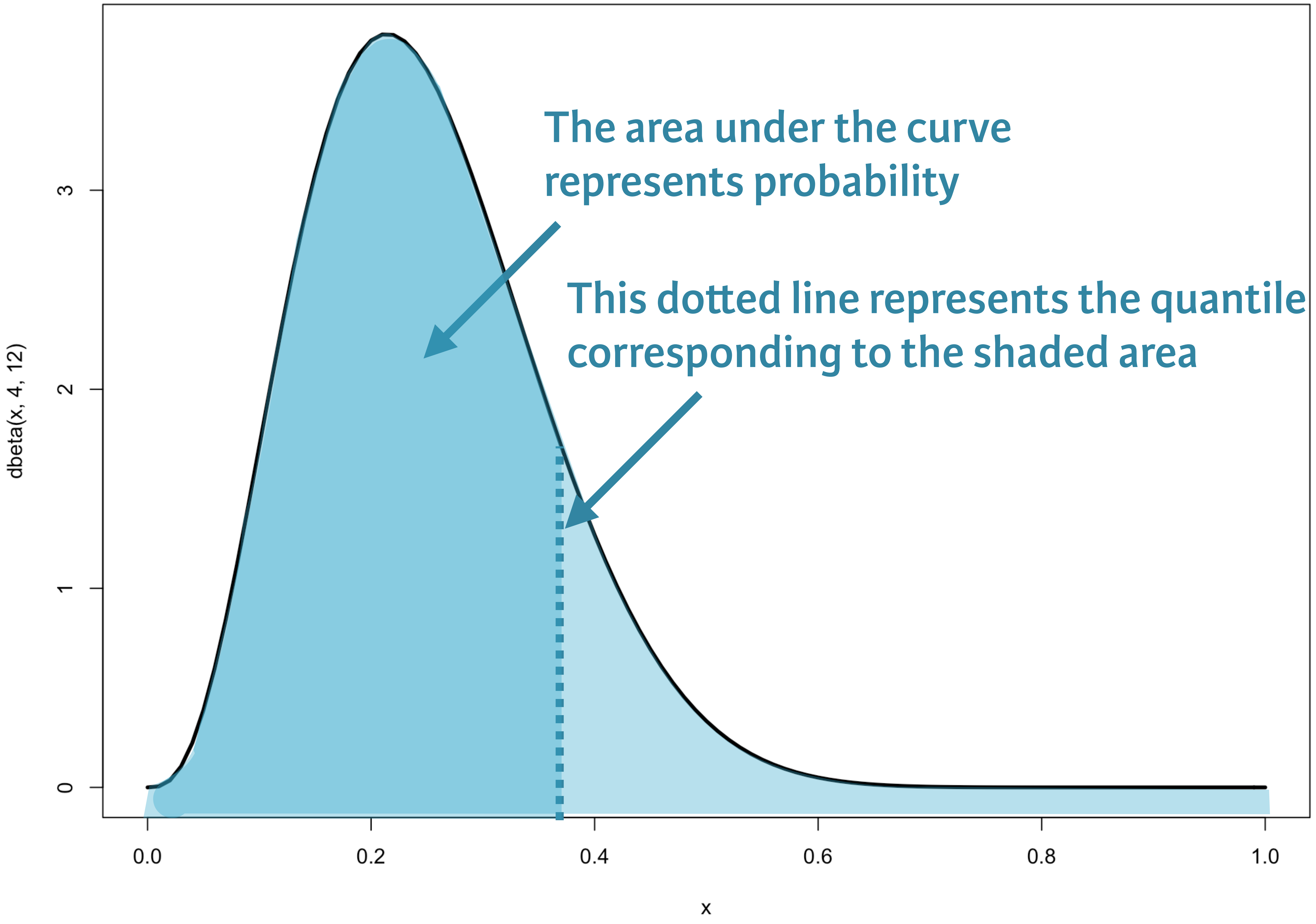
- **Define p**: proportion of all students who answer "Friday or Saturday"
- **Form opinion about p:** continuous on (0, 1)
    - Represent prior probabilities by a beta curve:

$$PRIOR = p^{a-1}(1-p)^{b-1}$$

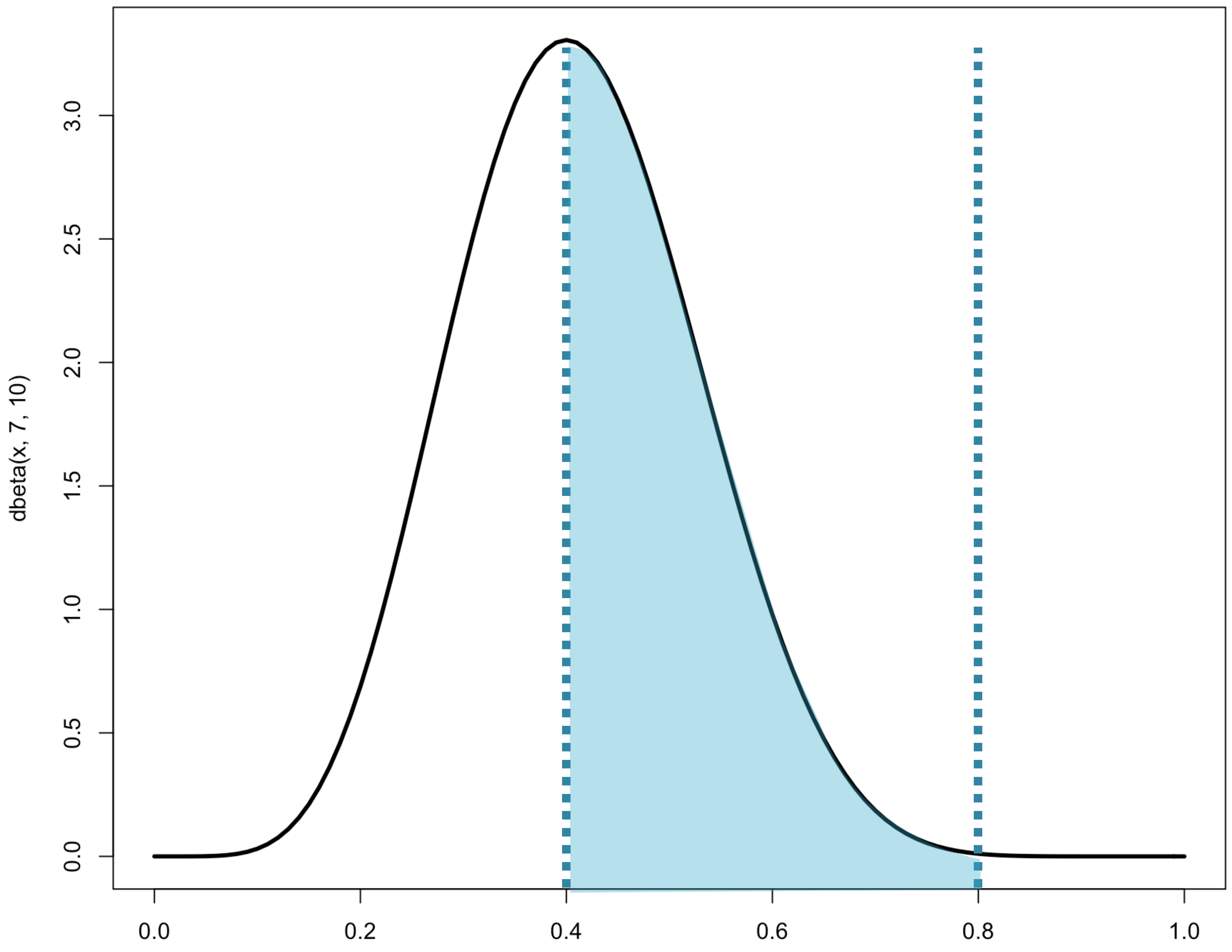
    - ***a*** and ***b*** are **shape parameters** of the beta curve

Example beta curve

Beta curve with shape parameters 4 and 12:

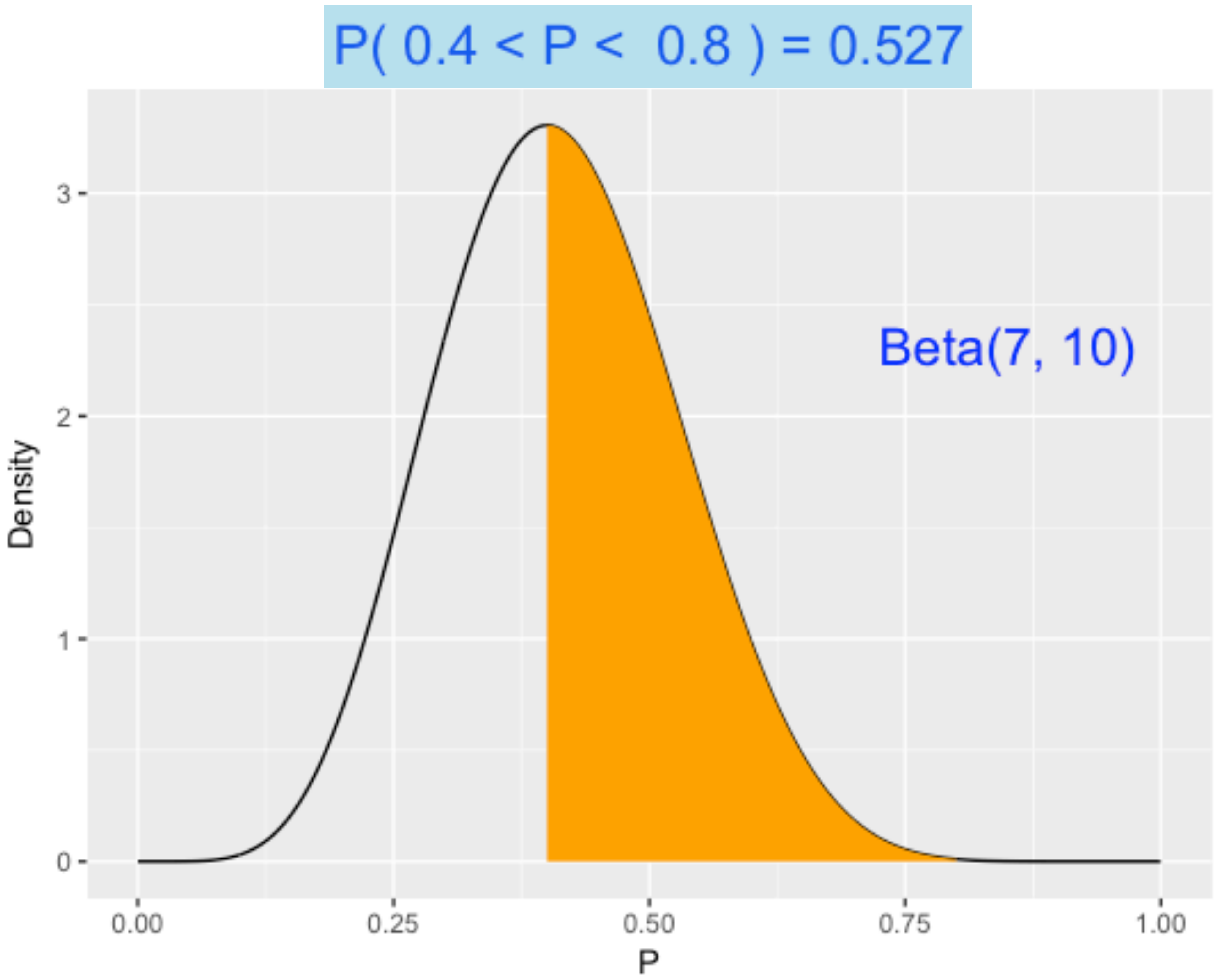


Probability for beta(7, 10) curve

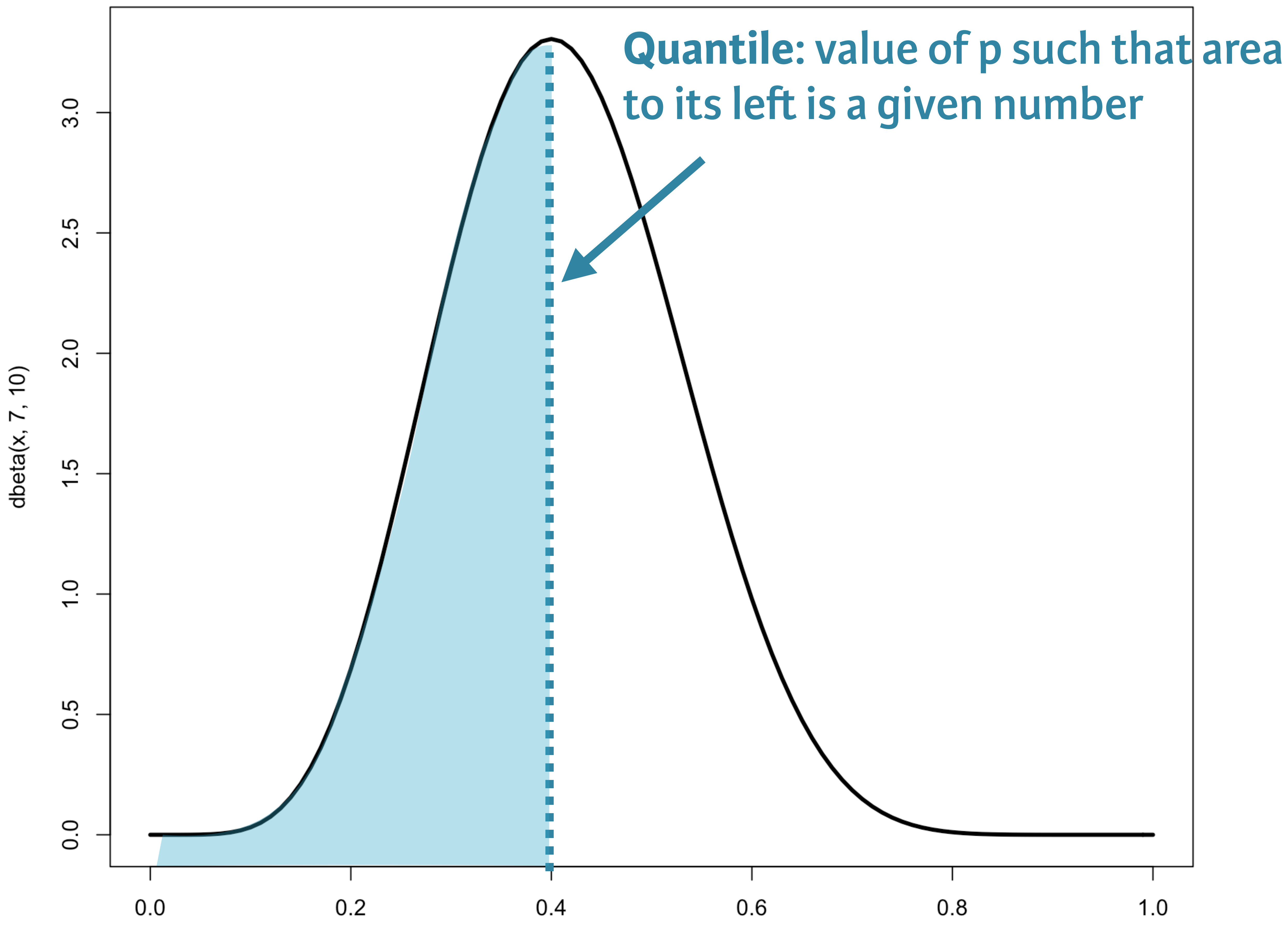


Probability for beta(7, 10) curve

```
> # Probability that p is between 0.4 and 0.8
> library(TeachBayes)
> beta_area(0.4, 0.8, c(7, 10))
```

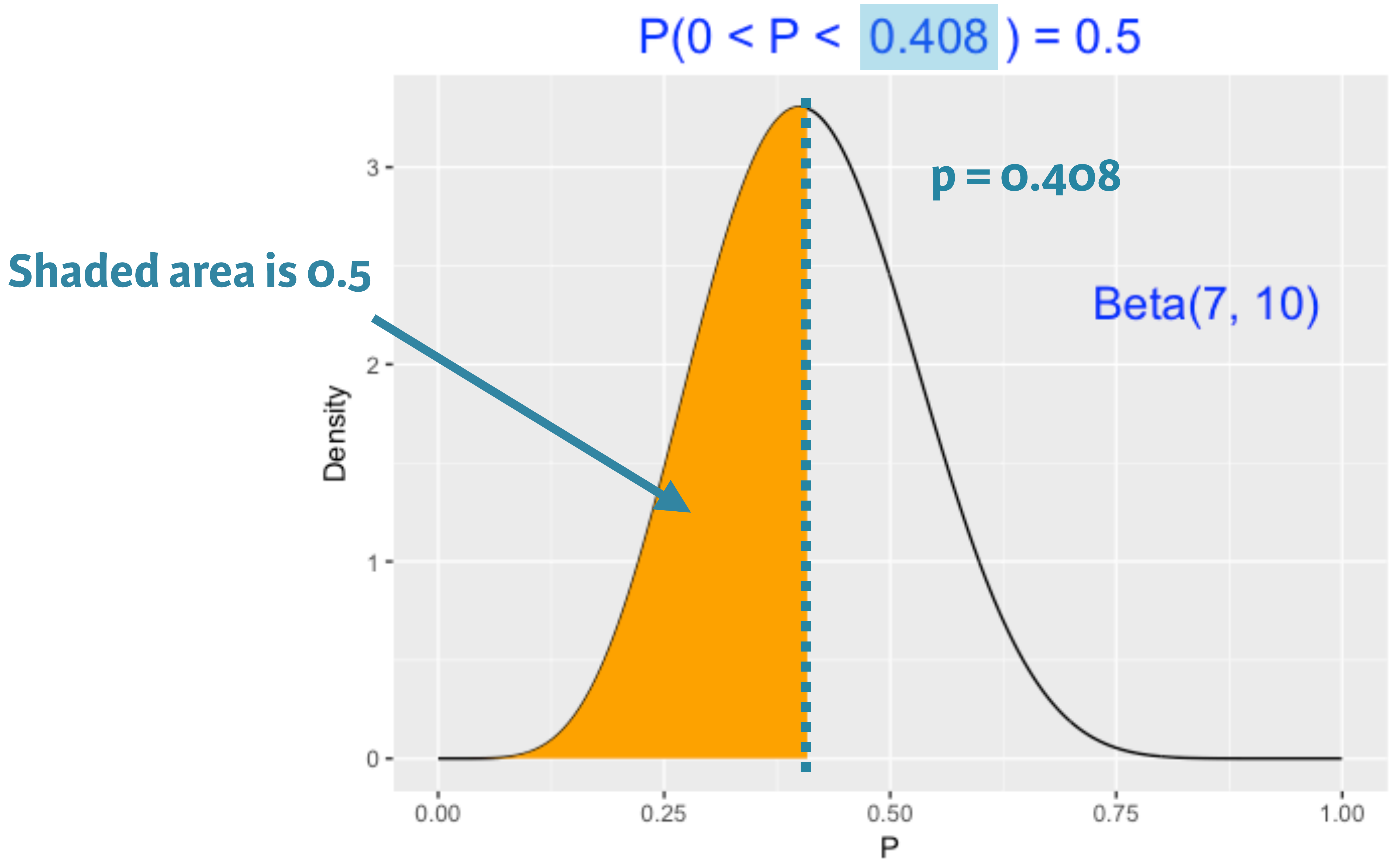


Quantile for beta(7, 10) curve



Quantile for beta(7, 10) curve

```
> # Finding the 0.50 quantile:
> library(TeachBayes)
> beta_quantile(0.5, c(7, 10))
```



Constructing a prior

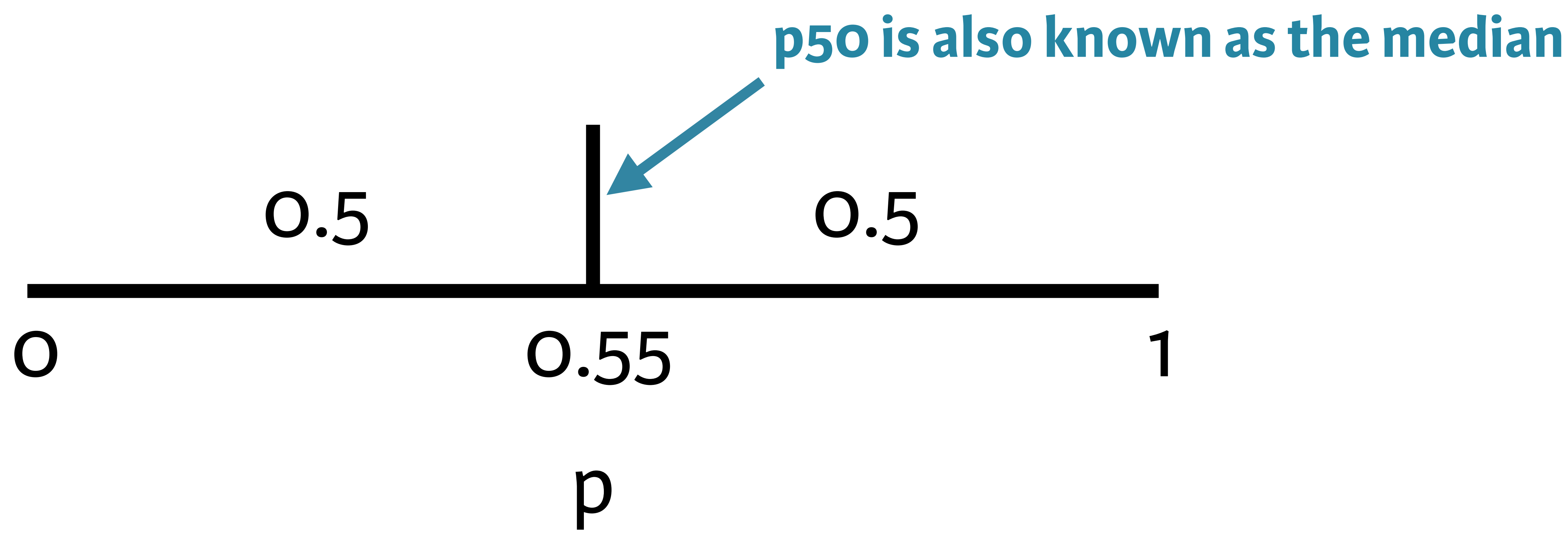
- Use beta curve to represent prior opinion about p
- Hard to pick values of ***a*** and ***b*** directly
- Instead, think of characteristics of the curve that are easier to specify (e.g. quantiles of the curve)

Finding a beta prior for p

- Specify the 0.5 and 0.9 quantiles
- Use `beta.select()` to find the parameters of the beta curve that match this information

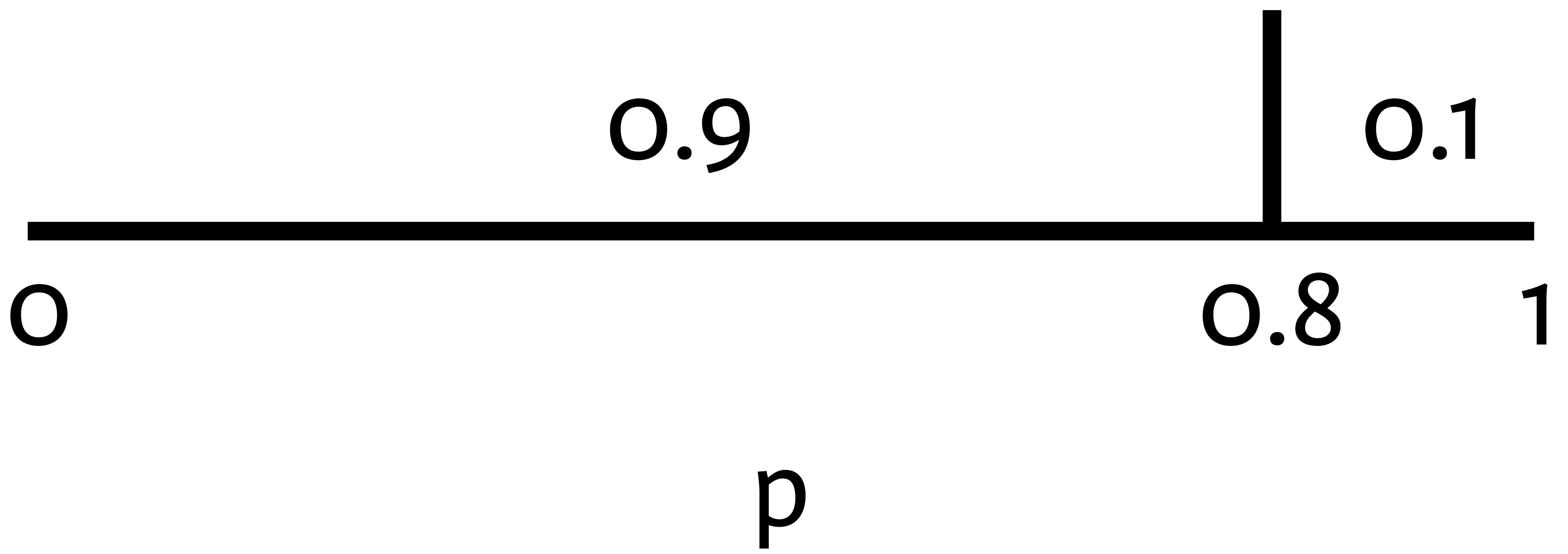
Specify 0.50 quantile

- **p50** represents 0.50 quantile for p, meaning p is equally likely to be smaller or larger than **p50**
- After some thought... decide that **p50** is 0.55



Specify 0.90 quantile

- **p90** represents 0.90 quantile for p, meaning p is likely (with probability 0.90) to be smaller than **p90**
- After more thought... decide that **p90** = 0.80



Find the matching beta curve

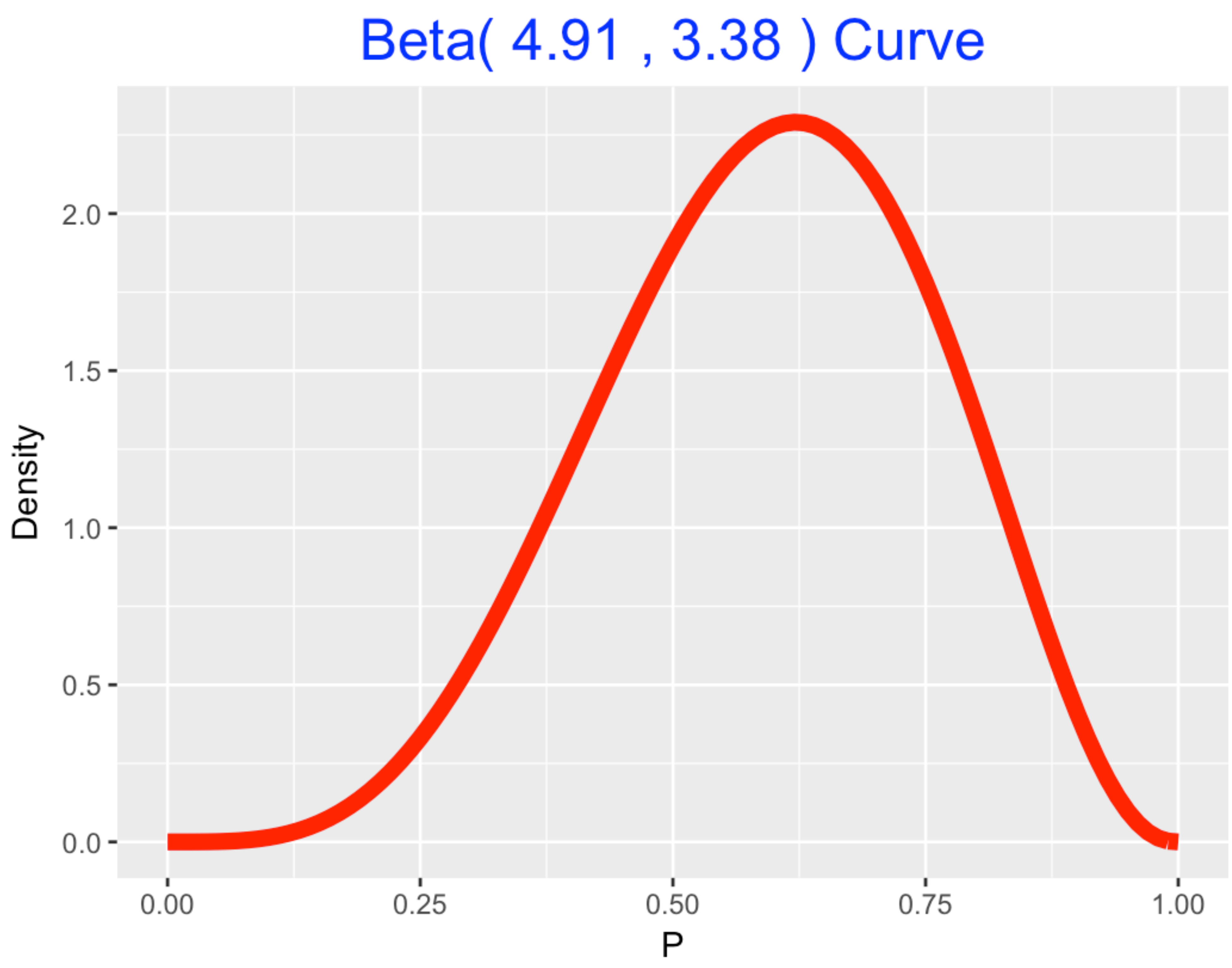
`beta.select()` finds shape parameters of the beta curve that has the same two quantiles

```
> # Specify 0.50 and 0.90 quantiles
> p50 <- list(x = 0.55, p = 0.5)
> p90 <- list(x = 0.80, p = 0.9)

> # Find the matching beta curve
> library(TeachBayes)
> beta.select(p50, p90)
[1] 4.91 3.38
```

corresponds to *a* and *b* shape parameters, respectively

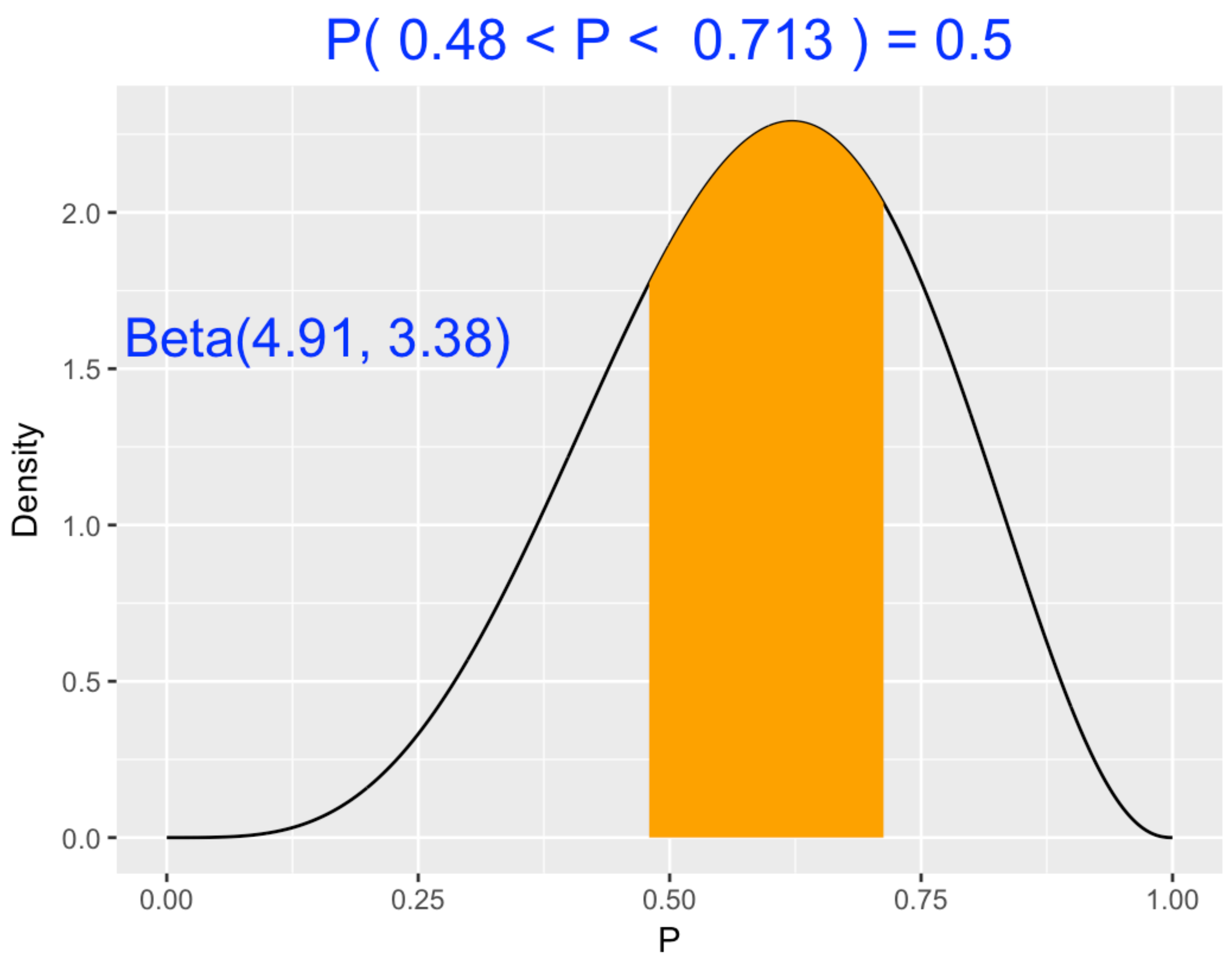
My beta(4.91, 3.38) curve for p

```
> # Plot beta prior obtained from beta.select()
> library(TeachBayes)
> beta_draw(c(4.91, 3.38))
```



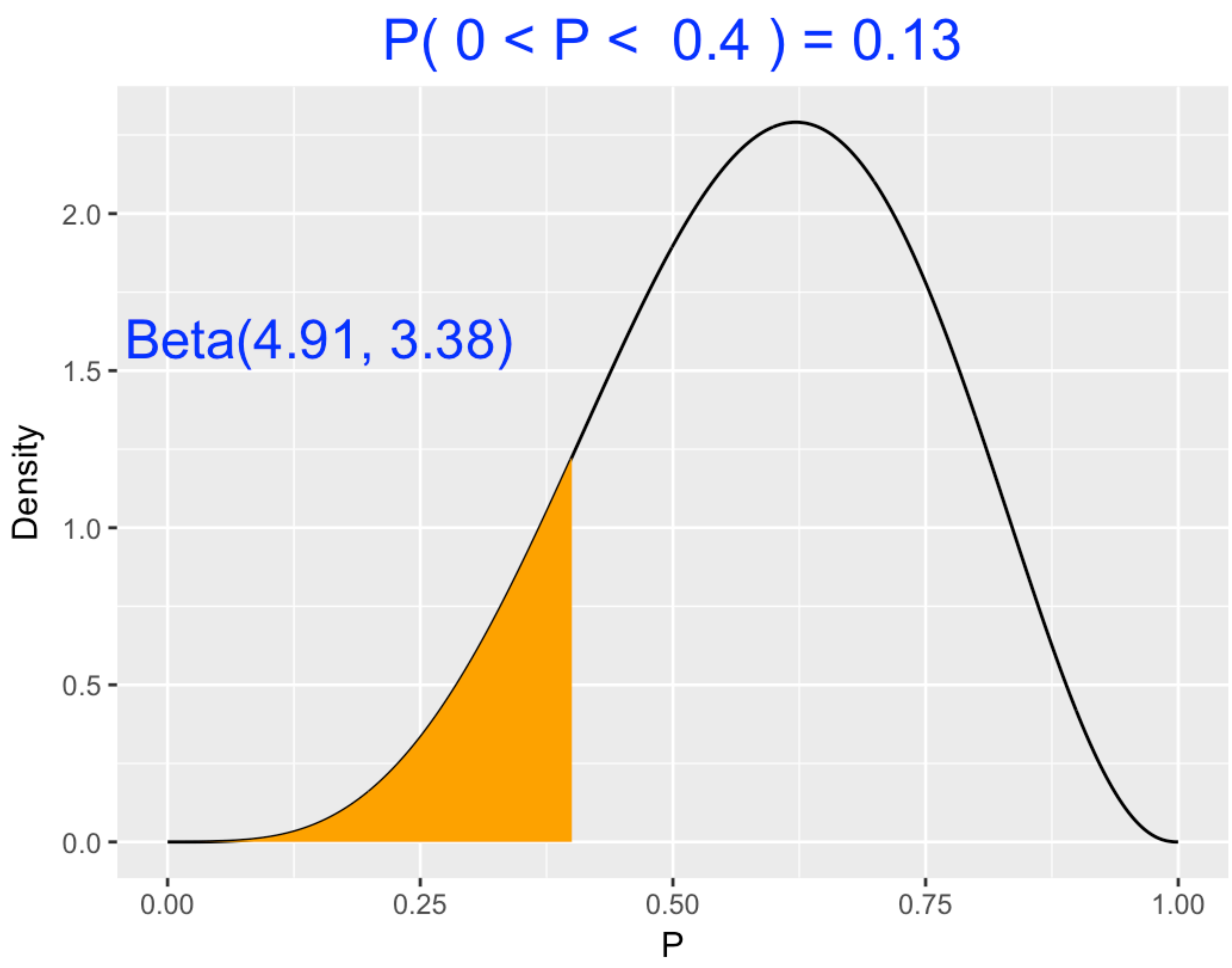
Reasonable prior?

```
> # Compute 50% probability interval
> library(TeachBayes)
> beta_interval(0.5, c(4.91, 3.38))
```



Reasonable prior?

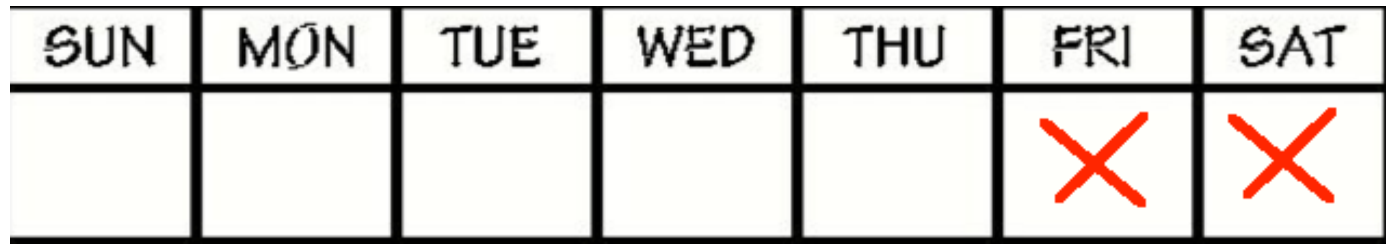
```
> # Compute probability p is smaller than 0.4
> library(TeachBayes)
> beta_area(0, 0.4, c(4.91, 3.38))
```



Let’s practice!

Updating the beta prior

Recall dining survey example



- **Interested in p**: proportion of all students who answer "Friday or Saturday"
- **Formed opinion about p:** continuous on (0, 1)
  - Represented prior probabilities by a beta curve:

$$PRIOR = p^{a-1}(1-p)^{b-1}$$

  - with 4.91 and 3.38 as ***a*** and ***b***
- **Survey results:** 12 of 20 say "Friday or Saturday"

Likelihood for binomial sampling

Chance of 12 successes in sample of 20:

$$LIKELIHOOD = \binom{20}{12} p^{12} (1-p)^{8}$$

Probability of success

Bayes' rule (a little math)

Compute posterior:

$$POSTERIOR \propto PRIOR \times LIKELIHOOD$$

means "is proportional to"

Combine beta prior and binomial sampling

Posterior is proportional to (beta prior) x (binomial likelihood):

$$POSTERIOR \propto [p^{4.91-1}(1-p)^{3.38-1}] \times [p^{12}(1-p)^{8}]$$

$$= p^{16.91-1}(1-p)^{11.38-1}$$

Recall that beta curves have this functional form: $p^{a-1}(1-p)^{b-1}$

So the posterior is also a beta curve

From a beta prior to a beta posterior

	Beta shape 1	Beta shape 2
Prior		
Data		
Posterior		

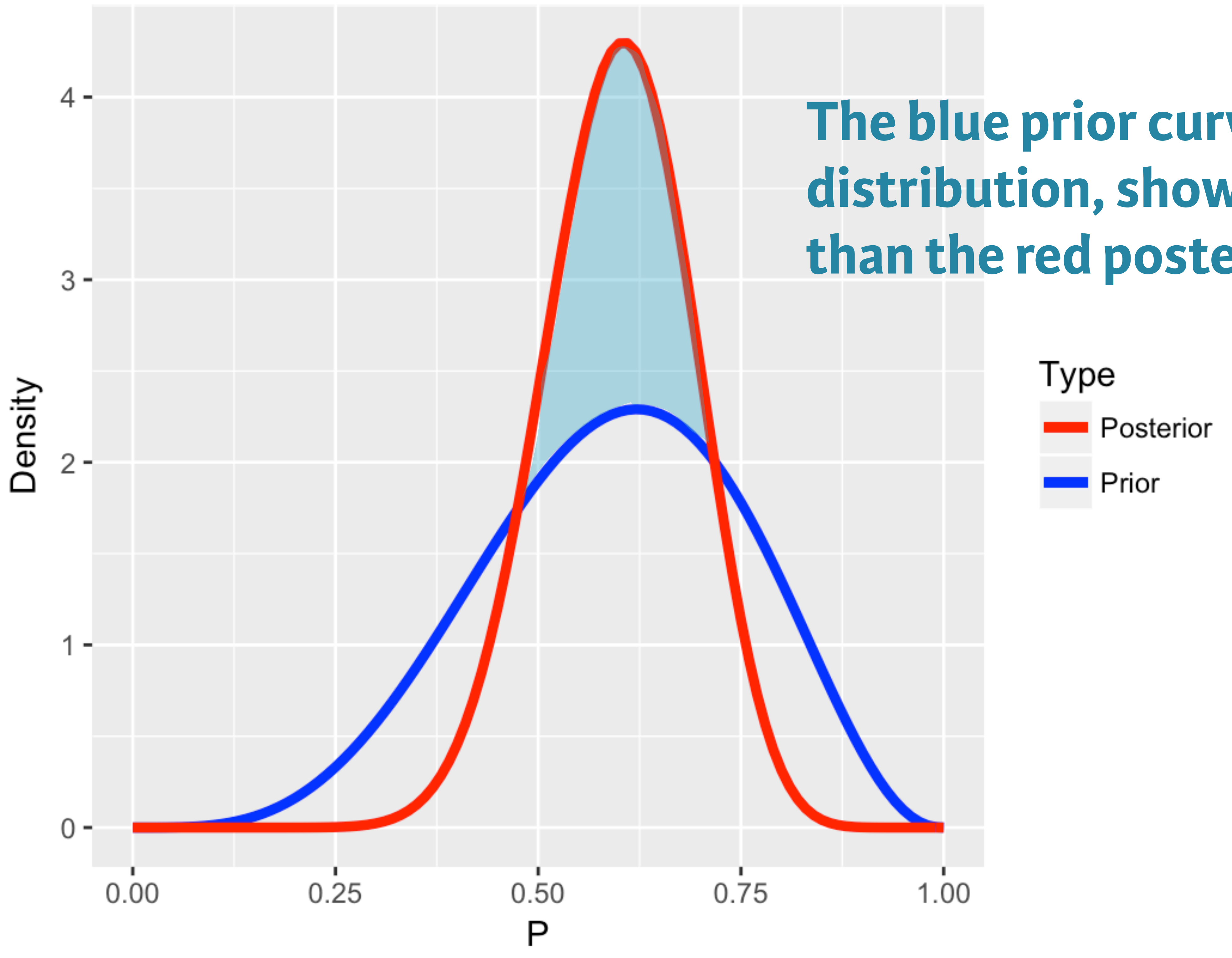
Example: beta prior to beta posterior

	Beta Shape 1	Beta Shape 2
Prior	4.91	3.38
Data	12	8
Posterior	4.91 + 12 = 16.91	3.38 + 8 = 11.38

```
> # Calculate shape parameters for beta posterior:
> prior_par <- c(4.91, 3.38)
> data <- c(12, 8)
> post_par <- prior_par + data
> post_par
[1] 16.91 11.38
```

Compare prior and posterior curves

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



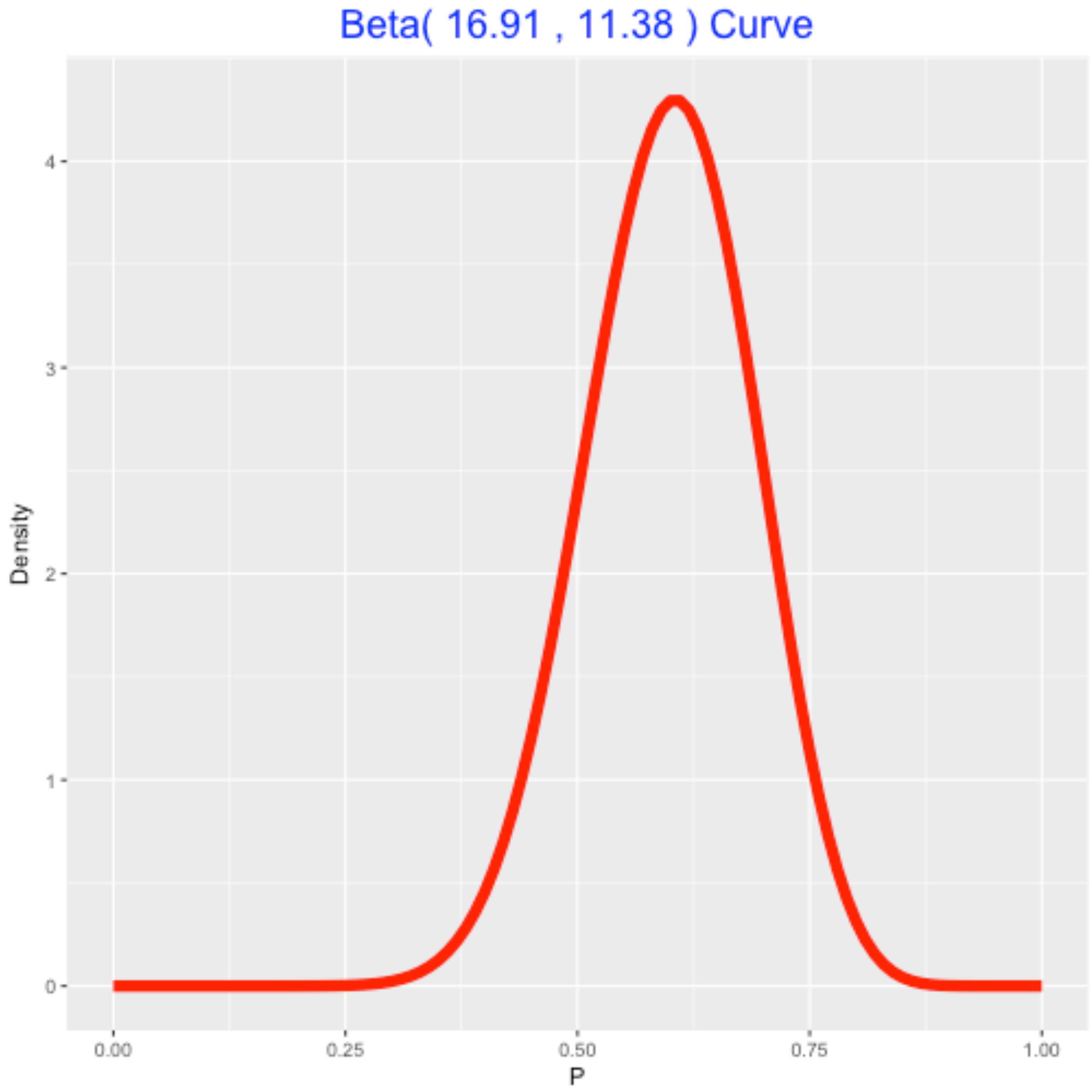
The blue prior curve shows a wider distribution, showing more uncertainty than the red posterior curve

Let's practice!

Bayesian inference

Recall dining survey example

- **Interested in p**: proportion of all students who answer "Friday or Saturday"
- **Current opinion about p:** posterior beta curve with 16.91 and 11.38 as ***a*** and ***b***



Bayesian inference

- Based on summarizing posterior beta curve
- Summary depends on type of inference:
    - **Testing problem**: interested in plausibility of values of p
    - **Interval estimation**: interested in interval likely to contain p

Fellow worker makes claim

- *"At least 75% of the college students prefer to eat on Friday or Saturday."*
- Is this a reasonable claim?
- Hypothesis: $p \geq 0.75$

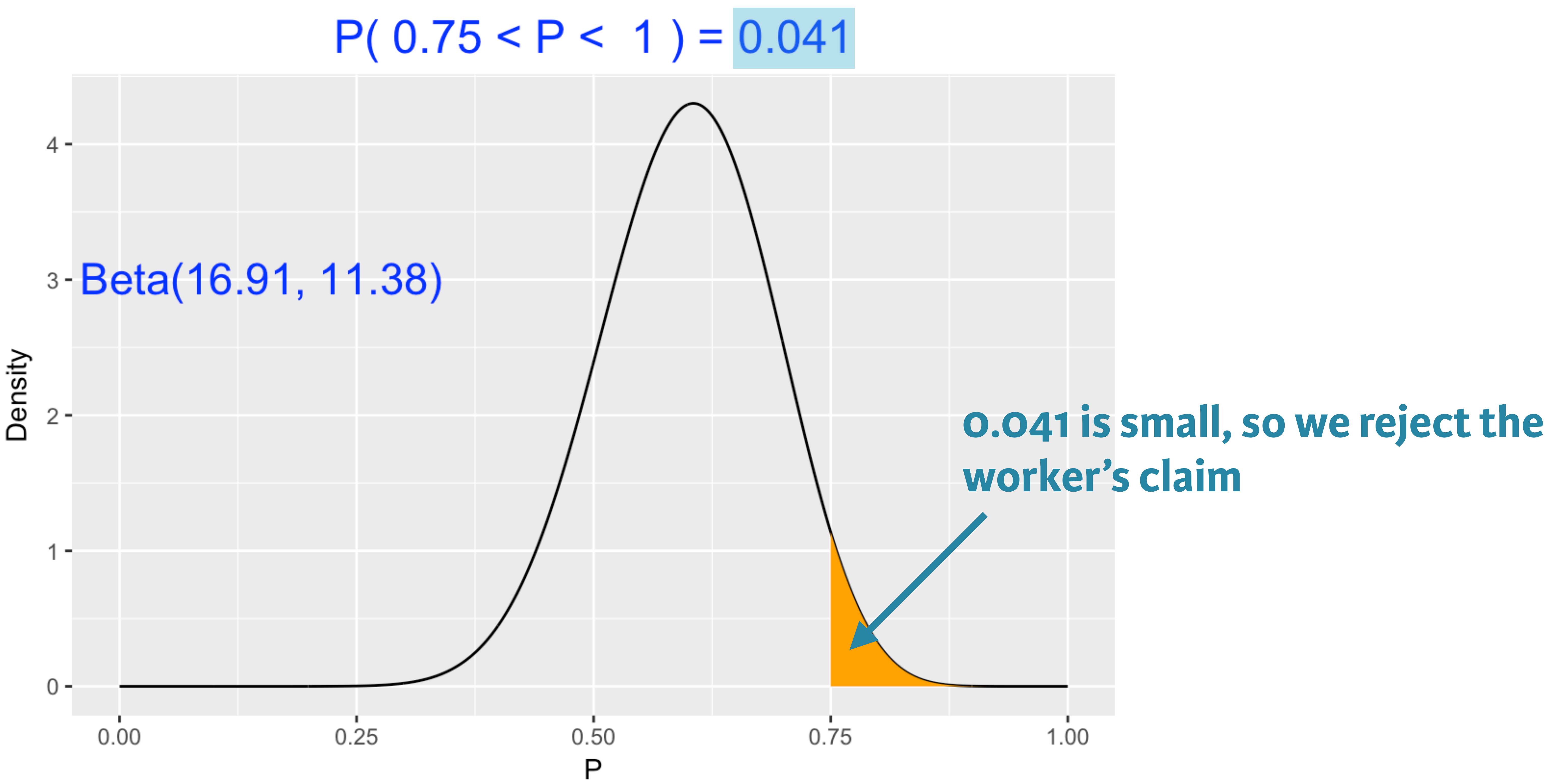
Bayesian approach

- The hypothesis is just an interval of values
- Find the area under posterior where $p \geq 0.75$
- If this probability is small, reject claim

Compute Prob(0.75 < p < 1)

```
> # Graph region where p >= 0.75 on posterior beta curve:
> library(TeachBayes)
> beta_area(0.75, 1, c(16.91, 11.38))
```

Note: 1 is maximum for beta curves

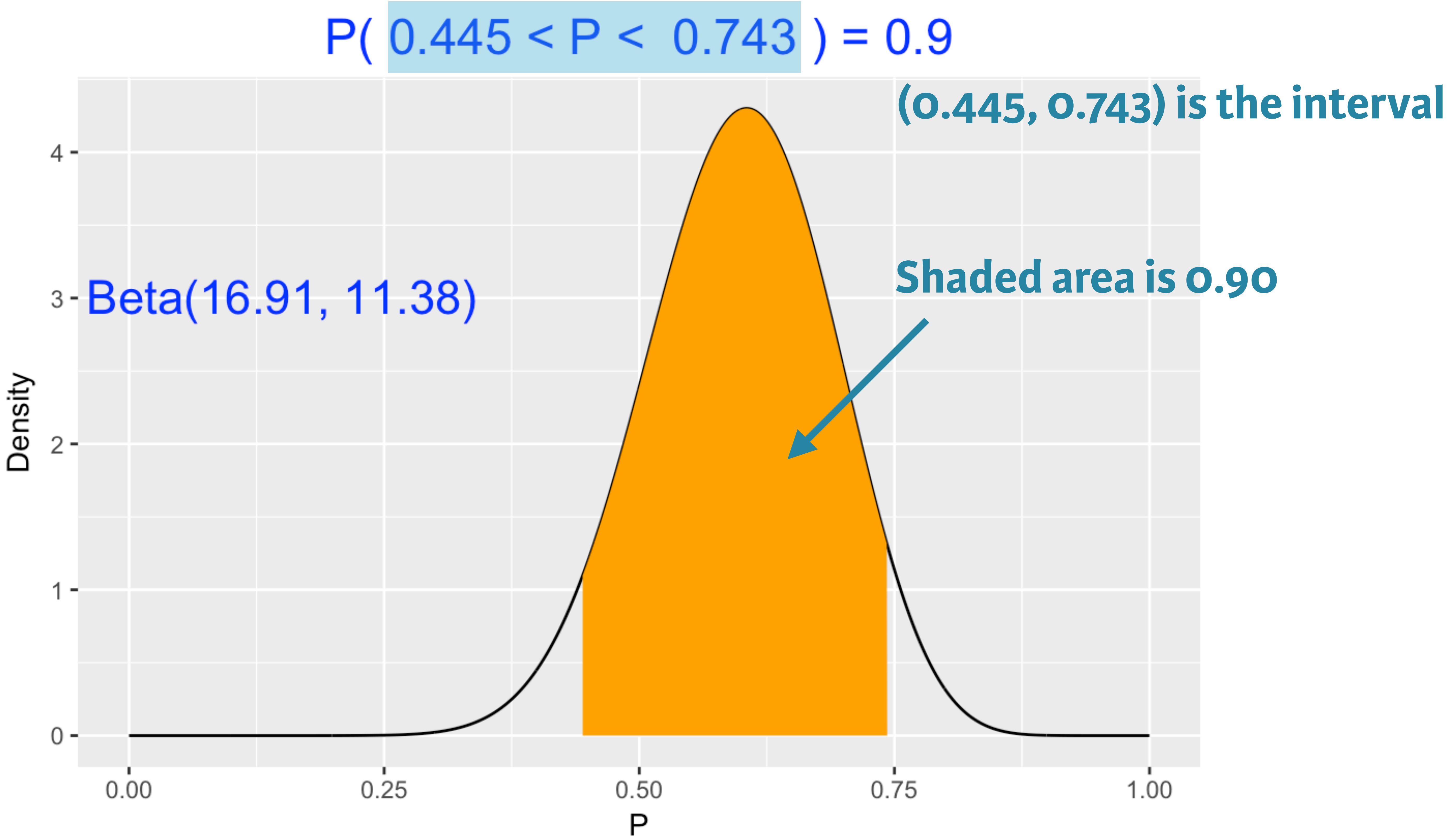


Interval estimate

- A 90% probability interval is an interval that contains 90% of the posterior probability
- Convenient to choose an “equal-tails” interval where 5% of the probability is in each tail

Compute 90% probability interval

```
> # Compute the 90% probability interval:
> library(TeachBayes)
> beta_interval(0.90, c(16.91, 11.38))
```



Interpreting Bayesian interval

- The probability p is in (0.44, 0.74) is exactly 0.90
- Differs from interpretation of classical confidence interval
- One does not know if p is in **one** 90% confidence interval
- “Confidence” is in repeated sampling

Compare with one classical method

- "Add 2 successes and 2 failures" method of Agresti and Coull
- Given y successes and sample size n, 90% interval is:

$$(\hat{p} - 1.645se, \hat{p} + 1.645se)$$

$$\hat{p} = \frac{y+2}{n+4}$$

$$se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n+4}}$$

Classical CI for our example

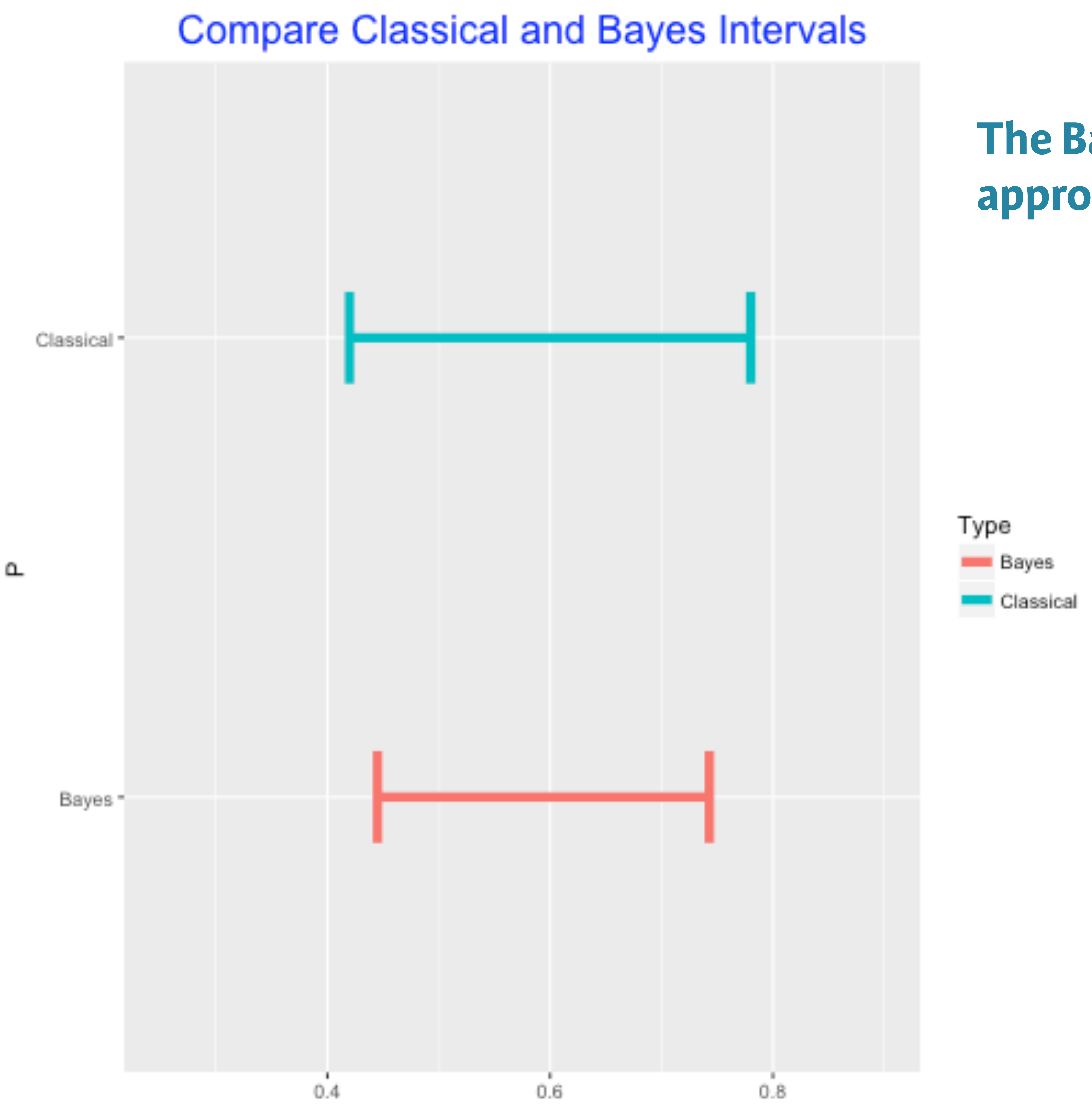
```
> y <- 12; n <- 20
> p_hat <- y/n; (se <- sqrt(p_hat * (1 - p_hat)/n))
[1] 0.1095
> (CI <- c(p_hat - 1.645 * se, p_hat + 1.645 * se))
[1] 0.4198725 0.7801275
```

$$y = 12, n = 20$$

$$\hat{p} = 12/20 = 0.6, se = \sqrt{0.6(0.4)/20} = 0.1095$$

$$(\hat{p} - 1.645se, \hat{p} + 1.645se)$$

Compare two intervals



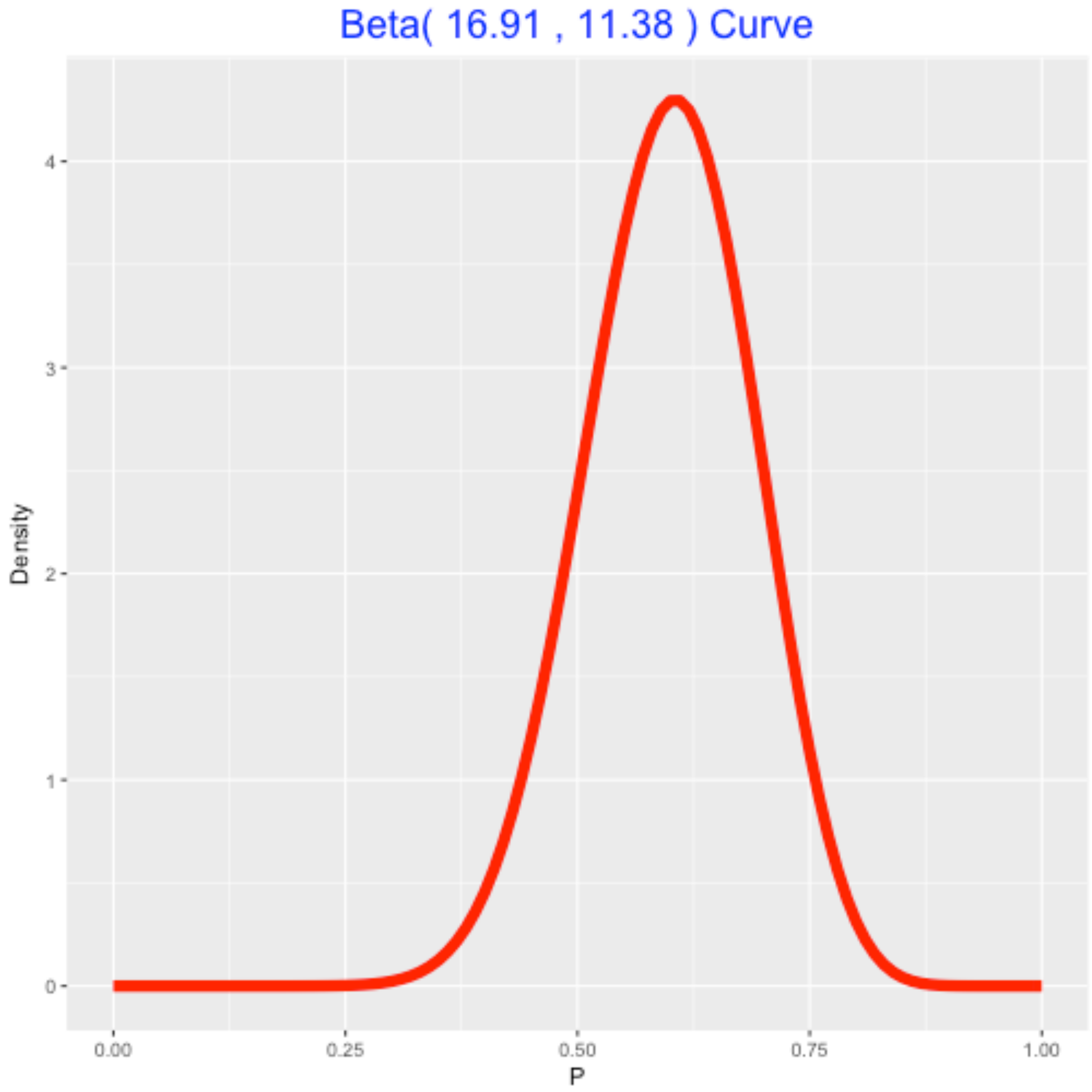
The Bayesian interval is shorter because the Bayesian approach combines data with prior information

Let's practice!

Posterior simulation

Recall dining survey example

- **Interested in p**: proportion of all students who answer "Friday or Saturday"
- **Current beliefs about p:** posterior beta curve with 16.91 and 11.38 as ***a*** and ***b***



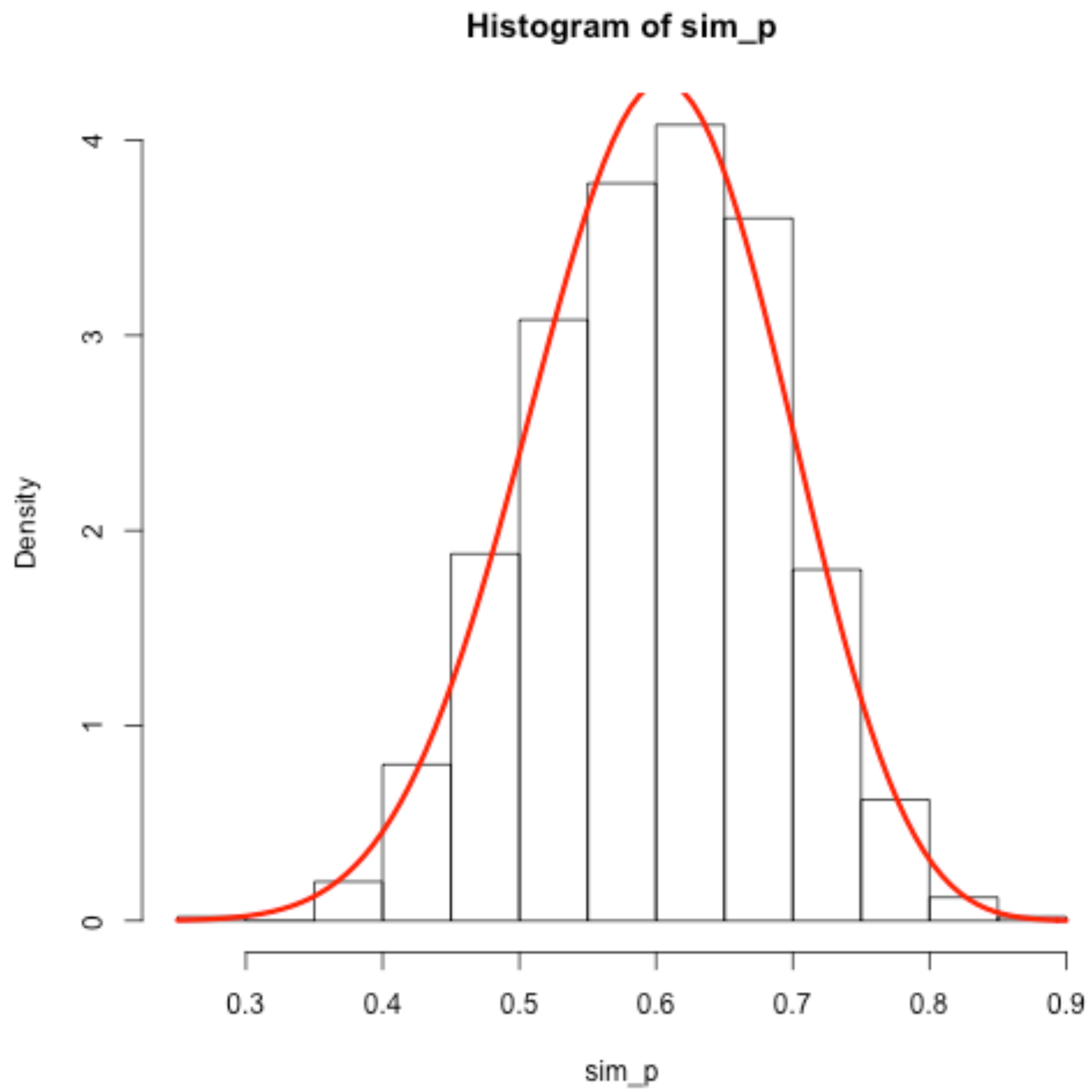
Bayesian inference using simulation

- **So far**: we summarized beta posterior density by computing probabilities and quantiles
- Simulate large number of values from the posterior
- Then summarize the posterior sample to do inference

Simulate using rbeta() function

```
> # Draw 1000 values from posterior: sim_p
> sim_p <- rbeta(1000, c(16.91, 11.38))

> # Check if simulations reflect posterior distribution
> hist(sim_p, freq = FALSE)
> curve(dbeta(x, 16.91, 11.38), add = TRUE, col = "red", lwd = 3)
```

Simulate using rbeta() function



Compute probabilities

```
> # Probability p < 0.5 approximated using simulation
> (prob <- sum(sim_p < 0.5) / 1000)
[1] 0.145

> # Exact answer from Beta(16.91, 11.38) posterior curve
> pbeta(0.5, 16.91, 11.38)
[1] 0.1448413
```

Recall you drew 1000 samples from the posterior curve

Compute quantiles

```
> # Find the sample quantiles of simulated values of p:
> quantile(sim_p, c(0.05, 0.95))
       5%       95%
0.4491962 0.7404108
```

Why simulate?

- We can compute exact posterior summaries using `pbeta()` and `qbeta()` functions
- Exact calculations are difficult for the posteriors of many Bayesian models
- Learn about *functions* of parameters of interest

Example: posterior of log odds ratio

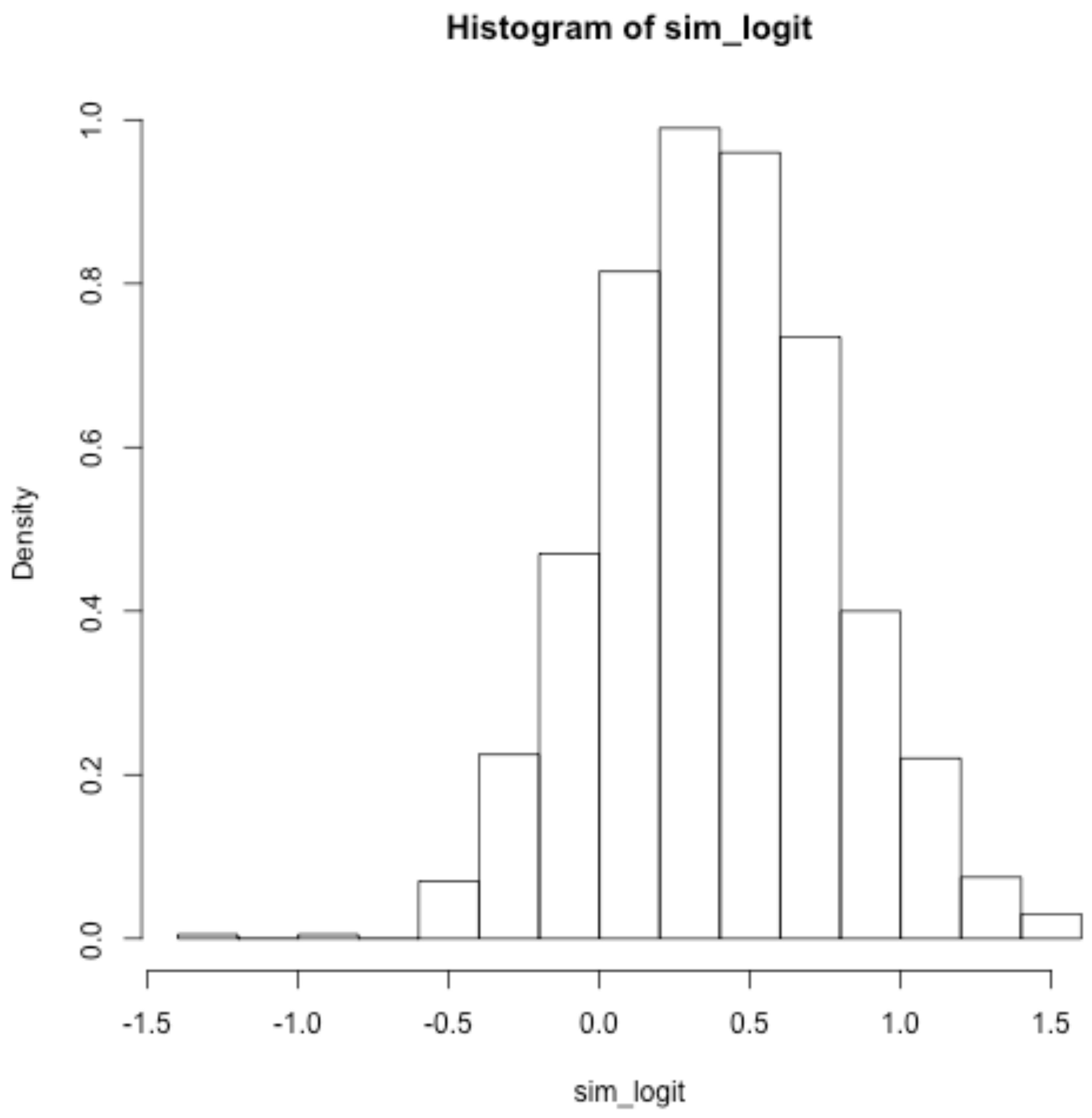
- Interested in 90% probability interval for $\log \frac{p}{1-p}$
- Useful parameter for categorical data analysis
- Easy to use simulation

```
> sim_p <- rbeta(1000, 16.91, 11.38)
> sim_logit <- log(sim_p / (1 - sim_p))
```

`sim_logit` is a sample from posterior distribution of the logit

Posterior of log odds ratio

```
> hist(sim_logit, freq = FALSE)
```



Posterior summaries

```
> # 90% probability interval for log odds ratio
> quantile(sim_logit, c(0.10, 0.90))
       10%         90%
-0.1027612  0.8877401
```

Let's practice!