INFSCI 2595

Fall 2019

Information Sciences Building: Room 403

Lecture 03

Last week, we introduced the Bernoulli distribution

$$p(x|\mu) = \text{Bernoulli}(x|\mu) = \mu^x (1-\mu)^{1-x}$$

- x is a **binary variable**, $x \in \{0, 1\}$
- μ is a probability and so is bounded: $0 \leq \mu \leq 1$

We stepped through the Maximum Likelihood Estimate (MLE) of μ given observations

- N **independent** observations, $\mathbf{x} = \{x_1, x_2, \dots, x_n, \dots, x_N\}$

- We observe $x = 1$ a total of M times.

- The MLE for the probability of the event is:

$$\boxed{\mu_{ML} = \frac{M}{N}}$$

But, let's ask a different question…

- Instead of asking, what's the probability $x = 1$ (the EVENT)…
- Let's ask, what's the probability the event occurs a **specific number of times out of a specific number of trials**?

In terms of our college football example from last week…

- What’s the probability of finding **exactly 1 Pitt fan out of 4 people**?

Did we calculate this probability last week?

- Based on the following independent observations:

Person	Fan	x
1	PSU	0
2	PSU	0
3	Pitt	1
4	PSU	0

Did we calculate this probability last week?

- Based on the following independent observations:

Person	Fan	x	$p(x\|\mu)$
1	PSU	0	$(1-\mu)$
2	PSU	0	$(1-\mu)$
3	Pitt	1	μ
4	PSU	0	$(1-\mu)$

Did we calculate this probability last week?

- Based on the following independent observations:

Person	Fan	$\boldsymbol{x}$	$\boldsymbol{p(x\|\mu)}$
1	PSU	0	$(1-\mu)$
2	PSU	0	$(1-\mu)$
3	Pitt	1	μ
4	PSU	0	$(1-\mu)$

$$\boxed{p(\mathbf{x}|\mu) = (1-\mu)(1-\mu)\mu(1-\mu)}$$

Did we calculate this probability last week?

- Based on the following independent observations:

Person	Fan	x	$p(x \mid \mu)$
4	PSU	0	$(1-\mu)$

Wait…is this the only way to observe 1 Pitt fan out of 4 people?

$$\boxed{p(\mathbf{x} \mid \mu) = (1-\mu)(1-\mu)\mu(1-\mu)}$$

No! Multiple **potential** sequences of 4 people consist of exactly 1 Pitt fan.

Person 1	Person 2	Person 3	Person 4
Pitt	PSU	PSU	PSU
PSU	Pitt	PSU	PSU
PSU	PSU	Pitt	PSU
PSU	PSU	PSU	Pitt

Rewrite each of the **potential** sequences in terms of the encoded variable x

x_1	x_2	x_3	x_4
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Calculate the probability of each **potential** sequence assuming independent observations

$p(x_1\|\mu)$	$p(x_2\|\mu)$	$p(x_3\|\mu)$	$p(x_4\|\mu)$
μ	$(1-\mu)$	$(1-\mu)$	$(1-\mu)$
$(1-\mu)$	μ	$(1-\mu)$	$(1-\mu)$
$(1-\mu)$	$(1-\mu)$	μ	$(1-\mu)$
$(1-\mu)$	$(1-\mu)$	$(1-\mu)$	μ

Each of the **potential** sequences have the same probability!

$p(\mathbf{x}\|\mu)$
$\mu \cdot (1-\mu)^3$
$\mu \cdot (1-\mu)^3$
$\mu \cdot (1-\mu)^3$
$\mu \cdot (1-\mu)^3$

The probability of observing exactly 1 Pitt fan out of 4 people:

- Sum together the probabilities of each **potential** sequence:

$$\boxed{4 \cdot \mu \cdot (1 - \mu)^3}$$

- Next, what's the probability of finding **exactly 2 Pitt fans out of 4 people**?

List all **potential** sequences with 2 Pitt fans

x_1	x_2	x_3	x_4
1	1	0	0
0	1	1	0
0	0	1	1
1	0	1	0
0	1	0	1
1	0	0	1

Calculate the probability of each **potential** sequence assuming independent observations

x_1	x_2	x_3	x_4
μ	μ	$(1-\mu)$	$(1-\mu)$
$(1-\mu)$	μ	μ	$(1-\mu)$
$(1-\mu)$	$(1-\mu)$	μ	μ
μ	$(1-\mu)$	μ	$(1-\mu)$
$(1-\mu)$	μ	$(1-\mu)$	μ
μ	$(1-\mu)$	$(1-\mu)$	μ

Calculate the probability of each **potential** sequence assuming independent observations

<table>
<tr><th>x_1</th><th>x_2</th><th>x_3</th><th>x_4</th></tr>
<tr><td colspan="4">$\mu^2(1-\mu)^2$</td></tr>
<tr><td colspan="4">$\mu^2(1-\mu)^2$</td></tr>
<tr><td colspan="4">$\mu^2(1-\mu)^2$</td></tr>
<tr><td colspan="4">$\mu^2(1-\mu)^2$</td></tr>
<tr><td colspan="4">$\mu^2(1-\mu)^2$</td></tr>
<tr><td colspan="4">$\mu^2(1-\mu)^2$</td></tr>
</table>

The probability of observing exactly 2 Pitt fans out of 4 people:

- Sum together the probabilities of each **potential** sequence:

$$\boxed{6 \cdot \mu^2 \cdot (1 - \mu)^2}$$

How many potential sequences exist?

- Assume 4 people (trials).

- A person can be either a Pitt fan or a PSU fan (binary outcome).

$$2^4 = 16$$

Sequence ID	x_1	x_2	x_3	x_4
1	0	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1
6	1	1	0	0
7	0	1	1	0
8	0	0	1	1
9	1	0	1	0
10	0	1	0	1
11	1	0	0	1
12	1	1	1	0
13	0	1	1	1
14	1	1	0	1
15	1	0	1	1
16	1	1	1	1

Sequence ID	x_1	x_2	x_3	x_4	Times $x = 1$
1	0	0	0	0	0
2	1	0	0	0	1
3	0	1	0	0	
4	0	0	1	0	
5	0	0	0	1	
6	1	1	0	0	2
7	0	1	1	0	
8	0	0	1	1	
9	1	0	1	0	
10	0	1	0	1	
11	1	0	0	1	
12	1	1	1	0	3
13	0	1	1	1	
14	1	1	0	1	
15	1	0	1	1	
16	1	1	1	1	4

Calculate the probability of observing $x = 1$ exactly 0, 1, 2, 3, and 4 times.

Times $x = 1$	$p(\mathbf{x}\|\mu)$
0	$1 \cdot \mu^0 \cdot (1-\mu)^4$
1	$4 \cdot \mu^1 \cdot (1-\mu)^3$
2	$6 \cdot \mu^2 \cdot (1-\mu)^2$
3	$4 \cdot \mu^3 \cdot (1-\mu)^1$
4	$1 \cdot \mu^4 \cdot (1-\mu)^0$

WHAT PATTERNS DO YOU SEE??

Times $x = 1$	$p(\mathbf{x}\|\mu)$
0	$1 \cdot \mu^0 \cdot (1-\mu)^4$
1	$4 \cdot \mu^1 \cdot (1-\mu)^3$
2	$6 \cdot \mu^2 \cdot (1-\mu)^2$
3	$4 \cdot \mu^3 \cdot (1-\mu)^1$
4	$1 \cdot \mu^4 \cdot (1-\mu)^0$

WHAT PATTERNS DO YOU SEE??

The exponent on μ equals the number of times $x = 1$.

The number of times $x = 1$, corresponds to the number of times we observed the EVENT.

Define the number of EVENTS to be m.

Times $x = 1$	$p(\mathbf{x}\|\mu)$
0	$1 \cdot \mu^{\boxed{0}} \cdot (1-\mu)^4$
1	$4 \cdot \mu^{\boxed{1}} \cdot (1-\mu)^3$
2	$6 \cdot \mu^{\boxed{2}} \cdot (1-\mu)^2$
3	$4 \cdot \mu^{\boxed{3}} \cdot (1-\mu)^1$
4	$1 \cdot \mu^{\boxed{4}} \cdot (1-\mu)^0$

WHAT PATTERNS DO YOU SEE??

The exponent on $(1-\mu)$ equals the number of TRIALS minus the number of EVENTS.

Corresponds to the number of times we did not observe the EVENT.

Define as $N-m$.

m	$p(\mathbf{x}\|\mu)$
0	$1 \cdot \mu^m \cdot (1-\mu)^{\boxed{4}}$
1	$4 \cdot \mu^m \cdot (1-\mu)^{\boxed{3}}$
2	$6 \cdot \mu^m \cdot (1-\mu)^{\boxed{2}}$
3	$4 \cdot \mu^m \cdot (1-\mu)^{\boxed{1}}$
4	$1 \cdot \mu^m \cdot (1-\mu)^{\boxed{0}}$

WHAT PATTERNS DO YOU SEE??

What about the coefficient out front?

Rewrite using:

$$\binom{4}{0} = 1, \binom{4}{1} = 4, \binom{4}{2} = 6$$
$$\binom{4}{3} = 4, \binom{4}{4} = 1$$

m	$p(\mathbf{x}\|\mu)$
0	$\boxed{1} \cdot \mu^m \cdot (1-\mu)^{N-m}$
1	$\boxed{4} \cdot \mu^m \cdot (1-\mu)^{N-m}$
2	$\boxed{6} \cdot \mu^m \cdot (1-\mu)^{N-m}$
3	$\boxed{4} \cdot \mu^m \cdot (1-\mu)^{N-m}$
4	$\boxed{1} \cdot \mu^m \cdot (1-\mu)^{N-m}$

WHAT PATTERNS DO YOU SEE??

What about the coefficient out front?

Rewrite using:

$$\binom{4}{0} = 1, \binom{4}{1} = 4, \binom{4}{2} = 6$$
$$\binom{4}{3} = 4, \binom{4}{4} = 1$$

Which can be generalized using:

$$\binom{N}{m}$$

$\boldsymbol{m}$	$\boldsymbol{p(\mathbf{x}\|\mu)}$
0	$\boxed{1} \cdot \mu^m \cdot (1-\mu)^{N-m}$
1	$\boxed{4} \cdot \mu^m \cdot (1-\mu)^{N-m}$
2	$\boxed{6} \cdot \mu^m \cdot (1-\mu)^{N-m}$
3	$\boxed{4} \cdot \mu^m \cdot (1-\mu)^{N-m}$
4	$\boxed{1} \cdot \mu^m \cdot (1-\mu)^{N-m}$

The probability distribution of m events out of N trials, given event probability μ:

$$\boxed{p(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}}$$

$$m \in \{0, \dots, N\}$$

Known as the Binomial distribution!

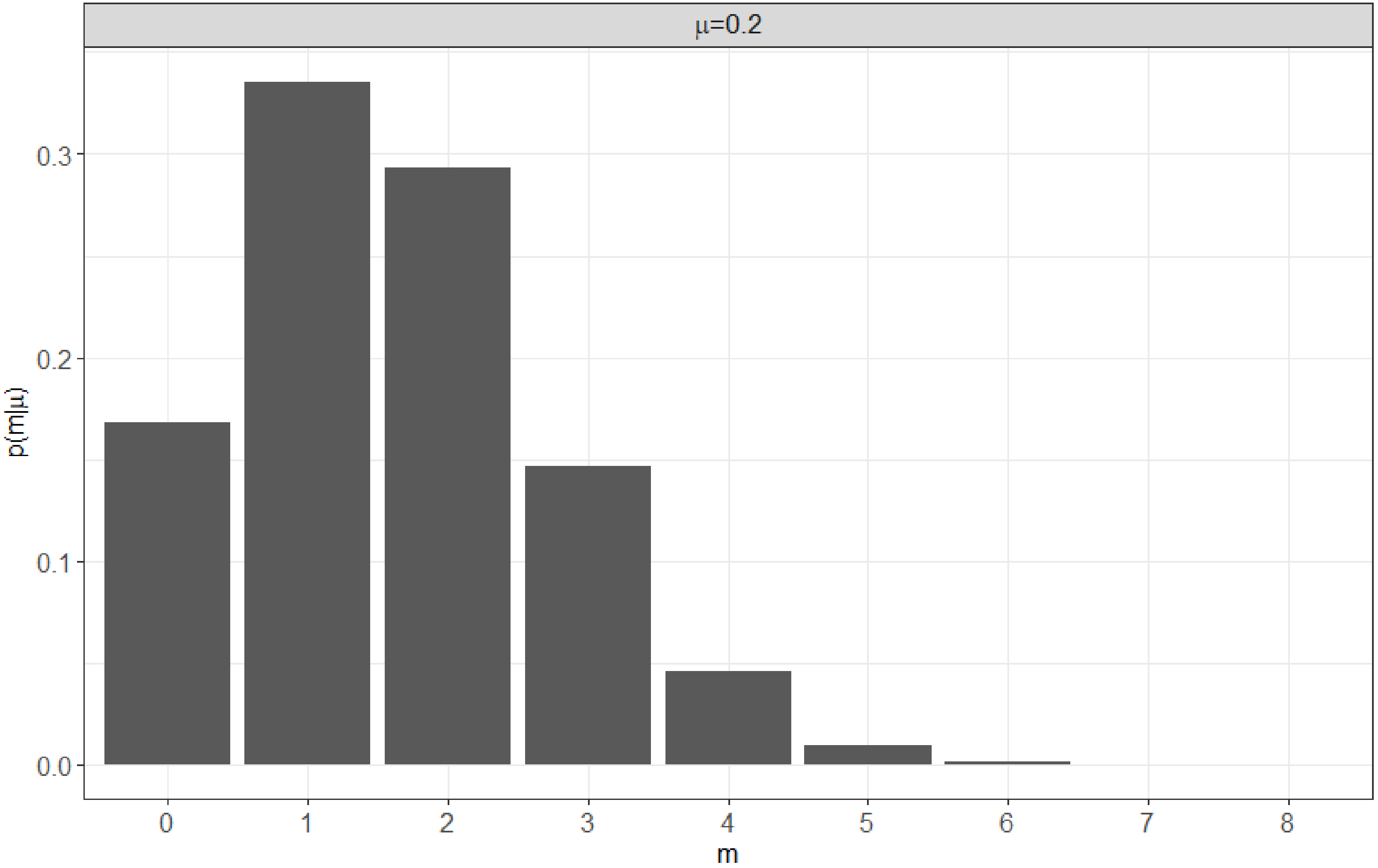
We derived the Binomial distribution starting from Bernoulli observations

- The Binomial distribution is a sequence of INDEPENDENT Bernoulli trials.

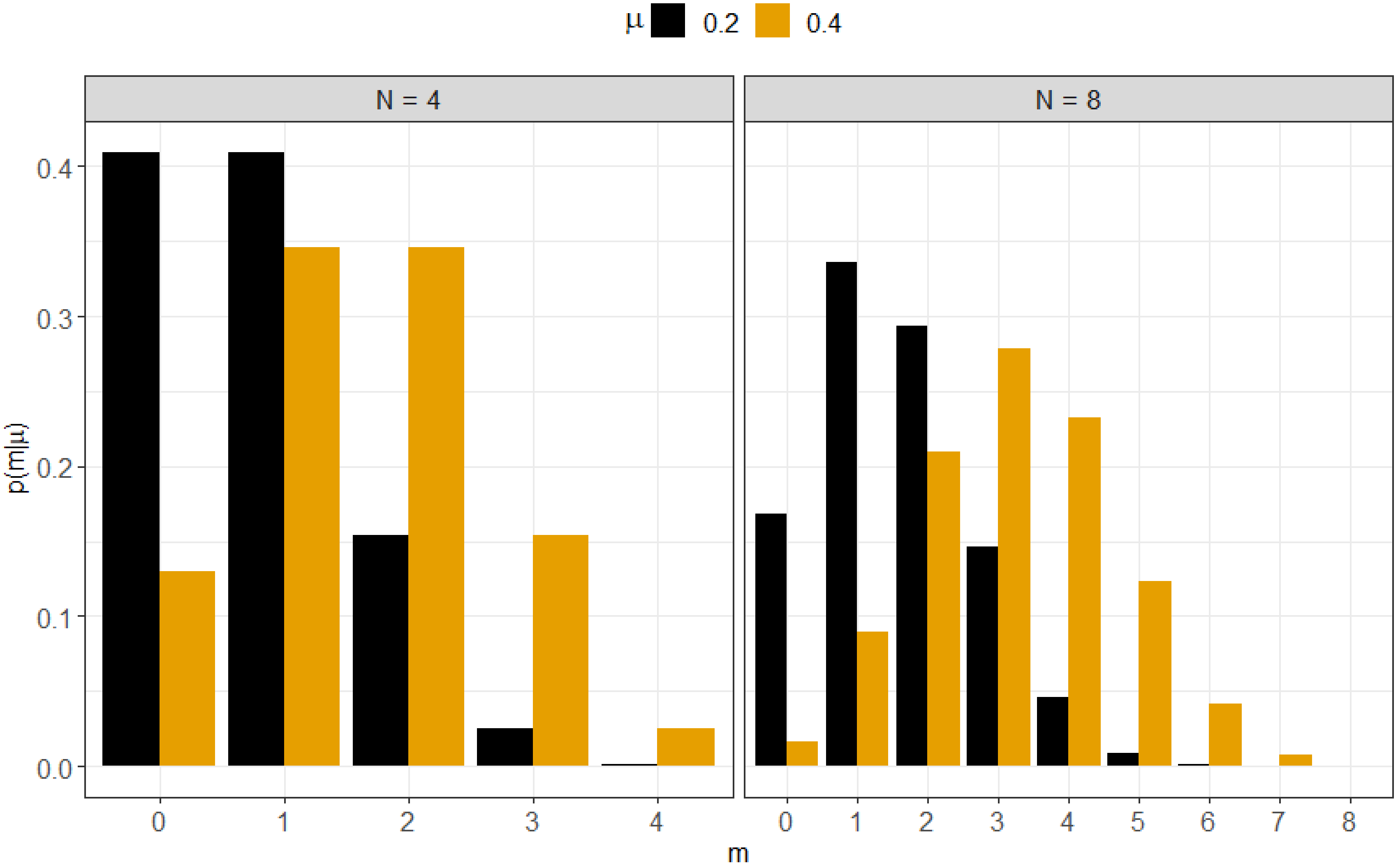
- We recover the Bernoulli distribution with $N = 1$. Thus, $m = \{0,1\}$.

- The Bernoulli is therefore a special case of the Binomial distribution.

Binomial distribution for $N = 8$ and $\mu = 0.2$



Binomial distribution two different N's and two different μ's

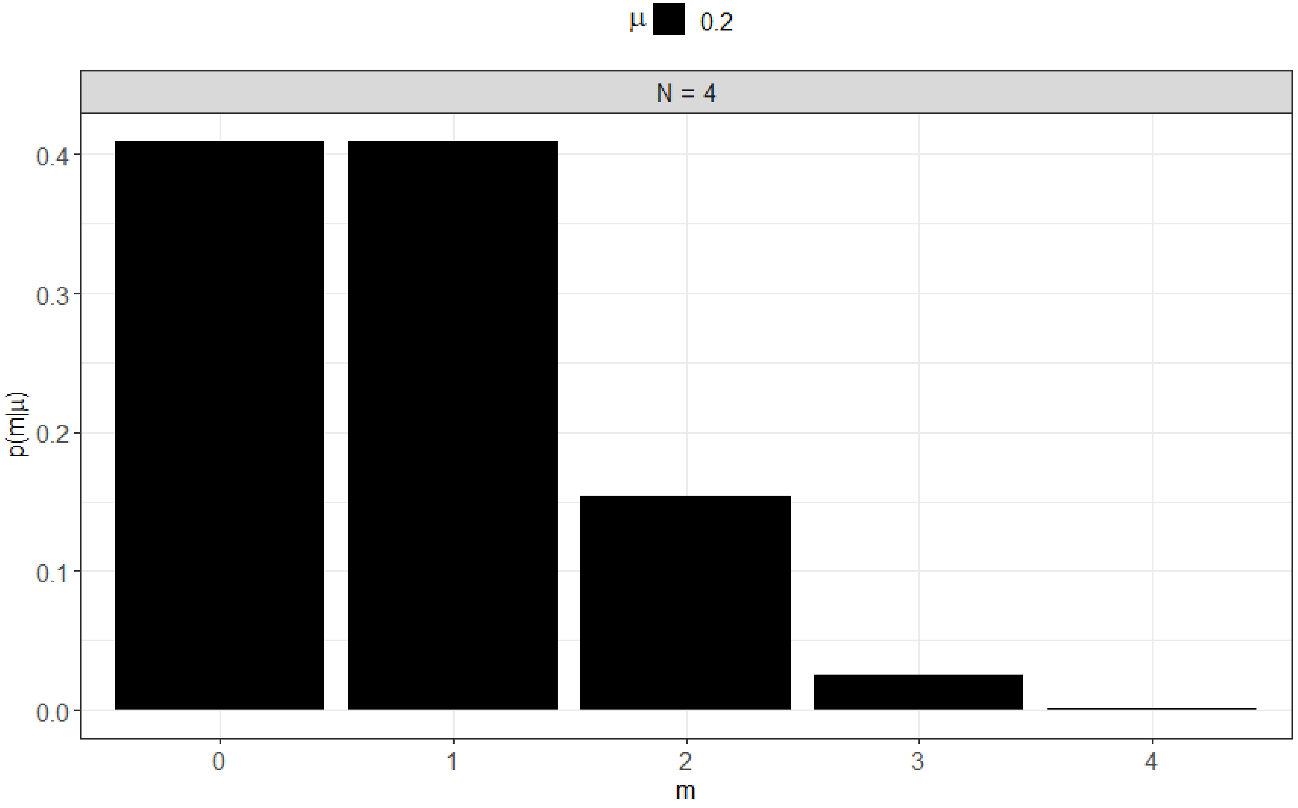


Back to our college football example, remember that the game is located at State College…

- If we ask 4 people, and **assume** the TRUE probability of a Pitt fan is $\mu = 0.2$…

- The probability of finding 0 Pitt fans is ≈40%!

- The probability of finding 2 Pitt fans is small but not negligible at ≈15%.

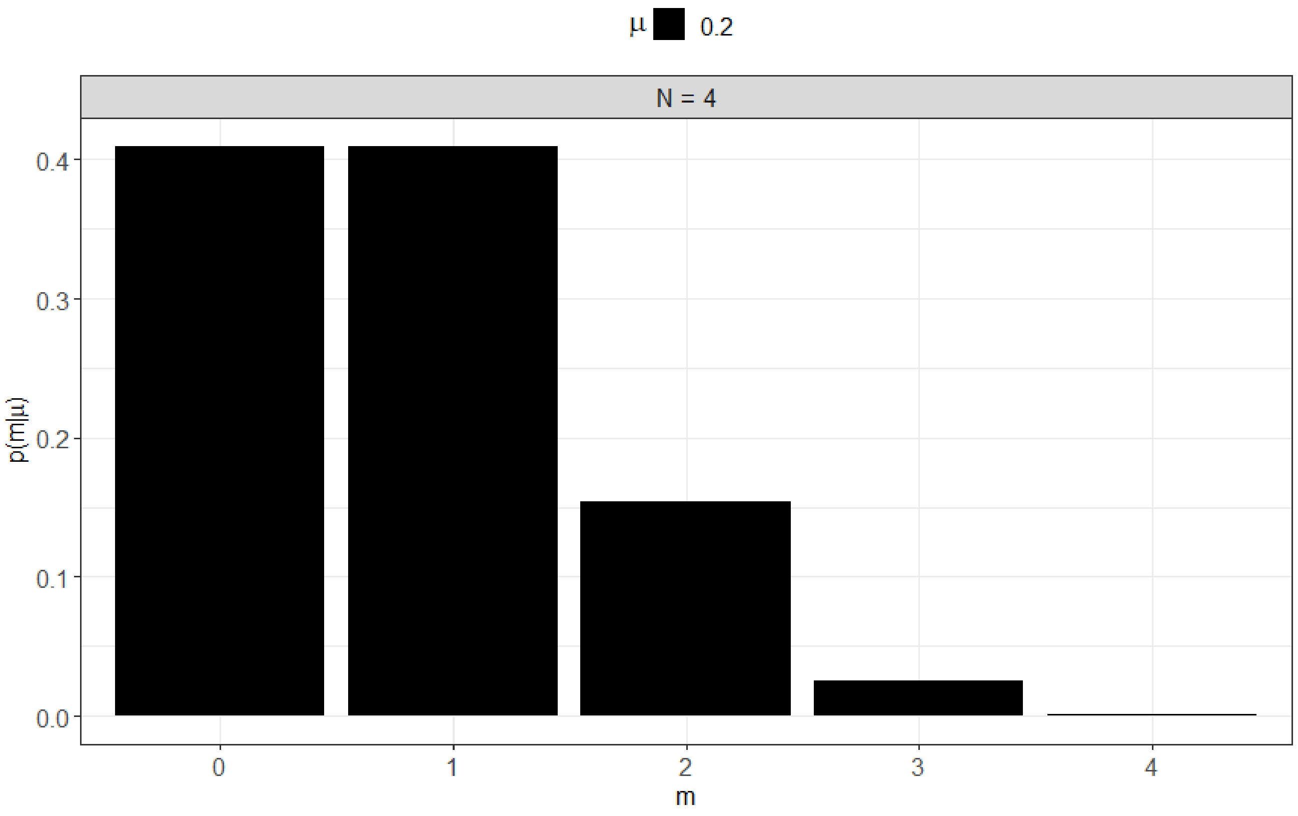


Back to our college football example, remember that the game is located at State College…

- If 0 out of 4 people are Pitt fans, our MLE for the probability would be $\mu_{ML} = 0$.

- If 2 out of 4 people are Pitt fans, our MLE for the probability would be $\mu_{ML} = 0.5$.

- Both estimates are not unrepresentative of $\mu_{TRUE} = 0.2$!



Our MLE is unreliable in this <u>small</u> data situation!

- How can we overcome this limitation?

Our MLE is unreliable in this small data situation!

- How can we overcome this limitation?

- Ask more people (collect more data)…but what if we cannot do that?

Our MLE is unreliable in this small data situation!

- How can we overcome this limitation?
- Ask more people (collect more data)…but what if we cannot do that?
- Could we make use of additional information?

Remember, the game is a **home** game for Penn State

- Thus, it is safe to anticipate more PSU fans than Pitt fans to be present at the game.
- How can we make use of this information in our analysis?

Remember, the game is a **home** game for Penn State

- Thus, it is safe to anticipate more PSU fans than Pitt fans to be present at the game.

- How can we make use of this information in our analysis?

Bayesian statistics!

Bayesian formulation for estimating μ

- We want to update our prior belief about μ based on observations.

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

Bayesian formulation for estimating μ

- We want to update our prior belief about μ based on observations.

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

Based on the binomial distribution as the likelihood

$$p(\mu|m, N) \propto \text{Binomial}(m|N, \mu)p(\mu)$$

Bayesian formulation for estimating μ

- We want to update our prior belief about μ based on observations.

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

Based on the binomial distribution as the likelihood

$$p(\mu|m, N) \propto \text{Binomial}(m|N, \mu)p(\mu)$$

Or, based on independent Bernoulli trials as the likelihood

$$p(\mu|\mathbf{x}) \propto \prod_{n=1}^{N} \{\text{Bernoulli}(x_n|\mu)\}\, p(\mu)$$

Bayesian formulation for estimating μ

- We want to update our prior belief about μ based on observations.

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

$$\boxed{p(\mu|m, N) \propto \text{Binomial}(m|N, \mu)p(\mu)}$$

Use this formulation now.

$$p(\mu|\mathbf{x}) \propto \prod_{n=1}^{N} \{\text{Bernoulli}(x_n|\mu)\}\, p(\mu)$$

Bayesian formulation for estimating μ

- We want to update our prior belief about μ based on observations.

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

$$p(\mu|m, N) \propto \text{Binomial}(m|N, \mu) \boxed{p(\mu)}$$

We know how to write out the likelihood…but what about the prior, $p(\mu)$?

How can we specify a prior belief about μ?

We will use a <u>BETA</u> distribution to encode our PRIOR belief on the probability, μ

Beta distribution

- The beta distribution is a probability density function (pdf) for continuous variables **BOUNDED** between 0 and 1.

Beta distribution

- The beta distribution is a probability density function (pdf) for continuous variables **BOUNDED** between 0 and 1.
- It is a flexible distribution capable of a wide variety of shapes.

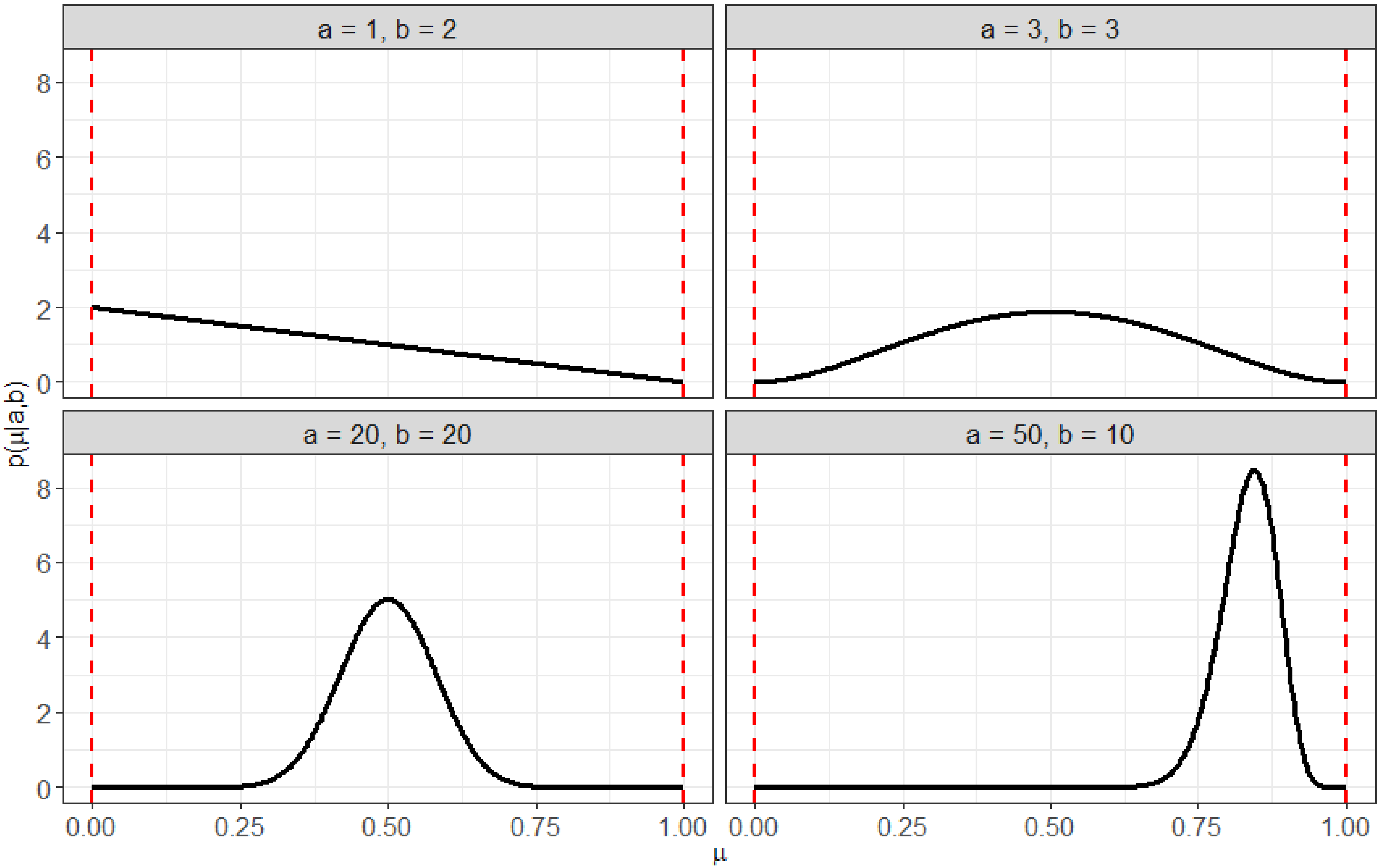
Beta distribution

- The beta distribution is a probability density function (pdf) for continuous variables **BOUNDED** between 0 and 1.

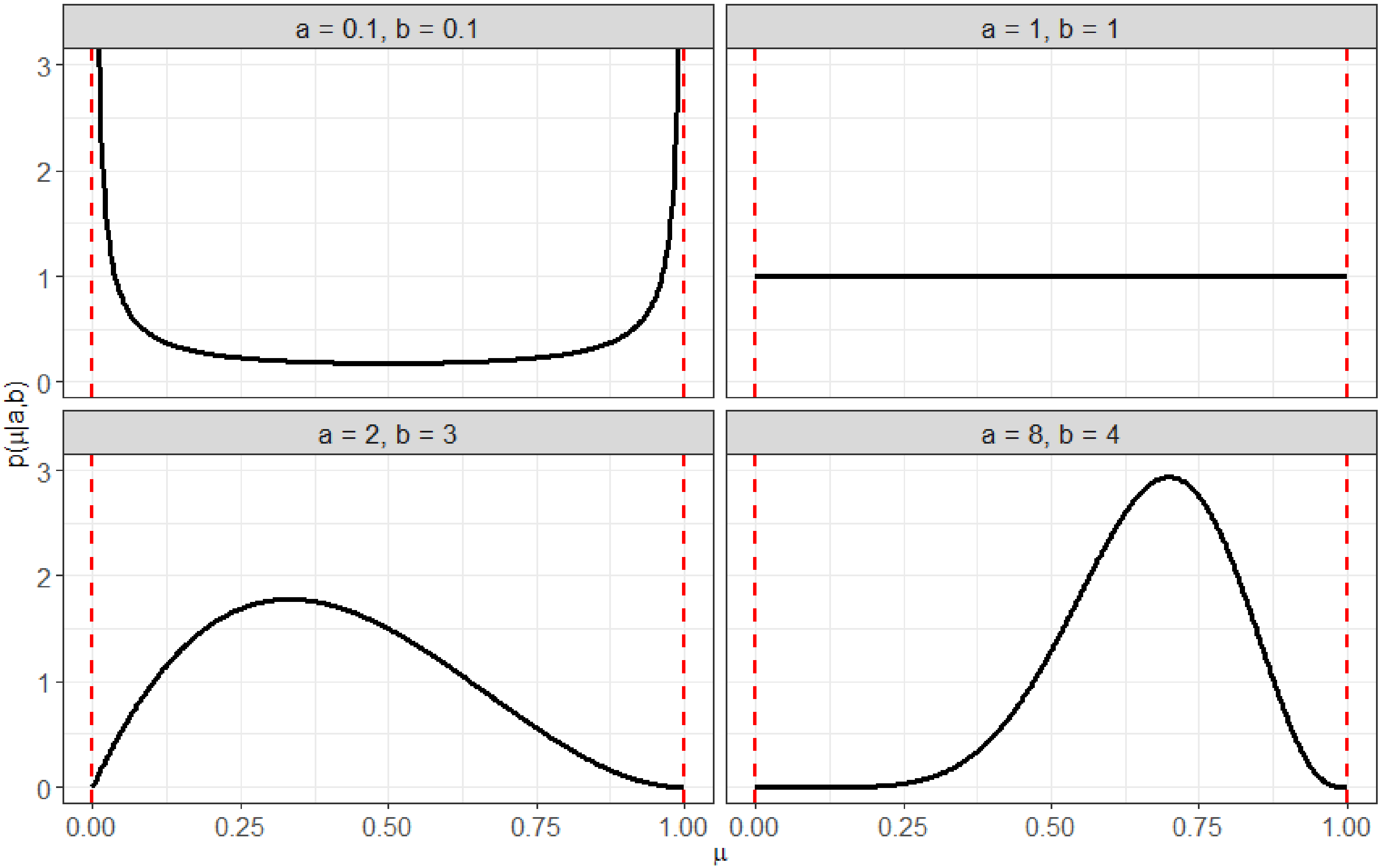
- It is a flexible distribution capable of a wide variety of shapes.

- The shape is controlled by the hyperparameters α and β.
  - The Bishop book denotes these two parameters as a and b.

Example shapes of the beta distribution



Example shapes of the beta distribution



The beta pdf…

$$p(\mu|a,b) = \mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

The beta pdf…looks rather familiar…

$$p(\mu|a,b) = \mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

- Focus on the terms involving μ:

$$\mathrm{Beta}(\mu|a,b) \propto \mu^{a-1}(1-\mu)^{b-1}$$

The beta distribution has the same functional form as the Binomial distribution!

$$\mathrm{Beta}(\mu|a,b) \propto \mu^{a-1}(1-\mu)^{b-1}$$

$$\Updownarrow$$

$$\mathrm{Binomial}(m|\mu,N) \propto \mu^{m}(1-\mu)^{N-m}$$

- The beta distribution is the <u>**conjugate prior**</u> of the binomial likelihood.

The beta distribution has the same functional form as the Binomial distribution!

$$\mathrm{Beta}(\mu|a,b) \propto \mu^{a-1}(1-\mu)^{b-1}$$

$$\Updownarrow$$

$$\mathrm{Binomial}(m|\mu,N) \propto \mu^{m}(1-\mu)^{N-m}$$

- A conjugate prior is useful because the resulting posterior distribution will have the same functional form as the prior.

The beta distribution has the same functional form as the Binomial distribution!

$$\mathrm{Beta}(\mu|a,b) \propto \mu^{a-1}(1-\mu)^{b-1}$$

$$\Updownarrow$$

$$\mathrm{Binomial}(m|\mu,N) \propto \mu^{m}(1-\mu)^{N-m}$$

- The posterior will therefore also be a **<u>beta distribution</u>**!

Posterior distribution on μ

$$p(\mu|m,N) = \text{Beta}\big(\mu|a+m, b+(N-m)\big)$$

Posterior distribution on μ

$$p(\mu|m,N) = \mathrm{Beta}\Big(\mu|\underbrace{a+m}_{a_{new}}, \underbrace{b+(N-m)}_{b_{new}}\Big)$$

$$\Downarrow$$

$$p(\mu|m,N) = \mathrm{Beta}(\mu|a_{new}, b_{new})$$

Beta distribution hyperparameter interpretations

- a is added to the number of Pitt fans, m, or more generally the number of observed EVENTS.

- b is added to the number of PSU fans, $N - m$, or more generally the number of times we did NOT observe the EVENT.

Beta distribution hyperparameter interpretations

- a is added to the number of Pitt fans, m, or more generally the number of observed EVENTS.
  - a is therefore the *a priori* number of EVENTS!!

- b is added to the number of PSU fans, $N - m$, or more generally the number of times we did NOT observe the EVENT.
  - b is therefore the *a priori* number of NON-EVENTS!!

We could have reached the same interpretations by considering the mean…

- The expected value (mean) of the Beta distribution is:

$$\mathbb{E}[\mu|a,b] = \frac{a}{a+b}$$

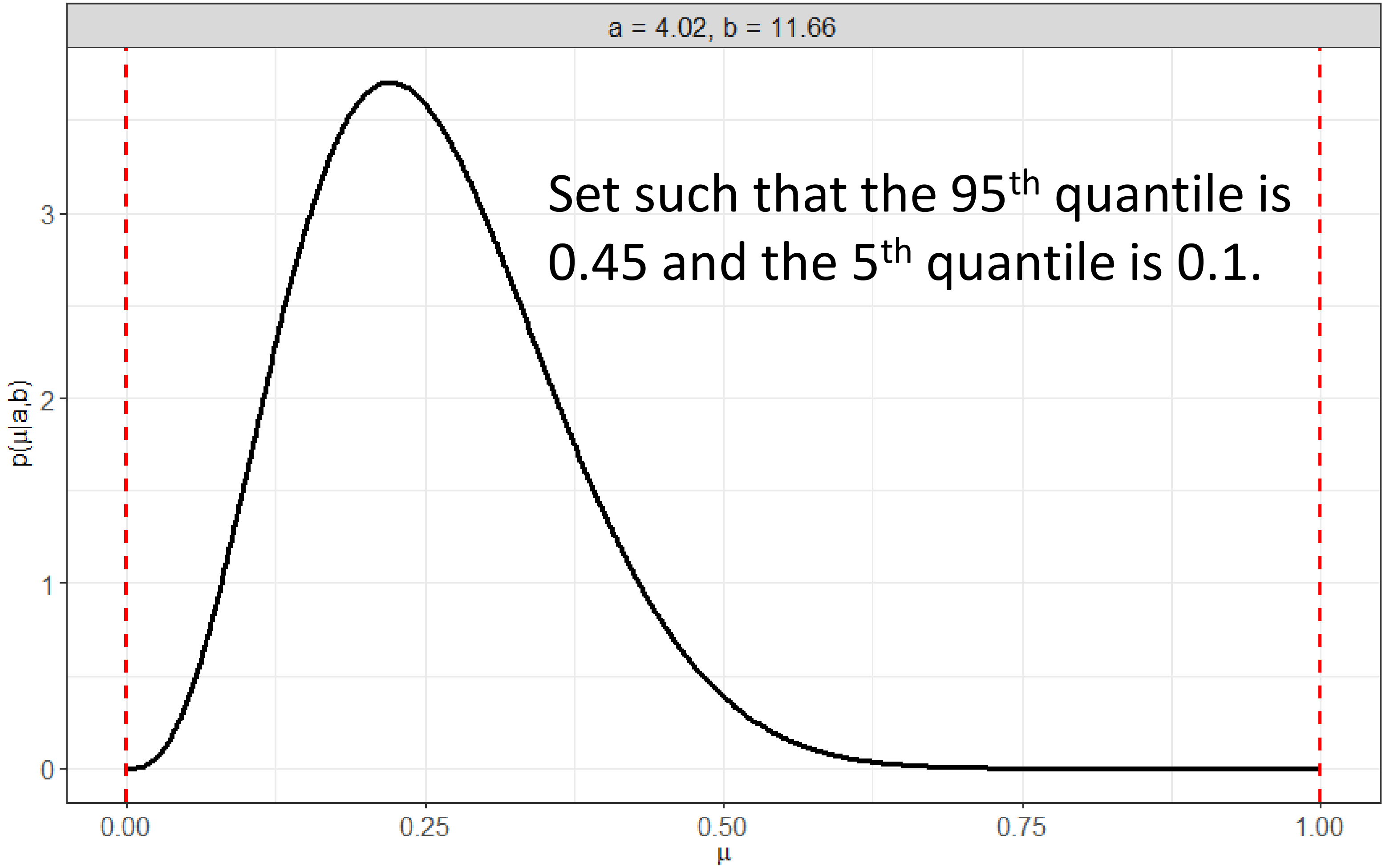
We could have reached the same interpretations by considering the mean…

- The expected value (mean) of the Beta distribution is:

$$\mathbb{E}[\mu|a,b] = \frac{a}{a+b} \Rightarrow \frac{\text{Number of events!}}{\text{Number of trials!}}$$

b is therefore the number of NON-EVENTS, or times $x = 0$!

Set our prior such that we feel the probability of finding a Pitt fan is greater than 0 but less than 0.5



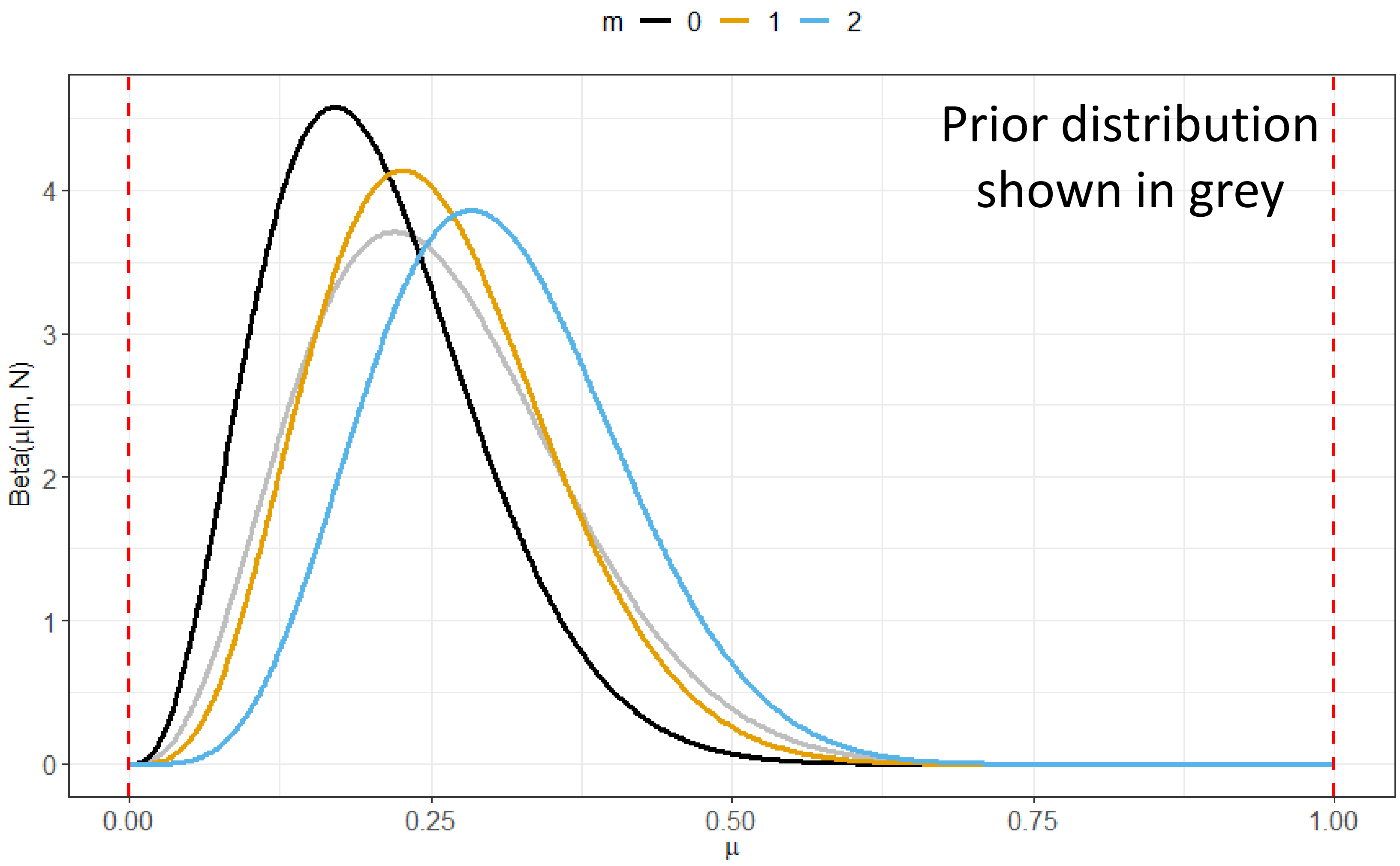
We will update our belief about μ under three different circumstances

- As we saw, the posterior distribution on μ given the observations is a Beta distribution.

- Let's compare the resulting Beta distributions based on observing $m = 0, 1, \text{and } 2.$

- Thus, what's our **updated belief** if we found 0 Pitt fans, vs 1 Pitt fan, vs 2 Pitt fans.

μ posterior distribution given m and $N = 4$



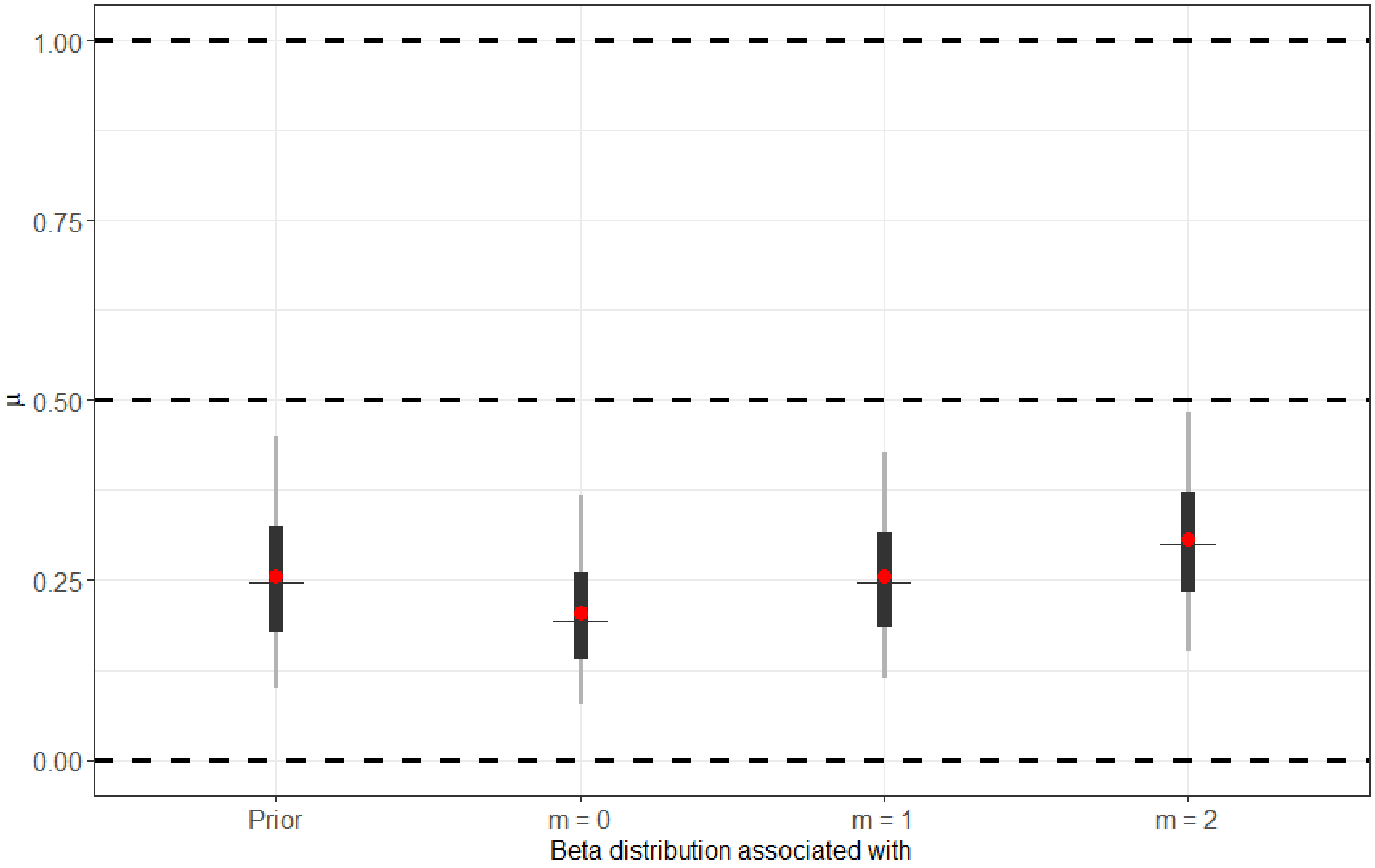
Summarize the Beta distributions

- Calculate summary statistics for each Beta distribution.

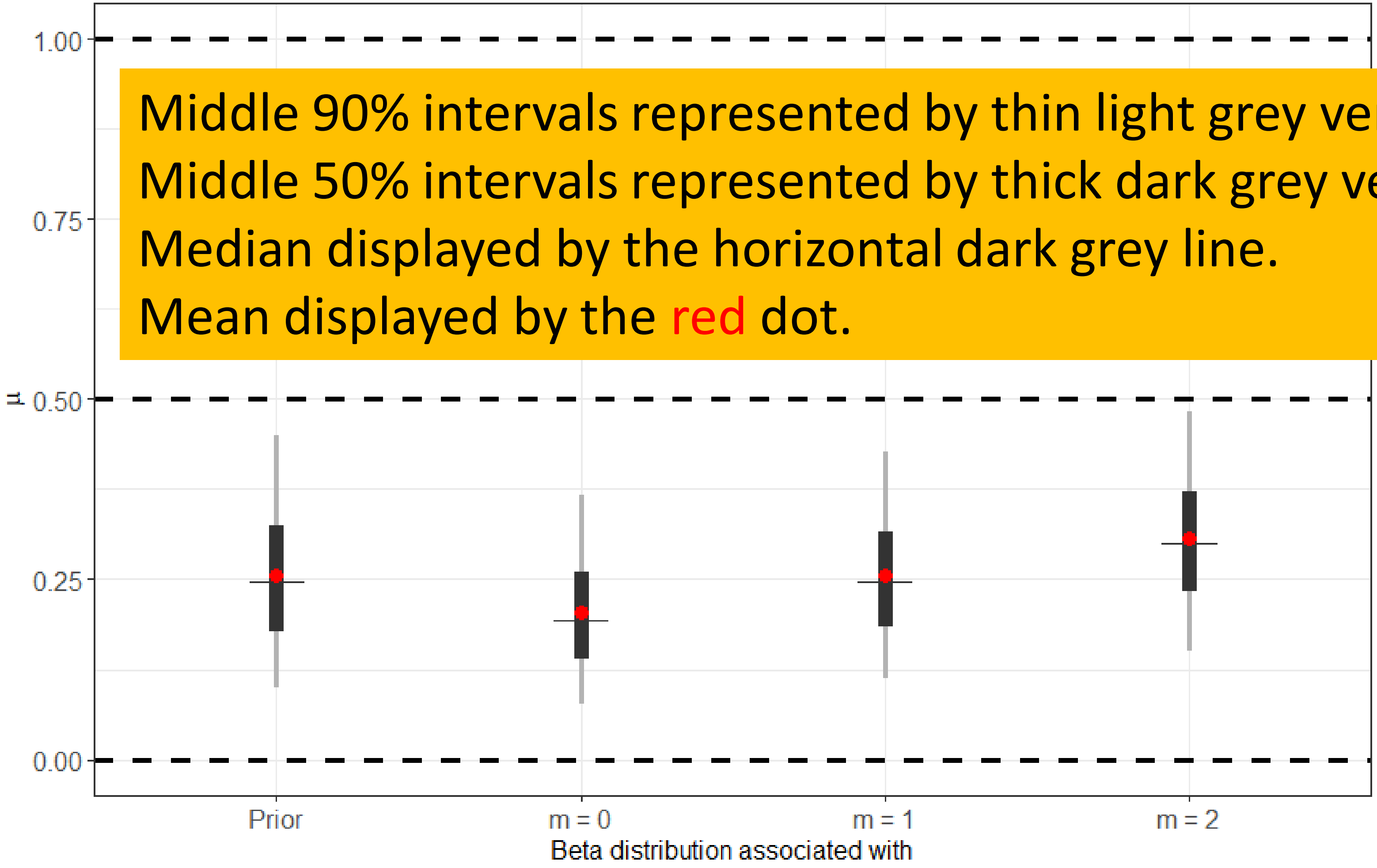
- Represent uncertainty with **<u>credible intervals</u>**:
  - Middle 50% interval – spans the 25^{th} through 75^{th} quantiles
  - Middle 90% interval – spans the 5^{th} through 95^{th} quantiles

- Represent the central tendency two ways:
  - Median – the 50^{th} quantile
  - Mean (average value)

Visualize the Beta distribution summary statistics

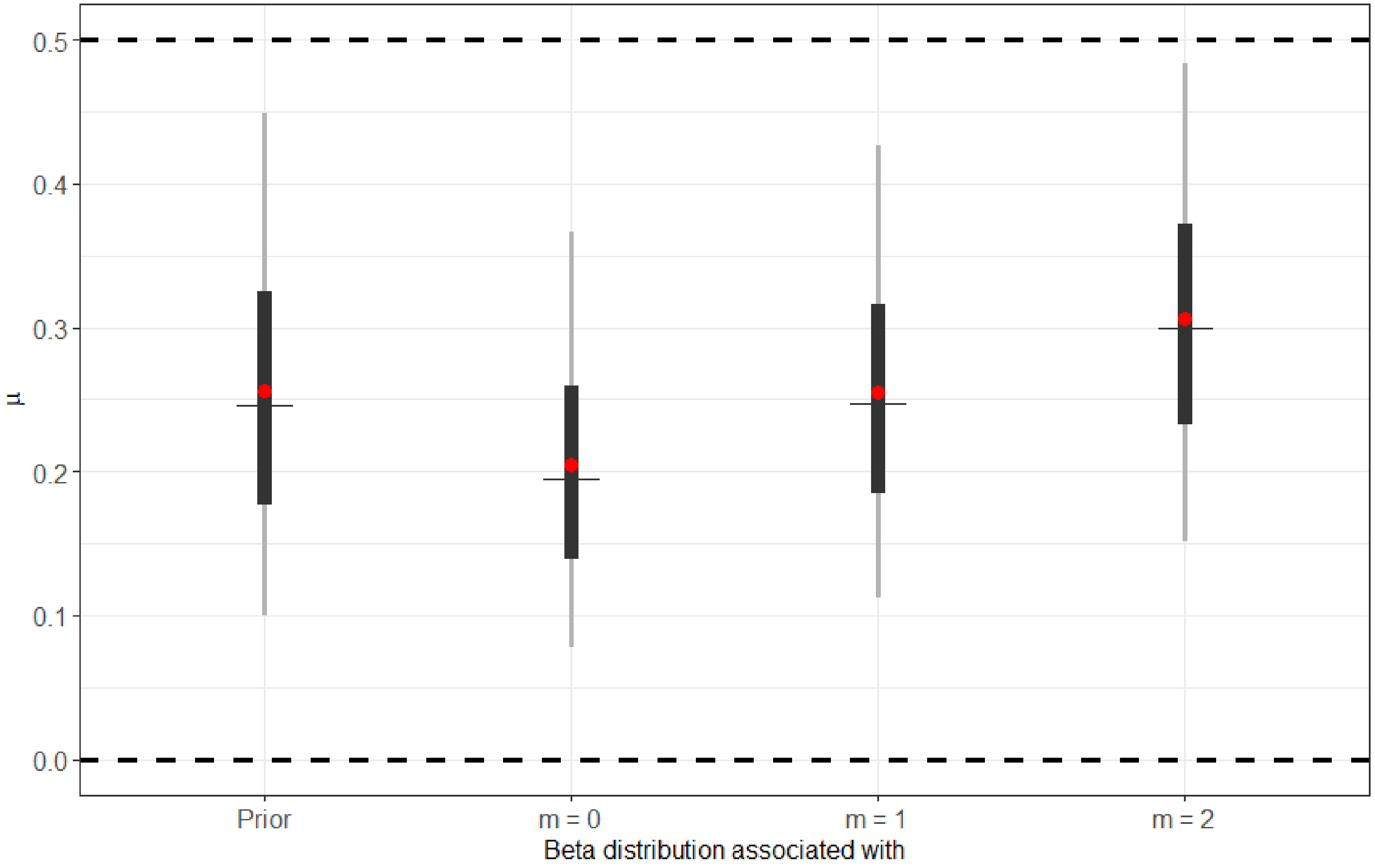


Visualize the Beta distribution summary statistics



Middle 90% intervals represented by thin light grey vertical lines.
Middle 50% intervals represented by thick dark grey vertical lines.
Median displayed by the horizontal dark grey line.
Mean displayed by the red dot.

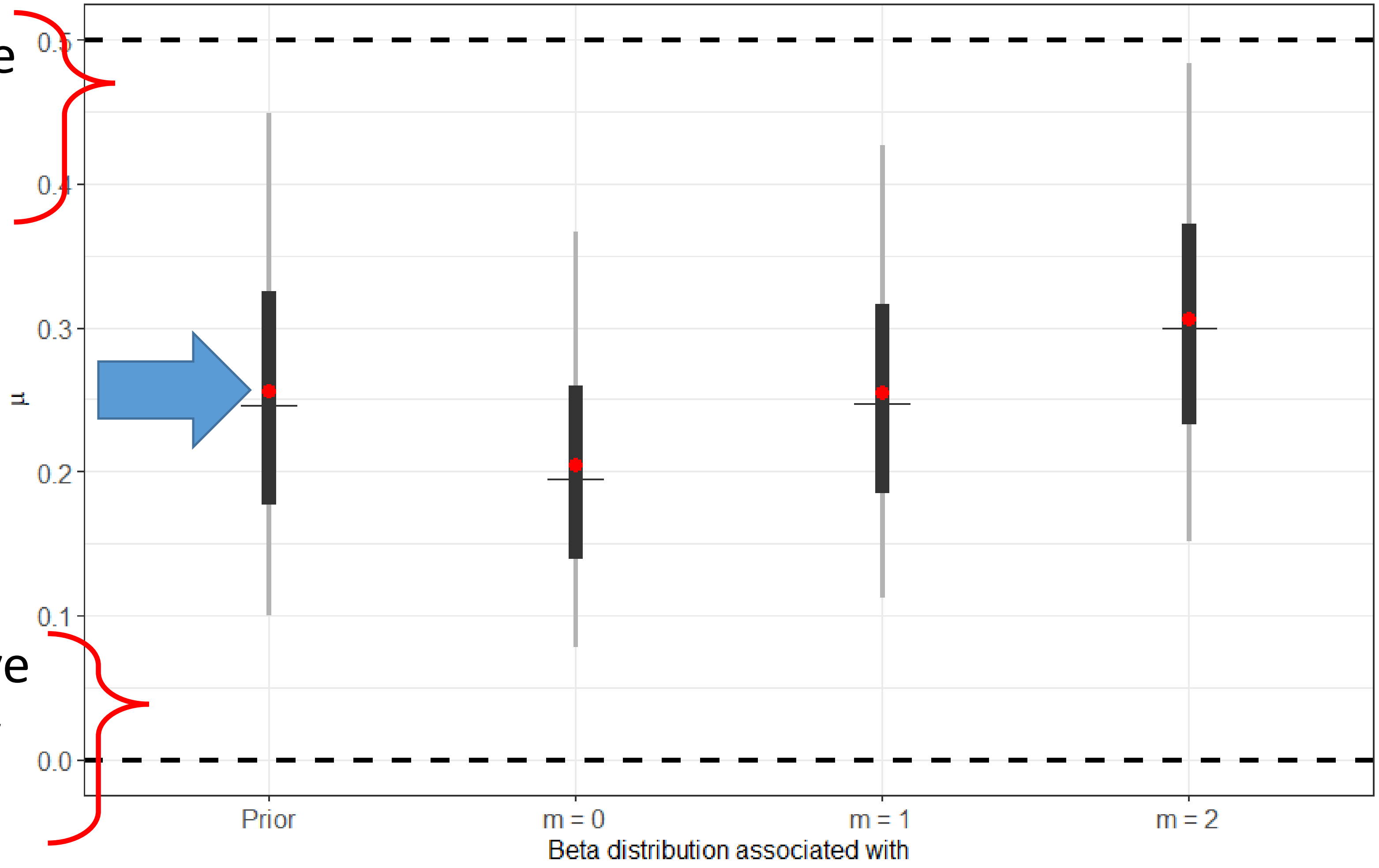
Zoom in



0.5
0.4
0.3
0.2
0.1
0.0
μ
Prior
m = 0
m = 1
m = 2
Beta distribution associated with

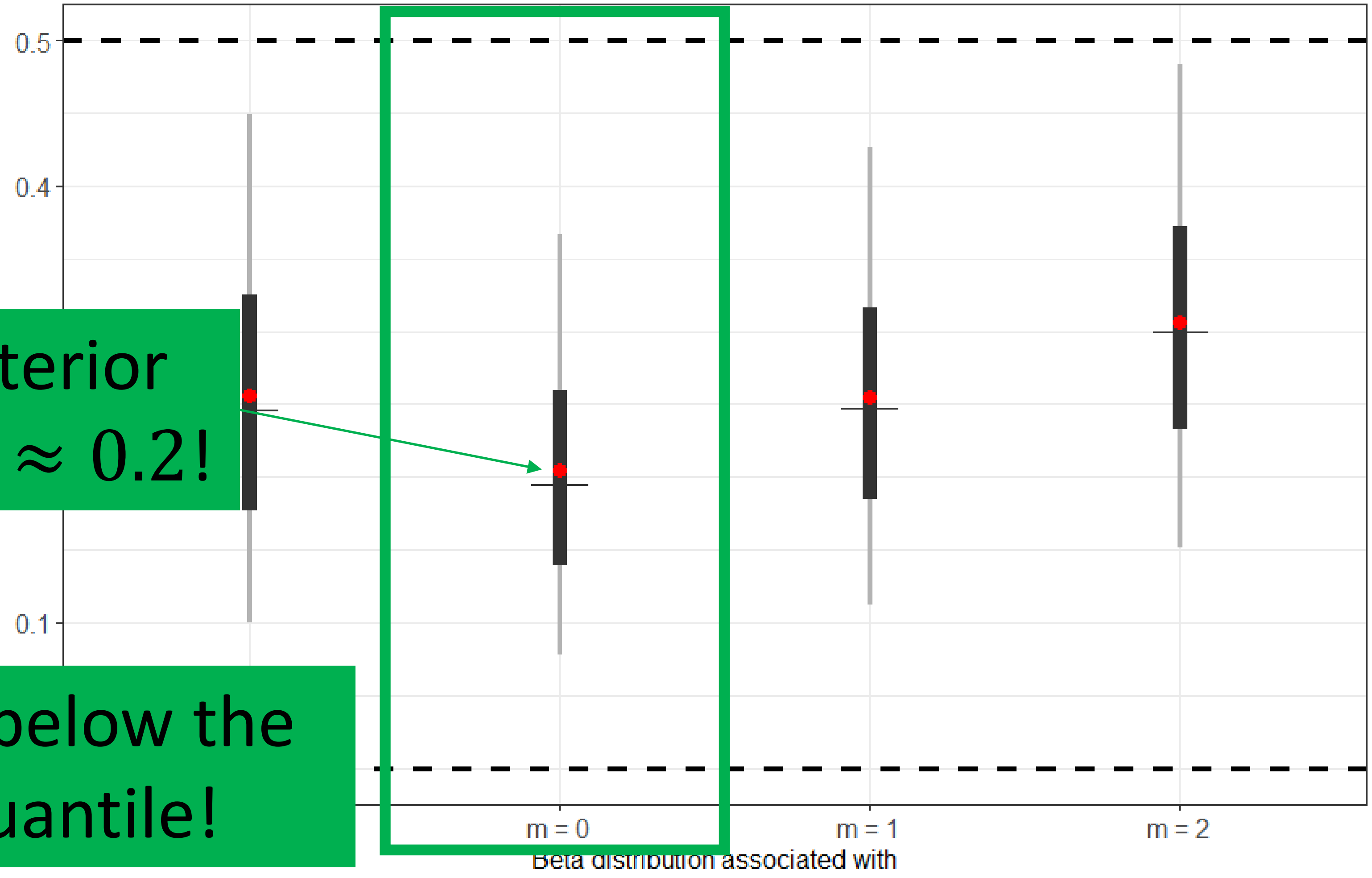
A priori we believe the mean is ≈ 0.25

A priori we believe
$\mu > 0.5$ is very
unlikely



A priori we believe
$\mu < 0.05$ is very
unlikely

If we observed $m = 0$ out of $N = 4$



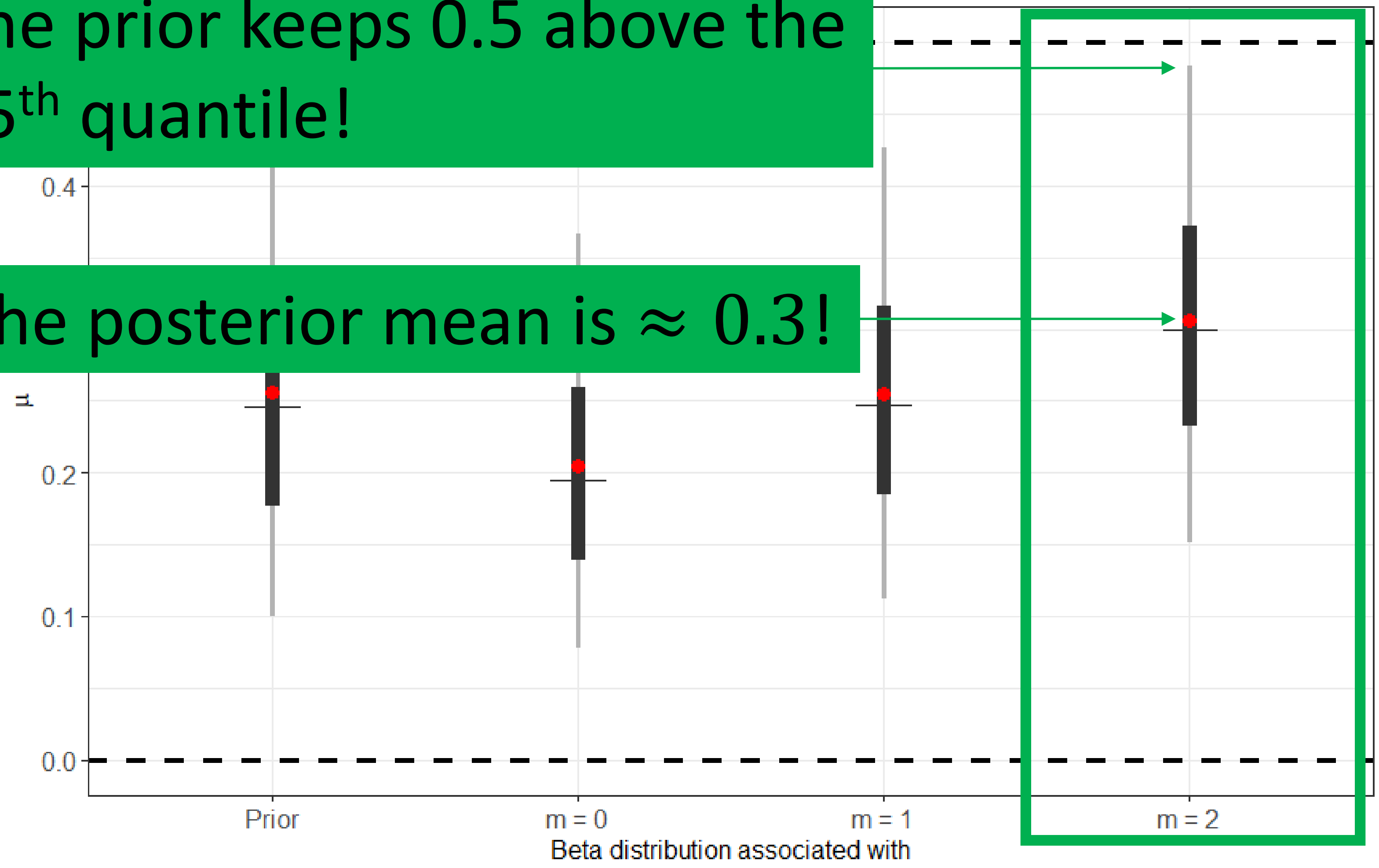
The posterior mean is ≈ 0.2!

Zero is below the 5th quantile!

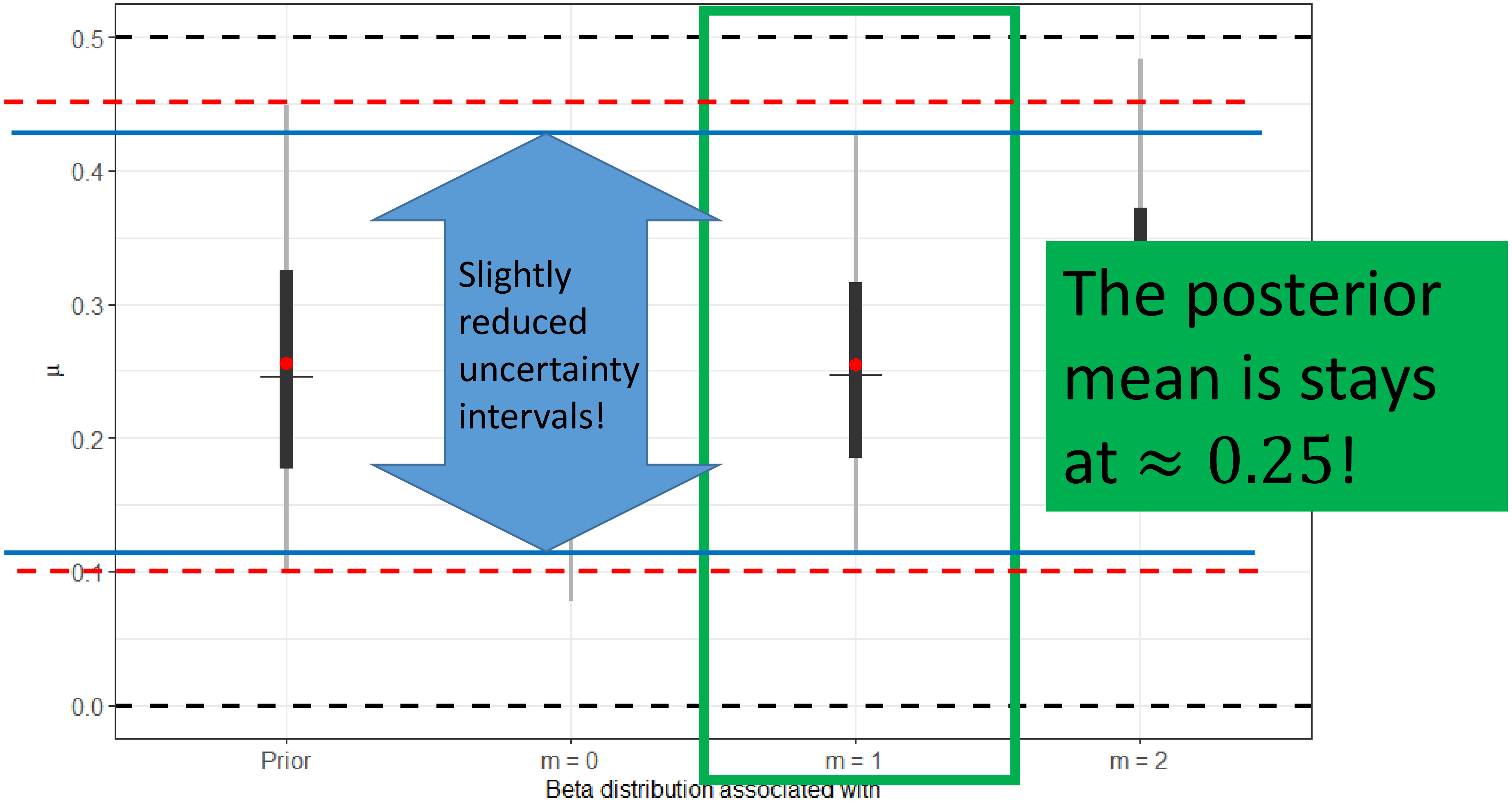
If we observed $m = 2$ out of $N = 4$

The prior keeps 0.5 above the 95th quantile!

The posterior mean is ≈ 0.3!



If we observed $m = 1$ out of $N = 4$



We introduced discussing uncertainty from a Bayesian framework

- However, classical or frequentist statistics also have ways for estimating uncertainty.

- Uncertainty usually represented by **confidence intervals**.

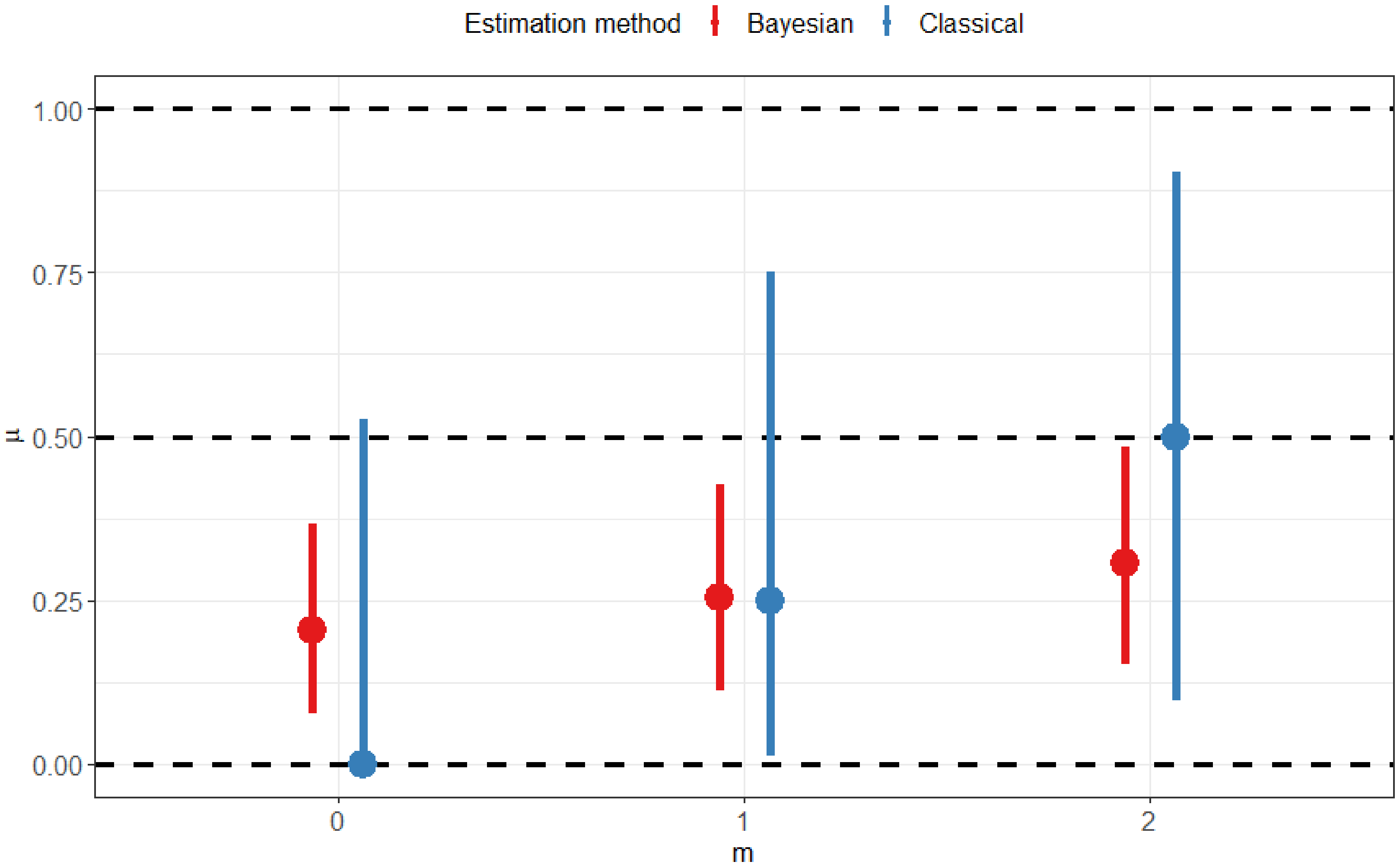
- How do 90% confidence intervals compare with the *posterior credible intervals* in our college football example?

Confidence interval calculation

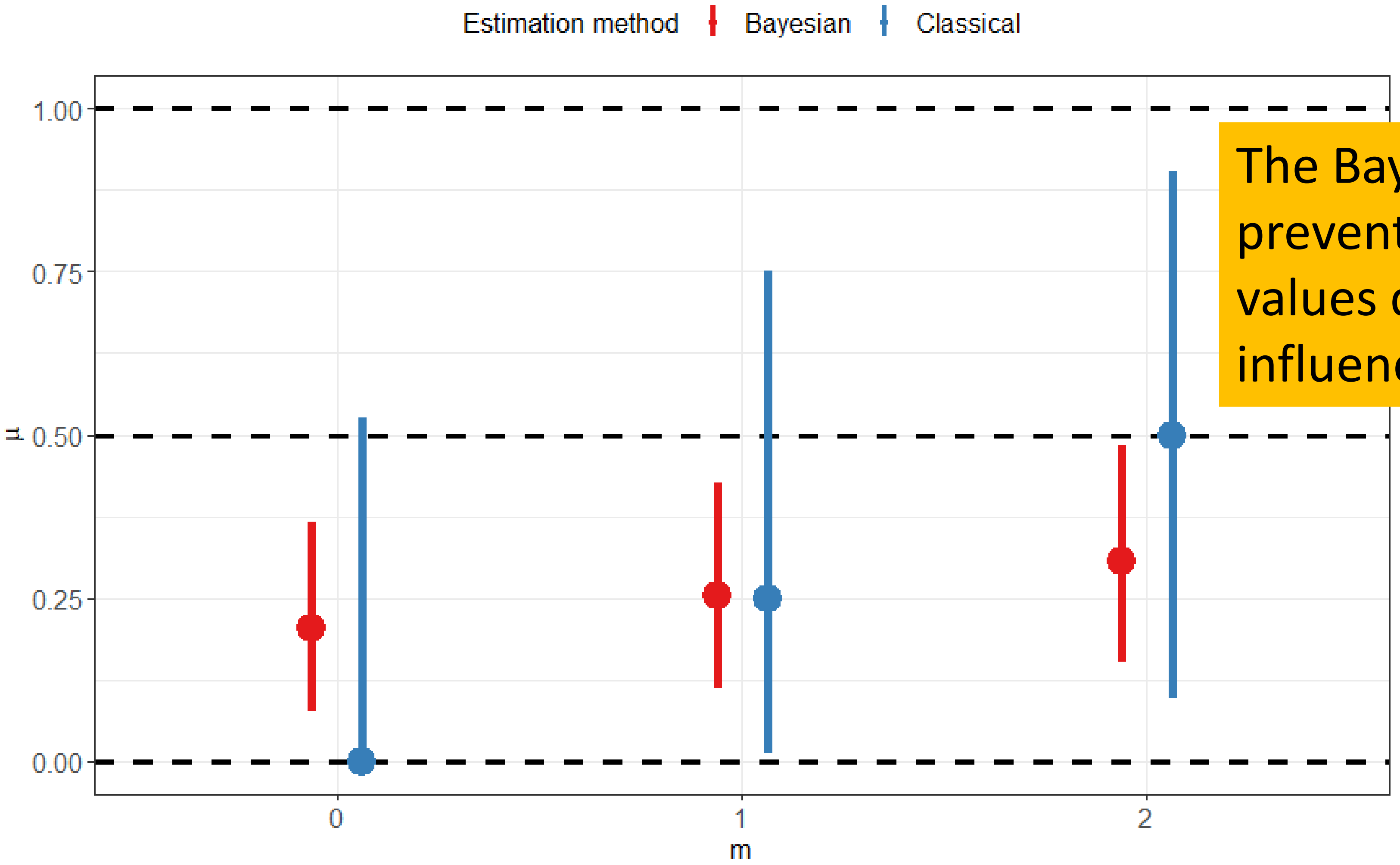
- The 90% confidence intervals are calculating using the Clopper-Pearson method, through `R`'s `binom.test()` function.

- Please see `?binom.test` for more discussion around the method.

90% credible intervals (red) compared with the 90% confidence intervals (blue)



90% credible intervals (red) compared with the 90% confidence intervals (blue)



The Bayesian approach prevents unrealistic values due to the influence of the prior!!