Introduction to Machine Learning

Week 4 - Bayesian Inference and Prior Distributions

Spring 2025
Instructor: Dr. Patrick Skeba

We've discussed two probability distributions

Functions that describe the likelihood of the data given values of the unknown parameter

- Bernoulli: $p(x \mid \mu) = \mu^x (1-\mu)^{1-x}$ for $x \in \{0,1\}$
- Binomial: $p(m \mid N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$

Application of binomial distribution

Estimating how many people like SW:TLJ from multiple review aggregators, instead of individual people

- Let K be the number of review aggregators.
  The total likelihood can then be expressed as:

$$p(\mathbf{m} \mid \mathbf{N}, \mu) = \prod_{k=1}^{K} p(m_k \mid N_k, \mu)$$

$$= \prod_{k=1}^{K} \binom{N_k}{m_k} \mu^{m_k} (1-\mu)^{N_k - m_k}$$

$$\hat{\mu}_{MLE} = \frac{\sum_{k=1}^{K} m_k}{\sum_{k=1}^{K} N_k}$$

	Like (m)	Dislike (N-m)	Total (N)
Rotten Tomatoes	199	286	485
Metacritic	5,678	13,245	18,923
StarFanatic	23	2	25
IHateStarWars.com	1	44	45

$$\mu_{MLE} = \frac{199 + 5{,}678 + 23 + 1}{485 + 18{,}923 + 25 + 45} = 0.303$$

What would've happened if we only considered data from one of the fan/hate sites?

StarFanatic MLE: $\frac{23}{25} = 0.92$ **IHateStarWars.com MLE:** $\frac{1}{45} = 0.02$

	Like (m)	Dislike (N-m)	Total (N)
Rotten Tomatoes	199	286	485
Metacritic	5,678	13,245	18,923
StarFanatic	23	2	25
IHateStarWars.com	1	44	45

The estimates from these sites are far away from what we would expect!

What can we do when our training data is too small or unreliable?

Get more data

Bayesian statistics! Introduce a prior belief on μ that establishes which values are most likely *before observing any training data.*

"about half the people I know like TLJ…I don't really think IHateStarWars seems very accurate…"

Bayes formulation for estimating μ

Start with an initial hypothesis, update based on the data

$$p(\mu \mid x) \propto p(x \mid \mu) p(\mu)$$

Prob. of data, given the unknown parameter(s)

a priori probability of parameters, irrespective of data

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

Prob. of parameters, given the data

Total probability of data, integrated over the range of μ

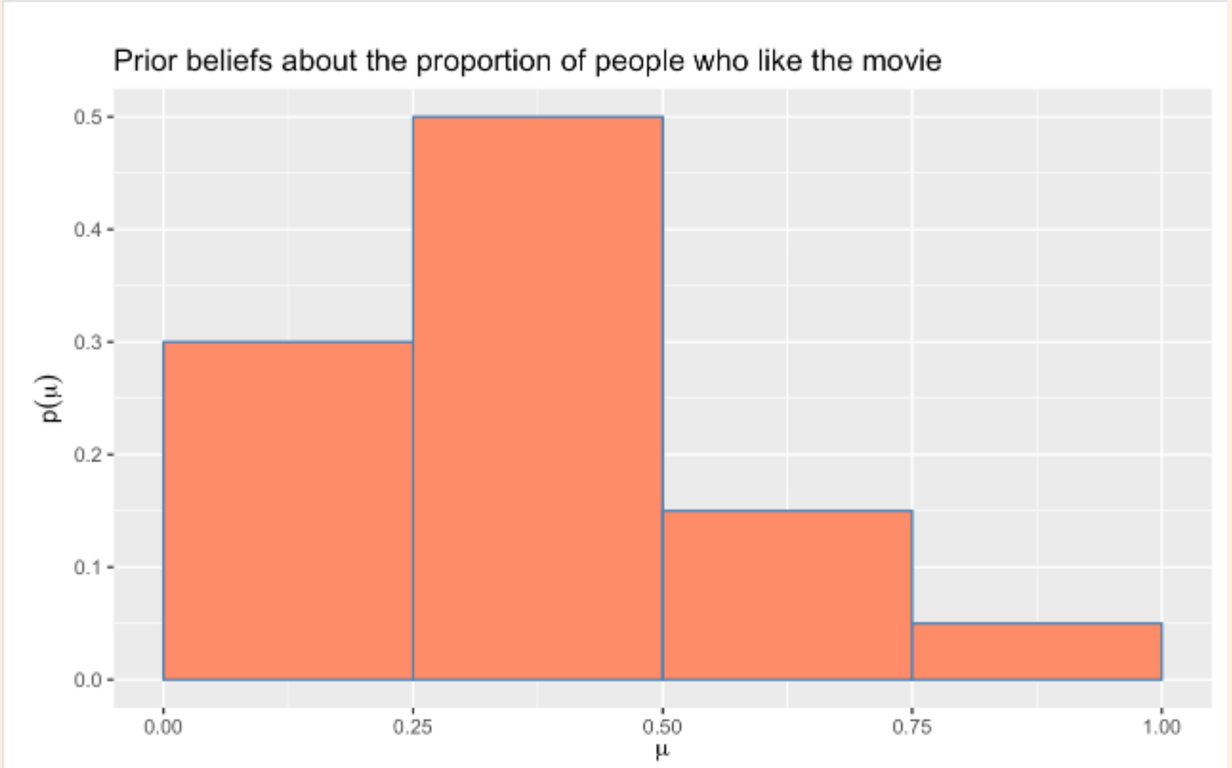
Encode prior beliefs about which values of the parameters are most likely

create a probability distribution over all possible values that μ can take

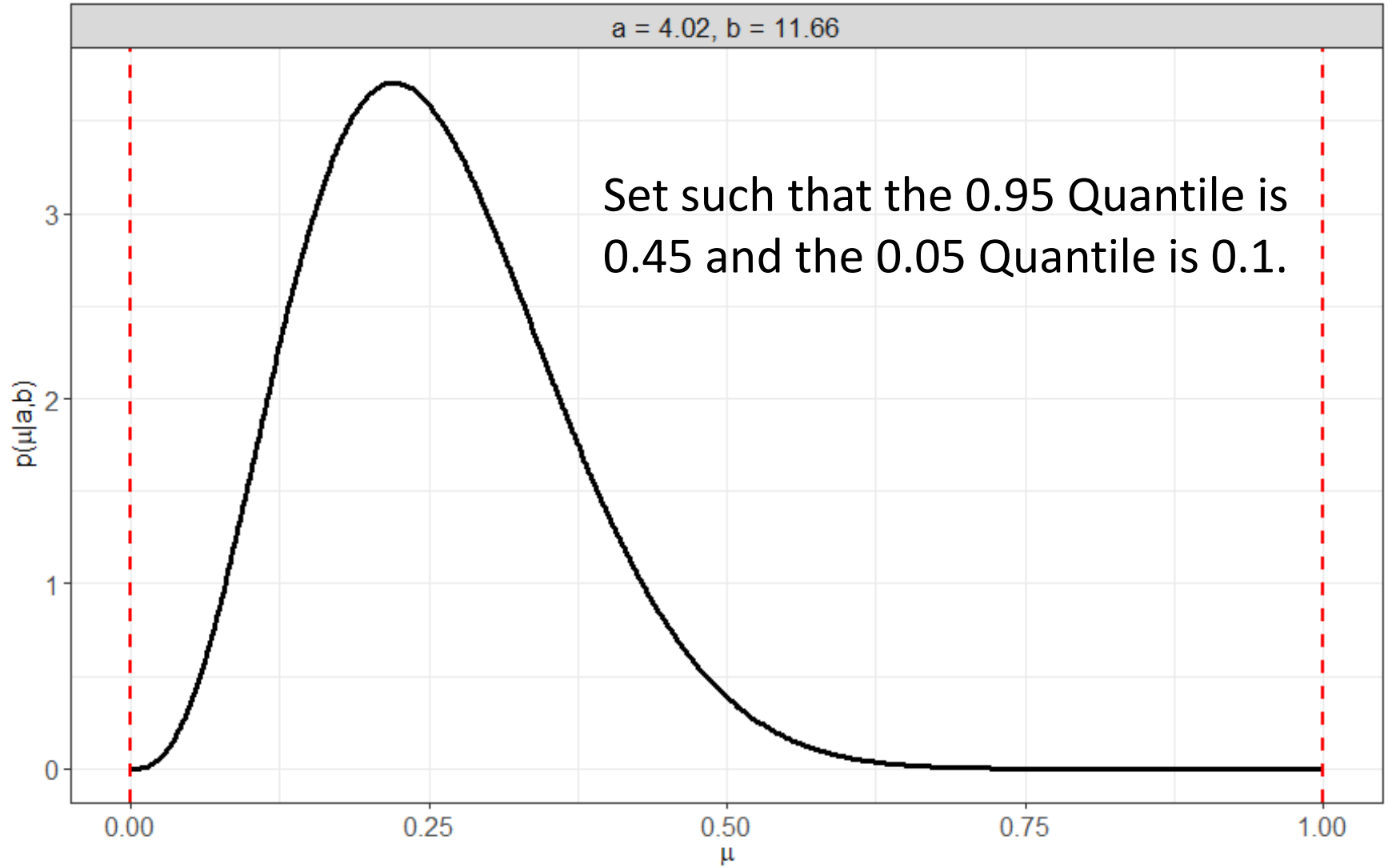
$$p(\mu) = \begin{cases} 0, & \mu < 0 \text{ or } \mu > 1 \\ .3, & 0 \leq \mu < 0.25 \\ .5, & 0.25 \leq \mu < 0.5 \\ .15, & 0.5 \leq \mu < 0.75 \\ .05, & 0.75 \leq \mu \leq 1 \end{cases}$$

$$p(\mu \mid m, N) \propto p(m \mid N, \mu) p(\mu)$$

```
tibble(mu0 = c(0.125,0.375,0.625,0.875), # bucket midpoints
       `p(mu0)` = c(0.30,0.50,0.15,.05)) |>
  ggplot(aes(x=mu0,y=`p(mu0)`)) +
  geom_bar(stat = 'identity',width=0.25,fill='salmon1',color='steelblue') +
  labs(title = "Prior beliefs about the proportion of people who like the movie",
       x = latex2exp::TeX(r"($\mu$)"),
       y = latex2exp::TeX(r"($p(\mu)$)"))
```



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



Maximum A Posteriori (MAP) Estimate for μ

Consider both how likely the data is given the parameters, AND how likely the parameters themselves are based on prior information

$$\hat{\mu}_{MAP} = \arg\max_{\mu} p(x \mid \mu) p(\mu)$$

- **The Evidence (denominator in Bayes Theorem) is a normalizing constant - therefore, we can ignore it for the purposes of optimization.**
- **The optimal parameters in a Bayesian formulation are a compromise between the data and your prior beliefs**
- **You as the analyst specify the prior distribution, it is separate from the model's training data**
- **The prior represents a kind of initial hypothesis, against which you compare the estimates from the data to update your belief about the "true" parameter value**

We can define *prior distributions* over the parameters

When our likelihood is a Binomial/Bernoulli distribution, the "conjugate prior" is the Beta Distribution

Gamma function: extension of factorial to complex numbers

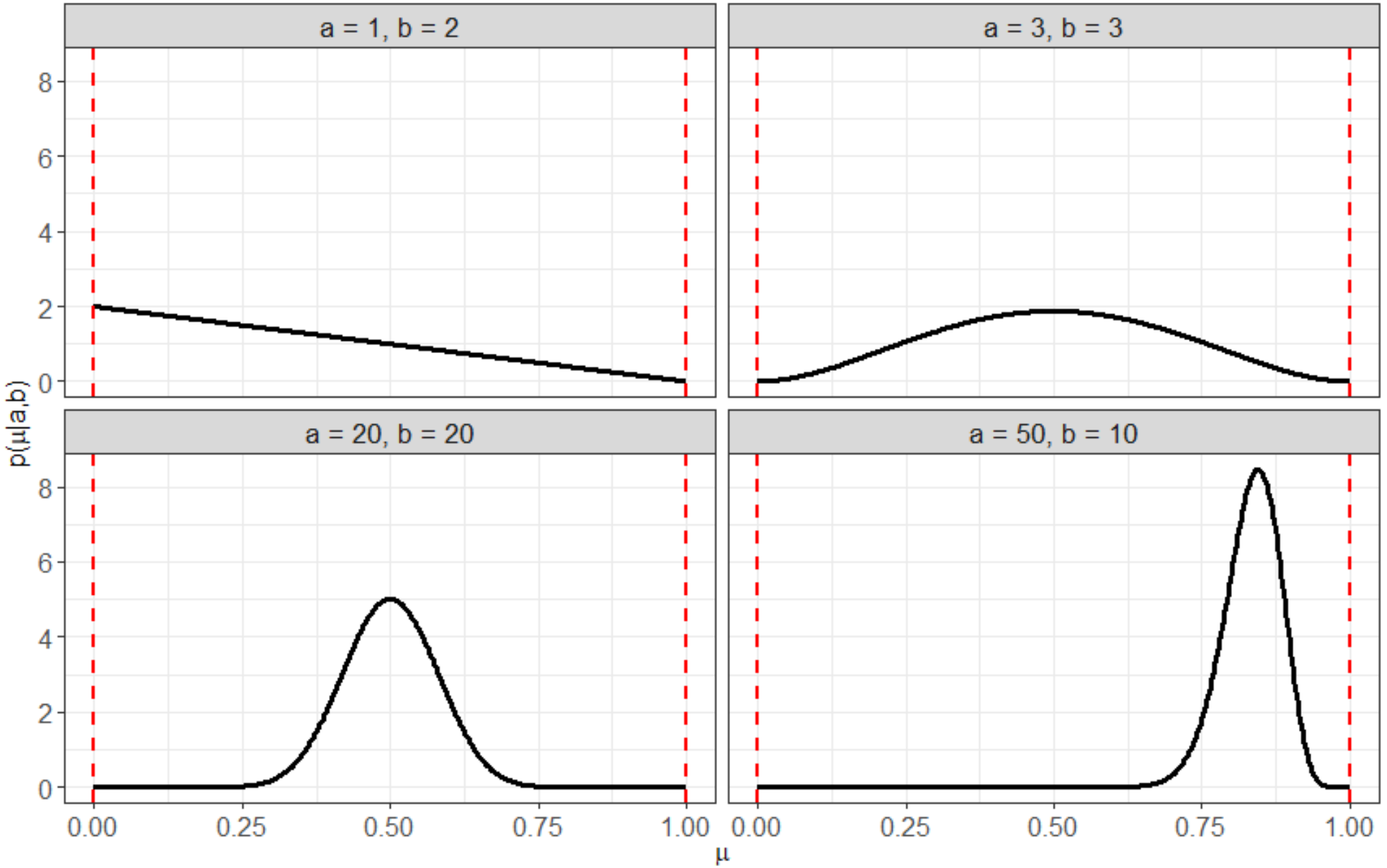
- $$\texttt{Beta}(\mu \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

- **The Beta distribution is a probability density function for continuous variables bounded between 0 and 1. It's a very flexible distribution that can take on many different shapes.**

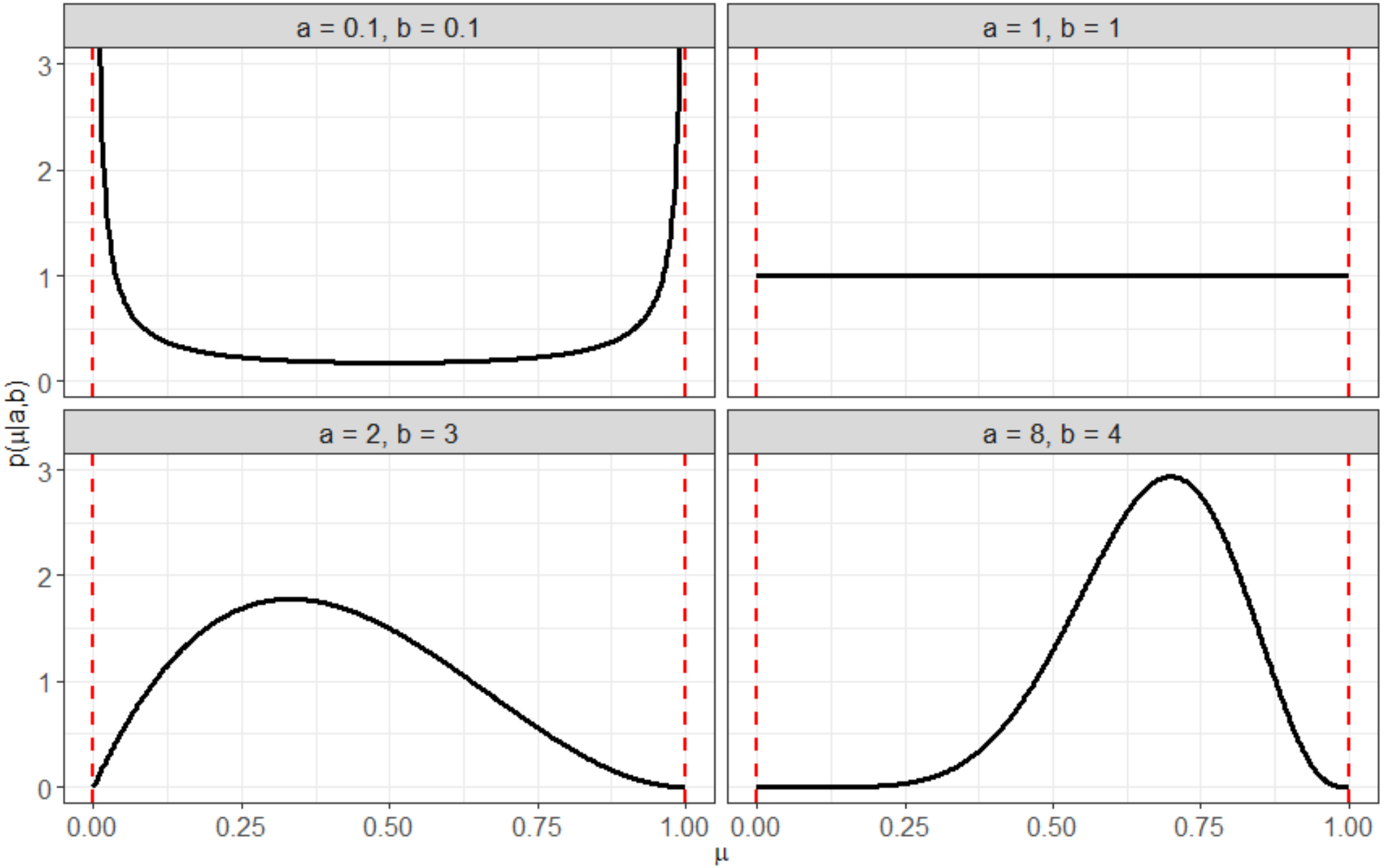
- $\texttt{Posterior}(\mu \mid m, N, a, b) \propto \texttt{Binomial}(m \mid N, \mu) \times \texttt{Beta}(\mu \mid a, b)$

- **Parameters a, b can be interpreted as the number of prior events and non-events, respectively.**

Example shapes of the beta distribution



Example shapes of the beta distribution



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

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

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

Combine the likelihood and prior to produce the (unnormalized) posterior distribution

"How likely is each value of μ given my prior belief and the data?"

$$p(\mu \mid m, N, a, b) \propto p(m \mid N, \mu) p(\mu \mid a, b)$$
$$\propto \mu^m (1-\mu)^{N-m} \mu^{a-1} (1-\mu)^{b-1}$$
$$\propto \mu^{m+a-1} (1-\mu)^{N-m+b-1}$$

Posterior: $\texttt{Beta}(\mu \mid m + a, N - m + b)$

$$a_{new} = m + a,\ b_{new} = N - m + b$$

$$p(\mu \mid m, N, a, b) = \texttt{Beta}(\mu \mid a_{new}, b_{new})$$

When the prior is conjugate to the likelihood, the posterior has the same functional form as the prior!
If the prior is a Beta, then the Posterior will be a Beta

Including a prior can help to mitigate an unreliable MLE

Contextualize a biased movie-review site with outside survey results

- Consider a relatively small dataset, like a fan-site with only 5 users. 4 out of the 5 say Yes, and 1 says No.
  - $\hat{\mu}_{MLE} = 4/5 = 0.8$
- However, a survey on the radio said 39 out of 176 people liked the movie
  - Prior parameters: $a = 39,\ b = 137$
  - Prior mode: $\dfrac{a-1}{a+b-2} = 0.218$
- $\texttt{Posterior} \propto \mu^4(1-\mu)^1\mu^{38}(1-\mu)^{136}$
  - $\hat{\mu}_{MAP} = \dfrac{a_{new}-1}{a_{new}+b_{new}-2} = 0.234$

Derive the Posterior Mode (MAP Estimate)

First, combine likelihood and prior distributions

$$Posterior \propto Likelihood \times Prior$$

$$\begin{aligned} p(\mu \mid m, N, a, b) &\propto p(m \mid N, \mu) p(\mu \mid a, b) \\ &\propto \mu^{m}(1-\mu)^{N-m}\mu^{a-1}(1-\mu)^{b-1} \\ &\propto \mu^{m+a-1}(1-\mu)^{N-m+b-1} \quad \text{combine powers of like bases} \\ &\propto \mu^{a_{new}-1}(1-\mu)^{b_{new}-1} \end{aligned}$$

- **As usual, we'll work directly with the log-posterior:**

$$\log p(\mu \mid a_{new}, b_{new}) \propto (a_{new}-1)\log \mu + (b_{new}-1)\log(1-\mu)$$

Derive the Posterior Mode (MAP Estimate)

Second, calculate the derivate of the log-posterior with respect to the unknown parameter μ

$$\log p(\mu \mid a_{new}, b_{new}) \propto (a_{new} - 1)\log \mu + (b_{new} - 1)\log(1 - \mu)$$

$$\frac{\partial}{\partial \mu} \log p(\mu \mid a_{new}, b_{new}) \propto \frac{\partial}{\partial \mu} \left[(a_{new} - 1)\log \mu + (b_{new} - 1)\log(1 - \mu)\right]$$

$$= \frac{a_{new} - 1}{\mu} - \frac{b_{new} - 1}{1 - \mu}$$

$$= \frac{(1 - \mu)(a_{new} - 1) - \mu(b_{new} - 1)}{\mu(1 - \mu)}$$

$$= \frac{a_{new} - 1 - \mu a_{new} + \mu - \mu b_{new} + \mu}{\mu(1 - \mu)}$$

$$= \frac{a_{new} - 1 - \mu(a_{new} + b_{new} - 2)}{\mu(1 - \mu)}$$

Derive the Posterior Mode (MAP Estimate)

Finally, set the derivative equal to zero and solve for the unknown parameter

$$0 = \frac{a_{new} - 1 - \mu(a_{new} + b_{new} - 2)}{\mu(1-\mu)}$$

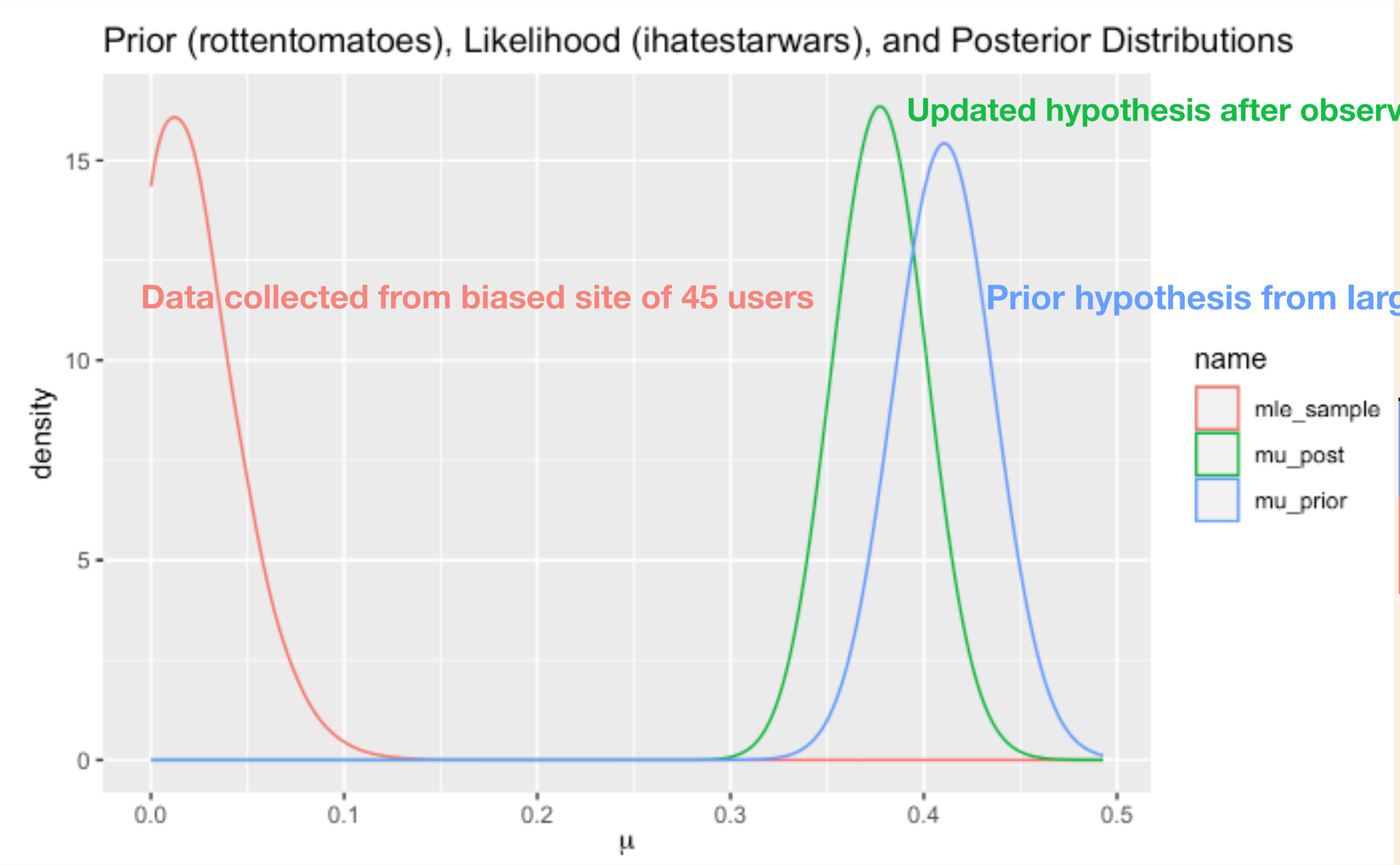
$$0 = a_{new} - 1 - \mu(a_{new} + b_{new} - 2)$$

$$\mu(a_{new} + b_{new} - 2) = a_{new} - 1$$

$$\hat{\mu}_{MAP} = \frac{a_{new} - 1}{a_{new} + b_{new} - 2} = \frac{m + a - 1}{N + a + b - 2}$$

Visualize three major components of Bayesian Posterior

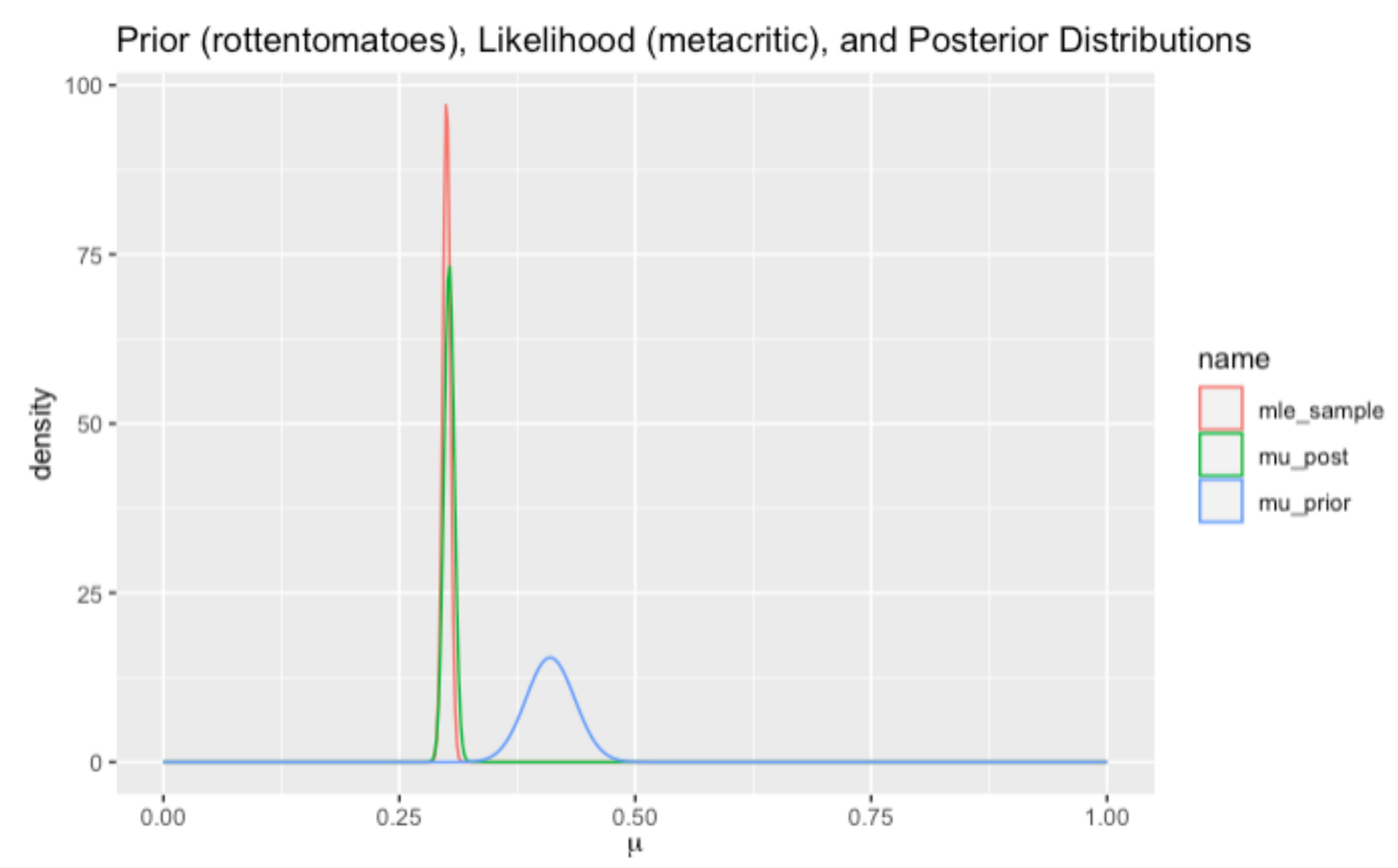
Probability of mu given the data, given prior information, and given both the data AND prior information



Role	Like	Dislike	Total
Rotten Tomatoes	Prior	199	286	485
IHateStarWars.com	Lik (data)	1	44	45

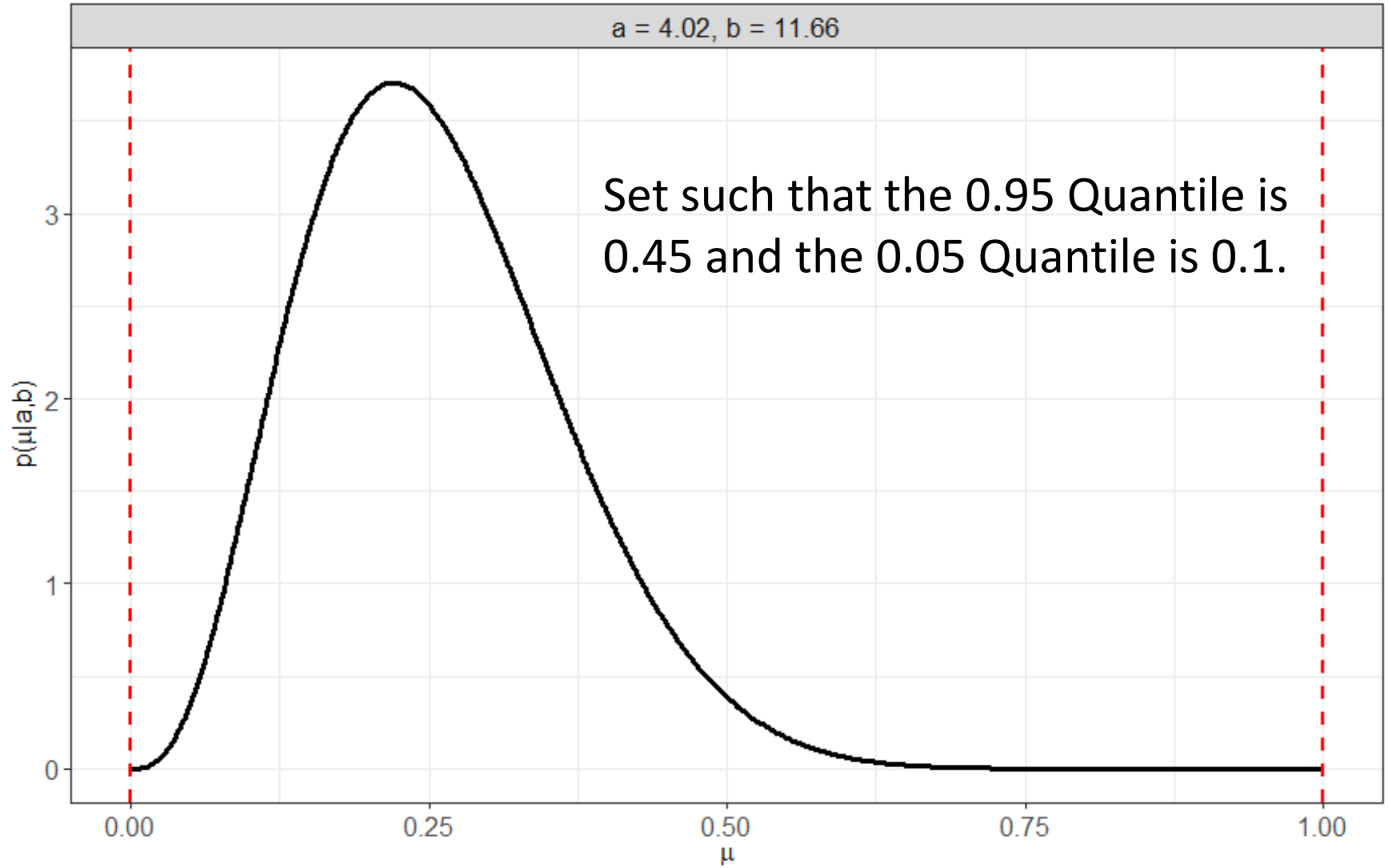
What if we used the same prior, but collected a lot more data for the likelihood?

The posterior will look more like the likelihood!



	Role	Like	Dislike	Total
Rotten Tomatoes	Prior	199	286	485
Metacritic	Lik (data)	5,678	13,245	18,923

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



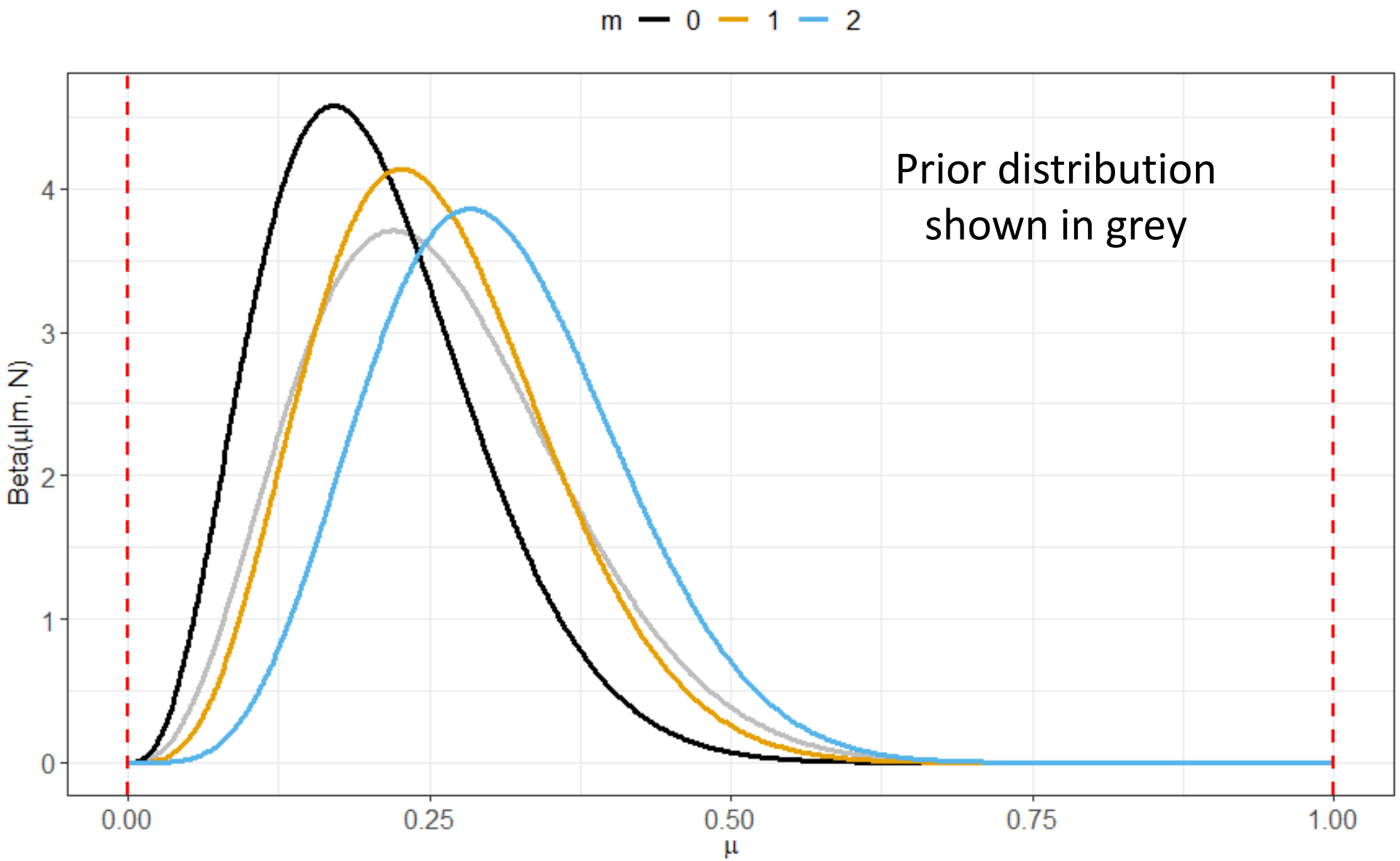
We will update our belief about μ under three different circumstances

- 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 Yes responses, vs 1 Yes response, vs 2 Yes responses.

μ posterior distribution given m and $N = 4$



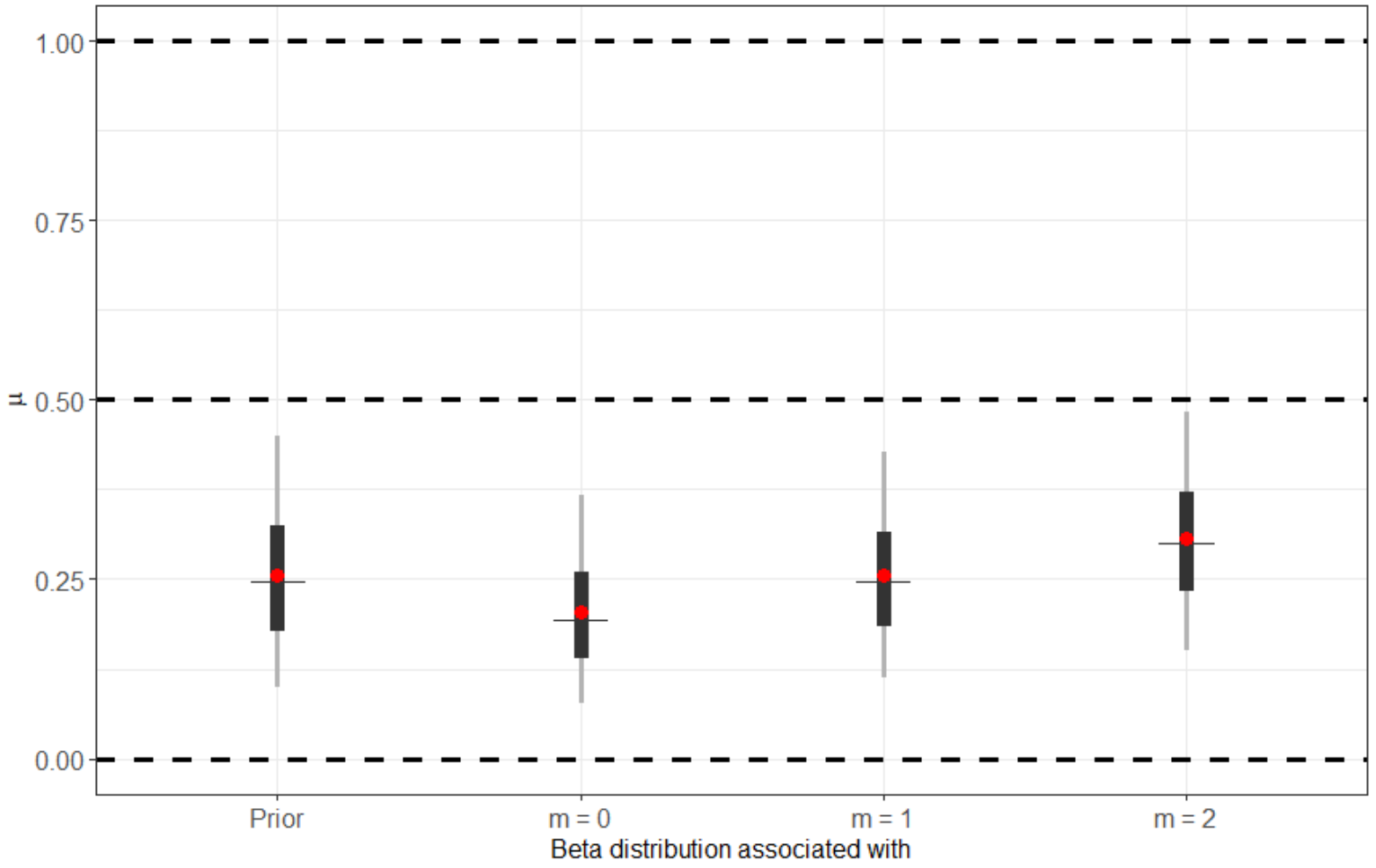
Summarize the Beta distributions

- Calculate summary statistics for each Beta distribution.

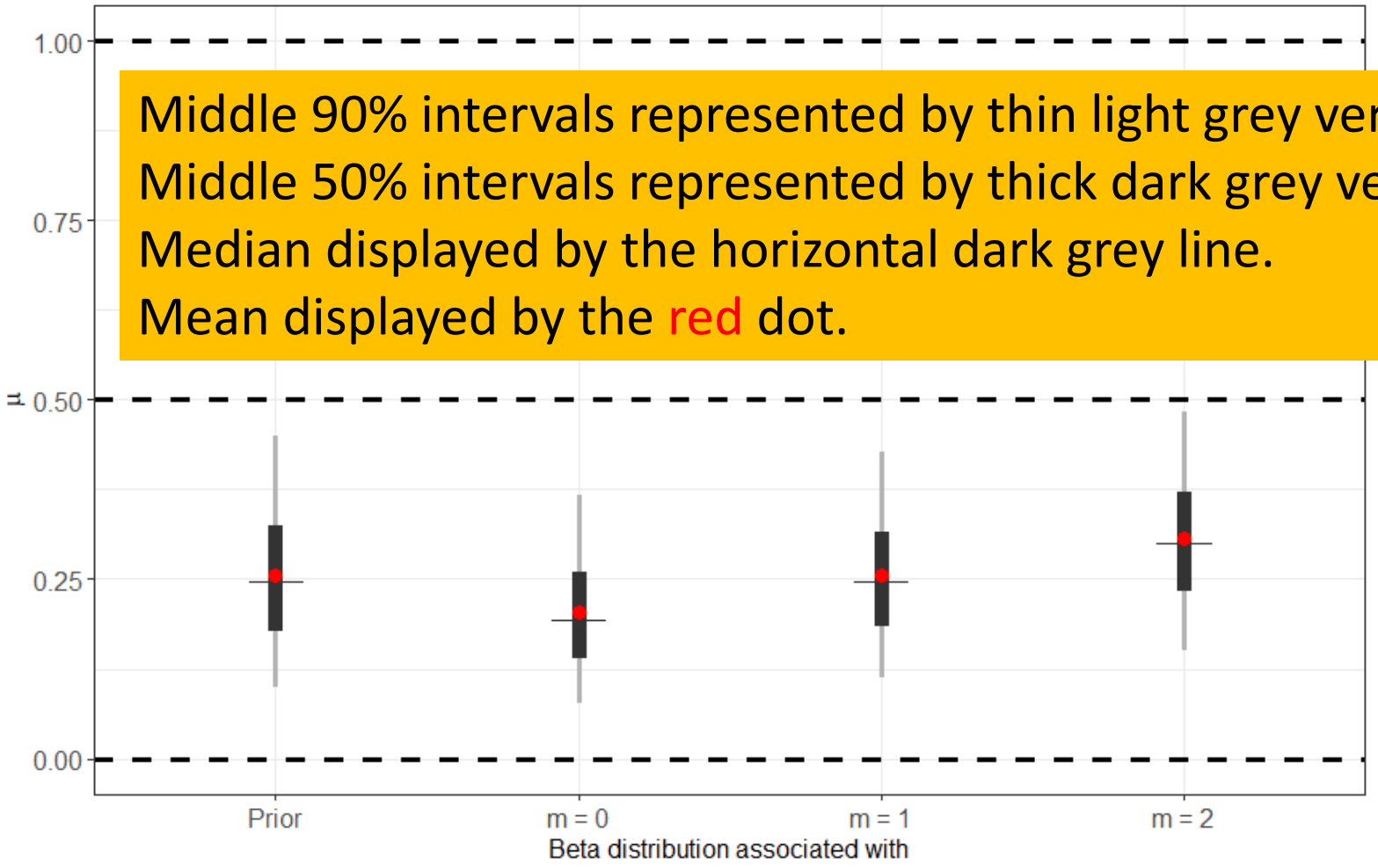
- Represent uncertainty with **credible intervals**:
  - Middle 50% interval – spans the 25th through 75th quantiles
  - Middle 90% interval – spans the 5th through 95th quantiles

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

Visualize the Beta distribution summary statistics

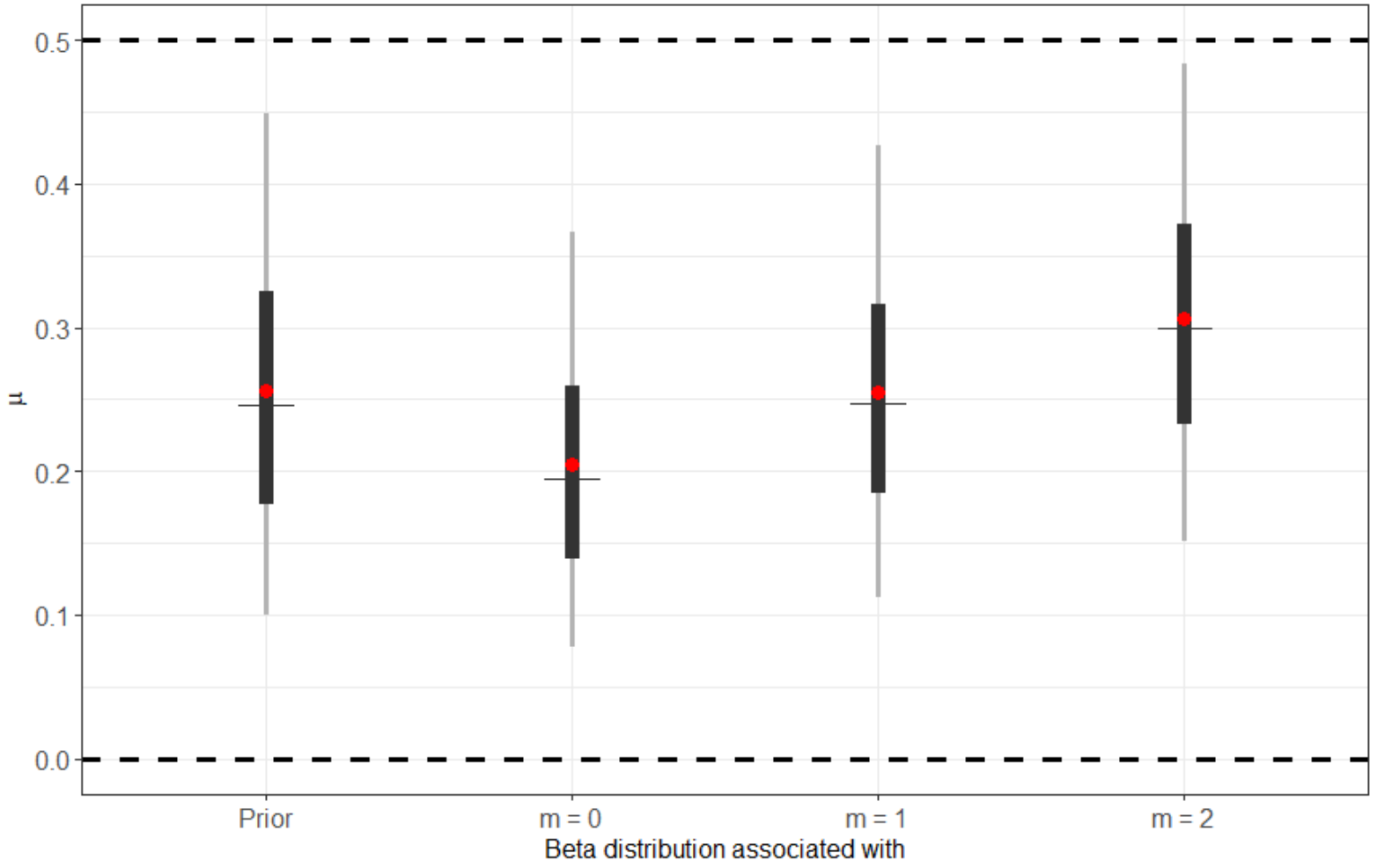


Visualize the Beta distribution summary statistics



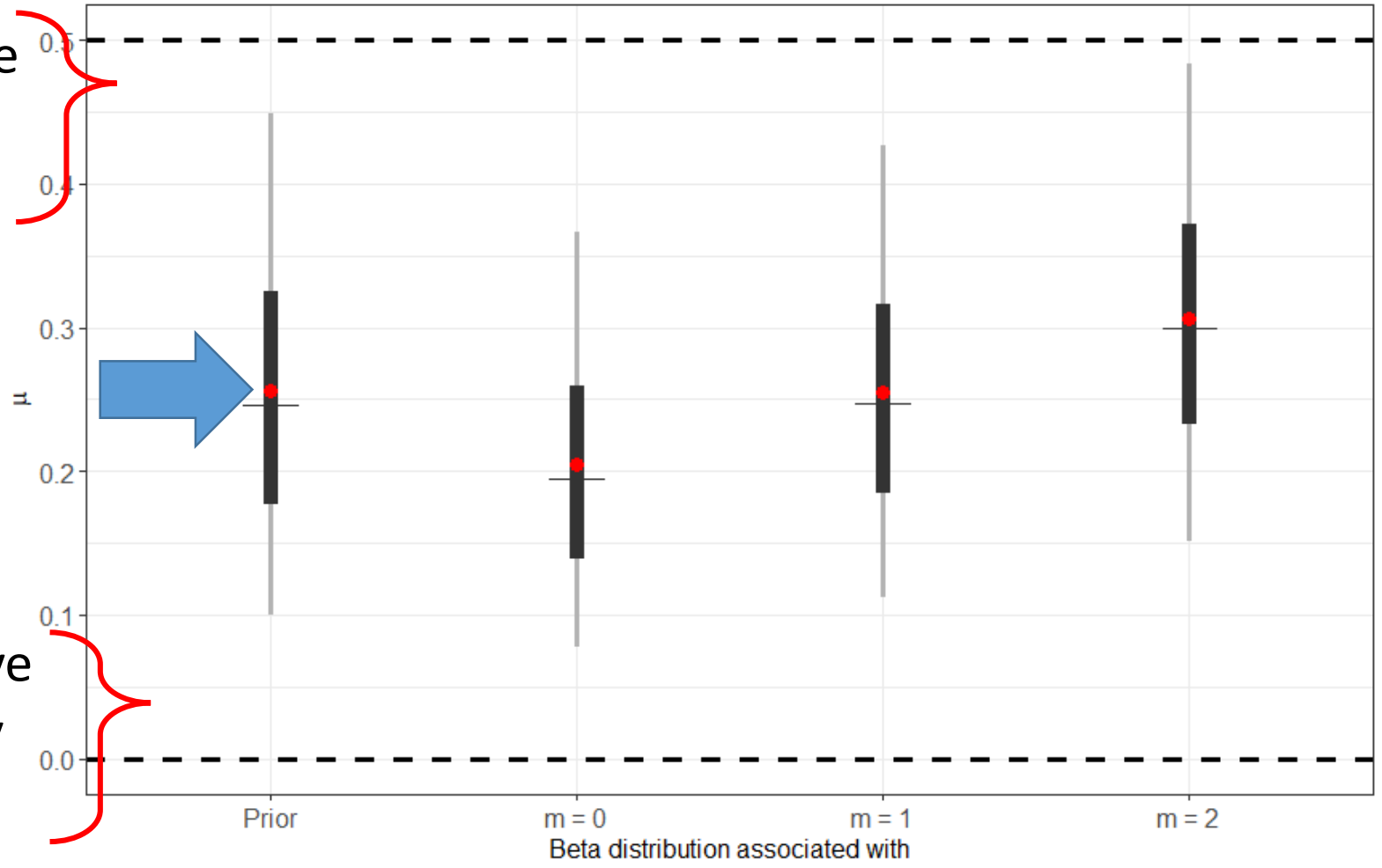
Middle 90% intervals represented by thin light grey vertical lines.
Middle 50% intervals represented by thick dark grey vertical lines.
Median displayed by the horizontal dark grey line.
Mean displayed by the red dot.

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

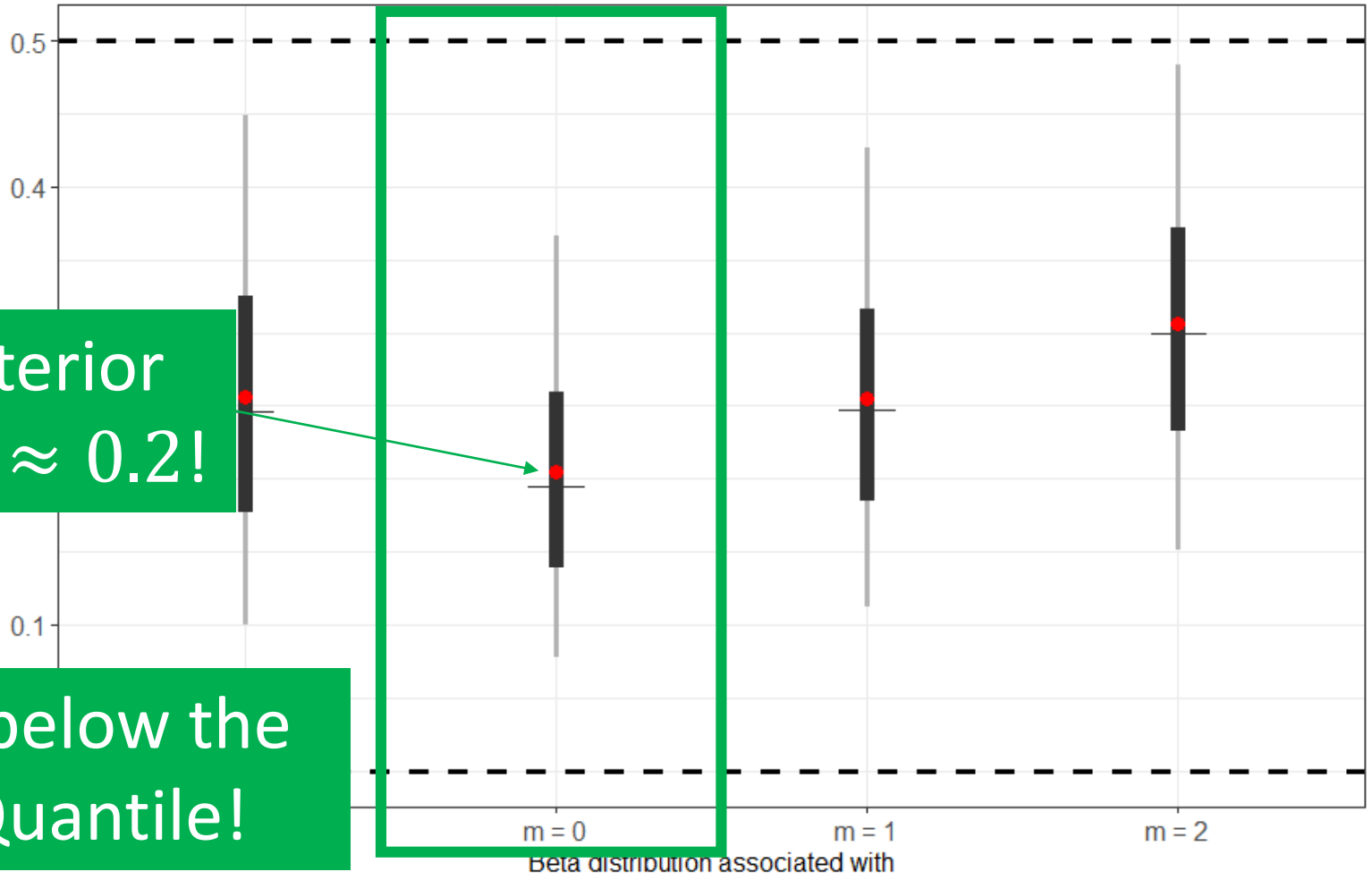
A priori we believe the mean is ≈ 0.25

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 $N = 4$



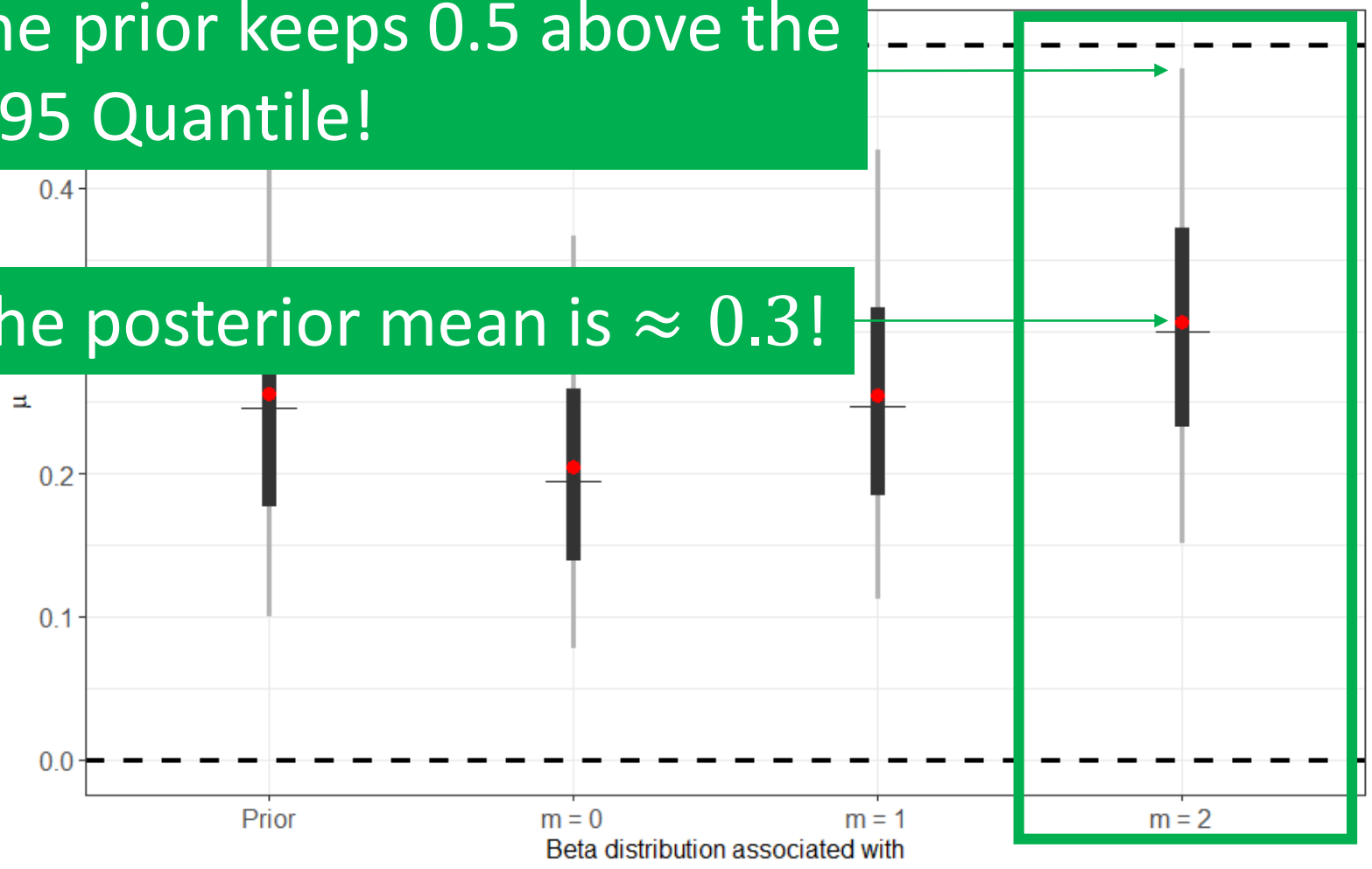
The posterior mean is ≈ 0.2!

Zero is below the 0.05 Quantile!

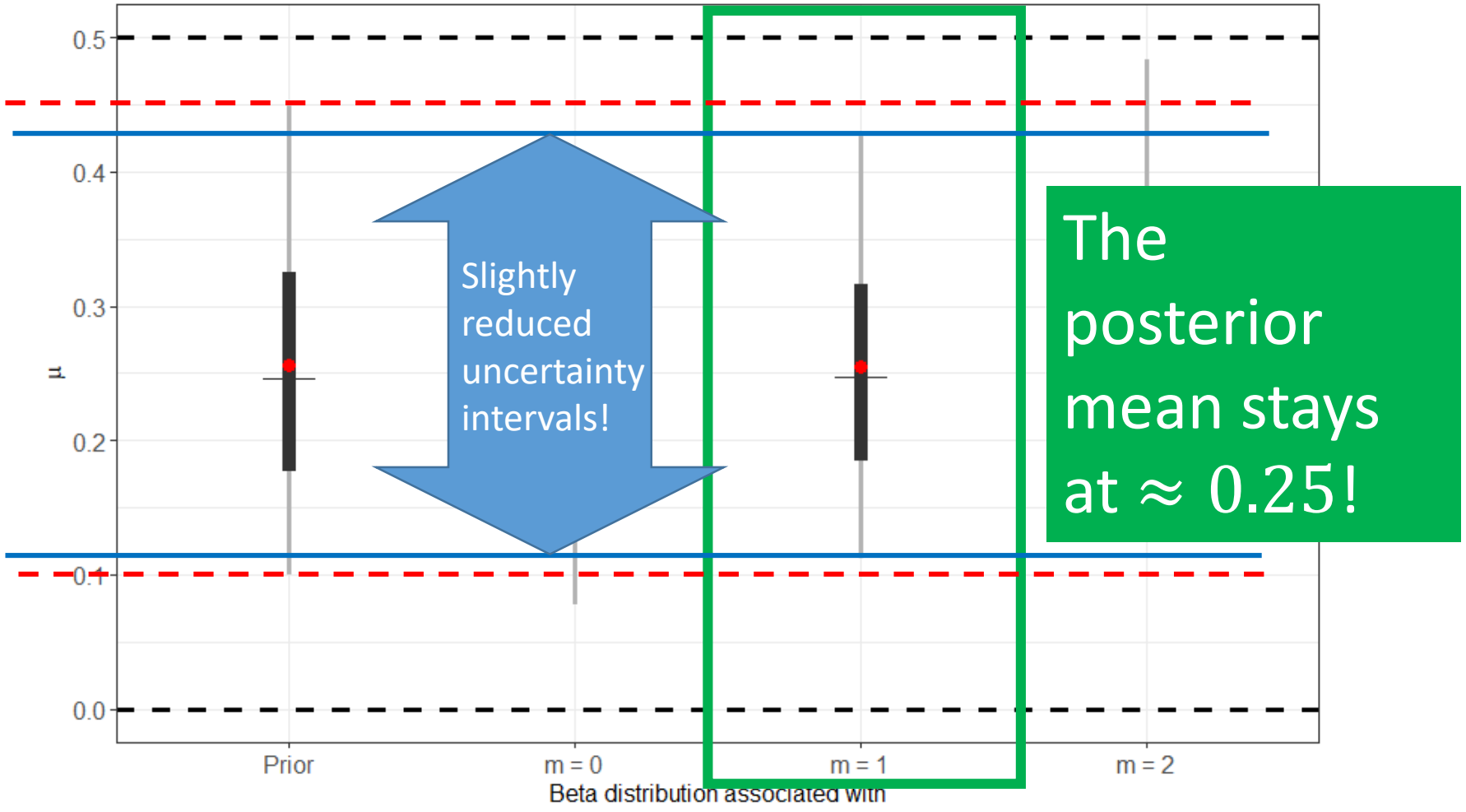
IF we observed $m = 2$ out of $N = 4$

The prior keeps 0.5 above the 0.95 Quantile!

The posterior mean is ≈ 0.3!



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



The posterior mean stays at ≈ 0.25!

We introduced discussing uncertainty from a Bayesian perspective

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

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

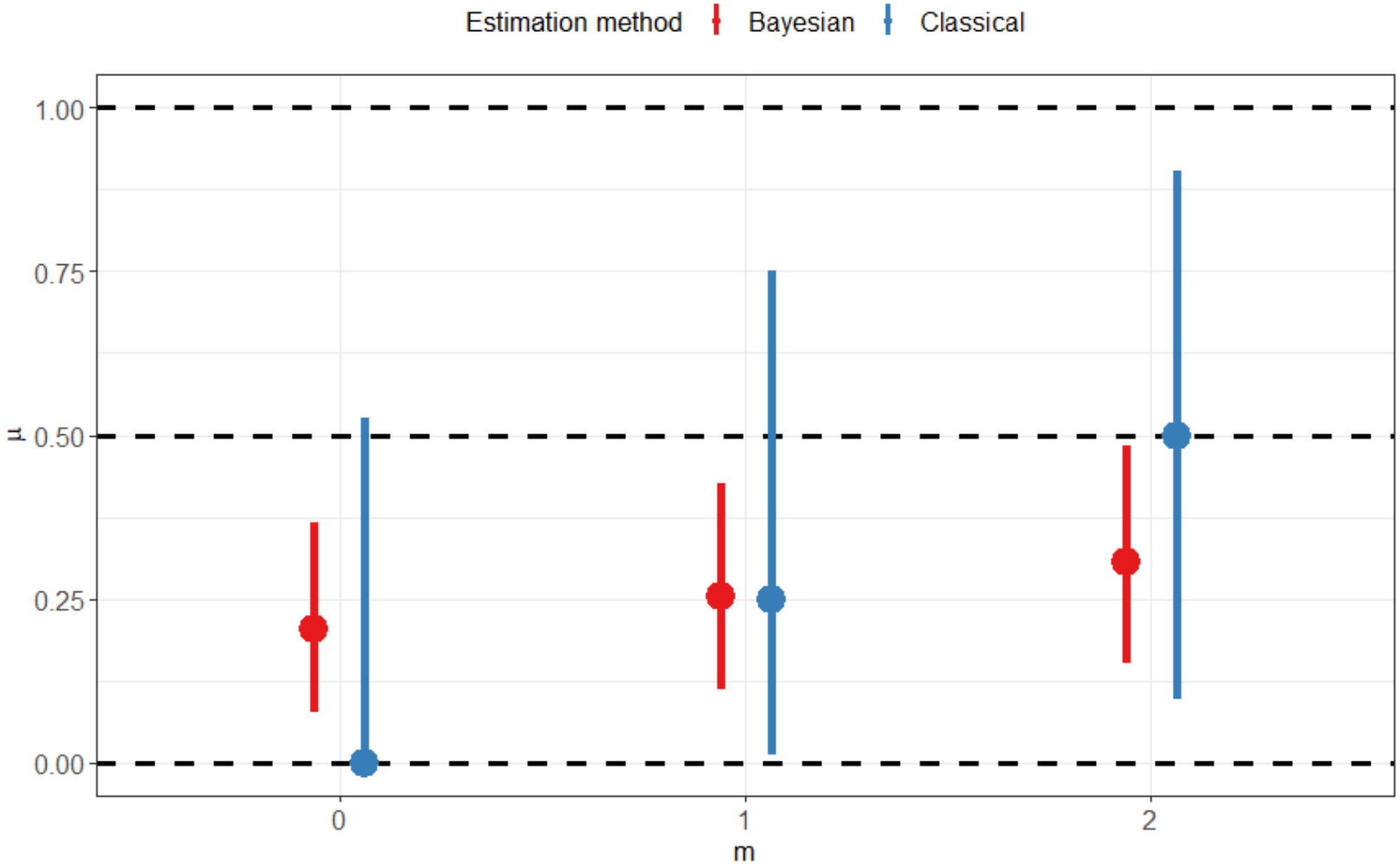
- How do 90% confidence intervals compare with the *posterior credible intervals* in our example?

Confidence interval calculation

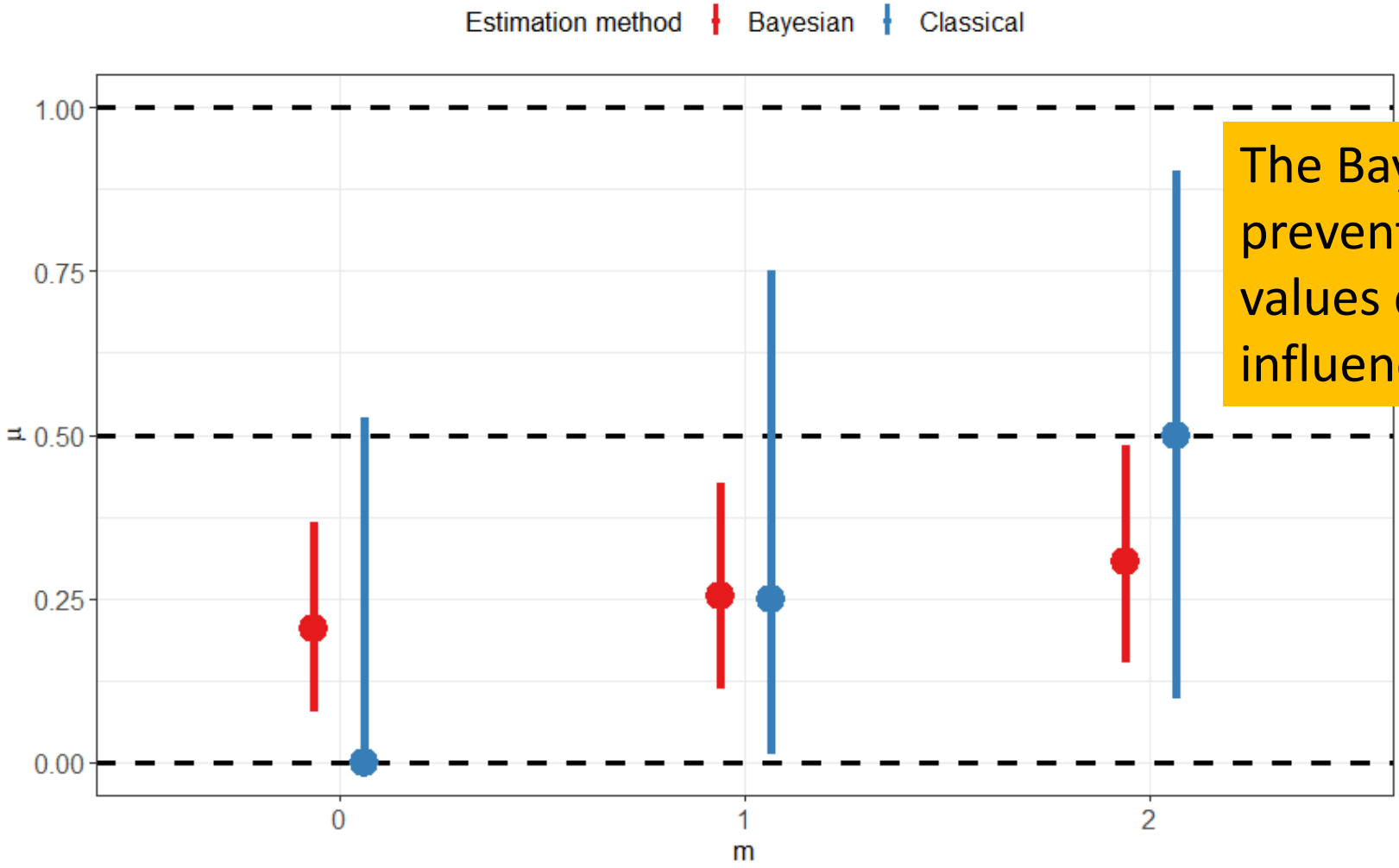
- The 90% confidence intervals are calculating using the Clopper-Pearson method, through R's `binom.test()` function.

- 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)



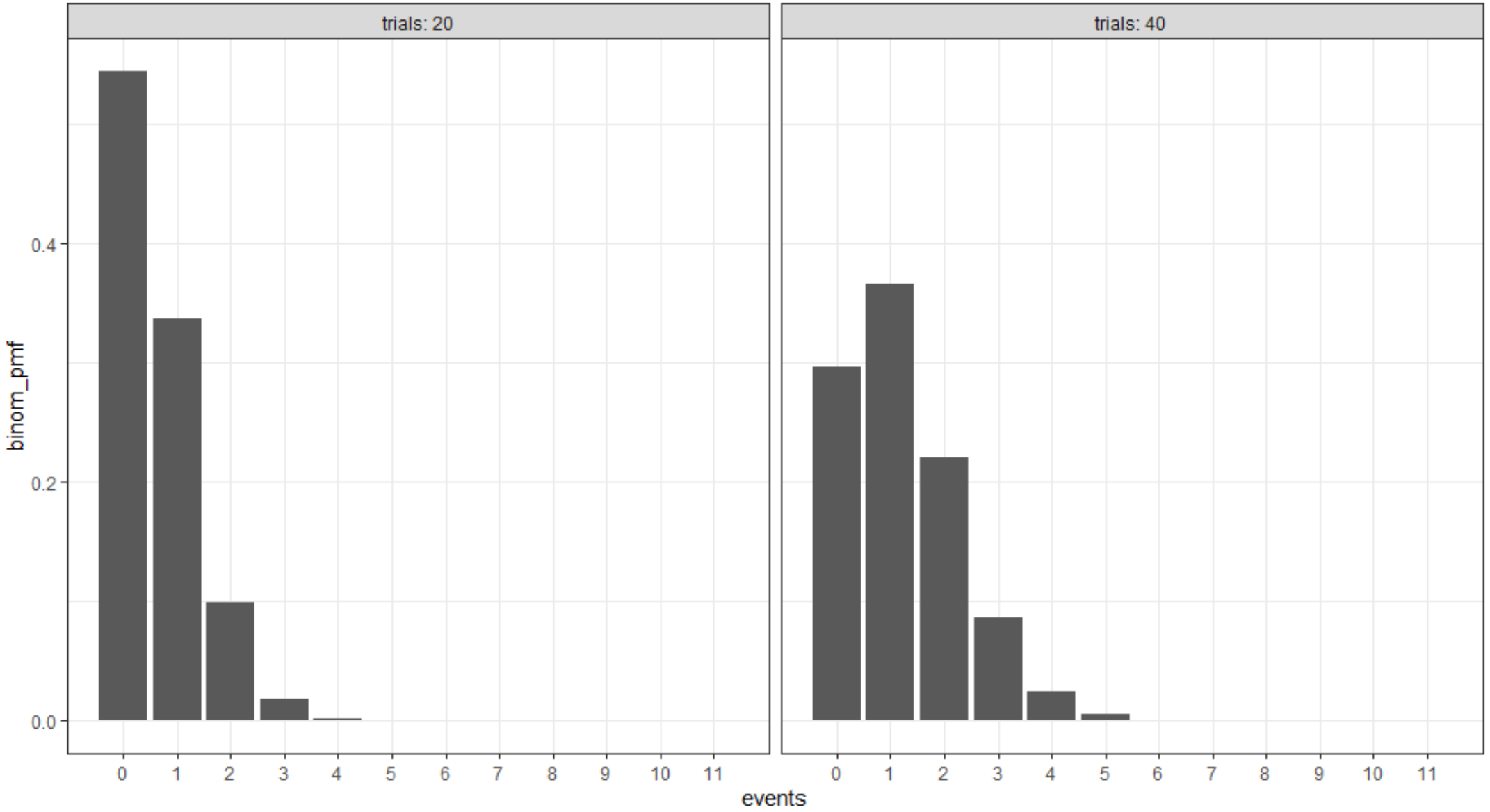
The Bayesian approach prevents unrealistic values due to the influence of the prior!!

Why is the Bayesian approach useful?

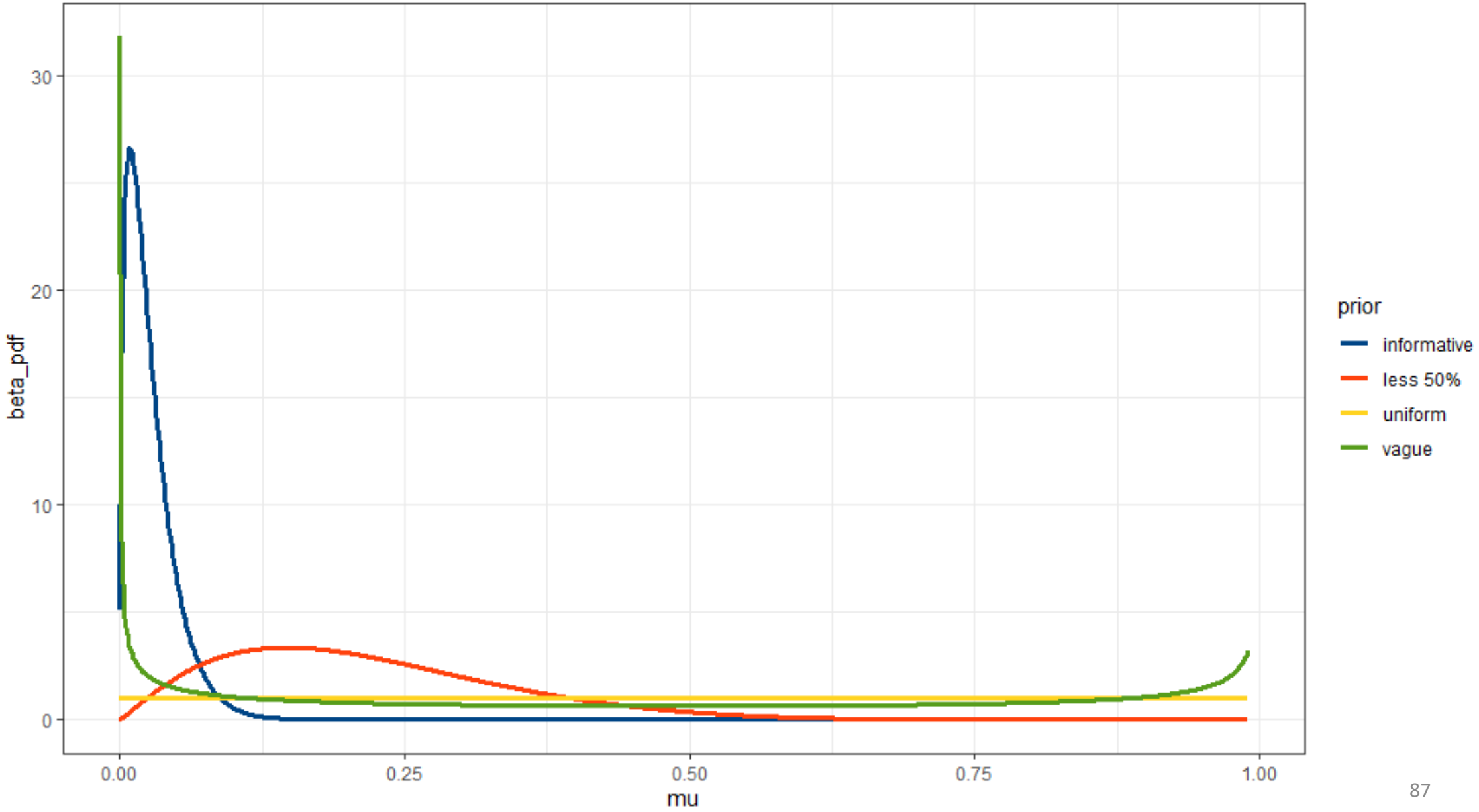
- Consider sampling a population to identify a rare event…

- For example, what if the TRUE event probability of this rare event…is just 3%...

If the TRUE probability is 3%, what's the probability of observing….



Consider 4 different prior distributions

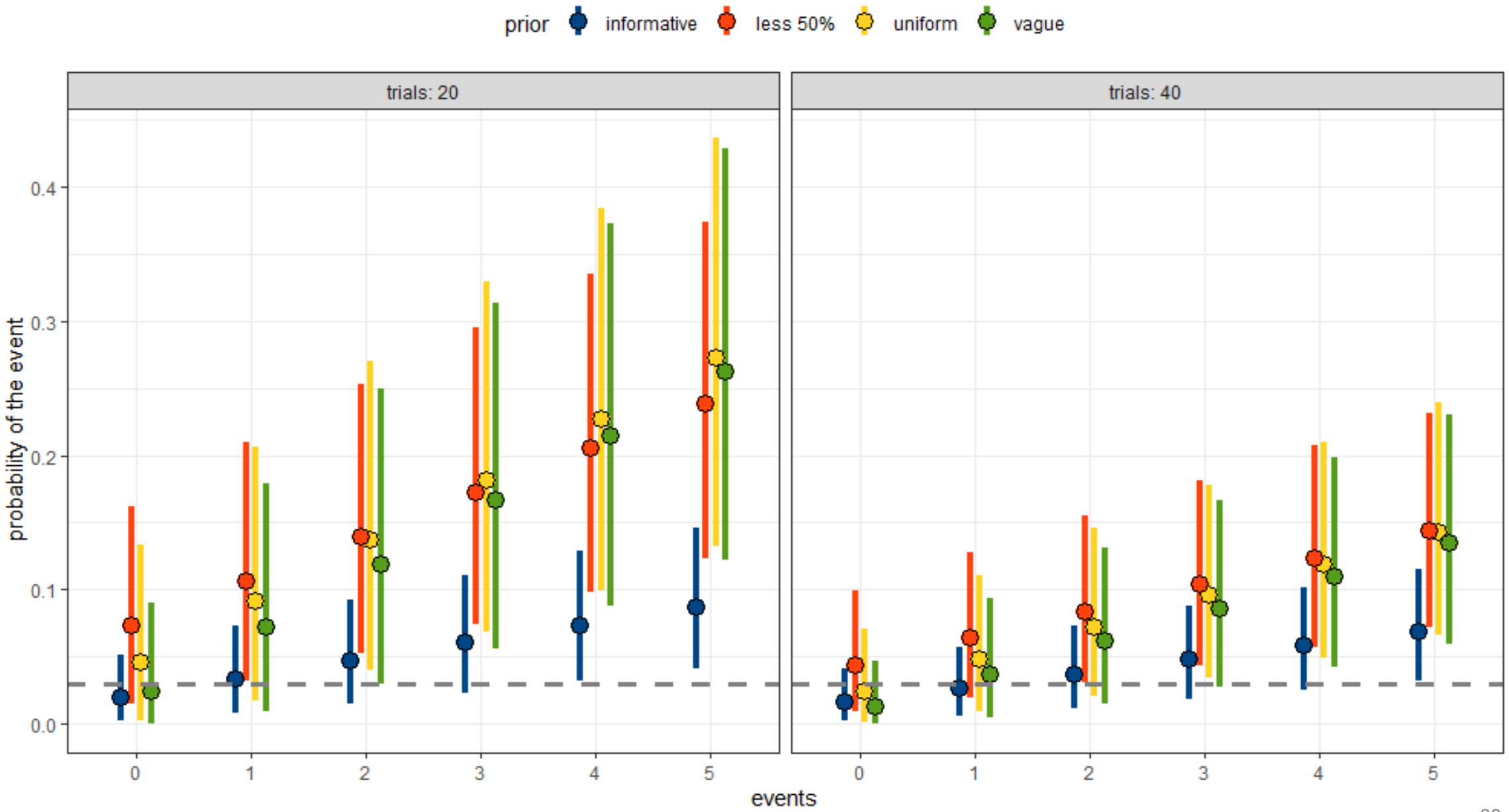


How does the posterior behave under different circumstances?

- Compare the posterior summaries:
  - Middle 90% credible intervals (5^{th} through 95^{th} quantiles)
  - Mean values

- Under different prior specifications, trial size, and observed number of events.

- Interested in understanding how the posterior distribution changes if we would observe a specific number of events out of a specific number of trials


prior
informative
less 50%
uniform
vague
trials: 20
trials: 40
probability of the event
0.0
0.1
0.2
0.3
0.4
0
1
2
3
4
5
events

Summary

Likelihood and Prior distributions for μ, the probability of an event (1) in a binary (0/1) classification task

- The likelihood function is a probability distribution over the unknown parameter μ given a set of training samples
  - Bernoulli: if x is a binary variable, then $p(x \mid \mu) = \mu^x (1-\mu)^{1-x}$
  - Binomial: if m is the number of events and N is the total number of trials, then $p(m \mid N, \mu) = \binom{N}{m} \mu^m (1-\mu) N - m$
- The prior is a distribution over the unknown parameter, given some hyperparameters that encode an initial hypothesis as to the likelihood of each value of μ
  - Beta: for non-negative, real numbers a, b, $p(\mu \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}$
    - a, b can be interpreted as the prior number of events and non-events respectively. e.g. "before I conducted this experiment, I learned prior researchers found a 1s and b 0s"

Summary

The posterior distribution combines the data and the prior. Represents an updated hypothesis about the most likely parameter estimates after observing a new set of training data.

- **The Beta distribution is the conjugate prior of Binomial, Bernoulli, and Beta likelihoods. This means that when we calculate the product between the likelihood and the prior, the result can be rewritten as a Beta distribution itself.**

- **The posterior is proportional to the product of the likelihood and the prior. The Bayesian Evidence is a normalizing constant, so we tend to ignore it for optimization which is why I say "proportional to"**

$$p(\mu \mid m, N, a, b) \propto \left[\prod_{n=1}^{N} \texttt{Bernoulli}(\texttt{x}_\texttt{n} \mid \mu)\right] \times \texttt{Beta}(\mu \mid a, b)$$

$$\propto \texttt{Binomial}(m \mid N, \mu) \times \texttt{Beta}(\mu \mid a, b)$$

$$\propto \texttt{Beta}(\mu \mid a + m, b + N - m)$$

- **Compared to the MLE estimate, the MAP takes into consideration not only which value of μ best matches the data, but how likely each value of μ is given information you collected prior to analyzing the present data**