CS1675/2075: Introduction to Machine Learning

Week 3: Probability Recap and Distribution Fitting

Fall 2024

Probability intro

Bayesian statistics play an important role in many machine learning algorithms

- Based on Bayes Theorem - ‘best’ parameter values are those with the highest probability *given* the training data (input-output pairs)

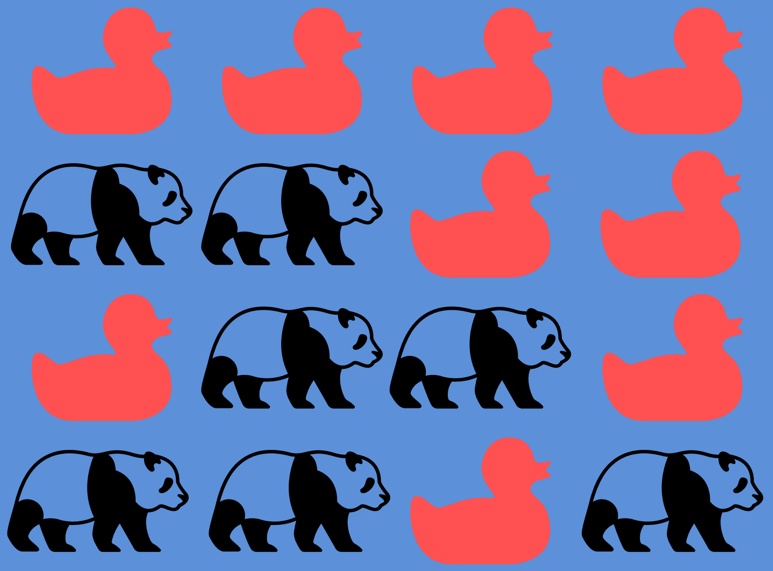
- Parameters are described by probability distributions, giving both mean and variance estimates. Ordinary Least Squares regression, by contrast, only gives the best coefficient values according to MSE and uncertainty must be calculated *post-hoc*

- Prior distributions let us incorporate knowledge about the data that is known prior to model training

- In my opinion, a Bayesian approach to ML gives a general, robust intuition towards the data science process.

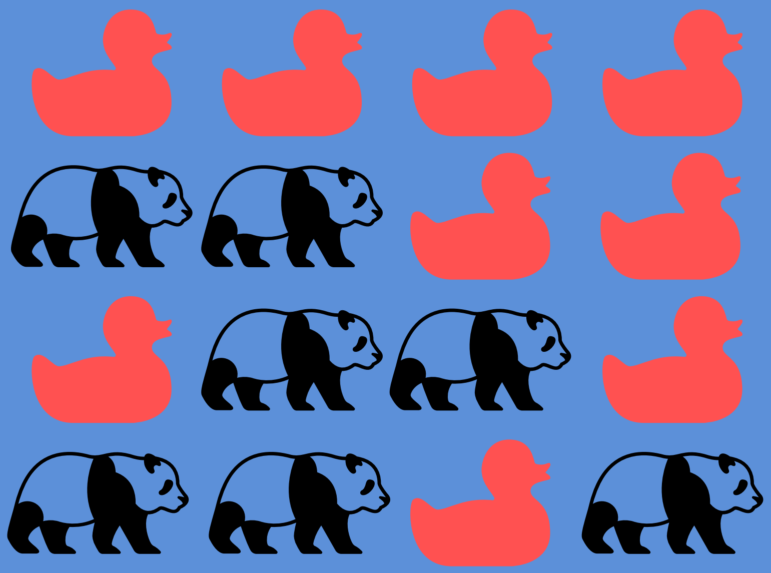
Probability Basics

- The probability of an event is the proportion of times an event occurs out of the total number of trials



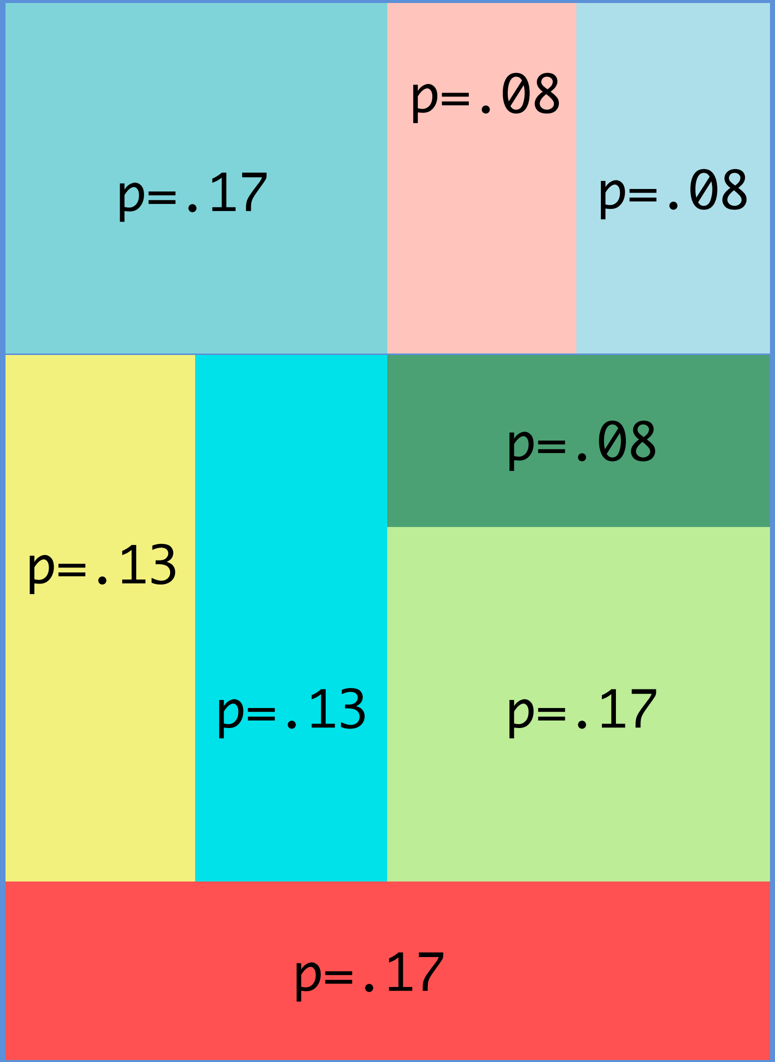
Probability Basics

- The probability of an event is the proportion of times an event occurs out of the total number of trials



- Let 🦆 be the "event" class, with probability $p(y = 1)$
- 🐼 is then the "non-event", $p(y = 0)$
- The *sample probability* of each class is the number of times it appears in the set, divided by its total size

Continuous distributions: integral from negative to positive infinity is equal to 1



p=.17

p=.08

p=.08

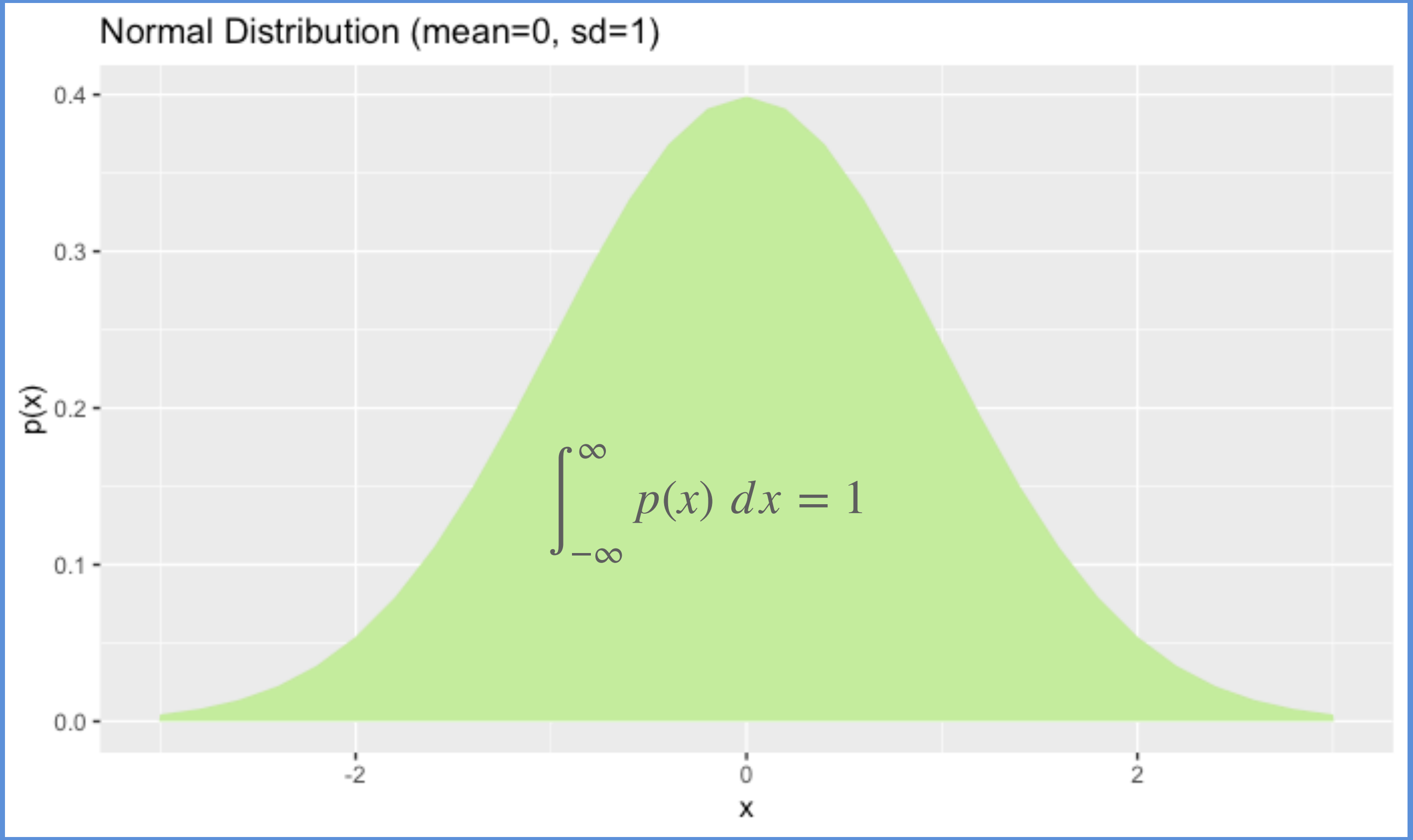
p=.13

p=.08

p=.13

p=.17

p=.17

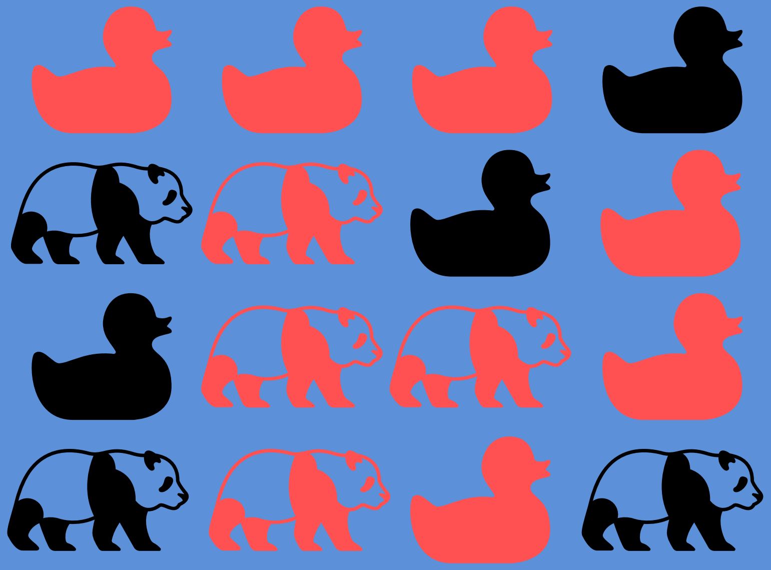


$$\int_{-\infty}^{\infty} p(x)\, dx = 1$$

Joint Probabilities

- Now, consider two variables: animal and color. A random sample from this set will have one of four combinations of features

- The joint probability of two (or more) variables is the probability of both variables taking particular values



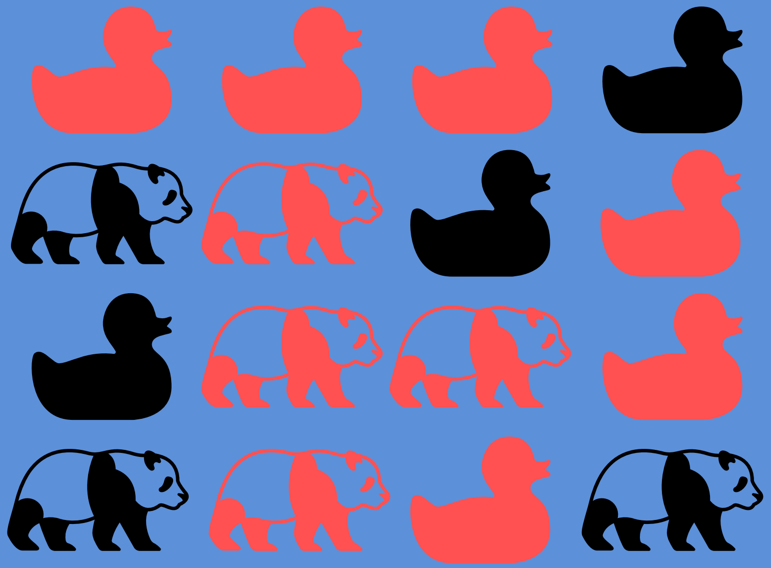
Counts	Red	Black
Panda	4	3
Duck	6	3

Joint Probabilities

- The joint probability between two (or more) variables is the proportion of all trials in which they take a specific combination of values

- *e.g.* if I randomly select one of the objects below, what are the odds that it is *both* red *and* a panda?

- For 2 variables with 2 possible values each, there are 4 combinations that can be made



Counts	Red	Black
Panda	4	3
Duck	6	3

- $p(P, R) = 4/16$
- $p(P, B) = 3/16$
- $p(D, R) = 6/16$
- $p(D, B) = 3/16$

Joint Probabilities - Sum Rule

Counts	Red	Black
Panda	4	3
Duck	6	3

- $p(P) = p(P, R) + p(P, B) = 7/16$
- $p(D) = p(D, R) + p(D, B) = 9/16$

Counts	Red	Black
Panda	4	3
Duck	6	3

- $p(R) = p(R, P) + p(R, D) = 10/16$
- $p(B) = p(B, P) + p(B, D) = 6/16$

The marginal probability of a variable can be calculated by summing across its intersections with all levels of a second variable

Joint Probabilities - Sum Rule

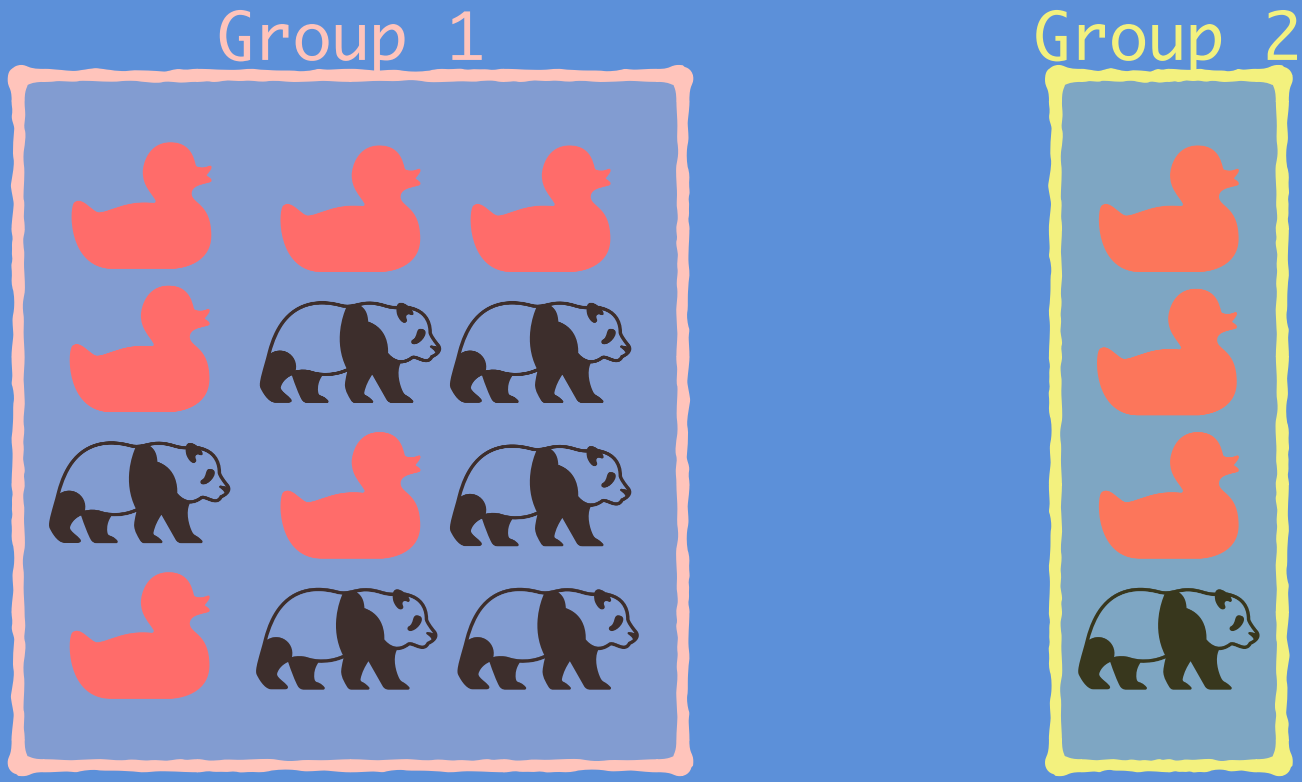
- For two random variables A and B, the total probability of A is equal to some of its joint probabilities across all values of B.
  - $p(A = a) = \sum_{B} p(A = a, B = b)$, discrete B
  - $p(A = a) = \int_{-\infty}^{\infty} p(a, b)\ db$, continuous B

Joint Probabilities - Independent Events

- If A and B are independent events - that is, observing information about A does not affect the probability of B - then the joint probability is the product of the marginal probabilities
    - $p(A, B) = p(A)p(B)$ **if and only if** A and B are independent
- Examples include:
    - Repeated random samples (e.g. if you flip a coin twice, the outcome of the second flip is 50/50 heads/tails regardless of the first coin)
    - Unrelated events (e.g. probability that a coin flip is heads **and** it is not raining outside)

Conditional Probability

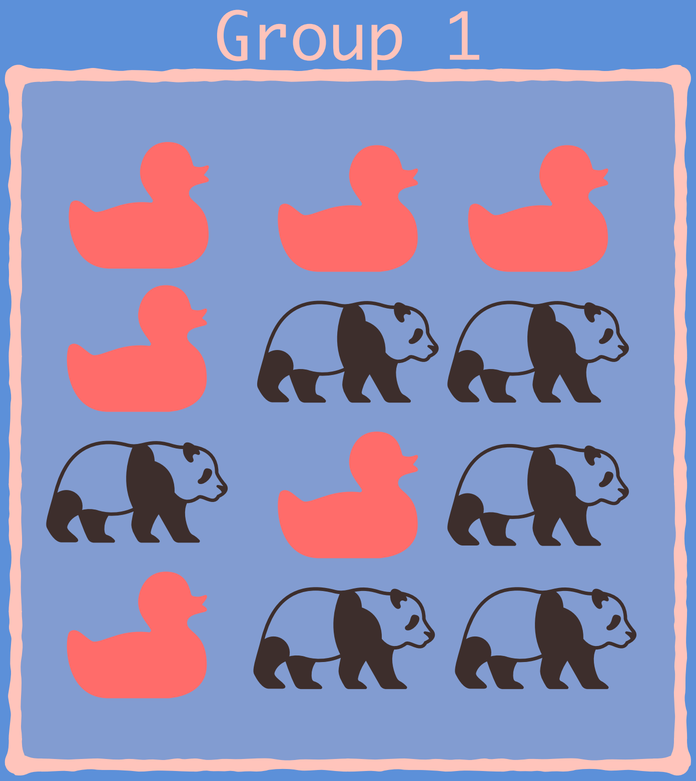
- Let us split the samples into two separate groups
- Now the probability of randomly selecting a panda or duck is dependent on whether you choose from Group 1 or Group 2



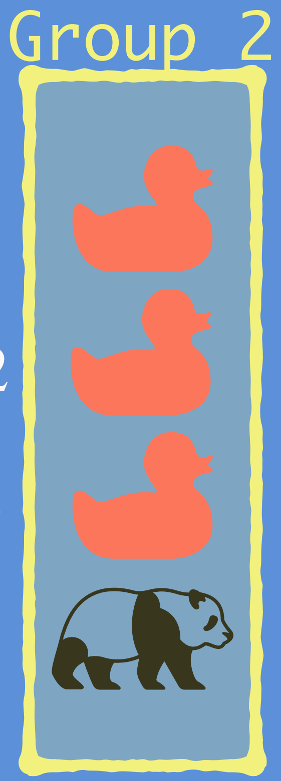
Conditional Probability

- What are the probabilities of each animal **conditioned on** either Group 1 or Group 2?

- What are the odds I randomly select a duck, **given that** I am choosing from Group 1?



- $p(D \mid G = 1) = 6/12$
- $p(P \mid G = 1) = 6/12$



- $p(D \mid G = 2) = 3/4$
- $p(P \mid G = 2) = 1/4$

Conditional Probability

Marginal counts	Group 1	Group 2
Panda	6	1
Duck	6	3

- $p(A = P \,|\, G = 1) = \dfrac{p(P{,}1)}{p(P{,}1) + (D{,}1)} = \dfrac{p(P{,}1)}{p(1)}$
- $p(A \,|\, G) = \dfrac{p(A, G)}{p(G)}$
- $p(A, G) = p(A \,|\, G)P(G)$ - chain rule

Types of probabilities

- Marginal: probability of an event A occurring, irrespective of any other variables
- Joint: probability of two events A and B occurring together
- Conditional: probability of A given that B has *already occurred*

Bayesian Theorem

Joint symmetry: $p(A, B) = p(B, A)$

Applying the chain-rule: $p(A \mid B)P(B) = p(B \mid A)p(A)$

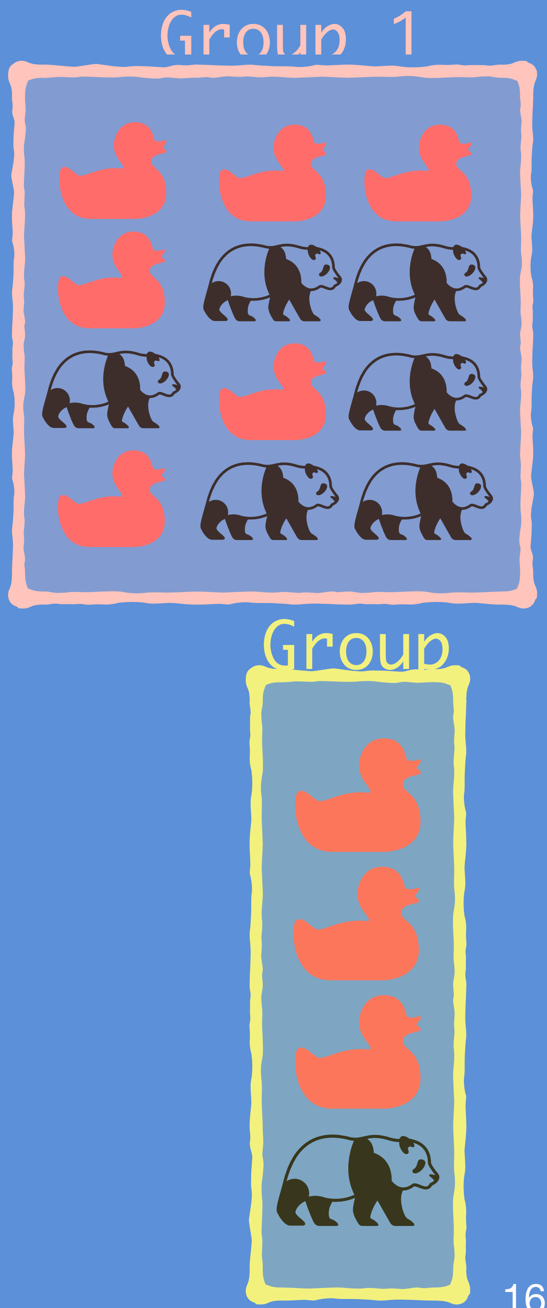
Rearranging: $p(A \mid B) = \dfrac{p(B \mid A)p(A)}{p(B)}$

ML Formulation

- Suppose we randomly sample an animal from one of the two groups. What is the probability that the sample came from Group 2, *given* that it was a duck? $p(1 \mid D)$

- The group is the **response** and the animal type is the **input**

- $p(y = 1 \mid x = D) = \dfrac{p(D \mid 1)p(1)}{p(D)} = \dfrac{1/2 \cdot 3/4}{9/16} = 2/3$



We give special names to each component of Bayes Theorem

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

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

The posterior represents updated beliefs about the prior based on new evidence

Classic example: diagnostic test for a disease

- Let $Dis, Test \in \{0,1\}$ be binary variables representing diseases state, and test result respectively

- 95% of people who have the disease test positive
  $p(Test = 1 \,|\, Dis = 1) = 0.95$, and there is a false-positive rate of 1%

- The rate of the disease in the population is 1 in 100,000
  $p(Dis) = .00001$

- What is the probability that a person has the disease, **given that** they test positive? $p(Dis = 1 \,|\, Test = 1)$

Classic example: diagnostic test for a disease

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

- Likelihood: probability of the evidence given the outcome (how likely is a pos/neg test given that one does/does not have the disease)
- Prior: probability of the outcome, irrespective of any evidence (how common is the disease, not considering the test at all)
- Evidence: total probability of the evidence, summed across all outcome conditions (how likely is a positive test, regardless of whether they have the disease or not)
- Posterior: probability of the outcome given the evidence (likelihood of disease state after observing the results of the test)

Classic example: diagnostic test for a disease

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

- Likelihood: $p(Test = 1 \,|\, Dis = 1) = 0.95$
- Prior: $p(Dis = 1) = 0.00001$
- Evidence:
  - $p(Test = 1) = p(Test = 1 \,|\, Dis = 1)p(Dis = 1) + p(Test = 1 \,|\, Dis = 0)p(Dis = 0)$
  - $p(Test = 1) = 0.95 * 0.00001 + 0.01 * 0.9999 = 0.01$
- Posterior:$p(Dis = 1 \,|\, Test = 1) = \dfrac{0.95 * 0.00001}{0.01} = 0.001$

Bernoulli Distribution

Let’s start to use the rules and concepts of probability to describe behavior

- We are interested in determining the PROBABILITY of an EVENT.

- We are not training a PREDICTIVE MODEL just yet…

- We simply collect observations of an EVENT.

For example, let’s say we wish to ask someone did they like **Star Wars Episode VIII The Last Jedi**?

- ~~I did!!!~~

We will use an encoding similar to that discussed with classification

- We will encode the general variable x to represent a person’s response.

- If a person liked the movie set $x = 1$
- If a person did NOT like the movie set $x = 0$

The PROBABILITY someone liked the movie, $x = 1$, will be denoted by the parameter μ

- The above statement can be written as:

$$p(x = 1|\mu) = \mu$$

The PROBABILITY someone liked the movie, $x = 1$, will be denoted by the parameter μ

- The above statement can be written as:

$$p(x = 1|\mu) = \mu$$

- Since we have either $x = 1$ or $x = 0$, the probability that $x = 0$ is:

$$p(x = 0|\mu) = 1 - \mu$$

The probability mass function over possible outcomes can be more compactly written as:

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

Referred to as the Bernoulli distribution after Jacob Bernoulli

Note: μ is a probability and so is bounded: $0 \leq \mu \leq 1$

The **probability mass function** over possible outcomes can be more compactly written as:

Check when $x = 1$

$$p(x = 1|\mu) = \mu^1(1 - \mu)^0 = \mu \times 1 = \mu$$

Referred to as the Bernoulli distribution after Jacob Bernoulli

Note: μ is a probability and so is bounded: $0 \leq \mu \leq 1$

The probability mass function over possible outcomes can be more compactly written as:

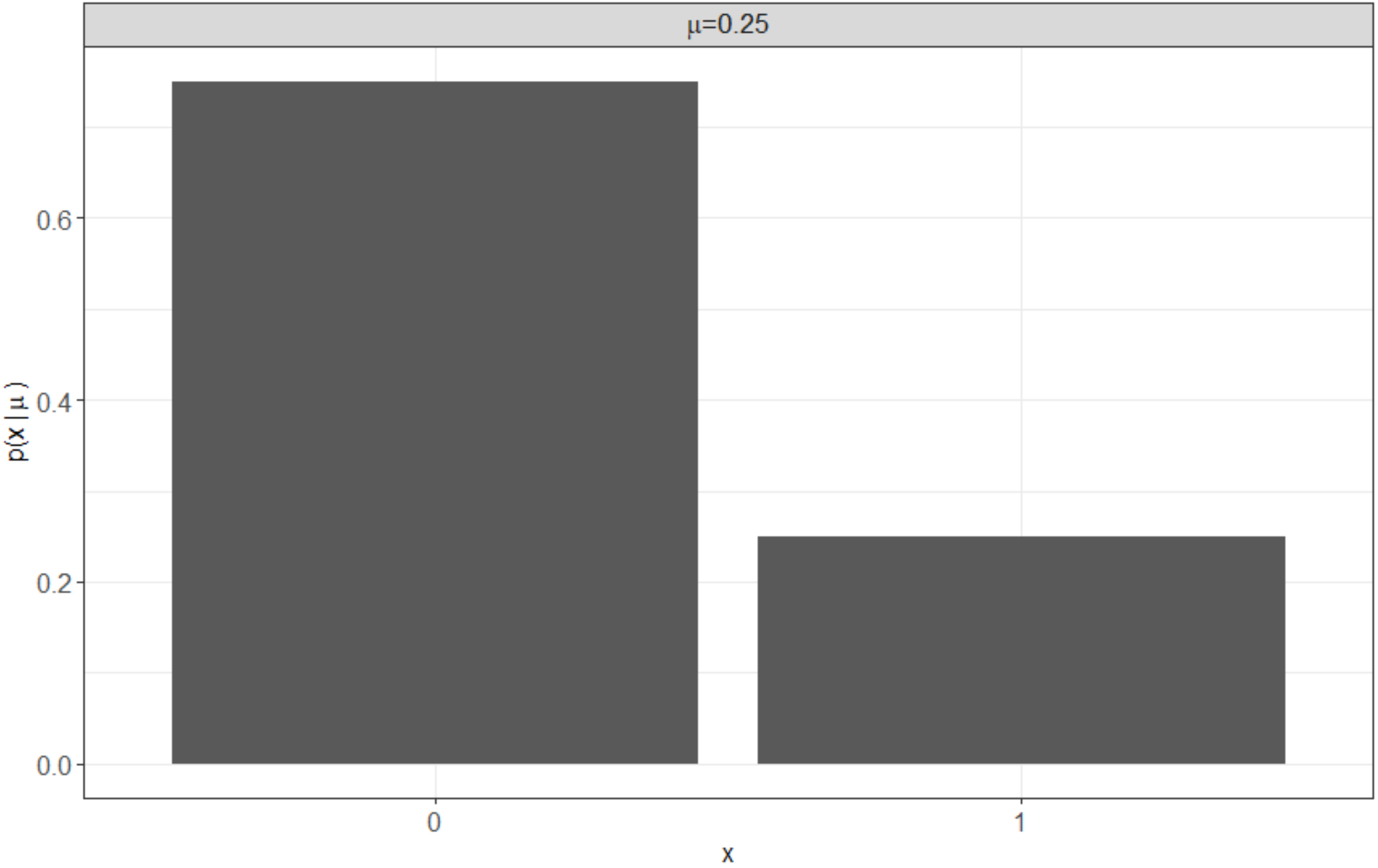
Check when $x = 0$

$$p(x = 0|\mu) = \mu^0(1 - \mu)^1 = 1 \times (1 - \mu) = 1 - \mu$$

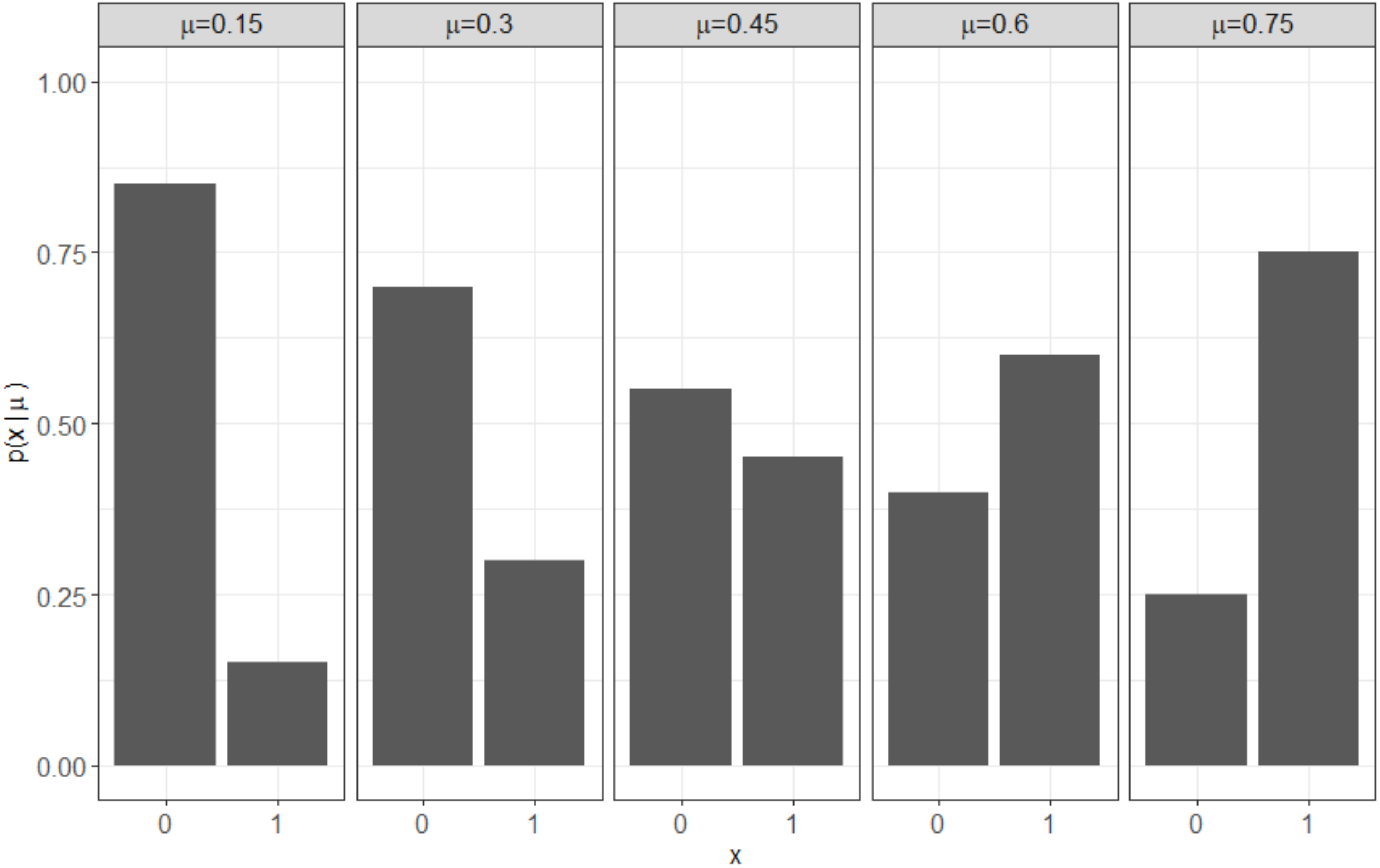
Referred to as the Bernoulli distribution after Jacob Bernoulli

Note: μ is a probability and so is bounded: $0 \leq \mu \leq 1$

The Bernoulli distribution PMF for a single μ



Bernoulli PMF for multiple μ values



The Bernoulli distribution can be used to represent many different real-life applications

- Today we introduced the idea in terms of liking a movie.
- Typically, the Bernoulli distribution is introduced with the canonical **coin flip** example.
- Applicable in sports, engineering, manufacturing, medicine, marketing, etc…

The Bernoulli distribution can be used to represent many different real-life applications

- Today we introduced the idea in terms of liking a movie.
- Typically, the Bernoulli distribution is introduced with the canonical **coin flip** example.
- Applicable in sports, engineering, manufacturing, medicine, marketing, etc…
- The Bernoulli distribution is applicable for many **BINARY OUTCOME** situations.

Back to our movie example...

- Suppose we ask 4 random people if they liked the movie:

Person	Response
1	No
2	No
3	Yes
4	No

In terms of the **encoded** variable x:

Person	Response	x
1	No	0
2	No	0
3	Yes	1
4	No	0

What is the probability of the sequence we observed?

- Start with the joint distribution:

$$p(x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 0|\mu)$$

- This looks rather difficult to work with…
- How can we evaluate this **JOINT** probability?

Independence

- Suppose we have two random variables, A and B.

- Their **JOINT** probability is: $p(A, B)$

- If the two random variables are independent, we can ***factor*** their **JOINT** probability as the product of their **MARGINALS**:

$$p(A, B) = p(A)p(B)$$

Independence and the PRODUCT rule

$$p(A, B) = p(A \mid B)p(B)$$

- If A and B are independent, they are unrelated to each other!

- Knowing something about B does not tell us anything about A!

$$p(A \mid B) = p(A)$$

- Product rule simplifies to: $p(A, B) = p(A \mid B)p(B) = p(A)p(B)$

Back to our problem of asking 4 people…

- Assuming the people are **INDEPENDENT** we can **FACTOR** the JOINT distribution!

$$p(x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 0|\mu) = p(x_1 = 0 \mid \mu)p(x_2 = 0 \mid \mu)\ p(x_3 = 1 \mid \mu)\ p(x_4 = 0 \mid \mu)$$

The complicated looking JOINT distribution has become the product of 4 easier distributions!

Each distribution on the Right Hand Side (RHS) is a **BERNOULLI**!

Substitute in the expression from the Bernoulli PMF for each person

$$p(x = 0 \mid \mu) = (1 - \mu)$$

Substitute in the expression from the Bernoulli PMF for each person

$$p(x = 1 \mid \mu) = \mu$$

Substitute in the expression from the Bernoulli PMF for each person

$$p(x_1 = 0 \mid \mu) p(x_2 = 0 \mid \mu)\, p(x_3 = 1 \mid \mu)\, p(x_4 = 0 \mid \mu) = (1 - \mu) \times (1 - \mu) \times \mu \times (1 - \mu)$$

Now generalize from asking 4 people, to asking N people

- Denote the sequence of responses as: $\mathbf{x} = \{x_1, x_2, \dots, x_n, \dots, x_N\}$

- Assuming each person is **independent** we can factor the joint distribution into the product of N independent distributions:

$$p(\mathbf{x}|\mu) = \prod_{n=1}^{N} (\mu^{x_n}(1-\mu)^{1-x_n}) = \prod_{n=1}^{N} \left(\mathrm{Bernoulli}(x_n|\mu)\right)$$

Likelihood Function!!!

What happens when μ is unknown?

- If we have the observed sequence $\mathbf{x} = \{x_1, x_2, \dots, x_n, \dots, x_N\}$ we can estimate μ from the data.

- Which value of μ is best?

- The value that **MAXIMIZES** the likelihood!

$$\hat{\mu} = \mu_{MLE} = \underset{\mu \in [0,1]}{\operatorname{argmax}}\, p(\mathbf{x}|\mu)$$

Rather than working with the likelihood directly…we will maximize the *log-likelihood*

- The log-likelihood for the N independent observations:

$$\log[p(\mathbf{x} \mid \mu)] = \log\left(\prod_{n=1}^{N} (\mu^{x_n}(1-\mu)^{1-x_n})\right)$$

Rather than working with the likelihood directly…we will maximize the *log-likelihood*

- PRODUCT of N indepedent likelihoods becomes the SUMMATION of N independent log-likelihoods:

$$\log[p(\mathbf{x} \mid \mu)] = \sum_{n=1}^{N} \left(\log\left[\mu^{x_n}(1-\mu)^{(1-x_n)}\right]\right)$$

Rather than working with the likelihood directly…we will maximize the *log-likelihood*

- Use the properties of the natural log to further simplify the expression:

$$\log[p(\mathbf{x}|\mu)] = \sum_{n=1}^{N}\left(\log[\mu^{x_n}] + \log\left[(1-\mu)^{(1-x_n)}\right]\right)$$

Rather than working with the likelihood directly…we will maximize the *log-likelihood*

- Use the properties of the natural log to further simplify the expression:

$$\log[p(\mathbf{x}|\mu)] = \sum_{n=1}^{N} (x_n \log[\mu] + (1 - x_n)\log[1 - \mu])$$

Rearrange the log-likelihood

$$\log[p(\mathbf{x} \mid \mu)] = \sum_{n=1}^{N}(x_n \log[\mu]) + \sum_{n=1}^{N}\big((1 - x_n) \log[1 - \mu]\big)$$

We are ASSUMING the event probability is CONSTANT.
Thus, $\log[\mu]$ and $\log[1 - \mu]$ do NOT depend on the observation!!

Both terms can be pulled in front of their respective summation series.

PLEASE NOTE we will work with changing event probabilities later in the semester!

Rearrange the log-likelihood

$$\log[p(\mathbf{x}|\mu)] = \log[\mu] \sum_{n=1}^{N} \{x_n\} + \log[1-\mu] \sum_{n=1}^{N} \{1 - x_n\}$$

Rearrange the log-likelihood

$$\log[p(\mathbf{x}|\mu)] = \log[\mu] \underbrace{\sum_{n=1}^{N} \{x_n\}}_{\text{Number of people that liked the movie, or more generally number of events } M.} + \log[1-\mu] \underbrace{\sum_{n=1}^{N} \{1 - x_n\}}_{\text{Number of people that did NOT like the movie, or more generally number of times we \textbf{did not} observe the event } N - M.}$$

Rearrange the log-likelihood

$$\boxed{\log[p(\mathbf{x}|\mu)] = \log[\mu] \times M + \log[1 - \mu] \times (N - M)}$$

To optimize, calculate the derivative of the log-likelihood with respect to $\boldsymbol{\mu}$

$$\frac{\partial}{\partial\mu}\{\log[p(\mathbf{x}|\mu)]\} = \underbrace{\frac{\partial}{\partial\mu}\{\log[\mu]\times M\}}_{\displaystyle\frac{M}{\mu}} + \underbrace{\frac{\partial}{\partial\mu}\{\log[1-\mu]\times(N-M)\}}_{\displaystyle\frac{-(N-M)}{1-\mu}}$$

Set the derivative equal to zero and solve for μ_{MLE}

$$\frac{\partial}{\partial\mu}\{\log[p(\mathbf{x}|\mu)]\} = 0 = \frac{M}{\mu_{MLE}} - \frac{N-M}{1-\mu_{MLE}}$$

$$\Downarrow$$

$$\frac{(1-\mu_{MLE}) \times M - \mu_{MLE} \times (N-M)}{\mu_{MLE} \times (1-\mu_{MLE})} = 0$$

$$\Downarrow$$

$$M - \mu_{MLE}M - \mu_{MLE}N + \mu_{MLE}M = 0 \rightarrow M - \mu_{MLE}N = 0$$

The maximum likelihood estimate (MLE) for μ is just based on counting!

$$\boxed{\mu_{MLE} = \frac{M}{N} = \frac{1}{N}\sum_{n=1}^{N}\{x_n\}}$$

Binomial Distribution

Earlier, we introduced the Bernoulli distribution

$$\boxed{p(x \mid \mu) = \mathrm{Bernoulli}(x \mid \mu) = \mu^{x}(1-\mu)^{1-x}}$$

- 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, $\mathbf{x} = \{x_1, x_2, \ldots, x_n, \ldots, x_N\}$

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

- The MLE for the probability of the event is:

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

Let's ask a different question…

- Instead of asking, what is the probability $x = 1$ (the EVENT)…

- Let's ask, what is the probability the event occurs a **<u>specific number of times out of a specific number of trials</u>**?

In terms of our movie example…

- What's the probability of finding **exactly 1 out of 4 people that liked the movie**?

Wait…didn’t we calculate this already?

- Based on the following independent observations:

Person	Response	x
1	No	0
2	No	0
3	Yes	1
4	No	0

Wait…didn’t we calculate this already?

- Based on the following independent observations:

Person	Response	$\boldsymbol{x}$	$\boldsymbol{p(x\|\mu)}$
1	No	0	$(1-\mu)$
2	No	0	$(1-\mu)$
3	Yes	1	μ
4	No	0	$(1-\mu)$

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

Wait…didn’t we calculate this already?

- Based on the following independent observations:

Person	**Response**	$\boldsymbol{x}$	$\boldsymbol{p(x\|\mu)}$
4	No	0	$(1-\mu)$

Is this the only way to observe 1 Yes out of 4 people?

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

No! Multiple **potential** sequences of 4 people consisting of exactly 1 Yes.

Sequence	Person 1	Person 2	Person 3	Person 4
1	Yes	No	No	No
2	No	Yes	No	No
3	No	No	Yes	No
4	No	No	No	Yes

No! Multiple **potential** sequences of 4 people consisting of exactly 1 Yes.

Sequence	Person 1	Person 2	Person 3	Person 4
1	Yes	No	No	No
2	No	Yes	No	No
3	No	No	Yes	No
4	No	No	No	Yes

The sequence we worked with last time was just 1 out of 4 possible sequences for observing 1 event out of 4 trials!

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

Sequence	x_1	x_2	x_3	x_4
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1

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

Sequence	$p(x_1\|\mu)$	$p(x_2\|\mu)$	$p(x_3\|\mu)$	$p(x_4\|\mu)$
1	μ	$(1-\mu)$	$(1-\mu)$	$(1-\mu)$
2	$(1-\mu)$	μ	$(1-\mu)$	$(1-\mu)$
3	$(1-\mu)$	$(1-\mu)$	μ	$(1-\mu)$
4	$(1-\mu)$	$(1-\mu)$	$(1-\mu)$	μ

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

Sequence	$p(x_1 \mid \mu)$	$p(x_2 \mid \mu)$	$p(x_3 \mid \mu)$	$p(x_4 \mid \mu)$
1	μ	$(1-\mu)$	$(1-\mu)$	$(1-\mu)$
2	$(1-\mu)$	μ	$(1-\mu)$	$(1-\mu)$
3	$(1-\mu)$	$(1-\mu)$	μ	$(1-\mu)$
4	$(1-\mu)$	$(1-\mu)$	$(1-\mu)$	μ

The probability of each sequence of observations, $p(\mathbf{x} \mid \mu)$, is the **product** of the 4 probabilities, $\prod_{n=1}^{N}\left(p(x_n \mid \mu)\right)$, because we have assumed independent observations

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

Sequence	$p(\mathbf{x}\vert\mu)$
1	$\mu \times (1-\mu)^3$
2	$\mu \times (1-\mu)^3$
3	$\mu \times (1-\mu)^3$
4	$\mu \times (1-\mu)^3$

The probability of observing exactly 1 Yes out of 4 people:

- **Sum** together the probabilities of each **<u>potential</u>** sequence:

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

- Next, what’s the probability of finding **exactly 2 Yes responses out of 4 people**?

List all **potential** sequences with 2 Yes responses

Sequence	x_1	x_2	x_3	x_4
1	1	1	0	0
2	0	1	1	0
3	0	0	1	1
4	1	0	1	0
5	0	1	0	1
6	1	0	0	1

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

Sequence	x_1	x_2	x_3	x_4
1	μ	μ	$(1-\mu)$	$(1-\mu)$
2	$(1-\mu)$	μ	μ	$(1-\mu)$
3	$(1-\mu)$	$(1-\mu)$	μ	μ
4	μ	$(1-\mu)$	μ	$(1-\mu)$
5	$(1-\mu)$	μ	$(1-\mu)$	μ
6	μ	$(1-\mu)$	$(1-\mu)$	μ

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

<table>
<tr><th>Sequence</th><th>x_1</th><th>x_2</th><th>x_3</th><th>x_4</th></tr>
<tr><td>1</td><td colspan="4">$\mu^2(1-\mu)^2$</td></tr>
<tr><td>2</td><td colspan="4">$\mu^2(1-\mu)^2$</td></tr>
<tr><td>3</td><td colspan="4">$\mu^2(1-\mu)^2$</td></tr>
<tr><td>4</td><td colspan="4">$\mu^2(1-\mu)^2$</td></tr>
<tr><td>5</td><td colspan="4">$\mu^2(1-\mu)^2$</td></tr>
<tr><td>6</td><td colspan="4">$\mu^2(1-\mu)^2$</td></tr>
</table>

The probability of observing exactly 2 Yes responses out of 4 people:

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

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

How many <u>potential</u> sequences exist?

- Assume 4 people (trials).

- A person can be either a Yes or a No (binary outcome).

$$2^4 = 16$$

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

WHAT PATTERNS DO YOU SEE??

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

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.

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

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 $N-m$.

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

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	$\boxed{1} \times \mu^m \times (1-\mu)^{N-m}$
1	$\boxed{4} \times \mu^m \times (1-\mu)^{N-m}$
2	$\boxed{6} \times \mu^m \times (1-\mu)^{N-m}$
3	$\boxed{4} \times \mu^m \times (1-\mu)^{N-m}$
4	$\boxed{1} \times \mu^m \times (1-\mu)^{N-m}$

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:

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

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

Counting combinations

$$\binom{N}{m} = \frac{N!}{m! \cdot (N - m!)}$$

“N choose m” - how many ways can we choose m events out of N total trials? The order or sequence of the events does not matter (consider the $m!$ term in the denominator)

How many unique subsets of size m can be made from the total set N

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!

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

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

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

The `dbinom()` function in `R` calculates the Binomial PMF

`dbinom(x, size, prob)` $\Rightarrow$ `x` $\leftrightarrow m$, `size` $\leftrightarrow N$, `prob` $\leftrightarrow \mu$

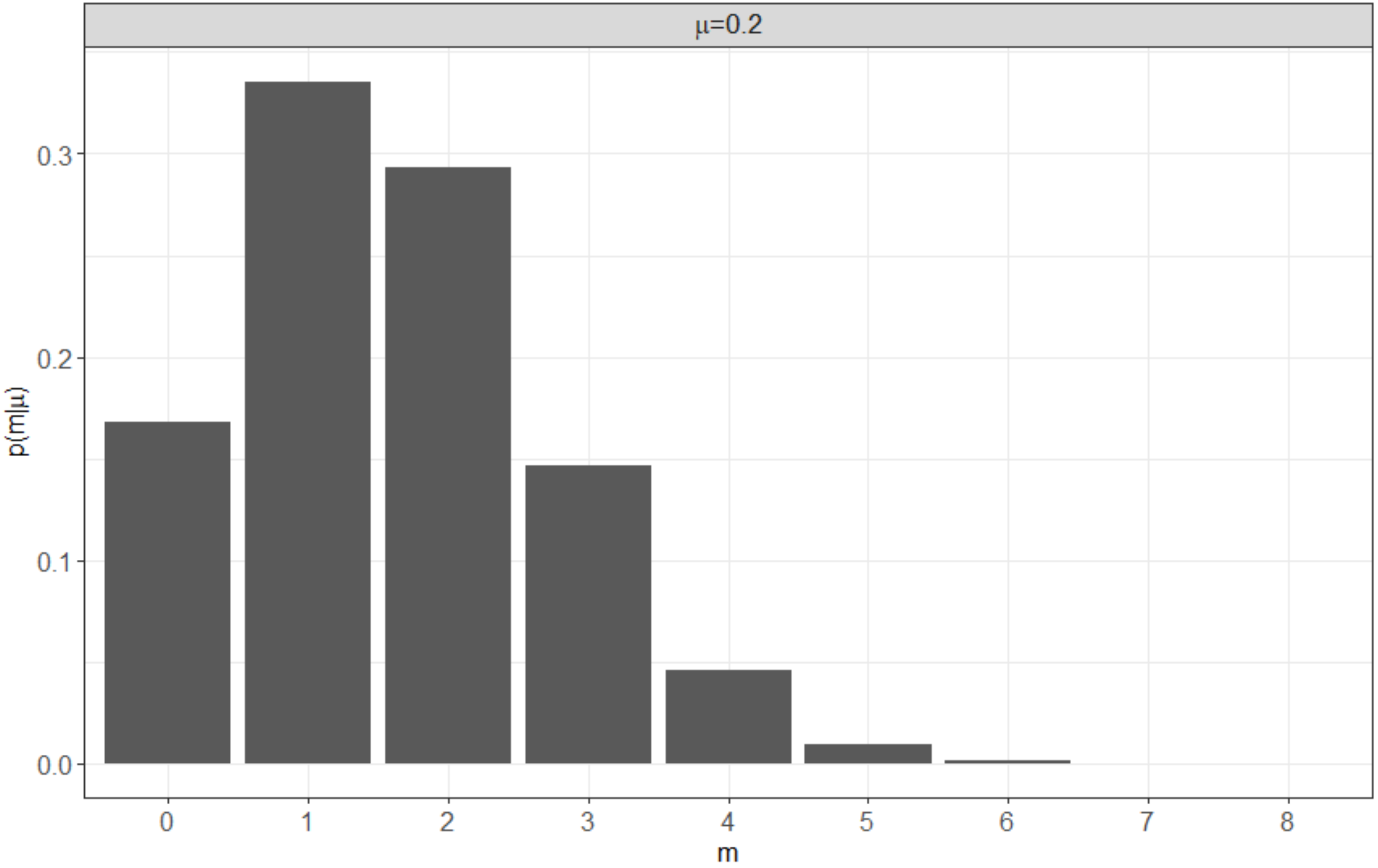
We derived the Binomial distribution starting from Bernoulli observations

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

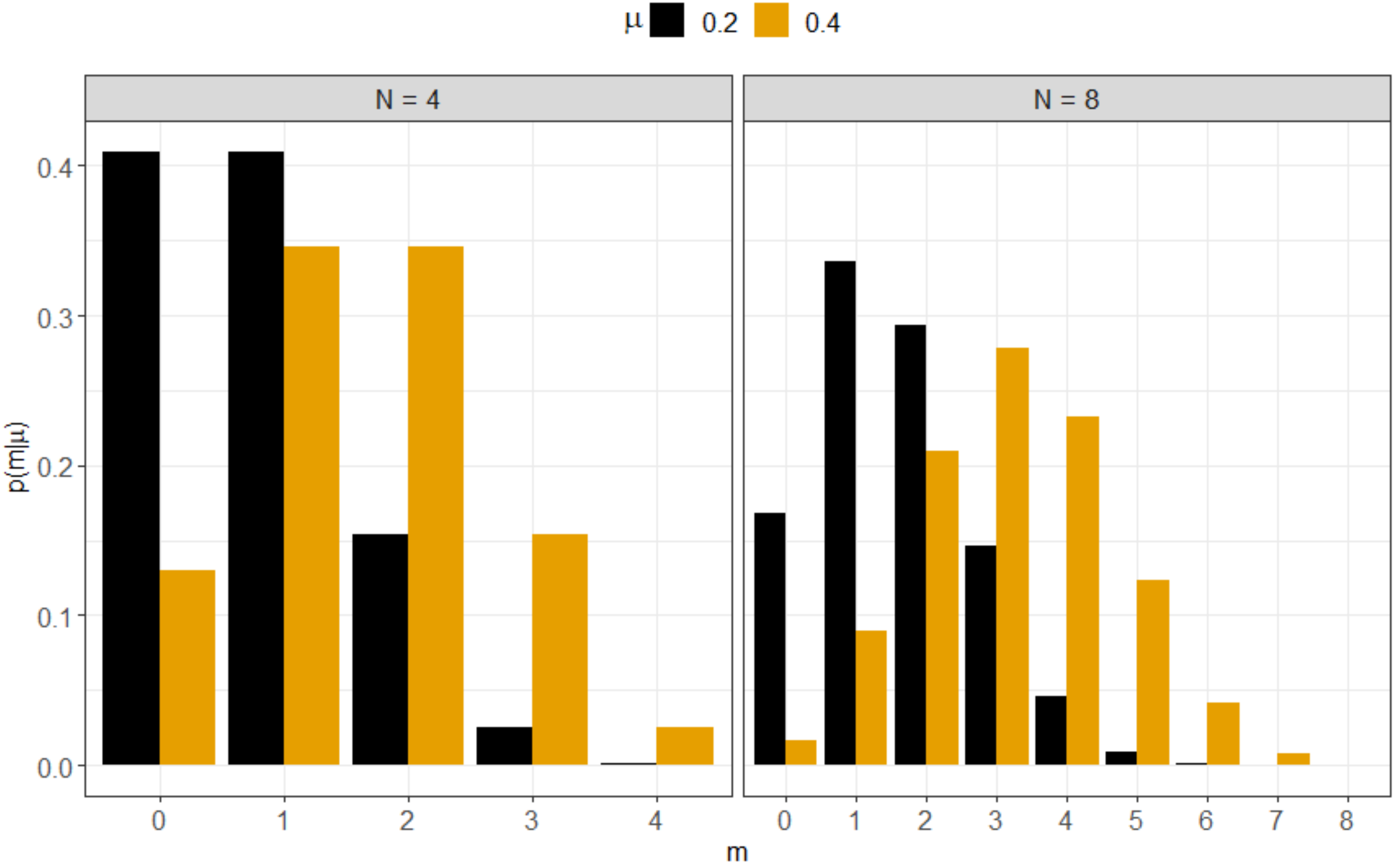
- We recover the Bernoulli distribution with $N = 1$. Thus, $m = \{0,1\}$.

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

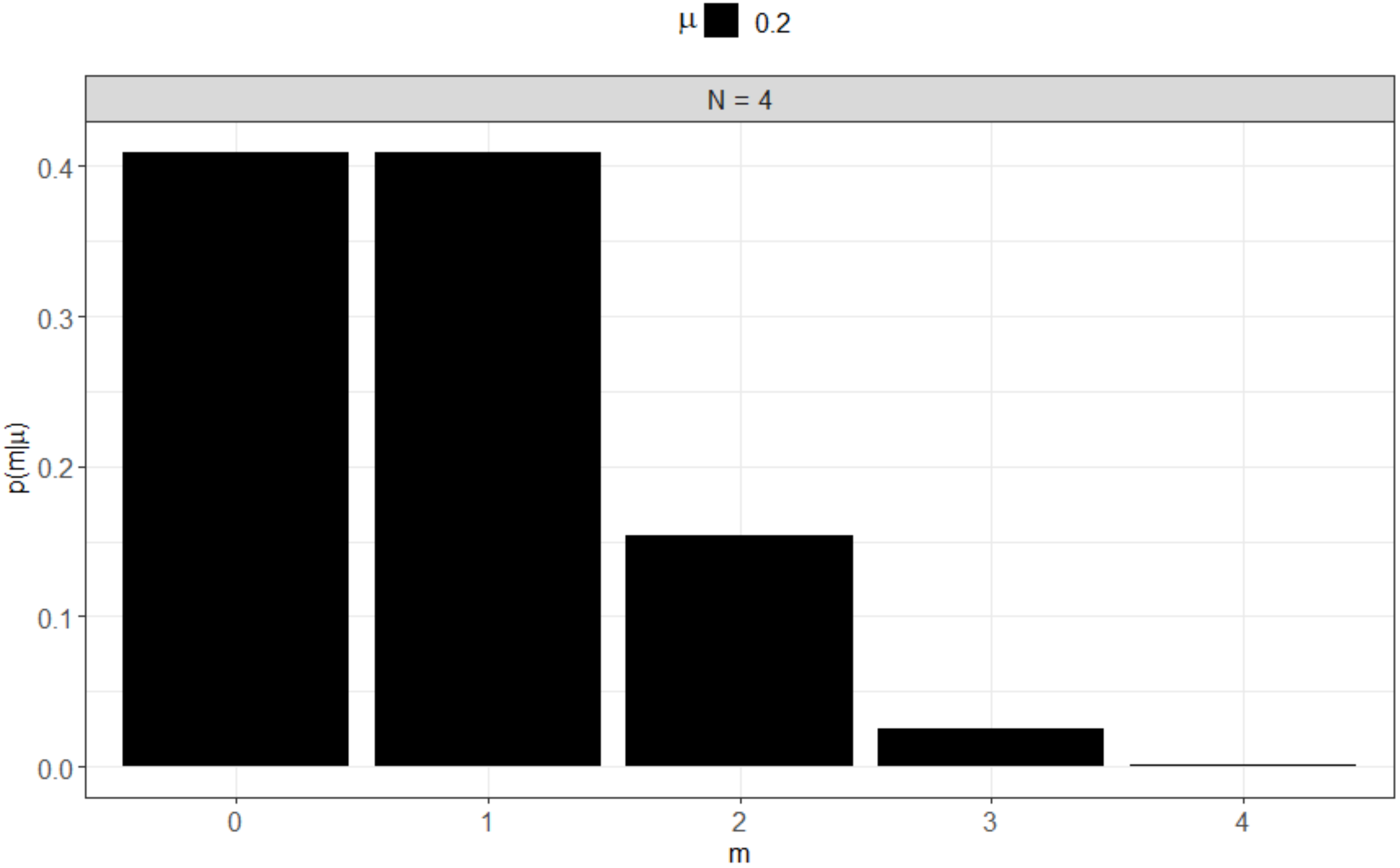
Binomial distribution for $N = 8$ and $\mu = 0.2$



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

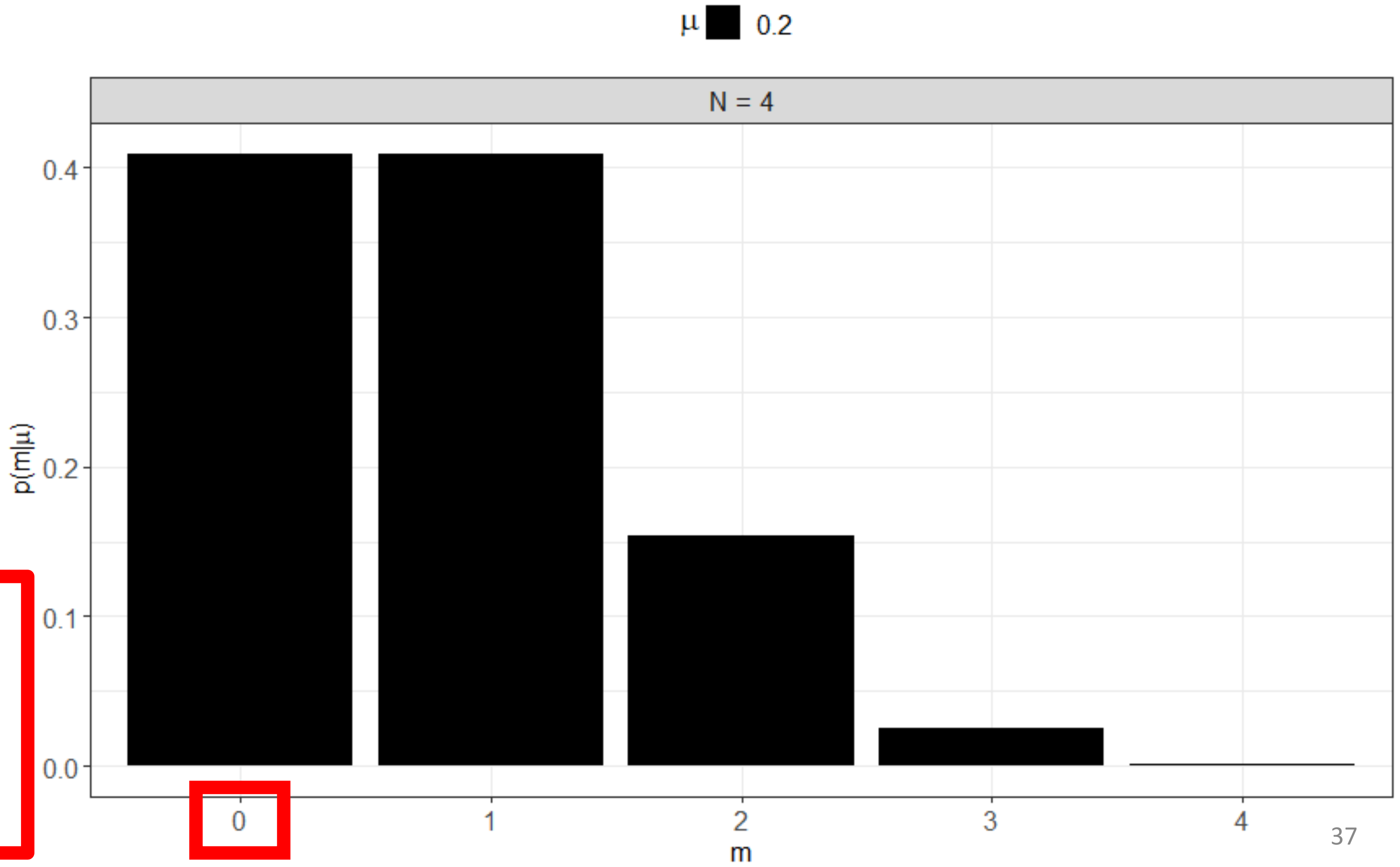


Back to our movie example…let’s assume the TRUE probability of Yes is $\mu_{TRUE} = 0.2$



Back to our movie example…let's assume the TRUE probability of Yes is $\mu_{TRUE} = 0.2$

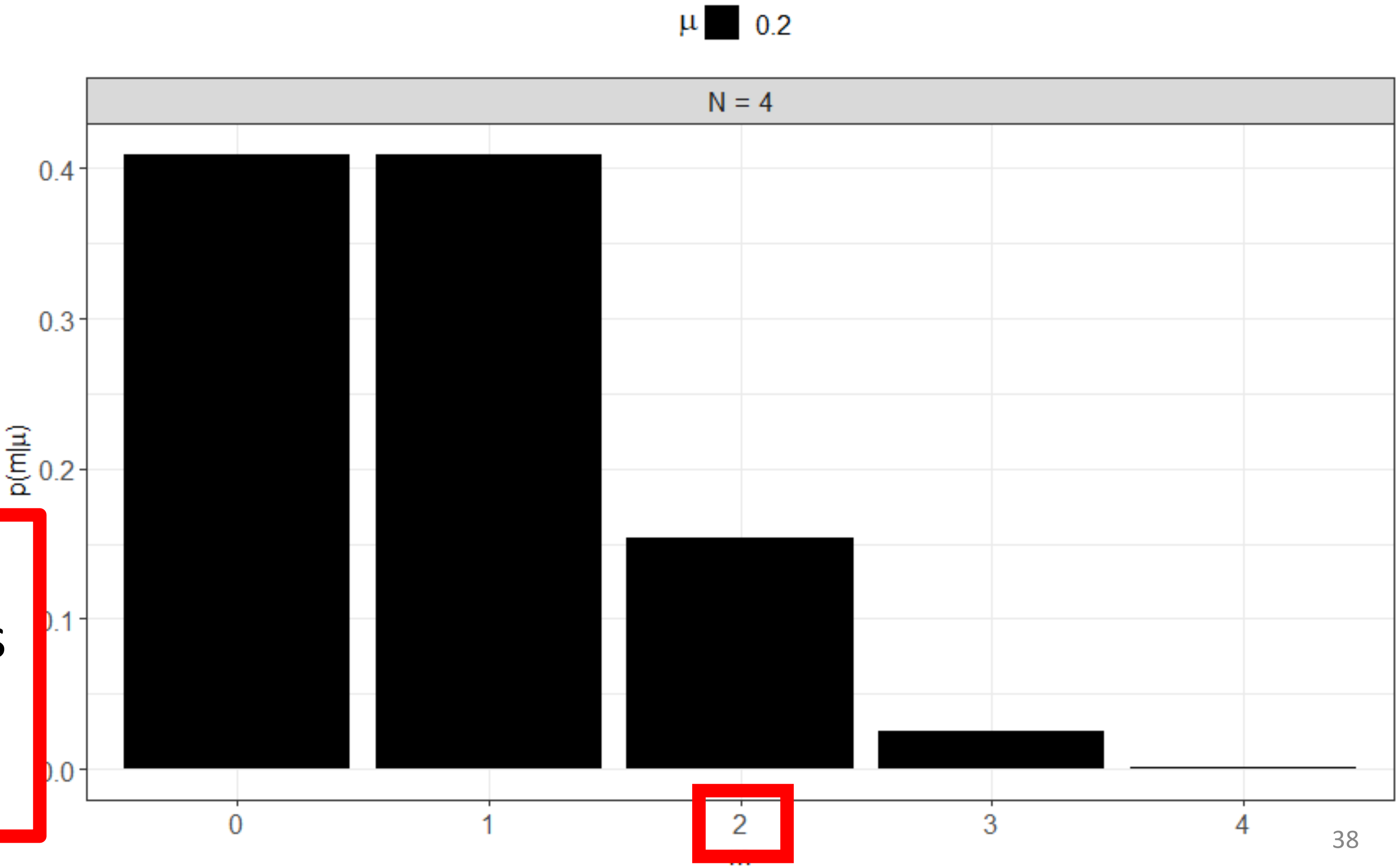
We ask 4 people, so we have 4 trials: $N = 4$



The probability of finding 0 Yes responses is ≈40%!

Back to our movie example…let's assume the TRUE probability of Yes is $\mu_{TRUE} = 0.2$

We ask 4 people, so we have 4 trials: $N = 4$



The probability of finding 2 Yes responses is small but not negligible at ≈15%.

We conduct an experiment where we ask 4 random people on the street…

- If 0 out of 4 people say Yes, our MLE for the probability would be?

- If 2 out of 4 people say Yes, our MLE for the probability would be?

We conduct an experiment where we ask 4 random people on the street…

- If 0 out of 4 people say Yes, our MLE for the probability would be $\mu_{MLE} = 0$.

- If 2 out of 4 people say Yes, our MLE for the probability would be $\mu_{MLE} = 0.5$.

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

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?

Maybe I’ve read a lot of reviews and blog posts…I think I know something about what people think about the movie

- It seemed like most people did not like the movie…
- 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.

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

The Evidence (marginal likelihood) is a NORMALIZING CONSTANT.

Let’s focus on the NUMERATOR for now…we will return to the EVIDENCE later in the semester.

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 BETA distribution to encode our PRIOR belief on the probability, μ

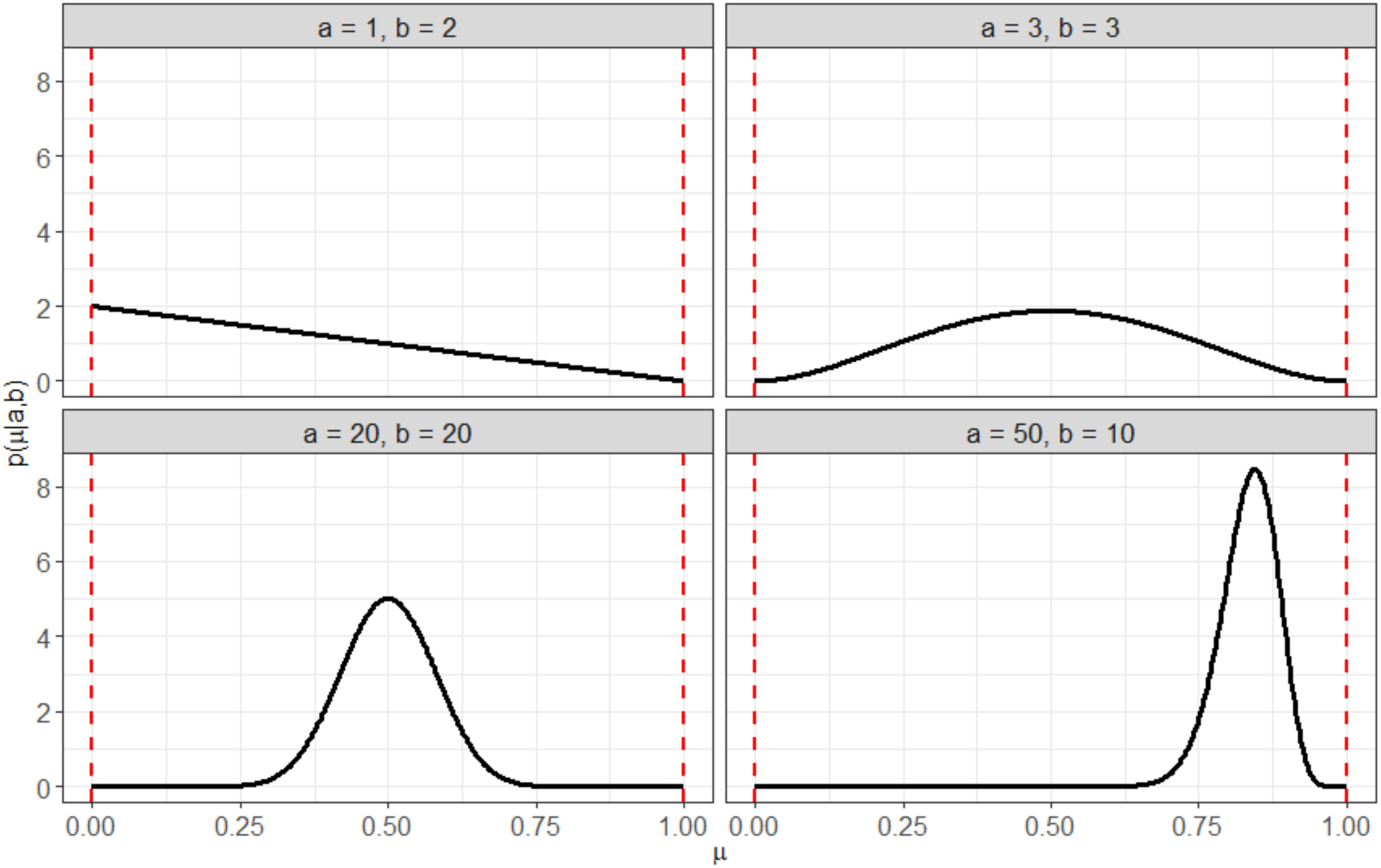
Beta distribution

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

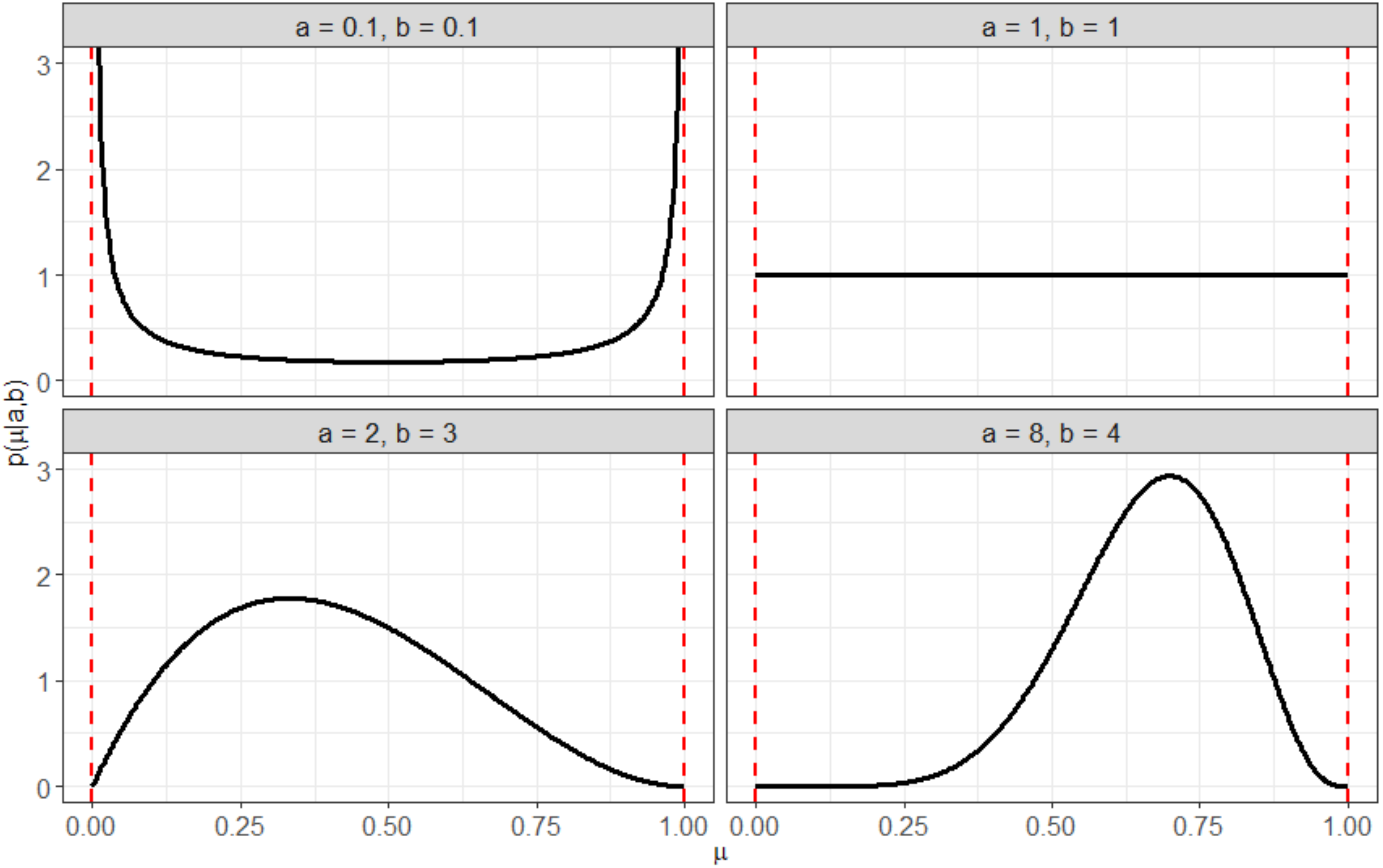
- It is a flexible distribution capable of a wide variety of shapes.

- The shape is controlled by the hyperparameters α and β.
  - The PRML 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 **conjugate prior** of the binomial likelihood.

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

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

⇕

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

- A conjugate prior is useful because the resulting **posterior** distribution will have the **<u>same functional form as the prior</u>**.

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 is a **beta distribution**!

Posterior distribution on μ

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

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 Yes responses, m, or more generally the number of observed EVENTS.

- b is added to the number of No responses, $N - m$, or more generally the number of times we did NOT observe the EVENT.

Beta distribution hyperparameter interpretations

- a is added to the number of Yes responses, 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 No responses, $N - m$, 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:

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

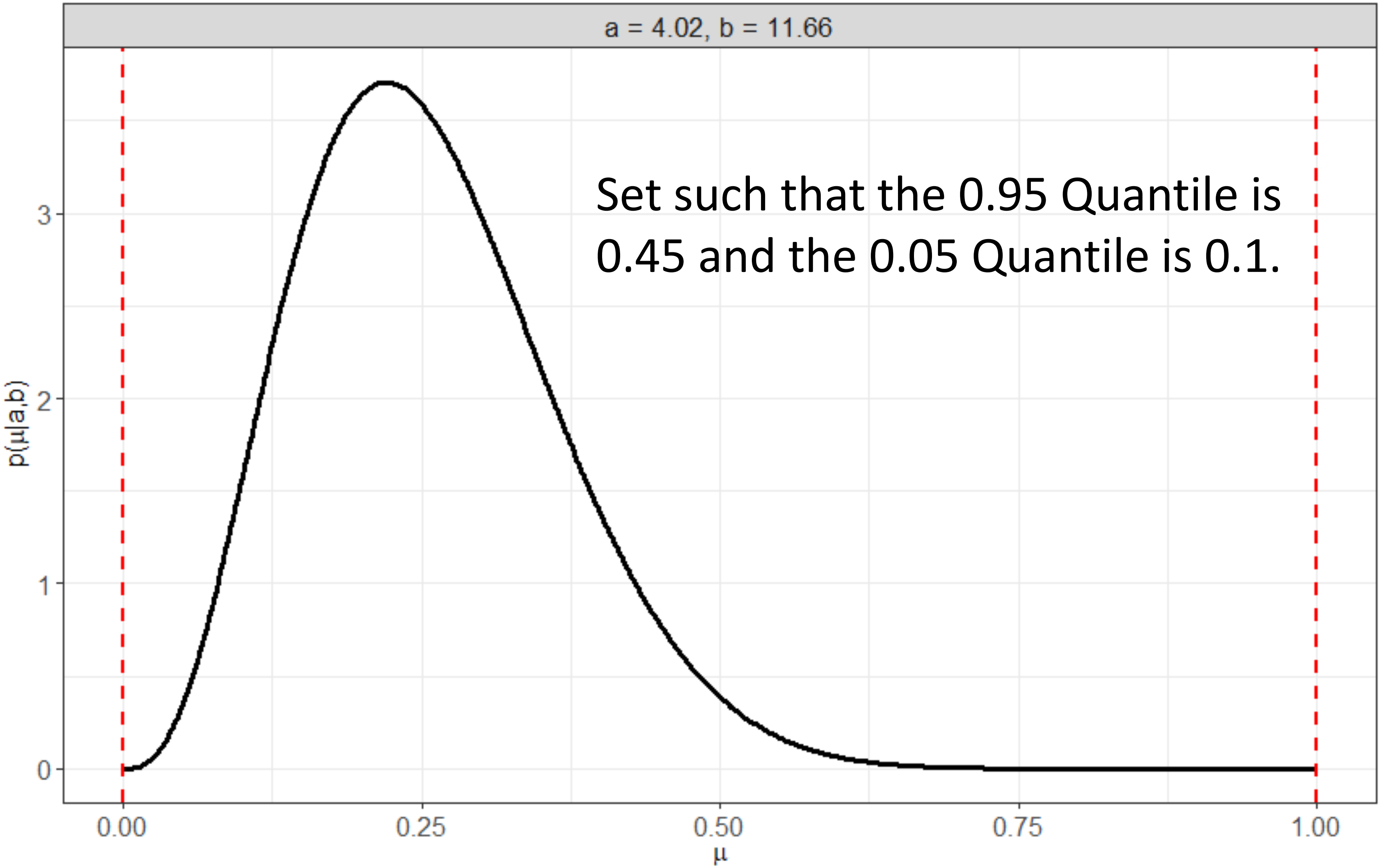
We could have reached the same interpretations by considering the mean…

- 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 Yes response is greater than 0 but less than 0.5



Set such that the 0.95 Quantile is 0.45 and the 0.05 Quantile is 0.1.

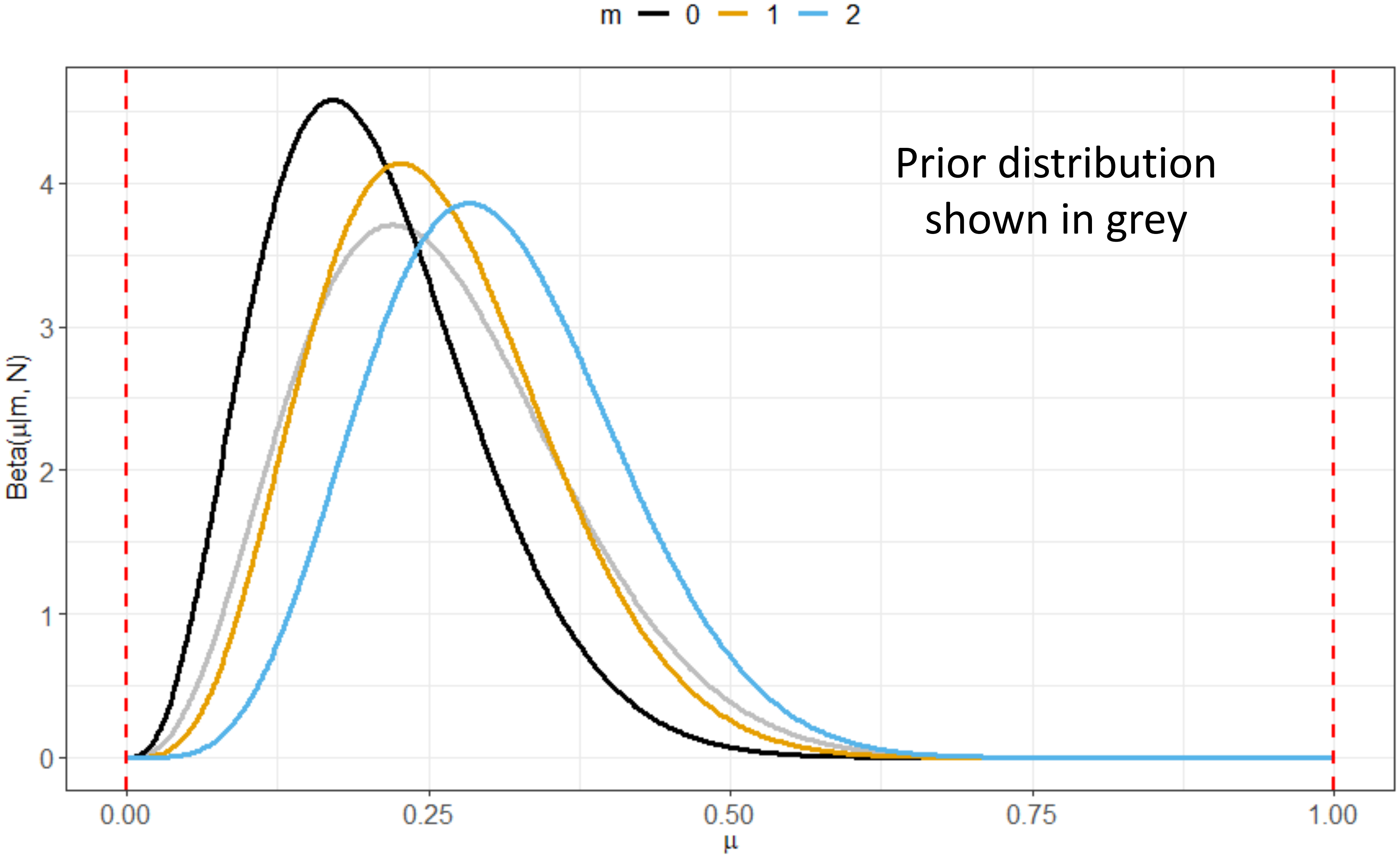
We will update our belief about μ under three different circumstances

- 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 **<u>updated belief</u>** **if** we found 0 Yes responses, vs 1 Yes response, vs 2 Yes responses.

μ posterior distribution given m and $N = 4$



Prior distribution shown in grey

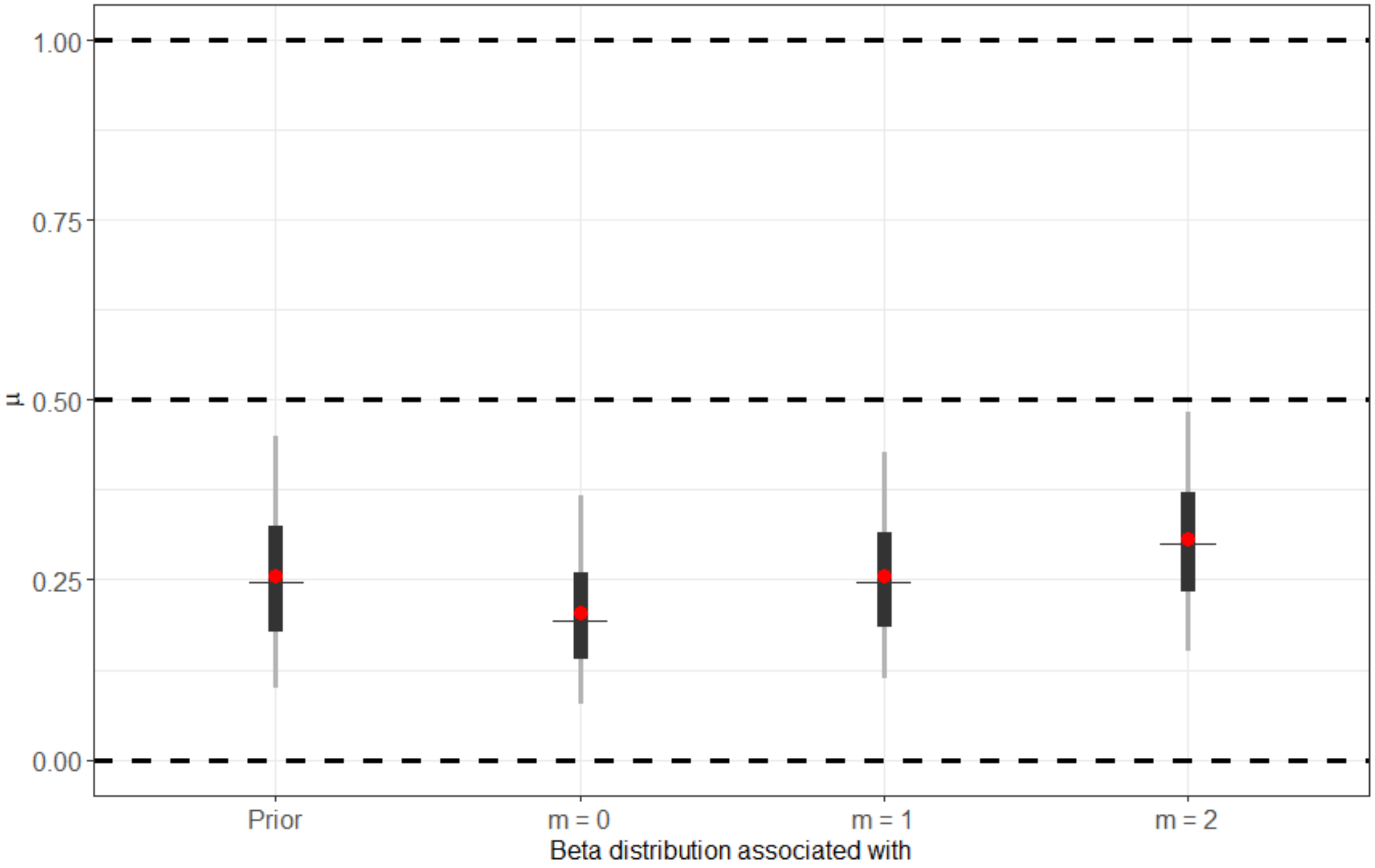
Summarize the Beta distributions

- Calculate summary statistics for each Beta distribution.

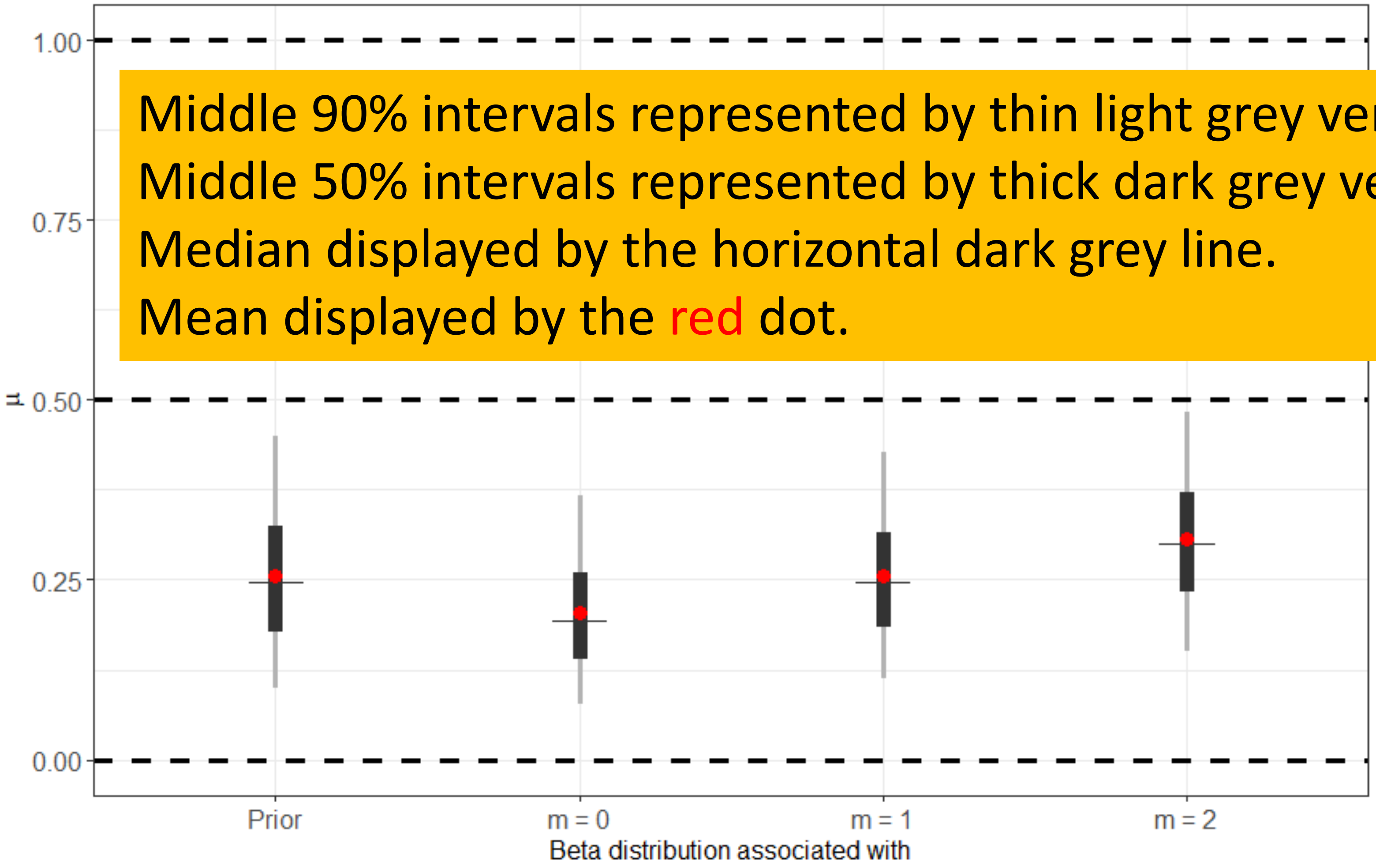
- Represent uncertainty with **credible intervals**:
  - Middle 50% interval – spans the 25th through 75th quantiles
  - Middle 90% interval – spans the 5th through 95th quantiles

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

Visualize the Beta distribution summary statistics

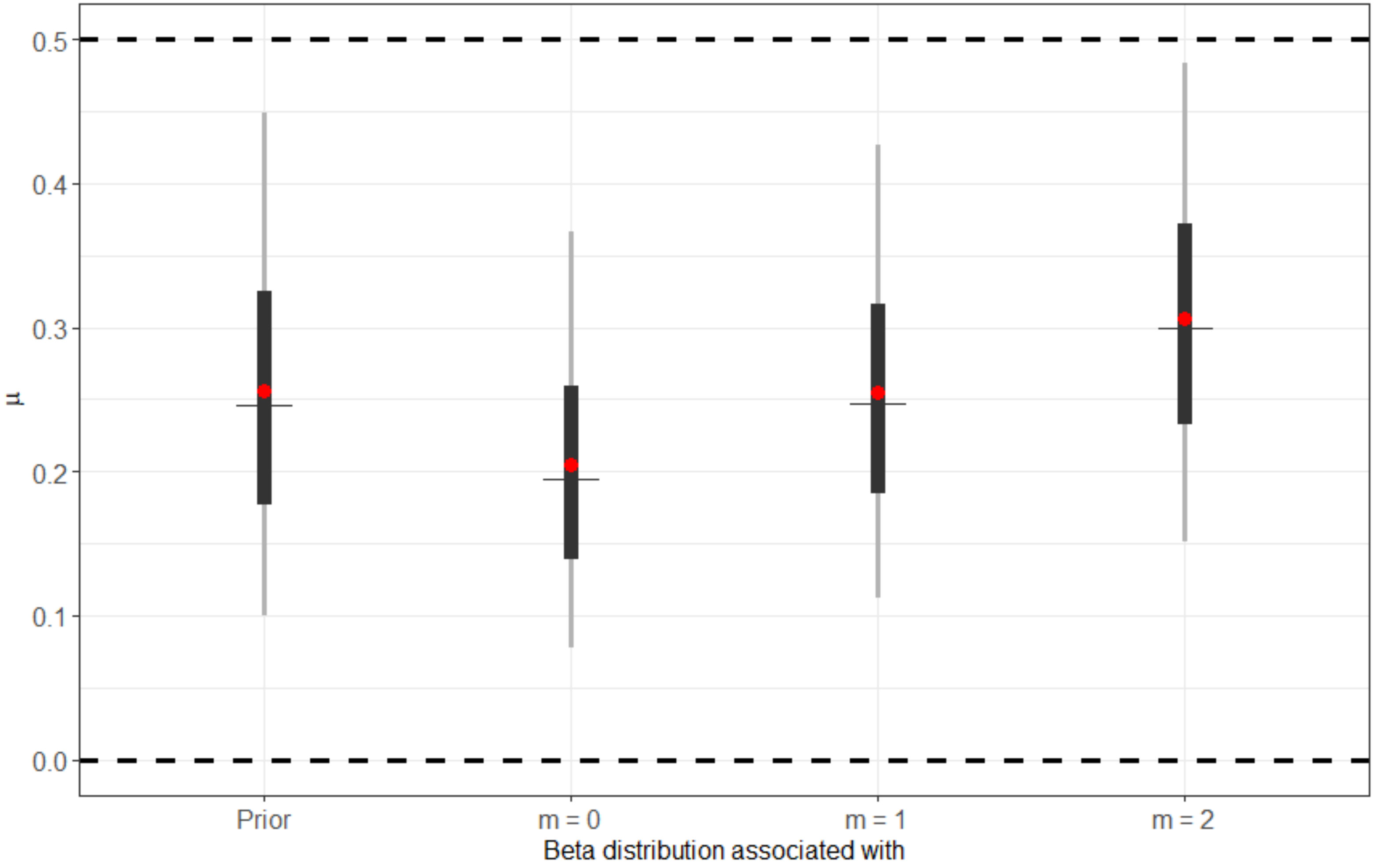


Visualize the Beta distribution summary statistics



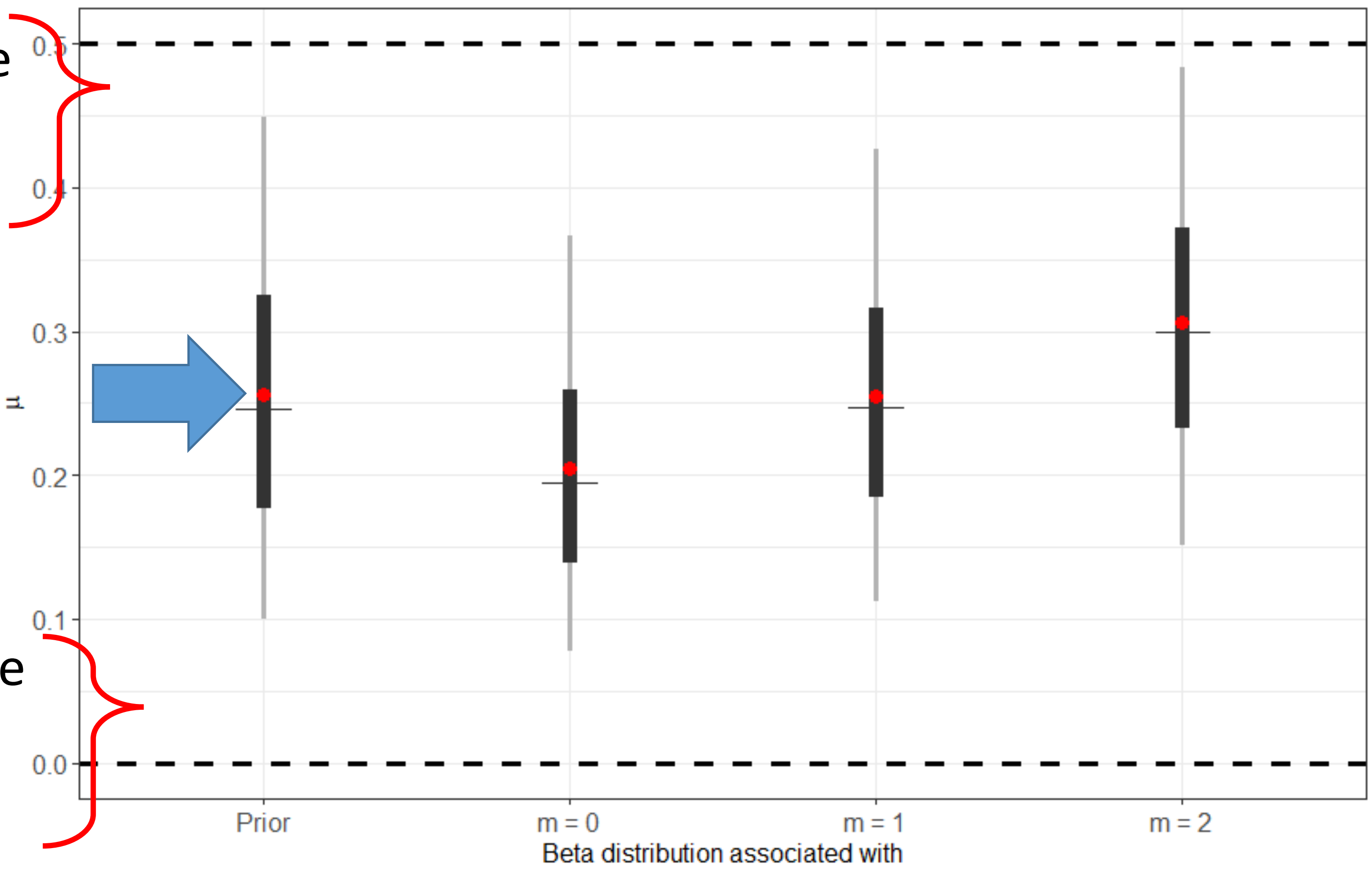
Middle 90% intervals represented by thin light grey vertical lines.
Middle 50% intervals represented by thick dark grey vertical lines.
Median displayed by the horizontal dark grey line.
Mean displayed by the red dot.

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

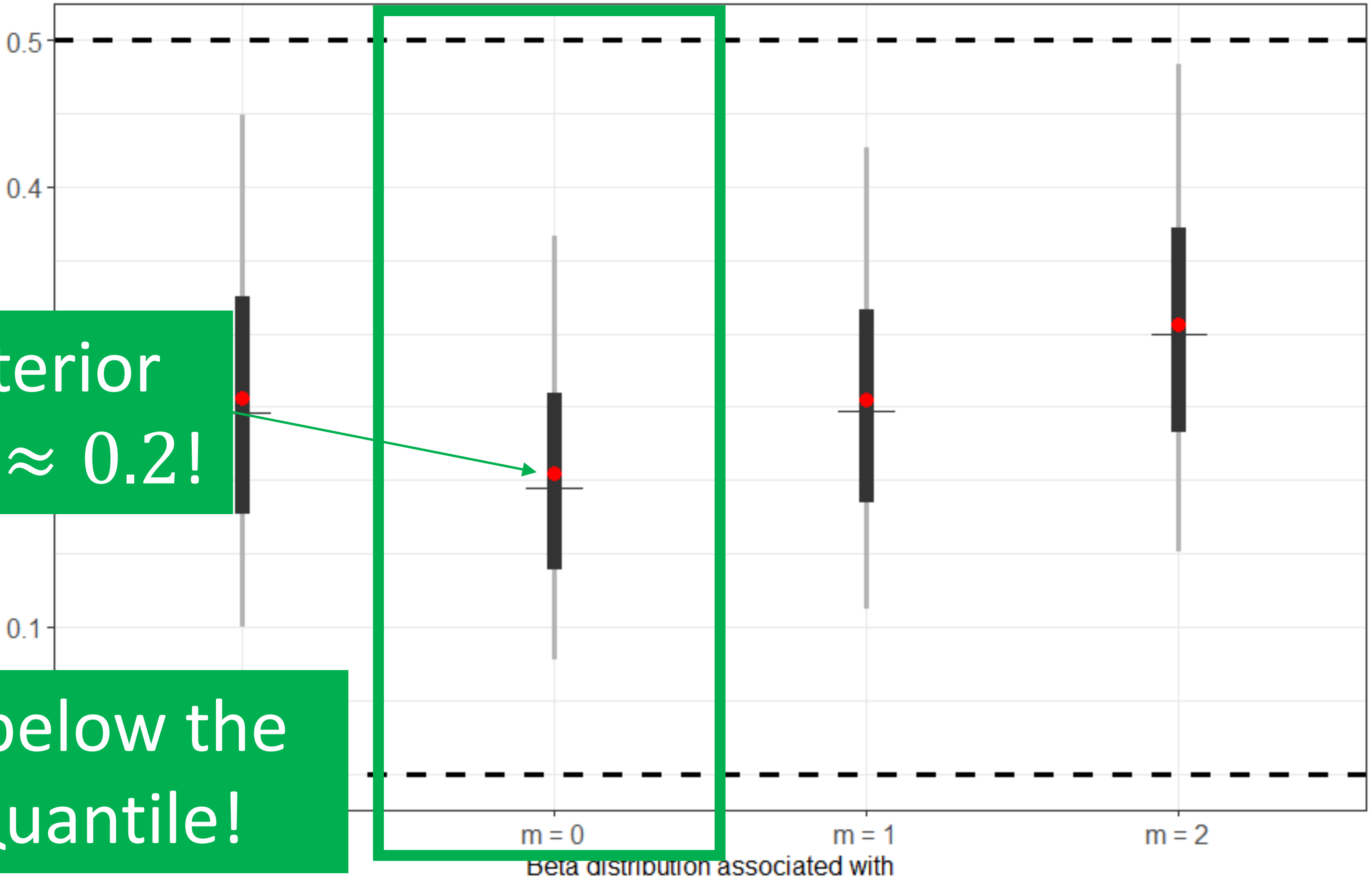
A priori we believe the mean is ≈ 0.25



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

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

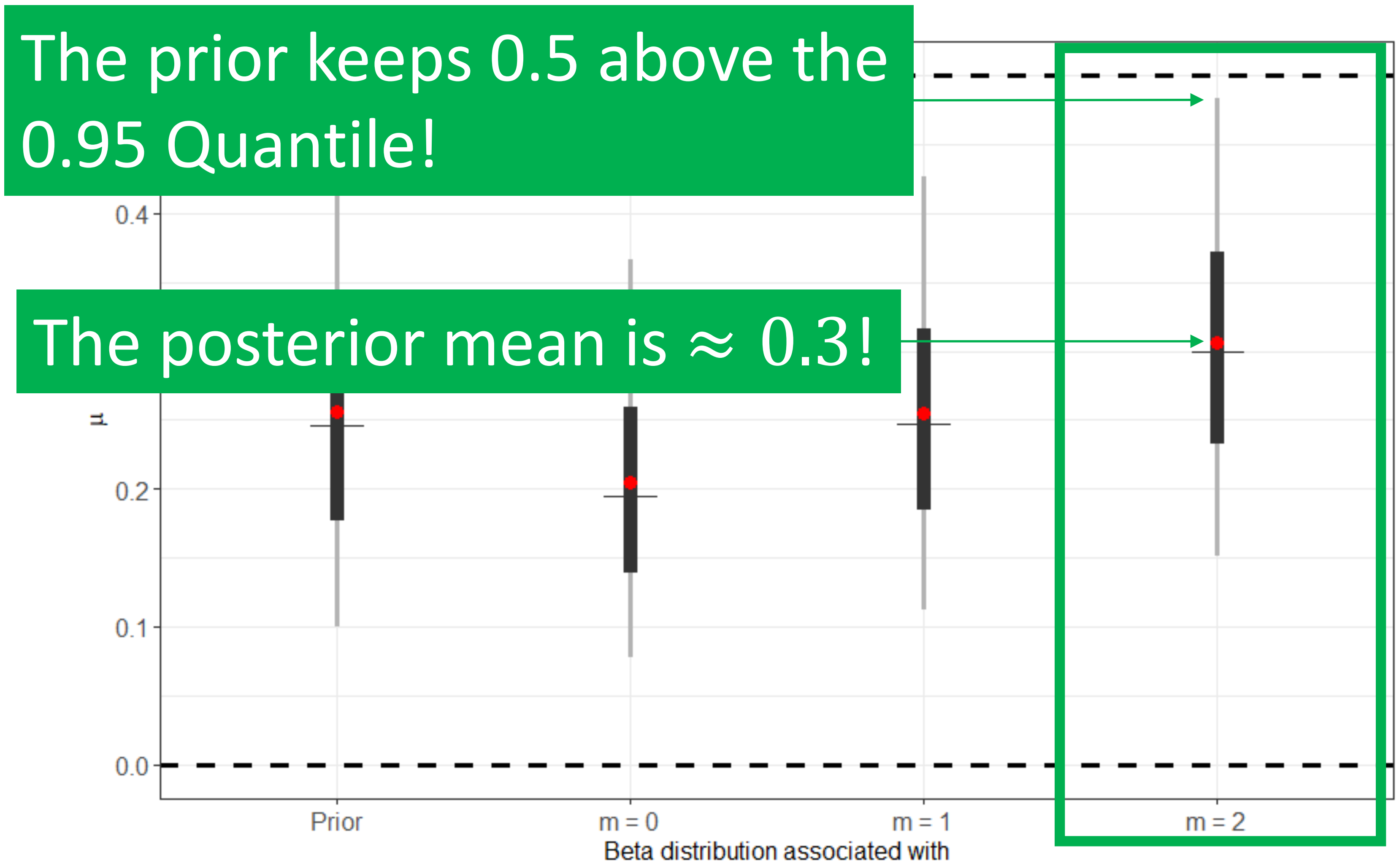
IF we observed $m = 0$ out of $N = 4$



The posterior mean is ≈ 0.2!

Zero is below the 0.05 Quantile!

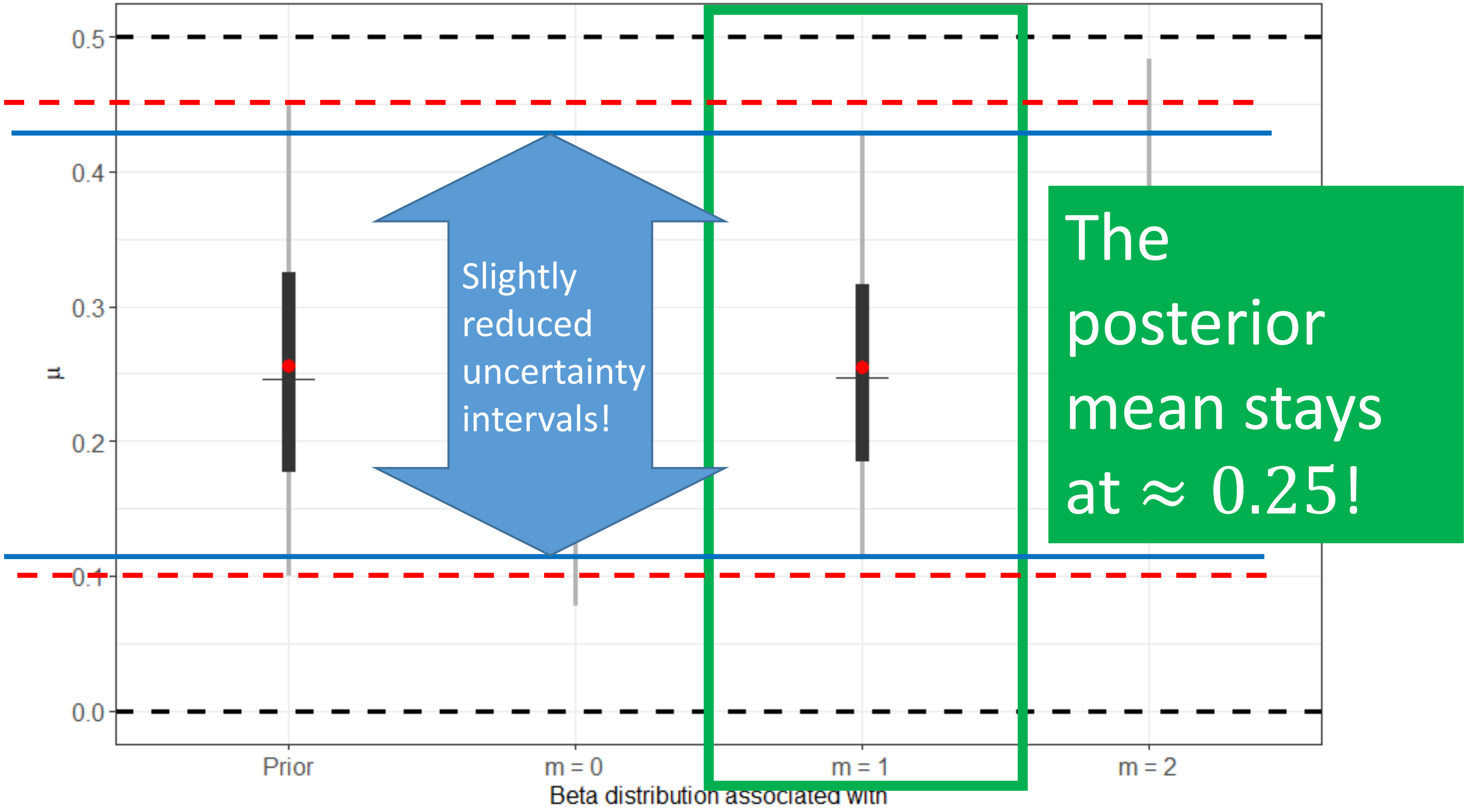
IF we observed $m = 2$ out of $N = 4$



The prior keeps 0.5 above the 0.95 Quantile!

The posterior mean is ≈ 0.3!

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



Slightly reduced uncertainty intervals!

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**.

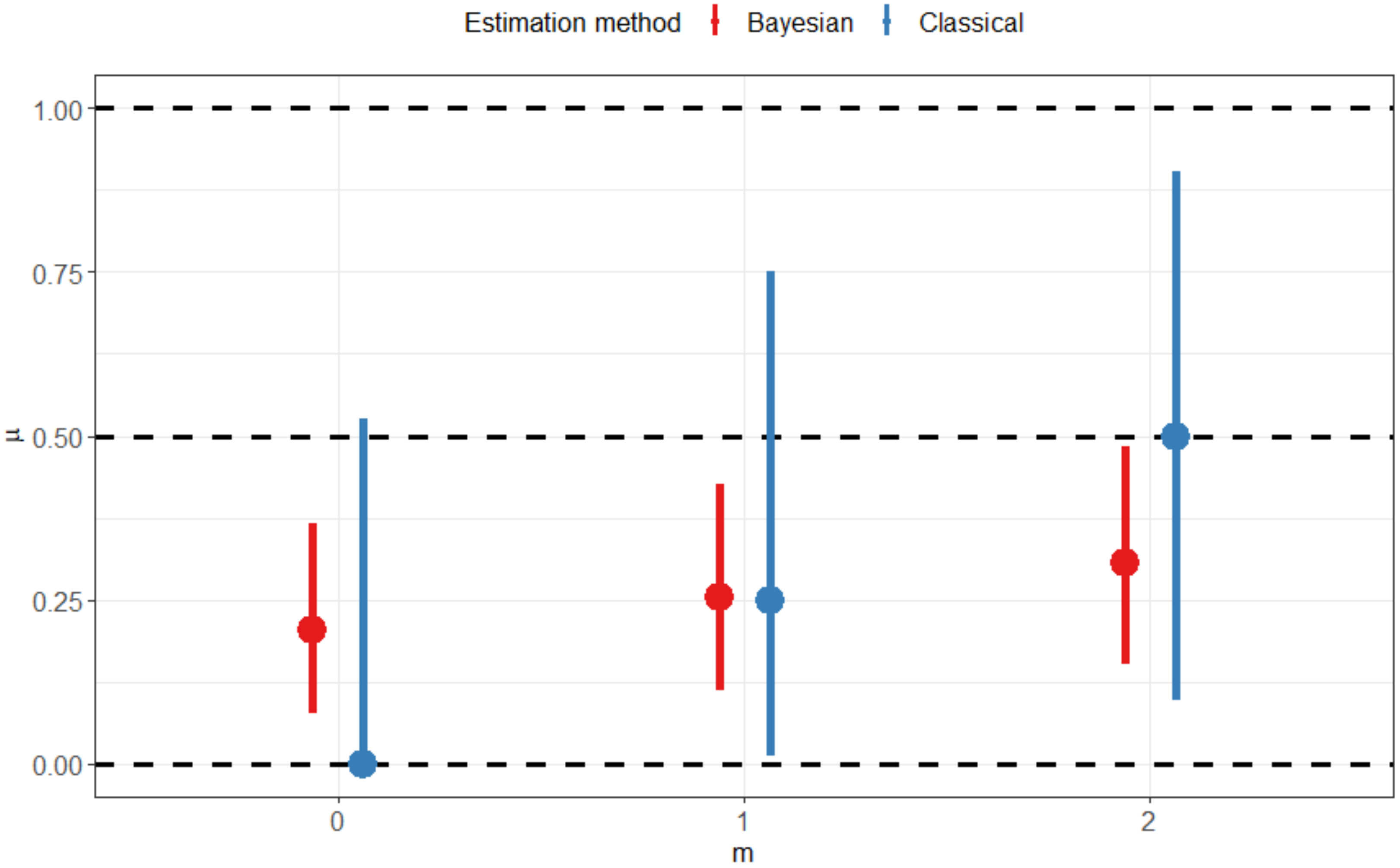
- How do 90% confidence intervals compare with the *posterior credible intervals* in our example?

Confidence interval calculation

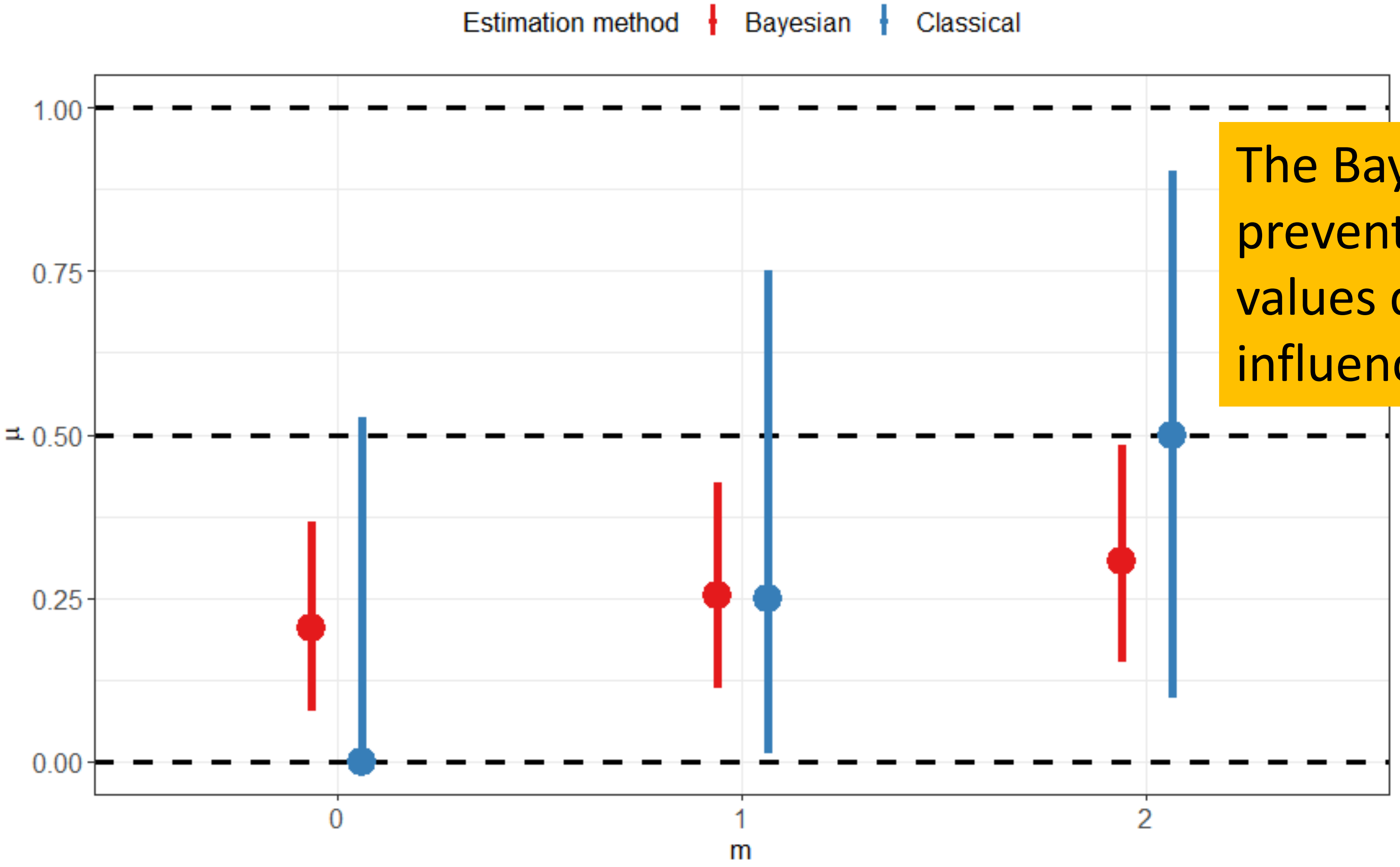
- The 90% confidence intervals are calculating using the Clopper-Pearson method, through `R`’s `binom.test()` function.

- 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)



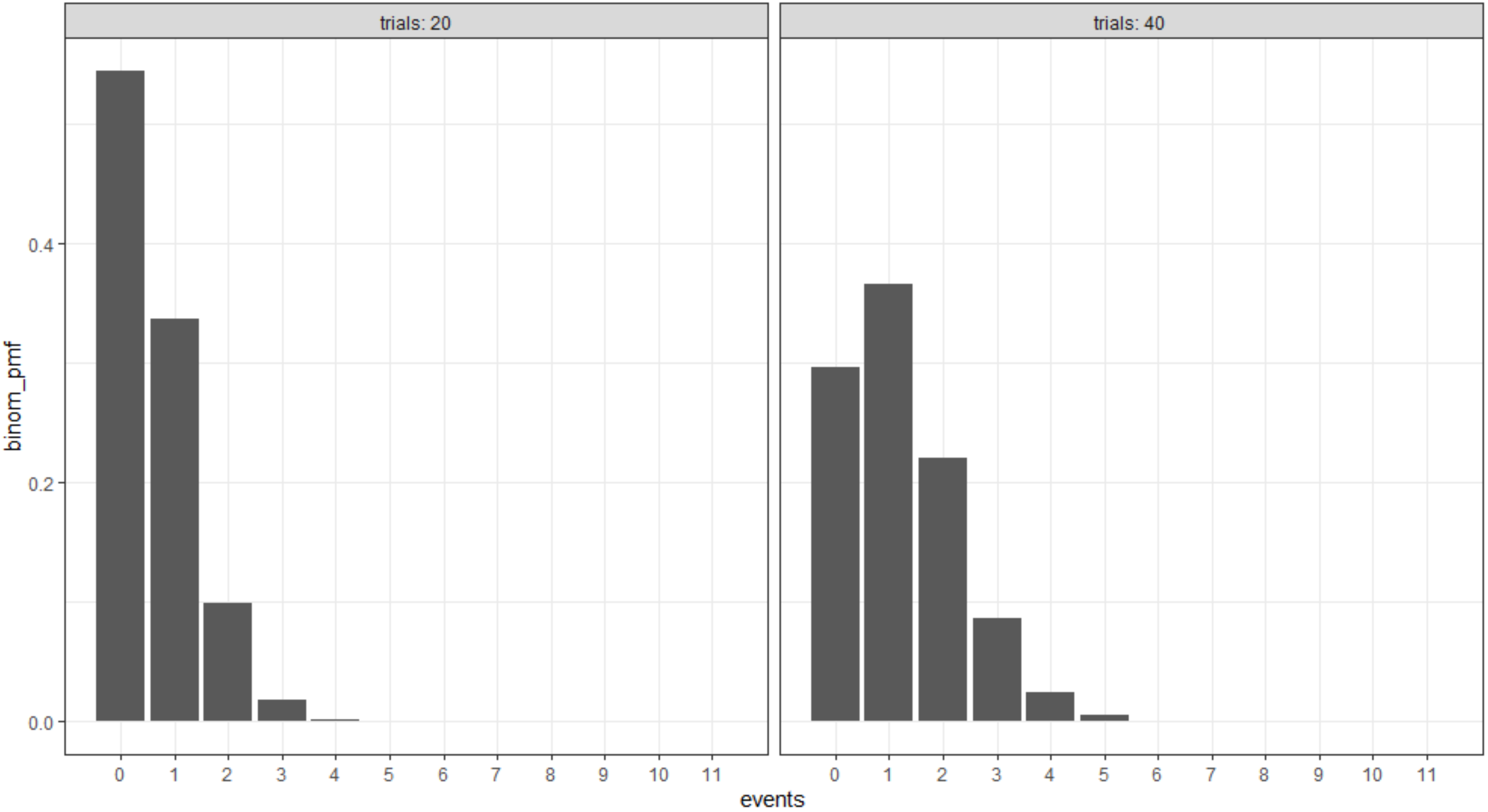
The Bayesian approach prevents unrealistic values due to the influence of the prior!!

Why is the Bayesian approach useful?

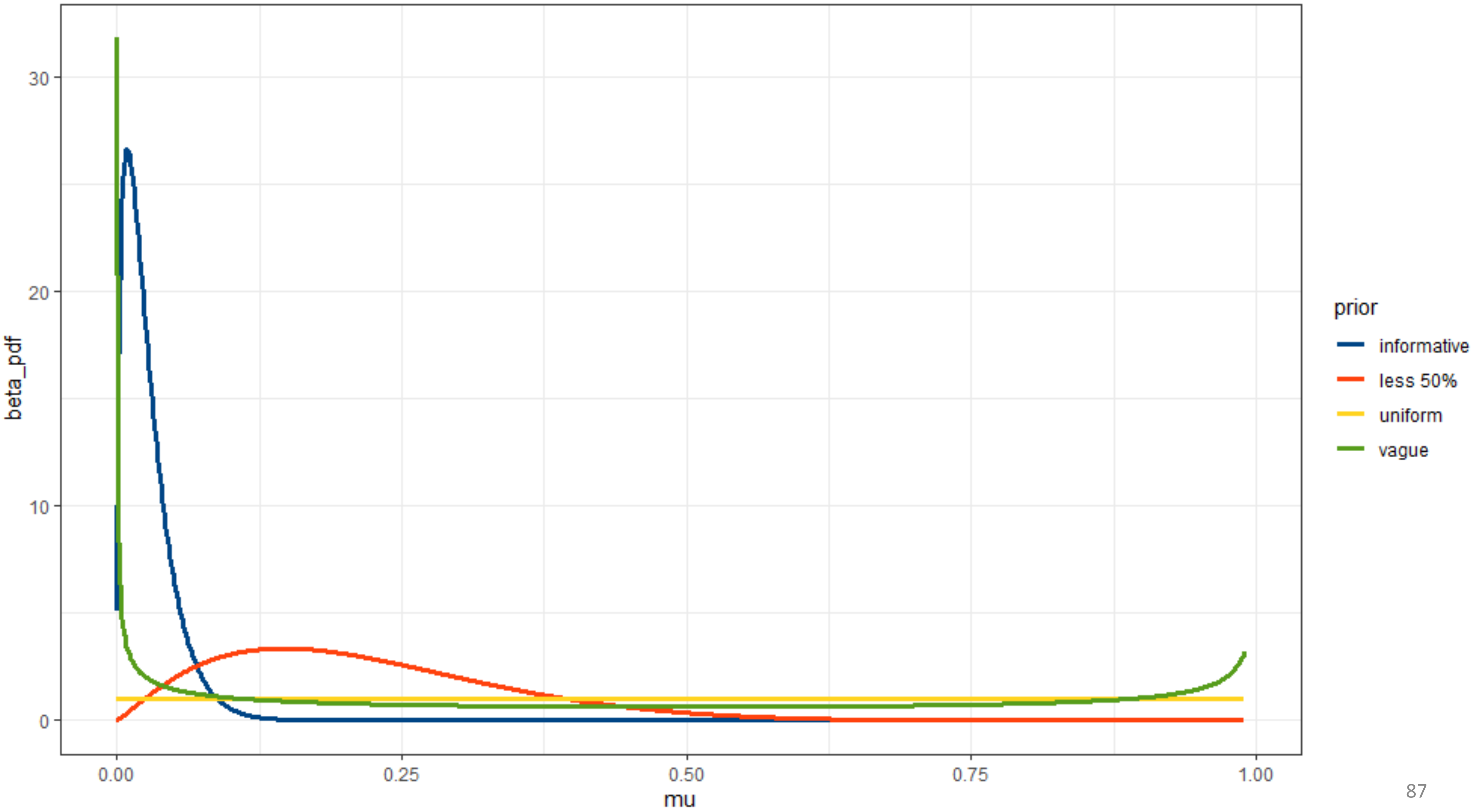
- Consider sampling a population to identify a rare event…

- 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

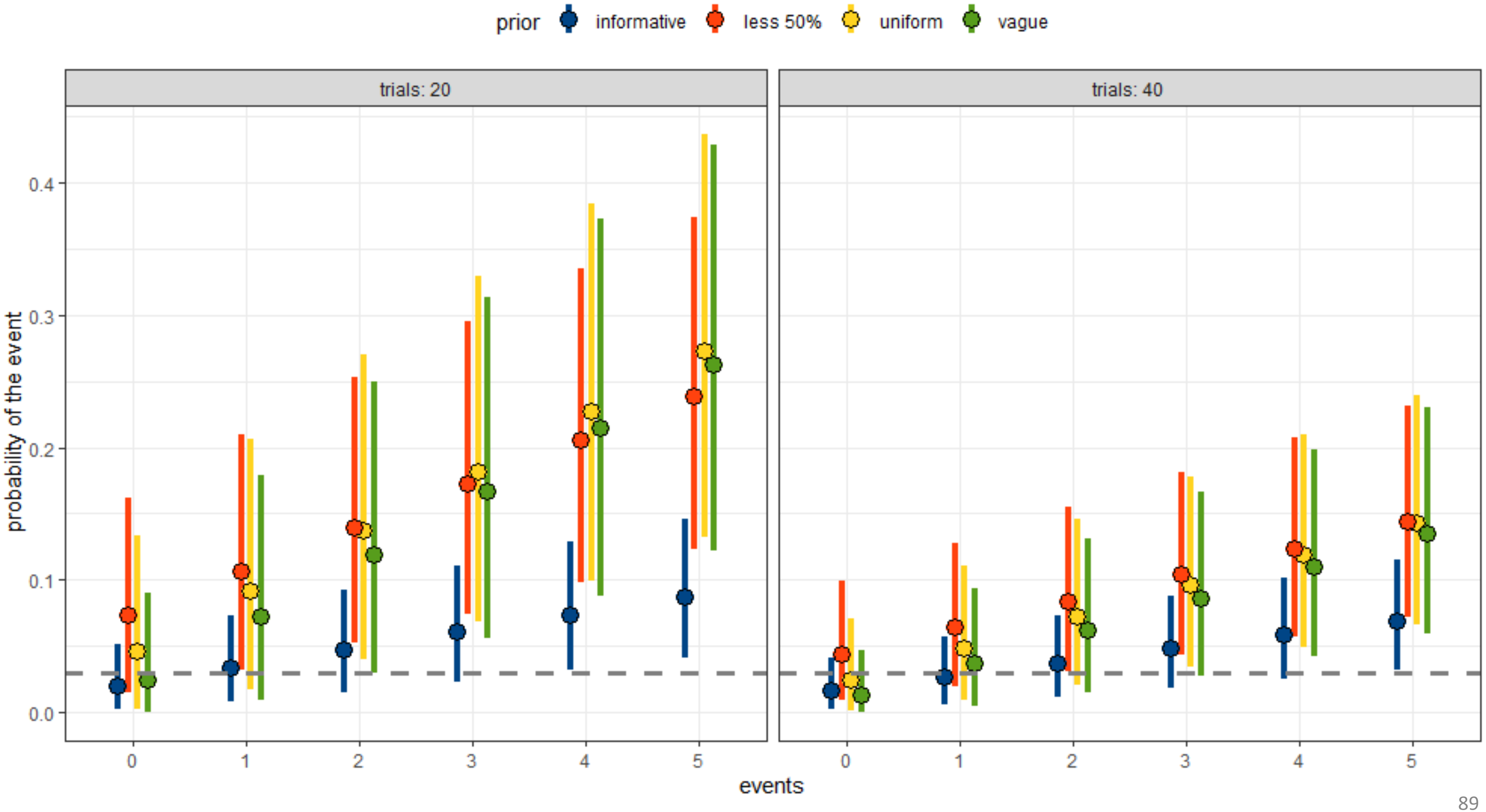


How does the posterior behave under different circumstances?

- Compare the posterior summaries:
  - Middle 90% credible intervals (5th through 95th 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