CS1675/2075: Introduction to Machine Learning

Week 4: Modeling Data with Probability Distributions (Bernoulli)

Fall 2024

Introduction to Machine Learning

Week 4 - Distribution Fitting Intro

Spring 2025
Instructor: Dr. Patrick Skeba

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

Task: determine the probability of an EVENT

- For now, we will only collect observations of EVENT and NON-EVENT. That is, we only have labels and no input features.
- Say I wish to ask somebody: "did you like the movie Inside Out 2?"
- We will use an encoding similar to that discussed with classification:
  - The general variable x will represent a person's response.

Encoding a single response

Use a binary variable x to indicate whether or not someone like the movie

- If a person did like the movie, set $x = 1$
- If a person did NOT like the movie, set $x = 0$
- The **PROBABILITY** someone liked the movie will be denoted by the parameter μ
  - $p(x = 1 \,|\, \mu) = \mu$
- Since we have either $x = 1$ or $x = 0$, the probability that someone did not like the movie can be expressed:
  - $p(x = 0 \,|\, \mu) = 1 - \mu$

Probability mass function (discrete variables)

Express the probability over all possible values of $x \in \{0,1\}$

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

- The **Bernoulli distribution** has one parameter, μ, that describes the probability of an EVENT. It is a discrete distribution defined over the binary variable x
  - μ is a probability and so is bounded $0 \leq \mu \leq 1$.
- x is binary and can only take the values 0 or 1.

Probability mass function (discrete variables)

Express the probability over all possible values of $x \in \{0,1\}$

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

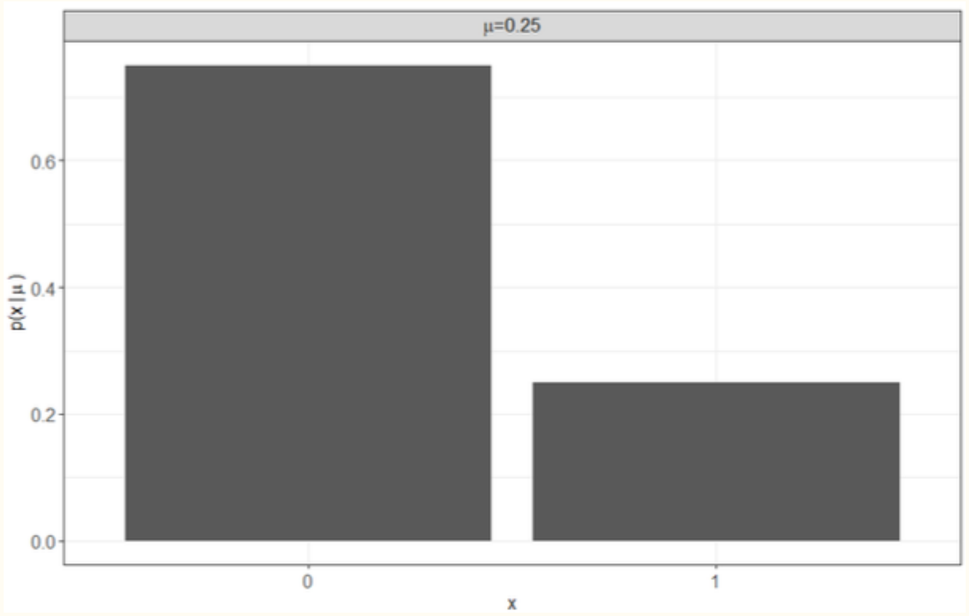
Check when $x = 1$

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

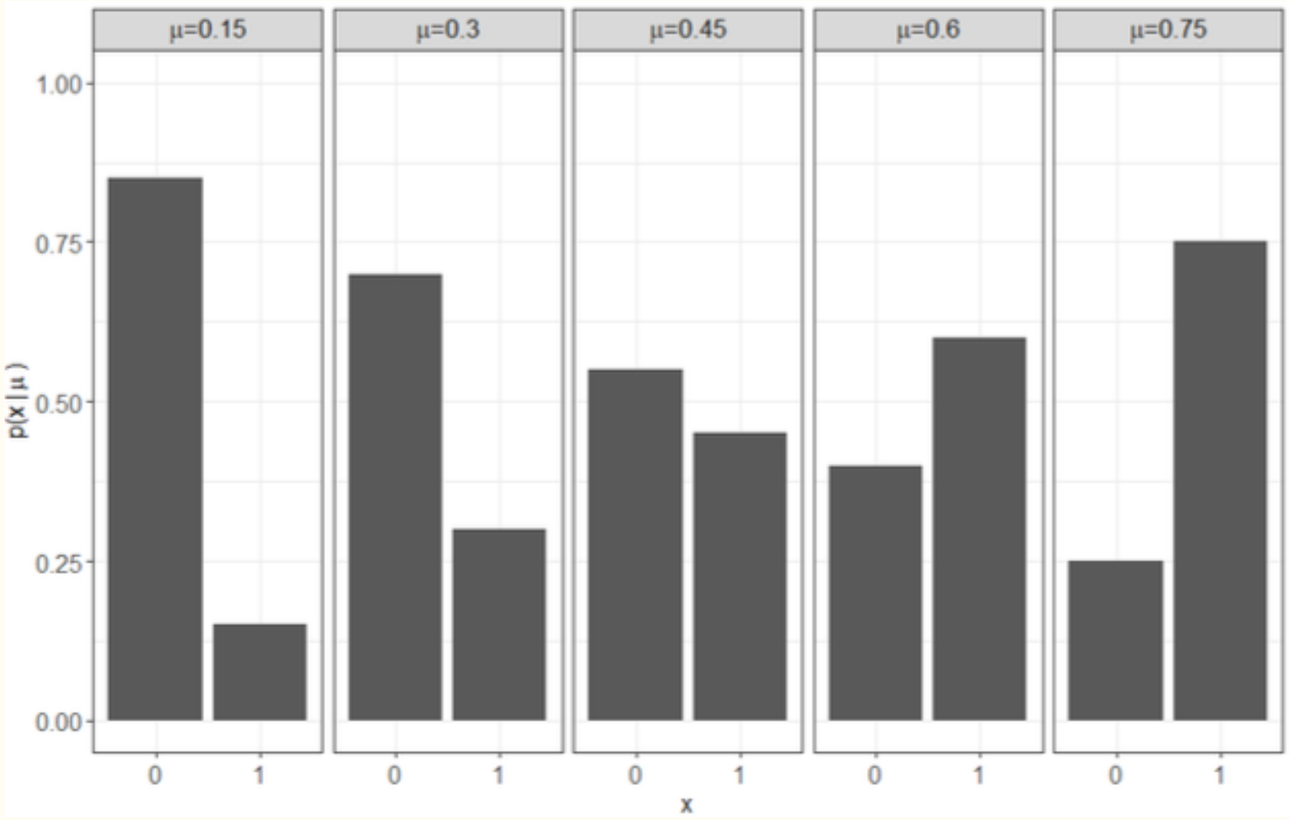
Check when $x = 0$

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

Bernoulli distribution PMF for a single value of μ



Bernoulli distribution PMF for multiple values of μ



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

Useful for many binary outcomes

- Today we're looking at (binary) movie ratings, but Bernoulli is most often introduced with the canonical coin-flip example.

- Immediately, we are asking "given a known event probability μ, how likely is it that we observe a particular response."

- The machine learning task, which we will get to shortly, is to **estimate** the value of μ from observations of the output variable x

- **Foreshadowing**: more useful ML models will estimate the value of μ based on input features and tunable parameters, separately for each sample

Back to the movie example…

We need a sample size bigger than one!

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

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

What is the probability of the *sequence* we observed?

Designate each respondent with a subindex

- Start with the joint distribution:
  - $p(x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 0 \mid \mu)$
  - How can we evaluate this joint probability?

Conditional independence!

Assume each observation is unaffected by the results of any other observations

- In the coin-flip example: each time I toss the coin up, the probability that it comes down heads is 50%.
- **Misunderstanding conditional independence is how you fall into the gambler's fallacy!**

Some roulette games display the last N numbers that came up on the wheel. This information is completely worthless, provided the wheel is fair.

Conditional independence!

Assume each observation is unaffected by the results of any other observations

- For our movie example, imagine we just have a box full of anonymous ballots marking “Like” or “Dislike.”
  - The order in which I read the responses is **irrelevant**
  - Observing the result of one ballot tells me nothing about the next one I draw.
- If two (or more) random variables are conditionally independent, we can **factor** their joint probability into the product of the **marginals:**
  - $p(x_1, x_2 \mid \mu) = p(x_1 \mid \mu) p(x_2 \mid \mu)$

Independence and the Product Rule

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

- If A and B are independent, then $p(A \mid B) = p(A)$
  - i.e. the probability of A is the same regardless of what value B takes
  - Knowing something about B does not tell us anything about A
- Product rule simplifies to:
  - $p(A, B) = p(A \mid B)p(B) = p(A)p(B)$

Back to the problem of asking 4 people

Assume independence and factor the joint distribution

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

- Each distribution on the right hand side is a **Bernoulli**

Substitute into the expression the Bernoulli pmf for each person

Assume independence and factor the joint distribution

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

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

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

Now generalize from asking 4 people to N people

- Denote the sequence of responses: $\mathbf{x} = \{x_1, x_2, x_3, \ldots, x_N\}$
- Assuming each person is **independent** we can factor the joint distribution into the product of N independent distributions:

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

What happens when μ is unknown?

Estimate it from the sequence of responses!

- If we have observed the sequence, $\mathbf{x} = \{x_1, x_2, x_3, \ldots, x_N\}$, we can estimate or **infer** the value of μ
- Which value is best?
- The value that **MAXIMIZES** the likelihood!

$$\hat{\mu} = \mu_{MLE} = \arg \max_{\mu \in [0,1]} p(\mathbf{x} \mid \mu)$$

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

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

Number of people that liked the movie, or more generally number of events M.

Number of people that did NOT like the movie, or more generally number of times we **did not** observe the event $N - M$.

Rearrange the log-likelihood

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

To optimize, calculate the derivative of the log-likelihood with respect to μ

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

$$\frac{M}{\mu} \qquad \frac{-(N-M)}{1-\mu}$$

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

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

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

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

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

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

Binomial Distribution

Earlier, we introduced the Bernoulli distribution

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

- 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:

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

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 **specific number of times out of a specific number of trials**?

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	x	$p(x\|\mu)$
1	No	0	$(1-\mu)$
2	No	0	$(1-\mu)$
3	Yes	1	μ
4	No	0	$(1-\mu)$

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

Wait…didn't we calculate this already?

- Based on the following independent observations:

Person	Response	x	$p(x\|\mu)$
4	No	0	$(1-\mu)$

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

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

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	$p(\mathbf{x}\|\mu)$
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 **potential** sequence:

$$4 \times \mu \times (1 - \mu)^3$$

- 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

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

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

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

$$6 \times \mu^2 \times (1 - \mu)^2$$

How many **potential** 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 $x = 1$	$p(\mathbf{x}\|\mu)$
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??

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.

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 $(1-\mu)$ 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	$1 \times \mu^m \times (1-\mu)^4$
1	$4 \times \mu^m \times (1-\mu)^3$
2	$6 \times \mu^m \times (1-\mu)^2$
3	$4 \times \mu^m \times (1-\mu)^1$
4	$1 \times \mu^m \times (1-\mu)^0$

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

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

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:

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

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

Known as the **Binomial** distribution!

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

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

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

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

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

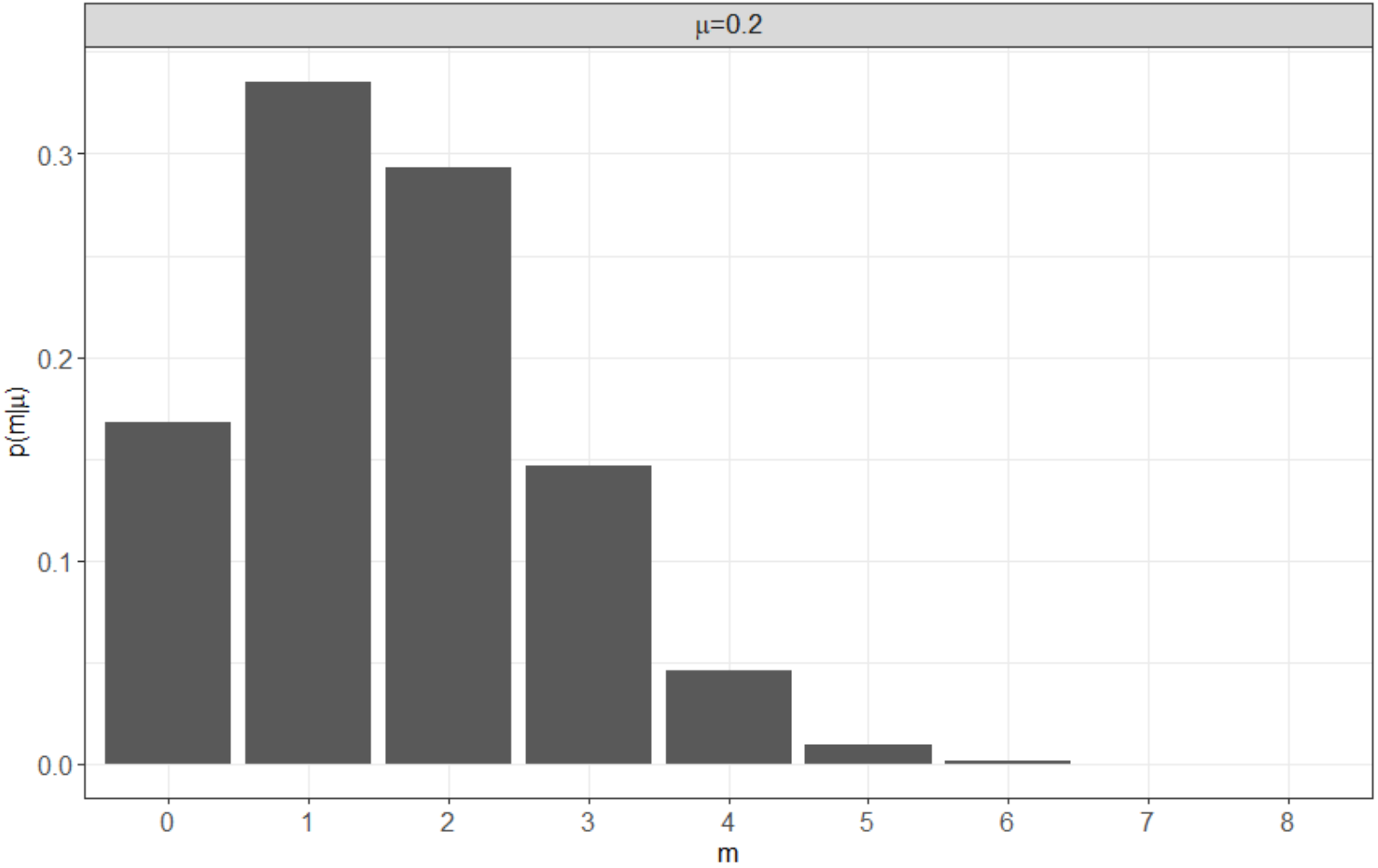
We derived the Binomial distribution starting from Bernoulli observations

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

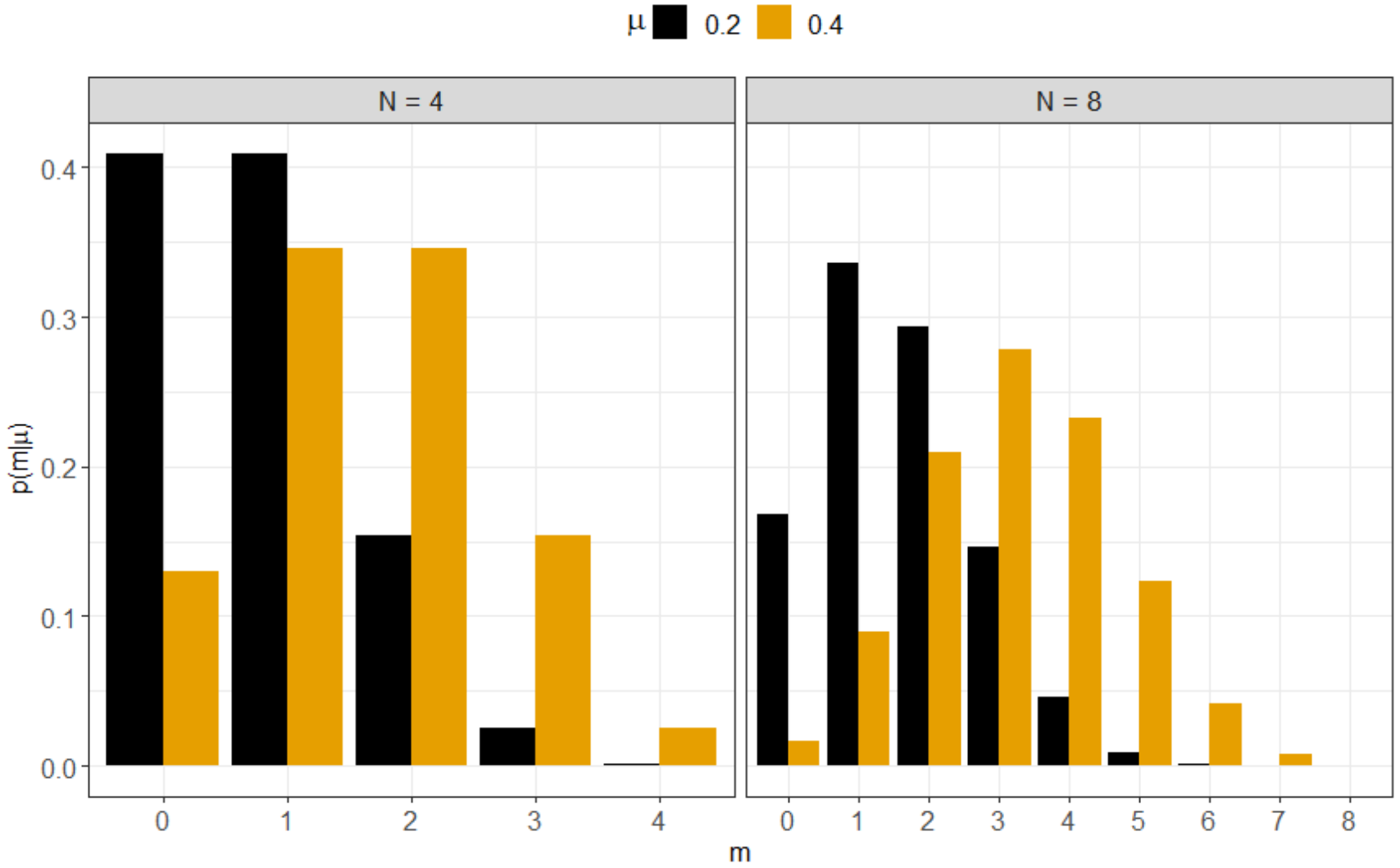
- We recover the Bernoulli distribution with $N = 1$. Thus, $m = \{0,1\}$.

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

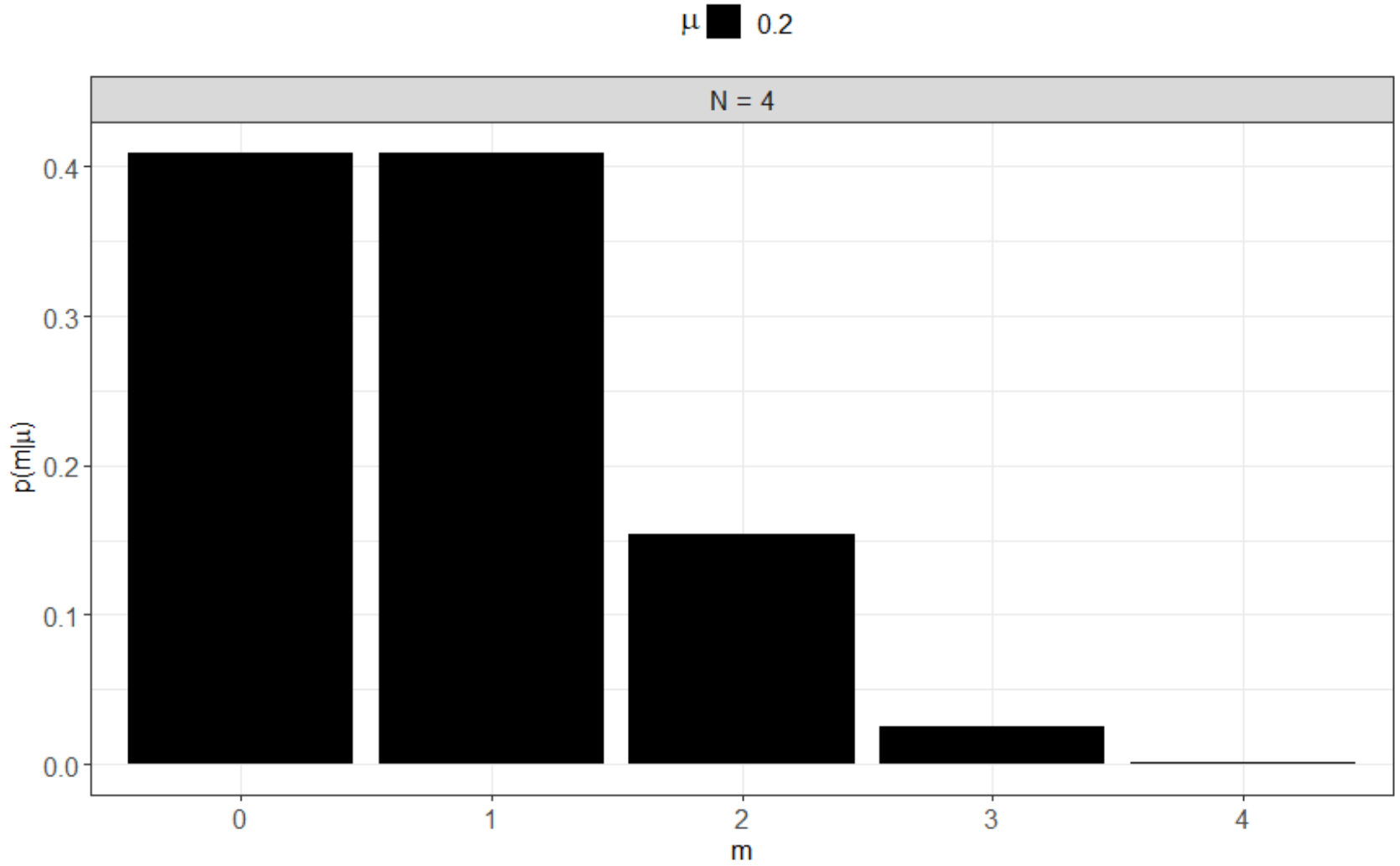
Binomial distribution for $N = 8$ and $\mu = 0.2$



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

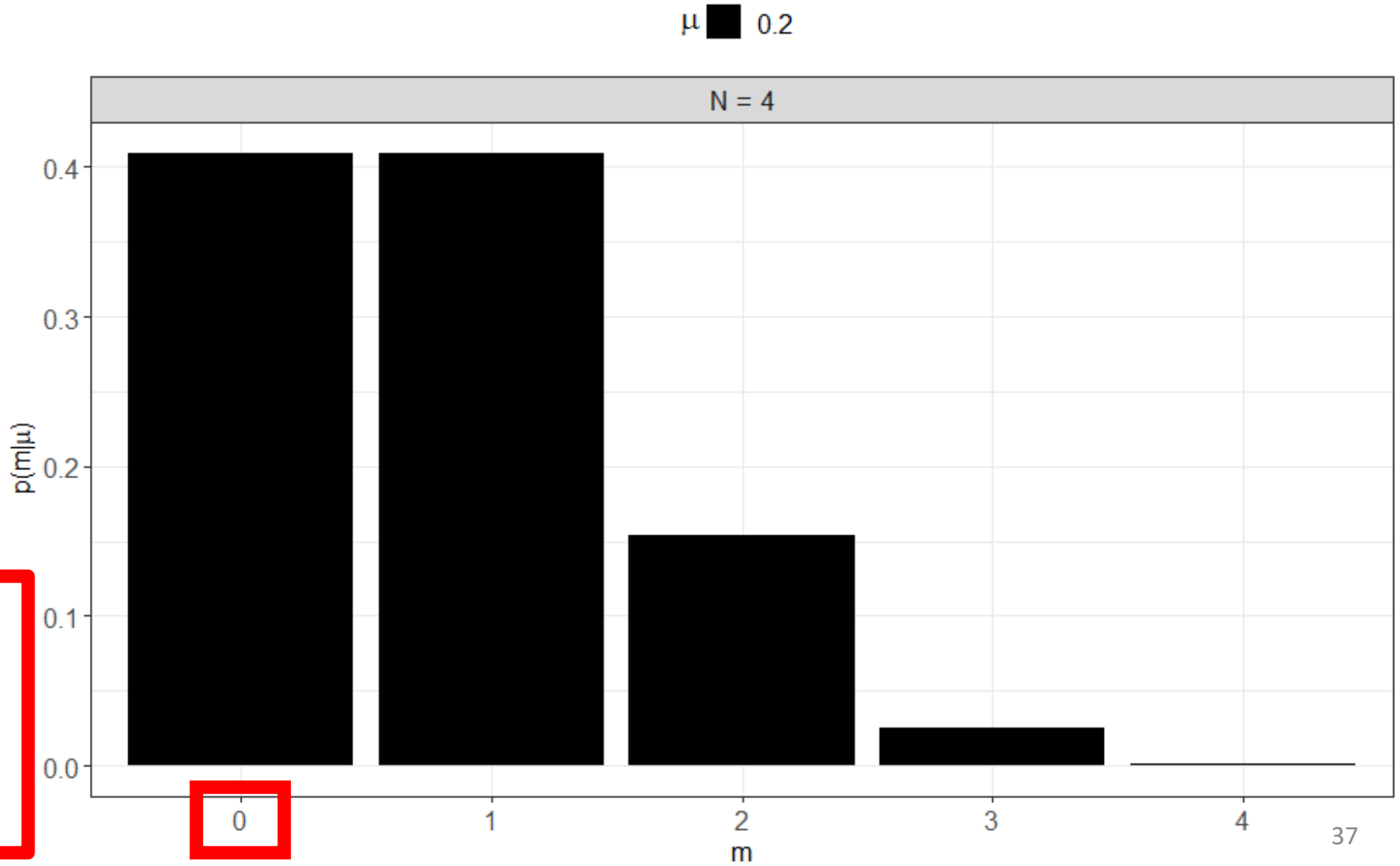


Back to our movie example…let's assume the TRUE probability of Yes is $\mu_{TRUE} = 0.2$



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

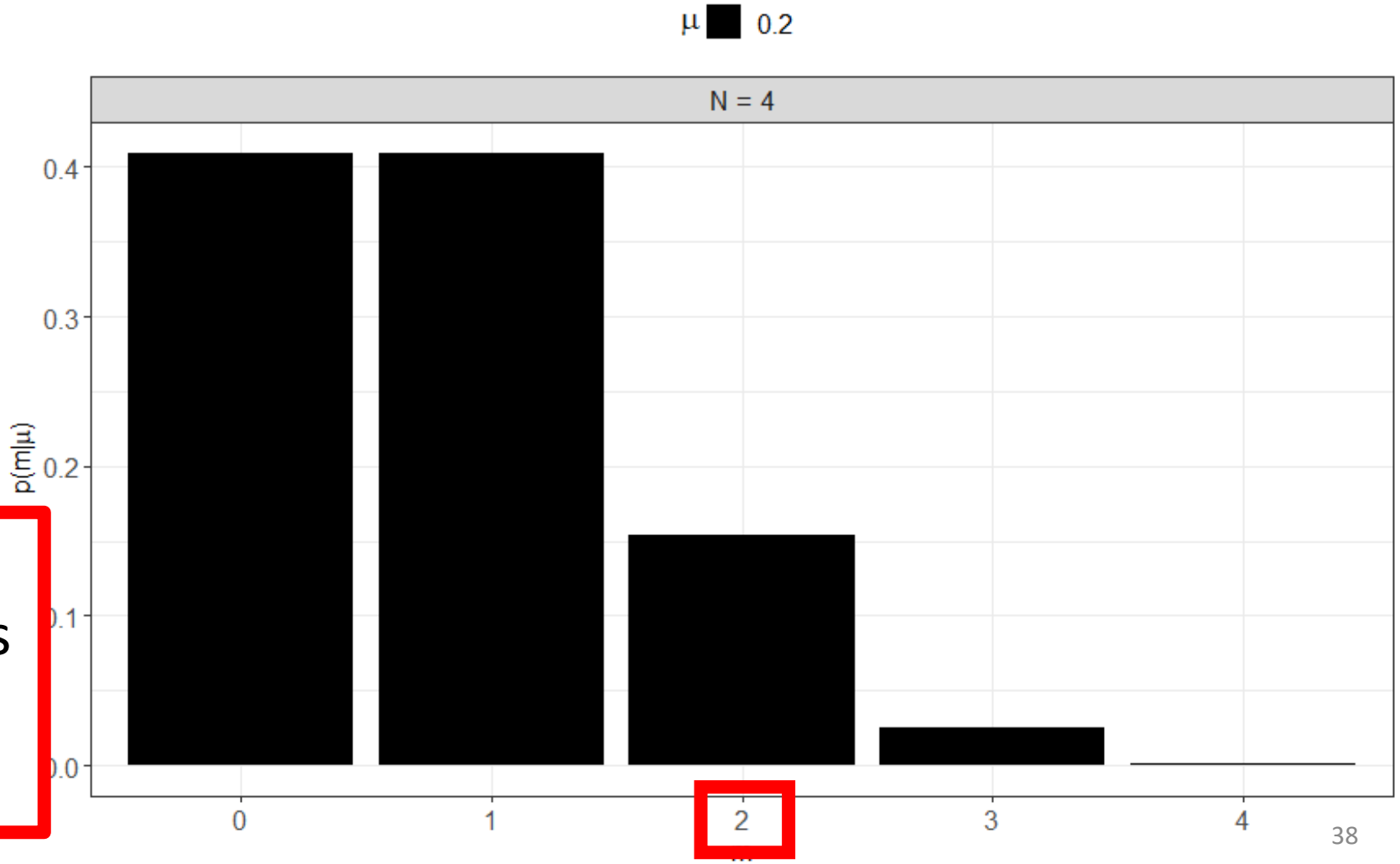
We ask 4 people, so we have 4 trials: $N = 4$



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?