Distributions, Populations, and Samples

Ling250/450: Data Science for Linguistics
C.M. Downey
Spring 2025

Random variables

Random variables

- We often want to know the **probability** of a **specific outcome**
  - E.g. **how likely** is it to get heads on a coin flip? To get a 6 on a die roll? That the Bills will win the Superbowl next year?

Random variables

- We often want to know the **probability** of a **specific outcome**
  - E.g. **how likely** is it to get heads on a coin flip? To get a 6 on a die roll? That the Bills will win the Superbowl next year?
- We model this with a **random variable**, representing possible outcomes
  - Example: X is the result of rolling a die
  - $X = 1$ means the rolled number is 1; $X = 2$ means rolling a 2; etc.
  - $P(X = 1)$ is the **probability that we roll a 1** (1/6 chance)

Random variables

- We often want to know the **probability** of a **specific outcome**
  - E.g. **how likely** is it to get heads on a coin flip? To get a 6 on a die roll? That the Bills will win the Superbowl next year?
- We model this with a **random variable**, representing possible outcomes
  - Example: X is the result of rolling a die
  - $X = 1$ means the rolled number is 1; $X = 2$ means rolling a 2; etc.
  - $P(X = 1)$ is the **probability that we roll a 1** (1/6 chance)
- If the **sum of probabilities** equals 1, this defines a **probability distribution**

Probability distributions

- A probability distribution expresses the **likelihood of all possible outcomes**, which must **add up to 1** (conceptually the same as 100%)
- Can often be expressed in a **table** (if the variable is **discrete**)

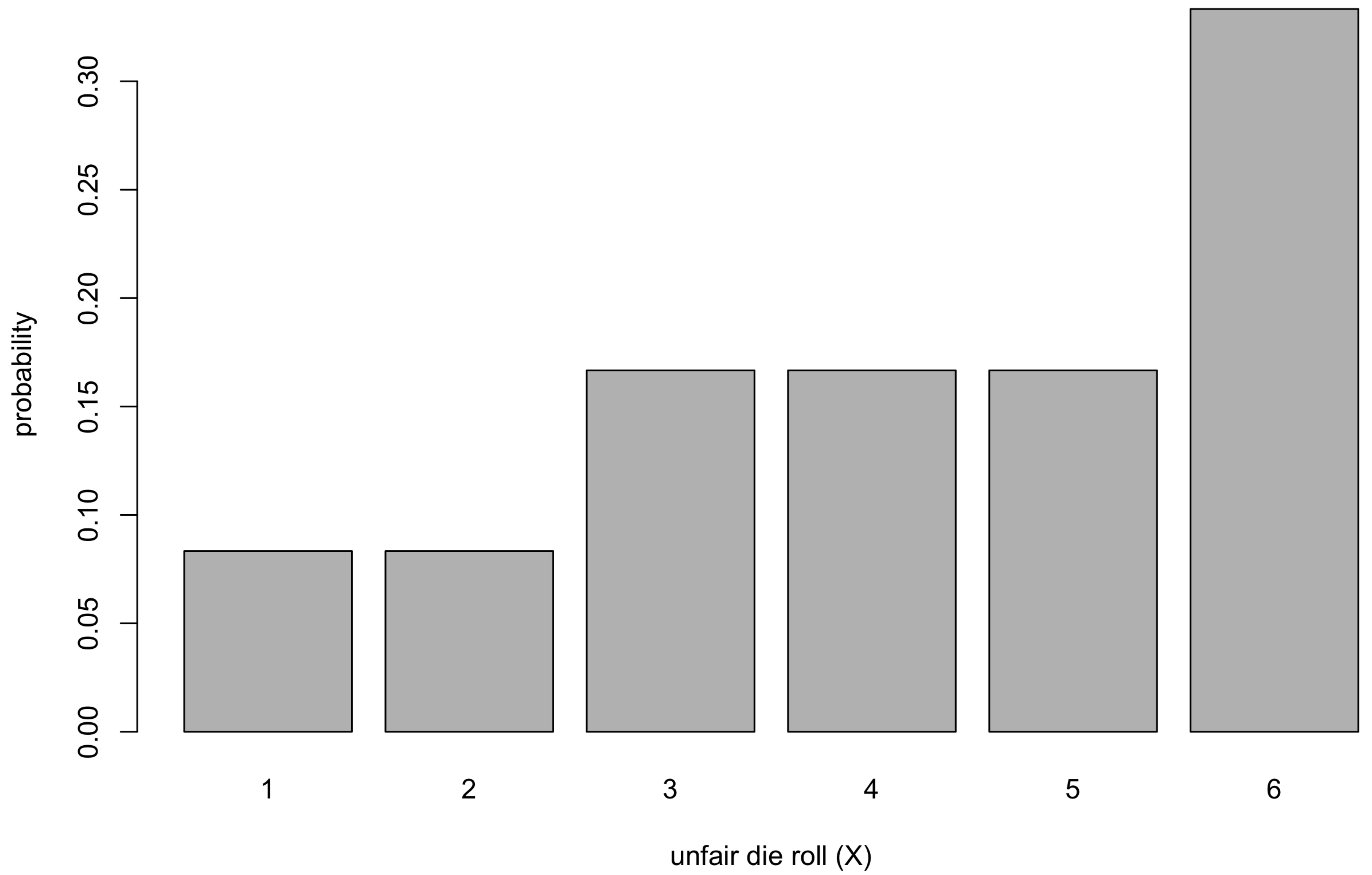
Outcome	Probability
X = 1	1/6
X = 2	1/6
X = 3	1/6
X = 4	1/6
X = 5	1/6
X = 6	1/6

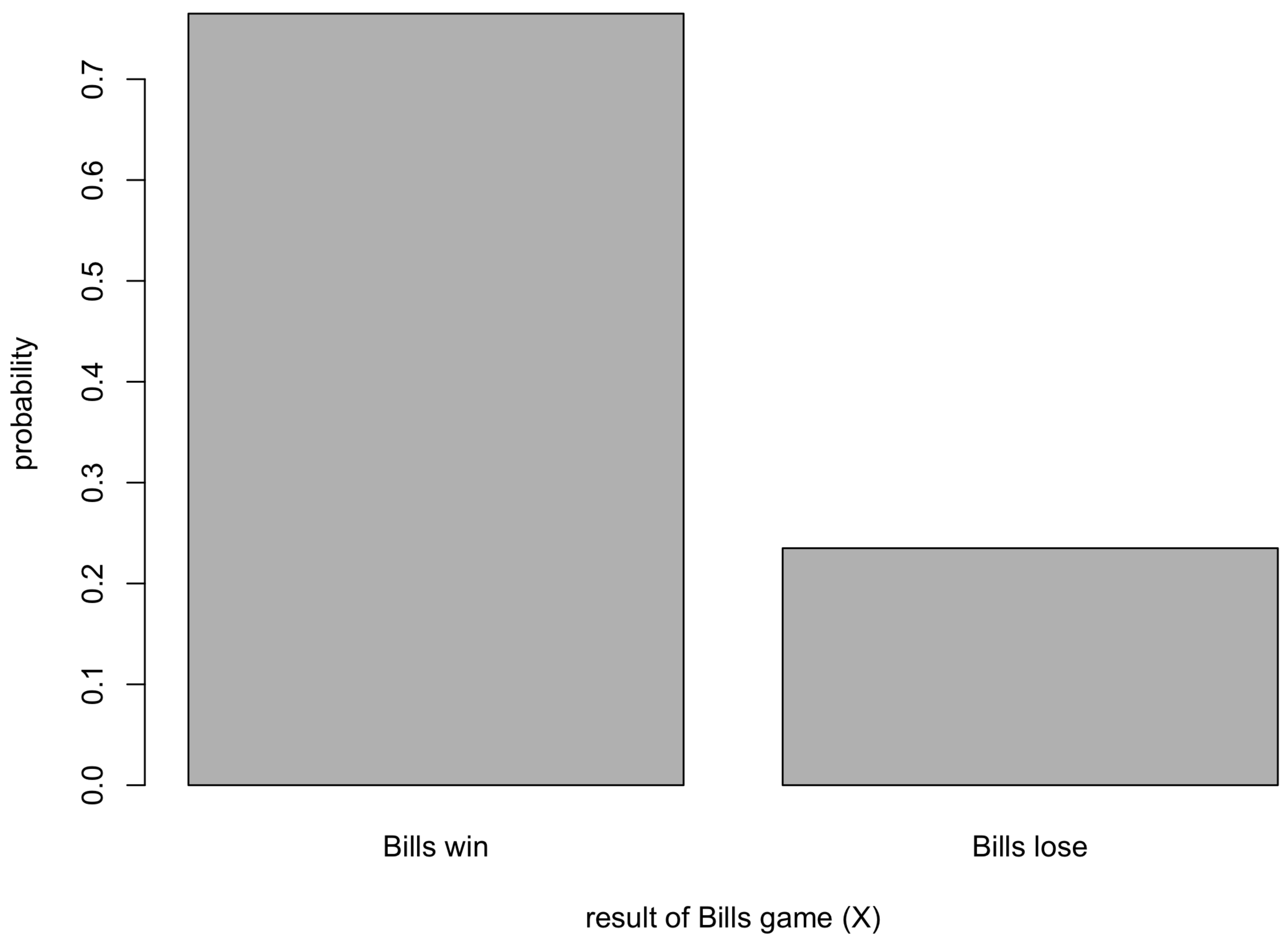
Outcome	Probability
X = heads	1/2
X = tails	1/2

Outcome	Probability
X = Bills win	0.765
X = Bills lose	0.235

Visualizing distributions

- Discrete distributions can also be visualized as a **bar plot**
- This is often called a **Probability Mass Function**





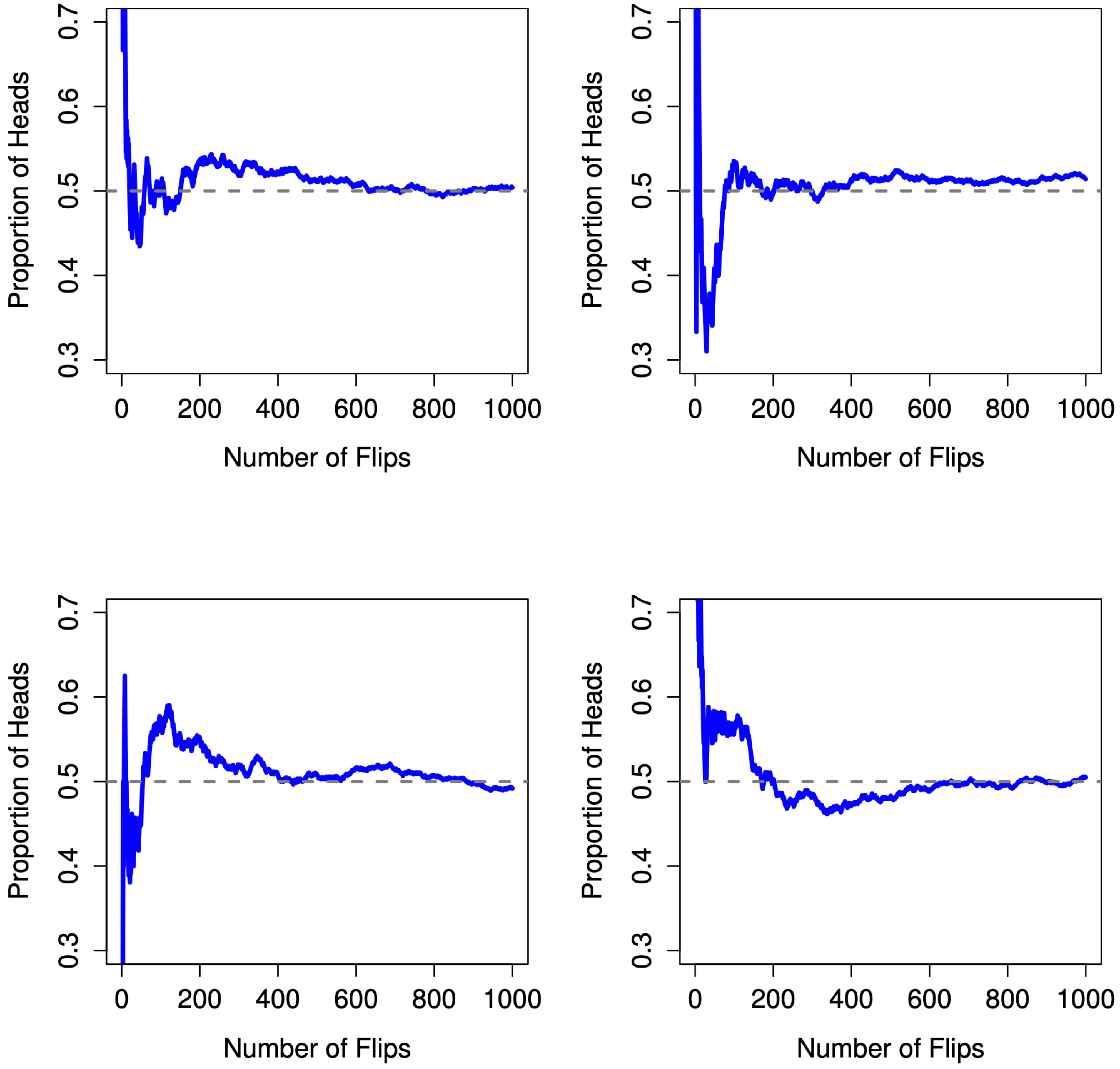
What is a probability anyway?

What is a probability anyway?

- **What do we mean** when we say "the probability of flipping heads is 0.5"?
  - This is **more complex** than we often give it credit for
  - Maybe something like "we **expect** both outcomes to be **equally likely**"

What is a probability anyway?

- **What do we mean** when we say "the probability of flipping heads is 0.5"?
  - This is **more complex** than we often give it credit for
  - Maybe something like "we **expect** both outcomes to be **equally likely**"
- **Frequentist** perspective: it is the **overall frequency** of an outcome when the experiment is **repeated many times**
  - Example: simulating many coin flips



Note on independence

Note on independence

- Two events are considered **independent** if **neither outcome influences the other** (oversimplification)

Note on independence

- Two events are considered **independent** if **neither outcome influences the other** (oversimplification)
- **IF** two events are independent, their **joint probability** is $P(X = x) \cdot P(Y = y)$
  - Example: the probability of getting heads, then heads again is $0.5 \cdot 0.5 = 0.25$
  - This is because we typically **assume coin-flips are independent**

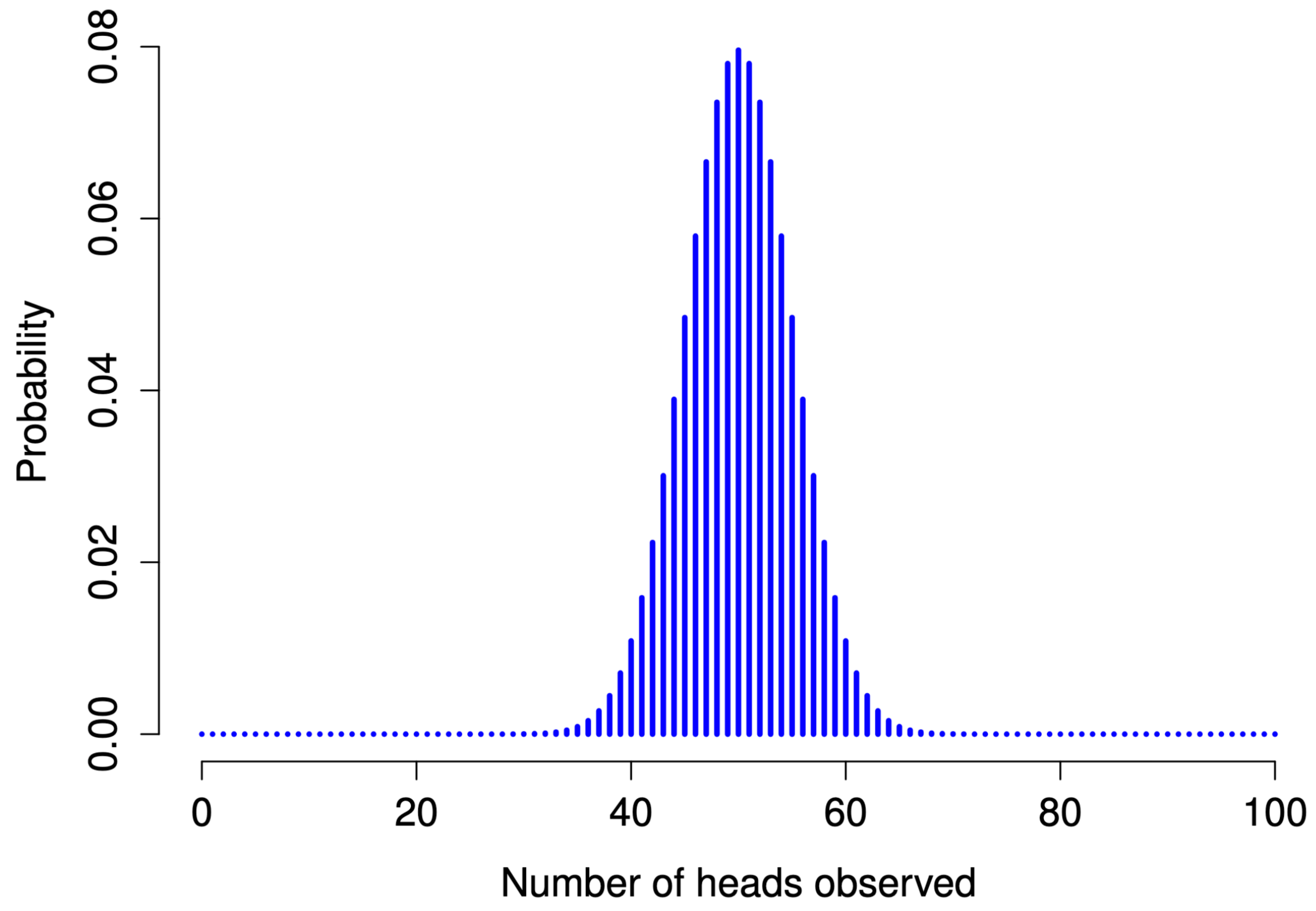
Note on independence

- Two events are considered **independent** if **neither outcome influences the other** (oversimplification)
- **IF** two events are independent, their **joint probability** is $P(X = x) \cdot P(Y = y)$
  - Example: the probability of getting heads, then heads again is $0.5 \cdot 0.5 = 0.25$
  - This is because we typically **assume coin-flips are independent**
- Is the result of the Bills' second game independent of their first?
  - **NO!** This is a bad assumption for many reasons (e.g. injuries, morale)

Important distributions

Binomial Distribution

- Distribution of **number of "successful" outcomes** in a set of **repeated experiments**
- Example: if we flip a coin 100 times, **how many heads** should we expect?
- "Success" just means **one of two possible outcomes** (e.g. heads vs. tails; win vs. lose; pass vs. fail)



Binomial parameters

$$X \sim \text{Binomial}(\theta, N)$$

Binomial parameters

- A Binomial Distribution is defined by several **parameters**

$$X \sim \text{Binomial}(\theta, N)$$

Binomial parameters

- A Binomial Distribution is defined by several **parameters**
  - **Size / Number of trials** (N): the number of times we repeat the event

$$X \sim \text{Binomial}(\theta, N)$$

Binomial parameters

- A Binomial Distribution is defined by several **parameters**
  - **Size / Number of trials** (N): the number of times we repeat the event
  - **Success probability** (θ): the **overall frequency** of whatever we define "success" to be (e.g. heads probability; win probability)

$$X \sim \text{Binomial}(\theta, N)$$

Binomial parameters

- A Binomial Distribution is defined by several **parameters**
  - **Size / Number of trials** (N): the number of times we repeat the event
  - **Success probability** (θ): the **overall frequency** of whatever we define "success" to be (e.g. heads probability; win probability)
  - **Random variable** (X): the variable representing the **overall number of successes**

$$X \sim \text{Binomial}(\theta, N)$$

Binomial parameters

- A Binomial Distribution is defined by several **parameters**
  - **Size / Number of trials** (N): the number of times we repeat the event
  - **Success probability** (θ): the **overall frequency** of whatever we define "success" to be (e.g. heads probability; win probability)
  - **Random variable** (X): the variable representing the **overall number of successes**
- The following can be read "X is distributed according to the Binomial Distribution with success rate θ and N total trials"

$$X \sim \text{Binomial}(\theta, N)$$

Binomial formula

$$P(X) = \frac{N!}{X!(N-X)!}\theta^X(1-\theta)^{(N-X)}$$

Ignore this part for now!

Binomial formula

- P(X): probability of getting X heads over all trials

$$P(X) = \frac{N!}{X!(N-X)!}\theta^X(1-\theta)^{(N-X)}$$

Ignore this part for now!

Binomial formula

- P(X): probability of getting X heads over all trials
- θ: probability of getting a single heads

$$P(X) = \frac{N!}{X!(N-X)!}\theta^{X}(1-\theta)^{(N-X)}$$

Ignore this
part for now!

Binomial formula

- P(X): probability of getting X heads over all trials
- θ: probability of getting a single heads
- $(1-\theta)$: probability of **not** getting heads (getting tails)

$$P(X) = \frac{N!}{X!(N-X)!}\theta^X(1-\theta)^{(N-X)}$$

Ignore this part for now!

Binomial formula

- P(X): probability of getting X heads over all trials
- θ: probability of getting a single heads
- $(1-\theta)$: probability of **not** getting heads (getting tails)
- θ^X: probability of getting heads X times total

$$P(X) = \frac{N!}{X!(N-X)!}\theta^X(1-\theta)^{(N-X)}$$

Ignore this part for now!

Binomial formula

- P(X): probability of getting X heads over all trials
- θ: probability of getting a single heads
- $(1-\theta)$: probability of **not** getting heads (getting tails)
- θ^X: probability of getting heads X times total
- $(1-\theta)^{(N-X)}$: probability of getting tails the other $(N-X)$ times

$$P(X) = \frac{N!}{X!(N-X)!}\theta^X(1-\theta)^{(N-X)}$$

Ignore this part for now!

Binomial Coefficient

$$P(X) = \frac{N!}{X!(N-X)!}\theta^X(1-\theta)^{(N-X)}$$

Binomial Coefficient

- What is this part I told you to ignore at first?
  - Called the **Binomial Coefficient**

$$P(X) = \frac{N!}{X!(N-X)!}\theta^X(1-\theta)^{(N-X)}$$

Binomial Coefficient

- What is this part I told you to ignore at first?
  - Called the **Binomial Coefficient**
- $\theta^X(1-\theta)^{(N-X)}$ gives the probability for a **specific** sequence of X heads and $N-X$ tails
  - Ex: the sequence [H, T, T, H] has probability $\theta^2(1-\theta)^2$
  - [T, T, H, H] has the **exact same probability**

$$P(X) = \frac{N!}{X!(N-X)!}\theta^X(1-\theta)^{(N-X)}$$

Binomial Coefficient

- What is this part I told you to ignore at first?
  - Called the **Binomial Coefficient**
- $\theta^X(1-\theta)^{(N-X)}$ gives the probability for a **specific** sequence of X heads and $N-X$ tails
  - Ex: the sequence [H, T, T, H] has probability $\theta^2(1-\theta)^2$
  - [T, T, H, H] has the **exact same probability**
- Binomial Coefficient gives the **number of distinct ways** to get X heads and $(N-X)$ tails

$$P(X) = \frac{N!}{X!(N-X)!}\theta^X(1-\theta)^{(N-X)}$$

On equations

On equations

- I'm showing the Binomial formula because it's **relatively** intuitive

On equations

- I'm showing the Binomial formula because it's **relatively** intuitive
- It is generally **not important to remember the formula**
  - It is **especially not important** to remember the formula for the Binomial Coefficient! Don't worry about it!

On equations

- I'm showing the Binomial formula because it's **relatively** intuitive
- It is generally **not important to remember the formula**
  - It is **especially not important** to remember the formula for the Binomial Coefficient! Don't worry about it!
- **More important**: remember what the parts of the equation **mean!**

On equations

- I'm showing the Binomial formula because it's **relatively** intuitive
- It is generally **not important to remember the formula**
    - It is **especially not important** to remember the formula for the Binomial Coefficient! Don't worry about it!
- **More important**: remember what the parts of the equation **mean!**
- Understanding the gist of it is more important than memorizing

Binomial Distribution in R

Binomial Distribution in R

- Specific **probabilities** from the binomial distribution can be found with the command `dbinom()`

Binomial Distribution in R

- Specific **probabilities** from the binomial distribution can be found with the command `dbinom()`
- Syntax: `dbinom(x=, size=, prob=)`
    - `x`: the **outcome value** for which you want the probability (e.g. number of heads)
    - `size`: the total **number of trials** (N)
    - `prob`: the individual **probability of "success"** (θ)

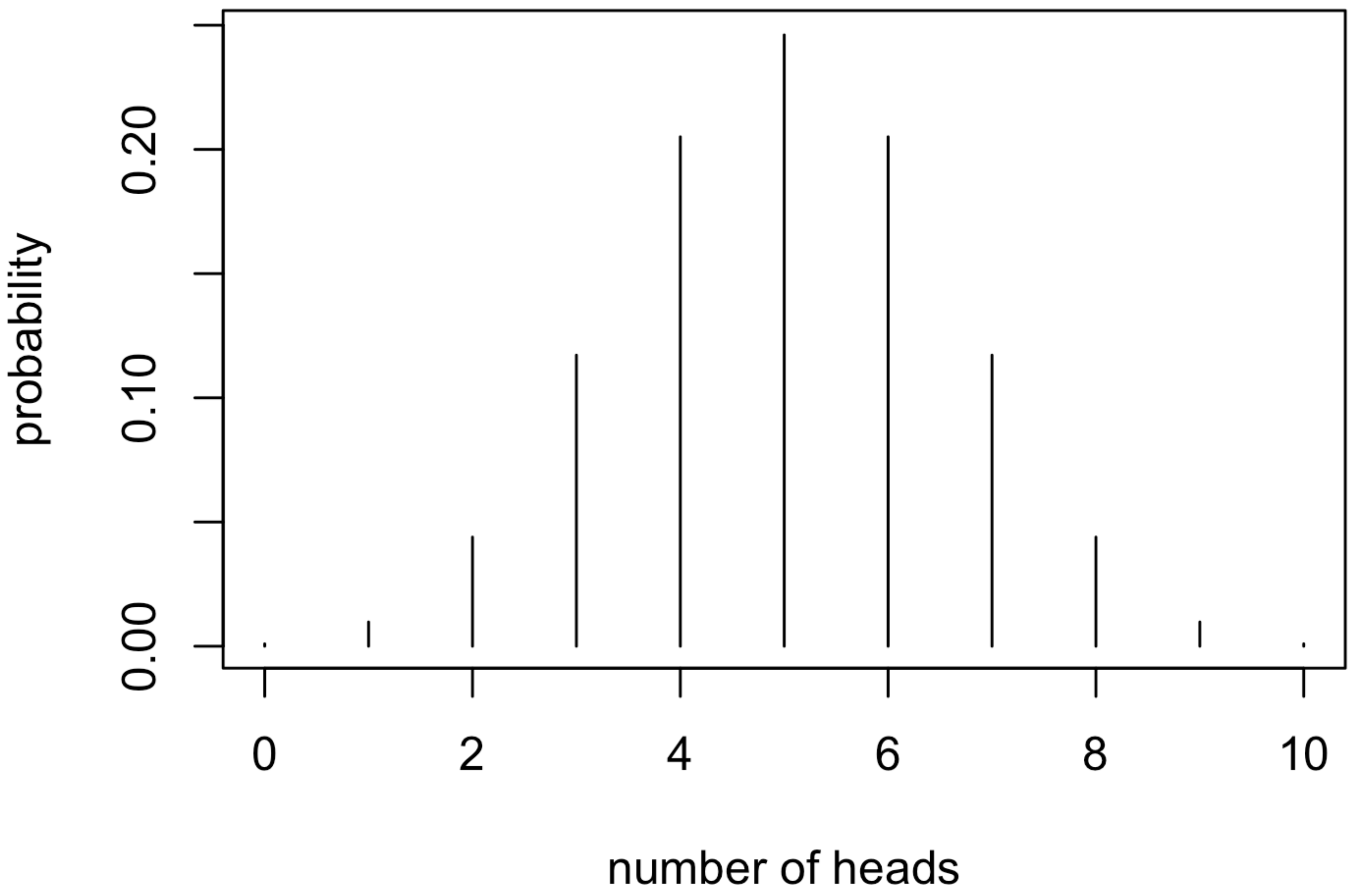
Binomial Distribution in R

- Specific **probabilities** from the binomial distribution can be found with the command `dbinom()`
- Syntax: `dbinom(x=, size=, prob=)`
  - `x`: the **outcome value** for which you want the probability (e.g. number of heads)
  - `size`: the total **number of trials** (N)
  - `prob`: the individual **probability of "success"** (θ)
- Probability that we get 50 heads out of 100 flips of a fair coin:

```
> dbinom(x=50, size=100, prob=0.5)
[1] 0.07958924
```

Binomial Distribution in R

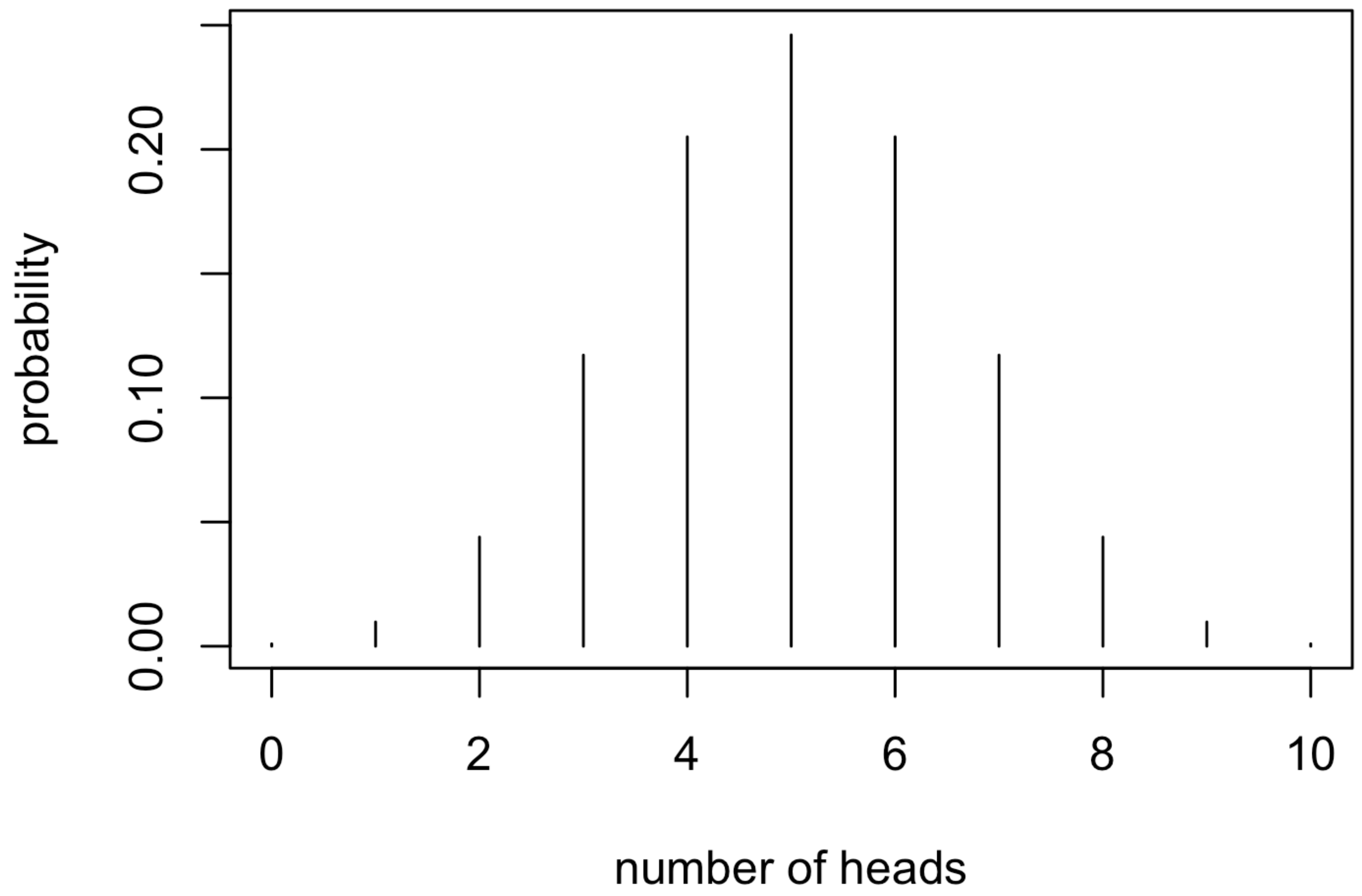
```
> ten_binom = dbinom(x=c(0:10), size=10, prob=0.5)
> ten_binom
 [1] 0.0009765625 0.0097656250 0.0439453125 0.1171875000
 [5] 0.2050781250 0.2460937500 0.2050781250 0.1171875000
 [9] 0.0439453125 0.0097656250 0.0009765625
> plot(x=c(0:10), y=ten_binom, type='h', xlab="number of heads", yl
ab="probability")
```



Binomial Distribution in R

- If you provide a **vector** as the argument for x, you will **get a vector out**
  - I.e. the probability for **each** of the input values

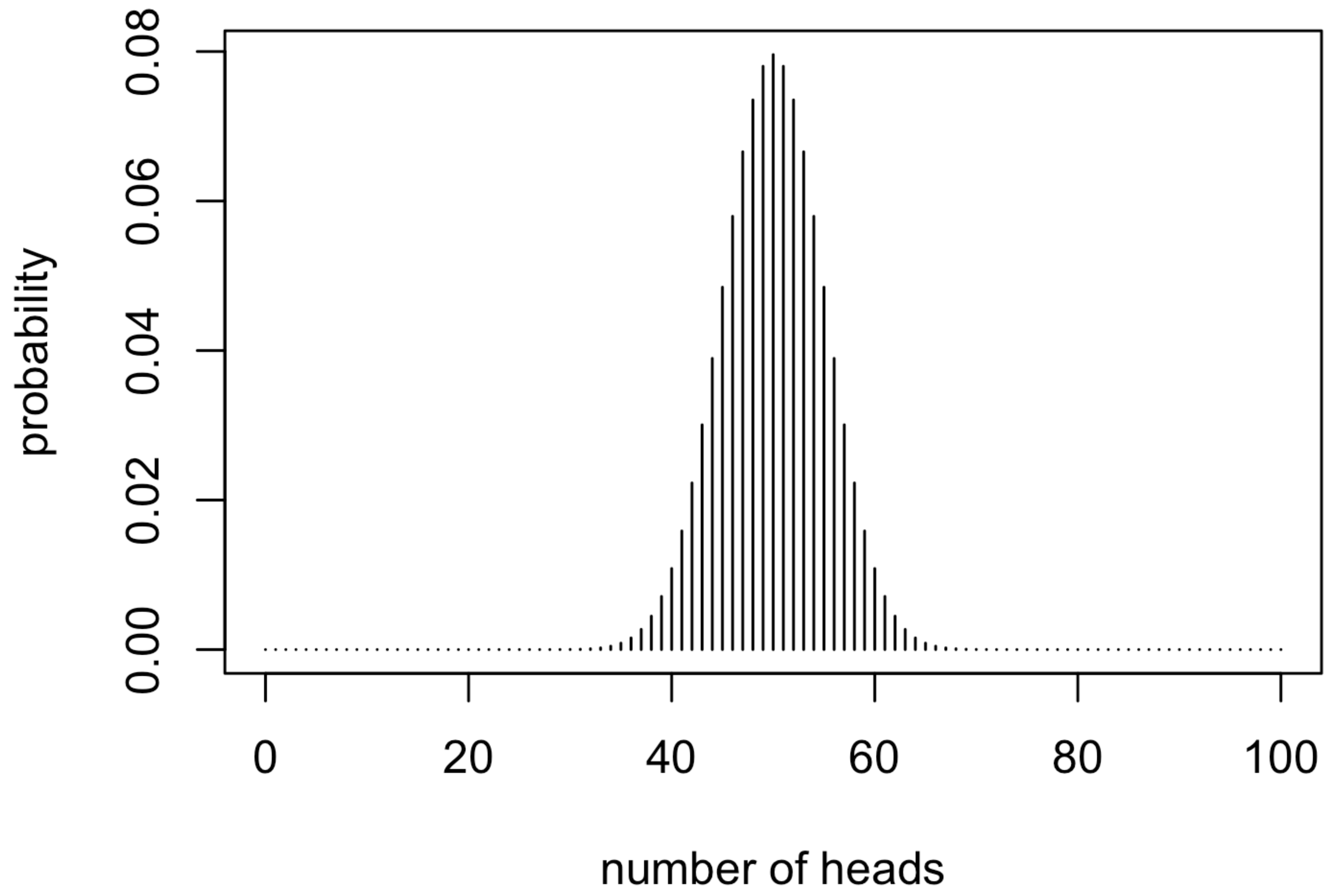
```
> ten_binom = dbinom(x=c(0:10), size=10, prob=0.5)
> ten_binom
 [1] 0.0009765625 0.0097656250 0.0439453125 0.1171875000
 [5] 0.2050781250 0.2460937500 0.2050781250 0.1171875000
 [9] 0.0439453125 0.0097656250 0.0009765625
> plot(x=c(0:10), y=ten_binom, type='h', xlab="number of heads", yl
ab="probability")
```



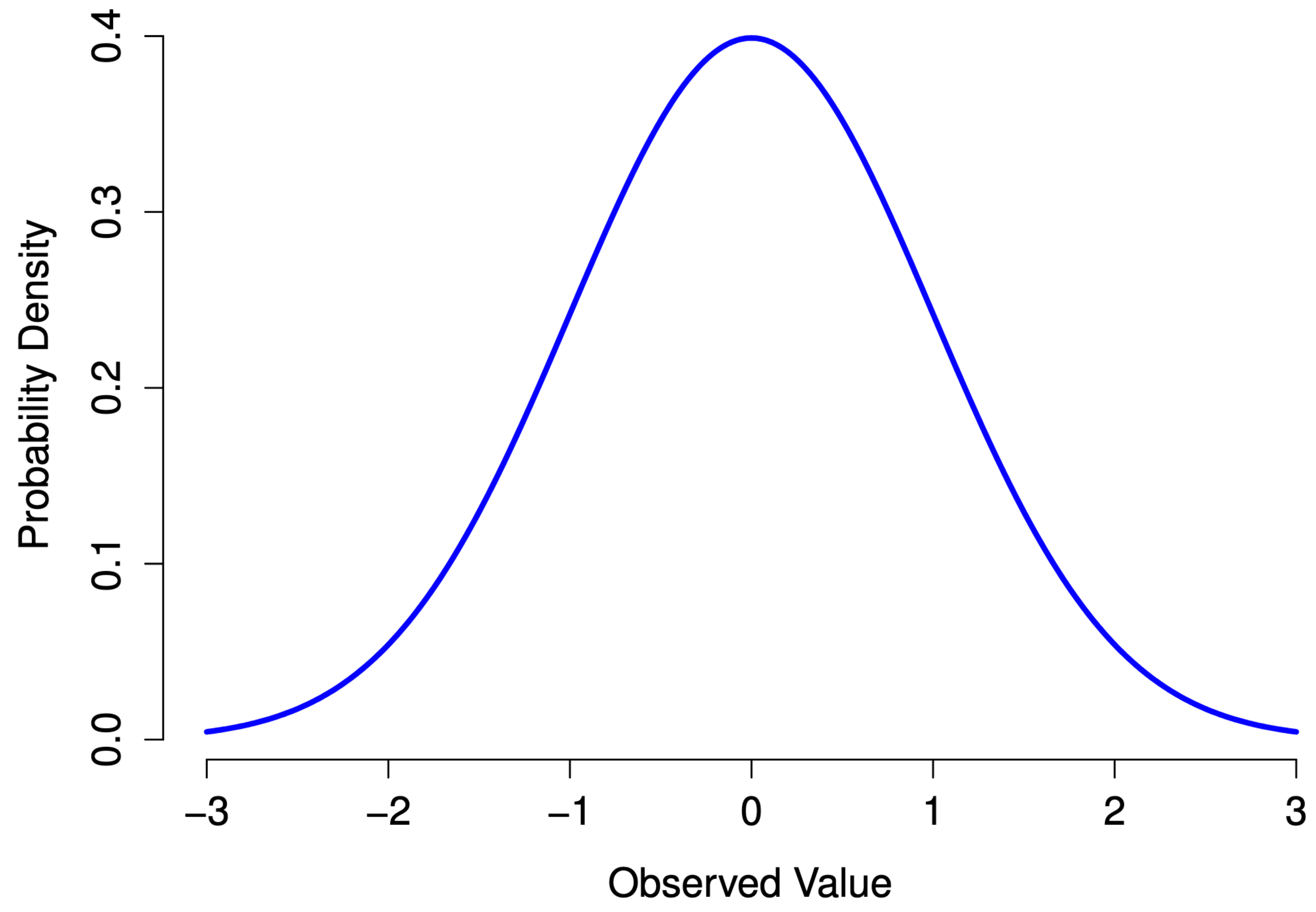
Binomial Distribution in R

- If you provide a **vector** as the argument for x, you will **get a vector out**
  - I.e. the probability for **each** of the input values
- This is the easiest way to **plot** a distribution in R
  - (There are better-looking ways though)

```
> hundred_binom = dbinom(x=c(0:100), size=100, prob=0.5)
> plot(x=c(0:100), y=hundred_binom, type='h', xlab="number of head
s", ylab="probability")
```

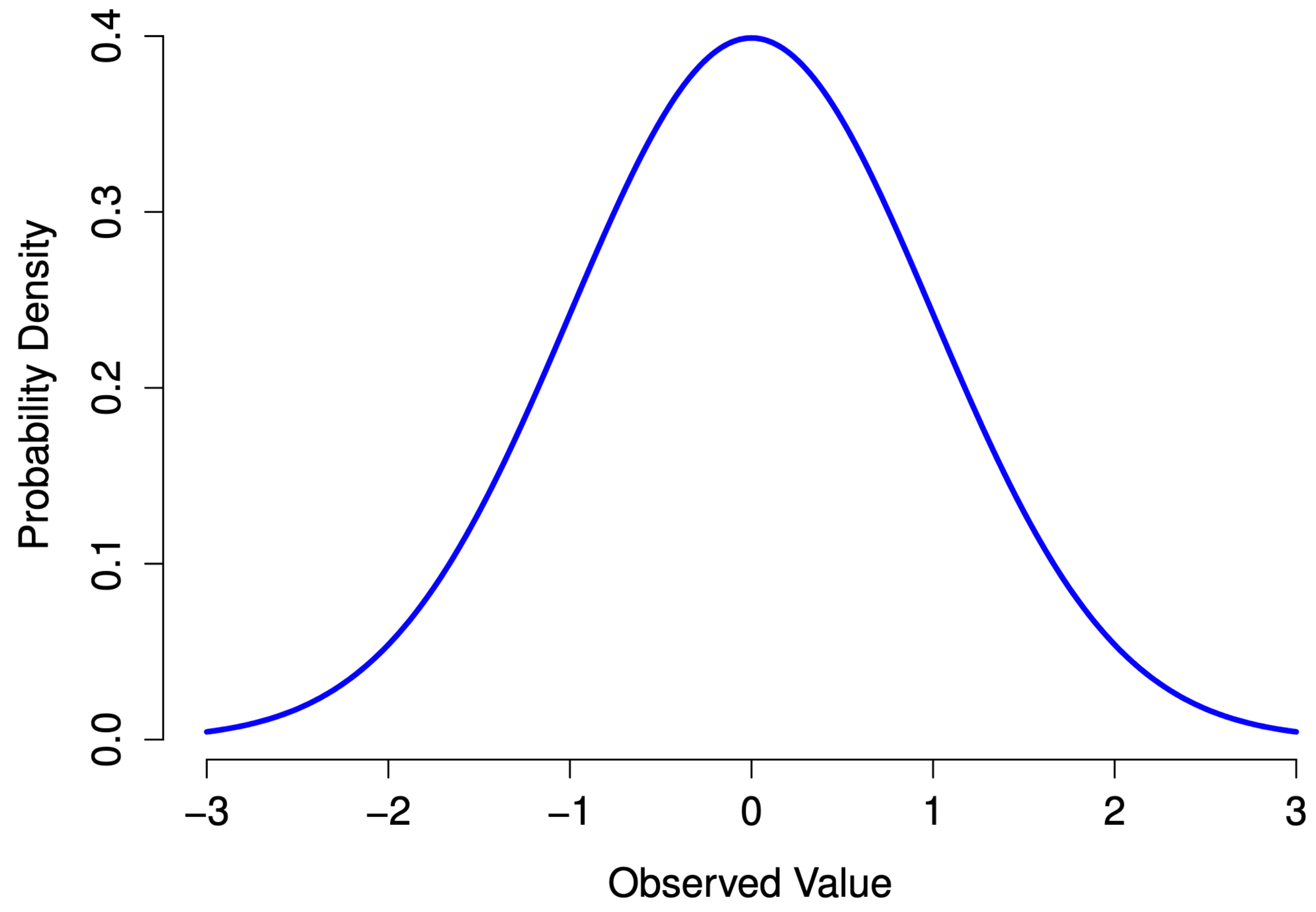


Normal Distribution



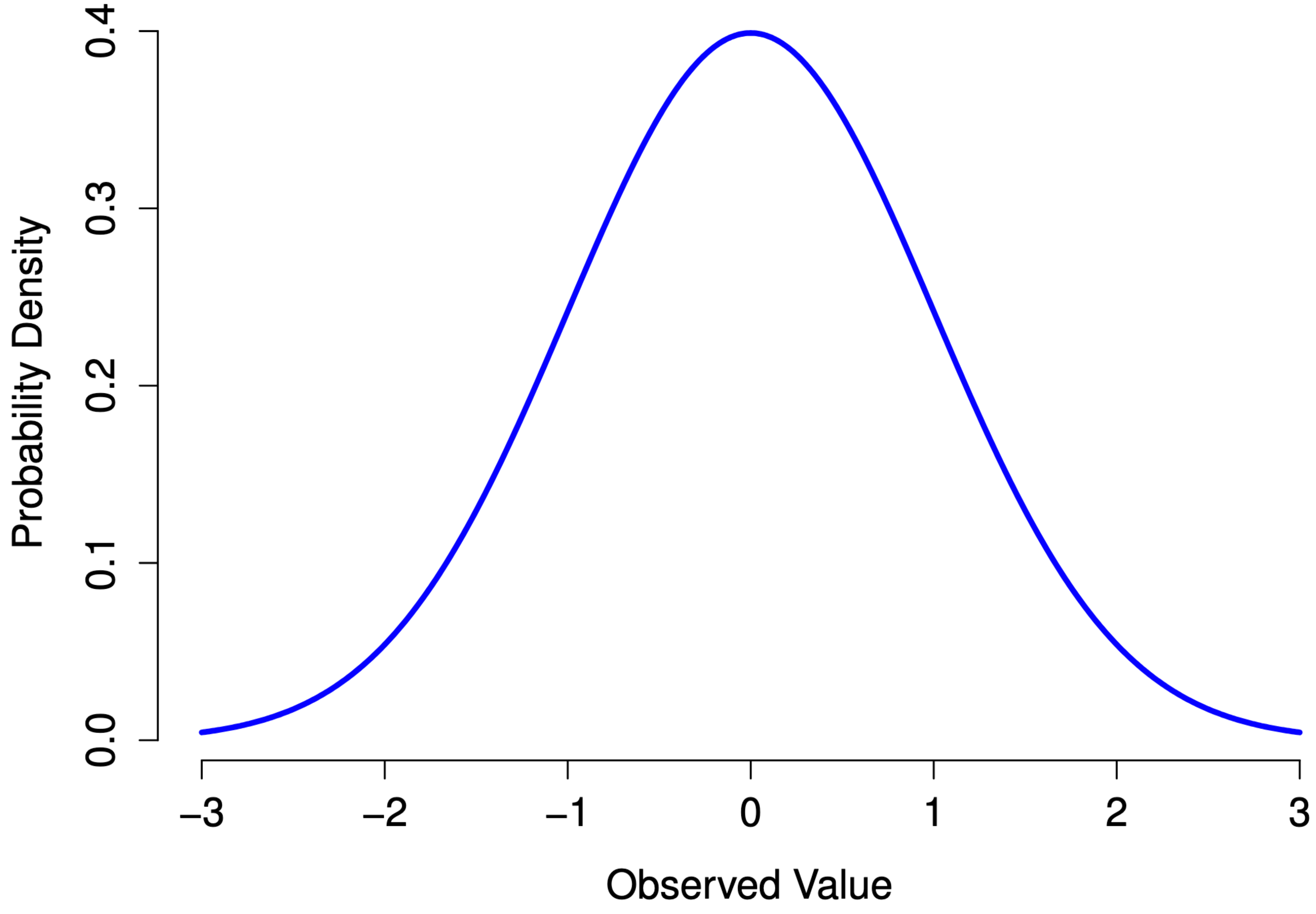
Normal Distribution

- The Normal Distribution is the **most important** distribution in statistics



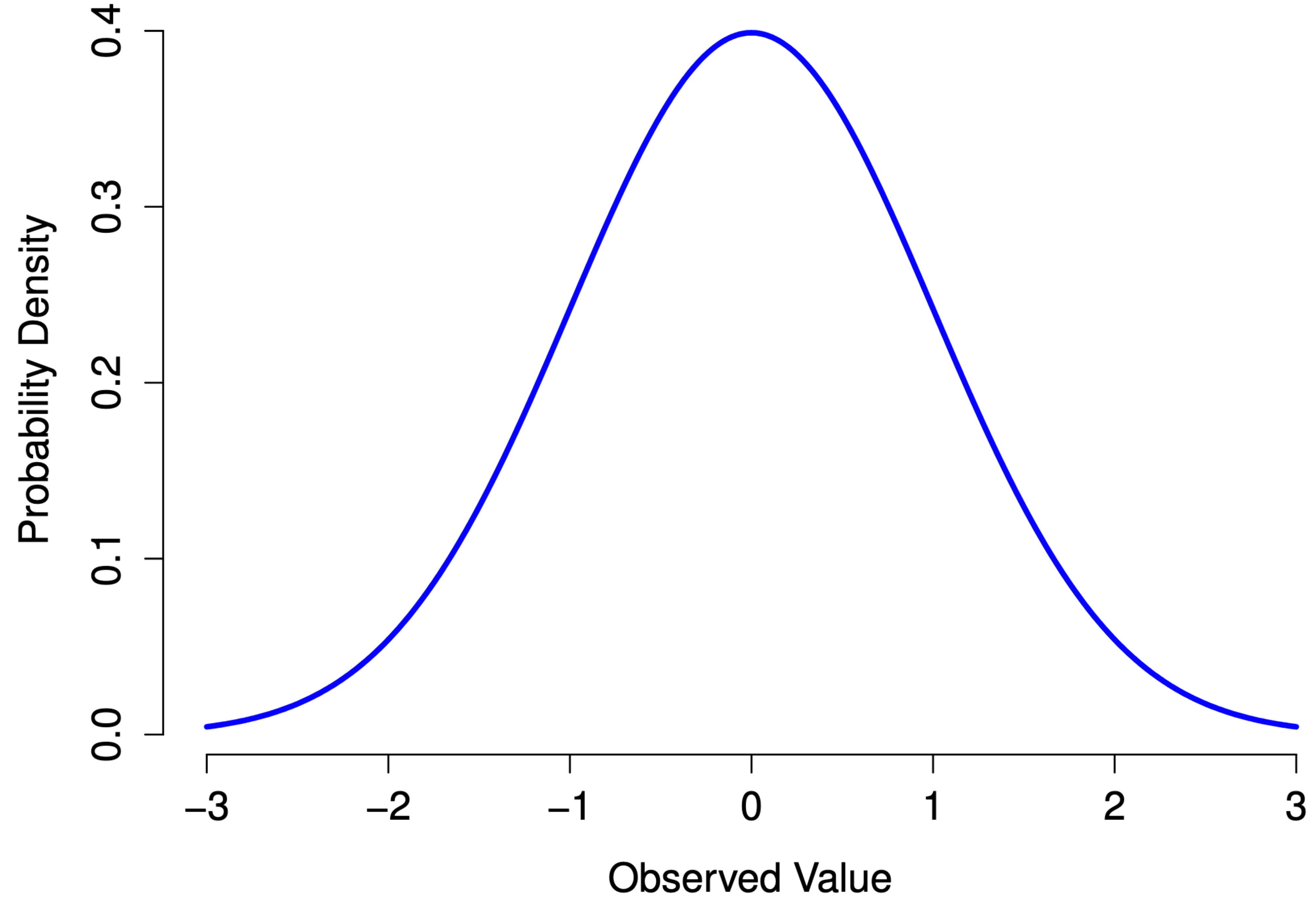
Normal Distribution

- The Normal Distribution is the **most important** distribution in statistics
- **Not** the easiest to understand though
    - Showing you the equation would probably confuse you



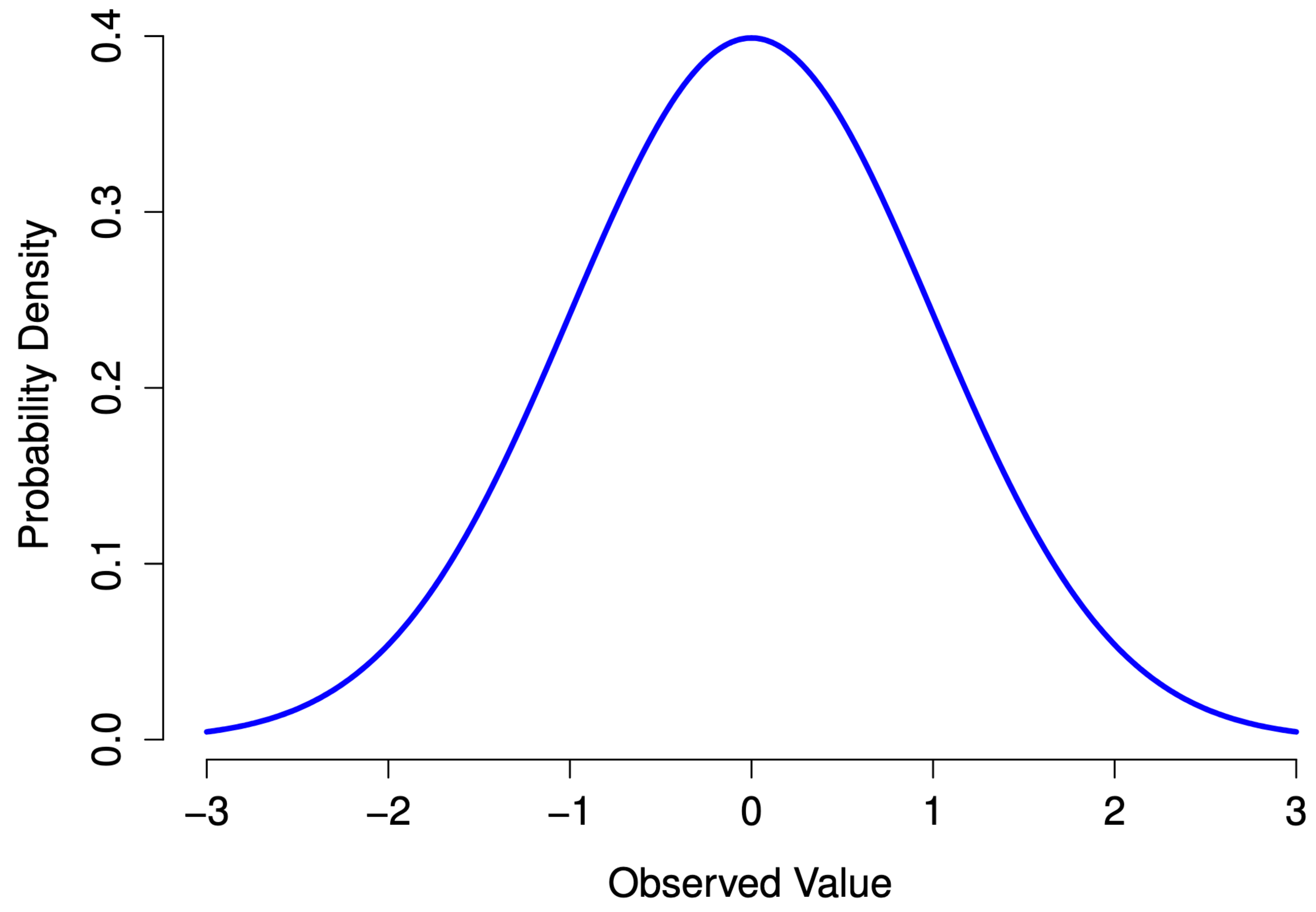
Normal Distribution

- The Normal Distribution is the **most important** distribution in statistics
- **Not** the easiest to understand though
  - Showing you the equation would probably confuse you
- Also called the **"Gaussian Distribution"** or **"bell curve"**

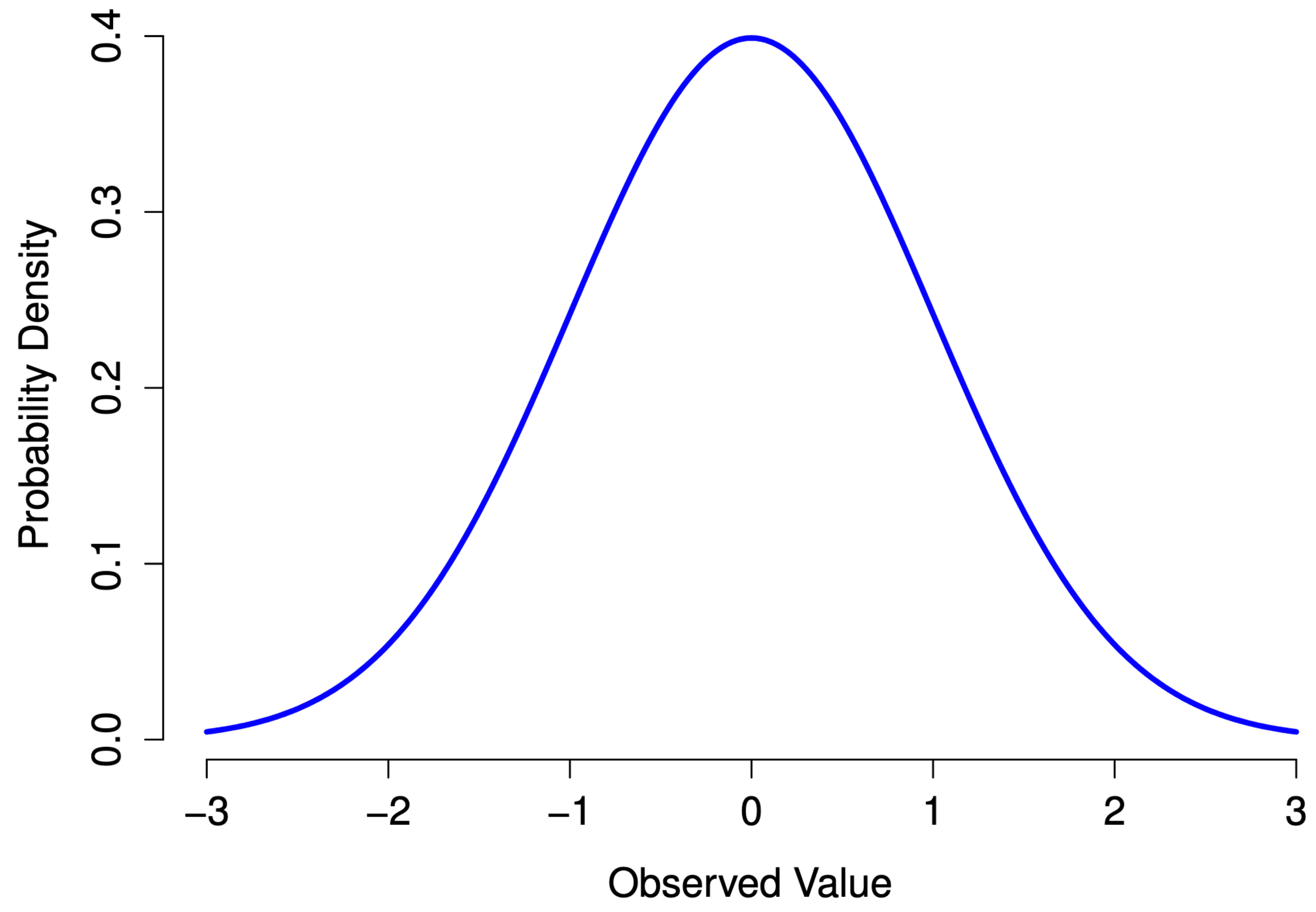


Normal Distribution

- The Normal Distribution is the **most important** distribution in statistics
- **Not** the easiest to understand though
  - Showing you the equation would probably confuse you
- Also called the **"Gaussian Distribution"** or **"bell curve"**
- Has important **math properties**, and **occurs in nature**

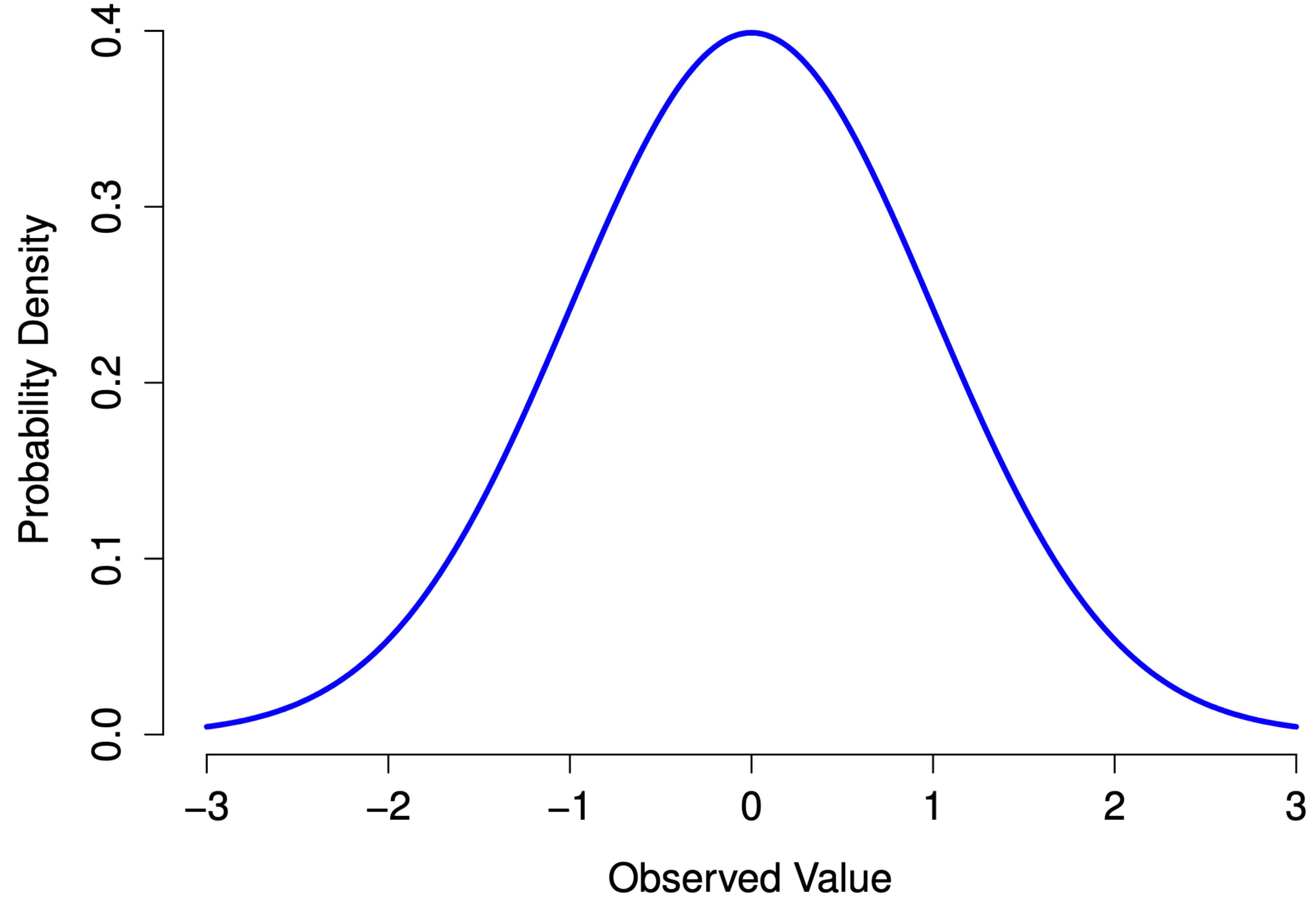


Normal Distribution



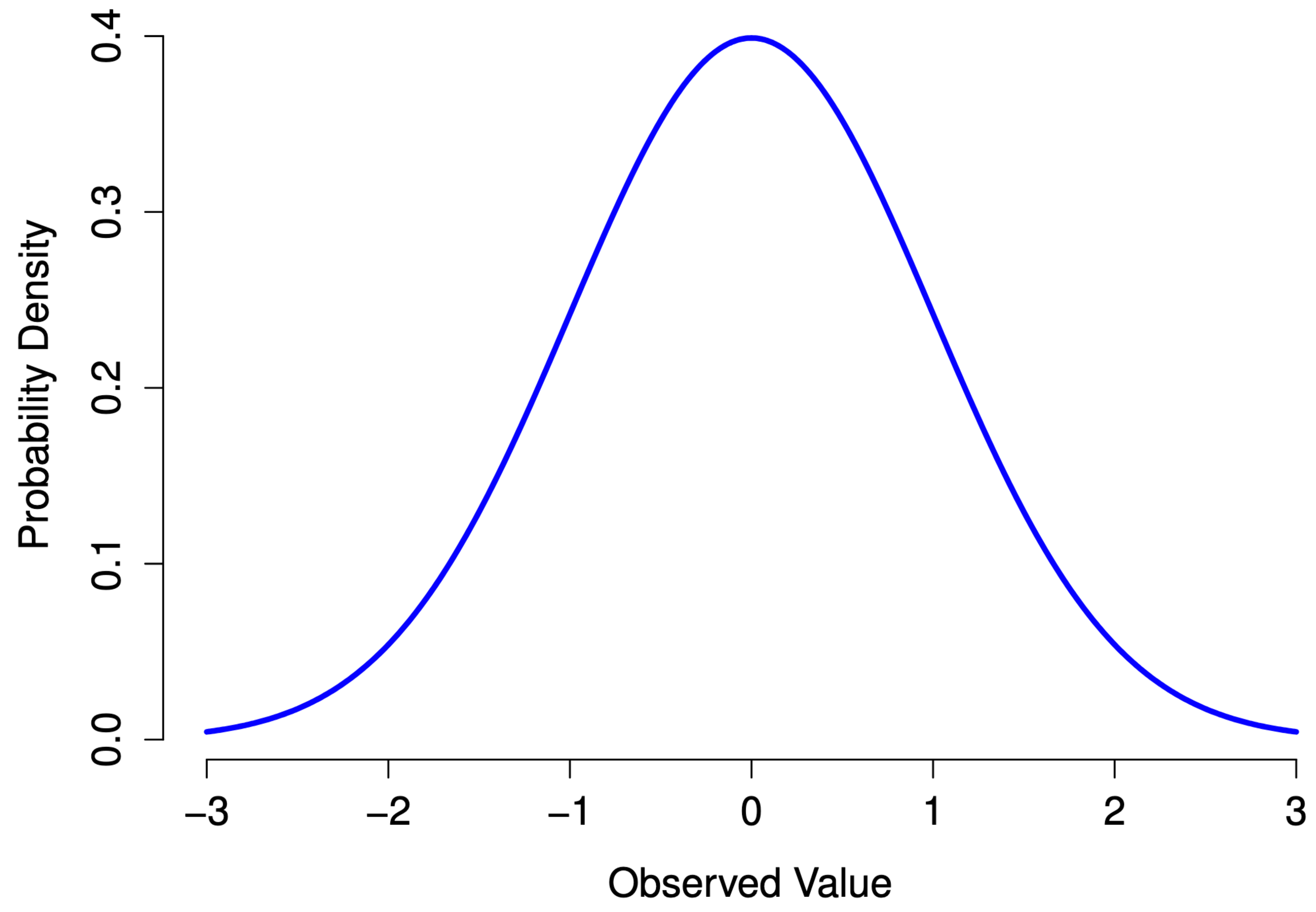
Normal Distribution

- Parameterized by its **mean** (μ) and **standard deviation** (σ)



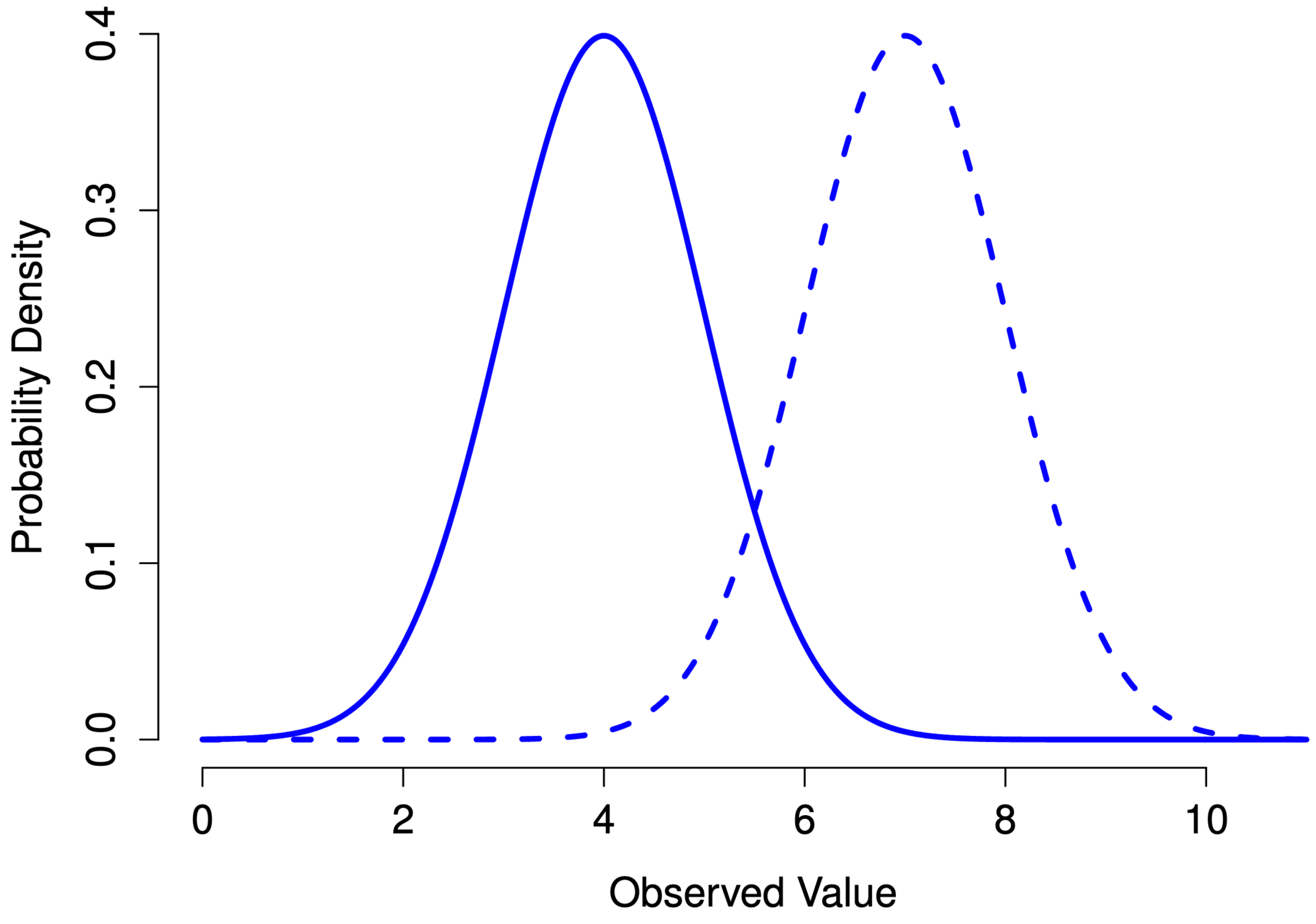
Normal Distribution

- Parameterized by its **mean** (μ) and **standard deviation** (σ)
  - Notated $X \sim \text{Normal}(\mu, \sigma)$



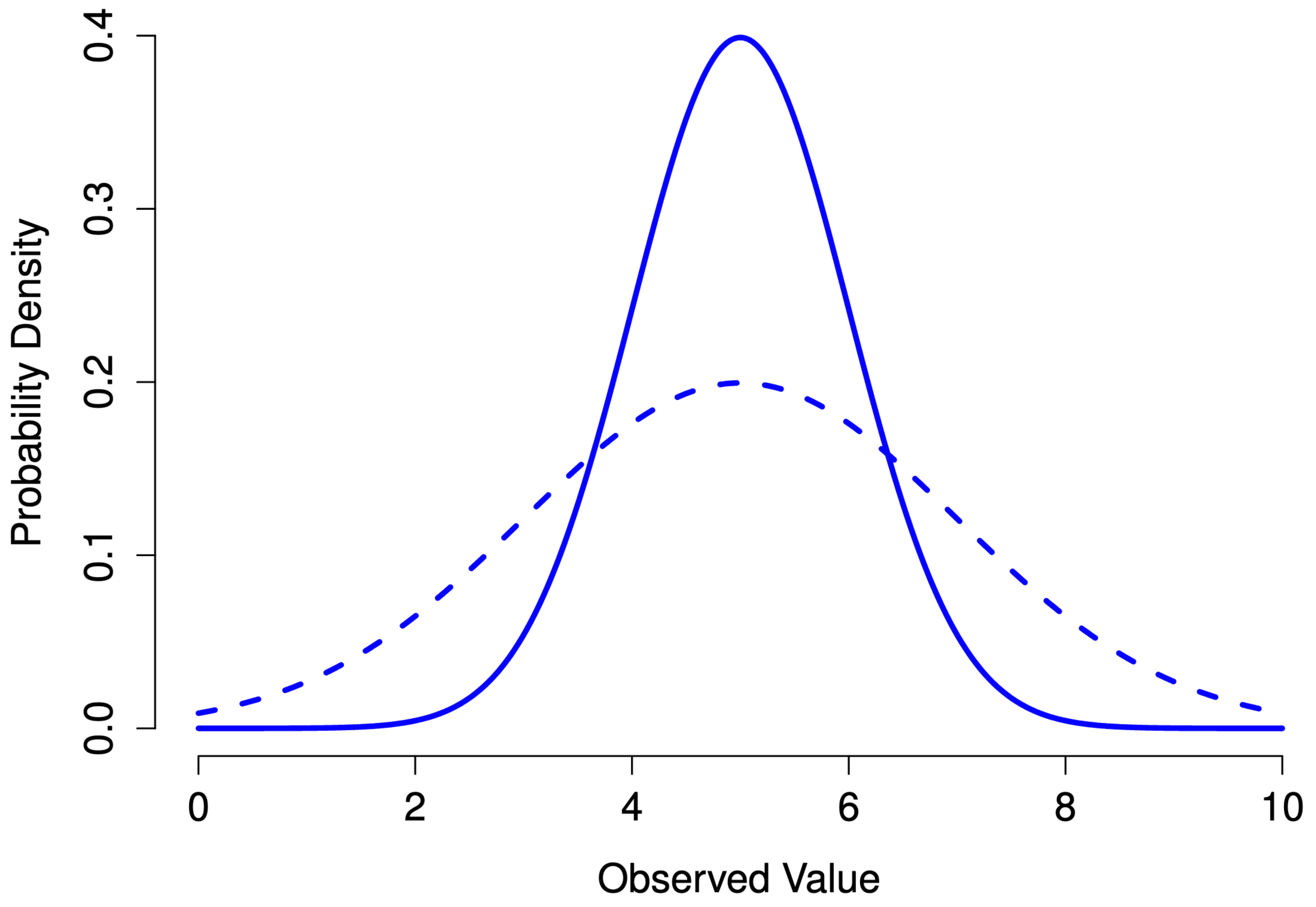
Normal Distribution

- Parameterized by its **mean** (μ) and **standard deviation** (σ)
  - Notated $X \sim \text{Normal}(\mu, \sigma)$
  - Different μ **shifts** the curve left/right



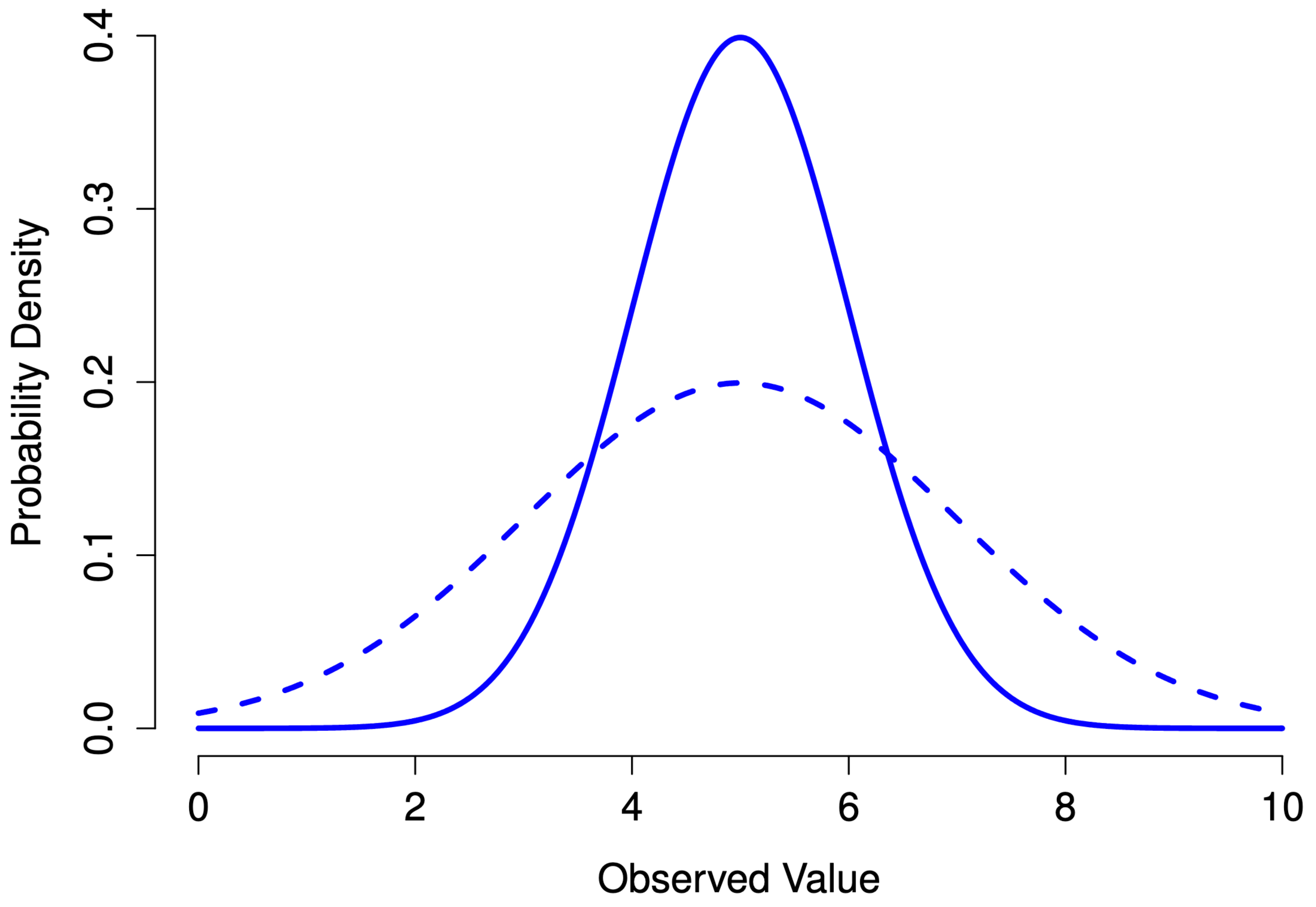
Normal Distribution

- Parameterized by its **mean** (μ) and **standard deviation** (σ)
  - Notated $X \sim \text{Normal}(\mu, \sigma)$
  - Different μ **shifts** the curve left/right
  - Higher σ **"flattens"** the curve



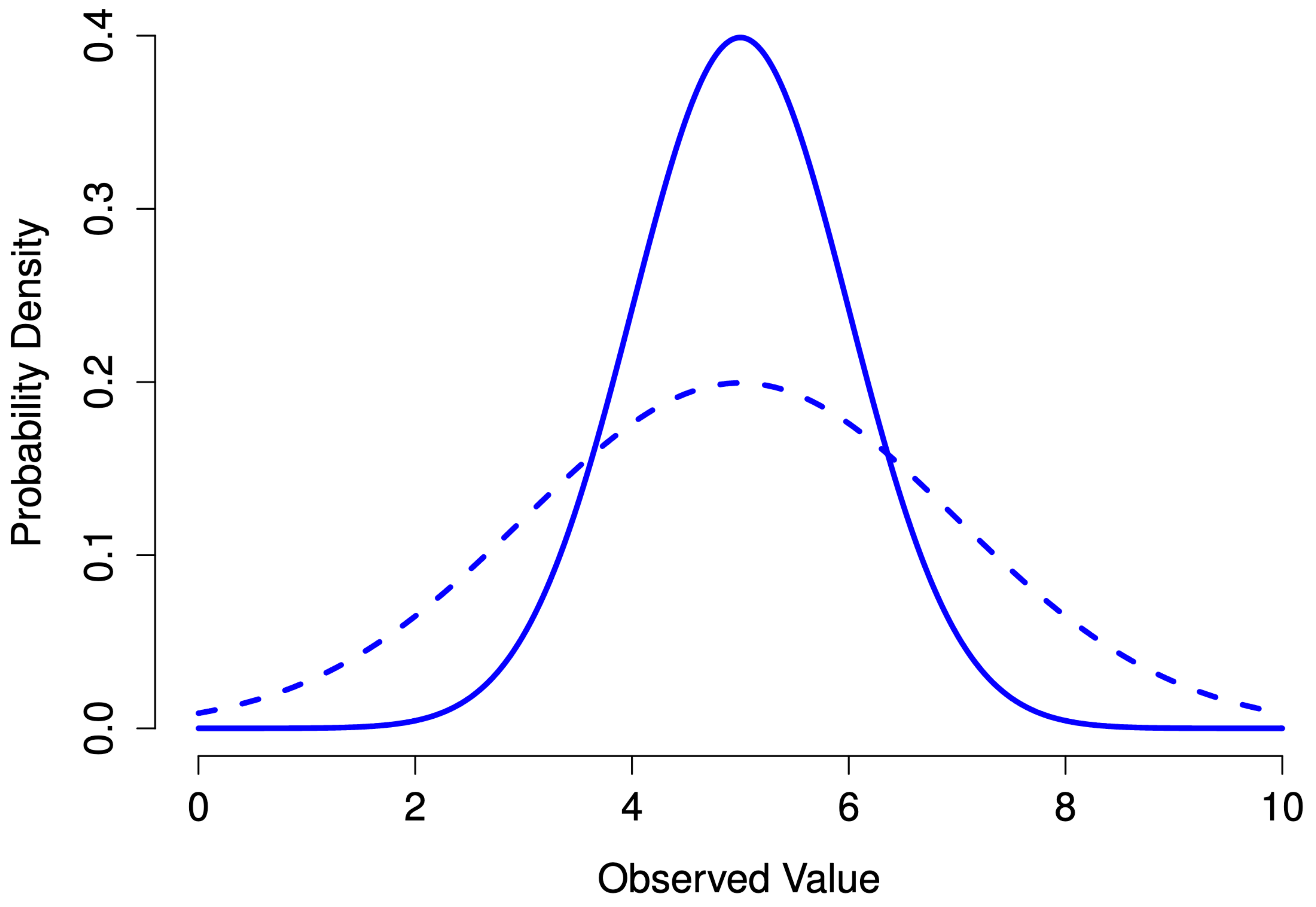
Normal Distribution

- Parameterized by its **mean** (μ) and **standard deviation** (σ)
  - Notated $X \sim$ Normal(μ, σ)
  - Different μ **shifts** the curve left/right
  - Higher σ **"flattens"** the curve
- Unlike the Binomial, the Normal is a **continuous** distribution



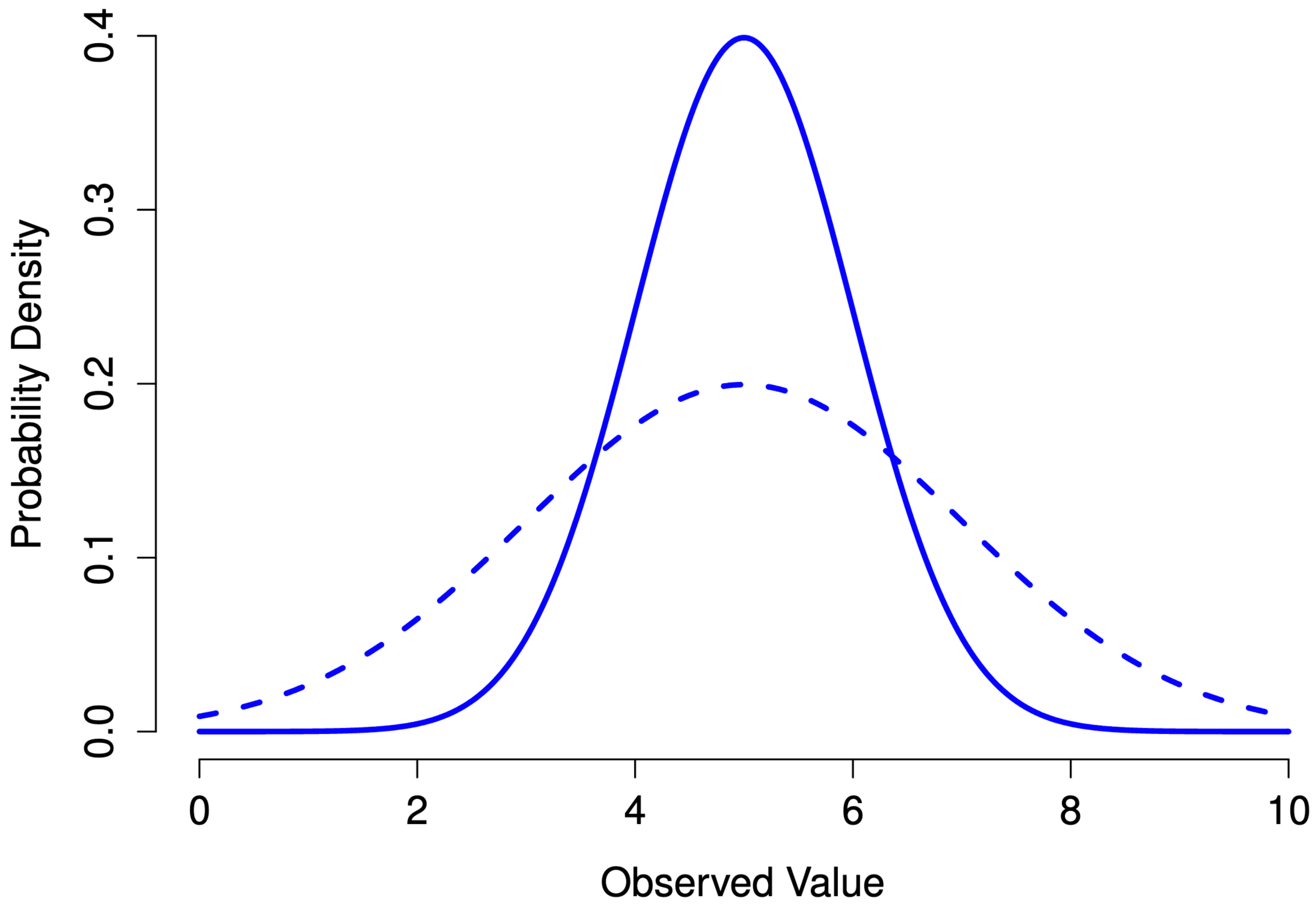
Normal Distribution

- Parameterized by its **mean** (μ) and **standard deviation** (σ)
  - Notated $X \sim \text{Normal}(\mu, \sigma)$
  - Different μ **shifts** the curve left/right
  - Higher σ **"flattens"** the curve
- Unlike the Binomial, the Normal is a **continuous** distribution
  - X can take on any real value



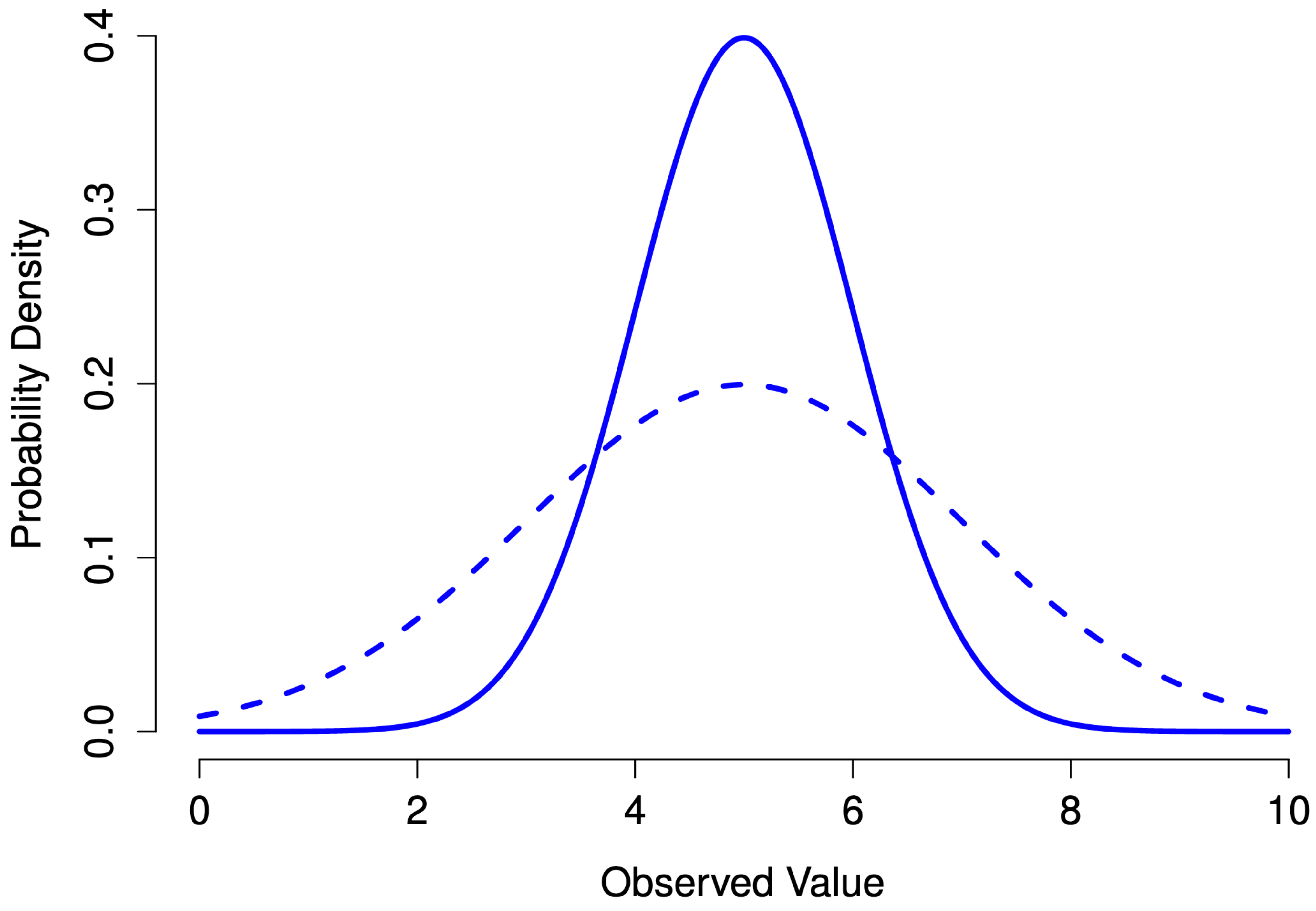
Normal Distribution

- Parameterized by its **mean** (μ) and **standard deviation** (σ)
  - Notated $X \sim \text{Normal}(\mu, \sigma)$
  - Different μ **shifts** the curve left/right
  - Higher σ **"flattens"** the curve
- Unlike the Binomial, the Normal is a **continuous** distribution
  - X can take on any real value
  - The y-axis is **not** actually probability, but **probability density**

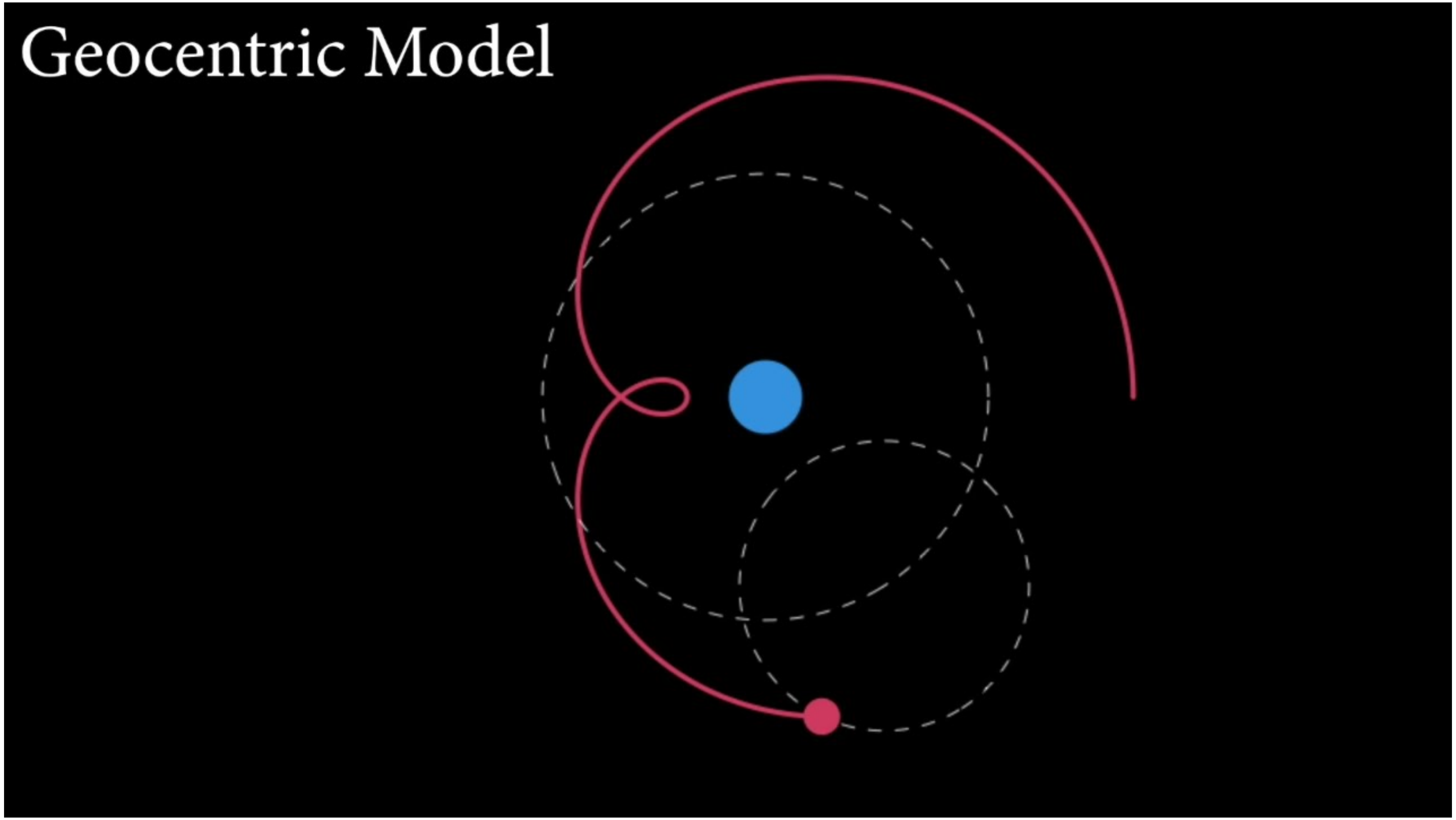


Normal Distribution

- Parameterized by its **mean** (μ) and **standard deviation** (σ)
  - Notated $X \sim \text{Normal}(\mu, \sigma)$
  - Different μ **shifts** the curve left/right
  - Higher σ **"flattens"** the curve
- Unlike the Binomial, the Normal is a **continuous** distribution
  - X can take on any real value
  - The y-axis is **not** actually probability, but **probability density**
  - The **area under the curve** adds to 1

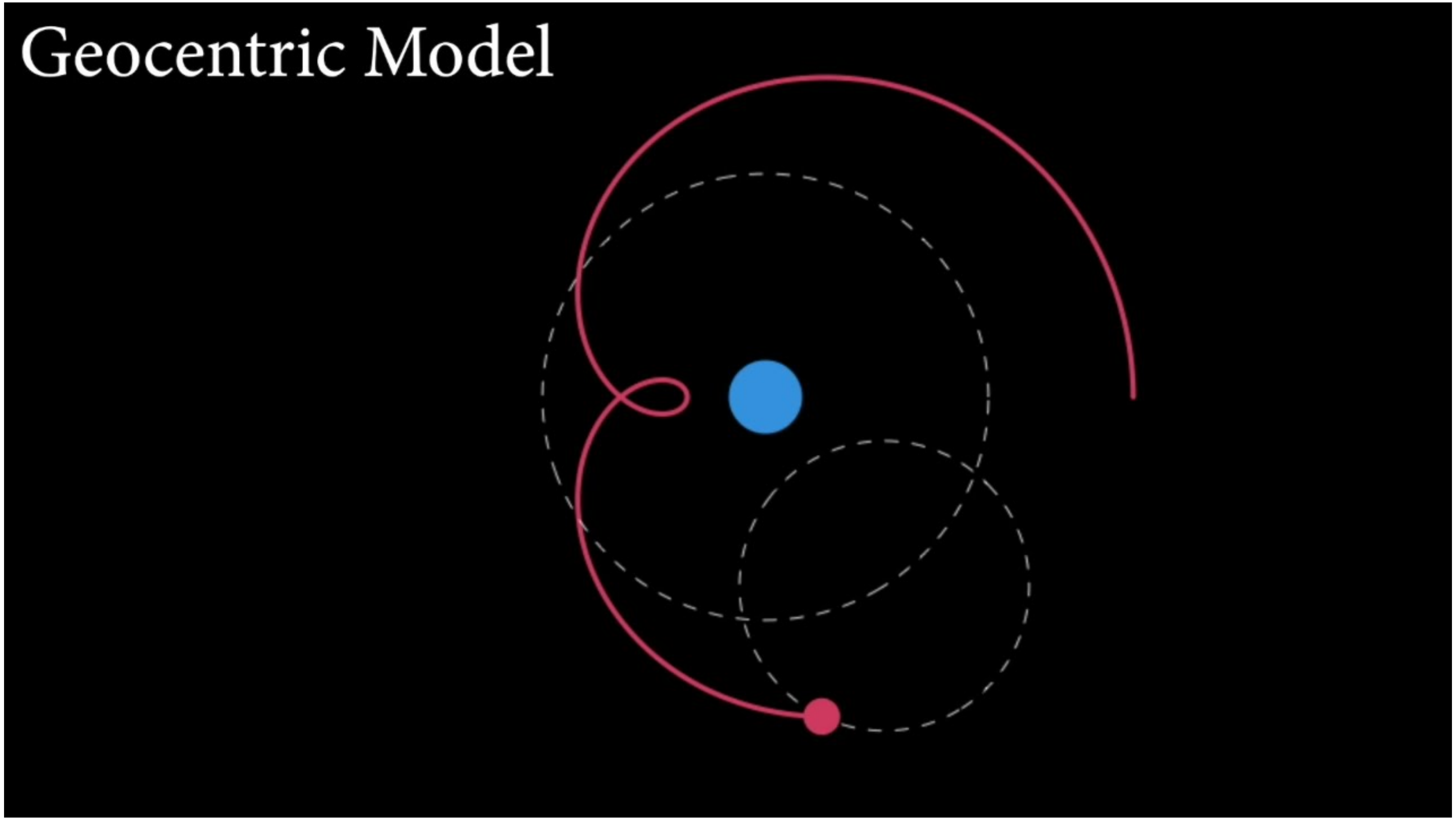


Visualizing the Normal Distribution



10:42-14:32

Visualizing the Normal Distribution



10:42-14:32

Normal Distribution in R

Normal Distribution in R

- Similarly to `dbinom()`, we have `dnorm()`
  - Returns a **probability density** rather than a probability!
  - These are not as interpretable, and can be **greater than 1!**

```
> dnorm(x=0, mean=0, sd=0.1)
[1] 3.989423
```

Normal Distribution in R

- Similarly to `dbinom()`, we have `dnorm()`
  - Returns a **probability density** rather than a probability!
  - These are not as interpretable, and can be **greater than 1!**

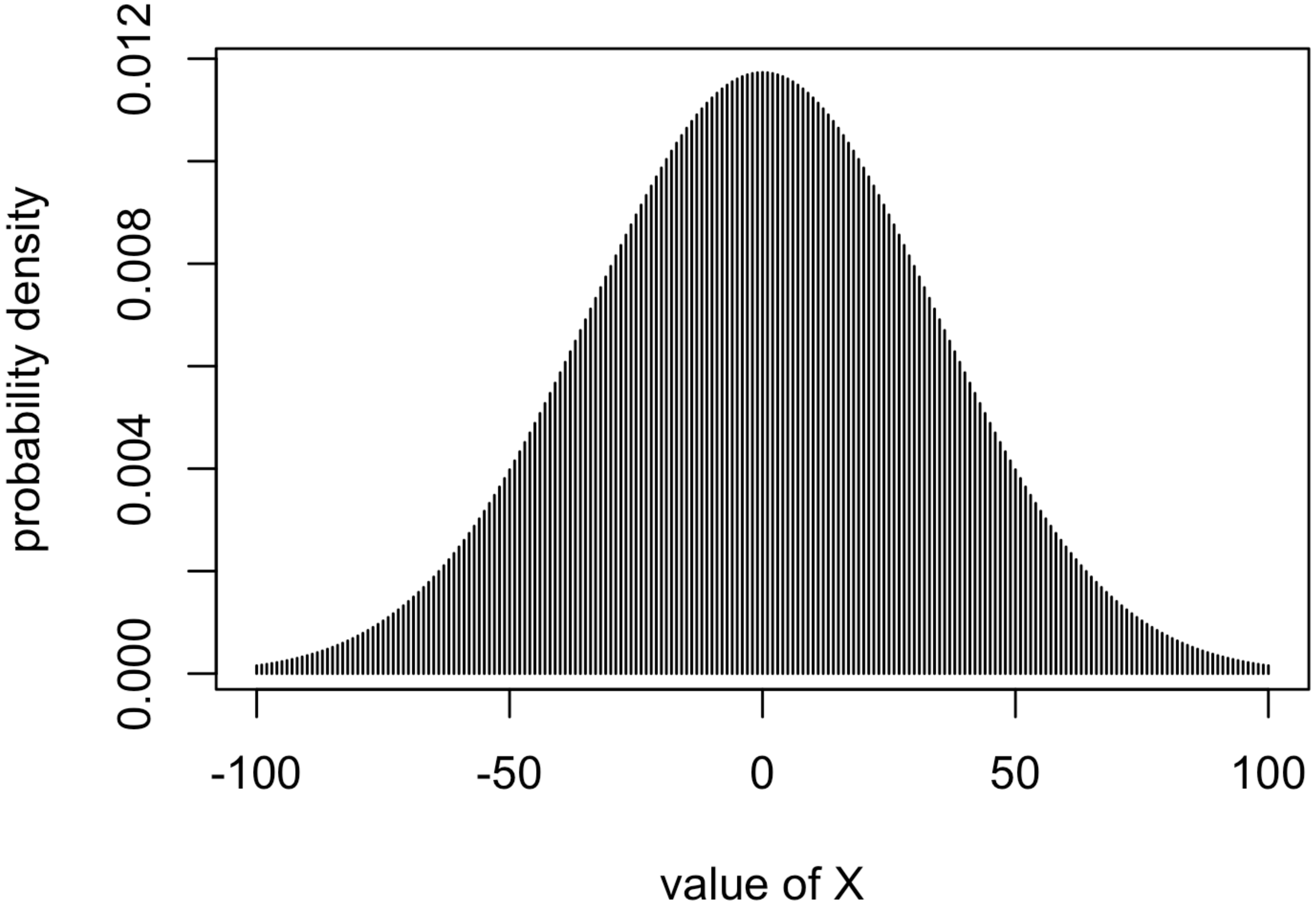
```
> dnorm(x=0, mean=0, sd=0.1)
[1] 3.989423
```

- For continuous variables, we want the probability for X being in a **certain range**
  - `pnorm()` gives us the **cumulative probability** up to a certain value
  - `pnorm(1.0, mean=0, sd=1) == 0.84` means that 84% of the **probability mass** falls **below 1.0** for this Normal Distribution
  - `pnorm(1.5) - pnorm(1.0)` gives us the probability that $1.0 \leq X \leq 1.5$ (9.2%)
  - `pbinom()` does the same thing for the Binomial Distribution

```
> pnorm(1.5, mean=0, sd=1) - pnorm(1.0, mean=0, sd=1)
[1] 0.09184805
```

Plotting the Normal Distribution

```
> norm_values = dnorm(c(-100:100), mean=0, sd=34)
> plot(x=c(-100:100), y=norm_values, type='h', xlab="value of X", y
lab="probability density")
```



Why learn these distributions?

Why learn these distributions?

- We often want to **characterize the distribution** of data "in the wild"

Why learn these distributions?

- We often want to **characterize the distribution** of data "in the wild"
- We often do this by **modeling the data** with one of these well-studied distributions
  - The model is **useful** insofar as it **fits, explains, or predicts** the data
  - Data in the wild doesn't always perfectly fit the model!

Why learn these distributions?

- We often want to **characterize the distribution** of data "in the wild"
- We often do this by **modeling the data** with one of these well-studied distributions
  - The model is **useful** insofar as it **fits, explains, or predicts** the data
  - Data in the wild doesn't always perfectly fit the model!
- It is often useful to **assume** some aspects of the data are **Normally distributed** (regression in particular assumes **model error** is Normal)
  - This is not always the case! But we'll make simplifying assumptions in this class

Samples and Populations

Descriptive vs. Inferential Statistics

Descriptive vs. Inferential Statistics

- Descriptive statistics: "concisely summarize **what we do know**" (i.e. from the data that we have)

Descriptive vs. Inferential Statistics

- Descriptive statistics: "concisely summarize **what we do know**" (i.e. from the data that we have)
- Inferential statistics: "learn about **what we do not know** from what we do"

Descriptive vs. Inferential Statistics

- Descriptive statistics: "concisely summarize **what we do know**" (i.e. from the data that we have)
- Inferential statistics: "learn about **what we do not know** from what we do"
- The data we have is **what we draw inferences from**, but usually **not what we draw inferences about**

Descriptive vs. Inferential Statistics

- Descriptive statistics: "concisely summarize **what we do know**" (i.e. from the data that we have)
- Inferential statistics: "learn about **what we do not know** from what we do"
- The data we have is **what we draw inferences from**, but usually **not what we draw inferences about**
    - Data is considered to be a **sample** taken from a larger **population**

Descriptive vs. Inferential Statistics

- Descriptive statistics: "concisely summarize **what we do know**" (i.e. from the data that we have)
- Inferential statistics: "learn about **what we do not know** from what we do"
- The data we have is **what we draw inferences from**, but usually **not what we draw inferences about**
  - Data is considered to be a **sample** taken from a larger **population**
  - The **population** is what we want to learn about

Descriptive vs. Inferential Statistics

- Descriptive statistics: "concisely summarize **what we do know**" (i.e. from the data that we have)
- Inferential statistics: "learn about **what we do not know** from what we do"
- The data we have is **what we draw inferences from**, but usually **not what we draw inferences about**
  - Data is considered to be a **sample** taken from a larger **population**
  - The **population** is what we want to learn about
  - Ex: our formant data in `vowels.csv` only **represents 44 speakers**, but we use it to **make generalizations** about English speakers in general!

Sampling a population

Sampling a population

- A sample is by definition **incomplete**: the population of interest is often **arbitrarily large** (e.g. "all U.S. adults" or "all linguistics majors")

Sampling a population

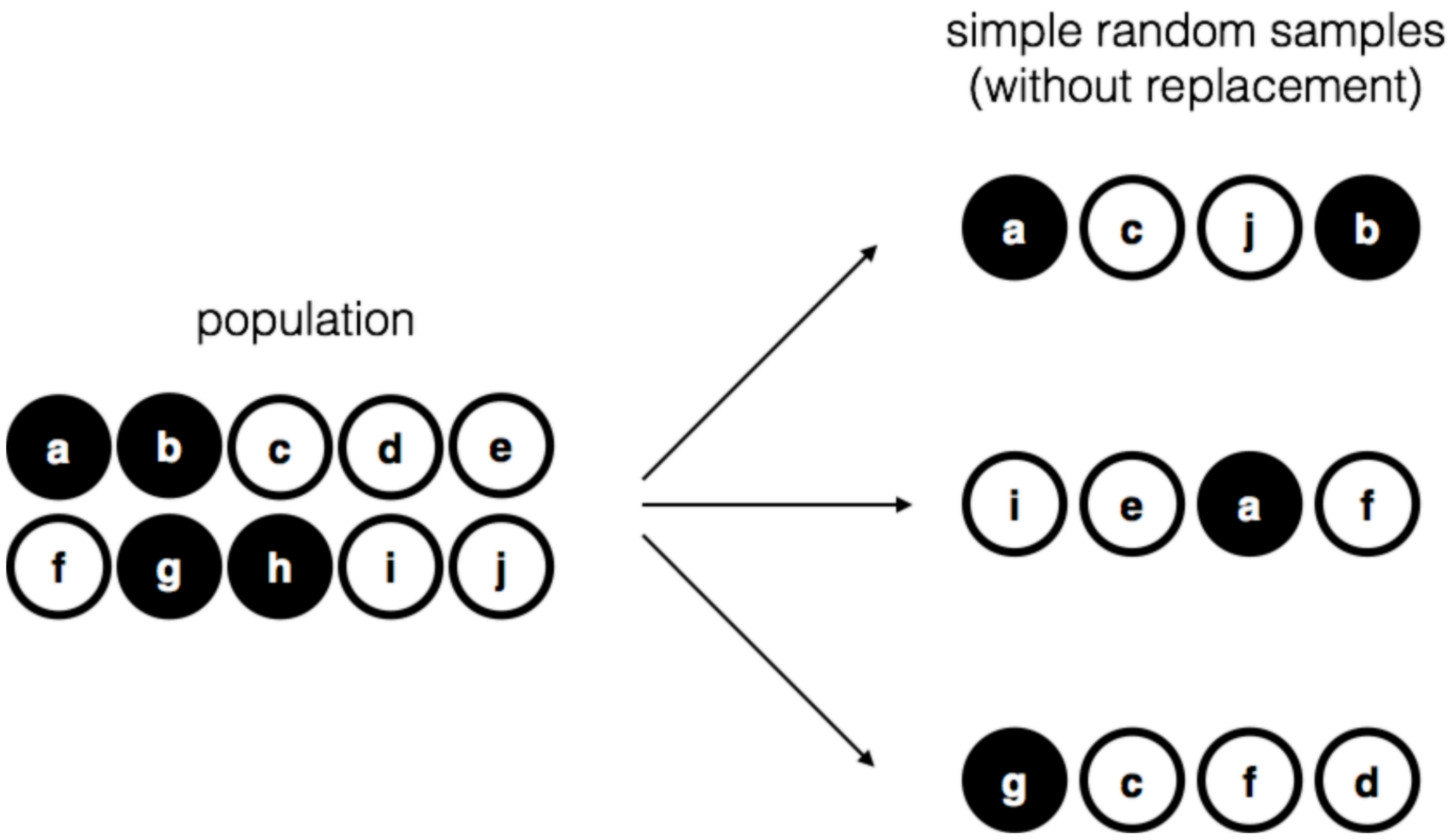
- A sample is by definition **incomplete**: the population of interest is often **arbitrarily large** (e.g. "all U.S. adults" or "all linguistics majors")
- The population is usually more **abstract** than the sample
  - We have a **full accounting** of all speakers of a certain language in a study
  - For most languages, there is **no full accounting** of "all language speakers"

Sampling a population

- A sample is by definition **incomplete**: the population of interest is often **arbitrarily large** (e.g. "all U.S. adults" or "all linguistics majors")
- The population is usually more **abstract** than the sample
  - We have a **full accounting** of all speakers of a certain language in a study
  - For most languages, there is **no full accounting** of "all language speakers"
- In **experimental design**, it's important to:
  - Be **precise** about the **population of interest**
  - Be conscious of how the sample **does or does not represent** the population

Random sampling

- The **ideal** way to sample from a population is **random sampling**
  - I.e. **all members** of the population have an **equal chance of being sampled**
  - This is **essentially impossible** in practice!



Sampling bias

Sampling bias

- In real life, sampling **almost always introduces bias**

Sampling bias

- In real life, sampling **almost always introduces bias**
- Example: election polling
  - High-quality U.S. political polls are almost always **administered by phone**
  - The type of person willing to answer a strange number and do a survey is **not representative** of the average American!
  - Known bias in election polling towards **older** and **more politically engaged** people

Sampling bias

- In real life, sampling **almost always introduces bias**
- Example: election polling
  - High-quality U.S. political polls are almost always **administered by phone**
  - The type of person willing to answer a strange number and do a survey is **not representative** of the average American!
  - Known bias in election polling towards **older** and **more politically engaged** people
- There are methods to **adjust for sampling bias**, but there is almost no way to be rid of it!

Law of Large Numbers

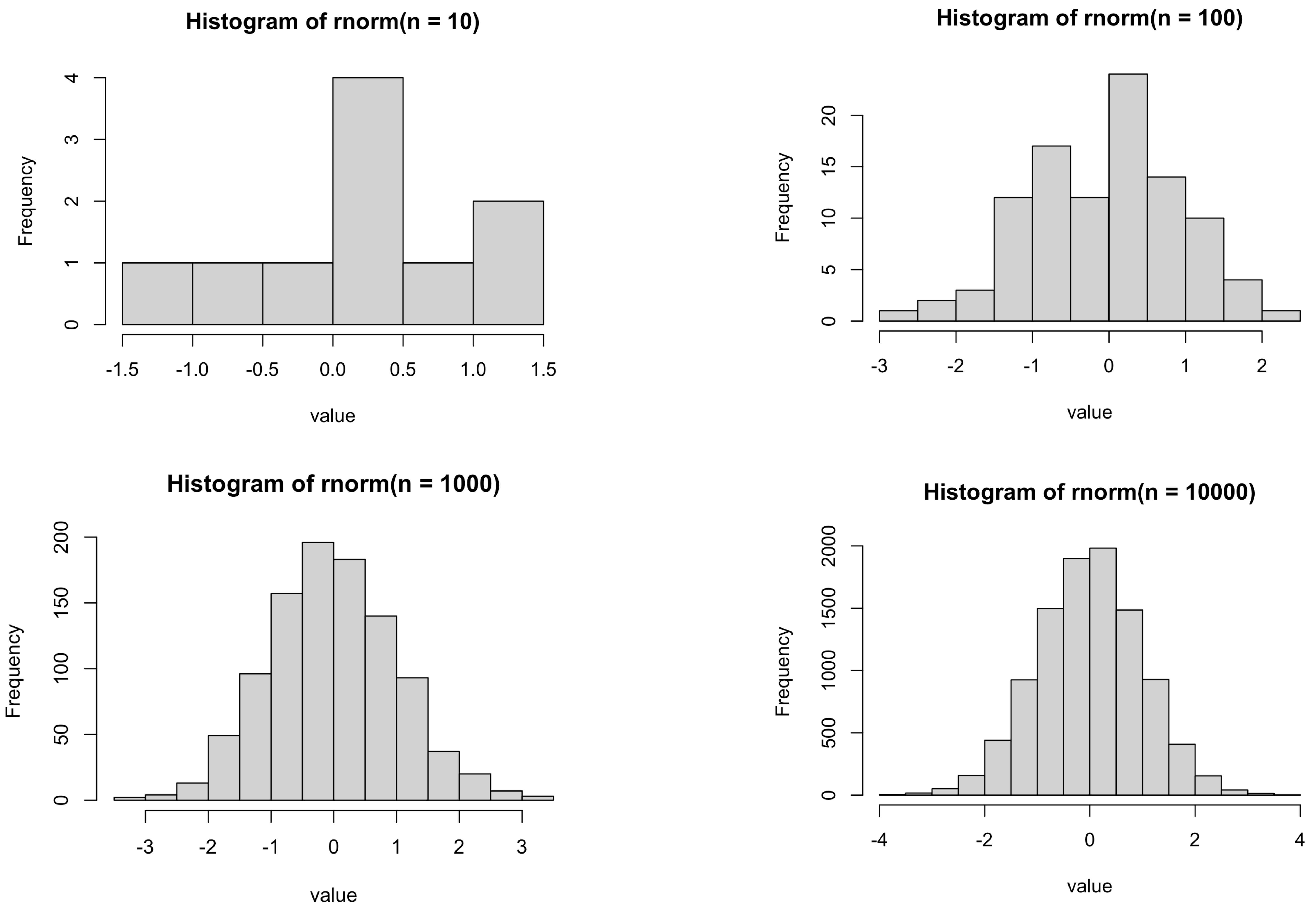
Law of Large Numbers

- The **Law of Large Numbers** is a mathematical proof that most people understand intuitively:
  - The **larger the sample size**, the better the sample **approximates the population**

Law of Large Numbers

- The **Law of Large Numbers** is a mathematical proof that most people understand intuitively:
  - The **larger the sample size**, the better the sample **approximates the population**
- We can observe this by **sampling from a distribution** in R
  - `rnorm()` and `rbinom()` give **random samples** from their respective distributions
  - `rnorm(n=100, mean=0, sd=1)` draws 100 random samples from the Normal
  - Putting the samples into `hist()` shows us the **sample distribution**

Law of Large Numbers



Sampling distributions

Sampling distributions

- What sample size is **large enough** to adequately represent the population?
  - Tricky question, but we can begin to address it with **sampling distributions**

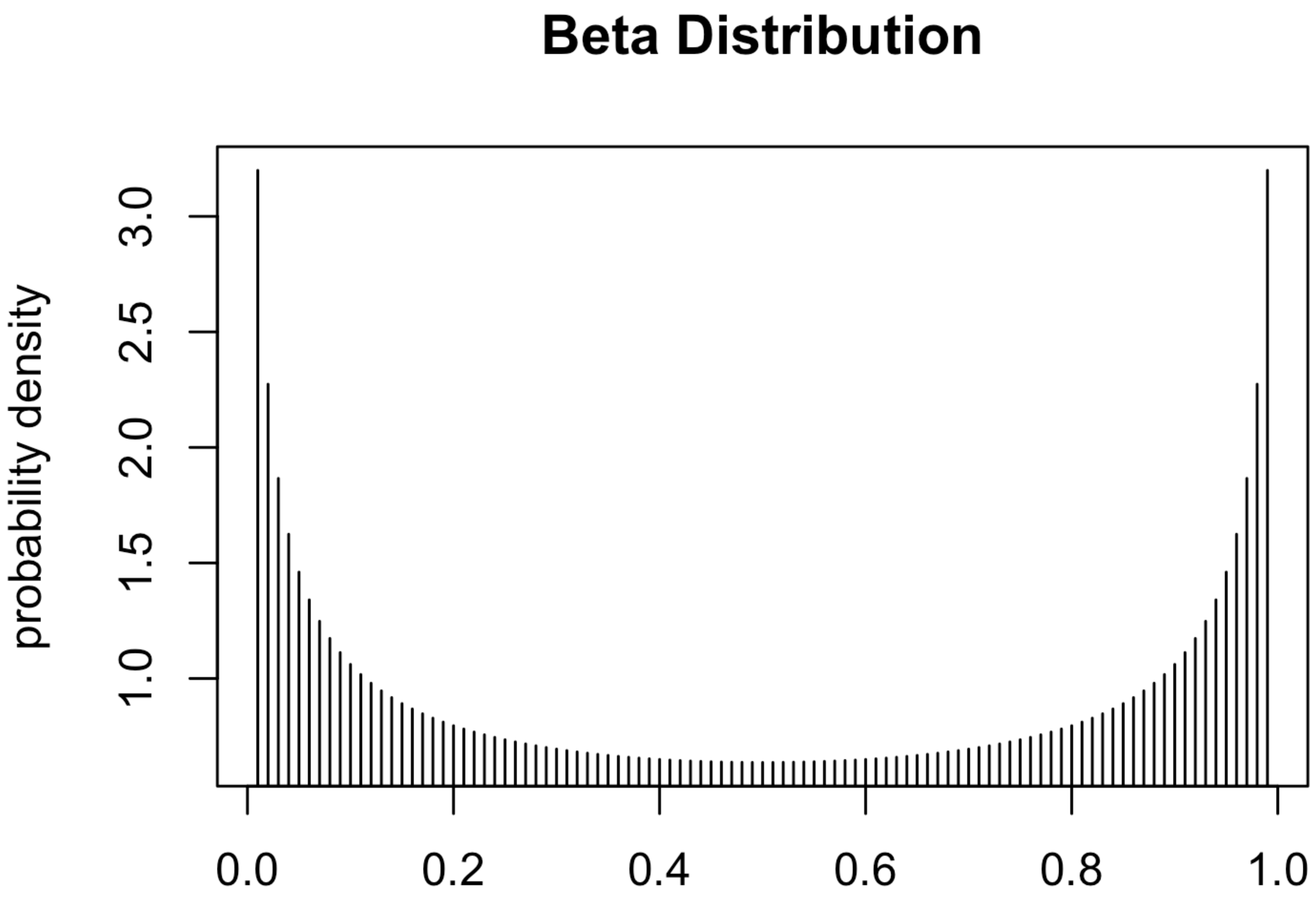
Sampling distributions

- What sample size is **large enough** to adequately represent the population?
  - Tricky question, but we can begin to address it with **sampling distributions**
- Sampling distributions are formed by **sampling many times** from the same population, using a **certain sample size**
  - Example: repeatedly sample **five values** from some population
  - Calculate and plot the **mean for each sample**
  - What you get is the **"Sampling Distribution of the Mean"**

Sampling distributions

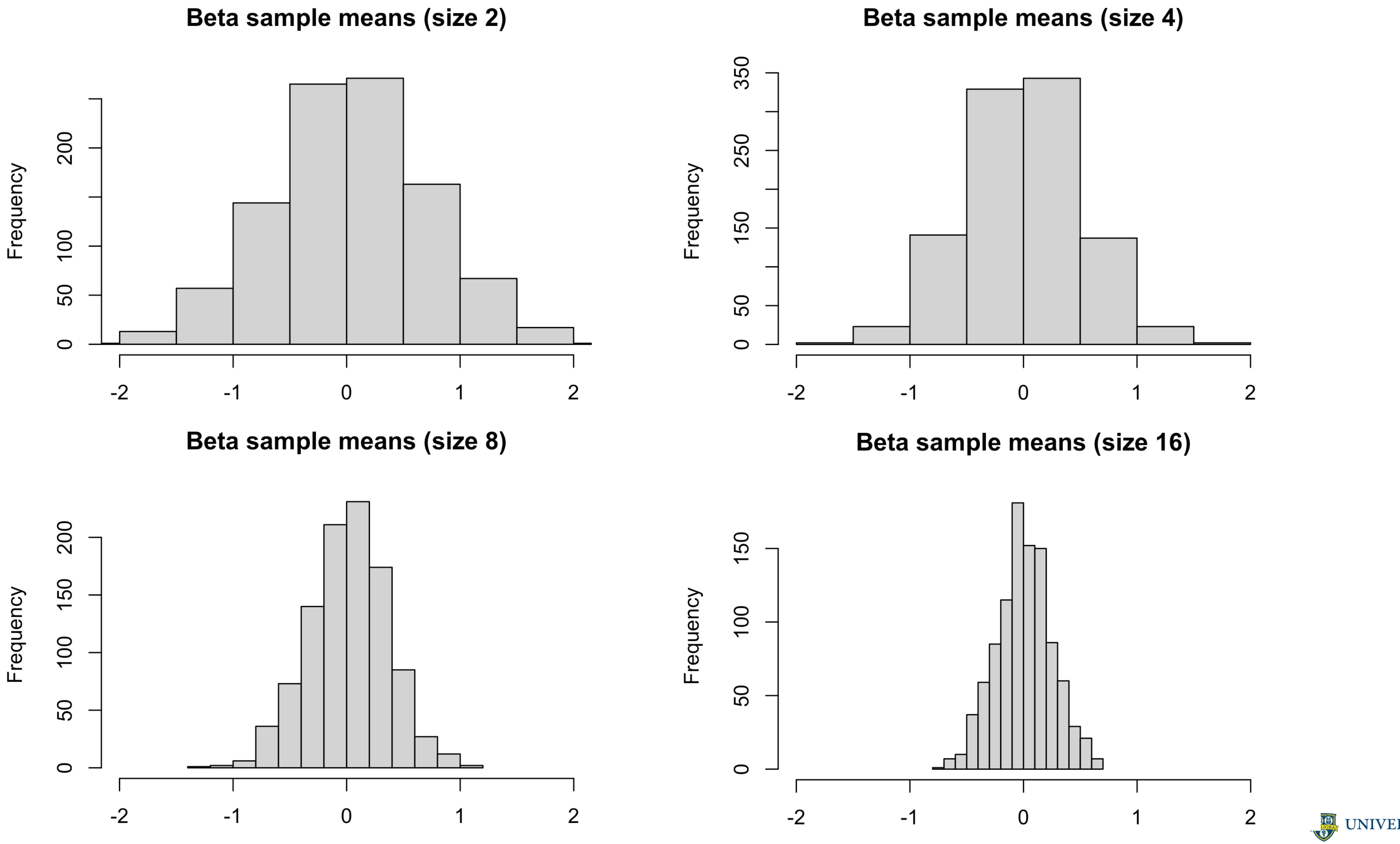
- What sample size is **large enough** to adequately represent the population?
  - Tricky question, but we can begin to address it with **sampling distributions**
- Sampling distributions are formed by **sampling many times** from the same population, using a **certain sample size**
  - Example: repeatedly sample **five values** from some population
  - Calculate and plot the **mean for each sample**
  - What you get is the **"Sampling Distribution of the Mean"**
- The Sampling Distribution of the Mean is **always Normal**, even if the original distribution is not!

Example: Beta Distribution



(not Normally distributed)

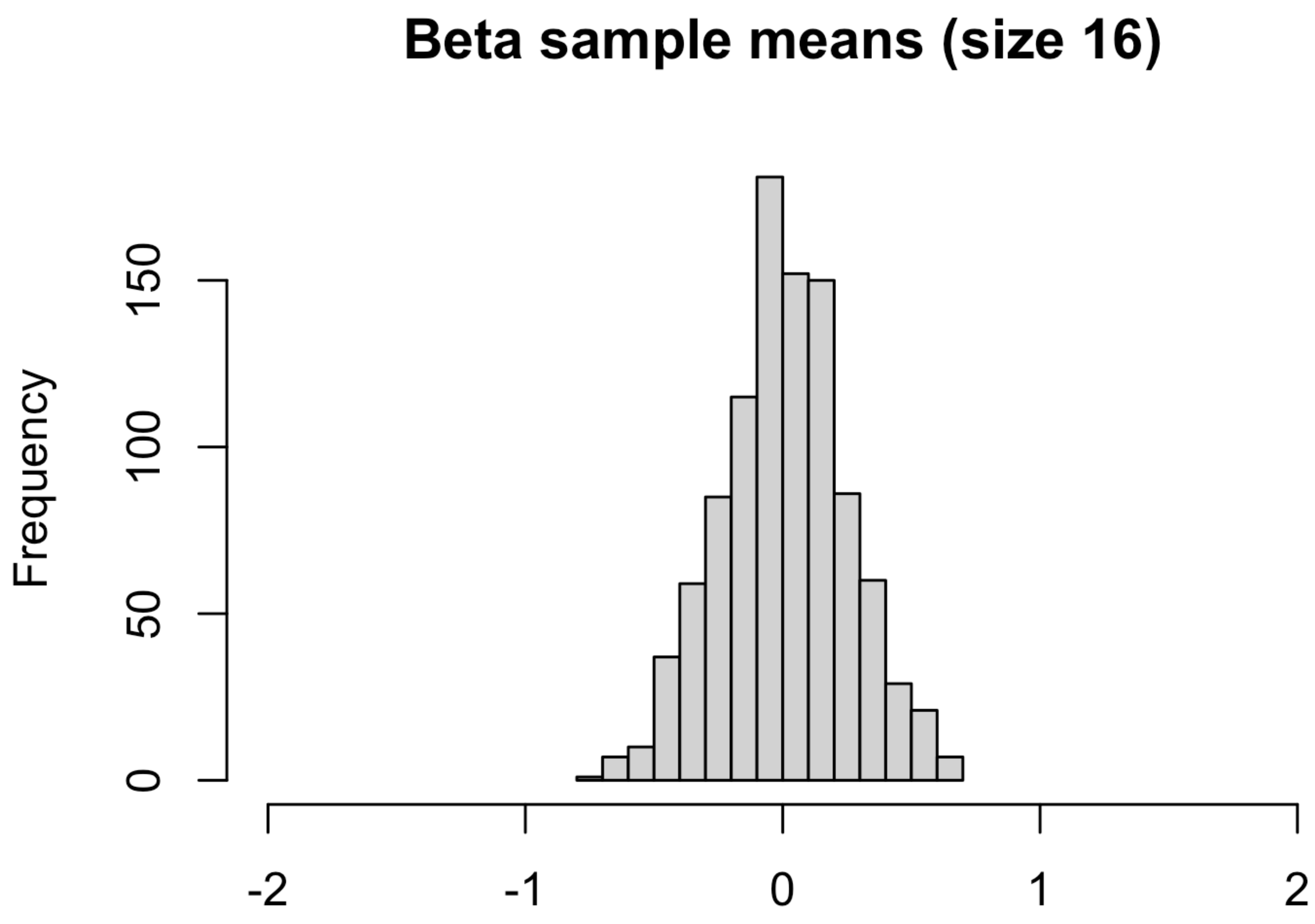
Example: Beta Distribution



Why does this matter?

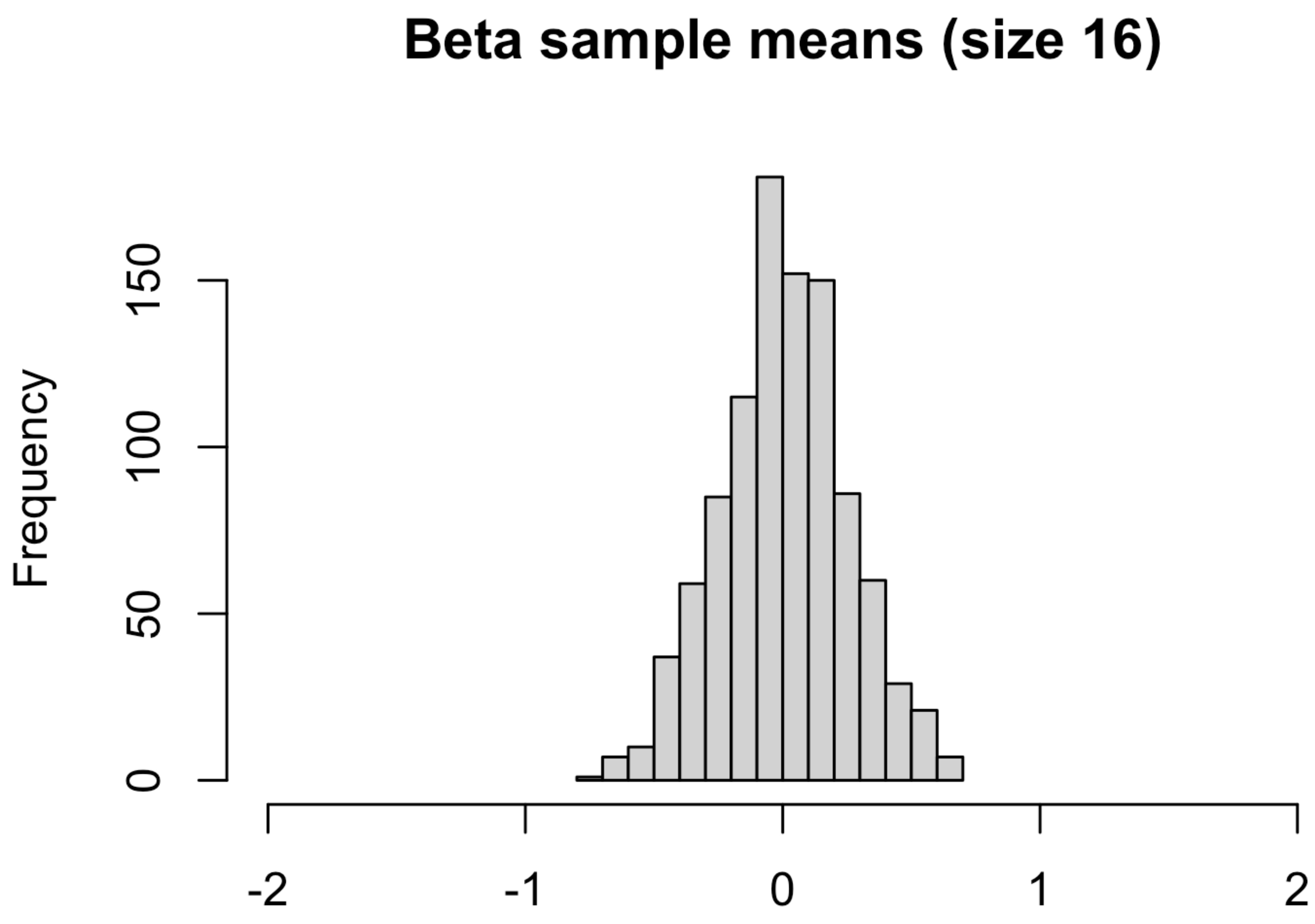
Why does this matter?

- The Sampling Distribution of the Mean shows us that:
  - For any underlying population,
  - For any sample size,
  - We know by **approximately how much** the **sample mean** deviates from our **population mean**



Why does this matter?

- The Sampling Distribution of the Mean shows us that:
  - For any underlying population,
  - For any sample size,
  - We know by **approximately how much** the **sample mean** deviates from our **population mean**
- This relationship between sample mean and population mean is called the **Standard Error of the Mean** (SEM)
  - This is part of what is known as the **Central Limit Theorem**



$$SEM = \frac{\sigma}{\sqrt{N}}$$

population standard deviation