The Central Limit Theorem

(CLT)

Prof. Maria Tackett

Click for PDF of slides

Sample Statistics and Sampling Distributions

Variability of sample statistics

- We've seen that each sample from the population yields a slightly different sample statistic (sample mean, sample proportion, etc.)
- Previously we've quantified this value via simulation
- Today we talk about some of the theory underlying **sampling distributions**, particularly as they relate to sample means.

Statistical inference

- Statistical inference is the act of generalizing from a sample in order to make conclusions regarding a population.
- We are interested in population parameters, which we do not observe. Instead, we must calculate statistics from our sample in order to learn about them.
- As part of this process, we must quantify the degree of uncertainty in our sample statistic.

Sampling distribution of the mean

Suppose we're interested in the mean resting heart rate of students at Duke, and are able to do the following:

Sampling distribution of the mean

Suppose we're interested in the mean resting heart rate of students at Duke, and are able to do the following:

1. Take a random sample of size n from this population, and calculate the mean resting heart rate in this sample, $\bar{X}_1$

Sampling distribution of the mean

Suppose we're interested in the mean resting heart rate of students at Duke, and are able to do the following:

1. Take a random sample of size n from this population, and calculate the mean resting heart rate in this sample, $\bar{X}_1$

2. Put the sample back, take a second random sample of size n, and calculate the mean resting heart rate from this new sample, $\bar{X}_2$

Sampling distribution of the mean

Suppose we're interested in the mean resting heart rate of students at Duke, and are able to do the following:

1. Take a random sample of size n from this population, and calculate the mean resting heart rate in this sample, $\bar{X}_1$

2. Put the sample back, take a second random sample of size n, and calculate the mean resting heart rate from this new sample, $\bar{X}_2$

3. Put the sample back, take a third random sample of size n, and calculate the mean resting heart rate from this sample, too...

Sampling distribution of the mean

Suppose we're interested in the mean resting heart rate of students at Duke, and are able to do the following:

1. Take a random sample of size n from this population, and calculate the mean resting heart rate in this sample, $\bar{X}_1$
2. Put the sample back, take a second random sample of size n, and calculate the mean resting heart rate from this new sample, $\bar{X}_2$
3. Put the sample back, take a third random sample of size n, and calculate the mean resting heart rate from this sample, too...

...and so on.

Sampling distribution of the mean

After repeating this many times, we have a dataset that has the sample averages from the population: $\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_K$ (assuming we took K total samples).

Sampling distribution of the mean

After repeating this many times, we have a dataset that has the sample averages from the population: $\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_K$ (assuming we took K total samples).

Can we say anything about the distribution of these sample means (that is, the **sampling distribution** of the mean?)

(Keep in mind, we don't know what the underlying distribution of mean resting heart rate looks like in Duke students!)

The Central Limit Theorem

A quick caveat...

For now, let's assume we know the underlying standard deviation, σ, from our distribution

The Central Limit Theorem

For a population with a well-defined mean μ and standard deviation σ, these three properties hold for the distribution of sample average $\bar{X}$, assuming certain conditions hold:

The Central Limit Theorem

For a population with a well-defined mean μ and standard deviation σ, these three properties hold for the distribution of sample average $\bar{X}$, assuming certain conditions hold:

1. The mean of the sampling distribution of the mean is identical to the population mean μ.

The Central Limit Theorem

For a population with a well-defined mean μ and standard deviation σ, these three properties hold for the distribution of sample average $\bar{X}$, assuming certain conditions hold:

1. The mean of the sampling distribution of the mean is identical to the population mean μ.

2. The standard deviation of the distribution of the sample averages is $\sigma/\sqrt{n}$.

    - This is called the **standard error** (SE) of the mean.

The Central Limit Theorem

For a population with a well-defined mean μ and standard deviation σ, these three properties hold for the distribution of sample average $\bar{X}$, assuming certain conditions hold:

1. The mean of the sampling distribution of the mean is identical to the population mean μ.

2. The standard deviation of the distribution of the sample averages is $\sigma/\sqrt{n}$.

    - This is called the **standard error** (SE) of the mean.

3. For n large enough, the shape of the sampling distribution of means is approximately **normally distributed**.

The normal (Gaussian) distribution

The normal distribution is unimodal and symmetric and is described by its density function:

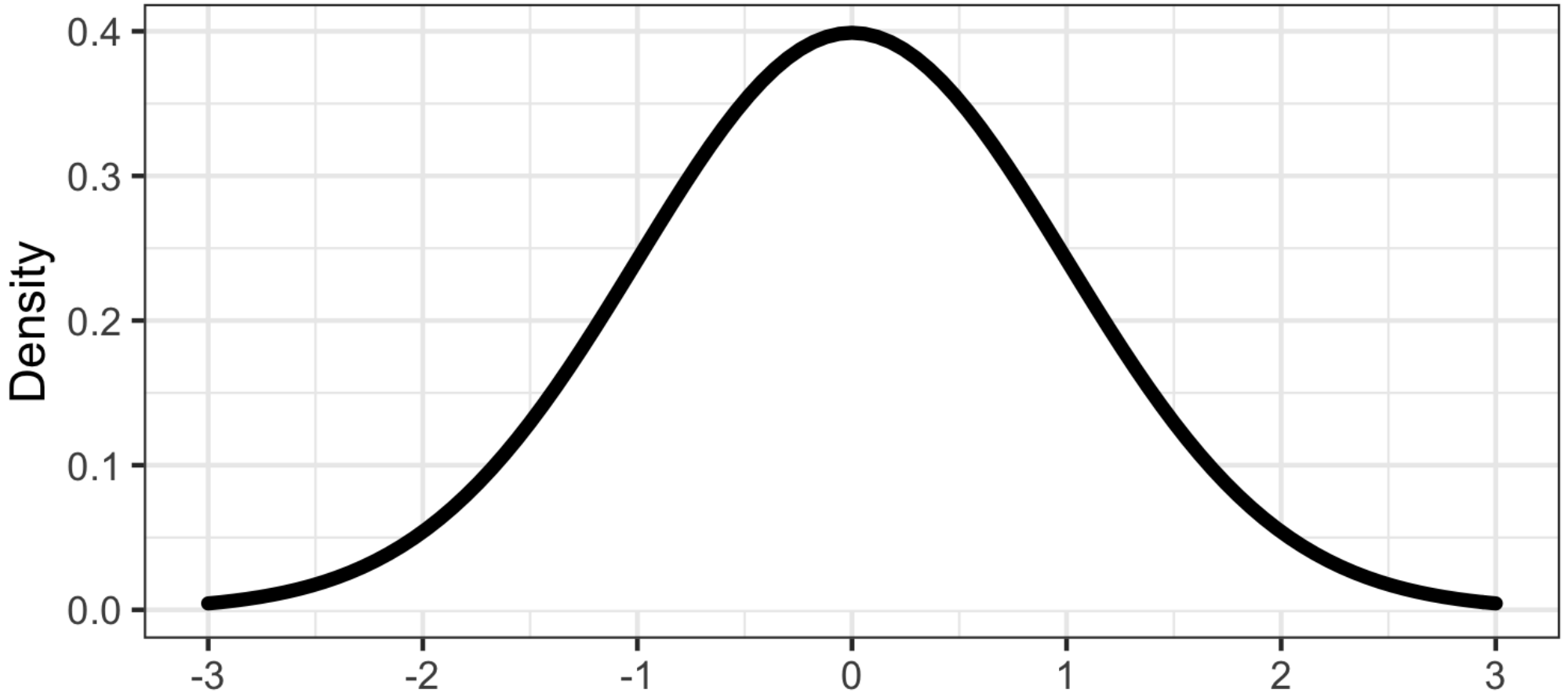
If a random variable X follows the normal distribution, then

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right\}$$

where μ is the mean and σ^2 is the variance (σ is the standard deviation)

We often write $N(\mu, \sigma)$ to describe this distribution.

The normal distribution (graphically)



Wait, *any* distribution?

The central limit theorem tells us that ***sample averages*** are normally distributed, if we have enough data and certain assumptions hold.

This is true *even if our original variables are not normally distributed*.

Click here to see an interactive demonstration of this idea.

Conditions for CLT

We need to check two conditions for CLT to hold: independence, sample size/distribution.

Conditions for CLT

We need to check two conditions for CLT to hold: independence, sample size/distribution.

✅ **Independence:** The sampled observations must be independent. This is difficult to check, but the following are useful guidelines:

- the sample must be randomly taken
- if sampling without replacement, sample size must be less than 10% of the population size

Conditions for CLT

We need to check two conditions for CLT to hold: independence, sample size/distribution.

✅ **Independence:** The sampled observations must be independent. This is difficult to check, but the following are useful guidelines:

- the sample must be randomly taken
- if sampling without replacement, sample size must be less than 10% of the population size

If samples are independent, then by definition one sample's value does not "influence" another sample's value.

Conditions for CLT

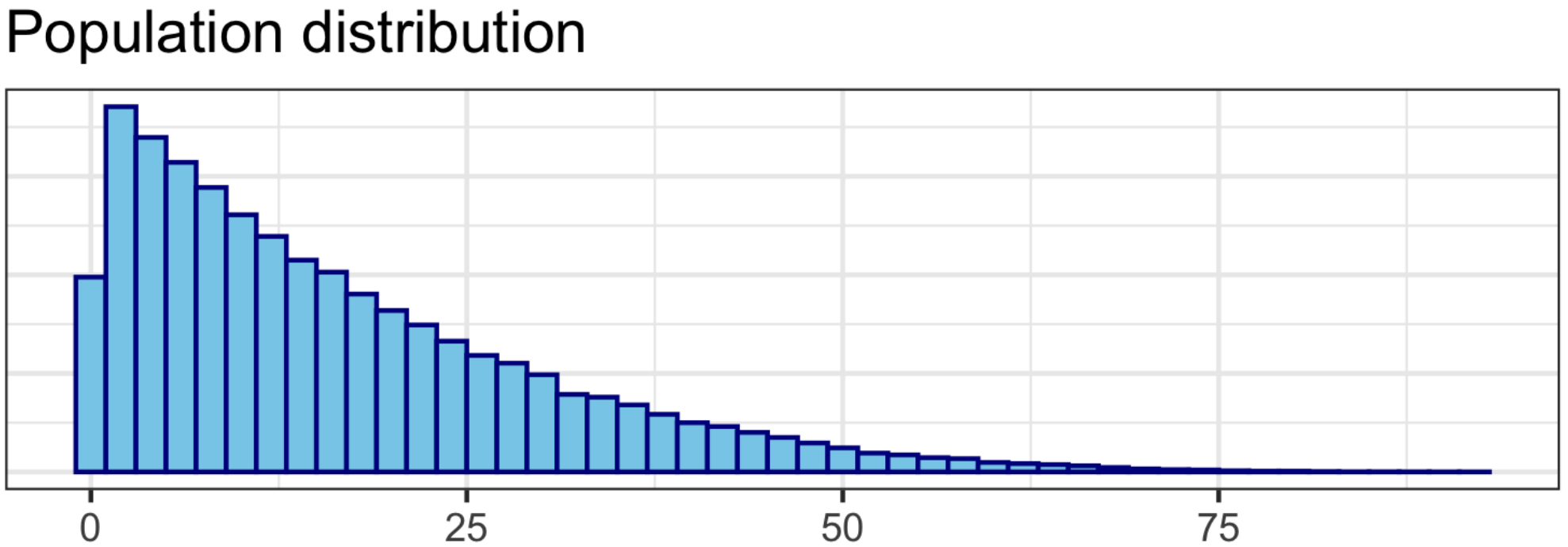
✅ Sample size / distribution:

- if data are numerical, usually n > 30 is considered a large enough sample for the CLT to kick in
- if we know for sure that the underlying data are normally distributed, then the distribution of sample averages will also be exactly normal, regardless of the sample size
- if data are categorical, at least 10 successes and 10 failures.

Let's run our own simulation

Underlying population (not observed in real life!)

```
rs_pop <- tibble(x = rbeta(100000, 1, 5) * 100)
```



Sampling from the population - 1

```
set.seed(1)
samp_1 <- rs_pop %>%
  sample_n(size = 50) %>%
  summarise(x_bar = mean(x))
```

```
samp_1
```

```
## # A tibble: 1 x 1
##   x_bar
##   <dbl>
## 1  15.9
```

Sampling from the population - 2

```
set.seed(2)
samp_2 <- rs_pop %>%
  sample_n(size = 50) %>%
  summarise(x_bar = mean(x))
```

```
samp_2
```

```
## # A tibble: 1 x 1
##   x_bar
##   <dbl>
## 1  17.1
```

Sampling from the population - 3

```
set.seed(3)
samp_3 <- rs_pop %>%
  sample_n(size = 50) %>%
  summarise(x_bar = mean(x))
```

```
samp_3
```

```
## # A tibble: 1 x 1
##   x_bar
##   <dbl>
## 1  19.2
```

Sampling from the population - 3

```
set.seed(3)
samp_3 <- rs_pop %>%
  sample_n(size = 50) %>%
  summarise(x_bar = mean(x))
```

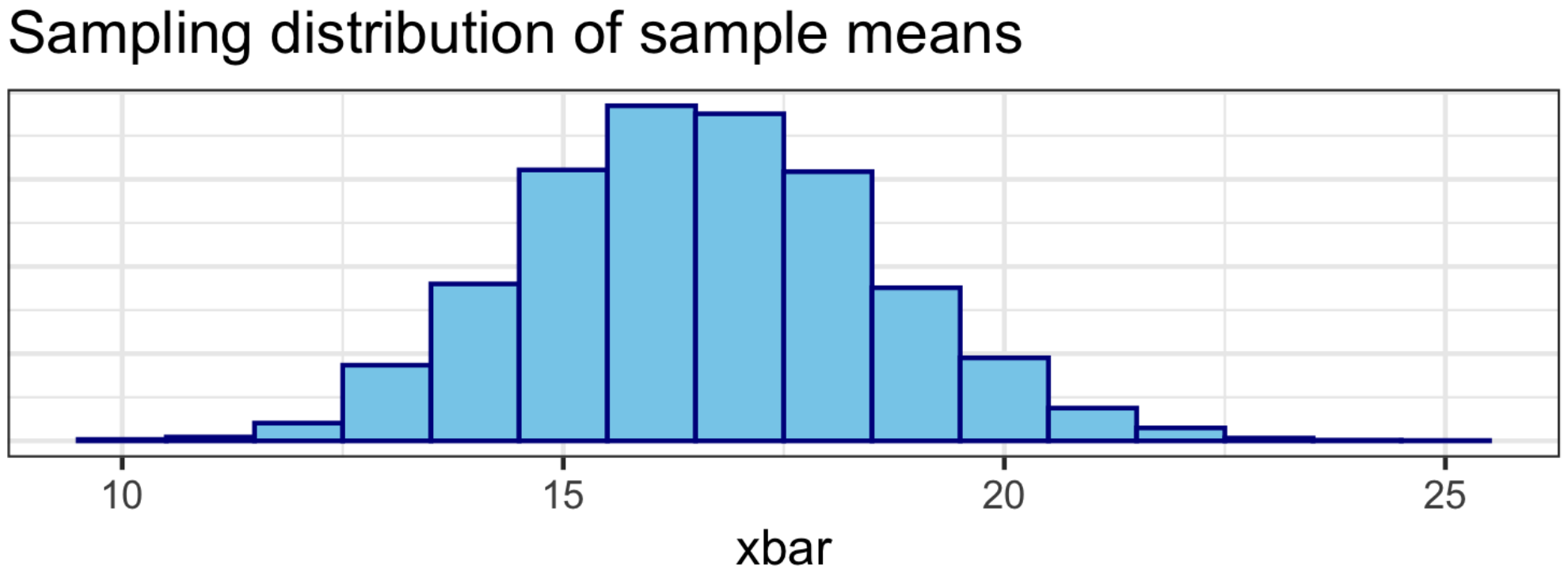
```
samp_3
```

```
## # A tibble: 1 x 1
##   x_bar
##   <dbl>
## 1  19.2
```

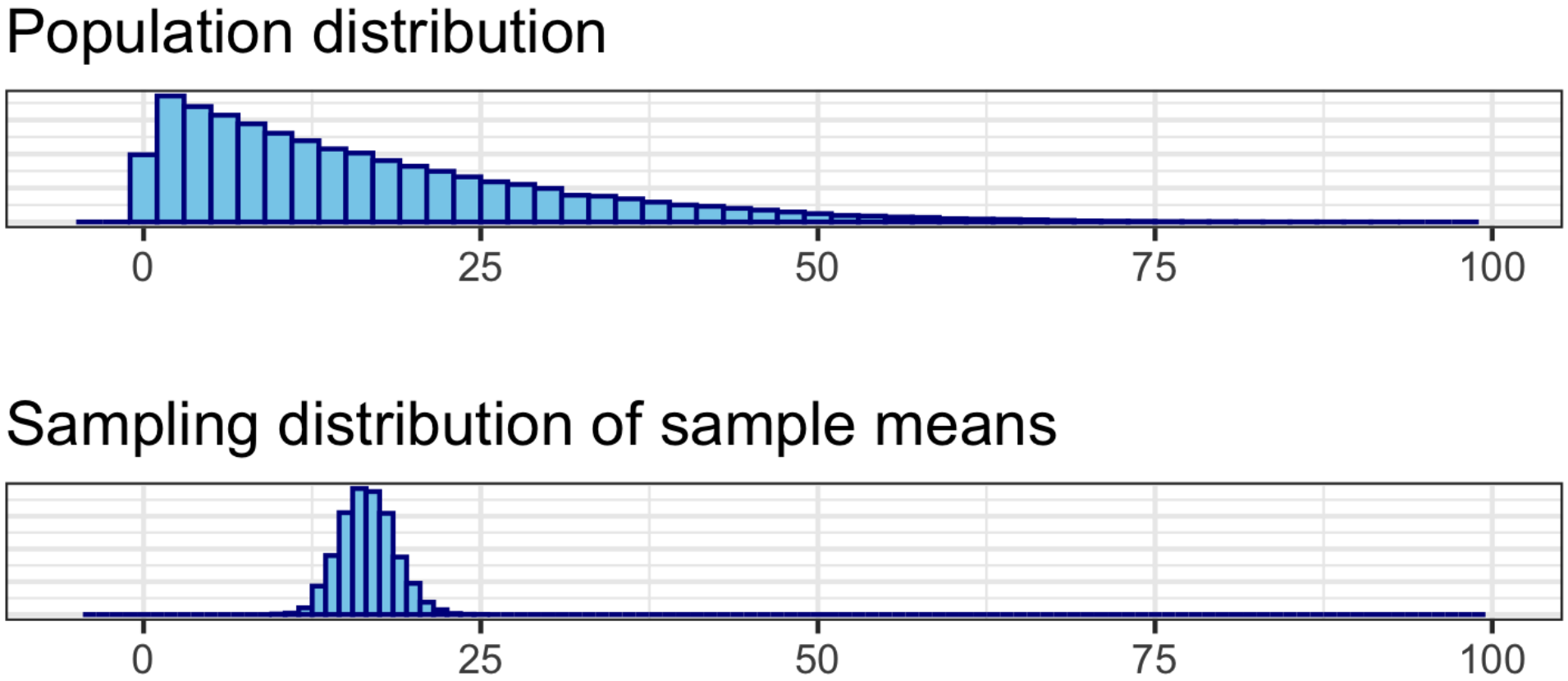
keep repeating...

Sampling distribution

```
set.seed(092620)
sampling <- rs_pop %>%
  rep_sample_n(size = 50, replace = TRUE, reps = 5000) %>%
  group_by(replicate) %>%
  summarise(xbar = mean(x))
```



How do the shapes, centers, and spreads of these distributions compare?



Recap

- If certain assumptions are satisfied, regardless of the shape of the population distribution, the sampling distribution of the mean follows an approximately normal distribution.

Recap

- If certain assumptions are satisfied, regardless of the shape of the population distribution, the sampling distribution of the mean follows an approximately normal distribution.
- The center of the sampling distribution is at the center of the population distribution.

Recap

- If certain assumptions are satisfied, regardless of the shape of the population distribution, the sampling distribution of the mean follows an approximately normal distribution.
- The center of the sampling distribution is at the center of the population distribution.
- The sampling distribution is less variable than the population distribution (and we can quantify by how much).

Recap

- If certain assumptions are satisfied, regardless of the shape of the population distribution, the sampling distribution of the mean follows an approximately normal distribution.
- The center of the sampling distribution is at the center of the population distribution.
- The sampling distribution is less variable than the population distribution (and we can quantify by how much).

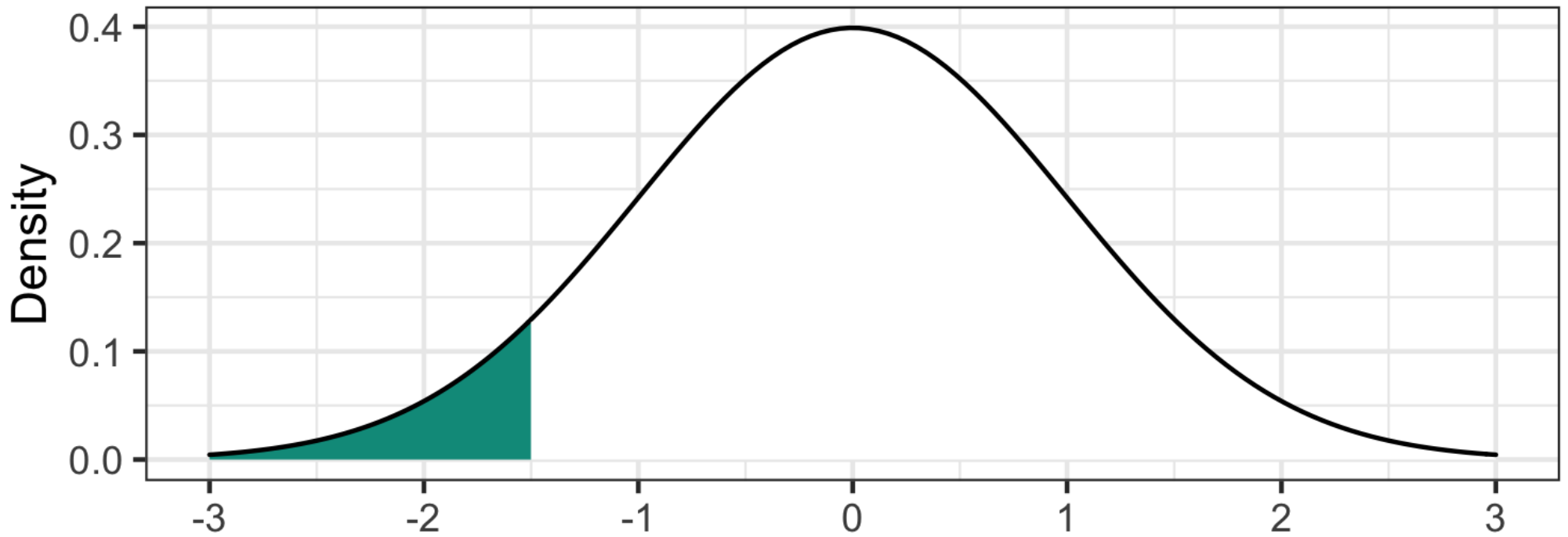
What is the standard error, and how are the standard error and sample size related? What does that say about how the spread of the sampling distribution changes as n increases?

Finding probabilities in R

Probabilities under N(0,1) curve

```
# P(Z < -1.5)
pnorm(-1.5)
```

```
## [1] 0.0668072
```

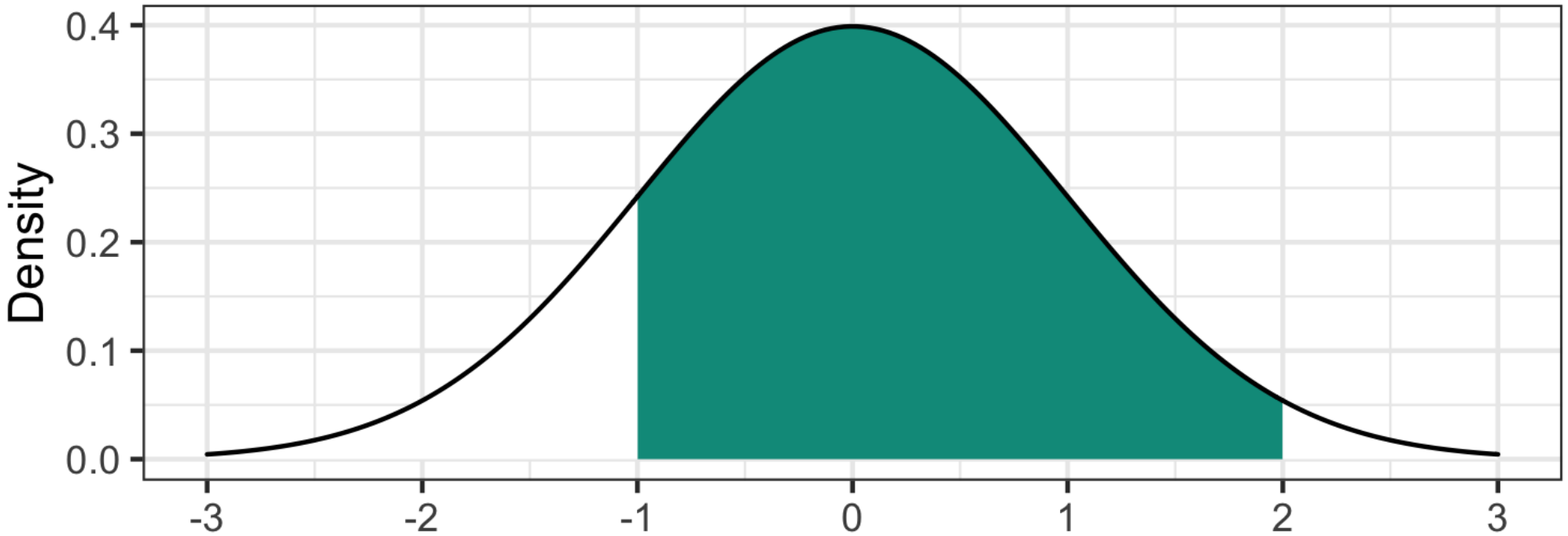


Probability between two values

If $Z \sim N(0, 1)$, what is $P(-1 < Z < 2)$?

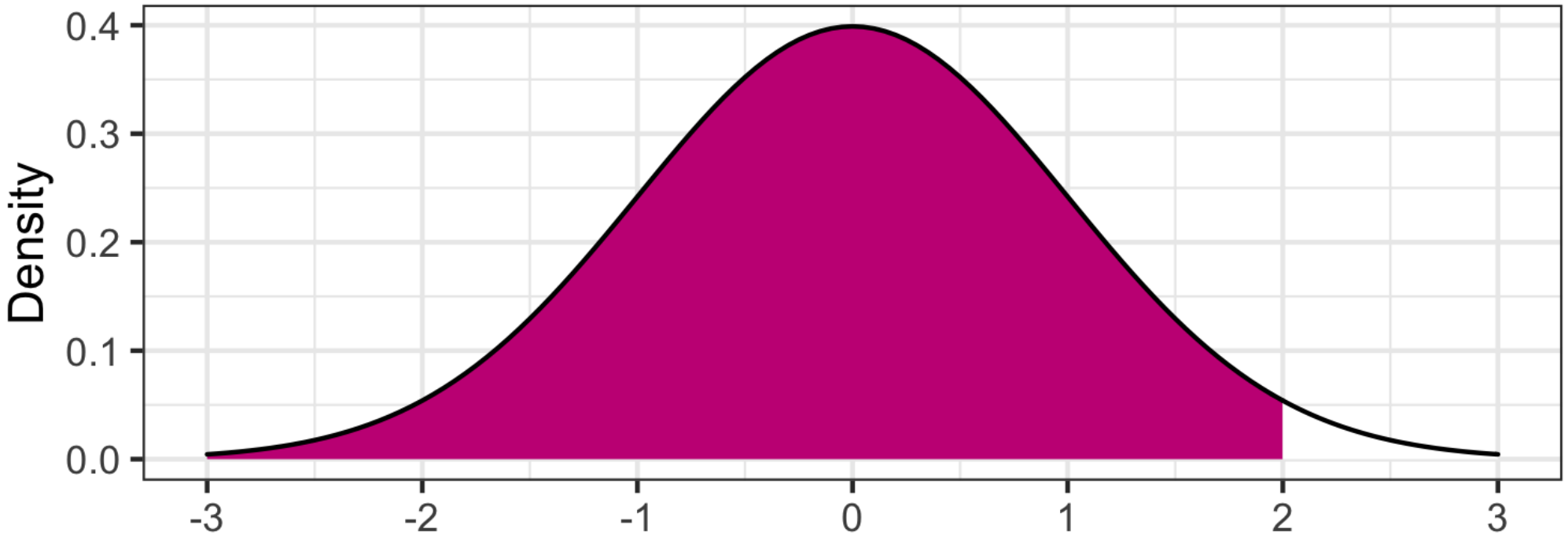
Probability between two values

If $Z \sim N(0, 1)$, what is $P(-1 < Z < 2)$?



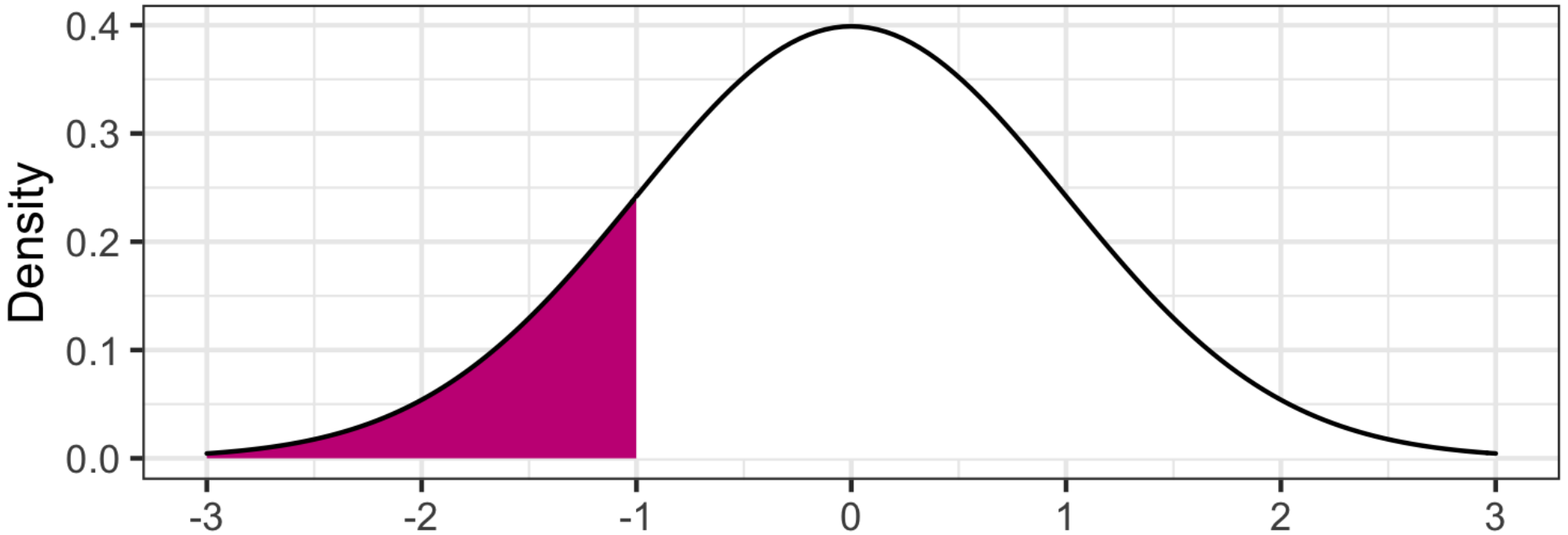
Probability between two values

If $Z \sim N(0, 1)$, what is $P(-1 < Z < 2)$?



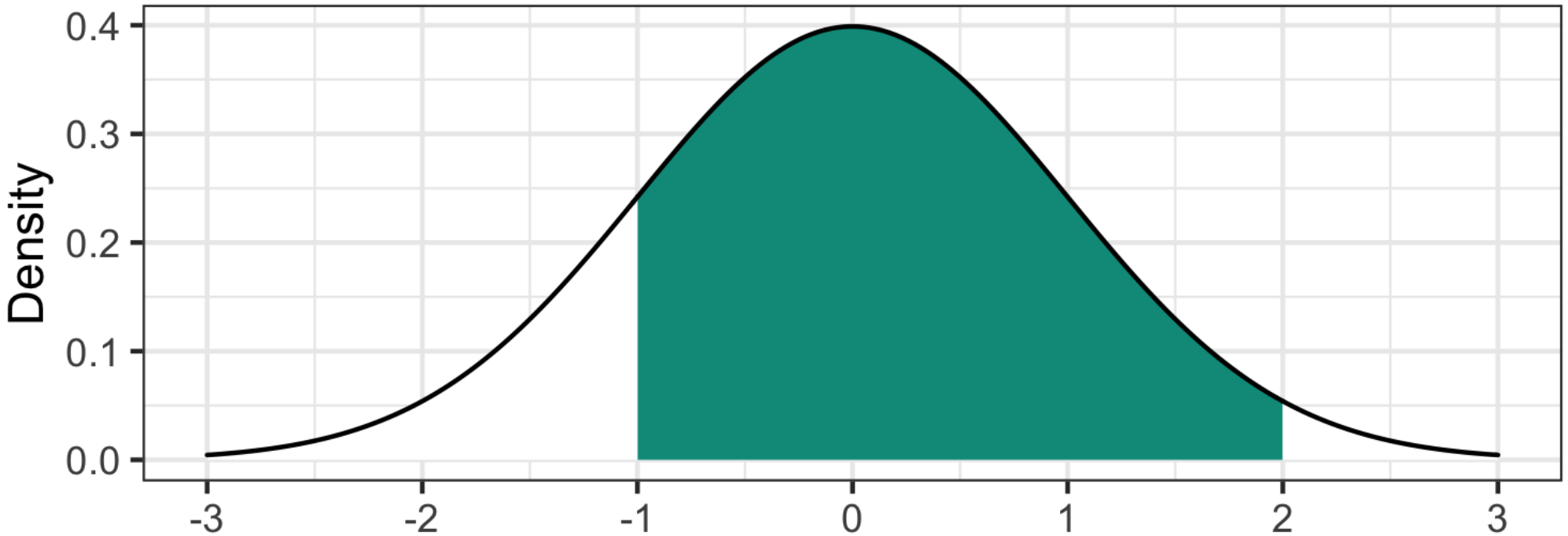
Probability between two values

If $Z \sim N(0, 1)$, what is $P(-1 < Z < 2)$?



Probability between two values

If $Z \sim N(0, 1)$, what is $P(-1 < Z < 2)$?



Probability between two values

If $Z \sim N(0, 1)$, what is $P(-1 < Z < 2)$?

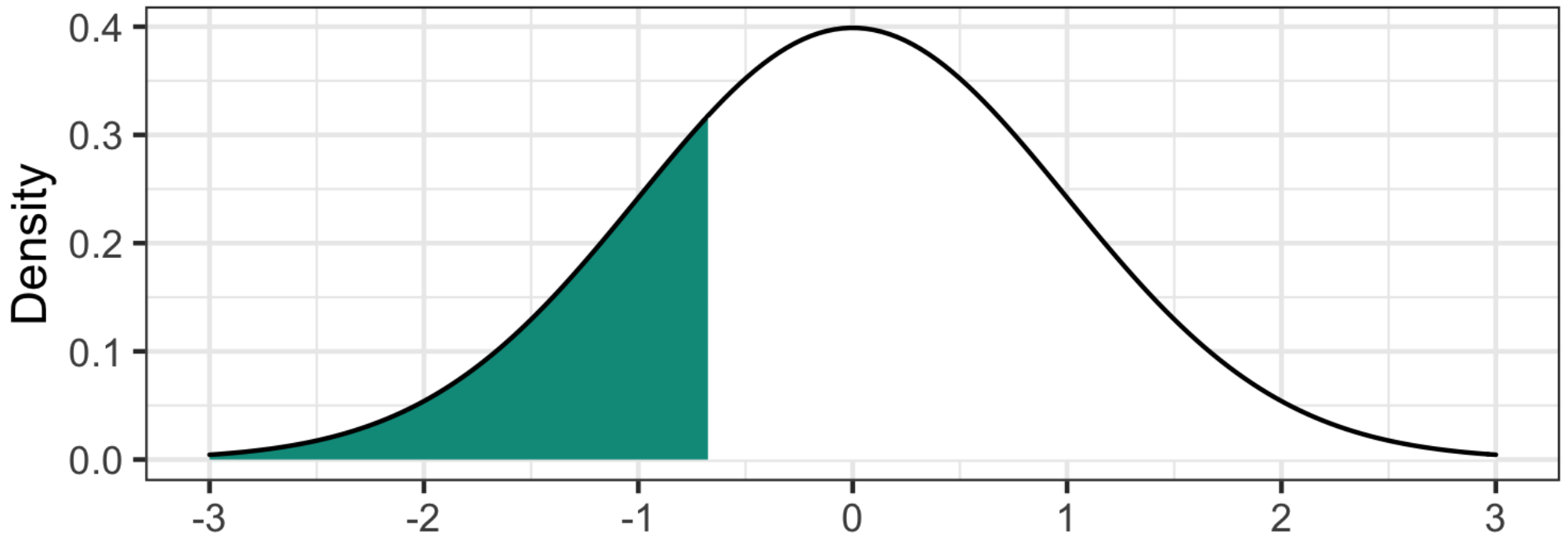
```
pnorm(2) - pnorm(-1)
```

```
## [1] 0.8185946
```

Finding cutoff values under N(0,1) curve

```
# find Q1
qnorm(0.25)
```

```
## [1] -0.6744898
```



Looking ahead...

We will use the Central Limit Theorem and the normal distribution to conduct statistical inference.