Inference with the CLT

Prof. Maria Tackett

Click for PDF of slides

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 distribution of the sample statistic is nearly normal

✅ The distribution is centered at the (often unknown) population parameter

✅ The variability of the distribution is inversely proportional to the square root of the sample size

Why do we care?

Knowing the distribution of the sample statistic $\bar{X}$ can help us

Why do we care?

Knowing the distribution of the sample statistic $\bar{X}$ can help us

- estimate a population parameter as point **estimate $\pm$ margin of error**
  - the **margin of error** is comprised of a measure of how confident we want to be and how variable the sample statistic is

Why do we care?

Knowing the distribution of the sample statistic $\bar{X}$ can help us

- estimate a population parameter as point **estimate $\pm$ margin of error**
  - the margin of error is comprised of a measure of how confident we want to be and how variable the sample statistic is
- test for a population parameter by evaluating how likely it is to obtain to observed sample statistic when assuming that the null hypothesis is true
  - this probability will depend on how variable the sampling distribution is

Inference based on the CLT

Inference based on the CLT

If necessary conditions are met, we can also use inference methods based on the CLT. Suppose we know the true population standard deviation.

Inference based on the CLT

If necessary conditions are met, we can also use inference methods based on the CLT. Suppose we know the true population standard deviation.

Then the CLT tells us that $\bar{X}$ approximately has the distribution $N\left(\mu, \sigma/\sqrt{n}\right)$.

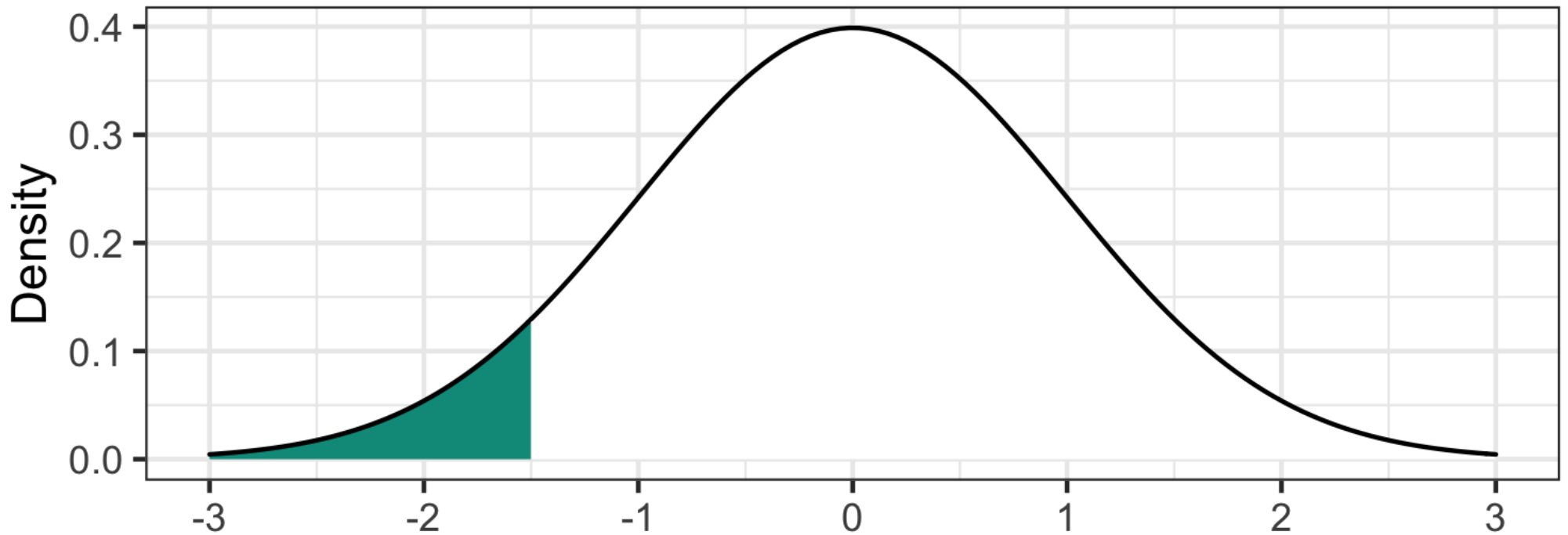
That is,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Probabilities under N(0,1) curve

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

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

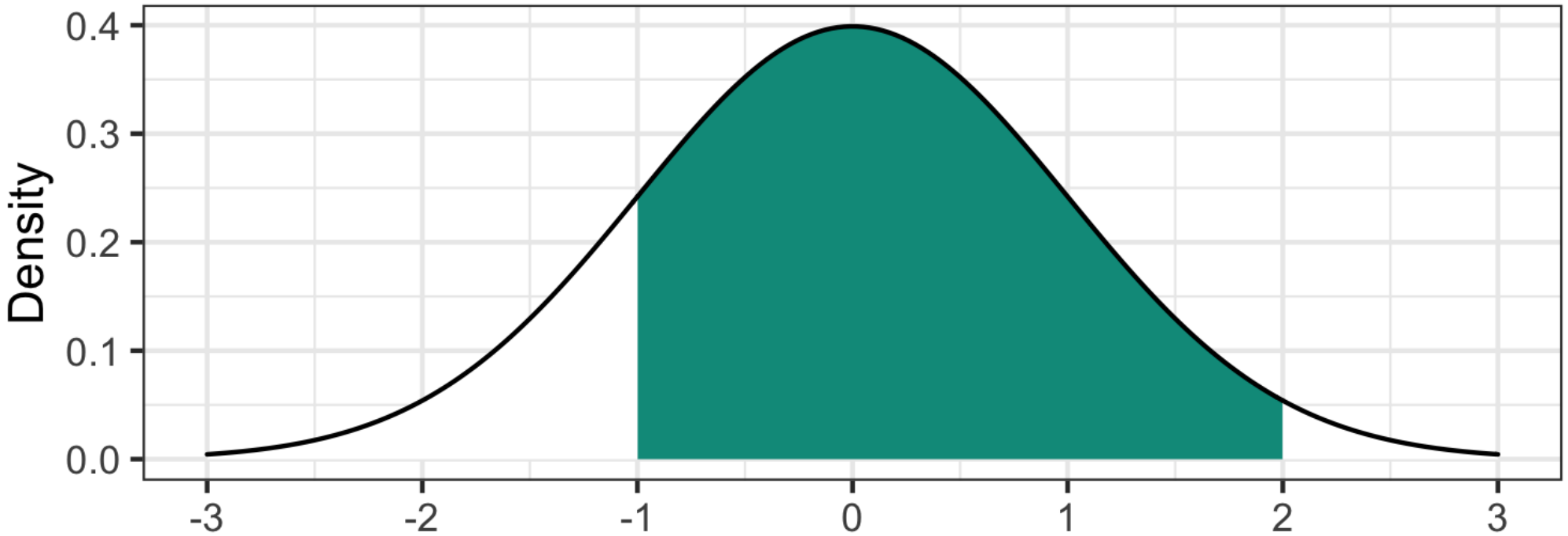


Probability between two values

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

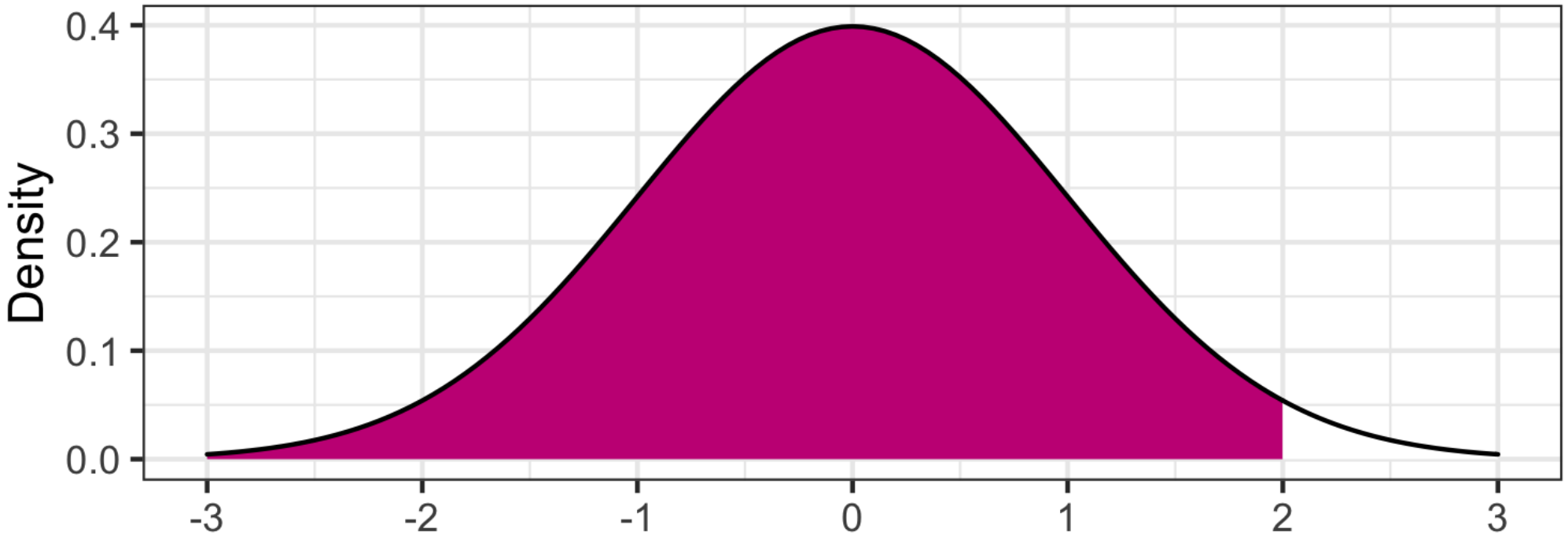
Probability between two values

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



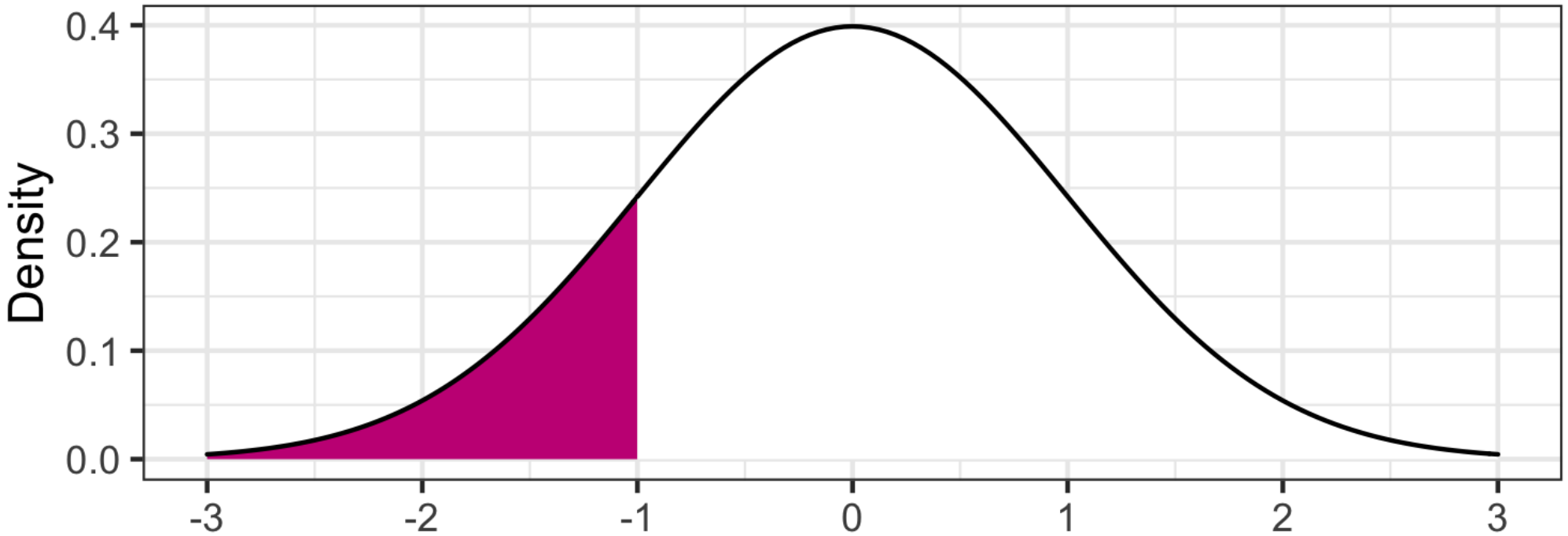
Probability between two values

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



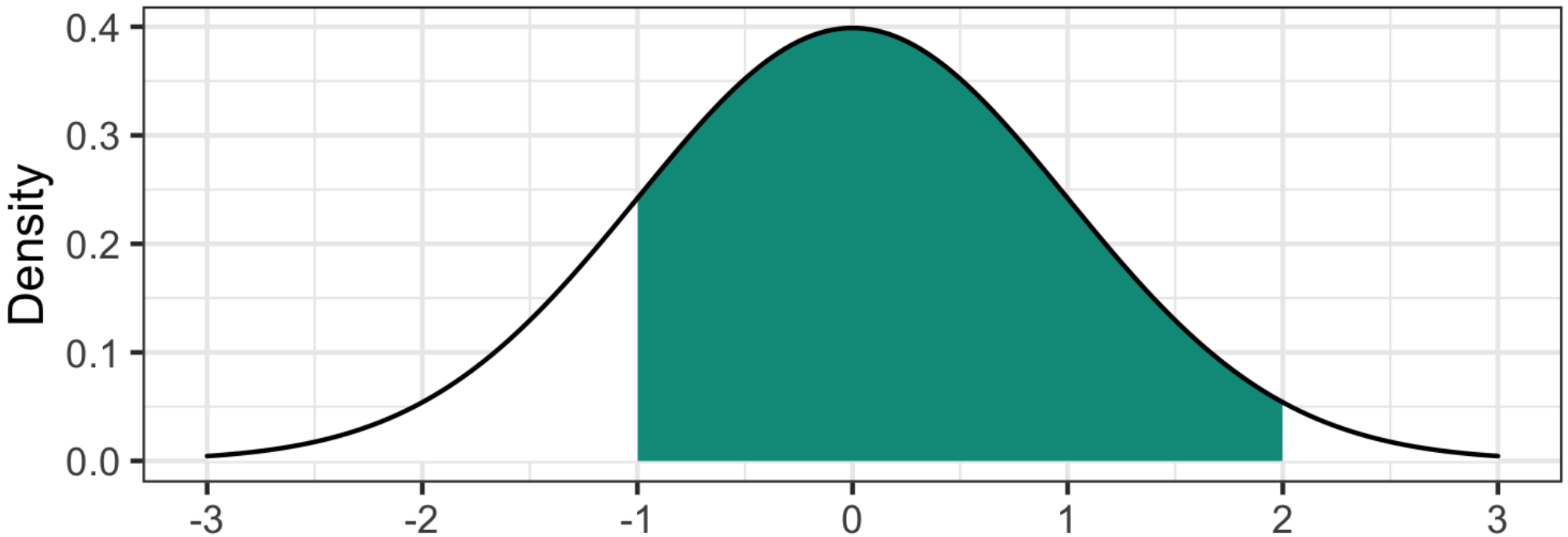
Probability between two values

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 $P(-1 < Z < 2)$?

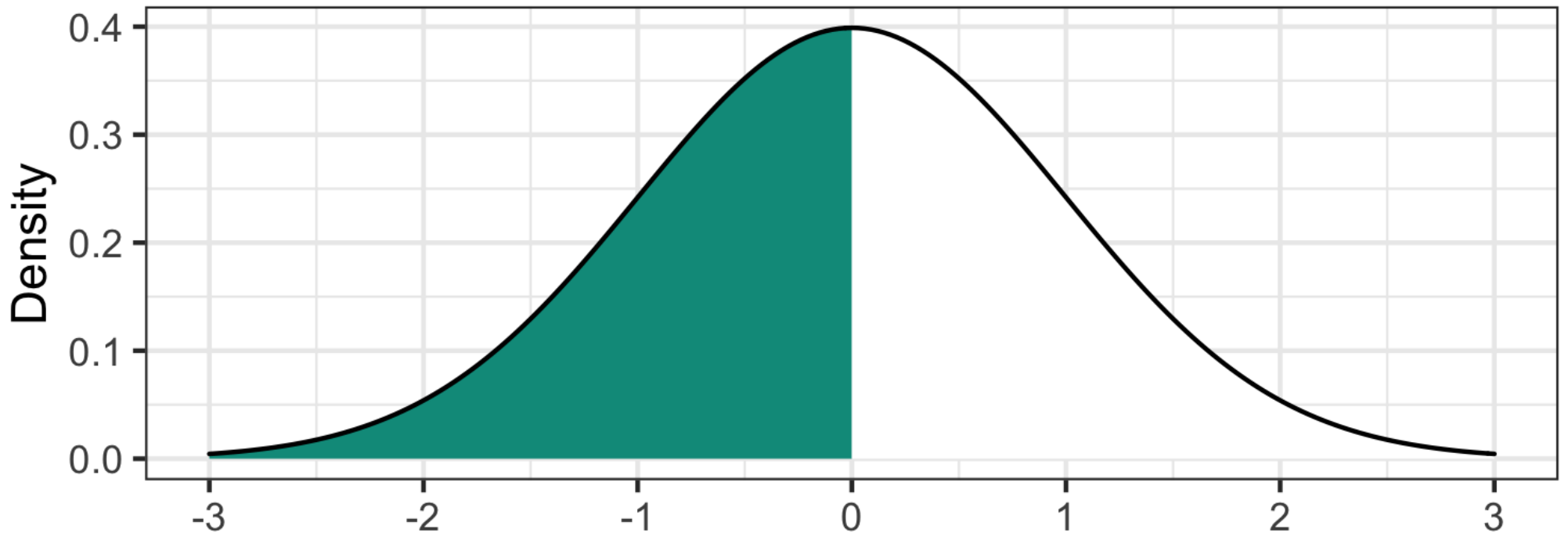
```
pnorm(2) - pnorm(-1)
```

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

Finding cutoff values under N(0,1) curve

```
# find the median, Q2
qnorm(0.5)
```

```
## [1] 0
```



What if σ isn't known?

T distribution

- In practice, we never know the true value of σ, and so we estimate it from our data with s.

We can make the following test statistic for testing a single sample's population mean, which has a **t-distribution with n-1 degrees of freedom**:

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

T distribution

The t-distribution is also unimodal and symmetric, and is centered at 0

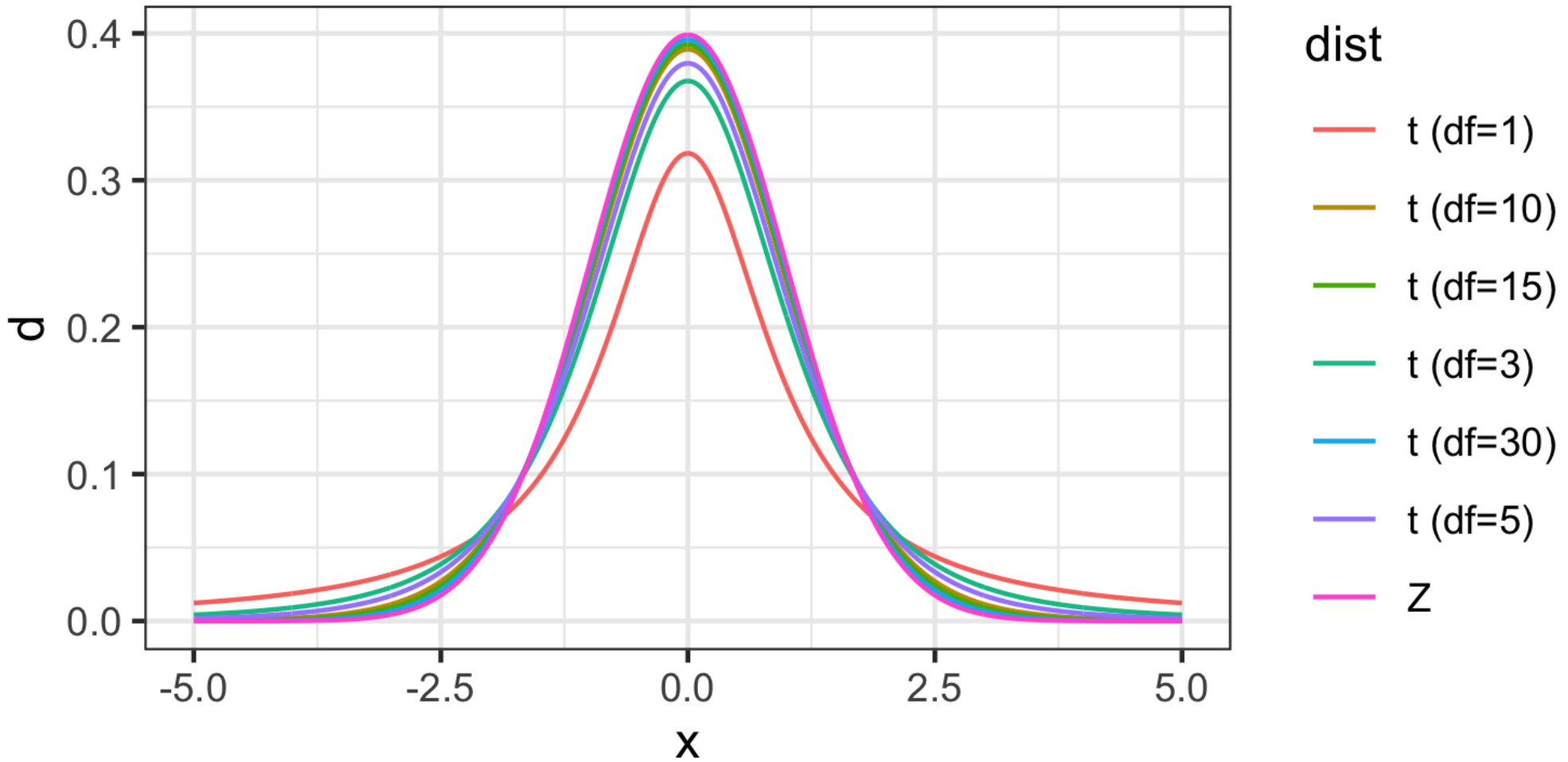
T distribution

The t-distribution is also unimodal and symmetric, and is centered at 0

Thicker tails than the normal distribution

- This is to make up for additional variability introduced by using s instead of σ in calculation of the SE

T vs Z distributions



T distribution

Finding probabilities under the t curve:

```
#P(t < -1.96)
pt(-1.96, df = 9)
```

```
## [1] 0.0408222
```

```
#P(t > -1.96)
pt(-1.96, df = 9,
   lower.tail = FALSE)
```

```
## [1] 0.9591778
```

T distribution

Finding probabilities under the t curve:

```
#P(t < -1.96)
pt(-1.96, df = 9)
```

```
## [1] 0.0408222
```

```
#P(t > -1.96)
pt(-1.96, df = 9,
   lower.tail = FALSE)
```

```
## [1] 0.9591778
```

Finding cutoff values under the t curve:

```
# Find Q1
qt(0.25, df = 9)
```

```
## [1] -0.7027221
```

```
# Q3
qt(0.75, df = 9)
```

```
## [1] 0.7027221
```

Resident satisfaction in Durham

`durham_survey` contains resident responses to a survey given by the City of Durham in 2018. These are a randomly selected, representative sample of Durham residents.

Questions were rated 1 - 5, with 1 being "highly dissatisfied" and 5 being "highly satisfied."

Resident satisfaction in Durham

`durham_survey` contains resident responses to a survey given by the City of Durham in 2018. These are a randomly selected, representative sample of Durham residents.

Questions were rated 1 - 5, with 1 being "highly dissatisfied" and 5 being "highly satisfied."

Is there evidence that, on average, Durham residents are generally satisfied (score greater than 3) with the quality of the public library system?

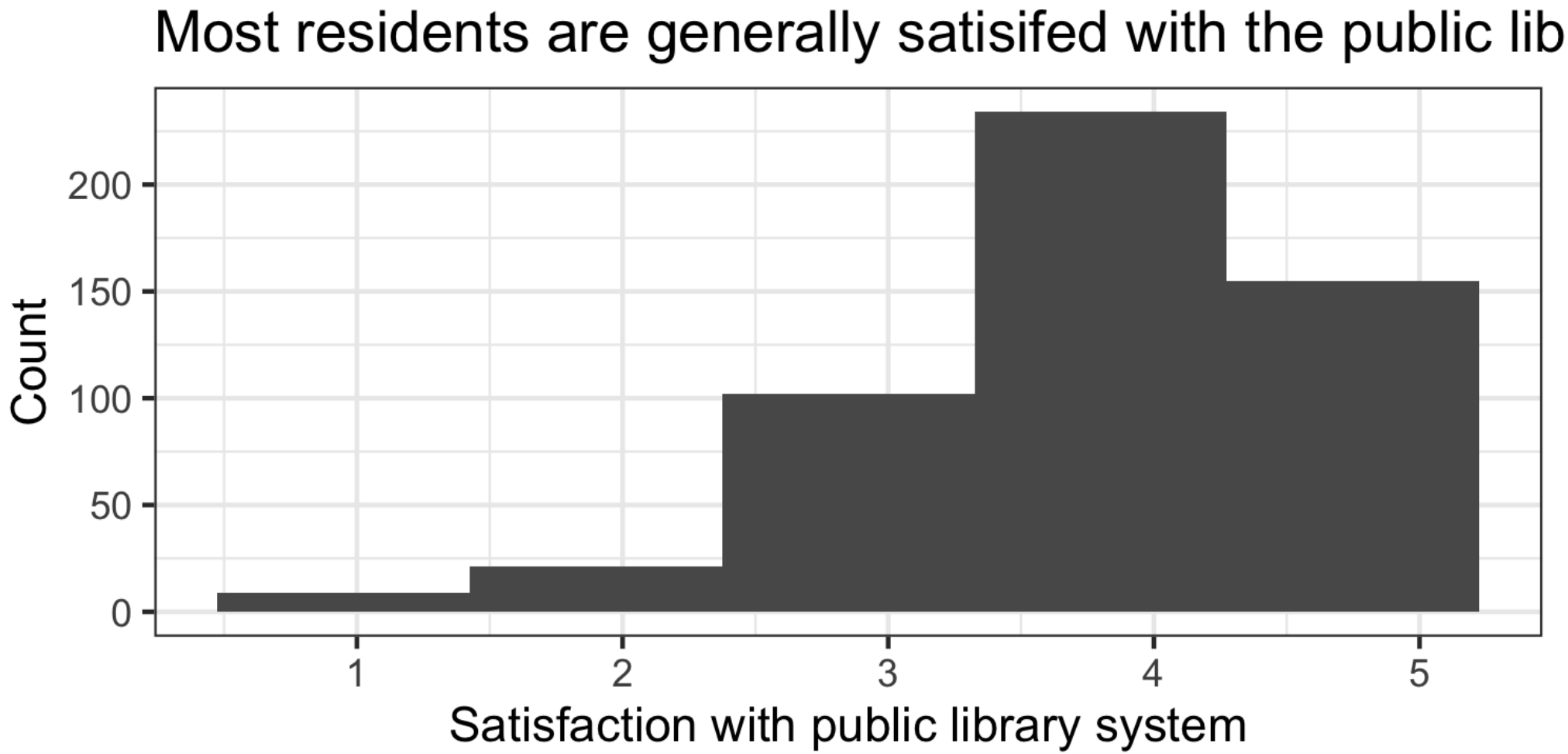
Exploratory Data Analysis

```
durham <- read_csv("data/durham_survey.csv") %>%
  filter(quality_library != 9)
```

```
durham %>%
  summarise(x_bar = mean(quality_library),
            med = median(quality_library),
            sd = sd(quality_library), n = n())
```

```
## # A tibble: 1 x 4
##   x_bar   med    sd     n
##   <dbl> <dbl> <dbl> <int>
## 1  3.97     4 0.900   521
```

Exploratory Data Analysis



Hypotheses

What are the hypotheses for evaluating if Durham residents, on average, are generally satisfied with the public library system?

Hypotheses

What are the hypotheses for evaluating if Durham residents, on average, are generally satisfied with the public library system?

$$H_0 : \mu = 3$$

$$H_a : \mu > 3$$

Conditions

What conditions must be satisfied to conduct this hypothesis test using methods based on the CLT? Are these conditions satisfied?

Conditions

What conditions must be satisfied to conduct this hypothesis test using methods based on the CLT? Are these conditions satisfied?

Independence?

Conditions

What conditions must be satisfied to conduct this hypothesis test using methods based on the CLT? Are these conditions satisfied?

Independence?

✅ The residents were randomly selected for the survey, and 521 is less than 10% of the Durham population (~ 270,000).

Conditions

What conditions must be satisfied to conduct this hypothesis test using methods based on the CLT? Are these conditions satisfied?

Independence?

✅ The residents were randomly selected for the survey, and 521 is less than 10% of the Durham population (~ 270,000).

Sample size / distribution?

Conditions

What conditions must be satisfied to conduct this hypothesis test using methods based on the CLT? Are these conditions satisfied?

Independence?

✅ The residents were randomly selected for the survey, and 521 is less than 10% of the Durham population (~ 270,000).

Sample size / distribution?

✅ 521 > 30, so the sample is large enough to apply the Central Limit Theorem.

Calculating the test statistic

Summary statistics from the sample:

```
## # A tibble: 1 x 3
##    xbar     s     n
##   <dbl> <dbl> <int>
## 1  3.97 0.900   521
```

Calculating the test statistic

Summary statistics from the sample:

```
## # A tibble: 1 x 3
##    xbar     s     n
##   <dbl> <dbl> <int>
## 1  3.97 0.900   521
```

And the CLT says:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

Calculating the test statistic

Summary statistics from the sample:

```
## # A tibble: 1 x 3
##    xbar     s     n
##   <dbl> <dbl> <int>
## 1  3.97 0.900   521
```

And the CLT says:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

How many standard errors away from the population mean is the observed sample mean?

Calculations

Calculations

```
(se <- durham_summary$s / sqrt(durham_summary$n)) # SE
```

```
## [1] 0.03944416
```

Calculations

```
(se <- durham_summary$s / sqrt(durham_summary$n)) # SE
```

```
## [1] 0.03944416
```

```
(t <- (durham_summary$xbar - 3) / se) # Test statistic
```

```
## [1] 24.57372
```

Calculations

```
(se <- durham_summary$s / sqrt(durham_summary$n)) # SE
```

```
## [1] 0.03944416
```

```
(t <- (durham_summary$xbar - 3) / se) # Test statistic
```

```
## [1] 24.57372
```

```
(df <- durham_summary$n - 1) # Degrees of freedom
```

```
## [1] 520
```

Calculations

```
(se <- durham_summary$s / sqrt(durham_summary$n)) # SE
```

```
## [1] 0.03944416
```

```
(t <- (durham_summary$xbar - 3) / se) # Test statistic
```

```
## [1] 24.57372
```

```
(df <- durham_summary$n - 1) # Degrees of freedom
```

```
## [1] 520
```

```
pt(t, df, lower.tail = FALSE) # P-value, P(T > t |H_0 true)
```

```
## [1] 2.247911e-89
```

Conclusion

The p-value is very small, so we reject H_0.

Conclusion

The p-value is very small, so we reject H_0.

The data provide sufficient evidence at the $\alpha = 0.05$ level that Durham residents, on average, are satisfied with the quality of the public library system.

Conclusion

The p-value is very small, so we reject H_0.

The data provide sufficient evidence at the $\alpha = 0.05$ level that Durham residents, on average, are satisfied with the quality of the public library system.

Would you expect a 95% confidence interval to include 3?

Confidence interval for a mean

General form of the confidence interval

$$point\ estimate \pm critical\ value \times SE$$

Confidence interval for a mean

General form of the confidence interval

$$point\ estimate \pm critical\ value \times SE$$

Confidence interval for the mean

$$\bar{x} \pm t^*_{n-1} \times \frac{s}{\sqrt{n}}$$

Calculate 95% confidence interval

$$\bar{x} \pm t^*_{n-1} \times \frac{s}{\sqrt{n}}$$

Calculate 95% confidence interval

$$\bar{x} \pm t^*_{n-1} \times \frac{s}{\sqrt{n}}$$

```
# Critical value
t_star <- qt(0.975, df)
```

Calculate 95% confidence interval

$$\bar{x} \pm t^*_{n-1} \times \frac{s}{\sqrt{n}}$$

```
# Critical value
t_star <- qt(0.975, df)
```

```
# Point estimate
point_est <- durham_summary$xbar
```

Calculate 95% confidence interval

$$\bar{x} \pm t^*_{n-1} \times \frac{s}{\sqrt{n}}$$

```
# Critical value
t_star <- qt(0.975, df)
```

```
# Point estimate
point_est <- durham_summary$xbar
```

```
# Confidence interval
round(point_est + c(-1,1) * t_star * se, 2)
```

```
## [1] 3.89 4.05
```

Interpret 95% confidence interval

```
round(point_est + c(-1,1) * t_star * se, 2)
```

```
## [1] 3.89 4.05
```

Interpret this interval in context of the data.

Interpret 95% confidence interval

```
round(point_est + c(-1,1) * t_star * se, 2)
```

```
## [1] 3.89 4.05
```

Interpret this interval in context of the data.

We are 95% confident that the true mean rating for Durham residents' satisfaction with the library system is between 3.89 and 4.05.

Inference with the CLT using `infer`

CLT-based hypothesis testing in `infer`

$$H_0 : \mu = 3 \text{ vs } H_a : \mu > 3$$

CLT-based hypothesis testing in `infer`

$$H_0 : \mu = 3 \text{ vs } H_a : \mu > 3$$

```
durham %>%
  t_test(response = quality_library,
         mu = 3,
         alternative = "greater",
         conf_int = FALSE)
```

```
## # A tibble: 1 x 4
##   statistic  t_df  p_value alternative
##       <dbl> <dbl>    <dbl> <chr>
## 1      24.6   520 2.25e-89 greater
```

CLT-based confidence intervals in `infer`

Calculate a 95% confidence interval for the mean satisfaction rating.

CLT-based confidence intervals in `infer`

Calculate a 95% confidence interval for the mean satisfaction rating.

```
durham %>%
  t_test(response = quality_library,
         alternative = "two-sided",
         conf_int = TRUE, conf_level = 0.95)
```

```
## # A tibble: 1 x 6
##   statistic  t_df p_value alternative lower_ci upper_ci
##       <dbl> <dbl>   <dbl> <chr>          <dbl>    <dbl>
## 1      101.   520       0 two.sided       3.89     4.05
```

Other built-in functionality in R

- There are more built in functions for doing some of these tests in R.
- However a learning goal is this course is not to go through an exhaustive list of all CLT based tests and how to implement them
- Instead the goal is to understand how these methods are / are not like the simulation based methods we learned about earlier

Other built-in functionality in R

- There are more built in functions for doing some of these tests in R.
- However a learning goal is this course is not to go through an exhaustive list of all CLT based tests and how to implement them
- Instead the goal is to understand how these methods are / are not like the simulation based methods we learned about earlier

What is similar, and what is different, between CLT based test of means vs. simulation based test?