Inference for 1 Mean

Prof. Wells

STA 209, 4/24/23

Outline

In this lecture, we will. . .

Outline

In this lecture, we will. . .

- Investigate the t distribution.
- Create confidence intervals and perform hypothesis tests using t distribution for sample means.

Section 1

The t-distribution

Distribution of Sample Means

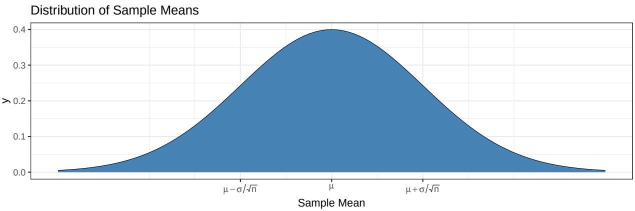
- For quantitative data, the parameter of interest is often the population mean μ, which may be estimated using a sample mean $\bar{x}$.

Distribution of Sample Means

- For quantitative data, the parameter of interest is often the population mean μ, which may be estimated using a sample mean $\bar{x}$.
- By the Central Limit Theorem, the distribution of $\bar{x}$ is approximately Normal, with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$
  - where n is the sample size and σ is the population standard deviation

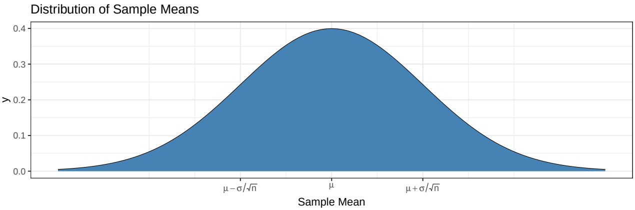
Distribution of Sample Means

- For quantitative data, the parameter of interest is often the population mean μ, which may be estimated using a sample mean $\bar{x}$.
- By the Central Limit Theorem, the distribution of $\bar{x}$ is approximately Normal, with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$
  - where n is the sample size and σ is the population standard deviation



Distribution of Sample Means

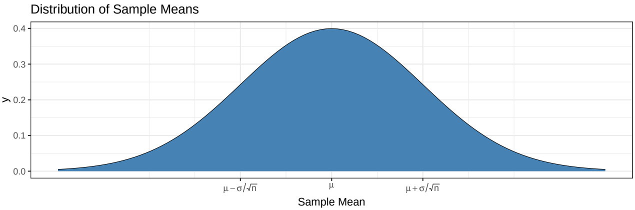
- For quantitative data, the parameter of interest is often the population mean μ, which may be estimated using a sample mean $\bar{x}$.
- By the Central Limit Theorem, the distribution of $\bar{x}$ is approximately Normal, with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$
  - where n is the sample size and σ is the population standard deviation



- Note that smaller σ and larger n both correspond to smaller standard error.

Distribution of Sample Means

- For quantitative data, the parameter of interest is often the population mean μ, which may be estimated using a sample mean $\bar{x}$.
- By the Central Limit Theorem, the distribution of $\bar{x}$ is approximately Normal, with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$
  - where n is the sample size and σ is the population standard deviation



- Note that smaller σ and larger n both correspond to smaller standard error.
- As n increases, Normal approximation becomes more accurate, even if population is skewed.

Practical Considerations

- Suppose we have a single sample that we want to use to estimate μ. The standard error for $\bar{x}$ is

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$

Practical Considerations

- Suppose we have a single sample that we want to use to estimate μ. The standard error for $\bar{x}$ is
$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$
  - But if we are estimating the population mean μ, we usually don't know the value of σ.

Practical Considerations

- Suppose we have a single sample that we want to use to estimate μ. The standard error for $\bar{x}$ is
$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$
    - But if we are estimating the population mean μ, we usually don't know the value of σ.
- Instead, we can use the sample's standard deviation s to estimate σ:
$$SE(\bar{x}) \approx \frac{s}{\sqrt{n}}.$$

Practical Considerations

- Suppose we have a single sample that we want to use to estimate μ. The standard error for $\bar{x}$ is
$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$
  - But if we are estimating the population mean μ, we usually don't know the value of σ.
- Instead, we can use the sample's standard deviation s to estimate σ:
$$SE(\bar{x}) \approx \frac{s}{\sqrt{n}}.$$
- But this adds a new complication! The standardized statistic
$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$
no longer follows a standard Normal distribution!

Practical Considerations

- Suppose we have a single sample that we want to use to estimate μ. The standard error for $\bar{x}$ is
$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$
  - But if we are estimating the population mean μ, we usually don't know the value of σ.
- Instead, we can use the sample's standard deviation s to estimate σ:
$$SE(\bar{x}) \approx \frac{s}{\sqrt{n}}.$$
- But this adds a new complication! The standardized statistic
$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$
no longer follows a standard Normal distribution!
  - This is because both $\bar{x}$ and s are now random variables, and so each adds variability to z.

Practical Considerations

- Suppose we have a single sample that we want to use to estimate μ. The standard error for $\bar{x}$ is
$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$
  - But if we are estimating the population mean μ, we usually don't know the value of σ.
- Instead, we can use the sample's standard deviation s to estimate σ:
$$SE(\bar{x}) \approx \frac{s}{\sqrt{n}}.$$
- But this adds a new complication! The standardized statistic
$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$
no longer follows a standard Normal distribution!
  - This is because both $\bar{x}$ and s are now random variables, and so each adds variability to z.
- Instead, the standardized statistic z follows a t-distribution

Practical Considerations

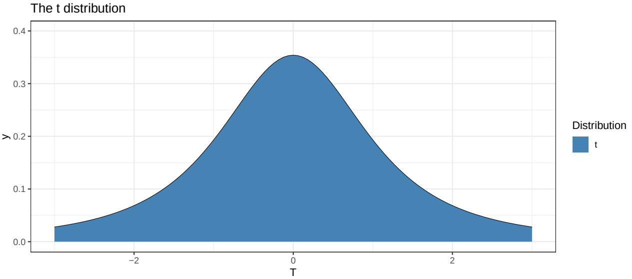
- Suppose we have a single sample that we want to use to estimate μ. The standard error for $\bar{x}$ is
$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$
  - But if we are estimating the population mean μ, we usually don't know the value of σ.
- Instead, we can use the sample's standard deviation s to estimate σ:
$$SE(\bar{x}) \approx \frac{s}{\sqrt{n}}.$$
- But this adds a new complication! The standardized statistic
$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$
no longer follows a standard Normal distribution!
  - This is because both $\bar{x}$ and s are now random variables, and so each adds variability to z.
- Instead, the standardized statistic z follows a t-distribution
  - The t-distribution was first studied in 1908 by William Gosset, who published under the pseudonym *Student*.

The t-distribution

- Like the standard Normal distribution, a t-distribution is symmetric, single-peaked, bell-shaped and centered at 0.

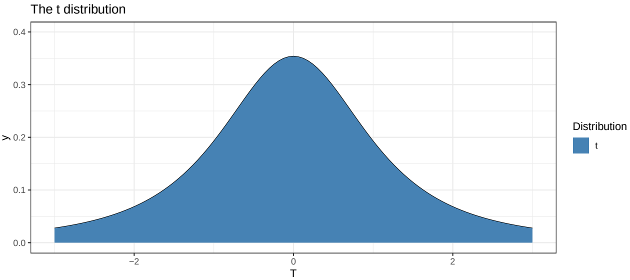
The t-distribution

- Like the standard Normal distribution, a t-distribution is symmetric, single-peaked, bell-shaped and centered at 0.



The t-distribution

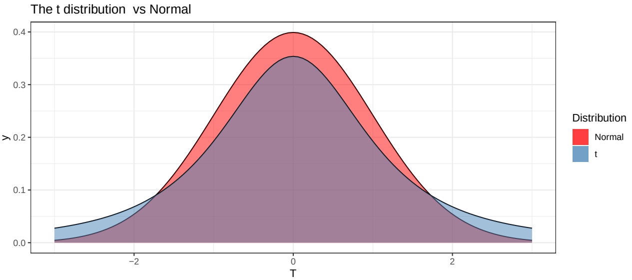
- Like the standard Normal distribution, a t-distribution is symmetric, single-peaked, bell-shaped and centered at 0.



- But a t-distribution has *heavier tails* than the Normal distribution.

The t-distribution

- Like the standard Normal distribution, a t-distribution is symmetric, single-peaked, bell-shaped and centered at 0.



- But the t-distribution has *heavier tails* than the Normal distribution

Degrees of Freedom

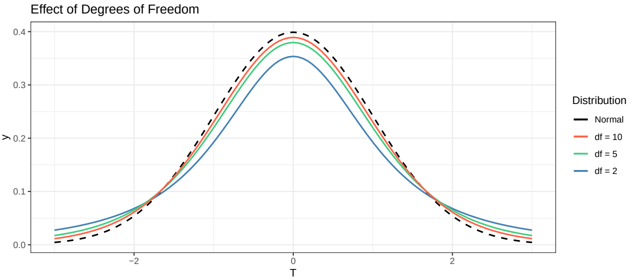
- The t-distribution is actually a parameterized *family* of distributions.

Degrees of Freedom

- The t-distribution is actually a parameterized *family* of distributions.
- The parameter for the t-distribution is called the degrees of freedom df and determines the **heaviness of tails**

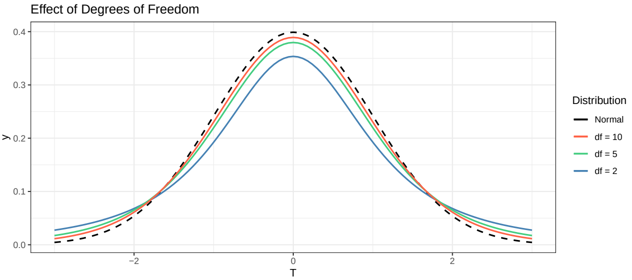
Degrees of Freedom

- The t-distribution is actually a parameterized *family* of distributions.
- The parameter for the t-distribution is called the degrees of freedom *df* and determines the **heaviness of tails**



Degrees of Freedom

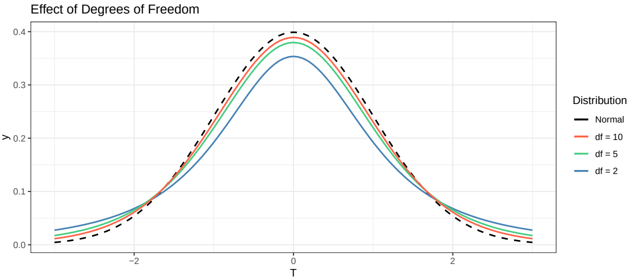
- The t-distribution is actually a parameterized *family* of distributions.
- The parameter for the t-distribution is called the degrees of freedom df and determines the **heaviness of tails**



- As degrees of freedom increases, the t distribution gets closer to the Normal distribution.

Degrees of Freedom

- The *t*-distribution is actually a parameterized *family* of distributions.
- The parameter for the *t*-distribution is called the degrees of freedom *df* and determines the **heaviness of tails**



- As degrees of freedom increases, the *t* distribution gets closer to the Normal distribution.
    - For $df \geq 30$, the *t* distribution is nearly indistinguishable from the Normal

Distribution of Sample Means using t-distribution

Theorem

Suppose a sample of size n is collected from a population with mean μ. The distribution of the sample mean $\bar{x}$ has the following characteristics:

- ***Center***: *The mean is equal to μ*
- ***Spread***: *The standard error is equal to $\frac{s}{\sqrt{n}}$ (where s is the sample st. dev.)*
- ***Shape***: *The standardized statistic follows approximately a t-distribution with $n-1$ degrees of freedom.*

For small sample sizes ($n \leq 30$), the t-distribution is only a good approximation if the population distribution is approximately Normal.

Section 2

Statistical Inference

The Origin Story

A batch of stout beer is best when it has an *original gravity* (OG) close to 1.071. The particular OG of a batch depends on a number factors (like temperature, rest time, recipe, etc.).

If we can only obtain a small number of measurements from the batch, how can we quantify whether the deviations we observe are due to random sampling, and not an actual deviation in OG?

Confidence Intervals

The t-procedures for Confidence Intervals

A $C\%$ confidence interval for a population mean μ using a sample of size n is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where $\bar{x}$ and s are the mean and standard deviation of the sample, and where t^* is the critical value for $C\%$ confidence in the t-distribution with $n-1$ degrees of freedom.

The t-procedures are appropriate if $n \leq 30$ and the population is approximately Normal, or if $n > 30$.

Confidence Interval for Original Gravity

Suppose we obtain the following 5 OG measurements from a batch of beer:

Confidence Interval for Original Gravity

Suppose we obtain the following 5 OG measurements from a batch of beer:

```
## [1] 1.067 1.060 1.077 1.072 1.067
```

Confidence Interval for Original Gravity

Suppose we obtain the following 5 OG measurements from a batch of beer:

```
## [1] 1.067 1.060 1.077 1.072 1.067
```

Goal: Create a 95% confidence interval for the true OG of the batch.

Confidence Interval for Original Gravity

Suppose we obtain the following 5 OG measurements from a batch of beer:

```
## [1] 1.067 1.060 1.077 1.072 1.067
```

Goal: Create a 95% confidence interval for the true OG of the batch.

- Since our sample size is small ($n \leq 30$), we need to make sure our population is approximately Normal.

Confidence Interval for Original Gravity

Suppose we obtain the following 5 OG measurements from a batch of beer:

```
## [1] 1.067 1.060 1.077 1.072 1.067
```

Goal: Create a 95% confidence interval for the true OG of the batch.

- Since our sample size is small ($n \leq 30$), we need to make sure our population is approximately Normal.
  - Fortunately, the only variability here is due to measurement errors, which are known to be approximately Normally distributed.

Confidence Interval for Original Gravity

Suppose we obtain the following 5 OG measurements from a batch of beer:

```
## [1] 1.067 1.060 1.077 1.072 1.067
```

Goal: Create a 95% confidence interval for the true OG of the batch.

- Since our sample size is small ($n \leq 30$), we need to make sure our population is approximately Normal.
  - Fortunately, the only variability here is due to measurement errors, which are known to be approximately Normally distributed.
- Our sample mean and standard deviation are

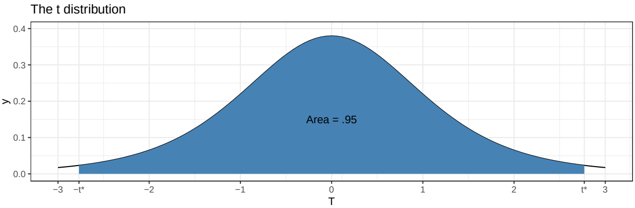
```
##    xbar        s
## 1 1.069 0.006348
```

The Critical Value

- We need the t^* critical value for 95% confidence from the t-distribution with $df = 4$.

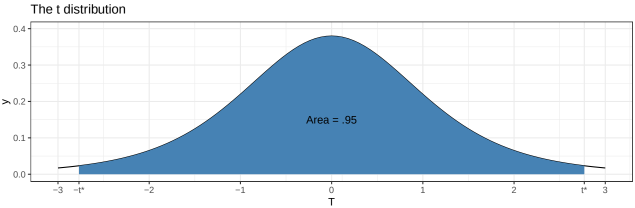
The Critical Value

- We need the t^* critical value for 95% confidence from the t-distribution with $df = 4$.



The Critical Value

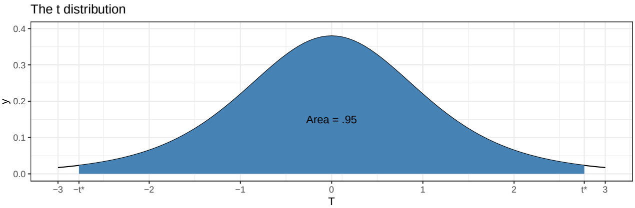
- We need the t^* critical value for 95% confidence from the t-distribution with $df = 4$.



- Note that t^* is the 0.975 quantile for the t-distribution with 4 degrees of freedom.

The Critical Value

- We need the t^* critical value for 95% confidence from the t-distribution with $df = 4$.



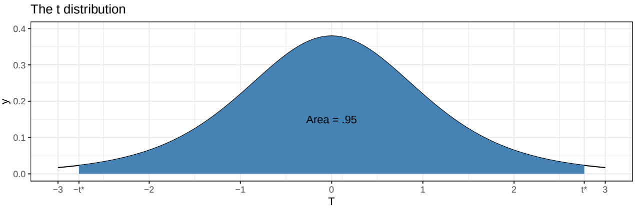
- Note that t^* is the 0.975 quantile for the t-distribution with 4 degrees of freedom.
- Use the function qt to get quantiles for a t-distribution (just like qnorm for Normal)

```
qt(p = 0.975, df = 4)
```

```
## [1] 2.776
```

The Critical Value

- We need the t^* critical value for 95% confidence from the t-distribution with $df = 4$.



- Note that t^* is the 0.975 quantile for the t-distribution with 4 degrees of freedom.
- Use the function qt to get quantiles for a t-distribution (just like qnorm for Normal)

```
qt(p = 0.975, df = 4)
```

```
## [1] 2.776
```

- Note that the $t*$ critical value of 95% confidence is larger than the z^* critical value

```
qnorm(p = 0.975)
```

```
## [1] 1.96
```

The Confidence Interval

- The 95% confidence interval given by

$$\text{statistic} \pm t^* \cdot SE \qquad \bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

The Confidence Interval

- The 95% confidence interval given by

$$\text{statistic} \pm t^* \cdot SE \qquad \bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

- Previously, we found

```
##    xbar        s n t_star
## 1 1.069 0.006348 5  2.776
```

The Confidence Interval

- The 95% confidence interval given by

$$\text{statistic} \pm t^* \cdot SE \qquad \bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

- Previously, we found

```
##     xbar        s n t_star
## 1 1.069 0.006348 5  2.776
```

- Putting all these values into place, our confidence interval is

$$1.069 \pm 2.776 \cdot \frac{0.006348}{\sqrt{5}} \qquad \text{or} \qquad 1.069 \pm 0.0079$$

The Confidence Interval

- The 95% confidence interval given by

$$\text{statistic} \pm t^* \cdot SE \qquad \bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

- Previously, we found

```
##    xbar        s n t_star
## 1 1.069 0.006348 5  2.776
```

- Putting all these values into place, our confidence interval is

$$1.069 \pm 2.776 \cdot \frac{0.006348}{\sqrt{5}} \qquad \text{or} \qquad 1.069 \pm 0.0079$$

- Thus, the range of plausible values for the OG of the beer is $(1.061, 1.076)$ at 95% confidence.

The Confidence Interval

- The 95% confidence interval given by

$$\text{statistic} \pm t^* \cdot SE \qquad \bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

- Previously, we found

```
##     xbar        s n t_star
## 1 1.069 0.006348 5  2.776
```

- Putting all these values into place, our confidence interval is

$$1.069 \pm 2.776 \cdot \frac{0.006348}{\sqrt{5}} \qquad \text{or} \qquad 1.069 \pm 0.0079$$

- Thus, the range of plausible values for the OG of the beer is $(1.061, 1.076)$ at 95% confidence.
- As $\mu = 1.071$ is within this interval, it is plausible that the batch has the desired OG.

Comparison using infer

If we instead use infer...

```
set.seed(1908)
beer %>%
  specify(response = OG) %>%
  generate(reps = 5000, type = "bootstrap" ) %>%
  calculate(stat = "mean") %>%
  get_ci(level = .95, type = "percentile")
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    1.064    1.074
```

- The bootstrap interval is a bit narrower than the theory-based interval:

```
## # A tibble: 1 x 2
##   theory_lower_ci theory_upper_ci
##             <dbl>           <dbl>
## 1           1.061           1.076
```

Comparison using infer

If we instead use infer...

```
set.seed(1908)
beer %>%
  specify(response = OG) %>%
  generate(reps = 5000, type = "bootstrap" ) %>%
  calculate(stat = "mean") %>%
  get_ci(level = .95, type = "percentile")
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    1.064    1.074
```

- The bootstrap interval is a bit narrower than the theory-based interval:

```
## # A tibble: 1 x 2
##   theory_lower_ci theory_upper_ci
##             <dbl>           <dbl>
## 1           1.061           1.076
```

- Our sample size is small $n = 5$. So if the model assumptions are true (measurement errors are **exactly Normal**) then the theory-based confidence interval is a better estimate
  - The bootstrap method has intervals that are too narrow, and so will have success rate lower than 95%

Hypothesis Tests

The t-test for Single Mean

To test $H_0 : \mu = \mu_0$ against $H_a : \mu \neq \mu_0$ (or 1-sided alternatives), use the t-statistic

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

where $\bar{x}$ and s are the mean and standard deviation of the sample with size n. The distribution of t is approximated by the t-distribution with $n - 1$ degrees of freedom.

The t-procedures are appropriate if $n \leq 30$ and the population is approximately Normal, or if $n > 30$.

Hypothesis Test for OG

Consider the previous sample of 5 OG measurements from a batch of beer:

Hypothesis Test for OG

Consider the previous sample of 5 OG measurements from a batch of beer:

```
## [1] 1.067 1.060 1.077 1.072 1.067
```

Hypothesis Test for OG

Consider the previous sample of 5 OG measurements from a batch of beer:

```
## [1] 1.067 1.060 1.077 1.072 1.067
```

Goal: Determine whether this sample gives evidence that the OG isn't 1.071.

Hypothesis Test for OG

Consider the previous sample of 5 OG measurements from a batch of beer:

```
## [1] 1.067 1.060 1.077 1.072 1.067
```

Goal: Determine whether this sample gives evidence that the OG isn't 1.071.

- Our Null and Alternate Hypotheses are

$$H_0 : \mu = 1.071 \qquad H_a : \mu \neq 1.071$$

Hypothesis Test for OG

Consider the previous sample of 5 OG measurements from a batch of beer:

```
## [1] 1.067 1.060 1.077 1.072 1.067
```

Goal: Determine whether this sample gives evidence that the OG isn't 1.071.

- Our Null and Alternate Hypotheses are

$$H_0 : \mu = 1.071 \qquad H_a : \mu \neq 1.071$$

- Our sample mean and standard deviation are

```
##    xbar        s
## 1 1.069 0.006348
```

Hypothesis Test for OG

Consider the previous sample of 5 OG measurements from a batch of beer:

```
## [1] 1.067 1.060 1.077 1.072 1.067
```

Goal: Determine whether this sample gives evidence that the OG isn't 1.071.

- Our Null and Alternate Hypotheses are

$$H_0 : \mu = 1.071 \qquad H_a : \mu \neq 1.071$$

- Our sample mean and standard deviation are

```
##    xbar        s
## 1 1.069 0.006348
```

- Therefore, our t-statistic is

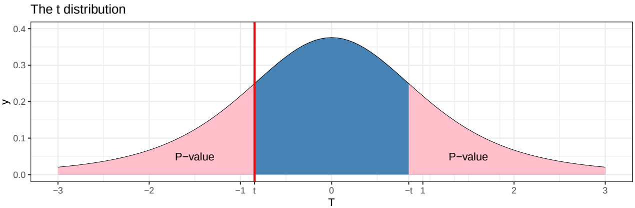
$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{1.0686 - 1.071}{\frac{0.0063}{\sqrt{5}}} = -0.845$$

The P-Value

- Plotting our t-statistic against the theoretical t-distribution with $df = 4$

The P-Value

- Plotting our t-statistic against the theoretical t-distribution with $df = 4$



- The exact P-value is twice the area left of t:

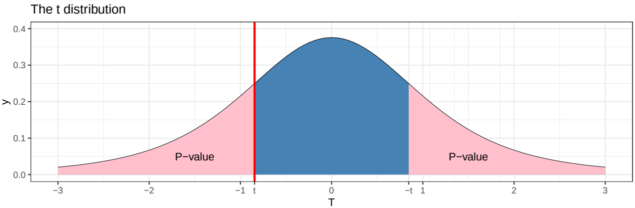
```
2*pt( q = -.845 , df = 4)
```

```
## [1] 0.4457
```

- At significance $\alpha = 0.05$, we do not have enough evidence to reject the null hypothesis.

The P-Value

- Plotting our t-statistic against the theoretical t-distribution with $df = 4$



- The exact P-value is twice the area left of t:

```
2*pt( q = -.845 , df = 4)
```

```
## [1] 0.4457
```

- At significance $\alpha = 0.05$, we do not have enough evidence to reject the null hypothesis.
    - Our sample is consistent with a true mean OG of $\mu = 1.071$ (of course, this doesn't prove that H_0 is correct!)
    - With a sample size of $n = 5$, we have limited ability to detect a true difference

Comparison using infer

If we instead use infer...

```
set.seed(1908)
beer %>%
  specify(response = OG) %>%
  hypothesize(null = "point", mu = 1.071) %>%
  generate(reps = 5000, type = "bootstrap" ) %>%
  calculate(stat = "mean") %>%
  get_p_value(obs_stat = 1.069, direction = "both")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.4984
```

Comparison using infer

If we instead use infer...

```
set.seed(1908)
beer %>%
  specify(response = OG) %>%
  hypothesize(null = "point", mu = 1.071) %>%
  generate(reps = 5000, type = "bootstrap" ) %>%
  calculate(stat = "mean") %>%
  get_p_value(obs_stat = 1.069, direction = "both")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.4984
```

- The bootstrap p-value is a bit larger than the theory-based p-value:

```
## [1] 0.4457
```

Theory vs. Simulation for 1 Mean

- In order to compute accurate confidence intervals and p-values, we need to ensure that appropriate conditions are met.
  - The strictness of these conditions depends on sample size.

Theory vs. Simulation for 1 Mean

- In order to compute accurate confidence intervals and p-values, we need to ensure that appropriate conditions are met.
  - The strictness of these conditions depends on sample size.
- Theory-based can be used if...
  - $n \geq 2$, and the population is Normal (or nearly so)
  - $n \geq 30$, and the population appears at most moderately skewed
  - $n \geq 60$, and the population is not extremely skewed

Theory vs. Simulation for 1 Mean

- In order to compute accurate confidence intervals and p-values, we need to ensure that appropriate conditions are met.
  - The strictness of these conditions depends on sample size.
- Theory-based can be used if...
  - $n \geq 2$, and the population is Normal (or nearly so)
  - $n \geq 30$, and the population appears at most moderately skewed
  - $n \geq 60$, and the population is not extremely skewed
- Simulation methods can be used if
  - $n \geq 15$, and the population is Normal or at most slightly skewed
  - $n \geq 30$, and the population appears at most moderately skewed
  - $n \geq 60$, and the population is not extremely skewed

Theory vs. Simulation for 1 Mean

- In order to compute accurate confidence intervals and p-values, we need to ensure that appropriate conditions are met.
    - The strictness of these conditions depends on sample size.
- Theory-based can be used if...
    - $n \geq 2$, and the population is Normal (or nearly so)
    - $n \geq 30$, and the population appears at most moderately skewed
    - $n \geq 60$, and the population is not extremely skewed
- Simulation methods can be used if
    - $n \geq 15$, and the population is Normal or at most slightly skewed
    - $n \geq 30$, and the population appears at most moderately skewed
    - $n \geq 60$, and the population is not extremely skewed
- In general, for small sample sizes, neither method should be used if population does not appear Normal. But if it is Normal, theory-based methods will be more accurate.

Theory vs. Simulation for 1 Mean

- In order to compute accurate confidence intervals and p-values, we need to ensure that appropriate conditions are met.
  - The strictness of these conditions depends on sample size.
- Theory-based can be used if...
  - $n \geq 2$, and the population is Normal (or nearly so)
  - $n \geq 30$, and the population appears at most moderately skewed
  - $n \geq 60$, and the population is not extremely skewed
- Simulation methods can be used if
  - $n \geq 15$, and the population is Normal or at most slightly skewed
  - $n \geq 30$, and the population appears at most moderately skewed
  - $n \geq 60$, and the population is not extremely skewed
- In general, for small sample sizes, neither method should be used if population does not appear Normal. But if it is Normal, theory-based methods will be more accurate.
- For moderate sample sizes with moderate skew, simulation-based methods will be more accurate