Chapter 6: Inferential Methods for a Single Numerical Variable

As seen in earlier chapters, research hypotheses involving either a single categorical variable or two categorical variables require us to test claims about population proportions. When a research question involves a single numerical variable, however, we end up testing claims about the population mean (or average) based on a sample. In this chapter, we will consider inferential methods (hypothesis tests and confidence intervals) for the population mean of a single numerical variable. Consider the following example.

Notation for Means

Sample (these are statistics)

- $\bar{x}$ is the sample mean (think of this as $\hat{p}$ from proportions)
- s is the sample standard deviation

Population (these are parameters)

- μ is the population mean (think of this as π from proportions)
- σ is the population standard devation

Example 6.1: Time Perception Impaired by Nicotine Withdrawal

A study conducted by researchers at Pennsylvania State University investigated whether time perception, a simple indication of a person's ability to concentrate, is impaired during nicotine withdrawal. The study results were presented in the paper *Smoking Abstinence Impairs Time Estimation Accuracy in Cigarette Smokers* (Klein, Corwin, and Stine 2003). After a 24-hour smoking abstinence, 20 daily smokers were asked to estimate how much time had passed during a 45-second period. Suppose the resulting data on perceived elapsed time (in seconds) were

summarizied and visualized as shown below (these results are artificial but are similar to the actual findings).

Research Question: Is there evidence the mean perceived elapsed time for all smokers suffering from nicotine withdrawal is significantly greater than the actual 45 seconds?

```
# A tibble: 6 x 4
  status     sex      age time_passed
  <chr>      <chr>  <dbl>       <dbl>
1 withdrawal female  33.0        49.0
2 withdrawal male    42.6        47.9
3 withdrawal female  41.8        52.5
4 withdrawal female  37.5        54.0
5 withdrawal male    38.0        41.7
6 withdrawal male    48.0        50.7
```

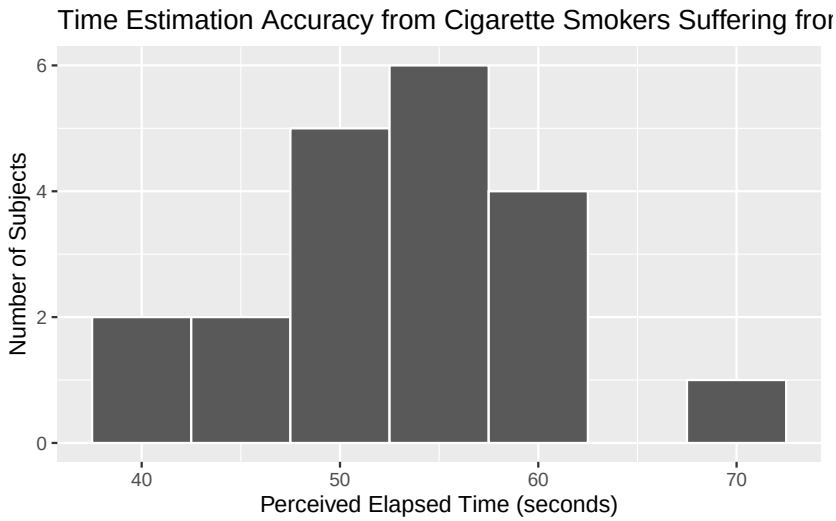
We can summarize the data with:

```
favstats( ~ time_passed, data = nicotine)
```

```
      min       Q1   median       Q3      max     mean       sd  n missing
 39.11611 48.74039 53.25595 56.65965 69.36787 52.65206 7.268201 20       0
```

and visualize the data with:

```
ggplot(data = nicotine,
       mapping = aes(x = time_passed)) +
  geom_histogram(binwidth = 5) +
  labs(title = "Distribution of our data",
       y = "Number of Subjects",
       x = "Perceived Elapsed Time (seconds)")
```



1. Identify the following in context of the scenario:

- Population:

- Sample:

- Variable of interest:

- Data Type:

2. What is the mean of the observed data? The standard deviation?

3. If another sample of $n = 20$ subjects were obtained, would these new subjects have a mean exactly the same as the mean from this sample? Why or why not?

4. Given your answer to the previous question, do you think it is appropriate to use only this sample mean to make inferences about the mean perceived elapsed time being greater then 45 seconds in the population of all smokers subjected to nicotine withdrawal? Explain.

The Distribution of the Sample Mean

The sample mean is a random quantity that changes from sample to sample. This randomness gives the sample mean its own distribution, called the distribution of the sample mean, which tells us:

- The possible values the sample mean can assume.
- How often each value will occur.

Understanding this distribution is key to making decisions about the population mean for a single numerical variable. Let's explore how sample means behave by simulating repeated samples.

Exploring the Distribution of the Sampling Mean

We will use an applet (https://emily-robinson.shinyapps.io/06-CLT-app/) to simulate drawing random samples from a population and observe the distribution of sample means.

The goal is to simulate multiple samples of $n = 20$ drawn from a hypothetical population of smokers experiencing nicotine withdrawal.

Step 1: Population Parameters We begin by setting up a hypothetical population of **all** smokers suffering from nicotine withdraw.

In this simulation study, what are the following parameters for your population?

$\mu =$ _____ – Where does this value of 45 seconds come from? *yours might differ slightly*

$\sigma =$ _____

What does each dot on this plot represent?

Step 2: Repeated Samples In practice, we usually don't know the true population mean (μ). Researchers often take random samples to estimate this mean. In our case, we'll simulate multiple random samples of 20 subjects.

- *Sample 1:* Click "Draw Samples". Record the sample mean and sample standard deviation:

$\bar{x}_{sample1} =$ _____

$s_{sample1} =$ _____

What does each dot on this plot represent?

- *Sample 2:* Click "Draw Samples" again. Record the new values:

$\bar{x}_{sample2} =$ _____

$s_{sample2} =$ _____

- *Sample 3:* Click "Draw Samples" once more. Record:

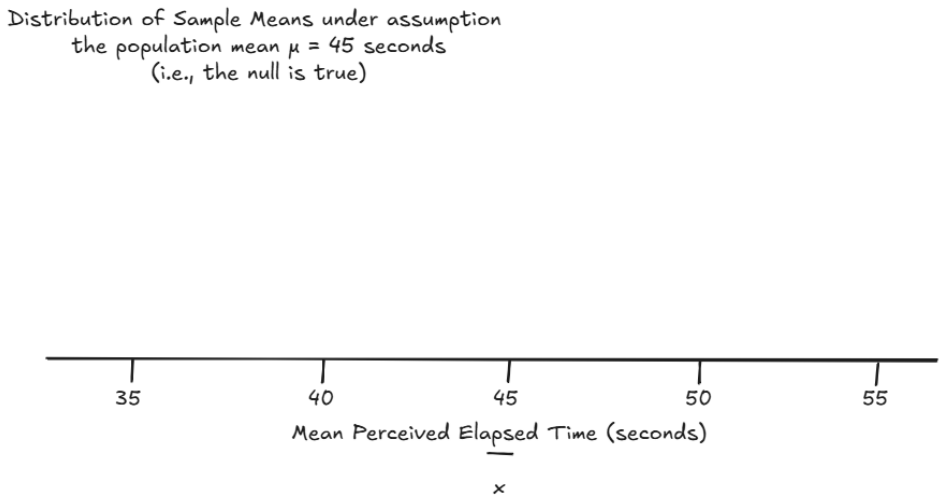
$\bar{x}_{sample3} =$ _____

$s_{sample3} =$ _____

Compare: How do the means and standard deviations from these three samples differ?

Step 3: Distribution of Sample Means: Each of these sample means is a single point from the *distribution of sample means*. Let's now visualize the distribution of sample means:

Add a dot on the plot for each sample mean from the three samples you took above.



Next: To get a better sense of this distribution, increase the "Number of Samples" to 100 and hit "Draw Samples" to generate more sample means. Record the overall mean and standard deviation from the "Distribution of Sample Means".

- Distribution Mean = _____
- Distribution SD = _____

What does each dot on this plot represent?

Step 4: Hypothesis Testing The researchers hypothesized that the mean perceived elapsed time for smokers during nicotine withdrawal was greater than 45 seconds. Let's formally define the hypotheses:

> H_0: The mean perceived elapsed time for *all* smokers suffering from nicotine withdrawal is *equal* to 45 seconds.
>
> H_A: The mean perceived elapsed time for *all* smokers suffering from nicotine withdrawal is *greater* than 45 seconds.

The "Distribution of Sample Means" assumes the null hypothesis is true. It shows what we expect sample means to look like if the true population mean really is 45 seconds.

Step 5: Observed Data Recall, in the actual research study, the mean perceived elapsed time for 20 subjects was 52.65 seconds. Sketch this observed value on the "Distribution of Sample Means" you generated.

Estimate the p-value: Based on the simulation, what is the probability of obtaining a sample mean of 52.65 seconds or greater, assuming the null hypothesis is true?

Estimated p-value = _____

Step 6: Understanding the Central Limit Theorem (CLT)

Rather than always using simulations, statisticians rely on a theoretical result called the *Central Limit Theorem (CLT)*. The CLT allows us to make inferences about the population mean without needing to take repeated samples.

> **i** The Central Limit Theorem (CLT)
>
> The Central Limit Theorem (CLT) allows us to approximate the *distribution of sample means* as normal, provided the sample size is large enough.
>
> - **Mean of Sample Means**: The *distribution of sample means* will be centered around the population mean μ. This means that, on average, the sample mean will be a good estimate of the population mean.
> - **Standard Error (SE):** The standard deviation of the *distribution of sample means* is called the standard error. It tells us how much the sample means tend to vary from the population mean:
>
> $$SE = \frac{\sigma}{\sqrt{n}}$$
>
> - **When is the CLT valid?** For most situations, the CLT applies well if the sample size is 30 or larger. If the sample size is smaller than 30, the population should be normally distributed. For populations that are strongly skewed or have outliers, larger samples may be necessary for the CLT to hold.

- Check the Shape: Does the "Distribution of Sample Means" look approximately normal?
- Recall from Step 1: The population parameters were $\mu =$ _____ and $\sigma =$ _____.
- Calculate the Standard Error (SE): We call the standard deviation of the distribution of sample means the standard error. It represents how much variability we expect in the sample means if we repeatedly take samples from the population.

  $SE = \frac{\sigma}{\sqrt{n}} =$ _____.

Step 7: Comparing Simulation Results and Theory

Compare the values from your simulation with the theoretical values based on the CLT:

- From Step 3: Mean = _____, SD = _____
- From Step 6: SE = _____

How does simulated standard deviation compare to the theoretical SE?

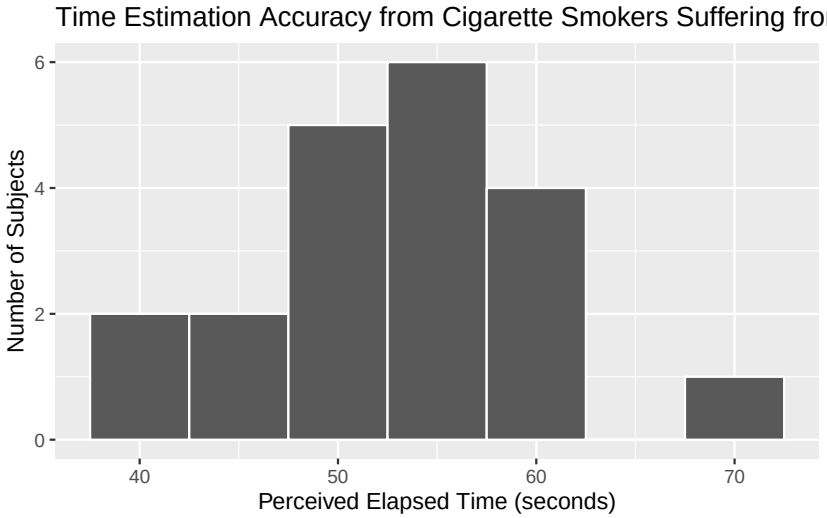
> **i** Key Statistical Concepts
>
> The Central Limit Theorem allows us to estimate the population mean using a single sample, knowing that:
>
> 1. The distribution of sample means will be approximately normal if the sample size is large enough.
> 2. The sample means will be centered around the true population mean, μ.
> 3. The variability of the sample means is determined by the standard error $\frac{\sigma}{\sqrt{n}}$, which decreases as the sample size increases.
>
> You can access a fun applet for exploring Sampling Distributions and the CLT further at https://onlinestatbook.com/stat_sim/sampling_dist/index.html.

In the next section, we will use the CLT to perform hypothesis tests using a procedure called the t-test.

The t-test for a Single Population Mean

Back to **Example 6.1:** Recall that the researchers wanted to show that the mean perceived elapsed time for smokers suffering from nicotine withdrawal was greater than the actual 45 seconds that had elapsed. The data collected in the study were summarized as follows:

min	Q1	median	Q3	max	mean	sd	n	missing
39.11611	48.74039	53.25595	56.65965	69.36787	52.65206	7.268201	20	0



Setup

1. What is the parameter of interest?

2. Set up the null and alternative hypotheses

H_0:

H_A:

Find the t-statistic and the p-value

To determine whether or not the distance between μ (the hypothesized population mean, null value) and (the mean from our observed sample) is larger than what we would expect by random chance, we will use the following statistic:

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

Why use this statistic? Because this quantity measures the position of our observed sample mean on the null model, just like the Z-score discussed in the previous chapter.

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{\text{sample mean - null value}}{\text{standard error}}$$

Note that this is very much like the Z-score, with one minor exception. We don't know the true population standard deviation, σ, so we estimate it with the standard deviation calculated from the 20 observed subjects in the study (this estimate is commonly denoted by s).

This t-statistic comes from what is called a t-distribution. The amount of variability in a t-distribution depends on the sample size n. Therefore, this distribution is indexed by its *degrees of freedom* (df).

> **Degrees of Freedom for a Single Mean**
>
> For inference on a single mean, $df = n - 1$.

To find the p-value associated with this test statistic, we must remember that this is an upper-tailed test (we are trying to find evidence that the mean is greater than 45 seconds). So, the p-value will be the probability we would observe a sample mean (or a t-statistic) greater than that obtained in the actual study by chance alone, assuming the null hypothesis is true:

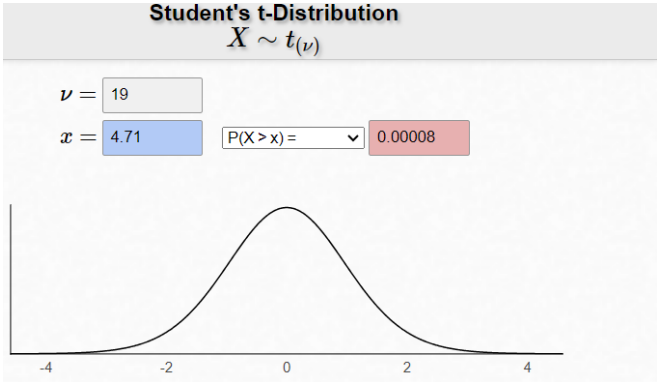


Figure 1: https://homepage.divms.uiowa.edu/~mbognar/applets/t.html

Note that in practice, we can use R to test our research question:

```
t_test(x = nicotine,                                    ①
       response = time_passed,                          ②
       mu = 45,                                         ③
       alternative = "greater",                         ④
       )
```

① Use the `t_test()` function from the `infer` package and designate your `x =` data set.
② Tell the function the `response =` variable of interest,
③ and the `mu =` null value (μ_0).
④ Specify the direction of the `alternative =` hypothesis.

```
# A tibble: 1 x 7
  statistic  t_df   p_value alternative estimate lower_ci upper_ci
      <dbl> <dbl>     <dbl> <chr>          <dbl>    <dbl>    <dbl>
1      4.71    19 0.0000765 greater         52.7     49.8      Inf
```

Conclusion

3. Write a conclusion in context of the problem.

Checking the Normality Assumption:

For the t-test to be valid, at least one of the following conditions must be met:

- Either the sample size is sufficiently large (greater than 30 or so), OR
- The distribution of the observed data is approximately normal (which would indicate that the population is normally distributed so that the Central Limit Theorem would apply even with a small sample size)

For Example 6.1, we have a sample size of 20 subjects, which is not sufficiently large. So we must check whether the data seem to come from a normal distribution.

4. Look back at the histogram of the original data. Does this seem to indicate that this is a reasonable assumption?

Example 6.2: Time Perception for Smokers NOT Suffering from Nicotine Withdrawal

For the data given in Example 6.1, we found evidence that the mean perceived elapsed was in fact greater than the actual 45 seconds that had elapsed. This study alone, however, doesn't really say that the nicotine withdrawal was what impaired one's perception of time. Why not?

Suppose that the researchers also studied 22 subjects who were smokers that did NOT abstain from smoking prior to the data collection (so, they were not suffering from nicotine withdrawal).

Research Question: Is there evidence the mean perceived elapsed time for all smokers **NOT** suffering from nicotine withdrawal is significantly greater than the actual 45 seconds?

```
head(nowithdrawal_data)
```

```
# A tibble: 6 x 4
  status        sex      age time_passed
  <chr>         <chr>  <dbl>       <dbl>
1 no withdrawal female  38.8        52.6
2 no withdrawal female  27.3        41.4
3 no withdrawal male    45.0        51.7
4 no withdrawal female  30.0        48.0
5 no withdrawal male    32.3        52.4
6 no withdrawal female  37.5        38.0
```

Carry out the formal t-test to address this research question.

1. Set up the null and alternative hypotheses

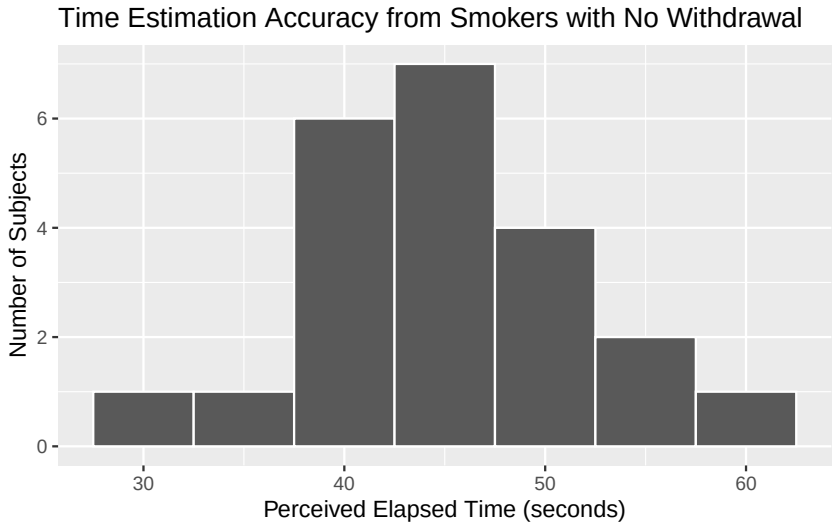
   H_0 :

   H_A :

2. Check normality assumptions for using the CLT.

```
favstats( ~ time_passed, data = nowithdrawal_data)
```

```
     min      Q1   median       Q3      max     mean       sd  n missing
30.86482 41.2379 45.86635 49.45927 61.21396 45.22004 6.826975 22       0
```



3. Find the t-statistic *(practice calculating this out)* and the p-value.

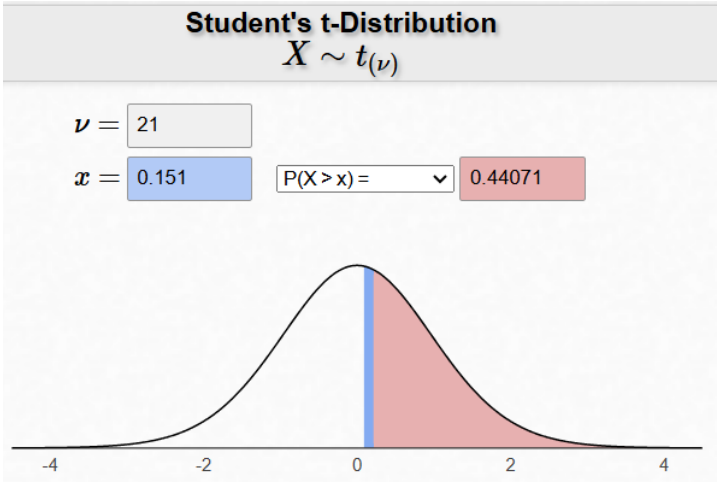


Figure 2: https://homepage.divms.uiowa.edu/~mbognar/applets/t.html

```
t_test(x = nowithdrawal_data,
       response = time_passed,
       mu = 45,
       alternative = "greater"
       )
```

```
# A tibble: 1 x 7
  statistic  t_df p_value alternative estimate lower_ci upper_ci
      <dbl> <dbl>   <dbl> <chr>          <dbl>    <dbl>    <dbl>
1     0.151    21   0.441 greater         45.2     42.7      Inf
```

4. Write a conclusion in the context of the problem.

Confidence Interval for a Single Population Mean

In **Example 6.1**, we found evidence that the mean perceived elapsed time for smokers suffering from nicotine withdrawal differed from the actual 45 seconds of time that had elapsed. Our next question is obvious: HOW MUCH does it differ? To answer this question, we must construct a confidence interval.

Recall our discussion of confidence intervals from earlier in the quarter:

This procedure does NOT require any hypotheses concerning our population parameter of interest (the population mean, in this case). We will use both our sample data (in particular, the observed mean) and what we know about the *distribution of sample means* to obtain a range of likely values for our population mean.

> ⚠ Warning
>
> - A confidence interval allows us to *estimate the population parameter* of interest (recall that the hypothesis test does NOT allow us to do this). Therefore, when available, a confidence interval should always accompany the hypothesis test.
> - The confidence interval does not require any hypothesized value for the population parameter. Instead, we center the confidence interval on the sample mean. Consider the following example.

Example 6.3: Estimated Perceived Time from Nicotine Withdrawal

Our goal is to construct a 95% confidence interval for the mean perceived elapsed time for smokers suffering from nicotine withdrawal (Example 6.1). To do this, we will center our distribution of sample means on the observed sample mean. Then, we will find the lower and upper endpoints that separate the middle 95% of the distribution from the rest (since we are constructing a 95% confidence interval).

Recall the general form of a confidence interval:

$$\text{point estimate} \pm \text{multiplier} \times \text{SE}.$$

Thus, the formula for calculating the endpoints of a confidence interval for a mean is given as follows:

$$\bar{x} \pm \text{t-quantitle} \times \left(\frac{s}{\sqrt{n}}\right)$$

The appropriate t-quantile can be found using R *(typically I will give you a table to choose from)* To find this value, you need the following information:

```
qt(0.975, df = 19)
```

```
[1] 2.093024
```

- confidence level =

- df =

Also, recall our summary statistics for time passed from Example 6.1:

```
favstats( ~ time_passed, data = nicotine)
```

```
     min       Q1   median       Q3      max     mean       sd  n missing
39.11611 48.74039 53.25595 56.65965 69.36787 52.65206 7.268201 20       0
```

1. Use this information to find the endpoints of the confidence interval:

- Lower endpoint $= \bar{x} -$ t-quantitle $\times \left(\frac{s}{\sqrt{n}}\right)$

- Upper endpoint $= \bar{x} +$ t-quantitle $\times \left(\frac{s}{\sqrt{n}}\right)$

Note that we can ask R to provide the endpoints of the 95% confidence interval for this mean when we use the `t_test` function with a `two-sided` test. We can change the confidence level with the `conf_level` argument.

```
t_test(x = nicotine,
       response = time_passed,
       mu = 45,
       alternative = "two.sided",                                    ④
       conf_level = 0.95                                             ⑤
       )
```

④ Change the `alternative =` to a two-sided test (CI are always two-sided),
⑤ and specify the `conf_level =`.

```
# A tibble: 1 x 7
  statistic  t_df  p_value alternative estimate lower_ci upper_ci
      <dbl> <dbl>    <dbl> <chr>          <dbl>    <dbl>    <dbl>
1      4.71    19 0.000153 two.sided       52.7     49.3     56.1
```

2. Interpret the meaning of this interval. What does this interval tell us about the true mean perceived elapsed time for all smokers that are suffering from nicotine withdrawal?

3. Does this interval agree with what you learned from the hypothesis test? Explain.

Key Statistical Concepts

While the main purpose of *confidence intervals* is to provide an *estimated range* of plausible values for μ, the interval should agree with conclusions found from hypothesis testing.

- If μ_0 falls within the confidence interval, then μ_0 is a plausible value for μ. Therefore, we would *Fail to Reject* the null hypothesis.
- If μ_0 falls outside the confidence interval, then μ_0 is not a plausible value for μ. Therefore, we would *Reject* the null hypothesis.

4. How would your calculations change if you wanted to obtain a 90% confidence interval, instead?

```
qt(0.90, df = 19)
```

```
[1] 1.327728
```

```
qt(0.95, df = 19)
```

```
[1] 1.729133
```

```
qt(0.975, df = 19)
```

```
[1] 2.093024
```

```
qt(0.995, df = 19)
```

```
[1] 2.860935
```

Klein, Laura Cousino, Elizabeth J Corwin, and Michele M Stine. 2003. "Smoking Abstinence Impairs Time Estimation Accuracy in Cigarette Smokers." *Psychopharmacology Bulletin* 37 (1): 90–95.