

Chapter 6: Inferential Methods for a Single Numerical Variable

As seen in Chapters 1-4, research hypotheses involving either a single categorical variable or two categorical variables require us to test claims about proportions. When a research question involves a single numerical variable, however, we end up testing claims about the mean, or average. In this chapter, we will consider inferential methods (hypothesis tests and confidence intervals) for the mean of a single numerical variable. Consider the following example.

i 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 deviation

Example 6.1: Time Perception Impaired by Nicotine Withdrawal

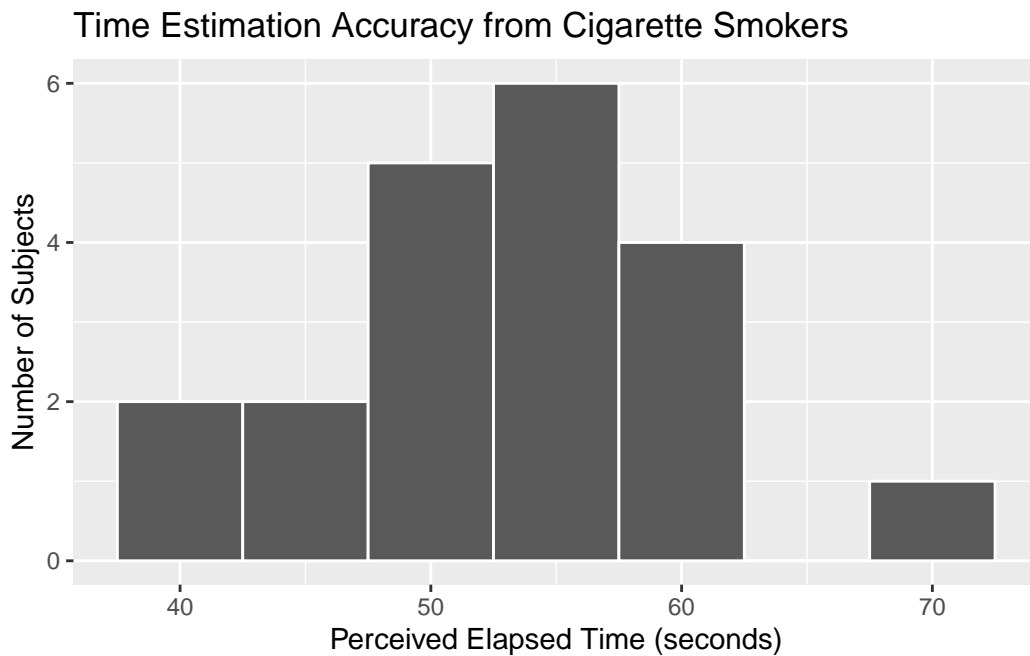
A study conducted by researchers at Pennsylvania State University investigated whether time perception, a simple indication of a person's ability to concentrate, is impaired during nicotine withdrawal. The study results were presented in the paper *Smoking Abstinence Impairs Time Estimation Accuracy 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 summarized and visualized as shown below (these results are artificial but are similar to the actual findings).

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

We can summarize the data with:

```
      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:



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 in the greater population of all smokers subjected to nicotine withdrawal? Explain.

The Distribution of the Sample Mean

The sample mean is a random quantity; that is, it changes from sample to sample. Therefore, the sample mean actually has its own distribution. This distribution will tell us two things:

- What values the sample mean can assume
- How often it will assume these values

This distribution is referred to as the *distribution of the sample mean*. An understanding of this distribution allows us to make decisions about a population mean for a single numerical variable.

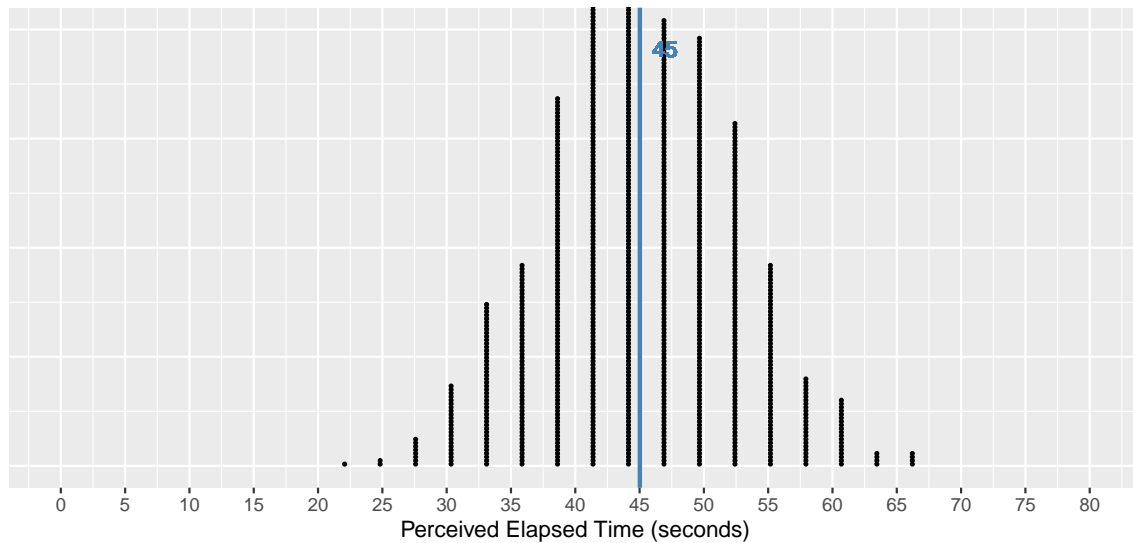
Before we discuss the procedure for inference, let's consider the next example to gain a better understanding of how the distribution of sample means works and how we use this distribution to make a decision concerning our research question.

Exploring the Distribution of the Sampling Mean

Before we discuss the procedure for inference, let's consider the next example to gain a better understanding of how the distribution of sample means works and how we use this distribution to make a decision concerning our research question.

Suppose we set up a hypothetical population of ALL smokers suffering from nicotine withdrawal. This population has been purposefully created so that the mean perceived elapsed time is exactly 45 seconds (i.e., the null is true).

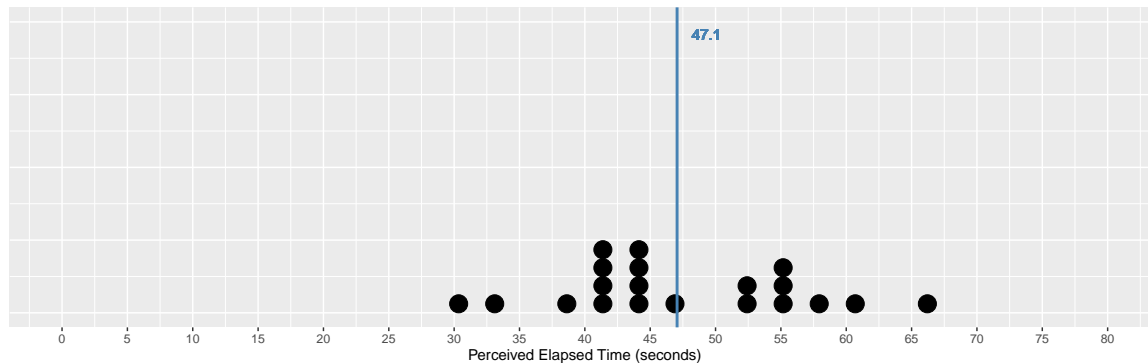
Distribution of 'all' smokers



5. In this simulation study, what is our value of μ , the true population mean?

Note that in reality, the true population mean is usually an unknown quantity which we are trying to estimate. Since it is not feasible to collect data on the entire population of smokers suffering from nicotine withdrawal, the researchers took a random sample of 20 subjects in order to estimate the average perceived elapsed time. Let's see what happens when we take various samples of size 20 from this population.

Sample 1

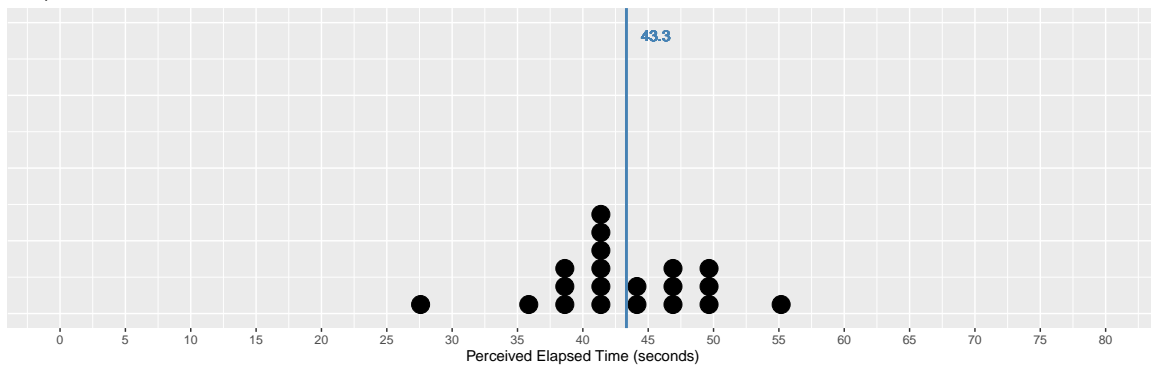


6. What is the average (or mean) perceived elapsed time of the 20 subjects in this “fake” study?

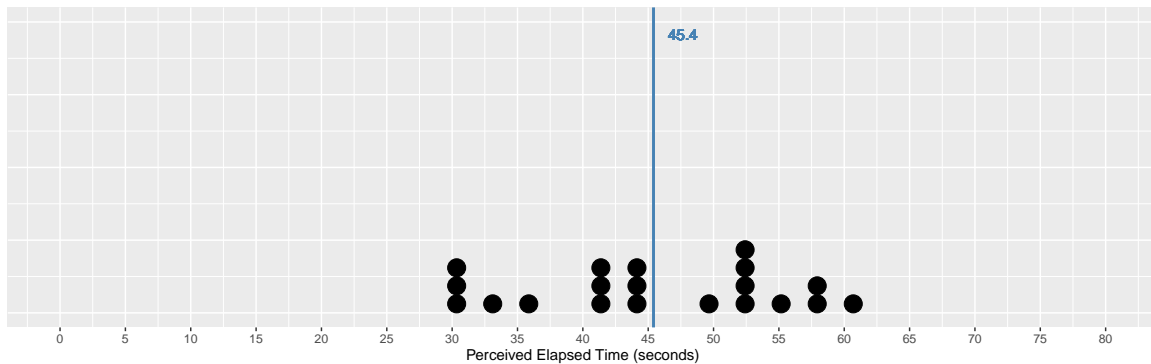
7. Does this necessarily mean that the average perceived elapsed time is greater than 45 seconds for all smokers suffering from nicotine withdrawal? What would you say to a researcher who tried to use only this sample mean to draw this conclusion?

Even though in reality we would carry out a study only once, we will take a sample of 20 subjects from this population over and over again so that we get an idea of how much the sample mean could change from sample to sample. Our second and third random samples of 20 subjects and their sample means are shown below:

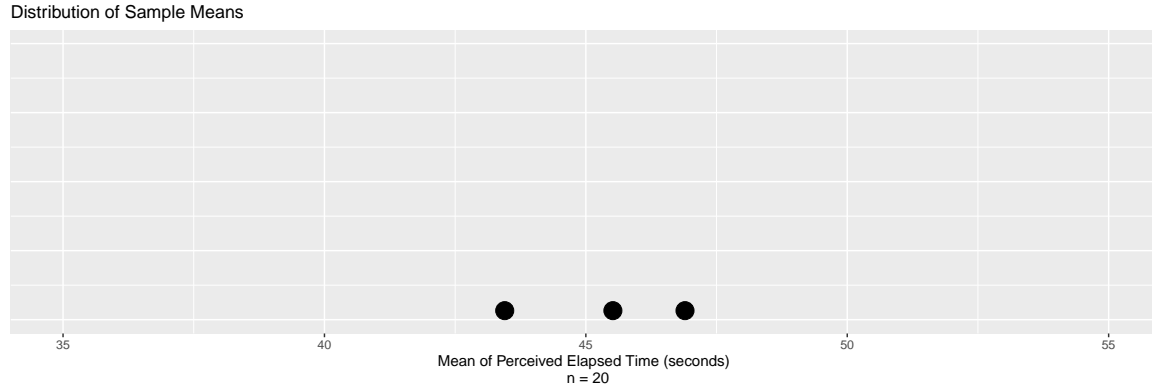
Sample 3



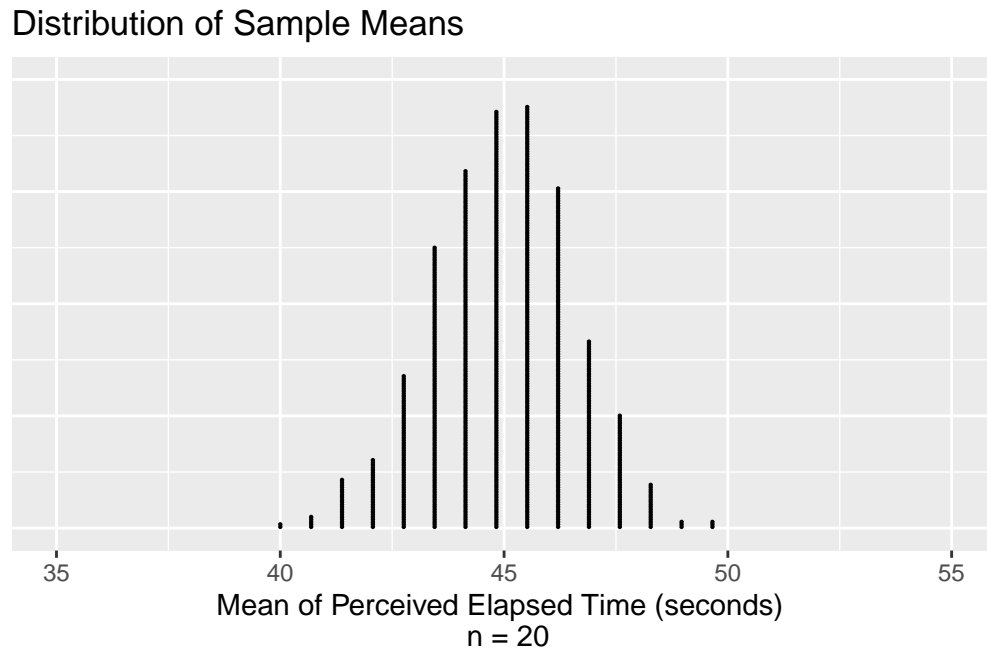
Sample 3



We can start to collect these sample means in a new plot to create the *distribution of sample means*:



To get a really good idea of what values are likely to occur by chance when taking random samples of size 20 from the population, we should take more than just three samples. The graphic below shows the process we used to create the distribution of sample means and the final plot of the means of 1,000 random samples.



This is a distribution of sample means. Recall that these 1000 samples (each consisting of 20 subjects) came from a population with a true mean perceived elapsed time of 45 seconds. So, this distribution gives us a good idea of what sample means from 20 subjects are likely (or unlikely) to occur by chance *if the true mean is 45 seconds*.

Next, note that the researchers wanted to show that the mean perceived elapsed time for smokers suffering from nicotine withdrawal was in fact greater than 45 seconds. The null and alternative hypotheses to address this research question are given as follows.

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.

Note that the distribution of sample means was created *assuming the null hypothesis is true*. To test this hypothesis, we compare our actual observed mean to this null distribution. If the sample mean from the actual research study is not likely to occur by chance according to this null model (i.e., if it is an outlier on this null distribution), then we have evidence against the null model and in support of the research question.

8. Recall that in the actual research study, the mean perceived elapsed time for the 20 subjects studied was 52.65 seconds. Sketch this observed value on the null distribution above.

There are two explanations for this observed mean of 52.65 seconds: either (1) the true mean perceived elapsed time for smokers suffering from nicotine withdrawal really is greater than 45 seconds, or (2) their time perception is not impaired (i.e., the true mean is actually 45 seconds) and the sample mean was greater than 45 seconds simply because of random chance.

9. Was a sample mean of 52.65 seconds likely to occur by chance if the true population mean is actually 45 seconds? What does this imply about the research question?

10. Give an approximate p-value based on this simulation study.

The Central Limit Theorem (CLT)

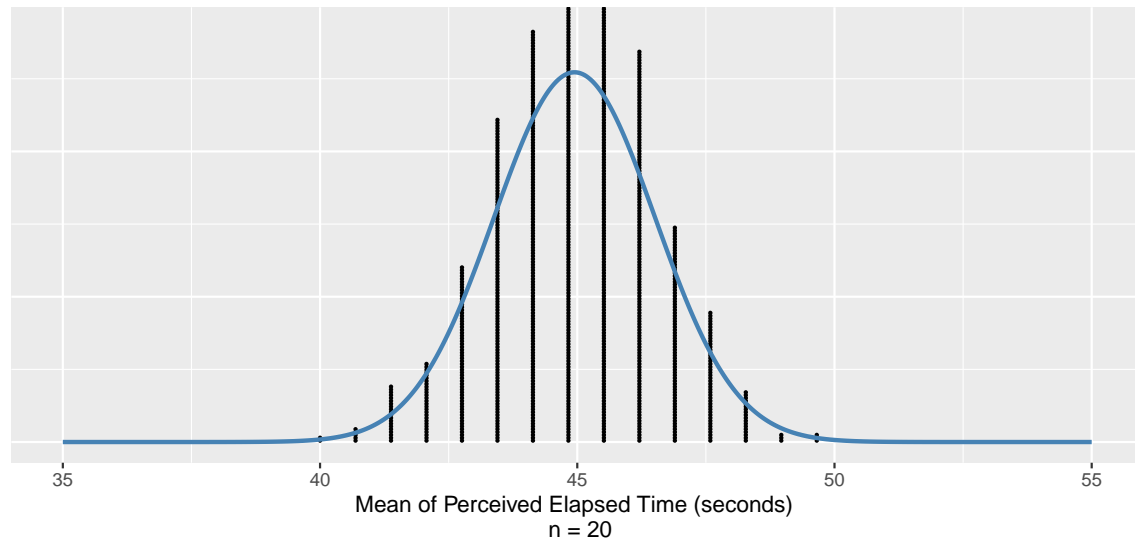
Note that a statistician would not necessarily carry out a simulation study such as this to answer a research question. Instead, one could use a “short-cut” known as a t-test to investigate a research question concerning a single population mean. This short-cut is a result of something known as the **central limit theorem**, which states the following:

Consider a random sample of n observations from ANY population with mean μ and standard deviation σ . When n (the number of subjects in the sample) is sufficiently large, the distribution of sample means will be approximately a NORMAL distribution with a mean of μ and a standard deviation of $\frac{\sigma}{\sqrt{n}}$. This approximation gets better as the sample size (n) increases.

We can see the Central Limit Theorem applied to the sample means calculated from our hypothetical population from Example 6.2 as follows:

- We see that the distribution of sample means is approximately normal:

Distribution of Sample Means



- We see that this normal distribution is centered at *about* the true population mean, $\mu = 45$.

min	Q1	median	Q3	max	mean	sd	n	missing
39.69695	43.91675	44.97946	46.05596	49.90892	44.94626	1.567717	1000	0

- The standard deviation of all observations in our hypothetical population is $\sigma = 7.17$ seconds. According to the central limit theorem, then, the standard deviation of the distribution of sample means is given by

$$\frac{\sigma}{\sqrt{n}} = \frac{7.17}{\sqrt{20}} = 1.60$$

Finally, recall from the previous chapter that given the mean and standard deviation of a distribution, we can determine whether a given observation is an outlier or not based on its position on this distribution. This will help us decide whether the sample mean *actually observed* in the research study is an outlier on the distribution that assumes the null hypothesis is true.

The only question that remains is this: How large does n (the number of subjects in a study) have to be in order for us to use the short-cut provided by the Central Limit Theorem?

i Sample Size for the Central Limit Theorem

For almost all populations, samples of size $n \geq 30$ subjects will be sufficient to say that the distribution of sample means is approximately normal. However, if the distribution is very skewed, the sample size may have to be much larger than 30 in order for the central limit theorem to apply.

In summary, we can use the Central Limit Theorem to help us create a procedure for comparing a population mean to some hypothesized null value.

This works because:

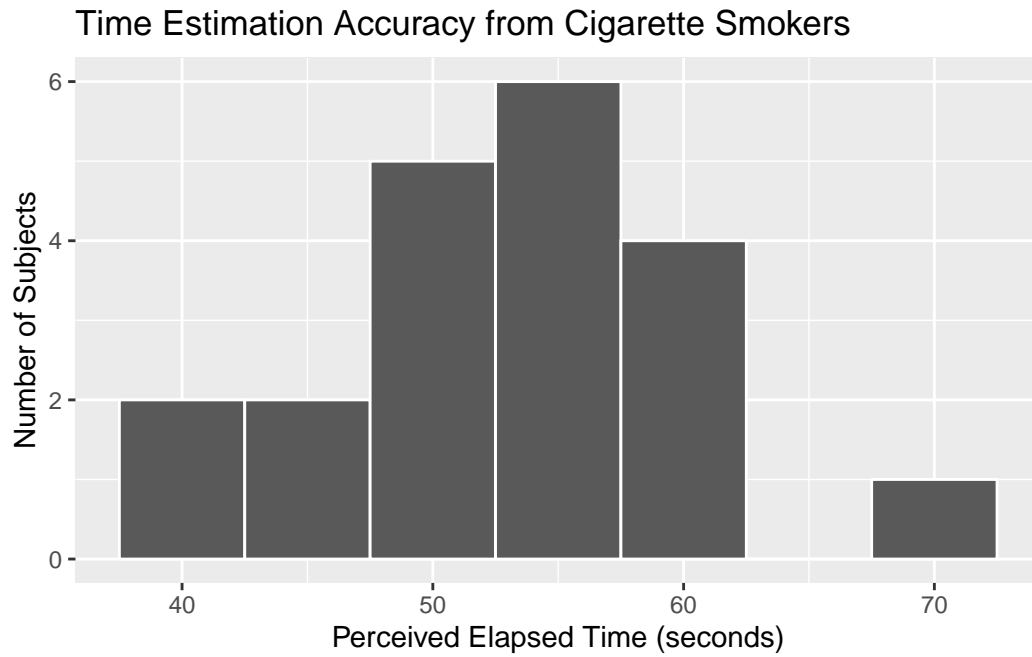
- We know the distribution of sample means will be approximately normal if either (i) the original population is normally distributed, or (ii) our sample size is sufficiently large.
- We know the distribution of sample means will be centered at the true population mean (which we can set to some hypothesized value in the null hypothesis).
- We know that the variability in the distribution of sample means is given by (i.e., the variability decreases as the sample size gets larger, which we can see in the above examples).

In the next section, we put all of the pieces together to create what is known as 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.

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{\text{observed sample mean} - \text{mean of the distribution of sample means}}{\text{standard deviation of the distribution of sample means}}$$

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 (the greater the sample size, the smaller the standard deviation of the distribution of sample means). Therefore, this distribution is indexed by its *degrees of freedom* (df).

i 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:

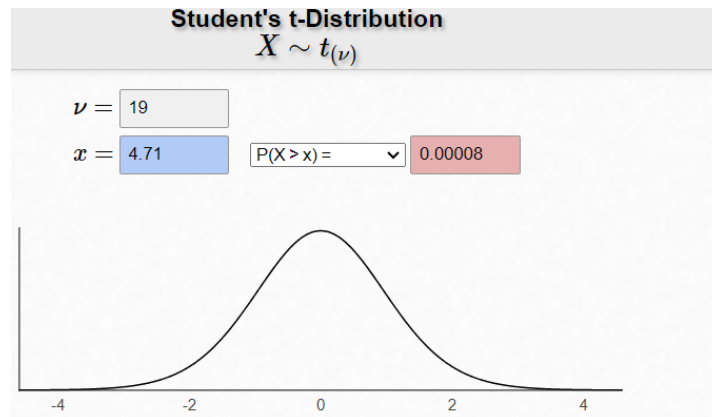


Figure 1: <https://homepage.divms.uiowa.edu/~mbognar/applets/t.html>

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

```
library(infer)
t_test(x = nicotine,
       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      4.71   19 0.0000765 greater         52.7    49.8    Inf
```

Conclusion

3. Write a conclusion in context of the problem.

i 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 Non-Smokers

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?

```
# A tibble: 6 x 4
  status    sex    age time_passed
  <chr>    <chr> <dbl>      <dbl>
1 non-smoker female 38.8        52.6
2 non-smoker female 27.3        41.4
3 non-smoker male   45.0        51.7
4 non-smoker female 30.0        48.0
5 non-smoker male   32.3        52.4
6 non-smoker female 37.5        38.0
```

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

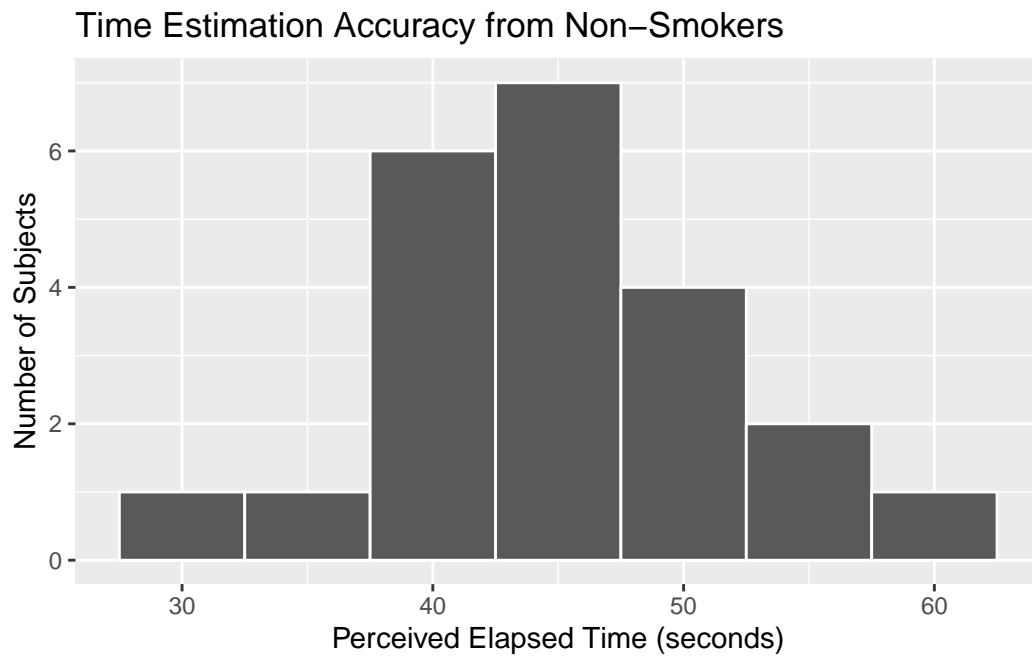
1. Set up the null and alternative hypotheses

$$H_0 :$$

$$H_A :$$

2. Check normality assumptions.

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.

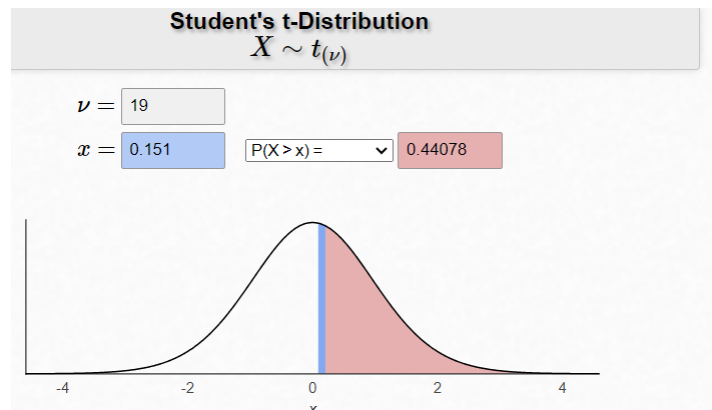


Figure 2: <https://homepage.divms.uiowa.edu/~mbognar/applets/t.html>

```
library(infer)
t_test(x = nonsmokers,
       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 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 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).

The formula for calculating the endpoints of this confidence interval is given as follows:

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

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

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

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",
       conf_int = TRUE,
       conf_level = 0.95)
```

```
# 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.

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.