LESSON: POPULATION MEAN INFERENCE FOR NORMALLY DISTRIBUTED DATA AND UNKNOWN POPULATION STANDARD DEVIATION

This lesson includes an overview of the subject, instructor notes, and example exercises using Minitab.

Population Mean Inference for Normally Distributed Data and Unknown Population Standard Deviation

Lesson Overview

Given normally distributed data and unknown population standard deviation, we will construct confidence intervals and conduct hypothesis tests on the true mean μ of a population.

Prerequisites

This lesson requires knowledge from the ***Normal Distribution***, ***Sampling Distribution of*** $\bar{X}$ (e.g. Central Limit Theorem), ***Population Mean Confidence Intervals for Large Samples***, and ***Population Mean Hypothesis Testing for Large Samples*** lessons. The distribution of $\bar{X}$ is either exactly or approximately normally distributed for large sample sizes n, in which n > 30. For small sample sizes, the Central Limit Theorem does not provide us with information about the distribution of $\bar{X}$. The ***Population Mean Confidence Intervals for Large Samples*** lesson teaches how to construct a basic one sample mean confidence interval and interpret it. The ***Population Mean Hypothesis Testing for Large Samples*** lesson displays how to set up a one or two-sided hypothesis test and conduct and interpret a basic one sample mean hypothesis test. To construct confidence intervals or perform hypothesis testing in this lesson, wel need to use Minitab to determine probabilities and critical values associated with a t distribution.

Learning Targets

This lesson teaches students to:

- Recognize a *t* distribution and its relationship to the normal distribution.
- Determine probabilities associated with a *t* distribution using Minitab.
- Select the correct distribution, *t* or normal, when performing inference on one population mean.
- Check sample data to determine if it is from a normal distribution.
- Use Minitab to construct a confidence interval and conduct a hypothesis test for one population mean given a normal distribution and an unknown population standard deviation.

Time Required

It will take the instructor at least 60 minutes in class to do this lesson since there are four main concepts: the introduction of the *t* distribution, a normality test for sample data, selection of the correct distribution, normal or *t*, for inference on one population mean, and the construction of a confidence interval and hypothesis test using the *t* distribution. Students need a variety of practice examples to further cement details about confidence interval construction and hypothesis testing that was first introduced in the ***Population Mean … Large Samples*** lessons. The new learning objective in this section is differentiating between when to use a normal or *t* distribution for inference on one population mean. The exercises on the activity sheet will also take at least 60 minutes. Some of the examples and exercises from the ***Population Mean … Large Samples*** lessons will be reused in this lesson to show the similar procedures for hypothesis testing and confidence interval construction with a *t* distribution compared to a normal distribution.

Materials Required

- Minitab desktop (20 or higher) or Minitab web app
- Minitab worksheet of sample data, entitled ***InferenceOnMean_tDistribution_Lesson.mtw***

Assessment

The activity sheet contains exercises for students to assess their understanding of the learning targets for this lesson.

Possible Extensions

The instructor may want to do the lesson ***Two Sample t Test for Population Means*** or ***Paired t Test for Population Means*** after this lesson so that students extend their knowledge of inference to two samples of data.

Instructor Notes with Examples

The Distribution $\frac{\bar{X}-\mu}{s/\sqrt{n}}$

- In practice, the **population standard deviation σ is seldom known**, especially if the population mean μ is unknown. It can be estimated by the sample standard deviation *s*. As *n* increases, *s* becomes a better approximation of σ.
- What effect does **dividing by *s* instead of σ** have on the distribution $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$?
- Historically, many early statisticians felt that **replacing σ by *s*** had no effect on the shape of the distribution $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ and it remained normal. Sometimes they were right – but only when *n* was large!
- When ***n* is small, replacing σ by *s* matters** and changes the way we construct a confidence interval for μ or perform hypothesis testing on μ.
- Deriving the exact shape and **probability density function *f*(*x*) for $\frac{\bar{X}-\mu}{s/\sqrt{n}}$** was one of the major statistical accomplishments of the early 20th century.

The History Behind the Student *t* distribution

- Credit for recognizing that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ and $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ have different distribution shapes and probability density functions goes to **William Sealy Gossett**.
- After graduating in 1899 from Oxford with a First Class degree in Chemistry, Gossett took a position at **Arthur Guinness, Son & Co., Ltd.,** a firm that brewed (and still does!) **a thick dark ale known as stout.** His employer forbade him to use his real name, so he published papers under the name "Student."
- Given the task of making the art of brewing more scientific, Gossett quickly realized that any experimental studies that might be undertaken would face two obstacles:
  - **Sample sizes would be small**, often *n* = 4 or 5.
  - There would **never be any way to know the value of the true population standard deviation σ** associated with any given set of measurements.
- When Gossett needed to draw inferences about μ, he found himself working with $\frac{\bar{X}-\mu}{s/\sqrt{n}}$. The distribution of $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ seemed to have the same general bell shape appearance of $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, but its **tails were "thicker,"** just like Guinness beer in my opinion!

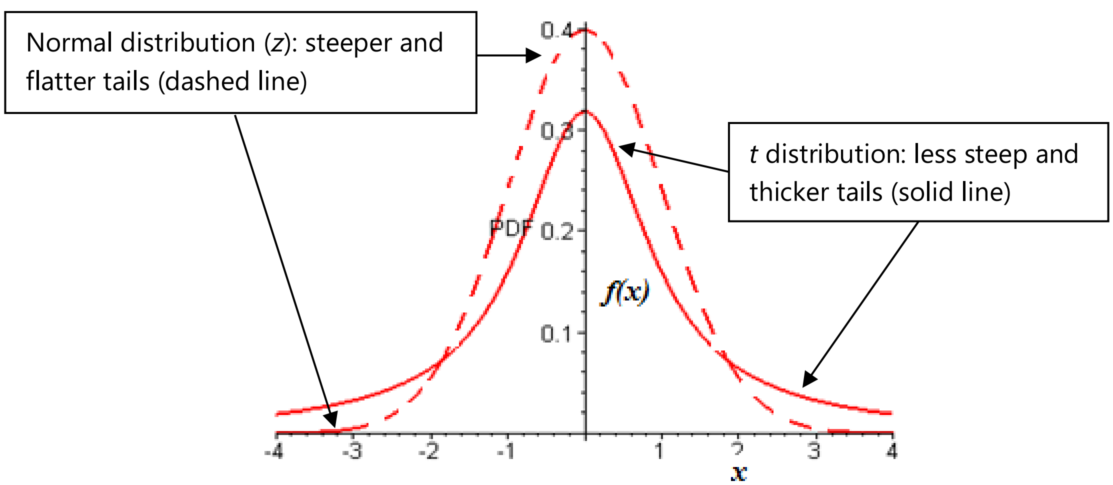
- Gossett called the quotient $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ a *t*-ratio and derived a formula for its probability density function:

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\cdot\Gamma\left(\frac{k}{2}\right)}\cdot\frac{1}{\left((x^2/k)+1\right)^{\frac{k+1}{2}}} \quad \text{for} \quad -\infty < x < \infty.$$

where $k = n - 1$, where n is the sample size, and $\Gamma(*)$ is the gamma function, defined as:

$$\Gamma(z+1) = \int_0^\infty x^{z} e^{-x}\, dx$$

- The function $f(x)$ is a valid probability density function for sample sizes n = 2, 3, 4, …
- For large n, z table values and t table values are "close."



Note: The *t* distribution has "heavier" tails than the normal distribution; i.e. there is more area under the curve in the tails of a *t* distribution than a normal distribution.

The random variable $t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ has a *t* distribution with $n - 1$ **degrees of freedom**, where n is the sample size.

When should the *t* distribution be used versus the normal distribution in hypothesis testing and building confidence intervals?

- The data comes from a normal distribution, AND
- The population standard deviation σ is unknown

Confidence Intervals

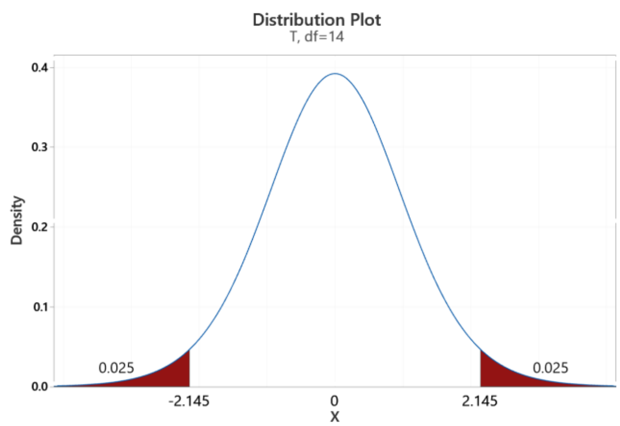
Recall from the ***Population Mean Confidence Intervals for Large Samples*** that we used the following notation with the standard normal distribution *Z* to represent **critical values.**

- We let $z_{\alpha/2}$ denote a *z*-score with α/2 probability to its right and $-z_{\alpha/2}$ denote a *z*-score with α/2 probability to its left.

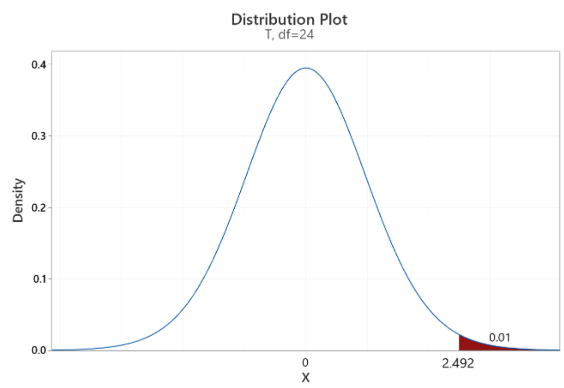
Similarly for the *t* distribution, the **critical value** $\boldsymbol{t_{\alpha/2,df}}$ is the value such that the area to the ***right of it*** is **α/2**, where *df* represents degrees of freedom or *n* – 1. Also, $\boldsymbol{-t_{\alpha/2,df}}$ is the critical value such that the area to the ***left* of it** is **α/2**.

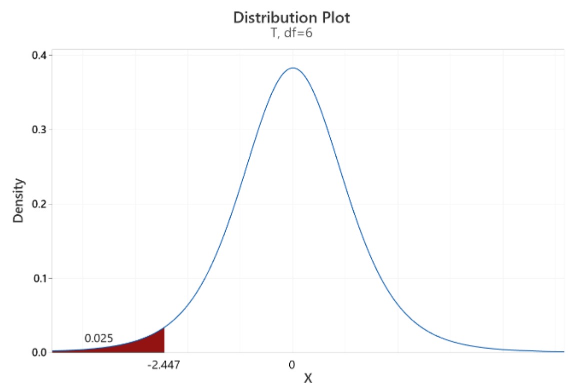
Examples:

- The value $\boldsymbol{t_{0.025,14}}$ is the value on a *t* distribution with *n* = 15 or 14 degrees of freedom (df) such 0.025 probability is to its right. The value of $\boldsymbol{t_{0.025,14}}$ is **2.145** as displayed in the probability plot below.
- The value $\boldsymbol{-t_{0.025,14}}$ is the value on a *t* distribution with 14 df such 0.025 probability is to its left; it's value is **-2.145** as displayed in the probability plot below.



- The value $t_{0.01,24}$ is the value on a *t* distribution with 24 df that has 0.01 probability to its right; it's value is: $t_{0.01,24} \approx 2.492$.
- The value $-t_{0.025,6}$ is the value on a *t* distribution with 6 df that has 0.025 probability to its left; it's value is: $-t_{0.025,6} \approx -2.447$.





Determining Critical Values for a *t* Distribution

In the ***Normal Distribution*** lesson, we used a *z* table, as well as Minitab, to determine *z* critical values. Because the *t* distribution is dependent on the sample size of the data, *t* tables are more difficult to interpret than a standard normal table. In this lesson, we'll use only Minitab to determine critical values for a *t* distribution.

In the section called "Inverse Normal Calculations" in the ***Normal Distribution*** lesson, we discussed how to use Minitab's **Probability Distribution Plot** function to determine *z* values corresponding to a given proportion α in the right, left, or both tails of a normal distribution. The instructions below mimic that procedure but for a *t* distribution.

To determine the value *t* that corresponds to the proportion α = 0.01 in the right tail of a *t* distribution with *df* = 24 degrees of freedom, we do the following in Minitab:

Minitab desktop (20 or higher)

1. Choose **Graph > Probability Distribution Plot**.
2. Choose **View Probability**, then click **OK**.
3. From **Distribution**, choose **t**. Type the *df* value *24* in the **Degrees of Freedom** textbox.
4. Click the **Shaded Area** tab. Under **Define Shaded Area By**, choose **Probability**.
5. Click **Right Tail**, since we want the *t* value corresponding to α = *0.01* in right tail. In **Probability**, type the probability value *0.01*.
6. Click **OK**.

Minitab web app

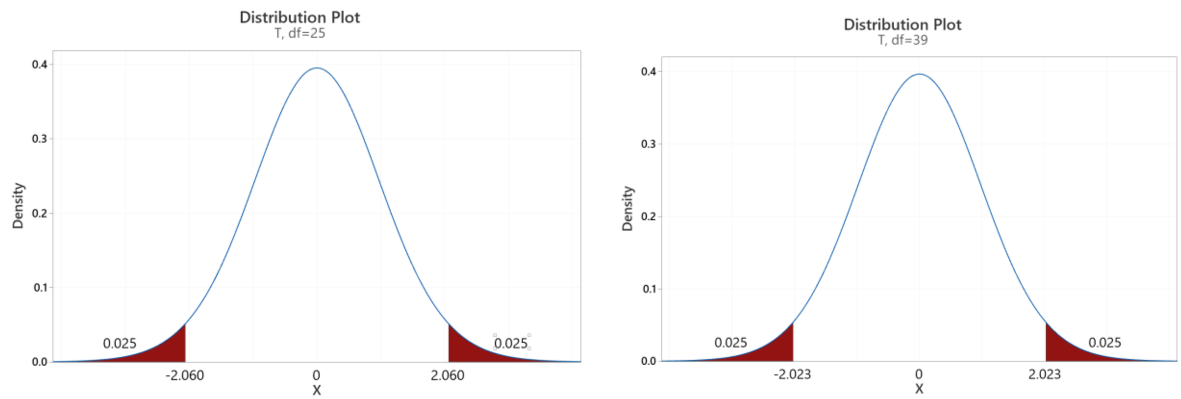
1. Choose **Graph > Probability Distribution Plot**.
2. Choose **View Probability**.
3. From **Distribution**, choose **t**. Type the *df* value *24* in the **Degrees of Freedom** textbox.
4. Click **Options**. Under **Shade the area corresponding to the following**, select **A specified probability**.

5 Click **Right Tail**, since we want the *t* value corresponding to α = *0.01* in right tail. In **Probability**, type the probability value *0.01*.
6 Click **OK** in each dialog box.

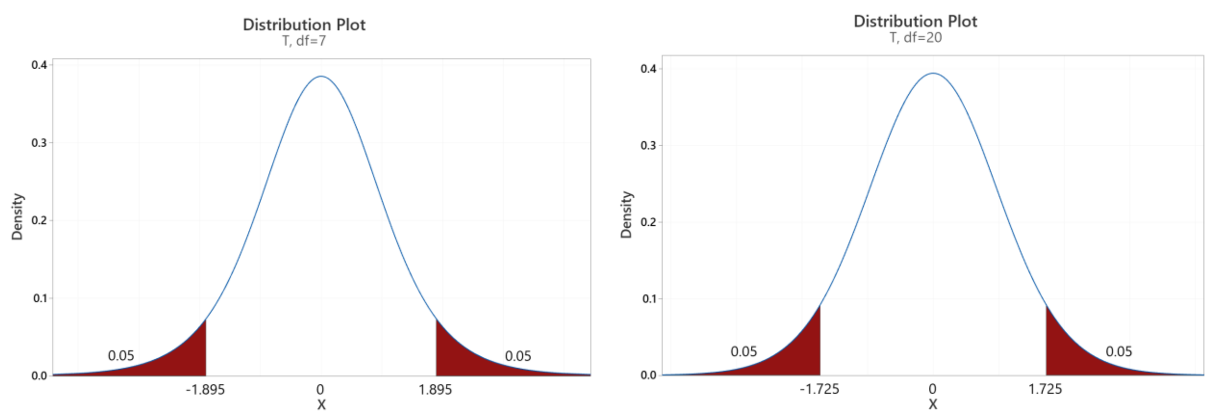
As displayed in the *t* distribution plot in the third example in the previous section, the *t* value that is returned is $t_{0.01,24} \approx 2.492$.

Additional Examples

Example 1. Determine 95% *t* critical values for a sample size of $n = 26$ and a sample size of $n = 40$. Notice that as *n* gets larger, the *t* critical value is smaller due to the more "flattened" tails of the *t* distribution for larger *n*. Recall also that $\pm z_{0.025} \approx \pm 1.96$.



Example 2. Determine 90% *t* critical values for a sample size of $n = 8$ and a sample size of $n = 21$. Recall that $\pm z_{0.025} \approx \pm 1.645$. The *t* critical value for a sample size of $n = 50$ is $\pm t_{0.025,49} \approx \pm 1.667$.



Confidence Interval for μ Given Normally Distributed Data and Unknown σ

We are now ready for the formula for a confidence interval for the population mean μ given normally distributed data and unknown population standard deviation. Since σ is unknown, we'll use the sample standard deviation *s* as an estimate of it. By the Central Limit Theorem, the distribution of the sample mean $\bar{X}$ is normally distributed with mean μ and standard deviation $\frac{s}{\sqrt{n}}$. As discussed at the beginning of this lesson, the random variable $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ has a *t* distribution with *n* – 1 degrees of freedom.

The formula for a two-tailed 100(1-α)% confidence interval for μ given normally distributed data and *unknown* population standard deviation is *very* similar to the formula for a two-tailed 100(1-α)% confidence interval for μ given a large sample size and *known* population standard deviation as shown in the ***Population Mean Confidence Intervals for Large Samples*** lesson.

> *A* ***two-tailed 100(1-α)% confidence interval for μ*** when the data is normally distributed and σ is unknown is:
>
> $$\left[\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right]$$

Example 3. When I work from home at night, my dog likes to sit on the couch with me. I've been trying to determine the true mean amount of time μ that it takes her to fall asleep after positioning herself snuggly beside me. I've taken a random sample of her times to fall asleep on *n* = 7 different occasions. The data is from a normally distributed population and the sample mean and sample standard deviation for these 7 times are $\bar{x}$ = 10.2 minutes and *s* = 2.1 minutes, respectively. Determine a 99% confidence interval for the true mean amount of time for her to fall asleep.

Solution: We are told that the sleep times come from a normal distribution. We are provided with the sample mean $\bar{x}$ and the sample standard deviation *s* for sample size *n* = 7. Since we are building a two-sided 99% confidence interval for the population mean, we need to determine the critical values $\pm t_{0.005,6}$. Using the two-tailed confidence interval for μ provided before Example 3, we have:

$$\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \cong 10.2 \pm 3.707 \cdot \frac{2.1}{\sqrt{7}} \rightarrow \mathbf{[7.258, 13.142]\ minutes.}$$

We can also compute this confidence interval directly using a Minitab function.

To compute this interval using Minitab:

1 Choose **Stat > Basic Statistics > 1-Sample t**.
2 From the drop-down list, select **Summarized data**.
3 In **Sample size**, enter *7*.
4 In **Sample mean**, enter *10.2*.
5 In **Standard deviation**, enter *2.1*.
6 Click **Options**. In **Confidence level**, enter *99*.
7 Click **OK**.

Minitab yields the following 99% confidence interval: [7.257, 13.143].

Example 4. Getting a wisdom tooth pulled can be quite a painful ordeal. In a random sample of n = 120 surgeries to remove a wisdom tooth at a dentist's office, the mean surgery time was $\bar{x}$ = 146.3 minutes with a standard deviation of s = 22.6 minutes. Determine a 95% confidence interval for the true mean amount of time for a wisdom tooth to be pulled by the dentist.

Solution: Unlike Example 3, we are not told that the time to remove a wisdom tooth is normally distributed. According to the Central Limit Theorem, we can assume that the distribution of the sample means, $\bar{X}$, is approximately normally distributed since the sample size is large: n = 120. Since we know the standard deviation of the sample and not the population, then we use Minitab's 1-Sample t function to determine the 95% confidence interval for μ. It is:

Descriptive Statistics

N	Mean	StDev	SE Mean	95% CI for μ
120	146.30	22.60	2.06	**(142.21, 150.39)**

μ: population mean of Sample

Example 5. A particular brand of diet margarine was analyzed to determine the true mean percentage of polyunsaturated fatty acid μ per stick. A random sample of ten sticks of this type of margarine resulted in the following data:

16.81 17.22 17.40 16.95 16.54 17.16 16.75 18.20 16.25 17.12

Determine a 95% confidence interval for the true mean level of polyunsaturated fatty acid per stick.

Solution: Unlike Example 3, we are not told that the time to remove a wisdom tooth is normally distributed, and unlike Example 4, we don't have a large sample size that would allow us to assume that the distribution of the sample means is normally distributed. Although we have a small sample size (n < 30), it is still possible that the sample data is from normal distribution even if we aren't told that it is. We need some way to check whether it is from a normal distribution. To do so, we need to introduce a **normality test**.

We'll return to this example if we can conclude that the sample data is from a normal distribution. If it is, then by the Central Limit Theorem, the distribution of the sample means is normally distributed. If it is not, then we need another statistical tool to create a confidence interval for the data. We won't discuss that tool in this lesson.

Normality Test

The **normality test** and **normal probability plot** are usually the best tools for judging normality of a data set. The null and alternative hypotheses for a normality test are:

H_0: The data follows a normal distribution
H_a: The data does not follow a normal distribution

The test results indicate whether we should reject or fail to reject the null hypothesis that the data is from a normally distributed population. Minitab has several different normality tests with the default test being the Anderson-Darling normality test. The Anderson-Darling (AD) test statistic measures how well the data follows a normal distribution. For a specified data set and distribution, the better the distribution fits the data, the smaller the AD test statistic and the larger the *p*-value will be. If the *p*-value is less than α, say $\alpha = 0.05$, then we reject the null hypothesis that the data comes from a normal distribution. Otherwise, if the *p*-value is greater than α, then we fail to reject the null hypothesis and conclude that we do not have enough evidence to refute normality.

For Example 5, we are going to use the Anderson-Darling normality test to determine whether the data meets the assumption of normality to continue moving forward with a 1-Sample *t* test.

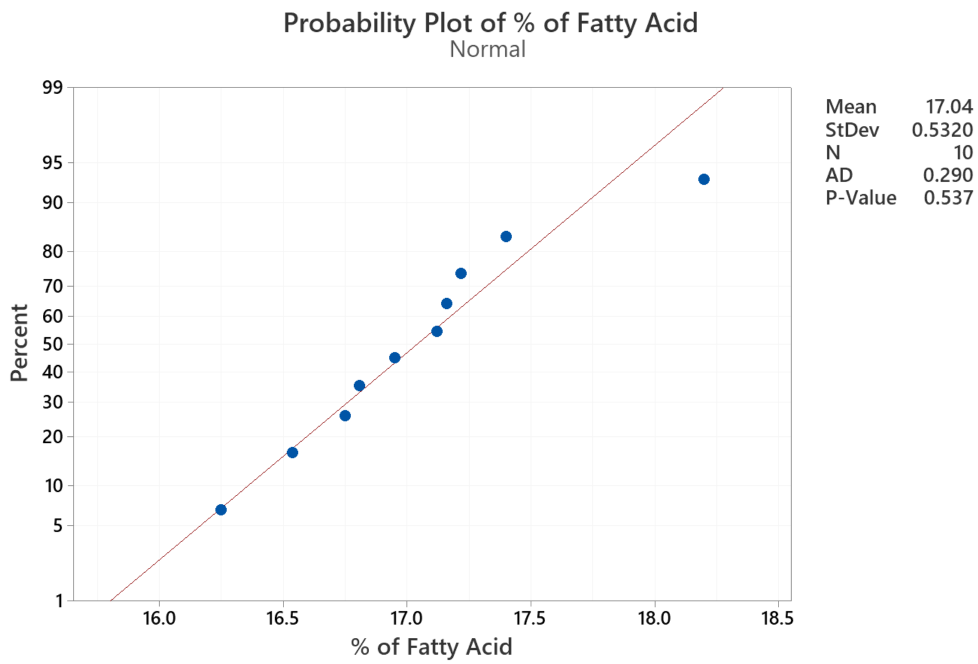
Example 5 (continued). Does the random sample of the percentage of polyunsaturated fatty acid in these ten sticks of margarine follow a normal distribution?

16.81	17.22	17.40	16.95	16.54	17.16	16.75	18.20	16.25	17.12

Solution (continued). Since the sample size is small, we'll check if the data follows a normal distribution by using the Anderson-Darling normality test. The Minitab instructions for producing an AD test statistic and probability plot are:

1. Choose **Stat > Basic Statistics > Normality Test**.
2. In the **Variable** text box, select the column containing the sample data of interest.
3. Under **Test for Normality**, the **Anderson-Darling** test is selected by default.
4. Click **OK**.

For the "Percentage of Fatty Acid" example, Minitab returns the following graphic along with the AD test statistic and its p-value.



Given that the resulting p-value is ~0.537, we would not reject that the "Percentage of Fatty Acid" sample data comes from a normal distribution at level of significance α = 0.05. Knowing this information allows us to finish constructing a 95% confidence interval for the true mean level of polyunsaturated fatty acid per stick of margarine since we can assume that $\bar{X}$ is normally distributed. Using Minitab, the 95% 1-Sample t confidence interval is:

Descriptive Statistics

N	Mean	StDev	SE Mean	95% CI for μ
10	17.040	0.532	0.168	**(16.659, 17.421)**

μ: population mean of % of Fatty Acid

Hypothesis Testing

Just as constructing a 1-sample t confidence interval is like constructing a 1-sample z confidence interval, the same holds for conducting a hypothesis test with a t distribution compared to a normal distribution. The decision to use a normal or a t distribution is based on the information that we are given in the problem situation.

If we have a:

- Large sample size n and known population standard deviation σ (rare), then standardizing the test statistic $\bar{x}$ will result in z-score from a normal distribution. That is:

$$z_0 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- Large sample size *n* and unknown population standard deviation (highly likely), then standardizing the test statistic $\bar{x}$ will result in *t*-score from a *t* distribution. Since we don't know the value of σ, we use the sample standard deviation *s* as an estimate. That is:

$$t_0 = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

- Small sample size *n*, unknown population standard deviation, and sample data from a normal distribution, then standardizing the test statistic $\bar{x}$ will result in *t*-score from a *t* distribution. We again use *s* as an estimate of σ. The standardized test statistic t_0 has the same formula as shown above.

- Small sample size *n*, unknown population standard deviation, and sample data from a non-normal distribution, then we need another method for performing a hypothesis test for one population mean μ. This method is not discussed in this lesson. Please seek help from the nearest statistician or online Minitab support.

Hypothesis Testing for μ Given Normally Distributed Data and Unknown σ

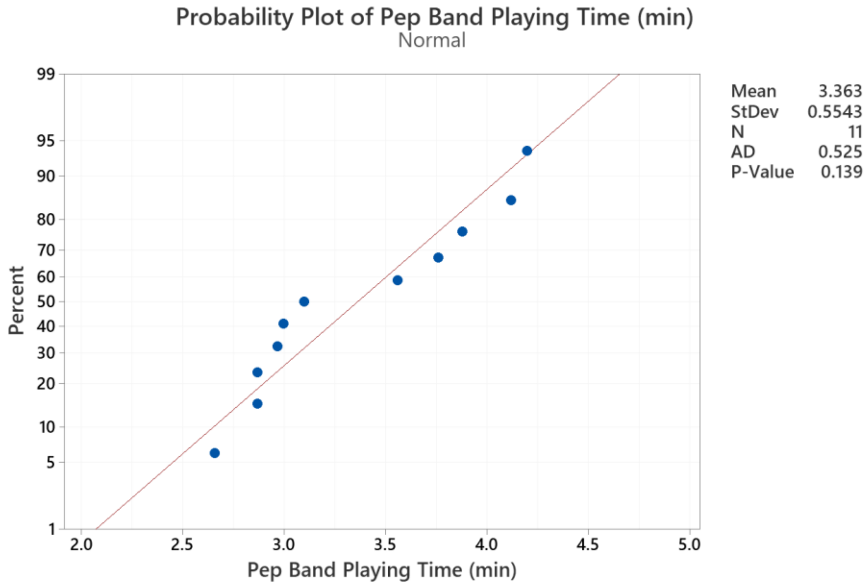
Example 6. During NCAA Division III men's basketball games, the pep band plays during time-outs. We randomly collected data on pep-band playing times from *n* = 11 recent games this year. Is the true mean amount of playing time μ during these time-outs more than 3 minutes? Here is the data for the randomly selected playing times:

3.10 4.20 3.00 2.97 2.66 3.56 3.76 3.88 2.87 4.12 2.87

Solution: We are being asked to perform the following hypothesis test:

H_0: μ = 3 versus H_a: μ > 3.

Since *n* is small, we cannot use the Central Limit Theorem to claim that $\bar{X}$ is normally distributed. First, we need to use the AD normality test to determine if the *n* = 11 data points are from a normal distribution. Below is the output from performing the normality test in Minitab:

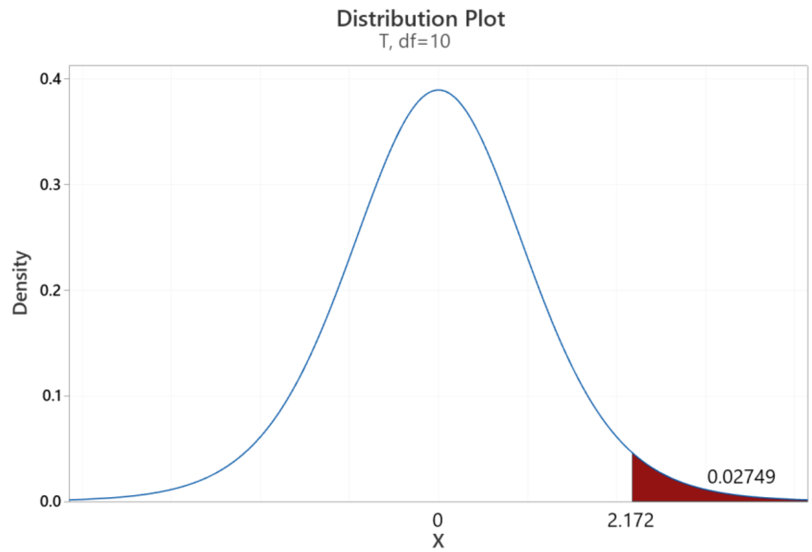


At significance level α = 0.05, we would fail to reject that the data comes from a normal distribution since the *p*-value is greater than α. Since we are assuming the sample data is from a normal distribution, then we can further assume that $\bar{X}$ is normally distributed.

Given that $\bar{X}$ is normally distributed and the population standard deviation is unknown, then we'll use the sample standard deviation $s \approx 0.5543$ as an approximation of σ. The correct standardized test statistic is from the *t* distribution with 10 degrees of freedom. Its value is:

$$t_0 = \frac{3.363 - 3}{\frac{0.5543}{\sqrt{11}}} \cong 2.172.$$

The *p*-value associated with t_0 can be obtained by using **Graph > Probability Distribution Plot** and selecting a *t* distribution with *df* = 10. It is ~0.02749.



Given this small *p*-value, we would reject H_0: μ = 3 in favor of H_a: μ > 3 at significance level α = 0.05. If H_0: μ = 3 was true, the probability of obtaining a test statistic value of 3.363 is 0.02749, which is unlikely.

We could have performed this hypothesis test by using Minitab's 1-Sample t test as follows:

1 Choose **Stat > Basic Statistics > 1-Sample t**.
2 In **One or more samples, each in a column**, enter *Pep Band Playing Time (min)*.
3 Select **Perform hypothesis test**. In **Hypothesized mean**, enter 3.
4 Select **Options**.
5 In **Alternative hypothesis**, choose **Mean > hypothesized mean.**
6 Click **OK** in each dialog box.

Minitab returns the *t*-score and *p*-value as previously determined by hand. In addition, the Minitab output allows us to see the alternative hypothesis.

Test

Null hypothesis	H_0: μ = 3
Alternative hypothesis	H_1: μ > 3

T-Value	P-Value
2.17	0.028

Example 7. In a random sample of n = 99 different textbooks from an institution's bookstore, their mean weight is $\bar{x}$ = 3.2 pounds with a standard deviation of s = 0.5 pounds. Is the true mean weight of textbooks sold in the bookstore equal to 3 pounds?

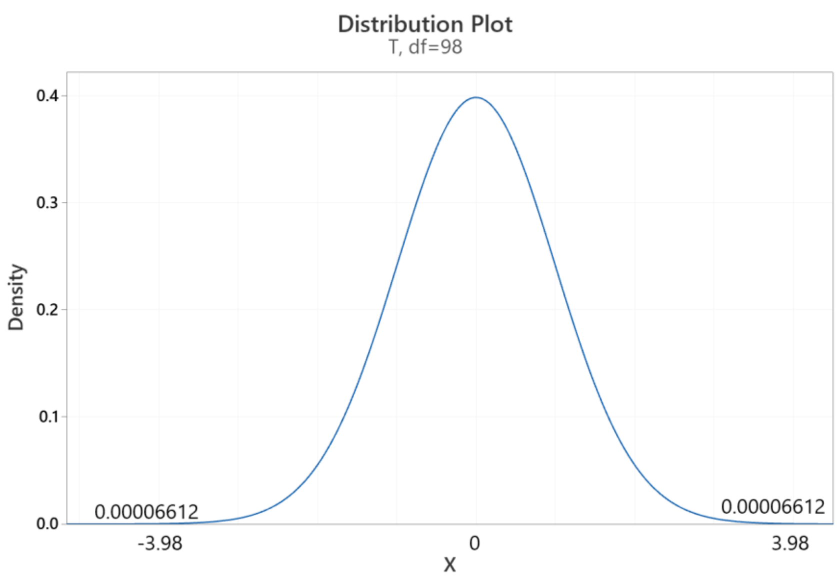
Solution: Since the sample size n = 99 is large, then the Central Limit Theorem guarantees us that $\bar{X}$ is normally distributed. Since we don't know the population standard deviation of the textbooks, then we'll use the sample standard deviation s = 0.05 to estimate it. We are being asked to perform the following hypothesis test:

$$H_0: \mu = 3 \text{ versus } H_a: \mu \neq 3.$$

Assuming H_0 is true, the standardized test statistic is:

$$t_0 = \frac{3.2 - 3}{\frac{0.5}{\sqrt{99}}} \cong 3.980.$$

Since the alternative hypothesis is "not equal to," then the p-value is ~2 · 0.00006612 = 0.000124. At significance level α = 0.05, we can reject H_0: μ = 3 in favor of H_a: μ ≠ 3.



Performing this hypothesis test in Minitab and using *Summarized data* for a 1-Sample t test, we get the same result.

Test

Null hypothesis	H_0: μ = 3
Alternative hypothesis	H_1: μ ≠ 3

T-Value	P-Value
3.98	0.000