

## 8.2: A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case

In practice, we rarely know the population **standard deviation**. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation  $s$  as an estimate for  $\sigma$  and proceeded as before to calculate a **confidence interval** with close enough results. This is what we did in an example in the previous section. The point estimate for the standard deviation,  $s$ , was substituted in the formula for the confidence interval for the population standard deviation. In this case there were 80 observations, well above the suggested 30 observations to eliminate any bias from a small sample. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Gosset (1876–1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing  $\sigma$  with  $s$  did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the **Student's  $t$ -distribution**. The name comes from the fact that Gosset wrote under the pen name "A Student."

Up until the mid-1970s, some statisticians used the **normal distribution** approximation for large sample sizes and used the Student's  $t$ -distribution only for sample sizes of at most 30 observations.

If you draw a simple random sample of size  $n$  from a population with mean  $\mu$  and unknown population standard deviation  $\sigma$  and calculate the  **$t$ -score**

$$t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)} \quad (8.2.1)$$

then the  $t$ -scores follow a **Student's  $t$ -distribution with  $n-1$  degrees of freedom**. The  $t$ -score has the same interpretation as the  $z$ -score. It measures how far in standard deviation units  $\bar{x}$  is from its mean  $\mu$ . For each sample size  $n$ , there is a different Student's  $t$ -distribution.

The **degrees of freedom,  $n-1$** , come from the calculation of the sample standard deviation  $s$ . Remember when we first calculated a sample standard deviation we divided the sum of the squared deviations by  $n-1$ , but we used  $n$  deviations ( $x_i - \bar{x}$  values) to calculate  $s$ . Because the sum of the deviations is zero, we can find the last deviation once we know the other  $n-1$  deviations. The other  $n-1$  deviations can change or vary freely. **We call the number  $n-1$  the degrees of freedom ( $df$ )** in recognition that one is lost in the calculations. The effect of losing a degree of freedom is that the  $t$ -value increases and the confidence interval increases in width.

### Properties of the Student's $t$ -Distribution

- The graph for the Student's  $t$ -distribution is similar to the standard normal curve and at infinite degrees of freedom it is the normal distribution.
- The mean for the Student's  $t$ -distribution is 0 and the distribution is symmetric about 0, again like the standard normal distribution.
- The Student's  $t$ -distribution has more probability in its tails than the standard normal distribution because the spread of the  $t$ -distribution is greater than the spread of the standard normal. So the graph of the Student's  $t$ -distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The exact shape of the Student's  $t$ -distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's  $t$ -distribution becomes more like the graph of the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean  $\mu$  and unknown population standard deviation  $\sigma$ . This assumption comes from the Central Limit Theorem because the individual observations in this case are the  $\bar{x}$ s of the sampling distribution. The size of the underlying population is generally not relevant unless it is very small. If it is normal then the assumption is met and doesn't need discussion.

A probability table for the Student's  $t$ -distribution is used to calculate  $t$ -values at various commonly-used levels of confidence. The table gives  $t$ -scores that correspond to the confidence level (column) and degrees of freedom (row). When using a  $t$ -table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails. Notice that at the bottom the table will show the  $t$ -value for infinite degrees of

freedom. Mathematically, as the degrees of freedom increase, the  $t$  distribution approaches the standard normal distribution. You can find familiar Z-values by looking in the relevant  $\alpha$  column and reading values in the last row.

The Student's  $t$  distribution has one of the most desirable properties of the normal: it is symmetrical. What the Student's  $t$  distribution does is spread out the horizontal axis so it takes a larger number of standard deviations to capture the same amount of probability. In reality there are an infinite number of Student's  $t$  distributions, one for each adjustment to the sample size. As the sample size increases, the Student's  $t$  distribution become more and more like the normal distribution. When the sample size reaches 30 the normal distribution is usually substituted for the Student's  $t$  because they are so much alike. This relationship between the Student's  $t$  distribution and the normal distribution is shown in Figure 8.2.1.

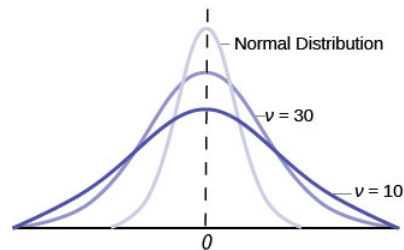


Figure 8.2.1

This is another example of one distribution limiting another one, in this case the normal distribution is the limiting distribution of the Student's  $t$  when the degrees of freedom in the Student's  $t$  approaches infinity. This conclusion comes directly from the derivation of the Student's  $t$  distribution by Mr. Gosset. He recognized the problem as having few observations and no estimate of the population standard deviation. He was substituting the sample standard deviation and getting volatile results. He therefore created the Student's  $t$  distribution as a ratio of the normal distribution and Chi squared distribution. The Chi squared distribution is itself a ratio of two variances, in this case the sample variance and the unknown population variance. The Student's  $t$  distribution thus is tied to the normal distribution, but has degrees of freedom that come from those of the Chi squared distribution. The algebraic solution demonstrates this result.

Development of Student's  $t$ -distribution:

$$1. T = \frac{Z}{\sqrt{\frac{X^2}{v}}}$$

Where  $Z$  is the standard normal distribution and  $X^2$  is the chi-squared distribution with  $v$  degrees of freedom.

$$2. T = \frac{\frac{(\bar{X} - \mu)}{\sigma}}{\sqrt{\frac{\frac{s^2}{(n-1)}}{\frac{\sigma^2}{(n-1)}}}}$$

by substitution, and thus Student's  $t$  with  $v = n - 1$  degrees of freedom is:

$$3. T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Restating the formula for a confidence interval for the mean for cases when the sample size is smaller than 30 and we do not know the population standard deviation,  $\sigma$ :

$$\bar{x} - t_{\nu, \alpha/2} \left( \frac{s}{\sqrt{n}} \right) \leq \mu \leq \bar{x} + t_{\nu, \alpha/2} \left( \frac{s}{\sqrt{n}} \right)$$

Here the point estimate of the population standard deviation,  $s$ , has been substituted for the population standard deviation,  $\sigma$ , and  $t_{\nu, \alpha/2}$  has been substituted for  $z_{\alpha/2}$ . The Greek letter  $\nu$  (pronounced "nu") is placed in the general formula in recognition that there are many Student  $t_{\nu}$  distributions, one for each sample size. The symbol  $\nu$  is the symbol for the degrees of freedom of the distribution and depends on the size of the sample. Often df is used to abbreviate degrees of freedom. **For this type of problem**, the degrees of freedom is  $\nu = n - 1$ , where  $n$  is the sample size. To look up a probability in the Student's  $t$  table we have to know the degrees of freedom in the problem.

### ? Example 8.2.1

The average earnings per share (EPS) for 10 industrial stocks randomly selected from those listed on the Dow-Jones Industrial Average was found to be  $\bar{x} = 1.85$  with a standard deviation of  $s = 0.395$ . Calculate a 99% confidence interval for the average EPS of all the industrials listed on the *DJIA*.

$$\bar{x} - t_{\nu, \alpha/2} \left( \frac{s}{\sqrt{n}} \right) \leq \mu \leq \bar{x} + t_{\nu, \alpha/2} \left( \frac{s}{\sqrt{n}} \right)$$

### Answer

To help visualize the process of calculating a confidence interval we draw the appropriate distribution for the problem. In this case this is the Student's  $t$  because we do not know the population standard deviation and the sample is small, less than 30.

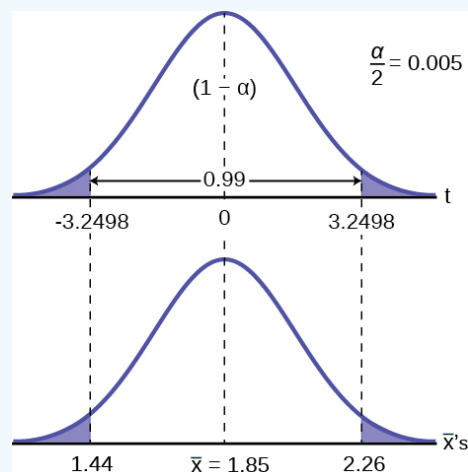


Figure 8.2.2

To find the appropriate  $t$ -value requires two pieces of information, the level of confidence desired and the degrees of freedom. The question asked for a 99% confidence level. On the graph this is shown where  $(1 - \alpha)$ , the level of confidence, is in the unshaded area. The tails, thus, have .005 probability each,  $\alpha/2$ . The degrees of freedom for this type of problem is  $n - 1 = 9$ . From the Student's  $t$  table, at the row marked 9 and column marked .005, is the number of standard deviations to capture 99% of the probability, 3.2498. These are then placed on the graph remembering that the Student's  $t$  is symmetrical and so the  $t$ -value is both plus or minus on each side of the mean.

Inserting these values into the formula gives the result. These values can be placed on the graph to see the relationship between the distribution of the sample mean,  $\bar{X}$ , and the Student's  $t$  distribution.

$$\mu = \bar{X} \pm t_{\alpha/2, df=n-1} \frac{s}{\sqrt{n}} = 1.85 \pm 3.2498 \frac{0.395}{\sqrt{10}} = 1.85 \pm 0.406$$

$$1.445 \leq \mu \leq 2.257$$

We state the formal conclusion as :

With 99% confidence level, the average *EPS* of all the industries listed at *DJIA* is from \$1.44 to \$2.26.

### ? Exercise 8.2.2

You do a study of hypnotherapy to determine how effective it is in increasing the number of hours of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results. Construct a 95% confidence interval for the mean number of hours slept for the population (assumed normal) from which you took the data.

8.2; 9.1; 7.7; 8.6; 6.9; 11.2; 10.1; 9.9; 8.9; 9.2; 7.5; 10.5

This page titled [8.2: A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [OpenStax](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.