

## 8.3: A Single Population Mean using the Student t-Distribution

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. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Goset (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 "Student."

Up until the mid-1970s, some statisticians used the normal distribution approximation for large sample sizes and only used the Student's  $t$ -distribution only for sample sizes of at most 30. With graphing calculators and computers, the practice now is to use the Student's  $t$ -distribution whenever  $s$  is used as an estimate for  $\sigma$ . If you draw a simple random sample of size  $n$  from a population that has an approximately a normal distribution with mean  $\mu$  and unknown population standard deviation  $\sigma$  and calculate the  $t$ -score

$$t = \frac{\bar{x} - \mu}{\left( \frac{s}{\sqrt{n}} \right)}, \quad (8.3.1)$$

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  $\bar{x}$  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  $s$ . In 2.8, we used  $n$  deviations ( $x - \bar{x}$  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).

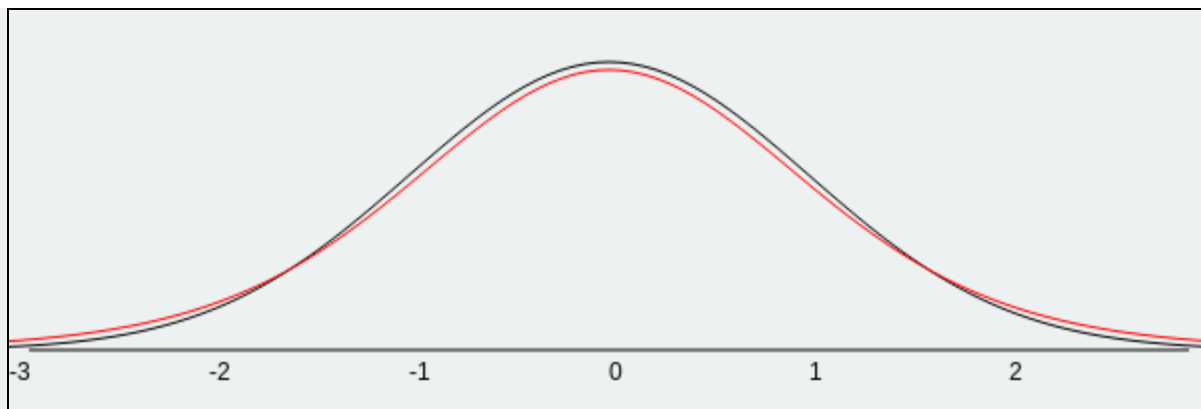
*For each sample size  $n$ , there is a different Student's  $t$ -distribution.*

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

- The graph for the Student's  $t$ -distribution is similar to the standard normal curve.
- The mean for the Student's  $t$ -distribution is zero and the distribution is symmetric about zero.
- 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  $\sigma$ . The size of the underlying population is generally not relevant unless it is very small. If it is bell shaped (normal) then the assumption is met and

### Visually Compare the Student's $t$ -Distribution with the normal distribution

The graphs below show the standard normal distribution in black and the student's  $t$  distribution in red. Move the slider to see how changing the sample size, and thus the degrees of freedom, changes the student's  $t$  distribution.



$n = 10, df = 9$

Calculators and computers can easily calculate any Student's  $t$ -probabilities.

The notation for the Student's  $t$ -distribution (using  $T$  as the random variable) is:

- $T \sim t_{df}$  where  $df = n - 1$ .
- For example, if we have a sample of size  $n = 20$  items, then we calculate the degrees of freedom as  $df = n - 1 = 20 - 1 = 19$  and we write the distribution as  $T \sim t_{19}$ .

If the population standard deviation is not known, the error bound for a population mean is:

- $E = (t_{\alpha/2}) \left( \frac{s}{\sqrt{n}} \right)$ ,
- $t_{\alpha/2}$  is the  $t$ -score with area to the right equal to  $\frac{\alpha}{2}$ ,
- use  $df = n - 1$  degrees of freedom, and
- $s$  = sample standard deviation.

The format for the confidence interval is:

$$(\bar{x} - E, \bar{x} + E). \quad (8.3.2)$$





### Example 8.3.1: Acupuncture

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given. Use the sample data to construct a 95% confidence interval for the mean sensory rate for the population (assumed normal) from which you took the data.

8.6; 9.4; 7.9; 6.8; 8.3; 7.3; 9.2; 9.6; 8.7; 11.4; 10.3; 5.4; 8.1; 5.5; 6.9

#### Answer

Use [Confidence Interval for a Mean With Data](#).

Enter the data separated by commas (The semicolons need to be replaced with commas.)

Enter 0.95 for CL.

Hit Calculate

The 95% confidence interval is (7.3006, 9.1527)

With 95% confidence, the population mean sensory rate for people who receive acupuncture is between 7.3006 and 9.1527.

### CONFIDENCE INTERVAL Calculator WITH STATISTICS (Sigma unknown)

Fill in the sample size ( $n$ ), the sample mean ( $\bar{x}$ ), the sample standard deviation ( $s$ ), and the confidence level (CL). Write the confidence level as a decimal. For example, for a 95% confidence level, enter 0.95 for CL. Then hit Calculate and assuming the population is normally distributed, the confidence interval will be calculated for you.

n:	$\bar{x}$ :	s:	CL:
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

Calculate	Lower Bound	Upper Bound
	<input type="text"/>	<input type="text"/>

Type in the values from the data set separated by commas, for example, 2,4,5,8,11,2. Then type in the confidence level, CL, and hit Calculate. Write the confidence level as a decimal. For example, for a 95% confidence level, enter 0.95 for CL.

Data:

CL:	<input type="button" value="Calculate"/>	$\bar{x}$ :	s:	Lower Bound	Upper Bound
<input type="text"/>		<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

### Exercise 8.3.2

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

#### Answer

(8.1634, 9.8032)

### Example 8.3.2: The Human Toxome Project

The Human Toxome Project (HTP) is working to understand the scope of industrial pollution in the human body. Industrial chemicals may enter the body through pollution or as ingredients in consumer products. In October 2008, the scientists at HTP tested cord blood samples for 20 newborn infants in the United States. The cord blood of the "In utero/newborn" group was tested for 430 industrial compounds, pollutants, and other chemicals, including chemicals linked to brain and nervous system toxicity, immune system toxicity, and reproductive toxicity, and fertility problems. There are health concerns about the effects of some chemicals on the brain and nervous system. The table below shows how many of the targeted chemicals were found in each infant's cord blood.

79	145	147	160	116	100	159	151	156	126
137	83	156	94	121	144	123	114	139	99

Use this sample data to construct a 90% confidence interval for the mean number of targeted industrial chemicals to be found in an infant's blood.

#### Answer

[Use Confidence Interval for a Mean With Data.](#)

Enter the data separated by commas (The semicolons need to be replaced with commas.)

Enter 0.90 for CL.

Hit Calculate

The 90% confidence interval is (7.3006, 9.1527)

We estimate with 90% confidence that the population mean number of all targeted industrial chemicals found in cord blood in the United States is between 117.412 and 137.488.

### Exercise 8.3.2

A random sample of statistics students were asked to estimate the total number of hours they spend watching television in an average week. The responses are recorded in Table. Use this sample data to construct a 98% confidence interval for the mean number of hours statistics students will spend watching television in one week.

0	3	1	20	9
5	10	1	10	4
14	2	4	4	5

#### Answer

Use [Confidence Interval for a Mean With Data](#).

Enter the data separated by commas (The semicolons need to be replaced with commas.)

Enter 0.98 for CL.

Hit Calculate

The 98% confidence interval is (2.397, 9.869)

We estimate with 98% confidence that the population mean number of all hours that statistics students spend watching television in one week is between 2.397 and 9.869.

## Reference

1. "America's Best Small Companies." Forbes, 2013. Available online at <http://www.forbes.com/best-small-companies/list/> (accessed July 2, 2013).
2. Data from *Microsoft Bookshelf*.
3. Data from <http://www.businessweek.com/>.
4. Data from <http://www.forbes.com/>.
5. "Disclosure Data Catalog: Leadership PAC and Sponsors Report, 2012." Federal Election Commission. Available online at <http://www.fec.gov/data/index.jsp> (accessed July 2, 2013).
6. "Human Toxome Project: Mapping the Pollution in People." Environmental Working Group. Available online at <http://www.ewg.org/sites/humantoxome...tero%2Fnewborn> (accessed July 2, 2013).
7. "Metadata Description of Leadership PAC List." Federal Election Commission. Available online at <http://www.fec.gov/finance/disclosur...pPacList.shtml> (accessed July 2, 2013).

## Glossary

### Degrees of Freedom ( $df$ )

the number of objects in a sample that are free to vary

### Normal Distribution

a continuous random variable (RV) with pdf  $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$ , where  $\mu$  is the mean of the distribution and  $\sigma$  is the standard deviation, notation:  $X \sim N(\mu, \sigma)$ . If  $\mu = 0$  and  $\sigma = 1$ , the RV is called **the standard normal distribution**.

### Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation:  $s$  for sample standard deviation and  $\sigma$  for population standard deviation

### Student's $t$ -Distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student; the major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as  $n$  get larger.
- There is a "family of  $t$ -distributions: each representative of the family is completely defined by the number of degrees of freedom, which is one less than the number of data.

## Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at <http://cnx.org/contents/30189442-699...b91b9de@18.114>.

This page titled [8.3: A Single Population Mean using the Student t-Distribution](#) 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.