Confidence Intervals for means

Suppose we have a large population we would like to study. We are interested in estimating the mean value of a quantity (a random variable) associated with the population from sample. We first suppose that we know the standard deviation $\sigma$ of the quantity.

We take a random sample of size $n$ from the population. We use the mean of the sample $\bar{x}$ to estimate the unknown parameter $\mu$. Note that every time we take a different sample, we get a slightly different estimate for $\mu$.

Thanks to central limit theorem, sample means, while varying, are very often normally distributed with mean $\mu$ and standard deviation $\frac{\sigma}{\sqrt{n}}$. This means, for example, that 95% of the time the sample mean will be with $2\frac{\sigma}{\sqrt{n}}$ of the population mean $\mu$. One way of thinking about this last statement is that if we go $2\frac{\sigma}{\sqrt{n}}$ on either side of our sample mean $\bar{x}$, then this interval will very often contain the true population mean. We call this interval $\left[ \bar{x} - 2\frac{\sigma}{\sqrt{n}}, \bar{x} + 2\frac{\sigma}{\sqrt{n}} \right]$, a 95% confidence interval for the population mean.

IMPORTANT: Confident about what?

It is important to understand the probability 95% really refers to. It means the following: if you were to sample the population repeatedly with random samples of size $n$ and each time form an interval in this way, about 95% of these intervals would contain the population mean $\mu$. Notice that you get a different confidence interval each time you sample. The probability refers to the entire sampling process. It is incorrect (or at least confusing) to say “there is a 95% chance that the mean $\mu$ is contained
\[ \left[ \bar{x} - 2 \frac{\sigma}{\sqrt{n}}, \bar{x} + 2 \frac{\sigma}{\sqrt{n}} \right] \].” A correct statement would be “95% of the time intervals formed like this one (by taking a random sample and going \( \pm 2 \frac{\sigma}{\sqrt{n}} \) on either side of the sample mean) will contain the population mean \( \mu \).”

Example. Suppose we are measuring the PCB concentration of Coho salmon caught in Lake Michigan. It is reasonable to assume that the PCB concentrations are normally distributed within an individual species. Previous experience with measurements using the same technique yielded standard deviation of 0.8 parts per million; we can safely assume that the standard deviation of PCB levels for Coho salmon is also 0.08 ppm. Suppose we catch 10 fish and measure their PCB concentrations:

11.2 12.4 10.8 11.6 12.5 10.1 11.0 12.2 12.4 10.6 (ppm)

Construct a 95% confidence interval for the mean PCB level in Coho salmon Lake Michigan based on this sample.

Answer. The sample mean is 11.48 ppm. The 95% confidence interval based on this sample is

\[ 11.48 - 2 \times 0.8/\sqrt{10} \approx 10.97 \text{ ppm} \text{ to } 11.48 + 2 \times 0.8/\sqrt{10} \approx 11.99 \text{ ppm}. \]
We say that “we are 95% confident that the mean PCB concentration for Coho salmon is between 10.43 ppm and 10.53 ppm.” What we mean by this is the following: if we were to do the same experiment over and over again (sampling 10 fish) and each time we did the experiment we form a different confidence interval based on the sample, 95% of the intervals formed will contain the mean PCB level for the entire population of Coho salmon.

What if the standard deviation of the parent population is also not known?

Realistically, if we don’t know the mean of the population parameter, we are unlikely to know the standard deviation. What do we do if we don’t know \( \sigma \)?

If you think about this problem for a while, we almost have no choice but to estimate \( \sigma \) by using the sample standard deviation (which we denote by \( s \)). In other words, instead of

\[ \left[ \bar{x} - 2 \frac{\sigma}{\sqrt{n}}, \bar{x} + 2 \frac{\sigma}{\sqrt{n}} \right] \]

we use

\[ \left[ \bar{x} - 2 \frac{s}{\sqrt{n}}, \bar{x} + 2 \frac{s}{\sqrt{n}} \right] \]
The problem with this idea is that we know right off the bat that it is not going to be correct because \( s \) also varies each time we take a sample. Because of this variation, we are in fact going to miss more the 5% of the time at least when the sample size is small. It is a somewhat subtle effect. For example, about half the time, \( s \) is going to be a little smaller than the true \( \sigma \). In those cases the interval will be a little smaller than they should be, and we are going to miss more often. On the other hand, about half the time, \( s \) is going to be a little larger than the true \( \sigma \). In those cases the interval will be a little wider than they should be, and we are going to contain the population mean slightly more often. However, in the end, careful calculations show that

\[
[\bar{x} - 2 \frac{s}{\sqrt{n}}, \bar{x} + 2 \frac{s}{\sqrt{n}}]
\]

is not an exact 95% confidence interval for \( \mu \).

W. S. Gossett in 1908 solved the problem of getting an exact 95% confidence interval when \( \sigma \) is unknown in the case that the parent population is normally distributed. He did it by figuring out what the question mark is

\[
[\bar{x} - \, \frac{s}{\sqrt{n}}, \bar{x} + \, \frac{s}{\sqrt{n}}]
\]

that would give a 95% confidence interval. Extra for experts: He did so by calculating the probability distribution of

\[
\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}
\]

He called the distribution \( t \). To this day, we call these distributions \( t \)-distributions. Because he published the result under the pseudonym, “Student,” they are often called “Student’s \( t \)-distributions.” \( t \)-distributions are rather similar to in shape to normal distributions. The graph below shows a normal distribution in black and the \( t \)-distribution with 3 degrees of freedom in red. The key difference is that the tails of \( t \)-distributions are longer.

The answer is as follows:
If the parent population is normally distributed, then an exact 95% confidence interval for the mean is

\[
\left( \bar{x} - t \frac{s}{\sqrt{n}}, \bar{x} + t \frac{s}{\sqrt{n}} \right)
\]

where \( t \) is the value so that the area between \(-t \) and \( t \) on the t-distribution with \( n-1 \) degrees of freedom is 95%.

Nowadays, we use computers to calculate the correct \( t \). Not too long ago, we looked up the correct value from tables. You can use Excel or Minitab to calculate the correct \( t \) as follows. In Excel, use TINV(0.05, degrees of freedom) where the “degrees of freedom” is the sample size minus 1. In Minitab, use Calc->Probability Distributions->t distribution

Minitab will also calculate the entire confidence interval directly using Stat->Basic Stats->One sample t.

Example
Recall the data for the PCB levels of Coho salmon were

11.2 12.4 10.8 11.6 12.5 10.1 11.0 12.2 12.4 10.6 (ppm)

Suppose we did not feel we could safely assume the standard deviation is 0.8 ppm as we did in the above example. However, based on a normal probability plot of the data

![Probability Plot of PCB levels](image)

it is reasonable to assume that the data comes from normally distributed population. So, the Student 95% confidence interval should be an accurate 95% confidence interval for the mean PCB levels of the fish.

The sample mean is 11.48 ppm; the sample standard deviation is 0.8638 ppm. The appropriate t-value in this case is TINV(0.05, 9) = 2.262. So the 95% confidence interval for the mean based on this sample is $11.48 - 2.262 \times 0.8638/\sqrt{10} \approx 10.86$ to $11.48 + 2.262 \times 0.8638/\sqrt{10} \approx 12.10$. You can get this result in one step in Minitab by using Stat->Basic Statistics->One sample t.

**Helpful facts that make life easier when the sample size is large**

**Helpful fact #1.** The t-distributions get closer and closer to the normal distribution as the sample size gets large. In particular, once the sample size is 25 or 30, there is only a very miniscule difference in the confidence intervals calculated with the correct t-value or just using a normal distribution. For example, suppose you have a sample of size 30 with sample mean 50 and sample standard deviation 3. The correct 95% confidence interval for the population mean based on this sample is

$50 - 2.042 \times 3/\sqrt{30} \approx 48.88$ to $50 + 2.042 \times 3/\sqrt{30} \approx 51.12$

If we use a confidence interval based on the normal distribution we would get

$50 - 1.96 \times 3/\sqrt{30} \approx 48.93$ to $50 + 1.96 \times 3/\sqrt{30} \approx 51.07$

The difference here is that the correct confidence interval is about 0.05 wider on each side which in most cases is completely irrelevant. So if the sample size is large, we can just as well use normal
distributions instead of t-distributions. When publishing results, one uses the t-distribution intervals however.

**Even more helpful fact #2.** Unfortunately, the confidence intervals are only correct when the parent population is assumed to be normally distributed. However, even if the parent population is not normally distributed, if the sample size is large (25 or 30 or larger), the Central Limit Theorem assures us the sampling distribution is *approximately* normally distributed. The t-distributions are also very close to being normal distributions when the sample size is large. **So in practice, if the sample size is large (25 or so in most cases), the confidence intervals formed with the t-distribution are good approximate confidence intervals even when the parent population is not known to be normally distributed.**

**What to watch out for**

The case you have to watch out for is when you are taking small samples from a population which you cannot assume to be normally distributed.

For example consider the following data set from a research paper (Mena, E. A. et al “Inflammatory intermediates produced by tissues encasing silicone breast implants,” *Journal of Investigative Surgery* 8: 31-42, 1995) on silicone breast implants.

<table>
<thead>
<tr>
<th>Interleukin-6 levels (pg/mL/10 g of tissue)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>231</td>
</tr>
<tr>
<td></td>
<td>308287</td>
</tr>
<tr>
<td></td>
<td>33291</td>
</tr>
<tr>
<td></td>
<td>124550</td>
</tr>
<tr>
<td></td>
<td>17075</td>
</tr>
<tr>
<td></td>
<td>22955</td>
</tr>
<tr>
<td></td>
<td>95102</td>
</tr>
<tr>
<td></td>
<td>5649</td>
</tr>
<tr>
<td></td>
<td>840585</td>
</tr>
<tr>
<td></td>
<td>58924</td>
</tr>
</tbody>
</table>

The sample size is 10. Can we safely assume that the data comes from a normally distributed population? No. A normal probability plot verifies this assertion:
Can we safely create a confidence interval for the mean using the t-distribution? No. We have to use other techniques. The two techniques most commonly used to deal with this situation (which I will not go into at the moment) are: transformation of the variable and non-parametric methods.