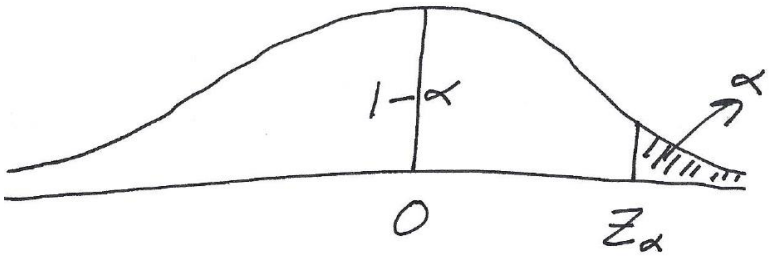
Section 8: Confidence Intervals Based on a Single Sample

Confidence in Inferential Statistics

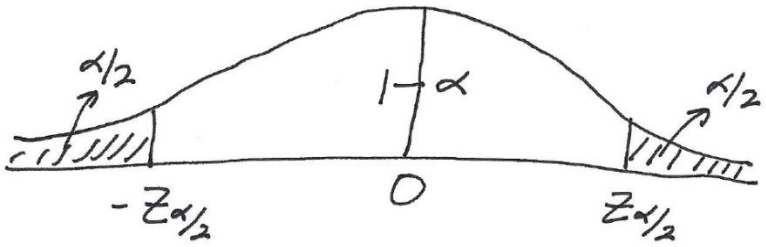
When conducting inferential statistics, analysis will be either one-tailed or two-tailed. A one-tailed test means that the error is in one tail of the sampling distribution and two-tailed means that half of the error is in each tail of the sampling distribution. Since we have already discussed the central limit theorem, we know that our sampling distribution is normal in certain instances. We will start our discussion of confidence intervals in these situations where the sampling distribution is normal.

In inferential statistics, the quantity α, known as **level of significance**, is the likelihood of being wrong. The complement of α, $1-\alpha$, is known as our **level of confidence**. For example if we want to be 95% confident in our conclusion then $\alpha = .05$ (or 5%). There is a 95% chance that we are correct and a 5% chance we are wrong. As you will see, these values come from probability we can calculate using the sampling distribution.

If we conduct a one-tailed test and the sampling distribution is normal then we will need to find the value z_α. The z_α value is the z from the standard normal table with the area in the tail being α. Also, the remaining area is $1-\alpha$, the confidence level. Below is a drawing of what this looks like.



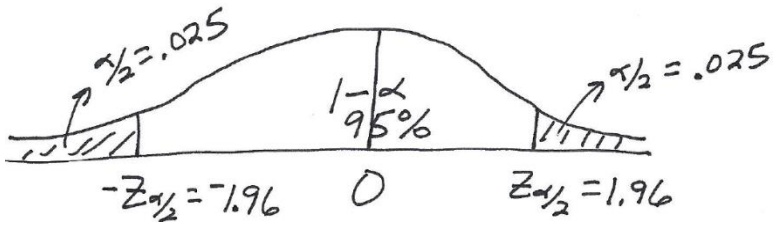
If we conduct a two-tailed test and the sampling distribution is normal then we will need to find the value $z_{\alpha/2}$. The $z_{\alpha/2}$ value is the z from the standard normal table with the area in the tail being $\frac{\alpha}{2}$. Since this is a two tailed test, half of α is in each tail. As before, the remaining area is $1-\alpha$, the confidence level. <u>All confidence intervals are two-tailed.</u> This is because when you have an interval there is a chance the actual value is below the interval and a chance that the actual value is above the interval (there is a possible error on both sides of the interval). Below is a drawing of what this looks like.



The confidence level is something a researcher can choose. Obviously, we do not want to choose a 50% confidence level. This would mean there is a 50% chance of being right and a 50% chance of being wrong. A confidence level this low does not give useful information about the parameter being estimated. We want to pick a high confidence level. Common confidence levels used in statistical analysis are 90%, 95%, and 99%. The $z_{\alpha/2}$ values for these confidence levels are identified below.

For 90% Confidence, $z_{\alpha/2} = z_{.05} = 1.645$
For 95% Confidence, $z_{\alpha/2} = z_{.025} = 1.96$
For 99% Confidence, $z_{\alpha/2} = z_{.005} = 2.576$

These $z_{\alpha/2}$ values come from the standard normal distribution as seen in the picture presented earlier. To give a specific example we will look at 95% confidence. For 95% confidence, $\alpha = .05$, $\frac{\alpha}{2} = .025$, and $1 - \alpha = .95$ as seen in the graph below. For the standard normal distribution 95% of observations are between -1.96 and 1.96. Therefore, $z_{\alpha/2} = 1.96$. To see where this value comes from, look at the standard normal distribution table and find the area between -1.96 and 1.96. If you look up 1.96, you will see the area is .4750 which is half the area we want. Double this value and you get .95 or 95%. All other $z_{\alpha/2}$ values are calculated in the same way. I will only ask you to do 90%, 95% and 99% confidence intervals, but if you wanted to choose a different confidence level, you can figure out the $z_{\alpha/2}$ value in this way.



Ultimately, the above graph illustrates that when our sampling distribution is normal, 95% of statistics are within 1.96 standard deviations of the parameter they estimate.

A common question: Can we do a 100% confidence interval? The answer is technically yes, but the interval is of no use. To get 100% of the area, $z_{\alpha/2} = \infty$. This should make since with the x-axis being an asymptote for the normal distribution. Therefore, a 100% confidence interval is always just $-\infty$ to ∞ (the entire real number line). We knew the parameter was somewhere on the real number line before we ever collected any data. In order to narrow down where the parameter is, the confidence level must be something less than 100%.

Identifying and Estimating the Parameter

Why not simply use a statistic to estimate a parameter? Is interval estimation really necessary? The mean will be utilized in order illustrate the importance of interval estimation.

When doing inferential statistics, the population mean, μ, is an unknown parameter. We wish to estimate μ based on a sample. The statistic $\bar{x}$ estimates the parameter μ. We call $\bar{x}$ a **point estimate** because its value is a specific point on the real number line. Unfortunately, if we sample from a continuous probability distribution, $P(\bar{x} = \mu) = 0$ (the probability for any single point in a continuous distribution is 0). Therefore, we are sure the estimate is wrong if we just use a point estimate. Thus, statisticians prefer interval estimates which are referred to as confidence intervals.

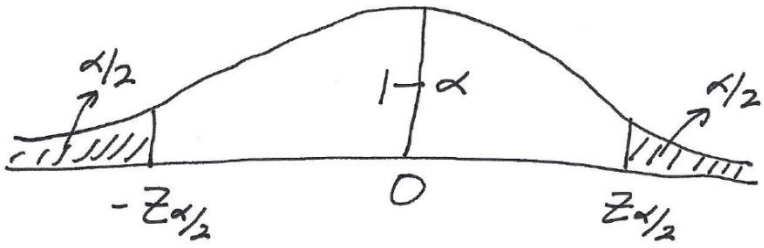
To create a confidence interval, we must calculate the **error tolerance**, denoted *E*. Error tolerance (also known as margin of error) is calculated from the standard error (standard deviation of the sampling distribution). The number of standard deviations we need depends on how confident we want to be as presented earlier. This means that error tolerance depends on three quantities: sample size, confidence level, and the amount of variability in the population.

Derivation of Confidence Interval for μ with known σ

A confidence interval is derived from the underlying sampling distribution. In this case, we know the underlying sampling distribution is normal because of the following transformation.

When σ is known $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ is approximately a standard normal distribution.

Combining this fact with the picture below, we can see where confidence intervals come from.



From this picture we can write the following probability statement:

$$P\left(-z\alpha/_2 < Z < z\alpha/_2\right) = 1 - \alpha$$

Next we plug in for *Z* using the transformation presented above and solve so that the parameter, μ, is isolated in the center of the inequality. The algebraic steps follow.

$$P\left(-z\alpha_{/2} < Z < z\alpha_{/2}\right) = 1-\alpha$$
$$P\left(-z\alpha_{/2} < \frac{\bar{x}-\mu}{\sigma_{/\sqrt{n}}} < z\alpha_{/2}\right) = 1-\alpha$$
$$P\left(-z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}} < \bar{x}-\mu < z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$
$$P\left(-\bar{x} - z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{x} + z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$
$$P\left(\bar{x} + z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}} > \mu > \bar{x} - z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$
$$P\left(\bar{x} - z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

We know that the population mean, μ, will be between $\bar{x} - z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}}$ and $\bar{x} + z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}}$ with a confidence level of $1-\alpha$. Notice that to get this interval, the same quantity is being subtracted and added to $\bar{x}$. This quantity is *E*, the error tolerance.

Confidence Interval Formula for μ with known σ

$$E = z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}} = z\left(\frac{\sigma}{\sqrt{n}}\right)$$
$$\bar{x} \pm E$$

It is common to not write the $\frac{\alpha}{2}$ when writing the formula for *E*.

Looking at the formula, you should understand the following relationships:

- As the sample size increases, the error tolerance decreases giving a narrower interval. (Taking a larger sample means we get more information and less error. Thus, we are narrowing down the possible values of the parameter. Mathematically, increasing the denominator of the standard error makes the error tolerance smaller.)
- As the confidence level increases, the error tolerance increases giving a wider interval. (If there is more room in the interval, wider interval, then naturally you would be more confident. There is an important relationship to notice here. In confidence intervals there are really two kinds of error, α and *E*. If you increase the confidence level then you are decreasing α. That error must go somewhere, so it is just moved to *E*. Basically, in inferential statistics there are always two types of error. If you decrease one, then the other will increase. Mathematically, increasing the confidence level will increase *z* and give a wider interval.)
- In situations where the variability in the population is higher the error tolerance will also be higher if other factors are held constant. (More variability means more error so *E* increases which gives a wider interval. Mathematically, increasing the numerator of the standard error makes the error tolerance larger.)

Example

A sample of 100 visa accounts were studied for the amount of unpaid balance.

$n = 100$

$\bar{x} = 645$

$\sigma = 132$

a. Construct a 95% confidence interval for μ.
b. Interpret the 95% confidence interval for μ.
c. Construct a 99% confidence interval for μ.
d. Interpret the 99% confidence interval for μ.

Solution

a. $E = z\left(\frac{\sigma}{\sqrt{n}}\right) = 1.96\left(\frac{132}{\sqrt{100}}\right) = 25.87$

   $\bar{x} \pm E$

   645 ± 25.87

   $(619.13, 670.87)$

b. We are 95% confident that the population mean unpaid balance of visa accounts is between \$619.13 and \$670.87.

c. $E = z\left(\frac{\sigma}{\sqrt{n}}\right) = 2.576\left(\frac{132}{\sqrt{100}}\right) = 34.00$

   $\bar{x} \pm E$

   645 ± 34.00

   $(611.00, 679.00)$

d. We are 99% confident that the population mean unpaid balance of visa accounts is between \$611.00 and \$679.00.

Confidence Interval for μ with unknown σ

You may have noticed a problem with using $E = z\left(\frac{\sigma}{\sqrt{n}}\right)$. The issue is when we are estimating the population mean, μ, we will not know the population standard deviation, σ (the only way to know this quantity would be to have the population, and if we had the population we could just calculate the mean). In statistics, when we do not know a value, we estimate it. So, in this case we can estimate the population standard deviation, σ, with the sample standard deviation,*s*. The ramification of replacing σ by *s* is the sampling distribution is no longer standard normal. The sampling distribution will instead be a t-distribution.

If we sample from a normal distribution,

$t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ has a t-distribution with $n - 1$ degrees of freedom.

The assumption that we sample from a normal population is important for small *n* but not for large *n*.

The t-distribution is still continuous and symmetric about 0, but it is more variable and a slightly different shape than the standard normal distribution. Basically, we are estimating another quantity so there is more error in the estimate. The confidence interval formula for μ using the t-distribution follows:

$$E = t\left(\frac{s}{\sqrt{n}}\right)$$
$$\bar{x} \pm E$$

As you can see the only change in the formula is σ being replaced by *s* which means the *z* becomes a *t*. Another important thing to recognize here is the relationship between the *t* and the *z*. Notice that as *n* becomes large, the t-distribution can be approximated by the standard normal distribution. This is because the larger *n* becomes, the better *s* is at estimating σ. When *n* approaches ∞, *s* becomes σ so the *t* becomes a *z*. (The bottom row of the t-distribution table is *z*. If you need to look up a *z* for a confidence interval, this is the best place to look. The t-distribution table is on the webpage under tables.)

We use two subscripts to write *t* values. The first represents the area in the tail of the distribution (just like with the *z*) and the second represents the degrees of freedom (abbreviated df). We write $t_{\alpha,df}$ for a one-tailed value and $t_{\alpha/2,df}$ for a two-tailed value. The degrees of freedom is an index that identifies how much variability is in the distribution. As mentioned above this changes when *n* increases because *s* becomes a better estimate for σ.

Example
Find the two-tailed *t* value with a sample size of 11 and a confidence level of 95%.

Solution
$\frac{\alpha}{2} = \frac{.05}{2} = .025$
$df = n - 1 = 11 - 1 = 10$
$t_{.025,10} = 2.228$

For a t-distribution with 10 degrees of freedom, 95% of the area is between -2.228 and 2.228.

Make sure you understand how to find the value above in the t-distribution table. This table is laid out differently than the standard normal distribution table. Also, you should notice that the degrees of freedom go from 1 to 30 and then skip numbers. If your degrees of freedom are between 30 and 120, pick the closest value to estimate the *t*. If your degrees of freedom is above 120 then use the bottom row of the table. Finally, you should notice that the bottom row of the table, $df = \infty$, are *z* values.

Example
Mileage of tires in 1000's of miles
Sample: 42, 36, 46, 43, 41, 35, 43, 45, 40, 39
Compute and interpret a 95% confidence interval for μ.

Solution
$n = 10$
$\bar{x} = 41$
$s = 3.590$
$\frac{\alpha}{2} = \frac{.05}{2} = .025$
$df = n - 1 = 10 - 1 = 9$
$t_{.025,9} = 2.262$

$E = t\left(\frac{s}{\sqrt{n}}\right) = 2.262\left(\frac{3.590}{\sqrt{10}}\right) = 2.568$
$\bar{x} \pm E$
41 ± 2.568
$(38.432, 43.568)$

We are 95% confident that the population mean mileage of tires is between 38,432 and 43,568 miles.

Example
A random sample of $n = 20$ apples yields $\bar{x} = 9.2\ oz.$ and $s = 1.1\ oz.$
Find and interpret a 99% confidence interval for μ.

Solution
$\frac{\alpha}{2} = \frac{.01}{2} = .005$
$df = n - 1 = 20 - 1 = 19$
$t_{.005,19} = 2.861$

$E = t\left(\frac{s}{\sqrt{n}}\right) = 2.861\left(\frac{1.1}{\sqrt{20}}\right) = .704$
$\bar{x} \pm E$
$9.2 \pm .704$
$(8.496, 9.904)$

We are 99% confident that the population mean weight of apples is between 8.496 and 9.904 oz.

Confidence Interval for *p* when *n* is Large

For large *n*, $Z = \frac{\hat{p}-p}{\sqrt{\frac{p \cdot q}{n}}}$ is approximately standard normal.

Using the same principles as we did with the mean, the error tolerance is a *z* value times the standard error. We then add and subtract this from the statistic, $\hat{p}$, in order to create the interval. This is the basic form of all confidence intervals. We do have one issue here. The standard error is $\sqrt{\frac{p \cdot q}{n}}$ and we do not know *p*. Therefore, we have to estimate *p* with $\hat{p}$ (and estimate *q* with $\hat{q}$). This only works when our sample size is large because then we know the sampling distribution is approximately normal. The sampling distribution for small samples is not known.

$$E = z\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$$
$$\hat{p} \pm E$$

Example
A survey of 1,200 registered voters yields 540 plan to vote for the republican candidate.

p = percentage of all registered voters who plan to vote for the republican candidate

Find a 95% confidence interval for *p*.

Solution
$$\hat{p} = \frac{540}{1200} = 0.45 = 45\%$$
$$\hat{q} = 1 - \hat{p} = 1 - .45 = .55 = 55\%$$
$$\frac{\alpha}{2} = \frac{.05}{2} = .025$$
$$z_\alpha = 1.96$$

$$E = z\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}} = 1.96\sqrt{\frac{(.45)(.55)}{1200}} = .028 = 2.8\%$$
$$\hat{p} \pm E$$
$$45\% \pm 2.8\%$$
$$(42.2\%, 47.8\%)$$

We are 95% confident that the population percentage of voters who plan to vote for the republican candidate is between 42.2% and 47.8%.

Determining the Sample Size

In the design stages of statistical research, it is good to decide the confidence level you wish to use and to select the error tolerance you want for the project. This will allow you to decide how big the sample needs to be. In calculating sample size, we will only deal with the scenario of estimating μ with known σ. If you can do this one, the others work the same way.

To get the formula for sample size, all you have to do is solve the error tolerance formula for *n*. So in the case of estimating μ with known σ, the formula is as follows:

$$E = z\left(\frac{\sigma}{\sqrt{n}}\right)$$
$$\sqrt{n} \cdot E = z\left(\frac{\sigma}{\sqrt{n}}\right) \cdot \sqrt{n}$$
$$\frac{\sqrt{n} \cdot E}{E} = \frac{z \cdot \sigma}{E}$$
$$\sqrt{n} = \frac{z \cdot \sigma}{E}$$
$$n = \left(\frac{z \cdot \sigma}{E}\right)^2$$

Notice that in the above equation we will have to estimate σ. This is commonly done with a pilot study or preliminary sampling. Once we have an estimate for σ, we can figure out how large a sample to take.

Example

We wish to estimate the number of patient-visit hours per week physicians in solo practice spent. How large a sample is needed if we want to be 99% confident that our point estimate is within 1 hour of the population mean? Assume a standard deviation of 11.97 hours.

Solution

$$z_{.005} = 2.576$$

$$n = \left(\frac{z \cdot \sigma}{E}\right)^2 = \left(\frac{(2.576)(11.97)}{1}\right)^2 = 950.779 \approx 951$$

With sample size, any fractional part of the answer may be needed to get the error tolerance we want. Since we cannot sample part of an observation, it is important to always round up to the next whole number.