Confidence interval for the difference of means

So far, we've been looking at hypothesis testing for one population, but the goal of many statistical studies is to compare two populations, often in terms of their means.

The difference of means

For example, maybe we want to compare the effectiveness of two different diet plans. We could define two populations; one group that follows the first diet plan, and a second group that follows the second diet plan.

Then the population means are the mean weight lost on the first diet plan μ_1, and the mean weight lost on the second diet plan μ_2. We could take a sample from each population, in which case the point estimators are the mean of each sample, $\bar{x}_1$ and $\bar{x}_2$, respectively.

Now let's say that what we're actually interested in is the difference of means, $\mu_1 - \mu_2$. In other words, we want to know how much more or less weight we can expect to lose if we follow the first diet plan instead of the second diet plan.

In this lesson, we'll look at how to build a confidence interval around the difference of sample means. By the end of this lesson, we'll want to be able to make a statement like

> *"95 % of the confidence intervals I construct around the sample statistic $\bar{x}_1 - \bar{x}_2$ will contain the population parameter $\mu_1 - \mu_2$,"*

or in simpler terms

> *"I'm* $95\,\%$ *certain that the difference in population means* $\mu_1 - \mu_2$ *will fall within the confidence interval* $(\bar{x}_1 - \bar{x}_2) \pm$ *margin of error."*

Just like before when we were investigating the mean of only one population (by taking only one sample), when we build a confidence interval for the difference of means, we'll use one confidence interval formula when population standard deviations are known, and a different confidence interval formula when population standard deviations are unknown and/or the sample sizes are small $n_1,\, n_2 < 30$.

With known standard deviations

This won't usually be the case, but let's assume that we know the standard deviation of both populations, σ_1 and σ_2. We'll take a sample of size n_1 from the first population, and a sample of size n_2 from the second population. Then we'll calculate the mean of each sample to find $\bar{x}_1$ and $\bar{x}_2$.

As long as both of the original populations were normally distributed, and/ we take large enough samples $n_1,\, n_2 \geq 30$ (so that the Central Limit Theorem kicks in), then the sampling distributions of $\bar{x}_1$ and $\bar{x}_2$ will be normally distributed, and therefore the **sampling distribution of the difference of means** $\bar{x}_1 - \bar{x}_2$ will be normally distributed as well. The mean of the sampling distribution of the difference of means will be

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2}$$

In other words, the mean of the sampling distribution of the difference of means is the difference of the means of the sampling distributions of the sample means.

The standard error of the sampling distribution of the difference of means will be

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Then the formula for the confidence interval is

$$(a, b) = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sigma_{\bar{x}_1-\bar{x}_2}$$

$$(a, b) = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $\bar{x}_1 - \bar{x}_2$ is the difference of sample means, $z_{\alpha/2}$ is the critical z-value, σ_1 and σ_2 are the population standard deviations, and n_1 and n_2 are the sample sizes.

Let's work through an example where we calculate the confidence interval around a difference of means.

Example

A research team is testing the effect of a low-carb diet on people with type 2 diabetes. 400 people are assigned to group 1 and put on a low-carb diet, while another 400 people are assigned to group 2 and put on a standard diet. Given the sample means and population standard

deviations below, estimate a $95\,\%$ confidence interval for the difference between the mean drop in blood sugar levels.

Group 1 (Low carb)	Group 2 (Standard)
$n_1 = 400$	$n_2 = 400$
$\bar{x}_1 = 9.5$ mg/dL drop	$\bar{x}_2 = 3.2$ mg/dL drop
$\sigma_1 = 0.35$ mg/dL	$\sigma_2 = 0.28$ mg/dL

At a confidence level of $95\,\%$, we know $\alpha = 0.05$ and $\alpha/2 = 0.025$. Using a z-table, we need to find the z-score that corresponds to $1 - 0.025 = 0.9750$.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	**.9750**	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817

So $z_{\alpha/2} = 1.96$, and we can substitute everything we know into the confidence interval formula.

$$(a, b) = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(a, b) = (9.5 - 3.2) \pm 1.96\sqrt{\frac{0.35^2}{400} + \frac{0.28^2}{400}}$$

$$(a, b) = 6.3 \pm 1.96\sqrt{\frac{0.1225}{400} + \frac{0.0784}{400}}$$

$$(a, b) = 6.3 \pm 1.96 \cdot \frac{\sqrt{0.2009}}{20}$$

$$(a, b) \approx 6.3 \pm 0.044$$

$$(a, b) \approx (6.256, 6.344)$$

So we can say with $95\,\%$ confidence that the mean drop in blood sugar levels is somewhere between 6.256 mg/dL and 6.344 mg/dL higher in the low-carb diet group than the drop seen in the standard diet group.

With unknown standard deviations and/or small samples

When our population standard deviations σ_1 and σ_2 are unknown, we can use the sample standard deviations s_1 and s_2 in their place. When we do, we have to consider two possible scenarios.

Unequal population variances

If the sample variances are significantly unequal, then we assume that the population variances are unequal, and our confidence interval formula will be

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\text{with df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Note: round down when df is not an integer, so that the estimate is more conservative

Equal (or almost equal) population variances

If the sample variances are equal or almost equal, then we assume that the population variances are approximately equal as well, and we calculate a pooled variance by combining the two sample variances into one. The formula for **pooled variance** is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

and therefore **pooled standard deviation** is

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

As a rule of thumb, we can use pooled variance when the two samples were taken from the same population, or when neither sample variance is more than twice the other.

In other words, if we take our samples from the same population, and we have no reason to believe that their variances or standard deviations will be different, then we can use pooled variance and pooled standard

deviation. But even if we take our samples from different populations, if their variances and standard deviations turn out to be close enough in value (neither is more than twice the other), then we can use the pooled formulas.

The standard error of the sampling distribution of the difference of means will change to

$$\sigma_{\bar{x}_1-\bar{x}_2} = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

and our confidence interval formula will be

$$(a,b) = (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \times s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$\text{with df} = n_1 + n_2 - 2$$

Let's rework the same example from before, but this time we'll use smaller samples, such that we'll be forced to use one of these t-score confidence interval formulas.

Example

A research team is testing the effect of a low-carb diet on people with type 2 diabetes. 25 people are assigned to group 1 and put on a low-carb diet, while another 25 people are assigned to group 2 and put on a standard diet. Given the sample means and sample standard deviations below, estimate a $95\,\%$ confidence interval for the difference between the mean drop in blood sugar levels.

Group 1 (Low carb)	Group 2 (Standard)
$n_1 = 25$	$n_2 = 25$
$\bar{x}_1 = 9.5$ mg/dL drop	$\bar{x}_2 = 3.2$ mg/dL drop
$s_1 = 0.35$ mg/dL	$s_2 = 0.28$ mg/dL

Because $s_1 = 0.35$, the sample variance of group 1 is $s_1^2 = 0.35^2 = 0.1225$; and because $s_2 = 0.28$, the sample variance of group 2 is $s_2^2 = 0.0784$. So neither sample variance is more than twice the other, and we can say that the sample variances are approximately equal, and therefore that the population variances are approximately equal. Therefore, we can use the pooled standard deviation formula.

$$s_p = \sqrt{\frac{(25-1)0.35^2 + (25-1)0.28^2}{25+25-2}}$$

$$s_p = \sqrt{\frac{24(0.1225) + 24(0.0784)}{48}}$$

$$s_p = \sqrt{\frac{0.1225 + 0.0784}{2}}$$

$$s_p = \sqrt{\frac{0.2009}{2}}$$

$$s_p \approx 0.317$$

The number of degrees of freedom is given by

$$\text{df} = 25 + 25 - 2 = 48$$

When we look up these degrees of freedom for a $95\,\%$ confidence level in the student's t-table, we find 2.011. Now we can substitute what we know into the confidence interval formula.

$$(a, b) = (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(a, b) \approx (9.5 - 3.2) \pm 2.011 \times 0.317 \sqrt{\frac{1}{25} + \frac{1}{25}}$$

$$(a, b) \approx 6.3 \pm 2.011 \times 0.317 \sqrt{\frac{1}{25} + \frac{1}{25}}$$

$$(a, b) \approx 6.3 \pm 0.637363 \cdot \frac{\sqrt{2}}{5}$$

$$(a, b) \approx 6.3 \pm 0.180$$

$$(a, b) \approx (6.3 - 0.180, 6.3 + 0.180)$$

$$(a, b) \approx (6.12, 6.48)$$

So we can say with $95\,\%$ confidence that the mean drop in blood sugar levels is higher in the group that was on the low-carb diet than in the group that was on a standard diet, and that the difference between means will fall between 6.12 and 6.48 mg/dL.