Differences between Means

Prerequisites
Sampling Distribution of Difference between Means, Confidence Intervals, Confidence Interval on the Mean

It is much more common for a researcher to be interested in the difference between means than in the specific values of the means themselves. We take as an example the data from the "Animal Research" case study. In this experiment, students rated (on a 7-point scale) whether they thought animal research is wrong. The sample sizes, means, and variances are shown separately for males and females in Table 1.

Table 1. Means and Variances in Animal Research study.

Condition	n	Mean	Variance
Females	17	5.353	2.743
Males	17	3.882	2.985

As you can see, the females rated animal research as more wrong than did the males. This sample difference between the female mean of 5.35 and the male mean of 3.88 is 1.47. However, the gender difference in this particular sample is not very important. What is important is the difference in the *population*. The difference in sample means is used to estimate the difference in population means. The precision of the estimate is revealed by a *confidence interval*.

In order to construct a confidence interval, we are going to make three assumptions:

1. The two populations have the same variance. This assumption is called the assumption of *homogeneity of variance*.
2. The populations are *normally distributed*.
3. Each value is sampled *independently* from each other value.

The consequences of violating these assumptions are discussed in a later section. For now, suffice it to say that small-to-moderate violations of assumptions 1 and 2 do not make much difference.

A confidence interval on the difference between means is computed using

the following formula:

$$\text{Lower Limit} = M_1 - M_2 - (t_{CL})(S_{M_1-M_2})$$
$$\text{Upper Limit} = M_1 - M_2 + (t_{CL})(S_{M_1-M_2})$$

where $M_1 - M_2$ is the difference between sample means, t_{CL} is the t for the desired level of confidence, and $S_{M_1-M_2}$ is the estimated *standard error* of the difference between sample means. The meanings of these terms will be made clearer as the calculations are demonstrated.

We continue to use the data from the "Animal Research" case study and will compute a confidence interval on the difference between the mean score of the males and the mean score of the females. For this calculation, we will assume that the variances in each of the two populations are equal. This assumption is called the assumption of homogeneity of variance.

The first step is to compute the estimate of the standard error of the difference between means ($S_{M_1-M_2}$). Recall from the relevant section in the chapter on sampling distributions that the formula for the standard error of the difference in means in the population is:

$$\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}}$$

In order to estimate this quantity, we estimate σ^2 and use that estimate in place of σ^2. Since we are assuming the population variances are the same, we estimate this variance by averaging our two sample variances. Thus, our estimate of variance is computed using the following formula:

$$\text{MSE} = \frac{s_1^2 + s_2^2}{2}$$

where MSE is our estimate of σ^2. In this example,

$$\text{MSE} = (2.743 + 2.985)/2 = 2.864.$$

Since n (the number of scores **in each condition**) is 17,

$$S_{M_1-M_2} = \sqrt{\frac{2MSE}{n}} = \sqrt{\frac{(2)(2.864)}{17}} = 0.5805.$$

The next step is to find the t to use for the confidence interval (t_{CL}). To calculate t_{CL}, we need to know the *degrees of freedom*. The degrees of freedom is the number of independent estimates of variance on which MSE is based. This is equal to $(n_1 - 1) + (n_2 - 1)$ where n_1 is the sample size for the first group and n_2 is the sample size of the second group. For this example, $n_1 = n_2 = 17$. When $n_1 = n_2$, it is conventional to use "n" to refer to the sample size of each group. Therefore the degrees of freedom is 16 + 16 = 32.

Online: Calculator: Find t for confidence interval

From either the above calculator or a t table, you can find that the t for a 95% confidence interval for 32 df is 2.0369.

We now have all the components needed to compute the confidence interval. First, we know the difference between means:

$$M_1 - M_2 = 5.3523 - 3.8824 = 1.470$$

We know the standard error of the difference between means is

$$S_{M_1 - M_2} = 0.5805$$

and that the t for the 95% confidence interval with 32 df is

$$t_{CL} = 2.0369$$

Therefore the 95% confidence interval is

$$\text{Lower Limit} = 1.470 - (2.0369)(0.5805) = 0.29$$

$$\text{Upper Limit} = 1.470 + (2.0369)(0.5805) = 2.65$$

We can write the confidence interval as:

$$0.29 \leq \mu_f - \mu_m \leq 2.65$$

where μ_f is the population mean for females and μ_m is the population mean for males. This analysis provides evidence that the mean for females is higher than the mean for males, and that the difference between means in the population is likely to be between 0.29 and 2.65.

Formatting data for Computer Analysis

Most computer programs that compute t tests require your data be in a

specific form. Consider the data in Table 2.

Table 2. Example Data

Group 1	Group 2
3	5
4	6
5	7

Here there are two groups, each with three observations. To format these data for a computer program, you normally have to use two variables: the first specifies the group the subject is in and the second is the score itself. For the data in Table 2, the reformatted data look as follows.

Table 3. Reformatted Data

G	Y
1	3
1	4
1	5
2	5
2	6
2	7

To use Analysis Lab to do the calculations, you would copy the data and then

1. Click the "Enter/Edit User Data" button (You may be warned that for security reasons you must use the keyboard shortcut for pasting data).
2. Paste your data.
3. Click "Accept Data"
4. Set the Dependent Variable to Y
5. Set the Grouping Variable to G
6. Click the t-test confidence interval button.

The 95% confidence interval on the difference between means extends from -4.267 to 0.2670.

Computations for Unequal Sample Sizes (optional)

The calculations are somewhat more complicated when the sample sizes are

not equal. One consideration is that MSE, the estimate of variance, counts the sample with the larger sample size more than the sample with the smaller sample size. Computationally this is done by computing the sum of squares error (SSE) as follows:

$$SSE = \sum (X - M_1)^2 + \sum (X - M_2)^2$$

where M_1 is the mean for group 1 and M_2 is the mean for group 2. Consider the following small example:

Table 4. Example Data

Group 1	Group 2
3	2
4	4
5	

$M_1 = 4$ and $M_2 = 3$.

$SSE = (3-4)^2 + (4-4)^2 + (5-4)^2 + (2-3)^2 + (4-3)^2 = 4$

Then, MSE is computed by: MSE = SSE/df

where the degrees of freedom (df) are computed as before:
$df = (n_1 - 1) + (n_2 - 1) = (3-1) + (2-1) = 3$.
$MSE = SSE/df = 4/3 = 1.333$.

The formula

$$S_{M_1 - M_2} = \sqrt{\frac{2MSE}{n}}$$

is replaced by

$$S_{M_1 - M_2} = \sqrt{\frac{2MSE}{n_h}}$$

where n_h is the harmonic mean of the sample sizes and is computed as

follows:

$$n_h = \frac{2}{1/n_1 + 1/n_2} = \frac{2}{1/3 + 1/2} = 2.4.$$

and

$$S_{M_1 - M_2} = \sqrt{\frac{(2)(1.333)}{2.4}} = 1.054.$$

t_{CL} for 3 df and the 0.05 level = 3.182.

Therefore the 95% confidence interval is

Lower Limit = 1 - (3.182)(1.054)= -2.35

Upper Limit = 1 + (3.182)(1.054)= 4.35

We can write the confidence interval as:

$-2.35 \leq \mu_1 - \mu_2 \leq 4.35$