Comparing Means: The t-Test

Katherine Dorfman
UMass biology Department, 2019

The Mean (Average).

This is probably the most common measure of central tendency. It is calculated by dividing the sum of all the data values by the number of such values:

$$\text{mean} = \bar{x} = \frac{x_1 + x_2 + x_3 \cdots x_n}{n}$$

Excel will calculate the mean for you with the following formula:

=average(*data array***).**

data array: You type in (or click and drag over) the address of the data you want averaged.

Function wizard: You can also use the function wizard (f_x) to calculate the average: pick a cell to contain the average, click on *fx*, choose AVERAGE (you may have to hunt for it in the statistical menu), highlight the values to average, and click OK. The cell will contain the formula given above.

The Standard Deviation

The standard deviation is one of the most commonly used and easiest to understand measures of spread. It also has some nice properties that will be described below.

The standard deviation is something like the average of all the individual deviations from the mean. This is a tedious calculation to do, so we usually ask a computer to do it for us (although generations of students before you managed with nothing more than calculators, and slide rules before that).

The calculation is done as follows: each datum is subtracted from the mean of all data (these are the individual deviations). About half of these deviations will be negative, and half positive, and if you add them together, they cancel each other out. To correct this, the deviations are squared so they will all be positive. These squares are added together, and their sum is divided by the number of data (actually, one less than the number of data – sorry) to get the "average". Finally, the square root of this average is taken to correct for the squaring done earlier.

$$sd = \sqrt{\frac{\sum(\bar{x} - x_i)^2}{n-1}}$$

Where

sd is the standard deviation,

n is the number of data,

x_i is each individual measurement,

$\bar{x}$ is the mean of all measurements, and

Σ means the sum of

Excel will calculate the standard deviation for you with the following formula:

=STDEVA(*data array***).**

One nice feature of the standard deviation alluded to above is that it is measured in the same units as the mean, so it is meaningful to add it to or subtract it from the mean.

A second nice feature is that when the data are distributed symmetrically around the mean (that is, when they fall into the famous bell curve of song and legend), between one standard deviation above the mean and one below are found 68% of the data, and between two deviations above and two deviations below the mean are found 95% of the data. This property is illustrated in Figure 1.

The bigger the standard deviation, the wider the spread of data around the mean, as illustrated in Figure 2.

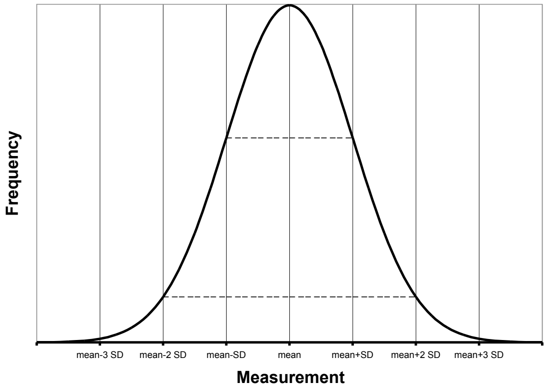


Figure 1. A “bell curve”, showing the symmetrical distribution around the mean. Horizontal lines indicate 1 and 2 standard deviations from the mean.

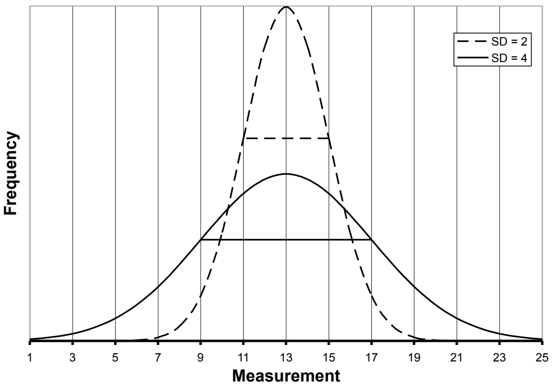


Figure 2. Two “bell curves”, with the same mean (13), but different standard deviations. Horizontal lines indicate 1 standard deviation from the mean.

Comparing the Means.

Usually, even when the means of two groups differ, there is some overlap between the two distributions. How different the two groups “really” are depends, therefore, not only on the difference between their means, but also on the extent of the overlap between their distributions.

Illustrating the relationship.

The standard deviation can help to make a more complete comparison between two sets of data. (Here’s where another nice feature of the standard deviation comes into play: it is in the same units as the mean, so they can be added together.) Figure 3 shows why it is important to report not only the difference in the means between your two groups, but also some measure of the variation in each one. The means of groups 1 and 2 differ from each other by the same amount as do 3 and 4, yet the error bars, illustrating the size of the standard deviation, indicate a much greater degree of overlap between groups 3 and 4.

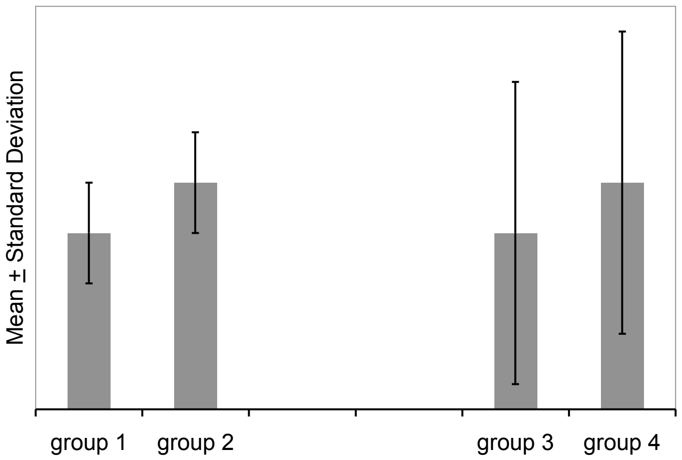


Figure 3. The effect of standard deviation on a comparison between means. Approximately two thirds of the values in each group lie within the one standard deviation error bars. There is much greater overlap between the measurements in groups 3 and 4 than between those in 1 and 2.

The easiest way to create a histogram comparing the means, with standard deviation as the error bars, requires making a little table in Excel that looks like this:

	mean	St dev
control		
experimental		

Choose appropriate descriptive labels for your two groups, as these will appear in your graph. Type the formula for the mean and standard deviation in the appropriate cells, i.e., =average(data array) and =stdeva(data array), respectively.

Highlight the cells containing the labels and the averages, shown above with a double outline, and use the chart wizard to make a bar graph. Then highlight the bars and format the data series. Make *custom* Y-error bars, using the two cells in which you have calculated standard deviation. ***Note: the "standard deviation" choice inside the Y-error bar dialog box is not what you think it is! Don't use it! You must calculate the standard deviation in a cell.*** (The dialog box has no idea which data went into making the means you are graphing, and cannot possibly calculate their standard deviation. Instead, it calculates the standard deviation of the values you are plotting – in this case the control mean and the experimental mean, and plots an error bar that length starting from the average of those values. I don't know when this would be useful.)

Quantifying the relationship: calculating "t".

The statistic "t" is a measure of the difference between two means, divided by the geometric mean of the standard errors of the population means (a sort of average of the standard deviations of the two populations). (The manner of calculating t depends on various characteristics of the experiment and the data, so this is why you must specify a "test type" before you ask Excel to calculate t for you.)

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}}$$

The value of t gets larger as the difference between the means gets larger; but this is counterbalanced by this measure of spread in the denominator. The greater the standard deviations, or the smaller the sample sizes (n), the bigger a difference in means is required to make t large. You can think of this as a ratio of signal (the difference between the means) to noise (the variation within the population).

Figure 4 illustrates the effect of standard deviation on the t statistic. The same difference between means can be significant or not, depending on the amount of variation in the populations being compared.

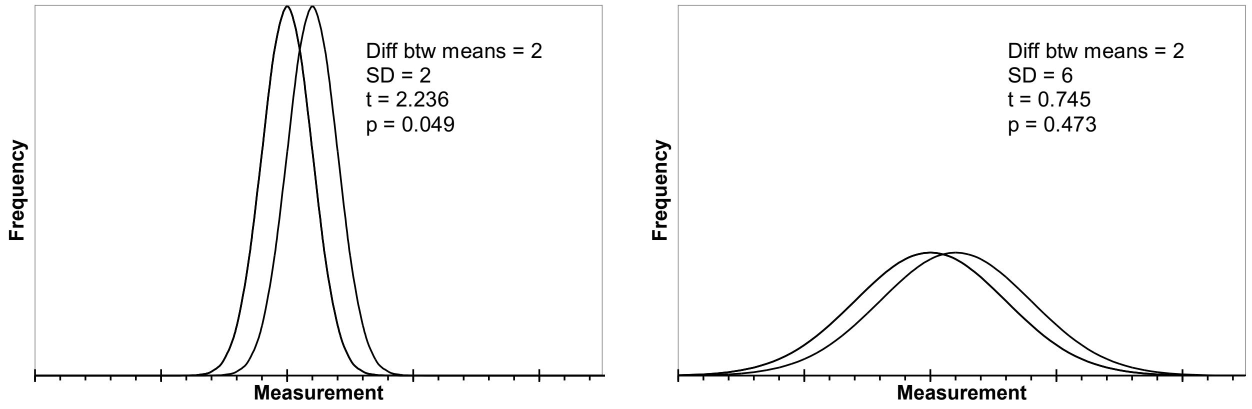


Figure 4. The effect of standard deviation on the t statistic. The means of both pairs of distributions differ by the same amount, yet the t statistic is 2.236 for the pair with the smaller standard deviations, and only 0.745 for those with the larger standard deviations. Notice the greater overlap between the curves on the right. (T values calculated for population size of 10 in each group.)

Interpreting the relationship.

Tests of significance always set up a straw man, the *null hypothesis*, which is that there is no difference between the groups, or no relationship between the variables. Then we find out how likely we are to get a result like our actual result if the null hypothesis were true. From this, we decide to accept or reject the null hypothesis.

If the category in question (*e.g*., control or treatment) had no effect on the variable being measured (*e.g*., number of cells in a suspension), you might as well be assigning categories at random. Imagine putting stickers on tubes of cells at random: some get labeled "group 1", others get labeled "group 2". Occasionally, such a process would result in all the tubes with dense populations being labeled "group 1", and all those with very few cells, "group 2". Such a result would give a rather large t.

How often would a t as large as the one you got in your actual experiment be expected to occur if the stickers were put on at random? The probability of getting a t as large as or larger than yours is called 'p', and fortunately, we can look up the p for a given t and number of measurements in any statistics book, on the internet, and inside of Excel.

If a t like yours is rather likely to occur by the chance procedure, you cannot rule chance out, and you must accept that there is no difference between the means of the two groups.

However, if such a t as yours would be exceedingly ***un***common in a chance process, you may reject the idea that there is no difference between the means, and conclude that there is a *statistically significant difference* between the two groups. The statistic tells you how often a result like yours might occur by chance alone; *it cannot tell you the probability that chance actually caused the difference you observed.*

Conventionally, if a t as big as or bigger than yours can be expected to occur less often than 5% of the time ($p < 0.05$) when there is nothing other than chance acting, we reject the null hypothesis, and conclude that there is a difference between the groups.

By this criterion, the two distributions illustrated on the left in Figure 4 are statistically significantly different from each other, while those on the right are not.

Calculating t with Excel (the quick and dirty method)

If you have both sets of data to compare, use the built-in t-test in Excel (which you can find the statistics category of the function menu:

=ttest(*data array 1, data array 2, number of tails, test type***)**

data array 1: the first set of data (enter the addresses, or click and drag over them)

data array 2: the second set

number of tails: 2 if you can't predict how one group will differ from the other, and think the means might vary in either direction, 1 otherwise. (Use 2 most of the time.)

test type: 1, 2, or 3

1 for paired data *i.e.,* two measurements on the same thing. In this test type, Excel takes the differences between the paired measurements, rather than between each one and the mean of the population. This works if you are comparing, for example, the absorbance at two different times *of the very same culture* (as opposed to the absorbance at the same time in two different cultures).

2 for unpaired data, but where both sets have the same standard deviation (don't use this one – it is unlikely to have two sets of data with identical standard deviations).

3 for unpaired data, with unequal standard deviations (two separate sets of measurements, such as number of cells in treated cultures vs. number of cells in control cultures). This is the most likely situation.

Excel will return the value of p that goes with this, but not tell you the actual t. In other words, Excel gives you just the punch line: what is the probability of grabbing two handfuls of data at random from a single population and winding up with two subsets as different from each other as your two data sets are from each other.

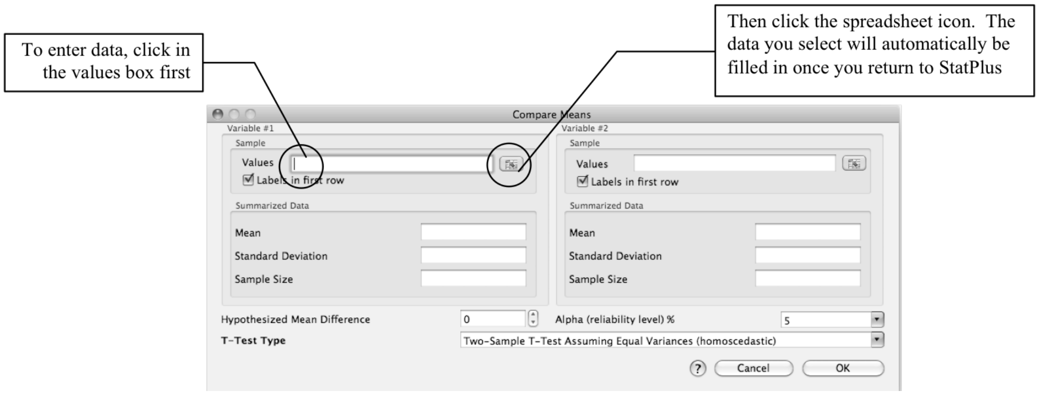
If you don't have both sets of data (if you are comparing your measurements to a standard reported in the literature only as mean plus or minus standard deviation), you have to do more work. Ask your instructor for guidance.

Calculating T With StatPlus.

If StatPlus is in the dock, open it by clicking on it.

If it isn't in the dock, open it from the applications folder. Doing so will put an icon for it ▦ in the dock, and put its main menu bar (StatPlus Spreadsheet Statistics Data Charts Help) at the top of the screen.

Choose Basic Statistics and Tables from the Statistics menu, and Comparing Means (t-test) from that menu. That should pull up a dialog box that looks like the one below.



To enter data, click in the values box first

Then click the spreadsheet icon. The data you select will automatically be filled in once you return to StatPlus

Compare Means

Variable #1

Sample

Values

☑ Labels in first row

Summarized Data

Mean

Standard Deviation

Sample Size

Variable #2

Sample

Values

☑ Labels in first row

Summarized Data

Mean

Standard Deviation

Sample Size

Hypothesized Mean Difference: 0

Alpha (reliability level) %: 5

T-Test Type: Two-Sample T-Test Assuming Equal Variances (homoscedastic)

To enter your data, click in the Values box, then click on the little spreadsheet icon, which should take you to your spreadsheet if it is open, or to a new Excel file. From that new blank file, you can open your data file. Select the data for the first variable, return to StatPlus, and repeat for the second variable.

Check the settings:

If you included labels in the selection make sure StatPlus knows there are labels in the first row.

Leave the summarized data blank.

Set the hypothesized mean difference to zero, (because the null hypothesis is that the two collections of data were chosen at random from the same population) and

Set the alpha (the maximum probability at which you will reject the null hypothesis) to 5% (= 0.05) by convention.

For T-Test Type, choose the two-sample test assuming unequal variances (heteroscedastic). You are not comparing paired data, and your two sets of values almost certainly do not have the same variance (standard deviation squared).

When you click OK, it chugs for awhile, then opens a new file called StatPlusMacResults.xlt, with something like this on a tab called Comparing Means:

Comparing Means [t-test assuming unequal variances (heteroscedastic)]			
Descriptive Statistics			
VAR	*Sample size*	*Mean*	*Variance*
	9	6.77778	1.94444
	8	9.375	1.125
Summary			
Degrees Of Freedom	15	*Hypothesized Mean Difference*	0.E+0
Test Statistics	4.34884	*Pooled Variance*	1.56204
Two-tailed distribution			
p-level	0.00057	*t Critical Value (5%)*	2.13145
One-tailed distribution			
p-level	0.00029	*t Critical Value (5%)*	1.75305
G-criterion			
Test Statistics	#N/A	*p-level*	#N/A
Critical Value (5%)	#N/A		
Pagurova criterion			
Test Statistics	4.34884	*p-level*	0.99938
Ratio of variances parameter	0.60573	*Critical Value (5%)*	0.0255

In this output table,

VAR: Probably short for *variable*. This should have the labels from your two columns of data. If not, type them in before you forget which comparison this is.

Sample Size: Sample size tells you the number of values in each group. Use it to do a reality check: do the sample sizes reported by StatPlus match the number of values in each of your categories? If not, you should repeat your data selection.

Mean: Is this the same as what you had Excel calculate for you on your spreadsheet? (Another reality check.)

Variance is standard deviation squared. It is another measure of the spread of data around the mean.

Degrees of freedom is the total number of data points minus 2 (it's the number of data points that are free to vary before the remaining ones are set).

The **Hypothesized Mean Difference** should be 0 (given as 0.E+0 in Excel-style scientific notation).

Test Statistics is your t value.

The Pooled Variance is an estimate of variance that assumes the true variance of the two samples is the same. You can ignore it.

The p-level (the probability of getting results this different from each other if the two sets of values were really drawn from the same pool) is given twice, once for a one-tailed distribution, and once for a two-tailed distribution.

In general, you should use the p value for the **two-tail** test, since this tests for both positive and negative differences between the means. The one-tail test gives you greater power to detect a difference in only one direction, but increases the possibility of getting a false indication of significant difference.

The **t Critical value (5%)** is the t associated with that probability. If your t is greater than the critical value, your p-value is less than 0.05. Remember, the bigger the difference between the two groups, the higher the t and the lower the probability of getting a t that big by random factors alone.

It is OK for our purposes to ignore the **G** and **Pagurova criteria**.

This output table is static. If you change anything about your data, you must re-do the analysis.

Excel reports the values with up to nine decimal places. ***This in no way obligates you to do the same!*** Round the values to the number of significant digits of your least precise measurement. You can do this by formatting the cells in Excel, or correcting the number when you paste a result into Word.

Multiple comparisons – a caution

What if you have more than two sets of data to compare (e.g., two experimental treatments and one control)? It is tempting to do multiple t-tests to find out which means are different from each other. However, this should only be undertaken with great caution, because the more comparisons you make, the more likely you are to find high t-values by chance alone. Remember, the p-value is a calculation of the likelihood of getting a t as high as yours even if the values in both groups were actually drawn from the same population (*i.e.*, by chance), so if you did 100 comparisons, you can predict that chance alone would produce about 5 "significant" results. See Adjusting the critical α, on p. 10.

Analysis of Variance

One improvement on multiple comparisons is to begin with an analysis of variance (ANOVA) on all the data first, to find out whether the variation between groups differs from the variation within each group. ANOVA tests the null hypothesis that there are no differences between the groups, that in fact, all the data are pulled from what is effectively a single population.

ANOVA uses the F-statistic, which is the ratio between the mean square (a measure of variation) between groups and the mean square within groups. The null hypothesis is that the variation between groups is no bigger than the variation within groups, or that $F \leq 1$. As with the t-test, a large F statistic occurs rarely when the groups are the same, so the associated p-value is a measure of how likely it is to get an F as large as yours by chance alone.

Use the one-way ANOVA in StatPlus, assuming you have only one variable (e.g., "treatment"), which consists of different conditions (e.g., "control", "treatment 1", "treatment 2", etc.). Set your data up in columns, with labels in the first row.

Select one-way ANOVA from the StatPlus Statistics menu, click in the "variables" box, click on the spreadsheet icon,find your datafile, and select your data including the labels. Let it run. It will produce a spreadsheet with a table like this:

Analysis of Variance (One-Way)

Summary

Groups	*Sample size*	*Sum*	*Mean*	*Variance*
control	7	20.	2.85714	1.80952
treatment 1	7	4.	0.57143	0.28571
treatment 2	7	9.	1.28571	0.57143

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*	*F crit*
Between Groups	19.14286	2	9.57143	10.76786	0.00084	4.90007
Within Groups	16.	18	0.88889			
Total	35.14286	20				

In this output table:

Groups should accurately label your sets of data. If not, repeat your data selection.

Sample size should accurately reflect the number of data points in each group. If not, repeat, with correct selection of data from your spreadsheet.

Sum should be what you get if you add up all the values in one group. The **Mean** should be the average of the data in the group, and **Variance** is standard deviation squared.

SS stands for the sum of (deviation) squares. It should remind you of the standard deviation. Squaring the difference between the individual values and the mean or between the group means and the overall mean gets rid of the negative values and magnifies the impact of a large deviation. $SS_{between}$ is the sum of the squares of the differences between each group mean and the total mean, times the number of scores in the group, and SS_{within} is the sum of the squares of the differences between every raw score and its sample mean.

$$SS_{between} = \sum(\bar{x}_i - \bar{x}_{total})^2 N$$

where

$\bar{x}_i$ = the mean of group *i*
$\bar{x}_{total}$ = the mean of all raw scores combined
N = the number of scores in group *i*

$$SS_{within} = \sum(x_i - \bar{x}_i)^2$$

where

x_i = raw score *i*
$\bar{x}_i$ = the mean of *i*'s group

Df, or degrees of freedom, is the number of values that are free to vary and still give you the same statistic. Within groups, this should be 1 less than the number of groups. Between groups, this should be the total number of samples minus the number of groups.

MS stands for mean square, or variance, and it is the ratio between SS and df. This yields something like an average, and corrects for the number of scores.

F is the ratio between $MS_{between}$ and MS_{within} ($\frac{MS_{between}}{MS_{within}}$). The null hypothesis is that the variation between groups is no bigger than the variation within groups, or that $F \leq 1$. A larger F indicates greater variance between groups than within each group.

The **p-level** is the probability of getting an F as big as yours by chance alone, even if all the data come from the same population.

The **F crit** is the F associated with the p-value (probably 0.05) you specified when you set up the test. It is the smallest F that would let you conclude your groups differ statistically from each other.

If you get a statistically significant result (i.e., a p less than 0.05) with ANOVA, you can test the treatment groups in pairs by the t-test. Doing so, however, increases the chance of finding spurious “significance.” You can guard against that in various ways. The two simplest are to adjust the α (that is, the maximum p-value you will use to reject the null hypothesis of no difference between the groups), or caution your reader that you have performed multiple comparisons.

Adjusting the critical α

The probability of getting a large t rises with the number of t-tests you perform. Therefore, you should tighten your decision criterion a little. The easiest way is to correct the critical α by dividing your usual α (0.05 for most purposes) by the number of pairwise comparisons you made. This protects you very well against claiming statistical significance where none exists, but not so well against claiming no difference when there is one.

For small numbers of comparisons, this is nearly identical to the next simplest correction, calculated as $1-(1-\alpha)^{1/n}$. As the number of comparisons grows, this correction stops the α from slipping into infinitesimal territory. This helps prevent you from being unable to detect any differences at all.

You can be very confident of any difference that passes either of these corrected α tests.

There are other ways to guard against spurious significance that are beyond the scope of this chapter, but which a statistician would be delighted to teach you.

Reporting Your Statistical Results.

Use the criterion above, that is, a significance level of $p \leq 0.05$, to decide whether the values in two sets of data can be called statistically different from one another.

Remember that the p value only tells you how likely it is to get a t as big as yours if the two samples really were picked from the same population, so resist the urge to say you’ve proved chance did or did not cause the difference you observe. Here are a few conventional expressions of such a result:

> The mean [whatever was measured] of Group 1 (5 ± 2, n = 10) is statistically lower than that of Group 2 (7 ± 2, n = 10) by the t-test (t = 2.236, p = 0.049).
>
> Group 2 had significantly higher [whatever was measured] by the t-test (t = 2.236, p = 0.049).
>
> There was no statistically significant difference between Groups 3 and 4 by the t-test (t = 0.745, p = 0.475).