# *How To Run Statistical Tests in Excel*

Microsoft Excel is your best tool for storing and manipulating data, calculating basic descriptive statistics such as means and standard deviations, and conducting simple mathematical operations on your numbers.  It can also run the five basic Statistical Tests.  It does have some limitations, however, and for certain tests you may have to turn to a more powerful statistical program like S-Plus or Minitab.

> <u>NOTE</u>:    The statistical tests are under the ***Tools menu… Data Analysis…***  If you do not see "Data Analysis" anywhere, you will have to "add in" the Analysis ToolPak, as follows: ***Tools… Add-Ins… Analysis ToolPak***.  If at RCC, your computer should find it via the network.  If at home, it will probably ask for your Microsoft Office CD.

## *Alert!  The example Data Sets given below were fabricated to fit the example experiments described in "Statistical Testing For Dummies"*

## **Data Organization and Descriptive Stats**

Initially you'll want to organize your raw data by ***treatment groups***, each in its own column, as shown below.  Later, however, for certain tests you'll have to "stack" the columns (e.g., for Regression and Two-Way ANOVA).  This is easy to do in Excel by copying and pasting cells.

|  | Untrimmed | | | Trimmed | | |
|---|---|---|---|---|---|---|
|  | High Marsh | Mid Marsh | Low Marsh | High Marsh | Mid Marsh | Low Marsh |
| Raw Data | 12 | 9 | 7 | 6 | 7 | 6 |
|  | 15 | 8 | 12 | 2 | 3 | 1 |
|  | 7 | 16 | 15 | 7 | 5 | 8 |
|  | 3 | 5 | 4 | 8 | 9 | 6 |
|  | 11 | 13 | 10 | 4 | 4 | 3 |
|  |  |  |  |  |  |  |
| *N* | *5* | *5* | *5* | *5* | *5* | *5* |
| *Mean* | *9.6* | *10.2* | *9.6* | *5.4* | *5.6* | *4.8* |
| *Std Dev* | *4.7* | *4.3* | *4.3* | *2.4* | *2.4* | *2.8* |
| *S.E.* | *2.1* | *1.9* | *1.9* | *1.1* | *1.1* | *1.2* |

After organizing your raw data this way, you'll want to calculate Descriptive Statistics for each column.  Excel has a readymade function for each of these except the last.  Use "COUNT" for Sample Size (N), "AVERAGE" for the Mean, and "STDEV" for the Standard Deviation.  The final stat is the ***Standard Error in the Mean***, which you calculate simply as the <u>standard deviation</u> divided by the <u>square root</u> ("SQRT" in Excel) of the <u>sample size</u>:

$$\text{S.E.  = Std Dev} / \sqrt{N}$$

**This is an important stat, as it's probably what you'll use for <u>Error Bars</u> on your graphs!**

*Hey!  Don't forget the "little black box" trick!  Once you plug in all the stat formulas under the first data column, you can simply highlight those cells, grab the little black box in the lower right corner, and drag to the right.  It carries the formulas across!*

## Standard t-test

1. Running this test is easy.  Excel wants your data in two columns, one for each group or treatment level.  Give each column a heading.  See example to the right.

| Control | Experimental |
|---------|--------------|
| 12 | 18 |
| 9 | 24 |
| 14 | 15 |
| 20 | 19 |
| 17 | 19 |
| 11 | 13 |
| 10 | 22 |
| 14 | 20 |

2. Under the *Tools* menu select *Data Analysis…* and choose "t-Test: Two-Sample Assuming Equal Variances."  OK.

3. Excel asks you to specify the range of cells containing the data.  Click the first red, white, & blue icon, then highlight your first column of cells, *including its heading*.  Enter.  Now click the second red, white, & blue icon, and highlight your second column, including the heading.  Enter.

4. Check the **Labels** box, so Excel knows you included headings atop each column.  OK.

5. Excel whips out an Output table.  You can quickly resize the columns by double-clicking up top between the A & B, between the B & C, and between the C & D.  There's lots of info here, but all you're really after are those P-values.  Use the **two-tailed p-value** if your original hypothesis predicted that the means would merely be different ($\neq$).  Usually, however, you will have specifically predicted one mean higher than the other ($<$ or $>$).  In that case (and if in fact the means match your prediction of greater than or less than), go with the smaller **one-tailed p-value**.

## Paired t-test

1. You can use the powerful paired t-test if (and only if) your study employed a "paired" design in which a pair of data were collected in parallel from each individual, "mirror image" style  …such as left-versus-right or before-versus-after.  Here again, Excel wants your data in two columns, one for each treatment level.  Give each column a heading.

2. Under the *Tools* menu select *Data Analysis…* and choose "t-Test: Paired Two Sample for Means."  OK.

3. Excel asks you to specify the range of cells containing the data.  Click the first red, white, & blue icon, then highlight your first column of cells, *including its heading*.  Enter.  Now click the second red, white, & blue icon, and highlight your second column, including the heading.  Enter.

4. Check the **Labels** box, so Excel knows you included headings atop each column.  OK.

5. Excel whips out an Output table.  You can quickly resize the columns by double-clicking up top between the A & B, between the B & C, and between the C & D.  There's lots of info here, but all you're really after are those P-values.  Use the **two-tailed p-value** if your original hypothesis predicted that the means would merely be different ($\neq$).  Usually, however, you will have specifically predicted one mean higher than the other ($<$ or $>$).  In that case (and if in fact the means match your prediction of greater than or less than), go with the smaller **one-tailed p-value**.

| Portside | Starboard |
|----------|-----------|
| 537 | 570 |
| 241 | 234 |
| 77 | 84 |
| 427 | 411 |
| 220 | 282 |
| 96 | 92 |
| 625 | 700 |
| 395 | 450 |

**One-Way ANOVA (Single Factor ANOVA)**

1. Here, too, Excel wants your data in side-by-side columns, one for each group or treatment level. Give each column a heading.

2. Under the *Tools* menu select *Data Analysis…* and choose "ANOVA: Single Factor." OK.

3. Excel asks you for a single range of cells containing ALL the data. Click the red, white, & blue icon, then highlight all three (or more) columns of cells, *including their headings.* Enter.

4. Check the **Labels** box, so Excel knows you included headings atop each column. OK.

| Red | Yellow | Blue |
|-----|--------|------|
| 5.1 | 2.9 | 5.4 |
| 4.9 | 3.4 | 5.9 |
| 5.3 | 3.7 | 6.2 |
| 4.4 | 2.7 | 5.2 |
| 5.5 | 2.5 | 5 |
| 5.6 | 3.4 | 5.9 |
| 3.9 | 2.1 | 4.6 |
| 4.2 | 2.3 | 4.8 |
| 4.7 | 4.1 | 6.6 |
| 5.6 | 2.1 | 4.6 |

5. Excel whips out an Output table. You can quickly resize the columns by double-clicking up top between the A & B, between the B & C, etc. There's lots of info here, but all you're really after is that "Between Groups" p-value. All data is naturally variable …or "noisy." The ANOVA test attempts to detect a "signal" of genuine difference amidst all that "noise." More precisely, it partitions the natural variance <u>within</u> the groups (the noise) from the variance <u>between</u> the groups (the signal). If the differences <u>between</u> the groups are substantially greater than the differences <u>within</u> the groups, then we say that there's a strong "signal-to-noise" ratio. And the stronger the signal-to-noise ratio, the lower the p-value!

**Important Note!** All an ANOVA test can tell you is whether there are statistically significant differences *somewhere* in the data as a whole. But it cannot tell you *just where* those differences lie. For example, run an ANOVA on the data above, and you'll get a very low p-value. This means that the independent variable (color of light) <u>does</u> affect the response variable (phytoplankton growth). But it doesn't tell you *which* colors affect growth differently from *which other* colors. You can plainly see that the yellow mean is different from the red and blue means, thus giving us our low p-value. But are the red and blue means different from each other (at 95%+ confidence)??? The ANOVA itself can only tell you that at least *one* group in there is different from *some other* group in there …but not which ones. Therefore **IF** (and only if) your Between Groups p-value falls below 0.05, then you will want to run a second test called a **"Multiple Comparisons"** test (like Tukey's test) in order to pinpoint just where the real differences lie. Unfortunately, this is something that Excel can't do for you, so you will have to turn to some other program such as S-Plus or Minitab. Consult teacher for help.

**<u>Linear Regression</u>**

| Depth (X) | Fish (Y) |
|-----------|----------|
| 1 | 43 |
| 1 | 55 |
| 1 | 58 |
| 1 | 79 |
| 1 | 53 |
| 1 | 49 |
| 2 | 60 |
| 2 | 33 |
| 2 | 44 |
| 2 | 39 |
| 2 | 41 |
| 2 | 50 |
| 3 | 38 |
| 3 | 34 |
| 3 | 19 |
| 3 | 29 |
| 3 | 24 |
| 3 | 31 |
| 4 | 18 |
| 4 | 16 |
| 4 | 5 |
| 4 | 25 |
| 4 | 17 |
| 4 | 19 |
| 5 | 0 |
| 5 | 2 |
| 5 | 7 |
| 5 | 4 |
| 5 | 0 |
| 5 | 5 |

1. To run a regression, you first need to stack your data as shown to the right. Independent variable goes on the left; response variable on the right. This probably isn't the way you originally arranged your data, but it's easy to stack it by copying and pasting. In effect, your setting your data up in ordered pairs (X,Y).

2. Under the ***Tools*** menu select ***Data Analysis…*** and choose "Regression." OK.

3. Excel asks you for a two ranges of cells, one containing the "Y" values (i.e., your response variable), and one containing the "X" values (i.e., your independent variable). Click each red, white, & blue icon, then highlight the appropriate columns of cells, *including their headings.* Enter.

4. Check the **Labels** box, so Excel knows you included headings atop each column. Also, check the **Line Fit Plots** to generate a graph of your data and a best fit line. OK.

5. You can quickly resize the columns by double-clicking up top between the A & B, between B & C, etc. There's lots of info here, but only four pieces of interest to you:

   o The **slope coefficient** (identified by the response variable; in this case "Depth") and the **intercept coefficient**. These respectively correspond to the slope (m) and the y-intercept (b) of your best fit line, and you can plug them into $y = m\,x + b$ to get the equation of that line.

   o The **p-value** for the *slope* (<u>not</u> the p-value for the y-intercept, which you usually don't care about). If $p < .05$, then you can reject the null hypothesis that the independent variable has no effect on the response variable. After all, a positive or negative slope is what you were after, and *the steeper it is, the stronger the relationship*.

   o The **R-Square value**. This is a number ranging from 0 to 1, and is a measure of *how tightly* your data points fit the best fit line. An R-square of 1.0 is a perfect fit, with every point falling right on the line, and zero means there's absolutely no pattern or fit whatsoever. In the example here, the regression returns an R-square of 0.86, or 86%. A scientist would say that the independent variable (depth) "explains 86%" of the variation in the response variable (fish).

6. Excel also gave you a graph of the data and the best fit line, but it's probably all scrunched together. Grab a corner and drag to make it bigger. To widen your plot even more, go ahead and delete the legend (click it, then hit delete). Finally, double click one of the best fit points (probably pink), then give it a solid line under the Patterns tab. How's it look?

*Note:* The data above come from a <u>replicated</u> experiment where fish were *repeatedly* sampled at a handful of depths at fixed, <u>regular</u> intervals. Regression also works fine when the treatments are spaced at <u>irregular</u> intervals. For example, the study might have instead used depths of 1, 2, 4, & 7. And you can also use a regression to analyze data from a <u>non-replicated</u> study. Suppose you're interested in whether fiddler crabs avoid the edges of a marsh due to the threat of predation. You count the number of burrows per square meter at *randomly* chosen distances from the waterline. You can now run a regression to see if there's a statistically significant correlation here. Once again, just stack your data in X-Y pairs, as in the table to the right.

| Dist to Edge (X) | Crab Burrows (Y) |
|---|---|
| 4.7 | 3 |
| 18.6 | 6 |
| 17.9 | 7 |
| 7.7 | 4 |
| 18.7 | 6 |
| 21.7 | 7 |
| 11 | 4 |
| 4.7 | 0 |
| 22.3 | 6 |
| 4.5 | 2 |
| 12.2 | 5 |
| 20.5 | 6 |
| 0.2 | 1 |
| 12.6 | 4 |
| 1.1 | 3 |
| 1.7 | 1 |
| 24.4 | 8 |

## Two-Way ANOVA

1. To run a Two-Way ANOVA, you first need to organize your data as shown to the right, with one independent variable's treatments across the top, and the other IV's treatments stacked atop one another. (Note: The numbers are staggered horizontally within the cells here – some left, some centered, some right – for visual purposes only; this is <u>not</u> something you have to do in order to run the test…)

2. Under the *Tools* menu select *Data Analysis…* and choose "ANOVA: Two-Factor With Replication." OK.

3. Excel asks you for a single ranges of cells containing your data. Click the red, white, & blue icon, then highlight ALL the cells containing your data, *including the labels and headings*. Enter.

|  | Untrimmed | Trimmed |
|---|---|---|
| High Marsh | 12 | 6 |
|  | 15 | 2 |
|  | 7 | 7 |
|  | 3 | 8 |
|  | 11 | 4 |
| Mid Marsh | 9 | 7 |
|  | 8 | 3 |
|  | 16 | 5 |
|  | 5 | 9 |
|  | 13 | 4 |
| Low Marsh | 7 | 6 |
|  | 12 | 1 |
|  | 15 | 8 |
|  | 4 | 6 |
|  | 10 | 3 |

4. In the **Rows per sample** box, enter your sample size *per group*. In the example here, N = 5. Note: to run a 2-way ANOVA in Excel, you must have "balanced" data, meaning that very group has the same number of numbers (no NA's). If your data is unbalanced, consult your teacher.

5. OK. Excel kicks out lots of info. What you're mainly after are the p-values down at the bottom. There are three of them. The "Sample" p-value tells you whether or not there are statistically significant differences between levels of the your first IV – the one you have organized horizontally by rows …in this case, High vs. Mid vs. Low

marsh. The "Columns" p-value tells you whether or not there are statistically significant differences between levels of the your second IV – the one you have organized vertically in columns …in this case, Untrimmed vs. Trimmed grass. The "Interaction" p-value tells you whether there was a statistically significant interaction between the two IV's. This is one of the great things about a 2-way ANOVA: it not only can analyze the influence of each IV on the RV, but also can sniff out *interactive* effects between those two IV's. For example, does trimming the grass affect the snails more at high elevation than it does at low elevation? Does elevation have more of an effect on snail density for trimmed grass than for untrimmed grass? The ability to detect an ***Interaction*** is one of the most powerful advantages of a 2-way experimental design.

*Important Note!* All an ANOVA test can tell you is whether there are statistically significant differences *somewhere* in the data. But it can't tell you *just where* those differences lie. **IF** (and only if) either your Sample or Columns p-value falls below 0.05, and if you have 3 or more treatment levels under each IV, then you will want to run a second test called a Multiple Comparisons test in order to pinpoint just where the real differences lie. Unfortunately, this is something that Excel can't do for you, so you will have to turn to some other program such as S-Plus. Consult teacher for help.