# Using Excel to do Statistical Tests and Make Graphs

## How to organize data

In Biology 180, you often will be working with a large "master" data set and will have to choose the variables and values that you want to compare. To do this, you'll need to copy the appropriate data from the large data set into a separate worksheet. Click and drag the cursor over the data to select it, then choose Copy from the Edit menu or press command-C on the keyboard. Go the tabs at the bottom-left of the master data worksheet, and click on one of the tabs to the right to open a blank worksheet. To paste the data you've copied, select the cell that will be the upper left point and choose Paste from the Edit menu or press command-V.

To sort data in Excel, highlight the data you want to work with. Then go to the Data menu and choose Sort; make appropriate selections in the Sort dialog box. Note that you can also sort by the values in 1 or 2 other columns if you wish, meaning that you could put all the male data together in numerical order and then all female data together in numerical order.

Your hypothesis probably falls into one of two basic categories: You may be interested in 1) comparing different treatments (e.g. gender, or country born), or 2) the relationship between continuous variables (e.g. age or number of housemates). In either case, you will copy the appropriate data into a new spreadsheet. To compare continuous variables, put the values for the independent (explanatory) variable in one column and the values for the dependent (response) variable in an adjacent column. To compare two treatments, create a left-hand column that identifies the treatment groups (e.g. High or Low, in a study of whether high- or low-sugar intake affects weight gain) and a column next to it, on the right, that contains the values (e.g. the amount of weight gained by each individual in the study).

## Doing *t*-tests and making Bar Graphs in Excel

Your data should be sorted so that values from the two groups you want to compare are in adjacent columns, with the first row containing a word or phrase identifying each variable. The following instructions will assume that the **data** is in **columns A and B**.

1. Set up a table for the t-Test:

   A. In cell D3, type "t-Test: Two-Sample Assuming Unequal Variances".

   B. In cells E4 and F4, type headings that describe the data groups. These instructions will use "Male" and "Female" as examples.

   C. In cells D5, D6, D7, and D8, type "Mean", "Observations", "Standard Error", and "P-Value (two-tailed)".

   D. You will have a table like this:

| t-Test: Two Sample Assuming Unequal Variances | | |
|---|---|---|
| | Male | Female |
| Mean | | |
| Observations | | |
| Standard Error | | |
| P-Value (two tailed) | | |

2. Enter in the following formulas *(Note: because all the numerical values in the columns are grades, the entire column is referenced in these formulas)*:

   A. E5 = AVERAGE(A:A)

   B. E6 = COUNT(A:A)

   C. Copy these formulas into the "female" column by selecting cells E5:E6 and dragging the bottom-right corner to column F.

   D. E8 = T.TEST(A:A,B:B,2,3). *"2" refers to two-tailed, and "3" refers to Two-sample unequal variances.*

   **What does the p-value tell you?**

**To make a bar graph of means with standard error bars:**

1. To make the graph, select cells E4:F5 (means and labels), then choose Charts from the green ribbon. Click on the Column Chart Type and choose Clustered Column and a chart should appear in the middle of the worksheet.

   A. Add a Chart title and X and Y axes labels by clicking on Chart Layout in Purple region of the green ribbon and selecting …
   Chart Title > Title above chart, then provide an informative title
   Axis Titles > Horizontal Axis Title > Title below axis, then type axis label
   Axis Titles > Vertical Axis Title > Rotated Title, then type axis label

   B. Remove the legend and gridlines by clicking each and then pressing Delete. Remove shading and use outline bars to reduce printer toner consumption.

   C. If necessary, move the graph underneath the P-value calculated previously.

2. To add standard error bars to the graph, first determine the standard errors for the data.

   A. Using the fact that Standard Error is equal to the Standard Deviation divided by the square root of the number of observations, enter the following formula in cell E7:

   = STDEV.S(A:A)/SQRT(E6)

   B. Copy the formula by selecting E7 and dragging the bottom-right corner to F7.

3. Add standard error bars:

   Double-click on one of the bars in the graph

   In the Format Data Series dialog box:

   – Select the Error Bars menu item

   – Choose Display > Both, and End Style > Cap

   – Click on Error Amount > Custom radio button and then click on Specify Values

   – In the Positive Error Value box select cells E7:F7 (**both** standard error values)

   – Repeat for Negative Error Value box (both cells), then click OK. Click OK again to finish the Format Data Series dialog.

   **What do these error bars tell you?**

4. To display the *p*-value on the graph, create a text box as follows: Click any empty cell, then Insert>Text box. Move the cursor over the graph; it changes to an "A." Click and drag a text box; type the p-value.
5. You can change the formatting of the axes by double-clicking on each and altering values in the dialog boxes. Try different things until you have a graph that is easy to interpret.

## Doing Linear Regression analysis and making Scatter plots in Excel

Your data should be sorted so that values from the variables you want to analyze are in adjacent columns, with the first row containing a word or phrase identifying each variable. These instructions assume that the independent (or explanatory, or "x") variable is in column A, and the dependent (response, or "y") variable is in B. Copy your data (or hand enter it) into those columns in the Regression Analysis sheet of the data file provided in lab.

1. Functions for Slope, Intercept, and R-squared (RSQ) are built into Excel. The values for each will appear in column E.

   **What does the $R^2$ values tell you?**

2. Microsoft has intentionally crippled the new Macintosh version of Excel (that is, removed features present in older versions), so we've added the information necessary to calculate a *p*-value, in columns F and G.

   **What does the *p*-value tell you?**

**To make a scatterplot:**

1. Select all of the data and column titles in columns A and B. *One trick is to start with the bottom cells and scroll up.*
2. Choose Charts from the green ribbon. Scroll down to Scatter Chart Type and choose Marked Scatter and a chart should appear in the middle of the worksheet.
   A. Add X and Y axes labels by clicking on Chart Layout in Purple region of the green ribbon and selecting Axis Titles > Horizontal Axis Title > Title below axis, and Axis Titles > Vertical Axis Title > Rotated Title. Type in titles, and remember to state units for each axis.
   B. Change the graph title to an informative one.
   C. As before, you can remove the legend and gridlines by clicking each and then pressing Delete.
   D. If necessary, move the graph underneath the regression values calculated previously.
3. Add the best-fit line: Click on Chart Layout in Purple region of the green ribbon and select Trendline > Linear Trendline. Double-click on the trendline, and in the Options Menu, select "Display R-squared value on Chart." Click OK and move the displayed text to a clear portion of the graph.
4. **Add the *p*-value** as explained above for bar charts (click an empty cell, Insert>Text box, drag a box0.

   The fact that Excel provides an easy means of displaying the $R^2$ value, but not the *p*-value, suggests that the *p*-value is unimportant. Not true! **What does the *p*-value tell you?**

## Doing a Chi-Squared test in Excel

Excel provides a "CHIDIST" function that returns a *p*-value from a chi-squared sum and degrees of freedom you provide. Details will be explained in lab.