

# Intro to Stats Project Help Guide

## Using MS Excel to Build a Data Set & Perform Regression Analysis

This help guide will demonstrate ways to build a data set file in Microsoft Excel (both 2003 and 2007 versions) together with needed data analysis formulas and steps.

### Key Steps

- Enter each individual survey as a **single row** in the spreadsheet.
- Each **column** represents a single **variable**.
- Use as few symbols as possible for each entry.
- Verify your data entry for **accuracy**.

The help guide will follow data collected using the following survey. The team studied the relationship between “nights out per week,” “hours of exercise per week,” and “classes missed per week.”

<b>Demographics</b> (check all that apply):				<b>Current Class Status</b> (check one):						
Male	Female			Freshman						
In a fraternity or sorority?	Yes	No		Sophomore						
Campus Resident?	Yes	No		Junior						
Corps of Cadets/ROTC?	Yes	No		Senior						
How many total hours did you spend at the gym/working out in the last week? (circle one)										
0	1	2	3	4	5	6	7	8	9	10+
How many nights last week did you go out? (circle one)										
0	1	2	3	4	5	6	7			
How many class meetings have you missed in the last week? (circle one)										
0	1	2	3	4	5+					

The survey is not perfect. Some folks may have missed more than 5 class meetings, and varsity athletes and some others may work out more than 10 hours in a week. When forming closed-response questions, be sure to include a large range of responses in case of outliers, even if the responses seem unlikely to you. These data were collected by students, and project worked reasonably well.

Directions are given for both Excel 2003 and Excel 2007 menu structures. The formulas are identical, and wizards are very similar, in most cases identical.

### **Making a Data Set in MS Excel**

Open the example data file ([2003](#) or [2007](#)). Note that as you scroll down, the top row remains fixed in place. Please use one row at the very top of the sheet for the titles of your variables.





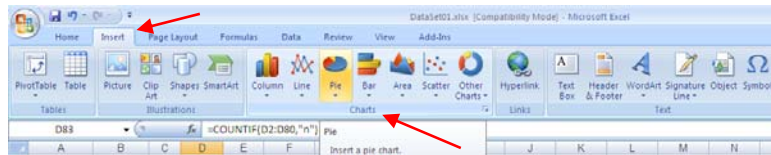
## Command Help: Chart Wizard

Drag-select Frequency Table

Click the INSERT menu, and  
select CHART.

Step through Wizard to format.

Excel 2007 has several “wizards” in the “Chart” area of the “Insert” tab (see below).



You can drag-select cells that non-contiguous using the CONTROL key. To make a pie chart with the ROTC Frequency Table, drag-select first column (the Yes / No cells plus the one above it). Press and hold the CONTROL key. Drag-select the ROTC cell and the two formula cells below it. (You must drag-select the same number of cells in each column.) Release the mouse. Then let up the CONTROL key. Perform “insert chart” as before, and the pie chart will exactly match the one you made for Greek-affiliation.

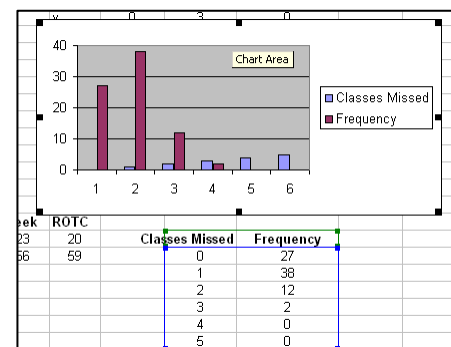
### Helpful Hints

1. Once you've made a chart or graph, right-click on any attribute. One of the options will be to FORMAT the attribute. Select this option and the wizard will open back up, allowing you to make any changes you wish.
2. If you are making side-by-side graphs to show the population compared to your sample demographics, you will need to make a Frequency Table for the population by hand, using the data from your university web site. Make sure you use the same color-coding for both graphs.
3. Technically, research graphs and charts are in black and white. Color can confuse the eye, making pieces of the graph appear either larger (darker, brighter colors) or smaller (lighter, pastel colors) than their actual area. Check with your professor to see if colors are allowed.

### Watch Out for Numeric Category Labels!

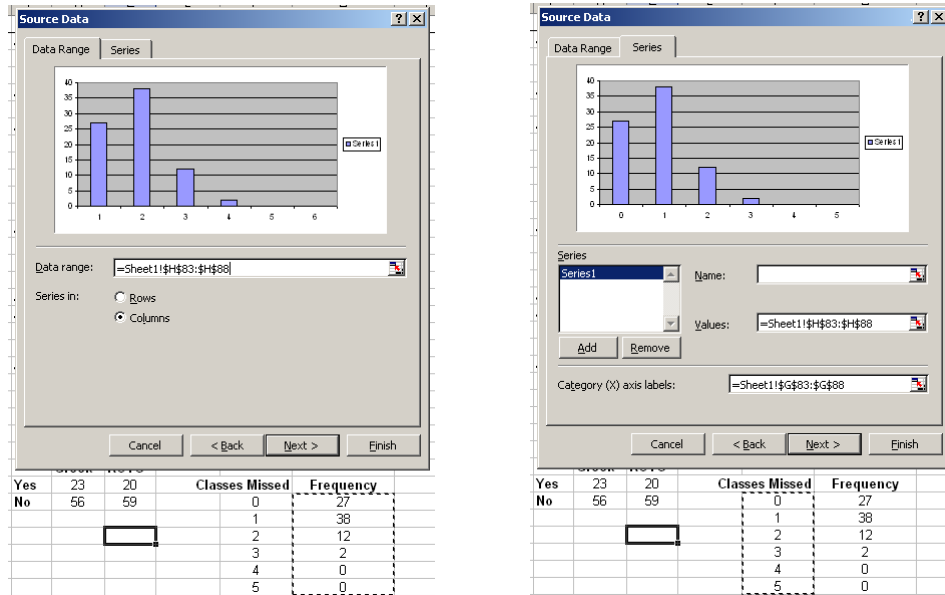
One common problem that occurs relates to Excel's habit of treating all numeric cells as data series. Let's make a Bar Graph for the “missed classes” Frequency Table. Drag-select the two columns, insert chart, and click FINISH. Instead of using the left-hand column for labels, Excel makes a side-by-side bar graph with two data series.

What we want is to use the Classes Missed column as x-axis labels in our chart. When we encounter problems with Excel's default graphs, the best approach is to start over using the “insert



chart” option first, with no data range pre-selected. Then we can move through the wizard, and “force” the correct columns of cells into the right parts of the graph. On the Source Data screen, drag-select the Frequency column for “Data Range.” Click the “Series” tab. The “Category (X) axis labels” field is at the bottom. Click in the white area next to it, then drag-select the Classes Missed column. The preview will adjust the bar graph with the correct labels. The steps are shown below.

An example file with a pie chart and bar graph is here: DataSet3 ([2003](#) or [2007](#)).



## Histograms vs. Bar Graphs

A histogram is a stylized bar graph with features that minimize possible visual distortions. Some features of histograms presented in research articles include:

1. Black-and-White graphics (no color)
2. The origin must be shown (Excel often “zooms on” on the “tops” of the bars, leaving out part of the y-axis).
3. All categories must have equal “bin width” or ranges.
4. Bars must be of equal width, and touching.

We suggest that our students simply reduce the “gap width” to zero. This creates a single-color bar graph with equal bar-widths and bars that touch. We also ask them to verify the y-axis includes the origin. While this is not an “authentic” histogram, it has most of the correct attributes and will avoid visual distortions.

### Helpful Hints

1. **Gap Width.** On a completed bar graph, right-click on any of the bars. Select “Format Data Series.” Click on the OPTIONS tab. Set the “gap width” field to zero, and click OK.
2. **Numeric Labels.** One workaround for numeric labels (see bar graph example above) is type “Zero,” “One,” “Two,” ... instead of numbers. Excel correctly interprets text as labels and allows you drag-select, then insert a chart, as with pie charts.

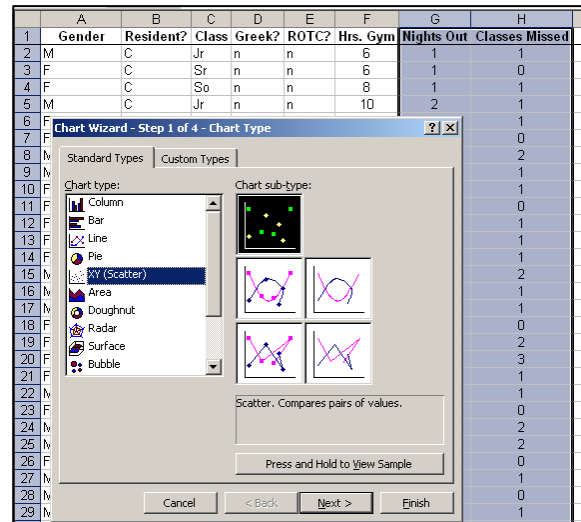
## Scatter Plots

Drag-select two columns. Excel's default setting forces the left-most column to be the x-variable. Perform Insert Chart, select the "XY (Scatter)" option, and move through the wizard. When you have finished your graph, you can right-click on any feature and modify it as desired. If the default x-variable is the opposite of what you want, start with Insert Chart and select the x-variable and y-variable individually.

In the example, the students were interested in predicting "classes missed" (dependent or y-variable) first using "nights out" and then with "workout hours" (independent or x-variable).

In the example shown at the right, start with the "nights out" variable since it is next to the "classes missed" column. Simply drag-select the two columns, but **do NOT include** the formulas and tabulations below the data. Insert chart, and follow the Chart Wizard through the steps.

The second scatter plot compares "workout hours" vs. "classes missed," which are non-contiguous blocks (must use CONTROL key). Drag-select the workout hours column, press and hold CONTROL, then drag-select the "classes missed" column. Release the mouse, release the CONTROL key, then Insert Chart.



DataSet4 ([2003](#) or [2007](#)) has both scatter plots completed and formatted.

### Helpful Hints

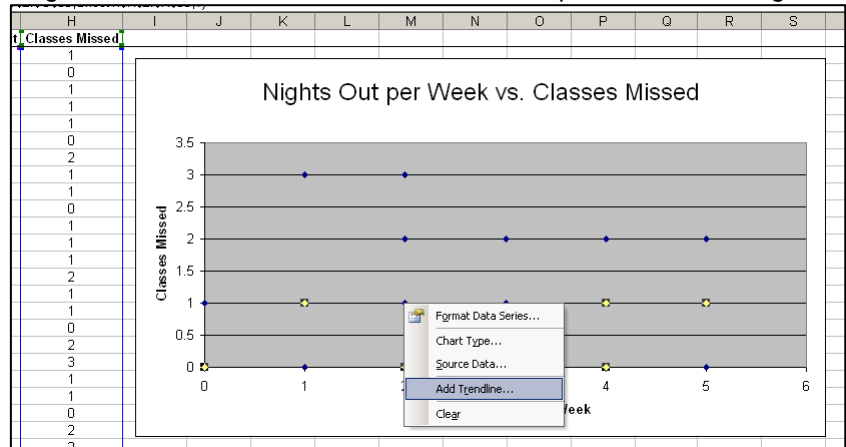
- Chart Area and Chart Options.** Right-click on the white space outside the plot. Use "Chart Options" to make x-axis and y-axis labels, a chart title and to format them. Use "Format Chart Area" option to change graph colors (background, foreground) and fonts. Remember to use color sparingly.
- Format Axis.** Right-click on the axis or axis labels. Select "Format Axis" option. You can "force" Excel to use ranges, values and tick marks you prefer.
- Explore with Right-Clicks.** If you don't like how your graph or chart looks, right-click on whatever piece bothers you and reformat. There are hundreds of options, not all of which appear in the main Chart Wizard.

## Regression

Excel has formulas to determine the regression variables, but a convenient chart option allows us to get nearly everything at once. Right-click on any data point. Select “Add Trendline” option. This will superimpose the Line of Best Fit on the scatter plot.

The first tab is “Type.” The default setting of “Linear” is correct – we’re performing a linear regression.

Click on the “Options” tab. At the bottom, check the two boxes marked “display equation on chart” and “display R-squared value on chart.” This will make a textbox with the equation for the regression line (e.g. Line of Best Fit) and the coefficient of determination ( $R^2$ ). These steps are shown below. Note that the textbox with the regression coefficients is usually pasted in on top of the scatter plot. You’ll want to drag-and-drop it somewhere on your chart where you can read it. Double-click the text box to format the text using a larger type or a different font.



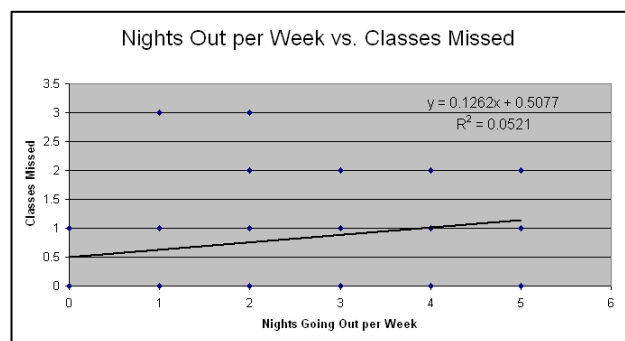
This block contains two screenshots of Excel's trendline options dialog boxes. The left screenshot is for Excel 2003, showing the "Add Trendline" dialog with the "Type" tab selected. The "Linear" option is selected, and the "Options" tab is also visible, with "Display equation on chart" and "Display R-squared value on chart" checked. The right screenshot is for Excel 2007, showing the "Format Trendline" task pane. The "Trendline Options" section has "Linear" selected, and the "Forecast" section has "Set Intercept = 0.0" checked. Red arrows point to the "Linear" option in both dialog boxes.

We now have the regression statistics, except for the correlation itself ( $r$ ). Find  $r$  by taking the square root of  $R^2$ . Use a calculator, or the Excel command: `=sqrt( )`. Doing so, we find  $r = 0.2283$ . Careful!! If the correlation is negative (downward sloping Line of Best Fit), we have to **add the negative sign** to the correlation coefficient.

## Regression Analysis

Three analysis steps will be reported (ask your professor and refer to class examples because opinions vary on exactly how to report and analyze them).

1. There is a weak, positive correlation between “nights going out per week” and



“classes missed per week” (  $r = 0.2283$  ).

2. “Nights going out per week” accounts for 5.2% of the variance in “classes missed per week.”
3. The slope of the regression line (  $m = 0.126$  ) indicates that, for each additional night students go out each week, classes missed increases by 0.126 (about one-eighth of a class).

DataSet5 ([2003](#) or [2007](#)) has trendlines added and regression statistics displayed on the chart. The second regression shows little or no correlation (  $r = 0.0911$  ) between “hours worked out per week” and “classes missed.”

The correlation (  $r$  ) provides a quick overview of the strength and direction of the relationship. The slope analysis explains in detail how changes in the predictor variable (x-variable) affect the dependent variable (y-variable). The coefficient of determination (  $R^2$  ) tells us the strength of the prediction.

In the analysis above, “nights going out per week” only explains about 5% of the variance in “classes missed.” While this might seem like a small amount, what it really means is that lots of other variables also influence how often students skip class. We might suggest variables like illness, number of hours worked, number of credit hours taken and student personality also influence absences. You can probably think of many more. The reason we capitalize  $R^2$  relates to multiple regression where we use several predictor variables in an equation. Then  $R^2$  tells us the overall predictive value of regression model which can include dozens of independent variables. We tend to focus on  $R^2$  and treat the correlation as less important, since most regression studies reported in the research literature and the popular media are referring to a multiple regression approach.