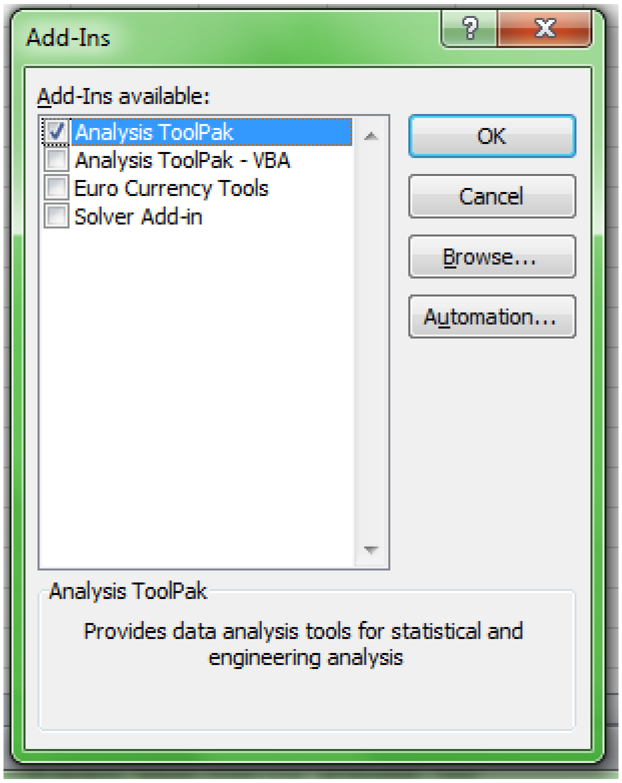
Data analysis

Data analysis in Excel using Windows 7/Office 2010

- Open the "*Data*" tab in Excel
- If "*Data Analysis*" is not visible along the top toolbar then do the following:
    - Right click anywhere on the toolbar and select "*Customize quick access toolbar…*"
    - On the left click on "*Add-Ins*"
    - Near the bottom, use the pull-down menu and select "*Excel Add-Ins*" and click "*Go*" to bring up this menu:



Add-Ins available:

- ☑ Analysis ToolPak
- □ Analysis ToolPak - VBA
- □ Euro Currency Tools
- □ Solver Add-in

Analysis ToolPak

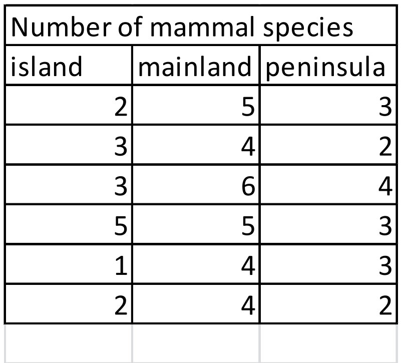
Provides data analysis tools for statistical and engineering analysis

-
- Select the "*Analysis ToolPak*" and click "*OK*".

Using one-way ANOVA in MS Excel

Introduction: When your observations fall into two or more categories of continuous or even discrete variables, you may be interested in asking if the groups differ from each other. Is fish diversity higher in phosphorus-enriched ponds than in low-phosphorus ponds? Does the abundance of forest-floor plants differ between clear-cut, tornado-damaged, and control plots of forest? Questions of this nature are answered using analysis of variance (ANOVA). It is worth mentioning that in the case of 2 categories you can run a *t* test or an ANOVA and the result will be the same.

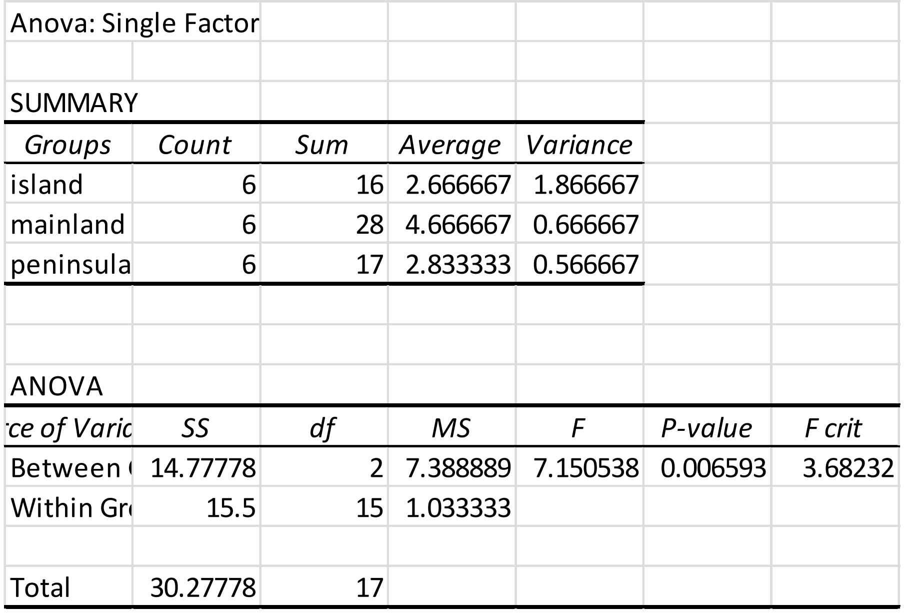
Analysis:



Number of mammal species		
island	mainland	peninsula
2	5	3
3	4	2
3	6	4
5	5	3
1	4	3
2	4	2

Table 1. Fake data for ANOVA

1. Organize your comparative data in adjacent columns (Table 1). There is no need to average them for analysis, and in fact averages will be calculated automatically during the ANOVA or *t* test.
2. From the "*Data*" tab, select "*data analysis*" (this must be added from the "addin" menu; see previous section).
3. Choose "*ANOVA single factor*"; click OK. Table 1 lists data from three habitats; so the *factor* of interest is habitat.
4. Click the tiny red arrow by "*input range*" and highlight all of the data including the column headings. Click the "Columns" button and check the "Labels in first row" box.
5. Select any of the output options that you like and hit "OK"
6. The output from the fake data should look like this:



Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
island	6	16	2.666667	1.866667		
mainland	6	28	4.666667	0.666667		
peninsula	6	17	2.833333	0.566667		
ANOVA						
ce of Varic	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between (	14.77778	2	7.388889	7.150538	0.006593	3.68232
Within Gr	15.5	15	1.033333			
Total	30.27778	17				

7. The conclusion based on the *p-value* would be that number of species differ significantly among the three habitats. Note that the ANOVA does not tell you which groups are different, although in this case it looks like more species are found on the mainland and there is no difference between the island and the peninsula.
8. Finally, if you are making a comparison between just 2 groups, you can use exactly the same procedure. Or you could choose to run a *t*-test and it will give you a result that is mathematically identical to that produced by an ANOVA run on 2 groups. We could go back to the fake data and ask if the island and peninsula differ from each other by running the test without including the mainland data column.

Graphing ANOVA-type data: Use the averages to draw a bar graph. Add standard error bars to the graph. Calculate those using this formula: *=stdev(A1:A6)/Sqrt(6)* (assuming your data are in cells A1 through A6 and you have 6 data points). More detailed instructions are provided in the graphing section of this manual.

Regression in MS Excel

Does blood pressure increase with age? Does shrub cover decrease with increasing canopy cover? Is there a relationship between phosphorus concentration and algal cell density in ponds? All of these questions can be addressed using regression.

Nature of the data

All of the datasets described above are *continuous*; that is to say, they vary over some range without breaks. They are not *categorical* (like male and female), that are not *discrete* (like number of people in a single car; you would not typically think about 3.5 people in a car). As the range of a discrete variable increases (number of plants per hectare for example), the larger number means that what in fact is a discrete variable can be treated as continuous.

Graphing

We typically graph such datasets using a scatter plot (Figure 1) . If we have a basis for considering for example that running speed impacts heart rate, then we would use running speed on the horizontal (*x*) axis, and heart rate on the vertical (*y*). In this case running speed is the *independent variable*. The *dependent*, or *response variable* is heart rate because we expect it to *depend on,* or *respond to* running speed.

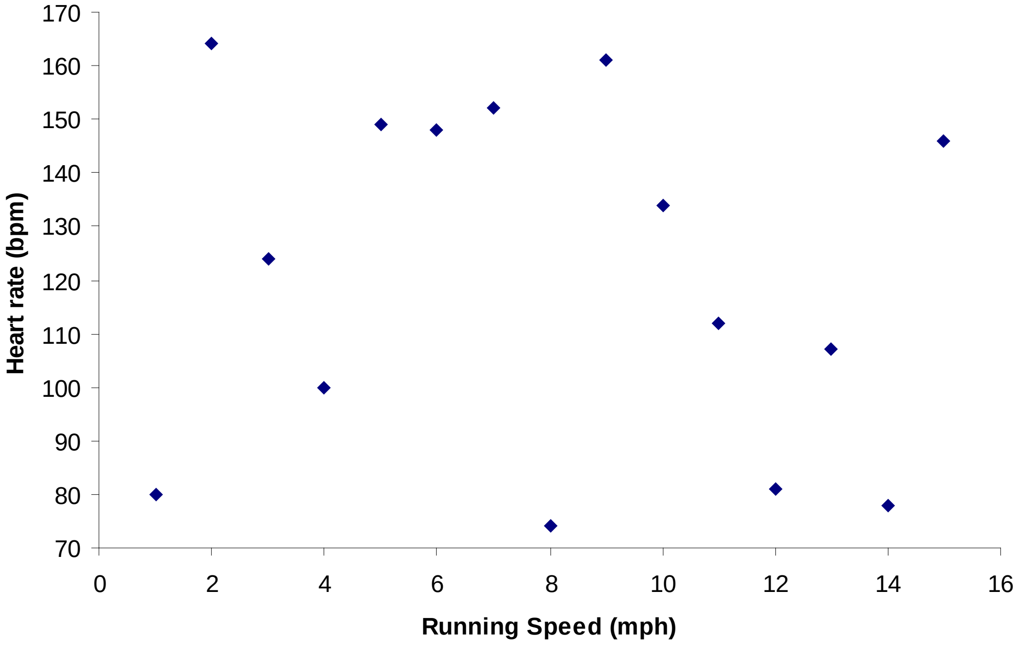
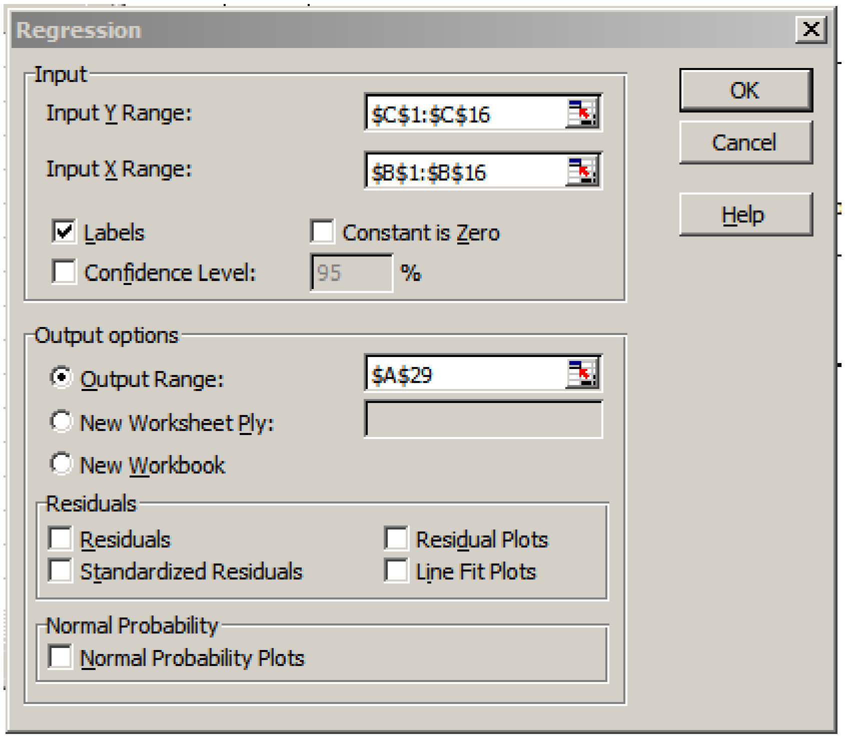


Figure 1. Fictional data representing the effects of running speed on heart rate.

Analysis: We might look at the pattern on the right and perceive a pattern, or not! As is the case with all statistics, the point is to remove subjectivity and have firm criteria for claiming a relationship. The analysis one would use for this sort of question is *regression*. There are many forms of regression for relationships of different shapes, but for our purposes we are considering only *linear regression*. In other words we are asking only if, and how well a straight line can describe the relationship between variables. In excel under the *Data tab,*select_ *data analysis, regression* to bring up this window:



Input

Input Y Range: $C$1:$C$16

Input X Range: $B$1:$B$16

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

● Output Range: $A$29

○ New Worksheet Ply:

○ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

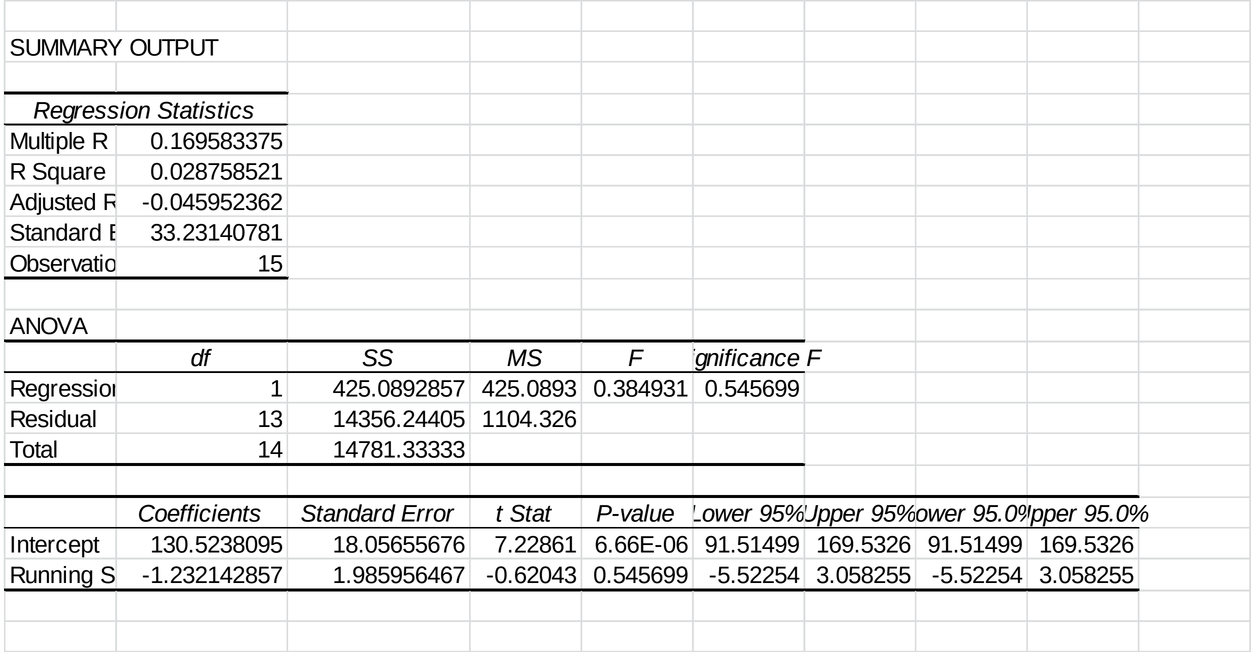
☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

The response variable goes in the Input Y Range and the independent variable goes in the Input X range. You can click on the tiny red arrow in each case and highlight the appropriate portion of the data (including labels). The output range simply is a place for the statistical output to go.

Output: Output from the preceding data set:



SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.169583375							
R Square	0.028758521							
Adjusted R	-0.045952362							
Standard E	33.23140781							
Observatio	15							
ANOVA								
	df	*SS*	*MS*	*F*	*gnificance F*			
Regressioı	1	425.0892857	425.0893	0.384931	0.545699			
Residual	13	14356.24405	1104.326					
Total	14	14781.33333						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	130.5238095	18.05655676	7.22861	6.66E-06	91.51499	169.5326	91.51499	169.5326
Running S	-1.232142857	1.985956467	-0.62043	0.545699	-5.52254	3.058255	-5.52254	3.058255

The number under *Significance F* is the *p* value. In this case the *p* value is greater than 0.05 and we can conclude that there is no relationship between running speed and heart rate.

Regression example 2: Along with other questions, Connon and Simberloff's (1978) paper examined the effect of sampling bias on collection data. They concluded that the number of collecting trips explained more of the variability in number of plant species observed on Galapagos Islands than did Island size or any other island feature measured. The data set:

Species	Collecting
57	10
31	6
3	1
25	4
2	1
18	6
10	6
8	1
2	1
96	13
94	12
40	7
5	2
54	13
346	27
47	7
2	1
102	10
108	9
12	6
69	10
290	28
237	24
440	38
61	11
283	29
45	6
16	3
21	5

And the statistical output:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.973547
R Square	0.947795
Adjusted R	0.945861
Standard E	27.01902
Observatic	29

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	357850.2	357850.2	490.1875	7.62E-19
Residual	27	19710.73	730.0272		
Total	28	377561			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-31.902	7.35061	-4.34005	0.000179	-46.9842	-16.8198	-46.9842	-16.8198
Collecting	11.61333	0.524536	22.14018	7.62E-19	10.53707	12.68959	10.53707	12.68959

Output	**Value**	**Standard interpretation**
p value	7.2 E-19	There is a very significant relationship between number of trips and number of species observed
Coefficient (of collecting trips)	11.61	The slope is positive telling us that as number trips increases, so does number of species seen. Negative slopes indicate the opposite trend.
R square	0.947	This measures how tight or strong the relationship is. In this case we can say that collecting trips explain 94.7% of the variability in number of species observed.

Graphing example 2: Connor and Simberloff's (1978) data set is presented graphically in the manual section on graphing. Compare how the data follow a tight linear pattern compared to the fake data on heart rate in this section.

Graphing

Figures in Community Ecology

All graphs, maps, photographs, and sketches are considered “Figures” and appear in a numbered sequence in the order cited in your paper. Any set of numbers and/or letters is considered a table and tables have their own numbered sequence (IE, even after three figures, your first table is still *Table 1*).

A good graph minimizes clutter and unnecessary ‘ink’. Use the MS Excel “Scatter Plot” option to make graphs displaying continuous data on the vertical and horizontal axis. The species area data for the upcoming lab report are a good example; area on the *X* axis; number of species on the *Y* axis. **Remove** all of the following items added by Microsoft excel: “Series 1”; background color; frames on right and top; grid lines; 3D effects.

Scatter plots

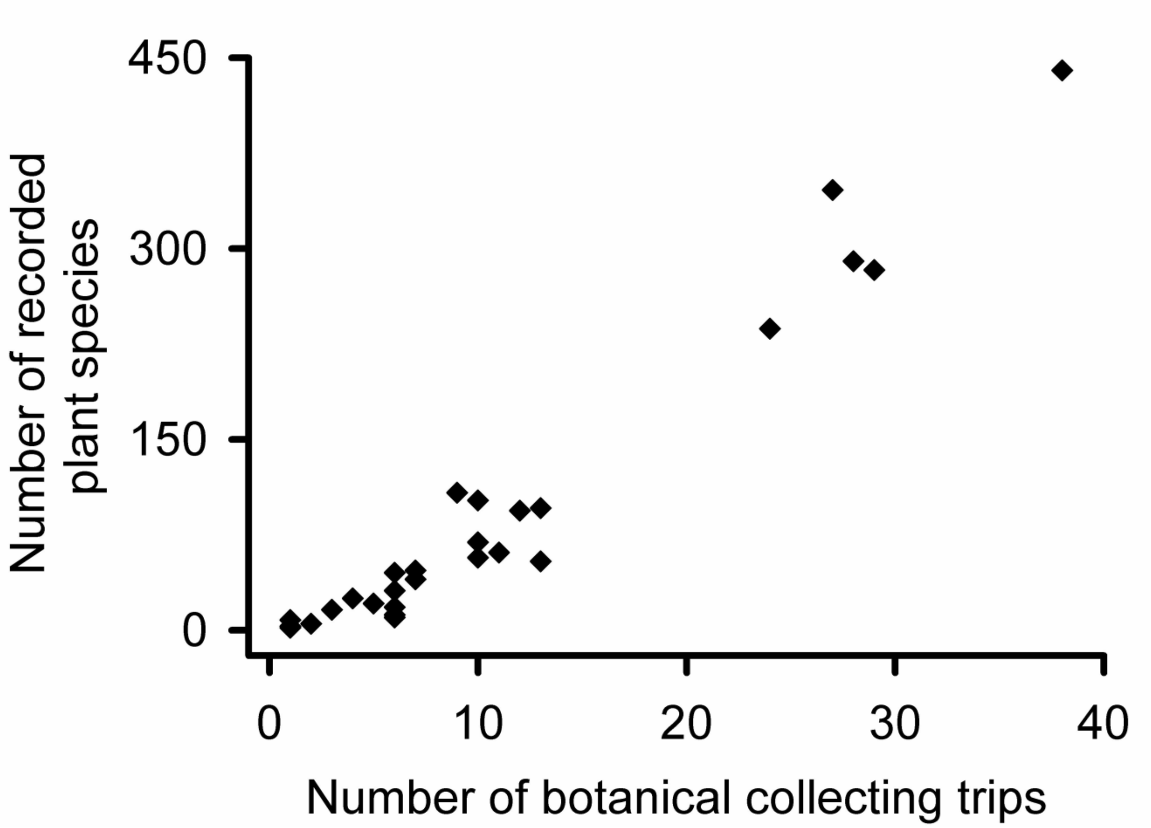


Figure 1. Illustrating the point that more sampling leads to more species observed. Connor & Simberloff (1978) analyzed data from collecting trips to the Galapagos Islands and found that number of collecting trips better explained number of species recorded than did island area, elevation, or isolation. Data extracted from Table 3 in Connor & Simberloff (1978).

The figure legend is always placed underneath and contains roughly a paragraph of information describing the figure content in sufficient detail that the figure stands alone. The *legend* inserted by MS excel is useful *only if two or more data sets are displayed* on one graph using symbols.

This figure contains data that span the nearly entire range presented. If we were presenting data from only the largest five islands we would adjust the horizontal axis to run from 20 to 40, and the vertical axis from 150 to 450. Note that the axis lines have been thickened and fonts enlarged beyond the default. **Important**: Graphs should not start at zero, zero if the data range fall between 75 and 85 (for example).

Bar graphs

We use bar graphs when presenting the averages of continuous variables (on the
We use bar graphs when presenting the averages of *continuous* variables (on the *Y* axis) from one or more *categories* on the horizontal axis.

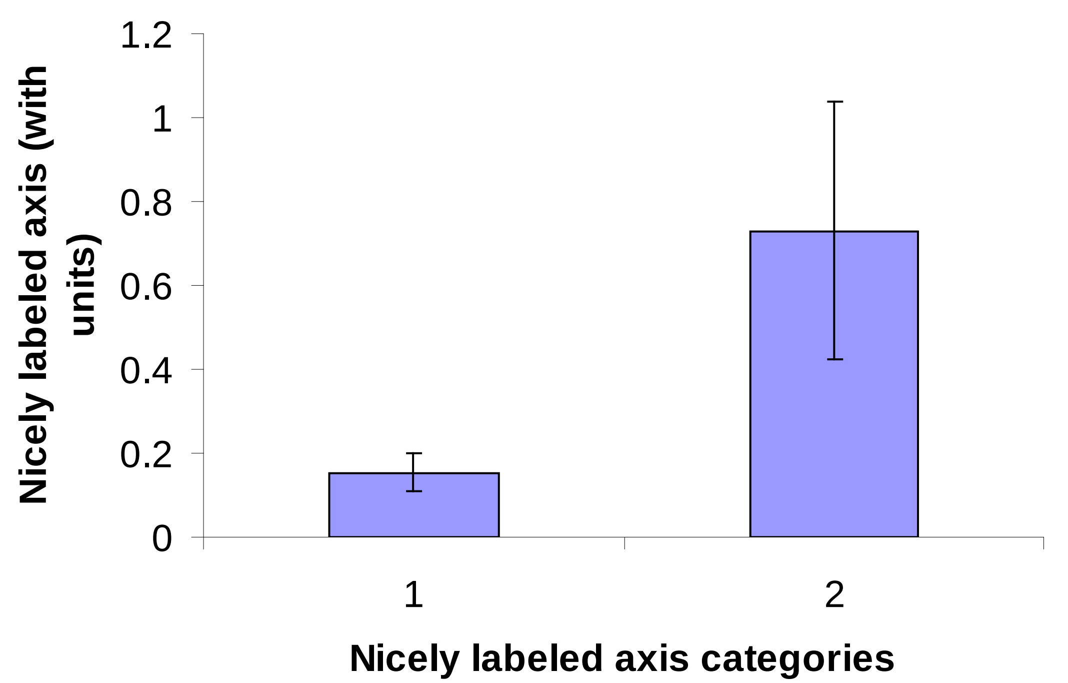


Figure 1. Very detailed title, 3-4 lines; place under the graph

The bar height equals the average of the response variables for treatments 1, and treatments 2. The error bars above and below the average in this case equal standard error; calculate these values as: (standard deviation)/(square root of the number of samples). The scale is appropriate to the data; if the averages were 150 and 200, I might start the axis at 100 rather than zero. **Important:** You should replace the numbers on the horizontal axis with names of sites or treatments (see example under adding error bars handout).

Adding error bars to bar graphs in excel

Introduction: Bar graphs are among the most common ways to present the averages of a set of treatments or conditions in community ecology and many other fields. Every average is based on raw data measured from a sample of several individuals. If I care about grass density in my lawn I might count the number of stems from several small quadrats and then calculate the average number of stems. The numbers of stems in each of my individual quadrats will be greater than or less than the average. In other words *there is variability in the raw data*. We might expect more variability in the heights of people than in the heights of Volkswagens. *Some data sets are more variable than others*. We use error bars above and below the average to depict that variability

How to measure variability: There are several metrics used to express variability. Standard deviation expresses the *variability in your sample* and is calculated in MS Excel using this Formula 1.

= stdev(A1:A6)…………………………………………….Formula 1

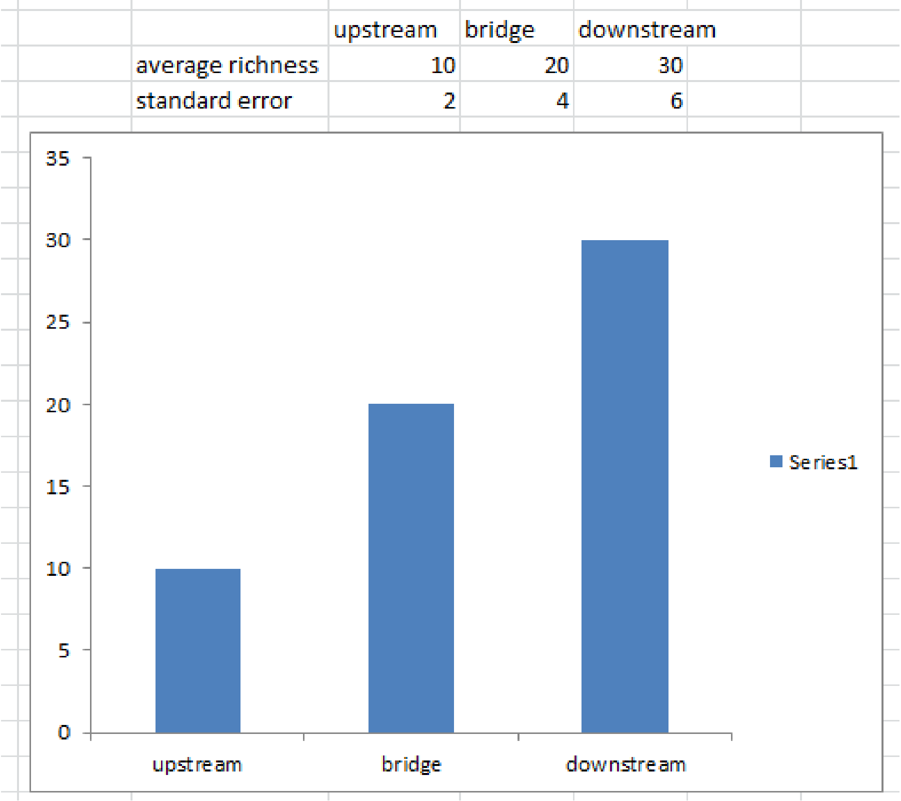
The formula calculates the standard deviation from the raw data you entered in the cells *A1* through *A6* in the spreadsheet. You can refer to any set of cells in the spreadsheet by changing the letters and numbers in parentheses in Formula 1. The disadvantage of standard deviation is that it increases in magnitude as your sample size decreases. Samples can be expensive or time consuming to collect and so we often need to work with small sample sizes. What we really need is a measure of variability in the entire population, and not just in our sample.

Standard error adjusts the value of standard deviation based upon the sample size using Formula 2

= stdev(A1:A6)/sqrt(n)…………………………….Formula 1

Where *n* = the number of replicates in your sample; don't enter the letter *n*, enter the number of samples you took or refer to a cell in the spreadsheet that contains that information. *Sqrt* calculates the square root of whatever value you use to replace *n* in Formula 2. **Standard error will be the preferred measure of variability used throughout this course**.

How to add the error bars to your bar graph:
Lay your data out as illustrated below. In this case the fake data represent the average number of insect species found several samples taken from each of three locations in a stream.



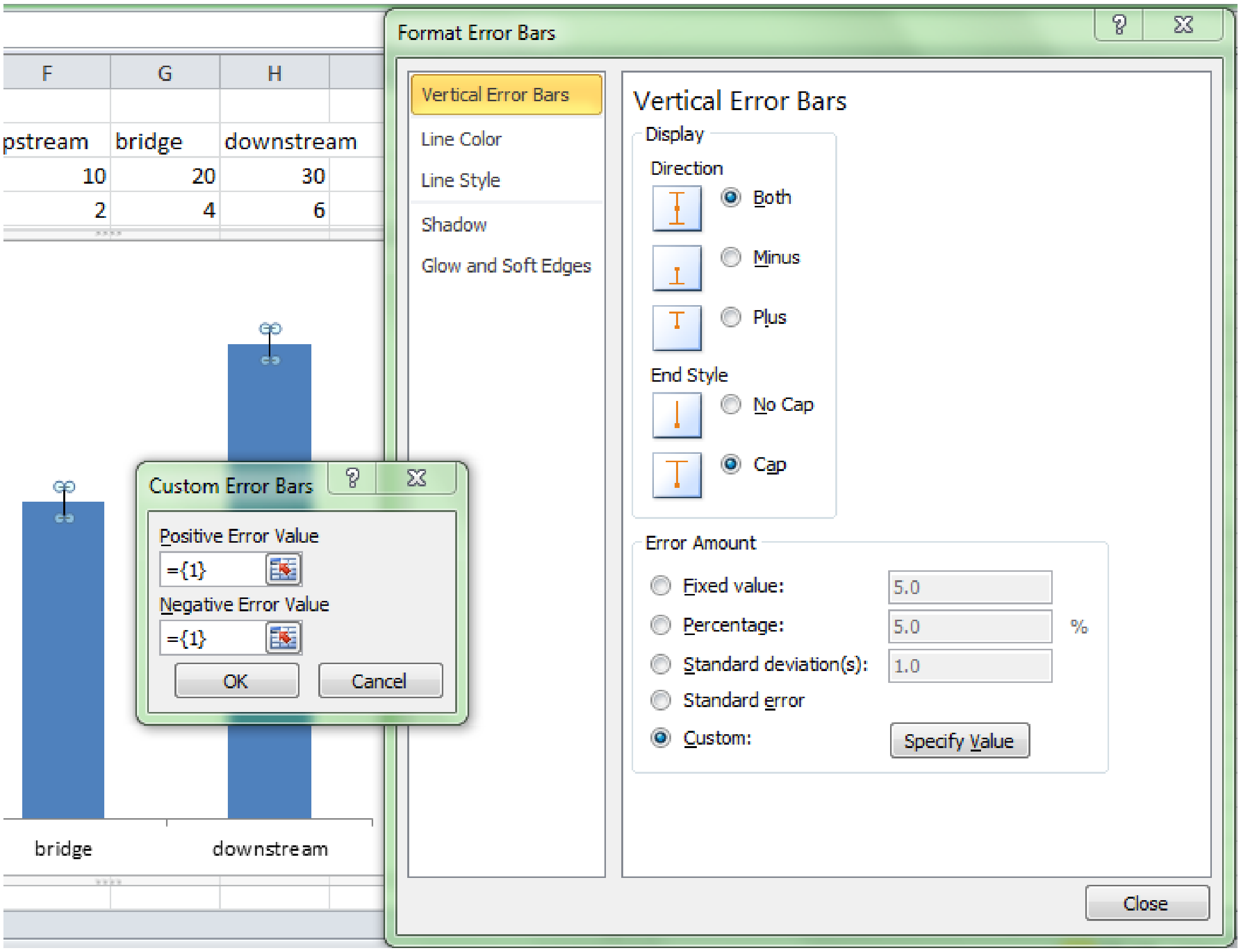
	upstream	bridge	downstream
average richness	10	20	30
standard error	2	4	6

Note:

- Standard error values are underneath the graphed averages.
- The graph has been moved in the spreadsheet so as not hide the numerical values.

1. Click anywhere on the chart - this will reveal the "*Chart Tools*" at the top of the window. Click "*Layout*"
2. Right click on any bar in the graph – 2 small windows will pop up – work in the smaller upper one. Click the little drop down arrow and select the data set to which you'd like to add error bars (*Series 1* unless you have renamed the data set).
3. Now, go up to "*Chart Tools*" at the top and select "*Error Bars*"/ "*More error Bar Options*" (because all of the other options offered are, to be perfectly honest, fake).

4. Click “*Custom*” and “*Specify Value*”.



F	G	H
pstream	bridge	downstream
10	20	30
2	4	6

Custom Error Bars

Positive Error Value: ={1}

Negative Error Value: ={1}

Format Error Bars

Vertical Error Bars

Display

Direction: ● Both ○ Minus ○ Plus

End Style: ○ No Cap ● Cap

Error Amount

○ Fixed value: 5.0

○ Percentage: 5.0 %

○ Standard deviation(s): 1.0

○ Standard error

● Custom: Specify Value

5. Next click the tiny red arrow in the box under “*Positive Error Bar*”; highlight the values for the standard errors that are lined up under the averages. Hit “*Enter*”!
6. Now, you would think that having selected “both”, that both the upper and lower error bars would be displayed; you would be wrong! Repeat the process for “*Negative Error Bars*”.
7. Click “*Close*”.
8. Truly beauteous error bars will now grace your bar graph!