Statistical Guide for Students
Environmental Science & Ecology
SUNY Brockport

Table of Contents

Introduction & Definitions

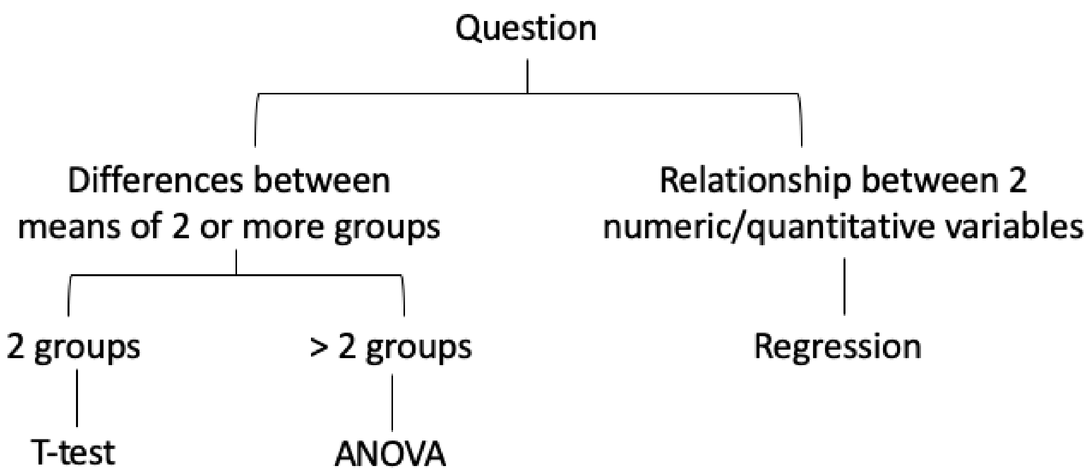
This guide provides basic background information on some of the most common statistical tests used in our ENV courses and includes details on t-tests, regressions, Chi-squared analysis, and analysis of variance (ANOVA). Microsoft Excel is used for each of the examples included in this guide.

First, we need to define a few terms that apply to statistical tests.

- **Test statistic**: the value calculated from the statistical equation. For the most part, you don't need to calculate the test statistic by hand. Statistical programs and Excel (which is used in this document) will calculate this value for you based on the data you enter.
    - Number of digits to report: 2 digits past the decimal point
- **P-value**: the probability value that tells you whether your data show significance. The value 0.05 is the most commonly used cut-off for whether or not your data show statistical differences/relationships. If the calculated p-value is less than 0.05, there is evidence that the data do show significant differences/relationships. If the calculated p-value is greater than 0.05, you generally must assume there are no statistically significant differences/relationships in your data.
    - Number of digits to report: 3 digits past the decimal point
- **Categorical variable**: A variable that is divided into two or more discrete groups. For example, the variable "tree species" can be divided into "oak", "maple", and "dogwood" if these are the three trees of interest in a study.
- **Quantitative variable**: A numeric variable, which can be discrete or continuous.
    - **Discrete variable**: Typically "count" data. For example, the number of species counted at a site, or the number of individuals in a population.
    - **Continuous variable**: A variable that can take on a range of values (i.e., there are not only 3 or 5 or 10 discrete options). For example, when looking at weights of gray squirrels, their weights would be considered a continuous variable (their weights can take on any value between, say, 0.2 to 0.5 kg).
- **Assumption of independence**: There are several assumptions necessary for these tests to provide accurate results. For the purposes of this document, we will only describe one here – the assumption that the data collected are independent. In other words, one observation does not influence the values of another observation. You can find a more detailed explanation of this assumption here: https://blog.minitab.com/blog/quality-business/common-assumptions-about-data-part-1-random-samples-and-statistical-independence

Decision Tree

Below is a decision tree to help you determine which of the three statistical tests included in this guide is most appropriate for your data. Keep in mind there are many other types of statistical tests – this decision tree (and guide) is not exhaustive.



T-test

When to use: Use this test when you want to determine if the continuous variables for two groups are different from each other.
Values to report: t-statistic, p-value
Examples of when to use a t-test:

- Are plant respiration rates at site #1 and site #2 different?
- Is there a difference between abundance of salamanders at site #1 vs. site #2?
- Do the numbers of fruits on apple trees differ from those on pear trees?

Note that each of the above scenarios has one categorical variable (e.g., site) and one quantitative variable (e.g., number of salamanders).

[Note: There are three types of t-tests. The most commonly used one (independent samples t-test) is described here. For information on the other two types of t-tests (paired sample and one sample t-tests) see https://blog.minitab.com/blog/adventures-in-statistics-2/understanding-t-tests-1-sample-2-sample-and-paired-t-tests]

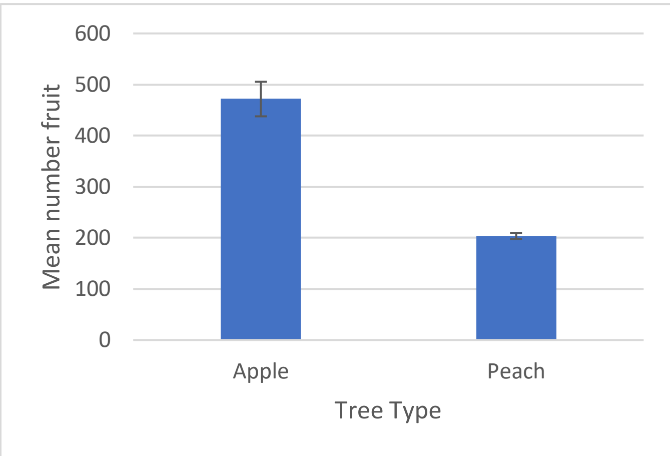


Fig. 1 This is the type of data on which you would want to use a t-test. You have two categories on the x-axis and want to compare the number of fruits of each species to see if one species produces significantly more fruit than the other. Note the error bars (standard error), which give the reader an understanding of the variability in the data for the two groups.

How to run a t-test

Tree	Height (m)	Fruit
Apple	4.1	429
Apple	4.2	451
Apple	4.3	468
Apple	4.8	498
Apple	4.9	512
Peach	4.4	198
Peach	4.4	199
Peach	4.7	211
Peach	4.8	201
Peach	4.9	208
Total Apple	22.3	2358
Total Peach	23.2	1017
Mean Apple	4.46	471.6
Mean Peach	4.64	203.4
SD Apple	0.3646917	33.84228
SD Peach	0.2302173	5.770615

Fig. 2 These are the data used to create the graph above. For this example, we are interested in the "fruit" data only. We are testing to see if fruit number is different in each species. In other words, does one species produce more fruits than the other?

- You may have to download the Analysis Toolpak in Excel if you haven't already.
- Go to Data → Data Analysis. When you choose Data Analysis you should see a box with a lot of options for different tests. Choose "**t-Test: Two sample assuming unequal variances**".
- Then, you'll need to choose the data you're comparing. Think back…what are you comparing? Number of fruits from the apple trees to number of fruits from the peach trees.
- For Variable 1 in the t-test box that pops up (see Fig. 3 below), highlight the apple fruit data (the raw data, not the averages or standard deviations). To highlight the data, click and drag from the first cell to the last in which the data are located (in the example data table below, this would be from "429" to "512"). Then for Variable 2, highlight the fruit numbers for the peach trees. This will tell the t-test to analyze if the number of fruits produced by each species is different from the other (i.e., do apple trees produce more fruits than peach trees?).
- You can leave everything else the same and click "OK".

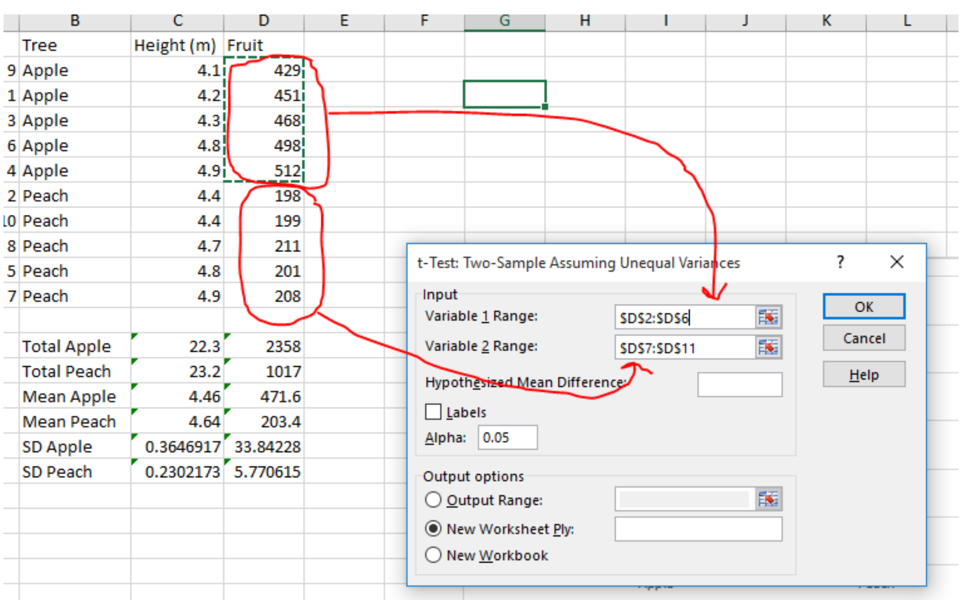


Fig. 3 Here are the details on what goes into the t-test dialogue box. Variable 1 will be apple fruit numbers. Variable 2 will be peach fruit numbers. (Which is Variable 1 vs. Variable 2 does not matter.)

- When you click "OK", the results will be put on a new blank worksheet and should look something like this:

Fruit numbers		
t-Test: Two-Sample Assuming Unequal Variances		
	Variable 1	*Variable 2*
Mean	471.6	203.4
Variance	1145.3	33.3
Observations	5	5
Hypothesized Mean Diff	0	
df	4	
t Stat	17.46869883	
P(T<=t) one-tail	3.15246E-05	
t Critical one-tail	2.131846786	
P(T<=t) two-tail	6.30493E-05	
t Critical two-tail	2.776445105	

Fig. 4 Resulting table of information from the t-test

- *Note: I added the red markings and the "Fruit numbers" title at the top. If you don't add the title, you may forget what the test is showing because there is no label that comes along with the test.*
- **The values circled in red are the important ones – the numbers you should report.**
- The bottom one ("P(T<=t) two-tail") is your p-value**. If this number is less than 0.05, the difference between the two sets of data you are comparing is significant.** In the case above, the "E-05" part means the decimal place is actually five spaces to the left (i.e., p = 0.00006305), so it is significant.

Regression

When to use: Use this test when you want to determine if two quantitative variables are related (i.e., both your x- and y-variables are quantitative, not categorical). Regression does not test for similarities or differences between groups but rather a relationship between the two (e.g., if one variable increases, does the other also increase?).
Values to report: r^2, p-value, equation of the regression line
Examples of when to use regression:

- Are the size of an apple tree and the number of fruits it produces related? In other words, do bigger trees produce more apples?
- Is there a relationship between age and mortality rate in gray squirrels?
- Is there a relationship between petal width and petal length in *Iris versicolor*?

Note that each of the above examples contain two variables that are quantitative (i.e., they are not separated into two or more categories – they are numbers on a scale).

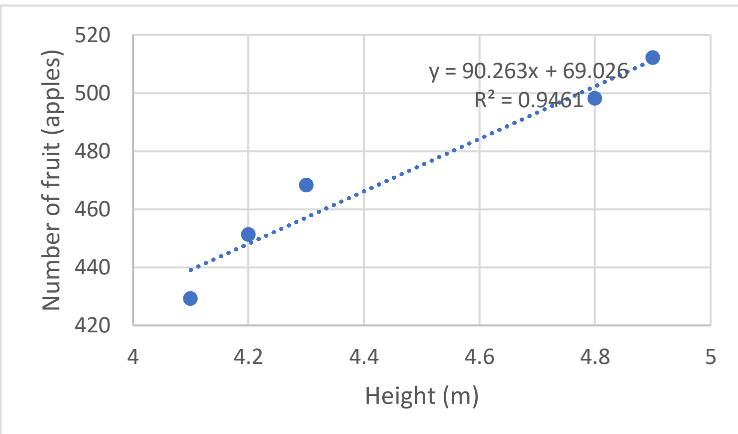


Fig. 5 This is the type of data on which you would use a regression test, rather than a t-test. The x-axis is a continuous variable (not categories like above). Both of the variables (x and y) are continuous.

How to run a regression analysis

- Remember to download the Analysis Toolpak if you haven't already (see above).
- Looking at the same data in Fig. 2 above (height of tree and number of fruits), we will focus on only the **apple data** to see if there is a relationship between height of the tree and the number of fruits it produces. In other words, do taller trees produce more fruits?
- **Important: the equation and r^2 value on the graph above are important and should be reported, but they do not tell you if the relationship is significant**. We'll need to use the Analysis ToolPak again to tell if there is a significant relationship (see next bullet point). The r^2 value indicates how well the line fits the data – the higher the r^2 value, the better the line fits. With a lot of variability in the data, you will notice a smaller r^2 compared to data that are less variable. R^2 can range from 0 to 1.
- Go to Data → Data Analysis. Choose "**Regression**".

- Now, you'll need to put in the data for the x and y variables. The x variable as you can see above is the Height and the y variable is the number of apples.

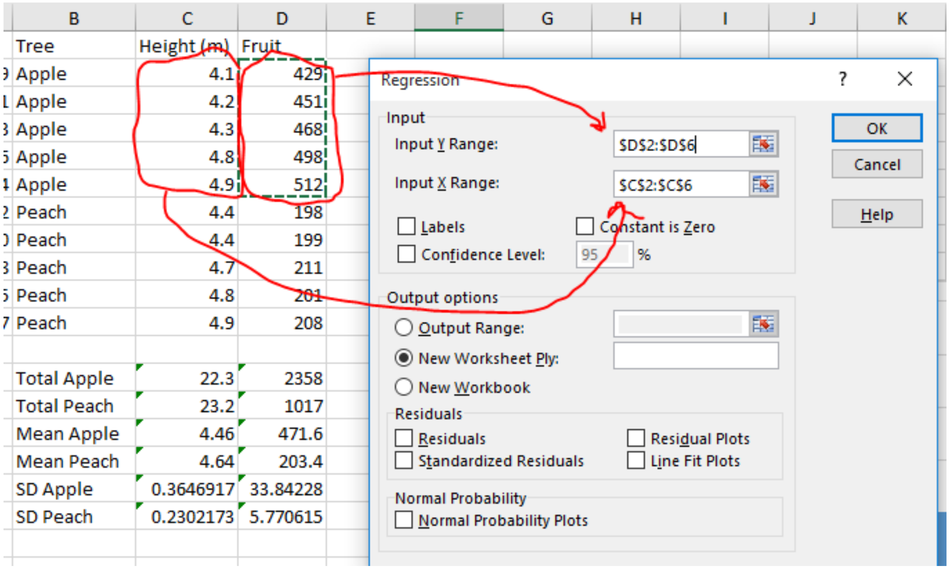


Fig. 6 This is an image of the data that will go into the Regression dialogue box to see if there is a significant relationship between height and fruit number.

- Click "OK" and you should see the regression results in a table on a new worksheet. Again, it will be helpful to label the results somehow so you know what you tested.
- All you need to report from this table is the p-value, which is labeled as "Significance F".

Fruit number vs. Tree height - Apples

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.972695
R Square	0.946136
Adjusted R Squa	0.928181
Standard Error	9.06942
Observations	5

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4334.437	4334.437	52.69551	0.005393973
Residual	3	246.7632	82.25439		
Total	4	4581.2			

	Coefficients	*Standard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	69.02632	55.6054	1.24136	0.302689	-107.9348958	245.9875	-107.935	245.9875
X Variable 1	90.26316	12.43437	7.259167	0.005394	50.69144941	129.8349	50.69145	129.8349

Fig. 7 Output from the regression test. You will need to report the R square value (listed in the first table) and the p-value which is circled in red here.

- The r^2 value can also be found in the output table ("R square" in the first table labeled "Regression Statistics") in addition to on the graph when this option is selected when adding the line to the graph.

- **Ignore the cells that are labeled as "p-value**" in the bottom table. These values do not tell you whether the relationship between the two variables is significant – they explain another aspect of the relationship that we are not interested in.
- **Note:** Be aware that fewer data points may mislead your interpretation of the results. For example, if there are only three points you can still run a regression and create a graph, but had more points been added, the results may have been different if those additional points were very different from the original three points.

Analysis of variance (ANOVA)

When to use: ANOVA is very similar to the t-test described above. The main difference is ANOVA can detect differences between more than two groups at a time (with the t-test you can only test one group of data vs. one other group of data). Use ANOVA when you want to compare three categorical sets of data.
Values to report: F-statistic, p-value
Examples of when to use ANOVA:

- Do the numbers of fruits on apple, pear, and peach trees differ?
- Does the abundance of chinook salmon differ among stream A, stream B, stream C, and stream D?
- Does bird abundance differ among the following habitats: grassland, forest, and agricultural field?

Note that each example above has one categorical and one continuous variable (just like the t-test example above).

How to run ANOVA

- Remember to download the Analysis Toolpak if you haven't already (see above).
- Below is an image of the example data for the first bullet point question above (fruit numbers of three different tree species).
- Enter your data in columns as shown in Fig. 9.
- Go to Data → Data Analysis. When you choose Data Analysis you should see a box with a lot of options for different tests. Choose "**ANOVA: Single factor**".

	A	B	C
1	ANOVA example		
2	FRUIT NUMBERS		
3	Apple	Pear	Peach
4	10	22	12
5	15	31	15
6	3	20	16
7	7	28	12
8	8	28	12
9	8	28	13
10	9	30	14
11	12	27	8
12	15	24	5
13	10	29	11

Fig. 9 Example data for ANOVA. Each column contains numbers of fruits for three different tree species (apple, pear, and peach trees).

- Then, you'll need to choose the data you're comparing. Think back…what are you comparing? (Number of fruits from the apple, pear and peach trees.)
- For Input Range in the ANOVA box that pops up (see Fig. 10 below), highlight all of your data including the column headers/titles (the raw data, not averages or standard

deviations if you've calculated those). To highlight the data, click and drag from the cell with "Apple" in it to the to the last cell in which the data are located (in the example data table below, this would be cell C13).

- Select "Columns" for the "Grouped by" option (this tells Excel that each column is a different category).
- Check the "Labels in first row" box to tell Excel that you have also highlighted the data labels (apple, pear, and peach). Doing this will make the output table a bit more clear.

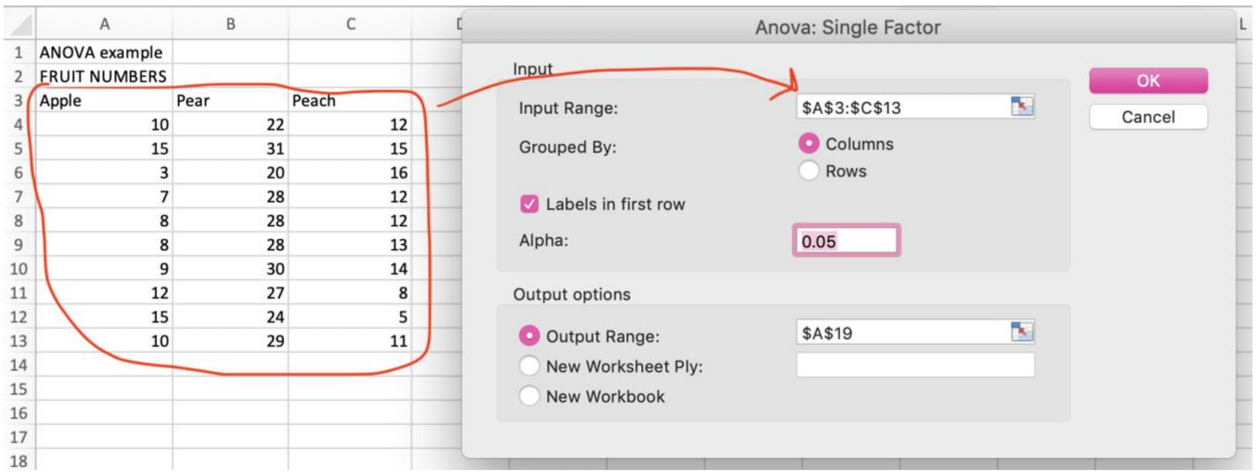


Fig. 10 After choosing "ANOVA: Single factor" from the Data Analysis options, you will see the Input window pop up. The Input Range should include all of your data. If you have also highlighted the column labels/headers (in this case, Apple, Pear, and Peach), check the "Labels in first row" box to let Excel know these are labels and not data.

- You can leave everything else the same and click "OK". The results will be presented on a new blank worksheet (unless you select another option under "Output options").
- In the upper output table labeled "SUMMARY" (see Fig. 11), you can see some descriptive statistics regarding the fruit numbers for each species. In the lower output table labeled "ANOVA" you can see the F-statistic ("F") and p-value that you need to report. In this case the p-value (1.967E-11) is below the 0.05 significance level; thus, the difference among the trees' fruit production is significant.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Apple	10	97	9.7	13.3444444
Pear	10	267	26.7	12.6777778
Peach	10	118	11.8	10.6222222

ANOVA

ource of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1718.066667	2	859.033333	70.327168	1.967E-11	3.35413083
Within Groups	329.8	27	12.2148148			
Total	2047.866667	29				

Fig. 11 Results from the ANOVA in Excel. The two values circled in red are the values you should report (F-statistic and p-value).

- The p-value above will tell you if there is an overall difference among the groups. However, to determine which groups differ from each other (i.e., do peach trees produce significantly more fruits than apple trees?), you will need to run an additional test called a Tukey test. Use one of the links below to run a Tukey test to determine which groups actually differ from each other:

  https://www.icalcu.com/stat/anova-tukey-hsd-calculator.html

  https://astatsa.com/OneWay_Anova_with_TukeyHSD/