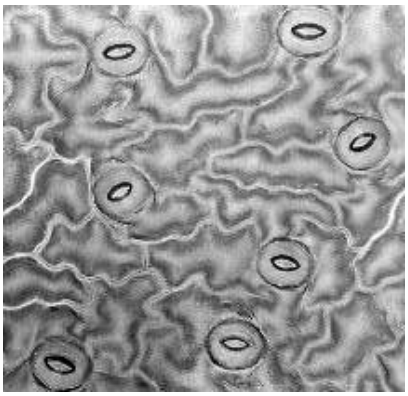
TIEE
Teaching Issues and Experiments in Ecology - Volume 1, January 2004

EXPERIMENTS

Environmental Correlates of Leaf Stomata Density

Bruce W. Grant and Itzick Vatnick
Biology, Widener University, Chester PA, 19013
grant@pop1.science.widener.edu
vatnick@pop1.science.widener.edu



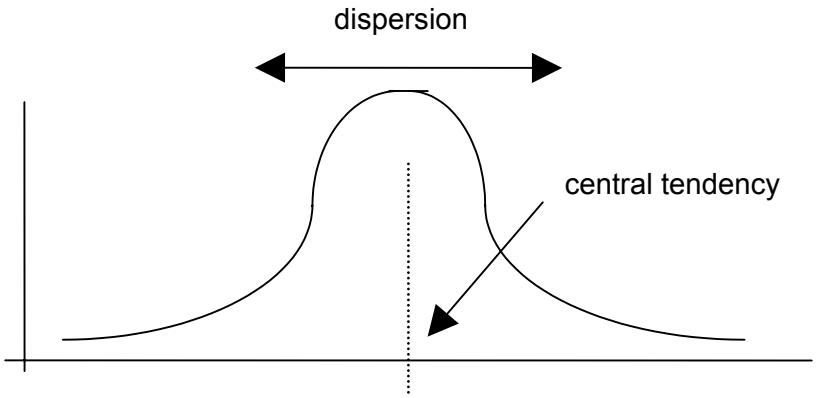
stomata viewed at 400x in nail polish impression from leaf underside © Marc Brodkin, 2000

Appendix 1. Guidelines for Statistical Analysis

Modern biological research emphasizes the collection of quantitative data on a variety of biological topics. Much of these data are highly variable. As a result, techniques of statistical analysis are very valuable in helping the biologist describe the variation within sets of data, express the degree of confidence that can be placed in average values, and objectively test hypotheses about data collected from different groups of subjects. This handout describes a number of techniques commonly used by biologists for these purposes and that you will use in the analysis of your stomata data.

A. Descriptive Statistics.

After a set of data is collected it then can be analyzed statistically in order to better determine whether the data support or reject a given hypothesis. The first procedure that is usually done is to calculate a set of parameters that describe two aspects of the data: (1) central tendency and (2) dispersion.



(1) Measures of Central Tendency. One type of statistics determines the central tendency of the data. The central tendency provides information on how the values of the data you collected cluster around some single middle value. There are three measures of central tendency that are used in the analysis of data, which are described below:

MODE = the most frequently observed value of the data

MEDIAN = the middle value when the data set is ordered in sequential rank (i.e. highest to lowest, or lowest to highest)

MEAN = average value. The mean is the most commonly used measure of central tendency. It is estimated using the sum of all the individual values (x_i) divided by the total number of individuals in the sample (n):

$$\textbf{MEAN} = \bar{X} = \sum_{i=1}^{n} \frac{x_i}{n} = (x_1 + x_2 + x_3 + x_4 + ... + x_N) / n$$

(2) Measures of Dispersion. Another set of statistics describes how spread out the data are.

RANGE = The highest value minus the lowest value.

VARIANCE. The variance is the sum of each of the differences or deviations between individual values and the mean value. The total difference is divided by the number of individuals in the sample minus one.

$$\textbf{VARIANCE: } \sigma^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1} = \sum_{i=1}^{n} \frac{(X_i^2 - n * \overline{X}^2)}{n-1}$$

STANDARD DEVIATION. The square root of the variance.

$$\textbf{STANDARD DEVIATION = S} = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{n} \frac{(X_i^2 - n * \overline{X}^2)}{n-1}}$$

STANDARD ERROR. The standard error is the standard deviation divided by the square root of the sample size.

$$\textbf{STANDARD ERROR} = \frac{S}{\sqrt{n}}$$

---

e.g. for the data: { 3, 3, 4, 5, 6, 6, 6, 6, 7, 8, 10 } that could represent a set of quiz scores,

MODE = 6,
MEDIAN = 6
MEAN = (3 + 3 + 4 + 5 + 6 + 6 + 6 + 6 + 7 + 8 + 10)/11 = 5.82
SAMPLE VARIANCE = 4.363636
STANDARD DEVIATION = 2.088932
STANDARD ERROR = 0.629837

FINDING DESCRIPTIVE STATISTICS USING MICROSOFT'S EXCEL

The computer makes data analysis easy. All you need is to enter your data into a spreadsheet and follow the simple steps below:

1. Under **Tools** click on **Add Ins** and then click on **Analysis ToolPack** and **OK.**

2. Look again under the **Tools** menu and a new option **Data Analysis** will appear at the bottom of the menu. Click on **Data Analysis**, and click on **Descriptive Statistics**.

3. Highlight your column of data. Hit the **Summary Statistics** box so that an **X** appears. Next, specify the **Output Range**, i.e. where you want to put the analysis output table, and finally hit **OK**.

4. The program will spew out a table of statistics that will look something like this:

Variable 1	
Mean	5.818182
Standard Error	0.629837
Median	6
Mode	6
Standard Deviation	2.088932
Sample Variance	4.363636
Kurtosis	0.338976
Skewness	0.454113
Range	7
Minimum	3
Maximum	10
Sum	64
Count	11
Confidence Level(95.000%)	1.234455

B. Statistical Testing: Comparisons of Means Using a STUDENT'S T-TEST

For this lab activity, we are going to carry the analysis of the data one step further and determine whether the hypothesis you proposed for the distribution of stomata in the two groups of leaves you collected should be accepted or rejected. To do this you should compare the means of your two experimental groups using a statistical test called Student's t-test.

The t-test is a statistical test used to determine if the means of two data sets are significantly different. In statistical terms, the t-test is used to determine if the two data sets you collected come from the same or different distributions. The t-test can only be used when comparing means of two samples. More than two requires a different test.

To perform a t-test, one calculates a "t-value" from the two data sets you wish to compare. The t-value is a measure of the ratio of "signal" to "noise" in your data. The "signal" in the numerator represents the difference between the means. In other words, if the means of your two samples are very different then the "signal" is large.

$$t = \frac{\text{signal}}{\text{noise}}$$

The "noise" in the denominator represents the total amount of variation in both samples and can be found by summing the standard deviations for each of the data sets (i.e. the pooled variation) divided by each data set's sample size. Admittedly, this is kind of a tricky idea, however it makes sense when you think about it because if there is a great deal of variation in either or both of your data sets, then it should be more difficult to tell their means apart. This is the whole idea behind the t-test (as well as behind a large class of statistical tests called parametric tests). The equation for the t-test (assuming unequal variances) is thus

$$t = \frac{\text{mean \#A - mean \#B}}{\text{pooled variation}} = \frac{\overline{X_A} - \overline{X_B}}{\sqrt{\frac{\sigma^2_A}{n_A} + \frac{\sigma^2_B}{n_B}}}$$

Calculation of the t-value can be done by hand, on a calculator, or on a computer (such as the computer program MS-EXCEL). To calculate the t-value by hand, all that is required beforehand is that one know the sample sizes, means, and variances, σ^2, for each group. As one can see from the equation above, as the difference between the means of your groups gets bigger, the t-value gets bigger. Also, as the pooled variation gets smaller, the t-value will get bigger.

Now, the next question to ask is - 'How big does the "t-value" have to be in order for one to conclude that the means are "significantly different"? This is a really important question, and the answer is at the heart of all statistical analyses. The answer depends on two things - how large is your sample? and how "confident" do you want to be that the averages are in fact different?

The effect of sample size can be easily seen in the equation for the t-value above. Note that the sample sizes (n_A and n_B) appear in the denominator of the denominator. Thus, as the n's get larger, the pooled variation gets smaller, and as you recall, the effect of this is that the t-value gets larger. The other way to look at the sample size is more formal and involves a term called "degrees of freedom." The number of "degrees of freedom - df" for a t-test equals the pooled sample size minus 2 (however if variances are not equal a more complicated approximation method is used). The "df" tells you in effect how well you can resolve your averages given the t-value you calculate – the higher the "df" the greater the resolution.

The second issue mentioned above in determining how big a t-value you need depends on how confident you want to be. If you wanted to be really confident that these means differ, then you had better look for a very large t-value. But, if you are only satisfied with a rather marginal level of confidence (say one in 20 that you're wrong when you say they differ) then you would be happy with a smaller t-value. The confidence level is denoted in the test by the "P" value, which stands for probability. Probability is expressed as a decimal, P = 0.05 is the same as P = 5%. If you happened to do a stats test and get a P value of exactly 0.05, then there is a 95% chance that your averages differ, however there also is a 5% chance that if you conclude that the averages differ you are in fact wrong. The 0.95 cutoff really is the minimum "criterion for significance", however, you can be 99% sure if you hold out for a larger t-value and use a P = 0.01 as your "criterion for significance." The acceptable probability level is always determined BEFORE performing the t-test (most experimenters use P = 0.05).

The computer makes data analysis easy. Now it is time to consider an example. All you need is to enter the data below in a spreadsheet (which you already did to perform descriptive statistical analyses above), and perform a t test (following the directions on whatever spreadsheet or stats package is available to you.

data set A	data set B
3	1
3	1
4	2
5	3
6	4
6	4
6	4
6	4
7	5
8	6
10	8

t-Test: Two-Sample Assuming Unequal Variances

	Data set A	data set B
Mean	5.8181	3.8181
Variance	4.3636	4.3636
Observations	11	11
Ho Mean Difference	0	
df	20	
t Stat	2.2453	
P(T<=t) one-tail	0.0181	
t Critical one-tail	1.7247	
P(T<=t) two-tail	0.0362	
t Critical two-tail	2.0859	

Q – are the means for data set A and B significantly different, and exactly what information in the table above tells you this?