A Brief Review of Statistical Concepts Needed For This Course

Mean. The mean is the average of a sample. We use it to say something about the typical value of a member of the sample. It is the sum of all values divided by the number of values: $\sum X/N$. Usually we denote it with a bar over the top, e.g. $\overline{X}$ [also sometimes you will see μ, for mean]. Another way of expressing this is: $\sum p_i x_i$, where p_i is the proportion of the sample in category i and x_i is the value of the sample. For example, if we have two categories of objects that are equally represented and the members of the first category have a value of 10 while the member of the second category have a value of 14, then the average is: .5 (10) + .5 (14) = 12.

Variance. The variance of a sample is a measure of the spread of the values in the sample. We use it to measure the extent of difference among the members of the sample. $\sum(X- \overline{X})^2/N$. Notice that we square the deviations of each value, X, from the mean, $\overline{X}$. This means that the variance has to be a positive number. Since we divide by the sample size, N, we are taking the average of the squared deviation. This is why the variance is sometimes called the 'mean square' [it is the 'mean squared deviation]. The standard deviation just the square root of the variance [it is sometimes, particularly in engineering contexts, called the 'root mean square' or RMS]. Often you will see the sum of squared deviations, $\sum(X- \overline{X})^2$, divided by N-1 rather than N. This has to do with how to obtain the best estimate of the variance (which turns out to be better estimated by N-1) in a sample. When the sample size is large (i.e. N is big, it doesn't matter much). The variance is usually denoted by $ó^2$ or s^2 with the standard deviation, then ó or s. You will also see something such as, Var X, especially when more than one variance can be measured.

The mean and variance are both attributes of a single distribution, or sample. The next three concepts have to do with relations between two measurements.

Covariance. The covariance is a measure of how two measurements (X and Y) change with one another (or covary). It is quite similar to the variance: $\sum(X- \overline{X})(Y-\overline{Y})/N$. Instead of squaring one deviation from the mean, you multiply the deviation in one measurement from the deviation in another measurement. Again you sum over all pairs of measurements and divide by the number of observations. This is another type of average; the average deviation in two measurements. It is possible for this mean to be positive or negative. If unusually large values of X (bigger than the average) occur with unusually large values of Y while small values of X occur with small values of Y, then the product $(X- \overline{X})(Y-\overline{Y})$ will be positive (either two positive or two negative numbers). If large values of X are associated with small values of Y, the covariance will be negative. If there is no consistent relation between the value of X and the value of Y the covariance will be near zero. Covariance is usually denoted by cov_{XY}.

Regression. In regression we are trying to predict the value of one variable by measuring another one. By convention we usually talk about measuring X and estimating Y. This is referred to as a regression of Y on X. The regression coefficient is calculated by cov_{XY} / Var X. The regression coefficient is the slope of the 'best-fitting' line relating the values of Y to the X's. There will be a positive slope with a positive covariance and a negative slope with a negative covariance.

Correlation. In correlation we are trying to express the relationship between two variables without trying to say that one measurement causes the other. The formula for the correlation coefficient is: $\text{cov}_{XY} /s_X s_Y$. The denominator is the product of the two standard deviations. The correlation will be positive or negative depending on the sign of the covariance. The correlation gives the same information as the covariance but the correlation is standardized so that it will always be in the range from -1 (a perfect negative correlation) to +1 (a perfect positive correlation).

Statistical Testing. We won't be doing much statistical testing of hypotheses in this course, but there is at least one method which you should be familiar with.

Chi-square. The chi-square test is one that you have probably encountered before, but you should re-familiarize yourself with it. The basic question is whether the number of items in various categories is different from expectations. We have observed numbers of items and expected numbers of items. The chi-square statistic is: $\sum(O-E)^2/E$ where O is the number of items observed in a category and E is the number of items expected in that category. For example, we may have three categories with 10, 20 and 15 items. These are the observed values. By some other method we know that the expected numbers are 12, 24 and 9. [How you get those expected numbers is a separate subject.] Then the chi- square statistic is $(10-12)^2/12 + (20-24)^2 /24 + (9-15)^2 /9 = 5.0$. Obviously, if the observed values are near the expected values, we will not be too surprised by value of the Chi-square. However, if the observed and expected values are very different, we will be more surprised. The significance level is a measure of our degree of surprise in a statistic. Obviously, if we have more categories we will tend to get larger values for the chi-square (we are just adding up more numbers), so we need to account for this effect. The 'degrees of freedom' accounts for this effect. With three categories, you have 2 degrees of freedom (with n categories you would have n-1 degrees of freedom). To find out if this is a surprisingly large value, you look up the value of a chi-square statistic for 2 degrees of freedom. When we look this up in a book of statistical tables we find that we expect to get a chi-square this large or larger somewhere between 5% and 10% of the time ($0.1 > p > 0.05$). Often, we use an arbitrary significance level of $p < 0.05$ in order for a value to be called 'significantly different.' It is important to realize that this is a convention that indicates our relative level of surprise at a result. This test is used, for example, when we are trying to decide if a set of genotypes is in Hardy-Weinberg equilibrium.