AN INTRODUCTION TO BIOSTATISTICS

This packet is designed to introduce you to the use of some key statistical measures in analyzing and drawing conclusions from experimental data in biology. In real scientific work, simply using your judgement in deciding whether or not experimental data show a cause and effect relationship is insufficient. Personal judgement is notoriously subjective and very prone to bias. Statistics provide an objective means for assessing what the data from an experiment do and do not show. Any scientific work you do will be made much more sound by the application of appropriate statistical methods.

What is a statistic? To explain the term, we need to first consider what is called a *population*. A population is a group of individuals: the individuals could be people, animals, batteries, peanuts from a crop, bolts, or just about anything else. A *parameter* is a characteristic or property of the individuals of the population. For example, if the population under consideration is all two-week-old babies in the city of Denver, one parameter of this population would be weight. Another would be height, and so on.

If for some reason we are studying the weight of young infants in Denver, we probably would first like to find their average weight. One way to do this would be to obtain the weights of absolutely every single young baby in Denver and then compute the average weight. However, this could prove to be a formidable task since there are so many babies -- we could be on the phone forever seeking medical records. But we might instead examine a smaller number of young infants and suppose that whatever average weight we determine from them is essentially the same as the average weight of the whole population of babies. This smaller group of babies that serves as an indicator for the whole population is called a *sample* and any property we obtain from the sample (such as weight) is a *statistic*.

A big part of statistics is assessing how reliably a characteristic obtained from a sample represents the entire population -- there certainly is no guarantee that what is true for the sample is also true for the population. Obtaining a truly representative sample is often the key step in determining a statistical description of a population. Suppose, for example, that you are testing the sugar content of the apples grown in an orchard. If you took 10 apples from the southeast corner of the orchard and tested their sugar content, what are the chances that the average sugar content of these apples is accurately representative of the average sugar content of the hundreds of apples in the orchard? It certainly is possible that the soil in this corner of the orchard is not very fertile and the apples growing on the trees there are sour. By measuring their sugar content, we would get a poor measure of the typical sweetness of the entire crop.

To avoid obtaining an inaccurate statistic from a sample, we must be very careful in properly selecting our sample. The first thing we can do is to examine as large of a sample as we can practically manage. The more individuals we use in our sample, the more likely it is that the sample represents the population well. This is why scientists like to gather as much data as they can in doing experimental work. Another thing we can do is to randomize our sampling

procedure, that is, randomize how we select individuals from the population to comprise our sample. Continuing with our apples in the orchard as an example, we might number the trees. We could throw dice or use some other random event to generate numbers which tell us which trees to take apples from. Then we could take several apples from different parts of the selected trees for analysis. The general process of sampling in statistics is so important that a great deal of literature is devoted to it.

In determining a statistic from a sample, we generally find that a series of different values is obtained. There is some lowest value and some highest value and values in between. The difference between the lowest value and the highest value is called the *range*. The middle of the range can be described by one of three different quantities. Called measures of central tendency, these quantities are the *mean, median,* and *mode* of a sample. The mean is the numerical average of the values obtained from the sample. The median is the middlemost values and the mode is the value which was obtained most often. For example, suppose we "measure" the age in years of a group of randomly selected students. We poll fifteen students with these results:

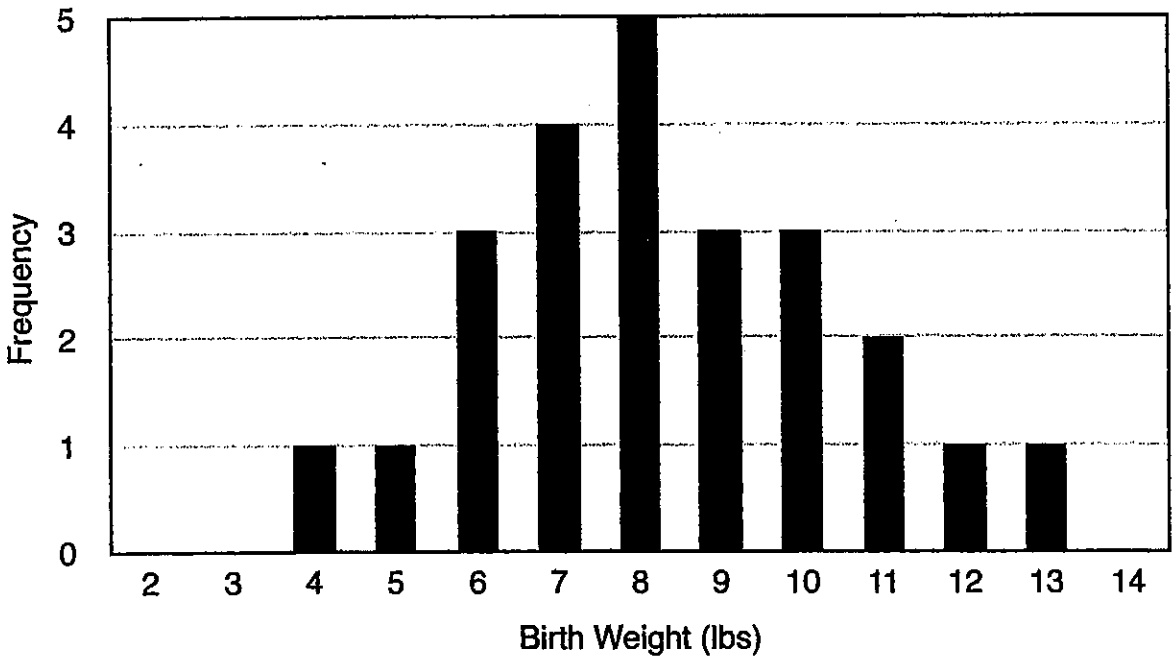
Student	Age (years)
1	15
2	13
3	15
4	16
5	15
6	15
7	13
8	12
9	13
10	15
11	15
12	16
13	16
14	14
15	14

The mean age of this sample of students is 14.47 years (obtained by adding up all of the ages and dividing by 15). The median age is 14 years since 14 is in the middle of the range of ages, 12...13...14...15...16. The mode of the ages is 15 years because 15 is the most frequently observed age in the sample.

When we obtain a series of values from a sample, it is useful to determine the frequency with which various values occur. When this information is graphed as a value on the x axis and the number of times that value was obtained (frequency) on the y axis, a *frequency distribution plot* results. When the variation in the values in the sample occurs mainly from chance alone, frequency distribution plots are fairly symmetrical bell-shaped curves. The logic behind such curves can be seen by returning to a simplified form of our example of young infant weights mentioned earlier. Say we sample 24 babies for their birth weights to the nearest pound and we obtain:

birth weight (lbs) (the value)	# of babies with that weight (the frequency)
2	0
3	0
4	1
5	1
6	3
7	4
8	5
9	3
10	3
11	2
12	1
13	1
14	0
	24 total

A frequency distribution for these values drawn as a bar graph looks like this:



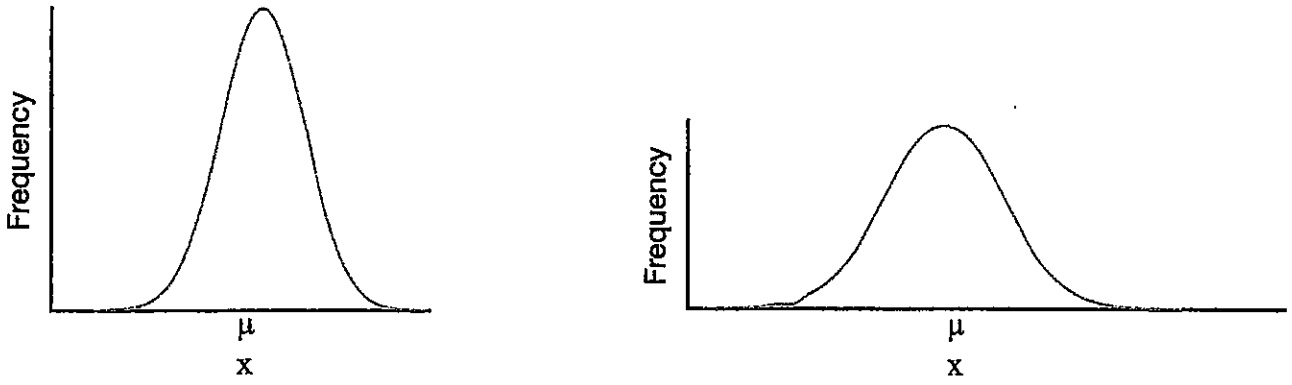
Notice how the distribution peaks in the middle and tails off at the either end. This simply means that most babies have weights near the average while very few babies have exceptionally high or exceptionally low weights. This type of bell-shaped behavior is very common for a distribution of values -- values near the average (middle) are more frequently observed than values much higher or lower than the average.

When probability alone is responsible for the variation of a parameter in a very large population (actually an infinitely large population), a special curve describes the frequency at which various values of the parameter are observed. This curve is called the *normal curve* and the frequency distribution it describes is called a *normal distribution.* The equation for the normal curve is:

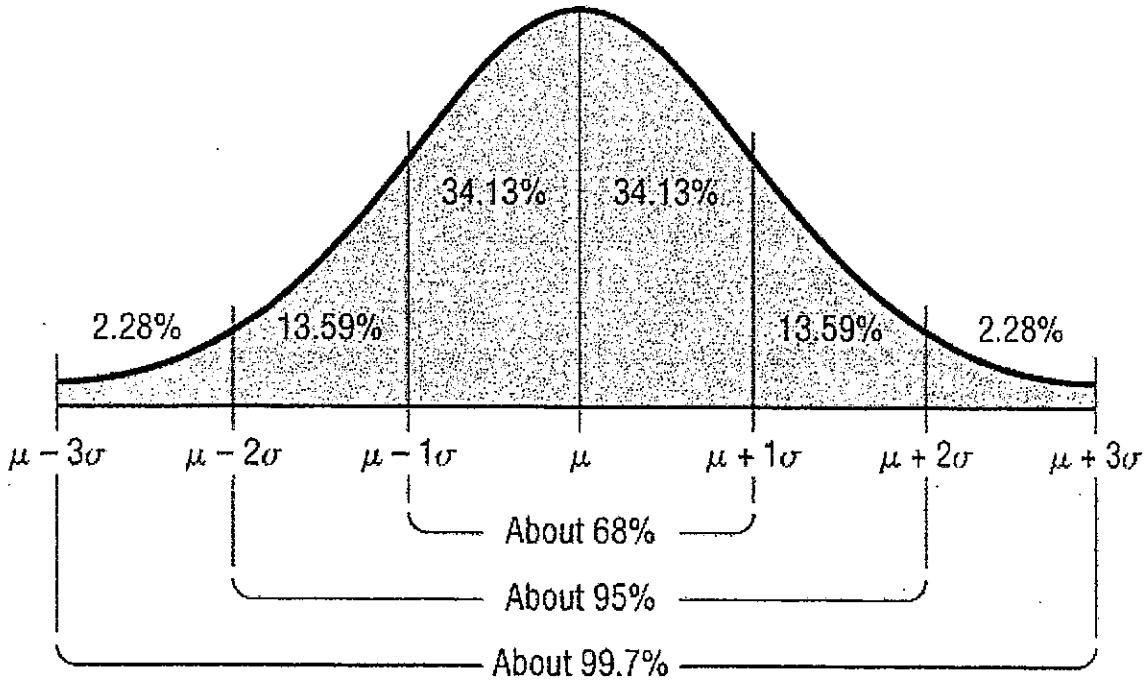
$$y = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$$

In this equation, y is the frequency of occurrence of a given value x (y represents how often the value x will be observed). The constant e is the base for natural logarithms, 2.718. The terms σ and μ have a very special importance. σ is called the *standard deviation* of the population while μ is the mean of all of the values of x observed for the population. Examine the normal curves shown on the next page. Note how both curves are symetrically centered about the mean value, μ (it makes sense that the most frequently observed value of x would be the mean value so the

curve peaks at μ). The standard deviation of each curve measures the spread or range of the values. The curve at left has a smaller range (the values are crowded more closely around the mean) and therefore has a smaller standard deviation than the curve at right.



It can be shown that 68.3% of all the values of x lie within one standard deviation on either side of the mean in a normal distribution while 95.5% of the values of x lie within two standard deviations above or below the mean and 99.7% of the values lie within three standard deviations above or below the mean:



Standard deviation is often used to express the *precision* of a set of measurements or results. When we measure the same thing many different times, our results will not agree perfectly from trial to trial because of the errors inherent in any measurement. We might find that our results agree rather poorly from trial to trial in which case we say that the precision of the measurement is poor. On the other hand, the results might differ by very little from one to the next and we would say that the precision is good. When the precision is good, the spread or range of the values is small and so is the standard deviation of the values. Thus a set of values with a small standard deviation is said to be precise. To illustrate the point, we can consider the data below for the determination of the percent of silver in an alloy by two different methods:

Trial	Method 1	Method 2
1	18.5%	16.2%
2	18.3%	15.9%
3	18.8%	18.4%
4	18.3%	17.0%
5	18.4%	19.8%
6	18.2%	20.8%

Method 1 is the more precise method because the results vary less than do those of method 2. Therefore the standard deviation of the results for method 1 would be less than for method 2.

It is important to note that the standard deviation of a parameter for a population can only be estimated if only a sample of the population is examined. This is also true for the mean value of the parameter. In other words, the mean and standard deviation of a sample may or may not be close to the mean and standard deviation for the population depending on how representative the sample is. Thus the mean and standard deviation of a sample are statistics. We use the symbols "s" and "$\bar{X}$" for the standard deviation and mean of a sample, respectively, while σ and μ are the corresponding symbols for the population.

A formula exists for finding the standard deviation of a sample. For samples with relative few values, this formula is:

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}}$$

In words, this formula for s says to sum the squares of the differences between the values and the mean value, divide by one less than the number of values, and take the square root. To illustrate

the use of this formula, we can use the data above for the percent of silver in an alloy. For method 1, the pertinent calculations are:

value	$(X-\bar{X})^2$
18.5%	0.006889
18.3%	0.013689
18.8%	0.146689
18.3%	0.013689
18.4%	0.000289
18.2%	0.047089

Sum = 110.5% **Sum = 0.228334**

$\bar{X}$ **= 110.5% ÷ 6**

$\bar{X}$ **= 18.417%**

s = (0.228334/(6-1))$^{1/2}$ = 0.2137%

Thus we could report the results of the analysis for the percent of silver as 18.41% +/- 0.21% where the +/- 0.21% tells us that roughly two-thirds of our results (ideally 68.3% of the results) fell within a range of 0.21% below or 0.21% above the mean of 18.42% silver. Note that it is customary to only keep one or two significant digits in the standard deviation and to round the mean result off to the decimal place where the standard deviation is rounded to. When we report both the mean result and the standard deviation, we not only indicate our best estimate of the quantity we're after (the mean value) but, knowing that any experimental result is never perfect, we also give an indication of just how uncertain we might be in stating the mean.

Exercise #1: On a separate sheet of paper, go through an analysis of the standard deviation of the second method for the percent of silver in the alloy and properly report the final result as we did above for method 1. Finally, explain any differences between the standard deviation for method 2 and the standard deviation for method 1. What does this imply about method 2 compared to method 1?

Departures from the Normal Distribution

When a frequency distribution plot for a measurement of a given population is prepared, it may or may not resemble a normal distribution curve. This is particularly true for small samples: random fluctuations can make the plot quite asymmetrical or jagged. As the size of the sample is made larger, the fluctuations tend to smooth out. However, it is also possible that factors other than probability affect the distribution of values in the sample (and the population). In such cases, the distribution curves can be asymmetrical (lean toward one side or tail more quickly on one side) or can be either more peaked or flattened than a true normal curve. These departures from a normal curve can be described with measures called *skewness* (asymmetry) and *kurtosis* (more peaked or flattened). We will not deal with these concepts here since they are probably more advanced than you will require in your scientific work at this point. Skewness and kurtosis are described in readily available references and can be researched if the need arises.

Indeterminate and Determinate Errors

Regarding analytical work, that is, work where you have tried to determine some quantity like the percent of silver in an alloy, it is also important to note the difference between indeterminate and determinate errors. *Indeterminate errors* are those of a completely random nature. Owing to their origin in chance, indeterminate errors are equally likely to lead to a high or low result. Thus indeterminate errors are responsible for producing a normal distribution in a series of repeated measurements of the same quantity. *Determinate errors*, on the other hand, are systematic -- they occur in one direction only. The effect of a determinate error is to move the normal curve to the left or right of the true value for the measurement so that the experimental mean, $\bar{X}$, no longer coincides with the true mean, μ, even for very large samples.

In principle, determinate errors can be corrected for although in practice this may prove difficult. An example would be a chemical analysis where it is always necessary to add a little too much of a reagent to get a color change to signify the end of a reaction. By carrying out repeated analyses on samples of known composition, it is possible to determine the amount of extra reagent needed and to correct for this error in experiments involving true unknowns.

One of the most difficult determinate errors to correct for is personal error. However, in some cases it can be done. In field studies, for example, biologists may need to estimate the time that an animal spends at a specific food source. While some observers might have anxious personalities and will always make a time estimate that is too short, other biologists might be more likely to generate an estimate that is too long. By comparing the results of many observers or by having observers carry out timings of known biological phenomena, recorders of biological societies can develop a personal error correction for each of the contributing observers and use it to adjust submitted data before compilation.

Indeterminate errors in a measurement affect the precision of the results. The precision, you recall, is indicated by the standard deviation of the set of results and is graphically visible in the width of the distribution curve -- a wide frequency distribution indicates poor precision and vice versa. The operation of determinate error, as we said earlier, shifts the experimental mean away from the true mean. This shift can be measured by what we call *percent error*. You can compute percent error according to:

$$\% \text{ error} = \frac{\text{experimental - accepted}}{\text{accepted}} \times 100$$

Note that % error is a measure of the *accuracy* of a result -- how close the experimental mean is to the "true" mean. Of course, such a calculation requires an accepted value. This may be obtained from the literature or perhaps a theoretical calculation. In real experimental work where new areas are under investigation, an accepted value may not be available.

Confidence Intervals

When a standard deviation has been determined for a sample, we can use it to indicate the spread of the results, as we saw for the example involving the percent of silver in an alloy. The spread of the results gives us an indication of the range within which we think the true value lies. For example, we know that 68.3% of the results are expected to fall within one standard deviation on either side of the mean in a normal distribution. However, we may wish to go even further in specifying the range within which we think the true value actually lies. We can use the standard deviation to determine what is called a *confidence interval.*

In calculating a confidence interval, we first must select a desired level of confidence, perhaps 95%. We can then use statistical tables of what are called t values to determine a range about the mean within which we are 95% sure the true value lies. In using the t tables, it is necessary to understand what we call *degrees of freedom* in a set of measurements.

A degree of freedom is defined to be any measurement that contributes independently to a set. Suppose, for example, that we have the set 2, 4, 8, 6, 10. This set has five independent measurements. But if we use the numbers to calculate a mean value as we often do in statistics, this extracts one degree of freedom. Here's why: The mean of 2, 4, 8, 6, and 10 is 6 since

$$\frac{2+4+8+6+10}{5} = 6$$

Once this mean has been found, all it takes is any four numbers from the set of five to find the fifth number. For example, suppose have 2, 4, 8, 6, and z where $\bar{X} = 6$. We'd have

$$\frac{2+4+8+6+z}{5} = 6 \quad \text{so } z = 30 - (2+4+8+6) = 10$$

Thus, once the mean is found it only takes four of the five numbers from the set to completely specify the set. As a result, we say that the set has four degrees of freedom. *Any set of n numbers with a mean of $\bar{X}$ will have n-1 degrees of freedom.*

Returning to the confidence interval, the formula we use to find the interval is

$$\text{confidence interval} = \bar{X} \text{ +/- (t) (s)/n}^{1/2}$$

where s is the standard deviation of the set of n measurements, $\bar{X}$ is the mean of the set and t is a value we obtain from a t table using the degrees of freedom and the desired level of confidence. The plus or minus notation means that the range within which the true value lies at a certain level of confidence is given by adding or subtracting the value given by the formula to or from the mean. The best way to illustrate all of this is to do an example, so we'll return to our analysis of the percent of silver in an alloy that we discussed earlier.

Suppose we would like to find the range about the mean for method 1 within which we are 95% confident that the true percentage of silver lies. We first recognize that there are six measurements and hence five degrees of freedom. Examining the table of t values (see the appendix), we see that t = 2.57 at the 95% level of confidence (also labeled the 0.05 level) and five degrees of freedom. Recalling that the standard deviation for method 1 was 0.21%, we can now find the confidence interval:

$$\begin{aligned}\text{confidence interval} &= \bar{X} \text{ +/- (t) (s)/n}^{1/2} \\ &= 18.42\% \text{ +/- } (2.57)(0.21\%)/(6)^{1/2} \\ &= 18.42\% \text{ +/- } 0.22\%\end{aligned}$$

Thus we would say we are 95% confident that the true percentage of silver in the alloy falls in the range 18.20% to 18.64%. You might wonder why we would go to all this trouble since the confidence interval is so similar to the standard deviation. The answer is that this is not always the case. Suppose that all we desire is the 90% level of confidence in our interval. Now we have:

$$\begin{aligned}\text{confidence interval} &= \bar{X} \text{ +/- (t) (s)/n}^{1/2} \\ &= 18.42\% \text{ +/- } (2.02)(0.21\%)/(6)^{1/2} \\ &= 18.42\% \text{ +/- } 0.17\%\end{aligned}$$

Note that in this latter calculation, there is a new t value since we are using a different level of confidence. The confidence interval is no longer nearly the same as the standard deviation. Also, notice how the interval is narrower for the 90% level than for the 95% level. This is because if we demand a higher level of confidence that the true value lies in the specified range, we have to provide a wider range to make sure we encompass the true value. If we choose a very high level of confidence for our interval, say 99.9% (the 0.0001 level), the interval gets significantly broader:

$$\text{confidence interval} = \bar{X} \text{ +/- (t) (s)/n}^{1/2}$$
$$= 18.42\% \text{ +/- (6.86)(0.21\%)/(6)}^{1/2}$$
$$= 18.42\% \text{ +/- 0.59\%}$$

Commonly, confidence intervals are determined at the 90% or 95% level in routine experimental work.

Exercise #2: On a separate sheet of paper, work out confidence intervals for method 2 in the percent of silver analysis: do both the 90% and the 95% levels. You will need your standard deviation for method 2 that you calculated earlier. You can obtain t values from the appendix.

The T Test

In doing real experimental work, it is commonly necessary to compare the behavior of an experimental group to a control group. Deciding whether or not a true difference exists between the experimental group and the control group is best done statistically and not by arbitrary judgement. The test that is commonly employed for this purpose is called the *t test* which, of course, requires a table of t values.

Perhaps the best way to explain the t test is to just start with an example of how it is done. Say we grow grass in a magnetic field as the experimental group and grass in no magnetic field as the control group. Random sampling of the lengths of the grass blades in each group yields:

length of blades of grass	
control (c)	experimental (e)
10.5 cm	5.8 cm
11.2 cm	7.2 cm
9.8 cm	5.2 cm
10.0 cm	6.1 cm
12.1 cm	8.0 cm
9.9 cm	6.2 cm
9.6 cm	5.4 cm
	5.9 cm

Looking a these data we might suspect that the effect of the magnetic field is to suppress the growth of the grass blades. Before we can draw such conclusions, however, we need to do a bit of number-crunching. First, we find the mean and the standard deviation of each set:

c	e
$\bar{X}$ = 10.44285714 cm	$\bar{X}$ = 6.225 cm
s = 0.907114735 cm	s = 0.936177638 cm
$\bar{X}$ = 10.44 cm +/- 0.91 cm	$\bar{X}$ = 6.22 cm +/- 0.94 cm

We can now proceed to assess whether or not the mean growth of the grass blades is significantly different between the two groups. In other words, we want to know if the magnet really stunted the growth of the experimental group. Even though the blades seem shorter in this group, we must objectively assess how possible it is that the apparent effect is just due to chance alone. This is where the t test comes into play.

Technically, a t test should only be employed when there is no real difference in the standard deviations of the experimental and control groups, meaning that a different kind of t test would be required if the two groups in question grew with very different amounts of variability. The t test, however, is said to be "robust", which means that it is fairly insensitive to differences in the standard deviations of the two groups in question. In most situations, then, the data can be pooled to provide a single estimate for the standard deviation of both sets together. Thus only very large differences in the standard deviations of the two groups would make this second version of the t test necessary.

In pooling the data to get a single estimate of the standard deviation for the experimental and control groups, we see that one degree of freedom was used up in computing the mean of each test. Thus the pooled set has two fewer degrees of freedom than the total number of measurements. That is, the degrees of freedom for the pooled data equal the total number of measurements minus two. The pooled standard deviation is then given by:

$$s^2 = \frac{s_1^2(N_1 - 1) + s_2^2(N_2 - 1)}{N_1 + N_2 - 2}$$

in which s^2 is the variance and s may be obtained as the square root of the variance.

Once s^2, the variance, is obtained, a t value is calculated according to:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

We now make the assumption that the means of the two data sets are *not* significantly different -- we adopt the null hypothesis. This would mean that the experimental variable (the magnetic field in our example) had no real effect and that any observed differences were just due to chance. Then we read a critical value of t from the table of t values at the appropriate degrees of

freedom and the desired level of confidence. If the computed t is less than the t(crit) from the table, then at the level of confidence chosen, the means are not significantly different. But if the computed t exceeds the t(crit), the means *are* significantly different -- the difference is real, not a fluke -- at the chosen level of confidence. If, for example, the computed t exceeds t(crit) at the 0.05 level, we would say that there is only a 5% chance that the means are not really different and a 95% chance that the means are significantly different.

Returning to our grass blades,

degrees of freedom = $N_1 + N_2 - 2 = 7 + 8 - 2 = 13$
where N_1 is the control group and N_2 is the experimental group

$$s^2 = \frac{s_1^2(N_1 - 1) + s_2^2(N_2 - 1)}{N_1 + N_2 - 2} = \frac{0.9071^2(7-1) + 0.9362^2(8-1)}{7+8-2} = 0.8517$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2}{N_1} + \frac{s^2}{N_2}}} = \frac{10.4429 - 6.2250}{\sqrt{\frac{0.8517}{7} + \frac{0.8517}{8}}} = 8.83$$

From a table of critical t values (see appendix), we find for 13 degrees of freedom that t(crit) = 2.160 at the 0.05 level and t(crit) = 3.01 at the 0.01 level. Since the computed t of 8.83 is greater than t(crit) at the 0.01 level, there is less than a 1% chance that the means are really the same and greater than a 99% chance that the means are significantly different. Thus we can say with a high level of confidence that the magnetic field had an effect on the growth of the grass blades in the experimental group.

We should point out one additional consideration in using the t test. Technically, the t test can either be a one tailed or two tailed test. The two tailed test is employed when the experimental mean could turn out to be either higher or lower than the control mean. The one tailed test is used with the experimental mean can only be observed in one direction relative to the control mean. A one tailed t test, for example, would be appropriate when the experimental mean could only be higher than the control mean.

Exercise #3: An experiment is performed at CCHS where air is sampled for bacteria in a standardized fashion in the gym and the cafeteria. The technique consists of exposing agar plates, incubating them, and counting bacterial colonies that grow on the agar. A series of replicate samples is taken for each location with the following results:

# of bacterial colonies		
sample	**gym**	**cafeteria**
1	80	98
2	75	94
3	77	89
4	85	89
5	76	91
6	87	87
7	—	96
8	—	97

Is there a significant difference in the airborne bacterial populations in the gym and the cafeteria? Perform a t test to see if there is a significant difference between the mean bacterial growth (# of colonies) in the two groups. Again, be sure to mention the level of confidence you selected in describing the results of the t test. Show your work clearly!

The Paired T-Test

Even though in a classic experimental set-up, we're comparing the mean of a control group to the mean of an experimental group through the use of a t-test, there are often situations in which no such comparison can be made. In other words, it's a situation where there is no formal control group because the treatment effect is being studied within individuals (as opposed to between individuals).

As an example, the effect of certain diet programs *could* be studied by comparing a large and randomized group of people who are not on the diet to a large and randomized group of people who are, but the treatment effect could be made more significant by comparing the pre-test weights of all of the participants before the diet to the post-test weights of these same people after the diet. In other words, an experimental set-up that compares individuals to themselves before and after treatment can be much more statistically powerful. This is called a *repeated measures* design and its power is due to small treatment differences within individuals being superimposed on large, stable differences between individuals.

Let's use the well-known "all bacon diet" as an example of the use of the paired t-test. In our hypothetical study, a group of twelve subjects weighed-in at the beginning of the study, then were subjected to six week's worth of nothing but bacon for breakfast, lunch, dinner, and all of the snacks in-between. At the end of the six weeks, the twelve participants were weighed once again. The results are shown below:

Pre-test and Post-test weights of 12 "all-bacon" dieters			
Subject	**Pre-Test**	**Post-Test**	**Difference**
1	65	62	-3
2	88	86	-2
3	125	118	-7
4	103	105	+2
5	90	91	+1
6	76	72	-4
7	85	81	-4
8	126	122	-4
9	97	95	-2
10	142	145	+3
11	132	132	0
12	110	105	-5
Mean	**103.3**	**101.2**	**-2.08**
SD	**24.0**	**24.8**	**3.03**

You'll note that the mean and standard deviation have been included in the table, not just for the pre-test and post-test data, but also for the difference scores (themselves having been calculated by subtracting the pre-test weight from the post-test weight). You will also note that the standard deviation of the pre-test and post-test weights is quite large, reflecting the large, stable differences between individual people in the group... and that the standard deviation of the difference scores is quite small, reflecting the small between-subject variation in the treatment effect.

We are now left wondering whether or not the difference scores reflect a significant or insignificant treatment effect. This is where the paired t-test comes in. Our null hypothesis in this case is that the all-bacon diet had no effect on the weights of our subjects. In more specific terms, the null hypothesis is that the true difference between pre-test and post-test weights is zero. In our example, however, the closest approximation we have for this true difference is 2.08 lbs (the mean of the differences). The question we must ask, then, is this: What is the likelihood that a difference of 2.08 lbs or greater could have occurred by chance in a sample size of 12 drawn from a population with a mean difference of 0 and a standard deviation of 3.03? To answer this question, we must generate a t value that reflects the chances that our result is just due to chance, and then compare this to the one-tailed t values in the appendix, as in the unpaired t-test. Here's our equation for the paired t-test:

$$t = \frac{d}{s/\sqrt{n}}$$

Here, the "d" is our observed difference between pre-tests and post-tests, or more specifically, the mean of the difference scores. The "s" is the standard deviation of the difference scores and "n" is the number of individual difference scores. Below is this formula put to use with the data from our all-bacon diet experiment:

$$t = \frac{d}{s/\sqrt{n}} = \frac{2.08}{3.03/\sqrt{12}} = 2.38$$

The t(crit) of a one-tailed t-test with 11 degrees of freedom (12 data - 1 mean) at the 0.05 level is equal to 1.80. Since the computed t of 2.38 is greater than the t(crit) at the 0.05 level, there is less than a 5% chance that the result we've obtained for the differences in pre-test and post-test weights is due just to chance and greater than a 95% chance that the differences are significant. So, the all-bacon diet worked!

It's important to note, also, that in some pre-test/post-test situations the paired t-test is not the best way to measure the significance of differences. In some situations, a control group is compared to an experimental group where pre-test and post-test data is obtained for both groups. In such cases, an *unpaired* t-test can be used on the *difference* scores (instead of on the means of the two groups).

The Chi-Square Test

The F and t tests are appropriate when the samples they describe are from normally distributed populations. In addition, these tests are only appropriate for variables that are continuous. What do we do when we have discrete categories rather than values that can vary continuously? For example, suppose a football referee comes to us with a coin he uses at the beginning of games for the coin flip. He tells us he suspects that when the coin is tossed, it does not land with an equal probability of being heads or tails -- in effect, the coin might "loaded" like weighted dice. To test this idea, we do an experiment. We toss the coin numerous times and record the number of heads and tails. Now, the outcome of each trial is a discrete variable: each toss can either be heads or tails but nothing else in between. Thus a t test could not be used to see if there is a significant difference in the number of heads vs the number of tails.

The statistical test that we must use to analyze our coin toss data is called a *chi-square test*; this is the test that permits ready analysis of discrete variables. In applying the chi-square test, it is generally necessary to have an expectation about the outcome of the experiment. The chi-square formula is based on comparing the expected results with the actual results of an experiment. In the case of the coin, we might expect it to land heads and tails with equal frequency, that is, we believe the coin is unbiased. Thus we predict that for 100 tosses of the coin, 50 will be heads and 50 will be tails. Now we carry out 100 tosses of the coin; suppose we get the results below:

# heads	# tails	total
57	43	100

We notice that the coin landed heads more frequently. Does this mean that the coin is "loaded"? Perhaps, but it is also possible that the uneven distribution of heads and tails is strictly due to chance. Calculating a chi-square gives us a way of statistically deciding if there is a significant difference between what we expected (a 50/50 distribution) and what we got (a 57/43 distribution). The chi-square formula is

$$x^2 = \sum \frac{(\text{observed - expected})^2}{\text{expected}}$$

Note that there will be a term in the summation for each category or attribute the variable can be assigned to; in the case of the coin, there are two categories, so there will be two terms in the chi-square test. For our coin experiment, chi-square is

$$x^2 = \frac{(57-50)^2}{50} + \frac{(43-50)^2}{50} = 1.96$$

We now go to a table of critical values of chi-square (see appendix). We must know the degrees of freedom, which are not determined as for a t test. For chi-square, the degrees of freedom

equal the number of attributes minus one (this is the number of summed terms in the chi-square equation minus one). For the coin, the degrees of freedom equal 2-1 or 1. Now, a greater value of chi-square indicates more of a discrepancy in our expected versus obtained results. We start off by assuming the null hypothesis -- we assume that there is no difference between the expected and actual results. The smaller the chi square, the more likely it is that we will fail to reject the null hypothesis. In the case of the coin, the chi-square of 1.96 at 1 degree of freedom falls between the 0.05 level and the 0.10 level according to the critical chi-square values in the table. This tells us that there is more than a 5% (1 in 20) chance that this high of a chi-square is just a fluke and the expected and actual results were really not significantly different. We would probably accept these odds as sufficient for failing to reject the null hypothesis. As far as we can tell, there is no clear difference between what we expected and what we got. Since what we expected was based on the assumption that the coin was not "loaded", we find no compelling evidence that the coin is biased, at least as far as this experiment is concerned.

In reviewing the results of our 100-toss experiment with the coin, we might decide that a chi-square of 1.96 at 1 degree of freedom is on the borderline. Maybe we aren't satisfied with between a 5% and 10% chance that the difference between the expected and actual results was a fluke. We would like an even smaller chi-square to be more confident in concluding that the coin is unbiased. One thing we could do is repeat the experiment with many more trials -- there is safety in numbers because statistical fluctuations tend to smooth out in large samples. So, we toss the coin 1000 times with, say, these results:

# heads	# tails	total
570	430	1000

Now we get a chi-square of 19.6:

$$x^2 = \frac{(570-500)^2}{500} + \frac{(430-500)^2}{500} = 19.6$$

At 1 degree of freedom, the chi-square table tells us that there is less than 1 chance in 1000 (the 0.001 level) that this high of a chi-square is just a fluke. The chi-square is so large that we reject the null hypothesis with a high degree of certainty and conclude that the actual results were in fact significantly different than the expected results. The coin has turned out to be biased after all. The reason that this conclusion was not reached in the experiment with 100 trials is that small samples are expected to show more random fluctuations. The percentage of heads was 57% in both experiments, but out of 1000 trials a deviation of 7% from a 50/50 distribution is not

nearly as likely to be random as in 100 trials. You can see that it is important to obtain as many data as you can to make it easier to draw clear conclusions.

Exercise #4: A horticulturist is performing experiments in inheritance in certain flowering plants. She has hypothesized a certain mechanism for the way flower color is genetically carried to offspring from cross-pollinated parent plants. In one experiment, her theory predicts that offspring from two cross-pollinated plants should be one-fourth red, one-fourth white and one-half pink. 200 seeds from the cross-pollinated plants are planted and the color of the flowers of the offspring are noted. The actual results are:

red flowers	43
pink flowers	110
white flowers	47
total	**200**

Do these data convincingly support the horticulturist's inheritance theory? Perform a chi-square test on these data to see. Show all pertinent work and be sure to mention levels of confidence in your answer. Remember that there is a table of critical chi-square values in the appendix.

TABLE C

Critical values for the *t*-test

	α for a 1-tailed test					
	.10	.05	.025	.01	.005	.0005
	α for a 2-tailed test					
df	.20	.10	.05	.02	.01	.001
1	3.078	6.314	12.706	31.820	63.656	636.615
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.141
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.849
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.768
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
35	1.306	1.690	2.030	2.438	2.724	3.591
40	1.303	1.684	2.021	2.423	2.704	3.551
45	1.301	1.679	2.014	2.412	2.690	3.520
50	1.299	1.676	2.009	2.403	2.678	3.496
55	1.297	1.673	2.004	2.396	2.668	3.476
60	1.296	1.671	2.000	2.390	2.660	3.460
70	1.294	1.667	1.994	2.381	2.648	3.435
80	1.292	1.664	1.990	2.374	2.639	3.416
90	1.291	1.662	1.987	2.368	2.632	3.402
100	1.290	1.660	1.984	2.364	2.626	3.390

TABLE F

Critical values for the chi-squared test

df	α = .10	.05	.025	.01	.005	.001
1	2.706	3.842	5.024	6.635	7.879	10.828
2	4.605	5.992	7.378	9.210	10.597	13.816
3	6.251	7.815	9.348	11.345	12.838	16.266
4	7.779	9.489	11.143	13.277	14.860	18.467
5	9.236	11.071	12.833	15.086	16.750	20.515
6	10.645	12.592	14.449	16.812	18.548	22.457
7	12.017	14.067	16.013	18.475	20.278	24.321
8	13.362	15.507	17.535	20.090	21.955	26.124
9	14.684	16.919	19.023	21.666	23.589	27.877
10	15.987	18.307	20.483	23.209	25.188	29.588
11	17.275	19.675	21.920	24.725	26.757	31.264
12	18.549	21.026	23.336	26.217	28.299	32.909
13	19.812	22.362	24.736	27.688	29.819	34.528
14	21.064	23.685	26.120	29.141	31.319	36.123
15	22.307	24.996	27.488	30.578	32.801	37.697
16	23.542	26.296	28.845	32.000	34.267	39.252
17	24.769	27.587	30.191	33.409	35.718	40.790
18	25.989	28.869	31.526	34.805	37.156	42.312
19	27.204	30.144	32.852	36.191	38.582	43.820
20	28.412	31.410	34.170	37.566	39.997	45.314
21	29.615	32.671	35.479	38.932	41.401	46.797
22	30.813	33.924	36.781	40.289	42.796	48.268
23	32.007	35.172	38.076	41.638	44.181	49.728
24	33.196	36.415	39.365	42.980	45.558	51.178
25	34.382	37.652	40.647	44.314	46.928	52.620
26	35.563	38.885	41.924	45.642	48.290	54.052
27	36.741	40.113	43.195	46.963	49.645	55.476
28	37.916	41.337	44.461	48.278	50.993	56.892
29	39.087	42.557	45.723	49.588	52.336	58.301
30	40.256	43.773	46.980	50.892	53.672	59.703