Appendix II
Testing Statistical Hypotheses

Many biological measurements vary among members of a study population. These variations may occur for any number of reasons, from differences in genetic expression to the effects of environmental variables. But when an investigator is trying to use measurements of a particular item of interest, that means that it is not enough to measure only one and assume that this is the correct value for the entire population. Instead, the investigator must measure as many individuals as possible to account for the variation in that population.

When two populations are being compared to each other in an attempt to see whether a measured item differs between the two populations, care must be taken to ensure that each population is represented as accurately as possible, and that the two are compared in a rigorous, logical fashion. This is the purpose of statistical analysis.

I. Data, parameters and statistics

Most investigations in the biological sciences today are quantitative. The investigator's goal is to collect biological observations which can be tabulated as numerical facts, also known as **data** (singular = **datum**). Biological research can yield several different types of data:

1. Attribute data. This simplest type consists of descriptive, "either-or" measurements, and usually describe the presence or absence of a particular attribute. The presence or absence of a genetic trait ("freckles" or "no freckles") or the type of genetic trait (type A, B, AB or o blood) are examples. Because this type of data has no specific sequence, it is considered *unordered* data.

2. Discrete numerical data. These data correspond to biological observations which are counted, and are integers (whole numbers). The number of leaves on each member of a group of plants, the number of breaths per minute in a group of newborns or the number of beetles per square meter of forest floor are all examples of numerical discrete data. Although these data are ordered, they do not describe physical attributes of the things being counted.

3. Continuous numerical data. The most quantitative data fall along a numerical continuum. The limit of resolution of such data is the accuracy of the methods and instruments used to collect them. Examples of continuous numerical data are tail length, brain volume, percent body fat...anything that varies on a continuous scale. Rates (such as decomposition of hydrogen peroxide per minute or uptake of oxygen during respiration over the course of an hour) are also numerical continuous data.

When you perform an experiment, be sure to determine which type of data you are collecting. The type of statistical test appropriate in any given situation depends upon the type of data!

When an investigator collects *numerical* data from a group of subjects, s/he must determine how and with what frequency the data vary. For example, if one wished to study the distribution of shoe size in the human population, one might measure the shoe size of a sample of the human population (say, 50 individuals) and graph the numbers with "shoe size" on the x-axis and "number of individuals" on the y-axis. The resulting figure shows the

frequency distribution of the data, a representation of how often a particular data point occurs at a given measurement.

Usually, data measurements are distributed over a range of values. Measures of the tendency of measurements to occur near the center of the range include the population **mean** (the average measurement), the **median** (the measurement located at the exact center of the range) and the **mode** (the most common measurement in the range).

It is also important to understand how much variation a group of subjects exhibits around the mean. For example, if the average human shoe size is "9," we must determine whether shoe size forms a very wide distribution (with a relatively small number of individuals wearing all sizes from 1 - 15) or one which hovers near the mean (with a relatively large number of individuals wearing sizes 7 through 10, and many fewer wearing sizes 1-6 and 11-15). Measurements of dispersion around the mean include the **range**, **variance** and **standard deviation**.

Parameters and Statistics

If you were able to measure the height of every adult male *Homo sapiens* who ever existed, and then calculate a mean, median, mode, range, variance and standard deviation from your measurements, those values would be known as **parameters**. They represent the *actual* values as calculated from measuring every member of a population of interest. Obviously, it is very difficult to obtain data from every member of a population of interest, and impossible of that population is theoretically infinite in size. However, one can estimate parameters by randomly sampling members of the population. Such an estimate, calculated from measurements of a subset of the entire population, is known as a **statistic**.

In general, parameters are written as Greek symbols equivalent to the Roman symbols used to represent statistics. For example, the standard deviation for a subset of an entire population is written as "s", whereas the true population parameter is written as σ.

Statistics and statistical tests are used to test whether the results of an experiment are significantly different from what is expected. What is meant by "**significant**?" For that matter, what is meant by "**expected**" results? To answer these questions, we must consider the matter of **probability**.

II. Experimental Design for Statistical Hypotheses

As you know from reading Appendix I, statistical hypotheses are stated in terms of two opposing statements, the **null hypothesis (H_o)** and the **alternative hypothesis (H_a)**. The null hypothesis states that there is no significant difference between two populations being compared. The alternative hypothesis may be either directional (one-tailed), stating the way in which the two populations will differ, or nondirectional (two-tailed), *not* specifying the way in which two populations will differ.

For example, if you were testing the efficacy of a new drug (Fat-B-Gon™) in promoting weight loss in a population of volunteer subjects, you would assemble two groups of volunteers who are as similar as possible in every aspect (age, sex, weight, health measurements, etc.), and divide them into two groups. One half of the subjects (the **treatment** group) would receive the drug, and the other half (the **control** group) would receive an inert substance, known as a **placebo**, that is cannot be physically distinguished from the actual drug. Both groups would be administered either the drug or the placebo in exactly the same way. , and the subjects

would not know whether they were getting the actual drug or the placebo. The subjects must not know whether they are in the treatment or control group (a **single-blind** study), as this will help to prevent the **placebo effect**, a measurable, observable change in health or behavior *not* attributable to a medication or other treatment, thought to be triggered by a subject's belief that the medication or treatment will have a particular effect. In some cases, not even the investigators know which subjects are in the treatment and control groups (a **double-blind** study). Thus, **the only difference between the treatment and control groups is the presence or absence of a single variable**, in this case, Fat-B-Gon^{tm}. Such careful design and execution of the experiment reduces the influence of **confounding effects,** uncontrolled differences between the two groups that might affect the results.

Over the course of the experiment, our investigators measure weight changes in each individual of both groups (Table A2-1). Because they cannot control for the obvious confounding effect of genetic differences in metabolism, the investigators must try to reduce the influence of that effect by using a **large sample size**--as many experimental subjects as possible--so that there will be a wide variety of metabolic types in both the treatment and control groups. **It is a general rule that the larger the sample size, the closer the approximation of the statistic to the actual parameter.** Even so, it is never wise to completely ignore the possibility of confounding effects. Honest investigators should mention them when reporting their findings.

Table A2-1. Change in weight (x) of subjects given Fat-B-Gon^{tm} (treatment) and placebo (control) food supplements over the course of one month. All weight changes were negative (weight loss). Mean weight change (x), the square of each data point (x^2) and the squared deviation from the mean $(x - x)^2$ are included for later statistical analysis.

control subjects	Dweight (kg) (= x)	$(\text{Dweight})^2$ $(= x^2)$	$(x - x)^2$	treatment subjects	Dweight (kg) (= x)	$(\text{D weight})^2$ $(= x^2)$	$(x - x)^2$
1	4.4	19.36	0.12	11	11.0	121.00	13.40
2	6.3	36.69	2.43	12	5.5	30.25	3.39
3	1.2	1.44	12.53	13	6.2	38.44	1.30
4	7.4	54.76	7.07	14	9.1	82.81	3.10
5	6.0	36.00	1.59	15	8.1	65.61	0.58
6	4.1	16.81	0.41	16	6.0	36.00	1.80
7	5.2	27.04	0.21	17	8.2	67.24	0.74
8	3.1	9.61	2.69	18	5.0	25.00	5.47
9	4.2	17.64	0.29	19	7.2	51.84	0.02
10	5.5	30.25	0.58	20	7.1	50.41	0.06
total (S)	47.4 (x=4.74)	249.6 $(=Sx^2)$	27.92 $(=S(x-x)^2)$	total (S)	73.4 (x = 7.34)	568.60 $(=Sx^2)$	29.86 $(=S(x-x)^2)$

III. Statistical tests

Let's return to our Fat-B-Gon^{tm} subjects. After the data have been collected, the subjects can go home and eat Twinkies$^{T.M.}$ and the investigators' analysis begins in earnest. They must now determine whether any difference in weight loss between the two groups is **significant** or simply due to random **chance**. To do so, the investigators must perform a **statistical test** on the data collected. The results of this test will enable them to either **ACCEPT** or **REJECT** the null hypothesis.

A. Mean, variance and standard deviation

You probably will be dealing most often with numerical continuous data, and so should be familiar with the definitions and abbreviations of several important quantities:

x	**= data point**	the individual values of a measured parameter (=x_i)
$\bar{x}$	**= mean**	the average value of a measured parameter
n	**= sample size**	the number of individuals in a particular test group
df	**= degrees of freedom**	the number of independent quantities in a system
s^2	**= variance**	a measure of individual data points' variability from the mean
s	**= standard deviation**	the positive square root of the variance

To calculate the **mean** weight change of either the treatment or control group, the investigators simply sum the weight change of all individuals in a particular group and divide it by the sample size.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Thus calculated, the mean weight change of our Fat-B-Gon^{tm} control group is 4.74 kg, and of the treatment group, 7.34 kg (Table A2-1).

To determine the degree of the subjects' variability from the mean weight change, the investigators calculate several quantities. The first is the **sum of squares (SS)** of the deviations from the mean, defined as:

$$SS = \Sigma (\bar{x} - x_i)^2$$

Whenever there is more than one test group, statistics referring to each test group are given a subscript as a label. In our example, we will designate any statistic from the control group with a subscript "c" and any statistic from the treatment group with a subscript "t." Thus, sum of squares of our control group (SS_c) is equal to 27.92 and SS_t is equal to 29.86 (Table A2-2).

The **variance (s^2)** of the data, the mean SS of each test group, is defined as:

$$s^2 = \frac{\sum (\bar{x} - x)^2}{n-1}$$

Calculate the variance for both the treatment and control Fat-B-Gon^{tm} groups. Check your answers against the correct ones listed Table A2-2.

Standard deviation (s), the square root of the variance:

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

Calculate the standard deviation for the treatment and control groups. Check your answers against the correct ones listed in Table A2-2.

B. Parametric tests

A **parametric test** is used to test the significance of **continuous numerical data** (e.g. - lizard tail length, change in weight, reaction rate, etc.). Examples of commonly used parametric tests are the Student's *t*-test and the ANOVA. You will be guided through the use of the Student t-test in the first two laboratories of this course, and so they will not be duplicated here.

C. Non-parametric tests

A **non-parametric test** is used to test the significance of *qualitative* data (e.g. numbers of purple versus yellow corn kernels, presence or absence of freckles in members of a population etc.). Both attribute data and discrete numerical data can be analyzed with non-parametric tests such as the Chi-square and Mann-Whitney U test. Although these tests are often simpler to perform, they are not as *powerful* as parametric tests. In other words, non-parametric tests less able than parametric tests to accurately predict whether unexpected results are due to random chance.

A Sample Non-parametric test: The Chi-square Test

A commonly used non parametric test is the Chi square ($\mathbf{X^2}$). Although this test has several complex permutations, we will use only the simplest formula to analyze genetic data from corn in the Mendelian Genetics laboratory, but you can also use it to test a wide variety of attribute or discrete data. (Complete instructions on how to perform this type of Chi-square test are included in that lab chapter.) The formula for calculating the Chi square statistic is as follows:

$$\mathbf{X^2 = \Sigma \frac{(O - E)^2}{E}}$$

In which:

O = the *observed* results

E = the *expected* results

Σ means the summation of C^2 values over every phenotypic category

In the Chi square test, **n** has a slightly different meaning than it has in parametric tests. In this case, n is the total number of categories possible. For example, if you are counting purple and yellow corn kernels, n = 2 (purple and yellow). If you are counting expression of two phenotypes, such as brown versus black fur *and* curly versus straight fur, n = 4 (black curly, black straight, brown curly and brown straight).

The **degrees of freedom (df)** in this Chi square test is equal to **n-1**.

IV. Probability and significance

The term "significant" is often used in every day conversation, yet few people know the statistical meaning of the word. In scientific endeavors, **significance** has a highly specific and important definition. Every time you read the word "significant" in this book, know that we refer to the following scientifically accepted standard:

The difference between an observed and expected result is said to be **statistically *significant*** if and only if:

> Under the assumption that there is no true difference, the probability that the observed difference would be at least as large as that actually seen is less than or equal to 5% (0.05).
>
> Conversely, under the assumption that there is no true difference, the probability that the observed difference would be smaller than that actually seen is greater than 95% (0.95).

Once an investigator has calculated a Chi-square or t-statistic, s/he must be able to draw conclusions from it. How does one determine whether deviations from the expected (null hypothesis) are significant?

As mentioned previously, depending upon the degrees of freedom, there is a specific probability value linked to every possible value of any statistic.

A. Determining the significance level of a parametric statistic

If were to perform an independent sample t- test (see Lab #1) on the Fat-B-Gon data listed previously, you should obtain values equal to those listed in Table A2-2, with a t-statistic equal to 4.05. The next step is to interpret what this statistic tells us about the difference in mean weight loss between the treatment and control groups. Is the difference significant, suggesting that Fat-B-Gon^{tm} is that mysterious factor "other than chance?" Or is the melting of unsightly cellulite at the pop of a pill just another poor biologist's fantasy of becoming fabulously wealthy? Once again, the answer lies in the table of critical values for the t-statistic, part of which is illustrated in Table A2-3.

Table A2-2. Treatment and control group statistics and overall statistics for weight loss in the Fat-B-Gon experiment.

statistic	**control**	**treatment**
mean (x)	4.74	7.34
sum of squares (SS)	27.9	29.9
variance (s2)	2.79	2.99
standard deviation (s)	1.66	1.82
overall statistics		
s^2_p	3.21	
$s_{x_t} - s_{x_c}$	0.642	
t	4.05	
degrees of freedom (df)	9	
P value (significance)	<---you fill in!	

To determine whether the t-statistic indicates rejection of the null hypothesis, do the following:

1. Locate the appropriate degrees of freedom in the far left column of Table A2-3..
2. Look across the df row to find a t value closest to the one you obtained from the Fat-B-Gon^{tm} data.

3. If the exact value does not appear on the table, note the two t values which most closely border your Fat-B-Gon value.

4. Find the P values which correspond to the two bordering values. Your P value lies between them. Fill in the P value for our Fat-B-Gon^{tm} experiment below and in Table A2-3.

_____ > P > _____

The Fat-B-Gon^{tm} *t* statistic (4.05) lies between 3.69 and 4.30 (df = 9) on the table of critical values. Thus, the probability that the weight difference in treatment and control groups is due to chance is between 0.005 (0.5%) and 0.002 (0.2%). This is *highly significant*, meaning that there is a 99.5% - 99.8% probability that the weight difference is due to the only variable between the two groups: Fat-B-Gon^{tm}! We can *reject* our original two-tailed hypothesis and *accept* the alternate hypothesis:

> **"There *is a significant difference* in the rate of weight loss between members of the population who use Fat-B-Gon^{tm} and those who do not use Fat-B-Gon^{tm}."**

Table A2-3. Partial table of critical values for the two-sample t-test. The second row of P values should be used for a two-tailed alternate hypothesis (i.e., one which does not specify the direction (weight *loss* or *gain*) of the alternate hypothesis). The first row of P values should be used for a one-tailed hypothesis (i.e., one which *does* specify the direction of the alternate hypothesis). A t-statistic to the right of the double bar indicates rejection of the two-tailed Fat-B-Gon^{tm} null hypothesis (at df = 9). (NOTE: This table is only a small portion of those available, some of which list df to 100 and beyond.)

P = 1-tail	0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
P = 2-tail	0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
df									
1	1.00	3.08	6.31	12.70	31.82	63.66	127.32	318.31	636.62
2	0.82	1.89	2.92	4.30	6.965	9.925	14.09	22.32	31.60
3	0.77	1.64	2.35	3.18	4.54	5.84	7.45	10.22	12.92
4	0.74	1.53	2.13	2.78	3.75	4.604	5.60	7.17	8.61
5	0.73	1.48	2.02	2.57	3.37	4.03	4.78	5.90	6.87
6	0.72	1.44	1.94	2.45	3.14	3.71	4.32	5.21	5.96
7	0.71	1.42	1.90	2.37	3.00	3.50	4.03	4.79	5.41
8	0.71	1.40	1.86	2.31	2.90	3.56	3.83	4.50	5.04
9	0.70	1.38	1.83	2.62	2.82	3.25	3.69	4.30	4.78
10	0.70	1.37	1.81	2.23	2.76	3.17	3.58	4.14	4.59

Notice that the *t*-value calculated for the Fat-B-Gon^{tm} data indicates rejection of even one-tailed hypothesis. However, because all honest researchers state their hypotheses *before* they see their results, Team Fat-B-Gon^{tm} should stick by their original hypothesis and let the direction of the data (i.e., all volunteers *lost* weight) speak for itself.

Remember that you must have a representative sample of the population--not a single experimental run--in order to perform the t-test (A single experiment cannot have a mean, variance or standard deviation.). Your probability value will come closer to the population parameter if your sample size is large. Hence, **it is best to use the pooled data from every group in a particular lab section if you perform this statistical test on your data**.

B. Determining significance level of a non-parametric statistic

First let us determine the probability value for our non parametric test, the Chi square. In this semester's laboratory on Mendelian Genetics, you will use the Chi Square to determine whether the proportion of physical types of offspring (purple or yellow corn kernels) in a single cohort is different from the expected. In the example presented in the chapter, data yield a $\mathbf{X}^2$ value equal to **1.333**. Because there are two independent categories (purple and yellow), **df = 2-1 = 1**.

1. In the far left column of Table A2-4, locate the appropriate df.

2. Go across the appropriate df row, and locate the Chi square value closest to the one we obtained with the example data. As you can see, 1.333 is not listed on the table. Rather, it lies between two values listed on the table, 1.323 and 2.706.

3. Go to the top row above each of the Chi square values bordering our example value. Above each is listed a corresponding probability (P) value.

4. The P value corresponding to 1.323 is 0.25; this means that a Chi square value of 1.323 indicates a 25% possibility that the deviation from the expected is due to chance. Thus, there is only a 75% chance that these deviations are due to some factor *other* than chance.

5. The P value corresponding to 2.706 is 0.10; this means that a Chi square value of 2.706 indicates a 10% probability that the deviation from the expected is due to chance, and a 90% probability that the deviation is due to some factor other than chance.

6. The probability value of our example Chi square lies between 0.25 and 0.10. This is most often expressed as

0.25 > P > 0.10

This P value is outside the accepted standards for statistical significance. The null hypothesis (the observed ratio of purple to yellow corn kernels will not differ from those predicted by Mendel's Laws) cannot be rejected.

Table A2-4. A *partial* table of the probability values for the Chi square statistic.

P =	0.999	0.995	0.990	0.975	0.950	0.900	0.750	0.50	0.25	0.10	0.05	0.02	0.01	0.005	0.001
df															
1	0.000	0.000	0.000	0.001	0.004	0.016	0.102	0.455	1.323	2.706	3.841	5.024	6.635	7.879	10.82
2	0.002	0.010	0.020	0.051	0.103	0.211	0.575	1.386	2.773	4.605	5.991	7.378	9.210	10.59	13.82
3	0.024	0.072	0.115	0.216	0.352	0.584	1.213	2.366	4.108	6.251	7.815	9.348	11.35	12.84	16.27
4	0.091	0.207	0.297	0.484	0.711	1.064	1.923	3.357	5.385	7.779	9.488	11.14	13.27	14.86	18.47
5	0.210	0.412	0.554	0.831	1.145	1.610	2.675	4.351	6.626	9.236	11.07	12.83	15.09	16.75	20.52