BIL 151
Data Analysis, Statistics, and Probability

By Dana Krempels, Ph.D. and Steven Green, Ph.D.

Most 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. In addition, there are variations in measurements that arise from the instruments used and from the way they're used, so the same event or object is unlikely to yield the same data twice, even from the same investigator let alone from different people and measuring tools. Hence, an investigator must measure as many individuals as possible to account for the variation in that population and must also employ analytic methods that account for errors, or variation, in the measurements themselves.

When a particular value is being compared in two different populations, care must be taken to ensure that each population is represented as accurately and rigorously as possible by the samples and by measurements taken on them. This is one purpose of statistical analysis.

I. Data, parameters and statistics

In biological science, investigators often collect observations that can be tabulated as numerical facts, also known as **data** (singular = **datum**). Important measurements include counts (frequency) and those that describe characteristics (length, mass, etc.). Data from a sample are often used to calculate estimates of the average values of the population of interest (**mean**, **mode**, and **median**) and dispersion around those average values (**variance**, and **standard deviation**). There are three different types of data (Figure 1).

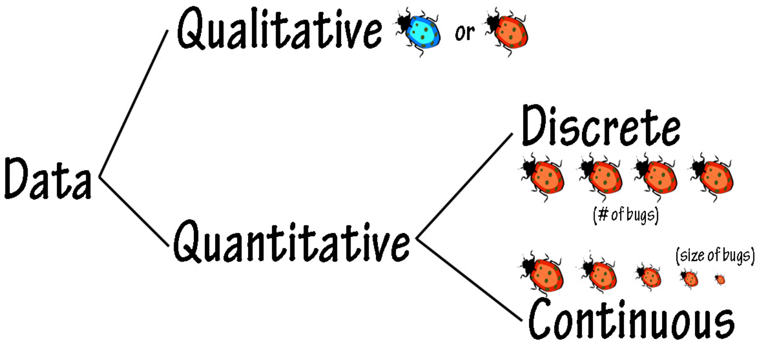

Figure 1. A simple diagram showing the basic distinguishing features of qualitative ("either/or") and quantitative data. Quantitative data can be either discrete (counted as integers) or continuous (measured along a continuum).

1. Qualitative/Attribute Data. These are 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 type of genetic trait (A, B, AB or o blood types) are examples. Because such data have no specific sequence, they are considered ***unordered***.

2. Quantitative/Numerical Data. These data correspond to numbers. Because they do have a specific sequence, they are considered ***ordered***. Numerical data may be of two types:

a. Discrete Quantitative/Numerical Data. These correspond to biological observations counted as 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 discrete numerical data. These data are ordered, but do not describe physical attributes of the things being counted.

b. Continuous Quantitative/Numerical Data. These are data that 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 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.

II. Mean, Mode and Median

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 subset that is a fair ("unbiased") representation of the human population (say, a sample of 3,000 individuals) and graph the numbers with "shoe size" on the **horizontal axis** (also known as the **abscissa**) and "number of individuals" on the **vertical axis** (also known as the **ordinate** when it represents counts). The resulting figure would show the **frequency distribution** of the data: a representation of how often a particular value occurs in the sample.

Measurements are never identical, so are distributed over a range of values. If they tend to cluster near the center of the range, then the sample can be fairly characterized by its arithmetic **mean** (the average value), its **median** (the value that places half the measurements above and half below) and its **mode** (the most common measurement in the range). These sample values can provide good estimates of the *actual* population values (**parameters**, defined below).

To calculate a **mean**, sum the values measured for all individuals in a population of interest and divide it by the sample size.

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

III. Range, Variance and Standard Deviation

When you are examining some measurable aspect of a population (such as a rate of reaction), the degree of variation around the mean should always be considered. In biological systems, there is almost always variation around the mean of any given value. In many biological studies, the estimation of degree of variation is as important, if not more important, than the mean.

Measurements of variation include the **range** as well as dispersion around the mean, namely the **variance** and **standard deviation**. The simplest of these is the range, which is defined as the highest value minus the lowest value. Unfortunately, the greater the sample size, the greater the range, and because it employs essentially only the two extreme values, a great deal of information about variation between those extremes is lost.

More useful are the variance and standard deviation, which are measures of deviations from the mean. The sample **variance (s^2)** is calculated as

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

In which **$\bar{x}$** is the sample mean, **x** is each individual value, and **n** is the sample size.

The **standard deviation (s)** is simply the square root of the variance:

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

Parameters vs. Statistics

If you were able to measure the second : fourth finger ratio of every adult *Homo sapiens* who ever existed, and then calculate a mean, median, mode, variance and standard deviation from those 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 if 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. But what do we mean when we say "**significant**?" For that matter, what is meant by "**expected**" results? To answer these questions, we must consider the matter of **probability**.

IV. Probability: What is meant by "expected results?"

The **probability** (**P value**) that an observed result is due to some factor other than chance is also known as **alpha** **(α).** By convention, α is usually set at 0.05, or 5%, which means that there is a 95% probability that a particular outcome is due to some factor <u>*other than*</u> random chance. In essence, α is a “cut off value” that defines the area(s) in a probability distribution where a particular value is <u>*unlikely*</u> to fall.

In some studies, a more rigorous α of 0.01 (1%) is required to reject the null hypothesis, and in some others, a more lenient α of 0.1 (10%) is allowed for rejection of the null hypothesis. For our study of mitosis, you will use an α level of 0.05.

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 lab manual, 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 a (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 statistic from collected data, s/he must be able to draw conclusions from it. How does one determine whether deviations from the expected (null hypothesis) are significant?

A **probability distribution** assigns a relative probability to any possible outcome.

V. Hypotheses: one tail or two?

The Scientific Process begins with three important steps:

1. Observation of some phenomenon that elicits a question/poses a problem.
2. Formulation of competing, testable hypotheses about that phenomenon
3. Prediction of all possible outcomes of experiments (or observations if experimentation is not possible) designed to test each hypothesis.

For example, as you wander in a field of wildflowers, you notice that individuals of Mexican Poppies (*Argemone mexicana*) have two distinct morphologies, one with spines and one without (Figure 4). You also notice that the individuals with spines are far more numerous than the smooth individuals. You might wonder: Is there a reason for the difference in their numbers? Are spiny individuals better able to deter herbivores with their spines than the smooth individuals? That is your **question**.

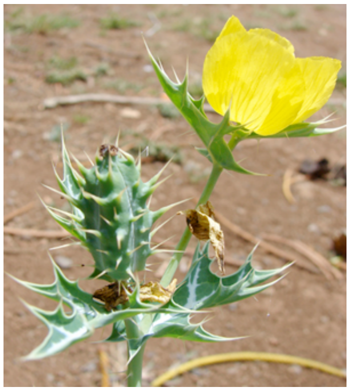

Figure 4. The Mexican Poppy, *Argemone mexicana*, has individuals with both spiny seed pods (shown) and smooth seed pods (not pictured). (http://commons.wikimedia.org)

The question can then be stated as a testable hypothesis. In this case, you might state "Individual *Argemone mexicana* with spiny seed pods suffer less seed predation than individuals with smooth seed pods." That is your **overall, or experimental hypothesis**.

The next step is to formulate competing statistical hypotheses. 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 precise way in which the two populations will differ, or nondirectional (**two-tailed**), *not* specifying the way in which two populations will differ. *(Note from Dr. Green: almost nobody justifies the use of one-tailed tests at all, ever, under any circumstance and they imply that you deliberately blind yourself to an outcome in an unexpected direction. Such an outcome may be the most interesting results you'll ever have!)*

For example:

Null hypothesis: There is no difference in herbivore damage between the spiny and smooth *Argemone mexicana* seed pods.

Your **alternative hypothesis** can be either **two-tailed**, or **one-tailed**.

Two-tailed alternative hypothesis: There is a difference in herbivore damage between the spiny and smooth *Argemone mexicana* seed pods.

That's fine. But since you already have a good basis to predict that spines protect the seeds from predation, it is arguably better (and more conclusive) to state a one-tailed hypothesis:

One-tailed alternative hypothesis: *Argemone mexicana* with spiny seed pods will suffer *less* herbivore damage than *Argemone mexicana* with smooth seed pods.

The difference is subtle, but important. *If you have a logical reason* (e.g., information from scientific literature or from your own preliminary observations) to state your alternative hypothesis as one-tailed, then don't be afraid to do so. The alternative is your hypothesis of interest, the one you are predicting you will fail to reject.

VI. Catalase Reaction: Statistical Analysis

You and your team have collected numerical data in the form of rates of reaction of catalase breaking down hydrogen peroxide under two different environmental conditions of your team's design. You should already have an overall hypothesis as well as null and alternative experimental hypotheses. You are now ready to apply statistics to your research.

A. Statistical Tests and Probability Distributions

Enough statistical tests and their associated probability distributions have been invented to fill many textbooks. Some of these, such as the Chi-square test, the Student's t-test, the Analysis of Variance (ANOVA), the Mann-Whitney U test and the Fisher's exact test may sound familiar to you. The specific probability distribution and statistical test appropriate in a given situation depend upon the type of data collected and the nature of your hypothesis.

In your catalase experiments, you performed multiple trials of catalase's speed at breaking down hydrogen peroxide under two different environmental conditions. The data were recorded as rates of reaction, and because you collected data from multiple trials under each of the two conditions, you can use those rates of reaction to determine whether the environmental variable had any real effect on your enzyme's speed. You will be comparing the mean rate of reaction at [condition #1/control] to the mean rate of reaction at [condition #2/treatment]. The most appropriate statistical test for analysis of means of continuous numerical data like this is the **student's t test**.

The probability distribution of the student's **t-distribution** is used to determine whether the observed difference between the means of two samples is unlikely to have arisen by chance if they were in fact drawn from the same population (or from populations with identical means). To make a very long and complex story short, you can use the mean, variance, and standard deviation of your reaction rates (treatment vs. control) to calculate a **t-statistic**. Every possible value of the t-statistic is linked to a certain probability that the observed difference in sample means is simply a matter of chance.

B. The student's t-test: A tool for determining whether there is a significant difference between two means

The student's t-test can be used to determine whether a difference between two means is **significant**. Note that "significant" in this sense is NOT the same as "biologically meaningful." It refers only to whether the observed difference is unlikely to be due to chance ("**statistically significant**"). Means may be calculated from observations that are either **paired** (as when individuals in a single group are subjected to "before and after" measurements, and data points are paired for each tested individual) or **independent** (as when individuals in two *similar* sample populations are measured, but each individual in each sample population is measured only once).

Paired designs are always best because they eliminate the added influence on a difference in means arising from variation between samples due solely to their containing different individuals. Although an experiment using paired observations will always yield more powerful results (with adequate sample size) than an experiment using independent observations, it is not always possible to pair observations. For example, in your enzyme experiments, it would be impossible to use exactly the same individual yeast and catalase molecules for both treatment and control runs. For this reason, you will use a statistical test appropriate for analyzing independent samples.

Slightly different calculations of the t-statistic must be used for paired means vs. independent means. **You will analyze your results with the independent sample t test.**

1. The Independent sample t-test

The independent sample t-test is designed to show whether there is a significant difference between the means of two **independent** samples, those of your (1) treatment and (2) control groups.

Use the following equation to calculate a t-statistic for your two independent mean (rates of reaction):

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{s_p^2 (1/n_1 + 1/n_2)}}$$

...in which

- **x_1** and and **x_2** are the means of your two groups
- **n_1** and **n_2** are the numbers of trials under each condition
- **s_p^2** is the **pooled variance**, calculated as:

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 \;+\; (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

...in which

- **s_1^2** is the variance of group 1
- **s_2^2** is the variance of group 1
- **df_1** is the degrees of freedom for group 1 **($df_1 = n_1 - 1$)**
- **df_2** is the degrees of freedom for group 2 **($df_2 = n_1 – 1$)**

Recall that the **variance (s^2)** of a mean is calculated as

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

In which

- **$\bar{x}$** is the mean
- **x** is each individual value
- **n** is the sample size

2. Degrees of Freedom

The last thing you must do is to determine the number of independent quantities in your system, or **degrees of freedom (df).** The degrees of freedom determine the significance level tied to every possible value of a statistic (such as the t-statistic). The degrees of freedom is the number of data points that are free to vary without changing the test statistic, and this changes depending on the type of statistic you are calculating. The **degrees of freedom** for the two-sample t-test with independent means is calculated as the sum of the degrees of freedom of each test group:

$$df = (n_1 - 1) + (n_2 - 1)$$

C. Determining the significance level of your t-statistic

Once you have calculated a t-statistic for your two mean rates of reaction, you must try to interpret what this statistic tells you about the difference between the two means. Is the difference significant, suggesting that there is something other than chance causing the variation in volume yielded by each method? The answer lies in the table of critical values for the t-statistic, part of which is illustrated in Table 3. Here’s how to find that answer.

1. Locate the appropriate degrees of freedom in the far left column.
2. Look across the df row to find a t value closest to the one you obtained.
3. If the exact value does not appear on the table, note the two t values which most closely border your value, or the single value that is either greater or smaller than your value.
4. Find the P value(s) that correspond to your bordering values, and enter them in the appropriate space(s) below:

__________ **> P >** __________

If the P value associated with your t-statistic is less than (or equal to) 0.05, there is less than 5% probability that the observed mean volume difference arose as a matter of chance. If this is the case, you must **reject** your null hypothesis and **fail to reject** your alternative hypothesis. If the P value is greater than 0.05, you **fail to reject** your null hypothesis.

Alternatively, simply find the t-statistic associated with P = 0.05 at your df. This is called the **critical value.** If your calculated t-statistic is larger than the critical value from the table, reject your null hypothesis.

2. Tying it all together

What is the value of your t-statistic? ______________________________

Your degrees of freedom? ______________________________

What is the P value associated with your t statistic at your df? ____________

Does your P value indicate that you should reject or fail to reject your null hypothesis?

____

It is not enough to simply state that you have rejected or failed to reject a null or alternative hypothesis. You also must explain your results logically, and—to the best of your ability—on a molecular level. Check the rubric that your lab instructor will be using to evaluate your team's PowerPoint, and plan accordingly. It is linked to the syllabus on the BIL 151 home page.

Table 3. Table of critical values for the two-sample t-test. The P levels (0.05) indicating rejection of the null hypothesis are shown in bold for both one-tailed and two-tailed hypotheses. (From Pearson and Hartley in *Statistics in Medicine* by T. Colton, 1974. Little, Brown and Co., Inc. publishers.)

2-tail -->	0.10	**0.05**	0.02	0.01	0.001
1-tail -->	**0.05**	0.02	0.01	0.005	0.0005
df					
1	6.314	12.706	31.821	63.657	636.619
2	2.920	4.303	6.965	9.925	31.598
3	2.353	3.182	4.541	5.841	12.941
4	2.132	2.776	3.747	4.604	8.610
5	2.015	2.571	3.365	4.032	6.859
6	1.934	2.447	3.143	3.707	5.959
7	1.895	2.365	2.998	3.499	5.405
8	1.860	2.306	2.896	3.355	5.041
9	1.833	2.262	2.821	3.250	4.781
10	1.812	2.228	2.764	3.169	4.587
11	1.796	2.201	2.718	3.106	4.437
12	1.782	2.179	2.681	3.055	4.318
13	1.771	2.160	2.650	3.012	4.221
14	1.761	2.145	2.624	2.977	4.140
15	1.753	2.131	2.602	2.947	4.073
16	1.746	2.120	2.583	2.921	4.015
17	1.740	2.110	2.567	2.898	3.965
18	1.734	2.101	2.552	2.878	3.922
19	1.729	2.093	2.539	2.861	3.883
20	1.725	2.086	2.528	2.845	3.850
21	1.721	2.080	2.518	2.831	3.819
22	1.717	2.074	2.508	2.819	3.792
23	1.714	2.069	2.500	2.807	3.767
24	1.711	2.064	2.492	2.797	3.745
25	1.708	2.060	2.485	2.787	3.725
26	1.706	2.056	2.479	2.779	3.707
27	1.703	2.052	2.473	2.771	3.690
28	1.701	2.048	2.467	2.763	3.674
29	1.699	2.045	2.462	2.756	3.659
30	1.697	2.042	2.457	2.750	3.646