Appendix 1. Excerpt from *The Statistics Primer* by M.A. Harris, J.M. Batzli, and E.V. Nordheim

Experimental Design & Summary Statistics Primer

Statistics and Science

From an early age, we are taught the importance of the scientific method in conducting research. However, just as we need a uniform and organized process by which to conduct studies, it is also necessary to utilize a standardized procedure, or statistics, to analyze and interpret our data. The statistical method provides scientists and researchers with a unifying language in which they can describe their data and support their conclusions.

Statistics and Biocore

In the Biocore labs, as in other research courses, you are given the opportunity to develop and study your own research questions that further expand on the current lab topic. Statistics will allow you to more appropriately describe your data and validate your conclusions. As you progress through the sequence of labs, your understanding of the relationship between statistics and science will grow, while improving your experimental designs.

Populations and Samples

In Biocore you will use data from small samples, which represent larger populations, to make conclusions about your questions and hypotheses. Here’s an example to help you understand the difference between samples and populations: Let’s say you are interested in the effects of chemical runoff from a local carwash into Willow Creek. Your lab instructors have suggested that *Daphnia magna*, a well-known indicator species, would serve as an excellent test species for this question. As a result, you have chosen all of the *D. magna* in Willow Creek to serve as your **population**.

> A **population** is “the set of all individuals of interest in a particular study.”[*]

Obviously, in Willow Creek it would be an impossible task to capture and examine the entire population of *D. magna*. A more logical approach would to be to study a smaller representation of the *D. magna* population and then generalize your conclusions to the whole population in Willow Creek. In other words, you need to select a **sample** of *D. magna* from the large population in Willow Creek.

> A **sample** is “a representative selection of a population that is examined to gain statistical information about the whole.”[a]

Figure 1 helps to illustrate the relationship between the population and the sample. The population is composed of all the possible

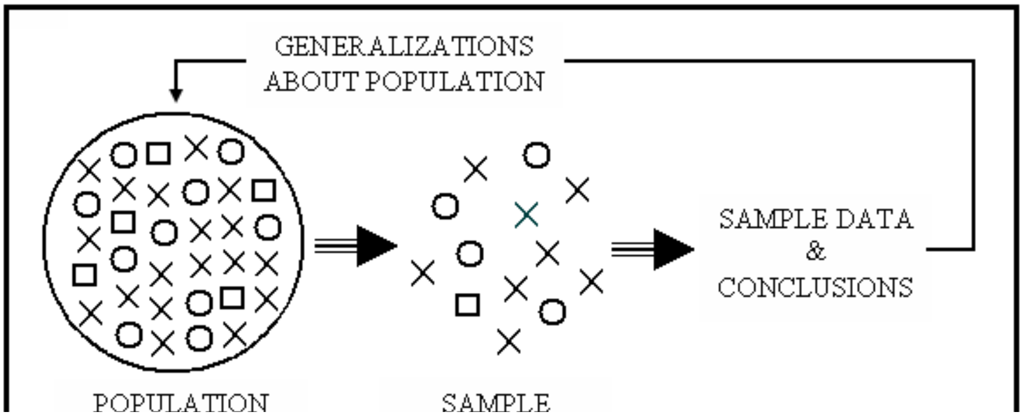

Fig. 1: An illustration of the relationship between the population and a sample. The population is made up of all the possible individuals, which are represented by the X’s O’s and □’s, of the group of interest. The various shapes represent the individual differences in the variable of interest (*e.g.*, heart rate, body size) within the population. The sample is a smaller subset of individuals

of Willow Creek. The sample is a smaller subset of *D. magna* from the population. If the sample is representative of the population, the study results can be used to generalize about the population.

Notice that the population has been defined as the *D. magna* in Willow Creek, as opposed to the entire species. It is important to carefully define the population of interest, taking into account your research question, experimental design, and resources. Table 1 provides some examples of appropriately defined samples from a specified population (keep in mind that these are only possible examples; sample sizes and definition may vary for a particular population).

Table 1: Three examples of a specified population and a possible representative sample.

Population	**Sample**
All 250 Galapagos Finches on the island of Isabela	75 Finches obtained from locations around the entire island via mark-recapture method
All of the yeast cells in a 500mL beaker	After stirring to evenly distribute the yeast in the liquid medium, 10 ml of yeast are removed as a representative sample of the beaker population
UW-Madison students (male and female) between the ages of 18 and 22	150 individuals from the UW campus who reflect the diversity in gender, fitness, age, and race within the population

Valuing Variability

Experimental results must be repeatable in order to be accepted as valid. This is done both by including replicate points in the design of the experiment and by repeating the entire experiment. Laboratory experiments are usually repeated many times before they are published. Field experiments are usually repeated or are carried out at more than one site. Replication is especially important in field experiments. Natural systems vary in space and through time. Replication helps the experimenter differentiate between natural variation and changes caused by the experimental treatment under review. Experimenters are looking for changes above and beyond the *natural variation*.

Natural variation is a combination of both environmental and individual variation. For example, yeast suspended in solution would be exposed to varying oxygen levels, depending on their location in a non-shaking flask. Those at the top of the flask would receive greater oxygen exposure than those at the bottom of the flask. This is clear example of natural, environmental variation. It is to be expected that the environment varies from location to location, even in a system as small and isolated as a flask, test tube, or beaker. In contrast, natural, individual variation refers to the differences between organisms within samples. For example, the amount of body fat on an individual can affect the readings of an electromyogram (EMG) signal measure at the skin's surface. (This is because a participant with a higher level of body fat will produce a less intense and more variable EMG signal than a participant with lower body fat due to the closer proximity of electrodes to muscle.)

How many replicates should I use?

When thinking about the number of replicates to use in an experiment, scientists will often rely on statistical estimates based on the variation observed in pilot study data. For Biocore, a good rule of thumb is that three samples is the absolute minimum for replication (you cannot estimate variability from only two samples). Ideally, you should measure many more than three replicates when you are attempting to generalize about a population, especially when you know that the population or the system you are working in has a great deal of environmental and/or individual variation (think 8-20 replicates!). Through replication, you hope to define how much of the observed variation is natural and how much is explained by your independent variable(s).

Types of Samples

Here, we introduce you to some analytical methods used by ecologists in designing and carrying out experiments. Ecological data is often full of variation because it is difficult to control all experimental parameters in the field. Therefore, it is very important to know your system, and/organism in order to anticipate variation when designing an experiment, and planning the statistical methods for analyzing data.

For example, let's say we wanted to know how repeatedly mowing a field early in the spring influenced the height of weedy Canada thistles. To run a controlled experiment, you mow half of your large field but leave the other half un-mowed as a control. This would be a **simple sample.** Many Canada thistles grow in both the mowed and un-mowed parts of the field later in the season. In order to make sure the plants you measure are 'representative' of all plants influenced by mowing (or non-mowing) you must take pains to make sure the sample is *not biased*, or that it is not unduly influenced by other environmental factors such as differences in soil moisture or nutrients across the field.

Simple Sample

Mow	Non-mow

But let's say now that the far west side of the field has slightly greater soil moisture (indicated by the gray shading in the systematic example below) than the center or eastern side of the plot. In order to account for this natural variation, we can alter the experimental design so that there are mow and non-mow areas in both the wetter and drier areas of the field, resulting in a **systematic sample**. This design attempts to *systematically control for the variation* in moisture.

Systematic Sample

Non-mow	Mow	Non-mow	Mow
Mow	Non-mow	Mow	Non-mow

If you are setting up a field experiment and there are no apparent environmental differences in your site, it would still be beneficial to split the area up into smaller plots. It would also be advantageous to randomize treatments to plots, resulting in a **random sample**, or to stratify the randomization so that plots are randomized within the rows, yielding a **stratified random sample** (notice that within each row there are 2 mow and 2 non-mow).

Random Sample

Non-mow	Non-mow	Non-mow	Mow
Mow	Mow	Mow	Non-mow

Stratified Random Sample

Non-mow	Non-mow	Mow	Mow
Mow	Non-mow	Mow	Non-mow

It is also important to note the applicability of theses sampling methods to the lab setting. For example, if you were attempting to determine the reaction rate of a particular enzyme, it would be necessary to test varying substrate concentrations. To make it simple, you choose two concentrations, 0M and 0.35M, and 4 replicates of each, for a total of 8 test tubes. The assay that tests the reaction rate requires the use of a warm water bath, but we don't know for certain that the water baths heat the water evenly. Either the test tubes could be placed in a simple sample arrangement or, if you wanted to control for temperature variance within the water bath, you could use systematic, random or stratified sampling.

If you are setting up test tubes in a rack and there are no apparent water temperature differences within the bath, it would still be beneficial to split the area up into smaller plots. It would also be advantageous to randomize treatments to plots, resulting in a **random sample**, or to stratify the randomization so that plots are randomized within the rows, yielding a **stratified random sample**.

Controls and Replicates

Controls are used to identify (and sometimes to correct for) results that are not due to variation in your experimental (independent) variable. For example, suppose that you have learned that controlled burning is sometimes an effective tool for combating invasive species, and you want to design an experiment to test whether burning will set back the purple loosestrife that is taking over your favorite wetland. You would not simply burn the wetland and see what happens because a reduction in loosestrife could be due to something other than your experimental treatment (burning). For example, it could be due to an influx of a particular insect or to the weather that year. It would be important to leave part of the wetland unburned as a control. You can then compare the burned and unburned parts.

In order to start making generalizations about the effectiveness of burning on control of purple loosestrife outside of your favorite wetland, you would have to locate several (>5) different wetlands that have similar density of purple loosestrife, and set up treatment (burn) and control (not burn) areas. If each wetland is independent of each other (not connected or influenced by one another), then "wetland" can be considered a sampling unit and each wetland is an independent **replicate.**

In chemistry you probably are used to controlling all the variables except the one you want to test. This usually is not possible in complex ecosystems (*e.g.*, the top of the hill may be different from the bottom of the hill; a bird may have deposited an unusually high concentration of seeds in one particular place). Ecologists try to minimize the effect of environmental *variation* within a site as much as they can by using **replication** in their experimental design. For example, in designing our experiments comparing three methods (plus control) for removing weeds from the Biocore Prairie prior to planting prairie species, we divided the area into twelve plots and randomly assigned one of the three treatments to each plot with three plots left untreated as controls. With 3 replicates of 3 treatments (plus control) we hoped to reduce the effect of known and unknown on-site variation to reveal obvious real differences among treatments.

Types of Variables and Measurement Scales

When conducting a study, your variables tend to fall into one of two categories—discrete or continuous. It is important to distinguish the types of variables you are working with, because the type of data influences the kind of analysis and the presentation that is most appropriate. **Discrete** variables are "separate, indivisible categories [and] no values can exist between two neighboring categories."* For example, if you were counting the number of plant species in a given plot, your data would only consist of whole numbers; that is you could not have 10.5 plant species, only 10 or 11 plant species. You would collect categorical data when you are surveying trees in a forest or plants on the prairie, finding 10 species of weeds and 18 species of prairie plants where the categories are "weeds" or "prairie plants".

Continuous variables "have an infinite number of possible values that fall between any two observed values and can be divisible into an infinite number of fractional parts."* In other words, continuous data are measured and the units of measurement can be infinitely subdivided (at least to the resolution of the measuring instrument), *e.g.*, length, weight, time. For instance, if you were measuring the amount of rainfall in inches for a given summer, you could find 20 inches of rainfall or 21.34 inches of rainfall. Table 2 provides some concrete examples of discrete and continuous data.

Table 2: Three examples of how data can be measured and interpreted using the four types of measurement scales.

Discrete	**Continuous**
Plant ID: Leaf type *(toothed, entire, divided, or lobed)*	Plant height in inches
Gender of *Drosophila* flies	Enzyme activity as measured by light absorption

(male or female)	
Number of water bugs in water sample	Amount of rainfall on the prairie in inches
Number shmooing yeast cells	Muscle activity as measured by an electromyogram (EMG)

Appreciating Raw Data

Previously, we discussed the idea of using the sample to represent the population for your study. After data collection, it is important to examine your raw (untransformed) data and determine whether the sample is indeed a fair representation of the population. Therefore it is often necessary to visualize your raw data graphically to see how data points are distributed around some mean value and to gain perspective about the population.

There are several types of graphs that can accomplish this. **Frequency distributions** illustrate the occurrence of a particular score or piece of data in a study. Suppose, for example, that we want to know the average stem length of red clover growing in a particular field. If we randomly select and measure a large number of plants and plot a **histogram** of the values, we find a bell-shaped curve called the **normal distribution (see Fig. 2)**.

Most populations in the natural world tend to follow a normal distribution, or a "bell –curve," where most individual values fall near and are equivalently distributed around some mean value as shown in Figure 2. Therefore, it is a common assumption that the population that you are studying will have a normal distribution. Accordingly, when you graphically depict your sample data, in most cases, it should closely resemble a normal curve. This is necessary and important since many statistical tests are based on the assumption that the data are approximately normally distributed. However, if this is not true of the data, corrections for various tests do exist.

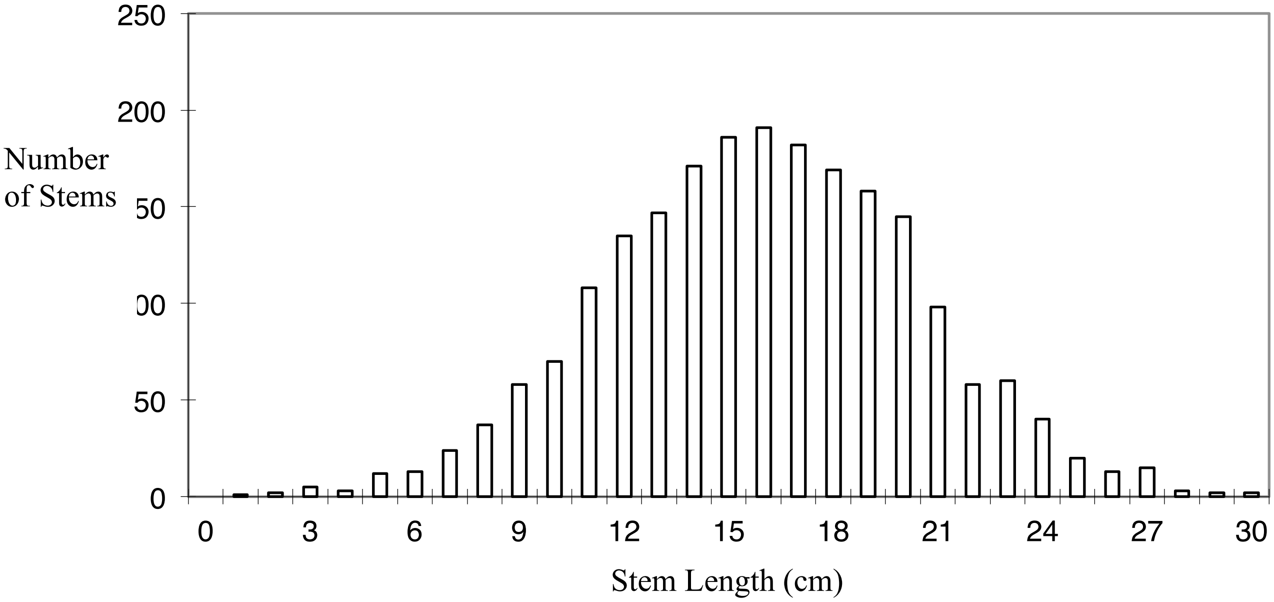

Figure 2: Frequency distribution of the stem lengths of a large, random sample (n=2128) of red clover (*Trifolium pratense*) in July 1999.

Another way to visualize continuous data is through an **x-y scatter plot**, which allows you to examine variation among replicates for any given treatment, or variation among individuals for any given feature. This can give insights into certain trends and/or possible relationships between variables. For example, in Figure 3 Galapagos finch beak length is plotted as a function of wing length for 35 birds. It is evident that, for any given wing length, beak length can vary among birds. (Look carefully at the eight birds whose wing length measured 67 mm: their beak lengths ranged from about 8.5 to 11.5 mm!)

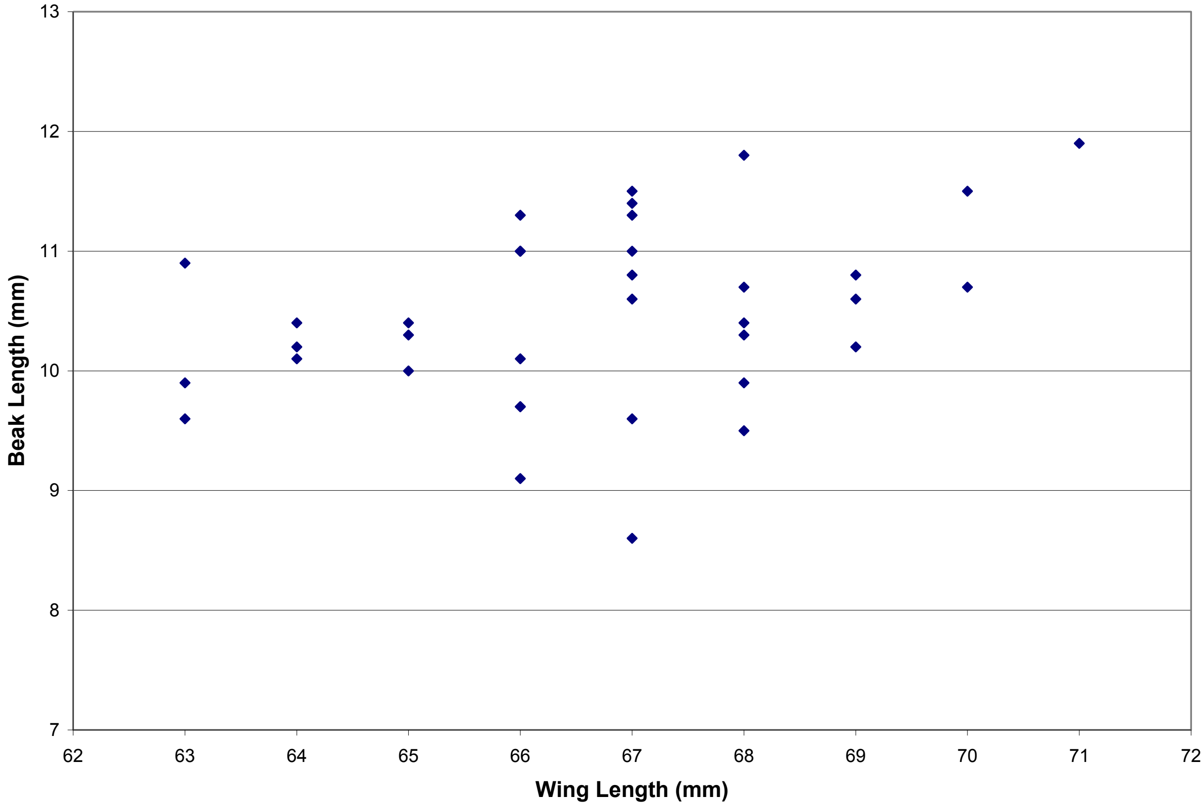

Figure 3: Scatterplot of Galapagos Island finch beak length as a function of wing length for 35 birds. Note that the x and y-axes do not begin at zero.

As you can see, a visual representation of the raw data can yield valuable information about the population that you are trying to study. If the variation within your sample is representative of the population variation, the distribution of the data will mirror that of the population. In reality, a sample may come close to this, but this is usually rare. Consequently, it is much more beneficial and accurate to use larger sample sizes and more replicates.

Descriptive Statistics

Two of the most useful measures for characterizing data from a sample are an indication of the ***central tendency*** and an indication of the ***dispersion*** of the data. Measures of the central tendency are the **mean** (average value), **mode** (most common value), and **median** (value at which half are greater and half are smaller). Locate these on Figure 1 above. The mean is given by summing each individual value (x_1, x_2, x_3, ...x_i) and dividing by n, the number individuals or samples measured.

$$Mean = \bar{x} = \frac{\sum x_i}{n}$$

One way to characterize the dispersion of the data is to give the **range** (spread from lowest to highest). The problem with using the range as a measure of dispersion is that it increases as we include more samples; furthermore, it tells us nothing about the average variation. You might think it useful to simply average the deviations from the mean, but this doesn't work because the deviations in the negative direction exactly cancel out the deviations in the positive direction and we end up with zero. We can, however, add up the **squares** of the deviations. After dividing by an indicator of our sample size, this gives the **variance (s^2)**.

$$Variance = \frac{\sum\left(x_i - \bar{x}\right)^2}{n-1}$$

The squared deviations are divided by n-1, rather than n, because dividing by n tends to underestimate the population variance when the sample size is small (*e.g.*, less than 30).

Think about what the variance tells you compared with what the range tells you.

One problem with variance is that its units don't make sense (*e.g.*, [beats/min.]^2 for heart rate). It is more common to use the square root of the variance, known as the **standard deviation**. You can think of the standard deviation as the average amount of individual deviation from the mean.

$$\textit{Standard deviation} = s = \sqrt{\frac{\sum\left(x_i - \bar{x}\right)^2}{n-1}} = \sqrt{\text{variance}}$$

For measurements that follow the normal distribution, 68% of all the observations lie within the range covered by one standard deviation on either side of the mean (*i.e.*, the mean ± 1 SD). 95% of the observations lie within 2 SD of the mean. Practically all (99.7%) of the observations lie within 3 SD of the mean.

We would like to be able to use data from our **samples** to make inferences about the **population** from which the samples were taken. The standard deviation tells us the variation among our samples; it does not allow us to compare 2 populations. For that we need to determine the standard error of the mean (see below).

Suppose that we measure many **sets** of samples from our population and calculate the **mean for each set**. These means will not be identical but will cluster around the average for the whole population. In fact, the means also show the bell-shaped pattern called the normal distribution. It should make sense to you that the amount of variation (dispersion) of the **means of the sample values** will not be as great as the dispersion of the **sample values** themselves. The standard deviation of the set of means is called the **standard error of the mean**. Standard errors are usually not obtained from a frequency distribution by repeated sampling but are estimated from only a single sample and represent the expected standard deviation for a large number of repeated samples. Usually, we do not have many sets of means, just one. Therefore, we calculate the standard error of the mean by dividing the standard deviation by $\sqrt{n}$.

$$\textit{Standard error of the mean} = s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

The beauty of the standard error of the mean is that it tells us a great deal about the population we are studying. The standard error of the mean allows us to determine, with a certain probability, the limiting values between which the "true" value of the population lies. For a sample of reasonable size (*e.g.*, 30 or more) that follows the normal distribution, the 95% confidence level is $\bar{x} \pm 1.96 s_{\bar{x}}$. This says that there is a 95% chance that the true mean of the population is within these limits. (1.96 is often rounded to 2). As can be seen from the formula for SE, this interval becomes smaller (tighter) as the sample size, n, gets larger.