# Basic Statistics



**INTRODUCTION TO STATISTICS** Statistics is the branch of mathematics that deals with the techniques for collecting, analyzing, and drawing conclusions from data (Snedecor and Cochran, 1980). Statistics are the analytical backbone of science. Without them, numbers tell us little. "Statistics" is one of those words that can strike fear into the hearts of the bravest teachers and students. But, it's not so bad when you consider that just calculating an average means you're doing statistics! And, as stated in the definition above, how you collect the data (i.e., as dictated by your experimental design) is also part of statistics. Which, when you think about it, makes sense. How data will be mathematically analyzed (or "the numbers crunched") depends on how those data were collected in the first place. That's why experimental design and statistics go hand in hand!

Once you have identified your question and hypothesis the next step is designing the experiment to test the hypothesis. In many experiments you will be manipulating or observing only a portion or *sample*, of the total available plants or animals. The total number, N, of a species in an area (i.e., your habitat) is the*population*. A subset of the population is a sample, usually denoted as n in equations. This subset of the population will be used to represent the whole population. The things associated with a population or sample, such as the mean, variance, and standard deviation, are called parameters. These will be explained below.

There are various ways to sample a population but we will go over the most commonly used, random sampling. You are probably most familiar with *random sampling* in the context of surveys of people on a variety of subjects. Those surveys are designed (if they are unbiased) to question as random a sample of the population as possible to get an accurate picture of opinions on a certain topic.

**Random sampling**— In experiments where a sample or subset of plants, leaves, insects, etc. will be treated, counted, or observed for collecting data, the sample taken should be random. If all the data are collected from one plant or area of the habitat it will be biased and not be a good indicator of what is really going on ecologically. An example in everyday life would be a researcher wanting to determine the average height of fifth graders in a particular school district. If she only measured boys her results would only apply to boys, not all fifth graders, and would thus be biased, not random. To collect unbiased data she would randomly choose the same number of boys and girls from each fifth grade class to measure. She could do this by assigning every child a number and then pulling numbers from a hat. These days, there are simple computer programs to do the picking.

Here is an ecological example: The experimental design calls for observing what food items red ants bring back to their colony as compared to black ants. You have too many ant colonies to observe all of them, so you pick a random sample of 5 colonies of each ant type to observe. An easy way to choose randomly is by giving each colony a number or letter on a slip of paper. Put these in a basket and pull 5 slips for each ant colony type. This way there is no bias toward any particular colonies.

Random sampling is done in many experiments. For example, in drug trials, fifty out of one hundredpeople are randomly chosen to receive the drug, while the other fifty receive a placebo.

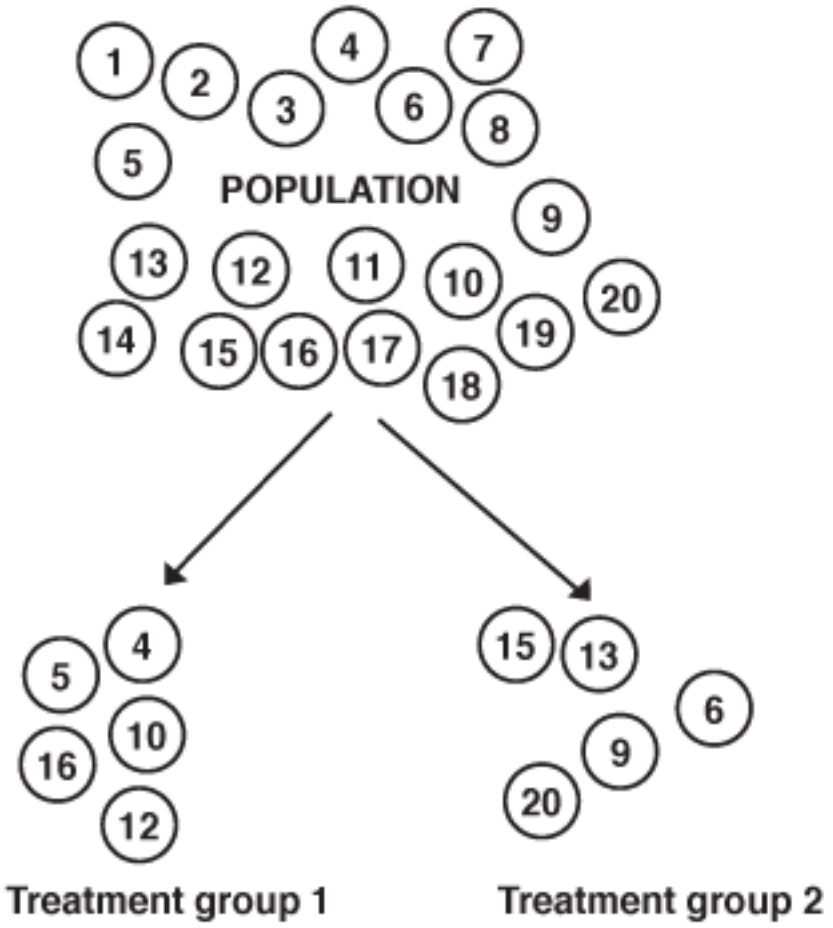



**Figure 1**. Drawing to illustrate random sampling.

Now, we know we need to randomly sample our population, but how many subjects, plants, plots, or whatever it is we are studying, do we need per treatment? Do we need more than one? Yes, you need more than one, and each one is called an *experimental unit or replicate*. By having more than one unit per treatment you are *replicating* the units.

**Experimental Unit**— The experimental unit is the plant, animal, patch of ground, or whatever is being subjected to the treatment, which is the independent variable. Example: Mesquite trees are being measured for growth differences caused by receiving different amounts of nitrogen. The treatment is the amount of nitrogen given to the different trees. Each tree is an experimental unit because it is the *unit*receiving the treatment.

Replication is a must! Replication means to have more than one experimental unit that will be subjected to the **independent variable** or treatment. The reasons for replicating are: (1) Organisms die or don't otherwise perform, and if you use only one the experiment is useless. With at least three or more experimental units you have a better chance of getting some data. (2) To calculate averages or other statistics, you must have more than one animal, plant, plot, etc., that is being manipulated and measured. Using only one of something means the experiment is only going to yield results for one individual, which will not give much of a picture of what's true for the population.

Example of no replication: Three plants are each given a different amount of water. Plant 1 receives

0.1L/day, Plant 2 receives 0.5L/day and Plant 3 receives 1L/day. Only one plant receives a particular amount of water each day. There is no replication, so no data analysis can be done and this is not a valid experiment.

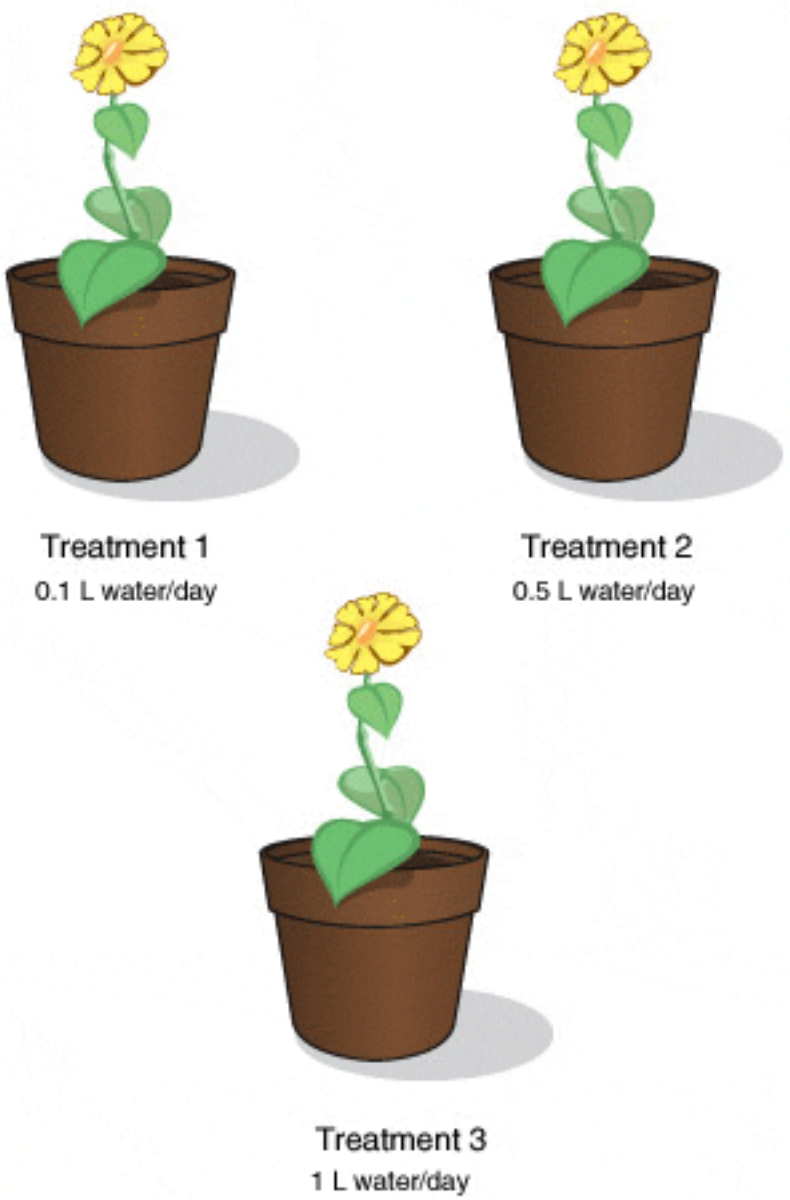



**Figure 2**. Having only one plant per treatment is an example of no replication.

Example with replication: A total of nine plants are in the experiment. Three plants receive 0.1L/day, three receive 0.5L/day, and three receive 1L/day. With three plants in each treatment group, you have three experimental units per treatment and can do data analyses such as averages.

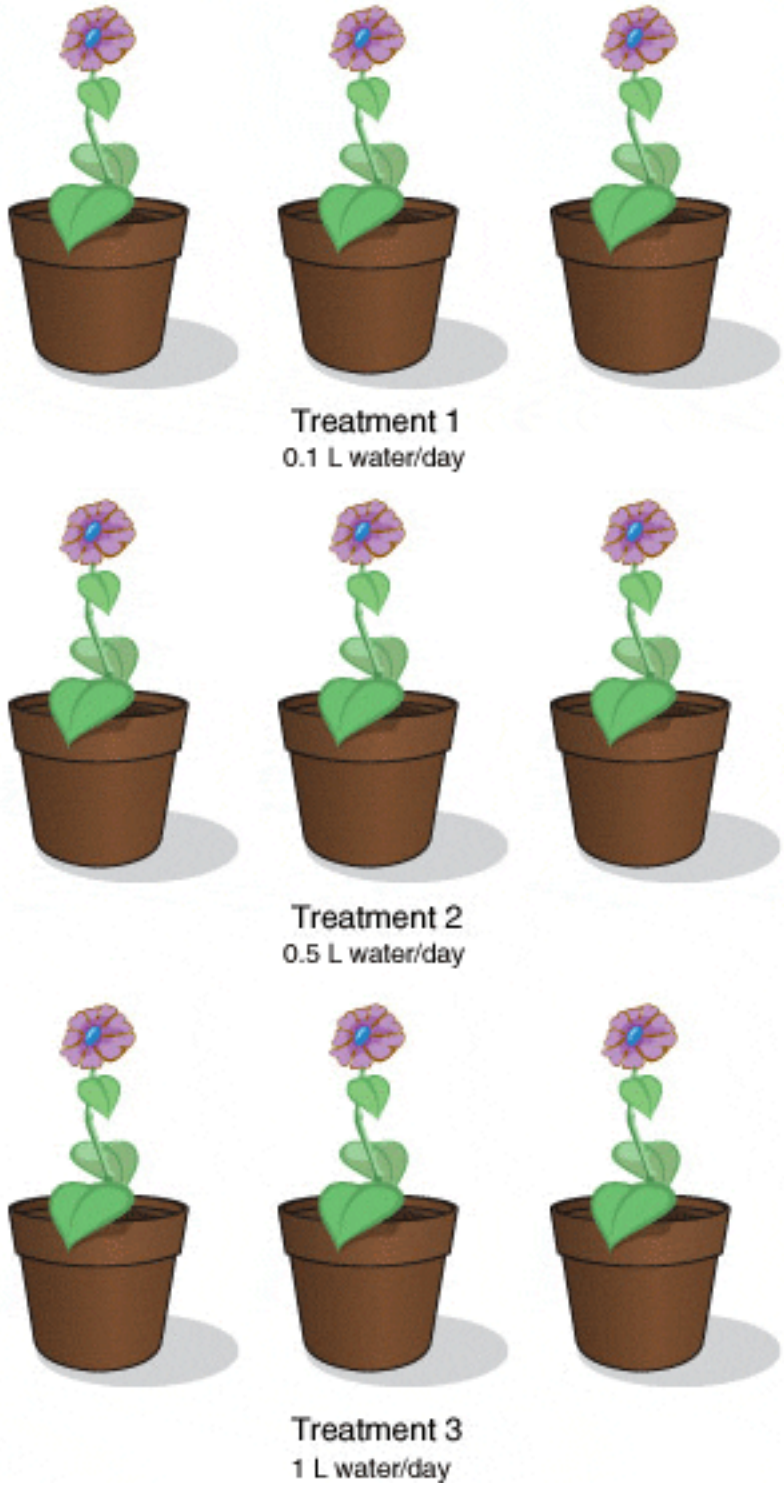



**Figure 3**. In this instance, the experimenter has replicated by having three plants per treatment group.

Pseudoreplication: Taking multiple measurements on the same experimental unit and treating each measurement as an independent data point is called pseudoreplication— not true replication.

Pseudoreplication should always be avoided because the results are not scientifically valid. Taking the same measurements on an experimental unit IS valid when the measurements are taken over time (e.g, once every week) AND the data are presented (See the **Line graph sample**) and analyzed as separate data points in a time sequence.

Example of pseudoreplication: Using one plant for an experiment measuring the effect of nitrogen on growth and counting each branch as a separate experimental unit or replicate, would be an example of pseudoreplication. You need to use multiple separate plants for each treatment. (See examples under Replication above.)

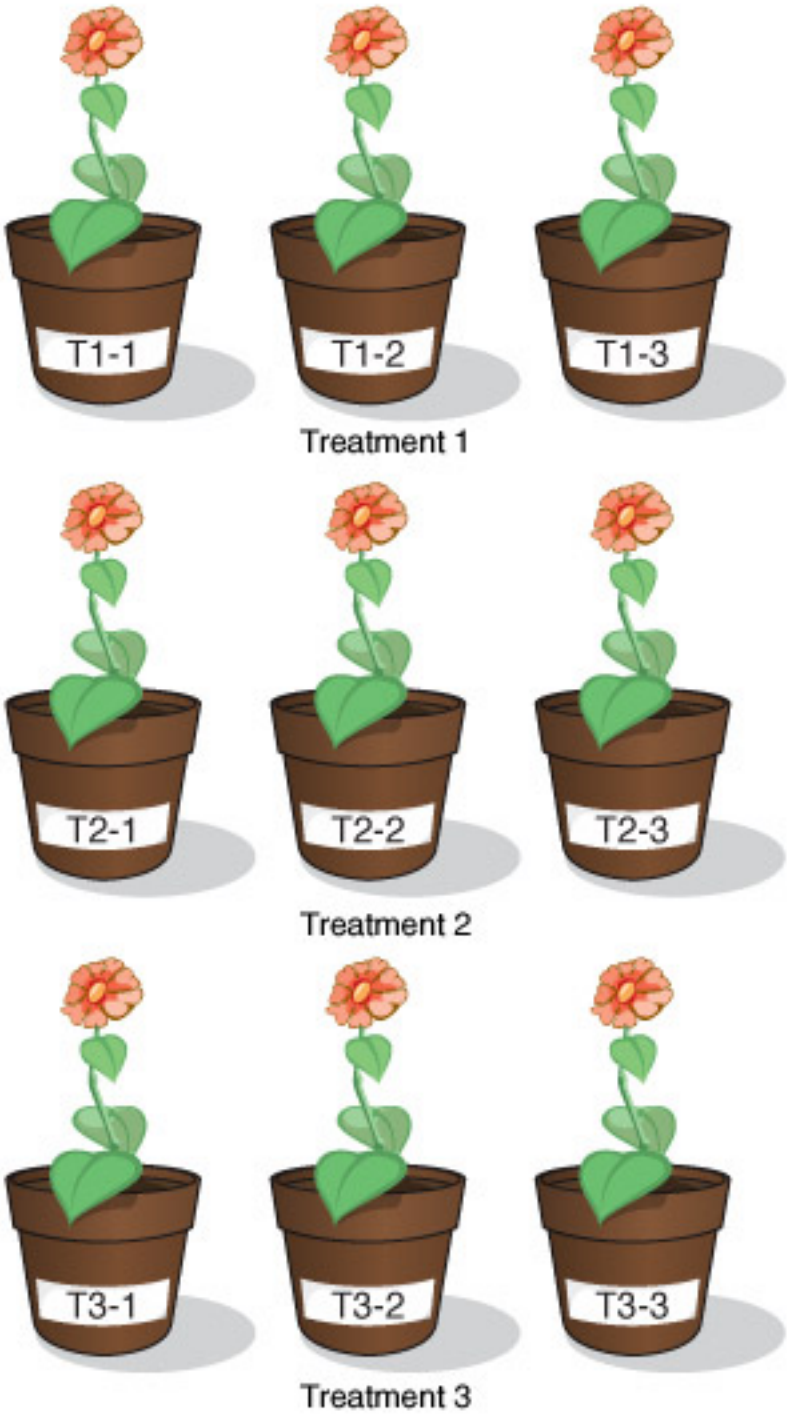
**Controlled experiment**— This is an experiment where only one variable or factor is manipulated and all other variables are held constant. An experiment is controlled if the only factor that is allowed to vary is the independent variable (treatment). All other factors are kept as constant as possible. Controlled experiments can be very difficult to accomplish in nature sometimes, but do the best you can. Nature can sometimes supply you with a ready-made experiment.

**Control**— Do not confuse this with a *controlled experiment*. A control is an experimental unit that is being subjected to all the same conditions as the units actually being treated, except for the control does not receive an actual treatment or receives only a placebo. Example: Nine pots are set up with one plant in each pot. Three pots receive nitrogen in the form of one liter of water mixed with five grams of nitrogen. The other three pots receive one liter of water mixed with ten grams of nitrogen. These are the two different experimental treatments, two levels of nitrogen. The three control pots receive only the one liter of water, but no nitrogen.

In human studies, which many of us are familiar with in the news, there is always a control group. If a new drug is being tested for example, some participants receive the drug while the others receive a placebo, a pill that looks like the drug, but is inert. This accomplishes two things for the researchers: (1) no participant knows if they are receiving the drug or a placebo, making this a *blind study*, and (2) they have a similar group of untreated people (placebo takers) against which to compare the results from those people on the drug.

In ecological studies, especially those in the field, it is not always possible to set up an exact control, but an attempt should be made.

**Blind study**—An experiment in which the people collecting and analyzing the data do not know which experimental units received which treatments. Only after the data are analyzed are the treatments revealed, or decoded. The purpose of this is to reduce any human bias toward an expected outcome. We are all human and it is human nature to be biased, consciously or subconsciously. It is ideal to run blind studies, but that can be even tougher because it means extra people working on the study. In a classroom, however, that can be feasible. One half of the class applies the treatments, the other half collect the data, without knowing which experimental units received which treatments. In the example above of the pots, if the pots have coded stickers on the bottom that only the treatment students understand, then the data takers will not know which plants are getting which treatment and that will reduce their bias (preconceived expectations), and the data will be more objective and reliable. Labels can be as simple as T1-1, T1-2, T1-3, T2-1...T2-3, and T3-1...T3-3. T1, T2 and T3 stand for the treatment (5 g N, 10 g N or 0 g N). The numerals after the dash number each pot within the treatment group.




Note that there are three replicates per treatment.

**Figure 4**. Drawings of pots and labels of treatment 1, treatment 2, control, with code labels.

Now you have the experimental units and treatments set up and are ready to collect data. It may seem straightforward, and usually is, but it always helps to plan ahead exactly what data you will collect, when, how, who will collect it, where it will be recorded, and in what format.

There are different ways to record data, such as numbers, drawings, or words.

**Counting (raw numbers)**— You start off collecting numerical data as counts, called raw numbers. Depending on the mathematical level of the students that may be as far as you need to go. They count something such as the number of flowers on the plants, write the numbers on a **data sheet** or in a science journal, and **graph** those or put them in a **table**.

**Pictures, drawings**— Sometimes the data collected is in the form of a drawing when recording variables such as shape and color. This is useful for all ages, but especially for pre-reading students. The numerical data can be recorded in the form of a drawing. Drawings are usually necessary for presentations to help explain to an audience what the experiment was, how it was conducted, and the results.

**Non-numerical data**— In some experiments the data to be collected is not numerical in nature. It might be color change, intensity of color, or some other qualitative measure such as high, low, or medium light.

**BASIC STATISTICAL FORMULAS**
(Note: Much of the following information is based on *Biometry*, second edition, by Robert R. Sokal and F. James Rohlf, 1981, W. H. Freeman and Co., and *Statistical Methods*, seventh edition, by George W. Snedecor and William C. Cochran, 1980, The Iowa State University Press.)

The following is a very brief primer on statistics. The intention is to whet your appetite and provide some familiarity with basic statistics. Any number of texts exist that can explain these and other statistical methods in detail.

**Averages (mean)**— An average or mean is calculated as the sum of the numbers for one group of treated organisms or plots of land (the experimental units), divided by the number of organisms or plots. Example: You run an experiment on 10 plants, with 5 plants being treated to nitrogen and 5 not receiving nitrogen. Each organism or plot is an experimental unit, so in this case you have 5 replicates in each treatment group. You let the plants grow and measure how tall they get after 2 weeks. See the data table below.

| | Plant Height in cm | Plant Height in cm |
|---|---|---|
| Treatment | No Nitrogen | Nitrogen |
| Plant 1 | 4 | 7 |
| Plant 2 | 7 | 8 |
| Plant 3 | 5 | 7 |
| Plant 4 | 9 | 9 |
| Plant 5 | 10 | 10 |
| Total | 35 | 41 |
| Average Height | 35/5 = 7 | 41/5 = 8.2 |

The average plant heights are 7 cm and 8.2 cm.

The statistical formula for the mean is given as: $\frac{\sum_{i=1}^{n} Y_i}{n} = \overline{Y}$

The number of items in the sample is n, i refers to each individual value, $Y_1$ is the individual value for each

sample item, and $\overline{Y}$ is the mean value. Using the table above for the Nitrogen treatment, n = 5, and $Y_1$ is each value, 7, 8, 7, 9 and 10.

The calculation is done this way: $$\frac{7+8+7+9+10}{5} = \frac{41}{5}$$

$\overline{Y}$ = 8.2 cm

**Median**— is a statistic of location. It is that value of the variable that has an equal number of items on either side of it. For example, in the nitrogen experiment above, the median plant height without nitrogen is found by first ordering the heights, 4, 5, 7, 9, 10. The median is 7 because it has the same number of observations above and below it. If there is an even number of observations (e.g. 4, 5, 7, 9) the median is halfway between the middle, in this case it would be 6 since that is halfway between 5 and 7. The median is commonly used in describing household income. If the median income in an area is $35,000, then half the households have incomes less than $35, 000 and half have incomes greater than $35,000.

**Mode**— Is the value represented by the greatest number of individuals in the sample. E.g. in a sample of twenty-five insects, five are beetles, seven are flies, ten are spiders, and three are moths. The mode is the spiders with ten individuals. The mode is the least used descriptive statistic.

**Range**— Range is a measure of dispersion. It is the difference between the smallest and largest items in a sample. The range of values in the table above for No Nitrogen is from 10 cm to 4 cm. The range is a good indicator of variance in small data sets, but as data sets get larger with many values, *variance* and*standard deviation* are used to determine the range of dispersion around the mean.

**Frequency distribution**— A frequency distribution is an arrangement of statistical data in order of the frequency of each size or value of the variable. The most well known frequency distribution is the bell-curve, also called a normal curve (Figure 6A). How the data are distributed is of great use in ecological experiments. Data can be normally distributed close to the mean (Fig. 6A); spread out around the mean (Fig. 6B); skewed in one direction (Fig. 6C); or even have more than one peak or mode (Fig. 6D).

**Figure 6.** The x-axis is insect species. Each species has been given a number instead of cluttering the axis with the species names.

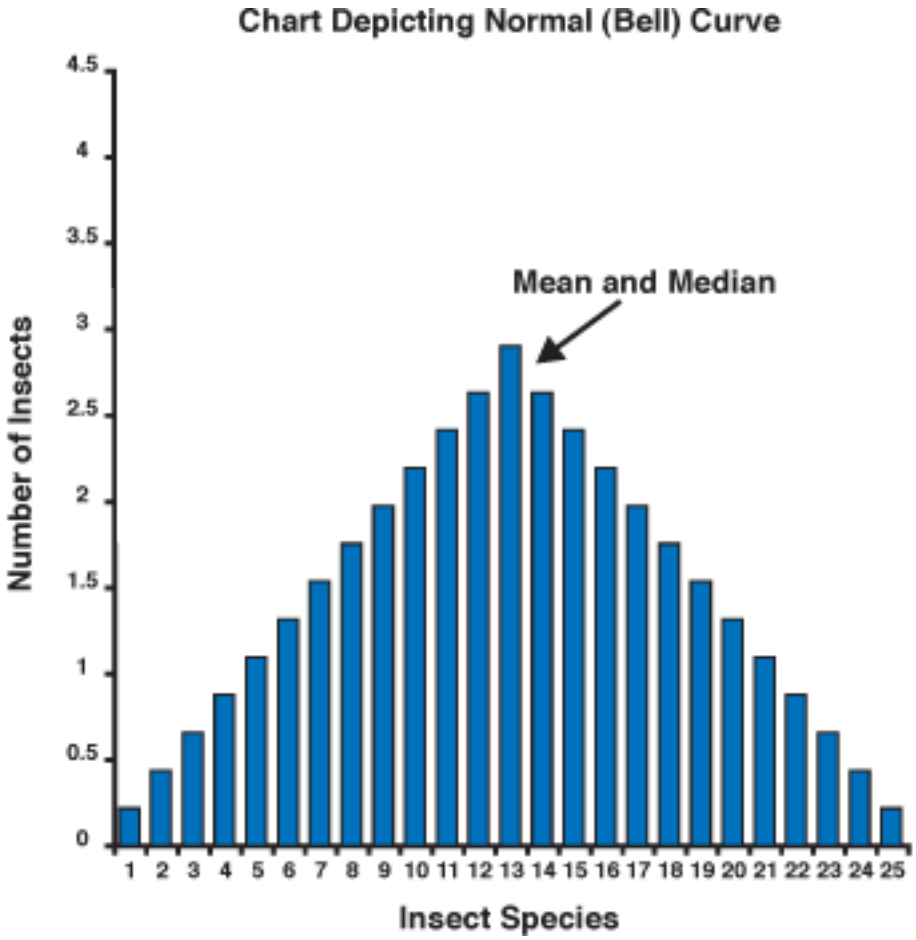



**Figure 6A.** Normal (bell) curve.

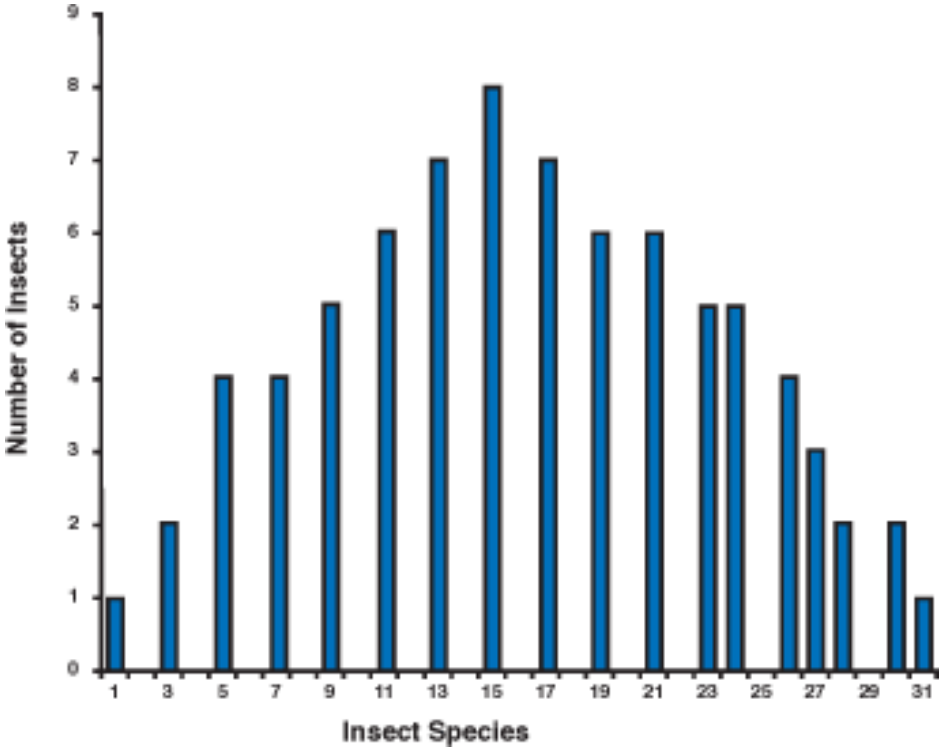



**Figure 6B.** Data spread out around the mean

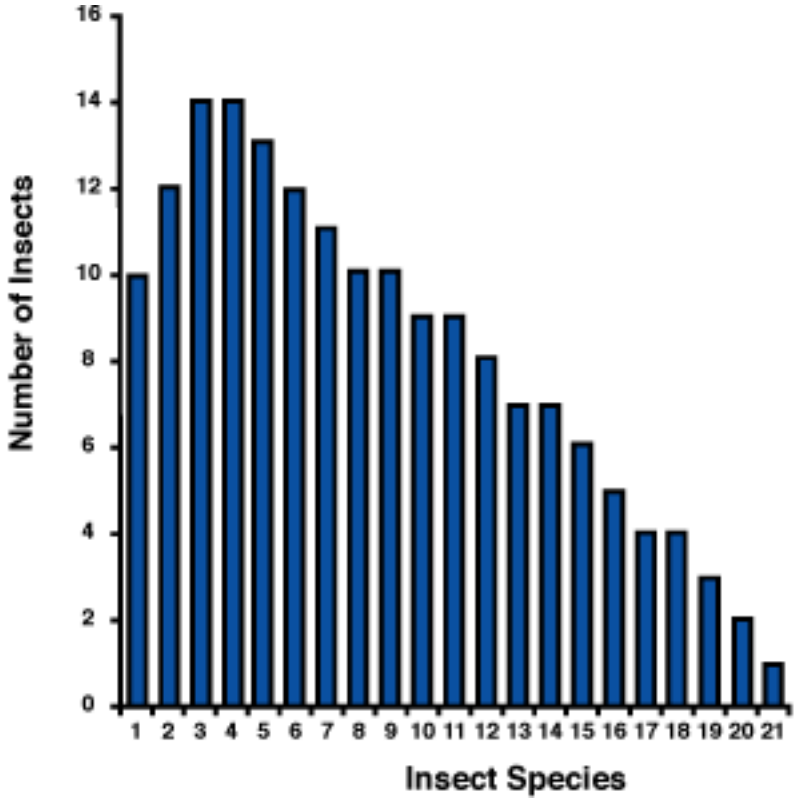



**Figure 6C.** Data skewed in one direction

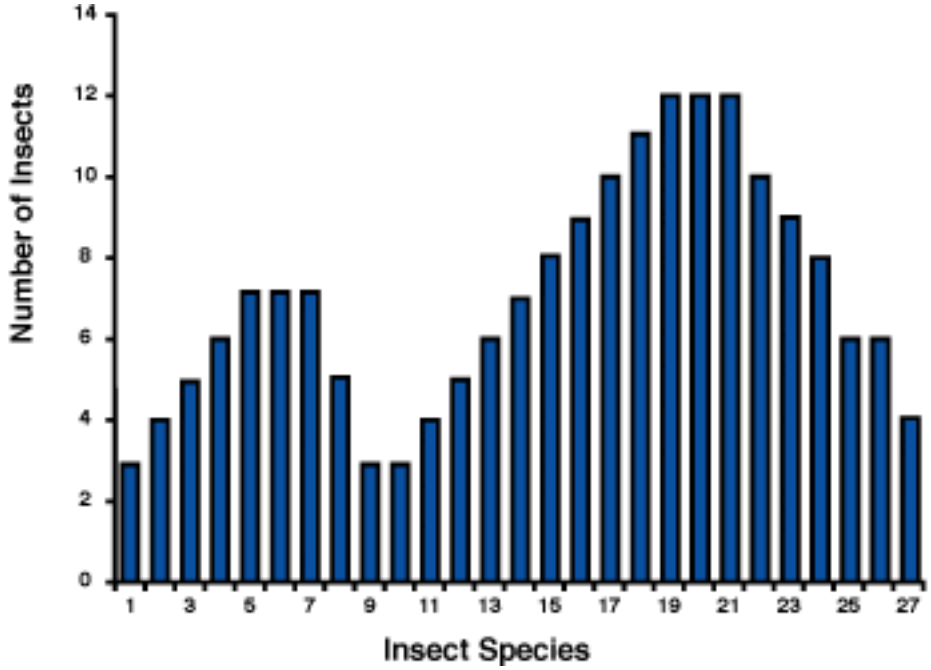



**Figure 6D.** Data with more than one peak or mode

**Variability**— Concerns the extent to which the values in a data set differ from the mean. Variability in a set of data is measured using variance and standard deviation calculations. Following are two sets of data, each having the same mean, but just by looking at the numbers you can see that Sample 1 values range from 3 to 21, while Sample 2 values only range from 10 to 14.

Sample set 1 values: 12, 6, 15, 3, 12, 6, 21, 15, 12, 18. Mean is 12.
Sample set 2 values: 12, 10, 12, 14, 13, 12, 11, 14, 12, 10. Mean is 12.

The formula for variance $s^2$ is:

$$s^2 = \frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1}$$

The calculation for variance using Sample set 1:

$$\frac{12^2+6^2+15^2+3^2+12^2+6^2+21^2+15^2+12^2+18^2-\frac{(12+6+15+3+12+6+21+15+12+18)^2}{10}}{10-1}$$

$$\frac{1784 - \frac{14400}{10}}{9} = \frac{1784 - 1440}{9}$$

$$s^2 = 32$$

The calculation for variance using Sample set 2:

$$\frac{12^2+10^2+12^2+14^2+13^2+12^2+11^2+14^2+12^2+10^2 - \frac{(12+10+12+14+13+12+11+14+12}{10}}{10-1}$$

$$\frac{1458 - \frac{14400}{10}}{9} = \frac{1458 - 1440}{9}$$

$$s^2 = 2$$

The mean is 12 for both sets, but the variance is 32 in set 1 and only 2 in set 2. The variance, or distance from the mean, is much greater in the first set.

**Standard Deviation**— The more commonly used statistic is standard deviation, which is calculated as the square root of the variance. The formula for standard deviation (s) is: $s = \sqrt{s^2}$ In Sample 1 the standard deviation is 5.6 cm and in Sample 2 it is 1.4 cm.

The standard deviation is often used on **graphs** along with the mean to visually represent the amount of variability in the data. The more variability in the data set, the less likely there is a difference between treatment groups. On a bar graph, if the error bars overlap, it is a good indication that the treatments are not different from each other, at least in the statistical sense.

**T-test**— This is a standard test used to compare the means of two treatment groups. It is listed in Microsoft Excel and stats software packages such as SAS. It is not possible here to go into an actual definition and description of the t-test. Please consult a statistics textbook for a detailed explanation.

For the curious, the basic definitional formula is: $t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$

The useful formula is: $t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\bar{x}_1 - \bar{x}_2}}$ where $s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{S^2 p(n_1 + n_2)}{n_1 n_2}}$ and

$$\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

$n_1$ is the sample size for treatment group 1 and $n_2$ is the sample size for treatment group 2 in an experiment. $s_1^2$ is the variance for treatment group 1 and $s_2^2$ is the variance for treatment group 2. $\overline{X}$ is the mean for each treatment group. $S^2p$ is the variances of each treatment group pooled together.

| | Hormone A | Hormone B | |
|---|---|---|---|
| | 57 | 89 | |
| | 120 | 30 | |
| | 101 | 82 | |
| | 137 | 50 | |
| | 119 | 39 | |
| | 117 | 22 | |
| | 104 | 57 | |
| | 73 | 32 | |
| | 53 | 96 | |
| | 68 | 31 | |
| | 118 | 88 | |
| | | | |
| Totals | 1067 | 616 | |
| n | 11 | 11 | |
| $\overline{X}$ | 97 | 56 | |
| $\sum \overline{X}$ | 111,971 | 42,244 | |
| $\sum X^2/n$ | 103,499 | 34,496 | |
| | 847.2 | 777.4 | |
| df | 10 | 10 | |
| t value | | | 3.38 |
| probability | | | 0.01584984 |

**Figure 7.** Results of a t-test.

**ANOVA**— Analysis of Variance or ANOVA, is one of the most commonly used statistical tests of the variability between more than two treatment groups. As with the t-test, it is not possible here to go into an actual definition and description. It is listed in Microsoft Excel and stats software packages such as SAS. Please consult a statistics textbook for a detailed explanation.

Treatment: Nitrogen Levels

| | 0ppm | 5ppm | 10ppm |
|---|---|---|---|
| Replicates | | Height (cm): | |
| Plant 1 | 5 | 5.5 | 6 |
| Plant 2 | 4.5 | 6 | 6.8 |
| Plant 3 | 4.7 | 6.2 | 7.1 |
| Totals | 14.2 | 17.7 | 19.9 |

Anova: Single Factor

| TREATMENT — Nitrogen Levels | | | |
|---|---|---|---|
| | 0 ppm | 5 ppm | 10 ppm |
| *Replicates:* | | Height (cm): | |
| *Plant 1* | 5 | 5.5 | 6 |
| *Plant 2* | 4.5 | 6 | 6.8 |
| *Plant 3* | 4.7 | 6.2 | 7.1 |
| Totals: | 14.2 | 17.7 | 19.9 |

Anova: Single Factor

SUMMARY

| *Groups* | *Count* | *Sum* | *Average* | *Variance* |
|---|---|---|---|---|
| Column 1 | 3 | 14.2 | 4.73333333 | 0.063333333 |
| Column 2 | 3 | 17.7 | 5.9 | 0.13 |
| Column 3 | 3 | 19.9 | 6.63333333 | 0.323333333 |

ANOVA

| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-Value* | *F crit* |
|---|---|---|---|---|---|---|
| Between Groups | 5.5088889 | 2 | 2.7544444 | 15.9935484 | 5.14324938 | 0.00394045 |
| Within Groups | 1.0333333 | 6 | 0.1722222 | | | |
| Total | 6.5422222 | 8 | | | | |

**Figure 8.** ANOVA results table.

**Correlation coefficient**— The correlation coefficient, r, is a measure of the closeness of the relationship between two variables. These are not **independent and dependent variables** , but just two variables that happen to exist in the same time or place. An example would be examining the correlation between the density of plants in a plot and the heights of the plants. The value of r can range from -1 to +1. As r approaches +1 the more positive the relationship. An r-value close to zero indicates no relationship and r-

values below zero indicate a negative relationship.

Density could be designated as $x_1$ and height as $x_2$. A sample data set could be: Heights are 4, 6, 8, 10, 10, 12, 15, 18 cm and densities are 1, 5, 10, 15, 20, 25, 30 and 35 plants per plot. Use the equation for correlation coefficient, r in the equations section below, or the correlation function in a software program.

$$r = \frac{\sum x_1 x_2}{\sqrt{\left(\sum x_1^2\right)\left(\sum x_2^2\right)}}$$

$$r = \frac{4\times1+6\times5+8\times10+10\times15+10\times20+12\times25+15\times30+}{\sqrt{\left(4^2+6^2+8^2+10^2+10^2+12^2+15^2+20^2\right)\left(1^2+5^2+10^2+15^2+2\right.}}$$

In this case, r = 0.98, which is close to 1, indicating a close correlation between density and plant height. If the r-value is positive, then there is a positive correlation that as density increases height increases.

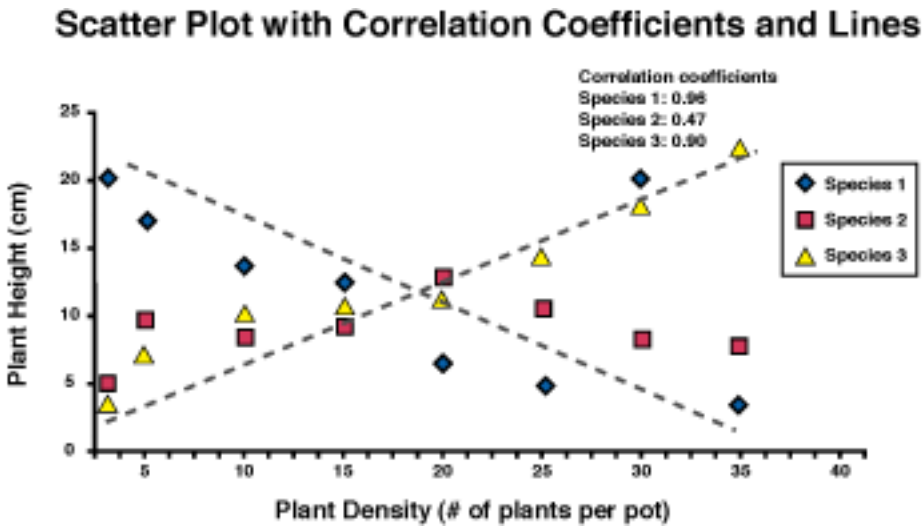



**Figure 9.**

**Outliers**— An outlier is a data point far from the rest of the data points. Outliers can occur for any number of reasons, such as human error, genetic abnormality, some unknown interference, etc. What to do with an outlier? That depends on the severity and the reason, if known. Depending on its value, an outlier can seriously affect the mean and thus any statistical tests. In some cases the data are analyzed both with and without the outlier value, and both are reported in the paper. The reader is then given all the information about the experiment and can come to her own conclusion

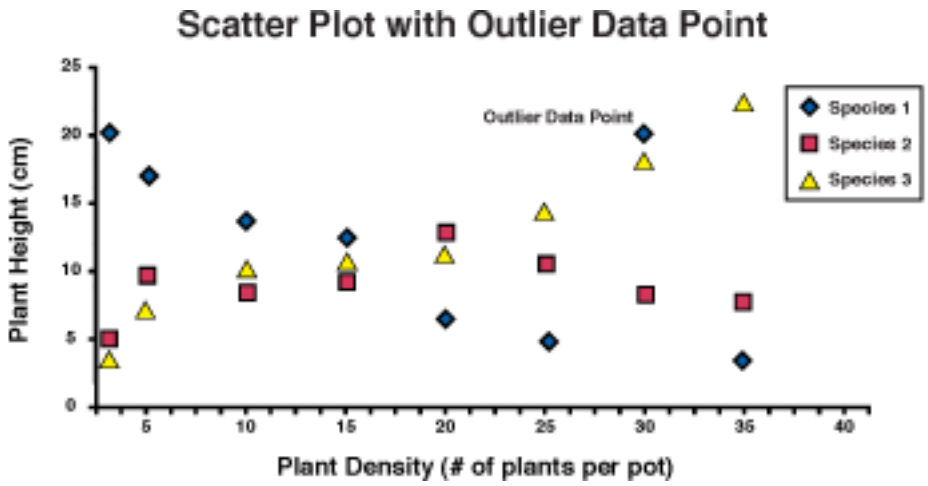



**Figure 10.**

**Probability**— Most statistical tests are based on probabilities. It is not possible here to go into an actual definition and description of probability. Please consult a statistics textbook for a detailed explanation.

**ECOLOGICAL EQUATIONS COMMONLY USED**

**Diversity, biodiversity** - Species richness is one measurement of the biodiversity in a habitat. The formula is:

$$d = \frac{(S-1)}{\log N}$$

*d* is the species richness index, *S* is the number of species and *N* is the number of individuals of all species in the population.

**Population Density** - is the measurement of the number of organisms per area. Density could be all organisms, only the individuals of a particular species, or a group of species. D = $\frac{N}{A}$

N is the number of individuals being counted and A is the area of the habitat under consideration. Area is calculated as A=(L)(W). If the volume of space is defined as the habitat space, then use V=(L)(W)(H) where L is length, W is width, and H is height.

**Birth and Death rates—**
$N_{now} = N_{then} + B - D + I - E$ or $N_{future} = N_{now} + B - D + I - E$

N is the number of individuals in the population, presently, in the past, or in the future.

B is number of births.
D is number of deaths
I is number of immigrants
E is number of emigrants

The equation above defines the main aim of ecology: to describe, explain, and understand the distribution

and abundance of organisms. (Begon, Harper and Townsend, 1999)

**Reproductive rate**, $R_o$

$$R_o = \sum l_x m_x$$

$l_x$ is the proportion of a cohort (members of a population of same age) surviving to reproductive age

$m_x$ is the number of offspring (eggs, seeds, young) produced per surviving individual

**Logistic equation**— For populations with continuous birth and death.

$$\frac{dN}{dt} = rN$$

This equation represents the speed at which a population increases in size, N, as time, t, progresses. dN is the change in the number of individuals, dt is the change in time, N is the number of individuals in the population, and r is the intrinsic rate of natural increase.

R is calculated as: $\frac{\ln R_o}{T}$ (Begon, Harper and Townsend, 1999)

**Lotka-Volterra model of interspecific competition**

$$\frac{dN}{dt} = rN\frac{(K-N)}{K}$$

Where dN is the change in number of individuals, dt is the change in time, N is the number of individuals in the population and K is the carrying capacity.

**Volume formulas**— Rectangle or cube: V = hlw; h is height, l is length, and w is width.

Sphere: $V = \frac{4\pi r^3}{3}$ r is the radius

**Velocity of a stream**— Calculate Velocity, V, by placing a floating object (a stick will do) in the water at the start point; time it with a stopwatch for 20 seconds. Mark the end point and measure the distance traveled. Calculate V as: $V = L/T$ where L is length or distance in meters and T is time in seconds.

**Streamflow or Discharge—** is the rate at which a volume of water flows past a point over a specified unit of time. The formula is: $Q = L^3/T$. L is length or distance in meters and T is time in seconds. If a 1 $m^3$ container is filled in 5 sec Q would equal 1/5 = 0.2 $m^3/s$.