Basic Statistical Stuff

I. Types of Variables:

A. Measurement Variables - can be expressed in a numerically ordered fashion

1. Continuous Variables - infinite # of possible measurements between any 2 measurements.

2. Discontinuous Variables - Fixed integers

- e.g., seed counts (can have 1 or 2 seeds but not 1.5 seeds)

B. Ranked Variables (may be continuous or discrete) - For example:

Individual	Wt	Rank
a	10.5	
b	7.3	
c	8.4	
d	11.6	

C. Attributes (Nominal Variables)

- Qualitative variable like ♂ or ♀, On/Off, Dead/Alive

II. Statistics of Location - Measures of "Central Tendency"

A. Mode = most frequent value in the sample

B. Median = value for which ½ of sample obs. are larger & ½ of sample obs. are smaller

1. list observation from low to high
2. middle observation in the list is the median

C. Mean (population estimate of μ)

1. = arithmetic mean
2. = $\sum x / n$
(Remember that n= the number of "things" in your sample)

D. When to use mode, median, or mean?

1. If you want to know a typical or representative value calculate the mode
2. If distribution of data is skewed and you need to know the center of the distribution then use median
3. If data are normally distributed and you want to have a relative and average value use mean

III. Dispersion Statistics

A. Range

1. Simplest measure of the spread of the data. Tells you the difference between the highest and lowest values.
2. Expressed as x_{max} to x_{min} (e.g., 3 to 15 grams)

B. Deviations from Mean

1. Sum of Squares: $SS = \sum (x - \bar{x})^2$

2. Variance (population estimate of σ^2): $s^2 = \frac{SS}{d.f.}$

a. where d.f. = degrees of freedom = n-1

b. A measure of spread that takes both deviations from your data (away from the mean) and frequency of these deviations into consideration

3. Standard Deviation: $\boldsymbol{s.d.} = \sqrt{s^2}$ or $\sqrt{\text{var.}}$

a. Reflection of the deviation from the mean and frequency of these deviations but is scaled to the data. In other words, Std Dev. is in same units as mean (i.e.,grams, number of plants, %, etc).

b. Or, average number of units that all observations in the sample differ from the mean.

c. What does it mean? It means that:

(1) 68.3% of observations fall within the range $\bar{x}$ ± 1 s.d.

(2) 95.5% observations in a sample fall within the range $\bar{x}$ ± 2 s.d.

(3) 99.7% observations in a sample fall within the range $\bar{x}$ ± 3 s.d.

(4) 95% observations in a sample fall within the range $\bar{x}$ ± 1.96 s.d.

4. Standard Error

a. Also a measure of deviation from the mean but takes into account size of the data set. (The bigger the data set the smaller the Std. Err)

b. It reflects th standard deviation of sample means.

(1) if you were to go out and collect a series of samples (each made of several observations) you could get a $\bar{x}$ and s^2 the group of samples

(2) Central Limits Theorum states that $\bar{x}$'s of a large group of samples will be normally distributed and $\bar{x}$ of these sample $\bar{x}$'s will = μ

c. Stderr = $s_{\bar{x}} = \sqrt{\dfrac{s^2}{n}}$ or $\dfrac{s.d.}{\sqrt{n}}$

d. When to use $\boldsymbol{s.d.}$ or stderr?

As sample size get larger the stderr becomes smaller even though the variance may not change. Therefore, if you have a large sample size (e.g., hundreds of observations) it is better to use the s.d. otherwise use stderr. Make sure to report number of observations (n) so that readers can convert between s.d and stderr.

IV. Confidence Limits around the Mean (Confidence Interval)

A. Range of numbers that includes μ, given a certain level of probability (confidence)

B. Based on **z-score**: C.I. = $\bar{x} \pm z_{(1\text{-}\alpha)}$ stderr

1. $z_{(1\text{-}.1)}$ or z for a 90% C.I. = 1.645

2. $z_{(1\text{-}.05)}$ or z for a 95% C.I. = 1.96

3. $z_{(1\text{-}01)}$ or z for a 99% C.I. = 2.57

C. Based on **t-value:** C.I. = $\bar{x} \pm t_{(\alpha,\ d.f.)}$ stderr

1. If you want a 95% C.I.

a. look up a t for α = .025 and sample d.f. for a 1-tailed t-table

b. look up a t for α = .05 and sample d.f. for a 2-tailed t-table

2. If you want a 90% C.I.

a. look up a t for α = .05 and sample d.f. for a 1-tailed t-table

b. look up a t for α = .10 and sample d.f. for a 2-tailed t-table

V. Sample Size Estimates Based on Variance (non-parametric approach)

A. Running Mean and Variance approach

1. First, go out and take a few samples

2. Plot each observation (1 through n) on the x-axis

3. Plot mean or variance on y-axis

4. When mean or variance begin to level-out an adequate sample size has probably been reached (you will not capture much more of the population variance with additional samples)

B. Statistical Approach (Equations based on sample mean and variance).

1. To use a statistical approach you will need a $\bar{x}$ and s^2.

    a. You can go out and take a few samples (i.e.,20) then estimate $\bar{x}$ and s^2.

    b. Or, you may be able to look at data colleted in a similar way on a similar site with similar precipitation pattern.

2. With these values in hand, there are several approaches you can use to estimate the number of samples you will need to collect for a statistically valid sample.

3. Steins two-stage sample (Steel and Torrie 1960)

$$n = \frac{t^2 s^2}{d^2} \text{ or } n = \left(\frac{t \times s}{d}\right)^2$$

    a. Where t^2 = value from t-table (with desired α and sample *d.f.*) squared

    b. α in this case is % confidence that if you go out and take a sample, your sample mean will be within ***d*** units of the true population mean (μ). Or, that your sample mean will be in the range of μ ± ***d***

    c. ***d*** = the half-width of the desired interval around the actual population mean (μ) in which you would like to insure that your sample mean ($\bar{x}$) falls.

    d. How do you set ***d*** ? Good question. Basically, you need to rely on your skill as an ecologist. Look at the sample you collected and see how many units you are willing to be from the real population mean. How much are you willing to "stomach"? For, example if you are measuring biomass production in g/m^2, and you are going to do a study, how many grams are you willing to let your estimate of production (your sample mean or $\bar{x}$) be from the actual production (if you could measure the population average or μ). Are you willing to be within 10 g of the population average? 20g? 30g?

4. Bonham approach: (pg 65-67 in Bonham 1989)

$$n = \frac{t^2 s^2}{(k\bar{x})^2}$$

    a. where t^2= the value from a t-table (for α and sample *d.f.*), squared

    b. s^2= the sample variance

    c. ***k*** = the proportion or precision of the true difference between the sample mean ($\bar{x}$) and the population mean (μ). If you want you sample mean to be within 10% of the population mean then ***k***= .10. For example, if you initial sample mean is 250 kg/ha and you want your study sample to have a mean within 25 kg/ha of the real population mean, your *k* would be .10.

5. West approach (West 1986)

$$n = \frac{(2ts)^2}{W^2}$$

   a. where *t*= the value from a t-table (for α and sample *d.f.*)

   b. *s* = standard deviation f the sample

   c. *W* = the total width of the desired interval around the population mean (μ) in which you would like your sample mean ($\bar{x}$) to fall. Note, that ***d*** and ***k*** in the equations above are half-widths of intervals (e.g., μ ± ***d***), however *W* is the total width, from low to high (about double ***d*** or ***k***).

6. If you have trouble setting ***d*** or ***k*** *in the previous equations, here is an ALTERNATE* equation based on confidence interval (based on Ott 1984)

   a. $$n = \frac{(t^2 s^2)}{(\frac{1}{2} c.i.)^2}$$

      (1) where t^2= the value from a t-table (for α and sample *d.f.*), squared

      (2) s^2= the sample variance

      (3) *c.i.* is the confidence interval you calculate based on the equations above. Your *c.i.* is usually expressed as $\bar{x}$ ± some value. For example, 40kg/ha ± 10 kg/ha. The total interval would be 20 kg/ha. However, in the equation you simply take the *c.i* you calculate (as a half-width interval) and divide it in half. For this example, 40kg/ha ± 10 kg/ha, the *c.i* would be 10. Therefore, the denominator of the equation would be $(½\ 10)^2 = (5)^2$=25.

References:

Bonham, C.D. 1989. Measurements of terrestrial vegetation. John Wiley & Sons. New York.

Steel, R.D. and J.H. Torrie. 1960. Principles and procedures of statistics. McGraw-Hill Book Co. New York.

Ott, L. 1984. An introduction to statistical methods and data analysis, 2nd ed. Duxbury Press, Boston, MA.

West, W.H. 1986. Statistical Analysis. *In*: Cooperrider, A.Y., R.J. Boyd, and H.R. Stuart (eds.). Inventory and Monitoring of Wildlife Habitat, Pub # BLM/YA/PT-87?001+6600.