In statistics, the word **population** refers to all possible values of a measurement about which one wishes to draw conclusions, e.g., the size of **all** salmon fry released by hatcheries in B.C. However, it is often difficult to measure an entire population, e.g., all the salmon fry released in B.C.; therefore, scientists may draw a **sample** (a smaller subset of organisms or measurements) from the larger population in order to make inferences about that population. It is important that the measurements we use in our sample are chosen at random; otherwise these samples may contain a bias that is not representative of the entire population.

We use statistics in order to draw conclusions from our measurements in the sample group studied. We can then apply these conclusions, with some probability, to the larger population. For example, if we sampled 55 salmon fry from a typical hatchery in B.C., we could draw conclusions about the average size of fry in the larger population (all the fry in that hatchery). This conclusion would be based on the average length of our sample fry.

Data that are **continuous** are analyzed with measures of **central tendency** and **variation** (see following discussion for definitions). Examples of continuous data are rates, organism or population size, time, etc. Means of two samples are compared statistically using a student t-test; means of more than two samples are compared statistically using an analysis of variance (ANOVA).

Data that are **discrete** are recorded as **frequencies** of observations and are analyzed using a **Chi-square (χ^2) test** (pronounced "ky" as in "sky").

The symbols used to represent these statistics may vary among authors (and therefore may differ depending on the statistics text you consult) but the underlying principles remain the same. Use the following ideas to analyze your data and draw conclusions about the larger population. **Once you have analyzed your data, you then will have a basis for discussing the meaning of your results.**

For more information on the statistics used in analyzing your data, you can consult any elementary statistics text, e.g., Sokal, R.P. and Rohlf, F.J. 1987 (or more recent edition). Introduction to Biostatistics. W. H. Freeman and Company, San Francisco, or Zar, J.H. 1999 (or more recent edition). Biostatistical Analysis. Prentice-Hall, Inc. Englewood Cliffs, N.J.

MEASURES OF CENTRAL TENDENCY AND VARIATION: MEAN, VARIANCE, STANDARD DEVIATION, 95% CONFIDENCE INTERVALS

Mean

The mean ($\bar{x}$), is the average of the data points in a treatment.

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

Means should be reported to the **same number of decimal places as the original data**.

Variance

The variance (s^2) is a measure of how much the data values scatter around the mean. A small variance indicates the data are tightly clustered around the mean. The larger the variance, the more scattered the data. The variance is calculated by summing all the squared deviations from the mean (a deviation is the difference between an individual measurement and the mean).

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

The numerator is referred to as the sum of squares (SS) and the denominator is the degrees of freedom (df). Note that if you are **reporting** variance values, they are expressed to the same number of decimal places as the data entries contain. If you are using them as an intermediate calculation, you should keep the full number of decimal places in your calculator. Calculate the variance for each treatment separately.

Standard deviation

The **standard deviation** (s) is a related statistic, which is the square root of the variance:

$$s = \sqrt{s^2}$$

Variance and standard deviation are very useful statistics for describing the variation of data in a sample. **If the data are normally distributed** (i.e., performing the experiment with a large number of subjects produces a bell-shaped curve), 68% of all data points fall within one standard deviation on either side of the mean.

95% Confidence Intervals (C.I.s)

The mean and standard deviation are statistics that describe the data points in a sample. As such, they are **estimates** of the true mean and standard deviation of the **entire** population of measurements. We can then use these estimates to calculate an interval or range of values within which the mean of the entire population (**true mean**) will likely fall. The level of confidence (e.g., 95%, 99%) indicates the probability that the true population mean will fall within these limits. For example, by calculating a 95% confidence interval, we can say that 95% of the time, the true mean will lie within this range. Most biologists use a 95% level of confidence.

The formula for calculating the **95% Confidence Interval** (C.I.) of the mean is:

$$\text{C.I.} = \overline{x} \pm 1.96\frac{s}{\sqrt{n}}$$

where *s* is the standard deviation and *n* is the sample size. This formula usually applies when sample sizes are large but it is a close-enough approximation for our purposes. Means and 95% Confidence Intervals are final calculations and should be reported to the **same number of decimal places as the original data**. Standard deviations and

variance values are intermediate calculations and you should retain as many decimal places as displayed on your calculators.

Means, standard deviations and 95% confidence intervals are statistics that describe the data points in a sample. In order to determine **which, if any**, means in a data set are significantly different, you must apply a statistical test, such as a t-test or analysis of variance (these files are also in the statistics folder, as is the file for calculating the chi-square statistic).