Basic Statistics Review

Students are strongly encouraged to take a course in basic statistics (STA 120 or community college equivalent) prior to taking Biometrics. A basic statistics course not only covers fundamental material critical for success in Biometrics, but also provides an extensive introduction to topics such as probability, which are important in the biological sciences. Below is a brief review of the material from basic statistics which is necessary for Biometrics. Readings cited below are from Zar (the textbook in Biometrics), and are crucial in understanding the material. Readings are cited from both the 4th and 5th editions of Zar. Either edition may be used.

Basic Statistics Review – Part One

Definitions; Data Scales; Frequency Distributions; Central Tendency
(See Zar 4th ed. or 5th ed.: pages 1-5; 6-15; 18-19)

In a scientific/statistical context, it is important to distinguish between **data** and **statistics**. In common usage, these terms are used interchangeably, but the distinction is critical in science.

Data (singular: datum) are numerical facts.
If a person is five feet, six inches tall – that is a datum – it's not a statistic.

Statistics are estimates of population parameters.
A **population** is defined biologically – essentially, the organisms being studied. Populations are large, and they are dynamic (their membership changes). A population often includes individuals who are no longer members, as well as individuals not yet in the population. For example, if we define a population as "all Cal Poly students", we may be including people who've left the university as well as future students. We may not want to restrict the population to people who are students at this instant.

A **population parameter** is a numerical feature of the population. If our population is all Cal Poly students, a numerical feature might be the average (arithmetic mean) height of all Cal Poly students. It is generally impossible (or at least difficult) to quantify a population parameter. Consider how difficult it would be to get the average height of all Cal Poly students – especially if the population includes past and future students! Since we can't quantify the average height, we **estimate** the average from a **random sample** of students. Our **estimate** of the **population parameter** is a **statistic.**

A **random sample** is a sample from the population which is representative of the population with respect to the trait being measured. Doing good sampling requires knowing the biology of what you're working with – it doesn't require mathematical or statistical skills. For example – if you're trying to estimate the average height of all Cal Poly students, you probably know it's not a good idea to hang around the gym after basketball practice, and ask the first 30 people coming out the door how tall they are. Basketball players are not representative of the student population with respect to height. Also – if there's a difference in height between men and women, you would want your sample to reflect the sex ratio in the Cal Poly population. Getting a good

(unbiased) random sample of students requires knowledge of students – not of math.

When you estimate a population parameter from a random sample, keep in mind that your estimate is almost certainly wrong. That is, your sample value won't be exactly the same as the value of the population parameter. If you get a random sample of 30 students' heights, the mean of those 30 will not be the same as the mean of all ~20,000 students. That's why it's called an "estimate". The fact that a sample estimate is not the same as the population parameter being estimated, is called **sampling error.** The word "error" here doesn't mean that somebody has made a mistake, it just means that you can't expect the sample value to be exactly the same as the population value.

As scientists, we must deal with sampling error. We always want to know what happens at the population level, but we must always work at the sample level. **Statistical analysis allows us to make conclusions at the population level using sample data.** Statistical analyses are a critical tool in understanding what our data are telling us about the biology of the system being studied.

Data

Although there are four generally recognized data scales, we will combine two scales (ratio and interval), and therefore discuss three scales below. It is important to think about the scale of your data, because part of the decision of what analysis should be done depends on the scale.

1. Nominal-scale, Categorical, Attributes

Subjects are classified by a quality rather than a number or ranking. For example, consider the variable "colors of shoes worn by students". Values of the variable might be white, black, brown, red, and so on. Although the colors are the actual values, we usually are interested in the **frequencies**, i.e. how often do the different values occur. How many students wear white shoes? How many wear black shoes? Other examples of nominal scale data include: eye color, gender, evergreen vs. deciduous trees, flower color.

2. Ordinal-scale

These are **ranks**, a relative ordering of measurements. (L. *ordinal* = order). For example, you might assign a value from 1 to 10 to all the movies you see; with 1 being the worst and 10 being the best. Other examples are letter grades (A, B, C, D, F), size classes of animals (1 = very small, 2 = small, 3 = medium, 4 = large, 5 = very large).

3. Interval-scale and Ratio-scale

These scales have a constant interval between successive values. For example, the interval between 70-75 degrees and 35-40 degrees is the same (5 degrees). Usually, the distinction between Interval- and Ratio- scales is not an issue in data analysis, and the same tests can be used for both scales. For the record, the distinction is:

> **Interval-scale** data do not have a true, biologically meaningful zero. Examples would be temperature measured on the Celsius scale, time of day, time of year, compass directions.
> **Ratio-scale** data have a true, biologically meaningful zero (there are no negative

values). Many of our common measurements are ratio-scales: e.g. centimeters, grams, °K, size, volume, weight, number (i.e. eggs/nest) are measured as ratio-scales.

Discrete vs. Continuous Data
Another way to categorize ratio-scale and interval-scale data is to distinguish between discrete and continuous data.

Discrete (meristic, discontinuous)
Variable can only take on integer values (counts). Examples: number of teeth per animal, eggs per nest, leaves per tree, aphids per leaf, base pairs per chromosome.

Continuous
Variable always has a value between **any** two other values, no matter how close together the values might be. Measurements of continuous variables are **always approximations** and depend on the precision of the instrument that you are using to measure. Examples: height, weight, age, systolic blood pressure.

Accuracy, Bias, Precision, and Significant Figures

Accuracy
The nearness of a measurement to the actual value of the variable being measured.

Bias
Systematic error that reduces accuracy.

Precision
Closeness of repeated measurements of the same quantity.

Relationship among Accuracy, Bias, and Precision
If there is no bias, then by increasing precision, you will **increase accuracy**. Thus, you want to **minimize bias** by appropriate experimental design and selection of experimental units, and **maximize precision** of measurement.

How to Measure and Report Data – Significant Figures
The last digit to the right in a measurement implies the precision with which the measurement was made. Also, the number of **significant figures** implies precision. The number of significant figures is the number of digits from the leftmost non-zero digit to the right-most digit that is not zero, or a zero considered to be exact. This is difficult to describe, but fairly simple in practice.

Examples
435.1 has 4 significant figures.

2.70 implies 3 significant figures (use 2.7 if you measured to the nearest tenth).

64,000 or 6.4×10^4 has 2 significant figures, but
6.40×10^4 has 3 significant figures!

If suitable instruments are available, **measure** enough significant figures so that the number of unit steps from the smallest to the largest measurement is between **30 & 300**.

Example
Suppose you are measuring the height of some plants, and the tallest plant is 185 cm, while the shortest plant is 162 cm. You should measure to the nearest 0.1 cm (185 – 162 = 23 units; but 185.0 – 162.0 = 230 units). If you report 185, you are implying that the true measure lies between 184.5 and 185.5. But if you report 185.0, you are implying that the true measure lies between 184.95 and 185.05.

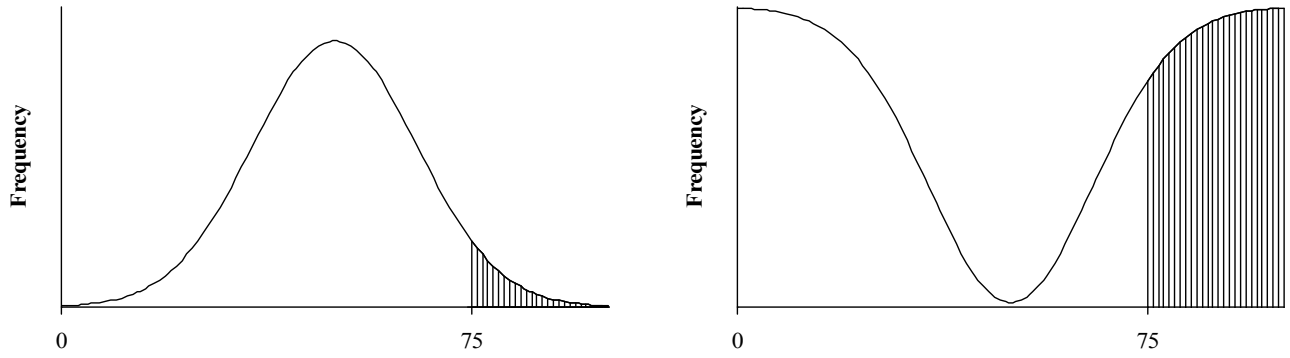
Frequency Distributions

Frequency distributions show us how often different values, or ranges of values, occur. Frequency distributions may be presented to us in the form of a table or a graph (see Zar for examples).

Frequency distributions are important because they allow us to determine probability.

Consider this number line: Suppose we were to ask the question: What is the probability of getting a value of this variable greater than or equal to 75? Note that it is not possible to answer this question. We would need to know how often values of 75 or larger occur. We must have the **frequency distribution.** We need to add a second axis (the y-axis) to our number line. The y-axis will be frequency – it will show us how often values of 75 or greater occur.

Here are two possible frequency distributions for the variable:



In the graph on the left, values of 75 or greater (shaded) occur only about 5% of the time. So the probability of getting a value of 75 or greater is about 0.05.

In the graph on the right, values of 75 or greater (shaded) occur about 1/3 of the time. The probability of getting a value of 75 or greater is about 0.33.

The critical thing to see is that, in order to determine the probability, we need to know what the frequency distribution looks like. The probability depends on the frequency distribution.

Central Tendency

Central tendency refers to the middle of the data set. What value best describes the middle of the data? In Zar, you will find a comprehensive discussion of estimates (statistics) of central tendency.

Arithmetic Mean

The most commonly used estimate in biology is the familiar **arithmetic mean**.

Note: in strict mathematical definitions, the terms **arithmetic mean** and **average** are not synonymous – there are several different types of averages, of which the arithmetic mean is only one. However, in popular usage, the scientific literature, and in this course, when the term **average** is used, it refers to the **arithmetic mean**. Similarly, the term **mean** will refer to the **arithmetic mean**. The mean should only be calculated on Interval-scale and Ratio-scale data. It is inappropriate to calculate the mean of ordinal-scale data (ranks).

Calculate the mean just as you learned in grade school. Add the values together and divide by how many values there are. For a sample of size n values of variable X, the formula is $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$. The symbol for the mean of X is called "X Bar".

Weighted Mean

You need to know how to calculate a **weighted mean**. This is the situation where each value of X has a numerical value (a **weight**) associated with it. A common situation where this occurs in Biometrics class is when we want to combine (pool) several samples together, and find the mean of all the data (the "grand mean"). For example, suppose you have the following three samples:

	n	Mean
Sample 1:	100	30
Sample 2:	12	15
Sample 3:	10	14

What is the mean of all the data (the "grand mean"), i.e. the mean of all 122 data points?

What you may **not** do here is simply calculate the "mean of the means",
i.e. $\frac{30+15+14}{3} = \frac{59}{3} = 19.7$. The grand mean is **NOT** 19.7.

The grand mean is much higher than 19.7. The means have to be weighted by their sample size. The mean of 30 must be "weighted" much more than the other means.
To calculate the weighted mean, multiply each value by its weight, sum the "weighted" values, and divide by the sum of the weights. Let w_i represent the weight of X_i. The formula is:

$$\overline{X} = \frac{\sum w_i X_i}{\sum w_i}.$$

For our example, the weights (w) are the sample sizes (n). So, to calculate the grand mean, we calculate the weighted mean of the means:

$$\overline{X} = \frac{\sum w_i X_i}{\sum w_i} = \frac{100\times 30 + 12\times 15 + 10\times 14}{100+12+10} = \frac{3000+180+140}{122} = \frac{3320}{122} = 27.2$$

The grand mean of our data is 27.2.

Median
A commonly encountered measure of central tendency is the **median**, or middle value. There are just as many data points above the median as there are below the median.

To determine the median, write the data points in order from smallest to largest and find the middle value.

Examples

5
8
10
11
15

If there are an odd number of data points, the median is the data point in the middle. Consider this sample of 5 data points (listed smallest to largest):

The median is 10. Notice there are two data points less than 10, and two data points greater than 10.

If there are an even number of data points, the median is halfway between the two data points in the middle. Consider this sample of 6 data points (listed smallest to largest):

5
8
10
11
15
18

The median is 10.5. Note that the two data points in the middle are 10 and 11, so the median is the number halfway between 10 and 11.

$$\frac{10+11}{2} = 10.5$$

Variability/Variation; Sum of Squares (SS); Variance; Standard Deviation; Standard Error; Coefficient of Variation

(See Zar 4th ed. pages 35-40; 76-78; or Zar 5th ed. Pages 37-42; 72-74)

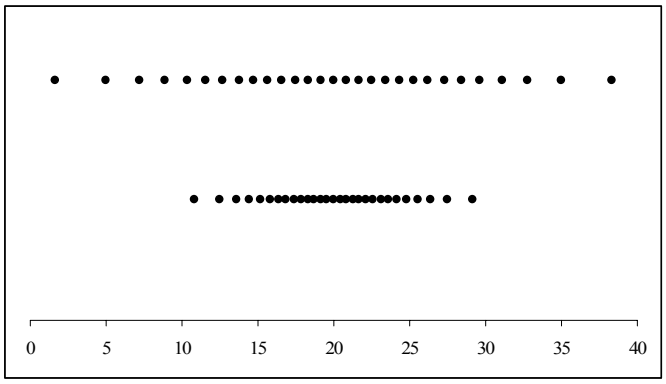
The central tendency of a population and a sample is very important; and most of us are comfortable with the arithmetic mean, simply because we've been exposed to it for so long – since grade school. Equally important, but not as familiar, is the **variability** or **variation** of a population and a sample.

Variability and **Variation**

The terms **variability and variation** simply refer to the property of being different. If we say: "There is variability in the height of Cal Poly students", or "Cal Poly students exhibit variation in height", all we've said is that students are not all the same height. These terms do not refer to any particular quantitative measure or estimate.

We now discuss quantitative estimates. As with the mean, all of these estimates should only be calculated on Ratio/Interval-scale data.

Variability and variation are quantified by describing the dispersion of the data about an estimate of central tendency. The mean is the commonly used central tendency estimate. If the data points are all close to the mean, then variability is low. If data points are dispersed widely about the mean then variability is high.



Examine the graph of two samples on the left. Both samples (each of size n=30) have a mean of 20. Variability is higher in the top sample than in the bottom sample. In the bottom line, the points are clustered more tightly about the mean of 20 than in the top line. That is, there is less dispersion in the bottom line.

Range

The range is the distance between the largest data point and the smallest data point. You simply subtract the smallest from the largest. For example, consider these 6 data points (listed from smallest to largest):

5 8 10 11 15 18

The range is 18 – 5 = 13.

The range is an estimate of variability, but it's not a very good one. It only uses two data points, regardless of how large the sample size. Also, the two data points it does use (the largest and the smallest) are the most likely to be inaccurate.

To quantify better estimates of variability, we want to use dispersion about the mean. The first thing to do is take each data point in the sample, and subtract the mean. $(X_i - \overline{X})$

This quantifies the distance of each point from the mean. Obviously, some points are greater than the mean, and some are less than the mean. When X_i is less than the mean, we will get a negative number. At this point, we're only interested in how far the point is from the mean, not whether it's negative or positive. We want to get rid of the negative signs. We will do this by squaring each deviation: $(X_i - \overline{X})^2$

> Note: why not use the absolute value? Wouldn't the absolute value get rid of the negative signs? Yes, it would have gotten rid of the signs, but you have to be careful with absolute value. It's not really a mathematical operation – you're just deciding to ignore negative signs. Squaring is an actual mathematical operation, and is better in this situation. In some situations, absolute values can be used (we will use them a few times in this course), but you have to be careful with the math.

Sum of Squares (SS)
Once we have calculated the squared deviation of each data point from the mean, the next thing we do is sum the squared deviations: $\sum_{i=1}^{n}(X_i - \overline{X})^2$ The sum of the squared deviations is called the **Sum of Squares** (abbreviated **SS**). SS is our first estimate (statistic) of variability.

Our SS formula yields a statistic. Itt is obtained from a sample of data and is an estimate of the SS for all the data in the population. Notice that our formula uses the sample mean $(\overline{X})$, not the population mean μ (Greek letter mu).

The above formula for SS is called the "**standard formula**".There is an alternative formula for the Sum of Squares, which is called a "**machine formula**". The "machine formula" is more accurate – there are fewer subtractions, and therefore less rounding error when done on a calculator or computer. Here's the "machine formula" for SS:

$$SS = \sum X_i^2 - \frac{(\sum X_i)^2}{n}$$

Be careful with the machine formula:
$\sum X_i^2$ means that you square all the data, then sum the squared values.
$(\sum X_i)^2$ means that you sum all the data, then square the sum.
You should be familiar with both formulas.

> Note: If you are comfortable with Excel formulas, you can use an Excel function to calculate the SS. For example, assume you have entered 10 data point into cells A1 through A10. Go to an empty cell, and enter =DEVSQ(A1:A10). Excel will return the sum of squares. But remember, you won't have Excel available on your exams. You must know how to calculate SS on your calculator!

The Sum of Squares (SS) is a measure of variability, but it is biased by the sample size. You couldn't compare SS values calculated on samples of different size. For example, if one sample was n = 100, the SS would be the sum of 100 squared deviations. If the other sample was n = 10, the SS would be the sum of 10 squared deviations. The n = 100 sample might have a larger SS just because there are more squared deviations being added together.

Variance

To fix the problem of bias, we calculate an unbiased estimate of variability by dividing the SS by n-1, where n is the sample size used. This quantity is called the **Sample Variance**, and is symbolized by $\mathbf{s^2}$. The formula is $s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$.

We always divide by n-1 when calculating the sample variance, never by just n. The quantity n-1 (called a **Degree of Freedom** or **DF**) gives a better estimate of the population variance – but at this point, we won't worry about why it's better. Just always use n-1.

It is important to know that the variance is the SS divided by the DF. In terms of formula:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{SS}{DF}$$

> Note: Excel has a function that will return the sample variance. =VAR(A1:A10) will return the sample variance of the numbers in cells A1 through A10.

Standard Deviation

The sample variance is an excellent estimate of variability. However, interpretation can be difficult, because the sample variance has the square of the original units of the data. For example, if you have data in grams, the variance has the unit "square grams". Who knows what a square gram is?? The solution here is obvious – just take the square root of the sample variance. The square root of the sample variance is called the **Standard Deviation**. Since the sample variance is s^2, the standard deviation is symbolized by s. For completeness, here's the formula:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}} = \sqrt{\frac{SS}{DF}}$$

> Note: The Excel function for standard deviation is =STDEV(A1:A10). This formula returns the standard deviation of the numbers in cells A1 through A10.

Standard Error of the Mean (SE)

An important and frequently encountered measure of variability is the **Standard Error of the Mean (SE)**. A "standard error" is a standard deviation of the values of a statistic. For example, suppose you had 25 different samples of data, and you calculated the mean of each sample. You could calculate the standard deviation of the 25 means. This is the standard error of the mean. Usually we don't have multiple samples, so we need to calculate an estimate of the **Standard**

Error of the Mean (SE) from a single sample of data. To do this, divide the standard deviation of the sample by the square root of the sample size. In terms of a formula:

$$SE = s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

The standard error of the mean will be discussed in lecture.

Coefficient of Variation (CV)

The Coefficient of Variation (CV) is defined as the standard deviation divided by the mean. It is usually multiplied by 100 so that it is expressed as a percentage. The formula is: $CV = 100 \times \frac{s}{\bar{X}}$

CV is a measure of **relative variability** which means that it is not affected by the magnitude of the numbers, or the units. The standard deviation and the mean are both in the original units of the data. When you divide the standard deviation by the mean, the units cancel out.

Below are body weights of 10 hummingbirds (in grams), and 10 house cats (in kilograms). Now, you already knew that cats are bigger than hummingbirds. When you look at the data, keep in mind the units are different. The cat numbers are about 1.6 times larger than the hummingbird numbers. However, if you made the units the same (i.e. changed the hummingbirds to kilograms; or the cats to grams), you'd see the cats have about 1600 times the body mass of hummingbirds.

The one quantity below you can compare without worrying about magnitude of the numbers or the units is the CV. Note that the CV for hummingbirds and cats is about the same, about 6.4% to 6.7%. That is, in both hummingbirds and house cats, the standard deviation of body weight is about 6.6% of the mean.

In **absolute** size, cats have a more variable body size than hummingbirds. But in **relative** size, the relative variability of cats and hummingbirds is similar.

	Hummingbirds (g)	House cats (kg)
	3.1	5.0
	2.9	5.3
	3.1	4.6
	3.3	5.3
	3.3	4.5
	3.4	5.4
	2.8	4.6
	3.1	5.0
	3.3	5.3
	2.9	5.0
Mean	**3.12**	**5.0**
Standard Deviation	**0.20**	**0.33**
Coefficient of Variation (CV)	**6.41%**	**6.67%**

Sample Estimates and Population Parameters - Variability

Remember that the SS, Variance, and Standard Deviation quantities above are all statistics – they are estimates of population parameters. We always use the above formulas with data – because we always work at the sample level, never at the population level. Nevertheless, we do need to be aware of the formulas for the population parameters. One practical consideration is that even cheap scientific calculators have functions for calculating variances and standard deviations. However, they calculate both sample estimates (which we want), and population values (which we don't want). So, we have to know what we're doing!

Here's a table showing sample estimates (statistics) and population parameters:

Quantity	Sample Estimate (Statistic)	Population Parameter
Sum of Squares	$\sum_{i=1}^{n}(X_i - \overline{X})^2$	$\sum_{i=1}^{N}(X_i - \mu)^2$
Variance	$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{SS}{DF}$	$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$
Standard Deviation	$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$
Standard Error of the Mean	$s_{\overline{X}} = \frac{s}{\sqrt{n}}$	$\sigma_{\mu} = \frac{\sigma}{\sqrt{N}}$

Notice that the population quantities are symbolized with Greek letters. μ is the population mean; σ^2 is the population variance; σ is the population standard deviation. Also notice that in the calculation of the population variance, that N (the total number in the population) is used in the denominator, not N-1.

Again, working biologists always calculate the sample estimate quantities – we will use only these in Biometrics class.

As you can see from the table above, the statistics of variability are all related to one another. It is important that you are able to go back and forth among the quantities. Some examples are in the Practice Problems that follow.

Want some practice? Download the **B211 Basic Stats Calculations** workbook. This is a small (less than 50 KB) Excel file that does these calculations. You can calculate and compare your answers to Excel. The workbook is available on Blackboard.

You can also download the workbook at Dr. Moriarty's BIO 211 web site:
http://www.csupomona.edu/~djmoriarty/b211/

Practice Problems (answers follow)

1. Calculate the sum of squares (SS) of the following set of values using the standard formula for the sum of squares.

	2.6
	1.9
	3.7
	2.5
	3.1
	2.8
Sum	16.6
Mean	2.77

2. Calculate the sum of squares (SS) of the set of values in No. 1, using the machine formula.

3. If the sample variance of 12 values is 13.52, what is the standard error of the mean (SE)?

4. If the sample standard deviation of 30 values is 1.79, what is the sum of squares (SS)?

5. Calculate the mean, median, sum of squares, variance, standard deviation, standard error of the mean, and coefficient of variation (CV) for the following 6 values: 16 22 11 14 13 31