

Chapter 3

Descriptive Statistics

Many investigations in biology today are quantitative, meaning they result in observations consisting of numerical facts called data (one fact is a datum). Objective methods are required to analyze and present these research data, and these methods are referred to as **statistics** and more specifically **biostatistics** given that the data are biological in nature. Basically, statistics refers to the organization, summarization, analysis, and interpretation of data. The organization and summarization of data is termed **descriptive statistics** and these typically are displayed in tables and figures (also known as graphs or charts). Analysis and interpretation of data is typically done to make some general conclusions and inferences and these are referred to as **inferential statistics**.

The purpose of most statistical analyses is to draw conclusions about a group of measurements of a variable being studied. The entire collection of measurements about which we wish to draw conclusions is referred to as the **population**. For example, we may wish to draw a conclusion about the oxygen consumption of a particular species of rodent, *Fattus rotundus*. All the *Fattus rotundus* oxygen consumptions therefore are the population under consideration.

In many cases, it is impossible to measure every individual in a population. Therefore, a **sample** or a subset of the population is obtained and measurements made on that sample. If valid conclusions are to be reached about the population by induction from the sample it is important that the sample be obtained in a random fashion. This requires that each member of the population has an equal and independent chance of being selected in the sample. Care and planning are required in order to assure that samples are taken randomly (discussed further in Chapter 4).

Two major types of measures describe and characterize populations in a statistical sense. One describes the general tendency for a majority of the measurements of a population to lie somewhere around the middle of the range. These are known as measures of **central tendency** or "location." The mean or the average is the most useful of these measures with others being the median and the mode. Another characteristic of populations of measurements is how dispersed they are around the average. Measures of the spread of measurements are referred to as measures of **dispersion** with the standard deviation being one of the most useful. To adequately summarize or describe a set of data, estimates of both central tendency and dispersion should be given.

When a measure of central tendency or dispersion is known for a population it is referred to as a **parameter**. However, as indicated above it is frequently impossible to measure all the individuals in a population and measurements are typically made on samples of the population. The sample is thus used to generate an estimate of the population parameter, and this estimate is referred to as a **statistic**. It is conventional to represent population parameters with Greek letters and sample statistics with Latin letters. As example, the mean or average for a population is indicated with the Greek letter mu (μ) while the sample estimate of the mean is indicated with an x with a bar over the top (\bar{x}). Similarly, the parameter standard deviation is indicated as sigma (σ) while the sample standard deviation is indicated with the letter s .

Calculation of Descriptive Statistics

Statistical notation appears complex but is easily understood with a little practice. Some notation is covered here and symbols used throughout this text will be defined.

Typically, an individual measurement taken from a population is referred to as X_i ; the subscript i can be any integer value up through N , the total number of measurements in the population. The population **mean**, a parameter, is denoted by the Greek letter μ (mu), and is calculated as the sum of all the individual measurements, or X_i values, divided by the size of the population, or in mathematical notation

$$\bar{\mu} = \frac{\sum_{i=1}^N X_i}{N} \quad [3.1]$$

The Greek letter Σ (capital sigma) means "summation." The symbols on the bottom of the sigma sign ("i=1") and on the top ("N") indicates that all the X's from i=1 to i=N are to be summed. Placing all these sub- and superscripts are tiresome so frequently they are dropped and the equation for a population mean would be written as simply

$$\bar{\mu} = \frac{\sum X}{N} \quad [3.2]$$

The best estimate of the population mean μ is the sample mean, which is typically denoted as \bar{x} ("x bar"). The size of a sample is denoted as n (lower case) and the calculation of the sample mean can be written formally and correctly as

$$\bar{x} = \frac{\sum_{i=1}^n X_i}{n} \quad [3.3]$$

The **population variance**, a measure of dispersion, is typically denoted as σ^2 (lower case Greek letter sigma). The population variance is calculated as

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \bar{\mu})^2}{N} \quad [3.4]$$

with the numerator term of $(X_i - \bar{\mu})$ referred to as a deviation from the mean. Note that if all the deviations are added together without squaring they will add up to zero. The population **standard deviation** (σ) is simply the square root of formula 3.4, or

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{\mu})^2}{N}} \quad [3.5]$$

The population variance can be referred to as the "mean squared deviation" for it is the mean of the squared deviation of individual values from the population mean. The standard deviation can be referred to as the "mean deviation" since it is the square root of the "mean squared deviation" or variance.

The numerator in formula 3.4, $\Sigma(X - \mu)^2$ in shorthand notation, is called the population **corrected sum of squares** (*SS*) and is an important calculation in many statistics treatments. It is more efficient to calculate the corrected sum of squares using the formula

$$SS = \sum_{i=1}^N X_i^2 - \frac{\left(\sum_{i=1}^N X_i\right)^2}{N} \quad [3.6]$$

Note that ΣX^2 and $(\Sigma X)^2$ in formula 2.6 will be different in value. The uncorrected sum of squares, ΣX^2 , indicates that each X has been squared and then the squared values are summed, while $(\Sigma X)^2$ indicates that the X values are summed first and then squared. Calculating the *SS* in this manner is more efficient when using a calculator to compute variance. Thus the formula for population variance (3.4) can be rewritten as

$$\sigma^2 = \frac{\sum_{i=1}^N X_i^2 - \frac{\left(\sum_{i=1}^N X_i\right)^2}{N}}{N} \quad [3.7]$$

The best, unbiased estimate of the population variance and standard deviation are the sample variance (s^2) and standard deviation (s), which are computed somewhat differently than the population parameters. Since these are sample statistics, n is used to indicate the sample size in place of N . Furthermore N in the denominator of formula 3.7 is replaced with $n-1$ (called the **degree of freedom** or simply *df*). Thus, the **sample variance** is computed as

$$s^2 = \frac{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}{n-1} \quad [3.8]$$

where the numerator, $\Sigma X^2 - [(\Sigma X)^2/n]$ in shortcut notation, is the sample *SS*. Thus, the sample variance formula (3.8) can also be written more simply as

$$s^2 = \frac{SS}{(n-1)} \quad [3.9]$$

The sample variance is simply the sum of squares divided by *df*. Variance is also referred to as a **mean square** or simply *MS*, and *MS*'s are always obtained by dividing *SS*'s by the

appropriate *df*. The **sample standard deviation**, *s*, is simply the square root of the sample variance, *s*², or

$$s = \sqrt{\frac{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}{n-1}} \quad [3.10]$$

Problem Set – Descriptive Statistics

3.1) Compute the mean, sum of squares, sample variance and sample standard deviation of the following sample of body masses (gram, abbreviated g).

10.1 12.4 13.9 11.6 10.9 13.3 12.7 12.9 9.8 10.7.

a) First, use the long-hand method for computing variance by setting up the following columns, and then filling them by making the appropriate calculations.

| <u>Col 1</u> | <u>Col 2</u> | <u>Col 3</u> |
|--------------|---------------|-------------------------------|
| X (g) | X - \bar{x} | (X - \bar{x}) ² |

Use the numerator of equation 2.4 to compute the SS. Sample variance needs to be obtained (divide by n-1).

b) You should be able to do the above by hand and using the excel template descriptive.xls

c) Make a table and error-bar graph for this data with labeled axis and captions. Part “c” will be turned in by the end of lab on Thursday 9/24/09.