

BIOL 458 Biometry

LAB 2 ANALYSIS OF DATA: DESCRIPTIVE STATISTICS

When you analyze data, you are doing so for a reason. Presumably, you want to use the information collected from sample data to infer the properties of some larger set of data - the population. Since the ultimate goal of data examination may be to test a specific hypothesis or to estimate attributes of some larger population of potential subjects, it is important that you become familiar with your data, so that its limitations are known and its usefulness can be fully realized.

It is best to carefully investigate properties of your data before you begin to conduct statistical tests and make statistical inferences. Describing your data is crucial in order to determine which statistical tests can be applied, and how those results must be interpreted. Failing to examine data prior to computing statistics can lead to incorrect conclusions.

The purpose of this exercise is to illustrate the kinds of information one can gain from a preliminary examination of data. You will analyze and characterize data sets, based on measures of central tendency and dispersion, shapes of frequency distributions, and other kinds of plots. This lab will also serve to increase your familiarity and self-confidence concerning the use of **SPSS** and the **SFSU** computing facilities.

1. NUMERICAL METHODS FOR DESCRIBING QUANTITATIVE DATA

1.1. Measures of Central Tendency and Location

Measures of central tendency such as: the mean, median, and mode, give you an idea of the average or 'typical' value of a set of observations. These measures locate the point on a real number line about which your data lies. Note that the mode, median, and mean of a perfectly symmetrical frequency distribution are identical.

The **MEAN** is the most commonly used measure of central tendency and location and is the mathematical name for the arithmetic average of a set of observations. Other values in the data set cluster around the mean. The mean can be viewed as the center of mass of the data and is usually depicted as \bar{x} .

$$\bar{x} = \sum_{i=1}^n x_i / n$$

The **MEDIAN** is the central point of a set of ranked observations. It is the number above and below which 50% of your observations lie. If the number of observations n is even, the median value falls midway between the two central ranks. For an odd number of observations, the median takes on the value of the central rank. (In other words, rank the numbers from lowest to highest; the middle number is the median.)

The **MODE** is the simplest and most readily calculable measure, taking on the 'value' associated with the class containing the highest number of observations. In other words, it is the 'value' that occurs the most. A frequency distribution with a single peak is called **unimodal**, those with two peaks **bimodal**.

1.2. Measures of Dispersion

Variability and dispersion are synonymous terms used to characterize the extent of differences among a set of observations relative to the group mean. The greater the variability among a set of observations, the more they will be spread out around the mean. Populations or samples having high variability will have a frequency distribution involving wider class intervals or more classes than a low-variability group measured on the same scale. Measures of dispersion, such as range, variance and standard deviation, are concerned with the 'spread' or degree of variability in a set of measurements.

The **RANGE** is the simplest measure of variability in the data set, which may be computed by subtracting the smallest observation from the largest. The range is the crudest measure of variability, but it nevertheless can be quite useful. Consider the daily temperature cycle for two locales, both of which experience a mean temperature of 85 degrees F. One cycle might occur in a very pleasant subtropical region where the diurnal temperature ranges from a low of 75 to a high of 95, with a range of 20. The other might be in a far less hospitable desert region, where the diurnal temperature can swing from 50 to 120 - a 70 range. Organisms in the desert environment with the greater temperature range would have to cope with a harsher environment, yet the two environments are characterized by the same mean temperature.

The most important measures of variability (variance and standard deviation) are based on the deviations of individual observations about a central value. For this purpose, the mean usually serves as the center.

The **VARIANCE** is the mean of the squared deviations about the central value. For a population, it is obtained by subtracting each observation from the mean, squaring the resulting differences to eliminate negative quantities, adding up to give the sum of squares, and finally dividing by the number of observations n .

$$\sigma^2 = \sum_{i=1}^n (x - \mu)^2 / n$$

For a sample (drawn from a population), this procedure is the same but the denominator of the expression is $(n - 1)$ instead of n .

$$s^2 = \sum_{i=1}^n (x - \bar{x})^2 / (n - 1)$$

An easier, although algebraically equivalent formula useful for computing the variance is:

$$s^2 = \left[\sum_{i=1}^n x^2 - \left(\left(\sum_{i=1}^n x \right)^2 / n \right) \right] / (n-1)$$

The **STANDARD DEVIATION**, an extremely important index to variability, is calculated simply by taking the square root of the variance.

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

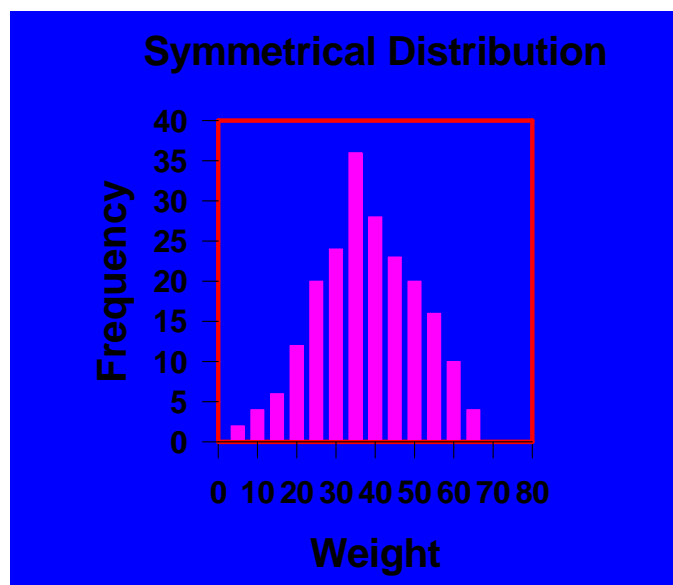
The **STANDARD ERROR** is the standard deviation divided by the square root of the sample size, n .

$$se = s / \sqrt{n}$$

2. GRAPHICAL METHODS FOR DESCRIBING QUANTITATIVE DATA

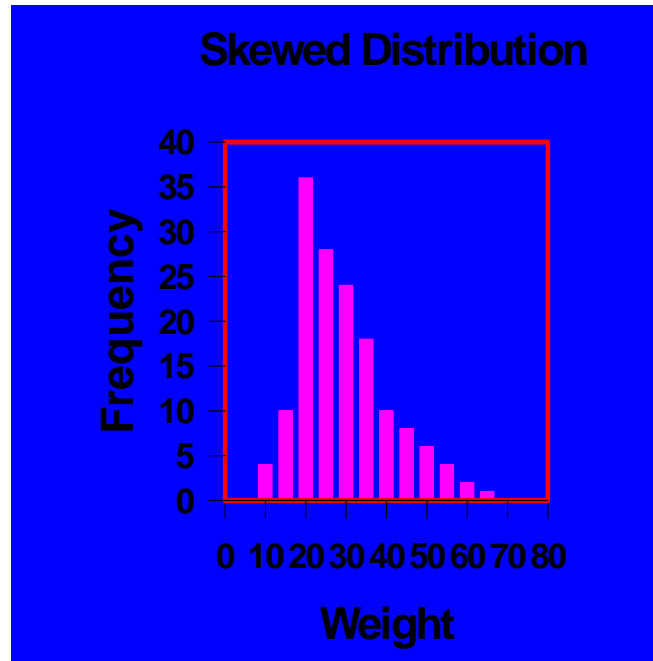
2.1. Histograms

A **HISTOGRAM** is basically a bar graph of a quantitative data set, showing the frequency for each particular class of data, the number of observations falling in each class. Looking at a histogram tells you something about how your data are distributed; it provides a visual summary of the distribution of your data. Two histograms may be compared if the units on the vertical axis are expressed in percentages, because this corrects for any difference in the number of observations.



Additional measures of the overall shape of a distribution can be described in terms of the concepts of **SKEWNESS** and **KURTOSIS**. A skewed frequency distribution is

asymmetric. Extreme positive or right skewness has a long 'tail' extending to the right; while negative or left skewness is less common, showing a long 'tail' to the left. Symmetrical distributions are not skewed. Kurtosis refers to the degree of peakedness of a frequency distribution.



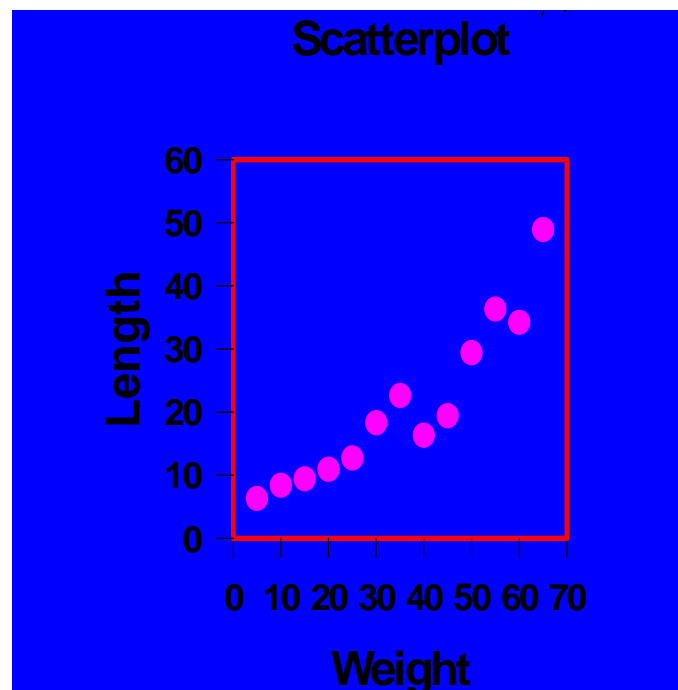
There are many ways in which a data set could be distributed. However, the most frequently encountered in statistics is the **NORMAL DISTRIBUTION**. This is often used partly because many physical measurements, such as human height, diameters of ball bearings, examination scores, and many more types of data closely approximate this type of distribution. But the high incidence of its natural occurrence is not the primary reason why the normal distribution is so important in statistics. For a wide range or parameters the normal distribution can be used to approximate other distributions. Many statistical tests *assume* that the sample data set was drawn from a population whose underlying distribution is normal. Thus before the test results can be deemed valid (or before you even apply the test in the first place), you must make certain that the data approximates a normal distribution.

If your variable is not normally distributed (and thus you cannot satisfy the assumptions for a statistical test you want to apply to the data), you may want to attempt a transformation of your data to try to get it to more closely approximate normality. In this class we will be concerned with the notion of a transformation. It is important to note that in Biology a transformation is not unreasonable, in fact many times it is justified. We do not transform data just to find a significant result where we did not find one before. Data may be transformed for a variety of reasons, but the primary reason is to change the scale of the observations to one in which the values more closely follow a normal distribution. When examining relationships between two variables, we often wish to apply tests that assume that the variables are linearly related. Transformation can also

be used to change the scale of one or both variables to make a non-linear relationship linear. See pages 273-281 in *Biostatistical Analysis by Zar* for further discussion of these transformations.

2.2. Scatterplots

Histograms provide much information about the distribution of a data set, but a single histogram can only describe a single variable. Scatter-plots are used to depict the relationship between variables. For example, the bivariate scatter-plot of Y versus X shows the relations between the dependent variable, Y and the independent variable, X .



LAB 2 ASSIGNMENT

Part 1: Questions

- 1) Given the following data set, find the mean, median, mode, and standard deviation. [5,29,45,46,51,53,55,55,63,72,98]
- 2) Give an example of a dataset that would be characterized by a) a negative (left) skewed distribution and b) a positive (right) skewed distribution.
- 3) Which measure of central tendency would you use to describe your data set, if you were describing it in a journal article, when visual inspection of your data set revealed a skewed distribution? Hint: Think about the values to which each measure is sensitive.

4) The following data set contains the original SAT math scores for five Biology applications: [592,755,600,734,584]. What is the sample range? What does it indicate?

Part 2: SPSS Exercise

1. Raw data can be quickly examined in **SPSS** using some of the techniques that we learned last week. The data for this week's lab are in the file "[DATA1-3](#)". This file contains three columns; each column is simply a set of data taken from a former student's thesis data. Using **SPSS**, describe each data set. Specifically, calculate summary statistics including mean, median, mode, range, and standard deviation. Make a histogram of the data, and interpret. What does this tell you about each variable? Write a paragraph describing each data set.

Hints for discussion: Is the data set highly variable? Is it skewed? Does it approximate normality? ETC.

Notes on the data files: 999 was used to indicate missing values.

2. Try to remedy any observed deviations from normality using logarithmic and square root transformations. Compare the distribution of the transformed variables to the untransformed variables, and discuss the impact of the transformation.

3. In the file "[INSTAR](#)," you will find data on the growth of the moth, *Noctua pronuba*. Column one is the instar (growth stage) and column two contains the length of the given instar. Plot these two variables against one another (stage – x variate, length – y variate). After examining this plot, take the natural log of length (**Compute**) and then plot the two variables again. Describe the new plot. What can be said about the relationship between instar and body length of this insect?

Further Instructions for the Lab Exercise

For this lab exercise you will potentially need to access several new commands including **Compute** from the **Transform** Menu, **Descriptives**, **Frequencies**, or **Explore**, from the **Descriptive Statistics** submenu of the **Analyze** Menu, and **Scatter** from the **Graph Menu**.

In Part 1 of the Lab exercise, you are asked to compute several measures of central tendency (mean, median, and mode), and a measure of variability (standard deviation). All or some of these quantities can be computed using the **Descriptives**, **Frequencies**, or **Explore** commands from the **Descriptive Statistics** submenu of the **Analyze** Menu. Try out the commands and see which you like best. They are somewhat redundant. After you click enough buttons to get to the appropriate command, click over the variable name to the "Variable" box and then choose any options you think necessary. Options are usually selected by first clicking a "button" to obtain another sub-window,

and you then click “boxes” or “radio buttons” on or off to select the options you want. Any boxes or radio buttons that are already chosen represent the “default” selection programmed into **SPSS**. You may unclick these choices and select other options.

In Part 2 of the lab exercise, you are asked to use both the **Scatter** and the **Compute** commands.

Scatter is in the ***Graph*** Menu. Once you have selected the **Scatter** command, chose Simple Scatter. A new sub-window opens and you can click over the variables into the appropriate boxes (y or x, vertical or horizontal axis) to generate the graph. The scatter plot will be generated after you click “OK.”

To transform a variable, select the **Compute** command from the ***Transform*** Menu. In the sub-window that appears, type a new variable name into the box that says “Target Variable.” In the “Numeric Expression” box, enter the transformation and the name of the variable to be transformed in a normal arithmetic expression. Use parentheses to insure that operations are done in the correct order and to enclose the variable name. For example, the natural log of variable **v1** would be entered as “**ln(v1)**.” You can use the arrow keys to click over variables names and function names into the “Numeric Expression” box, as well as the calculator keys to click over arithmetic operators. However, you can also just click the cursor into the “Numeric Expression” box and simply type these items in. Remember that **+**, **-**, and **/** are addition, subtraction, and division, respectively, but that asterisk ***** is multiplication and exponentiation is a double asterisk ****** (to square a value use ****2**, to cube a value use ****3**). Often used SPSS functions include; log base 10 (**lg10**), natural log, (**ln**), square root (**sqrt**), and arcsine (**arsin**).