STA 580 – Fall 2008 – Dr. Charnigo

Lecture 1

Welcome to STA 580

Statistics. If I had to give a single-sentence definition for the word "statistics", it might read like this: "Statistics is the science of making inferences about populations using data from samples." For example, considering a population of hypertensive patients, we might wonder what the average reduction in systolic blood pressure would be after six months on a certain medication that had just been developed. For various reasons (ethical, logistical, and financial), we could not possibly administer this medication to all hypertensive patients everywhere. But we might be able to set up a clinical trial and administer this medication to a sample of hypertensive patients. The data from this sample would allow us to estimate not only the average reduction in systolic blood pressure but also the variability in systolic blood pressure reduction, the fraction of patients who would experience adverse events while on the medication, and so forth.

Biostatistics. Many people use the word "biostatistics" instead of "statistics" when they want to emphasize that they are making inferences related to biological and medical phenomena. While your textbook is titled "Fundamentals of Biostatistics", many of the methods described therein can be applied to problems in other disciplines (engineering, operations research, economics, psychology...). However, since this course is geared predominantly to students in public health and allied fields, most of our examples will be from biology, medicine, and public health.

This semester you will begin learning to decide what statistical methods may or may not work in a given situation, to implement those methods that may work, and to interpret your findings coherently. Many of you will continue this learning process in other statistics courses such as CPH 630/STA 681, CPH 930, and STA 671. However, even those who do not pursue other statistics courses will benefit from taking STA 580.

Summarizing data. Suppose that we conducted the clinical trial and that I gave you a list of systolic blood pressure reductions for 500 patients receiving the new medication. You might find it difficult to absorb all of that information. On the other hand, if you could somehow summarize the data with a few numbers and a few graphs, you might find it easier to reach a conclusion. Today we will introduce various numerical and graphical techniques for summarizing data. Collectively, these techniques (or, more precisely, the results of their application) are referred to as descriptive statistics. With a suitable probabilistic framework (Lectures 2 and 3), these techniques will provide a foundation for making inferences about populations (Lectures 4 through 14).

Numerical Summaries: Central Tendency

Introduction. Suppose that our sample data consist of numerical values $x_1, x_2, \ldots, x_n$. We may want to identify the center of the sample in the following sense: what numerical values are typical or representative? Three measures of central tendency are the (arithmetic) mean, the median, and the mode. Note that some people, including the author of your textbook, refer to measures of central tendency as measures of location.

Mean. The mean (Definition 2.1) is denoted $\bar{x}$ and defined as

$$\sum_{i=1}^{n} x_i/n = (x_1 + x_2 + \cdots + x_n)/n.$$

The mean is an intuitive measure of central tendency, as we are just taking an average. However, the mean is sensitive to extreme values; since each sample value is directly incorporated into the mean, one or two extreme values may influence the mean to the extent that it is no longer representative. Even so, the mean is the most widely used measure of central tendency because of its simplicity and for a theoretical reason (Lecture 3).

Dependence on measurement units. The mean depends on the units of measurement, but the dependence is straightforward: if $y_i = c_1 x_i + c_2$ for some numbers c_1 and c_2, then (Equation 2.3) $\bar{y} = c_1\bar{x} + c_2$.

Example (mean). Refer to the "Before-After" column of Table 2.13 on page 39. Here, $x_1 = 49$, $x_2 = -10$, and so forth. We have

$$\bar{x} = \sum_{i=1}^{24} x_i/24 = 469/24 = 19.54.$$

Note that the mean is not a value that occurs in the sample, but it is amidst the values that occur. Now suppose that values are reported in g/dL instead of mg/dL, and refer to the g/dL-based values as $y_1, y_2, \ldots, y_n$. We have $y_1 = 0.049$, $y_2 = -0.010$, and so forth. Since $y_i = (1/1000)x_i + 0$, we know that $\bar{y} = (1/1000)19.54 + 0 = 0.01954$ without having to add up all of the y_i.

Median. Another measure of central tendency is the median. If the n sample values are ordered from smallest to largest, the median (Definition 2.2) is the $[n/2+1/2]^{th}$ value when n is odd and is the average of the $[n/2]^{th}$ and $[n/2+1]^{th}$ values when n is even. The median is insensitive to extreme values since it is defined so that (roughly) half of the sample values are smaller and half are larger.

Example (median). First we place the "Before-After" sample values in ascending order:

$-13\ -10\ -8\ 2\ 8\ 8\ 12\ 13\ 13\ 16\ 19\ 19\ 19\ 21\ 23\ 27\ 28\ 31\ 32\ 35\ 36\ 41\ 48\ 49.$

Since $n = 24$ is even, the median is the average of the 12^{th} and 13^{th} ordered values, namely 19.

Mode. A third measure of central tendency is the mode. The mode (Definition 2.3) is the most frequently occurring sample value. Note that the mode is not necessarily unique. Sometimes, each of the sample values occurs only once. In such a case, we may say that each value is a mode or that no mode exists. As a measure of central tendency, the mode is not too useful unless the number of possible sample values is rather limited.

Reading SAS output. Refer to page 1 of {TABLE213.pdf}. Find the box of "Basic Statistical Measures" and look for the "Location" columns. Reported here are the mean, median, and mode for the "Before-After" data. Using statistical software to calculate measures of central tendency when $n = 24$ may be more trouble than employing a desk calculator, but statistical software is highly advantageous once n is greater than about 50. In addition, there are other kinds of output in {TABLE213.pdf}. I will comment later about those that are of interest to us.

Numerical Summaries: Variability

Introduction. After we have located the center of the sample, we may want to quantify variability within the sample. What is a typical departure from the center of the sample, or, if you prefer, how spread out are the sample values? Five measures of variability or spread are the range, the interquartile range, the variance, the standard deviation, and the coefficient of variation.

Range. The range (Definition 2.5) is the difference between the largest and smallest values in a sample. The range is easy to compute but not too informative, except as a check for gross mistakes in data entry, because it is determined by the most extreme values. Moreover, the range is not directly comparable across samples of different sizes. To understand the latter point, imagine that you have computed the range for the "Before-After" data. If I now give you a 25^{th} sample value, what will happen to the range? If this 25^{th} sample value is greater than 49 (the largest value among the first 24) or less than -13 (the smallest value among the first 24), the range will increase. Otherwise, the range will stay the same. Thus, the range never becomes smaller as you acquire more data, and often it becomes larger. So, for instance, saying that a sample of size 24 with range 62 has "less spread" than a sample of size 100 with range 80 is potentially misleading.

Interquartile range. A better summary of variability is the interquartile range, defined as the difference between the 75^{th} percentile and the 25^{th} percentile. The 75^{th} percentile (or "upper quartile") is a number such that (roughly) 75% of the sample values are smaller, while the 25^{th} percentile (or "lower quartile") is a number such that (roughly) 25% of the sample values are smaller. Hence, the interquartile range is the distance separating numbers that contain the middle 50% of the sample values. One may compute the p^{th} percentile (Definition 2.6) by ordering the sample values from smallest to largest, letting k be the largest integer less than or equal to $(np/100)$, and

- taking the $(k+1)^{th}$ value if $(np/100)$ is not an integer;
- taking the average of the k^{th} and $(k+1)^{th}$ values if $(np/100)$ is an integer.

Note that the median is the 50^{th} percentile.

Example (interquartile range). For the "Before-After" data, the 75^{th} percentile is the average of the 18^{th} and 19^{th} ordered measurements $\{\ (np/100) = 18,\ k = 18\ \}$, or 31.5. The 25^{th} percentile is the average of the 6^{th} and 7^{th} ordered measurements $\{\ (np/100) = 6,\ k = 6\ \}$, or 10. So, the interquartile range is $31.5 - 10 = 21.5$.

Variance. If we use the mean as a measure of central tendency, we may consider measuring spread via the average of the sample values' squared deviations from the mean:

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

The larger the separation between sample values, the greater $\sum_{i=1}^{n} (x_i - \bar{x})^2$ will be. However, there is a good reason (Lecture 4) for using $n-1$ instead of n in the denominator. Hence, the variance is defined as (Definition 2.7)

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

and denoted s^2. A convenient shortcut formula for computation with a desk calculator is

$$s^2 = \frac{\Sigma_{i=1}^{n} x_i^2 - (\Sigma_{i=1}^{n} x_i)^2/n}{n-1}.$$

Standard deviation. The standard deviation (Definition 2.8), denoted s, is the square root of the variance, $\sqrt{s^2}$.

Dependence on measurement units. If $y_i = c_1 x_i + c_2$ for each i, then

$$s_y^2 = c_1^2 s_x^2.$$

Above, the subscripts y and x identify which values the variances are based on. Note that c_2 is irrelevant; shifting the data changes the center of the sample but not how spread out the values are. The factor of c_1^2 in the above formula reflects one reason that the standard deviation is more intuitive than the variance: the standard deviation is expressed in the same system of units as the sample values (e.g., mg/dL), whereas the variance is not (e.g., mg^2/dL^2).

Example (variance). For the "Before-After" data, $\Sigma_{i=1}^{24} x_i = 469$ and $\Sigma_{i=1}^{24} x_i^2 = 49^2 + (-10)^2 + \cdots + 12^2 = 15661$. So,

$$s_x^2 = \frac{15661 - 469^2/24}{23} = 282.4$$

and, with the y_i as defined earlier,

$$s_y^2 = (1/1000)^2\ 282.4 = 0.0002824.$$

Moreover, $s_x = \sqrt{282.4} = 16.81$ and $s_y = \sqrt{0.0002824} = 0.01681$.

Coefficient of variation. We can also express the standard deviation as a percentage of the mean, assuming that the latter is positive. We do this if we want to convey how much variation is in the sample values relative to the magnitudes of the sample values. For instance, a standard deviation of 10 may be regarded as fairly small if a typical sample value is 1000, but a standard deviation of 10 may be regarded as quite large if a typical sample value is 20. The coefficient of variation is defined as (Definition 2.9)

$$\frac{100\ s}{\bar{x}}\%$$

and denoted CV.

Non-dependence on measurement units. If $y_i = c_1 x_i$ for each i, with c_1 positive, then

$$CV_y = \frac{100\ s_y}{\bar{y}}\% = \frac{100\ c_1 s_x}{c_1 \bar{x}}\% = \frac{100\ s_x}{\bar{x}}\% = CV_x,$$

so that the coefficient of variation is unaltered. Hence, rescalings (such as conversions from mg/dL to g/dL) have no effect.

Reading SAS output. Refer to page 1 of {TABLE213.pdf}. Find the box of "Basic Statistical Measures" and look for the "Variability" columns. Reported here are the standard deviation, variance, range, and interquartile range. The coefficient of variation is reported in the "Moments" box at the top of page 1 as 86.00%, so that the standard deviation is 86.00% as large as the mean. Other quantities that we have encountered today are also present, along with some quantities that we are not concerned about today.

Graphical Summaries

Introduction. In addition to using the measures of central tendency and variability described above, we can summarize data (less concisely) via a frequency distribution or a grouped frequency distribution. These are tables in which we display how often each sample value occurs (Cf. Table 2.8) or how many sample values fall inside each of several intervals (Cf. Table 2.10). However, a graphical representation may make such information easier to absorb. Three common graphical techniques for summarizing data are constructing a histogram, creating a stem-and-leaf display, and producing a box plot.

Histogram. We have all seen histograms. A histogram for the "Before-After" data is displayed on page 3 of {TABLE213.pdf}. Pretending that the first twelve subjects in Table 2.13 were female and that the last twelve were male, I also prepared separate histograms for females (page 7) and males (page 10). If the intervals into which the sample values fall are not of equal sizes, we may call the picture a "bar graph" instead of a "histogram". In this case, it is conventional to keep the bars separated, as illustrated in Figure 2.1. Histograms (or bar graphs) may be prepared by hand using the procedure described in section 2.8, but using statistical software is preferable.

Stem-and-leaf display. A stem-and-leaf display is a rotated histogram in which the bars are replaced by numbers identifying the sample values. Figure 2.8 provides an illustration; also see page 2 of {TABLE213.pdf}. A procedure for preparing a stem-and-leaf display is described in section 2.8.

Box plot. A box plot prominently displays the 25^{th}, 50^{th}, and 75^{th} percentiles in addition to identifying outlying sample values with special symbols. Refer to Figure 2.8; also see page 4 of {TABLE213.pdf}. When placed side-by-side, box plots are useful for comparing groups. Refer to Figures 2.9 and 2.10; also see page 12 of {TABLE213.pdf}. A procedure for preparing a box plot is given in section 2.8.

Shape of a distribution. Graphical summaries help us to characterize the shape of a distribution of sample values (and to speculate about the shape of the corresponding population distribution). A distribution of sample values is often classified in one of three ways.

• Symmetric: The pattern to the left of the median is similar to the pattern to the right of the median.

• Right-skewed or positively-skewed: The pattern to the left of the median is more compressed than the pattern to the right of the median; in particular, there are a few extremely large values.

• Left-skewed or negatively-skewed: The pattern to the left of the median is less compressed than the pattern to the right of the median; in particular, there are a few extremely small values.

There is often special interest in whether a distribution of sample values is symmetric and bell-shaped in the sense that the bars of a histogram would form a pattern resembling a bell (Lecture 3).