SESSION 5

Descriptive Statistics

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

Descriptive statistics are typically distinguished from inferential statistics. With descriptive statistics you are simply describing what is or what the data shows. With inferential statistics, you are trying to reach conclusions that extend beyond the immediate data alone. For instance, we use inferential statistics to try to infer from the sample data what the population might think. Or, we use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study. Thus, we use inferential statistics to make inferences from our data to more general conditions; we use descriptive statistics simply to describe what's going on in our data.

Descriptive Statistics are used to present quantitative descriptions in a manageable form. In a research study we may have lots of measures. Or we may measure a large number of people on any measure. Descriptive statistics help us to simply large amounts of data in a sensible way. Each descriptive statistic reduces lots of data into a simpler summary. For instance, consider a simple number used to summarize how well a batter is performing in baseball, the batting average. This single number is simply the number of hits divided by the number of times at bat (reported to three significant digits). A batter who is hitting .333 is getting a hit one time in every three at bats. One batting .250 is hitting one time in four. The single number describes a large number of discrete events. Or, consider the scourge of many students, the Grade Point Average (GPA). This single number describes the general performance of a student across a potentially wide range of course experiences.

Every time you try to describe a large set of observations with a single indicator you run the risk of distorting the original data or losing important detail. The batting average doesn't tell you whether the batter is hitting home runs or singles. It doesn't tell whether she's been in a slump or on a streak. The GPA doesn't tell you whether the student was in difficult courses or easy ones, or whether they were courses in their major field or in other disciplines. Even given these limitations, descriptive statistics provide a powerful summary that may enable comparisons across people or other units.

Univariate Analysis

Univariate analysis involves the examination across cases of one variable at a time. There are three major characteristics of a single variable that we tend to look at:

- the distribution
- the central tendency

- the dispersion

In most situations, we would describe all three of these characteristics for each of the variables in our study.

The Distribution. The distribution is a summary of the frequency of individual values or ranges of values for a variable. The simplest distribution would list every value of a variable and the number of persons who had each value. For instance, a typical way to describe the distribution of college students is by year in college, listing the number or percent of students at each of the four years. Or, we describe gender by listing the number or percent of males and females. In these cases, the variable has few enough values that we can list each one and summarize how many sample cases had the value. But what do we do for a variable like income or GPA? With these variables there can be a large number of possible values, with relatively few people having each one. In this case, we group the raw scores into categories according to ranges of values. For instance, we might look at GPA according to the letter grade ranges. Or, we might group income into four or five ranges of income values.

Category	Percent
Under 35	9%
36-45	21
46-55	45
56-65	19
66+	6

Table 1. Frequency distribution table.

One of the most common ways to describe a single variable is with a ***frequency distribution***. Depending on the particular variable, all of the data values may be represented, or you may group the values into categories first (e.g., with age, price, or temperature variables, it would usually not be sensible to determine the frequencies for each value. Rather, the value are grouped into ranges and the frequencies determined.). Frequency distributions can be depicted in two ways, as a table or as a graph. Table 1 shows an age frequency distribution with five categories of age ranges defined. The same frequency distribution can be depicted in a graph as shown in Figure 2. This type of graph is often referred to as a histogram or bar chart.

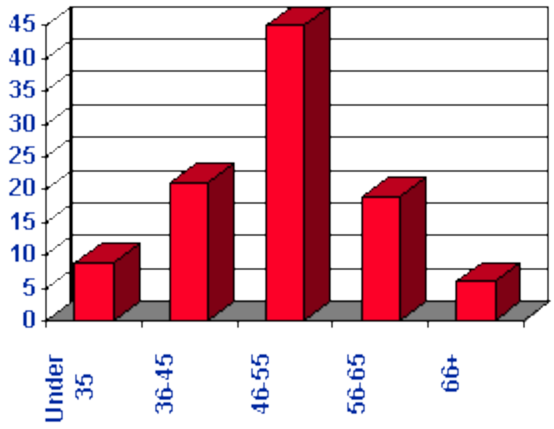

Table 2. Frequency distribution bar chart.

Distributions may also be displayed using percentages. For example, you could use percentages to describe the:

- percentage of people in different income levels
- percentage of people in different age ranges
- percentage of people in different ranges of standardized test scores

Central Tendency. The central tendency of a distribution is an estimate of the "center" of a distribution of values. There are three major types of estimates of central tendency:

- Mean
- Median
- Mode

The **Mean** or average is probably the most commonly used method of describing central tendency. To compute the mean all you do is add up all the values and divide by the number of values. For example, the mean or average quiz score is determined by summing all the scores and dividing by the number of students taking the exam. For example, consider the test score values:

15, 20, 21, 20, 36, 15, 25, 15

The sum of these 8 values is 167, so the mean is 167/8 = 20.875.

The **Median** is the score found at the exact middle of the set of values. One way to compute the median is to list all scores in numerical order, and then locate the score in the center of the sample. For example, if there are 500 scores in the list, score #250 would be the median. If we order the 8 scores shown above, we would get:

15,15,15,20,20,21,25,36

There are 8 scores and score #4 and #5 represent the halfway point. Since both of these scores are 20, the median is 20. If the two middle scores had different values, you would have to interpolate to determine the median.

The **mode** is the most frequently occurring value in the set of scores. To determine the mode, you might again order the scores as shown above, and then count each one. The most frequently occurring value is the mode. In our example, the value 15 occurs three times and is the model. In some distributions there is more than one modal value. For instance, in a bimodal distribution there are two values that occur most frequently.

Notice that for the same set of 8 scores we got three different values -- 20.875, 20, and 15 -- for the mean, median and mode respectively. If the distribution is truly normal (i.e., bell-shaped), the mean, median and mode are all equal to each other.

Dispersion. Dispersion refers to the spread of the values around the central tendency. There are two common measures of dispersion, the range and the standard deviation. The **range** is simply the highest value minus the lowest value. In our example distribution, the high value is 36 and the low is 15, so the range is 36 - 15 = 21.

The **Standard Deviation** is a more accurate and detailed estimate of dispersion because an outlier can greatly exaggerate the range (as was true in this example where the single outlier value of 36 stands apart from the rest of the values. The Standard Deviation shows the relation that set of scores has to the mean of the sample. Again lets take the set of scores:

15,20,21,20,36,15,25,15

to compute the standard deviation, we first find the distance between each value and the mean. We know from above that the mean is 20.875. So, the differences from the mean are:

15 - 20.875 = -5.875
20 - 20.875 = -0.875
21 - 20.875 = +0.125
20 - 20.875 = -0.875
36 - 20.875 = 15.125
15 - 20.875 = -5.875
25 - 20.875 = +4.125
15 - 20.875 = -5.875

Notice that values that are below the mean have negative discrepancies and values above it have positive ones. Next, we square each discrepancy:

-5.875 * -5.875 = 34.515625
-0.875 * -0.875 = 0.765625
+0.125 * +0.125 = 0.015625
-0.875 * -0.875 = 0.765625
15.125 * 15.125 = 228.765625

-5.875 * -5.875 = 34.515625
+4.125 * +4.125 = 17.015625
-5.875 * -5.875 = 34.515625

Now, we take these "squares" and sum them to get the Sum of Squares (SS) value. Here, the sum is 350.875. Next, we divide this sum by the number of scores minus 1. Here, the result is 350.875 / 7 = 50.125. This value is known as the **variance**. To get the standard deviation, we take the square root of the variance (remember that we squared the deviations earlier). This would be SQRT(50.125) = 7.079901129253.

Although this computation may seem convoluted, it's actually quite simple. To see this, consider the formula for the standard deviation:

$$\sqrt{\frac{\Sigma(X - \bar{X})^2}{(n - 1)}}$$

where:

X = each score
$\bar{X}$ = the mean or average
n = the number of values
Σ means we sum across the values

In the top part of the ratio, the numerator, we see that each score has the the mean subtracted from it, the difference is squared, and the squares are summed. In the bottom part, we take the number of scores minus 1. The ratio is the variance and the square root is the standard deviation. In English, we can describe the standard deviation as:

the square root of the sum of the squared deviations from the mean divided by the number of scores minus one

Although we can calculate these univariate statistics by hand, it gets quite tedious when you have more than a few values and variables. Every statistics program is capable of calculating them easily for you. For instance, I put the eight scores into SPSS and got the following table as a result:

N	8
Mean	20.8750

Median	20.0000
Mode	15.00
Std. Deviation	7.0799
Variance	50.1250
Range	21.00

which confirms the calculations I did by hand above.

The standard deviation allows us to reach some conclusions about specific scores in our distribution. Assuming that the distribution of scores is normal or bell-shaped (or close to it!), the following conclusions can be reached:

- approximately 68% of the scores in the sample fall within one standard deviation of the mean
- approximately 95% of the scores in the sample fall within two standard deviations of the mean
- approximately 99% of the scores in the sample fall within three standard deviations of the mean

For instance, since the mean in our example is 20.875 and the standard deviation is 7.0799, we can from the above statement estimate that approximately 95% of the scores will fall in the range of 20.875-(2*7.0799) to 20.875+(2*7.0799) or between 6.7152 and 35.0348. This kind of information is a critical stepping stone to enabling us to compare the performance of an individual on one variable with their performance on another, even when the variables are measured on entirely different scales

Learning Objectives

1. Compute mean
2. Compute median
3. Compute mode

In the previous section we saw that there are several ways to define central tendency. This section defines the three most common measures of central tendency: the mean, the median, and the mode. The relationships among these measures of central tendency and the definitions given in the previous section will probably not be obvious to you. Rather than just tell you these relationships, we will allow you to discover them in the simulations in the sections that follow.

This section gives only the basic definitions of the mean, median and mode. A further discussion of the relative merits and proper applications of these statistics is presented in a later section.

ARITHMETIC MEAN

The arithmetic mean is the most common measure of central tendency. It is simply the sum of the numbers divided by the number of numbers. The symbol "μ" is used for the mean of a population. The symbol "M" is used for the mean of a sample. The formula for μ is shown below:

$$\mu = \Sigma X / N$$

where ΣX is the sum of all the numbers in the population and
N is the number of numbers in the population.

The formula for M is essentially identical:

$$M = \Sigma X / N$$

where ΣX is the sum of all the numbers in the sample and
N is the number of numbers in the sample.

As an example, the mean of the numbers 1, 2, 3, 6, 8 is 20/5 = 4 regardless of whether the numbers constitute the entire population or just a sample from the population.

Table 1 shows the number of touchdown (TD) passes thrown by each of the 31 teams in the National Football League in the 2000 season. The mean number of touchdown passes thrown is 20.4516 as shown below.

$$\mu = \Sigma X / N$$
$$= 634/31$$
$$= 20.4516$$

Table 1. Number of touchdown passes.

37	33	33	32	29	28	28	23	22	22	22	21	21	21	20	20	19	19	18	18	18	18	16	15	14	14	14	12	12	9	6

Although the arithmetic mean is not the only "mean" (there is also a geometric mean), it is by far the most commonly used. Therefore, if the term "mean" is used without specifying whether it is the arithmetic mean,

the geometric mean, or some other mean, it is assumed to refer to the arithmetic mean.

MEDIAN

The *median* is also a frequently used measure of central tendency. The median is the midpoint of a distribution: the same number of scores is above the median as below it. For the data in Table 1, there are 31 scores. The 16th highest score (which equals 20) is the median because there are 15 scores below the 16th score and 15 scores above the 16th score. The median can also be thought of as the 50th *percentile*.

COMPUTATION OF THE MEDIAN

When there is an odd number of numbers, the median is simply the middle number. For example, the median of 2, 4, and 7 is 4. When there is an even number of numbers, the median is the mean of the two middle numbers. Thus, the median of the numbers 2, 4, 7, 12 is (4+7)/2 = 5.5. When there are numbers with the same values, then the formula for the third definition of the 50th percentile should be used.

MODE

The mode is the most frequently occurring value. For the data in Table 1, the mode is 18 since more teams (4) had 18 touchdown passes than any other number of touchdown passes. With continuous data such as response time measured to many decimals, the frequency of each value is one since no two scores will be exactly the same (see discussion of continuous variables). Therefore the mode of continuous data is normally computed from a grouped frequency distribution. Table 2 shows a grouped frequency distribution for the target response time data. Since the interval with the highest frequency is 600-700, the mode is the middle of that interval (650).

Table 2. Grouped frequency distribution.

Range	Frequency
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

Definition:

A cartel is a formal agreement among firms in an oligopolistic industry. Cartel members may agree on such matters as prices, total industry output, market shares, allocation of customers, allocation of territories, bid-rigging, establishment of common sales agencies, and the division of profits or combination of these.

Context:

Cartel in this broad sense is synonymous with "explicit" forms of collusion. Cartels are formed for the mutual benefit of member firms. The theory of "cooperative" oligopoly provides the basis for analyzing the formation and the economic effects of cartels. Generally speaking, cartels or cartel behaviour attempts to emulate that of monopoly by restricting industry output, raising or fixing prices in order to earn higher profits.

Dispersion (Measures of):

Measures of dispersion express quantitatively the degree of variation or dispersion of values in a population or in a sample . Along with measures of central tendency , measures of dispersion are widely used in practice as descriptive statistics . Some measures of dispersion are the standard deviation , the average deviation , the range , the interquartile range .

For example, the dispersion in the sample of 5 values (98,99,100,101,102) is smaller than the dispersion in the sample (80,90,100,110,120), although both samples have the same central location - "100", as measured by, say, the mean or the median . Most measures of dispersion would be 10 times greater for the second sample than for the first one (although the values themselves may be different for different measures of dispersion).

It is important from a practical standpoint that measures of dispersion are normally constructed to be shift invariant and scale invariant . If a measure is not scale invariant, for example, then the value of dispersion might depend on the units of measurement. For example, say the value of dispersion of prices of a particular CD-player model across a country is $10. If the measure of dispersion is scale-invariant and you convert all the prices from dollars to cents by multiplying them by 100, then the measure of dispersion will change from 10 (dollars) to 1000 (cents).