

# Statistics Handout

## I. DESCRIPTIVE STATISTICS

Suppose that we are curious about the SAT scores of incoming freshmen at Shoreline. After obtaining permission to access students' records, we go to the Registrar's Office and for each student that we select to be in our study (i.e., our *sample* of students), we record her or his total (Verbal + Math) SAT score. Now we want to examine our data. This will be no problem if we have selected 10 students to be in our study; we can look at their SAT scores and quickly get an idea of how well these students have performed. But, what if our sample includes 100 students? Or 1,000? Not only will it take us an extremely long time to go through all these scores, but it will also be difficult to get a meaningful or accurate "grasp" of all these data just by looking at them.

A similar example of information overload occurs in most psychological research. For example, suppose we want to investigate whether drinking alcohol affects how quickly people react to different stimuli (e.g., a flash of light, a sound, the sight of a pedestrian crossing in front you while driving a car simulator). We select 20 participants for our experiment: 10 will be assigned to drink alcohol (to reach a .10 blood alcohol level) and the other 10 will represent the control group, which drinks water. Each participant will perform a variety of reaction time tasks and also will perform each one several times. Even though there are only 20 participants in our experiment, we will still have a large set of data because we are recording multiple responses for each participant.

Fortunately, there are certain statistics that can help us condense large amounts of data into a few numbers that are easier to comprehend. These statistics are called **descriptive statistics** and they are used to convey summary information about any set of numbers or data. By using descriptive statistics we can get a quick estimate of what our data (e.g., students' SAT scores; people's reaction times) look like without having to examine every single number or "data point" that we have collected. (Statistics are our friends!) The rest of this handout will cover several of the most common descriptive statistics.

## II. MEASURES OF CENTRAL TENDENCY

Measures of central tendency are statistics that identify the center of a distribution of scores. The most common measures of central tendency are the mode, the median, and the mean. To make calculations easier, instead of discussing SAT scores or reaction times (which are recorded to a thousandth of a second), you can think of the following data as representing the number of times that a small, randomly selected sample of U.W. faculty members goes off on a boring tangent or irrelevant personal anecdote while lecturing to their classes.

A. The mode

The **mode** is a statistic that identifies *the most frequently occurring score in a distribution*. For example, in the following distribution, the mode is 6:

$$4 \ 6 \ 7 \ 8 \ 6 \ 3 \ 5 \ 9 \ 6 \quad \text{Mode} = 6 \text{ (it occurs 3 times)}$$

B. The median

The **median** is a statistic that identifies *the middle score in a distribution*. For example, in the distribution above, the median is also 6. To determine this, you must first rearrange the numbers and order them from lowest to highest:

$$3 \ 4 \ 5 \ 6 \ 6 \ 6 \ 7 \ 8 \ 9 \quad \text{Median} = 6$$

In a distribution with an *odd number* of scores, such as the one above, the median is simply the middle score. In this case it is 6 (a number which has 4 scores to the left of it and 4 scores to the right). In a distribution with an even number of scores, the median is found by taking the average of the two middle scores. For example, in the following distribution, the median is 5.5:

$$2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \quad \text{Median} = \frac{6 + 5}{2} = 5.5$$

C. The mean

The mean, more commonly known as the “average,” is the most common measure of central tendency in statistics. It is determined by dividing the sum of the scores ( $\Sigma N$ ) in a distribution by the total number of scores (N) in that distribution. For example, in the distribution below, the mean is 6:

$$4 \ 6 \ 7 \ 8 \ 6 \ 3 \ 5 \ 9 \ 6 \quad \text{Mean} = \frac{\Sigma N}{N} = \frac{4+6+7+8+6+3+5+9+6}{9} = 6$$

In the above distribution, the mode, median, and mean were all the same value (6). The number six thus represents the “center” of that distribution no matter how we measure it. However, the mode, median, and mean are rarely the same for a given distribution. Here are some examples for you to work out on your own. You can look up the answers on the last page of this handout.

		<u>Mode</u>	<u>Median</u>	<u>Mean</u>
Data Set #1:	4 6 7 2 7 9 3 7	_____	_____	_____
Data Set #2:	3 8 5 7 4 4 7 4	_____	_____	_____
Data Set #3:	8 9 7 8 1 5 6 8	_____	_____	_____

### III. MEASURES OF DISPERSION

A. Variance

Measures of central tendency provide useful information, but they do not always accurately represent the entire distribution. In addition to a measure of central tendency, it is usually

helpful to know something about the extent to which the scores in a given distribution differ (or deviate) from the mean. Another way to state this issue is to ask the question, "How much variation is there in a given set of scores?" Obviously, if all the scores in a set of data are the same (e.g., if every professor goes off on 6 boring tangents/irrelevant anecdotes during lecture) then there is no variation. This rarely happens in research (or in real life), however; people vary in their characteristics, behaviors, attitudes, and emotions. If everyone was the same, then the science of Psychology could proceed by merely studying one person. How dull. Variety, as they say, is the spice of life.

The concept of "variation" in a set of data is illustrated easily by the following example. Given your tremendous popularity you receive invitations to 3 parties that all fall on the same night. All you are told is that the mean age of the 8 guests at each party will be 22 years old. Given this information, which party do you wish to attend? What other information might be helpful in making your decision?

The ages of those attending the 3 different parties are as listed in the following table:

Person	Party 1	Party 2	Party 3
#1	20	34	2
#2	23	9	3
#3	22	10	2
#4	24	41	3
#5	22	38	4
#6	24	7	3
#7	20	31	78
#8	21	6	81
Mean:	22	22	22

Party 1 is going to be a night of who knows what with some young adults. Party 2 is going to be a dinner party with parents and their children. Party 3 is going to be Little Billy's third birthday party (hosted by his Grandma and Grandpa Jones). Therefore, although the mean age for each party is the same, these three data sets are very different from one another. Determining the mode and median for each party would give us a more complete picture, but it would be very helpful to have some statistic that tells us how much the ages of the guests at each party deviated or "varied" from the mean. Other than merely forming a subjective impression of our data, how do we determine this "variability" in scores?

The thought that immediately comes to mind is: "Let's simply add up how much each score differs from the mean." To do this, we could take the mean, which is 22, and then subtract the mean from each individual score. So, for example, for Person 1 at Party 1 we would have  $20 - 22 = -2$ . For Person 2 and Party 1 we would have  $23 - 22 = 1$ .

*At this point, it will facilitate your understanding of this concept if you do the following. For Party 1, continue with these calculations and subtract the mean from each of the 8 scores. Then total up these "deviation scores," being sure to pay attention to the plus and minus signs. After you get your total, then perform the same calculations for Party 2 and Party 3. Move on to the paragraph below after you have finished these calculations.*

Now you see the problem. For any set of data, if we simply subtract the mean from each individual score and add these deviation scores up, we will get a total of zero. Therefore, the average deviation will also come out to be zero (i.e., the sum of deviation scores divided by the number of scores, or  $0/8 = 0$  in our example.) Thus, we cannot determine the "average deviation" in this way. Astoundingly, to the great benefit of humankind, there are two statistics that provide very meaningful information about variability: the **variance** and the **standard deviation**. These are called **measures of dispersion** because they quantify the degree to which a set of scores, overall, differs from the mean.

The word "variance" -- because it sounds foreign, mystical, or technical -- sometimes produces a fear/anxiety response in students; some students sweat profusely, while others experience increased heart rate, nausea, or light-headedness. In extreme cases some students have heard voices telling them "Drop this course! Become an English major!" In case you are having any of the above reactions to the sight or sound of the word "variance," then do the following two things: 1) remember that Statistics are our friends, and 2) think of the word "variance" as a synonym for the word "variation."

Variance is defined as the *average of the squared deviations about the mean*. This is represented mathematically by the following equation, where  $X$  represents a single score within a distribution and  $N$  represents the total number of scores:

$$\text{Variance} = \frac{\sum(X - \text{Mean})^2}{N}$$

The variance is arrived at by performing the following computations in the listed order:

- Step 1. find the mean (sum of the scores divided by the number of scores)
- Step 2. compute the deviation scores (the difference between an individual score and the mean)
- Step 3. square each of the deviation scores
- Step 4. divide the sum of the deviation scores by the number of scores

For example, the variance for Party 1 would be computed as follows:

Person	Age	Mean	Deviation Score (Age-Mean)	Deviation Squared
#1	20	22	-2	4
#2	23	22	1	1
#3	22	22	0	0
#4	24	22	2	4
#5	22	22	0	0
#6	24	22	2	4
#7	20	22	-2	4
#8	21	22	-1	1
Average:	22	22	0	$18/8 = 2.25$

The answer to step 4 above, then, is 18 divided by 8, or 2.25. This is the **variance** of ages at Party 1.

### B. Standard Deviation

The **standard deviation** is the most commonly reported measure of dispersion. It is simply the square root of the variance. In the case of Party 1, the standard deviation is the square root of 2.25, or 1.5. Why take the square root of the variance? Well, remember that when we computed the variance we had to square each deviation score before adding these scores up. (As we saw earlier, if you add the deviation scores without squaring them, then you will always get a total of zero.) Therefore, because we squared the deviation scores to get the variance, we now take the square root of the variance in order to convert our numbers back to their original units of measurement. Thus, for Party 1 the mean is 22 years, the variance is 2.25, and the standard deviation is 1.5 years. *At this point I strongly encouraged you to calculate the variance and standard deviation for Party 2 and Party 3 to check your understanding of these statistics. The correct answers are on the last page of the handout.*

### C. The Range

The range is a relatively crude measure of dispersion. It represents the highest score in a distribution minus the lowest score. It is a crude measure of dispersion because the composition of other scores in the distribution (other than the high and low scores) have no effect on the range. For example, each of the 3 distributions below has a range of 8, despite the fact that the distributions are quite different. Each distribution, however, would have a different value for the variance (and hence, for the standard deviation).

10 7 6 5 4 3 2	Range = $10 - 2 = 8$ .
10 10 10 9 9 9 2	Range = $10 - 2 = 8$
10 4 4 3 2 2 2	Range = $10 - 2 = 8$ .

Variance and the standard deviation are more sensitive measures of dispersion, because they are influenced by each particular score in the distribution. Change even one score, and you'll change the values of the variance and standard deviation.

### D. Additional Comments: This Handout Versus The Text

You should be aware of the following differences between the formula used in this handout versus the one used in your textbook. On page 423 of your textbook, a formula is given for calculating the standard deviation. This formula calculates the standard deviation by dividing by  $n-1$ . The examples in this handout divide by  $n$ . Remember that  $n$  corresponds to the number of people attending each party (8). Why the difference and which divisor should you use? The answer depends on what you want to use the standard deviation for. Consider the following two examples.

Example 1. Assume that I wish to know how much variability in age there is for students in this class. To gather this information, I have the 25 students in this class complete an anonymous questionnaire on which they indicate their age. If I am only concerned with the variability of age for this specific group of people (and I am not interested in trying to generalize my findings to other students), then I would divide by  $n$ . In this instance, the standard deviation I am calculating would be considered a *descriptive statistic*, since I only want to describe the variability of this particular set of numbers. (Aside: Computer programs and calculators can both calculate this number for us very quickly so why do I bother to make you calculate it by hand? No, I don't enjoy inflicting pain and anguish on my students (Ok... maybe I do just a little!). The real reason for asking you to become familiar with this

way of calculating the standard deviation is that it is useful in helping you gain a better conceptual understanding of measures of dispersion.)

Example 2. Alternatively, assume that I wish to know the variability in age for all students at Shoreline. In general, what kind of spread is there in age for students attending Shoreline? It would be possible for me to contact every single student at Shoreline, ask how old he or she is, and calculate the standard deviation of this rather large set of numbers. But this would be very time consuming (and besides, I'm also very lazy!). What I could do instead is to determine the variability in age of a small sample of Shoreline students and then use this number to make an "educated guess" about the variability in age of the entire population of Shoreline students. So assume that I randomly sample 25 Shoreline students about their ages. Since I am using my sample to try to infer the variability in age for the whole population, when I calculate the standard deviation in this example, I divide by  $n-1$  instead of  $n$ . In this second case, the standard deviation I am calculating would be considered an *inferential statistic*, since I am using the standard deviation of my sample to make an inference about the overall standard deviation of the entire population of students. **For the purposes of this class, when you are asked to calculate the standard deviation, please divide by n.**

#### IV. Z-SCORES

Imagine that you and a friend are somewhat competitive and you want to compare how you each did on your last psychology exam. Unfortunately, you are taking different psychology classes (you are in Psychology 285/209 while your friend is in Psychology 236). What information would you need to find out who did better? You would probably start by comparing the points you each got correct on your respective exams. Let's say that you got 40 points correct and your friend got 20 points correct. Looks like you win, right? Not necessarily. You also need to know how many total points were on each exam. If your test had 50 possible points, and your friend's test had 25, you are still tied ( $40/50 = 20/25 = 80\%$ ). The next thing you might want to know is what were the respective class means on each of the tests. If I tell you that the mean for your test was 30 (out of 50) while the mean for your friend's test was 15 (out of 25), can you now tell who did better? It seems like you may have done better since you scored 10 points higher than the mean but your friend only scored 5 points higher than the mean. But remember your test had twice as many total points on it as your friend's test (50 vs. 25), so the difference of each of your scores from the means still seems roughly equivalent.

As you can see, making a comparison in this situation is a bit difficult. But thanks to the standard deviation, there is still a way to make a comparison. If you also know the standard deviation of the scores in each of the classes, this can allow you and your friend settle your dispute. If I tell you that the standard deviation for both classes is 10 points, who did better on their test relative to the rest of their class? Fortunately for you, it looks like you are the winner. Your scores is a full standard deviation above the mean ( $(40-30)/10 = 1$ ) but your friend was only a half of a standard deviation above the mean in his or her class ( $(25-20)/10 = 0.5$ ). Thus, using standard deviations, we can see that you outscored a larger percentage of your class than your friend did of his/her class.

What we have just done in this example is to calculate something called a z-score. A z-score counts the difference between an individual score and the mean in terms of standard deviations. For instance, we could say that you scored 10 points above the mean on your test, or equivalently, that you scored 1.0 standard deviation above the mean. This 1.0 represents

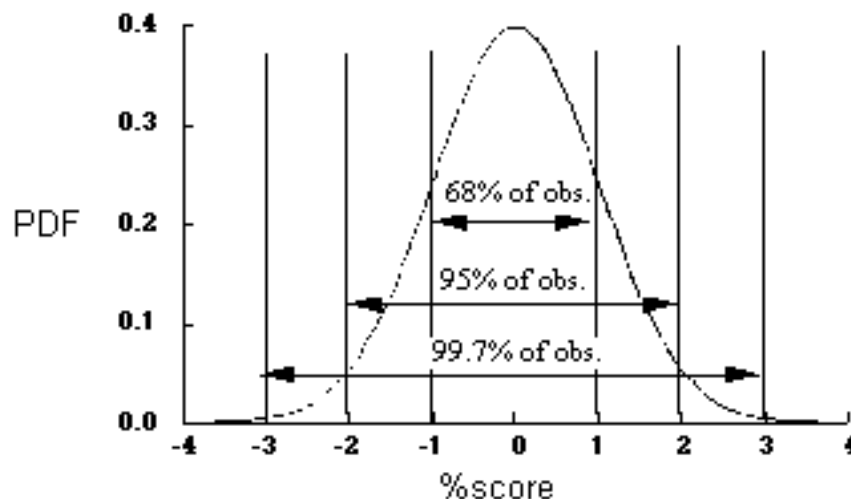
your z-score for this test. Similarly, the z-score for your friend was 0.5 standard deviation above the mean. He or she scored one half of a standard deviation better than the mean. It turns out that z-scores are a convenient way to compare scores from different groups that have used different numerical scales, as was the case in our example above (the scale used on your test was from 0-50 while the scale used in your friend's class was from 0-25). By converting our numbers to z-scores, this allows us to make meaningful comparisons between the numbers in each of these groups, something we couldn't do initially. Z-scores also come in handy when trying to understand our next topic, the normal distribution.

## V. THE NORMAL DISTRIBUTION

One of the most important concepts in statistics is a curve called the "normal distribution" or "normal curve." The normal distribution is an essential part of inferential statistics. While we will not be covering inferential statistics in this course, I would still like to introduce you to the normal curve and some of its unique properties. By learning a little about the normal curve now, you will be ahead of the game when you take your required statistics course.

The normal distribution is a bell shaped curve that is symmetric (that is, on either side of the mean it looks the same). A typical normal distribution is shown in Figure 1. The x-axis represents the all possible values or scores on some characteristic that we are interested in (e.g., SAT scores, reaction times, etc...). The y-axis represents the frequency or percentage of each score on the x-axis occurs. Because it is symmetric, the mean, median, and mode for the normal distribution are all equal to one another. The main reason that the normal curve is useful is because there are many variables out in the real world which distribute themselves normally (i.e. they approximate a normal distribution). For instance, IQ is one variable that is normally distributed; if we were to look at the distribution of IQs for all adults, this distribution would look pretty much like the theoretical normal curve pictured below.

Figure 1.



Another useful feature of the normal distribution concerns the standard deviation. Specifically, in any normally distributed set of numbers, the standard deviation can be used to divide the distribution into segments which contain fixed percentages of scores. For instance, we know that in a normal distribution, 34% of the scores fall between the mean and

one standard deviation above the mean (don't worry about how this percentage is calculated; it is simply a product of the mathematics of the normal distribution). Since the normal curve is symmetric, we also know that 34% of the scores will fall between the mean and one standard deviation below the mean. We can put these two pieces of information together to deduce that 68% of the scores in a normal distribution fall within one (plus or minus) standard deviation of the mean. Similar percentages also exist for segments further away from the mean and these are also shown in Figure 1.

To see why are these percentages useful, consider the following example. Assume that you took an IQ test and received a score of 130. Looking at this score, you assume that you did well but, being the competitive person you are, you want to know precisely how many people did as well or better than you. If we tell you that the average IQ score is 100 and that the standard deviation is 15 (both of which are true), you could determine, using the percentages listed in Figure 1, that only 2.5% of the people scored 130 or higher (Wow! You're pretty smart!). Alternatively, you could say that you scored better than 98.5% of the people. In either case, the percentages generated by the normal distribution allow you to more accurately determine how well you did.

Another reason that the normal curve is useful concerns *inferential statistics*, a topic that we will discuss briefly later in this course. Inferential statistics will be covered in greater detail when you take statistics. I am briefly covering this information now in hopes that it will make it easier to assimilate when you come across it again in the future.

## VI. ANSWERS

For Data Set #1:	mode = 7	median = 6.5	mean = 5.625
For Data Set #2:	mode = 4	median = 4.5	mean = 5.25
For Data Set #3:	mode = 8	median = 7.5	mean = 6.5

(If you have not yet done so, give yourself a chance to test your understanding of measures of dispersion by calculating the variance and standard deviation for the Party 2 and Party 3 data sets on page 3. After you are done, check the answers below.)

For Party 2, the variance is 204.5 and the standard deviation is 14.30.  
 For Party 3, the variance is 1103 and the standard deviation is 33.21.