Dr. Michael Passer
University of Washington

Psychology 209
Online Posting

Statistics Handout

I. DESCRIPTIVE STATISTICS

Suppose that we are curious about the SAT scores of incoming U.W. first-year students. After obtaining permission to access students' records, we go to the Registrar's Office and for each student we record her or his total (Verbal + Math) SAT score. Now we want to examine our data. If the entire first-year class only had 20 students, this would be easy. We would look at all 20 SAT scores and quickly get an idea of how well the first-year class performed. But in reality, each year the class of first-year students entering the U.W. from high school is about 4,000 students. Not only will it take us a long time to go through the 4,000 scores, it will also be difficult to get a meaningful or accurate grasp of all these data just by looking at them. Information overload!!! Even if the entering class had only 1,000 students, or even 100, it would be difficult to get a meaningful overall picture merely by looking at the individual scores.

Such information overload also occurs in most psychological research. For example, suppose we want to investigate whether drinking alcohol influences how quickly people react to different stimuli (e.g., react to a flash of light, a sound, or 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 .08 blood alcohol level) and the other 10 will represent the control group, which drinks water. Each participant will perform the reaction time task 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, called **descriptive statistics,** *summarize information about any set of 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. The rest of this handout will cover several of the most common descriptive statistics.

To make calculations easier, instead of discussing SAT scores or reaction times (which are recorded to a thousandth of a second), let's examine the number of times that a small, randomly selected sample of U.W. faculty members goes off on a boring tangent while lecturing to their classes. (Note that fascinating tangents, such as all those raised by your current Psych 209 instructor, are not included in these data.)

II. MEASURES OF CENTRAL TENDENCY

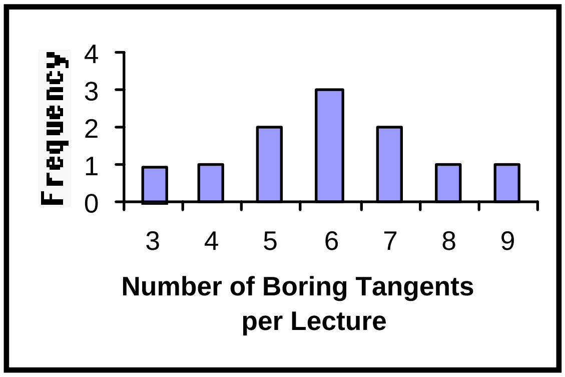
Measures of central tendency *are statistics that identify the center of a "distribution" of scores*. The term "distribution" refers to a set of scores that has been organized according to their values, either from "highest to lowest," or "lowest to highest." Thus, suppose we have 11 professors: Professor #1 goes off on 4 boring tangents during lecture, Professor #2 has 6 tangents, and so on. Here are the number of tangents for each of the 11 professors, respectively: 4 6 7 8 6 3 5 9 6 5 7

To organize these scores into a distribution, we would reorder them as:

3 4 5 5 6 6 6 7 7 8 9.

If we wished to, we could further organize these scores into a **frequency distribution**, *which shows how often each value occurs in a distribution*. For example, in the distribution of scores above, one professor had 3 tangents, one had 4 tangents, two professors each had 5 tangents, 3 professors each had 6 tangents, and so on. Frequency distributions can be constructed as tables or as graphs.

# of Boring Tangents	Frequency
3	1
4	1
5	2
6	3
7	2
8	1
9	1



Now on to measures of central tendency. The most common measures are the mode, the median, and the mean.

A. The mode

The **mode** *is the most frequently occurring score in a distribution*. For example, in the following set of data, the mode is 6:

3 4 5 5 6 6 6 7 7 8 9. Mode = 6 (it occurs 3 times)

B. The median

The **median** *identifies the middle score in a distribution*. For example, in the set of data above, the median is also 6. To determine the median for any set of scores, you must first order them from lowest to highest. Here, we've already done this in the process of creating our distribution.

3 4 5 5 6 6 6 7 7 8 9 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 the number in the sixth position (a number which has 5 scores to the left of it and 5 scores to the right of it). 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 of eight scores, the median is 5.5:

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

C. The mean

The **mean**, also known as the "average," is the most common measure of central tendency in statistics. It is determined, first, by calculating the sum of the scores (∑X) in a distribution, where "X" represents each individual score and the term "∑" (sigma) stands for "the sum of." Second, after adding all the scores, you divide by the total number of scores (N) in that distribution. For example, in the distribution on the next page, the mean is 6:

3 4 5 5 6 6 6 7 7 8 9 Mean = $\frac{\sum X}{N} = \frac{3+4+5+5+6+6+6+7+7+8+9}{11} = 6$

In the distribution for "number of tangents per lecture," the mode, median, and mean are all the same value (6). The number six thus represents the "center" of this distribution no matter how we measure it. However, in real life, 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 page 7 (Section III E) of this handout.

		Mode	Median	Mean
Data Set #1:	3 8 5 7 4 4 7 4	______	______	______
Data Set #2:	1 4 6 7 2 7 9 3 7	______	______	______
Data Set #3:	6 7 8 9 7 8 1 5 6 8	______	______	______

D. Advantages (+) and Disadvantages (—) of the Mode, Median, and Mean

Look at the salaries that employees make at Honest Al's company. You will see that, in this distribution, the mode, median, and mean are quite different.

	Employee	Annual Salary ($)
1.	Honest Al	205,000
2.	Al's mother	205,000
3.	Johnson	20,000
4.	Rodriguez	19,500
5.	Jones	18,500
6.	Chen	18,000
7.	Brown	17,500
8.	Carter	17,000
9.	Mullins	16,500
10.	Watson	16,000

What are the Mode, Median, and Mean?

Mode:

+ In a distribution, if your goal is to guess any individual score, the mode is your best bet (i.e., you will be correct most often).

— May be extremely unrepresentative of the overall distribution. ("Come work for Honest Al; the modal salary is $205,000." Such a deal!)

Median

+ Not influenced by extreme scores

— May fail to capture important information (If Johnson and Rodriguez get a huge salary raise and now make $80,000 each, the median remains the same and does not reflect this raise.)

Mean

+ Uses all the information in a distribution; every score value is taken into account. (The mean increases if Johnson and Rodriguez get a salary raise.)

— Influenced by extreme scores ("Come work for Honest Al; the average salary is $55,300." Yeah, right!)

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 during lecture) then there is no variation. But, when it comes to the human (and other) species, variation is the rule: we come in different sexes, heights, weights, nose shapes, and eye, hair, and skin colors. People also vary in their backgrounds, behaviors, personalities, 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
Sum	**176**	**176**	**176**
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 at Party 1 we would have 23 - 22 = 1.
At this point, to facilitate your understanding of this concept, do the following--it will only take a couple of minutes. 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 meaningful information about variability: the *variance* and the *standard deviation*. These statistics 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 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 simply 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:

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

To determine the variance of a set of scores, first create a Table with 5 columns, like the one shown below. Column 1 lists each person at the party (or each research participant in a study). Column 2 contains each person's score (in this case, each person's *age*). The mean of these scores goes in Column 3, and Column 4 and Column 5 will be used to calculate a "deviation score" and a "deviation squared" score for each person.

Once you lay out the table, follow these four simple steps.

Step 1. find the mean of Column 2 (add the scores, divide 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. add the squared deviation scores and divide 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
Sum	**176**	**176**	**0**	**18**
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 most common measure of dispersion is the **standard deviation** (**SD**), *which is defined as the square root of the variance.* Conceptually, you can think of the standard deviation roughly in this way: *in a distribution of scores, the standard deviation tells you "the typical distance" of those scores from the mean.* In other words, the standard deviation represents the "typical" or "standard" amount that scores in a distribution deviate from the mean.

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 encourage 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 2 2 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: "Descriptive" versus "Inferential" Statistics -- This Handout Versus The Text and Your Statistics Class

You should be aware that there is an **important difference** between the formulae in this handout, and the way you may calculate variance and the standard deviation when performing statistical tests of data in your statistics class. In this handout, after totaling the squared deviation scores, you **divided by N,** the number of people attending the party (8). When analyzing the data from a research study (e.g., do participants in the "drink alcohol condition" of an experiment have slower reaction times than participants in the "drink water condition"), the denominator in the formulae for calculating the variance and standard deviation is N - 1; thus it would be 7 in our example.

Without getting bogged down in a long explanation, you divide by N in order to get the variance and standard deviation for a specific group or set of scores (e.g., for the people attending this particular party). "N - 1" is the denominator in statistical tests where the goal of the statistical analysis is to use the variance and standard deviation of a sample (e.g., a sample of people) to estimate certain characteristics of the broader population to which the researcher wants to generalize the results. These statistical tests, *which take data from a sample and use them to estimate characteristics of a larger population, are called* **inferential statistics.** For example, we would use the reaction time scores from our participants in the experiment to make an estimate or *inference* about how alcohol influences the reaction time of college-aged people in general. **At present, for our purposes, you should divide by N.**

E. Answers for Sample Data Sets on Page 4

For Data Set #1:	mode = 4	median = 4.5	mean = 5.25
For Data Set #2:	mode = 7	median = 6	mean = 5
For Data Set #3:	mode = 8	median = 7	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, read the paragraph 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.

REVIEW QUESTIONS: CHECK YOUR UNDERSTANDING

1. What are "descriptive statistics" used for?
2. Conceptually, what is a "measure of central tendency?" Name three measures of central tendency.
3. What is a frequency distribution?
4. Conceptually, what is a "measure of dispersion?" Name three measures of dispersion.
5. What are inferential statistics used for?
6. Explain, conceptually, what the statistic measures;
   - mode
   - mean
   - variance
   - median
   - range
   - standard deviation

7. Use the PRACTICE ITEMS on the next page to check that you can correctly calculate each statistic from a simple set of data. **You will be asked to make these types of calculations on Exam 1.**

Calculating Basic Statistics: Practice Exam 1 Items

Ten students participated in a study on emotions. They wore electronic beepers and were randomly beeped four times a day for one week. At each beep, students wrote down the emotion they were experiencing. Listed below are the number of times each of the ten students reported being happy. Calculate the following statistics BY HAND. NO CALCULATORS. (If an answer requires a square root, just put the square root sign around the number.)

4, 2, 3, 8, 6, 2, 2, 6, 0, 7

A) Draw a frequency histogram of these data. (2 points for correct data placement)
. Be sure to label your axes properly. (2 points for appropriate labels))

B) What is the Mode? (1 point) _________

C) What is the Median? (1 point)________

D) What is the Mean? (2 points)_________ (SHOW YOUR WORK)

E) What is the Range? (1 Point) _________

F) What is the Variance? (2 Points) ________ (SHOW YOUR WORK)

G) What is the Standard Deviation (1 Point)

Answers to Practice Quiz

DON'T LOOK AT THESE ANSWERS UNTIL YOU HAVE WORKED THROUGH THE PRACTICE ITEMS.

Question A. The x-axis of your histogram should be labeled something similar to "Number of Times Students Reported Feeling Happy" or "Number of Times Happiness Experienced in One Week." If you just write down Happiness, you would lose some credit. "Happiness" could mean many things--it's not specific enough. For example, someone reading your histogram might think that "happiness" refers to the intensity (rather than frequency) of people's happiness. The equidistant data labels on the x-axis should read 0, 1, 2, 3, 4, 5, 6, 7, 8.

The y-axis should be labeled "Frequency" or "Number of Students," and the equidistant data labels along the y-axis should be 0, 1, 2, 3. You should have horizontal bars extending up from the x-axis corresponding to the frequency of each number in the data set (e.g., the bar representing number "2" along the x-axis should extend up to the level of number three on the y-axis because the number "2" occurs three times in the data set. You would have no bar (indicating a frequency of 0) for number 1 or for number 5 on the x-axis because no students reported this frequency of feeling happy.

A graph of this frequency histogram will be put on the course website by Friday of Week 3

Questions B through G. Here are the various statistics. The mode is 2.0, the median is 3.5, and the mean is 4.0. The range is 8.0, the variance is 6.2, and for the standard deviation you can just put 6.2 inside a square root sign (the standard deviation comes out to 2.49).