Central Tendency

When we have a set of data we might want to characterize the "center" of that data. There are three different ways to do this. We have three different names for the "center" of the data, namely, the mean, the median, and the mode. Each is computed in a different way. Each is appropriate in different situations. We will look at them one at a time.

The Mean: This is the value that you have been using as the average of a set of data. To compute the mean you add up all of the values and divide that total by the number of values. So, for the "list" of values [4, 8, 6, 13, 8, 2, 8], we would find the mean by adding all the values 4+8+6+13+8+2+8 to get 49, and then we divide that total by the number of values, in this case we had seven values so the mean is 49/7 which is 7.

There are a few things to note about the mean. First, as we saw in our example, the mean does not have to be one of the values in our data. Second, the mean is highly influenced by extreme values. If we were to add just one more value to our data and that value happened to be extremely different from the other values, say it is 233, then the total of all the values would be 282, we would have 8 values, and the mean would now be 282/8 or 35.25, a huge change from the original 7 and a value that has just one element of the new list that is greater than the new mean. Third, the mean is appropriate for cases where we are taking measurements based on a reliable and valid standard. Thus, finding the mean value of the noon temperatures in degrees Celsius over a period of a month is appropriate because we can use a standardized thermometer

to measure the temperature at noon (determined by a standardized clock). On the other hand, finding the mean value of the social security numbers of the people in the room, or finding the mean value of the credit card numbers of the people in a class is a misuse of the mean. We could do such a calculation but it would not tell us anything since those values are assigned as identifiers, as "names", not as measurements. One small touch of reality is required at this point. Just because something should not be done does not mean that it isn't done. For example, opinion surveys are often coded as numbers, perhaps as 1=disagree, 2=slightly disagree, 3=neutral, 4=slightly agree, and 5=agree. Then, if we have data from many such opinion surveys we could compute the mean value of the response to a question. There are two truths to this. First, it is totally inappropriate to do this because we have no reliable and valid standard for measuring an individual's opinion. Second, this computation of the mean of an opinion survey is done all the time. Computing the mean of opinion questionnaire questions is an example of popular usage outweighing reason.

The Median: The median is the midpoint of the data. That is, we need to sort the data and find the value that is in the middle of that sorted list. For the "list" of values [4, 8, 6, 13, 8, 2, 8], we would find the median by sorting the list to get [2, 4, 6, 8, 8, 8, 13] and then we would find the middle value, in this case that would be the fourth value which happens to be 8. Note that the median element has three elements that are less than or equal to it and three that are greater than or equal to it. In fact, since there are 7 items in the original list the middle item, once the list is sorted, will be in position (7+1)/2, that is 8/2, or

in position 4. For a moment, consider the same list with the element 13 removed. Now the list is [4, 8, 6, 8, 2, 8], and sorted it is [2, 4, 6, 8, 8, 8]. Our task is to find the middle value. We see that we have a problem doing this. There are six items in our list. Because the list has an even number of items there is no middle item. The "rule" to resolve this is that we take the median to be the average of the two middle items. In this case that would be the third and fourth items, namely 6 and 8. The average of 6 and 8 is (6+8)/2=7. We note that in this example the median value turned out to be a value that is not an element of the list. When we had an odd number of items in the list we always had a middle element. In fact, if the number of items in the list is n and if n is odd, then once the list is sorted the middle value will be in position (n+1)/2. If the number of items in the list is n and n is even, then once the list is sorted the median value will be the average of the item in position n/2 and the item in position (n/2)+1.

There are a few things to note about the median. First, as we just saw, for a list that has an even number of elements the median need not be a value in the list. Second, the median is not much affected by an extreme value. If we return to the original list, [2, 4, 6, 8, 8, 8, 13], and add an eighth item that is an extreme value, say 233, then the list becomes [2, 4, 6, 8, 8, 8, 13,233], a list with an even number of items so the median of the new list is the average of the items in position 8/2=4 and (8/2)+1=5. That is the new median will be the average of 8 and 8, which is 8. In this case, adding that extreme value did not change the median at all. Third, the funny rule about taking the average of the middle two values for a list with an even number of items is more of a trick problem on math tests than it is an

issue in real life. On a math test we will give students a small list and the odds are pretty good that the two middle values will not be equal. In the real world we generally deal with much larger lists and it is usually the case that the middle two values are just somewhere in a pile of identical values. For example, in a recent term there were 12,608 credit students at the college with ages ranging from 13 to 84. If we sort the list of ages then the median age would be the average of the item in position 12608/2=6304 and the item in position 6305. However, in our sorted list the age 23 is represented 683 times, starting in position 6037 and ending in position 6719. Our two calculated positions, 6304 and 6305, are deep inside the group of 23's and the average of 23 and 23 is just 23. Fourth, the median value only makes sense when we have a natural way to order, that is sort, the values. It makes sense to talk about the median age of students in a given term. It would not make sense to talk about a median phone number or a median credit card number.

The Mode: The mode is the element in the list that appears most frequently. In our original list, [4, 8, 6, 13, 8, 2, 8], the item 8 appears more often than any other item. Therefore, the mode of the list is 8. In the example of the age of credit students registered at the college in a particular term, it turns out that of the 12,608 students with an age on the system, there were 1192 students at age 18, 1333 students at age 19, 1130 students at age 20, and fewer students at all other ages. Thus, the mode age of students that term was 19. [We might recall that the median age was 23. Were we to add up all the

ages we would find that there was a total of 34928 years, which if we divide by 12608 would give us a mean age of 27.11992.]

There are a few things to note about the mode. First, the mode is not affected by extreme values. Second, although we can compute the mode for small data lists it really isn't very important. Adding just a few more items, and not even extreme items, to a small list can dramatically change the mode value. Third, in a relatively small list it is not at all unexpected to have multiple items be tied as the most frequent. In that case we give all of the values that appear that most frequent number of times as the values of the mode. Thus, in the list [2,3,4,3,6,3,7,6,1,4,4,3,7] the mode values are 3, 4, and 7. If a list has two such most frequent values then we say it is bi-modal. This is a particularly bad choice of terms since we also use that term, and we do so more often, when we are talking about a situation where we see that there are two peaks, not necessarily of the same height, in the frequency of different values. Let us look at a table of values of the ages of students in a hypothetical college:

Age of Student	Frequency of that age
17	250
18	375
19	765
20	542
21	345
22	256
23	412

24	573
25	674
26	459
27	231
28	212
29	134

The mode age is clearly 19. However, we note that the frequency drops from there to a low point at age 22. Then it rises again to a "local" high of 674 students at age 25. Then the frequency drops down again. This is the classic "bi-modal" distribution even though, overall, there is but one mode here, namely, 19. Fourth, the mode is most appropriate for values that are the result of counting. Consider the problem of finding the mode height of students at the college during a term. If we measure the students to the nearest foot then the mode is certainly going to be 5 feet and that will tell us nothing. If we were to measure the students to the nearest 0.00001"(one one-hundred thousandth of an inch) then we may not even have two students at the same height and every height would be a modal value with frequency 1. And, even if we had two students at the same measured height, having a modal value with frequency 2 when all other values have frequency 1 still tells us nothing. If we measure the student heights to the nearest inch we get a better feeling for the mode value but we have lost information that we would have had if we measure with more accuracy. In short, to use the mode with measurements may mean that we have to give up some accuracy. That does not mean that we cannot do it but we do need to be aware of the issue.