CHAPTER 2

CENTRAL TENDENCY

Once a set of data is organized, the next thing of general interest will normally be what is the most typical or "average" score in the distribution. When you have relatively large data sets, it is not usually possible to list out, for the consumer, each and every one of those values. However, being able to quickly summarize what the data set looks like in terms of a typical score is important and is the topic of central tendency. This section presents several different ways to indicate "average" and shows some relationships amongst these.

Mode

In Chapter 1, two sets of data: (1) for the "HardTest" and (2) for the "EasyTest", were presented. The tally output for the frequency distributions is again given below.

HardTest	COUNT	EasyTest	COUNT	EasyTest	Count
1	2	14	1	27	3
2	5	17	1	28	5
3	4	18	1	29	2
4	13	19	1	30	5
5	9	20	7	31	2
6	9	21	3	33	1
7	3	22	3	N=	50
8	2	23	6		
9	1	24	3		
10	2	25	3		
N=	50	26	3		

The simplest indicator of "average" would be the score that occurred most often or had the highest frequency. This is called the **MODE**. For the "HardTest" data, that answer is clearly 4. When you have a single peak point in a distribution (uni-modal), the mode is easy to find. However, for the "EasyTest" distribution, the modal pattern is not so clear. Technically, a score of 20 has the highest frequency and could be called the mode. However, the score of 23 has 6 frequencies and the score of 30 has 5 frequencies. Thus, there are two other scores, somewhat removed from 20, that have frequencies almost equal to the frequency for 20. Calling 20 the modal value somewhat ignores the fact that there are two other points in the distribution that have "comparable peaks". Therefore, in distributions where there is not one clear peak point, the modal picture becomes more ambiguous. This is one of the reasons why the mode is not the best measure of central tendency in that some distributions may have more than one peak point, rather than just one modal average value. But, the mode is a start. However, it is especially important to

report the modal values if there are clearly two or more. For example, the last Chapter showed a U shaped distribution which was clearly bi-modal. It would be important in that situation to state that the distribution was bimodal because the consumer would be mislead if we only presented one average value.

Median

Another way central tendency could be indicated would be to find the middle score value; ie, the one where 50 percent of N is above that point and 50 percent of N is below that point. This is the concept of the **MEDIAN** and was briefly mentioned when boxplots were discussed. The tally output above would assist us in finding this median value. Since both distributions have N=50, we want to count up to the point between the 25th and 26th frequency and see what score value is at that point. You can start at the bottom and work your way up, or the reverse. For the "HardTest" variable, if we start at the score of 1 and accumulate the frequencies, we would need to go up to the score of 5 before we would have accumulated at least 25 frequencies. For the scores of 1-4, we would only have accumulated 24 so we have to go into the next score interval (where 5's are) to find the middle frequency point. Thus, the easiest thing to do is to say that 5 is the median value. For the "EasyTest" data, we would have to count up to a score of 24 before we are able to accumulate at least 25 frequencies. Since we have only a total of 23 frequencies (not enough) up to a score of 23 and through the score interval of 24 we have 26 frequencies (too many), the middle frequency must be at a score value of 24. Thus, for the "EasyTest" variable, the median is 24.

The way most software packages (including Minitab) find the median is by the same process described above. First, the data are sorted from low to high. Then, if there is an odd number of total frequencies (N=27, 51, etc.), you would go directly to the score at the middle frequency value. If there is an even number of frequencies, like our N=50, then you would find the middle two scores and average them. If those middle two scores are the same, then that whole number is listed as the median. If they are different, say one is 24 and the other is 25, you would call the median (24 + 25)/2 = 24.5. Look at the sorted data below for the "HardTest" and "EasyTest" variables to see how you find the median. Since our N is even and =50, we would count up to the 25th and 26th scores. For the "HardTest" variable, both values are 5's and therefore the median would be listed as 5. For the "EasyTest" data, since both values are 24's, the median would be 24. I have underlined the area in the listing where we would find both medians.

HardSort

1	2	2	2	2	2	2	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4

4	4	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	7	7
7	8	8	9	10	10					

EasySort

14	17	18	19	20	20	20	20	20	20	20
21	21	21	22	22	22	23	23	23	23	23
23	24	24	24	25	25	25	26	26	26	27
27	27	28	28	28	28	28	29	29	30	30
30	30	30	31	31	33					

Mean

The third and most commonly used measure of central tendency, the one that comes to mind when we think of "average", is called the **MEAN**. The mean is the value that is found by adding up the scores and dividing by the number of scores. If you recall from high school, this was the way that was used to figure test score averages in classes. Thus, to find the mean, we need to add up all the score values and divide by the N. Look at the following data for the "HardTest" and "EasyTest" distributions.

HardTest: Sum of X = 241 N = 50

EasyTest Sum of X = 1224 N = 50

For finding the mean for HardTest, you would divide 241 by 50 resulting in a mean of 4.82. For the EasyTest data, you would divide 1224 by 50 to obtain a mean of 24.48. Most software packages, including Minitab would have a command for finding the mean or, would include the calculation of the mean in a command that lists out a variety of descriptive statistics. In Minitab, there is a command called **DESCRIBE** that will do that; see the output below.

Describe Command Output for HardTest and EasyTest Variables

	N	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN
HardTest	50	4.820	5.000	4.727	2.067	0.292
EasyTest	50	24.480	24.000	24.545	4.253	0.602

	MIN	MAX	Q1	Q3
HardTest	1.000	10.000	4.000	6.000
EasyTest	14.000	33.000	21.000	28.000

Not only does the describe command give the values for the mean and median, it also provides additional descriptive information on the data sets such as the low and high

values, and Q1 and Q3 (from the discussion of the boxplots in Chapter 1).

One important property of the mean is that the deviations around the mean add up to 0. Look at the simple example below. The mean of the scores is 15/3 = 5.

X	X - Mean
9	4
4	-1
2	-3
Sum =	0

You can either think of this as being that the sum of the deviations around the mean equals 0 or that the sum of the absolute deviations above the mean equals in total (4), the sum of the absolute deviations below the mean (4). In effect, the mean is a point where the deviations balance themselves out above and below that point.

Before leaving the mean, it would be helpful to mention something about the symbolism involved in the formula for the mean. If you were working with **POPULATION** data (a very rare event!), the formula and symbol for the mean would be as follows.

$$\mu_x = \frac{\sum X}{N}$$

The symbol on the left of the equal sign is called mu (mew) and is a common symbol for the population mean. The symbol that looks like a large E in the numerator is called sigma (upper case Greek letter) and simply means "add up". The cap N in the denominator represents the total number of values in the population. However, if you are using a **SAMPLE** and are attempting to estimate the population mean (a more typical situation), the formula would look as follows. The mean from a sample is commonly called X bar. Note that there is a small n in the denominator and this is the normal symbol for the size of a sample.

$$\bar{X} = \frac{\sum X}{n}$$

As you note, even though both formulas are for the mean, the symbolism is a little different. At the moment, this distinction is not particularly important. However, later in the text, the topic of inferential statistics will be introduced. In that context, distinctions between population values (parameters) and sample values (statistics) will be important.

Distributional Form and Central Tendency

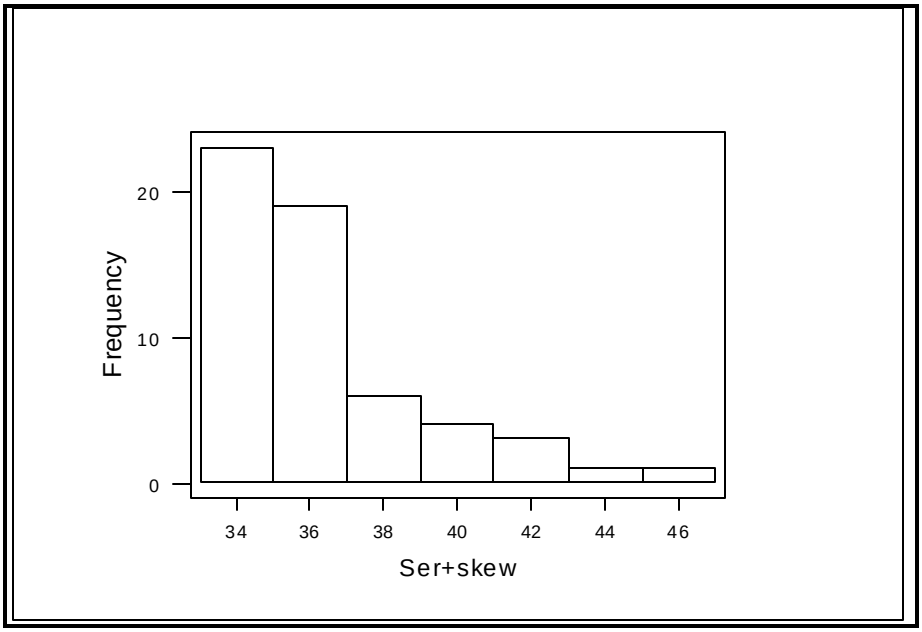
As mentioned in Chapter 1, some distributions are symmetrical and some are not. Of particular interest is the case of skewed or asymmetrical distributions and what impact that has on the values you obtain for the different measures of central tendency. Look at the seriously positively skewed distribution below.

Ser+skew	COUNT
34	23
36	19
38	6
40	4
42	3
44	1
46	1
	N = 57

Describe Command Output for Ser+skew Variable

	N	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN
Ser+skew	57	36.316	36.000	36.000	2.823	0.374

	MIN	MAX	Q1	Q3
Ser+skew	34.000	46.000	34.000	38.000



From the frequency distribution, it is easy to see that the modal value is 34; more frequencies occur at this point. Also, if we sort the data and then find the middle score, the median will be 36, as seen in the describe output. Finally, note that the mean is 36.316. Thus, what we have in a positively skewed distribution is: the mode is the lowest central tendency value, the median is next and in the middle, and the highest of the three is the mean. Note that the mean is pulled somewhat towards the trailing off end of the distribution. This pattern is not uncommon for distributions of data such as salaries or ages, which tend to be positively skewed. In situations like these, the mean tends to yield the highest central tendency value whereas the modal value tends to be the lowest. If we were working with a negatively skewed curve, just the opposite would occur. In negatively skewed distributions, especially seriously skewed ones, the mean would be pulled towards the lower end of the scale with the median being the next highest (again in the middle), and the modal value would be the highest.

A very simple measure of skewness would be a numerical comparison between the mean and the median as shown below.

Skewness = Mean - Median

For our data above, this value would be: 36.316 - 36 = + .316. If you use the order of Mean - Median, a resulting positive value would indicate a positive skew since, as mentioned above, the mean will be larger than the median. If the subtraction would have been negative, there would have been a negative skew to the distribution. Generally, in similarly scaled data sets, the larger the difference between the two, the more serious will be the skew. Chapter 5 presents a different and more sensitive measure of skewness but, for the time being, the one above is satisfactory.

Central Tendency and Constants

Last in our brief discussion of central tendency measures is the idea of manipulating the data set by some constant value. For example, what if you give a midterm test and then weight the scores by 2 since you want the midterm to count twice as much as other quizzes. What effect would multiplying every score by 2 have? Look at the following example.

	X	X + 2	X - 2	X * 2	X/2
	8	10	6	16	4
	6	8	4	12	3
	4	6	2	8	2
Mean	6	8	4	12	3

Note that when you operate on the original set of data (X) with some constant value, the mean will change by the amount of that constant. For example, if you multiply the scores

on a midterm test by 2 (double weight), the mean of the weighted test will be twice that of the mean of the original data set.

Useful Minitab Commands

MEAN MEDIAN DESC

Practice Problems

For each of the following sets of data, find the mode, median and mean. Also, use the skewness measure to see if the disbribution is symmetrical or not. If the distribution is skewed, tell how the 3 measures are different. Finally, what would happen to each mean if we divided each of the score values by 3?

Problem 1

19	17	17	17	16	16	15	20	18	18	20
18	17	19	20	18	18	17	18	18	17	18
19	18	18	18	18	18	20	16	18	16	14
17	19	18	17	19	16	17				

Problem 2

2	1	0	2	1	1	2	1	0	2	2	3	3
3	0	2	1	1	1	3	1	1	2	0	2	2
0	1	1	2									

CHAPTER 3

VARIABILITY

Examining the average value in a distribution is important to get a "feel" for the typical score. However, knowing what the average is does not convey any information about the dispersion or variability of the scores in a distribution. One thing that is very apparent when looking at sets of data such as test scores or heights or weights, is the fact that there is considerable variation in the values. For example, on a 30 item multiple-choice test in a college course, it is not uncommon for the scores to range from nearly 10 points to 30 points. Considering that everyone in the class was exposed more or less to the same material prior to the test, that amount of variation in test performance is interesting. The present section examines several ways to quantify variability.

Range

Look at the two sets of data below, named "LessVar" and "MoreVar" for reasons that will become obvious in a moment.

LessVar

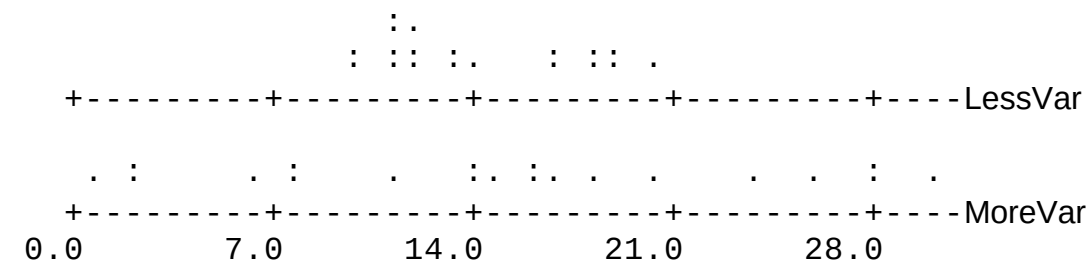
12 10 18 11 13 19 18 11 12 17
17 11 20 11 10 12 11 13 14 19

MoreVar

28 6 8 26 15 24 8 28 18 16
30 1 14 16 11 14 2 2 20 17

Since, in the present form, these are unorganized sets of data, it would be helpful to make a frequency distribution or examine a graphic to see if we can detect what is the level of variability, compared across the two sets of data. I have used a dotplot to show this pattern. Since the dots are more homogeneous or less spread out in the "LessVar" distribution, that is the one with less variability. Now you can see why I gave them the names that I did! On the scale, the "LessVar" distribution ranges from 10 or so to a little less than 21 whereas in the "MoreVar" distribution, the scores range from nearly 0 to 30. This can also be seem more precisely if we look at the tally output which gives us a frequency distribution. Note the the "LessVar" data goes exactly from 10 to 20 and the "MoreVar" distribution spreads from 1 to 30.

.



LessVar	COUNT	MoreVar	COUNT
10	2	1	1
11	5	2	2
12	3	6	1
13	2	8	2
14	1	11	1
17	2	14	2
18	2	15	1
19	2	16	2
20	1	17	1
N=	20	18	1
		20	1
		24	1
		26	1
		28	2
		30	1
		N=	20

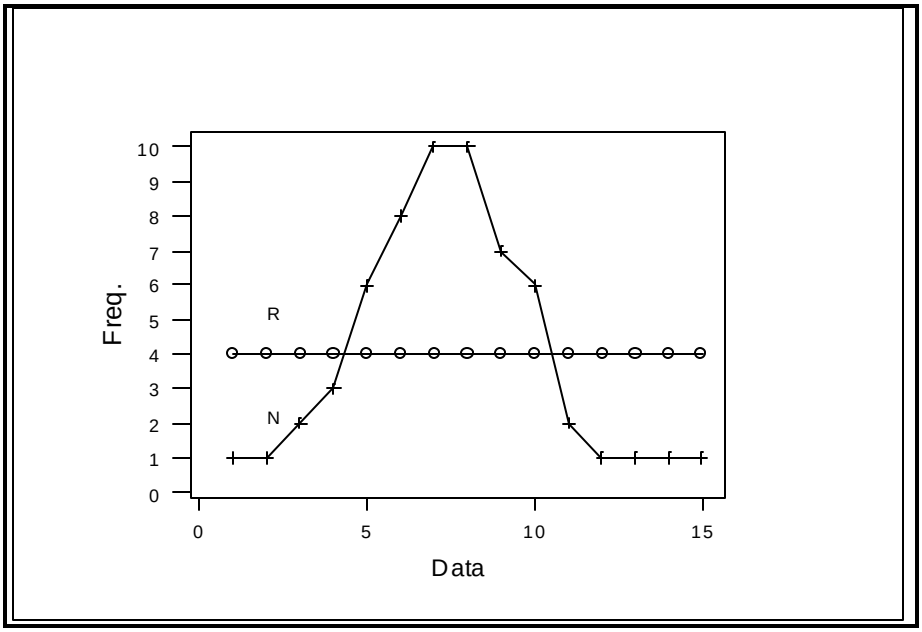
One way to formalize the variability would be to measure the distance from the bottom of the distribution to the top. Such a measure is called the **RANGE**. For the "LessVar" distribution, the range would be 20 - 10 = 10, and for the "MoreVar" distribution, the range would be 30 - 1 = 29.

Technically, the true range goes from the lowest possible point of the lowest score to the highest possible point of the highest score which would be from the lower limit of the lowest score to the upper limit of the highest score. Limits of scores are the dividing points between one number and the next adjacent number. In this context, the true range would be from 20.5 to 9.5 or 11 for the "LessVar" distribution and from 30.5 to .5 or 30 for the "MoreVar" distribution. Most software packages that have a range command do not calculate the true range and, that is fine by me since the range is not the best measure of variabilty to use anyway.

Variance and Standard Deviation

While the range is simple to calculate and easy to interpret, one major problem with

the range is the fact that it depends entirely on the two end points. Because of this, different patterns of frequecies between the two end points are ignored by the range. See the normal (N) and rectangular (R) distributions in the diagram below.



Note that the data in the rectangular distribution (R) are evenly spread out throughout the entire range of scores. However, for the normal distribution (N), while there are some extreme values at the far left and far right, the majority of the frequencies are located towards the middle. Clearly, the normal distribution has a different overall variability pattern than does the rectangular distribution. Graph N tends to be more homogeneous with respect to the middle (more compact) whereas graph R tends to be more heterogeneous with respect to the middle (more dispersed). Unfortunately, the range is not sensitive to this fact.

Because of problems with the range, we need a better measure of variability that would - if calculated - indicate that the normal distribution (N) had less overall variability than did the rectangular distribution (R). Since I alluded to the idea that N was more homogeneous with respect to the middle, perhaps we could develop a measure that quantifies the variations around the middle point. Since the mean is our best general purpose measure of "middleness" or average, why not look at variations around the mean?

The first step in working on a new and better measure of variability would be to calculate the deviation scores around the mean. To do this, we would find the mean, subtract that mean value from each score, and then list out the deviation values. For the LessVar distribution, the mean is 13.95 so, in that distribution, we would subtract 13.95 from each of the values. In the MoreVar distribution, the mean is 15.2 so, in that

distribution, we would subtract 15.2 from each of the values. As we will see in a moment, the development of a better measure of variability will also require the squares of the deviation values. So, to facilitate our work, I have also squared each of the deviation values in each distribution. Look at the columns of calculations below.

LessVar	MoreVar	DevLess	DevMore	SqDevLes	SqDevMor
12	28	-1.95	12.8	3.8025	163.84
10	6	-3.95	-9.2	15.6025	84.64
18	8	4.05	-7.2	16.4025	51.84
11	26	-2.95	10.8	8.7025	116.64
13	15	-0.95	-0.2	0.9025	0.04
19	24	5.05	8.8	25.5025	77.44
18	8	4.05	-7.2	16.4025	51.84
11	28	-2.95	12.8	8.7025	163.84
12	18	-1.95	2.8	3.8025	7.84
17	16	3.05	0.8	9.3025	0.64
17	30	3.05	14.8	9.3025	219.04
11	1	-2.95	-14.2	8.7025	201.64
20	14	6.05	-1.2	36.6025	1.44
11	16	-2.95	0.8	8.7025	0.64
10	11	-3.95	-4.2	15.6025	17.64
12	14	-1.95	-1.2	3.8025	1.44
11	2	-2.95	-13.2	8.7025	174.24
13	2	-0.95	-13.2	0.9025	174.24
14	20	0.05	4.8	0.0025	23.04
19	17	5.05	1.8	25.5025	3.24
	Sums	0	0	226.95	1535.2

To clarify one or two examples from the table, look at the first score in each of the "LessVar" and "MoreVar" columns. A score of 12 in "LessVar" is 12 - 13.95 = -1.95 points away from the mean of "LessVar" and the squared value of that deviation is 3.8025. The negative deviation simply means that 12 is below the mean. Note of course that even though the deviation is negative, the squared deviation is positive. For the first score of 28 in the "MoreVar" distribution, the deviation is 28 - 15.2 = 12.8 and the squared deviation is 163.84. What would the deviation and squared deviation be if in fact the score happened to be at the mean? Correct, 0.

One thing that might cross your mind would be to add up the deviations around the mean thinking that if scores deviate further from the mean, that the **SUM OF THE DEVIATIONS** should also be larger. But of course, recall in Chapter 2 on central tendency, it was pointed out that the sum of the deviaitions around the mean is always 0. This can be expressed in the following formula.

$$\sum (X - \bar{X}) = 0$$

Thus, the idea of using the sum of the deviations around the mean as an indicator of variability is not workable since, regardless of the real variability, the sum of the deviations around the mean will always be 0. However, this problem does not exist if we deal with the **SQUARED DEVIATIONS**. Since squared deviations are always positive, the squared deviation numbers above the mean do not cancel themselves out with the squared deviation numbers below the mean. Therefore, we could explore using the squared deviations as a method of expressing variability. What if we add up the squared deviations for both the LessVar and MoreVar variables? This would be accomplished by the simple addition of the last two columns above and, the results would be as follows.

Sum of SqDevLes = 226.95

Sum of SqDevMor = 1535.20

While each of the data sets above has 20 values, it is possible that some data sets will have different n's. When this happens, the sum of the squared deviations will tend to be larger if there are more squared deviations to add together, and smaller if there are fewer squared deviations to add together. Therefore, a way to eliminate this problem would be to take an average of the squared deviations. In the present case, the average of the squared deviations around the mean of the "LessVar" distribution is 226.95/19 = 11.9447 and for the "MoreVar" distribution, the average of the squared deviations is 1535.20/19 = 80.8. (**NOTE**: It is common practice to divide the sum of the squared deviations by n - 1 rather than n [20-1 = 19 in this case]. More on the reasoning for this later.) In a distribution where the squared deviations tend to be larger on the average, this indicates that the scores in that distribution tend to deviate more from the mean, on the average. This quantity, the **AVERAGE OF THE SQUARED DEVIATIONS AROUND THE MEAN**, is called the **VARIANCE**. Look at the formula for the variance that has just been used.

$$S_x^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

Recall in the formula for the mean, I indicated that if you are talking about the population mean, we would be using the population symbol mu and would be dividing by large N, which is the total number of values in the population. But, if you were using a sample to estimate the population mean, we used the symbol X bar and divided by n or the sample size. First, most software statistical packages assume that you have sample data are are trying to estimate population values. Well, you say OK but why divide by n-1, why not use n? Without going too much into this now, let me say that when you take a sample

from the population, it is not likely that you will obtain in your sample values that extend across the full range of scores in the population. Thus, variability in a sample tends to be smaller or less than the variability in the corresponding population. What this means is that if you calculate a measure of variability on the sample data as though it were the population, you would be underestimating the true variability in the population. Without knowing exactly why at this point, you should be able to see that dividing by n-1 will produce a somewhat larger value than division by n and therefore our variance value is made somewhat larger. Since this is what we expect in the population, the division by n-1 should produce an estimate that is closer to the true population variability. It is helpful to use a formula for calculating the variance that takes into account the fact that population variances tend to be larger than what appears to the eye in sample data. As was shown with the mean, note for reference purposes the population version of the variance formula.

$$\sigma_x^2 = \frac{\sum (X-\mu)^2}{N}$$

The symbol on the left of the equals sign is the lower case Greek letter **sigma**; hence, the population variance would be called sigma squared. Note that the symbol for the population mean is located in the numerator rather than the symbol for the sample mean. In addition, recall from the formula for the mean, the upper case Greek letter sigma was used to indicate the operation of "adding up the values". While the population variance formula is referred to as "sigma squared", please note that I will use the S squared variance notation (sample jargon) and will assume, as do statistical packages, that you are dealing with sample data even though your primary interest is in the corresponding population values (ie, parameters). Since the variance for the "LessVar" distribution is much less (11.9447) than for the "MoreVar" distribution (80.8), we conclude that the first set of data is much more homogeneous (compact around) with respect to the mean than the second set of data.

The variance is a good way to indicate the relative variability in two or more sets of comparable data. For example, if you had weight measurements for 3 different groups of children, then differences in the variances really reflect the fact that some groups have more variable weights around their means than do other groups. However, you must keep in mind that you can only come to conclusions like that if the data sets are similar. For example, to compare the variances on the variables of intelligence scores (IQ's) and grade point averages (GPA's) would not make sense since these two scales are radically different in the first place. With IQ scores naturally ranging from about 50 to 150, and GPA values ranging from 0 to 4, variances of IQ scores must necessarily be larger due to the properties of the measurement scale. To say that people show more variability (using the variance numbers) on intelligence than on grades in college makes no logical sense. Thus, you cannot simply look at the numbers but obviously must keep in mind the nature of the scales of the variables.

Another potential interpretation problem with the variance is the fact that variances are based on "squared" units of measurement.. Although some measurements, like inches, are easily interpretable in squared units, like squared inches, most measures are not easily interpreted in squared units. If we had weight data like pounds, then the interpretation of the variance of pounds would be in terms of "squared pounds". Have you seen one of those lately? I haven't! To avoid this problem, we could attempt to take the variance value and transform it in some way back to the original units of measure. Since the variance is sigma or S squared, all that needs to be done is to take the square root. The square root of the variance will return us back to the original units of measurement.

Square Root of Variance of LessVar = Square Root 11.9447 = 3.45611

Square Root of Variance of MoreVar = Square Root 80.8 = 8.9888

In statistical work, the **SQUARE ROOT OF THE VARIANCE** is called the **STANDARD DEVIATION**. The standard deviation is a measure of variability, based on the variance, that expresses dispersion in terms of the original score scale units. This is the most widely used measure of variability. Since the symbol for the variance based on sample data is S squared, the symbol for the square root or standard deviation will simply be S. Most software packages will have one or more commands that will calculate the variance and/or the standard deviation. In Minitab for example, there is a command to find the standard deviation and a describe command that will print out not only the standard deviation but also a variety of other statistics. Look at the output from the describe command for the LessVar and MoreVar variables.

Describe Output for LessVar and MoreVar Variables

	N	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN
LessVar	20	13.950	12.500	13.833	3.456	0.773
MoreVar	20	15.200	15.500	15.170	8.990	2.010

	MIN	MAX	Q1	Q3
LessVar	10.000	20.000	11.000	17.750
MoreVar	1.000	30.000	8.000	23.000

For most sets of data, it is important to report indicators of both central tendency and variability. Thus, for a set of data, we would - as a minimum - want to report something like the mean (as a good measure of average) and the standard deviation (as a good measure of dispersion). The reason for this is that providing information on central tendency does not necessarily tell you anything about variability (or vice versa). For example, two sets of data can have very similar means (say the same test in two groups could produce similar means) but the variabilities may be quite different. The reverse could also be true. Just because two test score distributions have quite different means does not necessarily indicate that they also differ in variability. In a sense, the factors of central tendency and variability are more or less independent of one another and that is why you

need to examine and present information on both.

Variability and Constants

In Chapter 2, we looked at what happens to central tendency if the data set is operated on by some constant. Therefore, before leaving the current topic, we should also examine the impact of constants on variability. What happens to variability if we operate on a set of data by adding, subtracting, multiplying or dividing by a constant? Look at the data below where I added 2, subtracted 2, multiplied by 2 and divided by 2.

X	X+2	X-2	X*2	X/2
3	5	1	6	1.5
2	4	0	4	1.0
1	3	-1	2	0.5

Look at the following.

	N	MEAN	STDEV	VAR
X	3	2.000	1.000	1.000
X+2	3	4.000	1.000	1.000
X-2	3	0.000	1.000	1.000
X*2	3	4.000	2.000	4.000
X/2	3	1.000	0.500	0.250

As you recall in central tendency, changing a set of data by a constant changes the mean by the amount of that constant. From the data above, you can easily see that here. But, what about variability? If we look at the standard deviation, we see that when you add or subtract a constant, nothing happens to variability. This makes sense. If everyone earns $500.00 more next year in salary, then the distance from the highest paid person to the lowest paid person is still the same; there is no change in the spread of the data. But, when multiplying by a constant, note that the standard deviation changes by the amount of the constant. If you multiply by 2, the standard deviation goes up by a factor of 2 and if you divide by 2, the standard deviation is half what it was. What would happen to the variance in that case? Since the variance is the square of the standard deviation, if the standard deviation changes from 1 to 2 (doubles), the variance changes from 1 to 4, or is 4 times larger (quadruples). If the standard deviation changes from 1 to 1/2 (is reduced by a factor of 2), the variance changes from 1 to .25 which is reduced to 1/4th the size. Whatever happens to the standard deviation in the case of multiplying or dividing by a constant, happens to the variance by a factor of the square of the constant. It should be obvious therefore that the variance is more impacted than the standard deviation. Look at the summary table below. Assume that k is the value of the constant.

Statistic	Add/Subtract	Multiply/Divide
Mean	k change	k change
SD	no change	k change
Var	no change	k^2 change

Useful Minitab Commands

RANGE STDEV

Practice Problems

For the following data sets, first use something similar to a dotplot placing both data sets on the same baseline to indicate which distribution has more or less variability. Then, find the range, variance (S squared) and standard deviation (S) for each set and compare the variabilities. What would the standard deviation and variance change to in both sets of data if you multiplied each data value in each set of data by a constant of 3?

Data Set 1

49 58 60 55 56 51 47 55 45 47 52
56 47 57 43 40 58 45 49 50

Data Set 2

46 45 46 52 49 52 50 49 50 51 52
49 47 51 45 49 45 53 46 49