CHAPTER 4

LINEAR COMBINATIONS AND COMPOSITE GROUPS

So far, we have applied measures of central tendency and variability to a single set of data or when comparing several sets of data. However, in some instances, we first combine data together in some special way. For example, in a course, the instructor will generally determine grades by adding together several test and/or project scores prior to assigning grades. This operation of adding together several scores for the same people is called forming a **LINEAR COMBINATION**. In other instances, we may have 2 or more separate groups of students (like Section 1 and Section 2 of the same course) and would like to look at the average or variability of both groups combined together as one overall group. Combining two or more groups together on the same variable is called forming a **COMPOSITE GROUP**. This section briefly examines what happens to central tendency and variability in these linear combination and composite group situations.

Linear Combinations

Look at the data below that represent 3 tests in a course and their totals. The "ExamTot" variable is what is called a linear combination since it is a simple linear addition of the other 3 columns of numbers. Obviously, the "ExamTot" variable contains numbers that are much larger than any single test. However, the real question is: what relationship is there between the descriptive characteristics of the separate tests and the descriptive characteristics of the "ExamTot" linear combination?

1stExam	2ndExam	3rdExam	ExamTot
17	28	29	74
14	26	29	69
19	19	25	63
28	23	25	76
22	30	24	76
23	20	19	62
20	29	28	77
20	27	28	75
25	25	23	73
26	22	21	69
28	23	23	74
30	24	25	79
24	20	26	70

Look at some of the describe output from Minitab.

Variable	N	Mean	Median	TrMean	StDev
1stExam	13	22.77	23.00	22.91	4.69
2ndExam	13	24.31	24.00	24.27	3.57
3rdExam	13	25.00	25.00	25.18	3.06
ExamTot	13	72.08	74.00	72.36	5.22

The first thing to notice is that the mean of the "ExamTot" linear combination variable is simply the addition of the 3 separate test score means (22.77 + 24.31 + 25 = 72.08). Thus, for central tendency, the mean of the linear combination is the sum of the means of the components that go into the linear combination. For variability, notice that the standard deviation for the "ExamTot" linear combination is larger in this instance than the standard deviation for any separate test. While this is the typical situation, it does not always have to be like that. Look at the following.

XDown	YUp	Total
10	5	15
9	7	16
7	8	15
8	6	14
5	9	14
6	10	16

And again, here is some output from the describe command in Minitab.

Variable	N	Mean	Median	TrMean	StDev
XDown	6	7.500	7.500	7.500	1.871
Yup	6	7.500	7.500	7.500	1.871
Total	6	15.000	15.000	15.000	0.894

Notice in this case that, while the mean of the "Total" variable is still the addition of the two separate means of "XDown" and "YUp", the variability or standard deviation of the linear combination "Total" variable is less than the standard deviations of the separate components. Thus, for variability, there is no hard and fast rule. As I will note later (Chapter 8), the interrelationship among the components will play a role in whether the linear combination variability is relatively more or less than the variabilities of the component parts.

Composite Groups

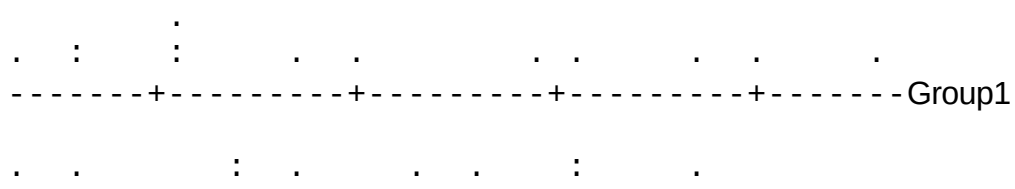
We will now examine what is referred to as the composite group situation. What if you had given the same short test to two different sections (1 and 2) of the same course

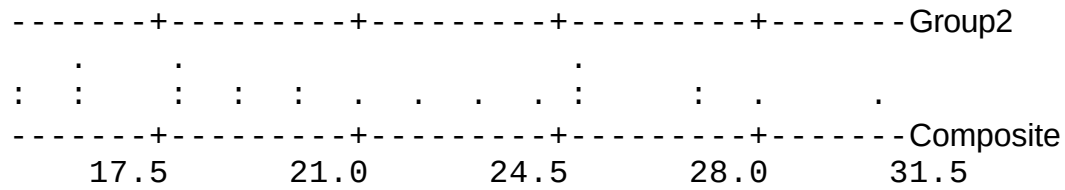
with the following results. Note: there are different numbers of students in the two groups.

Group1	Group2	Composite	
15	16	15	|
27	20	27	|
20	19	20	|
16	23	16	|
24	22	24	|
25	15	25	|
16	25	16	| ====> Group 1
28	27	28	|
18	19	18	|
21	25	21	|
18		18	|
30		30	|
18		18	|
		16	|
		20	|
		19	|
		23	|
		22	|
		15	| ====> Group 2
		25	|
		27	|
		19	|
		25	|

I have put the data for each group in separate columns but have also put the data together in a single new colum. Composite data situations take the data from two or more groups and recombine them into a new overall set of data. In effect, we are taking the data of the separate groups and "stacking" them into a new single column. In Minitab, there is a command called stack that easily accomplishes this.

One way to examine the composite set of data would be to compare it to the two separate sets of data. We could use something like a dotplot to do this as long as the sets of data are placed on the same scale. See the following.





Note that Group 1 is wider and Group 2 is somewhat narrower. When combining sets of data with similar means but different variabilities, the means of both the separate and composite groups will be about the same but the variability of the composite group will tend to be somewhere between the two separate group variabilities. See below.

Describe Output for the Variables

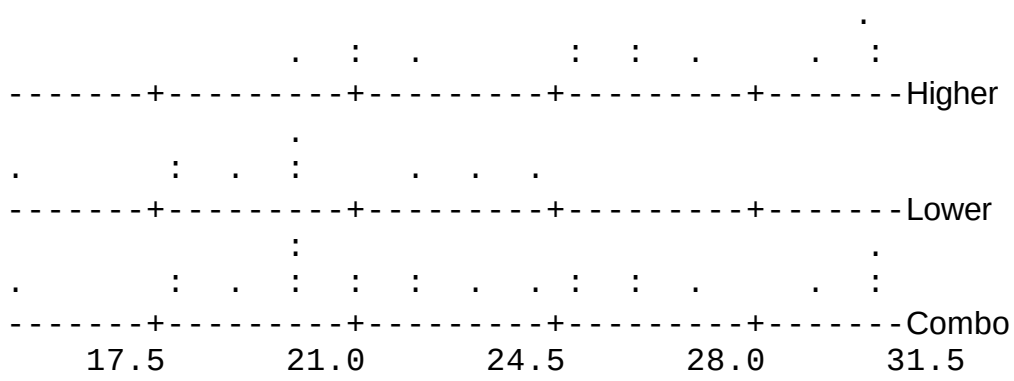
	N	MEAN	MEDIAN	TRMEAN	STDEV
Group1	13	21.23	20.00	21.00	5.04
Group2	10	21.10	21.00	21.12	3.98
Composite	23	21.17	20.00	21.04	4.50

Note that the mean of the composite set is about the same as the separate group means, but the standard deviation of the composite group (4.50) is somewhere between the standard deviation of Group 1 (5.04) and the standard deviation of Group 2 (3.98).

The above pattern does not always occur. Look at the following data sets (on the next page) that have been called "Higher" and "Lower". Again, I stacked the data into a new column and called the new variable "Combo". Dotplot diagrams using the same baseline for all variables can give us some clues about how the combined data compare to the separate groups along with some descriptive statistics. Note here that the "Higher" set is further up the scale and the "Lower" set is further down the scale. Obviously, the "Higher" set has a higher average score (hence the name "Higher"). But, in terms of variabilities, since we combine the two sets together, the new composite set will contain both the lower values from the "Lower" set and the higher values from the "Higher" set. The net effect of this is to make the composite set have more overall variability than either the "Higher" set or the "Lower" set above.

Higher	Lower	Combo
27	20	27
26	24	26
21	19	21
25	20	25

```
 26    20     26 |
 30    22     30 |
 29    18     29 | ====> Higher
 30    15     30 |
 20    18     20 |
 30    23     30 |
 22           22 |
 25           25 |
 21           21 |
             20  |
             24  |
             19  |
             20  |
             20  |
             22  | ====> Lower
             18  |
             15  |
             18  |
             23  |
```



The describe output from Minitab can be helpful too. See below.

Variable	N	Mean	Median	TrMean	StDev
Higher	13	25.54	26.00	25.64	3.64
Lower	10	19.90	20.00	20.00	2.64
Combo	23	23.08	22.00	23.14	4.27

Note that the standard deviation of the "Combo" composite variable is greater (4.27) than for the "Higher" variable (3.64) and the "Lower" variable (2.64). In this case, also see that the mean of the "Combo" variable is somewhere between the two separate group means. Thus, again, the composite characteristics depend upon the particular data sets combined and the relative comparison between or among the separate data sets.

Useful Minitab Commands

STACK

Practice Problems

For the data below, do the following.

1. For the Quiz1 and Quiz2 scores, add them together in a linear combination and then compare the individual means and standard deviations to the mean and standard deviation of the Total.

2. For the 1stGroup and the 2ndGroup data sets, stack them together into one overall composite set and then compare the means and standard deviations of the separate groups to the composite group or stacked set.

Quiz1	Quiz2	1stGroup	2ndGroup
20	16	50	52
20	18	56	54
18	13	53	50
16	17	60	53
10	14	54	53
19	17	56	51
12	16	53	48
17	20	58	53
11	16	59	48
14	18	60	48
		52	52
		51	48
			51
			50
			49

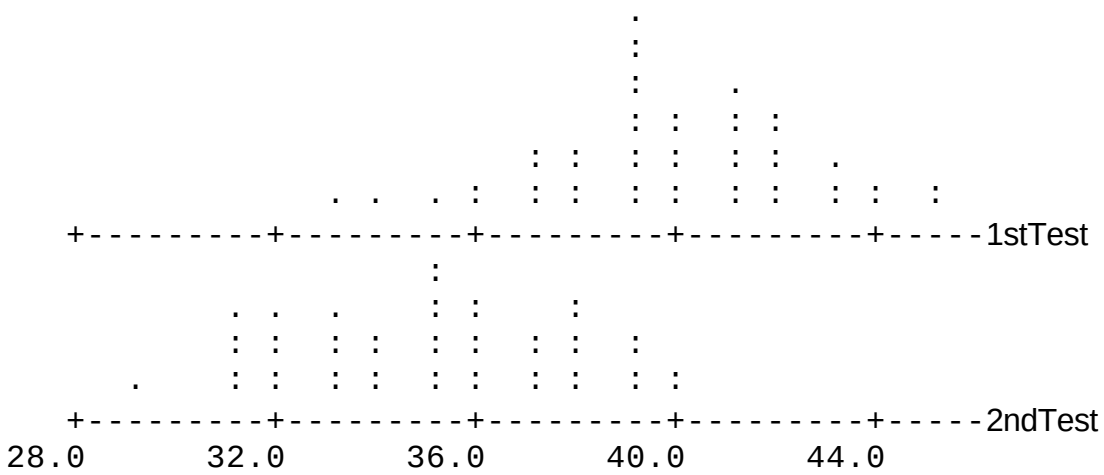
CHAPTER 5

MEASURES OF POSITION

In some instances, we would like to examine the location or position of specific values within a distribution of scores. For example, two tests are given and a student obtains the score of 20 on both. However, for one test, the mean is 19 which means that the score of 20 is above the mean. But, for the other test, the mean is 21, which means that the score of 20 is below the mean. We could simply say that on one test, the person obtained a score that was above the "average" whereas on the other test, their score was below "average". Unfortunately, this description of where the student fell with respect to the means of the two tests is not very precise. The present Chapter will review several methods for making position or location indications more precise.

Ranks

Perhaps the simplest way to indicate position would be to rank the scores. Look at the following dotplots and the describe output from Minitab for two tests.



Describe Output for the 1sTest and 2ndTest Variables

	N	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN
1stTest	50	39.780	40.000	39.841	2.644	0.374
2ndTest	50	35.060	35.000	35.068	2.766	0.391

The "1stTest" variable obviously contains score values that are higher up the scale

while the "2ndTest" variable has lower values. Values that are near 36 on "2ndTest" are almost at the bottom of the "1stTest" set of data. Thus, a person who obtained 36 on both would be near the bottom on the first but near the middle on the second.

The process of ranking the data requires us to first sort or order the data from low to high. We saw this process being used for finding the median. Therefore, using Minitab to order the "1stTest" and "2ndTest" variables, I have ordered sets of data as below.

Ordered Data for 1stTest

33	34	35	36	36	37	37	37	37	38	38
38	38	39	39	39	39	39	39	39	39	39
39	39	40	40	40	40	40	40	41	41	41
41	41	41	41	42	42	42	42	42	42	43
43	43	44	44	45	45					

Ordered Data for 2ndTest

29	31	31	31	31	31	32	32	32	32	32
33	33	33	33	33	34	34	34	34	35	35
35	35	35	35	35	35	36	36	36	36	36
36	37	37	37	37	38	38	38	38	38	38
39	39	39	39	40	40					

To rank scores, you start at the top or high point of the distribution. For example, in the "1stTest" distribution, we see that there are two scores of 45 at the top. Since they are at the top, they should both be given ranks of 1 but, because they are tied, the technical way to rank the values is to average the positions (1 and 2) and give them both ranks of 1.5. **FOR TIED SCORES, GIVE THE RANK THAT EQUALS THE AVERAGE POSITION**. For scores of 44, since they occupy the 3rd and 4th positions, each would get a tied rank of 3.5. We continue to work our way down through the data set until we come to the last value of 33 which, being untied and at the bottom, would be assigned the last rank = 50. Note: the lowest value will get a rank of N if it is untied. For the "2ndTest", the top two scores are 40 and, being tied in the first and second positions, will each receive ranks of 1.5. For the next score 39, where there are 4 tied values (positions 3rd, 4th, 5th and 6th), each would get tied ranks of 4.5. Again, working our way down to the bottom value, the score of 29 will be assigned a rank of 50.

Here is a sample of the ranks that would be assigned to the scores in the 1stTest and 2ndTest variables. What is shown represents the first 15 values that occured in the unordered sets of data that were originally placed in columns for analysis purposes.
Ranks for First 15 Values in the 1stTest and 2ndTest Variables

1stTest	RankTes1	2ndTest	RankTes2

39	32.0	39	4.5
40	23.5	38	9.5
39	32.0	33	37.0
42	10.5	32	42.0
39	32.0	37	14.5
45	1.5	31	47.0
39	32.0	36	19.5
41	17.0	38	9.5
41	17.0	40	1.5
39	32.0	36	19.5
39	32.0	32	42.0
43	6.0	29	50.0
43	6.0	39	4.5
40	23.5	38	9.5
37	43.5	38	9.5

Keep in mind that in our data sets of N = 50, scores with ranks close to 1 would be near the top, scores with ranks close to 25 would be near the middle, and scores with ranks close to 50 would be near to the bottom. For example, using one specific case, the score of 40 in the "1stTest" set has a rank of 23.5 which puts it near the middle whereas the same score of 40 in the "2ndTest" set results in a rank of 1.5 which is near the top. Comparing positions therefore shows that in these sets of data, the score of 40 is at quite a different location in the first set of data compared to the second set.

However, as simple as ranks are to find, one major problem with ranks is the fact that differences in ranks can be deceiving in terms of the distances between the original scores. Look at the following simple example to illustrate this point.

Original Score	Rank
22	1
21	2
15	3

The difference between the first two scores (22 - 21) is 1 point and the difference in the ranks is also 1 rank unit (1 - 2 = 1). But, for the scores of 21 and 15, where the score difference is 6 points, there still is only a 1 unit difference in the ranks. This problem occurs often and can be misleading. To assume that the raw score differences are the same between two pairs of ranked values that have the same difference in ranks can be quite inaccurate. Thus, while ranking is reasonable way for indicating relative position, one needs to be careful not to read quantitative meaning into ranks that does not exist.

Percentile Ranks

A second way to indicate position within a data set is called **PERCENTILE RANK**. A percentile rank, which sounds similar to a rank but is not, is simply an indication of **HOW MUCH OF THE DISTRIBUTION IS BELOW SOME SCORE VALUE**. For example, a score that has nearly 0 percent below it must be near the bottom of the distribution whereas a score that has nearly 100 percent below it must be near the top. Also, a score with nearly 50 percent below it must be in the middle and must be close to the median. The median has a percentile rank of 50.

To look at percentile ranks, we can use a variation of the tally command from Minitab that prints out several additional columns of data. See the following.

Expanded Tally Output for the 1stTest Variable

1stTest	COUNT	CUMCNT	PERCENT	CUMPCT
33	1	1	2.00	2.00
34	1	2	2.00	4.00
35	1	3	2.00	6.00
36	2	5	4.00	10.00
37	4	9	8.00	18.00
38	4	13	8.00	26.00
39	11	24	22.00	48.00
40	6	30	12.00	60.00
41	7	37	14.00	74.00
42	6	43	12.00	86.00
43	3	46	6.00	92.00
44	2	48	4.00	96.00
45	2	50	4.00	100.00
N=	50			

In addition to the score and frequency information, we can obtain from tally additional information that is useful for finding percentile ranks. The **CUMCNT** column adds up the frequencies from the bottom to the top so that when you reach the top, you have accumulated 50 frequencies (in this example). The last or **CUMPCT** column converts the accumulated frequency values into percentages out of the total of N=50. For example,

Expanded Tally Output for the 2ndTest Variable

2ndTest	COUNT	CUMCNT	PERCENT	CUMPCT
29	1	1	2.00	2.00
31	5	6	10.00	12.00
32	5	11	10.00	22.00
33	5	16	10.00	32.00

34	4	20	8.00	40.00
35	8	28	16.00	56.00
36	6	34	12.00	68.00
37	4	38	8.00	76.00
38	6	44	12.00	88.00
39	4	48	8.00	96.00
40	2	50	4.00	100.00
N=	50			

through the first score interval (which technically goes from the limits of 32.5 to 33.5), we have accumulated 1 out of 50 frequencies or 2 percent. Up through a score of 40 (which has an upper limit of 40.5), we have accumulated 30 out of 50 or 60 percent. By the time we have reached the tip top of the distribution (technically 45.5), we have accumulated 50 out of 50 or 100 percent. These cumpct values are the percentile ranks since they represent how much of N is below that score point in the distribution. For practical purposes, percentile ranks are reported as being either less than 1 up to greater than 99. Normally, you will not see percentile ranks of 0 or 100 listed (even though Minitab in its expanded tally output for the cumpct does put 100 as the top value).

Again, for comparison purposes, let's concentrate on the score of 40 in both the "1stTest" and "2ndTest" distributions. A score of 40 in the "1stTest" set of data has a percentile rank of approximately 60 whereas the same score of 40 in the "2ndTest" distribution has a percentile rank close to the maximum, nearly 100. Thus, 40 in the first set of data is near the middle of that distribution whereas 40 in the second set is near the top. Therefore, percentile ranks provide us with a second way to indicate position or location within a distribution. One real potential interpretation problem with percentile ranks is that it is very easy to confuse them with scores like "percentage correct" values on a test. The fact that someone obtained a percentile rank of 80 on a test does not mean that they answered 80 percent of the test questions correctly. All it means is that 80 percent of the group taking the test had scores less than the first person. It would be very unusual for a person's percentile rank on a test and his or her percentage correct score on the test to be the same.

Standardized Scores: z

A third way to indicate position would be to measure how many standard deviation units a particular score is away from the mean. Again, as a reference point, look at some of the describe output for the 1stTest and 2ndTest variables. .

	N	MEAN	STDEV
1stTest	50	39.780	2.644
2ndTest	50	35.060	2.766

For example, a score of 45 in the "1stTest" data set is (45 - 39.78)/2.644 = 1.97 standard deviations above the mean. However, for a score of 42, the distance is (42 - 39.78)/2.644 = .84 standard deviations above the mean. Thus, a score of 45 is further up in the distribution (above the mean) than a score of 42.

Thus, we can use the distance that a score is from the mean in terms of standard deviation units as a measure of position or location. Look at the following formula.

$$z_x = \frac{X - \overline{X}}{S_x}$$

The formula above is called a z score formula and a **z SCORE INDICATES THE NUMBER OF STANDARD DEVIATION UNITS A SCORE VALUE IS AWAY FROM THE MEAN**. If the score is above the mean, the z score will be positive (like the two calculated above). However, if the score is below the mean, then the z score will be negative. What will the z score be if, by chance, the score value is located exactly at the mean? Correct, 0. Hold that thought for a moment! For the 1stTest score distribution, what would the z score be for a score that happened to be exactly 2.644 points above the mean of 39.78? Since it would be one standard deviation unit above the mean, the z score would be 1. Again, hold that thought for a moment!

To calculate z scores, you first need to find the raw deviation scores around the mean (score - mean) and then divide each of these by the standard deviation. Most software packages have command or routines for easily calculating z scores. In Minitab for example, you could first make a column of the deviation scores and then next make another new column where we divide the deviation score by the standard deviation. Also, in Minitab, there is a command called **CENTER** that will automatically convert the original raw scores to z scores.

For example, let's assume that the data for the 1stTest and 2ndTest variables are located in worksheet columns C1 and C2. I could then use the center command in Minitab to convert the original values to z scores. Look at the work below. Note: The **MTB>** is called the Minitab prompt which allows the user to enter commands and have work done. The commands (just as examples) below take the data in C1 and C2, convert to z scores in each case, then place the z score conversions into new columns, C7 and C8. A few of the 50 values for each variable are printed below.

```
MTB> center c1 c7
MTB> center c2 c8
MTB> print c1 c7 c2 c8
```

1stTest	zonTes1	2ndTest	zonTes2
39	-0.29499	39	1.42457
40	0.08320	38	1.06300
39	-0.29499	33	-0.74482
42	0.83960	32	-1.10639
39	-0.29499	37	0.70144
45	1.97418	31	-1.46796
39	-0.29499	36	0.33987
41	0.46140	38	1.06300
41	0.46140	40	1.78613
39	-0.29499	36	0.33987

How can we use z scores to make comparisons of positions? Again, what if we focus on the scores of 40 in both the "1stTest" and "2ndTest" distributions? For the "1stTest" variable, a score of 40 has a z score of .08 which means it is only about 1/10th of a standard deviation unit above the mean. Since the z is close to 0, the score of 40 is close to the mean. However, in the "2ndTest" data set, a score of 40 has a z score of 1.78 which means that it is almost 2 standard deviation units above the mean. In this case, the larger positive z score means that 40 is much further away from the mean in the second set of data compared to the first. The closer the z score is to 0, the closer is the raw score to the mean. The further away from 0 the z score is, the further away from the mean is the raw score. For a summary of z score data, look at the following descriptive data.

	N	MEAN	STDEV
1stTest	50	39.780	2.644
zonTes1	50	0.000	1.000
2ndTest	50	35.060	2.766
zonTes2	50	0.000	1.000

The use of z scores is a nice way to indicate position within a distribution and a nice way to compare positions of the same scores in two or more distributions. Remember, positive z scores mean above the mean and negative z scores mean below the mean. What is the average of the z scores? You can see from the describe output that the means of the z scores are 0. This makes sense since any score that is at the mean will be converted to a z score of 0. But, what about the standard deviation? Note that the standard deviations are 1 on the z score scale. The reason for this is simple too. For the "1stTest" data, look at the following.

Raw Score	39.78	42.424	45.068
z Score	0	1	2

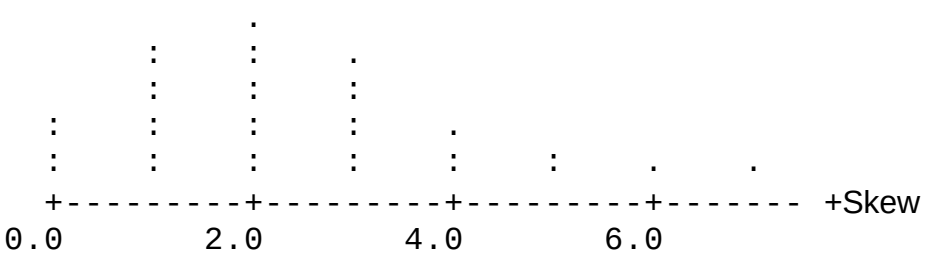
The distance from 39.78 to 42.424 is 2.644 or 1 standard deviation unit on the raw score scale. However, on the z score scale, that same distance is from 0 to 1 in z score units, or 1 unit. Thus, the distance of 2.644 on the raw score scale (one standard deviation) is equal to 1.00 on the z score scale (one standard deviation). Simply put, means of z scores are always 0 and standard deviations of z scores are always 1. These are constant values for z scores and you should try to remember these.

Skewness and z Scores

In both Chapters 1 and 2, the concept of skewness was mentioned. In Chapter 2 , a quick measure of skewness was given as the difference between the mean value and the median value. However, a better way for quantifying skewness is to use a measure involving z scores. In symmetrical distributions, one would see the same pattern of z scores on the right side of the mean as you would on the left side of the mean. However, in skewed distributions, this is not the case. For example, a positively skewed distribution will have more distance to the right of the mean than to the left of the mean, since the distribution stretches out more to the right. What this means is that there will be larger positive z scores that extend more to the right of the mean. Thus, while you may have z scores that go up to around +3 (to the right of the mean), you may only see z scores extend on the negative side down to about -2. Thus, on the average, z scores will tend to have higher positive values in positively skewed distributions than in a symmetrical distribution. Just the opposite occurs in a negatively skewed distribution. Thus, if we examine the pattern of z scores, we should obtain information on the skewness in the set of data. See the following formula.

$$\text{Skew}' \frac{(\sum z^3)}{N}$$

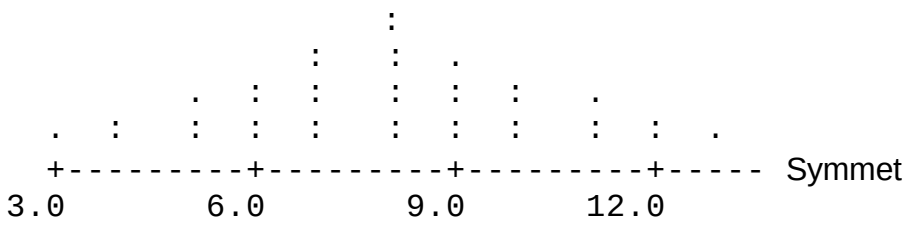
One measure of skewness, sometimes called **GAMMA 1**, is simply the average of the cubes of the z scores. In a positively skewed distribution for example, cubing the z's will produce larger cubes for the positive z's than it will for the negative z's. Thus, in a positively skewed distribution, the average of the cubes of the z's will be positive. Look at the following set of data.



Clearly, this set of data is positively skewed. To use the measure of skewness above, we need to convert the data to z scores, then cube the z scores, and then take an average. By using new columns, Minitab can easily accomplish that. See below for a few examples of the z and cubed z scores.

+Skew	z	zcube
0	-1.38078	-2.6326
0	-1.38078	-2.6326
0	-1.38078	-2.6326
0	-1.38078	-2.6326
1	-0.79142	-0.4957
1	-0.79142	-0.4957
1	-0.79142	-0.4957
.	.	.
.	.	.
5	1.56601	3.8405
5	1.56601	3.8405
6	2.15537	10.0130
7	2,74473	20.6775
	Mean =	0.77151

The average of the cubes of the z values is .77 and represents our skewness measure. If the value had been larger than .77, then there would have been more positive skew. If the skewness value had been less than .77, then there would have been less positive skew. Now, look at a second example.



Here is a distribution that looks nearly symmetrical. To apply the skewness measure here, again we need to calculate z scores and then the cubes of the z scores. Finally, we take an average. Again, letting Minitab do the work, here are a few values below.

Symmet	z1	z1cube
3	-2.11113	-9.40897
4	-1.68672	-4.79879
4	-1.68672	-4.79879
5	-1.26232	-2.01146
5	-1.26232	-2.01146
5	-1.26232	-2.01146
.	.	.
.	.	.
11	1.28409	2.11730
12	1.70849	4.98696
12	1.70849	4.98696
13	2.13289	9.70298
	MEAN =	0.029850

In this case, since the skewness value is only about .03, which means that the distribution is close to being symmetrical. A skewness value of exactly 0 would mean a perfectly symmetrical distribution.

Standardized Scores: Other Linear Transformations

It is important to point out that z scores are often used to convert data to other scales that we see in the literature. I will show you two such conversions here: big T scores and SAT type numbers. The purpose of this is to show you the method of conversion, which is sometimes called a **LINEAR SCORE TRANSFORMATION**. Actually, z scores are also linear transformations from the original score scale that happen to have a new mean of 0 and a new standard deviation of 1.

One useful linear transformation is called big T. T scores have an arbitrary mean of 50 and an arbitrary standard deviation of 10. Actually, if you think about it, if we added 50 to the mean of z scores and multiplied the standard deviation of 1 (for z) by 10, we would have the T scale. In fact, that is exactly what is done. Another transformation that is used in the area of college admissions testing programs (of the College Board variety) is called SAT that has had a mean of 500 (for one section of the test) and a standard deviation of 100. Actually, if we add 500 to the mean of the z scores and multiply the standard deviation of the z scores (which is 1) by 100, we would have the SAT scale. And, again, that is exactly what is done. Or, we could multiply the mean of the big T scale (50) by 10 to obtain 500, and multiply the standard deviation of the big T (10) scale by 10 to obtain 100. To convert orginal score datat to either the big T or SAT type scale transformations, we first need z scores. Recall that for the "1stTest" and "2ndTest" variables, we have already seen how these scores were converted to z values. So, to change either a 1stTest or 2ndTest score value to a T value, I need to muliply the z score times 10, and add a constant of 50 to it. For the SAT conversion, I need to multiply the z score times 100 and then add to that, a

constant of 500. Again, using Minitab makes this conversion easy if the z scores for each of the 1stTest and 2ndTest distribution values are located in a worksheet column. I have printed out some of the values below.

1stTest	zonTes1	TTest1	SATTest1
39	-0.29499	47.0501	470.501
40	0.08320	50.8320	508.320
39	-0.29499	47.0501	470.501
42	0.83960	58.3960	583.960
39	-0.29499	47.0501	470.501
45	1.97418	69.7418	697.418
39	-0.29499	47.0501	470.501
41	0.46140	54.6140	546.140
41	0.46140	54.6140	546.140

2ndTest	zonTes2	Ttest2	SATTest2
39	1.42457	64.2457	642.457
38	1.06300	60.6300	606.300
33	-0.74482	42.5518	425.518
32	-1.10639	38.9361	389.361
37	0.70144	57.0144	570.144
31	-1.46796	35.3204	353.204
36	0.33987	53.3987	533.987
38	1.06300	60.6300	606.300
40	1.78613	67.8613	678.613

Again, if we concentrate on the scores of 40 in both the 1stTest and 2ndTest distributions, we see that on the 1stTest variable, a value of 40 converts to a big T of about 50.8 and a SAT type value of about 508. For the same score of 40 on the 2ndTest varaible, the big T and SAT type values are 67.8 and 678 respectively. If we use 50 and 500 as the mean big T and SAT type values, then obviously 40 is much closer to the middle of the 1stTest distribution compared to the comparable position on the 2ndTest variable.

To summarize the different measures of position, look at the example of our score of 40 from both data sets. Here is what we have up to this point.

Data Set	Rank	Percentile Rank	z	T	SAT
1stTest	23.5	60	.08	50.8	508

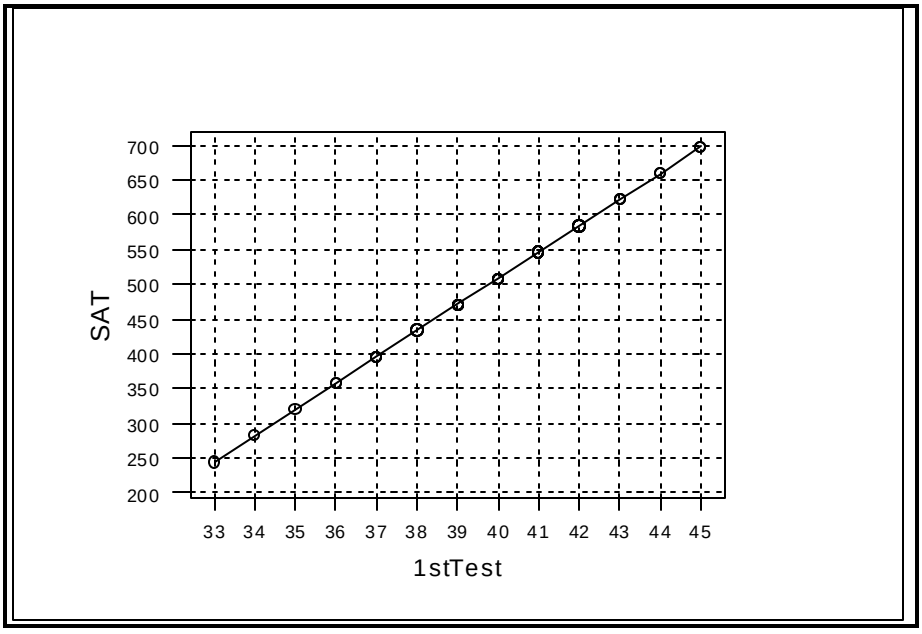
2ndTest	1.5	> 99	1.78	67.8	678

It is important to realize that conversions of the original data to transformed scales such as z or big T or SAT type, are only meant to locate the scores on a different scale. This in no way changes the meaning of the original data. For example, even though I have shown how you can change the original scores to SAT type numbers, that does not mean that a person who obtained that particular original score would have obtained that SAT number if he or she had taken the SAT test. All the transformation indicates is what the SAT number would look like if a person had obtained a score (on the SAT test) located in the same position as was the original value.

What may not be obvious from the conversion or transformation process is the guiding rule for doing so. That rule is a "linear transformation" process. The process of converting scores into z values, and then further converting them to things like big T or SAT type values, is merely changing the scale according to a straight line formula. The formula looks as follows. As was indicated above, the z score is multiplied by the new standard deviation, and then the constant of the new mean is added.

$$\text{New Score} = (\text{Old } z * \text{New } S) + \bar{\text{New } X}$$

To see this linear pattern, look at the plot of the "1stTest" scores and the transformed SAT type values. For example, our value of 40 on the baseline would convert to an SAT type number just a little above 500, and the values of 39 and 43 would convert to SAT type values of about 470 and 620 (guesstimating from the graph) respectively.



The purpose of a measure of position is to provide a quantitative way to indicate

the location of a score, relatively speaking, in a distribution. For the example of 40 above, we could look at ranks, or percentile ranks, or z scores, or big T values or finally (for the current material) SAT type values. Regardless of the specific method used, the main point here is that a score of 40 in the "1stTest" distribution is near the middle of the set whereas that same score of 40 is near the top in the "2ndTest" set of data. Conversions like z scores or T values or SAT type numbers give us alternative procedures for conveying information to consumers about the positions of scores within distributions of data.

Useful Minitab Commands

TALLY with ALL subcommand CENTER

Practice Problems

For the data below, find the ranks, percentile ranks, z scores, T and SAT type values for the score of 14 in both sets of data. Compare the relative positions in both distributions. Compute both the simple and gamma 1 measures of skewness for each: are they skewed and if so, how?

TestA

11	18	12	11	12	14	11	12	12	14	16
10	13	13	12	14	13	16	12	10	8	12
17	14	10	15	12	11	13	13			

TestB

16	14	16	11	17	13	15	16	15	16	13
14	14	17	16	17	19	14	14	15	11	16
10	14	15	18	16	13	14	16			