DOCUMENT RESUME

ED 342 790 TM 017 947

AUTHOR Bump, Wren M.
TITLE The Normal Curve Takes Many Forms: A Review of Skewness and Kurtosis.
PUB DATE Jan 91
NOTE 17p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 24-26, 1991).
PUB TYPE Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Data Collection; Equations (Mathematics); Functions (Mathematics); Graphs; *Mathematical Models; *Probability; *Raw Scores; *Statistical Distributions; Test Results
IDENTIFIERS Kurtosis; *Normal Distribution; Shape Statistics; *Skew Curves

ABSTRACT

The normal curve has long been important in statistics. Most interval variables yield normal or quasi-normal distributions when data are collected from large samples, and the normal "Z" distribution is also used as a test statistic (e.g., to test differences between two means when sample size is large, since "t" approaches "Z" as degrees of freedom increase). Thus, almost all statistics books discuss the normal curve. Nevertheless, many researchers do not fully understand some concepts related to the normal curve, such as skewness and kurtosis statistics, because these two statistics often receive cursory instructional treatment, given the press for instructional time. This paper illustrates that shape statistics remove the influence of distribution variability (i.e., shape statistics always initially involve the conversion of raw scores to "Z" form, SD=1=V, so that impact of variability is held constant). Nine figures illustrate the shape statistics, and one table lists raw scores and "Z" scores. An eight-item list of references is included. (Author/SLD)

The Normal Curve Takes Many Forms: A Review of Skewness and Kurtosis

Wren M. Bump
Texas A&M University 77843

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX, January 25, 1991.

Abstract

The normal curve has long been important in statistics. Most interval variables yield normal or quasi-normal distributions when data are collected from large samples, and the normal Z distribution is also used as a test statistic, e.g., to test differences between two means when sample size is large, since t approaches Z as degrees of freedom increase. Thus, almost all statistics books discuss the normal curve. Nevertheless, many researchers do not fully understand some concepts related to the normal curve, such as skewness and kurtosis statistics, because these two statistics often receive cursory instructional treatment, given the press for instructional time. This paper illustrates that shape statistics remove the influence of distribution variability, i.e., shape statistics always initially involve the conversion of raw scores to Z form (SD=1=V) so that the impact of variability is held constant.

The Normal Curve Takes Many Forms: A Review of Skewness and Kurtosis

The normal curve has many useful mathematical properties. For example, the percentage of people scoring within SD units from the mean is always known. Thus, 68% of scores fall between the mean and plus or minus one SD in a normal distribution, 95% fall between the mean plus or minus two SDs, and 99% fall between the mean and plus or minus three SDs (Gronlund, 1971, p. 387). These facts are useful to researchers because internally scaled data often are normally or nearly normally distributed. Put differently, when these and related rules work, scores can be considered distributed normally. Thus, it is useful to know when data constitute a normal distribution.

What is a normal curve and what does it look like?

The normal curve was investigated in the eighteenth century by mathematicians who were asked by gamblers, interested in winning gambling games, what the probabilities of certain outcomes were. Their chances of winning were represented by a curve (Downie & Heath, 1965). This work was elaborated by others and is widely used today. Downie and Heath point out the following assumptions made about the normal curve:

> In our educational and psychological work, we assume that certain traits are normally distributed. In actuality, probably no distribution ever takes on the absolute form of the normal distribution. Many of our frequency distributions are very close to the normal one, and we assume that they have a normal distribution. To the extent that our distributions differ from normal, error enters into our work. The normal curve is important not primarily because *scores* are assumed to be normally distributed, but because the *sampling* distributions of various statistics are known or assumed to be normal. (p. 69)

The first users of the normal curve believed that almost all human characteristics were distributed in a random fashion around an average value. These human characteristics included intellectual and moral qualities as well, and this thought, that somehow abilities are naturally distributed in a normal way, has carried over to mental measurement (Nitko, 1983). This is one reason why the normal curve is studied, analyzed, and thought to be important in statistics.

The normal curve has been described as a mathematical model defined by a particular equation that depends on two specific numbers: the **mean** and the **standard deviation**, signifying that many normal distributions exist and each has a different mean or standard deviation (Nitko, 1983). These two statistics are then used to calculate two additional statistics that are used to evaluate normalcy: skewness and kurtosis. These four elements, the **mean**, the **standard deviation**, the **skewness**, and the **kurtosis**, are called **the first four moments of a normal distribution.**

<u>Why are these statistics referred to as "moments"?</u>

A **moment** is a mechanical term for the measure of a force with reference to its tendency to produce rotation. This tendency to produce rotation is related to the amount of the force applied and the distance from the origin that this force is exerted (Mills, 1955). In Figure 1 we see eight pounds and two pounds representing the forces applied in a given situation. The eight pounds of pressure being exerted on the point one foot above the origin at zero is balanced by a force of two pounds being exerted four feet below the origin. The sum of the moments tending to cause rotation in one direction is equal to the sum of the moments tending to cause rotation in the opposite direction, so the object is balanced (Odell, 1957). If either of these points was moved or if the point of origin was moved, the sum of the forces that are measured by the moments would not be zero and the object would not be in balance.

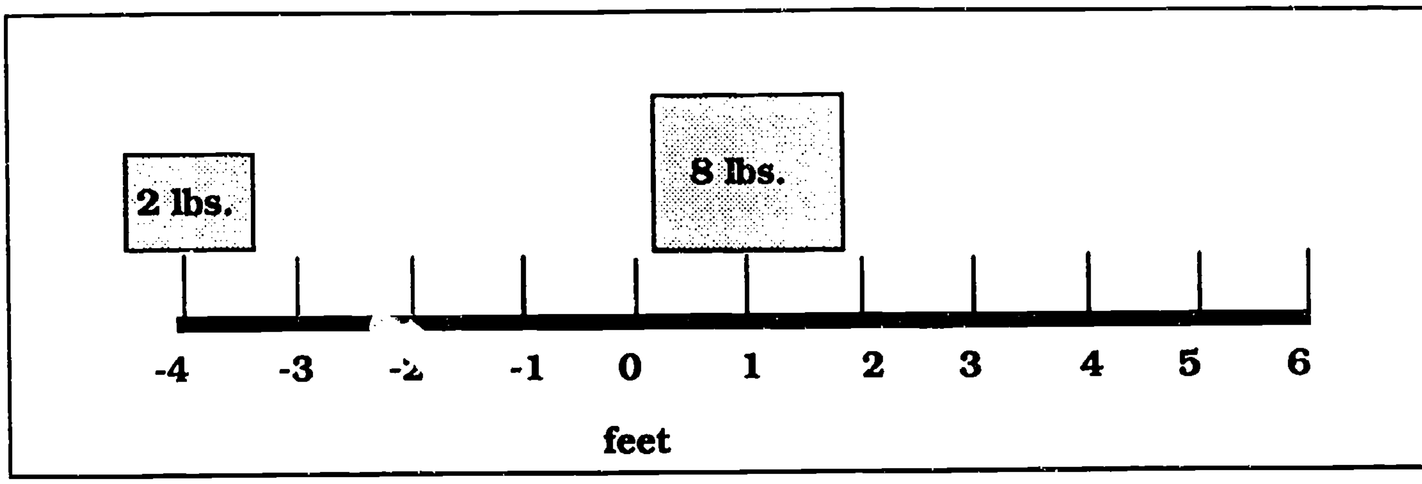

Figure 1. The relationship of weight and distance on the balance of the line.

In statistics, the term "moment" denotes class frequencies that are analogous to the forces exerted in the previous example. In Figure 2 we see a histogram for a test in which the mean is 104 and in which there are 90 grades. If each of the columns is thought of as a solid rectangle, with each column exerting a pressure on the x-axis, we can see the contribution of the forces. The "moment" contribution of each column is measured by the product of the class frequency (f) and the corresponding deviation (x) from the origin. The sum of the fx products, divided by the total frequencies, gives a net measure called the **first moment** or the **mean** (Mills, 1955).

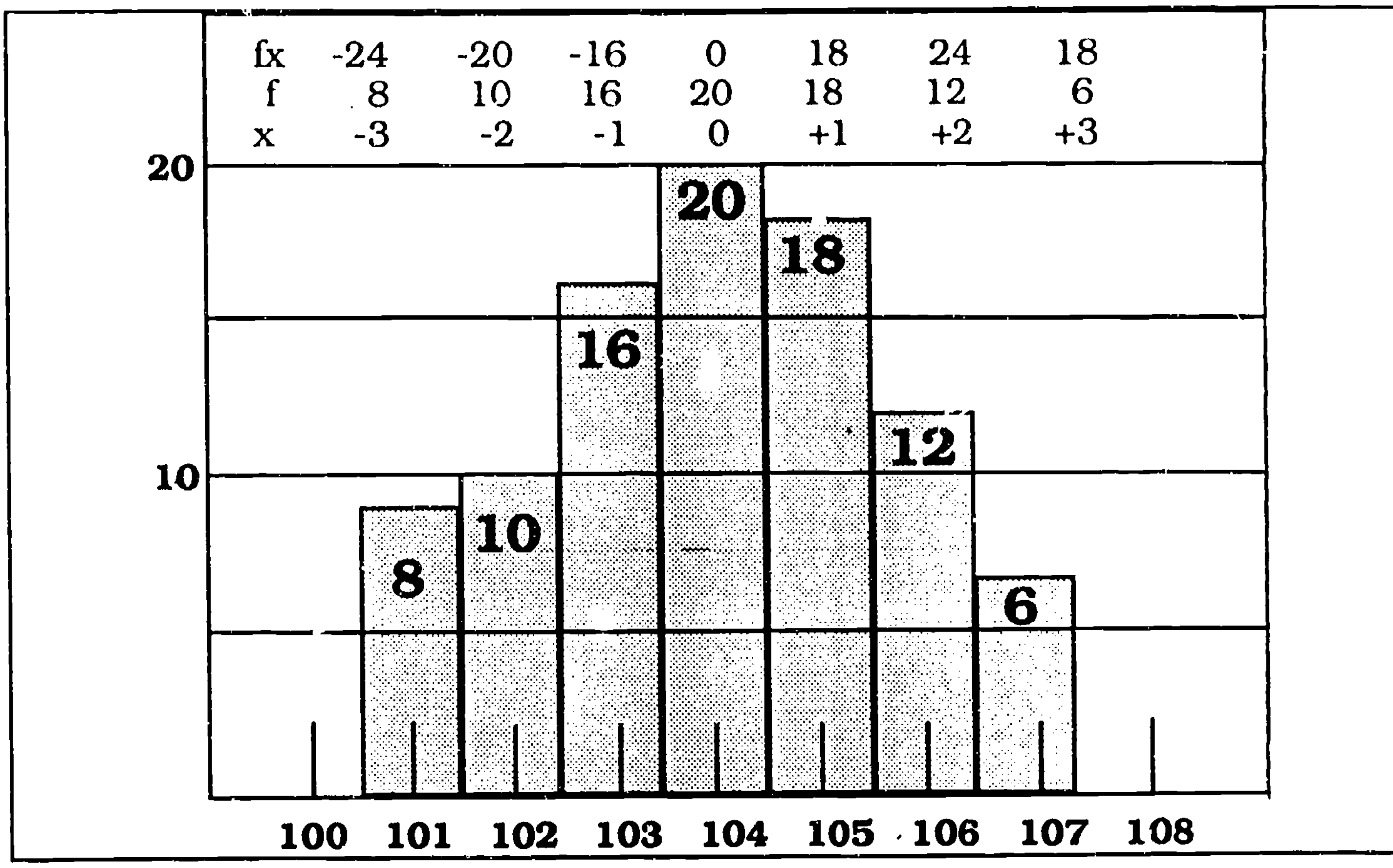

Figure 2. Class frequencies, pressures on the x-axis, and the balance around the mean.

The **second moment** of a statistical series is the **standard deviation.** It is a measure of the variation of the scores around the mean in standard units. Slight differences in patterns of variation are reflected in the moments which define the degree and character of the variation.

The third and fourth moments, skewness and kurtosis, are directly related to the standard deviation. Skewness and kurtosis quantitatively indicate the **nonnormal variation** in the statistical series. Skewness refers to the asymmetry of the curve and kurtosis refers to the tallness or flatness of the curve. Both of these moments depend upon the manner in which the scores scatter about the mean. A symmetrical curve provides a mirror image from a line drawn through the mean. But if the scatter is greater on one side of the mean than on the other side, the distribution is said to be skewed (Tate, 1965). When the distribution of scores extends from the mean further toward the **larger values** than smaller values of the distribution, the distribution is said to be **positively skewed** or **skewed right**. When the distribution of scores extends from the mean further toward the **smaller values** than larger values of the distribution, the distribution is said to be **negatively skewed** or **skewed left**.

The **skewness** is formulated from the **third moment** of the distribution because it reflects the average of the deviation scores raised to the **third power** divided by the standard deviation raised to the **third power** (Newell & Hancock, 1984). The formula for this is:

$$\text{Skewness} = \frac{\Sigma\,(X\text{-}X)^3/n}{SD^3}$$

When all the scores have been converted to z-scores (X=0; SD=1), we can use a much simpler formula and will always get the same answer for a given data set:

$$\text{Skewness} = \frac{\Sigma\,z^3}{n}$$

When there is a higher concentration of scores around the mean, the distribution is relatively narrow and the curve has positive kurtosis. When there is a low concentration of scores around the mean, the distribution is relatively broad and the curve has negative kurtosis. **Kurtosis** is called the **fourth moment** of the distribution because it is the ratio of the average of the deviation scores raised to the **fourth power** to the standard deviation also raised to the **fourth power**. Using this formula, the normal curve has a kurtosis value of 3, although the common practice of researchers and statistics packages now is to subtract 3 from the kurtosis value obtained so that zero represents the kurtosis value for a normal curve (Newell & Hancock, 1984) just as skewness of 0 implies no skewness relative to the normal distribution. A "tall" or "peaked" curve has a kurtosis value greater than 0, and a "flat" curve has a kurtosis value less than 0. The formula for kurtosis is:

$$\text{Kurtosis} = \frac{\Sigma (X\text{-}X)^4 / n}{SD^4} - 3$$

Alternatively, if the scores are converted to z-scores, we can use a much simpler formula that always yields the same answer for a given data set:

$$\text{Kurtosis} = \frac{\Sigma z^4}{n} - 3$$

<u>How does all this help us look at a curve and estimate if it is a normal curve?</u>

Asking if a certain curve *looks* like a normal curve is like asking, "Does that person look tall?", without knowing how fat that person is. So the real question to ask is, "Is the curve tall in comparison to its spreadoutness?" We cannot think about how tall someone looks without knowing how fat or skinny they are. So to compare people we would need to make them all the same width and then we could could easily see their variations and which ones vary from the norm. In order to compare test scores and their distribution, we need to make them all the same "width" by standardizing them. This can be done by converting the raw scores to z-scores.

Z-scores are the most basic standard scores and are used to derive other kinds of standard scores. Z-scores express raw score performance in terms of the number of SD units above or below the mean. "Knowing the z-score of any score enables us to determine the percentile rating of the score by comparing it to the properties of the standard normal distribution" (Moore,1983, p. 221). Table 1 presents the z-scores for the data (Thompson, 1991) that will be employed to illustrate these dynamics..

INSERT TABLE 1 ABOUT HERE

By using any spreadsheet application, the raw scores can be quickly converted to z scores. It is very helpful if the spreadsheet application that is used has graphing capabilities. The columns with range and frequency are used to obtain the histogram representing the curve produced by the scores.

Table 1 presents the sum of the 100 *z-scores*; by definition it is zero because the mean of *z-scores* is 0, and this only occurs if the sum of the scores is also 0. The next column shows z^2 and when these are summed and divided by the number of scores, the standard deviation (.9998) is found. The next column produces the value of skewness, i.e, z^3 summed (0.0000) and divided by the number of scores is 0.0000. The last column produces the value of kurtosis when z^4 is summed (285.3902) and divided by the number of scores, i.e., 2.8539.

Once these values have been obtained we can also graph the frequencies of the scores and then manipulate certain values to explore what happens to our curve. In the first set of figures presented in Figures 3a., 3b., and 3c, the standard deviation has been changed, but the mean is held constant at 50. Though the distributions appear to be different in shape, all three of the figures still represent normal curves.

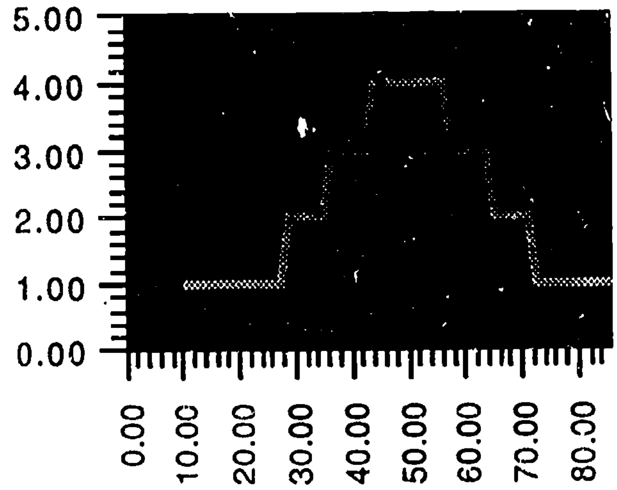

Figure 3a. The normal curve with a mean of 50 and SD of 15. (chosen by author)

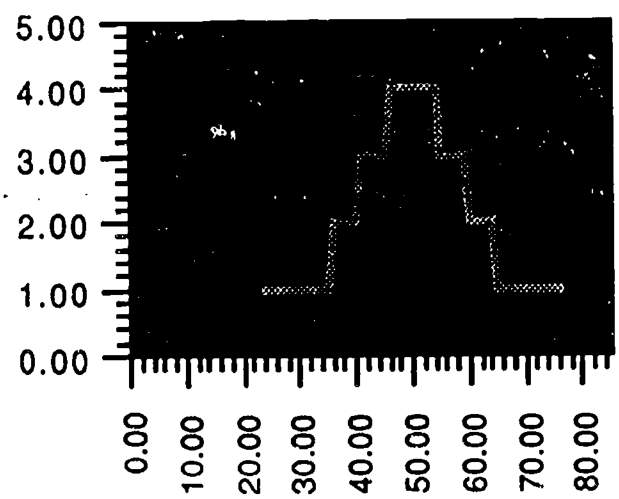

Figure 3b. The normal curve with a mean of 50 and SD of 10. (chosen by author)

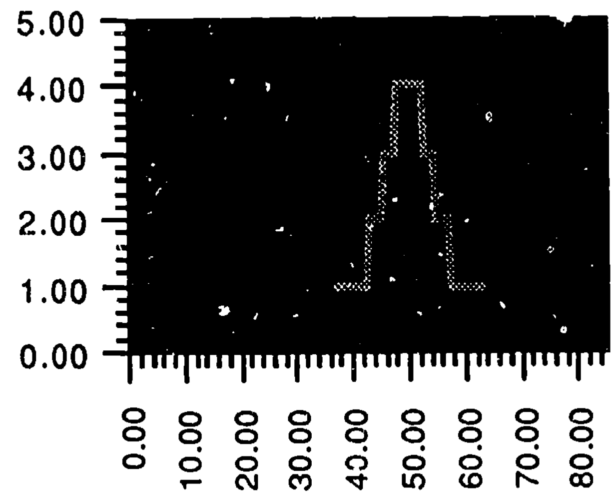

Figure 3c. The normal curve with a mean of 50 and SD of 5. (chosen by author)

We can also take our scores and see what happens when we change both the mean and the SD of the 100 scores from Table 1. Figure 4a. presents the same scores ($\bar{X}$=50 SD =10.05) presented in Figure 3b, for comparison purposes. In Figure 4b. the scores have been multiplied by 1.3. Notice that this spreads the scores out and increases the standard deviation, but the kurtosis and the skewness values remain the same because these values (skewness and kurtosis) are ratios to SD^3 or SD^4 and the criteria for a normal curve have still been met.

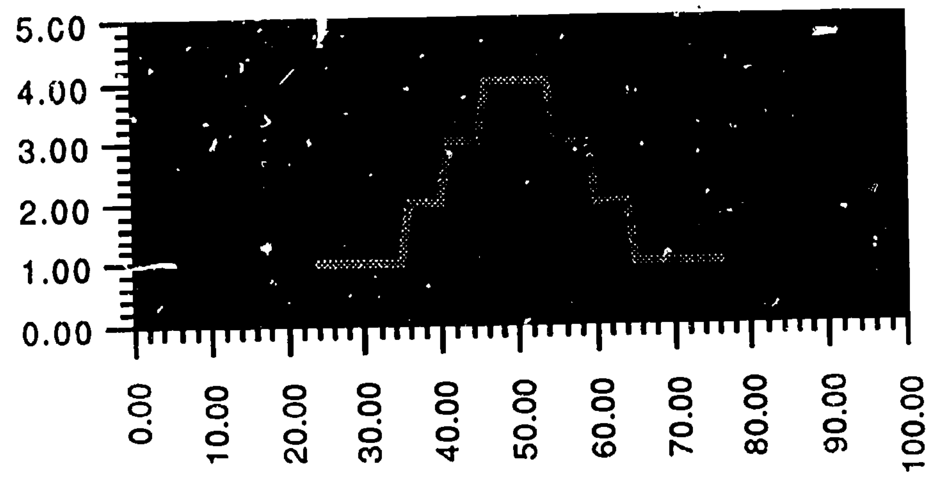

Figure 4a.The curve with the z scores multiplied by 1.00. **Multiply by 1** Mean: 50 SD (sum of z^2/n)=10.05 Skewness (sum of z^3/n)=0 Kurtosis(sum of z^4/n-3)=-.20

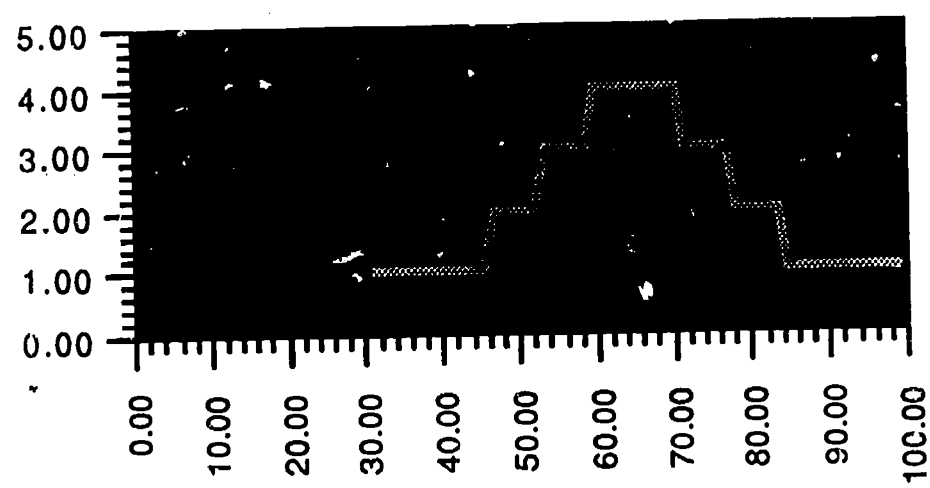

Figure 4b.The curve with the z scores multiplied by 1.30.**Multiply by 1.3** Mean: 65 SD (sum of z^2/n)=13.06 Skewness (sum of z^3/n)=0 Kurtosis(sum of z^4/n-3)=-.20

In Figures 4c. and 4d. a multiplicative constant less than 1 has been applied. The standard deviation goes down and the spread is narrower in both cases, but the kurtosis and the skewness values remain the same because, as emphasized before, the skewness and kurtosis values are computed by first converting scores to z form. Changes in the SD of the raw scores have no effect on SD of z, which is always 1.0.

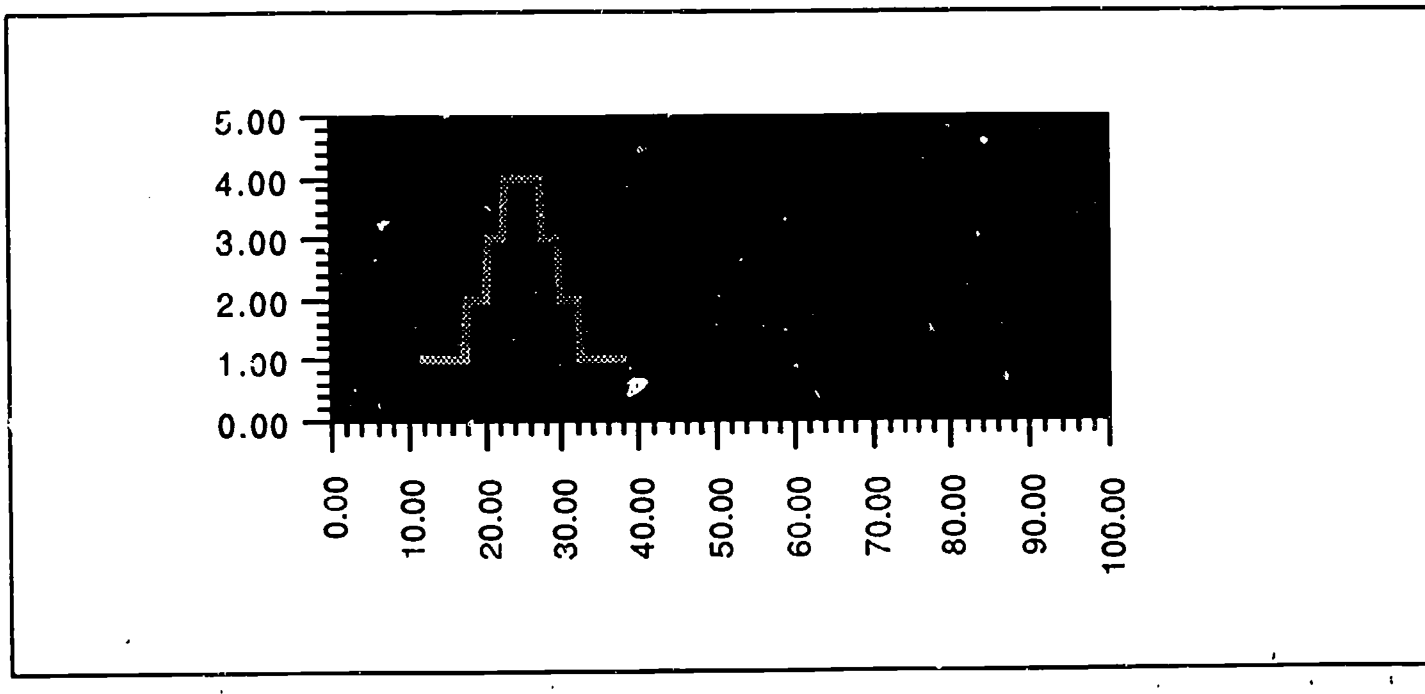

Figure 4c.The curve with the z scores multiplied by 0.50. **Multiply by 0.5** Mean: 25 SD (sum of z^2/n)=5.02 **Skewness (sum of z^3/n)=0 Kurtosis(sum of z^4/n-3)=-.20**

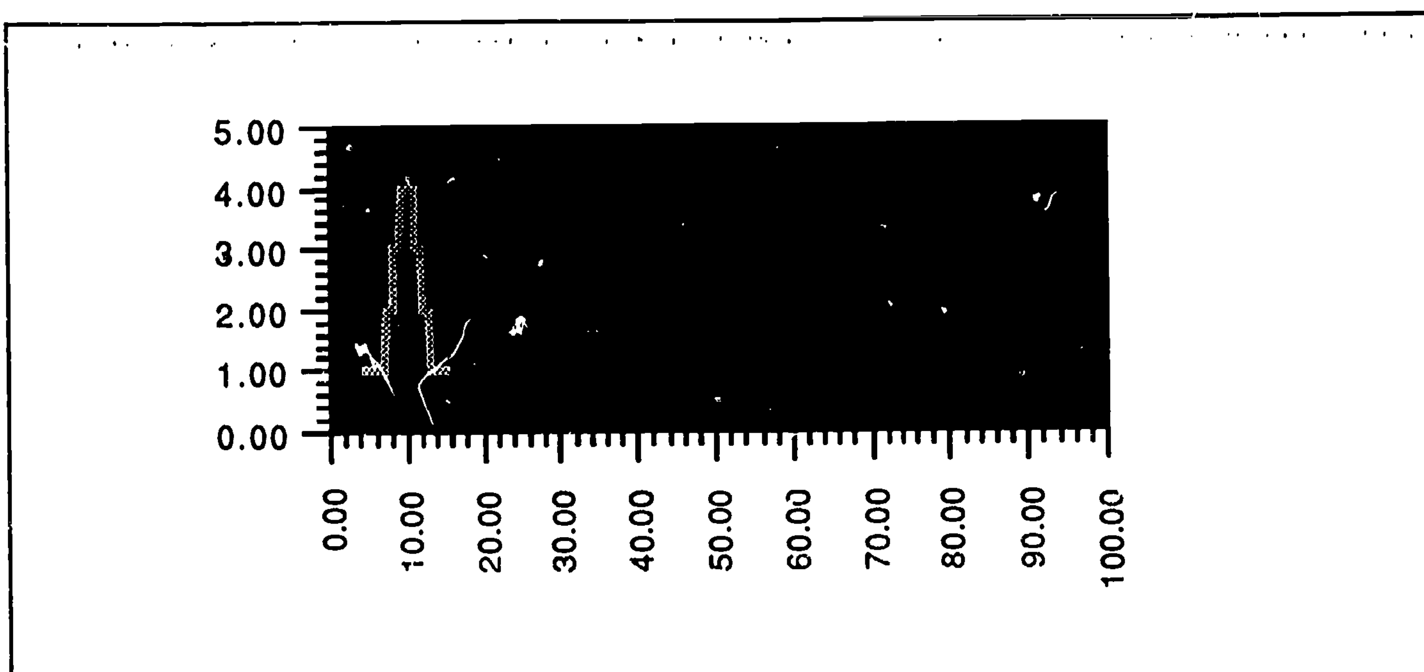

Figure 4d.The curve with the z scores multiplied by 0.20. **Multiply by 0.2** Mean: 10 SD(sum of z^2/n)=2.01 **Skewness (sum of z^3/n)=0 Kurtosis(sum of z^4/n-3)=-.20**

By converting the raw scores to z scores and applying shape statistics to them, we were able to see that some distributions that look tall were still normal and that some distributions that looked flat were also normal. So as a person "eyeballs" a distribution of scores to determine if it is normal, the spreadoutness of the scores in relation to the height must be considered. That is, "eyeballing" the distribution can lead to incorrect conclusions.

References

Downie, N.M. &Heath, R.W. (1965). *Basic statistical methods.* New York: Harper & Row.

Gronlund, N. E. (1971). *Measurement and evaluation in teaching.* New York: Macmillan Publishing Co., Inc.

Mills, F.C. (1955). *Statistical methods.* New York: Holt.

Moore, G.W. (1983). *Developing and evaluating educational research.* Boston: Little, Brown and Company.

Odell, C.W. (1957). *A first course in educational statistics.* Dubuque, Iowa: Wm. C. Brown, Co.

Newell, K.M. &Hancock, P.A. (1984). Forgotten moments: A note on skewness and kurtosis as influential factors in inferences extrapolated from response distributions. *Journal of Motor Behavior, 16,* 320-335.

Nitko, A. (1983). *Educational Tests and Measurements.* New York: Harcourt Brace Jovanovich, Inc.

Tate, M. W. (1965). *Statistics in education and psychology.* New York: Macmillan Co.

TABLE 1

Raw Scores	Range	Freq	Z Scores	Z^2	Z^3	Z^4
11.00	11.00	1	-2.60	6.7600	-17.5760	45.70
17.00	17.00	1	-2.20	4.8400	-10.6480	23.4256
20.00	20.00	1	-2.00	4.0000	-8.0000	16.0000
23.00	23.00	1	-1.80	3.2400	-5.8320	10.4976
24.50	24.50	1	-1.70	2.8900	-4.9130	8.3521
26.00	26.00	1	-1.60	2.5600	-4.0960	6.5536
27.50	27.50	1	-1.50	2.2500	-3.3750	5.0625
29.00	29.00	2	-1.40	1.9600	-2.7440	3.8416
29.00	30.50	2	-1.40	1.9600	-2.7440	3.8416
30.50	32.00	2	-1.30	1.6900	-2.1970	2.8561
30.50	33.50	2	-1.30	1.6900	-2.1970	2.8561
32.00	35.00	2	-1.20	1.4400	-1.7280	2.0736
32.00	36.50	3	-1.20	1.4400	-1.7280	2.0736
33.50	38.00	3	-1.10	1.2100	-1.3310	1.4641
33.50	39.50	3	-1.10	1.2100	-1.3310	1.4641
35.00	41.00	3	-1.00	1.0000	-1.0000	1.0000
35.00	42.50	3	-1.00	1.0000	-1.0000	1.0000
36.50	44.00	4	-0.90	0.8100	-0.7290	0.6561
36.50	45.50	4	-0.90	0.8100	-0.7290	0.6561
36.50	47.00	4	-0.90	0.8100	-0.7290	0.6561
38.00	48.50	4	-0.80	0.6400	-0.5120	0.4096
38.00	50.00	4	-0.80	0.6400	-0.5120	0.4096
38.00	51.50	4	-0.80	0.6400	-0.5120	0.4096
39.50	53.00	4	-0.70	0.4900	-0.3430	0.2401
39.50	54.50	4	-0.70	0.4900	-0.3430	0.2401
39.50	56.00	4	-0.70	0.4900	-0.3430	0.2401
41.00	57.50	3	-0.60	0.3600	-0.2160	0.1296
41.00	59.00	3	-0.60	0.3600	-0.2160	0.1296
41.00	60.50	3	-0.60	0.3600	-0.2160	0.1296
42.50	62.00	3	-0.50	0.2500	-0.1250	0.0625
42.50	63.50	3	-0.50	0.2500	-0.1250	0.0625
42.50	65.00	2	-0.50	0.2500	-0.1250	0.0625
44.00	66.50	2	-0.40	0.1600	-0.0640	0.0256
44.00	68.00	2	-0.40	0.1600	-0.0640	0.0256
44.00	69.50	2	-0.40	0.1600	-0.0640	0.0256
44.00	71.00	2	-0.40	0.1600	-0.0640	0.0256
45.50	72.50	1	-0.30	0.0900	-0.0270	0.0081
45.50	74.00	1	-0.30	0.0900	-0.0270	0.0081
45.50	75.50	1	-0.30	0.0900	-0.0270	0.0081
45.50	77.00	1	-0.30	0.0900	-0.0270	0.0081
47.00	80.00	1	-0.20	0.0400	-0.0080	0.0016
47.00	83.00	1	-0.20	0.0400	-0.0080	0.0016
47.00	89.00	1	-0.20	0.0400	-0.0080	0.0016
47.00		0	-0.20	0.0400	-0.0080	0.0016
48.50			-0.10	0.0100	-0.0010	0.0001
48.50			-0.10	0.0100	-0.0010	0.0001
48.50			-0.10	0.0100	-0.0010	0.0001
48.50			-0.10	0.0100	-0.0010	0.0001
50.00			0.00	0.0000	0.0000	0.0000
50.00			0.00	0.0000	0.0000	0.0000
50.00			0.00	0.0000	0.0000	0.0000
50.00			0.00	0.0000	0.0000	0.0000
51.50			0.10	0.0100	0.0010	0.0001
51.50			0.10	0.0100	0.0010	0.0001
51.50			0.10	0.0100	0.0010	0.0001
51.50			0.10	0.0100	0.0010	0.0001
53.00			0.20	0.0400	0.0080	0.0016
53.00			0.20	0.0400	0.0080	0.0016
53.00			0.20	0.0400	0.0080	0.0016
53.00			0.20	0.0400	0.0080	0.0016
54.50			0.30	0.0900	0.0270	0.0081
54.50			0.30	0.0900	0.0270	0.0081
54.50			0.30	0.0900	0.0270	0.0081
54.50			0.30	0.0900	0.0270	0.0081
56.00			0.40	0.1600	0.0640	0.0256
56.00			0.40	0.1600	0.0640	0.0256
56.00			0.40	0.1600	0.0640	0.0256
56.00			0.40	0.1600	0.0640	0.0256
57.50			0.50	0.2500	0.1250	0.0625
57.50			0.50	0.2500	0.1250	0.0625
57.50			0.50	0.2500	0.1250	0.0625
59.00			0.60	0.3600	0.2160	0.1296
59.00			0.60	0.3600	0.2160	0.1296
59.00			0.60	0.3600	0.2160	0.1296
60.50			0.70	0.4900	0.3430	0.2401
60.50			0.70	0.4900	0.3430	0.2401
60.50			0.70	0.4900	0.3430	0.2401
62.00			0.80	0.6400	0.5120	0.4096
62.00			0.80	0.6400	0.5120	0.4096
62.00			0.80	0.6400	0.5120	0.4096
63.50			0.90	0.8100	0.7290	0.6561
63.50			0.90	0.8100	0.7290	0.6561
63.50			0.90	0.8100	0.7290	0.6561
65.00			1.00	1.0000	1.0000	1.0000
65.00			1.00	1.0000	1.0000	1.0000
66.50			1.10	1.2100	1.3310	1.4641
66.50			1.10	1.2100	1.3310	1.4641
68.00			1.20	1.4400	1.7280	2.0736
68.00			1.20	1.4400	1.7280	2.0736
69.50			1.30	1.6900	2.1970	2.8561
69.50			1.30	1.6900	2.1970	2.8561
71.00			1.40	1.9600	2.7440	3.8416
71.00			1.40	1.9600	2.7440	3.8416
72.50			1.50	2.2500	3.3750	5.0625
74.00			1.60	2.5600	4.0960	6.5536
75.50			1.70	2.8900	4.9130	8.3521
77.00			1.80	3.2400	5.8320	10.4976
80.00			2.00	4.0000	8.0000	16.0000
83.00			2.20	4.8400	10.6480	23.4256
89.00			2.60	6.7600	17.5760	45.6976
		SUM	0.00	99.9800	0.0000	285.3902
		SUM/N	0.00	0.9998	0.0000	2.8539