CHAPTER 9

DESCRIPTIVE STATISTICS FOR UNIVARIATE DISTRIBUTIONS

The purpose of statistics is to assemble, describe, and infer important numerical characteristics of data sets. In this chapter a concrete data set, formerly collected and analyzed by this writer, will be subject to statistical scrutiny.

The focus of attention is on the dependent variable, in this case the salaries of American League baseball "starters." When explanatory or predictor variables are used, they are called independent variables. For the present we are concerned with the statistical description of a single ("univariate") variable. The data originally appeared in the popular press and listed annual salaries of a national organization.

The Frequency Distribution.

Although statistical observations are frequently punched on IBM cards and "run through" the computer (see chapter 7), an appreciation of statistical processing is more easily understood if we assume "hand processing." In this manner the logical steps in the analysis can be more easily grasped. Table 9.1 contains the individual salaries (in $1000's) of 138 employees. A glance at this ungrouped data arrangement probably leaves you befuddled. It would not be too much of an exaggeration to say that numerical chaos reigns. The initial step in dealing with raw data is to bring some semblance of order to them (note that the term data is plural; the singular of data is datum). Usually the data can be reduced through the construction of a frequency distribution, an arrangement of data showing the frequency with which different values of the variable occur.

(Table 9.1 here)

TABLE 9.1

SALARIES (in thousands of dollars) EARNED
BY 138 EMPLOYEES

40	40	190	40	75	45	30	150	25	80
130	150	115	55	60	100	85	145	30	140
28	195	50	85	165	95	50	150	100	25
140	200	70	75	75	80	60	35	40	19
40	65	60	40	50	150	30	20	75	30
60	150	30	19	150	40	250	20	50	27.5
55	40	40	90	50	60	200	20	80	25
160	50	160	45	135	45	60	75	85	19
19	80	35	50	175	32.5	135	20	19	19
180	200	125	45	19	45	135	50	200	19
160	60	35	90	140	40	120	100	100	30
175	50	70	40	40	165	100	40	100	90
80	200	75	75	120	30	250	30	50	
125	175	35	100	70	25	95	19	110	

These salaries are called "raw scores" or "ungrouped data."

TABLE 9.2

RANK DISTRIBUTION (Array) OF SALARIES SHOWN IN TABLE 1

250	165	135	100	80	60	50	40	30	20
250	160	135	100	80	60	50	40	30	20
200	160	135	100	75	60	50	40	30	20
200	160	130	95	75	60	50	40	30	19
200	150	125	95	75	60	45	40	30	19
200	150	125	90	75	60	45	40	30	19
200	150	120	90	75	55	45	40	30	19
195	150	120	90	75	55	45	40	28	19
190	150	115	85	75	50	45	35	27.5	19
180	150	110	85	70	50	40	35	25	19
175	145	100	85	70	50	40	35	25	19
175	140	100	80	70	50	40	35	25	19
175	140	100	80	65	50	40	32.5	25	
165	140	100	80	60	50	40	30	20	

To begin the condensation process, it is helpful to order the data from top to bottom or from the highest score(s) to the lowest score(s). This procedure culminates in what is termed a rank order distribution or an array such as appears in Table 9.2. Although we still have as many score values as was originally tallied, observation of the rank distribution (Table 9.2) already makes it easier to grasp some important statistical features of the data. Some of the meaningful statistical properties that can easily be identified from the rank distribution are: 1) the largest score, sometimes called the maximum, $250,000 in this distribution, can be noted; 2) the smallest score, sometimes called the minimum, $19,000 in this distribution, can be pinpointed; 3) the range (a measure of dispersion) can be determined by subtracting the smallest score from the largest score and adding one (e.g., the range for the salary data is $232,000; 4) the score which occurred most frequently, called the mode (a measure of central tendency), can be identified by counting the score value that was most frequent in occurrence. Here the mode is $40,000. In short, the purpose of constructing frequency distributions is to make the data more meaningful, manageable, and intelligible. The rank distribution is but one of several types of data distributions statisticians employ. Other commonly constructed ones are the ungrouped frequency distribution (Table 9.3) and the grouped frequency distribution (Table 9.4).

(Tables 9.2 and 9.3 here)

A Grouped Distribution.

Another way we can reduce the 138 scores even further is to collapse the scores into a range of values and indicate the frequency with which the scores in the various groupings occur. This is exactly what a grouped frequency distribution is, a reduction of the original raw scores into a range of score categories and then denoting how many scores take on the values encompassed by the exact limits of the score categories. To accomplish this feat, stat-

TABLE 9.3

UNGROUPED FREQUENCY DISTRIBUTION OF SALARIES FROM TABLE 1
WITH AS MANY CLASSES AS SALARY VALUES

Salary	f	Salary	f
250	2	100	7
200	5	95	2
195	1	90	3
190	1	85	3
180	1	80	5
175	3	75	7
165	2	70	3
160	3	65	1
150	6	60	7
145	1	55	2
140	3	50	10
135	3	45	5
130	1	40	13
125	3	35	4
120	2	32.5	1
115	1	30	8
110	1	28	1
		27.5	1
		25	4
		20	4
		19	9

TABLE 9.4

Grouped Frequency DISTRIBUTION OF SALARIES FROM TABLE 1

Class Interval (i)	f	%	F	cum%	Midpoints
244–268	2	1.45	138	100.01	256
219–243	0	0.00	136	98.56	231
194–218	6	4.35	136	98.56	206
169–193	5	3.62	130	94.21	181
144–168	12	8.70	125	90.59	156
119–143	11	7.97	113	81.89	131
94–118	11	7.97	102	73.92	106
69–93	21	15.22	91	65.95	81
44–68	25	18.12	70	50.73	56
19–43	45	32.61	45	32.61	31
TOTAL	138	100.01%			

isticians have advanced several guidelines that will be put forth in the form of questions: 1) How many score categories, or class intervals as they are often called, is desirable? The answer is that it depends upon the range of scores, and the meaningfulness of sub-classifying scores. However, it often turns out that somewhere between 10 and 20 class intervals makes for a meaningful tabular display of data. In the present case I decided upon 10 score categories. 2) How wide (how many different scores) should the class intervals be? This query is a function of the answer to question 1. In fact, both questions can be answered by employing the following formula:

$$\text{class interval width} = \frac{\text{range of scores}}{\text{number of class intervals}}$$

The quotient provides the appropriate width of the class intervals. Hence, for our data the range of 232 is divided by 10 (desired number of score categories) and the quotient tells us the approximate class interval width. That is,

$$\frac{232}{10} = 23$$

There is nothing magical or final about the number of class intervals or the width of the class intervals other than permitting a meaningful ordering of the data. 3) Having answered items 1 and 2 the final question becomes, where do we begin the class intervals and where do we end them? That is, what score do we commence with and with what score do we terminate? Frequently, it is desirable to begin with a score one below the smallest datum collected or the smallest datum itself and then proceed in multiples of the class interval size. For example, the smallest salary score is 19. We will proceed in multiples of 25's (rather than 23's) uhtil the highest score is contained within a class interval. Table 9.4 is a grouped frequency distribution of 10 class intervals all of which are 25 units wide.[2] It is highly recommended that all class intervals be of uniform size since statistical problems may occur when they are of uneven widths.

(Table 9.4 here)

Notice how more manageable and comprehensible the data are. Some of the key statistical features that can readily be seen are the modal (most frequently occurring class interval) class interval which ranges from 19-43. One cautionary note is in order. Precision is sacrificed when data are grouped. This loss of accuracy is termed grouping error (if you did not have access to Table 9.1 you would not know exactly where the score values lie). However, this loss is most often compensated for by facilitating statistical computations as well as the construction of graphs (which typically necessitate a grouped distribution) to pictorially represent the data.

The Important Statistical Properties of Frequency Distributions

Just as in any field of endeavor, to understand a phenomenon it is necessary to know its nature. The same holds true for collections of data such as that with which we are working. To better comprehend the nature of univariate frequency distributions, statisticians ask three salient questions: 1) What is the form or shape of the distribution? 2) What is the central tendency (or location) of the distribution? 3) What is the variation in the distribution? These three queries capture the most important statistical concepts that apply to univariate distributions. For each of these concepts--form, central tendency, and variation--both conceptual and operational definitions (computational formulae) will be provided. Computationally, each of these properties results in a single index number that is then interpreted relative to the original distribution.

The Form of a Distribution: From Numbers to Pictures

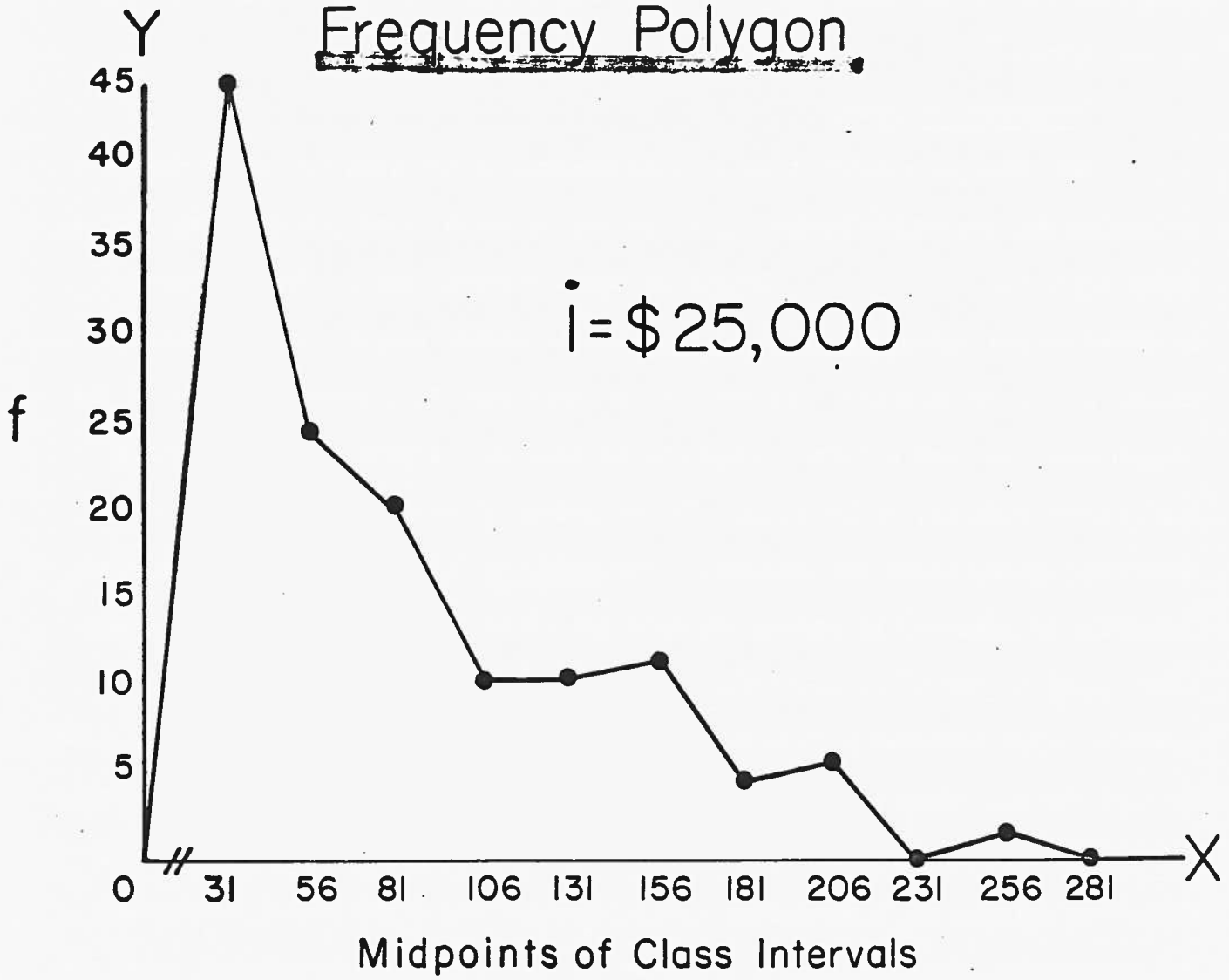
The form or shape of a distribution can be readily grasped by converting the numerical data into a pictorial display. That is, by constructing a graph, a geometric image of a data set, the shape of the salary distribution can be easily and clearly conveyed. There exist several different graphic techniques but

three of the most important ones--polygon, histogram, and ogive--will enable us to capture the statistical feature of the distribution known as form. Virtually all graphic techniques use one quadrant, the first, of the Cartesian coordinate system.[3] Hence, two axes form a perpendicular and intersect at a point called the origin. The vertical axis is called the ordinate and contains a listing of frequencies (as in the polygon and histogram) or cumulative frequencies (as in the ogive) while the horizontal axis is called the abscissa and contains a listing of the midpoints of class intervals (as in the polygon), exact limits of class intervals (as in the histogram), or upper exact limits of class intervals (as in the ogive).[4]

Frequency Polygon.

We commence construction of this graphic device (and the other two) by laying out two perpendicular lines in such a manner that the vertical axis is about 3/4's the length of the horizontal axis. This convention is known as the three-quarter high rule and is widely adopted in statistical circles. The frequencies, labeled with a lower case "f", proceed from 0 at the origin until we have a large enough one so that the largest frequency, 45 in the present example, is included. The midpoints of each class interval are uniformly spaced along the abscissa. The midpoint, the exact center of a class interval, is computed by adding one-half the class interval width to the lower exact limit of the class interval.[5] A dot is placed at the intersection of a specific class interval's midpoint and frequency of occurrence. This dot placing is done for all class intervals and finally the dots are connected by a series of straight lines from one adjacent dot to the next. To bring closure to the frequency polygon it is suggested that a midpoint below the smallest midpoint and a midpoint above the largest midpoint be addended to the horizontal axis, and connected with the adjacent dots to bring the graph to a close. Figure 9.1 represents the completed frequency polygon for the scores in Table 9.4.

Figure 9.1

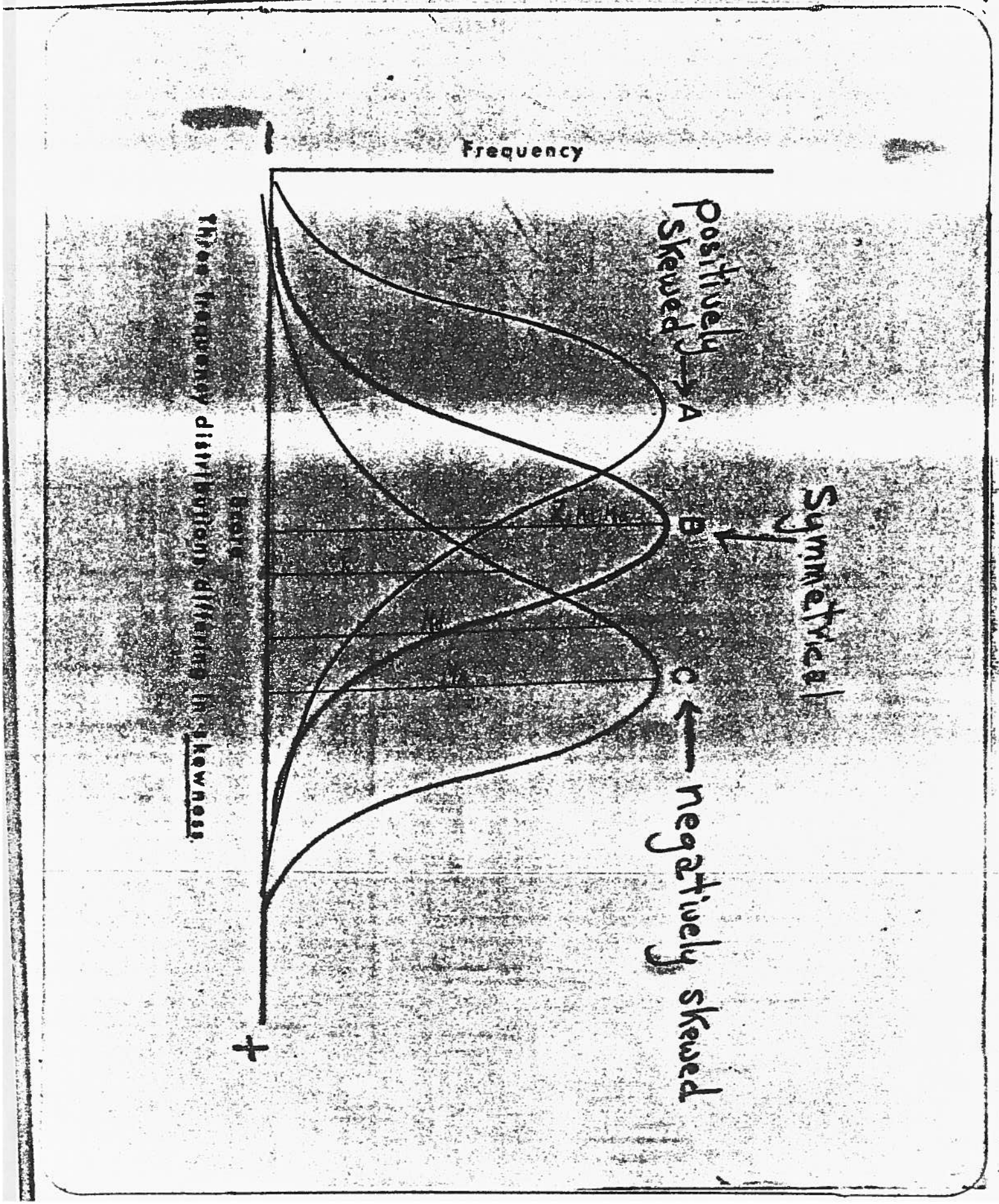

Frequency Polygon
Y
45
40
35
30
f
25
20
15
10
5
0
i = $25,000
31
56
81
106
131
156
181
206
231
256
281
X
Midpoints of Class Intervals

(Figure 9.1 here)

Any graphic procedure, like the frequency polygon, is generally not an end in itself. Instead, it conveys to the researcher a sense of an important statistical property of the distribution. What does the frequency polygon in Figure 9.1 tell us? In fact, the polygon provides us with a parsimonious view, albeit not as precise as we will eventually come to see, of the properties of symmetry (or skewness), kurtosis, central tendency, and dispersion.

Skew. If a curve is skewed it tells us that the scores are not evenly distributed on both sides of the exact center. If you were to fold the curve together at the highest point (the mode of the curve) the two sides would not be identical. On the other hand, if one side of the curve was a mirror image of the other the distribution would be termed symmetrical (see Figure 9.2B). Substantively, a skewed curve indicates scores piling up at one tail and a sparser concentration at the other tail. Look at Figure 9.1. Notice the tendency of scores to concentrate at the left end of the polygon and their tapering off at the right end. Consequently, this curve is skewed (to the right) and is referred to as a positively skewed curve (note that the tail moves to the right side of the abscissa and the mathmatician calls the left side negative and the right side positive; see also Figure 9.2A).[6] In terms of the salary scores this graph tells us that the scores tend to be relatively low. The conceptual opposite of a positively skewed curve is a negatively skewed curve as depicted in Figure 9.2C. This is said to be a left skew and would indicate that the tendency is for scores to concentrate at the upper end of the score categories. Finally, if the distribution were completely symmetrical, which means no skew exists, the "smooth curve" frequency polygon would look like Figure 9.2B. We have provided a verbal discussion and graphic representation of this property and later on will provide a mathematical description of this same feature.

Figure 9.2

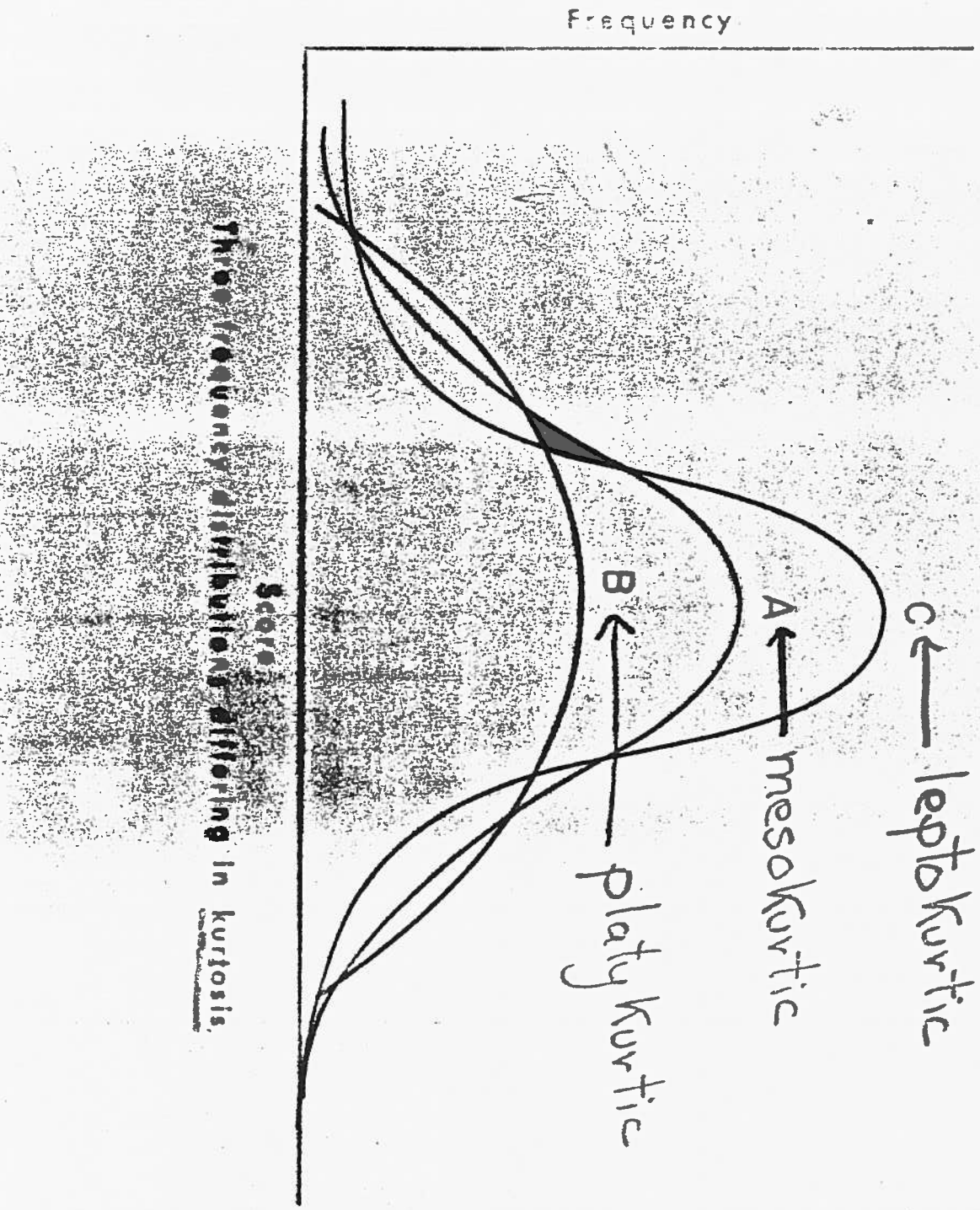

Frequency
Positively Skewed
A
B
Symmetrical
C
negatively skewed
Three frequency distributions differing in skewness
+

(Figure 9.2 here)

Kurtosis. Kutosis refers to how peaked or flat a distribution is relative to the normal curve which is used as the standard for judgement. The normal curve, the symmetrical bell-shaped curve, is referred to as mesokurtic (see figure 9.3A). If a curve is more peaked than this one it is said to be leptokurtic; (see Figure 9.3C) if more flat or depressed it is said to be platykurtic (see Figure 9.3B). Look at Figure9.3 for visual displays of three curves differing in kurtosis. Later on a statistical index for kurtosis will be computed and it will provide us with more precise mathematical description of kurtosis.

(Figure 9.3 here)

Figure 9.3


Frequency
C
leptokurtic
A
mesokurtic
B
platykurtic
Scores
The frequency distributions differing in kurtosis

Central Tendency. This characteristic of a distribution is commonly called the average, a most ambiguous term because several different measures of location actually exist. Central tendency is the score or scores around which the distribution tends to cluster. The highest point on the curve is the location measure called the mode, the 19-43 class interval in the present case. The other two common measures of central tendency--the arithmetic mean and the median--can not be exactly determined by inspection of the graph. However, it is possible to determine the relative position of these three indices from examining the graph. More specifically, in a symmetrical distribution the three statistics will coincide; in a positively skewed curve the mean will be pulled in the direction of the tail with the median inbetween the mean and the mode (which is always at the highest point of the curve); in a negatively skewed curve the mean will be farthest to the left (in the direction of the tail) and the median inbetween the mean and mode. See the relative positions of mean, median, and mode in Figure 9.4 for symmetrical, positively and negatively skewed curves.

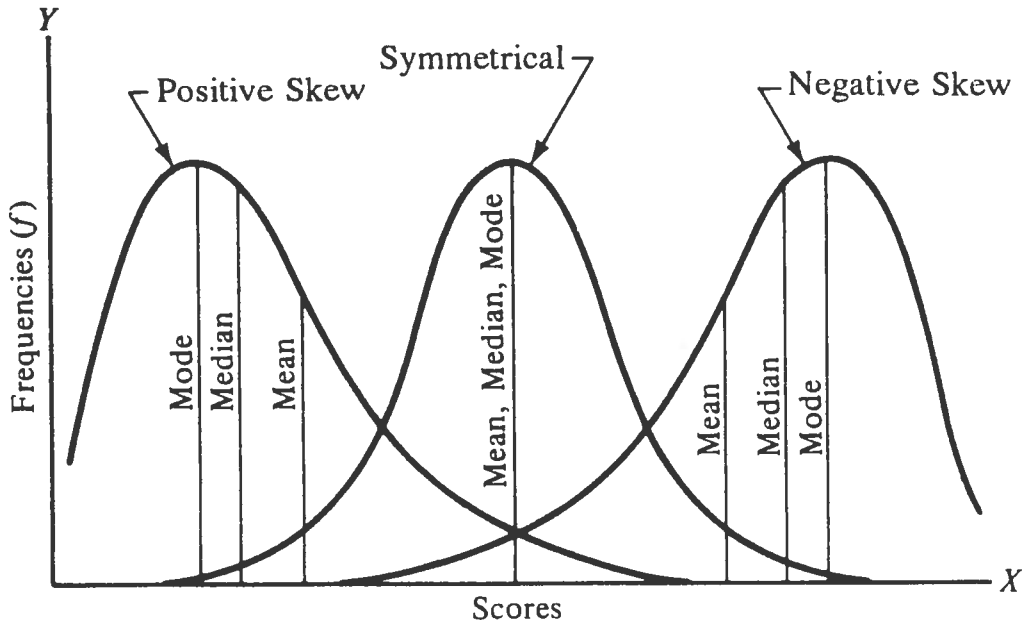
(Figure 9.4 here)

By inspecting Figure 9.1 we observe the mode to be 31 (the midpoint of the most frequently occurring class interval) and the mean to be larger than the median because the curve is skewed to the right. This intuitive approach will be corroborated shortly when indices of central tendency will be calculated.

Variability. The scores are not all concentrated at the central tendency but disperse or scatter from the distribution's center point. Variability indices measure the degree of spread among the scores in the distribution. One index of variation already discussed is the range, the distance between the largest and smallest scores in the distribution plus one. The range provides us with an indication of the variability of scores. In the frequency polygon the variation a la the range is 232. Like central tendency, there are a variety of statistical indices of variability and these will be shortly discussed and computed.

FIGURE 9.4

The Relationships among the Mean, Median, and Mode in Skewed and Symmetrical Distributions

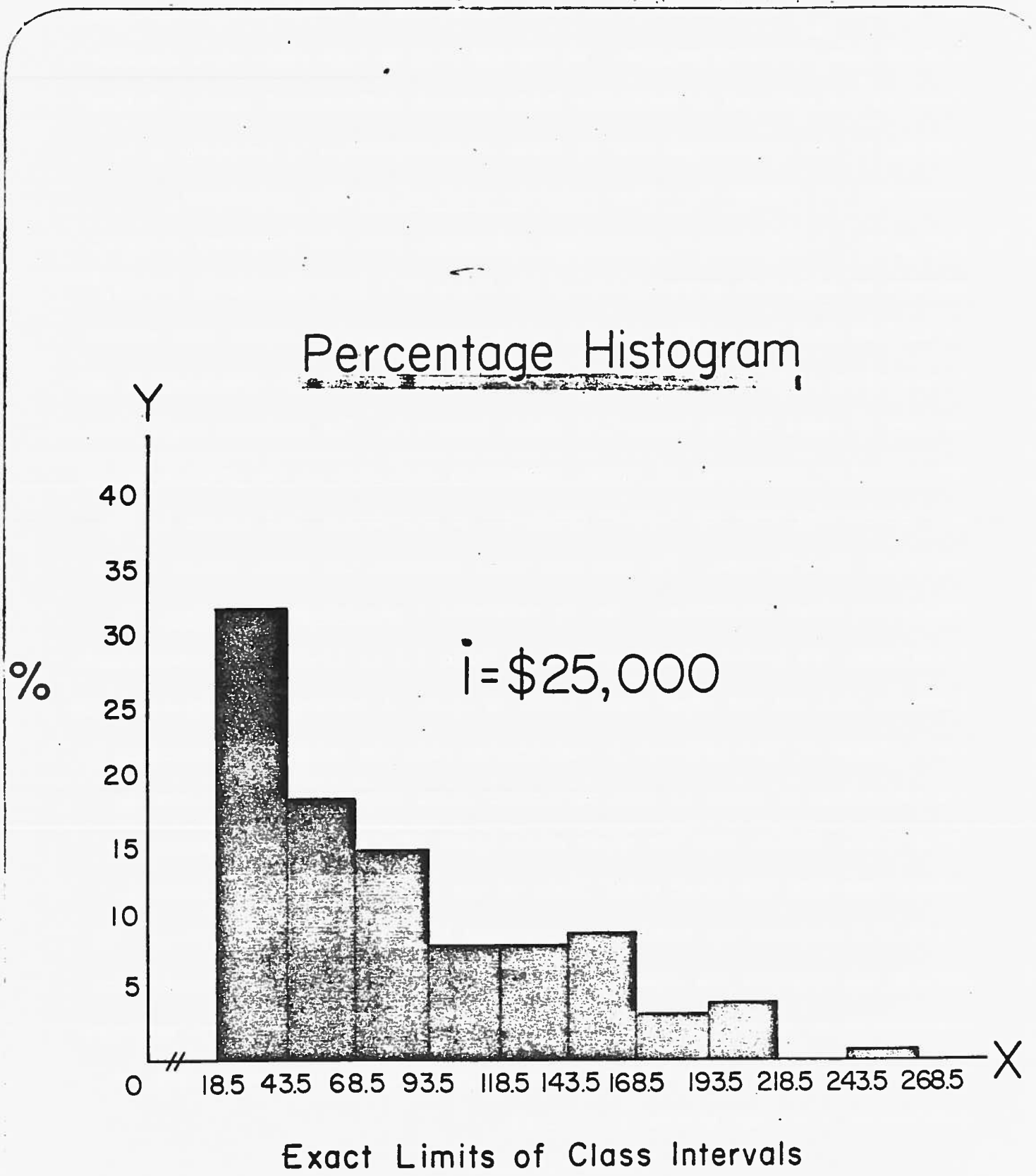


These are the four properties of a distribution to which statisticians are attuned. A rough estimate of their values can be obtained by inspecting a graph like the frequency polygon. We can and will be more exact when specific formulae are employed to determine their precise mathematical value. Before turning to the operational side of these characteristics it behooves us to construct and discuss two other commonly utilized graphic techniques.

Histogram. A frequency or percentage histogram, like the frequency polygon, conveys an intuitive appreciation of the shape of a distribution of scores.[7] It differs from the polygon insofar as bars or rectangles rather than points connected by straight lines are constructed according to the frequency or percentage of cases in the respective class intervals. The ordinate scale contains the number or percentage of cases, commencing with zero at the graph's origin and terminating with the largest frequency or percentage that exists in the empirical distribution. The abscissa scale contains the exact limits of the respective class intervals. Let us construct a percentage histogram. To illustrate, the lowest class interval in Table 9.4 is 19-43. The lower exact limit is 18.5 and the upper exact limit is 43.5. A rectangle encompassing the exact limits of that interval is constructed with the height corresponding to the percentage of cases in that particular class interval (see % column in Table 9.4). Notice that the upper exact limit of a given score category is coterminous with the lower exact limit of the next higher class interval. Figure 9.5 contains the completed percentage histogram for these scores. It conveys and contains the same information as the polygon (i.e., provides the reader with an appreciation of skewness, kurtosis, central tendency, and dispersion). Ordinarily not both of these graphic representations of data are constructed since one generally suffices.

(Figure 9.5 here)

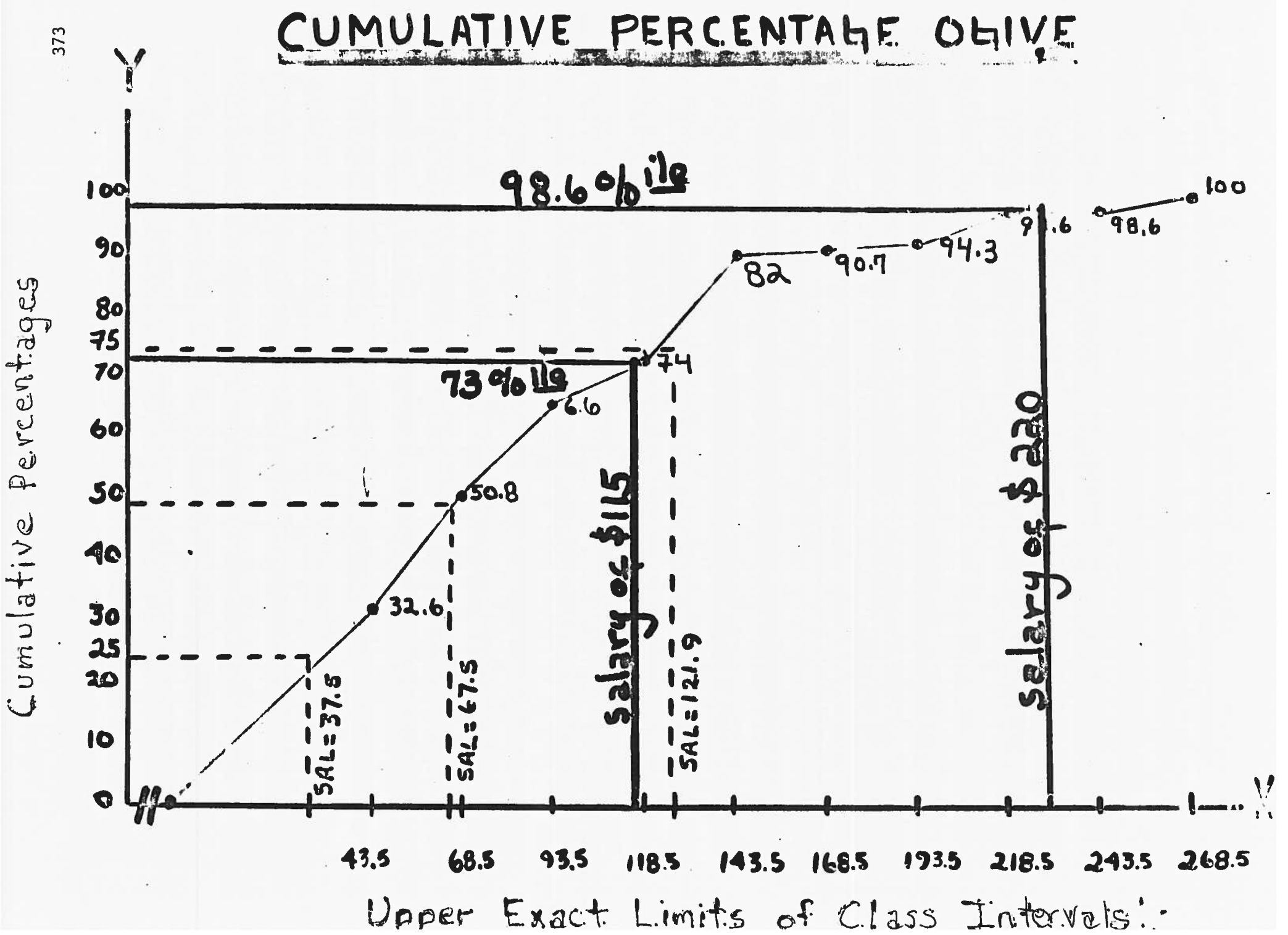
Figure 9.5


Percentage Histogram
Y
40
35
30
%
25
20
15
10
5
0
i=$25,000
18.5 43.5 68.5 93.5 118.5 143.5 168.5 193.5 218.5 243.5 268.5
X
Exact Limits of Class Intervals

Ogive. A cumulative frequency or cumulative percentage polygon, or ogive, conveys a different picture of the data distribution than the other two. It enables us to determine the number (cumulative frequency) or percentage (cumulative percentage) of cases above or below specified score categories in the distribution. By inspecting it we can tell how many players scored above and below certain points. To construct an ogive from the data in Table 9.4 requires some additional steps. Since the graph reflects cumulative percentages, we must construct a cumulative percentage column by successively adding the individual class interval percentages (see cumulative % column in Table 9.4). Cumulative percentages are denoted by "cum%." In the fifth column of Table 9.4 we indicate the cumulative percentages. To obtain them we begin the cumulation process with the smallest class interval. It has a percentage of 32.6 and we locate 32.6 in the column labeled "cum%." Then we add to that cum% the simple percentage in the above adjacent column. The 44-68 class interval has a percentage of 18.2 and that 18.2 is added to the former cumulative percentage to make 50.8 and 50.8 is placed in the cum% column across from 44-68. We proceed in this manner until the last simple percentage is added to the previous cumulated percentage. As a double check the cum% at the top of the table must correspond to 100% (although "rounding error" may yield a percentage slightly higher or lower than 100). The top cum% in Table 9.4 corresponds to 100%.

To construct a percentage ogive, cumulative percentages are located along the ordinate from 0 to 100 and the upper exact limits of the respective class intervals are located on the abscissa. A point is placed above the upper exact limit of each class interval in accordance with the cumulative percentage of that score category. The dots are then connected by a series of straight lines. Let us inspect the completed cumulative frequency polygon in Figure 9.6. It tells us such information as 51% of all salaries fall below $68,500; 74% of all salaries fall below $118,500; nearly 94½% of all salaries are less than $193,500; etc.

Figure 9.6


CUMULATIVE PERCENTAGE OGIVE
Y
100
90
80
75
70
60
50
40
30
25
20
10
0
98.6% ile
73% ile
82
90.7
94.3
98.6
100
74
66
50.8
32.6
SAL=37.5
SAL=67.5
SAL=121.9
Salary of $115
Salary of $220
Cumulative Percentages
43.5
68.5
93.5
118.5
143.5
168.5
193.5
218.5
243.5
268.5
Upper Exact Limits of Class Intervals

(Figure 9.6 here)

In summary, the raw scores (original data) appearing in Table 9.4 have been transformed into pictures, called graphs. There are many different kinds of graphic techniques but among the most common are those discussed and constructed here, the frequency polygon, percentage histogram, and cumulative percentage polygon (ogive). In general, graphs--such as the first two--convey to us the shape or form of the distribution of scores. The form of a distribution is one very important statistical property. Usually form means skew and kurtosis but the picture of the data also allows us to make inferences about both central tendency and dispersion. These, then, are the important properties of frequency distributions in which statisticians are interested: 1) symmetry (or skew), 2) kurtosis, 3) location, and 4) variation. In the next section mathematical formulae for computing statistical indices of these properties will be presented.

In this section we have enumerated the salient statistical properties of data distributions and have conveyed these characteristics through the construction of graphs to represent the data. Both the construction of frequency distributions and the graphic display of them have served the function of data reduction. The data have been condensed and distilled so that, as a whole, they become more comprehensible. For each of the properties of data sets--form, location, variation--there exist computational formulae which provide an even more succinct numerical description of the collected observations. In the next section we will focus upon the computation of what are called statistical indices of form (divided into skewness and kurtosis measures), central tendency, and variability.

Statistical Indices for Describing Form, Central Tendency and Dispersion.

The form or shape of a data set is reflected in two different statistical concepts: (1) symmetry (or skew) provides an indication of the symmetrical or asymmetrical nature of a distribution; and (2) kurtosis describes the peakedness

or flatness of a curve using the symmetrical mesokurtic normal curve as a criterion. Since an intuitive appreciation of these characteristics was developed in the last section, here mathematical formulae will serve as a cross-check on the verbal presentation as well as providing a more precise and exact indicator of the property in question.

Skewness. The index of skewness is called beta-one and is symbolized B_1. Computationally, it is determined by dividing the square of the third moment by the cube of the second moment or, in notation form:[8]

$$\text{beta-one } (B_1) = m^2{}_3/m^3{}_2$$

$$\text{where: } m_3 = \Sigma X^3 - [3\Sigma X \Sigma X^2/N] + [2(\Sigma X)^3/N^2]$$

$$m^2 = \frac{\Sigma X^2 - (\Sigma X)^2/N}{N}$$

Using this formula a B_1 value of .942 is produced. The skewness index runs from 0 (representing symmetry) to minus (representing negative skew) to plus (representing positive skew) values. Diagrammatically:

- values	0	+ values
negative skew	symmetrical	positive skew

To interpret the skewness index several guidelines are in order. If a data set were perfectly symmetrical there would be no skew to the curve by definition, and both the third moment and beta-one would be zero. On the other hand, if the curve is negatively skewed, B_1 will be negative; and if the curve is positively skewed B_1 will be positive.[9] In short, if the sign is positive the skew is to the right whereas if it is negative the skew is to the left. Our calculation of beta-one for the scores in Table 9.1 produced a value of +.942. The actual graphing of the data, albeit grouped, in Figure 9.1 displayed a concentration of scores at the low end of the scale and tapering off at the upper end.

Consequently, the graphic representation and the statistical index of skew are consistent with the index being more exact than the verbal analog.

Kurtosis. The index of kurtosis is called beta-two and is symbolized B_2. Computationally, it is determined by dividing the fourth moment by the square of the second moment or, in notation form:[10]

$$\text{beta-two } (B_2) = m_4/m^2_2$$

$$\text{where: } m_4 = \frac{\Sigma X^4 - [4\Sigma X \Sigma X^3/N] + 6(\Sigma X)^2 \Sigma X^2/N^2 - [3(\Sigma X)^4/N^3]}{N}$$

$$m_2 = \frac{\Sigma X^2 - (\Sigma X)^2/N}{N}$$

Using this formula a B_2 value of -.042 is produced.

To interpret the kurtotic property we can imagine a continuum with a middle point of three (which represents a symmetrical, mesokurtic, distribution). Values below three represent a playtykurtic (flatter than normal curve) distribution while those above three represent a leptokurtic (more peaked than normal curve) distribution. Hence, in diagrammatic form:

values<3	3	values>3
platykurtic	mesokurtic	leptokurtic

The frequency polygon in Figure 9.1 was judged to be platykurtic. The actual numerical index of kurtosis just computed produced a value of -.042 (slightly platykurtic) which is consistent with our inspection of the graphic representation of the data.

Central Tendency. Not one but several indices of central tendency are available. The most popular measures are the arithmetic mean (symbolized $\bar{X}$), the median (symbolized Md or Mdn), and the mode (symbolized Mo). The choice among these statistics is guided by three features: 1) the level of measurement of the data. Strictly speaking, the mean is generally only appropriate for interval-ratio data; the median for ordinal (and interval level data); and the mode is appropriate for

←——— data at any level of measurement; 2) the characteristics, particulary the form, of the data. While the mean is geared to extract relevant statistical information from interval/ratio level variables, regardless of the measurement level, when excessive skew exists the median is more appropriate. This is because the median is a positional measure, and is not affected by extreme scores (which is what produces the skew in the first place); and 3) the purpose of the statistic.

Arithmetic Mean. To be precise this location measure is called the "arithmetic" mean because other less commonly used means (e.g., harmonic, geometric, and countra-harmonic means) exist. For the raw scores in Table 9.1 the arithmetic mean is computed using the following computational formula:

$$\bar{X} = \Sigma X_i / N$$

where: Σ = the summation operator directing one to add what follows

X_i = a generic mathematical noun referring to each and every score in the distribution

N = the total number of cases

Substituting our data into the formula for the data in Table 9.1, we have:

$$\bar{X} = \frac{11,324}{138} = 82.06$$

The arithmetic mean for our data is 82.06. Note, though, that these summary values are in $1000's.

To interpret the arithmetic mean we may say that the typical or average salary is $82.06. It can be thought of as the point on the scale of scores that out the data in much the same way that a fulcrum on a teeter-totter balances balances the two ends of the board. It has two mathematical properties not possessed by the other indices of central tendency. In symbolic form these properties are:

1) $\Sigma(X_i - \bar{X}) = 0$ (the algebraic sum of deviations about the arithmetic mean equals zero)

2) $\Sigma(X_i - \bar{X})^2 = \text{minimum}$ (the sum of squared deviations about the mean is smaller than the sum of squared deviations about any other number)

Median. Two typical circumstances warrant using the median as the index of location: 1) when data are ordinal in nature, and 2) when a data distribution is characterized by inordinate skew. Under the latter circumstances the fact that it is a positional measure (i.e. based upon the middle position alone) suggests that the skew has little or no affect upon the statistic. The median is that statistical value that bisects a distribution of scores at the exact center so that one-half the scores fall above (i.e., are equal to or higher than) and one-half the scores fall below (i.e., are lower than) that point. Notice that we're talking about the number of scores rather than the arithmetic value of scores. This fact makes it an ordinal statistic. To compute the median in Table 9.1 we must first rank order the scores from high to low or vice versa as was done in Table 9.2. Once the scores are arrayed, if the number of scores is even, as it is here, the median is midway between the two middle scores. The two middle scores are 60 and 65. Substituting our data:

$$Mdn = \frac{60 + 65}{2} = \frac{125}{2} = 62.50$$

Suppose we had an odd number of scores (e.g., one more 250 which means the total number of cases is now 139). To compute the median for an odd number of observations the following formula is used. Hence,

$$\text{Mdn for an odd number of scores} = \frac{n + 1}{2} = \frac{139 + 1}{2} = \frac{140}{2} = 70^{\text{th}} \text{ score}$$

To interpret the median we may say that one-half the scores are above the 70th score and one-half the scores are below the 70th score. This interpretation should make clear why it is referred to as a positional statistic.

Mode. The measure of central tendency called the mode is the easiest to determine. In fact, although refined formulae exist, it does not demand calculation.[11] One merely inspects the distribution and determines the score, category, or value that most often occurs. In Table 9.1, since the score of 40 occurred most often, it is the mode of the distribution. In a grouped distribution (Table 9.4) the mode is the midpoint of the most frequently occurring

class interval or 31. The mode is sometimes called the probability average in the sense that it is that value most probable or likely to be observed. Notice that a score of 50 occurred almost as frequently as did a score of 19 and 30. In fact, the distribution is almost bimodal (two score values that occurred with equally large frequencies). Distributions of large data sets are sometimes trimodal (three scores that occurred with equally large frequencies) or even multi-modal (more than three scores have equally large frequencies).

Variability. There are several different genre of variability indices too. The most popular one is the standard deviation (symbolized s), but others such as the range, interquartile range (Q), average deviation (AD), and index of qualitative variation (IQV) are appropriate under certain conditions.

Standard Deviation. This statistic is appropriate to use with interval/ratio level data and has an interpretation possessed by no other index of variability which makes it so special. Additionally, a derivative of s, the variance (symbolized s^2 and suggesting the relationship between the two) is most frequently used in advanced statistics. To calculate the standard deviation the three-step method will be introduced. This approach enables one to see the interrelationships among three salient statistical concepts: 1) the sum of squares (symbolized Σx^2),[12] 2) the variance (s^2), and 3) the standard deviation (s).

Conceptually the sum of squares is the sum of the squared deviations about the mean. The operation $X_i - \bar{X}$ produces what is called a deviation x or mean deviate and is symbolized by the lower case x. Using this precedent the sum of squares quantity can be expressed in notation form as Σx^2.

When the sum of squares is divided by the total number of observations, N, the quotient is called the variance. Computationally,

$$s^2 = \Sigma x^2/N$$

The variance does not relate the dispersion in a data set in terms of the original units by virtue of the squaring process. To revert back to the original units the square root of the variance is derived yielding the statistic known as the standard deviation. Operationally,

$$s = \sqrt{\Sigma x^2/N}$$

For the data in Table 9.1 the sum of squares, variance, and standard deviation will now be computed. Using the above formula would be most cumbersome since the arithmetic mean would have to be subtracted from every score, then squared, and finally all squared deviations added. Instead of employing these operations a simpler computational formula for the sum of squares will be employed, namely,

$$\sum x^2 = \Sigma X^2 - (\Sigma X)^2/N$$

This operational definition directs us to square and sum all raw scores, sum and divide it by the total number of cases. Finally the latter quantity all raw scores and square that quantity is subtracted from the former and produces the sum of squares value for this data set. Performing this operation for the data in Table 9.1:

$$\Sigma x^2 = 1,359,970.5 - \frac{(11,324)^2}{138} = 430,746.04$$

To obtain the variance this quantity is divided by N or

$$s^2 = \frac{430,746.04}{138} = 3,121.35$$

Finally, the standard deviation is computed by extracting the square root of the variance or

$$s = \sqrt{3,121.35} = 55.86$$

The standard deviation for the salary data is $55.86.

To interpret this family of variability measures two different tactics will be used. First of all, both s and s^2 will be zero if the scores were all concentrated at the central tendency (mean) of the distribution (a rare phenomenon indeed) and reach a maximum when the scores are divided between

the scale's extremes. When the above remote conditions are non-existent, the variance can be interpreted in terms of a continuum ranging from a minimum value determined by dividing the range by the square root of twice the sample size to a maximum obtained by dividing the range squared by four.[13] In diagrammatic form:

minimum s^2	maximum s^2
$\frac{\text{range}}{\sqrt{2(N)}}$	$\frac{\text{range}^2}{4}$

The standard deviation could be interpreted in terms of a continuum ranging between values which are the square root of those above. For many distributions the total range encompasses about six standard deviation units. For the present data the following specific values may be substituted into the continuum and then our s or s^2 located on the scale of possible extreme values.

minimum s^2	maximum s^2
$\frac{232}{\sqrt{2(138)}} = 13.97$	$\frac{232^2}{4} = 13{,}456$
minimum s	maximum s
$\sqrt{13.97} = 3.74$	$\sqrt{13{,}456} = 116$

The second mode of interpreting the standard deviation necessitates a cursory discussion of one of the most important curves in statistics, the normal curve (N.C.). Most statisticians prefer to think of this bell-shaped curve in terms of standard scores (or z scores) rather than in terms of the original measurement units (e.g., tons, years, dollars). When conceptualized in this manner, the term standardized normal curve is applied to its description making it of much more general use than would otherwise be the case. Before discussing the properties of the normal curve apropos the

standard deviation, we must know what standard scores are and how they are used. Conceptually, a z-score represents the magnitude and direction a raw score deviates from the mean of a distribution in standard deviation units. Operationally, the following formula is used to obtain z-scores:

$$\text{z-score} = \frac{X_i - \bar{X}}{s}$$

For example, players with salaries of 140, 26, and 36 would have z-scores of:

$$z = (140 - 82.06)/55.86 = +1.04$$

$$z = (26 - 82.06)/55.86 = -1.00$$

$$z = (36 - 82.06)/55.86 = -0.82$$

Note that the standard scores represent how far above (e.g., + 1.04) or below (e.g., - 1.00; - .82)the mean a raw score lies in standard deviation units.

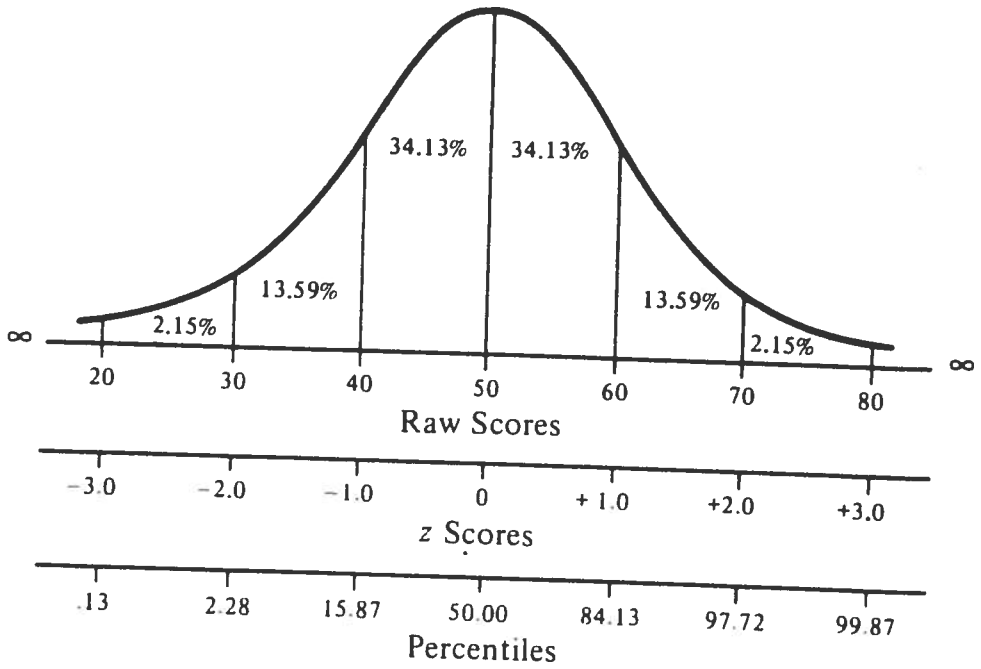
Figure 9.7 presents a standardized normal curve illustrating several important characteristics. The curve is symmetrical (i.e., possesses no skew), mesokurtic, and the three measures of central tendency coincide at the maximum height of the curve. It is described in terms of two parameters (a parameter is a characteristic of a population of elements) the arithmetic mean (μ) and the standard deviation (σ) and contains a total area (approximately) equal to one (1.00), unity, or 100 percent (when converted into percents). Finally, the N.C. is asymptotic, meaning the tails (right and left) extend to infinity without ever touching the abscissa of the graph.

(Figure 9.7 here)

Although the normal curve has a variety of uses, one important one is to think of it as a probability distribution. To illustrate, if a distribution of scores is perfectly symmetrical then a constant proportion (or percentage of cases) of the curve's area will fall between different points on the curve. Specifically, approximately 34% of the curve's area will lie

FIGURE 9.7

Relationships among Raw Scores, *z* Scores, and Percentiles of a Normally Distributed Variable



between a +1z (or +1s) or a -1z (or -1s) of the mean (or about 68% of its area between ± 1z (or ± 1s). About 13.6% of the area will lie between +1 and +2z (or -1 and -2z) and a little more than 27% of the curve's area between a +1z and a +2z or a -1z and a -2z. Finally, between a +2z and +3z or -2z and -3z there lies about 2% of the curve's area. For all practical purposes, nearly the entire area (99%+) lies between ± 3z or ± 3s. In this sense, the standard deviation is interpreted in relation to the N.C. and is useful in defining the characteristics of the curve.

Since the distribution of salaries is not completely symmetrical, the exact percentages presented above would not hold. Nevertheless, they are fairly good representations of the actual empirical distribution of scores obtained from the data set. [14]

In terms of the four statistical properties of data sets the standardized N.C. has a skew value (beta-one) of 0, a kurtosis value (beta-two) of 0, a mean of 0 and a standard deviation of 1. It has two primary uses in statistics. First, many empirical distributions are approximately normally shaped when graphed; second, and a very important consideration in inferential statistics, is the fact that many sampling distributions of statistics are normal (or approximately) which means the properties of and the knowledge we have about this special curve can be used in a probabilistic fashion.

Range. The index of variation known as the range does not require any detailed calculations. In most instances, it can readily be obtained by subtracting the extreme scores in the distribution and adding one as was previously done for the salary scores in which the maximum was 250 and the minimum was 19. Computationally, the range may be defined as the difference between the smallest score and the largest score plus one. Hence for our data:

$$250.00 - 19.00 + 1 = 232$$

One of the limitations of the range is that it only uses two scores, and two extreme scores at that. For this reason it is sometimes called the total or inclusive range. Because of this liability statisticians sometimes prefer to compute intermediate ranges such as the interquartile range or the interdecile range. For illustrative purposes the interquartile range (Q) will be computed and interpreted.

Interquartile Range. This intermediate range (Q) measures the distance encompassed by the middle 50% of the scores. It is appropriate to use with ordinal or higher level measurement data and provides the analyst with an indication of the range within which one-half the cases lie. Furthermore, it belongs to the same family of statistics as the median and is computed using a modification of that formula. Whereas the median score (actually Q2) was obtained by N/2, the first quartile (Q_1) is determined by N/4 and the third quartile (Q_3) by 3N/4. The difference between these two numbers is the interquartile range, i.e.,

$$Q = Q3 - Q1$$

For our data in Table 9.1:

$$Q = 120 - 40 = 80$$

The middle 50% of the cases contains a range of 80.

Index of Qualitative Variation. This measure of variability is designed for use with nominal level data. It enables the researcher to guage the spread or dispersion in attribute or qualitative data. The IQV provides us with a ratio between the variation that does exist to the maximum possible variation that could exist. To compute it the following formula is employed:

$$IQV = \left[\frac{\sum f_i f_j}{\frac{K(K-1)}{2}\left(\frac{N}{K}\right)^2} \right] \times 100 \qquad i \neq j$$

To obtain the actual number of differences that exist (the numerator in the formula) we multiply each attribute frequency (f_i) by every other attribute frequency (f_j) and sum (Σ) their products. To illustrate, we will calculate and interpret IQV for the following gender data. Suppose there were 43 and 105 males and females, respectively. To obtain the number of observed differences, we multiply 43 x 105 producing a value of 4515. To obtain the denominator, we substitute the appropriate values into the bottom half of the formula (e.g., K = number of categories and N = total number of cases):

$$\frac{2(2-1)}{2}\left(\frac{148}{2}\right)^2 = 5476$$

$$\frac{4515}{5476} = .82 \times 100 = 82\%$$

The index of qualitative variation for these data is .82 or 82%. To interpret it we imagine a continuum ranging from 0 (which represents no variation) to 100 (which represents maximum variation). We then make a judgment of the variability that does exist to what could exist in the data. It should be obvious that the interpretive features for this nominal-level coefficient lack the rigor and precision in comparison to say the standard deviation.

IQV range: 0% 10 20 30 40 50 60 70 80 90 100%

↑ no variation ↑ maximum variation

Summary

In this chapter some of the most commonly used descriptive statistics for univariate distributions were discussed. A concrete data set formerly collected and analyzed by this writer was subject to statistical scrutiny.

Beginning with a collection of 138 raw scores it was shown how the scores can be meaningfully assembled by constructing a frequency distribution, a rank order distribution (array), an ungrouped frequency distribution, and, finally, a grouped frequency distribution. Conventions for construction such distributions were enumerated.

Once data are meaningfully arranged the researcher typically attempts to identify their important statistical properties. To comprehend the nature of a univariate distribution (or frequency distribution) statisticians ask three salient questions: 1) What is the form or shape of the distribution? 2) What is the central tendency (or location) of the distribution? and 3) What is the variation in the distribution?

To answer query number 1, it was shown how the construction of various graphic techniques (e.g., polygon, histogram, and ogive) facilitate determining the data distribution's form or shape. Guidelines for constructing and interpreting these common graphic devices for univariate data were considered. From them an intuitive appreciation of form, central tendency, and variation could be inferred.

Because graphic devices generally lack the precision of specific statistical indices, various statistics for identifying the exact value of form, central tendency, and variation were discussed along with their computational formulae. Statistically speaking, beta-one (a skewness index) and beta-two (a kurtosis index) provide specific numerical values from which the form of the distribution can be assessed. The arithmetic mean, median, and mode were calculated and interpreted for their role in determining the central tendency of a data set. Similarly, the standard deviation, variance, interquartile range and index of qualitative variation were computed and inter-

preted since they are among the commonest indices of variability or dispersion.

Because the normal curve plays such an important role in statistics (both descriptive and inferential) the manner in which z-scores (standard scores) are related to it were considered. Also, the properties of the standardized normal curve where discussed with suggestions of how these properties can be efficiently utilized by the social researcher.

Important Concepts Discussed in This Chapter

Statistics
Frequency Distribution
Rank Order Distribution
Maximum
Minimum
Range
Class Intervals
Central Tendency
Form or Shape
Variation
Polygon
Histogram
Ogive
Positive Skew
Negative Skew
Symmetrical Curve
Asymmetrical Curve
Grouped Frequency Distribution
Arithmetic Mean
Median
Mode
Cumulative Frequencies
Exact Limits
Graphs
Skewness
Kurtosis
Beta-one
Beta-two
Sum of Squares
Variance
Standard Deviation
Standard Scores (z-scores)
Normal Curve
Parameter
Sampling Distribution
Interquartile Range
Index of Qualitative Variation

The Greek Alphabet

Greek Letter		*Name*	*English Equivalent*	*Greek Letter*		*Name*	*English Equivalent*
A	α	alpha	A	N	ν	nu	N
B	β	beta	B	Ξ	ξ	xi	X
Γ	γ	gamma	G	O	o	omicron	short O
Δ	δ	delta	D	Π	π	pi	P
E	ϵ	epsilon	short E	P	ρ	rho	R
Z	ζ	zeta	Z, DZ	Σ	σ	sigma	S
H	η	eta	long E	T	τ	tau	T
Θ	θ	theta	TH	Υ	υ	upsilon	U
I	ι	iota	I	Φ	ϕ	phi	PH
K	κ	kappa	K	X	χ	chi	CH
Λ	λ	lambda	L	Ψ	ψ	psi	PS
M	μ	mu	M	Ω	ω	omega	long O