CHAPTER 3
DESCRIPTIVE STATISTICS

Statistics is the collection of great masses of numerical information which is summarized and then analyzed for the purpose of making decisions; that is, the use of past information is used to predict future actions. In this chapter we will assume that the numerical data has been collected (by various processes) and will discuss distribution and measures of central tendency and variability.

FREQUENCY DISTRIBUTION

When we classify, by order, many variables into classes by size and put this information into table form, we have created a frequency distribution table.

ORDER

The grades in a mathematics class, as shown in table 3-1, are written in descending order; that is, the highest grade first, the next highest grade second, etc. We refer to this set of grades as an ordered array. If we had the grades shown in table 3-2, with the occurrences shown, we would refer to this as a frequency distribution; that is, we have shown the grades and the number of occurrences of each.

The data in table 3-2 are discrete variables and are naturally classified. By discrete variables, we mean a finite number of variables. By naturally ordered, we mean the variables are listed in increasing or decreasing order of value.

When dealing with continuous variables, we usually classify them; that is, we group them according to some class boundaries. If we were forming a frequency distribution table of weights of 100 individuals, we would determine the heaviest (231 lb) and lightest (109 lb), then divide the weights into from 10 to 20 classes. We subtracted 109 from 231 and found the difference to be 122. If we used 10 classes, each class interval would be

$$\frac{122}{10} = 12.2$$

Table 3-1.—Array of values.

Grades
99
98
97
97
96
94
92
90
88
88
83
80
80
78
76
71
71
68
60

Table 3-2.—Frequency distribution.

Grade	Occurrences	Frequency
99	I	1
98	II	2
96	IIII	4
92	~~IIII~~ II	7
90	~~IIII~~	5
88	~~IIII~~ ~~IIII~~ III	13
86	~~IIII~~ ~~IIII~~ I	11
83	~~IIII~~ II	7
80	~~IIII~~	5
78	IIII	4
75	III	3
60	I	1

If we used 20 classes, each class interval would be

$$\frac{122}{20} = 6.1$$

We may choose any number between 6.1 and 12.2, and we find it convenient to use 10 as the class interval. We know the smallest number must fall into the lowest class; therefore, we assign the lower limit of the first class as 108.5 which is one-half unit beyond the accuracy of the weights. This prevents any weight falling on a boundary. The first class interval is 108.5 (the lower boundary) plus the interval of 10 which gives 108.5 - 118.5. The next class interval is 118.5 - 128.5, etc.

Now as we determine each individual weight, we make a mark beside the proper class interval. We determine class marks by finding the midpoint of each class interval. The frequency column is the number of tally marks in the occurrence column. This is shown in table 3-3.

The class marks (x) indicate that we have assigned each weight in that class interval the weight of that class mark. The frequencies (f) indicate the number of occurrences.

HISTOGRAM

The information in table 3-3 would be easier to visualize if it were shown graphically. This is shown in figure 3-1.

The class boundaries are indicated on the horizontal axis and the frequencies are indicated on the vertical axis. If the width of the bars is considered unity, then the area of each rectangle is representative of the frequency. The total area of all the rectangles, then, represents the total frequency. Figure 3-1 is a histogram of the information in table 3-3.

POLYGON

Another method of representing the information in table 3-3 is shown in figure 3-2. This figure is developed by connecting the midpoints of the tops of adjacent rectangular bars of figure 3-1 together. These midpoints are in actual practice the class marks of the classes. The area under the curve of the polygon is the same as the area under the curve of the histogram. This may be seen by examining one of the rectangles and noting that there is the same amount of area, cut into a triangle, outside the bar as there is inside the bar. This is shown in the shaded area of figure 3-2.

Both the histogram and the polygon present a graphical representation of data which is easy to visualize. These are used to quickly compare one set of data with another set of data.

MEASURES OF CENTRAL TENDENCY

The previously mentioned frequency distribution tables dealt with the variables specifically. To summarize the tendencies of the variables we use the idea of central tendency. Several ways of describing the central tendency are by use of the arithmetic mean, the median, the mode, and the geometric and harmonic means. These are discussed in this section.

ARITHMETIC MEAN

We will use the term mean to indicate the arithmetic mean which is the commonly used idea of average.

Table 3-3.—Frequency distribution with class boundaries.

Class boundaries	Occurrences	Class marks: x	Frequencies
108.5 - 118.5	I	113.5	1
118.5 - 128.5	III	123.5	3
128.5 - 138.5	IIII	133.5	4
138.5 - 148.5	卌	143.5	5
148.5 - 158.5	卌 IIII	153.5	9
158.5 - 168.5	卌 卌 卌 II	163.5	17
168.5 - 178.5	卌 卌 卌 卌	173.5	20
178.5 - 188.5	卌 卌 卌	183.5	15
188.5 - 198.5	卌 卌	193.5	10
198.5 - 208.5	卌 III	203.5	8
208.5 - 218.5	卌	213.5	5
218.5 - 228.5	II	223.5	2
228.5 - 238.5	I	233.5	1

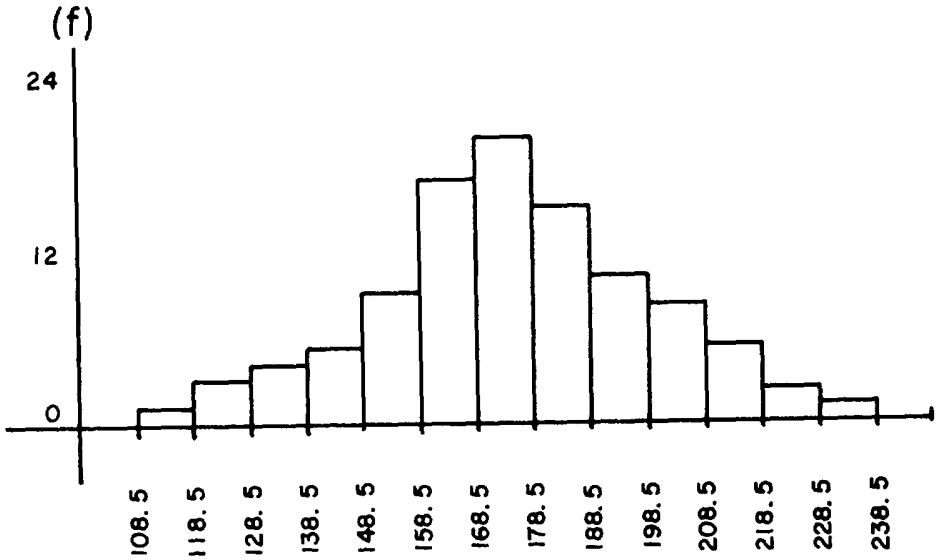

Figure 3-1.—Histogram.

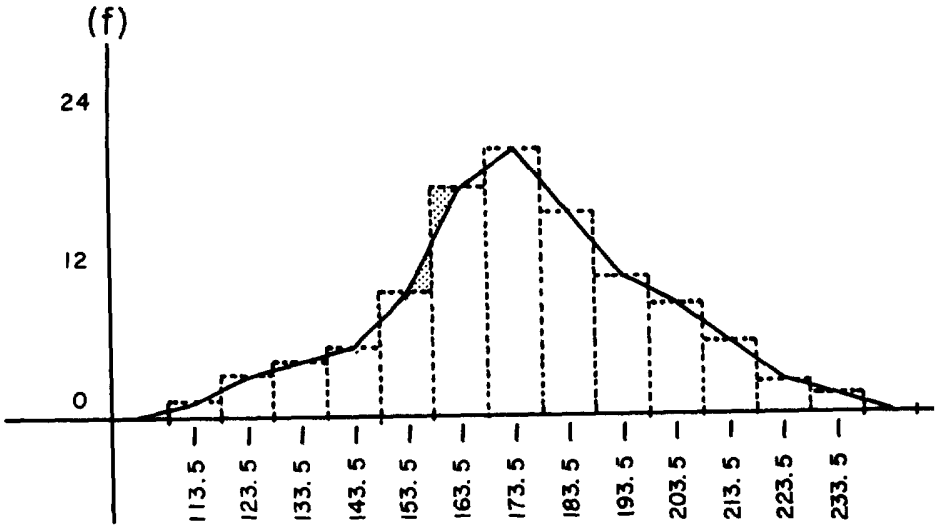

Figure 3-2.—Polygon.

The mean is defined as the sum of a group of values divided by the number of values.

If we have the test scores of 70, 66, 85, 95, 90, and 80, we find the mean by adding the scores and then dividing the sum by the number of scores we have; that is,

$$\begin{array}{r} 70 \\ 66 \\ 85 \\ 95 \\ 90 \\ \underline{80} \\ 486 \end{array}$$

and

$$486 \div 6 = 81$$

which is the mean.

If $\overline{X}$ is the mean, then

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

or

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Sigma (Σ) is the summation symbol and i = 1 to n indicates that the values of X_i from i = 1 to i = n are added. This sum is then divided by n, the number of scores involved.

EXAMPLE: Find the mean of 78, 92, 63, 76, 83, 82, and 79.

SOLUTION: In the formula

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

the sum of the scores is 553 and n equals 7, therefore,

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$= \frac{553}{7}$$

$$= 79$$

In cases such as shown in table 3-4, the mean could be found by adding each grade (notice there are three 72's, six 80's, etc.) then dividing by the total number of grades. It would be far easier to multiply each grade by the number of times it occurred and then adding these products to find the sum. This sum is then divided by the total number of grades. This is shown in the formula

$$\overline{X} = \frac{\sum_{i=1}^{n} fX_i}{n}$$

where

f = frequency of each grade

and

n = total number of grades

This gives

$$\overline{X} = \frac{1(60) + 3(72) + 6(80) + 4(92) + 2(96)}{16}$$

$$= \frac{60 + 216 + 480 + 368 + 192}{16}$$

$$= 82.25$$

Table 3-4.—Sample frequency distribution.

Grades	Frequency	$f(X_i)$
60	1	60
72	3	216
80	6	480
92	4	368
96	2	192
	16	1316

Computation of Mean by Coding

Our computations to this point have dealt with a relatively small number of values. When

the number of values becomes large, we may resort to the use of coding to determine the mean.

Before discussing the actual coding process, we will examine a few related type procedures. If we were to find the mean of the values 92, 87, 85, 80, 78, and 65, we could assume a mean of (80) and determine the deviation of each value from this mean as follows:

Value	Deviation
92	+ 12
87	+ 7
85	+ 5
80 assumed mean	0
78	- 2
65	- 15

We then algebraically add the deviations and find the sum to be + 7. Divide + 7 by 6, the number of values, and find the quotient to be +1.16. We then add, algebraically, the mean of the deviations (+1.16) to the assumed mean of the values (80) and find the actual mean of the values to be (80) + (+1.16) or 81.16.

If we were to find the mean of 90, 80, 70, 60, 50, and 40, we could assume the mean to be 70 and write

Value	Deviation
90	+ 20
80	+ 10
70 assumed mean	0
60	- 10
50	- 20
40	- 30

Then find the mean of the deviations to be

$$\frac{-30}{6} = -5$$

and the actual mean of the values is the algebraic sum of the assumed mean and the mean of the deviations which is

$$\begin{array}{ccccc} \text{assumed mean} & + & \text{mean of deviations} & = & \text{actual mean} \\ 70 & + & (-5) & = & 65 \end{array}$$

Notice in the preceding example that all deviations are multiples of ten. By dividing each by ten we would have deviations of 2, 1, 0, -1, -2, -3. Once again we shall assume 70 as the mean. We now have the following deviations and values:

Value	Deviation
90	+ 2
80	+ 1
70	0
60	- 1
50	- 2
40	- 3

The algebraic mean of the deviations is equal to

$$\frac{-3}{6} = -0.5$$

Now we must multiply -0.5 by ten to arrive at the same mean of deviations we found in the previous example. This may be done because the difference in the deviations is a constant, and this was due to the values having a constant difference.

As we have seen, the computation of the mean by use of the formula

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i f_i$$

is rather simple when the number of values and their frequencies are small. When X_i and f_i are large we may resort to the use of coding. We will use u to designate the coding variable. When we have a set of variables as shown in table 3-5, we represent the class marks by both positive and negative integers. The zero may be placed opposite any value near the middle of the distribution. We choose 163.5 as the value to correspond to u = 0.

To find the mean of the values, using the code, we find the algebraic sum of column u to be -3. The mean of the u's then, is

$$6\overline{)\,-3}\quad -0.5$$

Table 3-5.—Coded distribution.

Class boundaries	Frequency f	Class mark x	Code u
128.5 - 138.5	1	133.5	- 3
138.5 - 148.5	1	143.5	- 2
148.5 - 158.5	1	153.5	- 1
158.5 - 168.5	1	163.5	0
168.5 - 178.5	1	173.5	+ 1
178.5 - 188.5	1	183.5	+ 2

The differences in class marks is 10; therefore, we multiply (-0.5) by 10 and find this equal to -5. The -5 is added to the value corresponding to u equal 0. Thus,

$$163.5 + (-5) = 158.5$$

which is the mean of the class marks.

If we use x_o to designate the value corresponding to $u = 0$, and if we use C to indicate the class interval (difference between adjacent class marks), we may show the relationship between the x's and u's by

$$x_i = Cu_i + x_o$$

In table 3-5, the length of the class interval is 10; therefore,

$$x_i = 10u_i + x_o$$

We may verify this formula by choosing any class mark in table 3-5. Let us test

$$x_i = 143.5$$

Then

$$x_i = 10u_i + x_o$$

and

$$143.5 = 10\ (-2) + 163.5$$

$$= -20 + 163.5$$

$$= 143.5$$

To compute the value for $\bar{x}$ we substitute $\bar{x}$ for x_i and $\bar{u}$ for u_i in the formula

$$x_i = Cu_i + x_o$$

and find

$$\bar{x} = C\bar{u} + x_o$$

Then,

$$C = 10$$

$$\bar{u} = \frac{-3}{6} = -0.5$$

and

$$x_o = 163.5$$

therefore,

$$\bar{x} = C\bar{u} + x_o$$

$$= 10\ (-0.5) + 163.5$$

$$= -5 + 163.5$$

$$= 158.5$$

We may verify this by writing

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

$$= 133.5 + 143.5 + 153.5 + 163.5 + 173.5 + 183.5$$

$$= \frac{951.0}{6}$$

$$= 158.5$$

The reason we may substitute $\bar{x}$ for x_i and $\bar{u}$ for u_i is shown as follows:

We have shown that

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i f_i \qquad (1)$$

and

$$x_i = Cu_i + x_o \qquad (2)$$

Then by substituting (2) into (1) we have

$$\bar{x} = \left(\frac{1}{n}\right) \sum_{i=1}^{n} (Cu_i + x_o) f_i$$

$$= \left(\frac{1}{n}\right) \sum_{i=1}^{n} (Cu_i f_i + x_o f_i)$$

$$= \left(\frac{1}{n}\right) \sum_{i=1}^{n} Cu_i f_i + \left(\frac{1}{n}\right) \sum_{i=1}^{n} x_o f_i$$

$$= C \left(\frac{1}{n}\right) \sum_{i=1}^{n} u_i f_i + x_o \left(\frac{1}{n}\right) \sum_{i=1}^{n} f_i$$

Now,

$$\bar{u} = \left(\frac{1}{n}\right) \sum_{i=1}^{n} u_i f_i$$

and

$$\left(\frac{1}{n}\right) \sum_{i=1}^{n} f_i = 1, \text{ where } n = \sum_{i=1}^{n} f_i$$

therefore,

$$C \left(\frac{1}{n}\right) \sum_{i=1}^{n} u_i f_i$$

$$= C\bar{u}$$

and

$$x_o \left(\frac{1}{n}\right) \sum_{i=1}^{n} f_i = x_o$$

then

$$\bar{x} = C \left(\frac{1}{n}\right) \sum_{i=1}^{n} u_i f_i + x_o \left(\frac{1}{n}\right) \sum_{i=1}^{n} f_i$$

$$= C\bar{u} + x_o$$

The previous example which used table 3-5 dealt with values which all had a frequency of one. To compute the mean of the values shown in table 3-6 will involve varied frequencies and is done as follows:

We use the formula

$$\bar{x} = C\bar{u} + x_o \qquad (3)$$

and by inspection of table 3-6 find that

$$C = 10$$

and

$$x_o = 163.5$$

The next step is to determine u; that is,

$$\bar{u} = \left(\frac{1}{n}\right) \sum_{i=1}^{n} u_i f_i$$

where

$$n = 42$$

and

$$\sum_{i=1}^{n} u_i f_i = +9$$

Then,

$$\bar{u} = \frac{1}{42} \left(+ \frac{9}{1}\right)$$

$$= \frac{9}{42}$$

Table 3-6.—Coded frequency distribution.

Class boundaries	Frequency f	Class marks x	Code u	(Code) (Freq.) uf
128.5 - 138.5	2	133.5	- 3	- 6
138.5 - 148.5	4	143.5	- 2	- 8
148.5 - 158.5	7	153.5	- 1	- 7
158.5 - 168.5	11	163.5	0	0
168.5 - 178.5	9	173.5	+ 1	+ 9
178.5 - 188.5	6	183.5	+ 2	+ 12
188.5 - 198.5	3	193.5	+ 3	+ 9
	42			+ 9

Substituting into equation (3), find that

$$\bar{x} = 10\left(\frac{9}{42}\right) + 163.5$$

$$= \frac{90}{42} + 163.5$$

$$= 2.14 + 163.5$$

$$= 165.64$$

To show the usefulness of coding, we will now compute (the long way) the mean of the class marks of table 3-6 by use of the formula

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i f_i$$

$$= \frac{1}{42}\sum_{i=1}^{n} x_i f_i$$

and the products of the $x_i f_i$'s are

133.5 x 2 = 267.0

143.5 x 4 = 574.0

153.5 x 7 = 1074.5

163.5 x 11 = 1798.5

173.5 x 9 = 1561.5

183.5 x 6 = 1101.0

193.5 x 3 = 580.5

Therefore,

$$\frac{1}{42}\sum_{i=1}^{n} x_i f_i$$

$$= \frac{1}{42}(6957.0)$$

$$= \frac{6957.0}{42}$$

$$= 165.64$$

Notice that $\bar{x}$ was the same in each case, but in the first case far less computation was required.

PROBLEMS:

1. Find $\bar{x}$ in table 3-7 by completing the indicated columns and using coding, and check your answer by using the formula

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i f_i$$

2. Compute, by coding, $\bar{x}$ in table 3-3.

ANSWERS:

1. $\bar{x} = 182.42$

2. $\bar{x} = 174.3$

Table 3-7.—Practice problem.

Class boundaries	Frequency f	Class marks x	Code u	(Code) (Freq.) uf
138.2 - 148.2	4	143.2		
148.2 - 158.2	6	153.2		
158.2 - 168.2	13	163.2		
168.2 - 178.2	15			
178.2 - 188.2	17			
188.2 - 198.2	14			
198.2 - 208.2	11			
208.2 - 218.2	8			
218.2 - 228.2	2			
	90			

MEDIAN

In an ordered array of values, the value which has as many values above it as below it is called the median. In some cases the median may be a point rather than a value. This occurs when there is an even number of values and will be explained by a following example.

When we have a very large or small value, as compared to the other values in the array, the median is generally superior to the mean as a measure of central tendency. This is because the large or small value will cause the mean to move away from the major grouping of the values.

EXAMPLE: Compare the mean and median of the values 2, 3, 3, 5, 6, 9, 133.

SOLUTION: We find the mean to be

$$\frac{161}{7} = 23$$

The median is the middle number in the ordered array

133
9
6
5
3
3
2

which is the number 5. Notice that more of the values cluster above and below the 5 than about the mean of 23.

Also, the median may be used in cases where items are arranged according to merit rather than value. For example, workers may be rated by their ability; then, the median of the abilities of the workers is the rating of the middle worker in the array.

Generally, then, the median is a measure of position rather than a measure of value.

EXAMPLE: Find the median of the following values: 7, 6, 5, 3, 2, 2, 1.

SOLUTION: Arrange the values in an ordered array as

7
6
5
3
2
2
1

Then, by inspection, find the number which has as many values above it as below it: 3.

EXAMPLE: Find the median of the values 9, 9, 7, 6, 5, 4.

SOLUTION: The values in ordered array form are

9
9
7
6
5
4

We find no middle value; therefore, the median is the mean of the two middle values; that is,

$$\overline{X} = \frac{7 + 6}{2}$$

$$= 6.5$$

which is the median for the set of values given.

Notice that $\overline{X}$ in this case is the mean of only the two values 6 and 7 and not the entire array. We will designate the median by Md.

EXAMPLE: Find and compare the median of each of the following arrays; that is, A and B:

A	B
9	50
8	48
7	10
6	6
5	3
4	2
3	1

SOLUTION: The median of both A and B is 6. In this case the median should not be used to compare A and B because of the wide range of B values and the close grouping of A values. In this case the means would give a better comparison. In some cases the median will give a more realistic meaning to a set of values.

EXAMPLE: In a small organization the salaries of the 5 employees are

Employee	Salary
A	$ 7600
B	$ 4900
C	$ 4700
D	$ 4500
E	$ 4300

The median of the salaries is $ 4700. The mean of the salaries is

$$\overline{X} = \frac{7600 + 4900 + 4700 + 4500 + 4300}{5}$$

$$= \frac{26000}{5}$$

$$= \$\ 5200$$

Notice that with

$$Md = \$\ 4700$$

and

$$\overline{X} = \$\ 5200$$

the median is more representative than the mean because four of the five employees have salaries less than the mean.

MODE

In a distribution the value which occurs most often is called the mode. When two or more values occur most often, rather than just one value, there will be more than one mode.

EXAMPLE: Find the mode of the values 7, 9, 11, 7, 8, 6, 6, 7, 5.

SOLUTION: By inspection, the value which occurs most often is 7; therefore, the mode is 7. It is indicated by writing

$$Mo = 7$$

EXAMPLE: Find the mode of the values 18, 20, 17, 17, 16, 16, 15, 16, 17, 20.

SOLUTION: By inspection, the values which occur most often are 16 and 17; therefore,

$$Mo = 16 \text{ and } 17$$

In this case there are two modes.

RANGE

The range of a set of values or of an array is defined as the difference between the largest and smallest values. The range in the preceding example is the high value (20) minus the low value (15) which is

$$20 - 15 = 5 = r \text{ (range)}$$

PROBLEMS: Find the mean, median, mode, and range in the following:

1. 17, 19, 31, 21, 34, 6, 8, 9, 17
2. 100, 60, 80, 80, 60, 60, 70, 50
3. 7.2, 3.7, 6.2, 10.3, 11.9

ANSWERS:

1. $\overline{X} = 18$

   $Md = 17$

   $Mo = 17$

   $r = 28$

2. $\bar{X} = 70$

   Md = 65

   Mo = 60

   r = 50

3. $\bar{X} = 7.86$

   Md = 7.2

   Mo = none

   r = 8.2

GEOMETRIC MEAN

The geometric mean is sometimes used to average a set of percentages or other ratios. The geometric mean is found in the same manner as the arithmetic mean except that logarithms are used. The geometric mean is also useful when the variables have the characteristics of a geometric progression or sequence. The formula for the geometric mean is

$$G = \sqrt[n]{X_1 \cdot X_2 \cdot X_3 \cdots X_n}$$

This formula may be changed as follows, by taking logarithms on both sides:

then

$$\log G = \log \sqrt[n]{X_1 \cdot X_2 \cdot X_3 \cdots X_n}$$

$$= \log (X_1 \cdot X_2 \cdot X_3 \cdots X_n)^{\frac{1}{n}}$$

$$= \left(\frac{1}{n}\right) \log (X_1 \cdot X_2 \cdot X_3 \cdots X_n)$$

$$= \frac{\log (X_1 \cdot X_2 \cdot X_3 \cdots X_n)}{n}$$

$$= \frac{\log X_1 + \log X_2 + \log X_3 + \cdots + \log X_n}{n}$$

$$\log G = \frac{\sum_{i=1}^{n} \log x_i}{n}$$

By definition, the geometric mean of X is the antilogarithm of the arithmetic mean of log X.

EXAMPLE: Find the arithmetic mean and the geometric mean of the price:earning ratios for the stocks as shown in table 3-8.

SOLUTION: Find the logarithms of X and write the column of log X and find totals.

Table 3-8.—Geometric mean.

Stock	Price earning ratio (x)	Log of price:earning ratio (log x)
I	18.2	1.2601
II	17.3	1.2380
III	16.8	1.2253
IV	14.5	1.1614
V	31.2	1.4942
	98.0	6.3790

The arithmetic mean is

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{98}{5}$$

$$= 19.6$$

The geometric mean is

$$\log G = \frac{\sum_{i=1}^{n} \log x_i}{n}$$

$$= \frac{6.3790}{5}$$

$$= 1.2758$$

Then,

G = antilogarithm of log G

= antilog (1.2758)

= 18.8^{+}

PROBLEM: Find the arithmetic and geometric mean of the following:

Item	X	log X
I	32.6	1.5132
II	17.2	1.2355
III	9.6	0.9823
IV	21.7	1.3365
V	33.1	1.5198
VI	15.8	1.1987

ANSWER:

arithmetic mean is 21.66

geometric mean is 19.84

HARMONIC MEAN

In certain cases the harmonic mean serves a useful purpose. Although this mean is generally not found in statistics, it is a method of describing a set of numbers and will be explained.

When averages are desired where equal times are involved, the arithmetic mean of speeds is used and when equal distances are given the harmonic mean is useful.

EXAMPLE: An automobile travels for 3 hours at a rate of 60 miles per hour, then travels for 3 hours at a rate of 70 miles per hour. What is the average speed of the automobile?

SOLUTION: Equal times are involved; therefore, the arithmetic mean of the automobile speed is

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\bar{X} = \frac{60 + 70}{2}$$

$$= \frac{130}{2}$$

$$= 65$$

EXAMPLE: An automobile travels 200 miles at a rate of 60 miles per hour and the next 200 miles at a rate of 70 miles per hour. What is the average speed of the automobile?

SOLUTION: Equal distances are involved, resulting in unequal times for the two rates. Therefore, the harmonic mean is more accurate than the arithmetic mean. It is found as follows:

$$H = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \cdots + \frac{1}{X_n}}$$

$$= \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$$

$$= \frac{2}{\frac{1}{60} + \frac{1}{70}}$$

$$= \frac{2}{\frac{130}{4200}}$$

$$= 2\left(\frac{420}{13}\right)$$

$$= 64.6$$

The difference in the averages is explained by the fact that the automobile in the first example traveled a total distance of 390 miles in 6 hours or at an average speed of 65 miles per hour. The automobile in the second example traveled 400 miles in 3.33 hours plus 2.86 hours or 6.19 hours which is an average speed of 64.6 miles per hour.

The fallacy of using "averaging" (arithmetic mean) when times are unequal may be demonstrated even more dramatically by finding the "average" speed of an automobile which travels 595 miles at a speed of 60 miles per hour and travels the final 5 miles of a 600-mile trip at 20 miles per hour.

PROBLEM: An automobile travels 100 miles at a speed of 50 miles per hour, 100 miles at a speed of 45 miles per hour and 100 miles at a speed of 70 miles per hour. What is the average speed of the automobile?

ANSWER: 53.09 miles per hour.

MEASURES OF VARIABILITY

To this point we have discussed averages or means of sets of values. While the mean is a useful tool in describing a characteristic of a set of values, it does not indicate how the values are dispersed about the mean. That is, the values 20, 50, and 80 have the same mean as 45, 50, and 55 although in the first case the dispersion and range is much greater. In describing a set of values we need to know not only the mean but also how the values are dispersed about the mean.

Generally, when the dispersion is small, the average is a reliable description of the values; and if the dispersion is great, the average is not typical of the values, unless the number of values is very large.

MEAN DEVIATION

The mean deviation is defined as the arithmetic mean of the absolute values of the deviations from the mean. In the set of values 45, 50, and 55, the mean deviation, given by the formula

$$\text{M. D.} = \frac{1}{n} \sum_{i=1}^{n} |X_i - \overline{X}|$$

where the mean $(\overline{X})$ is 50, is shown as

values X	$\lvert X_i - \overline{X} \rvert$	= absolute value of deviations
55	$\lvert 55-50 \rvert$	= 5
50	$\lvert 50-50 \rvert$	= 0
45	$\lvert 45-50 \rvert$	= 5

and the mean of the absolute values of the deviations is

$$\frac{5 + 0 + 5}{3}$$

$$= \frac{10}{3}$$

$$= 3.33$$

Notice that the absolute value of the deviations is used because the mean of the deviations would be zero; that is,

$$X_i - \overline{X} = \text{deviation}$$

$$55-50 = 5$$

$$50-50 = 0$$

$$45-50 = -5$$

and the mean would be

$$\frac{5 + 0 - 5}{3}$$

$$= \frac{0}{3}$$

$$= 0$$

The mean deviation of the values 20, 50, and 80 is

$$\text{M. D.} = \frac{1}{n} \sum_{i=1}^{n} |X_i - \overline{X}|$$

$$= \frac{1}{3} (30 + 0 + 30)$$

$$= \frac{60}{3}$$

$$= 20$$

EXAMPLE: Find the mean deviation of the values 72, 60, 85, 90, 63, 80, 90, 93, and 87.

SOLUTION: Make an array with columns for X and $|X_i - \overline{X}|$ as follows:

X	$\lvert X_i - \overline{X} \rvert$
93	
90	
90	
87	
85	
80	
72	
63	
60	

Determine the mean as

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$= \frac{1}{9} (720)$$

$$= 80$$

then complete the column for $|X_i - \bar{X}|$ as

X	$\lvert X_i - \bar{X}\rvert$
93	13
90	10
90	10
87	7
85	5
80	0
72	8
63	17
60	20

Now,

$$\text{M. D.} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})$$

$$= \frac{1}{9}(13 + 10 + 10 + 7 + 5 + 0 + 8 + 17 + 20)$$

$$= \frac{90}{9}$$

$$= 10$$

PROBLEMS: Find the mean deviation for each set of values given.

1. 7, 9, 11, 20, 15

2. 2, 2, 3, 4, 5

ANSWERS:

1. 4.08

2. 1.04

To this point in our discussion of mean deviation we have dealt with arrays of values. If we desire to find the mean deviation of a frequency distribution, we need only modify the formula for mean deviation; that is,

$$\text{M. D.} = \frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{X}|$$

is written to include frequency (f_i) as follows:

$$\text{M. D.} = \frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{X}|\, f_i$$

EXAMPLE: Find the mean deviation of the following values:

16, 13, 15, 15, 13, 17, 13, 18, 20, 17, 12

SOLUTION: Write the frequency distribution as follows:

X	f	$\lvert X_i - \bar{X}\rvert$	$\lvert X_i - \bar{X}\rvert f_i$
20	1	4.6	4.6
18	1	2.6	2.6
17	2	1.6	3.2
16	1	0.6	0.6
15	2	0.4	0.8
13	3	2.4	7.2
12	1	3.4	3.4

where $\bar{X} = 15.4$ and

$$\text{M. D.} = \frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{X}|\, f_i$$

$$= \frac{1}{11}[1(4.6) + 1(2.6) + 2(1.6) + 1(0.6) + 2(0.4) + 3(2.4) + 1(3.4)]$$

$$= \frac{1}{11}(4.6 + 2.6 + 3.2 + 0.6 + 0.8 + 7.2 + 3.4)$$

$$= 2.04$$

PROBLEM: Find the mean deviation of the grades in table 3-4.

ANSWER: 8.31

STANDARD DEVIATION

While the mean deviation is a useful tool in statistics, the standard deviation is the most important measure of variability. The standard deviation is the square root of the mean of the squares of the deviations from the mean; that is,

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

The symbol σ (lower case sigma) designates the standard deviation. Notice that instead of using the absolute value of $X_i - \bar{X}$ as in the computation of the mean deviation, we square $X_i - \bar{X}$ and then find the square root of the sum of $(X_i - \bar{X})^2$ divided by n.

EXAMPLE: Find the standard deviation of the values 60, 70, 75, 65, 70, 80.

SOLUTION: Make a table as follows:

X	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
80	+ 10	100
75	+ 5	25
70	0	0
70	0	0
65	- 5	25
60	- 10	100
		250

Then,

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$= \sqrt{\frac{1}{6} (250)}$$

$$= \sqrt{41.7}$$

$$= 6.45$$

This indicates that the standard deviation of the values from the mean of 70 is 6.45.

EXAMPLE: Find the standard deviation of the values 2, 2, 3, 4, 5.

SOLUTION: Write

X	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
2	-1.2	1.44
2	-1.2	1.44
3	-0.2	0.04
4	+0.8	0.64
5	+1.8	3.24
		6.80

Therefore,

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$= \sqrt{\frac{1}{5} (6.80)}$$

$$= \sqrt{1.36}$$

$$= 1.166$$

In the two previous examples we used n as the divisor, but in many cases, especially where n is small, the formula is modified by the use of n - 1 in place of n; that is,

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

is the standard deviation for a large population and

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

is the formula for standard deviation when the population is small. In some cases, s is called the sample standard deviation. The latter is commonly used in statistics. In the previous example the value for s is

$$s = \sqrt{\frac{1}{4} (6.80)}$$

$$= \sqrt{1.70}$$

$$= 1.3$$

and gives a better estimate of the standard deviation of the population from which the sample was taken.

We have shown examples where the frequency of occurrence of each value was considered individually. To use the formula for standard deviation with a frequency distribution, we need only include f_i; that is,

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 f_i}$$

EXAMPLE: Find the standard deviation of the values 80, 75, 75, 70, 65, 65, 65, 60.

SOLUTION: Write

X	f
80	1
75	2
70	1
65	3
60	1

Then,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i f_i$$

$$= \frac{1}{8}[1(80) + 2(75) + 1(70) + 3(65) + 1(60)]$$

$$= \frac{1}{8}(80 + 150 + 70 + 195 + 60)$$

$$= 69.37$$

Now write the following tabulation:

X	f	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^2 f_i$
80	1	10.63	112.99	112.99
75	2	5.63	31.69	63.38
70	1	.63	.39	.39
65	3	-4.37	19.09	57.27
60	1	-9.37	87.79	87.79
				321.82

therefore,

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 f_i}$$

$$= \sqrt{\frac{1}{7}(321.82)}$$

$$= \sqrt{45.97}$$

$$= 6.78$$

In order to simplify our calculations, we resort to the use of coding as we did with the mean. We know that

$$X_i = Cu_i + X_o$$

and

$$\bar{X} = C\bar{u} + X_o$$

and by subtracting the second equation from the first equation we have

$$\begin{aligned} X_i &= Cu_i + X_o \\ (-)\quad \bar{X} &= C\bar{u} + X_o \\ \hline X_i - \bar{X} &= Cu_i + X_o - (C\bar{u} + X_o) \\ &= Cu_i + X_o - C\bar{u} - X_o \\ &= Cu_i - C\bar{u} \\ &= C(u_i - \bar{u}) \end{aligned}$$

Substitute

$$X_i - \bar{X} = C(u_i - \bar{u})$$

in the formula for standard deviation

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 f_i}$$

$$= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}[C(u_i - \bar{u})]^2 f_i}$$

$$= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}C^2(u_i - \bar{u})^2 f_i}$$

$$= C\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{u})^2 f_i}$$

In the previous example, because the difference between values is a constant which is 5, we may write

X	f_i	u_i	$u_i f_i$	$u_i - \bar{u}$	$(u_i - \bar{u})^2$	$(u_i - \bar{u})^2 f_i$
80	1	+2	+2	2.125	4.51	4.51
75	2	+1	+2	1.125	1.26	2.52
70	1	0	0	0.125	0.0156	0.0156
65	3	-1	-3	-0.875	0.76	2.28
60	1	-2	-2	-1.875	3.51	3.51
	8					12.84

where

$$\bar{u} = \frac{1}{n} \sum_{i=1}^{n} u_i f_i$$

$$= \frac{1}{8} (-1)$$

$$= -\frac{1}{8}$$

$$= -0.125$$

Then,

$$s = C \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (u_i - \bar{u})^2 f_i}$$

$$= C \sqrt{\frac{1}{n-1} (12.84)}$$

$$= 5 \sqrt{\frac{12.84}{7}}$$

$$= 5 \sqrt{1.834}$$

$$= 5 (1.354)$$

$$= 6.77$$

A simpler method for calculating the standard deviation is by changing the formula

$$s = C \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (u_i - \bar{u})^2 f_i}$$

by the following algebraic manipulations and, omitting the limits to simplify calculations, we have

$$\Sigma (u_i - \bar{u})^2 f_i$$

$$= \Sigma u_i^2 f_i - 2 \Sigma u_i \bar{u} f_i + \Sigma \bar{u}^2 f_i$$

$$= \Sigma u_i^2 f_i - 2 \bar{u} \Sigma u_i f_i + \bar{u}^2 \Sigma f_i$$

$$= \Sigma u_i^2 f_i - 2 \left(\frac{1}{n}\right) (\Sigma u_i f_i)^2 + \left(\frac{\Sigma u_i f_i}{n}\right)^2 \Sigma f_i$$

$$= \Sigma u_i^2 f_i - \left(\frac{2}{n}\right) (\Sigma u_i f_i)^2 + (\Sigma u_i f_i)^2 \frac{\Sigma f_i}{n^2}$$

$$= \Sigma u_i^2 f_i - \left(\frac{2}{n}\right) (\Sigma u_i f_i)^2 + (\Sigma u_i f_i)^2 \left(\frac{1}{n}\right)$$

$$= \Sigma u_i^2 f_i - \left(\frac{1}{n}\right) (\Sigma u_i f_i)^2$$

therefore,

$$s = C \sqrt{\frac{1}{n-1} \left[\sum_{i=1}^{n} u_i^2 f_i - \frac{1}{n} \left(\sum_{i=1}^{n} u_i f_i \right)^2 \right]}$$

The previous example is solved by writing

X	f	u	uf	$u^2 f$
80	1	+2	+2	4
75	2	+1	+2	2
70	1	0	0	0
65	3	-1	-3	3
60	1	-2	-2	4
			-1	13

and

$$s = C \sqrt{\frac{1}{n-1} \left[\Sigma u_i^2 f_i - \frac{1}{n} (\Sigma u_i f_i)^2 \right]}$$

$$= 5 \sqrt{\frac{1}{7} \left[13 - \frac{1}{8} (1) \right]}$$

$$= 5 \sqrt{\frac{1}{7} \left(12 \frac{7}{8} \right)}$$

$$s = 5\sqrt{\frac{1}{7}\ \frac{(103)}{8}}$$

$$= 5\sqrt{\frac{103}{56}}$$

$$= 5\sqrt{1.84}$$

$$= 5\ (1.36)$$

$$= 6.80$$

EXAMPLE: Find the standard deviation of the values

X	f	u	uf	u^2 f
80	3	+2	6	12
70	4	+1	4	4
60	7	0	0	0
50	6	-1	-6	6
40	4	-2	-8	16
30	2	-3	-6	18
	26		-10	56

SOLUTION: Write

$$s = C\sqrt{\frac{1}{n-1}\left[\Sigma\, u_i^2 f_i - \frac{1}{n}\left(\Sigma\, u_i f_i\right)^2\right]}$$

$$= 10\sqrt{\frac{1}{25}\left[56 - \frac{1}{26}\ (100)\right]}$$

$$= 10\sqrt{\frac{1}{25}\left(56 - \frac{100}{26}\right)}$$

$$= 10\sqrt{2.086}$$

$$= 10\ (1.444)$$

$$= 14.4$$

PROBLEM: Find the standard deviation, by coding, of

X	f
86	1
81	3
76	11
71	13
66	9
61	4
56	2

ANSWER: Approximately 6.6.

When calculating the standard deviation of ungrouped or raw values, we may use, instead of

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

the formula

$$s = \sqrt{\frac{\Sigma\, X^2 - \frac{(\Sigma\, X)^2}{n}}{n-1}}$$

where X is the symbol representing the original values.

EXAMPLE: Find the standard deviation of the values 80, 75, 75, 70, 65, 65, 65, and 60.

SOLUTION: Write

	X	X^2
	80	6400
	75	5625
	75	5625
	70	4900
	65	4225
	65	4225
	65	4225
	60	3600
Totals	555	38825

then,

$$s = \sqrt{\frac{\Sigma\, X^2 - \frac{(\Sigma\, X)^2}{n}}{n-1}}$$

$$= \sqrt{\frac{38825 - \frac{(555)^2}{8}}{7}}$$

$$= \sqrt{\frac{38825 - 38503}{7}}$$

$$= \sqrt{\frac{322}{7}}$$

$$= \sqrt{46}$$

$$= 6.78$$

which agrees with a previous problem in which we used the formula

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 f_i}$$

Notice that f_i is included in this formula. We could have grouped our values and used the formula

$$s = \sqrt{\frac{\Sigma X^2 f_i - \frac{(\Sigma X f_i)^2}{n}}{n-1}}$$

which is the formula for grouped values.

EXAMPLE: Compare the standard deviation of the values 82, 80, 80, 78, 77, 66, and 62 found by both

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 f_i}$$

and

$$s = \sqrt{\frac{\Sigma X^2 f_i - \frac{(\Sigma X f_i)^2}{n}}{n-1}}$$

SOLUTION: Write

X	f	Xf	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^2 f$	$X^2 f_i$
82	1	82	+7	49	49	6724
80	2	160	+5	25	50	12800
78	1	78	+3	9	9	6084
77	1	77	+2	4	4	5929
66	1	66	-9	81	81	4356
62	1	62	-13	169	169	3844
Totals		525			362	39737

where

$$\bar{X} = \frac{1}{n} \Sigma X_i f_i$$

$$= \frac{1}{7} (525)$$

$$= 75$$

Then,

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 f_i}$$

$$= \sqrt{\frac{1}{6} (362)}$$

$$= \sqrt{\frac{362}{6}}$$

$$= \sqrt{60.3}$$

$$= 7.76$$

and

$$s = \sqrt{\frac{\Sigma X^2 f_i - \frac{(\Sigma X f_i)^2}{n}}{n-1}}$$

$$= \sqrt{\frac{39737 - \frac{275625}{7}}{6}}$$

$$= \sqrt{\frac{362}{6}}$$

$$= \sqrt{60.3}$$

$$= 7.76$$