

## ***The Standard Deviation and Other Measures of Dispersion***

### **DISPERSION, OR VARIATION**

The degree to which numerical data tend to spread about an average value is called the *dispersion*, or *variation*, of the data. Various measures of this dispersion (or variation) are available, the most common being the range, mean deviation, semi-interquartile range, 10–90 percentile range, and standard deviation.

### **THE RANGE**

The *range* of a set of numbers is the difference between the largest and smallest numbers in the set.

**EXAMPLE 1.** The range of the set 2, 3, 3, 5, 5, 5, 8, 10, 12 is  $12 - 2 = 10$ . Sometimes the range is given by simply quoting the smallest and largest numbers; in the above set, for instance, the range could be indicated as 2 to 12, or 2–12.

### **THE MEAN DEVIATION**

The *mean deviation*, or *average deviation*, of a set of  $N$  numbers  $X_1, X_2, \dots, X_N$  is abbreviated MD and is defined by

$$\text{Mean deviation (MD)} = \frac{\sum_{j=1}^N |X_j - \bar{X}|}{N} = \frac{\sum |X - \bar{X}|}{N} = \overline{|X - \bar{X}|} \quad (1)$$

where  $\bar{X}$  is the arithmetic mean of the numbers and  $|X_j - \bar{X}|$  is the absolute value of the deviation of  $X_j$  from  $\bar{X}$ . (The *absolute value* of a number is the number without the associated sign and is indicated by two vertical lines placed around the number; thus  $|-4| = 4$ ,  $|+3| = 3$ ,  $|6| = 6$ , and  $|-0.84| = 0.84$ .)

**EXAMPLE 2.** Find the mean deviation of the set 2, 3, 6, 8, 11.

$$\text{Arithmetic mean } (\bar{X}) = \frac{2+3+6+8+11}{5} = 6$$

$$\text{MD} = \frac{|2-6| + |3-6| + |6-6| + |8-6| + |11-6|}{5} = \frac{|-4| + |-3| + |0| + |2| + |5|}{5} = \frac{4+3+0+2+5}{5} = 2.8$$

If  $X_1, X_2, \dots, X_K$  occur with frequencies  $f_1, f_2, \dots, f_K$ , respectively, the mean deviation can be written as

$$\text{MD} = \frac{\sum_{j=1}^K f_j |X_j - \bar{X}|}{N} = \frac{\sum f |X - \bar{X}|}{N} = \overline{|X - \bar{X}|} \quad (2)$$

where  $N = \sum_{j=1}^K f_j = \sum f$ . This form is useful for grouped data, where the  $X_j$ 's represent class marks and the  $f_j$ 's are the corresponding class frequencies.

Occasionally the mean deviation is defined in terms of absolute deviations from the median or other average instead of from the mean. An interesting property of the sum  $\sum_{j=1}^N |X_j - a|$  is that it is a minimum when  $a$  is the median (i.e., the mean deviation about the median is a minimum).

Note that it would be more appropriate to use the terminology *mean absolute deviation* than *mean deviation*.

## THE SEMI-INTERQUARTILE RANGE

The *semi-interquartile range*, or *quartile deviation*, of a set of data is denoted by  $Q$  and is defined by

$$Q = \frac{Q_3 - Q_1}{2} \quad (3)$$

where  $Q_1$  and  $Q_3$  are the first and third quartiles for the data (see Problems 4.6 and 4.7). The interquartile range  $Q_3 - Q_1$  is sometimes used, but the semi-interquartile range is more common as a measure of dispersion.

## THE 10-90 PERCENTILE RANGE

The *10-90 percentile range* of a set of data is defined by

$$10-90 \text{ percentile range} = P_{90} - P_{10} \quad (4)$$

where  $P_{10}$  and  $P_{90}$  are the 10th and 90th percentiles for the data (see Problem 4.8). The semi-10-90 percentile range,  $\frac{1}{2}(P_{90} - P_{10})$ , can also be used but is not commonly employed.

## THE STANDARD DEVIATION

The *standard deviation* of a set of  $N$  numbers  $X_1, X_2, \dots, X_N$  is denoted by  $s$  and is defined by

$$s = \sqrt{\frac{\sum_{j=1}^N (X_j - \bar{X})^2}{N}} = \sqrt{\frac{\sum (X - \bar{X})^2}{N}} = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\overline{(X - \bar{X})^2}} \quad (5)$$

where  $x$  represents the deviations of each of the numbers  $X_j$  from the mean  $\bar{X}$ . Thus  $s$  is the root mean square (RMS) of the deviations from the mean, or, as it is sometimes called, the *root-mean-square deviation*.

If  $X_1, X_2, \dots, X_K$  occur with frequencies  $f_1, f_2, \dots, f_K$ , respectively, the standard deviation can be written

$$s = \sqrt{\frac{\sum_{j=1}^K f_j (X_j - \bar{X})^2}{N}} = \sqrt{\frac{\sum f(X - \bar{X})^2}{N}} = \sqrt{\frac{\sum f x^2}{N}} = \sqrt{(X - \bar{X})^2} \quad (6)$$

where  $N = \sum_{j=1}^K f_j = \sum f$ . In this form it is useful for grouped data.

Sometimes the standard deviation of a sample's data is defined with  $(N - 1)$  replacing  $N$  in the denominators of the expressions in equations (5) and (6) because the resulting value represents a better estimate of the standard deviation of a population from which the sample is taken. For large values of  $N$  (certainly  $N > 30$ ), there is practically no difference between the two definitions. Also, when the better estimate is needed we can always obtain it by multiplying the standard deviation computed according to the first definition by  $\sqrt{N/(N - 1)}$ . Hence we shall adhere to the form (5) and (6).

## THE VARIANCE

The *variance* of a set of data is defined as the square of the standard deviation and is thus given by  $s^2$  in equations (5) and (6).

When it is necessary to distinguish the standard deviation of a population from the standard deviation of a sample drawn from this population, we often use the symbol  $s$  for the latter and  $\sigma$  (lowercase Greek *sigma*) for the former. Thus  $s^2$  and  $\sigma^2$  would represent the *sample variance* and *population variance*, respectively.

## SHORT METHODS FOR COMPUTING THE STANDARD DEVIATION

Equations (5) and (6) can be written, respectively, in the equivalent forms

$$s = \sqrt{\frac{\sum_{j=1}^N X_j^2}{N} - \left(\frac{\sum_{j=1}^N X_j}{N}\right)^2} = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\overline{X^2} - \bar{X}^2} \quad (7)$$

$$s = \sqrt{\frac{\sum_{j=1}^K f_j X_j^2}{N} - \left(\frac{\sum_{j=1}^K f_j X_j}{N}\right)^2} = \sqrt{\frac{\sum f X^2}{N} - \left(\frac{\sum f X}{N}\right)^2} = \sqrt{\overline{X^2} - \bar{X}^2} \quad (8)$$

where  $\overline{X^2}$  denotes the mean of the squares of the various values of  $X$ , while  $\bar{X}^2$  denotes the square of the mean of the various values of  $X$  (see Problems 4.12 to 4.14).

If  $d_j = X_j - A$  are the deviations of  $X_j$  from some arbitrary constant  $A$ , results (7) and (8) become, respectively,

$$s = \sqrt{\frac{\sum_{j=1}^N d_j^2}{N} - \left(\frac{\sum_{j=1}^N d_j}{N}\right)^2} = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2} = \sqrt{\overline{d^2} - \bar{d}^2} \quad (9)$$

$$s = \sqrt{\frac{\sum_{j=1}^K f_j d_j^2}{N} - \left(\frac{\sum_{j=1}^K f_j d_j}{N}\right)^2} = \sqrt{\frac{\sum f d^2}{N} - \left(\frac{\sum f d}{N}\right)^2} = \sqrt{\overline{d^2} - \bar{d}^2} \quad (10)$$

(See Problems 4.15 and 4.17.)

When data are grouped into a frequency distribution whose class intervals have equal size  $c$ , we have  $d_j = cu_j$  or  $X_j = A + cu_j$  and result (10) becomes

$$s = c \sqrt{\frac{\sum_{j=1}^K f_j u_j^2}{N} - \left( \frac{\sum_{j=1}^K f_j u_j}{N} \right)^2} = c \sqrt{\frac{\sum f u^2}{N} - \left( \frac{\sum f u}{N} \right)^2} = c \sqrt{\bar{u}^2 - \bar{u}^2} \quad (11)$$

This last formula provides a very short method for computing the standard deviation and should always be used for grouped data when the class-interval sizes are equal. It is called the *coding method* and is exactly analogous to that used in Chapter 3 for computing the arithmetic mean of grouped data. (See Problems 4.16 to 4.19.)

## PROPERTIES OF THE STANDARD DEVIATION

1. The standard deviation can be defined as

$$s = \sqrt{\frac{\sum_{j=1}^N (X_j - a)^2}{N}}$$

where  $a$  is an average besides the arithmetic mean. Of all such standard deviations, the minimum is that for which  $a = \bar{X}$ , because of Property 2 in Chapter 3. This property provides an important reason for defining the standard deviation as above. For a proof of this property, see Problem 4.27.

2. For normal distributions (see Chapter 7), it turns out that (as shown in Fig. 4-1):
  - (a) 68.27% of the cases are included between  $\bar{X} - s$  and  $\bar{X} + s$  (i.e., one standard deviation on either side of the mean).
  - (b) 95.45% of the cases are included between  $\bar{X} - 2s$  and  $\bar{X} + 2s$  (i.e., two standard deviations on either side of the mean).
  - (c) 99.73% of the cases are included between  $\bar{X} - 3s$  and  $\bar{X} + 3s$  (i.e., three standard deviations on either side of the mean).

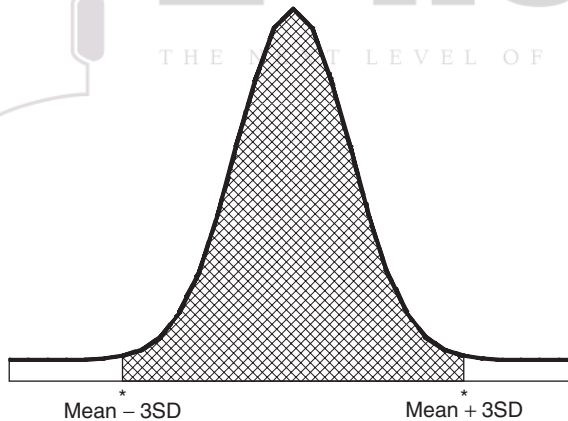
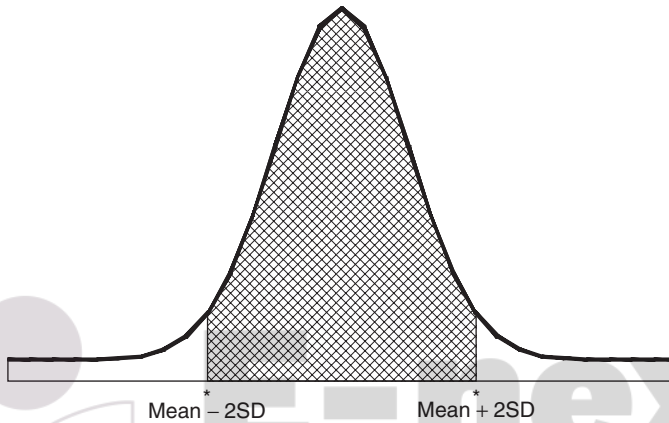
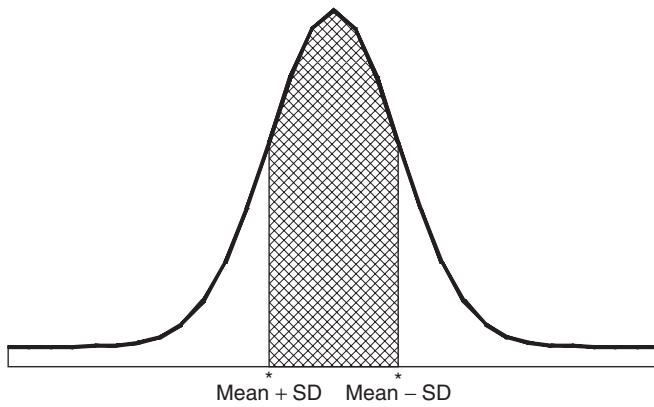
For moderately skewed distributions, the above percentages may hold approximately (see Problem 4.24).

3. Suppose that two sets consisting of  $N_1$  and  $N_2$  numbers (or two frequency distributions with total frequencies  $N_1$  and  $N_2$ ) have variances given by  $s_1^2$  and  $s_2^2$ , respectively, and have the *same* mean  $\bar{X}$ . Then the *combined*, or *pooled*, *variance* of both sets (or both frequency distributions) is given by

$$s^2 = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2} \quad (12)$$

Note that this is a weighted arithmetic mean of the variances. This result can be generalized to three or more sets.

4. Chebyshev's theorem states that for  $k > 1$ , there is at least  $(1 - (1/k^2)) \times 100\%$  of the probability distribution for any variable within  $k$  standard deviations of the mean. In particular, when  $k = 2$ , there is at least  $(1 - (1/2^2)) \times 100\%$  or 75% of the data in the interval  $(\bar{x} - 2S, \bar{x} + 2S)$ , when  $k = 3$  there is at least  $(1 - (1/3^2)) \times 100\%$  or 89% of the data in the interval  $(\bar{x} - 3S, \bar{x} + 3S)$ , and when  $k = 4$  there is at least  $(1 - (1/4^2)) \times 100\%$  or 93.75% of the data in the interval  $(\bar{x} - 4S, \bar{x} + 4S)$ .



**Fig. 4-1** Illustration of the empirical rule.

### CHARLIER'S CHECK

Charlier's check in computations of the mean and standard deviation by the coding method makes use of the identities

$$\sum f(u+1) = \sum fu + \sum f = \sum fu + N$$

$$\sum f(u+1)^2 = \sum f(u^2 + 2u + 1) = \sum fu^2 + 2\sum fu + \sum f = \sum fu^2 + 2\sum fu + N$$

(See Problem 4.20.)

## SHEPPARD'S CORRECTION FOR VARIANCE

The computation of the standard deviation is somewhat in error as a result of grouping the data into classes (grouping error). To adjust for grouping error, we use the formula

$$\text{Corrected variance} = \text{variance from grouped data} - \frac{c^2}{12} \quad (13)$$

where  $c$  is the class-interval size. The correction  $c^2/12$  (which is subtracted) is called *Sheppard's correction*. It is used for distributions of continuous variables where the “tails” go gradually to zero in both directions.

Statisticians differ as to *when* and *whether* Sheppard's correction should be applied. It should certainly not be applied before one examines the situation thoroughly, for it often tends to *overcorrect*, thus replacing an old error with a new one. In this book, unless otherwise indicated, we shall not be using Sheppard's correction.

## EMPIRICAL RELATIONS BETWEEN MEASURES OF DISPERSION

For moderately skewed distributions, we have the empirical formulas

$$\text{Mean deviation} = \frac{4}{5}(\text{standard deviation})$$

$$\text{Semi-interquartile range} = \frac{2}{3}(\text{standard deviation})$$

These are consequences of the fact that for the normal distribution we find that the mean deviation and semi-interquartile range are equal, respectively, to 0.7979 and 0.6745 times the standard deviation.

## ABSOLUTE AND RELATIVE DISPERSION; COEFFICIENT OF VARIATION

The actual variation, or dispersion, as determined from the standard deviation or other measure of dispersion is called the *absolute dispersion*. However, a variation (or dispersion) of 10 inches (in) in measuring a distance of 1000 feet (ft) is quite different in effect from the same variation of 10 in in a distance of 20 ft. A measure of this effect is supplied by the *relative dispersion*, which is defined by

$$\text{Relative dispersion} = \frac{\text{absolute dispersion}}{\text{average}} \quad (14)$$

If the absolute dispersion is the standard deviation  $s$  and if the average is the mean  $\bar{X}$ , then the relative dispersion is called the *coefficient of variation*, or *coefficient of dispersion*; it is denoted by  $V$  and is given by

$$\text{Coefficient of variation } (V) = \frac{s}{\bar{X}} \quad (15)$$

and is generally expressed as a percentage. Other possibilities also occur (see Problem 4.30).

Note that the coefficient of variation is independent of the units used. For this reason, it is useful in comparing distributions where the units may be different. A disadvantage of the coefficient of variation is that it fails to be useful when  $\bar{X}$  is close to zero.

STANDARDIZED VARIABLE; STANDARD SCORES

The variable that measures the deviation from the mean in units of the standard deviation is called a *standardized variable*, is a dimensionless quantity (i.e., is independent of the units used), and is given by

$$z = \frac{X - \bar{X}}{s} \tag{16}$$

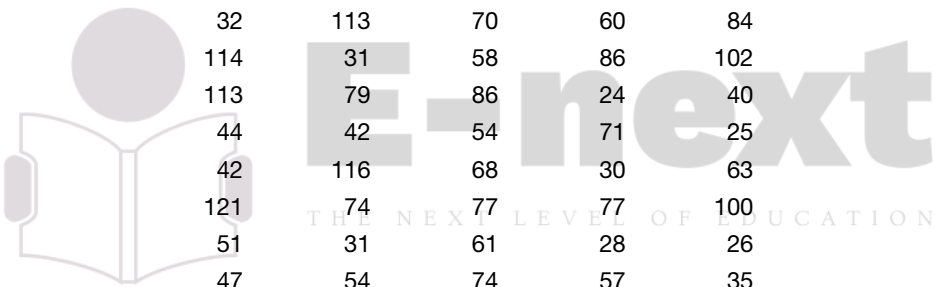
If the deviations from the mean are given in units of the standard deviation, they are said to be expressed in *standard units*, or *standard scores*. These are of great value in the comparison of distributions (see Problem 4.31).

SOFTWARE AND MEASURES OF DISPERSION

The statistical software gives a variety of measures for dispersion. The dispersion measures are usually given in descriptive statistics. EXCEL allows for the computation of all the measures discussed in this book. MINITAB and EXCEL are discussed here and outputs of other packages are given in the solved problems.

EXAMPLE 3.

- (a) EXCEL provides calculations for several measures of dispersion. The following example illustrates several. A survey was taken at a large company and the question asked was how many e-mails do you send per week? The results for 75 employees are shown in A1:E15 of an EXCEL worksheet.



32	113	70	60	84
114	31	58	86	102
113	79	86	24	40
44	42	54	71	25
42	116	68	30	63
121	74	77	77	100
51	31	61	28	26
47	54	74	57	35
77	80	125	105	61
102	45	115	36	52
58	24	24	39	40
95	99	54	35	31
77	29	69	58	32
49	118	44	95	65
71	65	74	122	99

The range is given by =MAX(A1:E15)-MIN(A1:E15) or 125 – 24 = 101. The mean deviation or average deviation is given by =AVEDEV(A1:E15) or 24.42. The semi-interquartile range is given by the expression =(PERCENTILE(A1:E15,0.75)-PERCENTILE(A1:E15,0.25))/2 or 22. The 10–90 percentile range is given by PERCENTILE(A1:E15,0.9)-PERCENTILE(A1:E15,0.1) or 82.6.

The standard deviation and variance is given by =STDEV(A1:E15) or 29.2563, and =VAR(A1:E15) or 855.932 for samples and =STDEVP(A1:E15) or 29.0606 and =VARP(A1:E15) or 844.52 for populations.

(b)

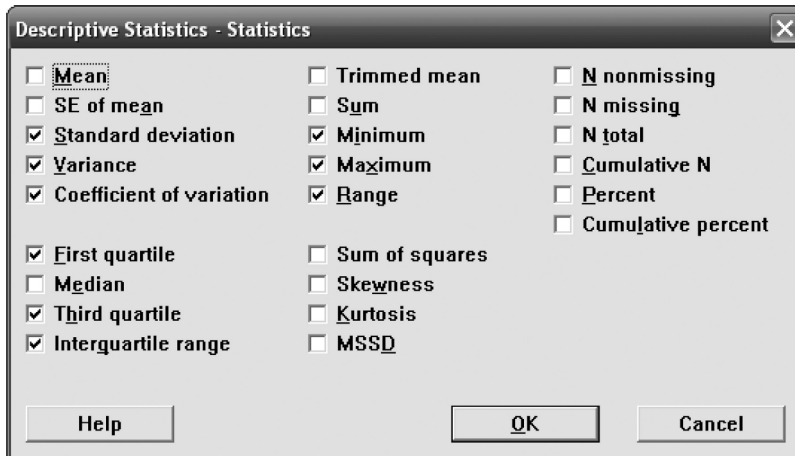


Fig. 4-2 Dialog box for MINITAB.

The dialog box in Fig. 4-2 shows MINITAB choices for measures of dispersion and measures of central tendency. Output is as follows:

#### Descriptive Statistics: e-mails

Variable	StDev	Variance	CoefVar	Minimum	Q1	Q3	Maximum	Range	IQR
e-mails	29.26	855.93	44.56	24.00	40.00	86.00	125.00	101.00	46.00



**E-next**  
THE NEXT LEVEL OF EDUCATION

## Solved Problems

### THE RANGE

**4.1** Find the range of the sets (a) 12, 6, 7, 3, 15, 10, 18, 5 and (b) 9, 3, 8, 8, 9, 8, 9, 18.

#### SOLUTION

In both cases, range = largest number – smallest number =  $18 - 3 = 15$ . However, as seen from the arrays of sets (a) and (b),

$$(a) \quad 3, 5, 6, 7, 10, 12, 15, 18 \quad (b) \quad 3, 8, 8, 8, 9, 9, 9, 18$$

there is much more variation, or dispersion, in (a) than in (b). In fact, (b) consists mainly of 8's and 9's.

Since the range indicates no difference between the sets, it is not a very good measure of dispersion in this case. Where extreme values are present, the range is generally a poor measure of dispersion.

An improvement is achieved by throwing out the extreme cases, 3 and 18. Then for set (a) the range is  $(15 - 5) = 10$ , while for set (b) the range is  $(9 - 8) = 1$ , clearly showing that (a) has greater dispersion than (b). However, this is not the way the range is defined. The semi-interquartile range and the 10–90 percentile range were designed to improve on the range by eliminating extreme cases.

**4.2** Find the range of heights of the students at XYZ University as given in Table 2.1.



## SOLUTION

There are two ways of defining the range for grouped data.

### First method

$$\begin{aligned}\text{Range} &= \text{class mark of highest class} - \text{class mark of lowest class} \\ &= 73 - 61 = 12 \text{ in}\end{aligned}$$

### Second method

$$\begin{aligned}\text{Range} &= \text{upper class boundary of highest class} - \text{lower class boundary of lowest class} \\ &= 74.5 - 59.5 = 15 \text{ in}\end{aligned}$$

The first method tends to eliminate extreme cases to some extent.

## THE MEAN DEVIATION

**4.3** Find the mean deviation of the sets of numbers in Problem 4.1.

### SOLUTION

(a) The arithmetic mean is

$$\bar{X} = \frac{12 + 6 + 7 + 3 + 15 + 10 + 18 + 5}{8} = \frac{76}{8} = 9.5$$

The mean deviation is

$$\begin{aligned}\text{MD} &= \frac{\sum |X - \bar{X}|}{N} \\ &= \frac{|12 - 9.5| + |6 - 9.5| + |7 - 9.5| + |3 - 9.5| + |15 - 9.5| + |10 - 9.5| + |18 - 9.5| + |5 - 9.5|}{8} \\ &= \frac{2.5 + 3.5 + 2.5 + 6.5 + 5.5 + 0.5 + 8.5 + 4.5}{8} = \frac{34}{8} = 4.25\end{aligned}$$

(b) 
$$\bar{X} = \frac{9 + 3 + 8 + 8 + 9 + 8 + 9 + 18}{8} = \frac{72}{8} = 9$$

$$\begin{aligned}\text{MD} &= \frac{\sum |X - \bar{X}|}{N} \\ &= \frac{|9 - 9| + |3 - 9| + |8 - 9| + |8 - 9| + |9 - 9| + |8 - 9| + |9 - 9| + |18 - 9|}{8} \\ &= \frac{0 + 6 + 1 + 1 + 0 + 1 + 0 + 9}{8} = 2.25\end{aligned}$$

The mean deviation indicates that set (b) shows less dispersion than set (a), as it should.

**4.4** Find the mean deviation of the heights of the 100 male students at XYZ University (see Table 3.2 of Problem 3.20).

### SOLUTION

From Problem 3.20,  $\bar{X} = 67.45$  in. The work can be arranged as in Table 4.1. It is also possible to devise a coding method for computing the mean deviation (see Problem 4.47).

Table 4.1

Height (in)	Class Mark ( $X$ )	$ X - \bar{X}  =  X - 67.45 $	Frequency ( $f$ )	$f X - \bar{X} $
60–62	61	6.45	5	32.25
63–65	64	3.45	18	62.10
66–68	67	0.45	42	18.90
69–71	70	2.55	27	68.85
72–74	73	5.55	8	44.40
			$N = \sum f = 100$	$\sum f X - \bar{X}  = 226.50$

$$MD = \frac{\sum f|X - \bar{X}|}{N} = \frac{226.50}{100} = 2.26 \text{ in}$$

- 4.5** Determine the percentage of the students' heights in Problem 4.4 that fall within the ranges (a)  $\bar{X} \pm MD$ , (b)  $\bar{X} \pm 2MD$ , and (c)  $\bar{X} \pm 3MD$ .

**SOLUTION**

- (a) The range from 65.19 to 69.71 in is  $\bar{X} \pm MD = 67.45 \pm 2.26$ . This range includes all individuals in the third class  $+\frac{1}{3}(65.5 - 65.19)$  of the students in the second class  $+\frac{1}{3}(69.71 - 68.5)$  of the students in the fourth class (since the class-interval size is 3 in, the upper class boundary of the second class is 65.5 in, and the lower class boundary of the fourth class is 68.5 in). The number of students in the range  $\bar{X} \pm MD$  is

$$42 + \frac{0.31}{3}(18) + \frac{1.21}{3}(27) = 42 + 1.86 + 10.89 = 54.75 \quad \text{or} \quad 55$$

which is 55% of the total.

- (b) The range from 62.93 to 71.97 in is  $\bar{X} \pm 2MD = 67.45 \pm 2(2.26) = 67.45 \pm 4.52$ . The number of students in the range  $\bar{X} \pm 2MD$  is

$$18 - \left(\frac{62.93 - 62.5}{3}\right)(18) + 42 + 27 + \left(\frac{71.97 - 71.5}{3}\right)(8) = 85.67 \quad \text{or} \quad 86$$

which is 86% of the total.

- (c) The range from 60.67 to 74.23 in is  $\bar{X} \pm 3MD = 67.45 \pm 3(2.26) = 67.45 \pm 6.78$ . The number of students in the range  $\bar{X} \pm 3MD$  is

$$5 - \left(\frac{60.67 - 59.5}{3}\right)(5) + 18 + 42 + 27 + \left(\frac{74.5 - 74.23}{3}\right)(8) = 97.33 \quad \text{or} \quad 97$$

which is 97% of the total.

**THE SEMI-INTERQUARTILE RANGE**

- 4.6** Find the semi-interquartile range for the height distribution of the students at XYZ University (see Table 4.1 of Problem 4.4).

**SOLUTION**

The lower and upper quartiles are  $Q_1 = 65.5 + \frac{2}{42}(3) = 65.64$  in and  $Q_3 = 68.5 + \frac{10}{27}(3) = 69.61$  in, respectively, and the semi-interquartile range (or quartile deviation) is  $Q = \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2}(69.61 - 65.64) = 1.98$  in. Note that 50% of the cases lie between  $Q_1$  and  $Q_3$  (i.e., 50 students have heights between 65.64 and 69.61 in).

We can consider  $\frac{1}{2}(Q_1 + Q_3) = 67.63$  in to be a measure of central tendency (i.e., average height). It follows that 50% of the heights lie in the range  $67.63 \pm 1.98$  in.

- 4.7** Find the semi-interquartile range for the wages of the 65 employees at the P&R Company (see Table 2.5 of Problem 2.3).

**SOLUTION**

From Problem 3.44,  $Q_1 = \$268.25$  and  $Q_3 = \$290.75$ . Thus the semi-interquartile range  $Q = \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2}(\$290.75 - \$268.25) = \$11.25$ . Since  $\frac{1}{2}(Q_1 + Q_3) = \$279.50$ , we can conclude that 50% of the employees earn wages lying in the range  $\$279.50 \pm \$11.25$ .

**THE 10-90 PERCENTILE RANGE**

- 4.8** Find the 10-90 percentile range of the heights of the students at XYZ University (see Table 2.1).

**SOLUTION**

Here  $P_{10} = 62.5 + \frac{5}{18}(3) = 63.33$  in, and  $P_{90} = 68.5 + \frac{25}{27}(3) = 71.27$  in. Thus the 10-90 percentile range is  $P_{90} - P_{10} = 71.27 - 63.33 = 7.94$  in. Since  $\frac{1}{2}(P_{10} + P_{90}) = 67.30$  in and  $\frac{1}{2}(P_{90} - P_{10}) = 3.97$  in, we can conclude that 80% of the students have heights in the range  $67.30 \pm 3.97$  in.

**THE STANDARD DEVIATION**

- 4.9** Find the standard deviation  $s$  of each set of numbers in Problem 4.1.

**SOLUTION**

$$(a) \quad \bar{X} = \frac{\sum X}{N} = \frac{12 + 6 + 7 + 3 + 15 + 10 + 18 + 5}{8} = \frac{76}{8} = 9.5$$

$$\begin{aligned} s &= \sqrt{\frac{\sum (X - \bar{X})^2}{N}} \\ &= \sqrt{\frac{(12 - 9.5)^2 + (6 - 9.5)^2 + (7 - 9.5)^2 + (3 - 9.5)^2 + (15 - 9.5)^2 + (10 - 9.5)^2 + (18 - 9.5)^2 + (5 - 9.5)^2}{8}} \\ &= \sqrt{23.75} = 4.87 \end{aligned}$$

$$(b) \quad \bar{X} = \frac{9 + 3 + 8 + 8 + 9 + 8 + 9 + 18}{8} = \frac{72}{8} = 9$$

$$\begin{aligned} s &= \sqrt{\frac{\sum (X - \bar{X})^2}{N}} \\ &= \sqrt{\frac{(9 - 9)^2 + (3 - 9)^2 + (8 - 9)^2 + (8 - 9)^2 + (9 - 9)^2 + (8 - 9)^2 + (9 - 9)^2 + (18 - 9)^2}{8}} \\ &= \sqrt{15} = 3.87 \end{aligned}$$

The above results should be compared with those of Problem 4.3. It will be noted that the standard deviation does indicate that set (b) shows less dispersion than set (a). However, the effect is masked by the fact that extreme values affect the standard deviation much more than they affect the mean deviation. This is to be expected, of course, since the deviations are squared in computing the standard deviation.

- 4.10** The standard deviations of the two data sets given in Problem 4.1 were found using MINITAB and the results are shown below. Compare the answers with those obtained in Problem 4.9.

```
MTB > print c1
set1
    12    6    7    3    15    10    18    5
MTB > print c2
set2
    9    3    8    8    9    8    9    18
MTB > standard deviation c1
```

#### Column Standard Deviation

Standard deviation of set1 = 5.21

```
MTB > standard deviation c2
```

#### Column Standard Deviation

Standard deviation of set2 = 4.14

### SOLUTION

The MINITAB package uses the formula

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}}$$

and therefore the standard deviations are not the same in Problems 4.9 and 4.10. The answers in Problem 4.10 are obtainable from those in Problem 4.9 if we multiply those in Problem 4.9 by  $\sqrt{N/(N-1)}$ . Since  $N = 8$  for both sets  $\sqrt{N/(N-1)} = 1.069045$ , and for set 1, we have  $(1.069045)(4.87) = 5.21$ , the standard deviation given by MINITAB. Similarly,  $(1.069045)(3.87) = 4.14$ , the standard deviation given for set 2 by MINITAB.

- 4.11** Find the standard deviation of the heights of the 100 male students at XYZ University (see Table 2.1).

### SOLUTION

From Problem 3.15, 3.20, or 3.22,  $\bar{X} = 67.45$  in. The work can be arranged as in Table 4.2.

$$s = \sqrt{\frac{\sum f(X - \bar{X})^2}{N}} = \sqrt{\frac{852.7500}{100}} = \sqrt{8.5275} = 2.92 \text{ in}$$

**Table 4.2**

Height (in)	Class Mark ( $X$ )	$X - \bar{X} = X - 67.45$	$(X - \bar{X})^2$	Frequency ( $f$ )	$f(X - \bar{X})^2$
60–62	61	–6.45	41.6025	5	208.0125
63–65	64	–3.45	11.9025	18	214.2450
66–68	67	–0.45	0.2025	42	8.5050
69–71	70	2.55	6.5025	27	175.5675
72–74	73	5.55	30.8025	8	246.4200
				$N = \sum f = 100$	$\sum f(X - \bar{X})^2 = 852.7500$

## COMPUTING THE STANDARD DEVIATIONS FROM GROUPED DATA

**4.12** (a) Prove that

$$s = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\bar{X^2} - \bar{X}^2}$$

(b) Use the formula in part (a) to find the standard deviation of the set 12, 6, 7, 3, 15, 10, 18, 5.

### SOLUTION

(a) By definition,

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

$$\begin{aligned}\text{Then } s^2 &= \frac{\sum (X - \bar{X})^2}{N} = \frac{\sum (X^2 - 2\bar{X}X + \bar{X}^2)}{N} = \frac{\sum X^2 - 2\bar{X}\sum X + N\bar{X}^2}{N} \\ &= \frac{\sum X^2}{N} - 2\bar{X} \frac{\sum X}{N} + \bar{X}^2 = \frac{\sum X^2}{N} - 2\bar{X}^2 + \bar{X}^2 = \frac{\sum X^2}{N} - \bar{X}^2 \\ &= \bar{X^2} - \bar{X}^2 = \frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2\end{aligned}$$

or

$$s = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\bar{X^2} - \bar{X}^2}$$

Note that in the above summations we have used the abbreviated form, with  $X$  replacing  $X_j$  and with  $\sum$  replacing  $\sum_{j=1}^N$ .

### Another method

$$\begin{aligned}s^2 &= \overline{(X - \bar{X})^2} = \overline{X^2 - 2X\bar{X} + \bar{X}^2} = \overline{X^2} - \overline{2X\bar{X}} + \overline{\bar{X}^2} = \overline{X^2} - 2\bar{X}\bar{X} + \bar{X}^2 = \overline{X^2} - \bar{X}^2 \\ (b) \quad \overline{X^2} &= \frac{\sum X^2}{N} = \frac{(12)^2 + (6)^2 + (7)^2 + (3)^2 + (15)^2 + (10)^2 + (18)^2 + (5)^2}{8} = \frac{912}{8} = 114 \\ \bar{X} &= \frac{\sum X}{N} = \frac{12 + 6 + 7 + 3 + 15 + 10 + 18 + 5}{8} = \frac{76}{8} = 9.5\end{aligned}$$

$$\text{Thus } s = \sqrt{\bar{X^2} - \bar{X}^2} = \sqrt{114 - 90.25} = \sqrt{23.75} = 4.87$$

This method should be compared with that of Problem 4.9(a).

**4.13** Modify the formula of Problem 4.12(a) to allow for frequencies corresponding to the various values of  $X$ .

### SOLUTION

The appropriate modification is

$$s = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2} = \sqrt{\bar{X^2} - \bar{X}^2}$$

As in Problem 4.12(a), this can be established by starting with

$$s = \sqrt{\frac{\sum f(X - \bar{X})^2}{N}}$$

Then

$$\begin{aligned}
 s^2 &= \frac{\sum f(X - \bar{X})^2}{N} = \frac{\sum f(X^2 - 2\bar{X}X + \bar{X}^2)}{N} = \frac{\sum fX^2 - 2\bar{X}\sum fX + \bar{X}^2\sum f}{N} \\
 &= \frac{\sum fX^2}{N} - 2\bar{X}\frac{\sum fX}{N} + \bar{X}^2 = \frac{\sum fX^2}{N} - 2\bar{X}^2 + \bar{X}^2 = \frac{\sum fX^2}{N} - \bar{X}^2 \\
 &= \frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2
 \end{aligned}$$

or

$$s = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2}$$

Note that in the above summations we have used the abbreviated form, with  $X$  and  $f$  replacing  $X_j$  and  $f_j$ ,  $\sum$  replacing  $\sum_{j=1}^K$ , and  $\sum_{j=1}^K f_j = N$ .

**4.14** Using the formula of Problem 4.13, find the standard deviation for the data in Table 4.2 of Problem 4.11.

**SOLUTION**

The work can be arranged as in Table 4.3, where  $\bar{X} = (\sum fX)/N = 67.45$  in, as obtained in Problem 3.15. Note that this method, like that of Problem 4.11, entails much tedious computation. Problem 4.17 shows how the coding method simplifies the calculations immensely.

**Table 4.3**

Height (in)	Class Mark ( $X$ )	$X^2$	Frequency ( $f$ )	$fX^2$
60–62	61	3721	5	18,605
63–65	64	4096	18	73,728
66–68	67	4489	42	188,538
69–71	70	4900	27	132,300
72–74	73	5329	8	42,632
$N = \sum f = 100$				$\sum fX^2 = 455,803$

$$s = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2} = \sqrt{\frac{455,803}{100} - (67.45)^2} = \sqrt{8.5275} = 2.92 \text{ in}$$

**4.15** If  $d = X - A$  are the deviations of  $X$  from an arbitrary constant  $A$ , prove that

$$s = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

**SOLUTION**

Since  $d = X - A$ ,  $X = A + d$ , and  $\bar{X} = A + \bar{d}$  (see Problem 3.18), then

$$X - \bar{X} = (A + d) - (A + \bar{d}) = d - \bar{d}$$

so that

$$s = \sqrt{\frac{\sum f(X - \bar{X})^2}{N}} = \sqrt{\frac{\sum f(d - \bar{d})^2}{N}} = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

using the result of Problem 4.13 and replacing  $X$  and  $\bar{X}$  with  $d$  and  $\bar{d}$ , respectively.

**Another method**

$$\begin{aligned}
 s^2 &= \overline{(X - \bar{X})^2} = \overline{(d - \bar{d})^2} = \overline{d^2 - 2d\bar{d} + \bar{d}^2} \\
 &= \overline{d^2} - 2\bar{d}^2 + \bar{d}^2 = \overline{d^2} - \bar{d}^2 = \frac{\sum fd^2}{N} - \left( \frac{\sum fd}{N} \right)^2
 \end{aligned}$$

and the result follows on taking the positive square root.

- 4.16** Show that if each class mark  $X$  in a frequency distribution having class intervals of equal size  $c$  is coded into a corresponding value  $u$  according to the relation  $X = A + cu$ , where  $A$  is a given class mark, then the standard deviation can be written as

$$s = c \sqrt{\frac{\sum fu^2}{N} - \left( \frac{\sum fu}{N} \right)^2} = c \sqrt{\overline{u^2} - \bar{u}^2}$$

**SOLUTION**

This follows at once from Problem 4.15 since  $d = X - A = cu$ . Thus, since  $c$  is a constant,

$$s = \sqrt{\frac{\sum f(cu)^2}{N} - \left( \frac{\sum f(cu)}{N} \right)^2} = \sqrt{c^2 \frac{\sum fu^2}{N} - c^2 \left( \frac{\sum fu}{N} \right)^2} = c \sqrt{\frac{\sum fu^2}{N} - \left( \frac{\sum fu}{N} \right)^2}$$

**Another method**

We can also prove the result directly without using Problem 4.15. Since  $X = A + cu$ ,  $\bar{X} = A + c\bar{u}$ , and  $X - \bar{X} = c(u - \bar{u})$ , then

$$s^2 = \overline{(X - \bar{X})^2} = \overline{c^2(u - \bar{u})^2} = \overline{c^2(u^2 - 2u\bar{u} + \bar{u}^2)} = c^2(\overline{u^2} - 2\bar{u}^2 + \bar{u}^2) = c^2(\overline{u^2} - \bar{u}^2)$$

and

$$s = c \sqrt{\overline{u^2} - \bar{u}^2} = c \sqrt{\frac{\sum fu^2}{N} - \left( \frac{\sum fu}{N} \right)^2}$$

- 4.17** Find the standard deviation of the heights of the students at XYZ University (see Table 2.1) by using (a) the formula derived in Problem 4.15 and (b) the coding method of Problem 4.16.

**SOLUTION**

In Tables 4.4 and 4.5,  $A$  is arbitrarily chosen as being equal to the class mark 67. Note that in Table 4.4 the deviations  $d = X - A$  are all multiples of the class-interval size  $c = 3$ . This factor is removed in Table 4.5. As a result, the computations in Table 4.5 are greatly simplified (compare them with those of Problems 4.11 and 4.14). For this reason, the coding method should be used wherever possible.

(a) See Table 4.4.

**Table 4.4**

Class Mark ( $X$ )	$d = X - A$	Frequency ( $f$ )	$fd$	$fd^2$
61	-6	5	-30	180
64	-3	18	-54	162
$A \rightarrow 67$	0	42	0	0
70	3	27	81	243
73	6	8	48	288
		$N = \sum f = 100$	$\sum fd = 45$	$\sum fd^2 = 873$

$$s = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\frac{873}{100} - \left(\frac{45}{100}\right)^2} = \sqrt{8.5275} = 2.92 \text{ in}$$

(b) See Table 4.5.

**Table 4.5**

Class Mark ( $X$ )	$u = \frac{X - A}{c}$	Frequency ( $f$ )	$fu$	$fu^2$
61	-2	5	-10	20
64	-1	18	-18	18
$A \rightarrow 67$	0	42	0	0
70	1	27	27	27
73	2	8	16	32
		$N = \sum f = 100$	$\sum fu = 15$	$\sum fu^2 = 97$

$$s = c \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = 3 \sqrt{\frac{97}{100} - \left(\frac{15}{100}\right)^2} = 3 \sqrt{0.9475} = 2.92 \text{ in}$$

**4.18** Using coding methods, find (a) the mean and (b) the standard deviation for the wage distribution of the 65 employees at the P&R Company (see Table 2.5 of Problem 2.3).

**SOLUTION**

The work can be arranged simply, as shown in Table 4.6.

$$(a) \quad \bar{X} = A + c\bar{u} = A + c \frac{\sum fu}{N} = \$275.00 + (\$10.00) \left( \frac{31}{65} \right) = \$279.77$$

$$(b) \quad s = c \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = (\$10.00) \sqrt{\frac{173}{65} - \left(\frac{31}{65}\right)^2} = (\$10.00) \sqrt{2.4341} = \$15.60$$

**Table 4.6**

$X$	$u$	$f$	$fu$	$fu^2$
\$255.00	-2	8	-16	32
265.00	-1	10	-10	10
$A \rightarrow 275.00$	0	16	0	0
285.00	1	14	14	14
295.00	2	10	20	40
305.00	3	5	15	45
315.00	4	2	8	32
		$N = \sum f = 65$	$\sum fu = 31$	$\sum fu^2 = 173$

**4.19** Table 4.7 shows the IQ's of 480 school children at a certain elementary school. Using the coding method, find (a) the mean and (b) the standard deviation.



Table 4.7

Class mark ( $X$ )	70	74	78	82	86	90	94	98	102	106	110	114	118	122	126
Frequency ( $f$ )	4	9	16	28	45	66	85	72	54	38	27	18	11	5	2

**SOLUTION**

The intelligence quotient is

$$IQ = \frac{\text{mental age}}{\text{chronological age}}$$

expressed as a percentage. For example, an 8-year-old child who (according to certain educational procedures) has a mentality equivalent to that of a 10-year-old child would have an IQ of  $10/8 = 1.25 = 125\%$ , or simply 125, the % sign being understood.

To find the mean and standard deviation of the IQ's in Table 4.7, we can arrange the work as in Table 4.8.

$$(a) \quad \bar{X} = A + c\bar{u} = A + c \frac{\sum fu}{N} = 94 + 4 \left( \frac{236}{480} \right) = 95.97$$

$$(b) \quad s = c \sqrt{u^2 - \bar{u}^2} = c \sqrt{\frac{\sum fu^2}{N} - \left( \frac{\sum fu}{N} \right)^2} = 4 \sqrt{\frac{3404}{480} - \left( \frac{236}{480} \right)^2} = 4\sqrt{6.8499} = 10.47$$

**CHARLIER'S CHECK**

**4.20** Use Charlier's check to help verify the computations of (a) the mean and (b) the standard deviation performed in Problem 4.19.

**SOLUTION**

To supply the required check, the columns of Table 4.9 are added to those of Table 4.8 (with the exception of column 2, which is repeated in Table 4.9 for convenience).

(a) From Table 4.9,  $\sum f(u+1) = 716$ ; from Table 4.8,  $\sum fu + N = 236 + 480 = 716$ . This provides the required check on the mean.

Table 4.8

$X$	$u$	$f$	$fu$	$fu^2$
70	-6	4	-24	144
74	-5	9	-45	225
78	-4	16	-64	256
82	-3	28	-84	252
86	-2	45	-90	180
90	-1	66	-66	66
$A \rightarrow$ 94	0	85	0	0
98	1	72	72	72
102	2	54	108	216
106	3	38	114	342
110	4	27	108	432
114	5	18	90	450
118	6	11	66	396
122	7	5	35	245
126	8	2	16	128
		$N = \sum f = 480$	$\sum fu = 236$	$\sum fu^2 = 3404$

Table 4.9

$u + 1$	$f$	$f(u + 1)$	$f(u + 1)^2$
-5	4	-20	100
-4	9	-36	144
-3	16	-48	144
-2	28	-56	112
-1	45	-45	45
0	66	0	0
1	85	85	85
2	72	144	288
3	54	162	486
4	38	152	608
5	27	135	675
6	18	108	648
7	11	77	539
8	5	40	320
9	2	18	162
$N = \sum f = 480$		$\sum f(u + 1) = 716$	$\sum f(u + 1)^2 = 4356$

- (b) From Table 4.9,  $\sum f(u + 1)^2 = 4356$ ; from Table 4.8,  $\sum fu^2 + 2\sum fu + N = 3404 + 2(236) + 480 = 4356$ . This provides the required check on the standard deviation.

### SHEPPARD'S CORRECTION FOR VARIANCE

- 4.21** Apply Sheppard's correction to determine the standard deviation of the data in (a) Problem 4.17, (b) Problem 4.18, and (c) Problem 4.19.

#### SOLUTION

- (a)  $s^2 = 8.5275$ , and  $c = 3$ . Corrected variance  $= s^2 - c^2/12 = 8.5275 - 3^2/12 = 7.7775$ . Corrected standard deviation  $= \sqrt{\text{correct variance}} = \sqrt{7.7775} = 2.79$  in.
- (b)  $s^2 = 243.41$ , and  $c = 10$ . Corrected variance  $= s^2 - c^2/12 = 243.41 - 10^2/12 = 235.08$ . Corrected standard deviation  $= \sqrt{235.08} = \$15.33$ .
- (c)  $s^2 = 109.60$ , and  $c = 4$ . Corrected variance  $= s^2 - c^2/12 = 109.60 - 4^2/12 = 108.27$ . Corrected standard deviation  $= \sqrt{108.27} = 10.41$ .

- 4.22** For the second frequency distribution of Problem 2.8, find (a) the mean, (b) the standard deviation, (c) the standard deviation using Sheppard's correction, and (d) the actual standard deviation from the ungrouped data.

#### SOLUTION

The work is arranged in Table 4.10.

$$(a) \quad \bar{X} = A + c\bar{u} = A + c \frac{\sum fu}{N} = 149 + 9 \left( \frac{-9}{40} \right) = 147.0 \text{ lb}$$

$$(b) \quad s = c\sqrt{u^2 - \bar{u}^2} = c\sqrt{\frac{\sum fu^2}{N} - \left( \frac{\sum fu}{N} \right)^2} = 9\sqrt{\frac{95}{40} - \left( \frac{-9}{40} \right)^2} = 9\sqrt{2.324375} = 13.7 \text{ lb}$$

$$(c) \quad \text{Corrected variance} = s^2 - c^2/12 = 188.27 - 9^2/12 = 181.52. \text{ Corrected standard deviation} = 13.5 \text{ lb.}$$

Table 4.10

$X$	$u$	$f$	$fu$	$fu^2$
122	-3	3	-9	27
131	-2	5	-10	20
140	-1	9	-9	9
$A \rightarrow 149$	0	12	0	0
158	1	5	5	5
167	2	4	8	16
176	3	2	6	18
		$N = \sum f = 40$	$\sum fu = -9$	$\sum fu^2 = 95$

- (d) To compute the standard deviation from the actual weights of the students given in the problem, it is convenient first to subtract a suitable number, say  $A = 150$  lb, from each weight and then use the method of Problem 4.15. The deviations  $d = X - A = X - 150$  are then given in the following table:

-12	14	0	-18	-6	-25	-1	7
-4	8	-10	-3	-14	-2	2	-6
18	-24	-12	26	13	-31	4	15
-4	23	-8	-3	-15	3	-10	-15
11	-5	-15	-8	0	6	-5	-22

from which we find that  $\sum d = -128$  and  $\sum d^2 = 7052$ . Then

$$s = \sqrt{d^2 - \bar{d}^2} = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2} = \sqrt{\frac{7052}{40} - \left(\frac{-128}{40}\right)^2} = \sqrt{166.06} = 12.9 \text{ lb}$$

Hence Sheppard's correction supplied some improvement in this case.

## EMPIRICAL RELATIONS BETWEEN MEASURES OF DISPERSION

- 4.23** For the distribution of the heights of the students at XYZ University, discuss the validity of the empirical formulas (a) mean deviation  $= \frac{4}{5}$ (standard deviation) and (b) semi-interquartile range  $= \frac{2}{3}$ (standard deviation).

### SOLUTION

- (a) From Problems 4.4 and 4.11, mean deviation  $\div$  standard deviation  $= 2.26/2.92 = 0.77$ , which is close to  $\frac{4}{5}$ .
- (b) From Problems 4.6 and 4.11, semi-interquartile range  $\div$  standard deviation  $= 1.98/2.92 = 0.68$ , which is close to  $\frac{2}{3}$ .

Thus the empirical formulas are valid in this case.

Note that in the above we have not used the standard deviation with Sheppard's correction for grouping, since no corresponding correction has been made for the mean deviation or semi-interquartile range.

## PROPERTIES OF THE STANDARD DEVIATION

- 4.24** Determine the percentage of the students' IQ's in Problem 4.19 that fall within the ranges (a)  $\bar{X} \pm s$ , (b)  $\bar{X} \pm 2s$ , and (c)  $\bar{X} \pm 3s$ .

**SOLUTION**

- (a) The range of IQ's from 85.5 to 106.4 is  $\bar{X} \pm s = 95.97 \pm 10.47$ . The number of IQ's in the range  $\bar{X} \pm s$  is

$$\left(\frac{88 - 85.5}{4}\right)(45) + 66 + 85 + 72 + 54 + \left(\frac{106.4 - 104}{4}\right)(38) = 339$$

The percentage of IQ's in the range  $\bar{X} \pm s$  is  $339/480 = 70.6\%$ .

- (b) The range of IQ's from 75.0 to 116.9 is  $\bar{X} \pm 2s = 95.97 \pm 2(10.47)$ . The number of IQ's in the range  $\bar{X} \pm 2s$  is

$$\left(\frac{76 - 75.0}{4}\right)(9) + 16 + 28 + 45 + 66 + 85 + 72 + 54 + 38 + 27 + 18 + \left(\frac{116.9 - 116}{4}\right)(11) = 451$$

The percentage of IQ's in the range  $\bar{X} \pm 2s$  is  $451/480 = 94.0\%$ .

- (c) The range of IQ's from 64.6 to 127.4 is  $\bar{X} \pm 3s = 95.97 \pm 3(10.47)$ . The number of IQ's in the range  $\bar{X} \pm 3s$  is

$$480 - \left(\frac{128 - 127.4}{4}\right)(2) = 479.7 \quad \text{or} \quad 480$$

The percentage of IQ's in the range  $\bar{X} \pm 3s$  is  $479.7/480 = 99.9\%$ , or practically 100%.

The percentages in parts (a), (b), and (c) agree favorably with those to be expected for a normal distribution: 68.27%, 95.45%, and 99.73%, respectively.

Note that we have not used Sheppard's correction for the standard deviation. If this is used, the results in this case agree closely with the above. Note also that the above results can also be obtained by using Table 4.11 of Problem 4.32.

- 4.25** Given the sets 2, 5, 8, 11, 14, and 2, 8, 14, find (a) the mean of each set, (b) the variance of each set, (c) the mean of the combined (or pooled) sets, and (d) the variance of the combined sets.

**SOLUTION**

- (a) Mean of first set  $= \frac{1}{5}(2 + 5 + 8 + 11 + 14) = 8$ . Mean of second set  $= \frac{1}{3}(2 + 8 + 14) = 8$ .  
 (b) Variance of first set  $= s_1^2 = \frac{1}{5}[(2 - 8)^2 + (5 - 8)^2 + (8 - 8)^2 + (11 - 8)^2 + (14 - 8)^2] = 18$ . Variance of second set  $= s_2^2 = \frac{1}{3}[(2 - 8)^2 + (8 - 8)^2 + (14 - 8)^2] = 24$ .  
 (c) The mean of the combined sets is

$$\frac{2 + 5 + 8 + 11 + 14 + 2 + 8 + 14}{5 + 3} = 8$$

- (d) The variance of the combined sets is

$$s^2 = \frac{(2 - 8)^2 + (5 - 8)^2 + (8 - 8)^2 + (11 - 8)^2 + (14 - 8)^2 + (2 - 8)^2 + (8 - 8)^2 + (14 - 8)^2}{5 + 3} = 20.25$$

**Another method** (by formula)

$$s^2 = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2} = \frac{(5)(18) + (3)(24)}{5 + 3} = 20.25$$

**4.26** Work Problem 4.25 for the sets 2, 5, 8, 11, 14 and 10, 16, 22.

### SOLUTION

Here the means of the two sets are 8 and 16, respectively, while the variances are the *same* as the sets of the preceding problem, namely,  $s_1^2 = 18$  and  $s_2^2 = 24$ .

$$\text{Mean of combined sets} = \frac{2 + 5 + 8 + 11 + 14 + 10 + 16 + 22}{5 + 3} = 11$$

$$s^2 = \frac{(2-11)^2 + (5-11)^2 + (8-11)^2 + (11-11)^2 + (14-11)^2 + (10-11)^2 + (16-11)^2 + (22-11)^2}{5+3} = 35.25$$

Note that the formula

$$s^2 = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2}$$

which gives the value 20.25, is *not* applicable in this case since the means of the two sets are *not* the same.

- 4.27** (a) Prove that  $w^2 + pw + q$ , where  $p$  and  $q$  are given constants, is a minimum if and only if  $w = -\frac{1}{2}p$ .  
 (b) Using part (a), prove that

$$\frac{\sum_{j=1}^N (X_j - a)^2}{N} \quad \text{or briefly} \quad \frac{\sum (X - a)^2}{N}$$

is a minimum if and only if  $a = \bar{X}$ .

### SOLUTION

- (a) We have  $w^2 + pw + q = (w + \frac{1}{2}p)^2 + q - \frac{1}{4}p^2$ . Since  $(q - \frac{1}{4}p^2)$  is a constant, the expression has the least value (i.e., is a minimum) if and only if  $w + \frac{1}{2}p = 0$  (i.e.,  $w = -\frac{1}{2}p$ ).

$$(b) \quad \frac{\sum (X - a)^2}{N} = \frac{\sum (X^2 - 2aX + a^2)}{N} = \frac{\sum X^2 - 2a \sum X + Na^2}{N} = a^2 - 2a \frac{\sum X}{N} + \frac{\sum X^2}{N}$$

Comparing this last expression with  $(w^2 + pw + q)$ , we have

$$w = a \quad p = -2 \frac{\sum X}{N} \quad q = \frac{\sum X^2}{N}$$

Thus the expression is a minimum when  $a = -\frac{1}{2}p = (\sum X)/N = \bar{X}$ , using the result of part (a).

## ABSOLUTE AND RELATIVE DISPERSION; COEFFICIENT OF VARIATION

- 4.28** A manufacturer of television tubes has two types of tubes,  $A$  and  $B$ . Respectively, the tubes have mean lifetimes of  $\bar{X}_A = 1495$  hours and  $\bar{X}_B = 1875$  hours, and standard deviations of  $s_A = 280$  hours and  $s_B = 310$  hours. Which tube has the greater (a) absolute dispersion and (b) relative dispersion?

**SOLUTION**

- (a) The absolute dispersion of  $A$  is  $s_A = 280$  hours, and of  $B$  is  $s_B = 310$  hours. Thus tube  $B$  has the greater absolute dispersion.
- (b) The coefficients of variation are

$$A = \frac{s_A}{\bar{X}_A} = \frac{280}{1495} = 18.7\% \quad B = \frac{s_B}{\bar{X}_B} = \frac{310}{1875} = 16.5\%$$

Thus tube  $A$  has the greater relative variation, or dispersion.

- 4.29** Find the coefficients of variation,  $V$ , for the data of (a) Problem 4.14 and (b) Problem 4.18, using both uncorrected and corrected standard deviations.

**SOLUTION**

- (a)  $V(\text{uncorrected}) = \frac{s(\text{uncorrected})}{\bar{X}} = \frac{2.92}{67.45} = 0.0433 = 4.3\%$   
 $V(\text{corrected}) = \frac{s(\text{corrected})}{\bar{X}} = \frac{2.79}{67.45} = 0.0413 = 4.1\%$  by Problem 4.21(a)
- (b)  $V(\text{uncorrected}) = \frac{s(\text{uncorrected})}{\bar{X}} = \frac{15.60}{79.77} = 0.196 = 19.6\%$   
 $V(\text{corrected}) = \frac{s(\text{corrected})}{\bar{X}} = \frac{15.33}{79.77} = 0.192 = 19.2\%$  by Problem 4.21(b)

- 4.30** (a) Define a measure of relative dispersion that could be used for a set of data for which the quartiles are known.
- (b) Illustrate the calculation of the measure defined in part (a) by using the data of Problem 4.6.

**SOLUTION**

- (a) If  $Q_1$  and  $Q_3$  are given for a set of data, then  $\frac{1}{2}(Q_1 + Q_3)$  is a measure of the data's central tendency, or average, while  $Q = \frac{1}{2}(Q_3 - Q_1)$ , the semi-interquartile range, is a measure of the data's dispersion. We can thus define a measure of relative dispersion as

$$V_Q = \frac{\frac{1}{2}(Q_3 - Q_1)}{\frac{1}{2}(Q_1 + Q_3)} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

which we call the *quartile coefficient of variation*, or *quartile coefficient of relative dispersion*.

- (b)  $V_Q = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{69.61 - 65.64}{69.61 + 65.64} = \frac{3.97}{135.25} = 0.0293 = 2.9\%$

**STANDARDIZED VARIABLE; STANDARD SCORES**

- 4.31** A student received a grade of 84 on a final examination in mathematics for which the mean grade was 76 and the standard deviation was 10. On the final examination in physics, for which the mean grade was 82 and the standard deviation was 16, she received a grade of 90. In which subject was her relative standing higher?

**SOLUTION**

The standardized variable  $z = (X - \bar{X})/s$  measures the deviation of  $X$  from the mean  $\bar{X}$  in terms of standard deviation  $s$ . For mathematics,  $z = (84 - 76)/10 = 0.8$ ; for physics,  $z = (90 - 82)/16 = 0.5$ . Thus the student had a grade 0.8 of a standard deviation above the mean in mathematics, but only 0.5 of a standard deviation above the mean in physics. Thus her relative standing was higher in mathematics.

The variable  $z = (X - \bar{X})/s$  is often used in educational testing, where it is known as a *standard score*.

4.32 The STATISTIX analysis of the data in Example 3 of this chapter gave the following output.

Statistix 8.0

Descriptive Statistics

Variable	SD	Variance	C.V.	MAD
e - mails	29.256	855.93	44.562	21.000

The MAD value is the *median absolute deviation*. It is the median value of the absolute differences among the individual values and the sample median. Confirm that the MAD value for this data equals 21.

SOLUTION

The sorted original data are:

24	24	24	25	26	28	29	30	31	31	31	32	32
35	35	36	39	40	40	42	42	44	44	45	47	49
51	52	54	54	54	57	58	58	58	60	61	61	63
65	65	68	69	70	71	71	74	74	74	77	77	77
77	79	80	84	86	86	95	95	99	99	100	102	102
105	113	113	114	115	116	118	121	122	125			

The median of the original data is 61.

If 61 is subtracted from each value, the data become:

-37	-37	-37	-36	-35	-33	-32	-31	-30	-30	-30	-29	-29
-26	-26	-25	-22	-21	-21	-19	-19	-17	-17	-16	-14	-12
-10	-9	-7	-7	-7	-4	-3	-3	-3	-1	0	0	2
4	4	7	8	9	10	10	13	13	13	16	16	16
16	18	19	23	25	25	34	34	38	38	39	41	41
44	52	52	53	54	55	57	60	61	64			

Now, take the absolute value of each of these values:

37	37	37	36	35	33	32	31	30	30	30	29	29	26	26
25	22	21	21	19	19	17	17	16	14	12	10	9	7	7
7	4	3	3	3	1	0	0	2	4	4	7	8	9	10
10	13	13	13	16	16	16	16	18	19	23	25	25	34	34
38	38	39	41	41	44	52	52	53	54	55	57	60	61	64

The median of this last set of data is 21. Therefore MAD = 21.

# Supplementary Problems

## THE RANGE

- 4.33** Find the range of the sets (a) 5, 3, 8, 4, 7, 6, 12, 4, 3 and (b) 8.772, 6.453, 10.624, 8.628, 9.434, 6.351.
- 4.34** Find the range of the maximum loads given in Table 3.8 of Problem 3.59.
- 4.35** Find the range of the rivet diameters in Table 3.10 of Problem 3.61.
- 4.36** The largest of 50 measurements is 8.34 kilograms (kg). If the range is 0.46 kg, find the smallest measurement.
- 4.37** The following table gives the number of weeks needed to find a job for 25 older workers that lost their jobs as a result of corporation downsizing. Find the range of the data.

13	13	17	7	22
22	26	17	13	14
16	7	6	18	20
10	17	11	10	15
16	8	16	21	11

## THE MEAN DEVIATION

- 4.38** Find the absolute values of (a)  $-18.2$ , (b)  $+3.58$ , (c)  $6.21$ , (d)  $0$ , (e)  $-\sqrt{2}$ , and (f)  $4.00 - 2.36 - 3.52$ .
- 4.39** Find the mean deviation of the set (a) 3, 7, 9, 5 and (b) 2.4, 1.6, 3.8, 4.1, 3.4.
- 4.40** Find the mean deviation of the sets of numbers in Problem 4.33.
- 4.41** Find the mean deviation of the maximum loads in Table 3.8 of Problem 3.59.
- 4.42** (a) Find the mean deviation (MD) of the rivet diameters in Table 3.10 of Problem 3.61.  
(b) What percentage of the rivet diameters lie between  $(\bar{X} \pm \text{MD})$ ,  $(\bar{X} \pm 2\text{MD})$ , and  $(\bar{X} \pm 3\text{MD})$ ?
- 4.43** For the set 8, 10, 9, 12, 4, 8, 2, find the mean deviation (a) from the mean and (b) from the median. Verify that the mean deviation from the median is not greater than the mean deviation from the mean.
- 4.44** For the distribution in Table 3.9 of Problem 3.60, find the mean deviation (a) about the mean and (b) about the median. Use the results of Problems 3.60 and 3.70.
- 4.45** For the distribution in Table 3.11 of Problem 3.62, find the mean deviation (a) about the mean and (b) about the median. Use the results of Problems 3.62 and 3.72.



- 4.46** Find the mean deviation for the data given in Problem 4.37.
- 4.47** Derive coding formulas for computing the mean deviation (*a*) about the mean and (*b*) about the median from a frequency distribution. Apply these formulas to verify the results of Problems 4.44 and 4.45.

### THE SEMI-INTERQUARTILE RANGE

- 4.48** Find the semi-interquartile range for the distributions of (*a*) Problem 3.59, (*b*) Problem 3.60, and (*c*) Problem 3.107. Interpret the results clearly in each case.
- 4.49** Find the semi-interquartile range for the data given in Problem 4.37.
- 4.50** Prove that for any frequency distribution the total percentage of cases falling in the interval  $\frac{1}{2}(Q_1 + Q_3) \pm \frac{1}{2}(Q_3 - Q_1)$  is 50%. Is the same true for the interval  $Q_2 \pm \frac{1}{2}(Q_3 - Q_1)$ ? Explain your answer.
- 4.51** (*a*) How would you graph the semi-interquartile range corresponding to a given frequency distribution?  
(*b*) What is the relationship of the semi-interquartile range to the ogive of the distribution?

### THE 10–90 PERCENTILE RANGE

- 4.52** Find the 10–90 percentile range for the distributions of (*a*) Problem 3.59 and (*b*) Problem 3.107. Interpret the results clearly in each case.
- 4.53** The tenth percentile for home selling prices in a city is \$35,500 and the ninetieth percentile for home selling prices in the same city is \$225,000. Find the 10–90 percentile range and give a range within which 80% of the selling prices fall.
- 4.54** What advantages or disadvantages would a 20–80 percentile range have in comparison to a 10–90 percentile range?
- 4.55** Answer Problem 4.51 with reference to the (*a*) 10–90 percentile range, (*b*) 20–80 percentile range, and (*c*) 25–75 percentile range. What is the relationship between (*c*) and the semi-interquartile range?

### THE STANDARD DEVIATION

- 4.56** Find the standard deviation of the sets (*a*) 3, 6, 2, 1, 7, 5; (*b*) 3.2, 4.6, 2.8, 5.2, 4.4; and (*c*) 0, 0, 0, 0, 1, 1, 1.
- 4.57** (*a*) By adding 5 to each of the numbers in the set 3, 6, 2, 1, 7, 5, we obtain the set 8, 11, 7, 6, 12, 10. Show that the two sets have the same standard deviation but different means. How are the means related?  
(*b*) By multiplying each of the numbers 3, 6, 2, 1, 7, and 5 by 2 and then adding 5, we obtain the set 11, 17, 9, 7, 19, 15. What is the relationship between the standard deviations and the means for the two sets?  
(*c*) What properties of the mean and standard deviation are illustrated by the particular sets of numbers in parts (*a*) and (*b*)?

- 4.58** Find the standard deviation of the set of numbers in the arithmetic progression 4, 10, 16, 22, ..., 154.
- 4.59** Find the standard deviation for the distributions of (a) Problem 3.59, (b) Problem 3.60, and (c) Problem 3.107.
- 4.60** Demonstrate the use of Charlier's check in each part of Problem 4.59.
- 4.61** Find (a) the mean and (b) the standard deviation for the distribution of Problem 2.17, and explain the significance of the results obtained.
- 4.62** When data have a bell-shaped distribution, the standard deviation may be approximated by dividing the range by 4. For the data given in Problem 4.37, compute the standard deviation and compare it with the range divided by 4.
- 4.63** (a) Find the standard deviation  $s$  of the rivet diameters in Table 3.10 of Problem 3.61.  
 (b) What percentage of the rivet diameters lies between  $\bar{X} \pm s$ ,  $\bar{X} \pm 2s$ , and  $\bar{X} \pm 3s$ ?  
 (c) Compare the percentages in part (b) with those which would theoretically be expected if the distribution were normal, and account for any observed differences.
- 4.64** Apply Sheppard's correction to each standard deviation in Problem 4.59. In each case, discuss whether such application is or is not justified.
- 4.65** What modifications occur in Problem 4.63 when Sheppard's correction is applied?
- 4.66** (a) Find the mean and standard deviation for the data of Problem 2.8.  
 (b) Construct a frequency distribution for the data and find the standard deviation.  
 (c) Compare the results of part (b) with that of part (a). Determine whether an application of Sheppard's correction produces better results.
- 4.67** Work Problem 4.66 for the data of Problem 2.27.
- 4.68** (a) Of a total of  $N$  numbers, the fraction  $p$  are 1's, while the fraction  $q = 1 - p$  are 0's. Prove that the standard deviation of the set of numbers is  $\sqrt{pq}$ .  
 (b) Apply the result of part (a) to Problem 4.56(c).
- 4.69** (a) Prove that the variance of the set of  $n$  numbers  $a, a + d, a + 2d, \dots, a + (n - 1)d$  (i.e., an arithmetic progression with the first term  $a$  and common difference  $d$ ) is given by  $\frac{1}{12}(n^2 - 1)d^2$ .  
 (b) Use part (a) for Problem 4.58. [Hint: Use  $1 + 2 + 3 + \dots + (n - 1) = \frac{1}{2}n(n - 1)$ ,  $1^2 + 2^2 + 3^2 + \dots + (n - 1)^2 = \frac{1}{6}n(n - 1)(2n - 1)$ .]
- 4.70** Generalize and prove Property 3 of this chapter.

## EMPIRICAL RELATIONS BETWEEN MEASURES OF DISPERSION

- 4.71** By comparing the standard deviations obtained in Problem 4.59 with the corresponding mean deviations of Problems 4.41, 4.42, and 4.44, determine whether the following empirical relation holds: Mean deviation =  $\frac{4}{5}$ (standard deviation). Account for any differences that may occur.

- 4.72** By comparing the standard deviations obtained in Problem 4.59 with the corresponding semi-interquartile ranges of Problem 4.48, determine whether the following empirical relation holds: Semi-interquartile range  $= \frac{2}{3}$  (standard deviation). Account for any differences that may occur.
- 4.73** What empirical relation would you expect to exist between the semi-interquartile range and the mean deviation for bell-shaped distributions that are moderately skewed?
- 4.74** A frequency distribution that is approximately normal has a semi-interquartile range equal to 10. What values would you expect for (a) the standard deviation and (b) the mean deviation?

#### **ABSOLUTE AND RELATIVE DISPERSION; COEFFICIENT OF VARIATION**

- 4.75** On a final examination in statistics, the mean grade of a group of 150 students was 78 and the standard deviation was 8.0. In algebra, however, the mean final grade of the group was 73 and the standard deviation was 7.6. In which subject was there the greater (a) absolute dispersion and (b) relative dispersion?
- 4.76** Find the coefficient of variation for the data of (a) Problem 3.59 and (b) Problem 3.107.
- 4.77** The distribution of SAT scores for a group of high school students has a first quartile score equal to 825 and a third quartile score equal to 1125. Calculate the quartile coefficient of variation for the distribution of SAT scores for this group of high school students.
- 4.78** For the age group 15–24 years, the first quartile of household incomes is equal to \$16,500 and the third quartile of household incomes for this same age group is \$25,000. Calculate the quartile coefficient of variation for the distribution of incomes for this age group.

#### **STANDARDIZED VARIABLES; STANDARD SCORES**

- 4.79** On the examinations referred to in Problem 4.75, a student scored 75 in statistics and 71 in algebra. In which examination was his relative standing higher?
- 4.80** Convert the set 6, 2, 8, 7, 5 into standard scores.
- 4.81** Prove that the mean and standard deviation of a set of standard scores are equal to 0 and 1, respectively. Use Problem 4.80 to illustrate this.
- 4.82** (a) Convert the grades of Problem 3.107 into standard scores, and (b) construct a graph of relative frequency versus standard score.

#### **SOFTWARE AND MEASURES OF DISPERSION**

- 4.83** Table 4.11 gives the per capita income for the 50 states in 2005.

**Table 4.11 Per Capita Income for the 50 States**

State	Per Capita Income	State	Per Capita Income
Wyoming	36,778	Pennsylvania	34,897
Montana	29,387	Wisconsin	33,565
North Dakota	31,395	Massachusetts	44,289
New Mexico	27,664	Missouri	31,899
West Virginia	27,215	Idaho	28,158
Rhode Island	36,153	Kentucky	28,513
Virginia	38,390	Minnesota	37,373
South Dakota	31,614	Florida	33,219
Alabama	29,136	South Carolina	28,352
Arkansas	26,874	New York	40,507
Maryland	41,760	Indiana	31,276
Iowa	32,315	Connecticut	47,819
Nebraska	33,616	Ohio	32,478
Hawaii	34,539	New Hampshire	38,408
Mississippi	25,318	Texas	32,462
Vermont	33,327	Oregon	32,103
Maine	31,252	New Jersey	43,771
Oklahoma	29,330	California	37,036
Delaware	37,065	Colorado	37,946
Alaska	35,612	North Carolina	30,553
Tennessee	31,107	Illinois	36,120
Kansas	32,836	Michigan	33,116
Arizona	30,267	Washington	35,409
Nevada	35,883	Georgia	31,121
Utah	28,061	Louisiana	24,820

The SPSS analysis of the data is as follows:

#### Descriptive Statistics

	N	Range	Std. Deviation	Variance
Income	50	22999.00	4893.54160	2E + 007
Valid N (listwise)	50			

Verify the range, standard deviation, and variance.