Technology *Step by Step*

Excel
Step by Step

Finding Measures of Central Tendency

Example XL3–1

Find the mean, mode, and median of the data from Example 3–11. The data represent the population of licensed nuclear reactors in the United States for a recent 15-year period.

104	104	104	104	104
107	109	109	109	110
109	111	112	111	109

1. On an Excel worksheet enter the numbers in cells A2–A16. Enter a label for the variable in cell A1.

On the same worksheet as the data:

2. Compute the mean of the data: key in **=AVERAGE(A2:A16)** in a blank cell.
3. Compute the mode of the data: key in **=MODE(A2:A16)** in a blank cell.
4. Compute the median of the data: key in **=MEDIAN(A2:A16)** in a blank cell.

These and other statistical functions can also be accessed without typing them into the worksheet directly.

1. Select the Formulas tab from the toolbar and select the Insert Function Icon f_x.
2. Select the Statistical category for statistical functions.
3. Scroll to find the appropriate function and click [OK].

	A	B	C
1	Number of Reactors		
2	104	107.7333	mean
3	104	104	mode
4	104	109	median
5	104		
6	104		
7	107		
8	109		
9	109		
10	109		
11	110		
12	109		
13	111		
14	112		
15	111		
16	109		

3–2 Measures of Variation

In statistics, to describe the data set accurately, statisticians must know more than the measures of central tendency. Consider Example 3–18.

Example 3–18

Objective 2

Describe data, using measures of variation, such as the range, variance, and standard deviation.

Comparison of Outdoor Paint

A testing lab wishes to test two experimental brands of outdoor paint to see how long each will last before fading. The testing lab makes 6 gallons of each paint to test. Since different chemical agents are added to each group and only six cans are involved, these two groups constitute two small populations. The results (in months) are shown. Find the mean of each group.

Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25

Solution

The mean for brand A is

$$\mu = \frac{\Sigma X}{N} = \frac{210}{6} = 35 \text{ months}$$

The mean for brand B is

$$\mu = \frac{\Sigma X}{N} = \frac{210}{6} = 35 \text{ months}$$

Since the means are equal in Example 3–18, you might conclude that both brands of paint last equally well. However, when the data sets are examined graphically, a somewhat different conclusion might be drawn. See Figure 3–2.

As Figure 3–2 shows, even though the means are the same for both brands, the spread, or variation, is quite different. Figure 3–2 shows that brand B performs more consistently; it is less variable. For the spread or variability of a data set, three measures are commonly used: *range, variance,* and *standard deviation.* Each measure will be discussed in this section.

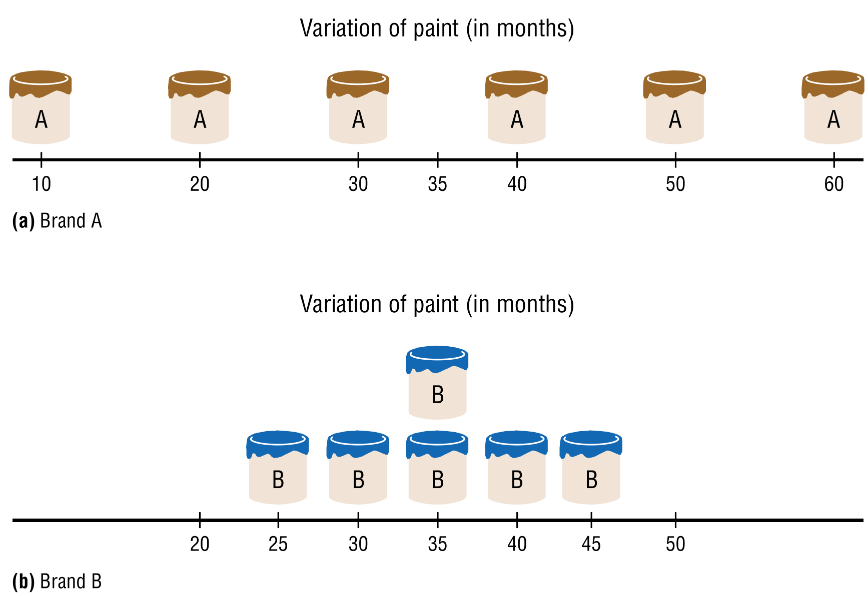
Range

The range is the simplest of the three measures and is defined now.

The **range** is the highest value minus the lowest value. The symbol R is used for the range.

R = highest value − lowest value

Figure 3–2
Examining Data Sets Graphically



Example 3–19

Comparison of Outdoor Paint

Find the ranges for the paints in Example 3–18.

Solution

For brand A, the range is

$$R = 60 - 10 = 50 \text{ months}$$

For brand B, the range is

$$R = 45 - 25 = 20 \text{ months}$$

Make sure the range is given as a single number.

The range for brand A shows that 50 months separate the largest data value from the smallest data value. For brand B, 20 months separate the largest data value from the smallest data value, which is less than one-half of brand A's range.

One extremely high or one extremely low data value can affect the range markedly, as shown in Example 3–20.

Example 3–20

Employee Salaries

The salaries for the staff of the XYZ Manufacturing Co. are shown here. Find the range.

Staff	Salary
Owner	\$100,000
Manager	40,000
Sales representative	30,000
Workers	25,000
	15,000
	18,000

Solution

The range is $R = \$100{,}000 - \$15{,}000 = \$85{,}000$.

Since the owner's salary is included in the data for Example 3–20, the range is a large number. To have a more meaningful statistic to measure the variability, statisticians use measures called the *variance* and *standard deviation.*

Population Variance and Standard Deviation

Before the variance and standard deviation are defined formally, the computational procedure will be shown, since the definition is derived from the procedure.

Rounding Rule for the Standard Deviation The rounding rule for the standard deviation is the same as that for the mean. The final answer should be rounded to one more decimal place than that of the original data.

Example 3–21

Comparison of Outdoor Paint

Find the variance and standard deviation for the data set for brand A paint in Example 3–18.

10, 60, 50, 30, 40, 20

Solution

Step 1 Find the mean for the data.

$$\mu = \frac{\Sigma X}{N} = \frac{10 + 60 + 50 + 30 + 40 + 20}{6} = \frac{210}{6} = 35$$

Step 2 Subtract the mean from each data value.

$10 - 35 = -25$ $\quad 50 - 35 = +15$ $\quad 40 - 35 = +5$

$60 - 35 = +25$ $\quad 30 - 35 = -5$ $\quad 20 - 35 = -15$

Step 3 Square each result.

$(-25)^2 = 625$ $\quad (+15)^2 = 225$ $\quad (+5)^2 = 25$

$(+25)^2 = 625$ $\quad (-5)^2 = 25$ $\quad (-15)^2 = 225$

Step 4 Find the sum of the squares.

$625 + 625 + 225 + 25 + 25 + 225 = 1750$

Step 5 Divide the sum by N to get the variance.

Variance $= 1750 \div 6 = 291.7$

Step 6 Take the square root of the variance to get the standard deviation. Hence, the standard deviation equals $\sqrt{291.7}$, or 17.1. It is helpful to make a table.

A Values X	B $X - \mu$	C $(X - \mu)^2$
10	−25	625
60	+25	625
50	+15	225
30	−5	25
40	+5	25
20	−15	225
		1750

Column A contains the raw data X. Column B contains the differences $X - \mu$ obtained in step 2. Column C contains the squares of the differences obtained in step 3.

The preceding computational procedure reveals several things. First, the square root of the variance gives the standard deviation; and vice versa, squaring the standard deviation gives the variance. Second, the variance is actually the average of the square of the distance that each value is from the mean. Therefore, if the values are near the mean, the variance will be small. In contrast, if the values are far from the mean, the variance will be large.

You might wonder why the squared distances are used instead of the actual distances. One reason is that the sum of the distances will always be zero. To verify this result for a specific case, add the values in column B of the table in Example 3–21. When each value is squared, the negative signs are eliminated.

Finally, why is it necessary to take the square root? The reason is that since the distances were squared, the units of the resultant numbers are the squares of the units of the original raw data. Finding the square root of the variance puts the standard deviation in the same units as the raw data.

When you are finding the square root, always use its positive value, since the variance and standard deviation of a data set can never be negative.

Historical Note

Karl Pearson in 1892 and 1893 introduced the statistical concepts of the range and standard deviation.

The **variance** is the average of the squares of the distance each value is from the mean. The symbol for the population variance is σ^2 (σ is the Greek lowercase letter sigma).

The formula for the population variance is

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

where

X = individual value
μ = population mean
N = population size

The **standard deviation** is the square root of the variance. The symbol for the population standard deviation is σ.

The corresponding formula for the population standard deviation is

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

Example 3–22

Comparison of Outdoor Paint

Find the variance and standard deviation for brand B paint data in Example 3–18. The months were

35, 45, 30, 35, 40, 25

Solution

Step 1 Find the mean.

$$\mu = \frac{\Sigma X}{N} = \frac{35 + 45 + 30 + 35 + 40 + 25}{6} = \frac{210}{6} = 35$$

Step 2 Subtract the mean from each value, and place the result in column B of the table.

Step 3 Square each result and place the squares in column C of the table.

A X	B $X - \mu$	C $(X - \mu)^2$
35	0	0
45	10	100
30	−5	25
35	0	0
40	5	25
25	−10	100

Interesting Fact

Each person receives on average 598 pieces of mail per year.

Step 4 Find the sum of the squares in column C.

$$\Sigma(X - \mu)^2 = 0 + 100 + 25 + 0 + 25 + 100 = 250$$

Step 5 Divide the sum by N to get the variance.

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N} = \frac{250}{6} = 41.7$$

Step 6 Take the square root to get the standard deviation.

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}} = \sqrt{41.7} = 6.5$$

Hence, the standard deviation is 6.5.

Since the standard deviation of brand A is 17.1 (see Example 3–21) and the standard deviation of brand B is 6.5, the data are more variable for brand A. *In summary, when the means are equal, the larger the variance or standard deviation is, the more variable the data are.*

Sample Variance and Standard Deviation

When computing the variance for a sample, one might expect the following expression to be used:

$$\frac{\Sigma(X - \overline{X})^2}{n}$$

where $\overline{X}$ is the sample mean and n is the sample size. *This formula is not usually used, however, since in most cases the purpose of calculating the statistic is to estimate the corresponding parameter.* For example, the sample mean $\overline{X}$ is used to estimate the population mean μ. The expression

$$\frac{\Sigma(X - \overline{X})^2}{n}$$

does not give the best estimate of the population variance because when the population is large and the sample is small (usually less than 30), the variance computed by this formula usually underestimates the population variance. Therefore, instead of dividing by n, find the variance of the sample by dividing by $n - 1$, giving a slightly larger value and an *unbiased* estimate of the population variance.

The formula for the sample variance, denoted by s^2, is

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

where

$\overline{X}$ = sample mean

n = sample size

To find the standard deviation of a sample, you must take the square root of the sample variance, which was found by using the preceding formula.

Formula for the Sample Standard Deviation

The standard deviation of a sample (denoted by s) is

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

where

X = individual value

$\overline{X}$ = sample mean

n = sample size

Shortcut formulas for computing the variance and standard deviation are presented next and will be used in the remainder of the chapter and in the exercises. These formulas are mathematically equivalent to the preceding formulas and do not involve using the mean. They save time when repeated subtracting and squaring occur in the original formulas. They are also more accurate when the mean has been rounded.

Shortcut or Computational Formulas for s^2 and s

The shortcut formulas for computing the variance and standard deviation for data obtained from samples are as follows.

Variance

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

Standard deviation

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

Examples 3–23 and 3–24 explain how to use the shortcut formulas.

Example 3–23

European Auto Sales

Find the sample variance and standard deviation for the amount of European auto sales for a sample of 6 years shown. The data are in millions of dollars.

11.2, 11.9, 12.0, 12.8, 13.4, 14.3

Source: *USA TODAY*.

Solution

Step 1 Find the sum of the values.

$$\Sigma X = 11.2 + 11.9 + 12.0 + 12.8 + 13.4 + 14.3 = 75.6$$

Step 2 Square each value and find the sum.

$$\Sigma X^2 = 11.2^2 + 11.9^2 + 12.0^2 + 12.8^2 + 13.4^2 + 14.3^2 = 958.94$$

Step 3 Substitute in the formulas and solve.

$$\begin{aligned} s^2 &= \frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)} \\ &= \frac{6(958.94) - 75.6^2}{6(6-1)} \\ &= \frac{5753.64 - 5715.36}{6(5)} \\ &= \frac{38.28}{30} \\ &= 1.276 \end{aligned}$$

The variance is 1.28 rounded.

$$s = \sqrt{1.28} = 1.13$$

Hence, the sample standard deviation is 1.13.

Note that ΣX^2 is not the same as $(\Sigma X)^2$. The notation ΣX^2 means to square the values first, then sum; $(\Sigma X)^2$ means to sum the values first, then square the sum.

Variance and Standard Deviation for Grouped Data

The procedure for finding the variance and standard deviation for grouped data is similar to that for finding the mean for grouped data, and it uses the midpoints of each class.

Example 3–24 Miles Run per Week

Find the variance and the standard deviation for the frequency distribution of the data in Example 2–7. The data represent the number of miles that 20 runners ran during one week.

Class	Frequency	Midpoint
5.5–10.5	1	8
10.5–15.5	2	13
15.5–20.5	3	18
20.5–25.5	5	23
25.5–30.5	4	28
30.5–35.5	3	33
35.5–40.5	2	38

Unusual Stat

At birth men outnumber women by 2%. By age 25, the number of men living is about equal to the number of women living. By age 65, there are 14% more women living than men.

Solution

Step 1 Make a table as shown, and find the midpoint of each class.

A Class	B Frequency f	C Midpoint X_m	D $f \cdot X_m$	E $f \cdot X_m^2$
5.5–10.5	1	8		
10.5–15.5	2	13		
15.5–20.5	3	18		
20.5–25.5	5	23		
25.5–30.5	4	28		
30.5–35.5	3	33		
35.5–40.5	2	38		

Step 2 Multiply the frequency by the midpoint for each class, and place the products in column D.

$1 \cdot 8 = 8 \qquad 2 \cdot 13 = 26 \qquad \ldots \qquad 2 \cdot 38 = 76$

Step 3 Multiply the frequency by the square of the midpoint, and place the products in column E.

$1 \cdot 8^2 = 64 \qquad 2 \cdot 13^2 = 338 \qquad \ldots \qquad 2 \cdot 38^2 = 2888$

Step 4 Find the sums of columns B, D, and E. The sum of column B is n, the sum of column D is $\Sigma f \cdot X_m$, and the sum of column E is $\Sigma f \cdot X_m^2$. The completed table is shown.

A Class	B Frequency	C Midpoint	D $f \cdot X_m$	E $f \cdot X_m^2$
5.5–10.5	1	8	8	64
10.5–15.5	2	13	26	338
15.5–20.5	3	18	54	972
20.5–25.5	5	23	115	2,645
25.5–30.5	4	28	112	3,136
30.5–35.5	3	33	99	3,267
35.5–40.5	2	38	76	2,888
	$n = 20$		$\Sigma f \cdot X_m = 490$	$\Sigma f \cdot X_m^2 = 13,310$

Step 5 Substitute in the formula and solve for s^2 to get the variance.

$$s^2 = \frac{n(\Sigma f \cdot X_m^2) - (\Sigma f \cdot X_m)^2}{n(n-1)}$$
$$= \frac{20(13{,}310) - 490^2}{20(20-1)}$$
$$= \frac{266{,}200 - 240{,}100}{20(19)}$$
$$= \frac{26{,}100}{380}$$
$$= 68.7$$

Step 6 Take the square root to get the standard deviation.

$$s = \sqrt{68.7} = 8.3$$

Be sure to use the number found in the sum of column B (i.e., the sum of the frequencies) for *n*. Do not use the number of classes.

The steps for finding the variance and standard deviation for grouped data are summarized in this Procedure Table.

Procedure Table

Finding the Sample Variance and Standard Deviation for Grouped Data

Step 1 Make a table as shown, and find the midpoint of each class.

A	B	C	D	E
Class	Frequency	Midpoint	$f \cdot X_m$	$f \cdot X_m^2$

Step 2 Multiply the frequency by the midpoint for each class, and place the products in column D.

Step 3 Multiply the frequency by the square of the midpoint, and place the products in column E.

Step 4 Find the sums of columns B, D, and E. (The sum of column B is *n*. The sum of column D is $\Sigma f \cdot X_m$. The sum of column E is $\Sigma f \cdot X_m^2$.)

Step 5 Substitute in the formula and solve to get the variance.

$$s^2 = \frac{n(\Sigma f \cdot X_m^2) - (\Sigma f \cdot X_m)^2}{n(n-1)}$$

Step 6 Take the square root to get the standard deviation.

Unusual Stat

The average number of times that a man cries in a month is 1.4.

The three measures of variation are summarized in Table 3–2.

Table 3–2 Summary of Measures of Variation

Measure	Definition	Symbol(s)
Range	Distance between highest value and lowest value	R
Variance	Average of the squares of the distance that each value is from the mean	σ^2, s^2
Standard deviation	Square root of the variance	σ, s

Uses of the Variance and Standard Deviation

1. As previously stated, variances and standard deviations can be used to determine the spread of the data. If the variance or standard deviation is large, the data are more dispersed. This information is useful in comparing two (or more) data sets to determine which is more (most) variable.
2. The measures of variance and standard deviation are used to determine the consistency of a variable. For example, in the manufacture of fittings, such as nuts and bolts, the variation in the diameters must be small, or the parts will not fit together.
3. The variance and standard deviation are used to determine the number of data values that fall within a specified interval in a distribution. For example, Chebyshev's theorem (explained later) shows that, for any distribution, at least 75% of the data values will fall within 2 standard deviations of the mean.
4. Finally, the variance and standard deviation are used quite often in inferential statistics. These uses will be shown in later chapters of this textbook.

Coefficient of Variation

Historical Note

Karl Pearson devised the coefficient of variation to compare the deviations of two different groups such as the heights of men and women.

Whenever two samples have the same units of measure, the variance and standard deviation for each can be compared directly. For example, suppose an automobile dealer wanted to compare the standard deviation of miles driven for the cars she received as trade-ins on new cars. She found that for a specific year, the standard deviation for Buicks was 422 miles and the standard deviation for Cadillacs was 350 miles. She could say that the variation in mileage was greater in the Buicks. But what if a manager wanted to compare the standard deviations of two different variables, such as the number of sales per salesperson over a 3-month period and the commissions made by these salespeople?

A statistic that allows you to compare standard deviations when the units are different, as in this example, is called the *coefficient of variation.*

The **coefficient of variation,** denoted by CVar, is the standard deviation divided by the mean. The result is expressed as a percentage.

For samples,

$$\text{CVar} = \frac{s}{\overline{X}} \cdot 100$$

For populations,

$$\text{CVar} = \frac{\sigma}{\mu} \cdot 100$$

Example 3–25

Sales of Automobiles

The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is $5225, and the standard deviation is $773. Compare the variations of the two.

Solution

The coefficients of variation are

$$\text{CVar} = \frac{s}{\overline{X}} = \frac{5}{87} \cdot 100 = 5.7\% \qquad \text{sales}$$

$$\text{CVar} = \frac{773}{5225} \cdot 100 = 14.8\% \qquad \text{commissions}$$

Since the coefficient of variation is larger for commissions, the commissions are more variable than the sales.