Mean, Variance, and Standard Deviation for Grouped Data

Mean for Grouped Data:

Calculating Mean for Grouped Data

$$\text{Mean for population data: } \mu = \frac{\Sigma mf}{N}$$

$$\text{Mean for sample data: } \bar{x} = \frac{\Sigma mf}{n}$$

where m is the midpoint and f is the frequency of a class.

The procedure for finding the mean for grouped data assumes that the mean of all the raw data values in each class is equal to the midpoint of the class. In reality, this is not true, since the average of the raw data values in each class usually will not be exactly equal to the midpoint. However, using this procedure will give an acceptable approximation of the mean, since some values fall above the midpoint and other values fall below the midpoint for each class, and the midpoint represents an estimate of all values in the class.

Procedure Table

Finding the Mean for Grouped Data

Step 1 Make a table as shown.

A	B	C	D
Class	**Frequency f**	**Midpoint X_m**	**$f \cdot X_m$**

Step 2 Find the midpoints of each class and place them in column C.

Step 3 Multiply the frequency by the midpoint for each class, and place the product in column D.

Step 4 Find the sum of column D.

Step 5 Divide the sum obtained in column D by the sum of the frequencies obtained in column B.

The formula for the mean is

$$\bar{X} = \frac{\Sigma f \cdot X_m}{n}$$

[*Note:* The symbols $\Sigma f \cdot X_m$ mean to find the sum of the product of the frequency (f) and the midpoint (X_m) for each class.]

Variance for Grouped Data:

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

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.

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,

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

For populations,

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

Example: Miles Run per Week

Find the mean for the distribution (shown here) of the miles that 20 randomly selected runners ran during a given week

Class boundaries	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2
	20

Solution:

Convert each frequency to a proportion or relative frequency by dividing the frequency for each class by the total number of observations.

For class 5.5–10.5, the relative frequency is $\frac{1}{20} = 0.05$; for class 10.5–15.5, the relative frequency is $\frac{2}{20} = 0.10$; for class 15.5–20.5, the relative frequency is $\frac{3}{20} = 0.15$; and so on.

Place these values in the column labeled Relative frequency.

Class boundaries	Midpoints	Relative frequency
5.5–10.5	8	0.05
10.5–15.5	13	0.10
15.5–20.5	18	0.15
20.5–25.5	23	0.25
25.5–30.5	28	0.20
30.5–35.5	33	0.15
35.5–40.5	38	0.10
		1.00

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

Find the sum of column D.

Divide the sum by n to get the mean.

$$\overline{X} = \frac{\Sigma f \cdot X_m}{n} = \frac{490}{20} = 24.5 \text{ miles}$$

Example: Miles Run per Week

Find the modal class for the frequency distribution of miles that 20 runners ran in one week, used in Example 2–7.

Class	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5 ← Modal class
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2

The modal class is 20.5–25.5, since it has the largest frequency. Sometimes the midpoint of the class is used rather than the boundaries; hence, the mode could also be given as 23 miles per week.

Properties and Uses of Central Tendency

The Mean

1. The mean is found by using all the values of the data.
2. The mean varies less than the median or mode when samples are taken from the same population and all three measures are computed for these samples.
3. The mean is used in computing other statistics, such as the variance.
4. The mean for the data set is unique and not necessarily one of the data values.
5. The mean cannot be computed for the data in a frequency distribution that has an open-ended class.
6. The mean is affected by extremely high or low values, called outliers, and may not be the appropriate average to use in these situations.

The Median

1. The median is used to find the center or middle value of a data set.
2. The median is used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution.
3. The median is used for an open-ended distribution.
4. The median is affected less than the mean by extremely high or extremely low values.

The Mode

1. The mode is used when the most typical case is desired.
2. The mode is the easiest average to compute.
3. The mode can be used when the data are nominal, such as religious preference, gender, or political affiliation.
4. The mode is not always unique. A data set can have more than one mode, or the mode may not exist for a data set.

Example: Miles Run per Week

Find the variance and the standard deviation for the frequency distribution of the data. 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

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.

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

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.

Example: 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.

The coefficients of variation are

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

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

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

Example: Pages in Women's Fitness Magazines

The mean for the number of pages of a sample of women's fitness magazines is 132, with a variance of 23; the mean for the number of advertisements of a sample of women's fitness magazines is 182, with a variance of 62. Compare the variations.

The coefficients of variation are

$$\text{CVar} = \frac{\sqrt{23}}{132} \cdot 100\% = 3.6\% \qquad \text{pages}$$

$$\text{CVar} = \frac{\sqrt{62}}{182} \cdot 100\% = 4.3\% \qquad \text{advertisements}$$

The number of advertisements is more variable than the number of pages since the coefficient of variation is larger for advertisements.