STATISTICS

- Statistics is concerned with scientific methods for collecting, organizing, summarizing, presenting, and analyzing data as well as with drawing valid conclusions and making reasonable decisions on the basis of such analysis.
- **There are two types of statistics.**

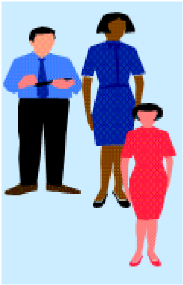
1- **Descriptive Statistic:** used to describe or explain a data set of the population that is being studied. It cannot be extrapolated or generalized to any other population or group. For example like something that tells us that the average age when most people get their first job.

2- **Inferential statistics:** it to generalizing from samples to populations, performing estimations and hypothesis testing, determining predictions. For example the relation between smoking and cancer.

Population: A collection, or set, of all individuals or objects or events whose properties are to be analyzed.

it is often impossible or impractical to observe the entire group, especially if it is large. Instead of examining the entire group, called the "population", one examines a small part of the group called a "sample".

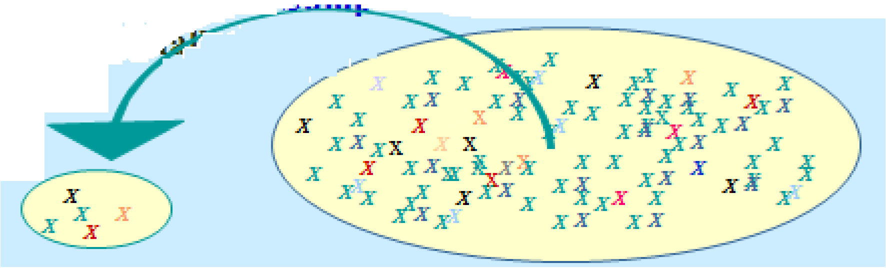
Sample: A portion or subset of the population.



Population

Sample

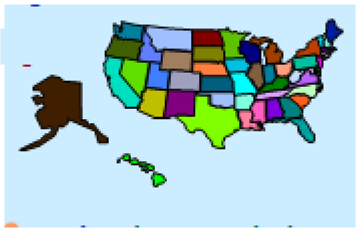
Random Sample: Each member of the population has an equal chance of being selected.



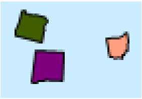
- **Parameters vs. Statistics**

A parameter is a number that describes the population characteristic. Usually its value is unknown.

Parameter: A number that describes a population characteristic. Example: Average gross income of all people in the United States in 2018.



Statistic: A number that describes a sample characteristic and it can be computed from the sample data . Example: *2018 gross income of people from a sample of three states.*



In practice, we often use a statistic to estimate an unknown parameter.

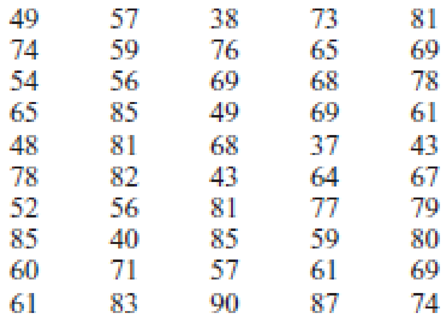
Data Organization

This chapter explains how to organize data by constructing frequency distributions and how to present the data by constructing graphs.

- ***raw data***.

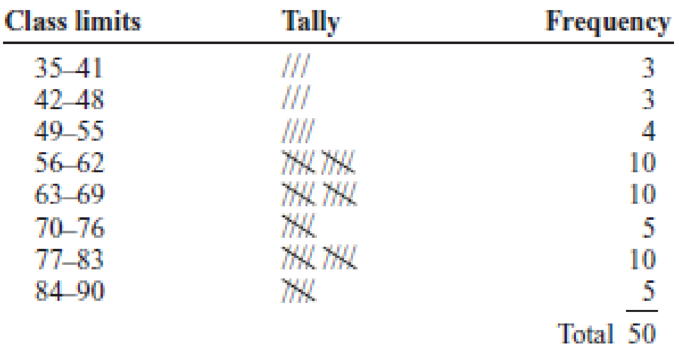
Raw data are collected in original form that have not been organized numerically. For example the following raw data:

Suppose a researcher wished to do a study on the ages of the top 50 wealthiest people in the world. The researcher first would have to get the data on the ages of the people. When the data are in original form, they are called **raw data** and are listed next.



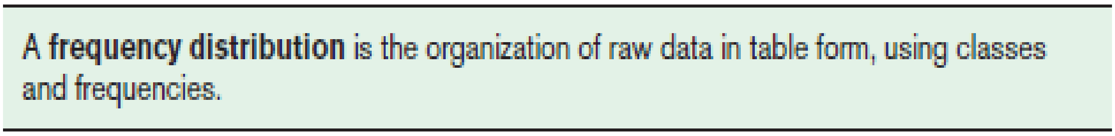
49	57	38	73	81
74	59	76	65	69
54	56	69	68	78
65	85	49	69	61
48	81	68	37	43
78	82	43	64	67
52	56	81	77	79
85	40	85	59	80
60	71	57	61	69
61	83	90	87	74

Since little information can be obtained from looking at raw data, the researcher organizes the data into what is called a *frequency distribution.* A frequency distribution consists of *classes* and their corresponding *frequencies.* Each raw data value is placedinto a quantitative or qualitative category called a **class.** The **frequency** of a class then is the number of data values contained in a specific class. A frequency distribution is shown for the preceding data set.



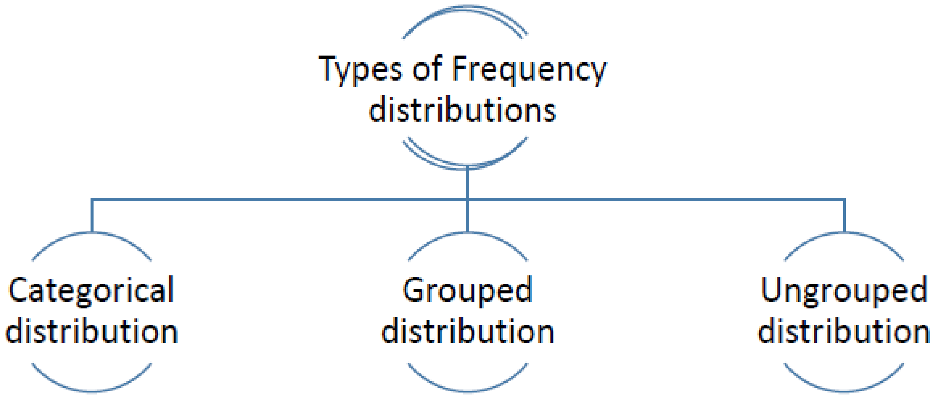
Class limits	Tally	Frequency
35–41	///	3
42–48	///	3
49–55	////	4
56–62	𝍸 𝍸	10
63–69	𝍸 𝍸	10
70–76	𝍸	5
77–83	𝍸 𝍸	10
84–90	𝍸	5
		Total 50

Now some general observations can be made from looking at the frequency distribution.For example, it can be stated that the majority of the wealthy people in the study are over 55 years old.



A **frequency distribution** is the organization of raw data in table form, using classes and frequencies.

Three types of frequency distributions that are most often used are the *categorical frequency distribution*, the *grouped frequency, and ungrouped frequency distribution.*



Types of Frequency distributions

Categorical distribution

Grouped distribution

Ungrouped distribution

1. **Categorical Frequency Distribution**

Example: Twenty-five army indicates were given a blood test to determine their blood type. The raw Data is:

A	B	B	AB	O
O	O	B	AB	B
B	B	O	A	O
A	O	O	O	AB
AB	A	O	B	A

Solution

Since the data are categorical, discrete classes can be used. There are four blood types: A, B, O, and AB. These types will be used as the classes for the distribution.

procedure for constructing a frequency distribution for categorical data is:

Step 1 Make a table as shown.

A	B	C	D
Class	**Tally**	**Frequency**	**Percent**
A			
B			
O			
AB			

Step 2 Tally the data and place the results in column B.

Step 3 Count the tallies and place the results in column C.

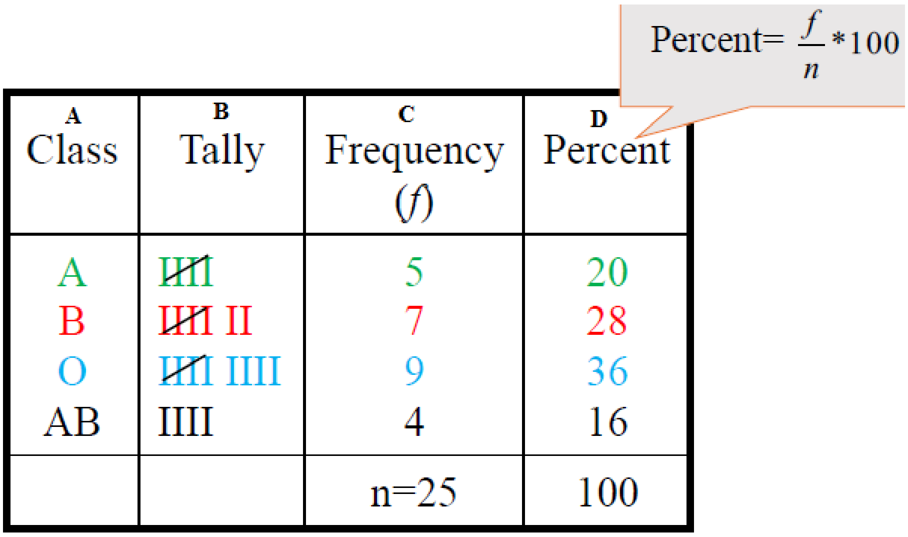
Step 4 Find the percentage of values in each class by using the formula

$$\% = \frac{f}{n} \cdot 100\%$$

where f = frequency of the class and n = total number of values. For example, in the class of type A blood, the percentage is

$$\% = \frac{5}{25} \cdot| 100\% = 20\%$$

Step 5 Find the totals for columns C (frequency) and D (percent). The completed table is shown.



Percent= $\frac{f}{n}$*100

A Class	B Tally	C Frequency (*f*)	D Percent
A	~~IIII~~	5	20
B	~~IIII~~ II	7	28
O	~~IIII~~ IIII	9	36
AB	IIII	4	16
		n=25	100

The *percentage* for a category is obtained by multiplying the relative frequency for that category

by 100. The sum of the percentages for all the categories will always equal 100 percent.

2. Grouped Frequency Distributions

When the range of the data is large, the data must be grouped into classes that are more than one unit in width, in what is called a grouped frequency distribution.

□ The smallest and largest possible data values in a class are the ***lower*** and ***upper class limits***. ***Class boundaries*** separate the classes.

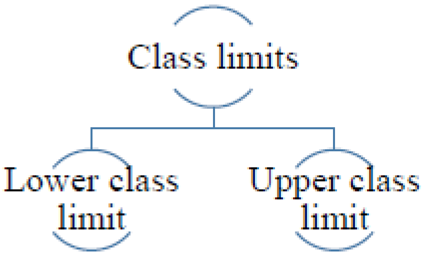
□ To find a class boundary, average the upper class limit of one class and the lower class limit of the next class.

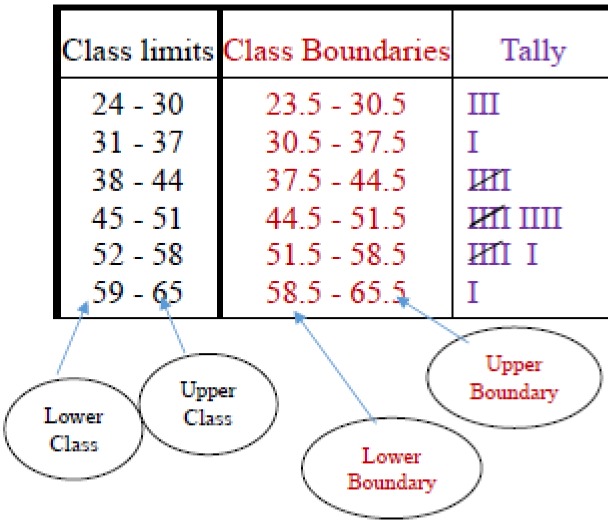
□ The **class width** can be calculated by subtracting □ successive lower class limits (or boundaries)

□ successive upper class limits (or boundaries)

□ upper and lower class boundaries .

□ The ***class midpoint Xm*** can be calculated by averaging □ upper and lower class limits (or boundaries)





Class limits	Class Boundaries	Tally	Frequency (*f*)
24 - 30	23.5 - 30.5	III	3
31 - 37	30.5 - 37.5	I	1
38 - 44	37.5 - 44.5	卌	5
45 - 51	44.5 - 51.5	卌 IIII	9
52 - 58	51.5 - 58.5	卌 I	6
59 - 65	58.5 - 65.5	I	1

Lower Class
Upper Class
Lower Boundary
Upper Boundary

- In the example, the values 24 and 30 of the first class are the **class limits.**

- The **lower class** limit is 24 and the **upper class** limit is 30.

- **The *Class boundaries*** are used to separate the classes. So that there are no gaps in the frequency distribution

- Class limits should have the same decimal place value as the data, but the class boundaries should have one additional place value and end in a 5.

For example: Class limit 7.8 – 8.8

Class boundary 7.75 – 8.85

➢**Lower boundary= lower limit - 0.05**

=7.8- 0.05 =7.75

➢**Upper boundary= upper limit + 0.05**

=8.8+0.05=8.85

Class width = lower of second class limit- lower of first class limit

Class width: 31 – 24 =7

The class midpoint Xm is found by adding the lower and upper class limit and dividing by 2.

$$X_m = \frac{lower\ limit + upper\ limit}{2}$$

In this example :

$$\frac{24+30}{2} = 27$$

Homework:

• Find the boundaries for the following class limits:

- 44 - 37
- 10.3 - 11.5
- 22.2 – 23.0
- 547.04 - 553.20

• Find the class width for the following class limits:

• 37 – 44

• 45 – 52

• 625 – 654

• 655 - 684

• Find the class width for the following class boundaries:

• 10.5 – 11.5

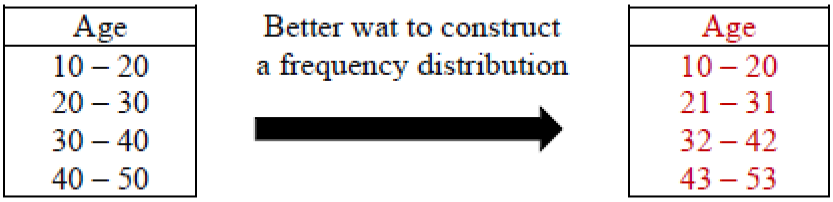
• 22.15 – 27.15

Rules for Classes in Grouped Frequency Distributions

1. There should be 5-20 classes.

2. The class width should be an odd number.

3. The classes must be mutually exclusive.



Age
10 – 20
20 – 30
30 – 40
40 – 50

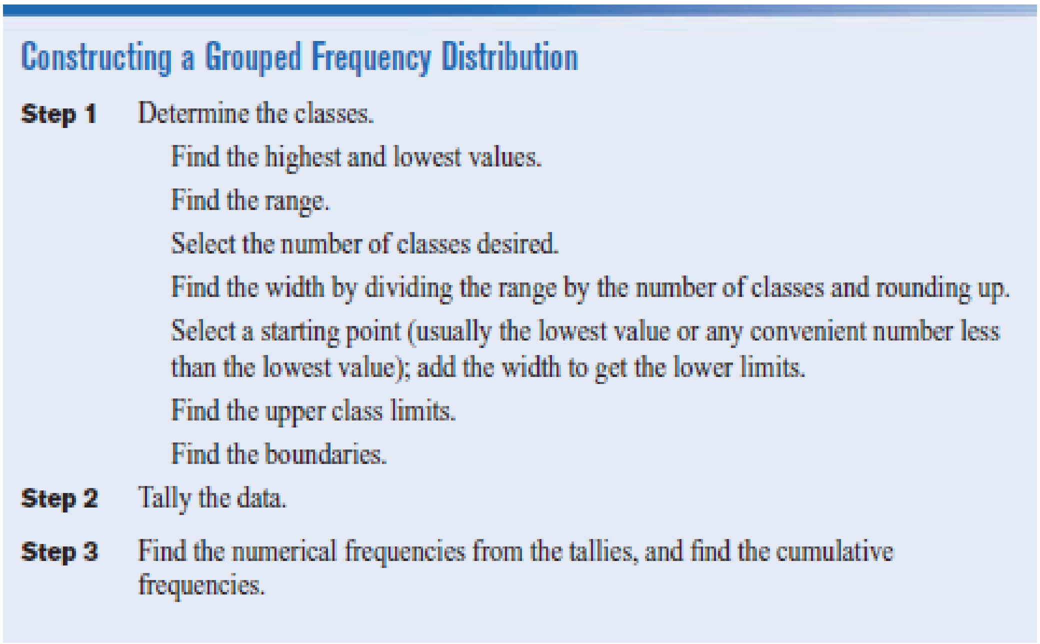
Better wat to construct a frequency distribution

Age
10 – 20
21 – 31
32 – 42
43 – 53

4. The classes must be continuous.

5. The classes must be exhaustive.

6. The classes must be equal in width (except in open-ended distributions).



Constructing a Grouped Frequency Distribution

Step 1 Determine the classes.

Find the highest and lowest values.

Find the range.

Select the number of classes desired.

Find the width by dividing the range by the number of classes and rounding up.

Select a starting point (usually the lowest value or any convenient number less than the lowest value); add the width to get the lower limits.

Find the upper class limits.

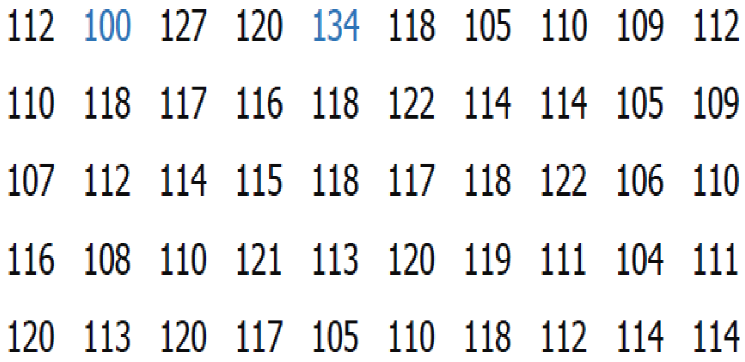
Find the boundaries.

Step 2 Tally the data.

Step 3 Find the numerical frequencies from the tallies, and find the cumulative frequencies.

Example 2–2

The following data represent the record high temperatures for each of the 50 states. Construct a grouped frequency distribution for the data using 7 classes.



112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

STEP 1 Determine the classes. Find the class width by dividing the range by the number of classes 7.

Range = High – Low

= 134 – 100 = 34

Width = Range / 7 = 34 /7 = 5

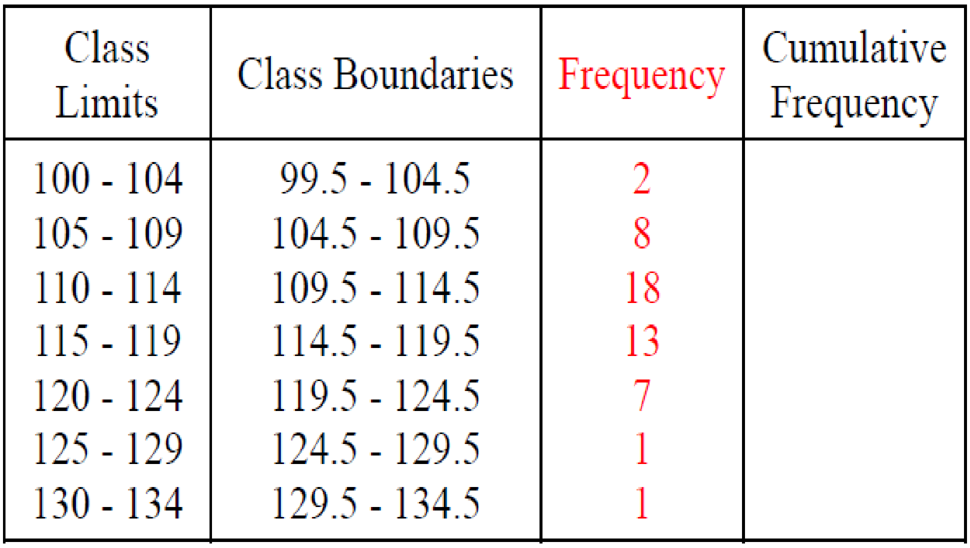
Note: Rounding Rule: Always round up if a remainder.

Round the answer up to the nearest whole number if there is a remainder: او

4.9 _ 5. (Rounding *up* is different from rounding *off.* Anumber is rounded up if there is any decimal remainder when dividing. For example, 85 _ 6 _ 14.167 and is rounded up to 15. Also, 53 _ 4 _ 13.25 and is rounded up to 14. Also, after dividing, if there is no remainder, you will need to add an extra class to accommodate all the data.)

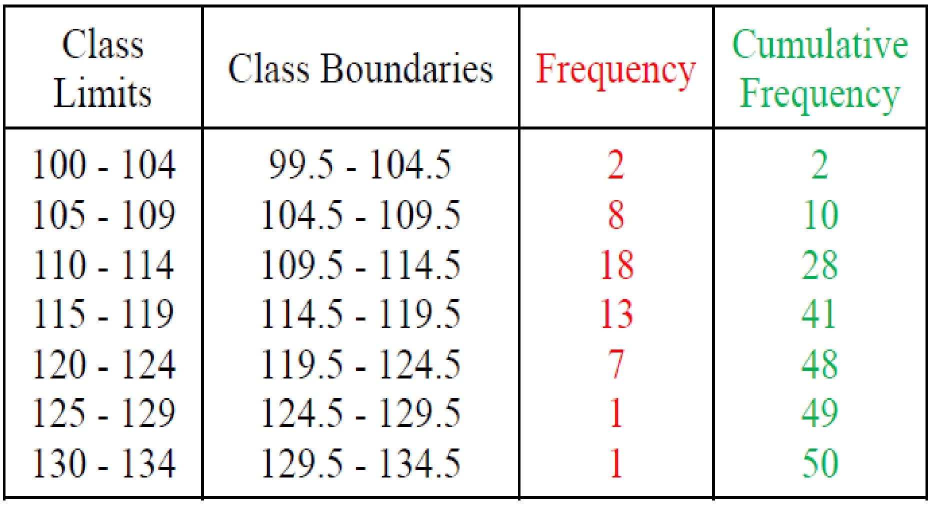
STEP 2 Tally the data.

STEP 3 Find the frequencies.



Class Limits	Class Boundaries	Frequency	Cumulative Frequency
100 - 104	99.5 - 104.5	2	
105 - 109	104.5 - 109.5	8	
110 - 114	109.5 - 114.5	18	
115 - 119	114.5 - 119.5	13	
120 - 124	119.5 - 124.5	7	
125 - 129	124.5 - 129.5	1	
130 - 134	129.5 - 134.5	1	

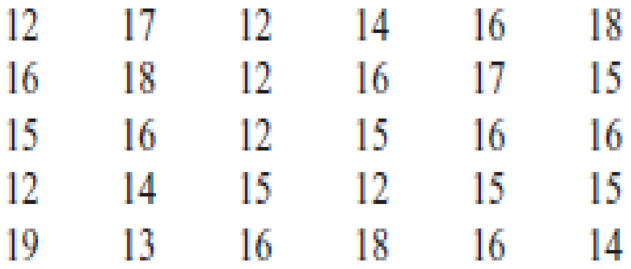
STEP 4 Find the cumulative frequencies by keeping a running total of the frequencies.



Class Limits	Class Boundaries	Frequency	Cumulative Frequency
100 - 104	99.5 - 104.5	2	2
105 - 109	104.5 - 109.5	8	10
110 - 114	109.5 - 114.5	18	28
115 - 119	114.5 - 119.5	13	41
120 - 124	119.5 - 124.5	7	48
125 - 129	124.5 - 129.5	1	49
130 - 134	129.5 - 134.5	1	50

Ungrouped Frequency distribution:

The data shown here represent the number of miles per gallon (mpg) that 30 selected
four-wheel-drive sports utility vehicles obtained in city driving.
Construct a frequency distribution, and analyze the distribution?



12	17	12	14	16	18
16	18	12	16	17	15
15	16	12	15	16	16
12	14	15	12	15	15
19	13	16	18	16	14

Solution

Step 1 Determine the classes.

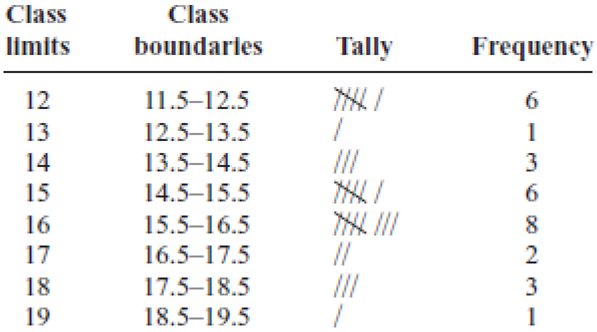
Since the range of the data set is small (19 - 12 = 7),
classes consisting of a single data value can be used.
They are 12, 13, 14, 15, 16, 17, 18, 19.

Note: If the data are continuous, class boundaries can be used. Subtract 0.5 from each class value to get the lower class boundary, and add 0.5 to each class value to get the upper class boundary.

Step 2 Tally the data.

Step 3 Find the numerical frequencies from the tallies, and find the cumulative frequencies.

The completed ungrouped frequency distribution is:



Class limits	Class boundaries	Tally	Frequency
12	11.5–12.5	𝍸 /	6
13	12.5–13.5	/	1
14	13.5–14.5	///	3
15	14.5–15.5	𝍸 /	6
16	15.5–16.5	𝍸 ///	8
17	16.5–17.5	//	2
18	17.5–18.5	///	3
19	18.5–19.5	/	1

In this case, almost one-half (14) of the vehicles get 15 or 16 miles per gallon.

The cumulative frequencies are:

	Cumulative frequency
Less than 11.5	0
Less than 12.5	6
Less than 13.5	7
Less than 14.5	10
Less than 15.5	16
Less than 16.5	24
Less than 17.5	26
Less than 18.5	29
Less than 19.5	30

- ## Graphic representation

After the data have been organized into a frequency distribution, they can be presented in graphic forms. The purpose of graphs in statistics is to convey the data to the viewer in pictorial form. It is easier for most people to comprehend the meaning of data presented graphically than data presented numerically in tables or frequency distributions.

Statistical graphs can be used to describe the data set or analyze it. Graphs are also useful in getting the audience's attention in a publication or a speaking presentation. They can be used to discuss an issue, reinforce a critical point, or summarize a data set.

They can also be used to discover a trend or pattern in a situation over a period of time.

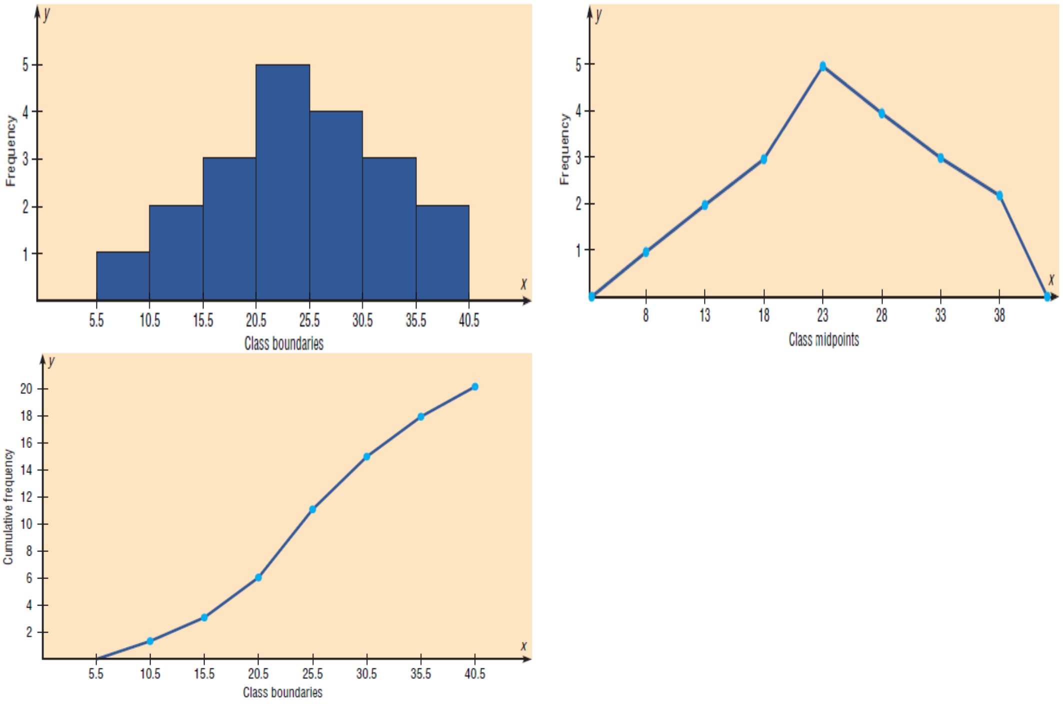
The three most commonly used graphs in research are:

1. The histogram.

2. The frequency polygon.

3. The cumulative frequency graph, or ogive (pronounced o-jive).

An example of each type of graph is shown in the following figure.



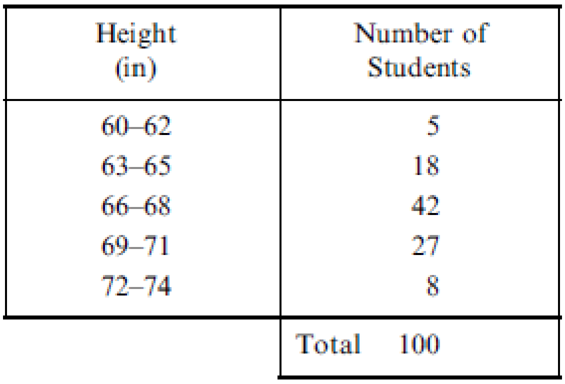
histogram **frequency polygon**

ogive

1- HISTOGRAMS

Histograms is a graphic representations that displays the data by using vertical bars of various heights to represent the frequencies distributions of classes.

ExampleL: Histogram corresponding to frequency distribution of heights in Table1 are shown in Figs. 1



Height (in)	Number of Students
60–62	5
63–65	18
66–68	42
69–71	27
72–74	8
	Total 100

Table1 Heights of 100 male students at X University

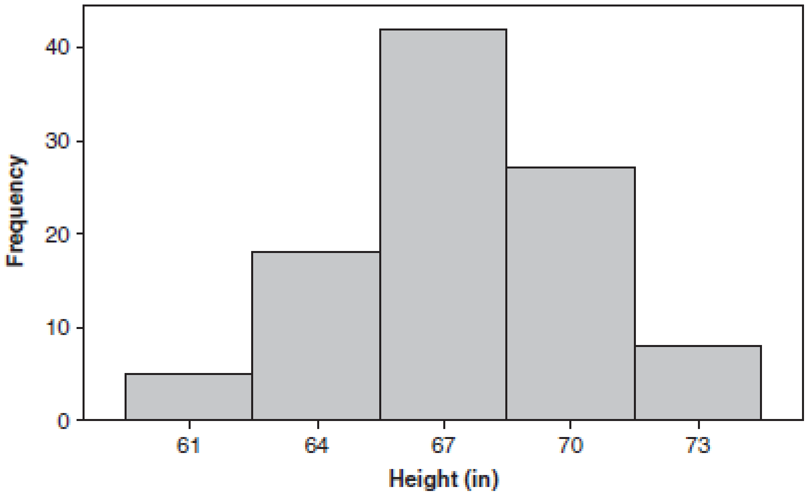
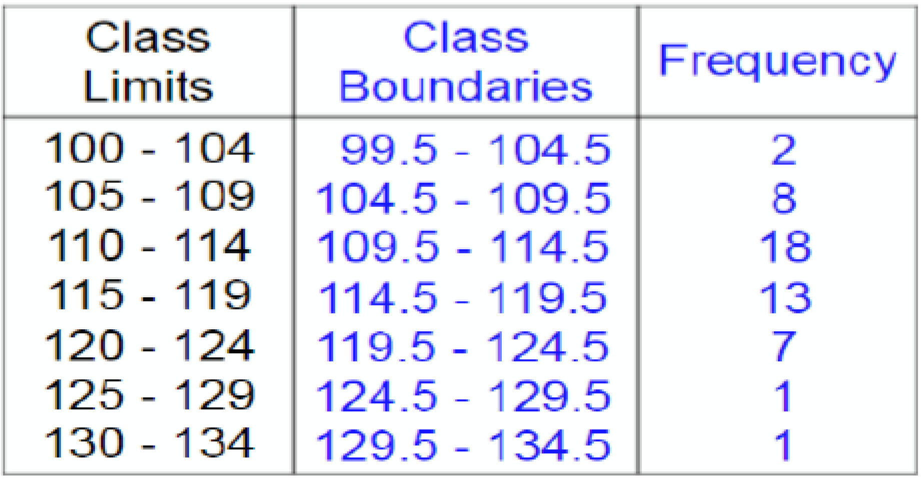


Figure1

Example: Construct a histogram to represent the data shown below for the record high temperatures.



Class Limits	Class Boundaries	Frequency
100 - 104	99.5 - 104.5	2
105 - 109	104.5 - 109.5	8
110 - 114	109.5 - 114.5	18
115 - 119	114.5 - 119.5	13
120 - 124	119.5 - 124.5	7
125 - 129	124.5 - 129.5	1
130 - 134	129.5 - 134.5	1

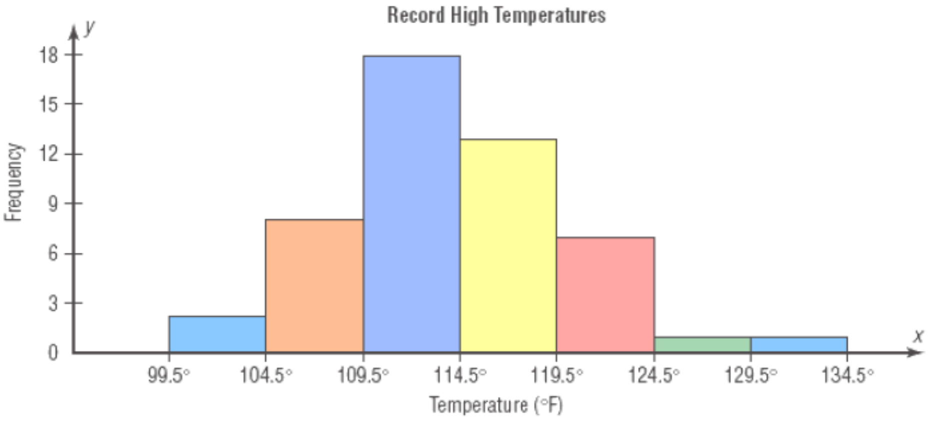
Solution

STEP 1 Draw and label the *x* and *y* axes. The *x* axis is always the horizontal axis,

and the *y* axis is always the vertical axis.

STEP 2 Represent the frequency on the *y* axis and the class boundaries on the *x* axis.

STEP 3 Using the frequencies as the heights, draw vertical bars for each class. As in the following figure:



2-The Frequency Polygon

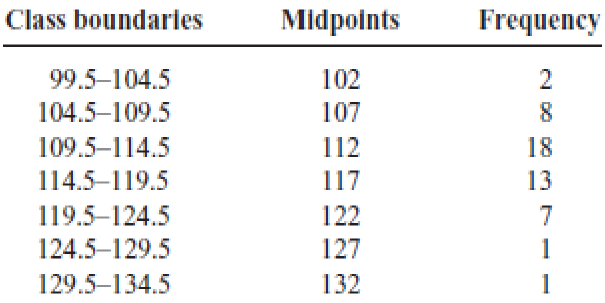
construct a frequency polygon

Solution

Step 1 Find the midpoints of each class. Recall that midpoints are found by adding the upper and lower boundaries and dividing by 2:

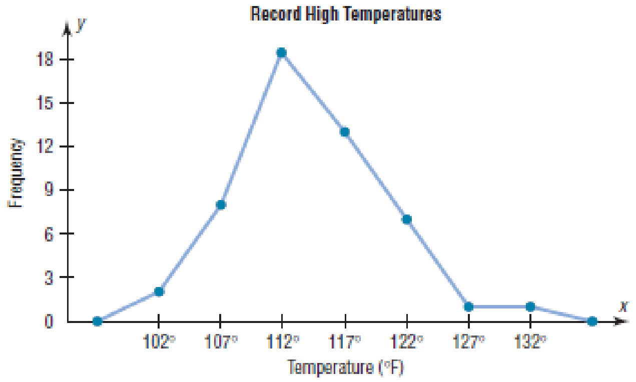
$$\frac{99.5 + 104.5}{2} = 102 \qquad \frac{104.5 + 109.5}{2} = 107$$

and so on. The midpoints are:



Class boundaries	Midpoints	Frequency
99.5–104.5	102	2
104.5–109.5	107	8
109.5–114.5	112	18
114.5–119.5	117	13
119.5–124.5	122	7
124.5–129.5	127	1
129.5–134.5	132	1

Step 2 Draw the *x* and *y* axes. Label the *x* axis with the midpoint of each class, and then use a suitable scale on the *y* axis for the frequencies.



Step 3 Using the midpoints for the *x* values and the frequencies as the *y* values, plot the points.

Step 4 Connect adjacent points with line segments.

The frequency polygon and the histogram are two different ways to represent the same data set.

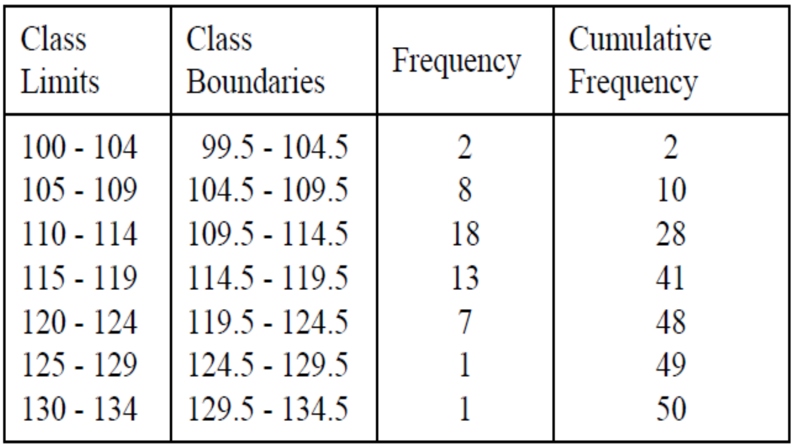
3- The Ogive

The third type of graph that can be used represents the cumulative frequencies for the classes. This type of graph is called the *cumulative frequency graph*, or *ogive*. The **cumulative frequency** is the sum of the frequencies accumulated up to the upper boundary of a class in the distribution.

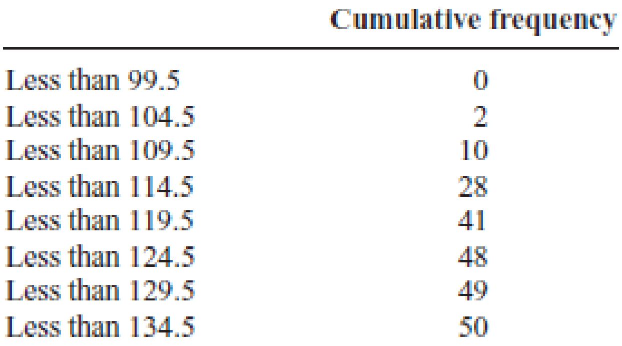
Construct an ogive for the frequency distribution

Solution

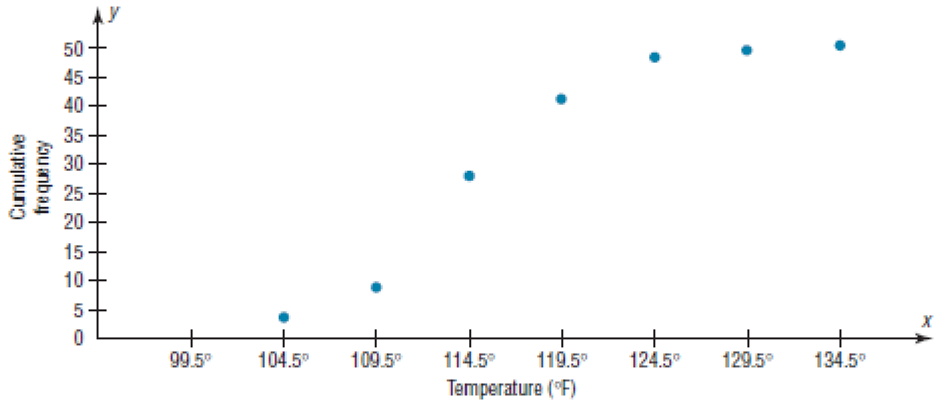
Ogives use upper class boundaries and cumulative frequencies of the classes.

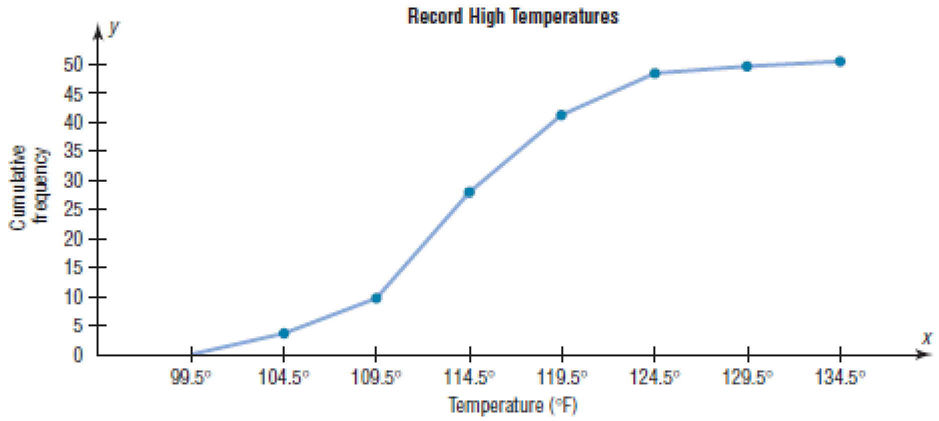


Class Limits	Class Boundaries	Frequency	Cumulative Frequency
100 - 104	99.5 - 104.5	2	2
105 - 109	104.5 - 109.5	8	10
110 - 114	109.5 - 114.5	18	28
115 - 119	114.5 - 119.5	13	41
120 - 124	119.5 - 124.5	7	48
125 - 129	124.5 - 129.5	1	49
130 - 134	129.5 - 134.5	1	50

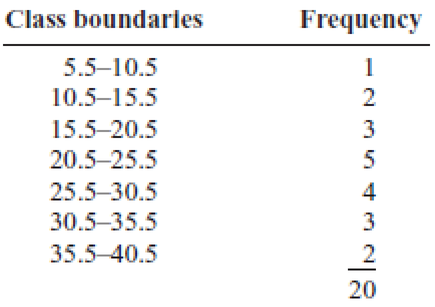


	Cumulative frequency
Less than 99.5	0
Less than 104.5	2
Less than 109.5	10
Less than 114.5	28
Less than 119.5	41
Less than 124.5	48
Less than 129.5	49
Less than 134.5	50





Homework: Construct a histogram, frequency polygon, and ogive using relative frequencies for the distribution of the miles that 20 randomly selected runners ran during a given week.

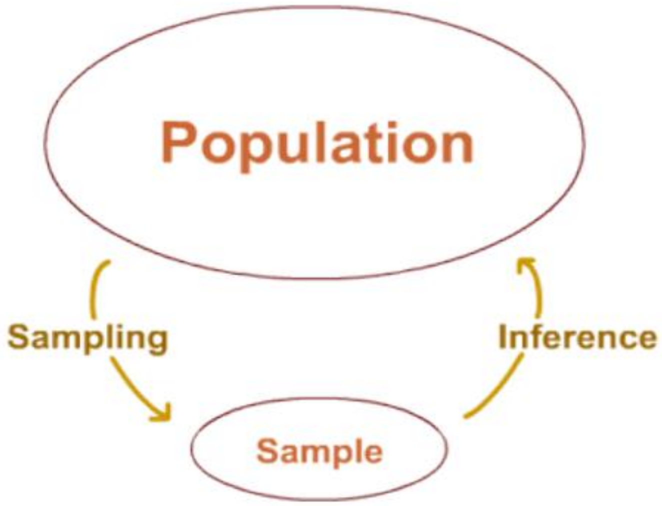


Class boundaries	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2
	20

Descriptive Measures

statisticians use samples taken from populations; however, when populations are small, it is not necessary to use samples since the entire population can be used to gain information. For example, suppose an insurance manager wanted to know the average weekly sales of all the company's representatives. If the company employed a large number of salespeople, say nationwide, he would have to use a sample and make an inference to the entire sales force. But if the company had only a few salespeople, say only 87 agents, he would be able to use all representatives' sales for a randomly chosen week and thus use the entire population.

Measures taken by using all the data values in the populations are called _parameters_. Measures obtained by using the data values of samples are called _statistics_.



MEASURES OF CENTRAL TENDENCY (MEAN, MEDIAN, AND MODE)

There are many different measures of central tendency. The three most widely used are ***mean, median,*** and ***mode.***

1- The mean

The *mean*, also known as the arithmetic average, is found by adding the values of the data and dividing by the total number of values. For example, the mean of 3, 2, 6, 5, and 4 is found by adding 3 + 2 + 6 + 5 + 4 = 20 and dividing by 5; hence, the mean of the data is 20 / 5 = 4. The values of the data are represented by *X*’s. In this data set, *X*1 = 3, *X*2 = 2, *X*3 = 6, *X*4 = 5, and *X*5 = ╚ 4. To show a sum of the total *X* values, the symbol **Σ** (the capital Greek letter sigma) is used, and **Σ***X* means to find the sum of the *X* values in the data set.

The mean is the sum of the values divided by the total number of values. The symbol represents the sample mean.

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

where n represents the total number of values in the sample.

For a population, the Greek letter (mu) is used for the mean.

$$\mu = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\Sigma X}{N}$$

where N represents the total number of values in the population.

Example 1: The data represent the number of different plans 10 HMO systems offer their enrollees. Find the mean:

84, 12, 27, 15, 40, 18, 33, 33, 14, 4.

Solution

$$\overline{X} = \frac{\Sigma X}{n} = \frac{84 + 12 + 27 + 15 + 40 + 18 + 33 + 33 + 14 + 4}{10}$$

$$\overline{X} = \frac{280}{10} = 28$$

The mean is 28 plans.

Example 2: The fat contents in grams for one serving of 11 brands of packaged foods, as determined by the U.S. Department of Agriculture, are given as follows. Find the mean.

6.5, 6.5, 9.5, 8.0, 14.0, 8.5, 3.0, 7.5, 16.5, 7.0, 8.0

Solution

$$\overline{X} = \frac{\Sigma X}{n} = \frac{6.5 + 6.5 + 9.5 + 8.0 + 14.0 + 8.5 + 3.0 + 7.5 + 16.5 + 7.0 + 8.0}{11}$$

$$= \frac{95}{11} = 8.64 \text{ grams}$$

Hence, the mean fat content is 8.64 grams.

The procedure for finding the mean for grouped data uses the midpoints of the classes. This procedure is shown next.

Example3: **Miles Run per Week**

Using the frequency distribution for Example 2–7, find the mean. The data represent the

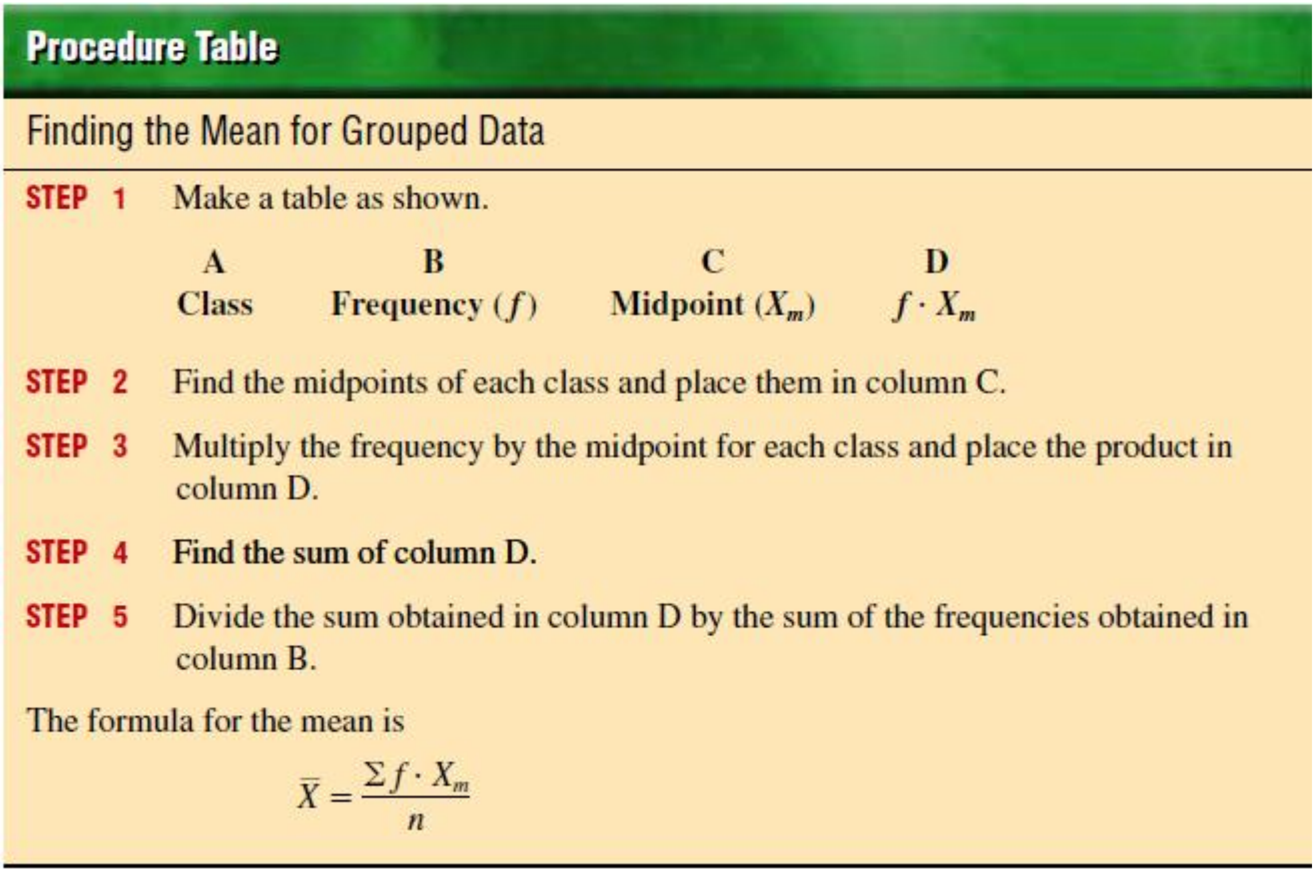
number of miles run during one week for a sample of 20 runners.

Solution

The procedure for finding the mean for grouped data is given here.

Step 1 Make a table as shown.

ص ١٠٧ اكملي



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

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

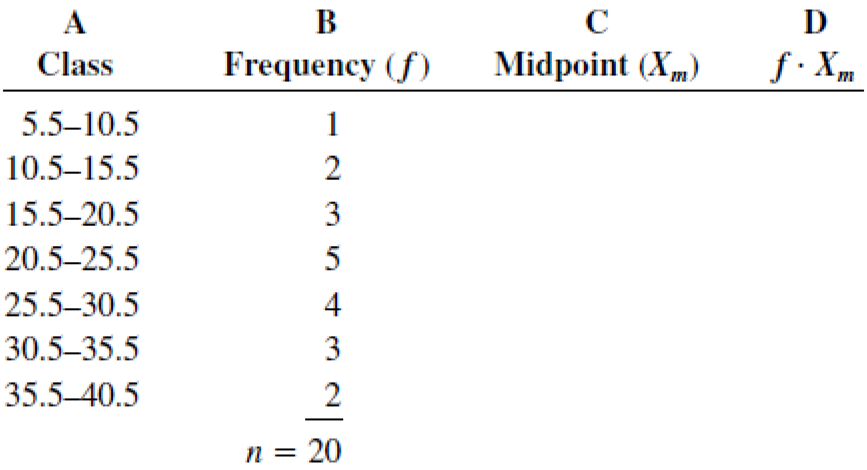
Example 3: Using the frequency distribution for Example 2–7 in Chapter 2, find the mean. The data

represent the number of miles run during one week for a sample of 20 runners.

Solution

The procedure for finding the mean for grouped data is given here.

STEP 1 Make a table as shown.



A Class	B Frequency (f)	C Midpoint (X_m)	D $f \cdot X_m$
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		
	$n = 20$		

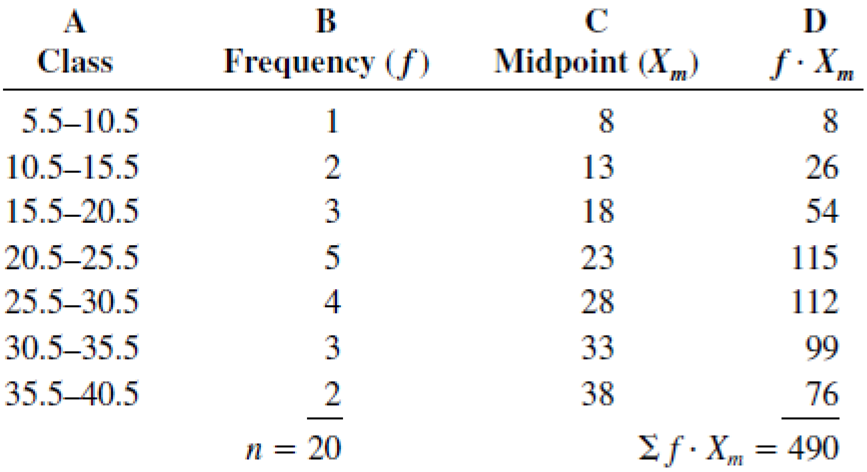
STEP 2 Find the midpoints of each class and enter them in column C.

$$X_m = \frac{5.5 + 10.5}{2} = 8, \qquad \frac{10.5 + 15.5}{2} = 13, \qquad \text{etc.}$$

STEP 3 For each class, multiply the frequency by the midpoint, as shown below, and place the product in column D.

$$1 \cdot 8 = 8, \qquad 2 \cdot 13 = 26, \qquad \text{etc.}$$

The completed table is shown here.



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
	$n = 20$		$\Sigma f \cdot X_m = 490$

STEP 4 Find the sum of column D, as shown above.

STEP 5 Divide the sum by *n* to get the mean.

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

2- **The median** is the halfway point in a data set. Before one can find this point, the data must be arranged in order. When the data set is ordered, it is called a **data array.** The median either will be a specific value in the data set or will fall between two values.

The median is the midpoint of the data array. The symbol for the median is MD.

Steps in computing the median of a data array

STEP 1 Arrange the data in order.

STEP 2 Select the middle point.

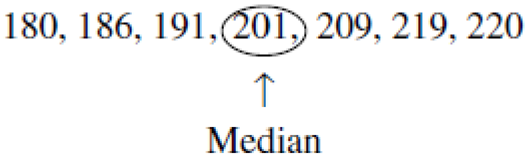
Example 4: The weights (in pounds) of seven army recruits are 180, 201, 220, 191, 219, 209, and 186. Find the median.

Solution

STEP 1 Arrange the data in order.

180, 186, 191, 201, 209, 219, 220

STEP 2 Select the middle value.



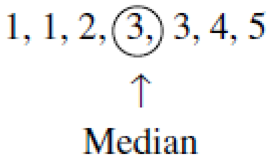
$180, 186, 191, \underset{\substack{\uparrow \\ \text{Median}}}{\boxed{201,}}\ 209, 219, 220$

Hence, the median weight is 201 pounds.

Example 5: Find the median for the ages of seven preschool children. The ages are

1, 1, 3, 4, 2, 3, 5,

Solution



$1, 1, 2, \underset{\substack{\uparrow \\ \text{Median}}}{\boxed{3,}}\ 3, 4, 5$

Hence, the median age is 3 years.

Each of these examples had an odd number of values in the data set; hence, the median was an actual data value. When there is an even number of values in the data set,the median will fall between two given values, as illustrated in the following examples.

Example 6: The number of tornadoes that have occurred in the United States over an eight-year period follows. Find the median.

684, 764, 656, 702, 856, 1133, 1132, 1303

Solution

$$656,\ 684,\ 702,\ 764,\ \underset{\substack{\uparrow \\ \text{Median}}}{}\ 856,\ 1132,\ 1133,\ 1303$$

Since the middle point falls halfway between 764 and 856, find the median by adding the two values and dividing by 2.

$$\text{MD} = \frac{764 + 856}{2} = \frac{1620}{2} = 810$$

The median number of tornadoes is 810.

Example 7: The ages of 10 college students are given below. Find the median.

18, 24, 20, 35, 19, 23, 26, 23, 19, 20

Solution

$$18,\ 19,\ 19,\ 20,\ 20,\ \underset{\substack{\uparrow \\ \text{Median}}}{}\ 23,\ 23,\ 24,\ 26,\ 35$$

$$\text{MD} = \frac{20 + 23}{2} = 21.5$$

Hence, the median age is 21.5 years.

Example 8: Six customers purchased the following number of magazines: 1, 7, 3, 2, 3, 4. Find the median.

Solution

$$1,\ 2,\ 3,\ \underset{\substack{\uparrow \\ \text{Median}}}{}\ 3,\ 4,\ 7, \qquad \text{MD} = \frac{3 + 3}{2} = 3$$

Hence, the median number of magazines purchased is 3.

3- The **mode**. is the value that occurs most often in the data set. It is sometimes said to be the most typical case.

A data set can have more than one mode or no mode at all.

Example 9: The following data represent the duration (in days) of U.S. space shuttle voyages for the

years 1992–94. Find the mode.

8, 9, 9, 14, 8, 8, 10, 7, 6, 9, 7, 8, 10, 14, 11, 8, 14, 11

Solution

It is helpful to arrange the data in order, although it is not necessary.

6, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 10, 10, 11, 11, 14, 14, 14

Since 8-day voyages occurred five times—a frequency larger than any other number— the mode for the data set is 8.

Example 10: Find the mode for the number of coal employees per county for 10 selected counties in Southwestern Pennsylvania.

110, 731, 1031, 84, 20, 118, 1162, 1977, 103, 752

Solution

Since each value occurs only once, there is no mode.

***Note**: Do not say that the mode is zero.* That would be incorrect, because in some data, such as temperature, zero can be an actual value.

Example 11: Eleven different automobiles were tested at a speed of 15 miles per hour for stopping

distances. The data, in feet, are shown below. Find the mode.

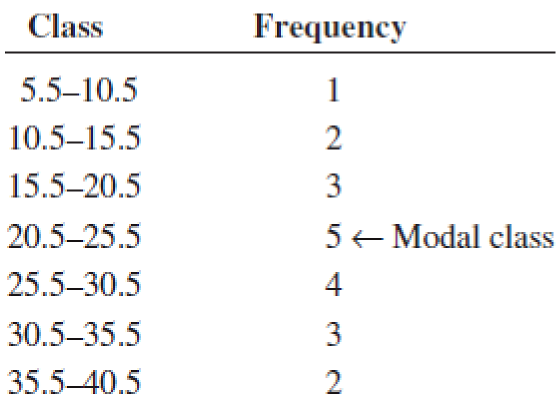
15, 18, 18, 18, 20, 22, 24, 24, 24, 26, 26

Solution

Since 18 and 24 both occur three times, the modes are 18 and 24 feet. This data set is said to be *bimodal.*

The mode for grouped data is the modal class. The **modal class** is the class with the largest frequency.

Example 12: Find the modal class for the frequency distribution of miles 20 runners ran in one week, used in the following table:



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

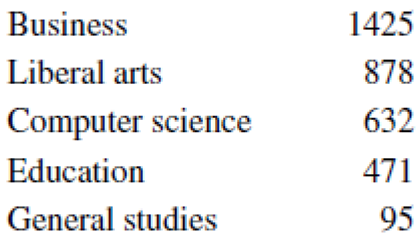
Solution

The modal class is 20.5–25.5, since it has the largest frequency.

The mode is the only measure of central tendency that can be used in finding the most typical case when the data are nominal or categorical.

Example 13: A survey showed the following distribution for the number of students enrolled in each

field. Find the mode.



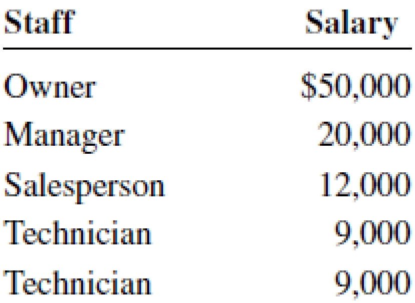
Business	1425
Liberal arts	878
Computer science	632
Education	471
General studies	95

Solution

Since the category with the highest frequency is business, the most typical case is a business major.

For a data set, the mean, median, and mode can be quite different. Consider the following example.

Example 14:A small company consists of the owner, the manager, the salesperson, and two technicians, all of whose annual salaries are listed here. (Assume that this is the entire population.)



Staff	Salary
Owner	$50,000
Manager	20,000
Salesperson	12,000
Technician	9,000
Technician	9,000

Find the mean, median, and mode.

Solution

$$\mu = \frac{\Sigma X}{N} = \frac{50{,}000 + 20{,}000 + 12{,}000 + 9{,}000 + 9{,}000}{5} = \$20{,}000$$

Hence, the mean is $20,000, the median is $12,000, and the mode is $9,000.

In this example, the mean is much higher than the median or the mode. This is because the extremely high salary of the owner tends to raise the value of the mean.

In this and similar situations, the median should be used as the measure of central tendency.

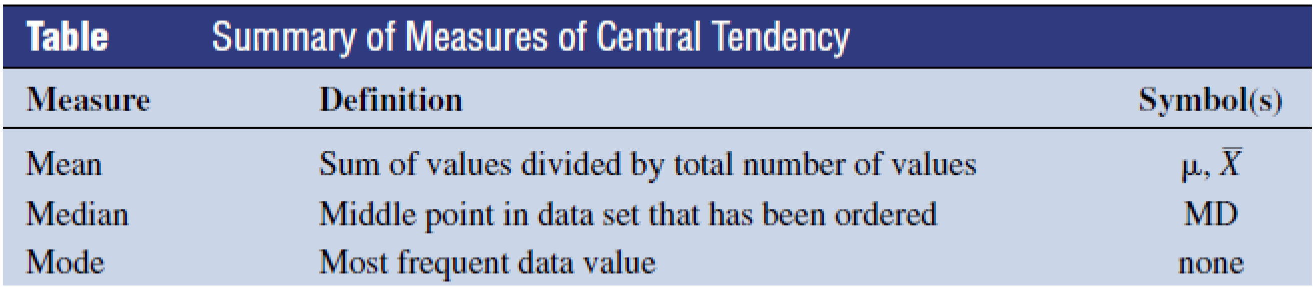
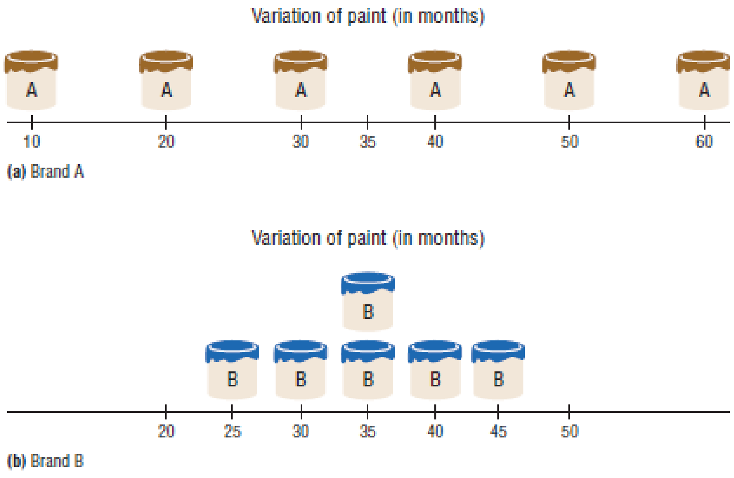


Table Summary of Measures of Central Tendency

Measure	Definition	Symbol(s)
Mean	Sum of values divided by total number of values	$\mu, \overline{X}$
Median	Middle point in data set that has been ordered	MD
Mode	Most frequent data value	none

Measures of Variation

To describe a distribution well, we need more than just the measures of center. We also need to know how the data is spread out or how it varies. As in the following Figure shows, even though the means are the same for both brands, the spread, or variation, is quite different. Two , measures are commonly used: variance, and standard deviation to measure the variation



(a) Brand A

(b) Brand B

□ The Variance: is the average of the squares of the distance each value is from the mean.

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

□ The **standard deviation** is the square root of the variance.

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

Step 1 Find the mean for the data. $\mu = \frac{\Sigma X}{N}$

Step 2 Find the Deviation for each data value. $X - \mu$

Step 3 Square each of the deviations. $(X - \mu)^2$

Step 4 Find the sum of the squares. $\Sigma(X - \mu)^2$

Example: Find the variance and standard deviation for the data set for Brand A paint.

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.

$$\begin{array}{lll} 10 - 35 = -25 & 50 - 35 = +15 & 40 - 35 = +5 \\ 60 - 35 = +25 & 30 - 35 = -5 & 20 - 35 = -15 \end{array}$$

Step 3 Square each result.

$$\begin{array}{lll} (-25)^2 = 625 & (+15)^2 = 225 & (+5)^2 \ = 25 \\ (+25)^2 = 625 & (-5)^2 \ = 25 & (-15)^2 = 225 \end{array}$$

Step 4 Find the sum of the squares.

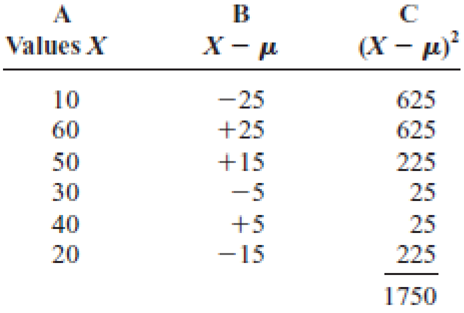
$$625 + 625 + 225 + 25 + 25 + 225 = 1750$$

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

$$\text{Variance} = 1750 \div 6 = 291.7$$

Step 6 Take the square root of the variance to get the standard deviation. Hence, the

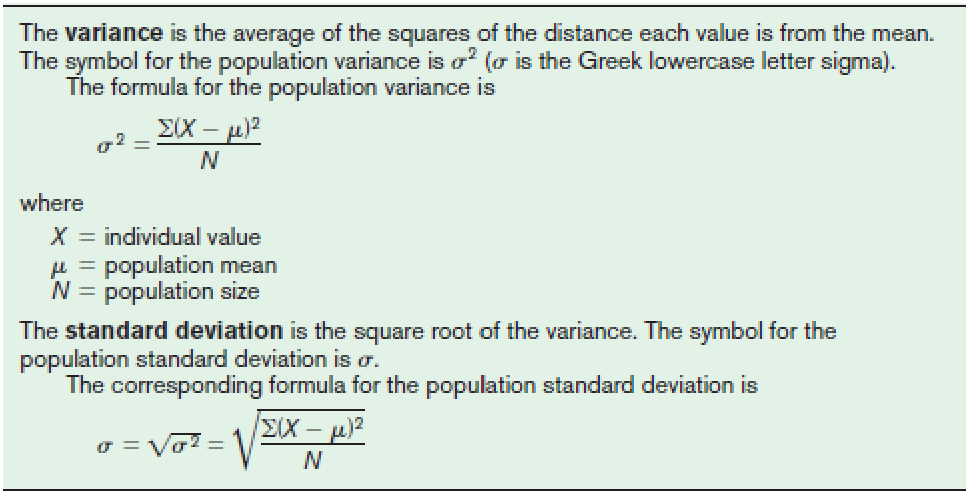
standard deviation equals , or 17.1. It is helpful to make a table.



A	B	C
Values X	$X - \mu$	$(X - \mu)^2$
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 **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 : Find the variance and standard deviation for brand B paint data:

The months were 35, 45, 30, 35, 40, 25

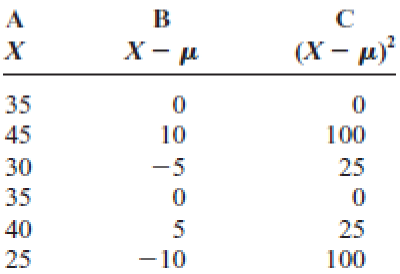
Solution

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

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 last Example) and the standard deviation of brand B is 6.5, the data are more variable for brand A. *In slummary, when the means are equal, the larger the variance or standard deviation is, the more variable thedata are.*

Example : 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

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.

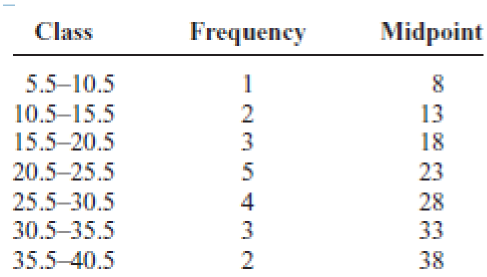
$$s = \sqrt{1.28} = 1.13$$

Hence, the sample standard deviation is 1.13.

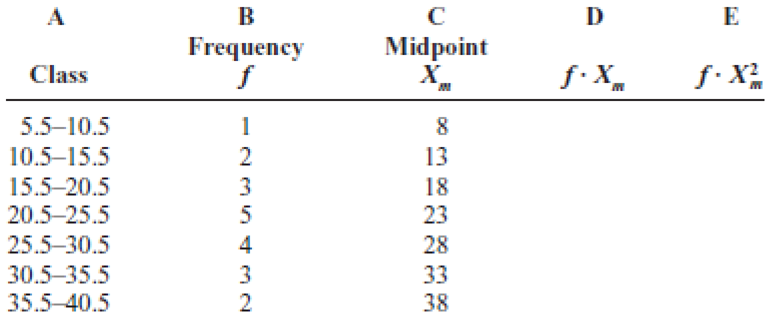
Variance and Standard Deviation for Grouped Data

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 : 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.

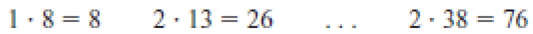


Class	Frequency	Midpoint
5.5–10.5	1	8
10.5–15.5	2	13
15.5–20.5	3	18
20.5–25.5	5	23
25.5–30.5	4	28
30.5–35.5	3	33
35.5–40.5	2	38



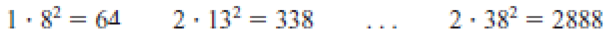
A	B	C	D	E
	Frequency	Midpoint		
Class	f	X_m	$f \cdot X_m$	$f \cdot X_m^2$
5.5–10.5	1	8		
10.5–15.5	2	13		
15.5–20.5	3	18		
20.5–25.5	5	23		
25.5–30.5	4	28		
30.5–35.5	3	33		
35.5–40.5	2	38		

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



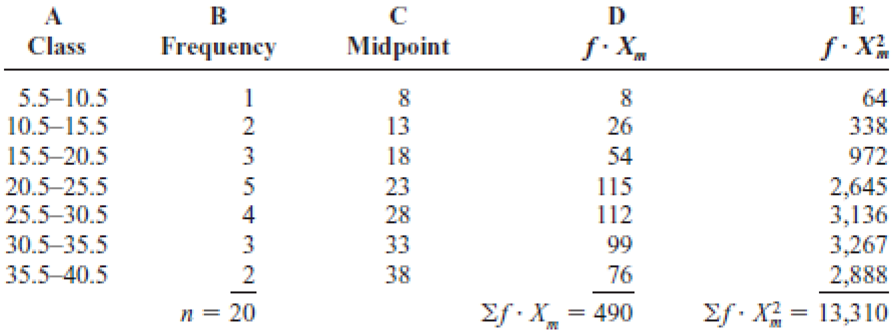
$1 \cdot 8 = 8 \qquad 2 \cdot 13 = 26 \qquad \ldots \qquad 2 \cdot 38 = 76$

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



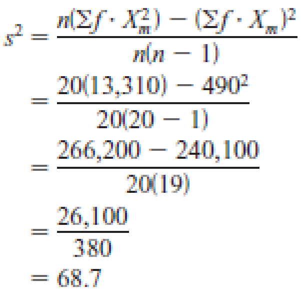
$1 \cdot 8^2 = 64 \qquad 2 \cdot 13^2 = 338 \qquad \ldots \qquad 2 \cdot 38^2 = 2888$

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 summation_*f* . *Xm*, and the sum of column E is _*f* .X^2 The completed table is shown.



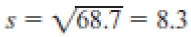
A Class	B Frequency	C Midpoint	D $f \cdot X_m$	E $f \cdot X_m^2$
5.5–10.5	1	8	8	64
10.5–15.5	2	13	26	338
15.5–20.5	3	18	54	972
20.5–25.5	5	23	115	2,645
25.5–30.5	4	28	112	3,136
30.5–35.5	3	33	99	3,267
35.5–40.5	2	38	76	2,888
	$n = 20$		$\Sigma f \cdot X_m = 490$	$\Sigma f \cdot X_m^2 = 13{,}310$

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

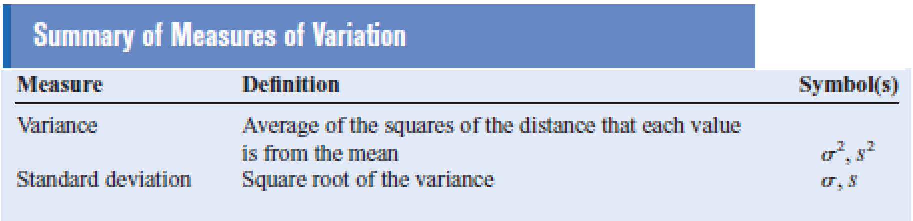


$$\begin{aligned} s^2 &= \frac{n(\Sigma f \cdot X_m^2) - (\Sigma f \cdot X_m)^2}{n(n-1)} \\ &= \frac{20(13{,}310) - 490^2}{20(20-1)} \\ &= \frac{266{,}200 - 240{,}100}{20(19)} \\ &= \frac{26{,}100}{380} \\ &= 68.7 \end{aligned}$$

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



$s = \sqrt{68.7} = 8.3$



Summary of Measures of Variation

Measure	Definition	Symbol(s)
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