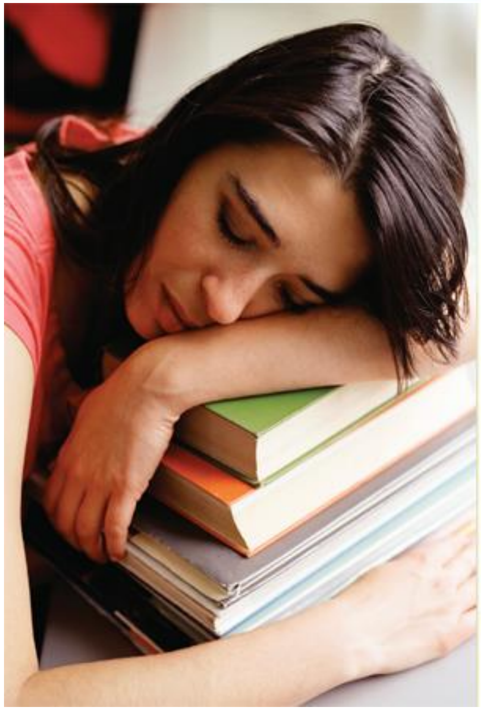
2 Descriptive Analysis and Presentation of Single-Variable Data



2.2 Frequency Distributions and Histograms

Frequency Distributions and Histograms

Frequency distribution A listing, often expressed in chart form, that pairs values of a variable with their frequency.

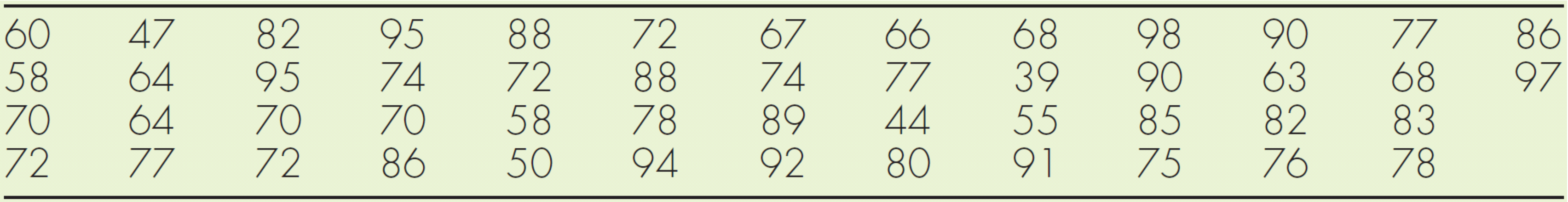
An **ungrouped frequency distribution**—“ungrouped” because each value of *x* in the distribution stands alone.

When a large set of data has many different *x* values instead of a few repeated values, we can group the values into a set of classes and construct a **grouped frequency distribution**.

Example 6 – *Grouping Data to Form a Frequency Distribution*

To illustrate this grouping (or classifying) procedure, let’s use a sample of 50 final exam scores taken from last semester’s elementary statistics class.

Table 2.6 lists the 50 scores.



60	47	82	95	88	72	67	66	68	98	90	77	86
58	64	95	74	72	88	74	77	39	90	63	68	97
70	64	70	70	58	78	89	44	55	85	82	83	
72	77	72	86	50	94	92	80	91	75	76	78	

Statistics Exam Scores **[TA02-06]**

Table 2.6

Example 6 – *Grouping Data to Form a Frequency Distribution*

cont'd

Procedure for Constructing a Grouped Frequency Distribution

1. Identify the high score (H = 98) and the low score (L = 39), and find the range:

$$\text{range} = H - L = 98 - 39 = 59$$

2. Select a number of classes (m = 7) and a class width (c = 10) so that the product (mc = 70) is a bit larger than the range (range = 59).

3. Pick a starting point. This starting point should be a little smaller than the lowest score, L.

Example 6 – *Grouping Data to Form a Frequency Distribution*

cont’d

Suppose we start at 35; counting from there by tens (the class width), we get 35, 45, 55, 65, . . ., 95, 105. These are called the **class boundaries**.

The classes for the data in Table 2.6 are:

35 or more to less than 45 → $35 \le x < 45$

45 or more to less than 55 → $45 \le x < 55$

55 or more to less than 65 → $55 \le x < 65$

65 or more to less than 75 → $65 \le x < 75$

⋮ $75 \le x < 85$

$85 \le x < 95$

95 or more to and including 105 → $95 \le x \le 105$

Example 6 – *Grouping Data to Form a Frequency Distribution*

cont'd

Notes

1. At a glance you can check the number pattern to determine whether the arithmetic used to form the classes was correct (35, 45, 55, . . . , 105).

2. For the interval $35 \leq x < 45$, 35 is the lower class boundary and 45 is the upper class boundary. Observations that fall on the lower class boundary stay in that interval; observations that fall on the upper class boundary go into the next higher interval, except for the last class.

Example 6 – *Grouping Data to Form a Frequency Distribution*

cont'd

3. The class width is the difference between the upper and lower class boundaries.

4. Many combinations of class widths, numbers of classes, and starting points are possible when classifying data. There is no one best choice. Try a few different combinations, and use good judgment to decide on the one to use.

Frequency Distributions and Histograms

Therefore, the following **basic guidelines** are used in constructing a grouped frequency distribution:

1. Each class should be of the same width.
2. Classes (sometimes called *bins*) should be set up so that they do not overlap and so that each data value belongs to exactly one class.
3. Use a system that takes advantage of a number pattern to guarantee accuracy.
4. When it is convenient, an even class width is often advantageous.

Frequency Distributions and Histograms

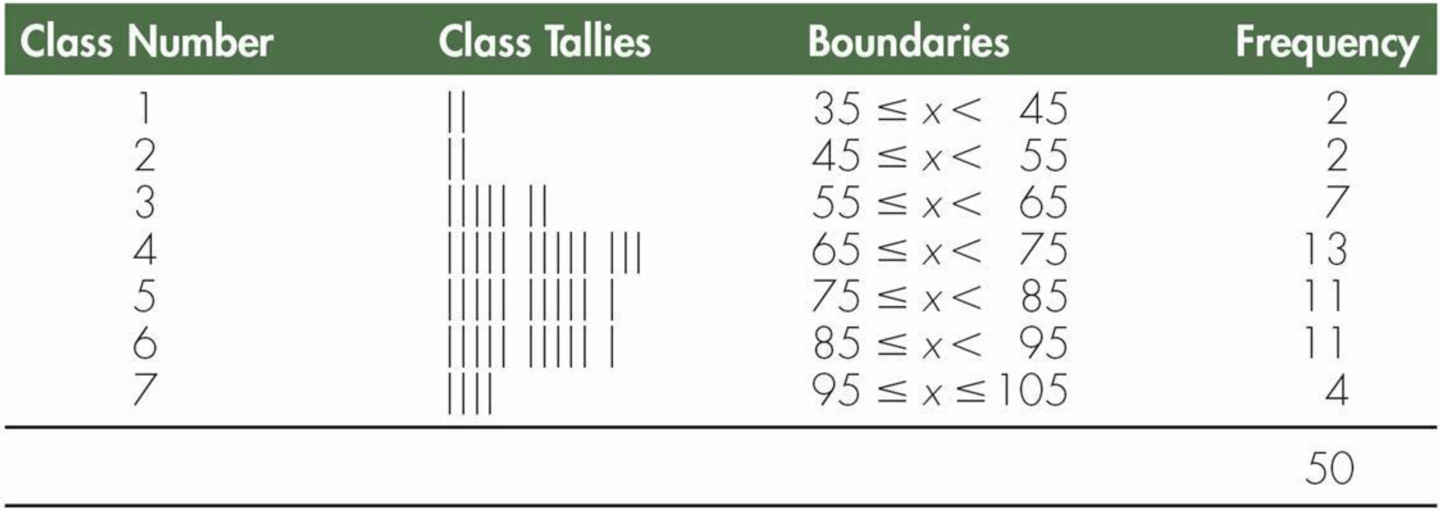
Once the classes are set up, we need to sort the data into those classes.

The method used to sort will depend on the current format of the data: If the data are ranked, the frequencies can be counted; if the data are not ranked, we will **tally** the data to find the frequency numbers.

When classifying data, it helps to use a standard chart.

Frequency Distributions and Histograms

See Table 2.7



Class Number	Class Tallies	Boundaries	Frequency
1	||	$35 \le x < 45$	2
2	||	$45 \le x < 55$	2
3	||||| ||	$55 \le x < 65$	7
4	||||| ||||| |||	$65 \le x < 75$	13
5	||||| ||||| |	$75 \le x < 85$	11
6	||||| ||||| |	$85 \le x < 95$	11
7	||||	$95 \le x \le 105$	4
			50

Standard Chart for Frequency Distribution

Table 2.7

Frequency Distributions and Histograms

Notes

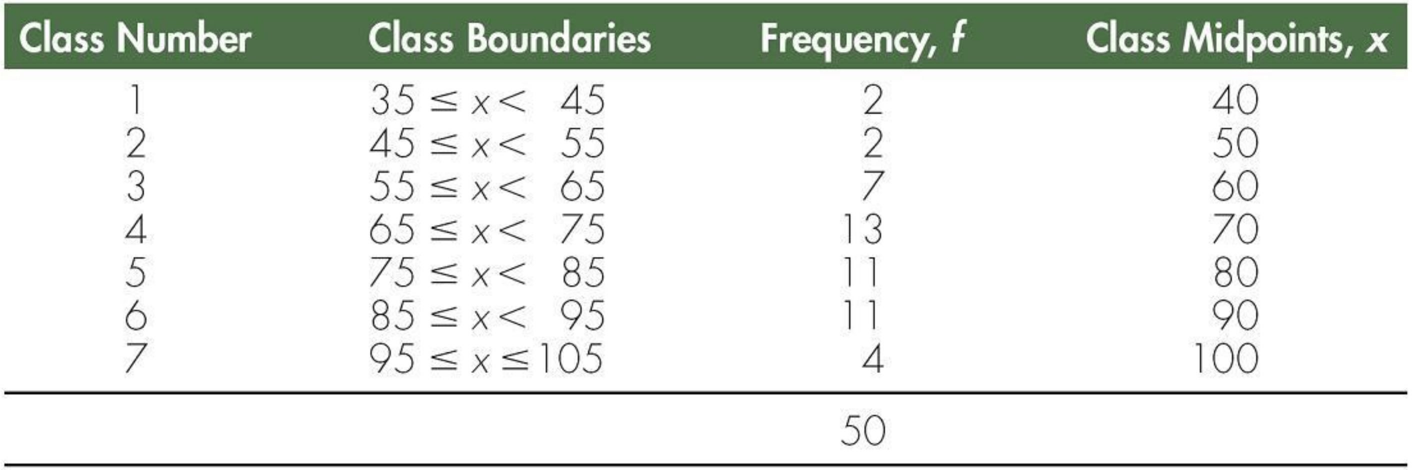
1. If the data have been ranked (list form, dotplot, or stem-and-leaf), tallying is unnecessary; just count the data that belong to each class.

2. If the data are not ranked, be careful as you tally.

3. The frequency, *f*, for each class is the number of pieces of data that belong in that class.

4. The sum of the frequencies should equal the number of pieces of data, $n(n = \Sigma f)$. This summation serves as a good check.

Frequency Distributions and Histograms



Class Number	Class Boundaries	Frequency, f	Class Midpoints, x
1	$35 \leq x < 45$	2	40
2	$45 \leq x < 55$	2	50
3	$55 \leq x < 65$	7	60
4	$65 \leq x < 75$	13	70
5	$75 \leq x < 85$	11	80
6	$85 \leq x < 95$	11	90
7	$95 \leq x \leq 105$	4	100
		50	

Frequency Distribution with Class Midpoints

Table 2.8

Note

Now you can see why it is helpful to have an even class width. An odd class width would have resulted in a class midpoint with an extra digit. (For example, the class 45–54 is 9 wide and the class midpoint is 49.5.)

Frequency Distributions and Histograms

Each class needs a single numerical value to represent all the data values that fall into that class.

The **class midpoint** (sometimes called the *class mark*) is the numerical value that is exactly in the middle of each class. It is found by adding the class boundaries and dividing by 2.

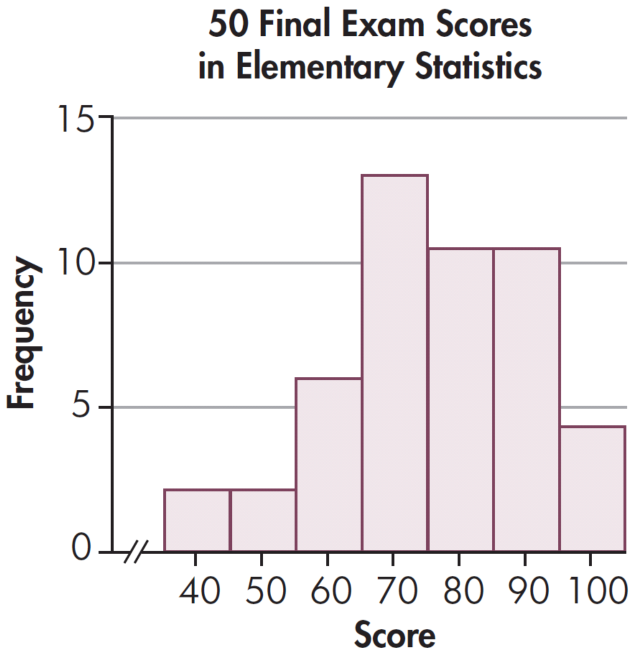
Frequency Distributions and Histograms

Histogram A bar graph that represents a frequency distribution of a quantitative variable. A histogram is made up of the following components:

1. A title, which identifies the population or sample of concern.
2. A vertical scale, which identifies the frequencies in the various classes.
3. A horizontal scale, which identifies the variable *x*. Values for the class boundaries or class midpoints may be labeled along the *x*-axis. Use whichever method of labeling the axis best presents the variable.

Frequency Distributions and Histograms

The frequency distribution from Table 2.8 appears in histogram form in Figure 2.10.



Frequency Histogram

Figure 2.10

Frequency Distributions and Histograms

Sometimes the **relative frequency** of a value is important. The relative frequency is a proportional measure of the frequency for an occurrence.

It is found by dividing the class frequency by the total number of observations.

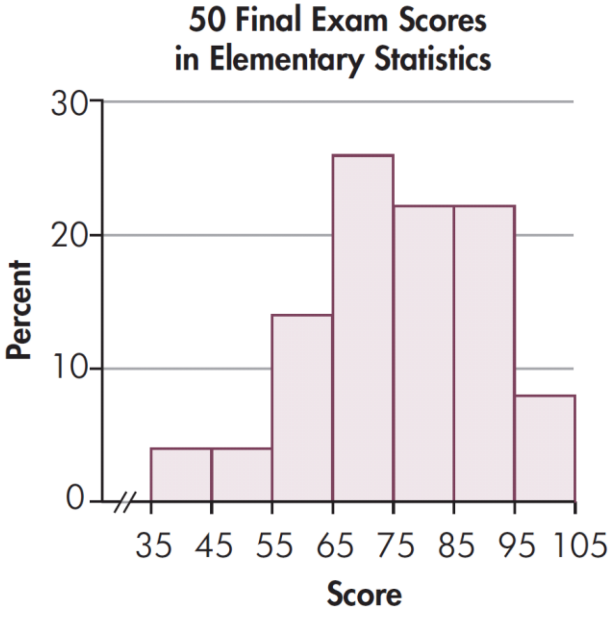
Relative frequency can be expressed as a common fraction, in decimal form, or as a percentage.

Relative frequencies are often useful in a presentation because most people understand fractional parts when expressed as percents.

Frequency Distributions and Histograms

Relative frequencies are particularly useful when comparing the frequency distributions of two different size sets of data.

Figure 2.11 is a **relative frequency histogram** of the sample of the 50 final exam scores from Table 2.8.

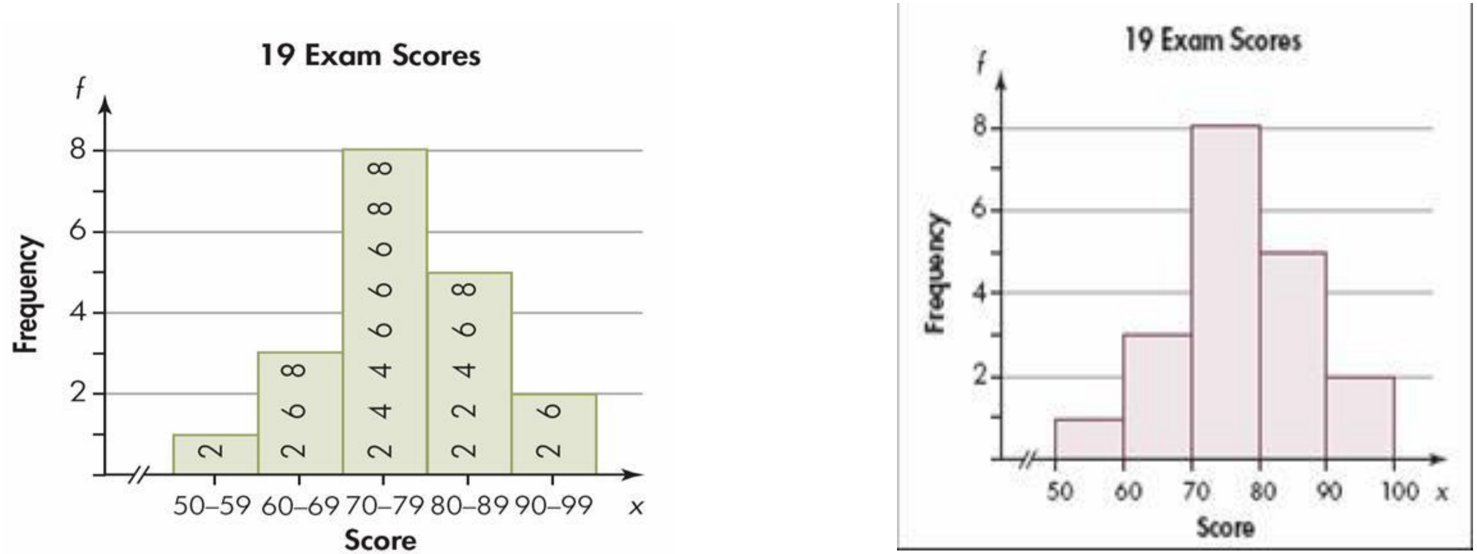


Relative Frequency Histogram

Figure 2.11

Frequency Distributions and Histograms

In Figure 2.12A the stem-and-leaf display has been rotated 90° and labels have been added to show its relationship to a histogram. Figure 2.12B shows the same set of data as a completed histogram.



(a) Modified Stem-and-Leaf Display

(b) Histogram

Figure 2.12

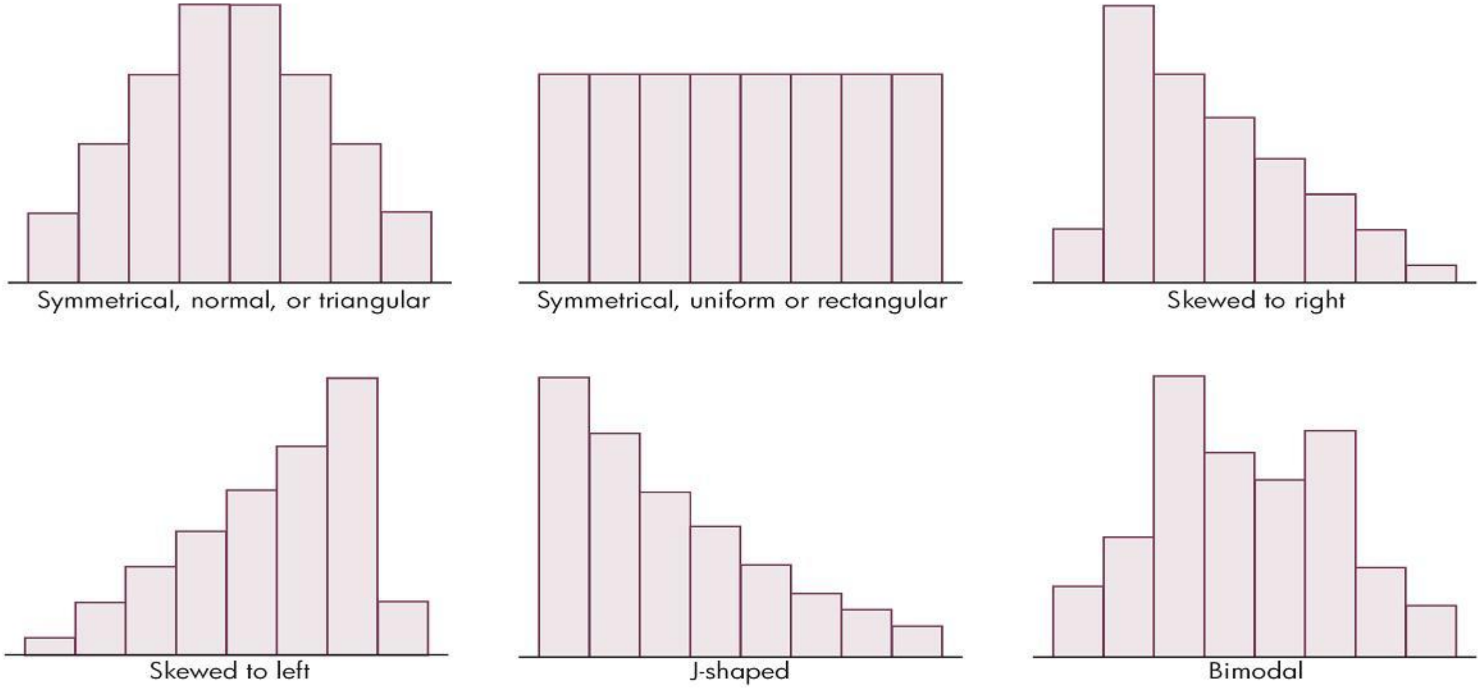
Frequency Distributions and Histograms

Histograms are valuable tools. For example, the histogram of a sample should have a distribution shape very similar to that of the population from which the sample was drawn.

If the reader of a histogram is at all familiar with the variable involved, he or she will usually be able to interpret several important facts.

Frequency Distributions and Histograms

Figure 2.13 presents histograms with specific shapes that suggest descriptive labels. Possible descriptive labels are listed under each histogram.



Symmetrical, normal, or triangular

Symmetrical, uniform or rectangular

Skewed to right

Skewed to left

J-shaped

Bimodal

Shapes of Histograms

Figure 2.13

Frequency Distributions and Histograms

Symmetrical Both sides of this distribution are identical (halves are mirror images).

Normal A symmetrical distribution is mounded up about the mean and becomes sparse at the extremes. (Additional properties are discussed later.)

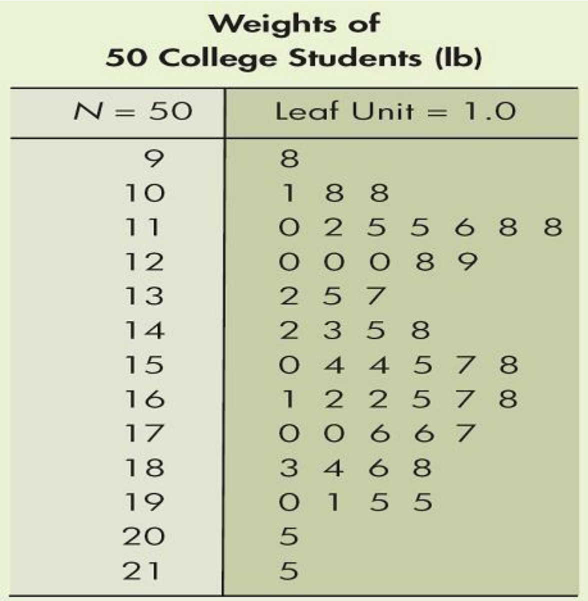
Uniform (rectangular) Every value appears with equal frequency.

Skewed One tail is stretched out longer than the other. The direction of skewness is on the side of the longer tail.

J-shaped There is no tail on the side of the class with the highest frequency.

Frequency Distributions and Histograms

Bimodal The two most populous classes are separated by one or more classes. This situation often implies that two populations are being sampled. (See Figure 2.7)



Weights of 50 College Students (lb)

N = 50	Leaf Unit = 1.0
9	8
10	1 8 8
11	0 2 5 5 6 8 8
12	0 0 0 8 9
13	2 5 7
14	2 3 5 8
15	0 4 4 5 7 8
16	1 2 2 5 7 8
17	0 0 6 6 7
18	3 4 6 8
19	0 1 5 5
20	5
21	5

Stem-and-Leaf Display

Figure 2.7

Frequency Distributions and Histograms

Notes

1. The **mode** is the value of the data that occurs with the greatest frequency.

2. The **modal class** is the class with the highest frequency.

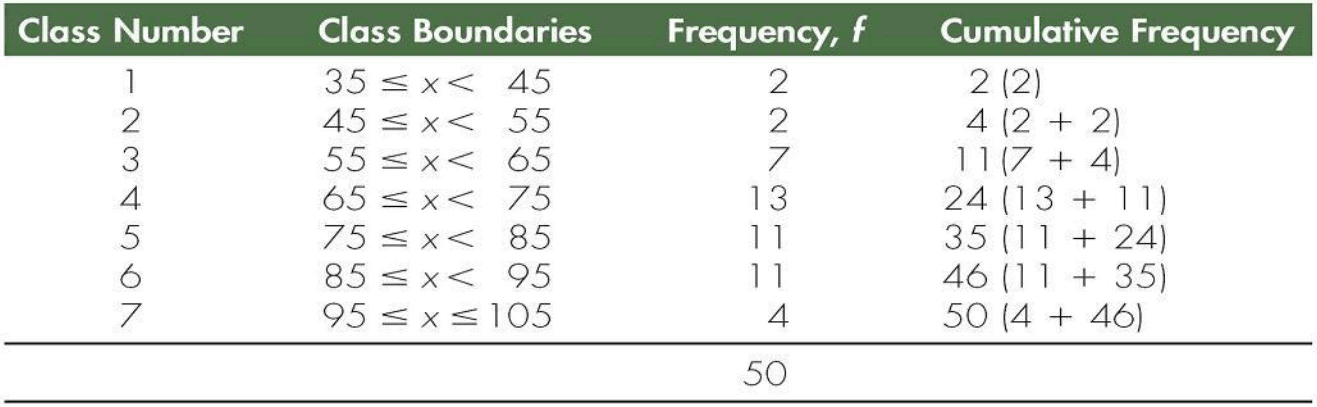
3. A **bimodal distribution** has two high-frequency classes separated by classes with lower frequencies. It is not necessary for the two high frequencies to be the same.

Another way to express a frequency distribution is to use a *cumulative frequency distribution*.

Frequency Distributions and Histograms

Cumulative Frequency Distribution A frequency distribution that pairs cumulative frequencies with values of the variable.

The **cumulative frequency** for any given class is the sum of the frequency for that class and the frequencies of all classes of smaller values.



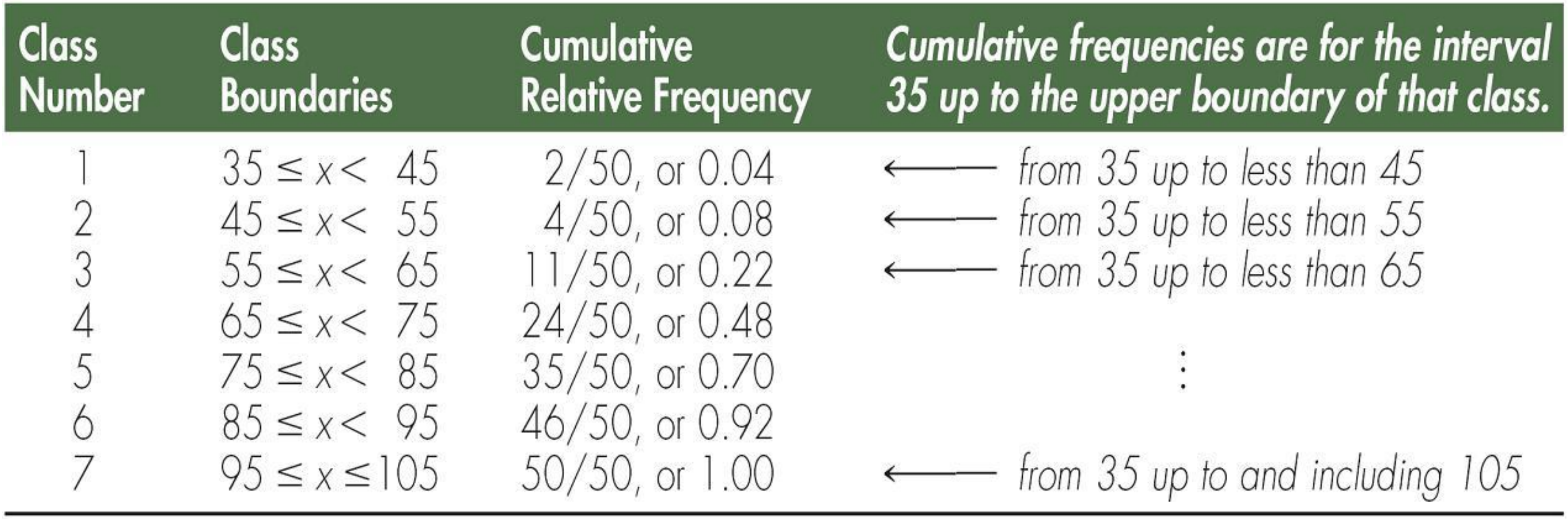
Class Number	Class Boundaries	Frequency, f	Cumulative Frequency
1	$35 \le x < 45$	2	2 (2)
2	$45 \le x < 55$	2	4 (2 + 2)
3	$55 \le x < 65$	7	11 (7 + 4)
4	$65 \le x < 75$	13	24 (13 + 11)
5	$75 \le x < 85$	11	35 (11 + 24)
6	$85 \le x < 95$	11	46 (11 + 35)
7	$95 \le x \le 105$	4	50 (4 + 46)
		50	

Using Frequency Distribution to Form a Cumulative Frequency Distribution

Table 2.9

Frequency Distributions and Histograms

The same information can be presented by using a *cumulative relative frequency distribution* (see Table 2.10). This combines the cumulative frequency and the relative frequency ideas.



Class Number	Class Boundaries	Cumulative Relative Frequency	Cumulative frequencies are for the interval 35 up to the upper boundary of that class.
1	$35 \le x < 45$	2/50, or 0.04	← from 35 up to less than 45
2	$45 \le x < 55$	4/50, or 0.08	← from 35 up to less than 55
3	$55 \le x < 65$	11/50, or 0.22	← from 35 up to less than 65
4	$65 \le x < 75$	24/50, or 0.48	
5	$75 \le x < 85$	35/50, or 0.70	⋮
6	$85 \le x < 95$	46/50, or 0.92	
7	$95 \le x \le 105$	50/50, or 1.00	← from 35 up to and including 105

Cumulative Relative Frequency Distribution

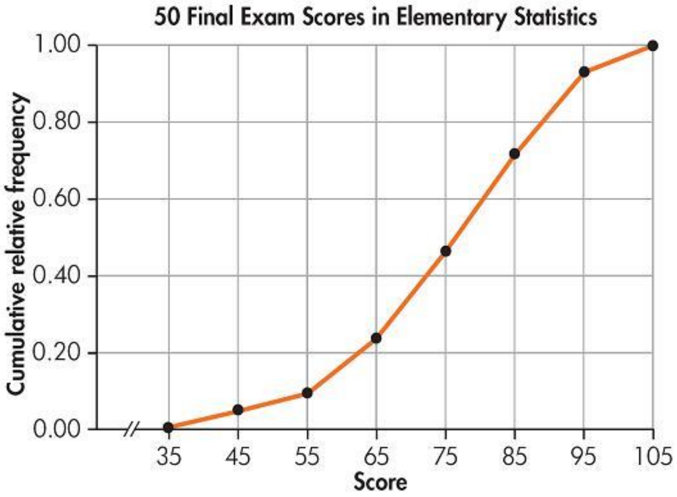
Table 2.10

Frequency Distributions and Histograms

Ogive A line graph of a cumulative frequency or cumulative relative frequency distribution. An ogive has the following components:

1. A title, which identifies the population or sample.

2. A vertical scale, which identifies either the cumulative frequencies or the cumulative relative frequencies. (Figure 2.14 shows an ogive with cumulative relative frequencies.)



Ogive

Figure 2.14

Frequency Distributions and Histograms

3. A horizontal scale, which identifies the upper class boundaries. (Until the upper boundary of a class has been reached, you cannot be sure you have accumulated all the data in that class. Therefore, the horizontal scale for an ogive is always based on the upper class boundaries.)

The ogive can be used to make percentage statements about numerical data much like a Pareto diagram does for attribute data.

Frequency Distributions and Histograms

For example, suppose we want to know what percent of the final exam scores were not passing if scores of 65 or greater are considered passing.

Following vertically from 65 on the horizontal scale to the ogive line and reading from the vertical scale, we could say that approximately 22% of the final exam scores were not passing grades.