1 Overview and Descriptive Statistics

1.2 Pictorial and Tabular Methods in Descriptive Statistics

Pictorial and Tabular Methods in Descriptive Statistics

Descriptive statistics can be divided into two general subject areas. In this section, we consider representing a data set using **visual techniques**.

Many visual techniques may already be familiar to you: **frequency tables, tally sheets, histograms, pie charts, bar graphs, scatter diagrams**, and the like. Here we focus on a selected few of these techniques that are most useful and relevant to probability and inferential statistics.

Notation

Notation

Some general notation will make it easier to apply our methods and formulas to a wide variety of practical problems.

The number of observations in a single sample, that is, the ***sample size***, will often be **denoted by** n, so that $n = 4$ for the sample of universities {Stanford, Iowa State, Wyoming, Rochester} and also for the sample of pH measurements {6.3, 6.2, 5.9, 6.5}.

If two samples are simultaneously under consideration, either m and n or n_1 and n_2 can be used to denote the numbers of observations.

Notation

Thus if {29.7, 31.6, 30.9} and {28.7, 29.5, 29.4, 30.3} are thermal-efficiency measurements for two different types of diesel engines, then $m = 3$ and $n = 4$.

Given a data set consisting of n observations on some variable x, the individual observations will be denoted by $x_1, x_2, x_3, \ldots, x_n$. The subscript bears no relation to the magnitude of a particular observation.

Thus x_1 will not in general be the smallest observation in the set, nor will x_n typically be the largest.

Notation

In many applications, x_1 will be the first observation gathered by the experimenter, x_2 the second, and so on. The *i*th observation in the data set will be denoted by x_i.

Stem-and-Leaf Displays

Stem-and-Leaf Displays

Consider a numerical data set $x_1, x_2, x_3, \ldots, x_n$ for which each x_i consists of at least two digits. A quick way to obtain an informative visual representation of the data set is to construct a *stem-and-leaf display.*

Constructing a Stem-and-Leaf Display

1. Select one or more leading digits for the stem values. The trailing digits become the leaves.
2. List possible stem values in a vertical column.
3. Record the leaf for each observation beside the corresponding stem value.
4. Indicate the units for stems and leaves someplace in the display.

Stem-and-Leaf Displays

If the data set consists of exam scores, each between 0 and 100, the score of 83 would have a stem of 8 and a leaf of 3.

For a data set of automobile fuel efficiencies (mpg), all between 8.1 and 47.8, we could use the tens digit as the stem, so 32.6 would then have a leaf of 2.6.

In general, a display based on between 5 and 20 stems is recommended.

Example

A common complaint among college students is that they are getting less sleep than they need. The article "Class Start Times, Sleep, and Academic Performance in College: A Path Analysis" (Chronobiology Intl., 2012: 318–335) investigated factors that impact sleep time. The stem-and-leaf display in Figure 1.4 shows the average number of hours of sleep per day over a two-week period for a sample of 253 students.

Example

cont'd

5L	00
5H	6889
6L	000111123444444
6H	55556778899999
7L	00001111111222222333333344444444
7H	55555555666666666667777778888888899999999999999
8L	000000000000111111222222222222222333333333344444444444
8H	5555555566666666677777788888888999999999
9L	00001111111222223334
9H	666678999
10L	00
10H	56

Stem: ones digit
Leaf: tenths digit

Example

cont'd

The first observation in the top row of the display is 5.0, corresponding to a stem of 5 and leaf of 0, and the last observation at the bottom of the display is 10.6.

The display suggests that a typical or representative sleep time is in the stem 8L row, perhaps 8.1 or 8.2.

The general shape of the display is rather symmetric.

Outliers ?

Example

stem-and-leaf displays for a random sample of lengths of golf courses (yards) that have been designated by Golf Magazine as among the most challenging in the United States. Among the sample of 40 courses, the shortest is 6433 yards long, and the longest is 7280 yards. The lengths appear to be distributed in a roughly uniform fashion over the range of values in the sample.

Example

cont'd

64	35	64	33	70	Stem: Thousands and hundreds digits		
65	26	27	06	83	Leaf: Tens and ones digits		
66	05	94	14				
67	90	70	00	98	70	45	13
68	90	70	73	50			
69	00	27	36	04			
70	51	05	11	40	50	22	
71	31	69	68	05	13	65	
72	80	09					

Example

cont'd

```
Stem-and-leaf of yardage  N = 40
Leaf Unit = 10
           4          64   3367
           8          65   0228
          11          66   019
          18          67   0147799
         (4)          68   5779
          18          69   0023
          14          70   012455
           8          71   013666
           2          72   08
```

Display from Minitab with truncated one-digit leaves

Stem-and-Leaf Displays

A stem-and-leaf display conveys information about the following aspects of the data:

- identification of a typical or representative value
- extent of spread about the typical value
- presence of any gaps in the data
- extent of symmetry in the distribution of values
- number and location of peaks
- presence of any outlying values

Dotplots

Dotplots

A dotplot is an attractive summary of numerical data when the data set is reasonably small or there are relatively few distinct data values. Each observation is represented by a dot above the corresponding location on a horizontal measurement scale.

When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically. As with a stem-and-leaf display, a dotplot gives information about location, spread, extremes, and gaps.

Example

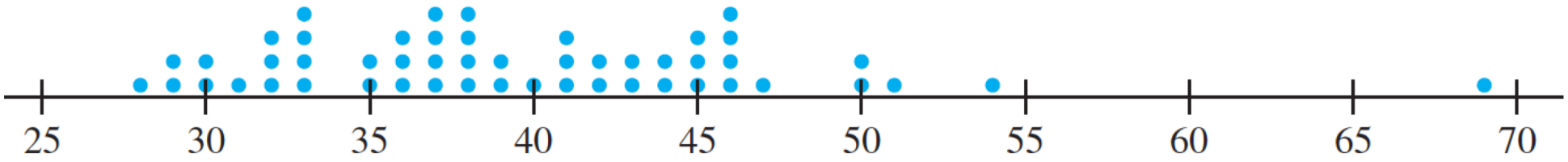
There is growing concern in the U.S. that not enough students are graduating from college. Here is data on the percentage of 25- to 34-year-olds in each state who had some type of postsecondary degree as of 2010 (listed in alphabetical order, with the District of Columbia included):

31.5	32.9	33.0	28.6	37.9	43.3	45.9	37.2	68.8	36.2	35.5
40.5	37.2	45.3	36.1	45.5	42.3	33.3	30.3	37.2	45.5	54.3
37.2	49.8	32.1	39.3	40.3	44.2	28.4	46.0	47.2	28.7	49.6
37.6	50.8	38.0	30.8	37.6	43.9	42.5	35.2	42.2	32.8	32.2
38.5	44.5	44.6	40.9	29.5	41.3	35.4				

Example

cont'd

The following shows a dotplot of the data. The most striking feature is the substantial state-to-state variability.



The largest value, for D.C., is obviously an extreme Outlier.

Histograms

Histograms

Some numerical data is obtained by counting to determine the value of a variable (the number of traffic citations a person received during the last year, the number of customers arriving for service during a particular period), whereas other data is obtained by taking measurements (weight of an individual, reaction time to a particular stimulus).

The prescription for drawing a histogram is generally different for these two cases.

Histograms

Definition

A numerical variable is **discrete** if its set of possible values either is finite or else can be listed in an infinite sequence (one in which there is a first number, a second number, and so on). A numerical variable is **continuous** if its possible values consist of an entire interval on the number line.

A discrete variable *x* almost always results from counting, in which case possible values are 0, 1, 2, 3, . . . or some subset of these integers. **Continuous variables arise from making measurements**. For example, if *x* is the pH of a chemical substance, then in theory *x* could be any number between 0 and 14: 7.0, 7.03, 7.032, and so on.

Histograms

Of course, in practice there are limitations on the degree of accuracy of any measuring instrument, so we may not be able to determine pH, reaction time, height, and concentration to an arbitrarily large number of decimal places.

However, from the point of view of creating mathematical models for distributions of data, it is helpful to imagine an entire continuum of possible values.

Consider data consisting of observations on a **discrete variable *x***. The **frequency** of any particular *x* value is the number of times that value occurs in the data set.

Histograms

The **relative frequency** of a value is the fraction or proportion of times the value occurs:

$$\text{relative frequency of a value} = \frac{\text{number of times the value occurs}}{\text{number of observations in the data set}}$$

Suppose, for example, that our data set consists of 200 observations on x = the number of courses a college student is taking this term. If 70 of these x values are 3, then

frequency of the x value 3: 70

Relative frequency of the x value 3: $\frac{70}{200} = .35$

Histograms

Multiplying a relative frequency by 100 gives a percentage; in the college-course example, 35% of the students in the sample are taking three courses.

The relative frequencies, or percentages, are usually of more interest than the frequencies themselves. In theory, the relative frequencies should sum to 1, but in practice the sum may differ slightly from 1 because of rounding.

A **frequency distribution** is a tabulation of the frequencies and/or relative frequencies.

Histograms

Constructing a Histogram for Discrete Data

First, determine the frequency and relative frequency of each x value. Then mark possible x values on a horizontal scale. Above each value, draw a rectangle whose height is the relative frequency (or alternatively, the frequency) of that value.

This construction ensures that the *area* of each rectangle is proportional to the relative frequency of the value. Thus if the relative frequencies of $x = 1$ and $x = 5$ are .35 and .07, respectively, then the area of the rectangle above 1 is five times the area of the rectangle above 5.

Example

Frequency Distribution for Hits in Nine-Inning Games

How unusual is a no-hitter or a one-hitter in a major league baseball game, and how frequently does a team get more than 10, 15, or even 20 hits?

Example

cont'd

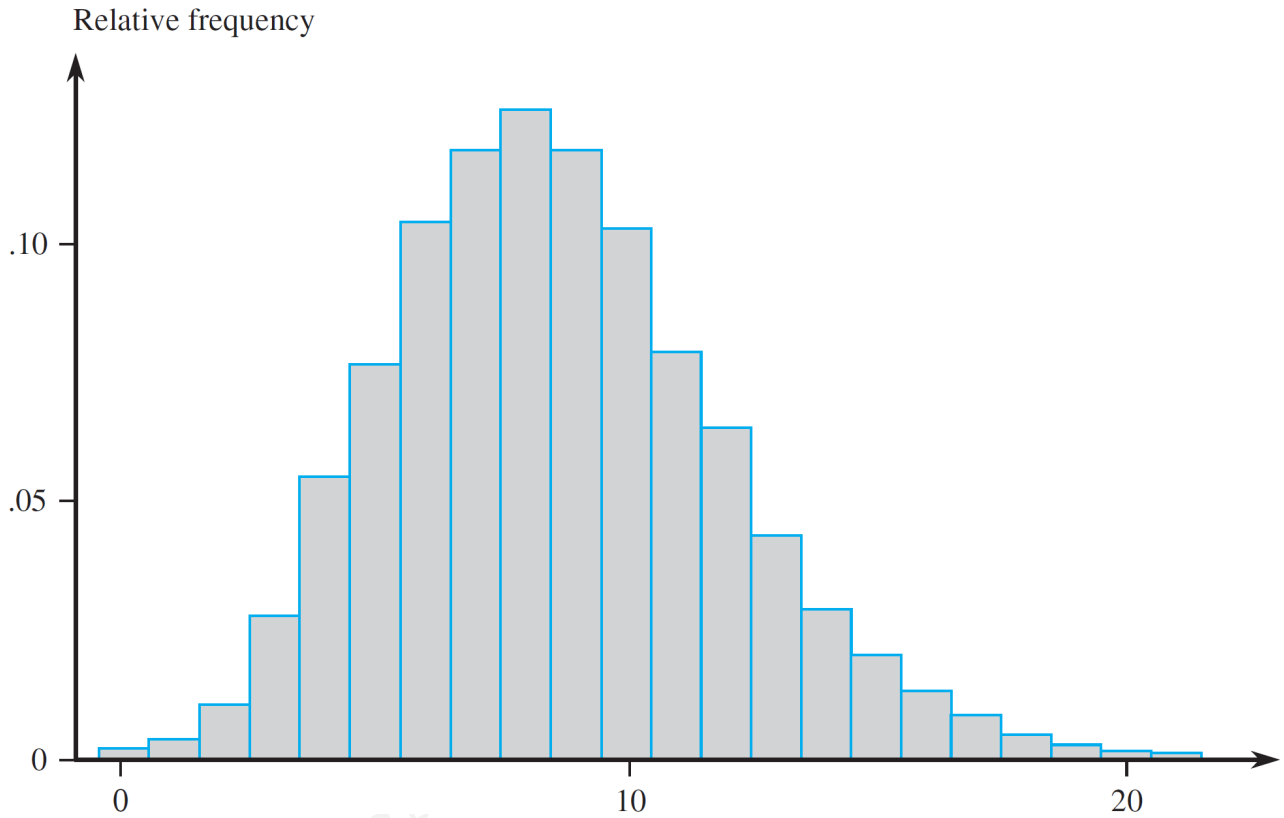
Frequency distribution for the number of hits per team per game for all nine-inning games played between 1989 and 1993.

Hits/Game	Number of Games	Relative Frequency	Hits/Game	Number of Games	Relative Frequency
0	20	.0010	14	569	.0294
1	72	.0037	15	393	.0203
2	209	.0108	16	253	.0131
3	527	.0272	17	171	.0088
4	1048	.0541	18	97	.0050
5	1457	.0752	19	53	.0027
6	1988	.1026	20	31	.0016
7	2256	.1164	21	19	.0010
8	2403	.1240	22	13	.0007
9	2256	.1164	23	5	.0003
10	1967	.1015	24	1	.0001
11	1509	.0779	25	0	.0000
12	1230	.0635	26	1	.0001
13	834	.0430	27	1	.0001
				19,383	1.0005

Example

cont'd

The corresponding histogram rises rather smoothly to a single peak and then declines. The histogram extends a bit more on the right (toward large values) than it does on the left—a slight "positive skew."



Histogram of number of hits per nine-inning game

Example

cont'd

Either from the tabulated information or from the histogram itself, we can determine the following:

$$\begin{aligned}\text{proportion of games with at most two hits} &= \text{relative frequency for } x = 0 + \text{relative frequency for } x = 1 + \text{relative frequency for } x = 2 \\ &= .0010 + .0037 + .0108 \\ &= .0155\end{aligned}$$

Example

cont'd

Similarly,

$$\text{proportion of games with between 5 and 10 hits (inclusive)} = .0752 + .1026 + \cdots + .1015$$

$$= .6361$$

That is, roughly 64% of all these games resulted in between 5 and 10 (inclusive) hits.

Histograms

Constructing a Histogram for Continuous Data: Equal Class Widths

Determine the frequency and relative frequency for each class. Mark the class boundaries on a horizontal measurement axis. Above each class interval, draw a rectangle whose height is the corresponding relative frequency (or frequency).

Constructing a Histogram for Continuous Data: Unequal Class Widths

After determining frequencies and relative frequencies, calculate the height of each rectangle using the formula

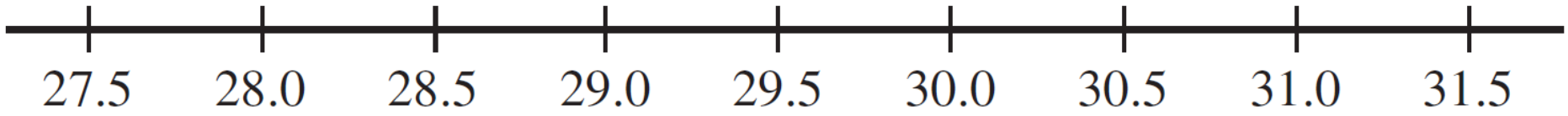
$$\text{rectangle height} = \frac{\text{relative frequency of the class}}{\text{class width}}$$

The resulting rectangle heights are usually called *densities*, and the vertical scale is the **density scale.** This prescription will also work when class widths are equal.

Histograms

Constructing a histogram for continuous data (measurements) entails subdividing the measurement axis into a suitable number of **class intervals** or **classes,** such that each observation is contained in exactly one class.

Suppose, for example, that we have 50 observations on x = fuel efficiency of an automobile (mpg), the smallest of which is 27.8 and the largest of which is 31.4. Then we could use the class boundaries 27.5, 28.0, 28.5, . . . , and 31.5 as shown here:



Histograms

One potential difficulty is that occasionally **an observation lies on a class boundary** so therefore does not fall in exactly one interval, for example, 29.0.

One way to deal with this problem is to use boundaries like 27.55, 28.05, . . . , 31.55. Adding a hundredths digit to the class boundaries prevents observations from falling on the resulting boundaries.

Another approach is to use the classes 27.5 – < 28.0, 28.02 – < 28.5, . . . , 31.0 – < 31.5. Then 29.0 falls in the class 29.0 – < 29.5 rather than in the class 28.5 – < 29.0.

Histograms, continuous data

Power companies need information about customer usage to obtain accurate forecasts of demands. Investigators from Wisconsin Power and Light determined energy consumption (BTUs) during a particular period for a sample of 90 gas-heated homes. An adjusted consumption value was calculated as follows:

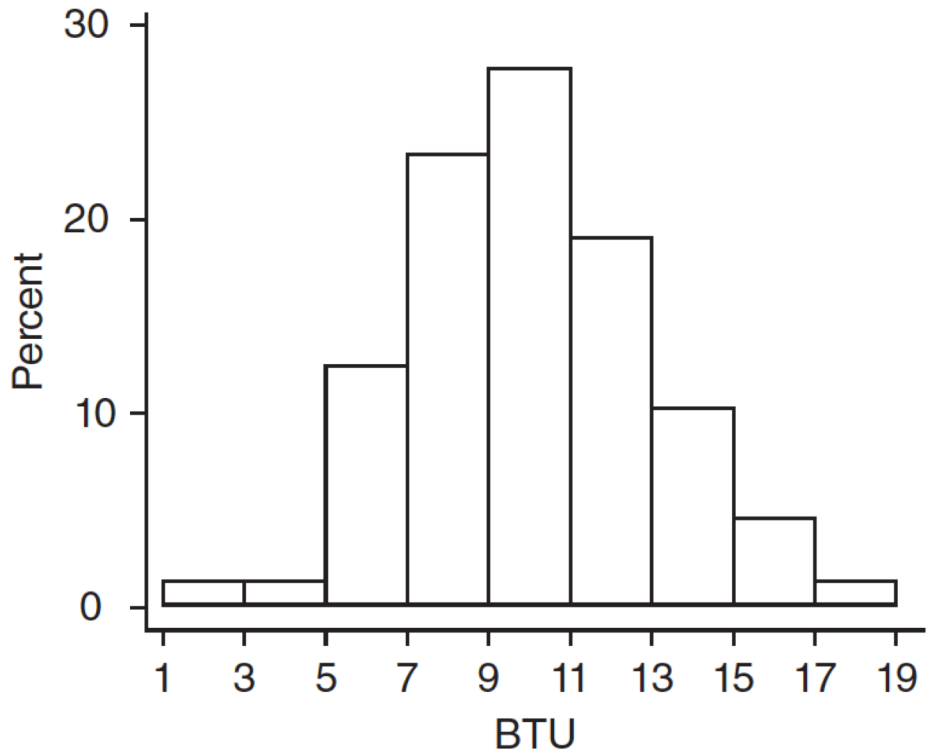
$$\text{adjusted consumption} = \frac{\text{consumption}}{(\text{weather, in degree days})(\text{house area})}$$

This resulted in the accompanying data (part of the stored data set FURNACE.MTW available in Minitab), which we have ordered from smallest to largest.

2.97	4.00	5.20	5.56	5.94	5.98	6.35	6.62	6.72	6.78
6.80	6.85	6.94	7.15	7.16	7.23	7.29	7.62	7.62	7.69
7.73	7.87	7.93	8.00	8.26	8.29	8.37	8.47	8.54	8.58
8.61	8.67	8.69	8.81	9.07	9.27	9.37	9.43	9.52	9.58
9.60	9.76	9.82	9.83	9.83	9.84	9.96	10.04	10.21	10.28
10.28	10.30	10.35	10.36	10.40	10.49	10.50	10.64	10.95	11.09
11.12	11.21	11.29	11.43	11.62	11.70	11.70	12.16	12.19	12.28
12.31	12.62	12.69	12.71	12.91	12.92	13.11	13.38	13.42	13.43
13.47	13.60	13.96	14.24	14.35	15.12	15.24	16.06	16.90	18.26

Histograms, continuous data

Class	1−<3	3−<5	5−<7	7−<9	9−<11	11−<13	13−<15	15−<17	17−<19
Frequency	1	1	11	21	25	17	9	4	1
Relative frequency	.011	.011	.122	.233	.278	.189	.100	.044	.011



Histograms, number of classes

There are no hard-and-fast rules concerning either the number of classes or the choice of classes themselves. Between 5 and 20 classes will be satisfactory for most data sets. Generally, the larger the number of observations in a data set, the more classes should be used. A reasonable rule of thumb is:

$$\text{number of classes} \approx \sqrt{\text{number of observations}}$$

Continuous data: unequal class Widths

Constructing a Histogram for Continuous Data: Unequal Class Widths

After determining frequencies and relative frequencies, calculate the height of each rectangle using the formula

$$\text{rectangle height} = \frac{\text{relative frequency of the class}}{\text{class width}}$$

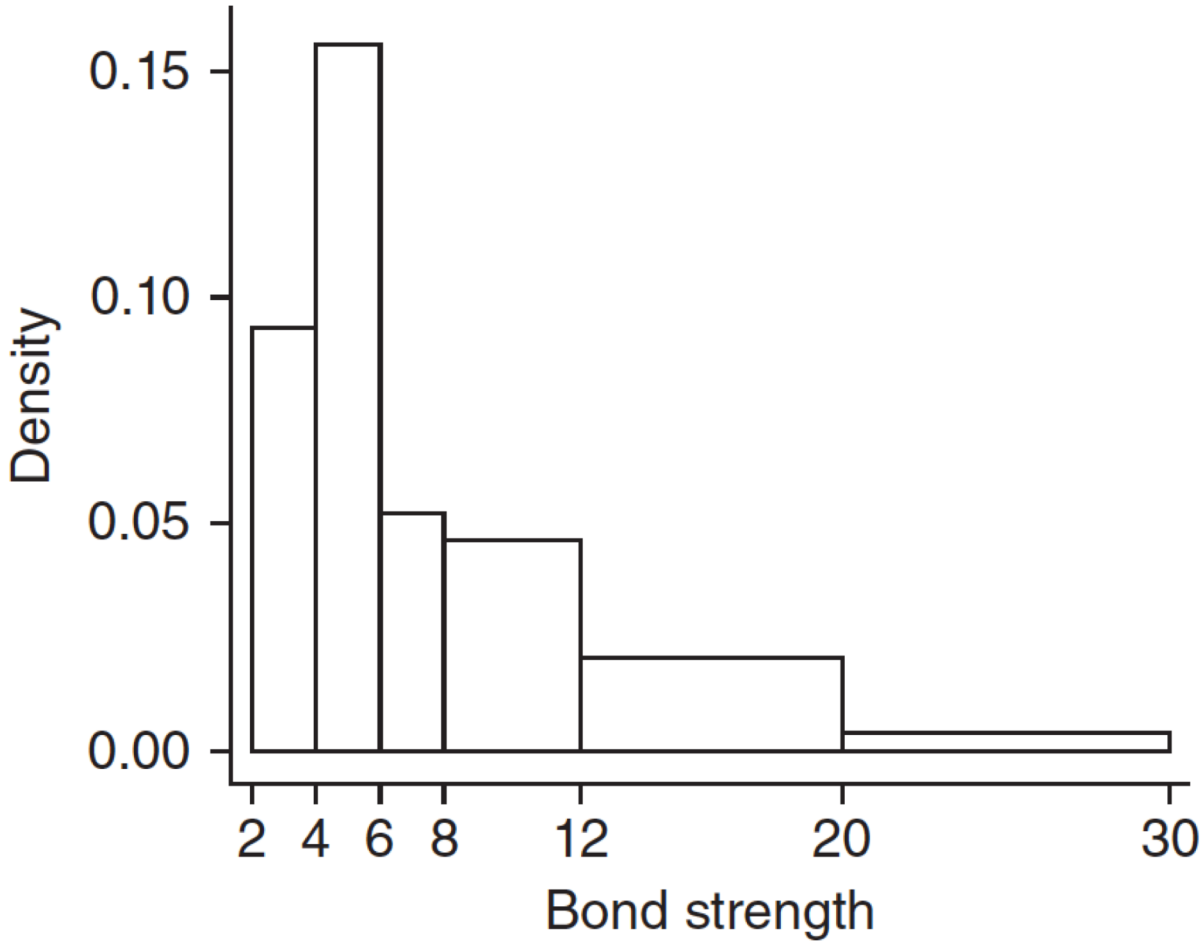
The resulting rectangle heights are usually called *densities*, and the vertical scale is the **density scale.** This prescription will also work when class widths are equal.

Example

11.5	12.1	9.9	9.3	7.8	6.2	6.6	7.0	13.4	17.1	9.3	5.6
5.7	5.4	5.2	5.1	4.9	10.7	15.2	8.5	4.2	4.0	3.9	3.8
3.6	3.4	20.6	25.5	13.8	12.6	13.1	8.9	8.2	10.7	14.2	7.6
5.2	5.5	5.1	5.0	5.2	4.8	4.1	3.8	3.7	3.6	3.6	3.6

Class	2−<4	4−<6	6−<8	8−<12	12−<20	20−<30
Frequency	9	15	5	9	8	2
Relative frequency	.1875	.3125	.1042	.1875	.1667	.0417
Density	.094	.156	.052	.047	.021	.004

Example



Histogram Shapes

Histogram Shapes

Histograms come in a variety of shapes. A **unimodal** histogram is one that rises to a single peak and then declines. A **bimodal** histogram has two different peaks.

Bimodality can occur when the data set consists of observations on two quite different kinds of individuals or objects.

For example, consider a large data set consisting of driving times for automobiles traveling between San Luis Obispo, California, and Monterey, California (exclusive of stopping time for sightseeing, eating, etc.).

Histogram Shapes

This histogram would show two peaks: one for those cars that took the inland route (roughly 2.5 hours) and another for those cars traveling up the coast (3.5–4 hours).

However, bimodality does not automatically follow in such situations. Only if the two separate histograms are "far apart" relative to their spreads will bimodality occur in the histogram of combined data.

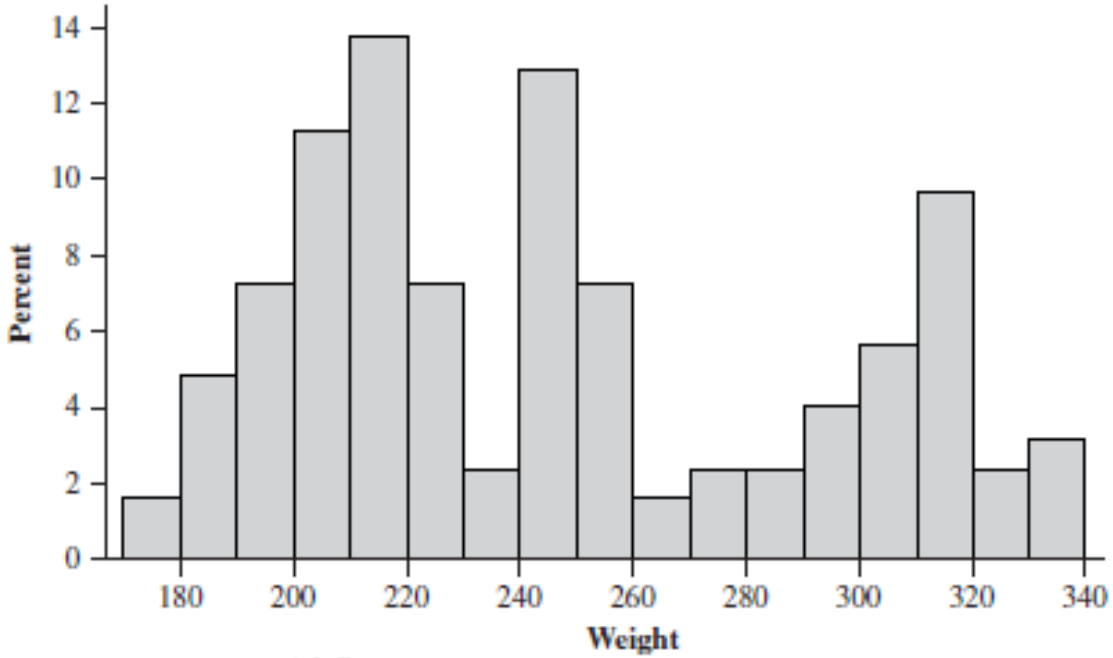
Thus a large data set consisting of heights of college students should not result in a bimodal histogram because the typical male height of about 69 inches is not far enough above the typical female height of about 64–65 inches.

Histogram Shapes

A histogram with more than two peaks is said to be **multimodal.** Of course, the number of peaks may well depend on the choice of class intervals, particularly with a small number of observations. The larger the number of classes, the more likely it is that bimodality or multimodality will manifest itself.

Example

Figure 1.11(a) shows a Minitab histogram of the weights (lb) of the 124 players listed on the rosters of the San Francisco 49ers and the New England Patriots (teams the author would like to see meet in the Super Bowl) as of Nov. 20, 2009.

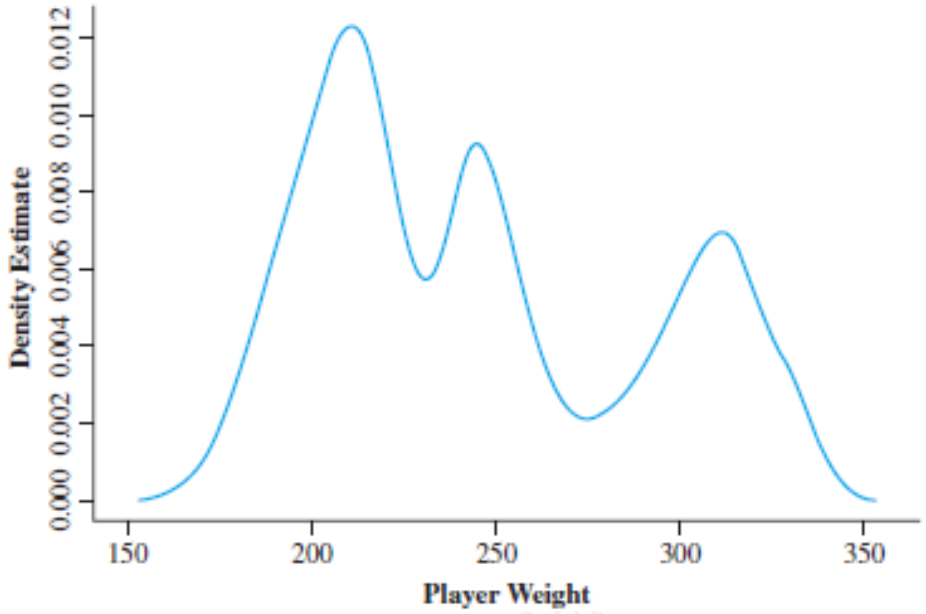


NFL player weights Histogram

Example

cont'd

Figure 1.11(b) is a smoothed histogram (actually what is called a *density estimate*) of the data from the R software package.



NFL player weights Smoothed histogram

Figure 1.11(b)

Example

cont'd

Both the histogram and the smoothed histogram show three distinct peaks; the one on the right is for linemen, the middle peak corresponds to linebacker weights, and the peak on the left is for all other players (wide receivers, quarterbacks, etc.).

A histogram is **symmetric** if the left half is a mirror image of the right half. A unimodal histogram is **positively skewed** if the right or upper tail is stretched out compared with the left or lower tail and **negatively skewed** if the stretching is to the left.

Example

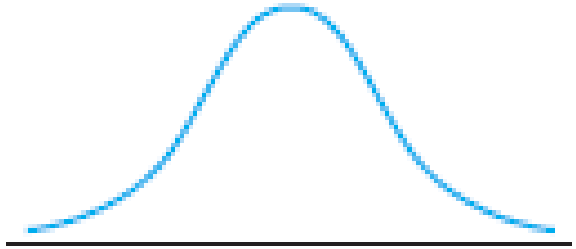
cont'd

A histogram is **symmetric** if the left half is a mirror image of the right half. A unimodal histogram is **positively skewed** if the right or upper tail is stretched out compared with the left or lower tail and **negatively skewed** if the stretching is to the left.

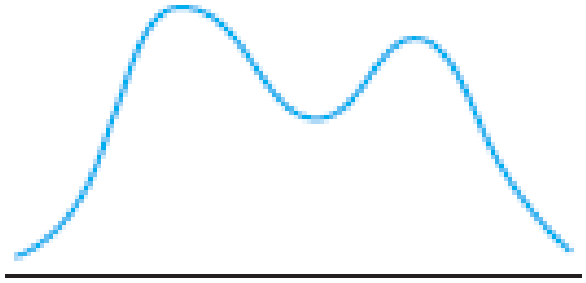
Example

cont'd

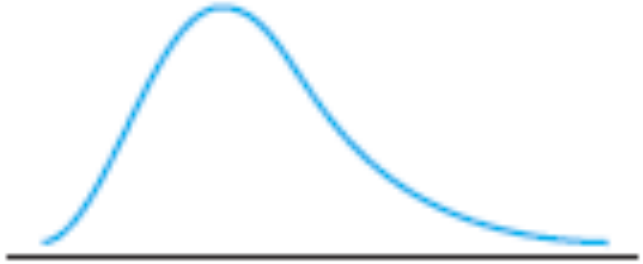
"smoothed" histograms, obtained by superimposing a smooth curve on the rectangles, that illustrate the various possibilities.



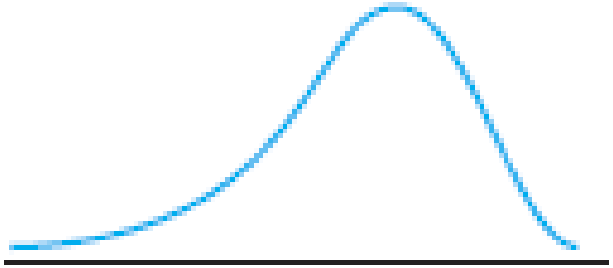
(a) symmetric unimodal



(b) bimodal



(c) Positively skewed



(d) negatively skewed

Smoothed histograms

Qualitative Data

Qualitative Data

Both a frequency distribution and a histogram can be constructed when the data set is *qualitative* (categorical) in nature.

In some cases, there will be a natural ordering of classes—for example, freshmen, sophomores, juniors, seniors, graduate students—whereas in other cases the order will be arbitrary—for example, Catholic, Jewish, Protestant, and the like.

With such categorical data, the intervals above which rectangles are constructed should have equal width.

Example

The Public Policy Institute of California carried out a telephone survey of 2501 California adult residents during April 2006 to ascertain how they felt about various aspects of K-12 public education. One question asked was "Overall, how would you rate the quality of public schools in your neighborhood today?"

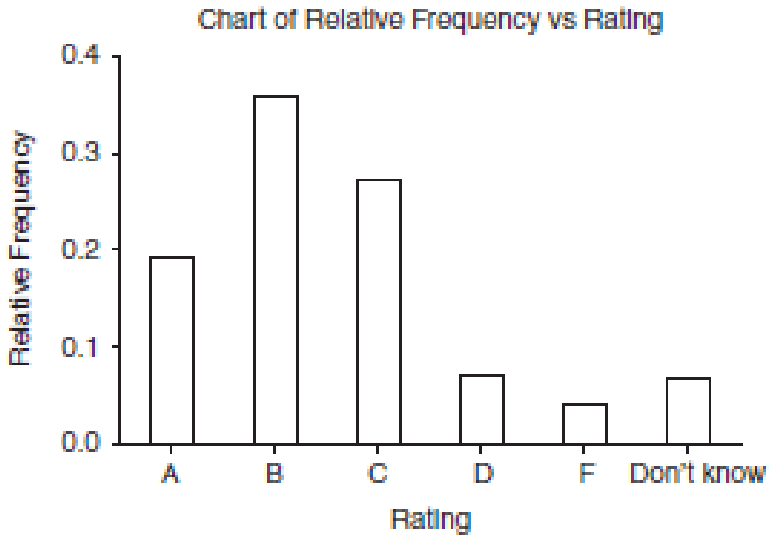
Example

cont'd

Frequencies and relative frequencies, and the corresponding histogram (bar chart).

Rating	Frequency	Relative Frequency
A	478	.191
B	893	.357
C	680	.272
D	178	.071
F	100	.040
Don't know	172	.069
	2501	1.000

Frequency Distribution for the School Rating Data



Histogram of the school rating data from Minitab