Chapter 2

Organization and Description of Data

Ram C. Kafle, Ph.D.
Assistant Professor of Statistics
Sam Houston State University

Describing the data

Describing a Data Set of Measurements

1. Summarization and description of the overall pattern.
   (a) Presentation of tables and graphs.
   (b) Noting important features of the graphed data including symmetry or departures from it.
   (c) Scanning the graphed data to detect any observations that seem to stick far out from the major mass of the data—the outliers.
2. Computation of numerical measures.
   (a) A typical or representative value that indicates the center of the data.
   (b) The amount of spread or variation present in the data.

Frequency Table

Outcome	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
2	6	6	6/20 = 0.3	0.3
3	3	6+3 = 9	3/20 = 0.15	0.3+0.15 = 0.45
4	4	9+4 = 13	4/20 = 0.2	0.45+0.2 = 0.65
5	3	13+3 = 16	3/20 = 0.15	0.65+0.15 = 0.8
6	1	16+1 = 17	1/20 = 0.05	0.8+0.05 = 0.85
7	3	17+3 = 20	3/20 = 0.15	0.85+0.15 = 1
Total	20		1	

Frequency Tables: Continuous Data

When we have a large set of quantitative data, it's useful to organize it into smaller intervals or *classes* and count how many data values fall into each class. A frequency table does just that.

A **frequency table** partitions data into classes or intervals and shows how many data values are in each class. The classes or intervals are constructed so that each data value falls into exactly one class.

Example 1 – *Frequency table*

A task force to encourage car pooling did a study of one-way commuting distances of workers in the downtown Dallas area. A random sample of 60 of these workers was taken. The commuting distances of the workers in the sample are given in Table 2-1. Make a frequency table for these data.

13	47	10	3	16	20	17	40	4	2
7	25	8	21	19	15	3	17	14	6
12	45	1	8	4	16	11	18	23	12
6	2	14	13	7	15	46	12	9	18
34	13	41	28	36	17	24	27	29	9
14	26	10	24	37	31	8	16	12	16

Example 1 – *Solution*

a. First decide how many classes you want. Five to 15 classes are usually used. If you use fewer than five classes, you risk losing too much information. If you use more than 15 classes, the data may not be sufficiently summarized.

b. Next, find the *class width* for the six classes.

HOW TO FIND THE CLASS WIDTH (INTEGER DATA)

1. Compute $\dfrac{\text{Largest data value} - \text{smallest data value}}{\text{Desired number of classes}}$
2. Increase the computed value to the next highest whole number.

Example 1 – *Solution*

cont'd

To find the class width for the commuting data, we observe that the largest distance commuted is 47 miles and the smallest is 1 mile. Using six classes, the class width is 8, since

$$\text{Class width } \frac{47 - 1}{6} \approx 7.7 \quad \text{(increase to 8)}$$

c. Now we determine the data range for each class.

The **lower class limit** is the lowest data value that can fit in a class. The **upper class limit** is the highest data value that can fit in a class. The **class width** is the difference between the *lower* class limit of one class and the *lower* class limit of the next class.

Example 1 – *Solution*

cont'd

d. There is a space between the upper limit of one class and the lower limit of the next class. The halfway points of these intervals are called *class boundaries*. These are shown in Table 2-2.

Procedure:

HOW TO FIND CLASS BOUNDARIES (INTEGER DATA)

To find **upper class boundaries,** add 0.5 unit to the upper class limits.

To find **lower class boundaries,** subtract 0.5 unit from the lower class limits.

Example 1 – *Solution*

cont'd

Table 2-2, shows the upper and lower class limits for the commuting distance data.

Class Limits Lower–Upper	Class Boundaries Lower–Upper	Tally	Frequency	Class Midpoint
1–8	0.5–8.5	𝍸 𝍸 ||||	14	4.5
9–16	8.5–16.5	𝍸 𝍸 𝍸 𝍸 |	21	12.5
17–24	16.5–24.5	𝍸 𝍸 |	11	20.5
25–32	24.5–32.5	𝍸 |	6	28.5
33–40	32.5–40.5	||||	4	36.5
41–48	40.5–48.5	||||	4	44.5

Frequency Table of One-Way Commuting Distances for 60 Downtown Dallas Workers (Data in Miles)

Table 2-2

Example 1 – *Solution*

cont'd

e. The center of each class is called the *midpoint* (or *class mark*). The midpoint is often used as a representative value of the entire class. The midpoint is found by adding the lower and upper class limits of one classand dividing by 2.

$$\text{Midpoint} = \frac{\text{Lower class limit} + \text{upper class limit}}{2}$$

Table 2-2 shows the class midpoints.

Frequency Tables

F. Relative frequency

$$\text{Relative frequency} = \frac{f}{n} = \frac{\text{Class frequency}}{\text{Total of all frequencies}}$$

Class	Frequency f	Relative Frequency f/n
1–8	14	14/60 ≈ 0.23
9–16	21	21/60 ≈ 0.35
17–24	11	11/60 ≈ 0.18
25–32	6	6/60 ≈ 0.10
33–40	4	4/60 ≈ 0.07
41–48	4	4/60 ≈ 0.07

Relative Frequencies of One-Way Commuting Distances

Table 2-3

Continuous Data Example

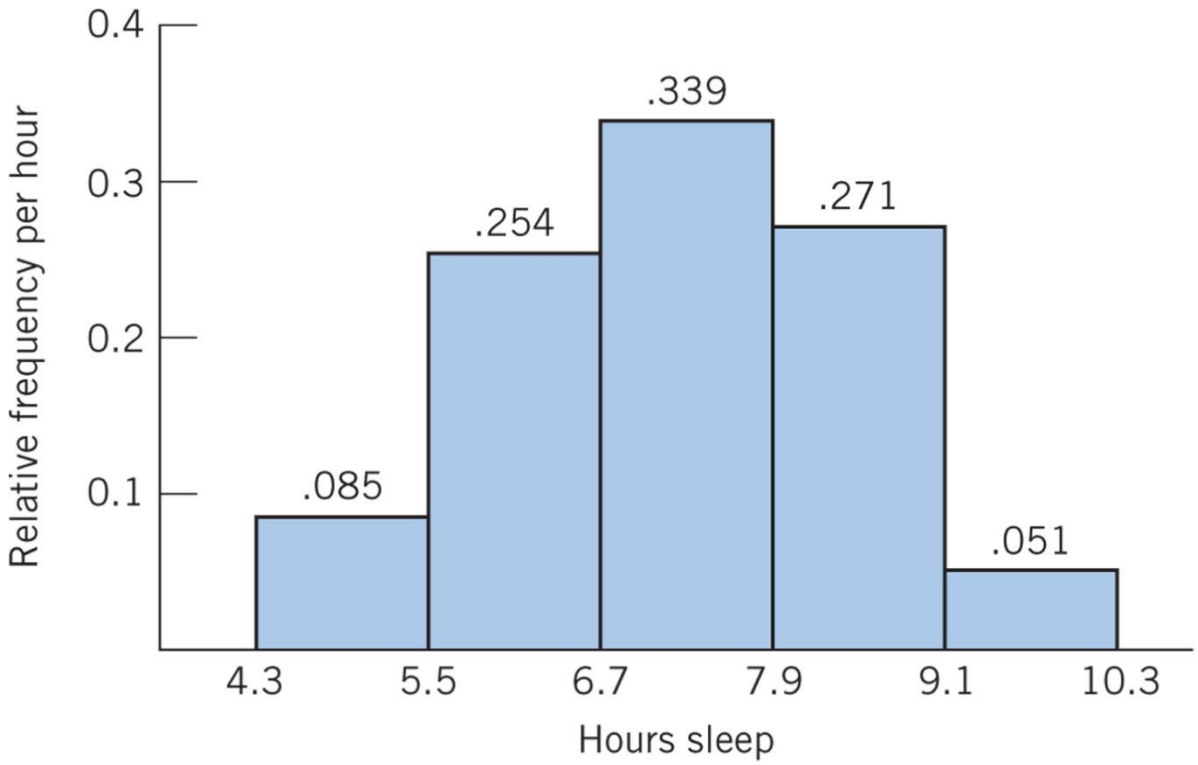
TABLE 4 Hours of Sleep for Fifty-nine Students

4.5	4.7	5.0	5.0	5.3	5.5	5.5	5.7	5.7	5.7
6.0	6.0	6.0	6.0	6.3	6.3	6.3	6.5	6.5	6.5
6.7	6.7	6.7	6.7	7.0	7.0	7.0	7.0	7.3	7.3
7.3	7.3	7.5	7.5	7.5	7.5	7.7	7.7	7.7	7.7
8.0	8.0	8.0	8.0	8.3	8.3	8.3	8.5	8.5	8.5
8.5	8.7	8.7	9.0	9.0	9.0	9.3	9.3	10.0	

TABLE 5 Frequency Distribution for Hours of Sleep Data (left endpoints included but right endpoints excluded)

Class Interval	Frequency	Relative Frequency
4.3–5.5	5	$\frac{5}{59} = .085$
5.5–6.7	15	$\frac{15}{59} = .254$
6.7–7.9	20	$\frac{20}{59} = .339$
7.9–9.1	16	$\frac{16}{59} = .271$
9.1–10.3	3	$\frac{3}{59} = .051$
Total	59	1.000

Histogram

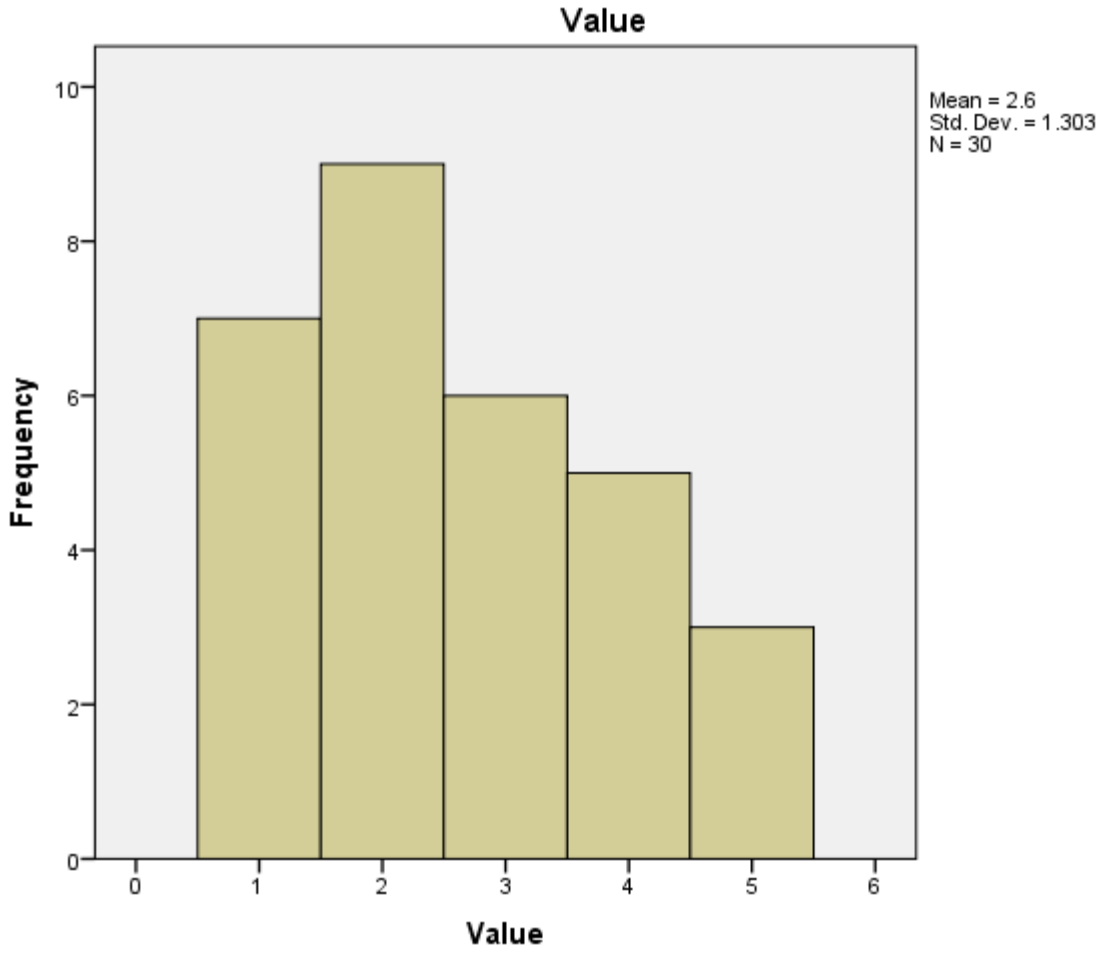

0.4
0.3
0.2
0.1
Relative frequency per hour
.085
.254
.339
.271
.051
4.3
5.5
6.7
7.9
9.1
10.3
Hours sleep

Discrete Data (Quantative)

TABLE 3 Frequency Distribution for Number (x) of Items Returned

Value x	Frequency	Relative Frequency
1	7	.233
2	9	.300
3	6	.200
4	5	.167
5	3	.100
Total	30	1.000

Histogram


Value
Mean = 2.6
Std. Dev. = 1.303
N = 30
Frequency
10
8
6
4
2
0
0
1
2
3
4
5
6
Value

Example

Class Interval	Limits		Midpoint	Boundaries	
	Lower	Upper		Lower	Upper
(14, 21)	14	21	17.5	13.5	21.5
(22, 29)	22	29	25.5	21.5	29.5
(30, 37)	30	37	33.5	29.5	37.5
(38, 45)	38	45	41.5	37.5	45.5
(46, 53)	46	53	49.5	45.5	53.5
(54, 61)	54	61	57.5	53.5	61.5
(62, 69)	62	69	65.5	61.5	69.5
Class Width	8				

Example

Construct a frequency table for the listed data:

2, 2, 6, 3, 5, 4, 4, 7, 5, 2, 7, 4, 5, 3, 2, 2, 5, 7, 2, 4

Example

Construct a frequency table for the listed data given below using five classes. Include the relative frequency, cumulative frequency, and cumulative relative frequency.

80	67	51	82	57	80	71	82	52
75	58	62	63	51	77	85	91	97
59	61	99	70	86	98	88	87	92
66	64	61						

E13: Construct a Histogram (HW)

Construct a histogram using the information in example 12.

Class Interval	Class Boundary	Freq
51 – 60	50.5 – 60.5	6
61 – 70	60.5 – 70.5	8
71 – 80	70.5 – 80.5	4
81 – 90	80.5 – 90.5	7
91 – 100	90.5 – 100.5	5

Distribution Shapes

Histograms are valuable and useful tools. If the raw data came from a random sample of population values, the histogram constructed from the sample values should have a distribution shape that is reasonably similar to that of the population.

Several terms are commonly used to describe histograms and their associated population distributions.

Distribution Shapes

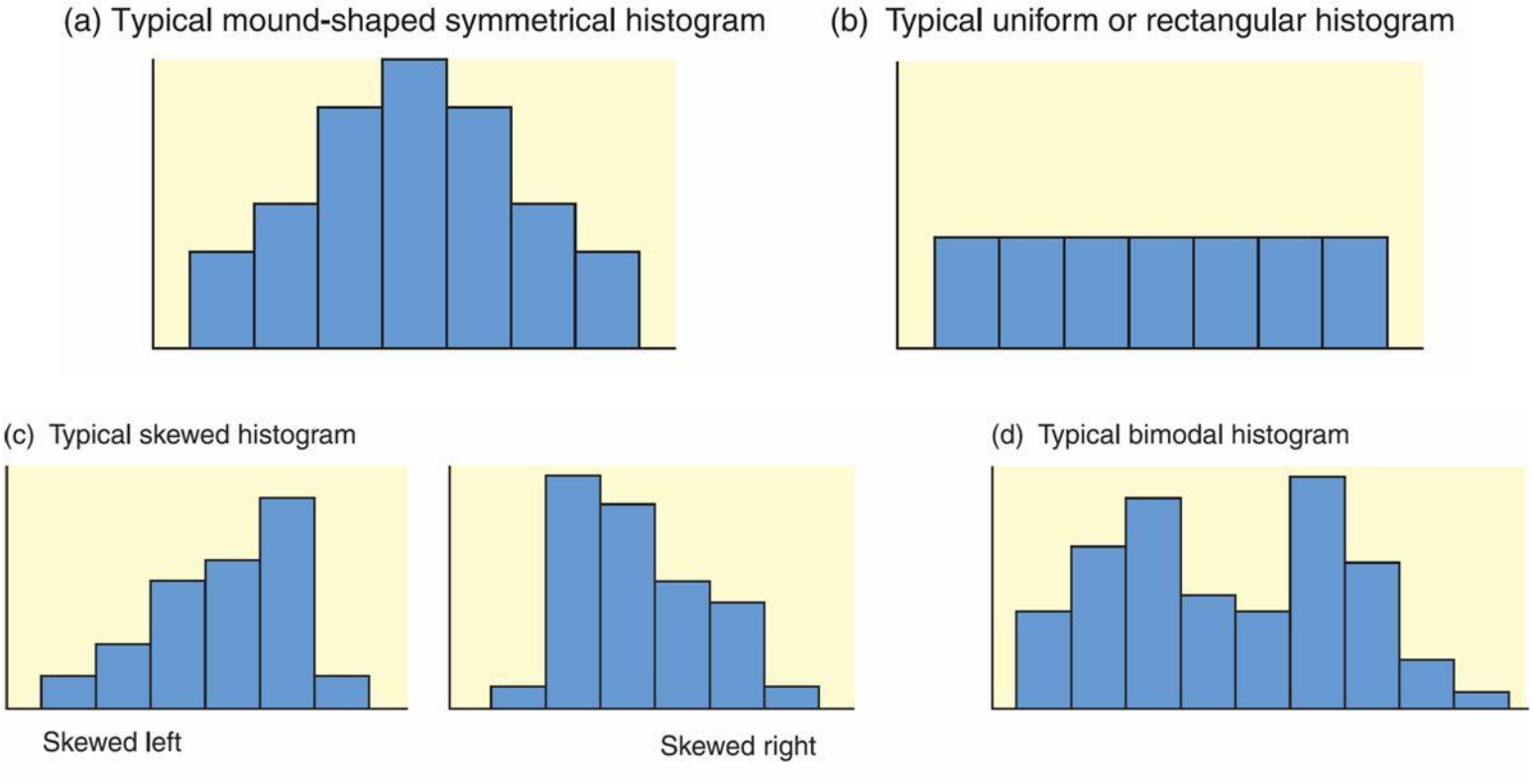
(a) ***Mound-shaped symmetrical:*** This term refers to a histogram in which both sides are (more or less) the same when the graph is folded vertically down the middle. Figure 2-8(a) shows a typical mound-shaped symmetrical histogram.

(b) ***Uniform or rectangular:*** These terms refer to a histogram in which every class has equal frequency. From one point of view, a uniform distribution is symmetrical with the added property that the bars are of the same height. Figure 2-8(b) illustrates a typical histogram with a uniform shape.

(c) ***Skewed left or skewed right:*** These terms refer to a histogram in which one tail is stretched out longer than the other. The direction of skewness is on the side of the *longer* tail. So, if the longer tail is on the left, we say the histogram is skewed to the left. Figure 2-8(c) shows a typical histogram skewed to the left and another skewed to the right.

(d) ***Bimodal:*** This term refers to a histogram in which the two classes with the largest frequencies are separated by at least one class. The top two frequencies of these classes may have slightly different values. This type of situation sometimes indicates that we are sampling from two different populations. Figure 2-8(d) illustrates a typical histogram with a bimodal shape.

Distribution Shapes



Types of Histograms

Figure 2-8

Cumulative-Frequency Tables and Ogives

Sometimes we want to study cumulative totals instead of frequencies. Cumulative frequencies tell us how many data values are smaller than an upper class boundary.

Once we have a frequency table, it is a fairly straightforward matter to add a column of cumulative frequencies.

The **cumulative frequency** for a class is the sum of the frequencies for *that class* and *all previous classes*.

Cumulative-Frequency Tables and Ogives

An *ogive (*pronounced “oh-ji ve”) is a graph that displays cumulative frequencies.

HOW TO MAKE AN OGIVE

1. Make a frequency table showing class boundaries and cumulative frequencies.
2. For each class, make a dot over the *upper class boundary* at the height of the cumulative class frequency. The coordinates of the dots are (upper class boundary, cumulative class frequency). Connect these dots with line segments.
3. By convention, an ogive begins on the horizontal axis at the lower class boundary of the first class.

Example 3 – *Cumulative-Frequency Table and Ogive*

Aspen, Colorado, is a world-famous ski area. If the daily high temperature is above 40°F, the surface of the snow tends to melt. It then freezes again at night.

This can result in a snow crust that is icy. It also can increase avalanche danger.

Example 3 – *Cumulative-Frequency Table and Ogive*

cont'd

Table 2-11 gives a summary of daily high temperatures (°F) in Aspen during the 151-day ski season.

Class Boundaries			
Lower	Upper	Frequency	Cumulative Frequency
10.5	20.5	23	23
20.5	30.5	43	66 (sum 23 + 43)
30.5	40.5	51	117 (sum 66 + 51)
40.5	50.5	27	144 (sum 117 + 27)
50.5	60.5	7	151 (sum 144 + 7)

High Temperatures During the Aspen Ski Season (°F)

Table 2-11

Example 3 – *Cumulative-Frequency Table and Ogive*

cont'd

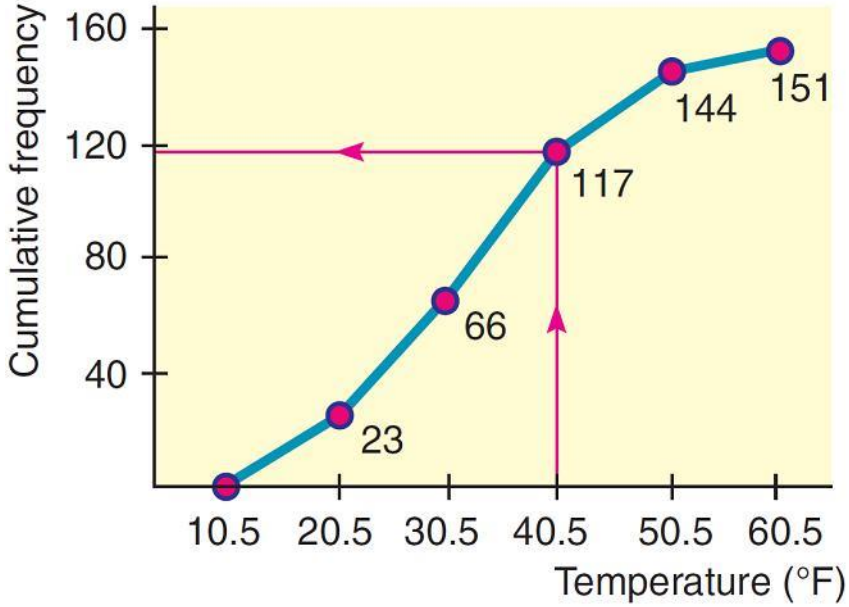
a. The cumulative frequency for a class is computed by adding the frequency of that class to the frequencies of previous classes. Table 2-11 shows the cumulative frequencies.

b. To draw the corresponding ogive, we place a dot at cumulative frequency 0 on the lower class boundary of the first class. Then we place dots over the *upper class boundaries* at the height of the cumulative class frequency for the corresponding class.

Example 3 – *Cumulative-Frequency Table and Ogive*

cont'd

Finally, we connect the dots. Figure 2-9 shows the corresponding ogive.



Ogive for Daily High Temperatures (°F) During Aspen Ski Season

Figure 2.9

Example 3 – *Cumulative-Frequency Table and Ogive*

cont'd

c. Looking at the ogive, estimate the total number of days with a high temperature lower than or equal to 40 °F.

Solution:

The red lines on the ogive in Figure 2-9, we see that 117 days have had high temperatures of no more than 40°F.

Bar Graphs, Circle Graphs, and Time-Series Graphs

Histograms provide a useful visual display of the distribution of data.

However, the data must be quantitative. In this section, we examine other types of graphs, some of which are suitable for qualitative or category data as well.

Let's start with *bar graphs*. These are graphs that can be used to display quantitative or qualitative data.

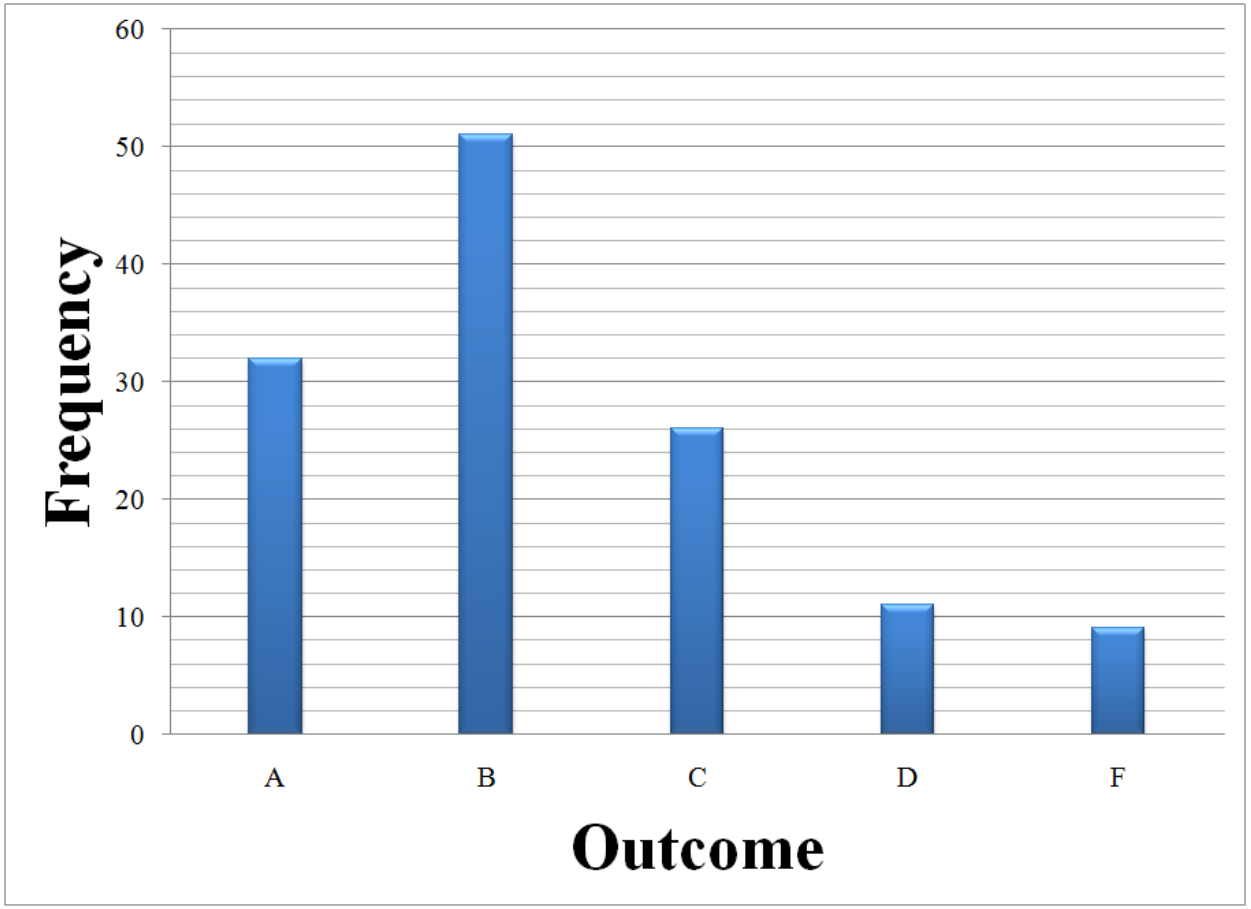
Bar Graphs, Circle Graphs, and Time-Series Graphs

Features of a Bar Graph

1. Bars can be vertical or horizontal.
2. Bars are of uniform width and uniformly spaced.
3. The lengths of the bars represent values of the variable being displayed, the frequency of occurrence, or the percentage of occurrence. The same measurement scale is used for the length of each bar.
4. The graph is well annotated with title, labels for each bar, and vertical scale or actual value for the length of each bar.

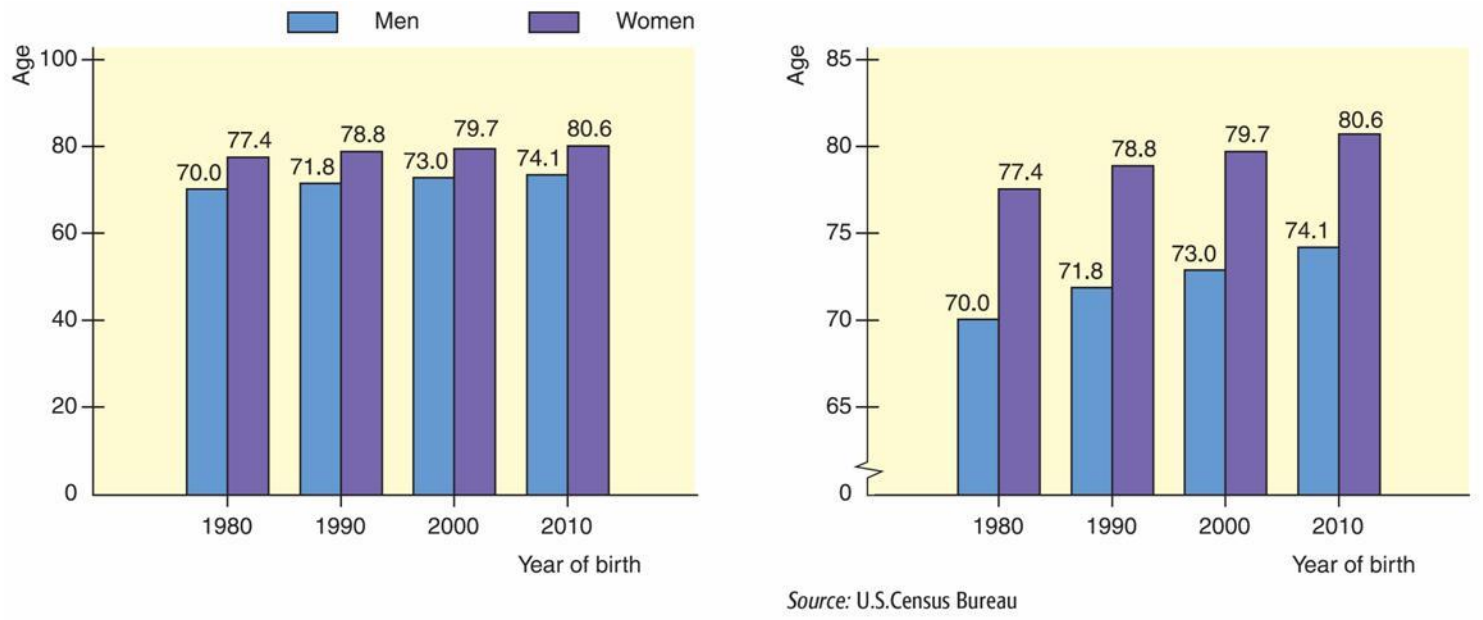
Example 5
Bar Chart

Outcome (Grade)	Frequency (No. of Student)
A	32
B	51
C	26
D	11
F	9



Example 4 – *Bar Graph*

Figure 2-11 shows two bar graphs depicting the life expectancies for men and women born in the designated year. Let's analyze the features of these graphs.



Bar Graphs, Circle Graphs, and Time-Series Graphs

An important feature illustrated in Figure 2-11(b) is that of a *changing scale*. Notice that the scale between 0 and 65 is compressed.

The changing scale amplifies the apparent difference between life spans for men and women, as well as the increase in life spans from those born in 1980 to the projected span of those born in 2010.

Changing Scale

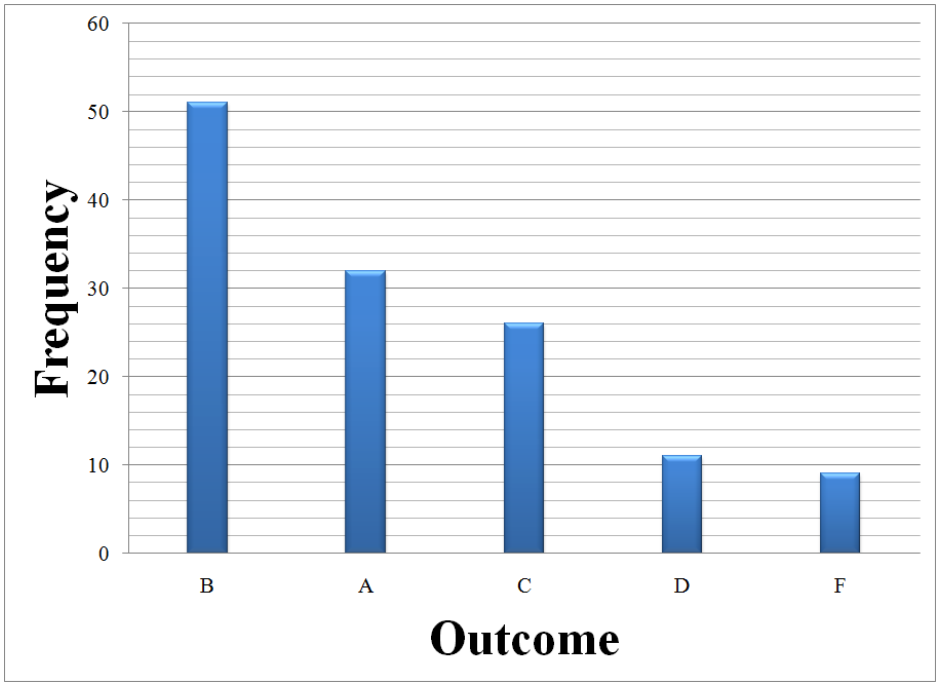
Whenever you use a change in scale in a graphic, warn the viewer by using a squiggle ⌇ on the changed axis. Sometimes, if a single bar is unusually long, the bar length is compressed with a squiggle in the bar itself.

E6
Pareto Chart

Outcome	Frequency
A	32
B	51
C	26
D	11
F	9

Sorted Outcome	Frequency
B	51
A	32
C	26
D	11
F	9

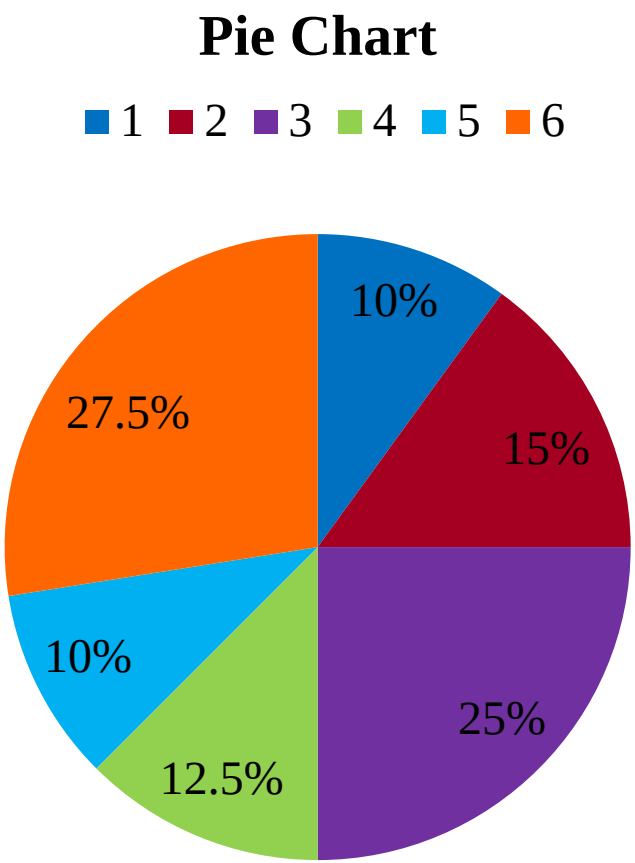
A **Pareto chart** is a bar graph in which the bar height represents frequency of an event. In addition, the bars are arranged from left to right according to decreasing height.



Example
Pie- Chart

Outcome(Color Code	Frequency
A	20
B	30
C	50
D	25
E	20
F	55
	n=200

Percentage	Degree
10	36 (=20/200)*360
15	54
25	90
12.5	45
10	36
27.5	99



Frequency Table (for qualitative data)

Summarizing an op
inion pole:

A campus press polled a sample of 280 undergraduate students in order t o study student attitude towards a pr- oposed change in the dormitory regul- ations. Each students was to respond as support, oppose, or neutral.

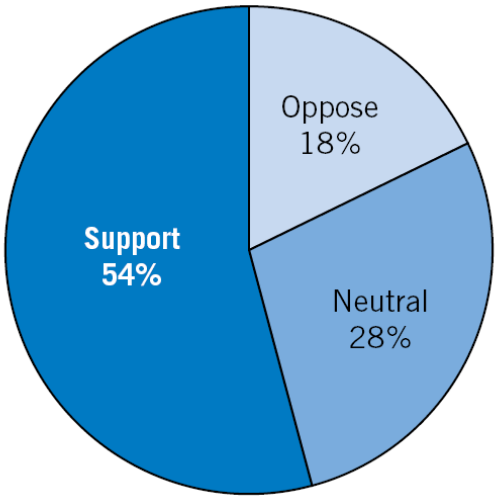
TABLE 1 Summary Results of an Opinion Poll

Responses	Frequency	Relative Frequency
Support	152	$\frac{152}{280} = .543$
Neutral	77	$\frac{77}{280} = .275$
Oppose	51	$\frac{51}{280} = .182$
Total	280	1.000

Bar Graphs, Circle Graphs, and Time-Series Graphs

These proportional segments are usually labeled with corresponding percentages of the total.

In a **circle graph** or **pie chart,** wedges of a circle visually display proportional parts of the total population that share a common characteristic.



Bar Graphs, Circle Graphs, and Time-Series Graphs

We will use a *time-series graph*. A time-series graph is a graph showing data measurements in chronological order.

To make a time-series graph, we put time on the horizontal scale and the variable being measured on the vertical scale. In a basic time-series graph, we connect the data points by line segments.

In a **time-series graph,** data are plotted in order of occurrence at regular intervals over a period of time.

Example 5 – *Time-Series Graph*

Suppose you have been in the walking/jogging exercise program for 20 weeks, and for each week you have recorded the distance you covered in 30 minutes. Your data log is shown in Table 2-14.

Week	1	2	3	4	5	6	7	8	9	10
Distance	1.5	1.4	1.7	1.6	1.9	2.0	1.8	2.0	1.9	2.0
Week	11	12	13	14	15	16	17	18	19	20
Distance	2.1	2.1	2.3	2.3	2.2	2.4	2.5	2.6	2.4	2.7

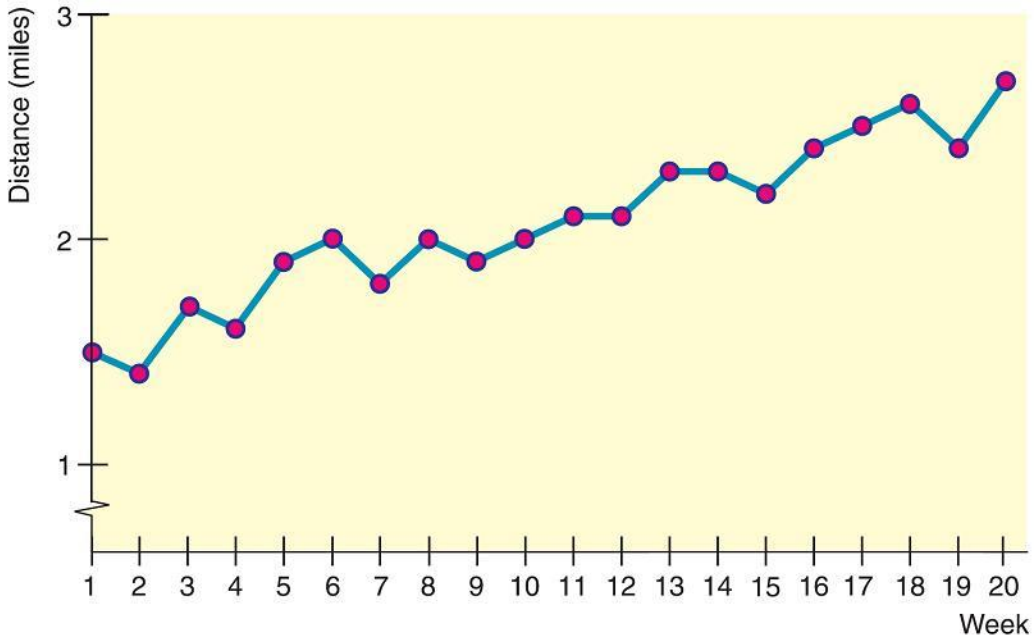
Distance (in Miles) Walked/Jogged in 30 Minutes

Table 2-14

Example 5(a) – *Solution*

cont'd

In a **time-series graph,** data are plotted in order of occurrence at regular intervals over a period of time.



Time-Series Graph of Distance (in miles) Jogged in 30 Minutes

Figure 2-14

Section 2.3

Stem-and-Leaf Displays

Focus Points

Construct a stem-and-leaf display from raw data.

Use a stem-and-leaf display to visualize data distribution.

Compare a stem-and-leaf display to a histogram.

Stem-and-Leaf Display

In this text, we will introduce the EDA techniques: stem-and-leaf displays.

A **stem-and-leaf display** is a method of exploratory data analysis that is used to rank-order and arrange data into groups.

We know that frequency distributions and histograms provide a useful organization and summary of data. However, in a histogram, we lose most of the specific data values.

Example 6 – *Stem-and-Leaf Display*

Many airline passengers seem weighted down by their carry-on luggage. Just how much weight are they carrying?

The carry-on luggage weights in pounds for a random sample of 40 passengers returning from a vacation to Hawaii were recorded (see Table 2-15).

30	27	12	42	35	47	38	36	27	35
22	17	29	3	21	0	38	32	41	33
26	45	18	43	18	32	31	32	19	21
33	31	28	29	51	12	32	18	21	26

Weights of Carry-On Luggage in Pounds

Table 2-15

Example 6 – *Stem-and-Leaf Display*

cont'd

In the stem-and-leaf display, we list each possible stem once on the left and all its leaves in the same row on the right, as in Figure 2-15(a).

3 | 2 represents 32 lb

Stem	Leaves
0	3 0
1	2 7 8 8 9 2 8
2	7 7 2 9 1 6 1 8 9 1 6
3	0 5 8 6 5 8 2 3 2 1 2 3 1 2
4	2 7 1 5 3
5	1

(a) Leaves Not Ordered

3 | 2 represents 32 lb

Stem	Leaves
0	0 3
1	2 2 7 8 8 8 9
2	1 1 1 2 6 6 7 7 8 9 9
3	0 1 1 2 2 2 2 3 3 5 5 6 8 8
4	1 2 3 5 7
5	1

(b) Final Display with Leaves Ordered

Stem-and-Leaf Displays of Airline Carry-On Luggage Weights

Figure 2-15

Example 6 – *Stem-and-Leaf Display* cont'd

Note that the lengths of the lines containing the leaves give the visual impression that a sideways histogram would present.

As a final step, we need to indicate the scale. This is usually done by indicating the value represented by the stem and one leaf.

Example 1
Stem-and-Leaf

The following information was obtained from an experiment:

1.25, 1.34, 1.47, 1.27, 1.29, 1.31, 1.35, 1.36, 1.42, 1.20, 1.45, 1.39

Leaf – the hundredths place

Stem – the units and tenths place

1.2	|	0, 5, 7, 9
1.3	|	1, 4, 5, 6, 9
1.4	|	2, 5, 7

Example 1
Stem-and-Leaf

TABLE 6 Examination Scores of 50 Students

75	98	42	75	84	87	65	59	63
86	78	37	99	66	90	79	80	89
68	57	95	55	79	88	76	60	77
49	92	83	71	78	53	81	77	58
93	85	70	62	80	74	69	90	62
84	64	73	48	72				

TABLE 7 Stem-and-Leaf Display for the Examination Scores

0	
1	
2	
3	7
4	289
5	35789
6	022345689
7	01234556778899
8	00134456789
9	0023589

What is wrong with the following stem-and-leaf below?

1 |	0, 5, 7, 9
2 |	1, 4, 5, 6, 9
4 |	2, 5, 7

The stem is missing the number, 3.