

# Chapter 2

## Descriptive Statistics: Tabular and Graphical Methods

---

In Chapter 1 we saw that although we can sometimes take a census of an entire population, we often must randomly select a sample from a population. When we have taken a census or a sample, we typically wish to describe the observed data set. In particular, we describe a sample in order to make inferences about the sampled population.

In this chapter we begin to study **descriptive statistics**, which is the science of describing the important characteristics of a data set. The techniques of descriptive statistics include **tabular and graphical methods**, which are discussed in this chapter, and **numerical methods**, which are

discussed in Chapter 3. We will see that, in practice, the methods of this chapter and the methods of Chapter 3 are used together to describe data. We will also see that the methods used to describe quantitative data differ somewhat from the methods used to describe qualitative data. Finally, we will see that there are methods—both graphical and numerical—for studying the relationships between variables.

We will illustrate the methods of this chapter by describing the cell phone usages, bottle design ratings, and car mileages introduced in the cases of Chapter 1. In addition, we introduce two new cases:



**The e-billing Case:** A management consulting firm assesses how effectively a new electronic billing system reduces bill payment times.

**The Brokerage Firm Case:** A financial broker examines whether customer satisfaction depends upon the type of investment product purchased.

# Chapter Outline

- 2.1 Graphically Summarizing Qualitative Data
- 2.2 Graphically Summarizing Quantitative Data
- 2.3 Dot Plots
- 2.4 Stem-and-Leaf Displays
- 2.5 Contingency Tables (Optional)
- 2.6 Scatter Plots (Optional)
- 2.7 Misleading Graphs and Charts (Optional)

## 2.1 Graphically Summarizing Qualitative Data

- With qualitative data, names identify the different categories
- This data can be summarized using a frequency distribution
- **Frequency distribution:** A table that summarizes the number (or frequency) of items in each of several non-overlapping classes

# Example : Describing Pizza Preferences

Pizza Preferences of 50 College Students				
Little Caesars	Papa John's	Bruno's	Papa John's	Domino's
Papa John's	Will's Uptown	Papa John's	Pizza Hut	Little Caesars
Pizza Hut	Little Caesars	Will's Uptown	Little Caesars	Bruno's
Papa John's	Bruno's	Papa John's	Will's Uptown	Papa John's
Bruno's	Papa John's	Little Caesars	Papa John's	Little Caesars
Papa John's	Little Caesars	Bruno's	Will's Uptown	Papa John's
Will's Uptown	Papa John's	Will's Uptown	Bruno's	Papa John's
Papa John's	Domino's	Papa John's	Pizza Hut	Will's Uptown
Will's Uptown	Bruno's	Pizza Hut	Papa John's	Papa John's
Little Caesars	Papa John's	Little Caesars	Papa John's	Bruno's

# The Resulting Frequency Distribution

A Frequency Distribution of Pizza Preferences	
Restaurant	Frequency
Bruno's	8
Domino's	2
Little Caesars	9
Papa John's	19
Pizza Hut	4
Will's Uptown	$\frac{8}{50}$

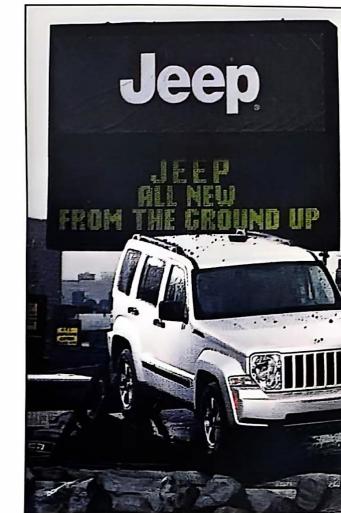
# Example 2.1: Describing 2006 Jeep Purchasing Patterns

- Jeep Sales

## EXAMPLE 2.1 Describing 2006 Jeep Purchasing Patterns

Jeep dealerships sold four different models in 2006—Jeep Commander, Jeep Grand Cherokee, Jeep Liberty, and Jeep Wrangler. At the beginning of each new model year, a dealership must decide how many vehicles of each model should be stocked in order to meet customer demand without tying up too much money in unneeded inventory. According to the sales managers at several Greater Cincinnati Jeep dealers, the most important factor influencing their stocking decisions for a new model year is the pattern of customer purchases in the prior model year. In order to study customer purchasing patterns in 2006, the sales manager for a Cincinnati Jeep dealership compiled a list of all 251 vehicles his dealership sold in 2006. These data are given in Table 2.1—in the table C = Commander, G = Grand Cherokee, L = Liberty, and W = Wrangler. (Note that Jeep introduced three new models in 2007—Jeep Compass, Jeep Patriot, and Jeep Wrangler Unlimited. Each sales manager we interviewed believed that these new models were aimed at new market segments, and that the pattern of sales for the four 2006 models would still help them to make stocking decisions for 2007.)

Unfortunately, the raw data in Table 2.1 do not reveal much useful information about the pattern of Jeep sales in 2006. In order to summarize the data in a more useful way, we can construct a frequency distribution. To do this we simply count the number of times each model appears in Table 2.1. We find that Commander (C) appears 71 times, Grand Cherokee (G) appears 70 times, Liberty (L) appears 80 times, and Wrangler (W) appears 30 times. The frequency distribution for the Jeep sales data is given in Table 2.2—it is a list of each of the four models along with their corresponding counts (or frequencies). The frequency distribution shows us how sales are distributed among the four models. The purpose of the frequency distribution is to make the data easier to understand—certainly looking at the frequency distribution in Table 2.2 is more informative than looking at the raw data in Table 2.1. We see that Jeep Liberty is the most popular model, Jeep Commander and Jeep Grand Cherokee are both slightly less popular than Jeep Liberty, and Jeep Wrangler is (by far) the least popular model.



# Relative Frequency

- The *relative frequency* of a class is the proportion or fraction of data that is contained in that class
  - Calculated by dividing the class frequency by the total number of data values
  - Relative frequency may be expressed as either a decimal or percent
  - A **relative frequency distribution** is a list of all the data classes and their associated relative frequencies

# Relative Frequency and Percent Frequency

- Relative frequency summarizes the proportion of items in each class
- For each class, divide the frequency of the class by the total number of observations
- Multiply times 100 to obtain the percent frequency

$$\text{Relative frequency of a class} = \frac{\text{frequency of the class}}{n}$$

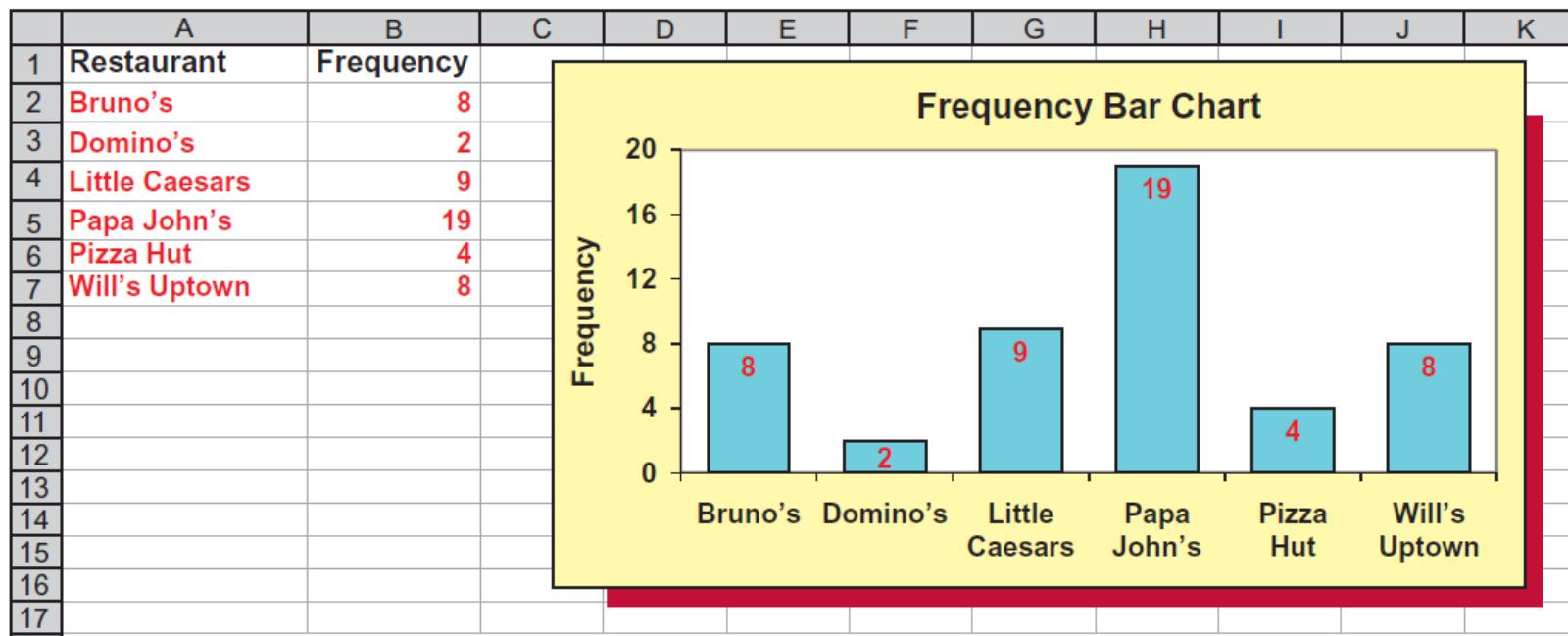
# The Resulting Relative Frequency and Percent Frequency Distribution

Relative Frequency and Percent Frequency Distributions for the Pizza Preference Data		
Restaurant	Relative Frequency	Percent Frequency
Bruno's	$8/50 = .16$	16%
Domino's	.04	4%
Little Caesars	.18	18%
Papa John's	.38	38%
Pizza Hut	.08	8%
Will's Uptown	$\frac{.16}{1.0}$	$\frac{16\%}{100\%}$

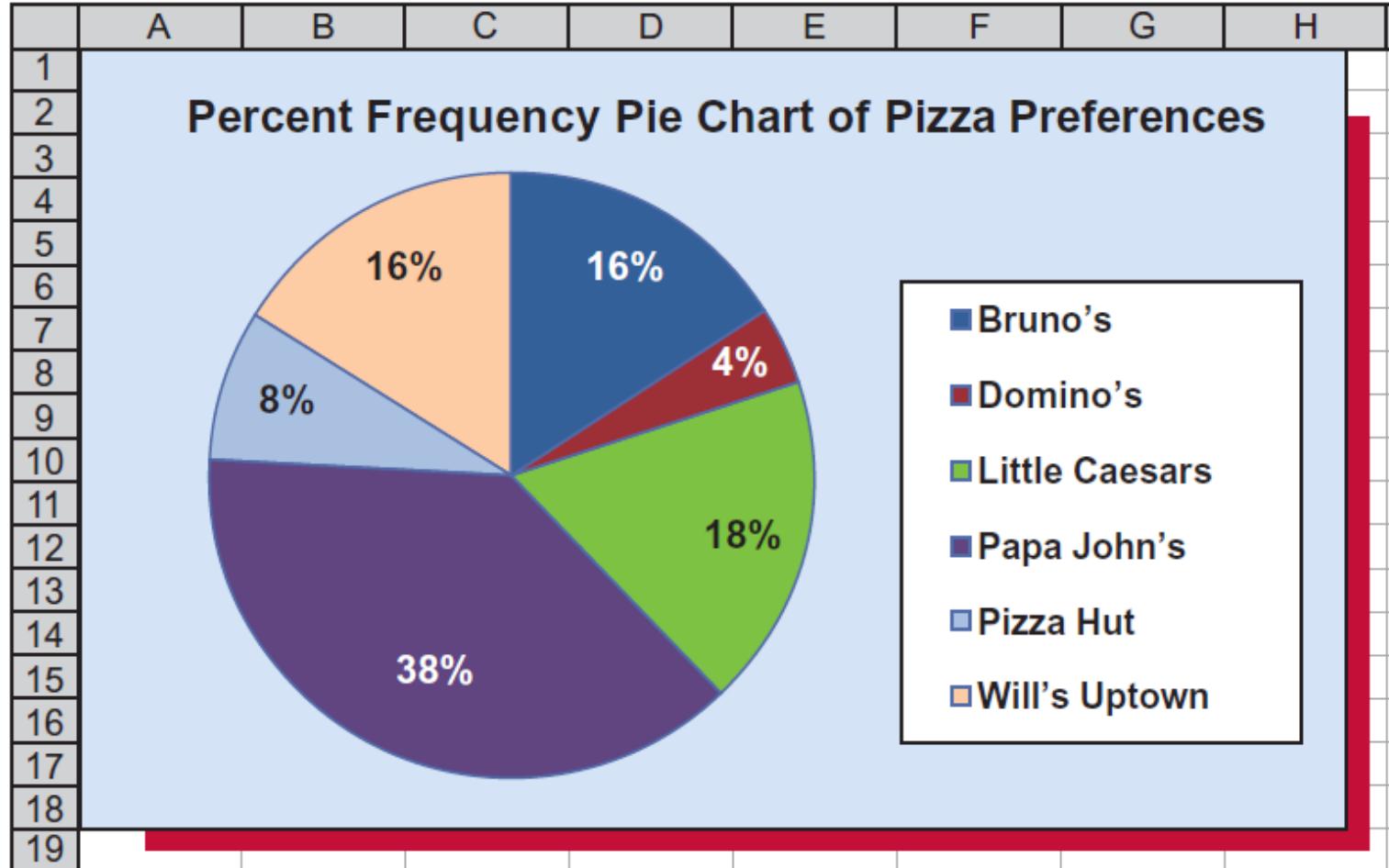
# Bar Charts and Pie Charts

- **Bar chart:** A vertical or horizontal rectangle represents the frequency for each category
  - Height can be frequency, relative frequency, or percent frequency
- **Pie chart:** A circle divided into slices where the size of each slice represents its relative frequency or percent frequency

# Excel Bar Chart of Pizza Preference



# Excel Pie Chart of Pizza Preference



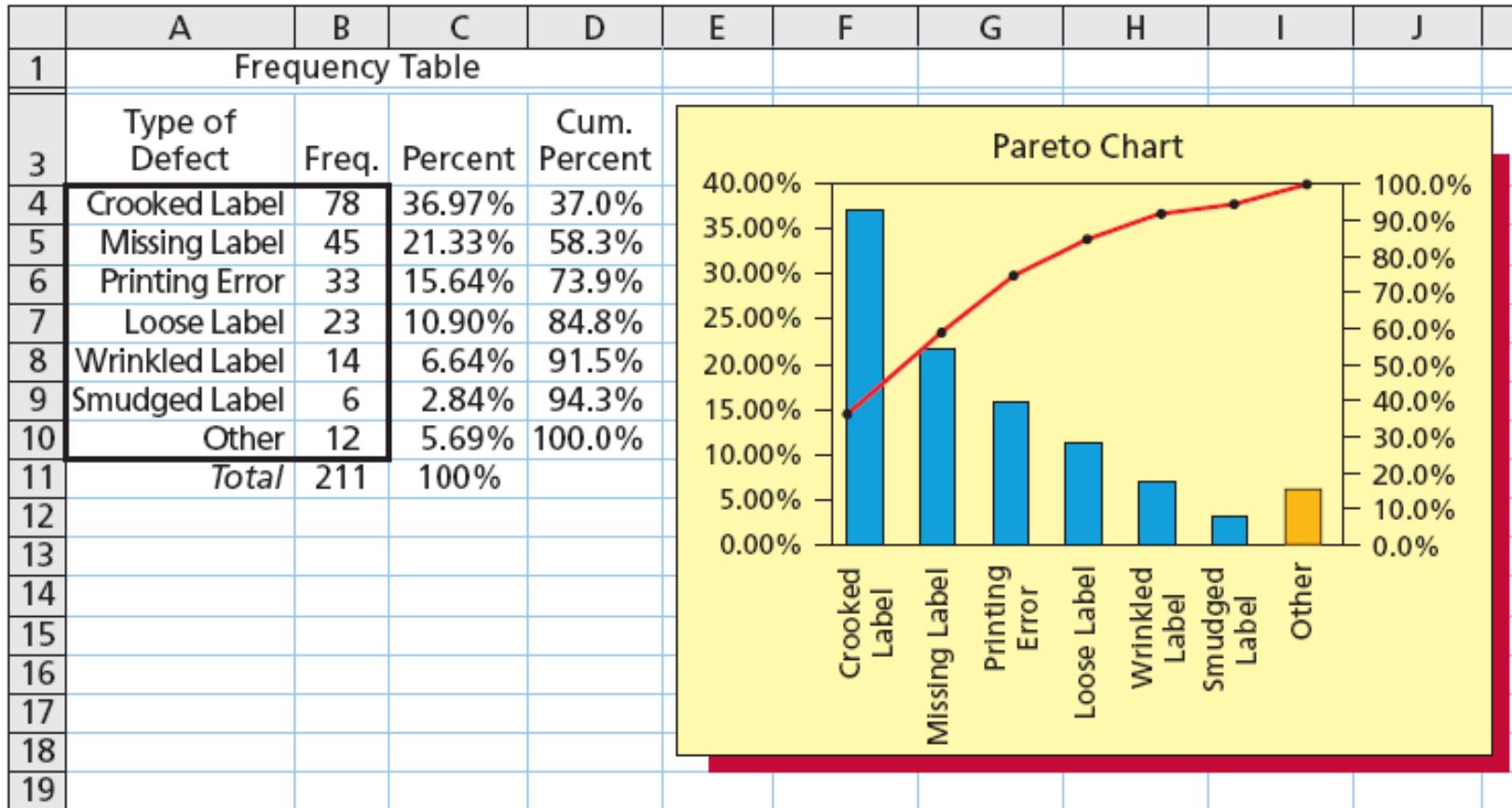
**The Pareto chart (Optional)** Pareto charts are used to help identify important quality problems and opportunities for process improvement. By using these charts we can prioritize problem-solving activities. The Pareto chart is named for Vilfredo Pareto (1848–1923), an Italian economist. Pareto suggested that, in many economies, most of the wealth is held by a small minority of the population. It has been found that the “**Pareto principle**” often applies to defects. That is, only a few defect types account for most of a product’s quality problems.

To illustrate the use of Pareto charts, suppose that a jelly producer wishes to evaluate the labels being placed on 16-ounce jars of grape jelly. Every day for two weeks, all defective labels found on inspection are classified by type of defect. If a label has more than one defect, the type of defect that is most noticeable is recorded. The Excel output in Figure 2.4 presents the frequencies and percentages of the types of defects observed over the two-week period.

with the highest frequency will be at the top of the table, the defect with the second-highest frequency below the first, and so forth. If an “other” category is employed, it should be placed at the bottom of the table. The “other” category should not make up 50 percent or more of the total of the frequencies, and the frequency for the “other” category should not exceed the frequency for the defect at the top of the table. If the frequency for the “other” category is too high, data should be collected so that the “other” category can be broken down into new categories. Once the frequency and the percentage for each category are determined, a cumulative percentage for each category is computed. As illustrated in Figure 2.4, the cumulative percentage for a particular category is the sum of the percentages corresponding to the particular category and the categories that are above that category in the table.

A Pareto chart is simply a bar chart having the different kinds of defects or problems listed on the horizontal scale. The heights of the bars on the vertical scale typically represent the frequency of occurrence (or the percentage of occurrence) for each defect or problem. The bars are arranged in decreasing height from left to right. Thus, the most frequent defect will be at the far left, the next most frequent defect to its right, and so forth. If an “other” category is employed, its bar is placed at the far right. The Pareto chart for the labeling defects data is given in Figure 2.4. Here the heights of the bars represent the percentages of occurrences for the different labeling defects, and the vertical scale on the far left corresponds to these percentages. The chart graphically illustrates that crooked labels, missing labels, and printing errors are the most frequent labeling defects.

# Pareto Chart



## 2.2 Graphically Summarizing Quantitative Data

- Often need to summarize and describe the shape of the distribution
- One way is to group the measurements into classes of a frequency distribution and then displaying the data in the form of a histogram

# Example 2.2 The e-Billing Case

## EXAMPLE 2.2 The e-billing Case: Reducing Bill Payment Times<sup>3</sup>

C



Major consulting firms such as Accenture, Ernst & Young Consulting, and Deloitte & Touche Consulting employ statistical analysis to assess the effectiveness of the systems they design for their customers. In this case a consulting firm has developed an electronic billing system for a Hamilton, Ohio, trucking company. The system sends invoices electronically to each customer's computer and allows customers to easily check and correct errors. It is hoped that the new billing system will substantially reduce the amount of time it takes customers to make payments. Typical payment times—measured from the date on an invoice to the date payment is received—using the trucking company's old billing system had been 39 days or more. This exceeded the industry standard payment time of 30 days.

A Sample of Payment Times (in Days) for 65 Randomly Selected Invoices

22	29	16	15	18	17	12	13	17	16	15
19	17	10	21	15	14	17	18	12	20	14
16	15	16	20	22	14	25	19	23	15	19
18	23	22	16	16	19	13	18	24	24	26
13	18	17	15	24	15	17	14	18	17	21
16	21	25	19	20	27	16	17	16	21	

The consulting firm will use this sample of 65 randomly selected invoices to estimate the average payment time for all 7,823 invoices. By examining the sample data, the consulting firm can assess the system's performance. For example, the average payment time for the 65 invoices in the sample was 19.8 days. This is less than the 30-day standard, so the consulting firm may conclude that the new system is effective. However, the consulting firm must also consider the variability of the payment times. If the payment times are highly variable, then the consulting firm may conclude that the new system is not effective. To do this, the consulting firm will form a frequency distribution of the data and will graph the distribution by constructing a histogram. Similar to the frequency distributions for qualitative data we studied in Section 2.1, the frequency distribution will divide the payment times into classes and will tell us how many of the payment times are in each class.

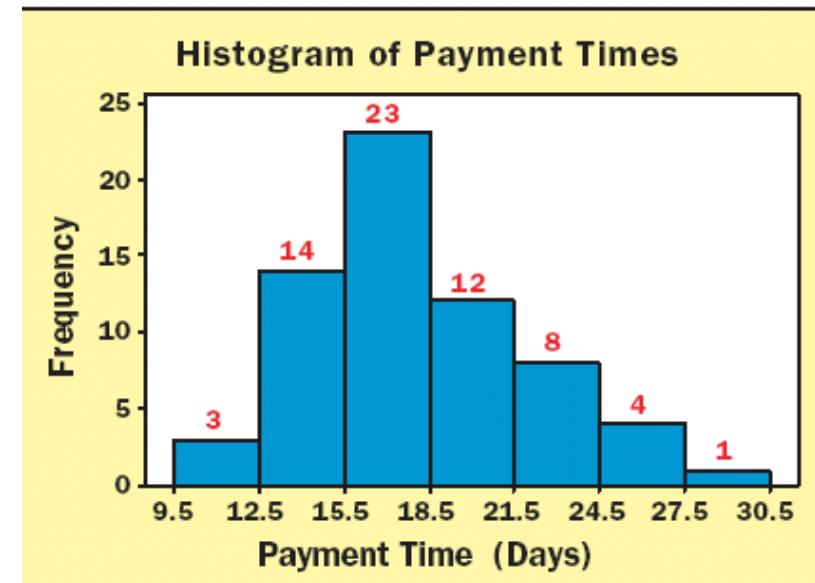
# Frequency Distribution and Histogram

A *frequency distribution* is a list of data classes with the count or “frequency” of values that belong to each class

- “Classify and count”
- The frequency distribution is a table

Show the frequency distribution in a *histogram*

- The histogram is a picture of the frequency distribution



# Steps in Constructing a Frequency Distribution

1. Find the number of classes
2. Find the class length
3. Form non-overlapping classes of equal width
4. Tally and count the number of measurements in each class
5. Graph the histogram

# Example 2.2 The e-Billing Case

## Paytime

Table 2.4

A Sample of Payment Times (in Days) for 65 Randomly Selected Invoices											
22	29	16	15	18	17	12	13	17	16	15	
19	17	10	21	15	14	17	18	12	20	14	
16	15	16	20	22	14	25	19	23	15	19	
18	23	22	16	16	19	13	18	24	24	26	
13	18	17	15	24	15	17	14	18	17	21	
16	21	25	19	20	27	16	17	16	21		

# Number of Classes

- Group all of the  $n$  data into  $K$  number of classes
- $K$  is the smallest whole number for which  $2^K \geq n$
- In Examples 2.2  $n = 65$ 
  - For  $K = 6$ ,  $2^6 = 64, < n$
  - For  $K = 7$ ,  $2^7 = 128, > n$
  - So use  $K = 7$  classes

# Class Length

- Find the length of each class as the largest measurement minus the smallest divided by the number of classes found earlier (K)
- For Example 2.2,  $(29-10)/7 = 2.7143$ 
  - Because payments measured in days, round to three days

$$\text{approximate class length} = \frac{\text{largest measurement} - \text{smallest measurement}}{\text{number of classes}}$$

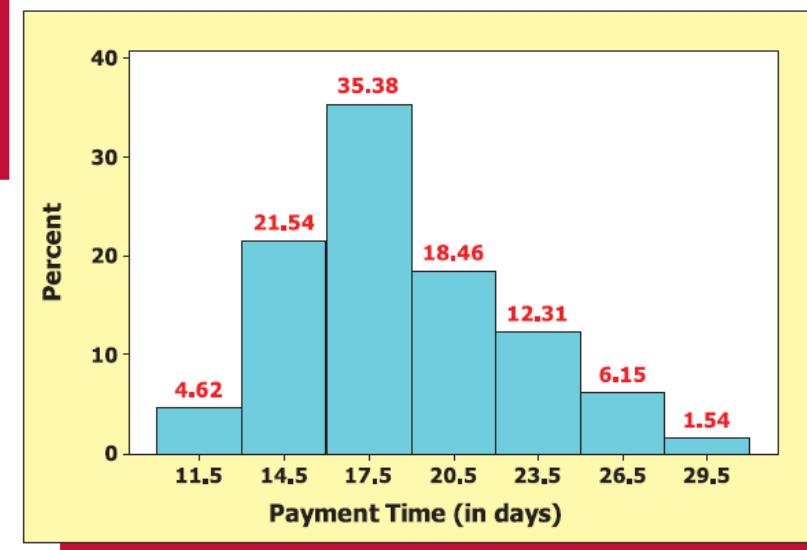
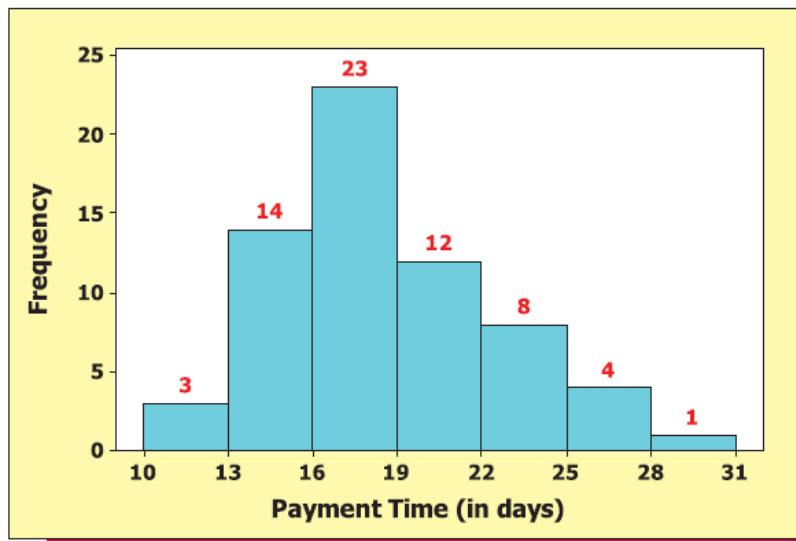
# Form Non-Overlapping Classes of Equal Width

- |         |                               |
|---------|-------------------------------|
| Class 1 | 10 days and less than 13 days |
| Class 2 | 13 days and less than 16 days |
| Class 3 | 16 days and less than 19 days |
| Class 4 | 19 days and less than 22 days |
| Class 5 | 22 days and less than 25 days |
| Class 6 | 25 days and less than 28 days |
| Class 7 | 28 days and less than 31 days |

# Tally and Count the Number of Measurements in Each Class

# Histograms

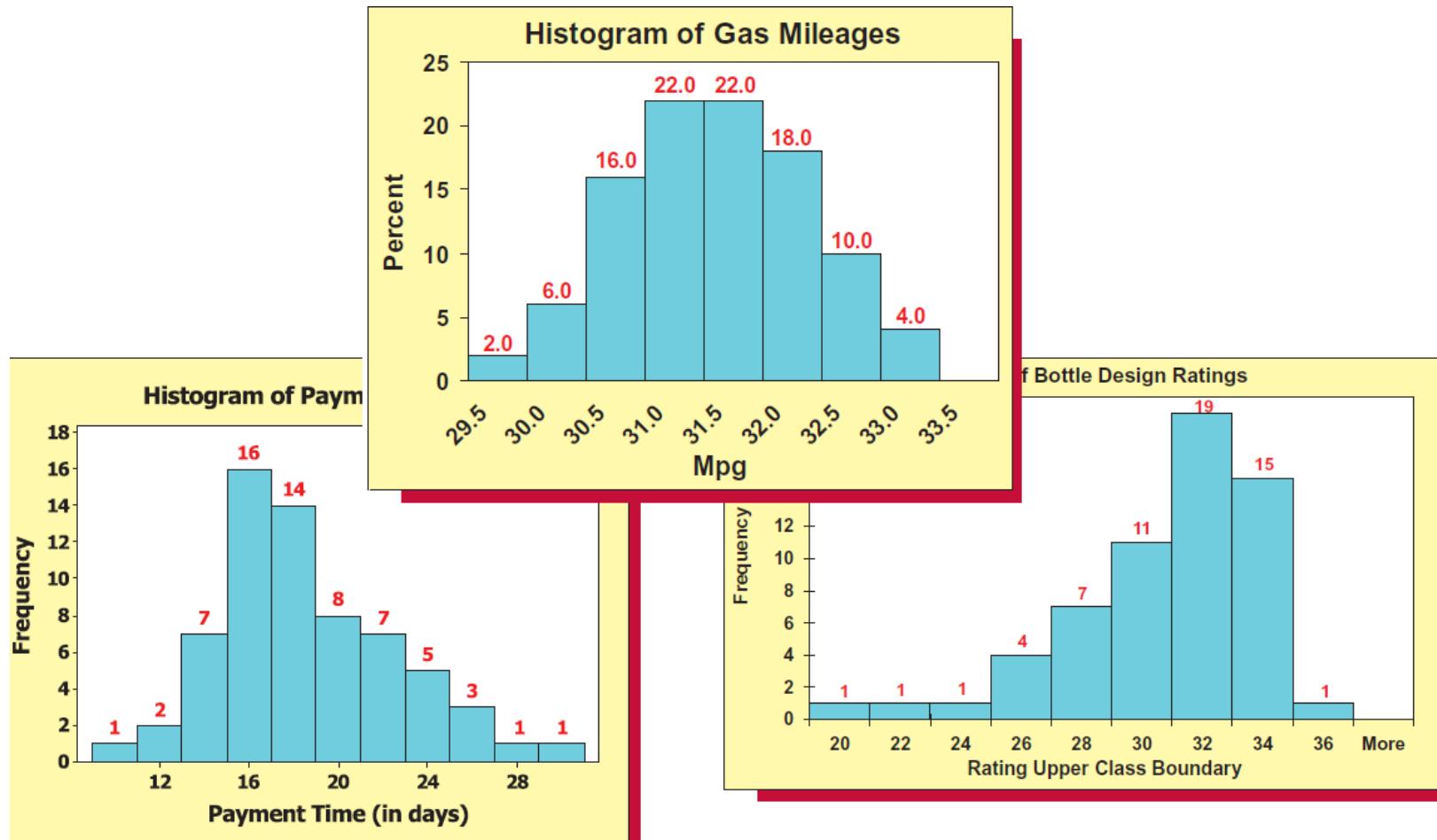
Figure 2.7 and 2.8



- 1 None of the payment times exceeds the industry standard of 30 days. (Actually, all of the payment times are less than 30—remember the largest payment time is 29 days.)
- 2 The payment times are concentrated between 13 and 24 days (57 of the 65, or  $(57/65) \times 100 = 87.69\%$ , of the payment times are in this range).
- 3 More payment times are in the class “ $16 < 19$ ” than are in any other class (23 payment times are in this class).

# Skewed Distributions

Figure 2.9, 2.10, 2.11

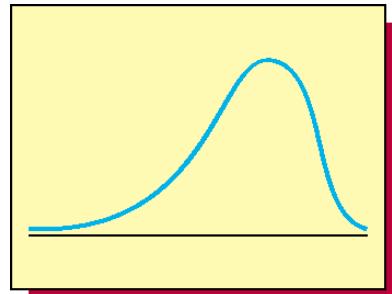


# Skewness

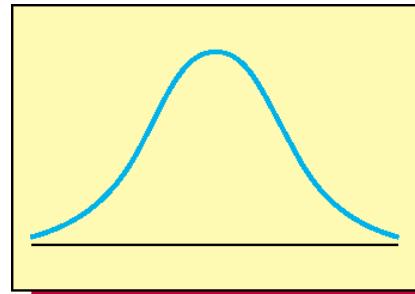
Skewed distributions are not symmetrical about their center. Rather, they are lop-sided with a longer tail on one side or the other.

- A population is distributed according to its relative frequency curve
- The skew is the side with the longer tail

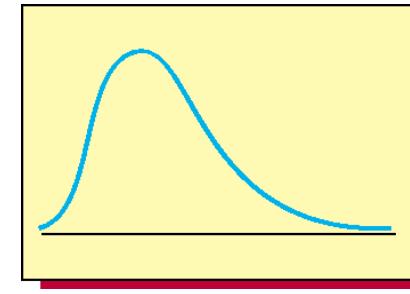
Left Skewed



Symmetric

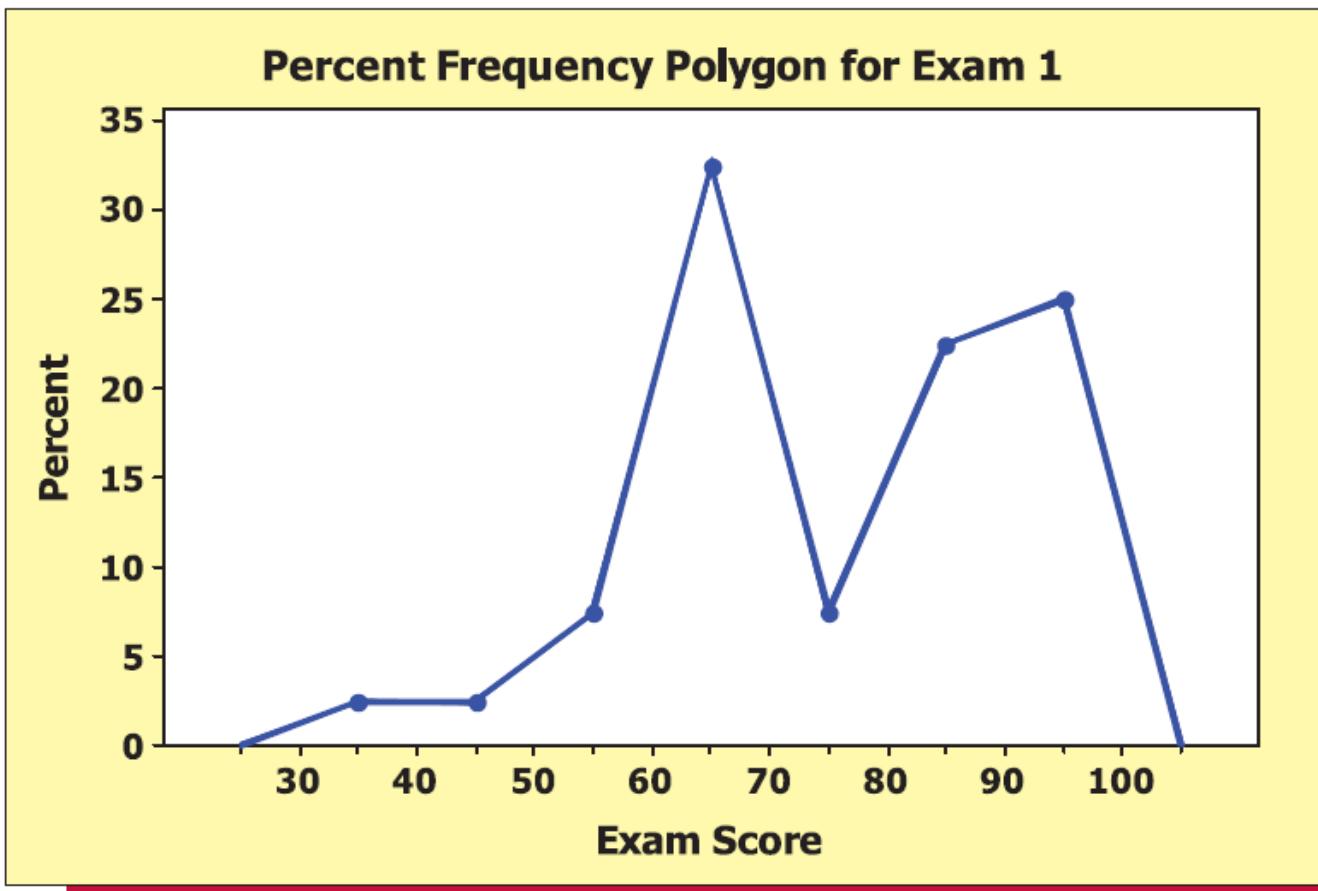


Right Skewed



# A Percent Frequency Polygon of the Exam Scores

Figure 2.12



# Several Frequency Distributions

Table 2.10

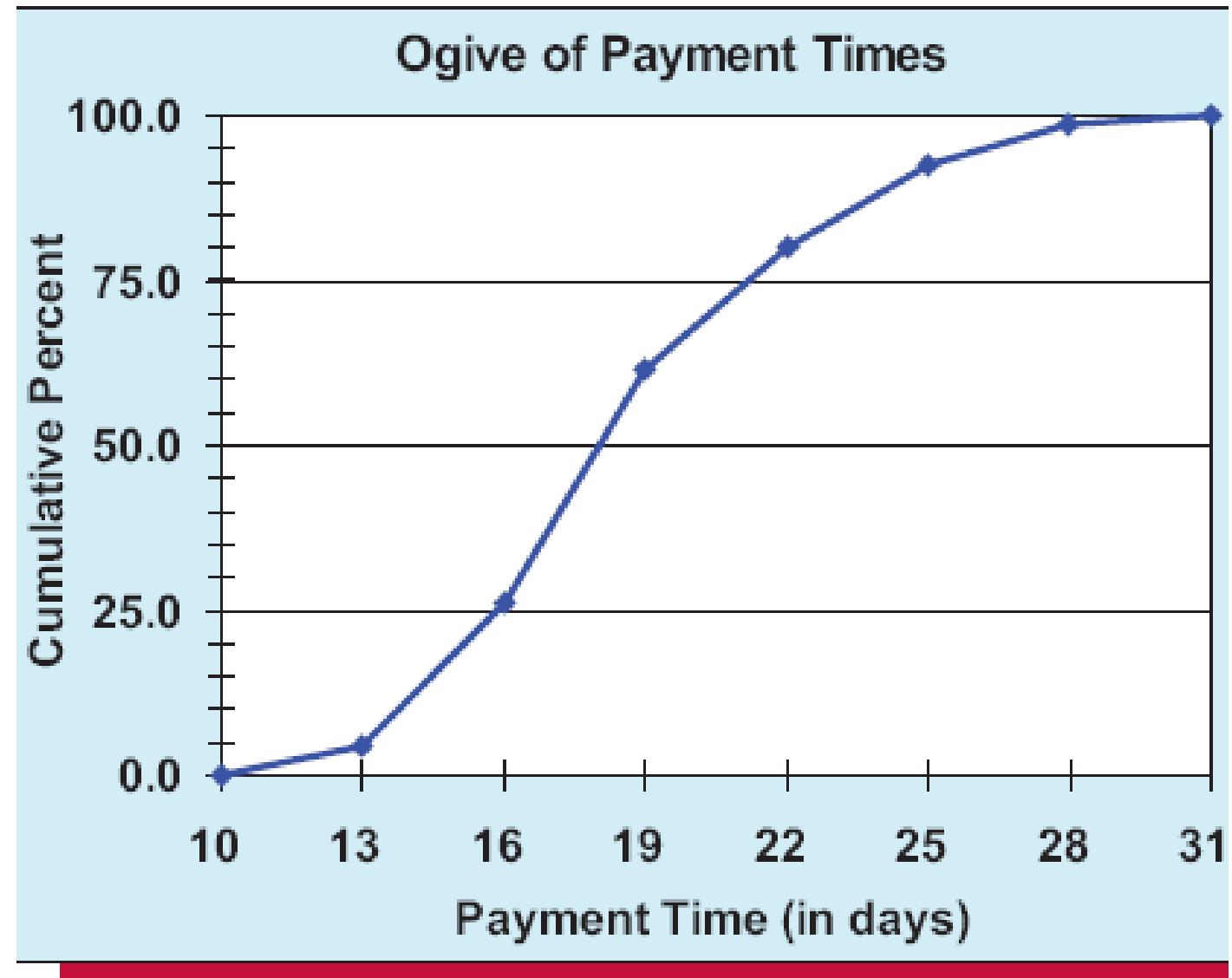
A Frequency Distribution, Cumulative Frequency Distribution, Cumulative Relative Frequency Distribution, and Cumulative Percent Frequency Distribution for the Payment Time Data				
(1) Class	(2) Frequency	(3) Cumulative Frequency	(4) Cumulative Relative Frequency	(5) Cumulative Percent Frequency
10 < 13	3	3	$3/65 = .0462$	4.62%
13 < 16	14	17	$17/65 = .2615$	26.15
16 < 19	23	40	.6154	61.54
19 < 22	12	52	.8000	80.00
22 < 25	8	60	.9231	92.31
25 < 28	4	64	.9846	98.46
28 < 31	1	65	1.0000	100.00

# Cumulative Distributions

- Another way to summarize a distribution is to construct a cumulative distribution
- To do this, use the same number of classes, class lengths, and class boundaries used for the frequency distribution
- Rather than a count, we record the number of measurements that are less than the upper boundary of that class
  - In other words, a running total

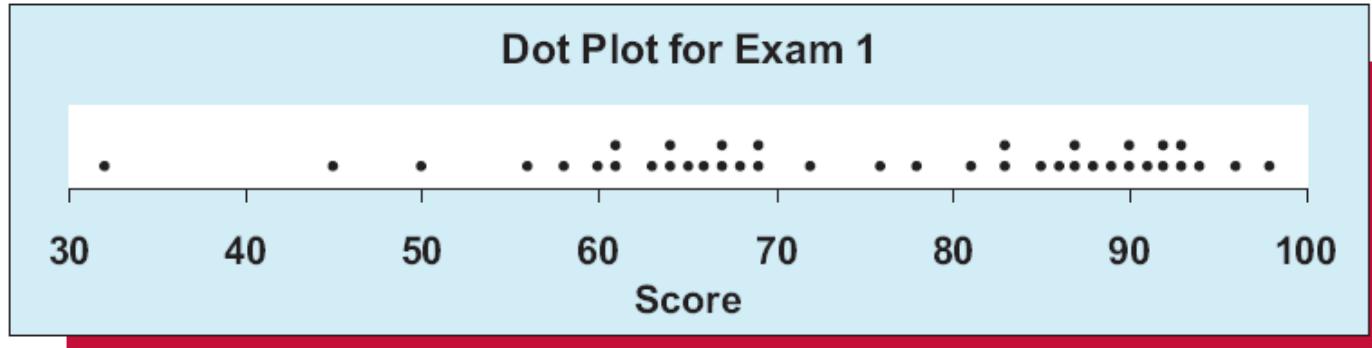
## Ogive

- **Ogive:** A graph of a cumulative distribution
- Plot a point above each upper class boundary at height of cumulative frequency
- Connect points with line segments
- Can also be drawn using:
  - Cumulative relative frequencies
  - Cumulative percent frequencies

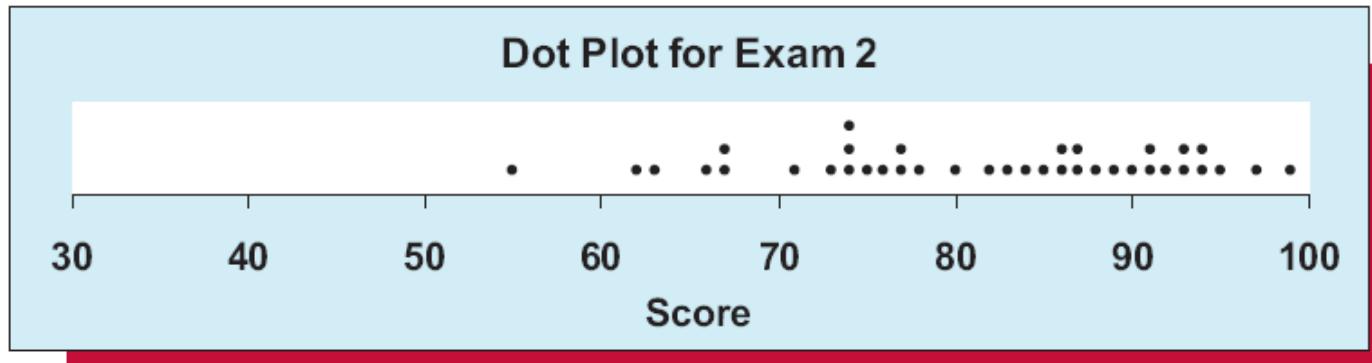


## 2.3 Dot Plots

(a) Dot Plot of Scores on Exam 1: Before Attendance Policy



(b) Dot Plot of Scores on Exam 2: After Attendance Policy



## 2.4 Stem-and-leaf Display

- The purpose of a stem-and-leaf display is to see the overall pattern of the data, by grouping the data into classes
  - To see:
    - the variation from class to class
    - the amount of data in each class
    - the distribution of the data within each class
  - Best for small to moderately sized data distributions

# Constructing a Stem-and-Leaf Display

- Decide what units will be used for the stems and the leaves. As a general rule, choose units for the stems so that there will be somewhere between 5 and 20 stems.
- Place the stems in a column with the smallest stem at the top of the column and the largest stem at the bottom.
- Enter the leaf for each measurement into the row corresponding to the proper stem. The leaves should be single-digit numbers (rounded values).
- If desired, rearrange the leaves so that they are in increasing order from left to right.

## The Car Mileage Case

- In this case study, we consider a tax credit offered by the federal government to automakers for improving the fuel economy of midsize cars.
- To find the combined city and highway mileage estimate for a particular car model, the EPA tests a sample of cars.

## EXAMPLE 2.4 The Car Mileage Case: Estimating Mileage

C



Table 2.14 presents the sample of 50 gas mileages for the new midsize model previously introduced in Chapter 1. To develop a stem-and-leaf display, we note that the sample mileages range from 29.8 to 33.3 and we place the leading digits of these mileages—the whole numbers 29, 30, 31, 32, and 33—in a column on the left side of a vertical line as follows.

29	
30	
31	
32	
33	

This vertical arrangement of leading digits forms the **stem** of the display. Next, we pass through the mileages in Table 2.14 one at a time and place each last digit (the tenths place) to the right of the vertical line in the row corresponding to its leading digits. For instance, the first three mileages—30.8, 31.7, and 30.1—are arranged as follows:

29	
30	8 1
31	7
32	
33	

We form the **leaves** of the display by continuing this procedure as we pass through all 50 mileages. After recording the last digit for each of the mileages, we sort the digits in each row from smallest to largest and obtain the stem-and-leaf display that follows:

29	8
30	1 3 4 5 5 6 7 7 8 8 8
31	0 0 1 2 3 3 4 4 4 4 5 5 6 6 7 7 7 8 8 9 9
32	0 1 1 1 2 3 3 4 4 5 5 7 7 8
33	0 3

As we have said, the numbers to the left of the vertical line form the stem of the display. Each number to the right of the vertical line is a leaf. Each combination of a stem value and a leaf value

**TABLE 2.14** A Sample of 50 Mileages for a New Midsize Model  GasMiles

30.8	30.8	32.1	32.3	32.7
31.7	30.4	31.4	32.7	31.4
30.1	32.5	30.8	31.2	31.8
31.6	30.3	32.8	30.7	31.9
32.1	31.3	31.9	31.7	33.0
33.3	32.1	31.4	31.4	31.5
31.3	32.5	32.4	32.2	31.6
31.0	31.8	31.0	31.5	30.6
32.0	30.5	29.8	31.7	32.3
32.4	30.5	31.1	30.7	31.4

The stem-and-leaf display of car mileages:

		$29 + 0.8 = 29.8$
29	8	
30	13445666889	
31	00123344455566777889	
32	0001122344556788	
33	3	$33 + 0.3 = 33.3$

- There are no rules that dictate the number of stem values
- Can split the stems as needed

## Another display of the same data using more classes

- Starred classes (\*) extend from 0.0 to 0.4
- Unstarred classes extend from 0.5 to 0.9

29	8
30*	1344
30	5666889
31*	001233444
31	55566777889
32*	0001122344
32	556788
33*	3

- Looking at the last stem-and-leaf display, the distribution appears almost “symmetrical”
  - ❑ The upper portion of the display...
    - Stems 29, 30\*, 30, and 31\*
  - ❑ ... is almost a mirror image of the lower portion of the display
    - Stems 31, 32\*, 32, and 33\*
  - ❑ But not exactly a mirror reflection
    - Maybe slightly more data in the lower portion than in the upper portion
    - ❑ Later, we will call this a slightly “left-skewed” distribution

## Example: The Payment Time Case: Reducing Payment Times

- In order to assess the effectiveness of the system, the consulting firm will study the payment times for invoices processed during the first three months of the system's operation.
- During this period, 7,823 invoices are processed using the new system. To study the payment times of these invoices, the consulting firm numbers the invoices from 0001 to 7823 and uses random numbers to select a random sample of 65 invoices. The resulting 65 payment times are given in Table

Table: A Sample of Payment Times (in Days) for 65 Randomly Selected Invoices.

22	29	16	15	18	17	12	13	17	16	15
19	17	10	21	15	14	17	18	12	20	14
16	15	16	20	22	14	25	19	23	15	19
18	23	22	16	16	19	13	18	24	24	26
13	18	17	15	24	15	17	14	18	17	21
16	21	25	19	20	27	16	17	16	21	

The leftmost column of numbers are the numbers are the amounts of values in each stem

- The number 8 in parentheses indicates that there are 8 payments in the stem for 17 days
- The number 26 (no parentheses) indicates that there are 26 payments made in 16 or less days

1	10	0	Shorter tail
1	11		
3	12	00	
6	13	000	
10	14	0000	
17	15	0000000	
26	16	000000000	
(8)	17	00000000	
30	18	000000	
24	19	00000	
19	20	000	
16	21	000	
13	22	000	
10	23	00	
8	24	000	
5	25	00	Longer tail
3	26	0	
2	27	0	
1	28		
1	29	0	

# The Payment Times: Results

- Looking at this display, we see that all of the sampled payment times are substantially less than the 39-day typical payment time of the former billing system.
- The stem-and-leaf display do not appear symmetrical. The “tail” of the distribution consisting of the higher payment times is longer than the “tail” of the distribution consisting of the smaller payment times.
- We say that the distribution is **skewed with a tail to the right.**

## 2.5 Contingency Tables (Crosstabulation)

- Classifies data on two dimensions
  - Rows classify according to one dimension
  - Columns classify according to a second dimension
- Requires three variable
  - The row variable
  - The column variable
  - The variable counted in the cells

## EXAMPLE 2.5 The Brokerage Firm Case: Studying Client Satisfaction

An investment broker sells several kinds of investment products—a stock fund, a bond fund, and a tax-deferred annuity. The broker wishes to study whether client satisfaction with its products and services depends on the type of investment product purchased. To do this, 100 of the broker’s clients are randomly selected from the population of clients who have purchased shares in exactly one of the funds. The broker records the fund type purchased by each client and has one of its investment counselors personally contact the client. When contacted, the client is asked to rate his or her level of satisfaction with the purchased fund as high, medium, or low. The resulting data are given in Table 2.16.

Looking at the raw data in Table 2.16, it is difficult to see whether the level of client satisfaction varies depending on the fund type. We can look at the data in an organized way by constructing a contingency table. A cross-tabulation of fund type versus level of client satisfaction is shown in Table 2.17. The classification categories for the two variables are defined along the left and top margins of the table. The three row labels—bond fund, stock fund, and tax deferred annuity—define the three fund categories and are given in the left table margin. The three column labels—high, medium, and low—define the three levels of client satisfaction and are given along the top table margin. Each row and column combination, that is, each fund type and level of satisfaction combination, defines what we call a “cell” in the table. Because each of the randomly selected clients has invested in exactly one fund type and has reported exactly one level of satisfaction, each client can be placed in a particular cell in the contingency table. For example, because client number 1 in Table 2.16 has invested in the bond fund and reports a high level of client satisfaction, client number 1 can be placed in the upper left cell of the table (the cell defined by the Bond Fund row and High Satisfaction column).

We fill in the cells in the table by moving through the 100 randomly selected clients and by tabulating the number of clients who can be placed in each cell. For instance, moving through the 100 clients results in placing 15 clients in the “bond fund—high” cell, 12 clients in the

**TABLE 2.16** Results of a Customer Satisfaction Survey Given to 100 Randomly Selected Clients Who Invest in One of Three Fund Types—a Bond Fund, a Stock Fund, or a Tax-Deferred Annuity 

Client	Fund Type	Level of Satisfaction	Client	Fund Type	Level of Satisfaction	Client	Fund Type	Level of Satisfaction
1	BOND	HIGH	35	STOCK	HIGH	69	BOND	MED
2	STOCK	HIGH	36	BOND	MED	70	TAXDEF	MED
3	TAXDEF	MED	37	TAXDEF	MED	71	TAXDEF	MED
4	TAXDEF	MED	38	TAXDEF	LOW	72	BOND	HIGH
5	STOCK	LOW	39	STOCK	HIGH	73	TAXDEF	MED
6	STOCK	HIGH	40	TAXDEF	MED	74	TAXDEF	LOW
7	STOCK	HIGH	41	BOND	HIGH	75	STOCK	HIGH
8	BOND	MED	42	BOND	HIGH	76	BOND	HIGH
9	TAXDEF	LOW	43	BOND	LOW	77	TAXDEF	LOW
10	TAXDEF	LOW	44	TAXDEF	LOW	78	BOND	MED
11	STOCK	MED	45	STOCK	HIGH	79	STOCK	HIGH
12	BOND	LOW	46	BOND	HIGH	80	STOCK	HIGH
13	STOCK	HIGH	47	BOND	MED	81	BOND	MED
14	TAXDEF	MED	48	STOCK	HIGH	82	TAXDEF	MED
15	TAXDEF	MED	49	TAXDEF	MED	83	BOND	HIGH
16	TAXDEF	LOW	50	TAXDEF	MED	84	STOCK	MED
17	STOCK	HIGH	51	STOCK	HIGH	85	STOCK	HIGH
18	BOND	HIGH	52	TAXDEF	MED	86	BOND	MED
19	BOND	MED	53	STOCK	HIGH	87	TAXDEF	MED
20	TAXDEF	MED	54	TAXDEF	MED	88	TAXDEF	LOW
21	TAXDEF	MED	55	STOCK	LOW	89	STOCK	HIGH
22	BOND	HIGH	56	BOND	HIGH	90	TAXDEF	MED
23	TAXDEF	MED	57	STOCK	HIGH	91	BOND	HIGH
24	TAXDEF	LOW	58	BOND	MED	92	TAXDEF	HIGH
25	STOCK	HIGH	59	TAXDEF	LOW	93	TAXDEF	LOW
26	BOND	HIGH	60	TAXDEF	LOW	94	TAXDEF	LOW
27	TAXDEF	LOW	61	STOCK	MED	95	STOCK	HIGH
28	BOND	MED	62	BOND	LOW	96	BOND	HIGH
29	STOCK	HIGH	63	STOCK	HIGH	97	BOND	MED
30	STOCK	HIGH	64	TAXDEF	MED	98	STOCK	HIGH
31	BOND	MED	65	TAXDEF	MED	99	TAXDEF	MED
32	TAXDEF	MED	66	TAXDEF	LOW	100	TAXDEF	MED
33	BOND	HIGH	67	STOCK	HIGH			
34	STOCK	MED	68	BOND	HIGH			

# Bond Fund Satisfaction Survey

Table 2.17

Fund Type	Level of Satisfaction			Total
	High	Medium	Low	
Bond Fund	15	12	3	30
Stock Fund	24	4	2	30
Tax Deferred Annuity	1	24	15	40
Total	40	40	20	100

# More on Contingency Tables

- Row totals provide a frequency distribution for the different fund types
- Column totals provide a frequency distribution for the different satisfaction levels
- Main purpose is to investigate possible relationships between variables

# Percentages

- One way to investigate relationships is to compute row and column percentages
- Compute row percentages by dividing each cell's frequency by its row total and expressing as a percentage
- Compute column percentages by dividing by the column total

# Row Percentage for Each Fund Type

Table 2.18

Fund Type	Level of Satisfaction			Total
	High	Medium	Low	
Bond Fund	50%	40%	10%	100%
Stock Fund	80%	13.33%	6.67%	100%
Tax Deferred	2.5%	60%	37.5%	100%

# Types of Variables

- In the bond fund example, we cross-tabulated two qualitative variables
- Can use a quantitative variable versus a qualitative variable or two quantitative variables
- With quantitative variables, often define categories

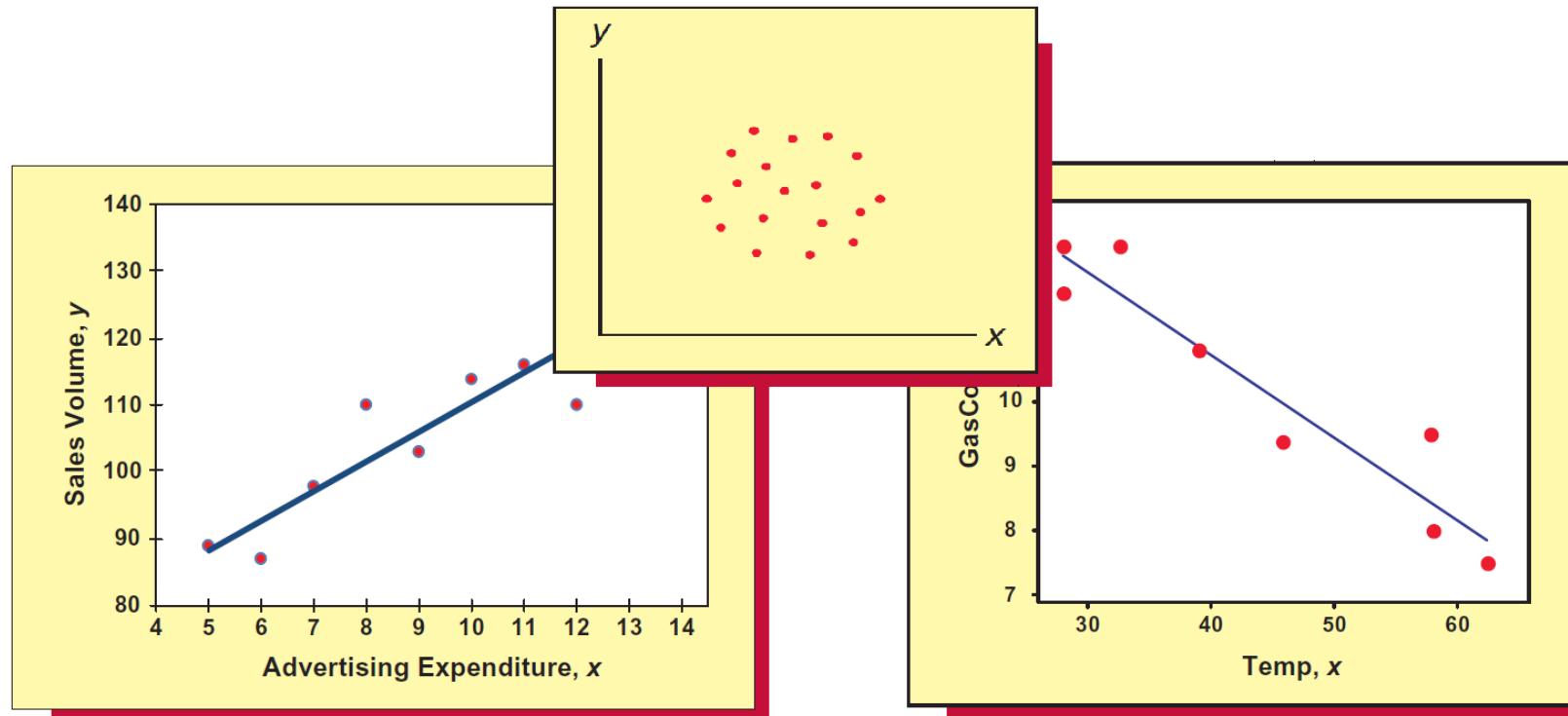
## 2.6 Scatter Plots

- Used to study relationships between two variables (numerical )
- Place one variable on the x-axis
- Place a second variable on the y-axis
- Place dot on pair coordinates

# Types of Relationships

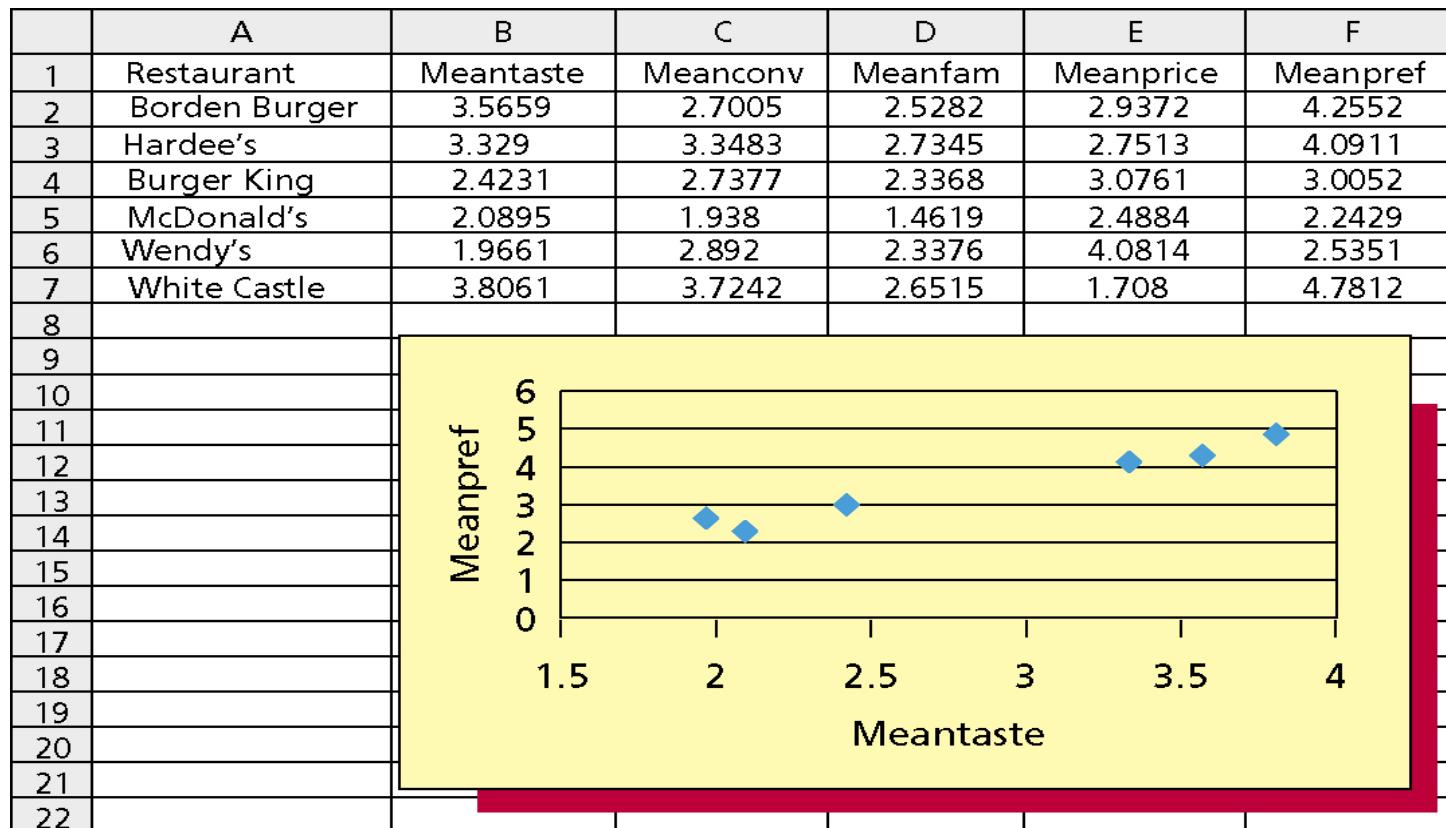
- **Linear:** A straight line relationship between the two variables
- **Positive:** When one variable goes up, the other variable goes up
- **Negative:** When one variable goes up, the other variable goes down
- **No Linear Relationship:** There is no coordinated linear movement between the two variables

# Scatters Plot Showing Relationships

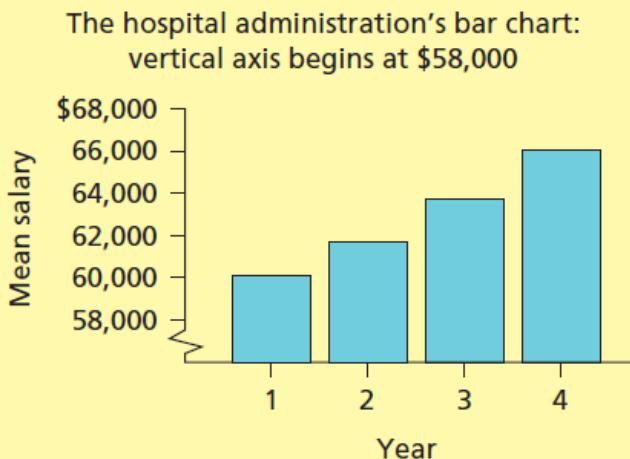


Visualize the data to see patterns, especially “trends”

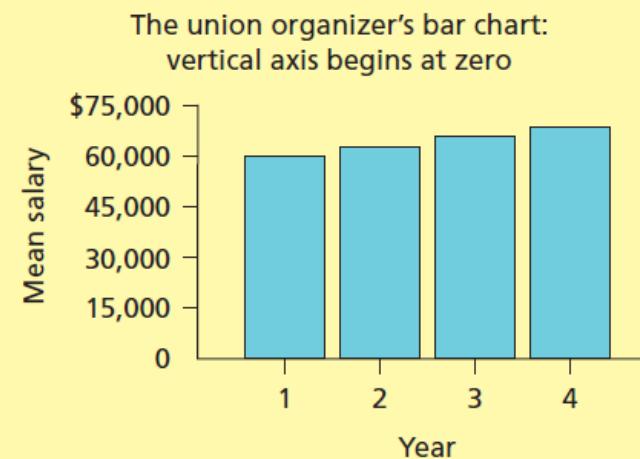
## Restaurant Ratings: Mean Preference vs. Mean Taste



# 2.7 Misleading Graphs and Charts



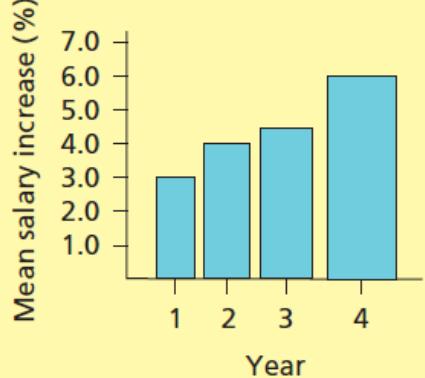
(a) Starting the vertical scale at \$58,000 makes the increases in mean salary look more dramatic.



(b) When the vertical scale starts at zero, the increases in mean salary look less impressive.

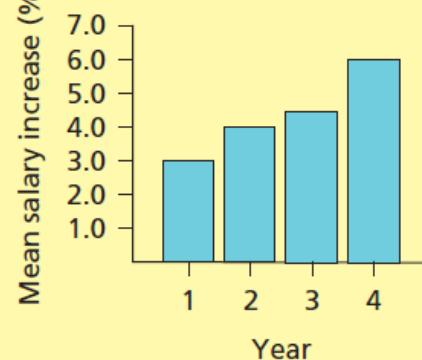
# Misleading Graphs and Charts: Horizontal Scale Effects

The hospital administration's bar chart:  
unequal bar widths



(a) Making the width of the bars proportional to the heights of the bars makes the improvements in the mean salary look more dramatic.

The union organizer's  
bar chart: equal bar widths



(b) The improvements in the mean salary increases look less impressive when the widths of the bars are the same.

# Summary

## Chapter Summary

We began this chapter by explaining how to summarize qualitative data. We learned that we often summarize this type of data in a table that is called a **frequency distribution**. Such a table gives the **frequency**, **relative frequency**, or **percent frequency** of items that are contained in each of several nonoverlapping classes or categories. We also learned that we can summarize qualitative data in graphical form by using **bar charts** and **pie charts** and that qualitative quality data are often summarized using a special bar chart called a **Pareto chart**. We continued in Section 2.2 by discussing how to graphically portray quantitative data. In particular, we explained how to summarize such data by using frequency distributions and histograms. We saw that a **histogram** can be constructed using frequencies, relative frequencies, or percentages, and that we often construct histograms using statistical software such as MINITAB or the analysis toolpak in Excel. We used histograms to describe the shape of a distribution and we saw that distributions are sometimes **mound shaped and symmetrical**, but that a distribution can also be **skewed (to the right or to the left)**. We also learned that a frequency distribution can

be graphed by using a **frequency polygon** and that a graph of a **cumulative frequency distribution** is called an **ogive**. In Sections 2.3 and 2.4 we showed how to summarize relatively small data sets by using **dot plots** and **stem-and-leaf displays**. These graphics allow us to see all of the measurements in a data set and to (simultaneously) see the shape of the data set's distribution. Next, we learned about how to describe the relationship between two variables. First, in optional Section 2.5 we explained how to construct and interpret a **contingency table**, which classifies data on two dimensions using a table that consists of rows and columns. Then, in optional Section 2.6 we showed how to construct a **scatter plot**. Here, we plot numerical values of one variable on a horizontal axis versus numerical values of another variable on a vertical axis. We saw that we often use such a plot to look at possible straight-line relationships between the variables. Finally, in optional Section 2.7 we learned about misleading graphs and charts. In particular, we pointed out several graphical tricks to watch for. By careful analysis of a graph or chart, one can avoid being misled.

Thank you!

## The number of classes $K$

- Group all of the  $n$  data into  $K$  number of classes
- Sturges rule

$$K = 1 + \frac{\lg n}{\lg 2}$$

$K$  is the smallest whole number for which  $2^K \geq n$

- In Examples 2.2 ,  $n = 65$ 
  - For  $K = 6$ ,  $2^6 = 64, < n$
  - For  $K = 7$ ,  $2^7 = 128, > n$
  - So use  $K = 7$  classes

## Class Length $L$

- Class length  $L$  is the step size from one to the next

$$L = \frac{\text{Largest value} - \text{smallest value}}{K}$$

- In Examples 2.2, The Payment Time Case, the largest value is 29 days and the smallest value is 10 days, so

$$L = \frac{29 - 10 \text{ days}}{7 \text{ classes}} = \frac{19 \text{ days}}{7 \text{ classes}} = 2.7143 \text{ days/class}$$

- Arbitrarily round the class length up to 3 days/class.

# Starting the classes

- The classes start on the smallest data value
  - This is the lower limit of the first class
- The upper limit of the first class is
  - $\text{smallest value} + (L - 1)$ 
    - In the example, the first class starts at 10 days and goes up to 12 days
- The second class starts at the upper limit of the first class + 1 and goes up  $(L - 1)$  more
  - The second class starts at 13 days and goes up to 15 days
- And so on

## Tallies and Frequencies: Example 2.2

Classes (days)	Tally	Frequency
10 to 12		3
13 to 15		4
16 to 18		5
19 to 21		3
22 to 24		3
25 to 27		3
28 to 30		1
		65

Check: All frequencies must sum to  $n$

# Relative Frequency: Example 2.2

Classes (days)	Frequency	Relative Frequency
10 to 12	3	$3/65 = 0.0462$
13 to 15	14	$14/65 = 0.2154$
16 to 18	23	0.3538
19 to 21	12	0.1846
22 to 24	8	0.1231
25 to 27	4	0.0615
28 to 30	$\frac{1}{65}$	<u>0.0154</u>
	65	1.0000

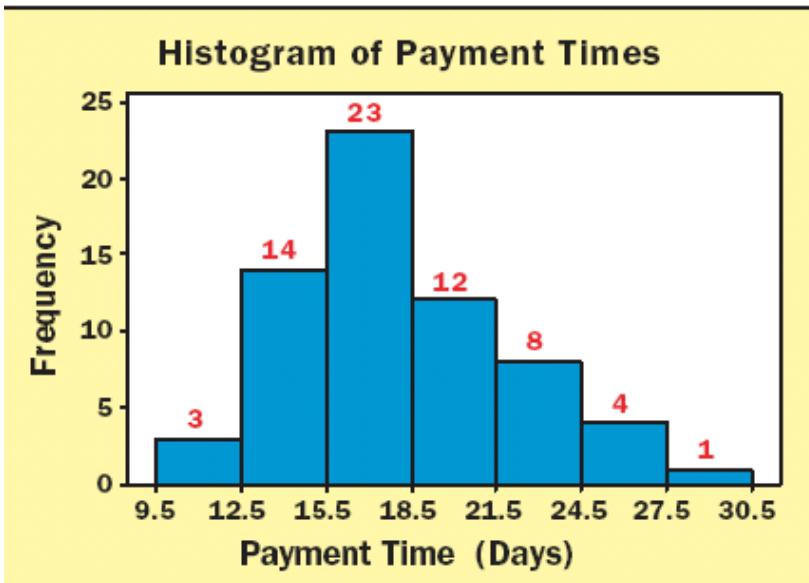
Check: All relative frequencies must sum to 1

<b>Classes</b>	<b>Frequency</b>	<b>Relative Frequency</b>	<b>Boundaries</b>	<b>Midpoint</b>
10 to 12	3	0.0462	9.5, 12.5	11
13 to 15	14	0.2154	12.5, 15.5	14
16 to 18	23	0.3538	15.5, 18.5	17
19 to 21	12	0.1846	18.5, 21.5	20
22 to 24	8	0.1231	21.5, 24.5	23
25 to 27	4	0.0615	24.5, 27.5	26
28 to 30	<u>1</u>	<u>0.0154</u>	27.5, 30.5	29
	65	1.0000		

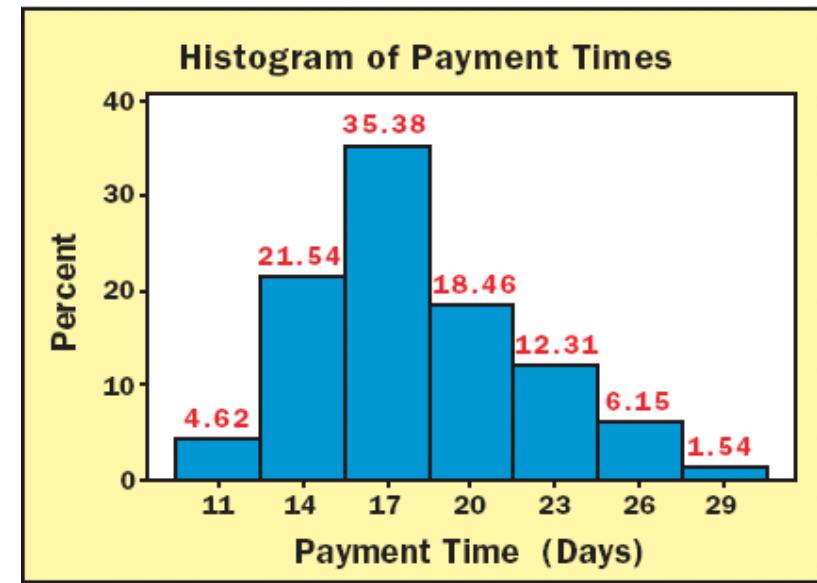
# Histogram

## Example 2.2: The Payment Times Case

Frequency Histogram



Relative Frequency Histogram



As with the earlier stem-and-leaf display, the tail on the right appears to be longer than the tail on the left.

# Histogram

- A graph in which rectangles represent the classes
- The base of the rectangle represents the class length
- The height of the rectangle represents
  - the frequency in a frequency histogram, or
  - the relative frequency in a relative frequency histogram

# Table: A Frequency Distribution and a Relative Frequency Distribution of the 49 Mileages

<b>Classes</b>	<b>Freq.</b>	<b>Relative Freq.</b>	<b>Boundaries</b>	<b>Midpoint</b>
• 29.8-30.3	3	0.0612	29.75, 30.35	30.05
• 30.4-30.9	9	0.1837	30.35, 30.95	30.65
• 31.0-31.5	12	0.2449	30.95, 31.55	31.25
• 31.6-32.1	13	0.2653	31.55, 32.15	31.85
• 32.2-32.7	9	0.1827	32.15, 32.75	32.45
• 32.8-33.3	3	0.0612	32.75, 33.35	33.05

## Histogram

