Chapter 2

Descriptive Statistics: Tabular and Graphical Methods

华鹏 Peng HUA

huapeng@hit.edu.cn

办公室 Office: G620

REVIEW AND OUTLOOK

- **Chapter 1**: Population and Sample

Population	the complete set of all individuals, objects, or elements (people, objects or events) that share common characteristics and are the subject of a statistical analysis.
Sample	A smaller subset of the population. Make inferences about the sampled population.

- **Chapter 2**: Descriptive statistics

Science of describing the important characteristics of a data set

Techniques include tabular, graphical (chapter 2) , and numerical (chapter 3) methods

CHAPTER 2 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)

CHAPTER 2 OUTLINE

2.1 Graphically Summarizing Qualitative Data

2.1 GRAPHICALLY SUMMARIZING QUALITATIVE DATA

- With qualitative data, names identify the different categories
- Qualitative 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

2.1 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	8
	50

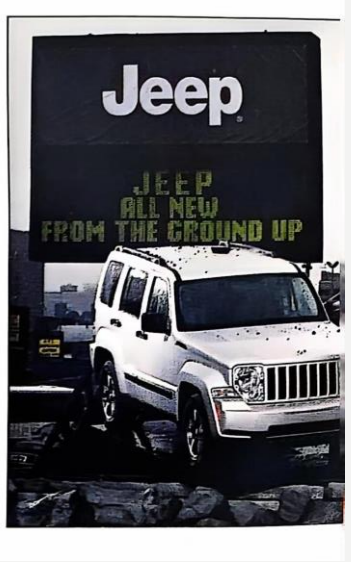
EXAMPLE 2.1: DESCRIBING 2006 JEEP PURCHASING PATTERNS

- Jeep Sales

Jeep sold four different models in 2006:
Commander (C), Grand Cherokee (G), Liberty (L), Wrangler (W)

A dealership must decide how many vehicles of each model should be stocked to meet the customer demand without tying up too much money in unneeded inventory.

251 vehicles sold in a dealership in 2006.

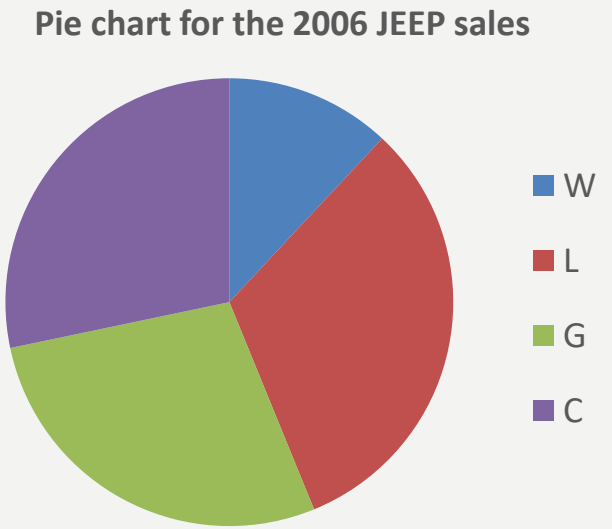


2.1 Relative Frequency

- 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

EXAMPLE 2.1: DESCRIBING 2006 JEEP PURCHASING PATTERNS

	Counts	Percentage (%)
W	30	11.95219124
L	80	31.87250996
G	70	27.88844622
C	71	28.28685259



2.1 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

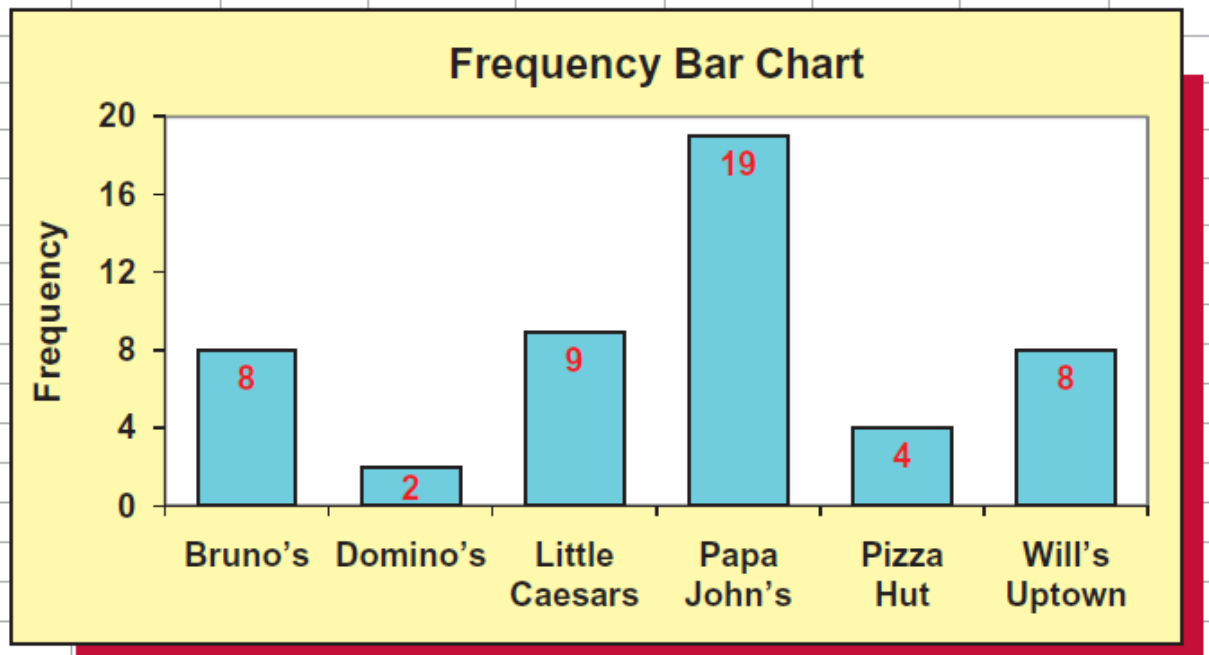
Restaurant	Relative Frequency	Percent Frequency
Bruno's	8/50 = .16	16%
Domino's	.04	4%
Little Caesars	.18	18%
Papa John's	.38	38%
Pizza Hut	.08	8%
Will's Uptown	.16	16%
	1.0	100%

2.1 BAR CHARTS AND PIE CHARTS

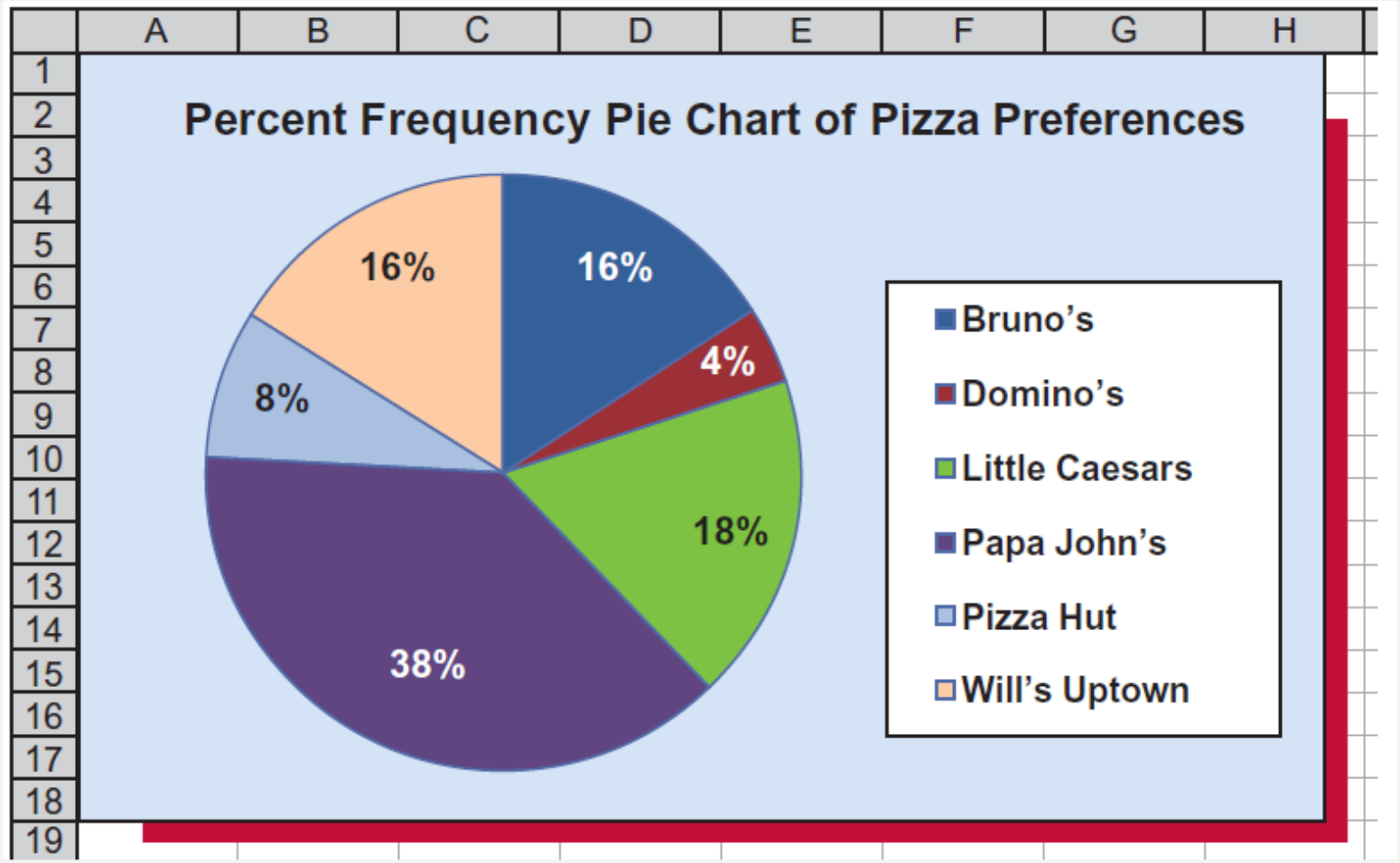
- **Bar chart**: A vertical or horizontal rectangle represents the frequency for each category
  - **Height can be frequency, relative frequency, or percent frequency**
- **Pie chart**: A circle divided into slices where the size of each slice represents its relative frequency or percent frequency

2.1 BAR CHART OF PIZZA PREFERENCE

	A	B
1	Restaurant	Frequency
2	Bruno's	8
3	Domino's	2
4	Little Caesars	9
5	Papa John's	19
6	Pizza Hut	4
7	Will's Uptown	8



2.1 PIE CHART OF PIZZA PREFERENCE

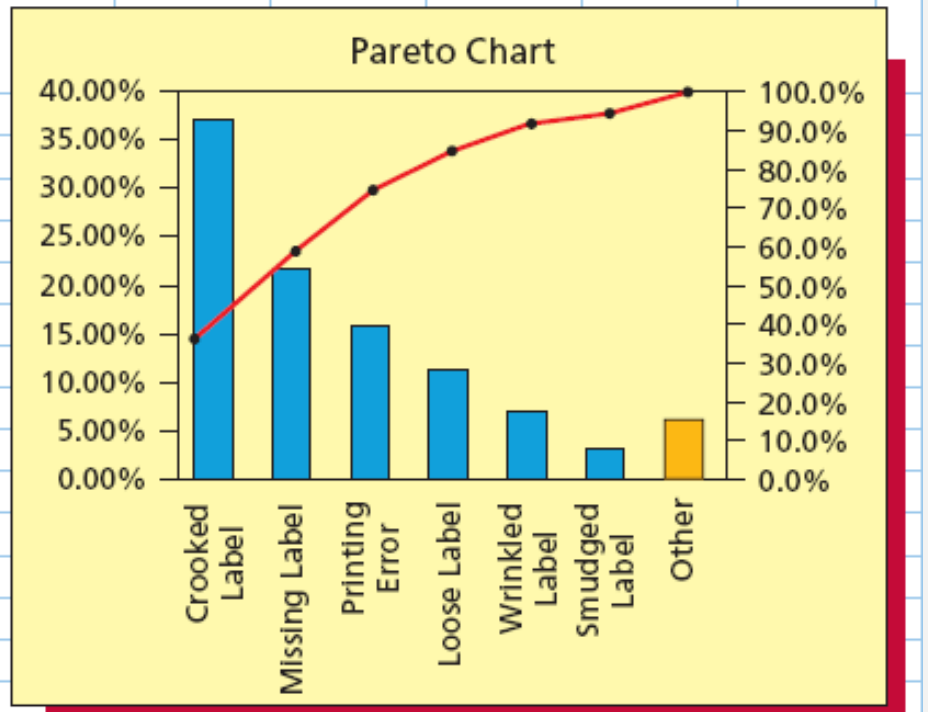


The Pareto chart (Optional)

- 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" oftenapplies to defects. That is, only a few defect types account for most of a product's quality problems.
- 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 frequeney of occurrence (or the percentage of occurrence) for each defect or problem. The bars are arranged in decreasing height from left to right.

PARETO CHART

	A	B	C	D
1	Frequency Table			
3	Type of Defect	Freq.	Percent	Cum. Percent
4	Crooked Label	78	36.97%	37.0%
5	Missing Label	45	21.33%	58.3%
6	Printing Error	33	15.64%	73.9%
7	Loose Label	23	10.90%	84.8%
8	Wrinkled Label	14	6.64%	91.5%
9	Smudged Label	6	2.84%	94.3%
10	Other	12	5.69%	100.0%
11	*Total*	211	100%	



CHAPTER OUTLINE

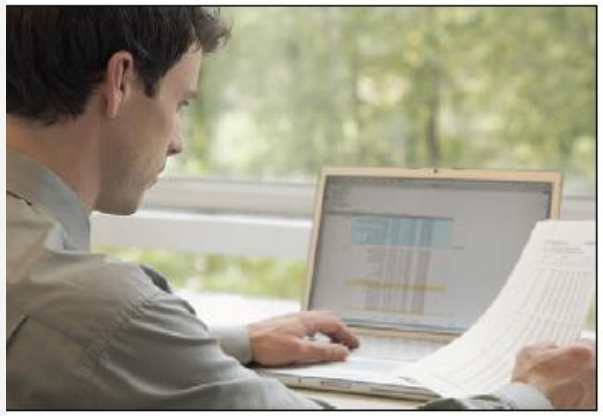
2.2 Graphically Summarizing Quantitative Data

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[3]



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.

EXAMPLE 2.2 THE E-BILLING CASE

Table 2.4

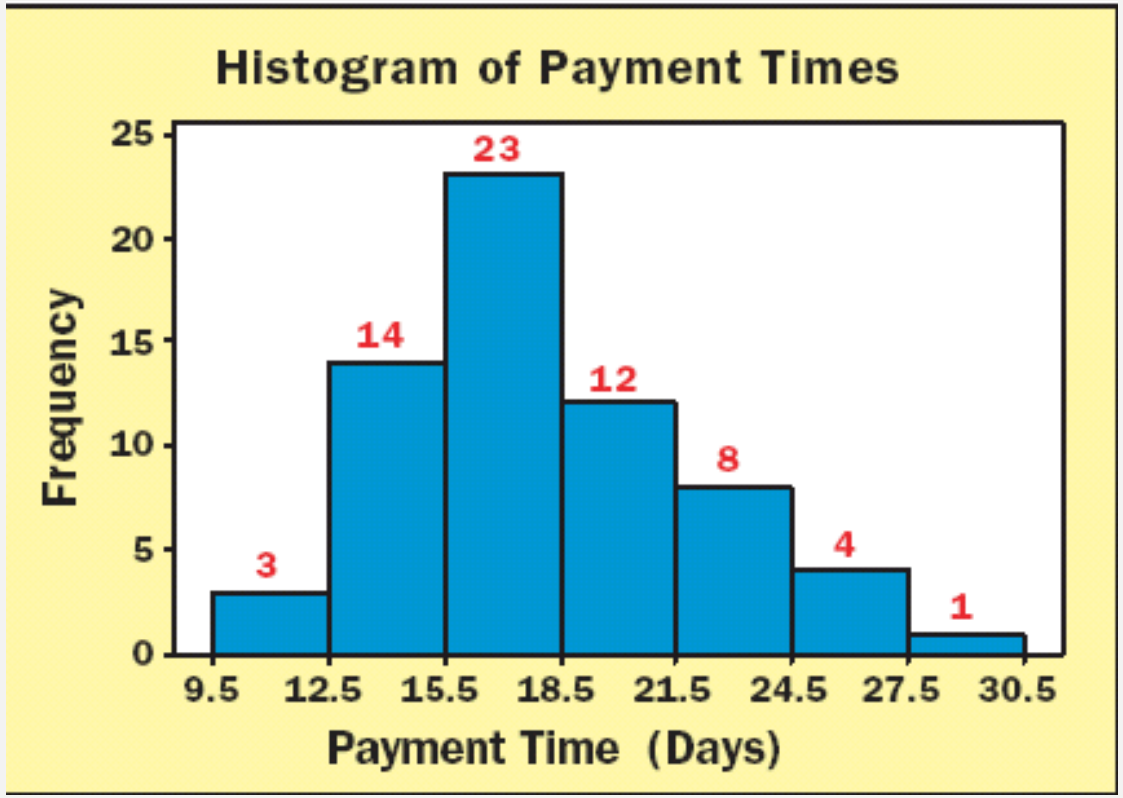
A Sample of Payment Times (in Days) for 65 Randomly Selected Invoices

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	

2.2 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

2.2 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

2.2 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

2.2 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

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

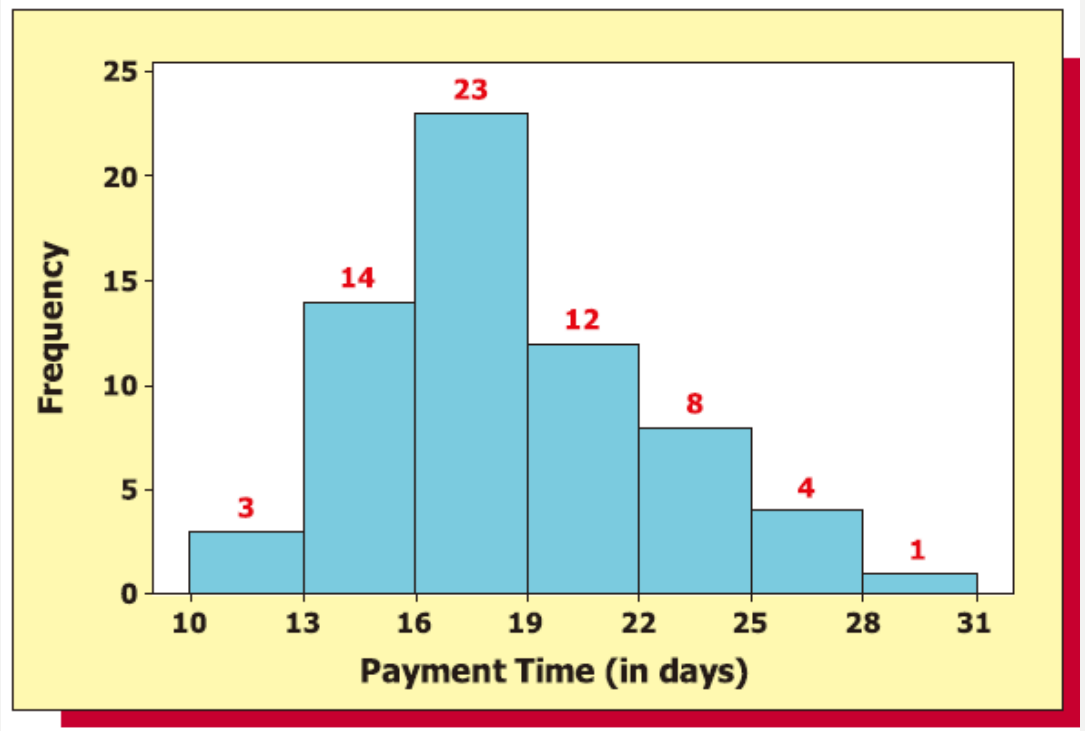
2.2 FORM NON-OVERLAPPING CLASSES OF EQUAL WIDTH

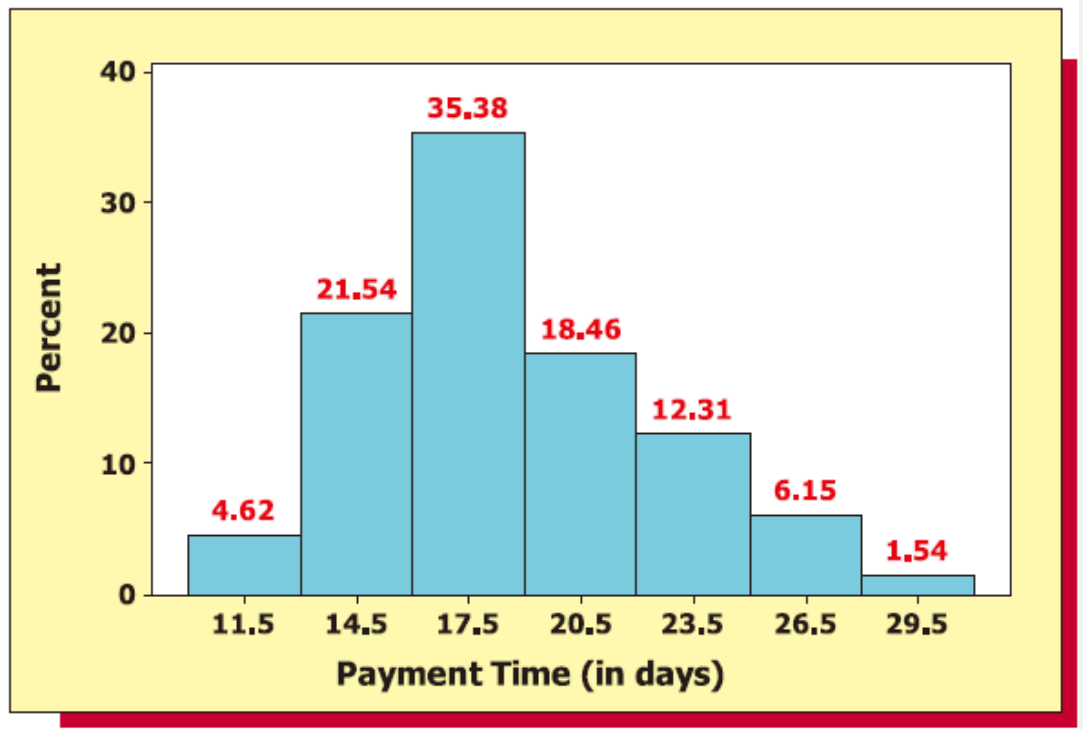
Class 1	10 days and less than 13 days
Class 2	13 days and less than 16 days
Class 3	16 days and less than 19 days
Class 4	19 days and less than 22 days
Class 5	22 days and less than 25 days
Class 6	25 days and less than 28 days
Class 7	28 days and less than 31 days

2.2 TALLY AND COUNT THE NUMBER OF MEASUREMENTS IN EACH CLASS

Class	First 4 Tally Marks	All 65 Tally Marks	Frequency
$10 < 13$		///	3
$13 < 16$		𝍸 𝍸 ////	14
$16 < 19$	//	𝍸 𝍸 𝍸 𝍸 ///	23
$19 < 22$	/	𝍸 𝍸 //	12
$22 < 25$	/	𝍸 ///	8
$25 < 28$		////	4
$28 < 31$		/	1

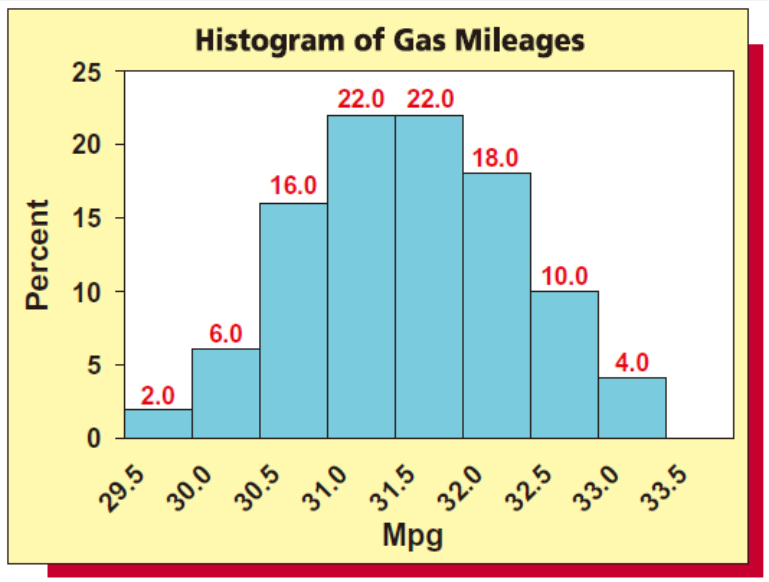
2.2 PLOT HISTOGRAMS





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) *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).

2.2 SKEWED DISTRIBUTIONS

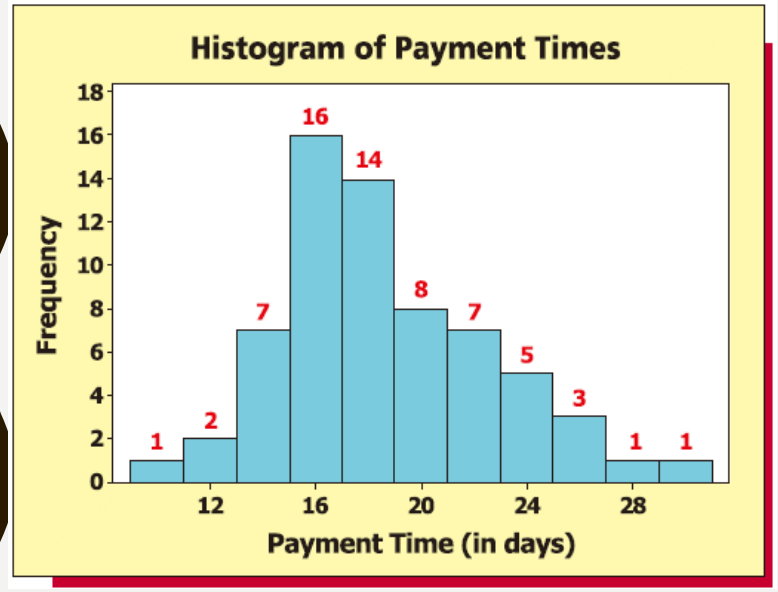


Symmetrical and Mound Shaped

Left and right tails of equal length

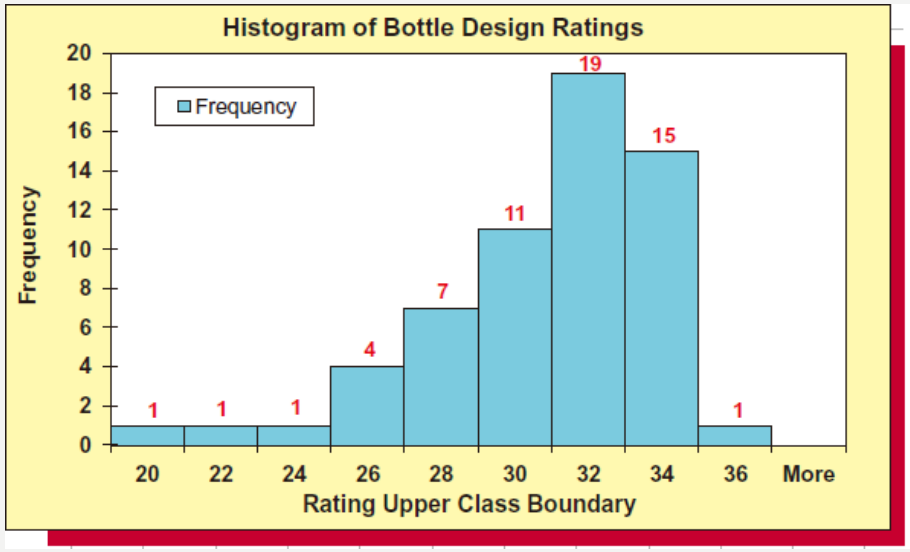
Skewed to the Right

Right tail longer than left



Skewed to the Left

Left tail longer than right

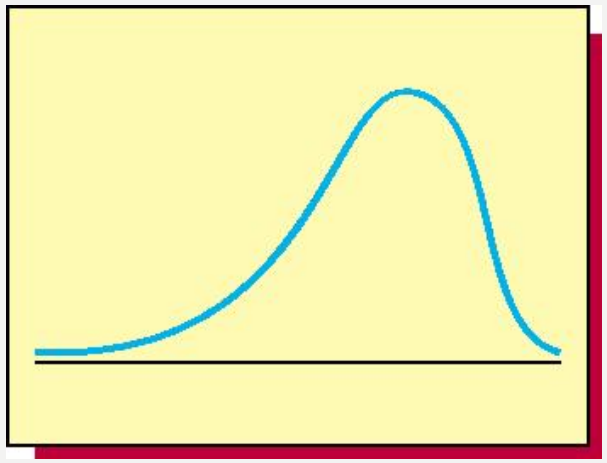


2.2 SKEWNESS

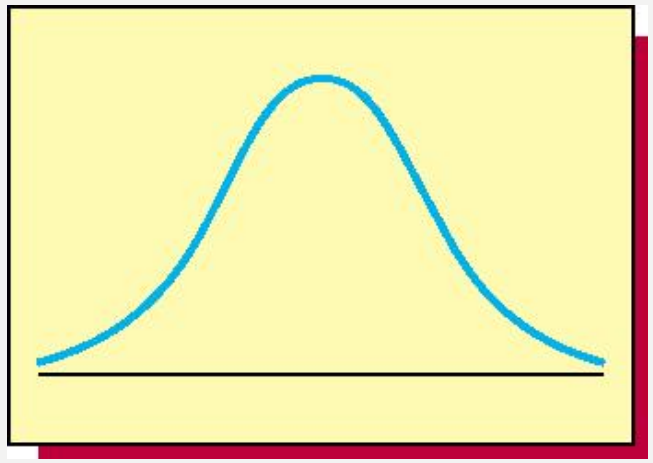
Skewed distributions are not symmetrical about their center. Rather, they are lop-sided with a longer tail on one side or the other.

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

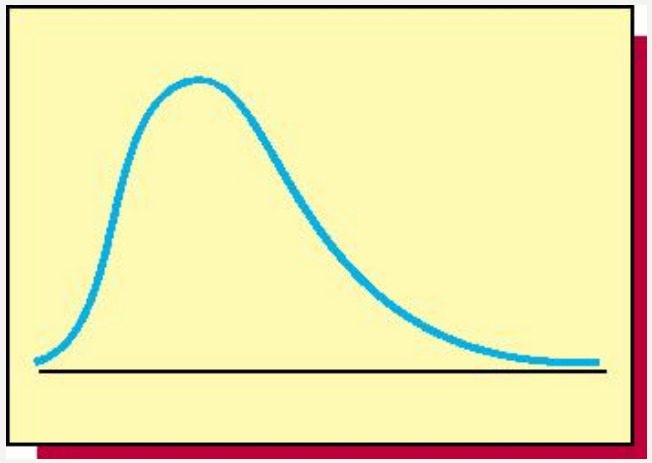
Left Skewed



Symmetric



Right Skewed

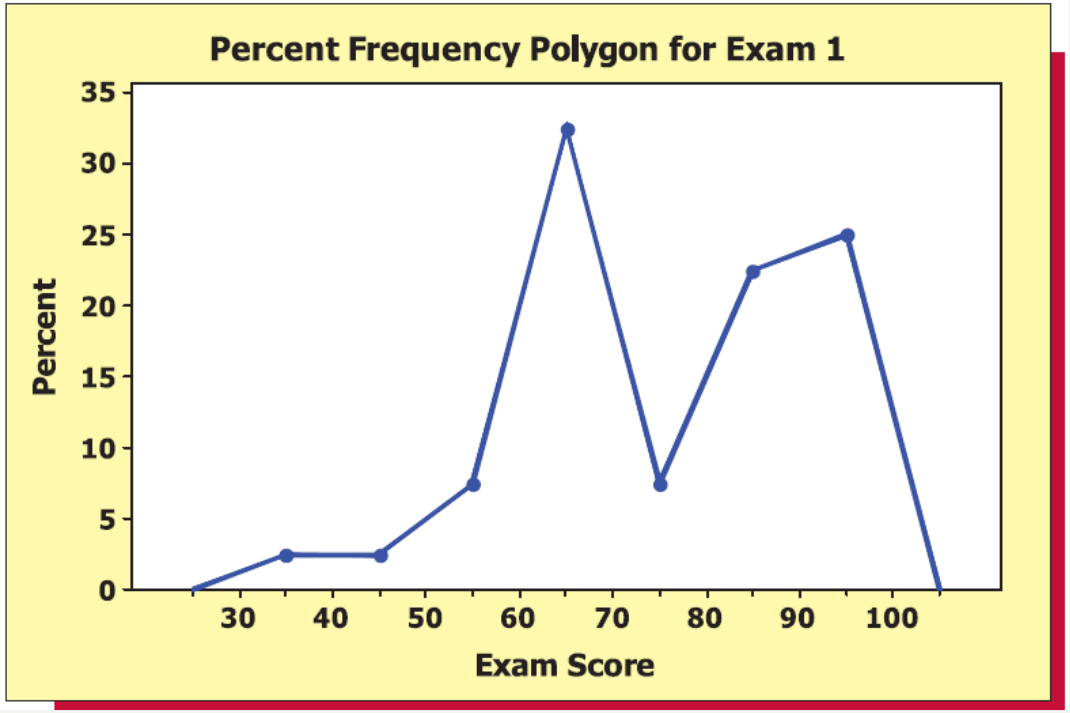


2.2 FREQUENCY POLYGON

TABLE 2.8 Exam Scores for the First Exam Given in a Statistics Class DS FirstExam

32	63	69	85	91
45	64	69	86	92
50	64	72	87	92
56	65	76	87	93
58	66	78	88	93
60	67	81	89	94
61	67	83	90	96
61	68	83	90	98

Figure 2.12



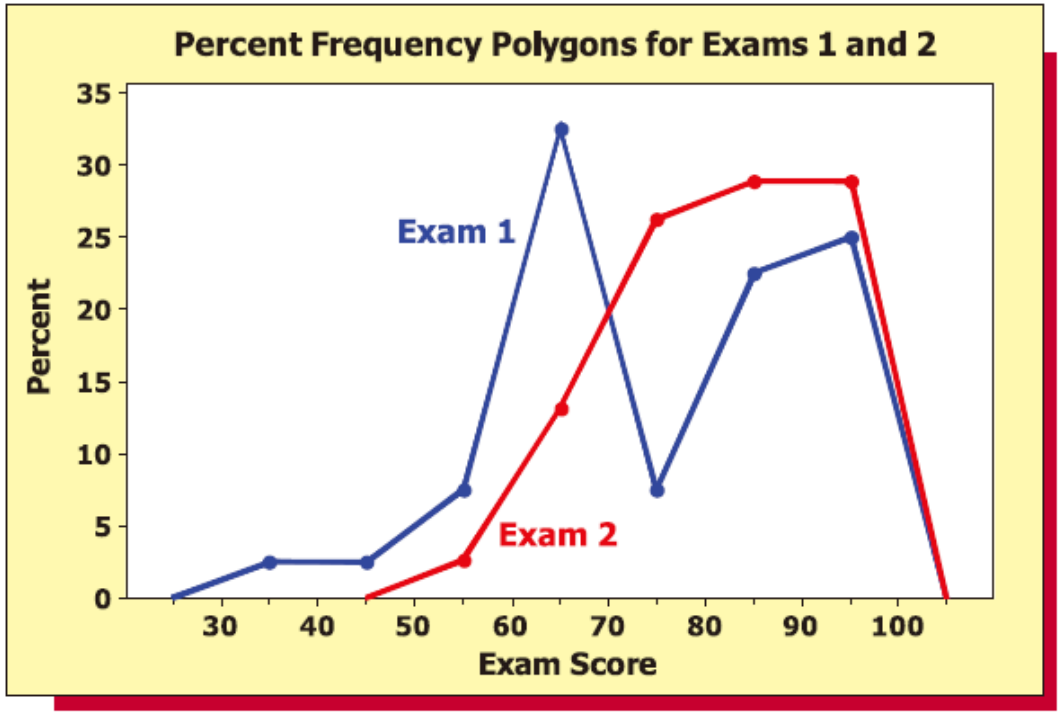
- Plot a point above each class midpoint at a height equal to the frequency of the class—the height can also be the class relative frequency or class percent frequency if so desired.
- Then we connect the points with line segments.

2.2 COMPARING TWO GRADE DISTRIBUTIONS

TABLE 2.9 Exam Scores for the Second Statistics Exam—after a New Attendance Policy DS SecondExam

55	74	80	87	93
62	74	82	88	94
63	74	83	89	94
66	75	84	90	95
67	76	85	91	97
67	77	86	91	99
71	77	86	92	
73	78	87	93	

FIGURE 2.13 Percent Frequency Polygons of the Scores on the First Two Exams in a Statistics Course



- A concentration of scores in the 85 to 95 range and another concentration of scores around 65 for Exam 1.
- After Exam 1, the instructor established an attendance policy that any student who missed a class was to be dropped from the course.
- The polygon for the second exam is single peaked—the attendance policy eliminated the concentration of scores in the 60s, although the scores are still somewhat skewed to the left.

2.2 CUMULATIVE DISTRIBUTIONS AND OGIVES

Table 2.10

A Frequency Distribution, Cumulative Frequency Distribution, Cumulative Relative Frequency Distribution, and Cumulative Percent Frequency Distribution for the Payment Time Data

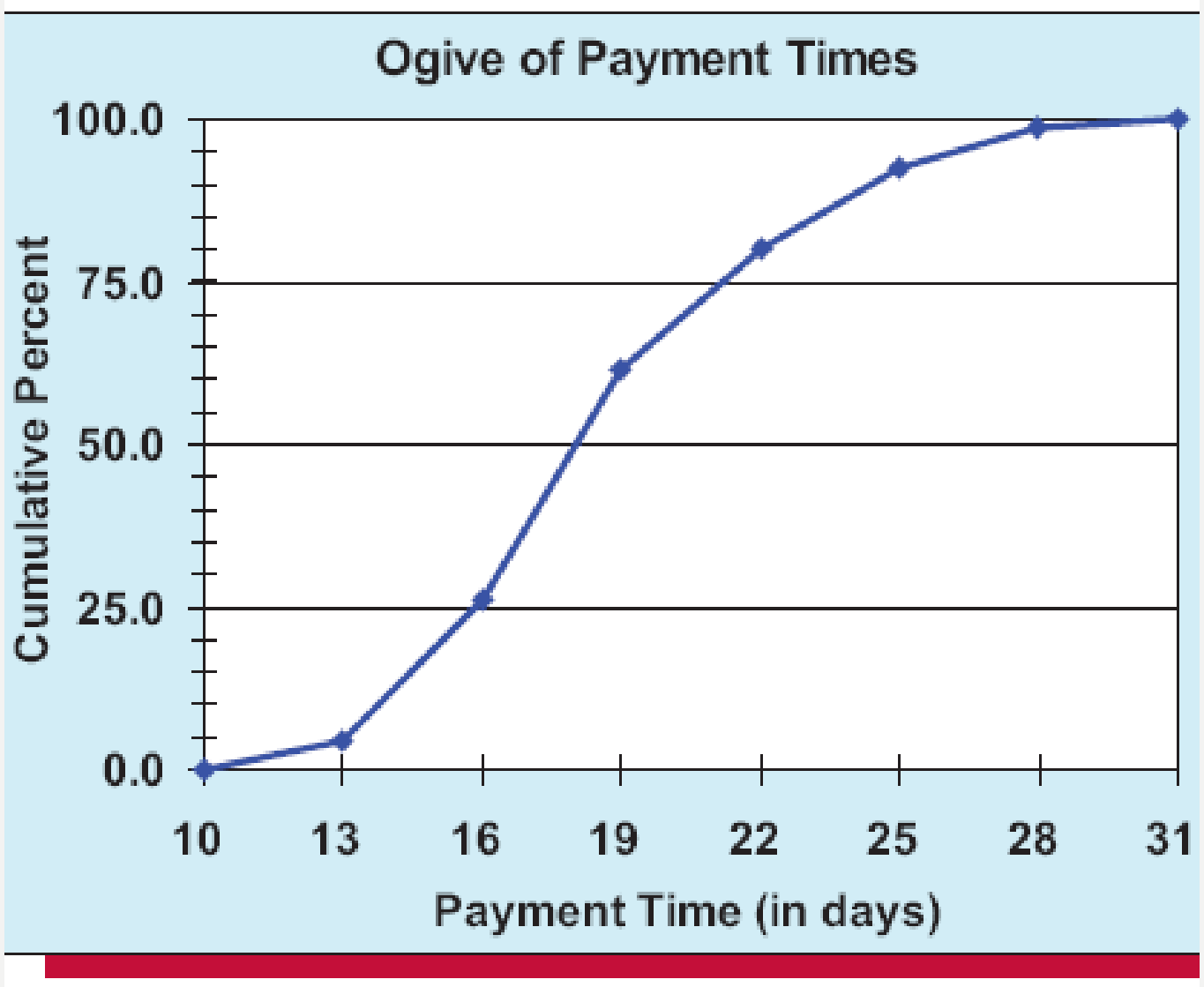
(1) Class	(2) Frequency	(3) Cumulative Frequency	(4) Cumulative Relative Frequency	(5) Cumulative Percent Frequency
$10 < 13$	3	3	3/65 = .0462	4.62%
$13 < 16$	14	17	17/65 = .2615	26.15
$16 < 19$	23	40	.6154	61.54
$19 < 22$	12	52	.8000	80.00
$22 < 25$	8	60	.9231	92.31
$25 < 28$	4	64	.9846	98.46
$28 <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

To construct a cumulative frequency distribution, we record for each class the number of measurements that are less than the upper boundary of the class.

In other words, a running total

2.2 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

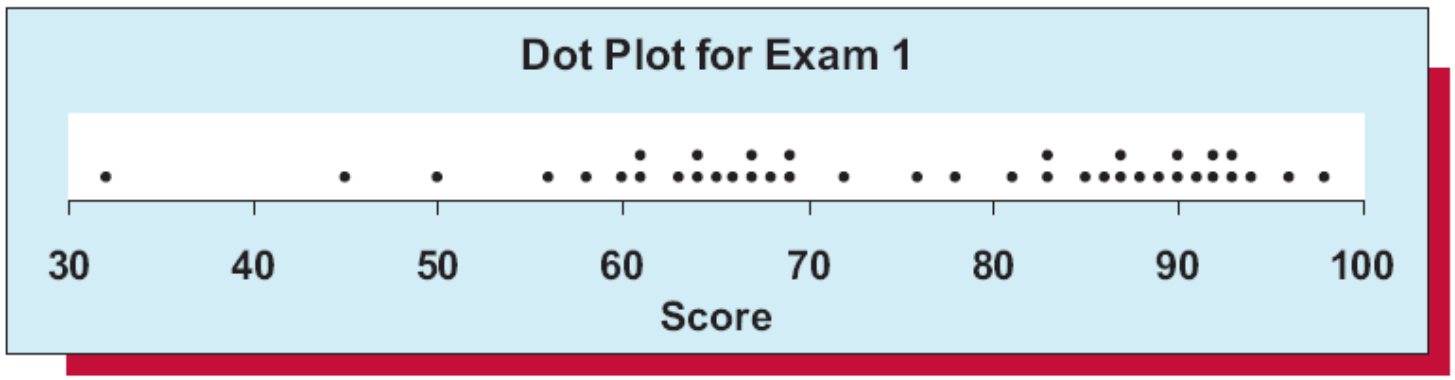


CHAPTER OUTLINE

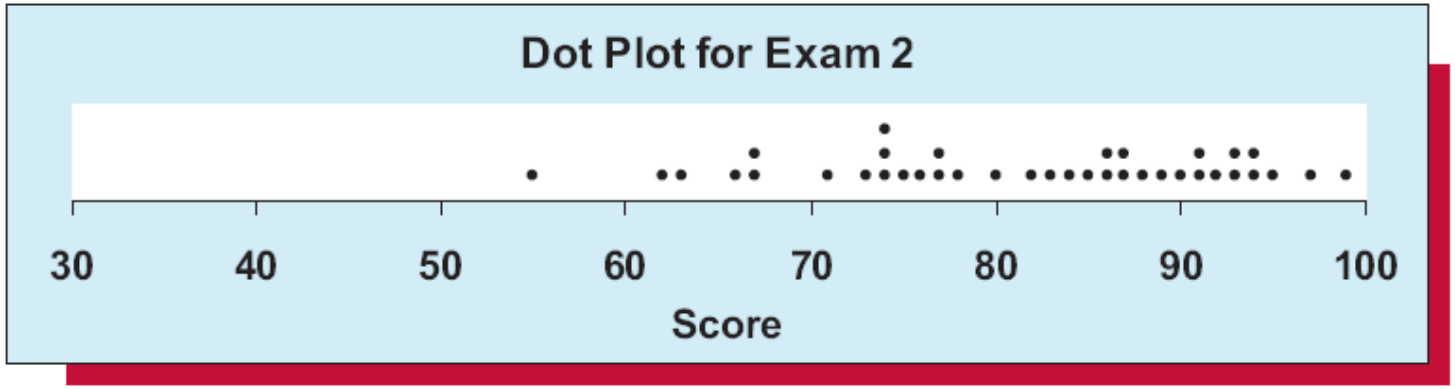
2.3 Dot Plots

2.3 DOT PLOTS

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



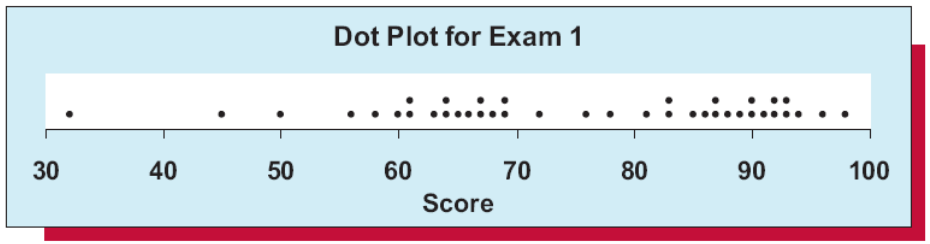
(b) Dot Plot of Scores on Exam 2: After Attendance Policy



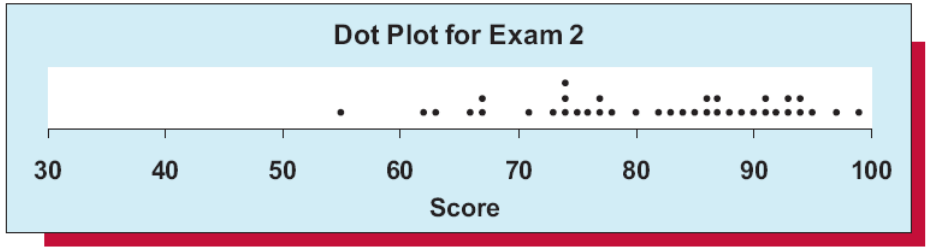
To make a dot plot we draw a horizontal axis that spans the range of the measurements in the data set. We then place dots above the horizontal axis to represent the measurements

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

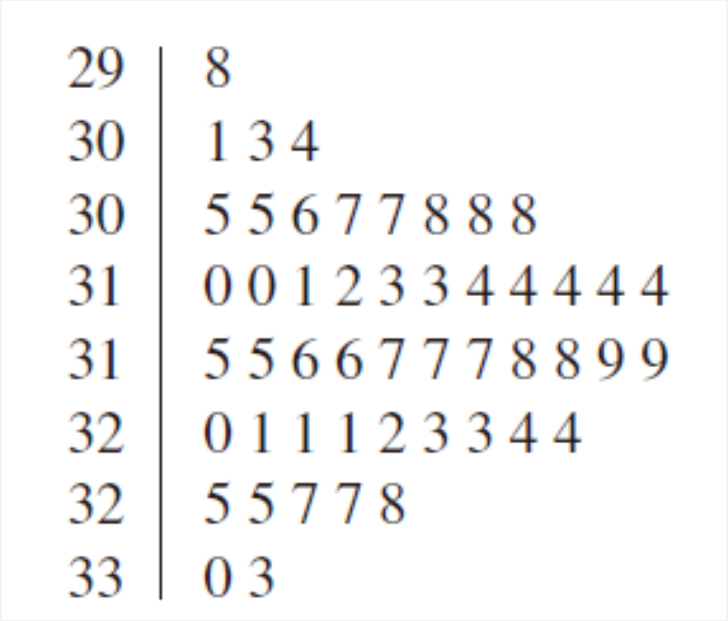


Dot plots are useful for detecting outliers, which are unusually large or small observations that are well separated from the remaining observations

CHAPTER OUTLINE

2.4 Stem-and-leaf Display

2.4 Stem-and-leaf Display



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

The purpose of a stem-and-leaf display is

- to see all of the measurements in the data set
- to see the shape of the data set's distribution
- to detect outliers, which are unusually large or small observations that are well separated from the remaining observations.
- best for small to moderately sized data distributions

2.4 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.

2.4 Example: 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.

2.4 Example: The Car Mileage Case

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

Step 1: Form The Stem

29	
30	
31	
32	
33	

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

Step 11: Form The Leaves

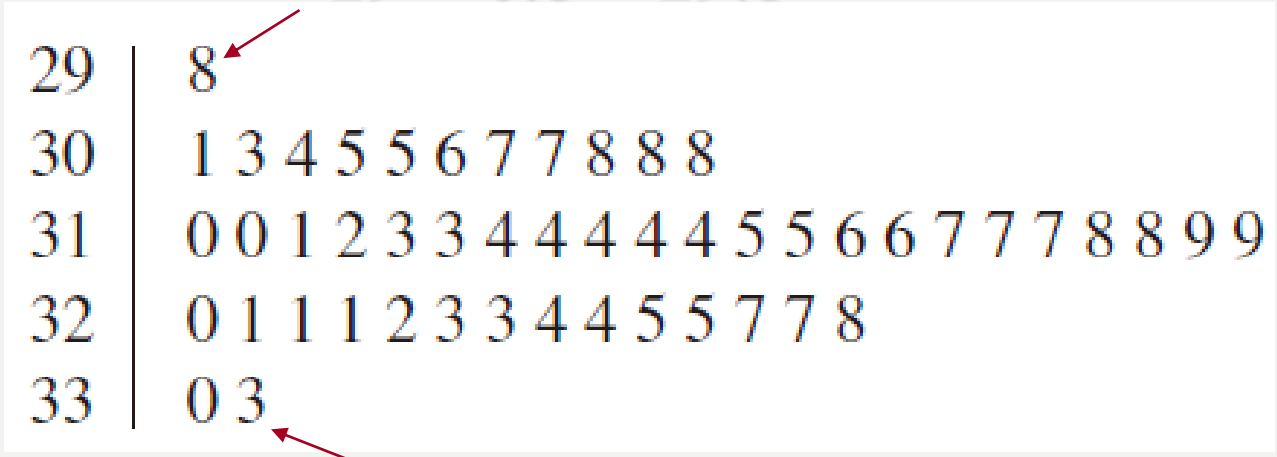
29	
30	8 1
31	7
32	
33	

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

2.4 Example: The Car Mileage Case

The stem-and-leaf display of car mileages:

29 + 0.8 = 29.8



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 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

33 + 0.3 = 33.3

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

2.4 Example: The Car Mileage Case

Stem-and-leaf display I

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 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

Another display of the same data using more classes

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

- Same data used for the two displays.
- Diffirent classification and displays give readers different impressions.
  E.g., the second display is more symmetrical. But not exactly a mirror reflection. Later, we will call this a slightly "left-skewed" distribution

2.4 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	

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

The leftmost column of numbers are the numbers are the cumulative 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

Payment Times

amounts of values;	Stem	Leaves	
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	Longer tail
19	20	000	
16	21	0000	
13	22	000	
10	23	00	
8	24	000	
5	25	00	
3	26	0	
2	27	0	
1	28		
1	29	0	

2.4 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 does 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.**

SUMMARY

- This chapter shows you how to summarize qualitative and quantitave data
- 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
- Bar charts and pie charts
- For quantitave data, a histogram can be constructed using frequencies, relative frequencies, or percentages
- Shape of a distribution is sometimes mound shaped and symmetrical, or skewed (to the right or to the left)
- Frequency polygon
- A graph of a cumulative frequency distribution is called an ogive
- Dot plots and stem-and-leaf displays (see all of the measurements in a data set and to (simultaneously) see the shape of the data set's distribution.)

THANK YOU!

CHAPTER 2 OUTLINE

2.5 Contingency Tables (Optional)

2.6 Scatter Plots (Optional)

2.7 Misleading Graphs and Charts (Optional)

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 DS Invest

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

	Level of Satisfaction			
Fund Type	High	Medium	Low	Total
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

	Level of Satisfaction			
Fund Type	High	Medium	Low	Total
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

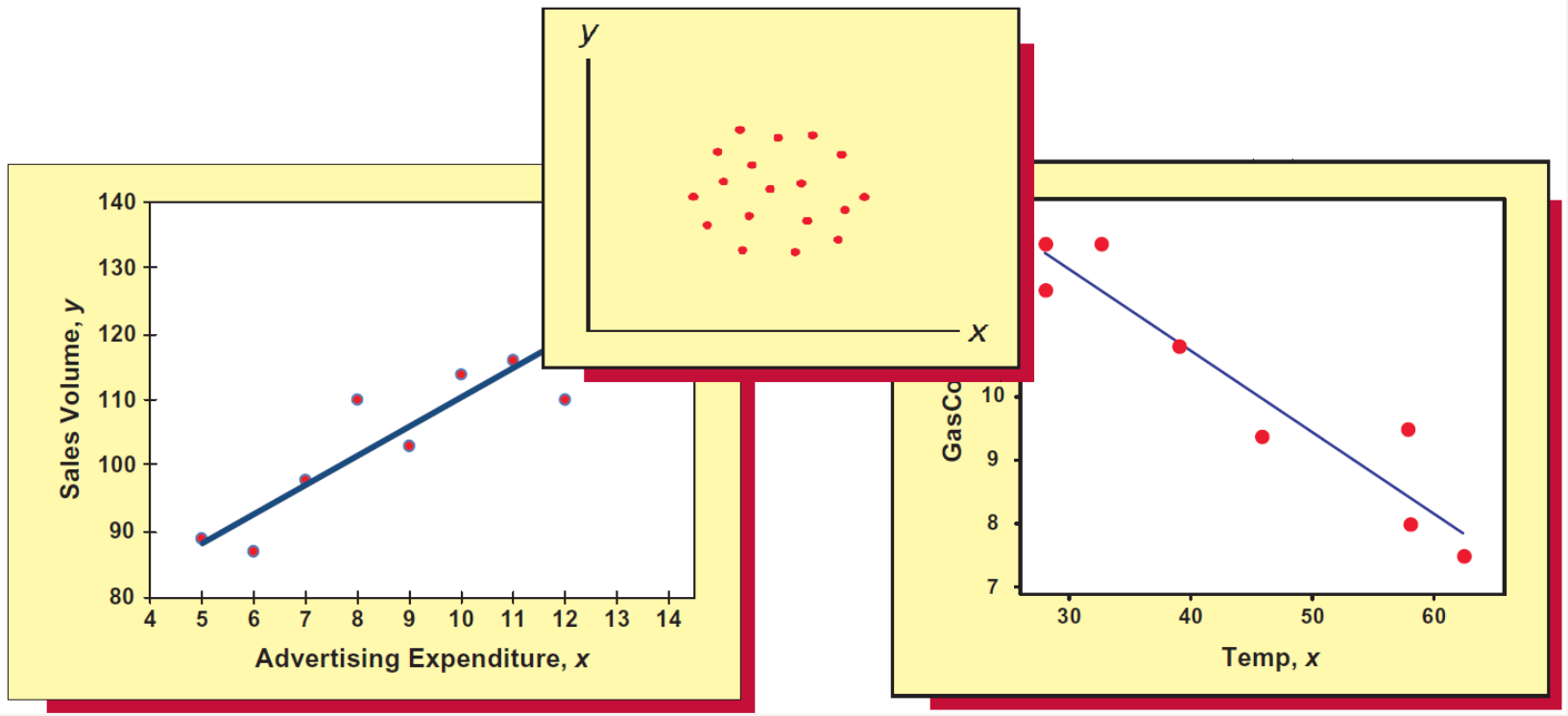
2.6 SCATTER PLOTS

- 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

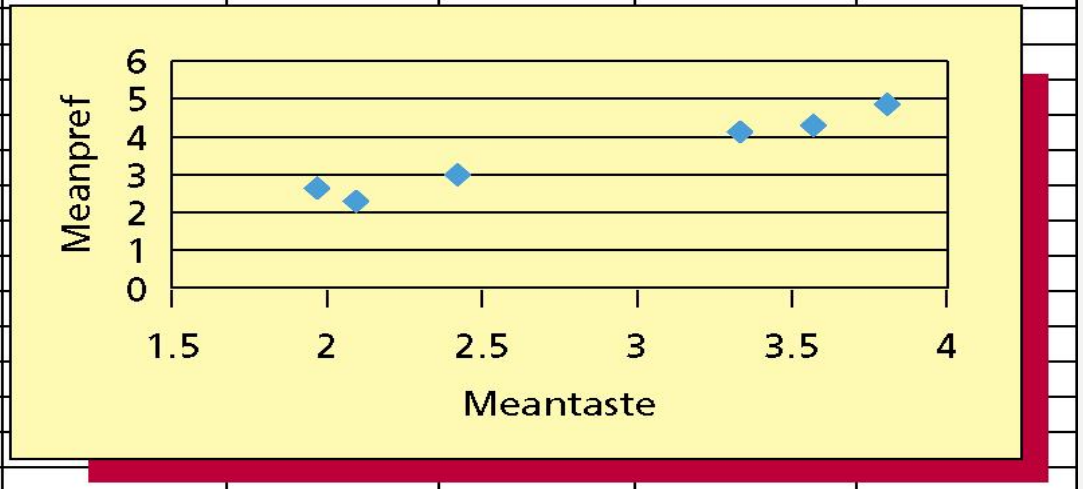
SCATTERS PLOT SHOWING RELATIONSHIPS



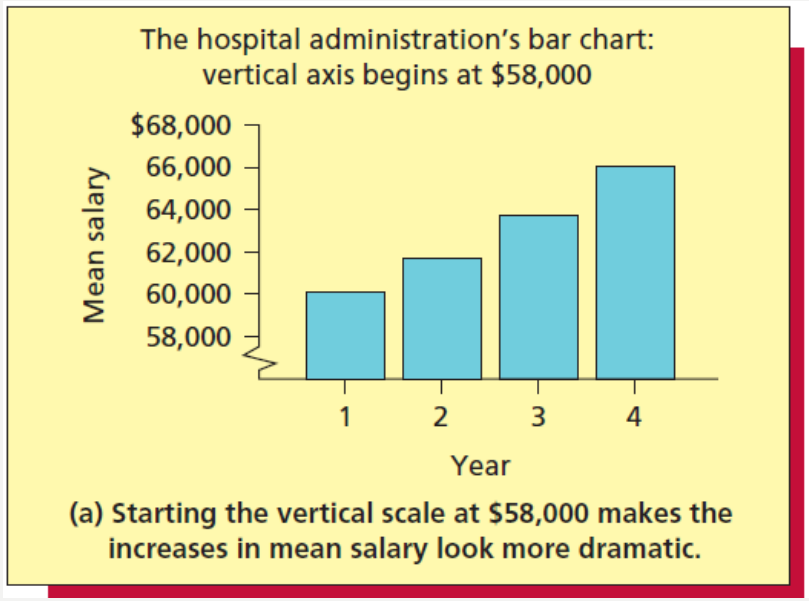
Visualize the data to see patterns, especially “trends”

Restaurant Ratings: Mean Preference vs. Mean Taste

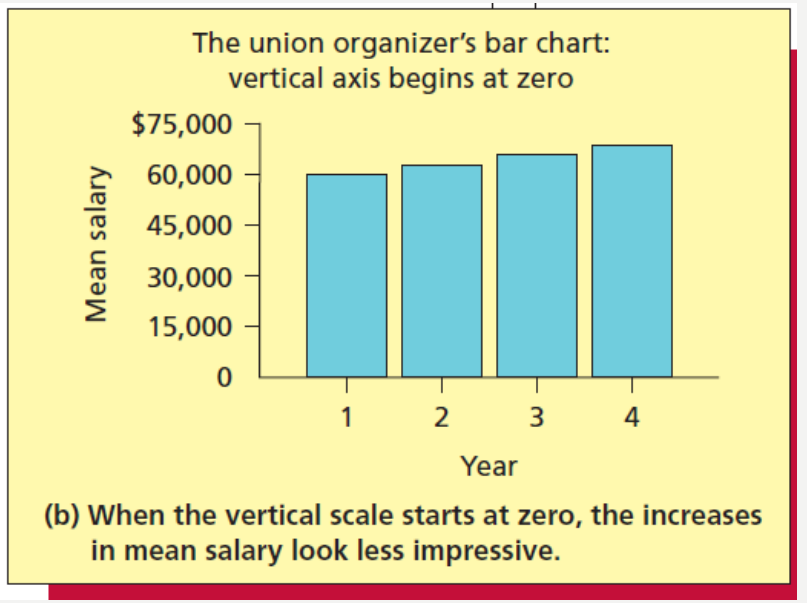
	A	B	C	D	E	F
1	Restaurant	Meantaste	Meanconv	Meanfam	Meanprice	Meanpref
2	Borden Burger	3.5659	2.7005	2.5282	2.9372	4.2552
3	Hardee’s	3.329	3.3483	2.7345	2.7513	4.0911
4	Burger King	2.4231	2.7377	2.3368	3.0761	3.0052
5	McDonald’s	2.0895	1.938	1.4619	2.4884	2.2429
6	Wendy’s	1.9661	2.892	2.3376	4.0814	2.5351
7	White Castle	3.8061	3.7242	2.6515	1.708	4.7812
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						



2.7 MISLEADING GRAPHS AND CHARTS



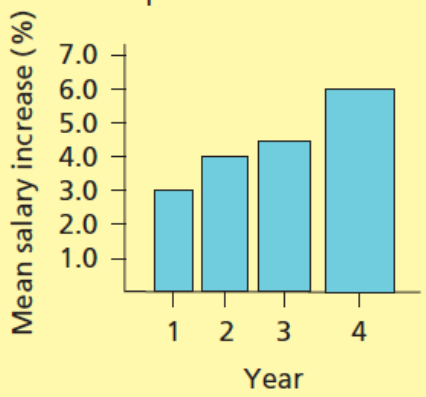
(a) Starting the vertical scale at $58,000 makes the increases in mean salary look more dramatic.



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

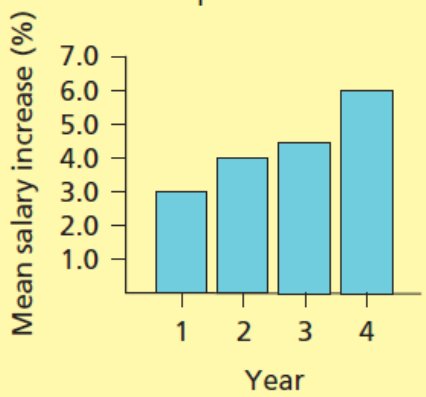
MISLEADING GRAPHS AND CHARTS: HORIZONTAL SCALE EFFECTS

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

$$K = 1 + \frac{lgn}{lg2}$$

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

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

- 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
- 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	𝍸	3
13 to 15	𝍸 𝍸 ||||	14
16 to 18	𝍸 𝍸 𝍸 𝍸 |||	23
19 to 21	𝍸 𝍸 ||	12
22 to 24	𝍸 |||	8
25 to 27	||||	4
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	1	0.0154
		65	1.0000

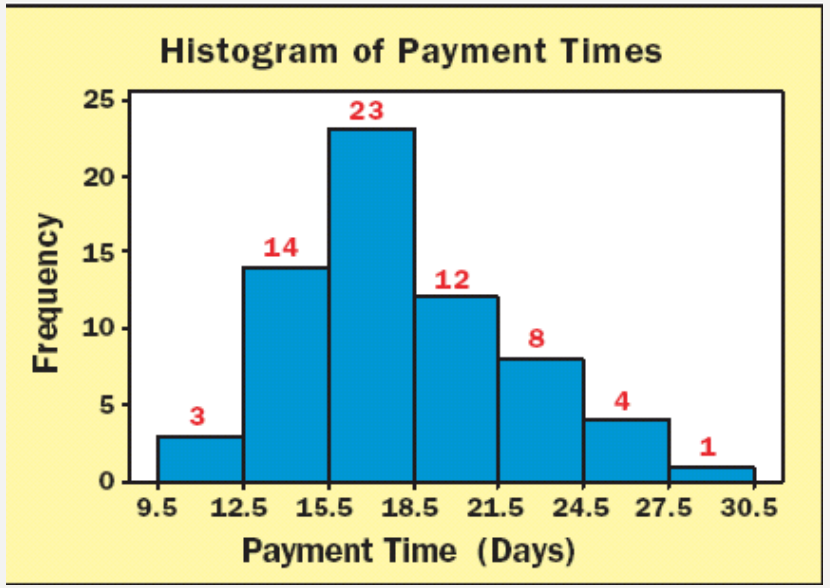
Check: All relative frequencies must sum to 1

Classes	Frequency	Relative Frequency	Boundaries	Midpoint
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	1	0.0154	27.5, 30.5	29
	65	1.0000		

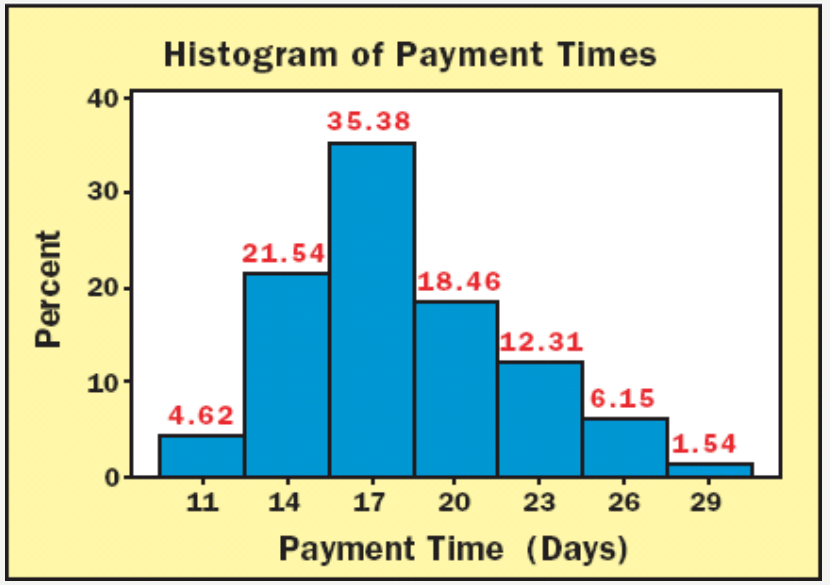
Histogram

Example 2.2: The Payment Times Case

Frequency Histogram



Relative Frequency Histogram



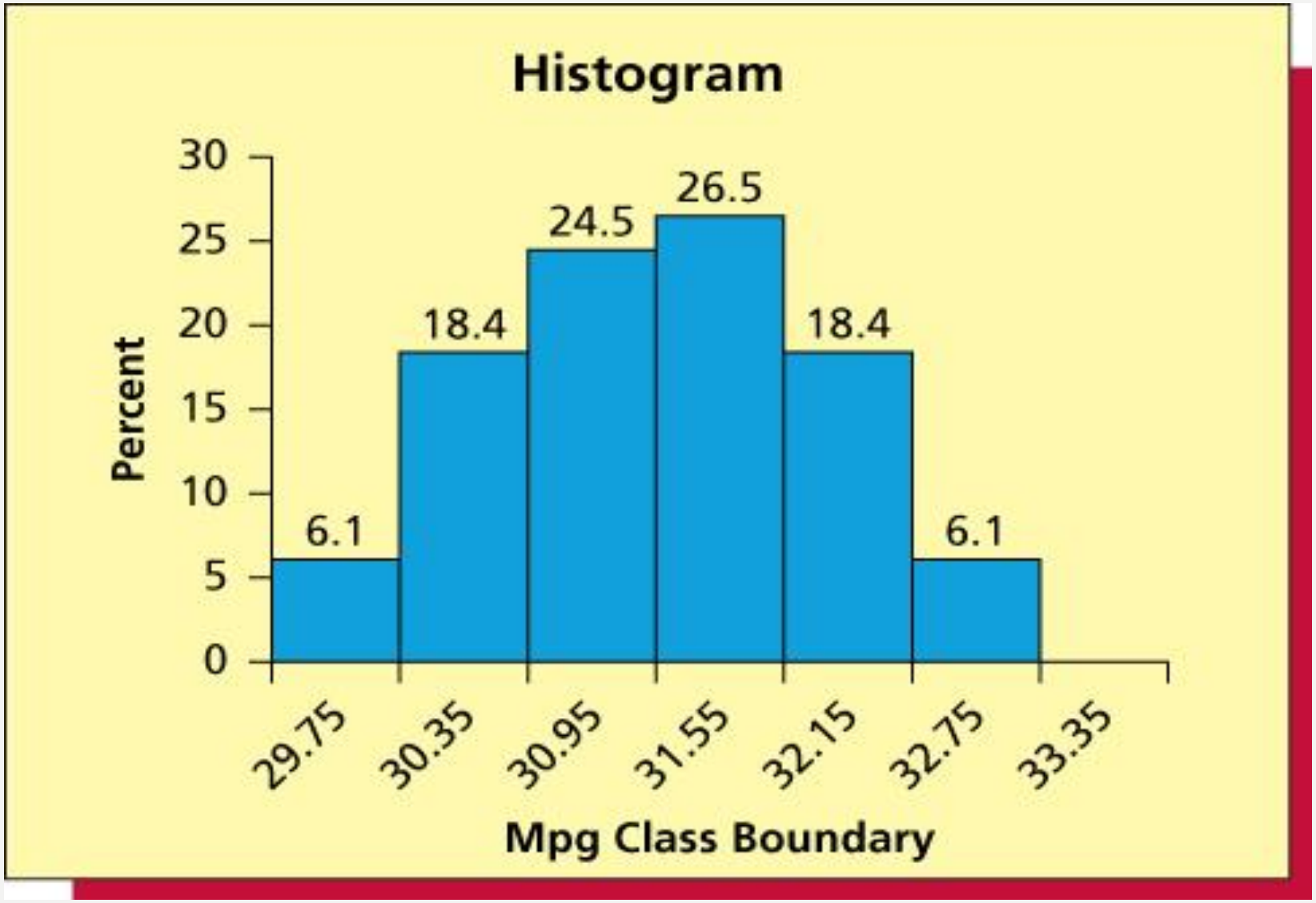
As with the earlier stem-and-leaf display, the tail on the right appears to be longer than the tail on the left.

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

Classes	Freq.	Relative Freq.	Boundaries	Midpoint
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
Percent
30
25
20
15
10
5
0
6.1
18.4
24.5
26.5
18.4
6.1
29.75
30.35
30.95
31.55
32.15
32.75
33.35
Mpg Class Boundary