CAL 282

Probability & Statistics

2013-2014 SPRING TERM

2nd Week

Assist.Prof.Dr. Yılmaz BAYAR

Introduction

- Looking at the numbers presented in a table does not have the same impact as presenting numbers in a well-drawn chart or graph.
- This week we will learn how to construct appropriate graphs to represent data and help you to get your point across to your audience.
- When conducting a statistical study, the researcher must gather data for the particular variable under study.
- Example: If a researcher wishes to study the number of people who were bitten by poisonous snakes in a specific geographic area over the past several years, he or she has to gather the data from various doctors, hospitals, or health departments.
- To describe situations, draw conclusions, or make inferences about events, the researcher must organize the data in some meaningful way.
- The most convenient method of organizing data is to construct a **frequency distribution (frekans dağılımı)**.
- After organizing the data, the researcher must present them so they can be understood by those who will benefit from reading the study.
- The most useful method of presenting the data is by constructing **statistical charts (grafik)** and **graphs (grafik)**.

Organizing Data

- **Raw data (ham veri)**, the data are in original form.
- **Frequency distribution** consists of **classes** and their corresponding frequencies.
- Each raw data value is placed into a quantitative or qualitative category called a **class**.
- The frequency of a class then is the number of data values contained in a specific class.
- **Example:** Suppose a researcher wished to do a study on the ages of the top 50 wealthiest people in the world.

49	57	38	73	81
74	59	76	65	69
54	56	69	68	78
65	85	49	69	61
48	81	68	37	43
78	82	43	64	67
52	56	81	77	79
85	40	85	59	80
60	71	57	61	69
61	83	90	87	74

Organizing Data

- **Example:** Suppose a researcher wished to do a study on the ages of the top 50 wealthiest people in the world.

Class limits	Tally	Frequency
35–41	///	3
42–48	///	3
49–55	////	4
56–62	𝍸 𝍸	10
63–69	𝍸 𝍸	10
70–76	𝍸	5
77–83	𝍸 𝍸	10
84–90	𝍸	5
	Total	50

- It can be stated that the majority of the wealthy people in the study are over 55 years old.

Organizing Data

- There are two types of frequency distributions that are most often used: **categorical frequency distribution** and the **grouped frequency distribution**
- **Categorical frequency distribution** is used for data that can be placed in specific categories, such as nominal- or ordinal-level data.
- For example, data such as political affiliation, religious affiliation, or major field of study would use categorical frequency distributions.
- When the range of the data is large, the data must be grouped into classes that are more than one unit in width, in what is called **a grouped frequency distribution**.

Organizing Data

- **Example:** Twenty-five army inductees were given a blood test to determine their blood type. The data set is given below. Construct a frequency distribution for the data.

A	B	B	AB	O
O	O	B	AB	B
B	B	O	A	O
A	O	O	O	AB
AB	A	O	B	A

A Class	B Tally	C Frequency	D Percent
A	𝍸	5	20
B	𝍸 //	7	28
O	𝍸 ////	9	36
AB	////	4	16
		Total 25	100

- For the sample, more people have type O blood than any other type

Organizing Data

- **Grouped frequency distribution**:
- The **lower class limit** is the smallest data value that can be included in the class.
- The **upper class limit** is the largest data value that can be included in the class.
- **Class boundaries** are used to separate the classes so that there are no gaps in the frequency distribution.
- Class limits should have the same decimal place value as the data, but the class boundaries should have one additional place value and end in a 5.
- If the values in the data set are whole numbers, find the boundaries by subtracting 0.5 from the lower class limit and adding 0.5 the upper class limit.
- If the data are in tenths, find the boundries by subtracting 0.05 from the lower class limit and adding 0.05 the upper class limit.
- **Class width** for a class in a frequency distribution is found by subtracting the lower (or upper) class limit of one class from the lower (or upper) class limit of the next class.
- **The researcher must decide how many classes to use and the width of each class**.

Organizing Data

- **Grouped frequency distribution**:
- There should be between 5 and 20 classes. it is of the utmost importance to have enough classes to present a clear description of the collected data.
- It is preferable but not absolutely necessary that the class width be an odd number. This ensures that the midpoint of each class has the same place value as the data. Class midpoint:

$$X_m = \frac{lower\ boundary + upper\ boundary}{2}$$

- The classes must be mutually exclusive. Mutually exclusive classes have nonoverlapping class limits so that data cannot be placed into two classes.
- The classes must be continuous. Even if there are no values in a class, the class must be included in the frequency distribution.
- The classes must be exhaustive. There should be enough classes to accommodate all the data.
- The classes must be equal in width. This avoids a distorted view of the data.

Organizing Data

- **Example:** A distribution of the number of hours that boat batteries lasted is the following.

Class limits	Class boundaries	Tally	Frequency
24–30	23.5–30.5	///	3
31–37	30.5–37.5	/	1
38–44	37.5–44.5	𝍸	5
45–51	44.5–51.5	𝍸 ////	9
52–58	51.5–58.5	𝍸 /	6
59–65	58.5–65.5	/	1
			25

Organizing Data

- **Example:** These data represent the record high temperatures in degrees Fahrenheit (F) for each of the 50 states. Construct a grouped frequency distribution for the data using 7 classes.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Source: *The World Almanac and Book of Facts.*

Organizing Data

- **Example:** These data represent the record high temperatures in degrees Fahrenheit (F) for each of the 50 states. Construct a grouped frequency distribution for the data using 7 classes.
- Determine the classes.

  Find the highest value and lowest value: H=134 and L=100.

  Find the range: R=highest value-lowest value= H-L=34

  Select the number of classes desired (usually between 5 and 20). In this case, 7 is arbitrarily chosen.

  Find the class width by dividing the range by the number of classes.

$$Width = \frac{R}{Number\ of\ classes} = \frac{34}{7} = 4.9$$

  Round the answer up to the nearest whole number if there is a remainder: 4.9≈ 5.

- Select a starting point for the lowest class limit. This can be the smallest data value or any convenient number less than the smallest data value. **In this case, 100 is used**.
- Add the width to the lowest score taken as the starting point to get the lower limit of the next class. Keep adding until there are 7 classes, as
- shown, 100, 105, 110, etc.

Organizing Data

- **Example:** These data represent the record high temperatures in degrees Fahrenheit (F) for each of the 50 states. Construct a grouped frequency distribution for the data using 7 classes.
- Subtract one unit from the lower limit of the second class to get the upper limit of the first class. Then add the width to each upper limit to get all the upper limits.
- 105-1=104. The first class is 100–104, the second class is 105–109, etc.
- Find the class boundaries by subtracting 0.5 from each lower class limit and adding 0.5 to each upper class limit: 99.5–104.5, 104.5–109.5, etc.
- Tally the data.
- Find the numerical frequencies from the tallies.

Organizing Data

- **Example:** These data represent the record high temperatures in degrees Fahrenheit (F) for each of the 50 states. Construct a grouped frequency distribution for the data using 7 classes.

Class limits	Class boundaries	Tally	Frequency
100–104	99.5–104.5	//	2
105–109	104.5–109.5	𝍸 ///	8
110–114	109.5–114.5	𝍸 𝍸 𝍸 ///	18
115–119	114.5–119.5	𝍸 𝍸 ///	13
120–124	119.5–124.5	𝍸 //	7
125–129	124.5–129.5	/	1
130–134	129.5–134.5	/	1
			$n = \Sigma f = 50$

Organizing Data

- **Cumulative frequency distribution** is a distribution that shows the number of data values less than or equal to a specific value (usually an upper boundary).
- The values are found by adding the frequencies of the classes less than or equal to the upper class boundary of a specific class.
- Cumulative frequencies are used to show how many data values are accumulated up
- to and including a specific class.
- **Example:** These data represent the record high temperatures in degrees Fahrenheit (F) for each of the 50 states.

	Cumulative frequency
Less than 99.5	0
Less than 104.5	2
Less than 109.5	10
Less than 114.5	28
Less than 119.5	41
Less than 124.5	48
Less than 129.5	49
Less than 134.5	50

Organizing Data

- The reasons for constructing a frequency distribution are as follows:

  To organize the data in a meaningful, intelligible way.

  To enable the reader to determine the nature or shape of the distribution.

  To facilitate computational procedures for measures of average and spread.

  To enable the researcher to draw charts and graphs for the presentation of data.

  To enable the reader to make comparisons among different data sets.

Histograms, Frequency Polygons, and Ogives

- The purpose of graphs in statistics is to convey the data to the viewers in pictorial form.
- It is easier for most people to comprehend the meaning of data presented graphically than data presented numerically in tables or frequency distributions.
- The three most commonly used graphs in research are **histogram, frequency polygons** and **cumulative frequency graph** or **ogives**.
- **Histogram (histogram)** is a graph that displays the data by using contiguous vertical bars (unless the frequency of a class is 0) of various heights to represent the frequencies of the classes.
- **Frequency polygon (frekans poligonu)** is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoints of the classes. The frequencies are represented by the heights of the points.
- The frequency polygon and the histogram are two different ways to represent the same data set. The choice of which one to use is left to the discretion of the researcher.
- **Cumulative frequency graph** or **ogive** is the sum of the frequencies accumulated up to the upper boundary of a class in the distribution.

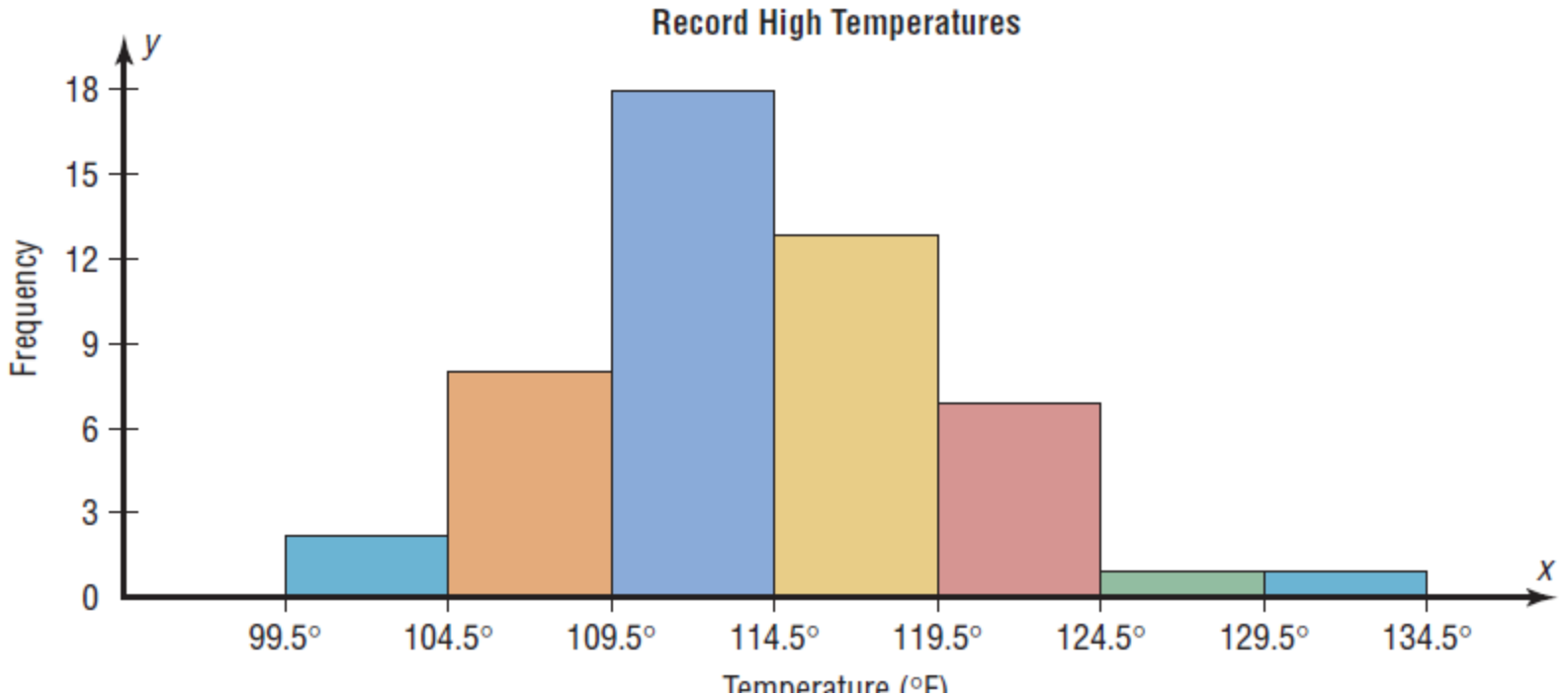
Histograms, Frequency Polygons, and Ogives

- **Ex:** Construct a histogram to represent the data shown for the record high temperatures for each of the 50 states.

Class boundaries	Frequency
99.5–104.5	2
104.5–109.5	8
109.5–114.5	18
114.5–119.5	13
119.5–124.5	7
124.5–129.5	1
129.5–134.5	1

Histograms, Frequency Polygons, and Ogives

- **Ex:** Construct a histogram to represent the data shown for the record high temperatures for each of the 50 states.



Histograms, Frequency Polygons, and Ogives

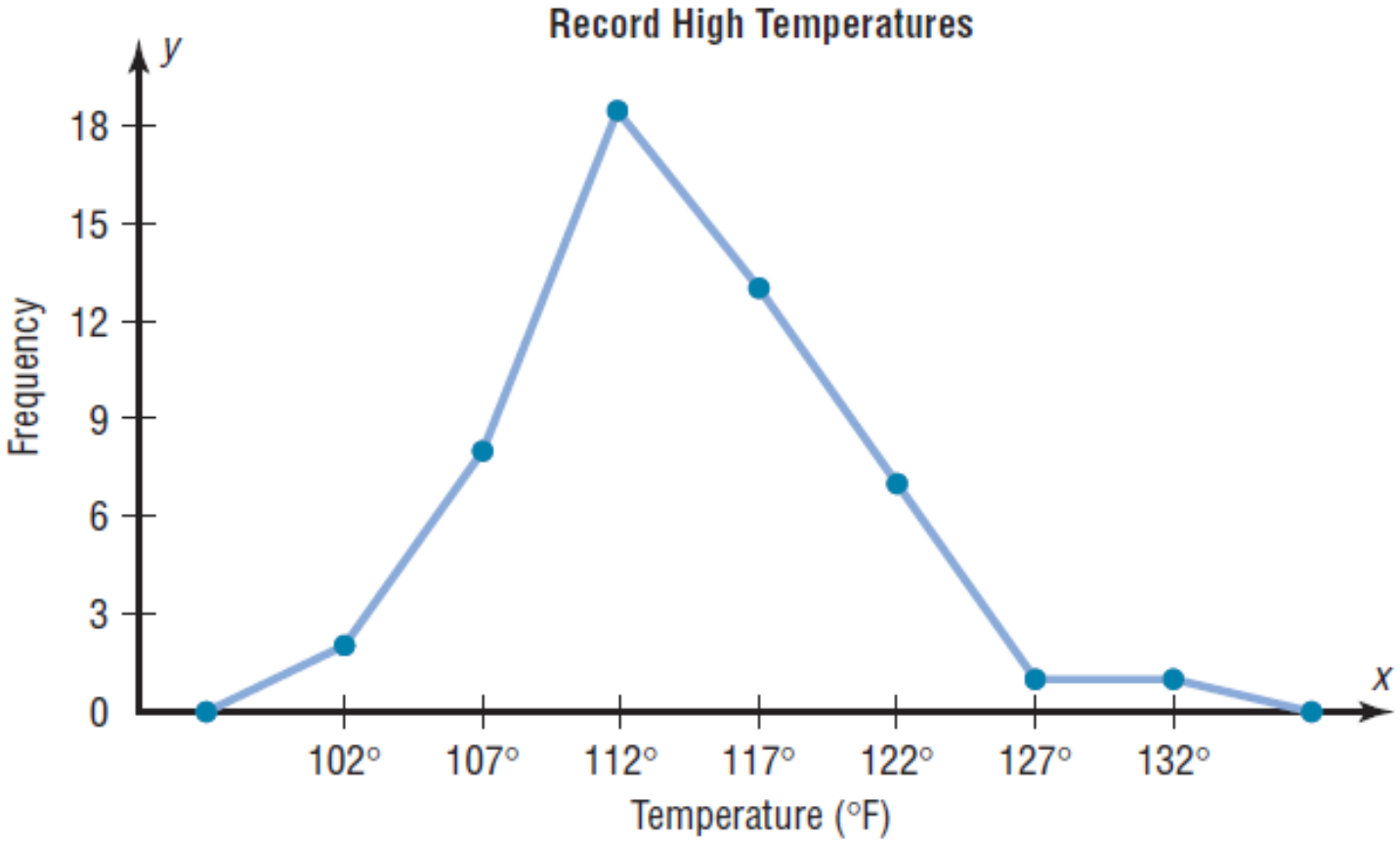
- **Ex:** Construct a histogram to represent the data shown for the record high temperatures for each of the 50 states. **(Frequency polygon)**
- Find the midpoints of each class. Recall that midpoints are found by adding the upper and lower boundaries and dividing by 2:

$$\frac{99.5 + 104.5}{2} = 102$$

Class boundaries	Midpoints	Frequency
99.5–104.5	102	2
104.5–109.5	107	8
109.5–114.5	112	18
114.5–119.5	117	13
119.5–124.5	122	7
124.5–129.5	127	1
129.5–134.5	132	1

Histograms, Frequency Polygons, and Ogives

- **Ex:** Construct a histogram to represent the data shown for the record high temperatures for each of the 50 states. **(Frequency polygon)**



Histograms, Frequency Polygons, and Ogives

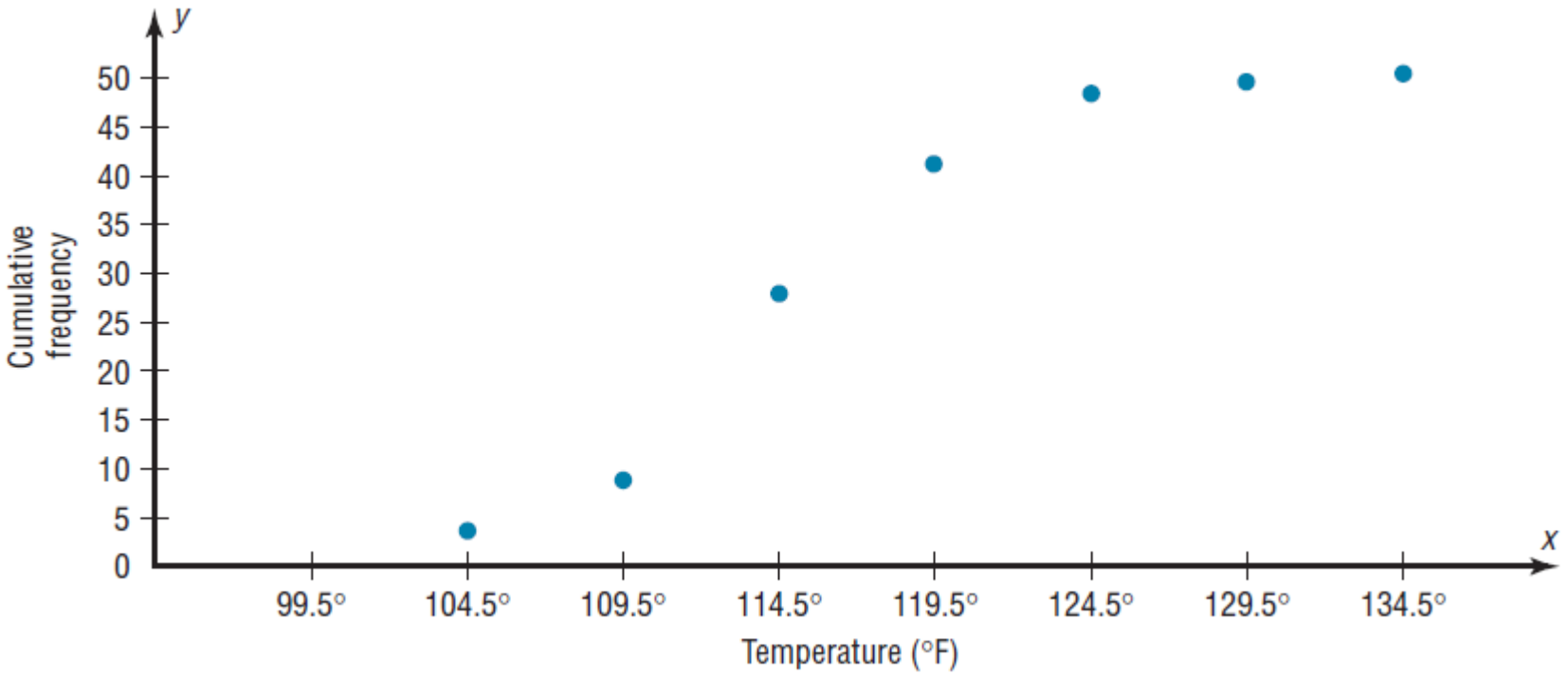
- **Ex:** Construct a histogram to represent the data shown for the record high temperatures for each of the 50 states. **(Ogive)**

	Cumulative frequency
Less than 99.5	0
Less than 104.5	2
Less than 109.5	10
Less than 114.5	28
Less than 119.5	41
Less than 124.5	48
Less than 129.5	49
Less than 134.5	50

Histograms, Frequency Polygons, and Ogives

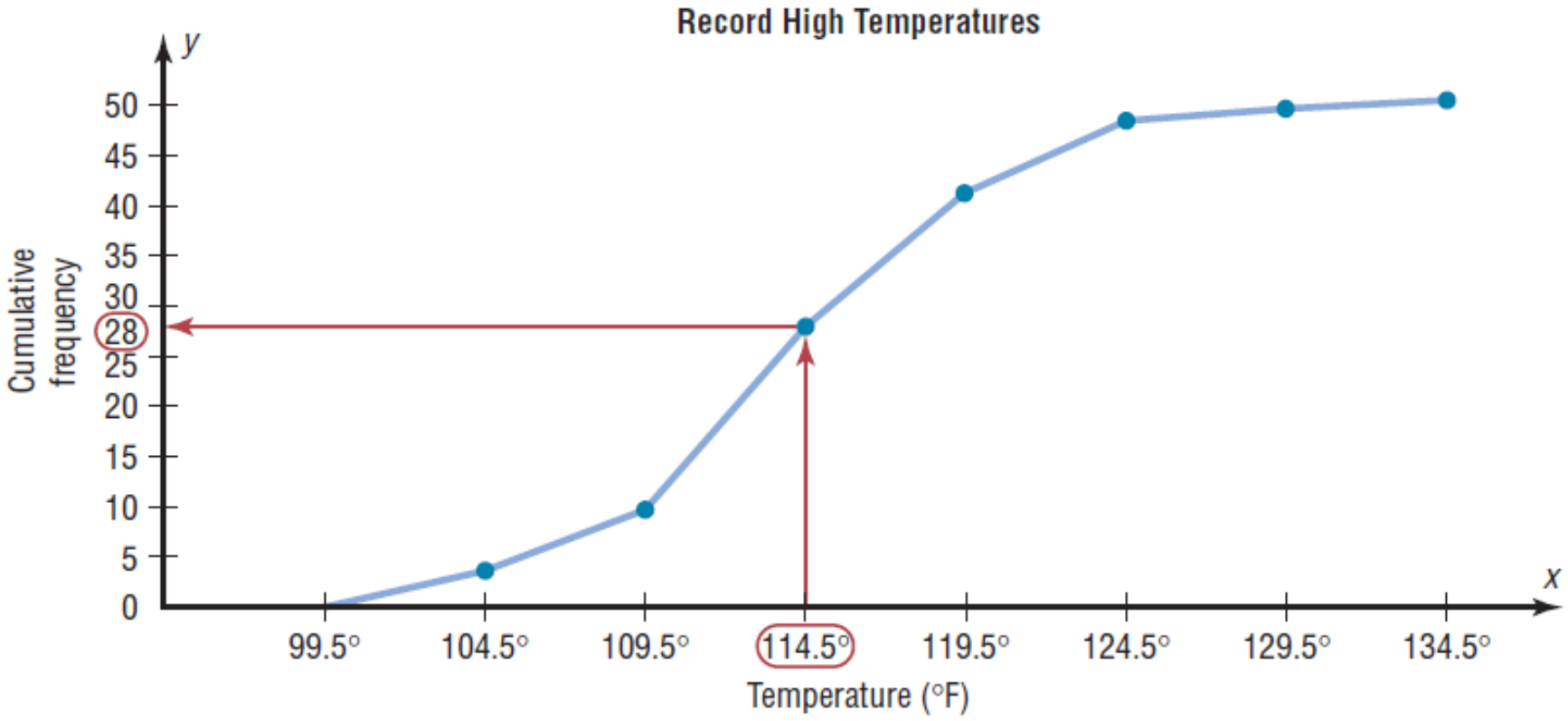
- **Ex:** Construct a histogram to represent the data shown for the record high temperatures for each of the 50 states. **(Ogive)**

Cumulative Frequency



Histograms, Frequency Polygons, and Ogives

- **Ex:** Construct a histogram to represent the data shown for the record high temperatures for each of the 50 states. **(Ogive)**



Relative Frequency Graphs

- The histogram, the frequency polygon, and the ogive shown previously were constructed by using frequencies in terms of the raw data. These distributions can be converted to distributions using **proportions** instead of raw data as frequencies. These types of graphs are called **relative frequency graphs**.
- Graphs of relative frequencies instead of frequencies are used when the proportion of data values that fall into a given class is more important than the actual number of data values that fall into that class.
- To convert a frequency into a proportion or relative frequency, divide the frequency for each class by the total of the frequencies. The sum of the relative frequencies will always be 1.

Relative Frequency Graphs

- **Ex.** Construct a histogram, frequency polygon, and ogive using relative frequencies for the distribution of the miles that 20 randomly selected runners ran during a given week.

Class boundaries	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2
	20

Class boundaries	Midpoints	Relative frequency
5.5–10.5	8	0.05
10.5–15.5	13	0.10
15.5–20.5	18	0.15
20.5–25.5	23	0.25
25.5–30.5	28	0.20
30.5–35.5	33	0.15
35.5–40.5	38	0.10
		1.00

- Convert each frequency to a proportion or relative frequency by dividing the frequency for each class by the total number of observations.
- For class 5.5–10.5, the relative frequency is $\frac{1}{20} = 0.05$; for class 10.5–15.5, $\frac{2}{2o} = 0.10$ and so on.

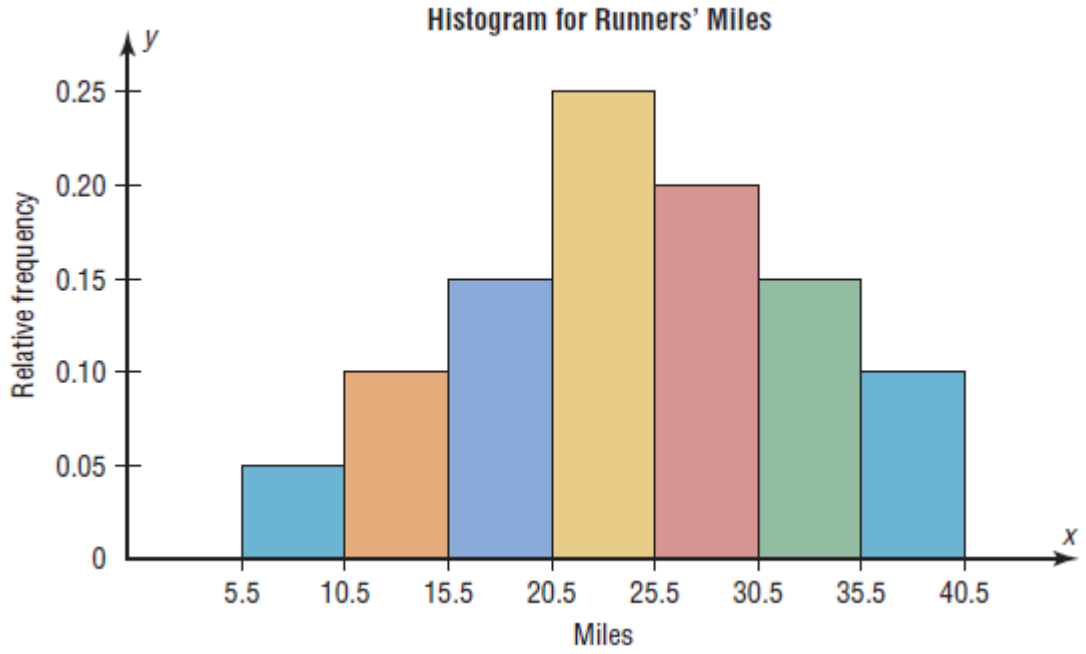
Relative Frequency Graphs

- **Ex.** Construct a histogram, frequency polygon, and ogive using relative frequencies for the distribution of the miles that 20 randomly selected runners ran during a given week.
- Cumulative relative frequencies are found by adding the frequency in each class to the total frequency of the preceding class. In this case, 0+0.05=0.05, 0.05+ 0.10= 0.15, 0.15+0.15=0.30, 0.30+0.25=0.55, etc.

	Cumulative frequency	Cumulative relative frequency
Less than 5.5	0	0.00
Less than 10.5	1	0.05
Less than 15.5	3	0.15
Less than 20.5	6	0.30
Less than 25.5	11	0.55
Less than 30.5	15	0.75
Less than 35.5	18	0.90
Less than 40.5	20	1.00

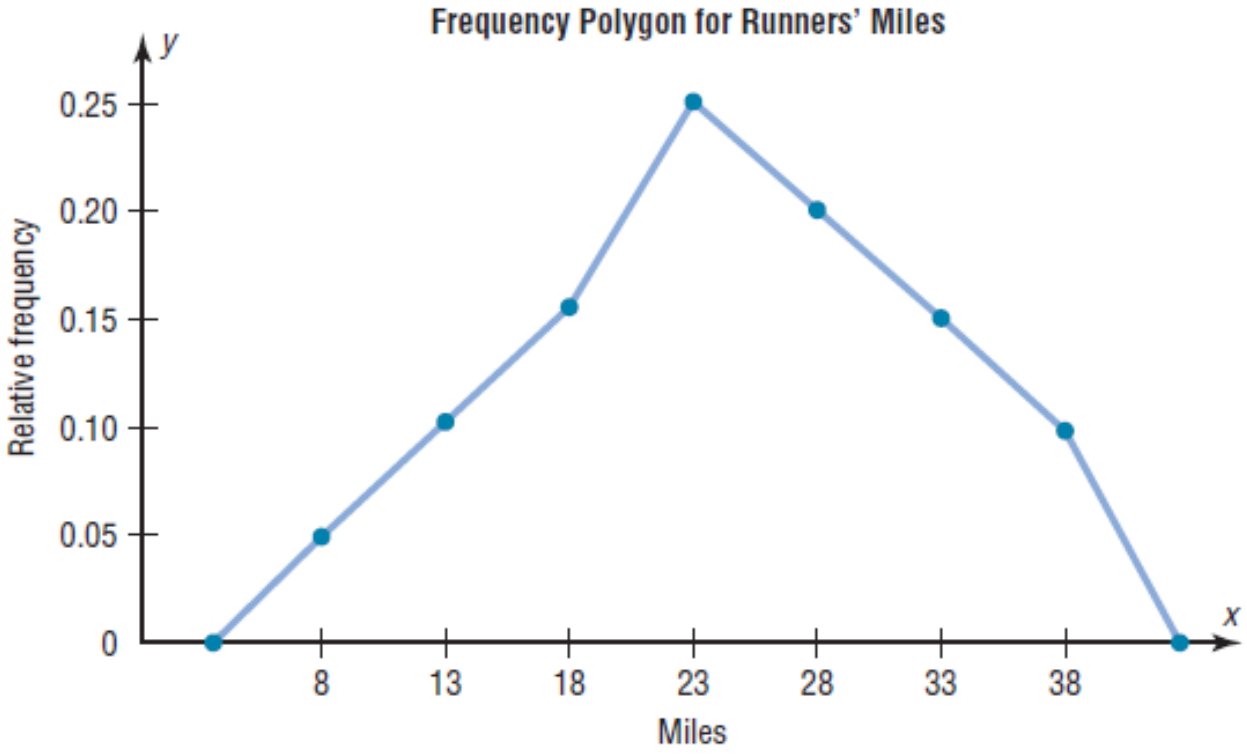
Relative Frequency Graphs

- **Ex.** Construct a histogram, frequency polygon, and ogive using relative frequencies for the distribution of the miles that 20 randomly selected runners ran during a given week.



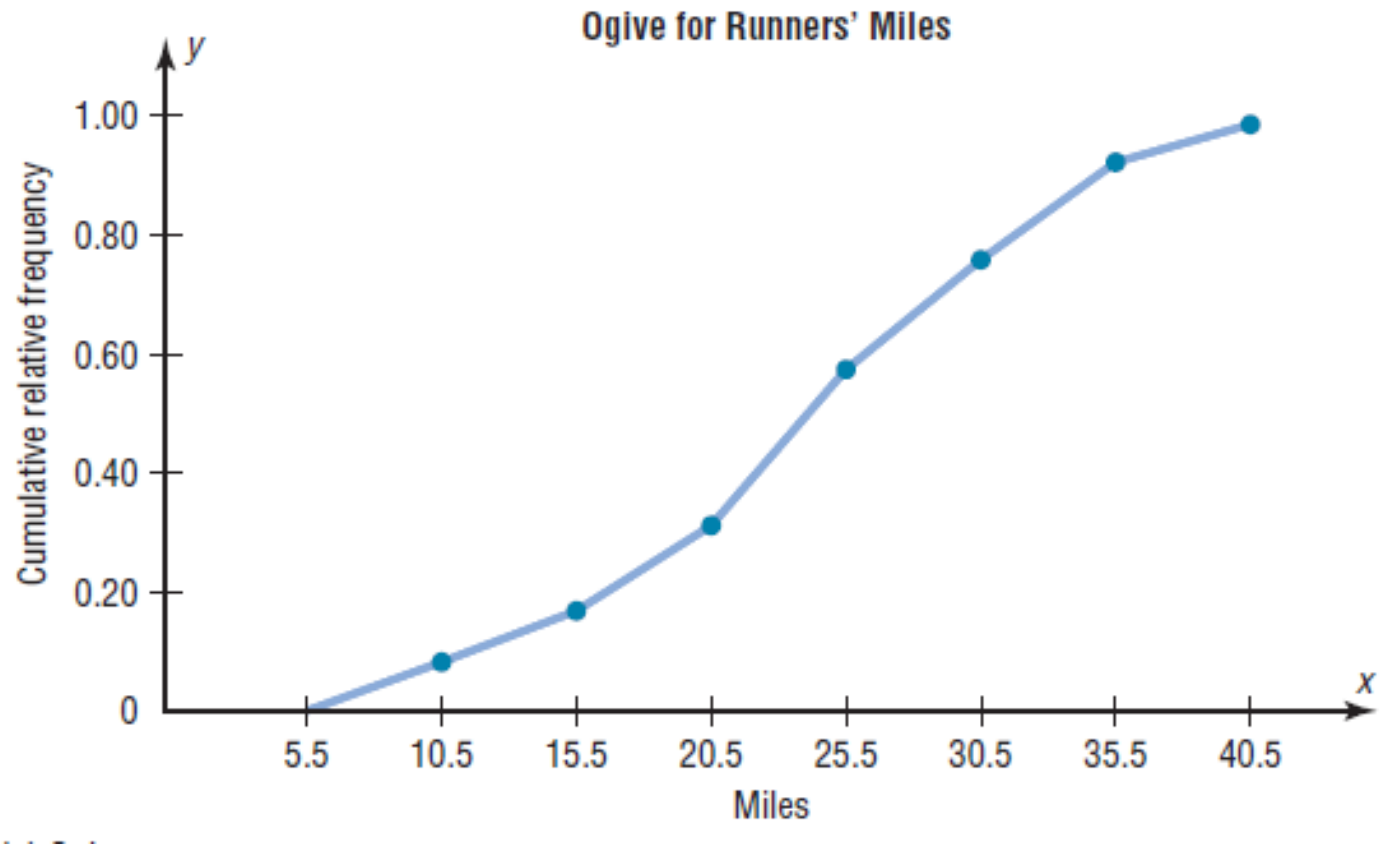
Relative Frequency Graphs

- **Ex.** Construct a histogram, frequency polygon, and ogive using relative frequencies for the distribution of the miles that 20 randomly selected runners ran during a given week.



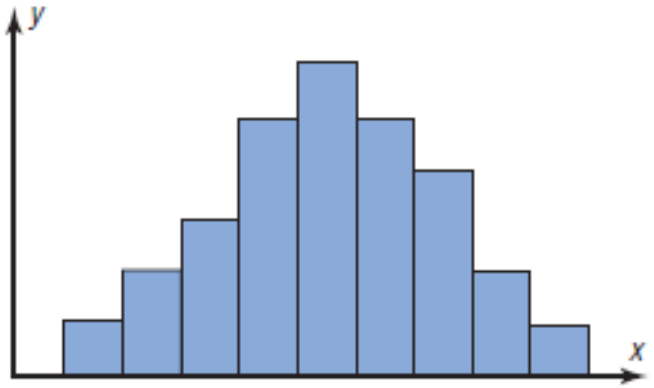
Relative Frequency Graphs

- **Ex.** Construct a histogram, frequency polygon, and ogive using relative frequencies for the distribution of the miles that 20 randomly selected runners ran during a given week.

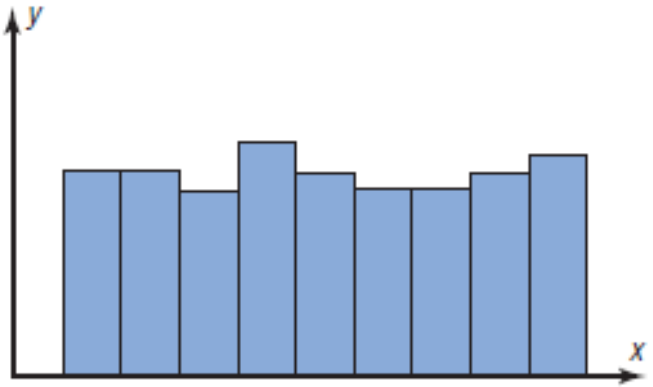


Distribution Shapes

- When one is describing data, it is important to be able to recognize the shapes of the distribution values.
- **The shape of a distribution also determines the appropriate statistical methods used to analyze the data**.
- A distribution can have many shapes, and one method of analyzing a distribution is to draw a histogram or frequency polygon for the distribution.
- The bell-shaped or mound-shaped (normal dağılım), the uniform shaped, the J-shaped, the reverse J-shaped, the positively or right-skewed shape, the negatively or left-skewed shape, the bimodal-shaped, and the U-shaped.

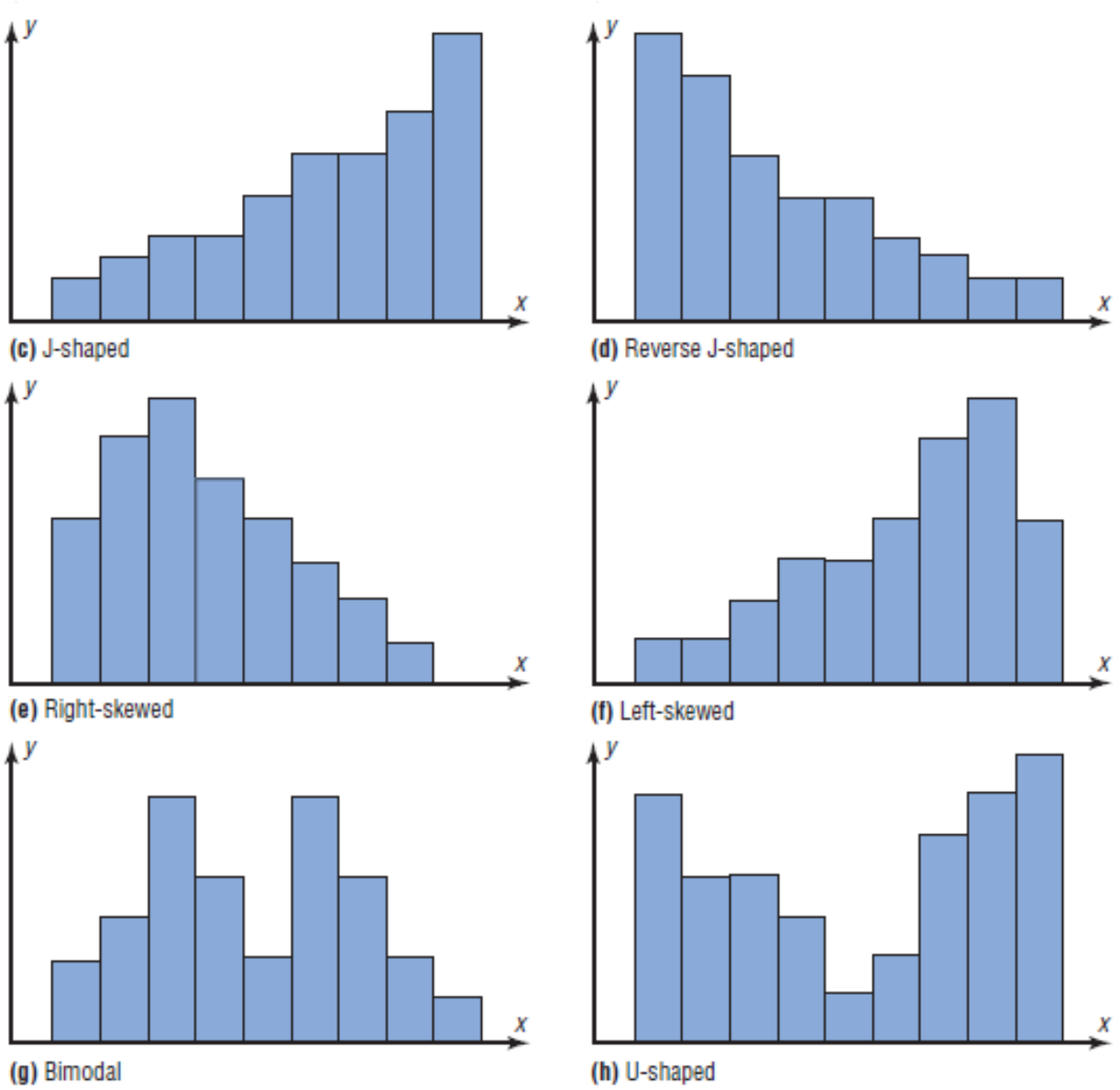


(a) Bell-shaped



(b) Uniform

Distribution Shapes



Distribution Shapes

- A bell-shaped distribution has a single peak and tapers off at either end. It is approximately symmetric; i.e., it is roughly the same on both sides of a line running through the center.
- A uniform distribution is basically flat or rectangular.
- J-shaped distribution is has a few data values on the left side and increases as one moves to the right. A reverse J-shaped distribution is the opposite of the J-shaped distribution.
- When the peak of a distribution is to the left and the data values taper off to the right, a distribution is said to be positively or **right-skewed**.
- When the data values are clustered to the right and taper off to the left, a distribution is said to be negatively or **left-skewed**.
- Distributions with one peak are said to be **unimodal**.
- When a distribution has two peaks of the same height, it is said to be **bimodal**.

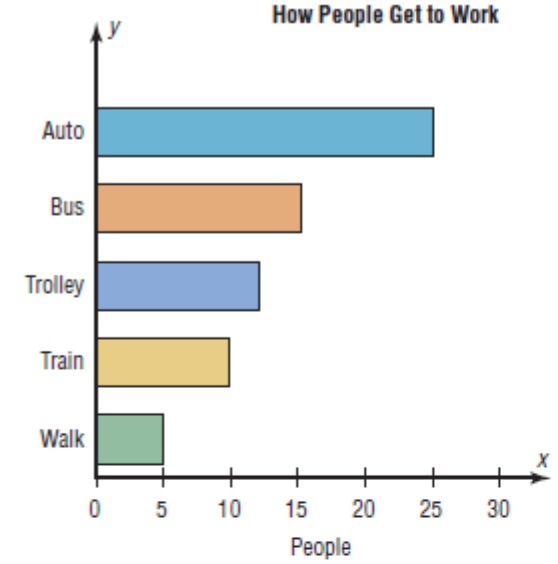
Distribution Shapes

- When you are analyzing histograms and frequency polygons, look at the shape of the curve.
- Does it have one peak or two peaks?
- Is it relatively flat, or is it U-shaped?
- Are the data values spread out on the graph, or are they clustered around the center?
- Are there data values in the extreme ends? These may be outliers.
- Are there any gaps in the histogram, or does the frequency polygon touch the x axis somewhere other than at the ends?
- Are the data clustered at one end or the other, indicating a skewed distribution?

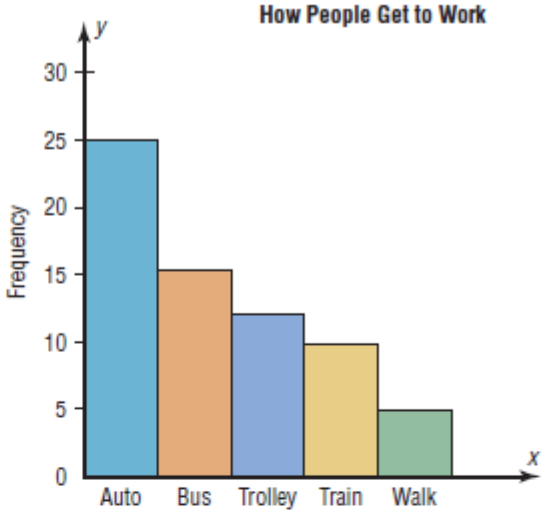
Other Types of Graphs

- **Bar graph** represents the data by using vertical or horizontal bars whose heights or lengths represent the frequencies of the data.
- **Pareto chart** is used to represent a frequency distribution for a categorical variable, and the frequencies are displayed by the heights of vertical bars, which are arranged in order from highest to lowest.
- **Time series graph** represents data that occur over a specific period of time.
- When you analyze a time series graph, look for a trend or pattern that occurs over the time period. For example, is the line ascending (indicating an increase over time) or descending (indicating a decrease over time)? Another thing to look for is the slope, or steepness, of the line. A line that is steep over a specific time period indicates a rapid increase or decrease over that period.
- **Pie graph** is a circle that is divided into sections or wedges according to the percentage of frequencies in each category of the distribution.

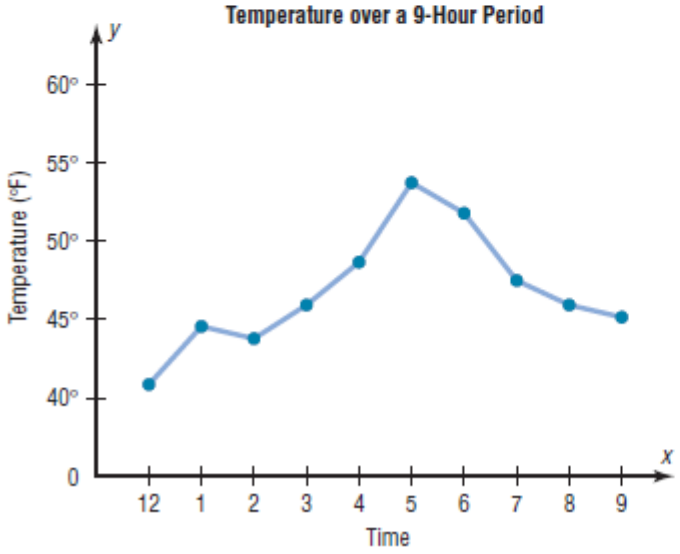
Other Types of Graphs



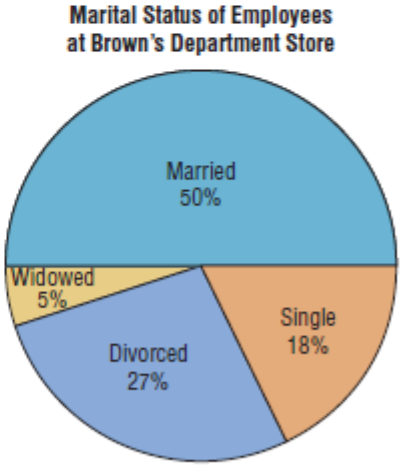
(a) Bar graph



(b) Pareto chart



(c) Time series graph



(d) Pie graph

- Bar graph, Pareto chart, time series graph, and pie graph