Chapter 6

Descriptive Statistics

Chapter 6: Random sampling & Data Description

Learning objectives

1. Numerical Summaries
2. Stem-and-Leaf Diagrams
3. Frequency distributions and histograms
4. Box Plots
5. Time Sequence Plots

1.Numerical Summaries

Numerical Summaries

Measures of central tendency

Sample mean

If the n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, **the sample mean** is

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Example

Let's consider the weight of the eight observations collected from the prototype engine connectors: 12.6, 12.9, 13.4, 12.3, 13.6, 13.5, 12.6 and 13.1

NUMERICAL SUMMARIES

MEASURES OF CENTRAL TENDENCY

Sample median

(1) The value that lies in the middle of the data when the data set is ordered.

(2) Measures the center of an ordered data set by dividing it into two equal parts.

(3) If the data set has an

(a) even number of entries: **median** is the mean of the two middle data entries.

(b) odd number of entries: **median** is the middle data entry.

NUMERICAL SUMMARIES

MEASURES OF CENTRAL TENDENCY

Example

The prices (in dollars) for a sample of roundtrip flights from Chicago, Illinois to Cancun, Mexico are listed. Find the median of the flight prices.

872 432 397 427 388 782 397

First order the data.

388 397 397 427 432 782 872

The median price of the flights is $427.

NUMERICAL SUMMARIES

MEASURES OF CENTRAL TENDENCY

Sample mode

(1) The data entry that occurs with the greatest frequency.

(2) If no entry is repeated the data set has no mode.

(3) If two entries occur with the same greatest frequency, each entry is a mode (bimodal).

MEASURES OF CENTRAL TENDENCY

Example

At a political debate a sample of audience members was asked to name the political party to which they belong. Their responses are shown in the table. What is the mode of the responses?

Political Party	Frequency, *f*
Democrat	34
Republican	56
Other	21
Did not respond	9

NUMERICAL SUMMARIES

MEASURES OF VARIATION

Sample Variance and sample standard deviation

(1) If $x_1, x_2, \ldots, x_n$, is a sample of n observations, the sample variance is

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

(2) The sample standard deviation, s, is the positive square root of the sample variance.

MEASURES OF VARIATION

Computing formula for σ^2

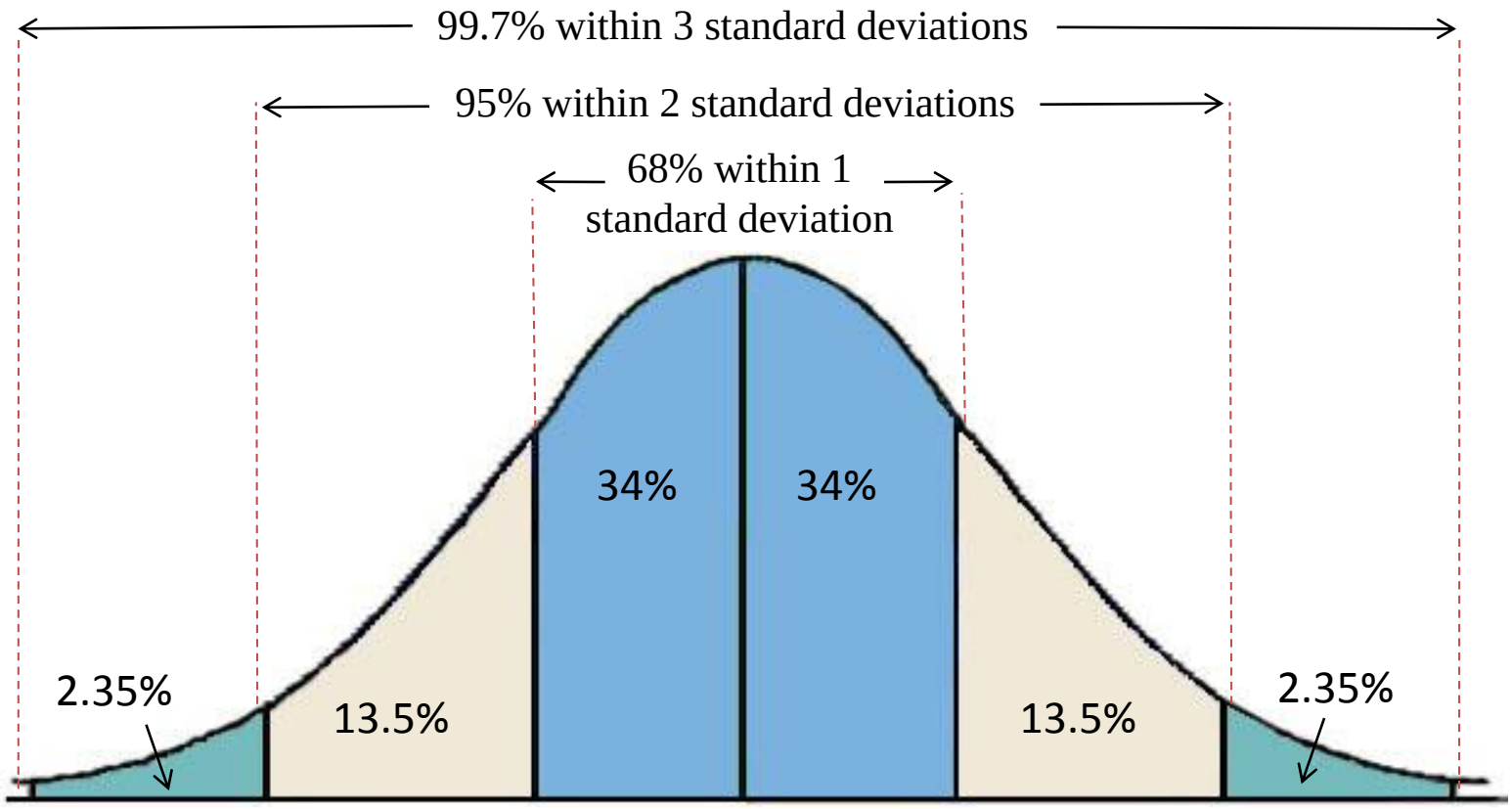
$$s^2 = \frac{n}{n-1}\left(\frac{\sum_{i=1}^{n} x_i^2}{n} - (\bar{x})^2\right)$$

Example

Let's consider the weight of the eight observations collected from the prototype engine connectors: 12.6, 12.9, 13.4, 12.3, 13.6, 13.5, 12.6 and 13.1

MEASURES OF VARIATION

Interpreting standard deviation: For data with a bell-shaped distribution



Numerical Summaries

MEASURES OF VARIATION

Sample range

- The difference between the maximum and minimum data entries in the set.

- The data must be quantitative.

- If the n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the sample range is

$$r = max(x_i) - min(x_i)$$

Numerical Summaries

Measures of position: quartiles

(1) **Fractiles** are numbers that partition (divide) an ordered data set into equal parts.

(2) **Quartiles** approximately divide an ordered data set into four equal parts.

(a) **First quartile, *Q*1**: About one quarter of the data fall on or below *Q*1.

(b) **Second quartile, *Q*2**: About one half of the data fall on or below *Q*2 (median).

(c) **Third quartile, *Q*3**: About three quarters of the data fall on or below *Q*3.

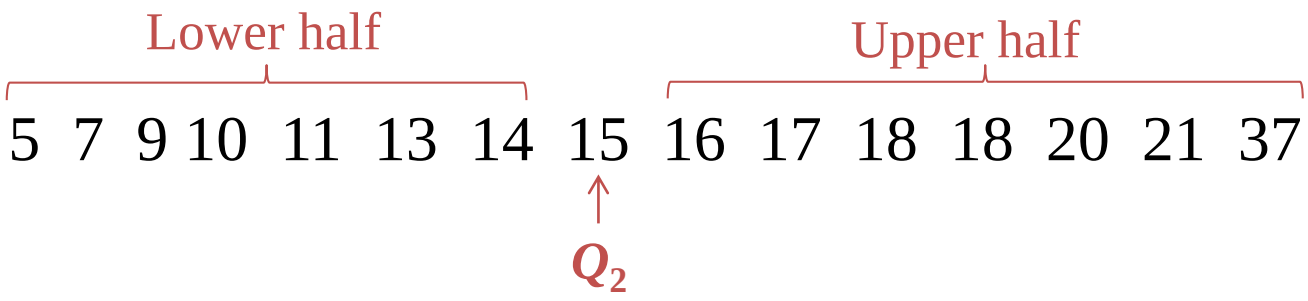
MEASURES OF POSITION: QUARTILES

Example

The test scores of 15 employees enrolled in a CPR training course are listed. Find the first, second, and third quartiles of the test scores.

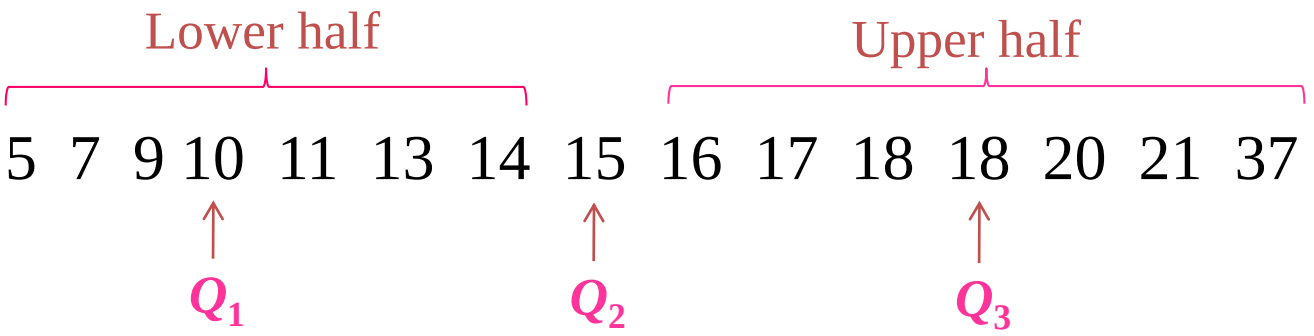
13 9 18 15 14 21 7 10 11 20 5 18 37 16 17

Q_2 divides the data set into two halves.



MEASURES OF POSITION: QUARTILES

The first and third quartiles are the medians of the lower and upper halves of the data set.



About one fourth of the employees scored 10 or less, about one half scored 15 or less; and about three fourths scored 18 or less.

Stem-and-Leaf Diagrams

Stem-and-Leaf Diagrams

Steps to Construct a Stem-and-Leaf Diagram

(1) Divide each number x_i into two parts: a **stem**, consisting of one or more of the leading digits, and a **leaf**, consisting of the remaining digit.

(2) List the stem values in a vertical column.

(3) Record the leaf for each observation beside its stem.

(4) Write the units for stems and leaves on the display

Example 1: Use the data in the table to make a stem-and-leaf diagrams.

Test Scores				
75	86	83	91	94
88	84	99	79	86

Example 1: Use the data in the table to make a stem-and-leaf diagrams.

Test Scores				
75	86	83	91	94
88	84	99	79	86

Test Scores

Stems	Leaves
7	5 9
8	3 4 6 6 8
9	1 4 9

Key: 7\|5 means 75

Example 2

Use the data in the table to make a stem-and-leaf plot.

Test Scores				
72	88	64	79	61
84	83	76	74	67

Example 2: Reading Stem-and-Leaf Plots

Find the least value, greatest value, mean, median, mode, Q1, Q3, and range of the data.

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

Key: 4\|0 means 40

3. Frequency distributions and histograms

Frequency distribution

Frequency Distribution

(1) The **frequency distribution** is a summary table in which the data are arranged into numerically ordered class groupings.

(2) You must give attention to selecting the appropriate *number* of **class groupings** for the table, determining a suitable *width* of a class grouping, and establishing the *boundaries* of each class grouping to avoid overlapping.

(3) To determine the **width of a class interval,** you divide the **range** (Highest value–Lowest value) of the data by the number of class groupings desired.

FREQUENCY DISTRIBUTION

Example

A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

24, 35, 17, 21, 24, 37, 26, 46, 58, 30,
32, 13, 12, 38, 41, 43, 44, 27, 53, 27

Frequency distribution

Histogram

(1) A graph of the data in a frequency distribution is called a **histogram.**

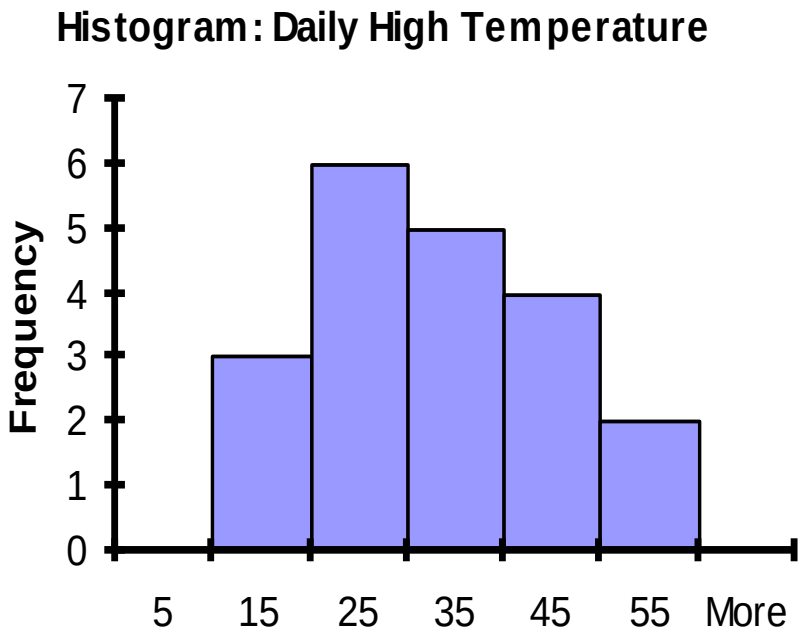
(2) The **class boundaries** (or **class midpoints**) are shown on the horizontal axis.

(3) The vertical axis is either **frequency, relative frequency,** or **percentage**.

(4) Bars of the appropriate heights are used to represent the number of observations within each class.

Frequency distribution

Class	Frequency	Relative Frequency
[10, 20)	3	0.15
[20, 30)	6	0.30
[30, 40)	5	0.25
[40, 50)	4	0.20
[50, 60)	2	0.10
Total	20	1.00



Frequency distribution

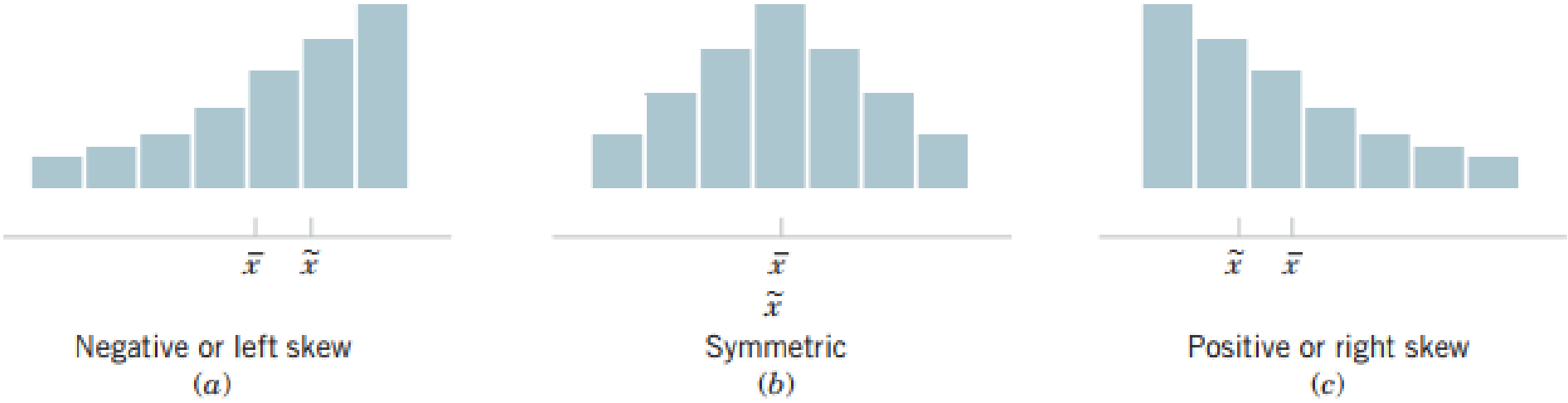
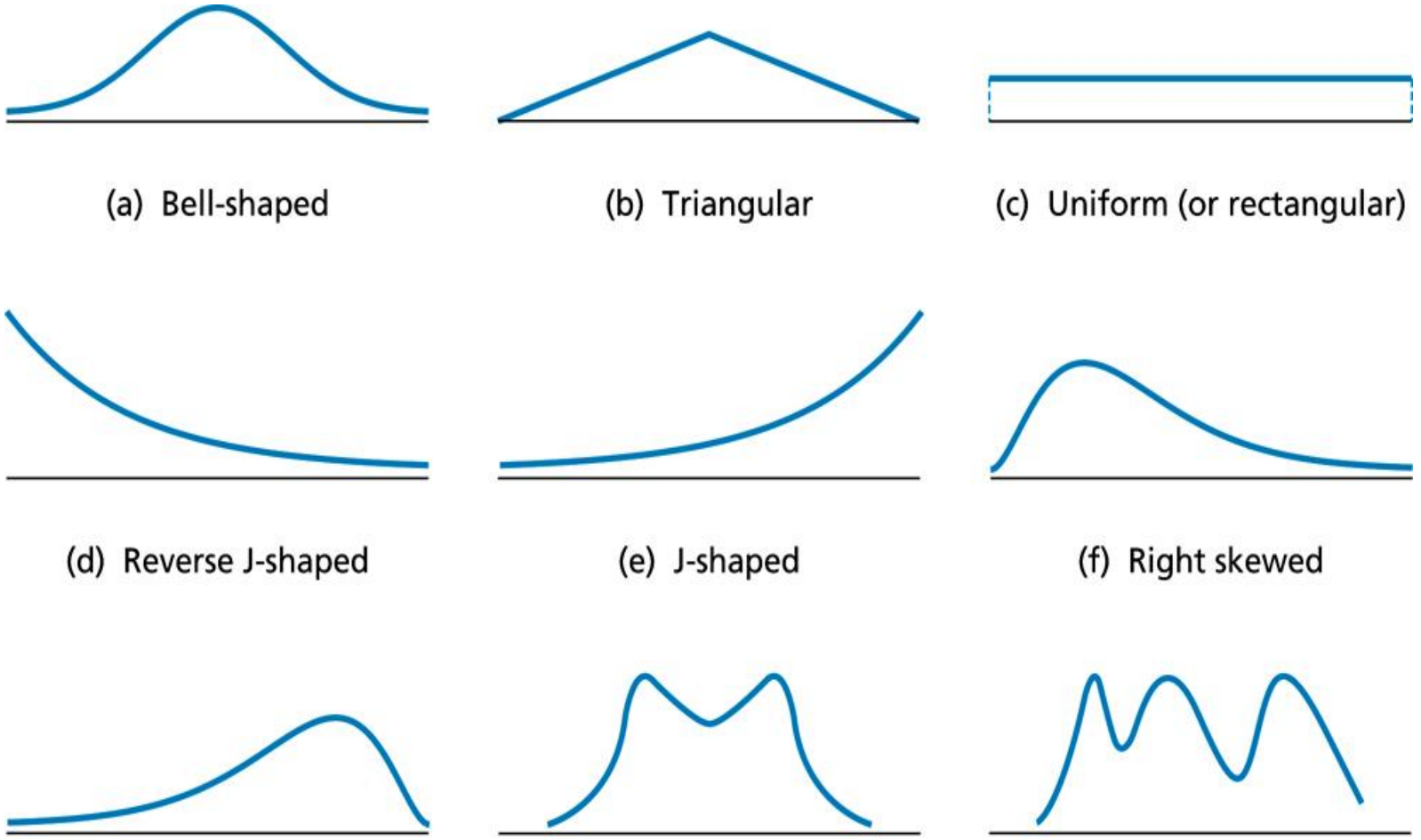


Figure 6-11 Histograms for symmetric and skewed distributions.

Frequency distribution

COMMON DISTRIBUTION SHAPES



Box Plots

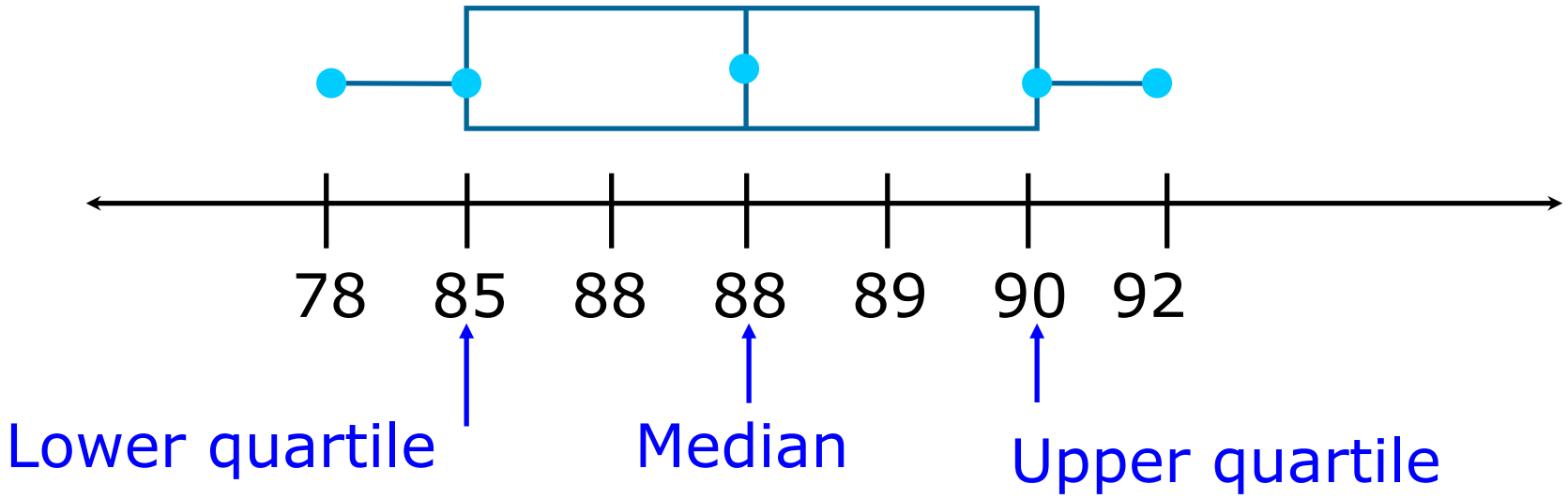
Box Plots

- A **box plot** shows the distribution of data. The middle half of the data is represented by a "box" with a vertical line at the median.
- The box extends to the **upper and lower quartiles.**
- The **upper quartile** is the median of the upper half of the data. The **lower quartile** is the median of the lower half of the data.
- The lower fourth and upper fourth quarters are represented by "whiskers" that extend to the **minimum** (least) and **maximum** (greatest) values.

Example: Making a Box Plots

Data set: 85,92,78,88,90,88,89

What is a box plot?
What are the upper and lower quartiles?
What are minimum and maximum values?



Example 2:

Use the given data to make a box plot.

a) 21, 25, 15, 13, 17, 19, 19, 21

b) 31, 23, 33, 35, 26, 24, 31, 29

Example: Making a Box Plots

The table below summarizes a cat breeder's records for kitten litters born in a given year. You can divide the data into four equal part using *quartiles.*

Litter Size	2	3	4	5	6
Number of Litters	1	6	8	11	1

Time Sequence Plots

2-5 Time Sequence Plots

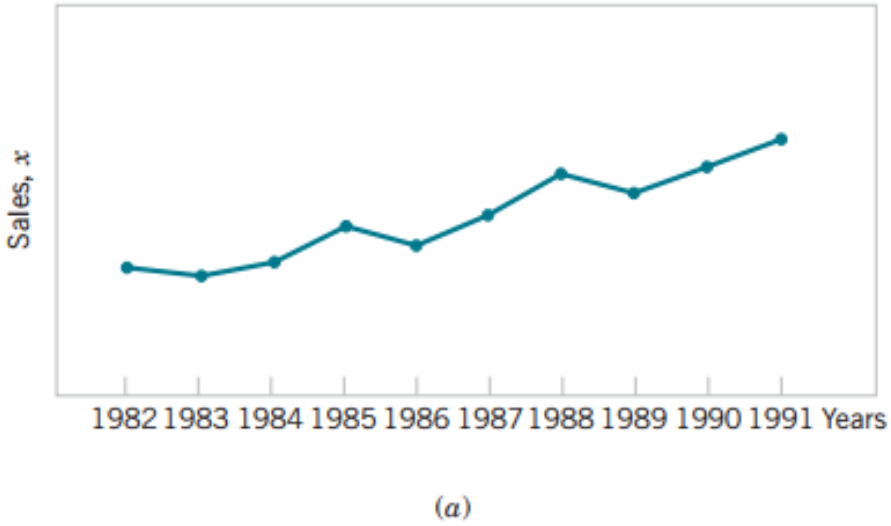
• A **time series** or **time sequence** is a data set in which the observations are recorded in the order in which they occur.

• A **time sequence plot** is a graph in which the vertical axis denotes the observed value of the variable (say *x*) and the horizontal axis denotes the time (which could be minutes, days, years, etc.).

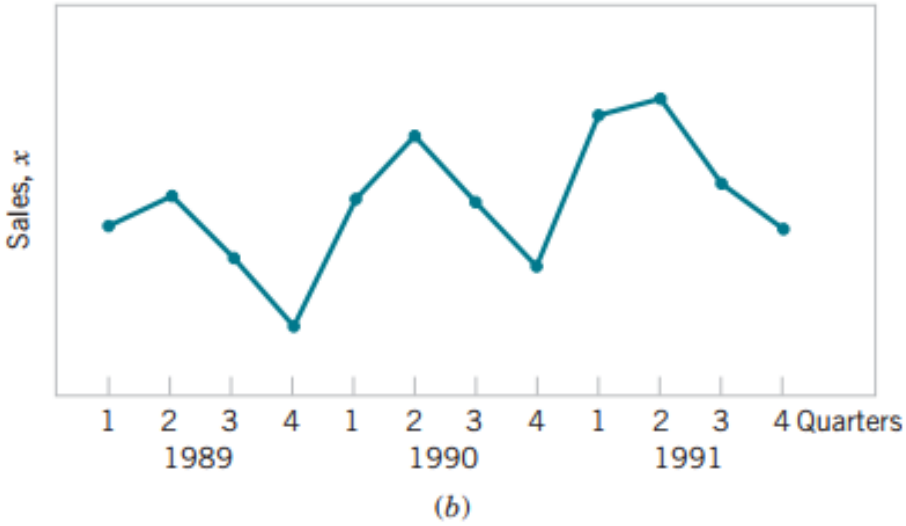
• When measurements are plotted as a time series, we often see

- **trends,**
- **cycles, or**
- **other broad features of the data**

2-5 Time Sequence Plots



(a)



(b)

FIGURE 6.16

Company sales by year (*a*). By quarter (*b*).

Leaf	Stem
5	24
7	23
189	22
8	21
7108	20
960934	19
0361410	18
8544162106	17
3073050879	16
471340886808	15
29583169	14
413535	13
103	12
580	11
15	10
7	9
7	8
6	7

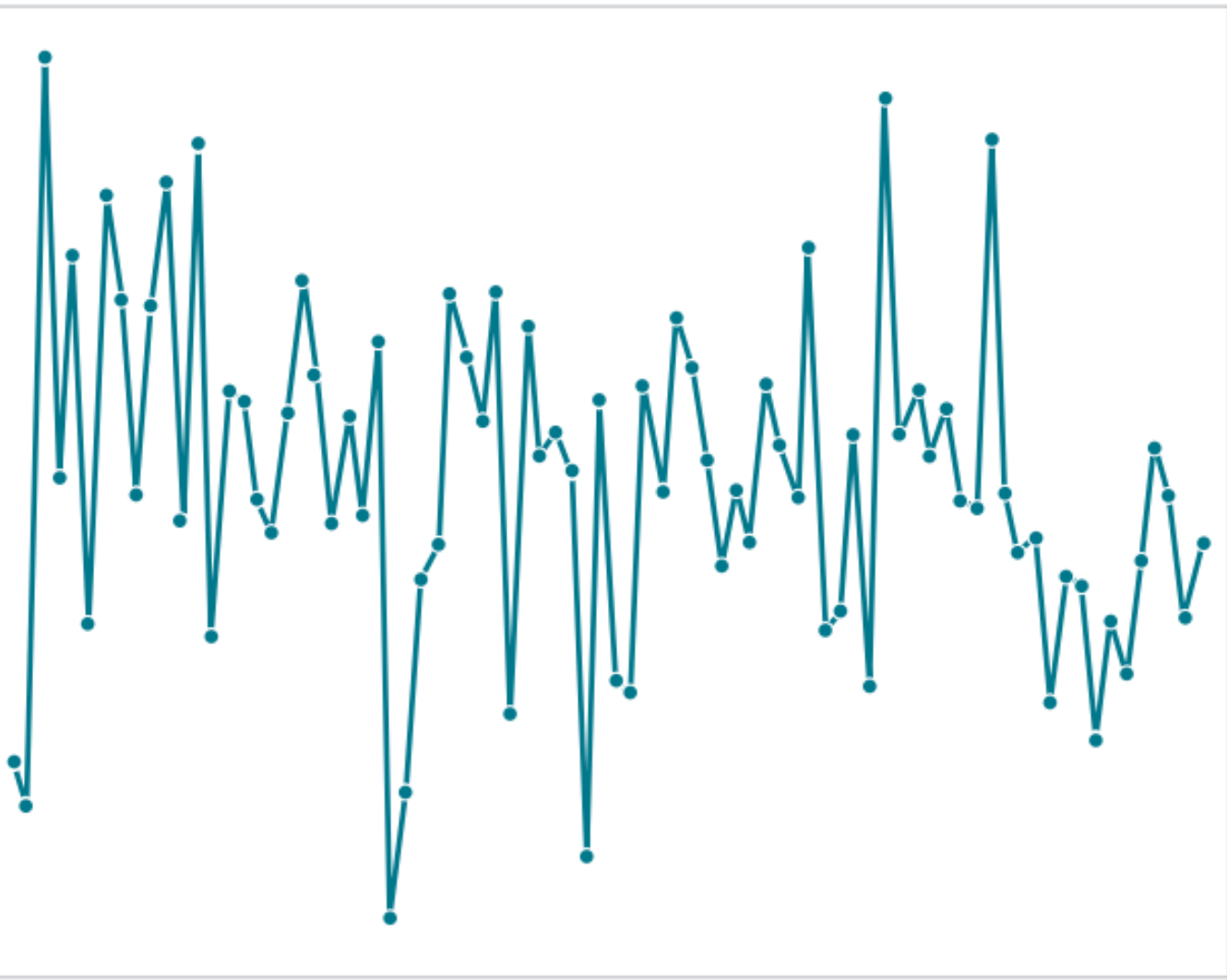


FIGURE 6.17

A digidot plot of the compressive strength data in Table 6.2.

2-5 Time Sequence Plots

Leaf	Stem
8	9e
6	9s
45	9f
2333	9t
0010000	9z
99998	8e
66677	8s
45	8f
23	8t
1	8z

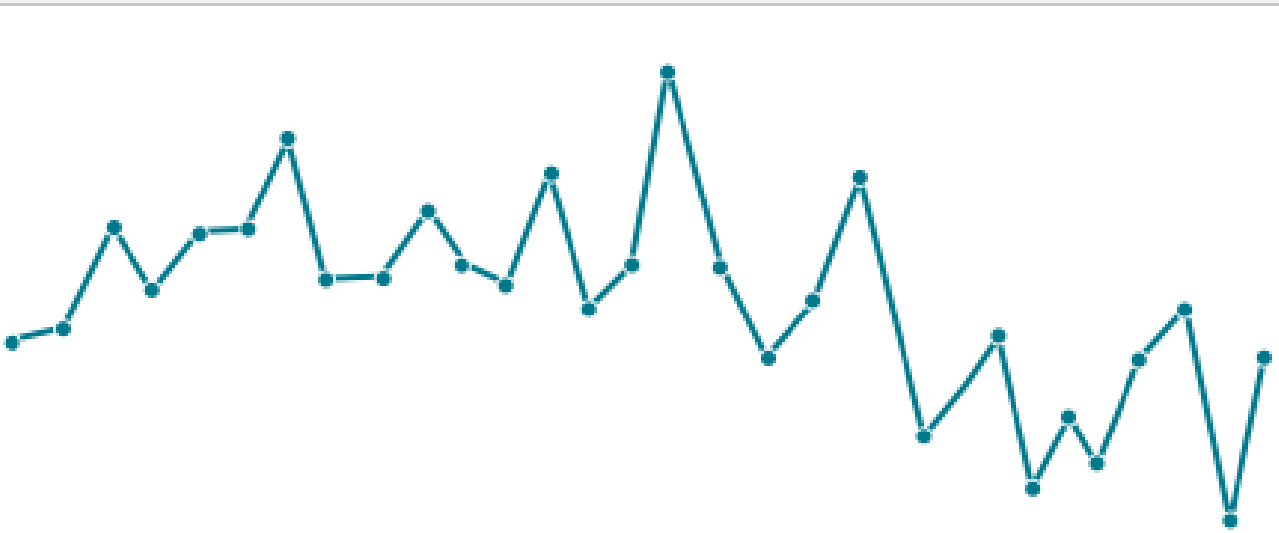


FIGURE 6.18

A digidot plot of chemical process concentration readings, observed hourly.