Chapter Three

Data Description

3-1 Introduction

- Measures of average are also called **measures of central tendency,** which is used to summarize data, and include **mean**, **median**, **mode**, **midrange**, and **weighted mean**.

- Measures that determine the spread of data values are called **measures of variation** and include **range**, **variance**, and **standard deviation**.

- **Measures of position** tell where a specific data value falls within a data set or its relative position in comparison with other data values and include **standard scores** and **quartiles**.

3-1 Introduction

- The measures of central tendency, variation, and position are part of what is called **traditional statistics**.
- Another type of statistics is called **exploratory data analysis**, which includes the **box plot** and **five-number summary**.
- A **statistic** is a measure calculated using the data values of a sample.
- A **parameter** is a measure calculated using all the data values of a specific population.

3-2 Measures of Central Tendency

- The **mean** is the sum of the values divided by the total number of values.

  - The Greek letter $\boldsymbol{\mu}$ (mu) is used to represent the **population mean**.

$$\boldsymbol{\mu} = \frac{\sum \boldsymbol{X}}{\boldsymbol{N}} = \frac{\boldsymbol{X}_1 + \boldsymbol{X}_2 + \cdots + \boldsymbol{X}_N}{\boldsymbol{N}}$$

  - The symbol $\boldsymbol{x}$ (x-bar) represents the **sample mean**.

$$\overline{\boldsymbol{x}} = \frac{\sum \boldsymbol{x}}{\boldsymbol{n}} = \frac{\boldsymbol{x}_1 + \boldsymbol{x}_2 + \cdots + \boldsymbol{x}_n}{\boldsymbol{n}}$$

3-2 Measures of Central Tendency

- Example:
  - Suppose that the number of burglaries reported in a specific year for nine cities are 61, 11, 1, 3, 2, 30, 18, 3, 7 Find the mean.

$$\bar{x} = \frac{\sum x}{n} = \frac{61 + 11 + 1 + 3 + 2 + 30 + 18 + 3 + 7}{9} = \frac{136}{9} = 15.1$$

3-2 Measures of Central Tendency

- The **median** (***MD***) is the halfway point in a data set.
- The median is found by arranging the data in order and selecting the middle point.

- Example if ***n*** is odd:
  - Suppose that the number of burglaries reported in a specific year for nine cities are 61, 11, 1, 3, 2, 30, 18, 3, 7 Find the median.

$$1, 2, 3, 3, \boxed{7}, 11, 18, 30, 61$$

$$\boxed{\boldsymbol{MD = 7}}$$

3-2 Measures of Central Tendency

- Example ***n*** is even:
  - Suppose that the number of burglaries reported in a specific year for eight cities are 11, 1, 3, 2, 30, 18, 3, 7 Find the median.

$$1, 2, 3, \boxed{3, 7}, 11, 18, 30$$

$$\boxed{\boldsymbol{MD = \frac{3+7}{2} = 5}}$$

3-2 Measures of Central Tendency

- The **<u>mode</u>** is the value that occurs most often in a data set.
  - A data set with one value that occurs with greatest frequency is said to be **<u>unimodal</u>**. e.g., 3, 4, 2, 6, 4, 1, 5 → ***mode=4***
  - A data set with two values that occur with greatest frequency is said to be **<u>bimodal</u>**. e.g., 3, 4, 2, 6, 4, 1, 2 → ***mode=2, 4***
  - A data set with more than two values that occur with greatest frequency is said to be **<u>multimodal</u>**.
    e.g., 6, 3, 4, 2, 6, 4, 1, 2, 5, 6, 4, 2 → ***mode=2, 4,6***
  - When all the values in a data set occur with the same frequency is said to have **<u>no mode</u>**.
    e.g., 3, 4, 2, 3, 6, 4, 1, 2, 1, 6 → ***no mode***
    3, 4, 2, 6, 5, 1, 7, 8 → ***no mode***

3-2 Measures of Central Tendency

- Example:
  - Suppose that the number of burglaries reported in a specific year for nine cities are 61, 11, 1, 3, 2, 30, 18, 3, 7 Find the mode.

$$61, 11, 1, \boxed{3}, 2, 30, 18, \boxed{3}, 7$$

$$\boldsymbol{mode = 3}$$

3-2 Measures of Central Tendency

- The **midrange** (***MR***) is a rough estimate of the middle and defined as the sum of the lowest and highest values in a data set divided by 2.

$$\boldsymbol{MR = \frac{Max - Min}{2}}$$

- Example:
  - Suppose that the number of burglaries reported in a specific year for nine cities are 61, 11, 1, 3, 2, 30, 18, 3, 7 Find the midrange.

$$\boxed{61}, 11, \boxed{1}, 3, 2, 30, 18, 3, 7$$

$$\boldsymbol{MR = \frac{Min + Max}{2} = \frac{1 + 61}{2} = \frac{62}{2} = 31}$$

3-2 Measures of Central Tendency

- The **weighted mean** is used when the values in a data set are not all equally represented.

- The **weighted mean of a variable *X*** is found by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights.

$$\bar{x}_w = \frac{x_1 w_1 + x_2 w_2 + \cdots + x_n w_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum xw}{\sum w}$$

Where $w_1, w_2, \ldots, w_n$ are the weights for $x_1, x_2, \ldots, x_n$

3-2 Measures of Central Tendency

- Example:
  - A student received 90 in English (3 credits), 70 in Statistics (3 credits), 80 in Biology (4 credits) and 60 in physical education (2 credits), find the student's average grade.

$$\bar{x}_w = \frac{(90)(3) + (70)(3) + (80)(4) + (60)(2)}{3 + 3 + 4 + 2}$$

$$= \frac{270 + 210 + 320 + 120}{12} = \frac{920}{12} = 76.67$$

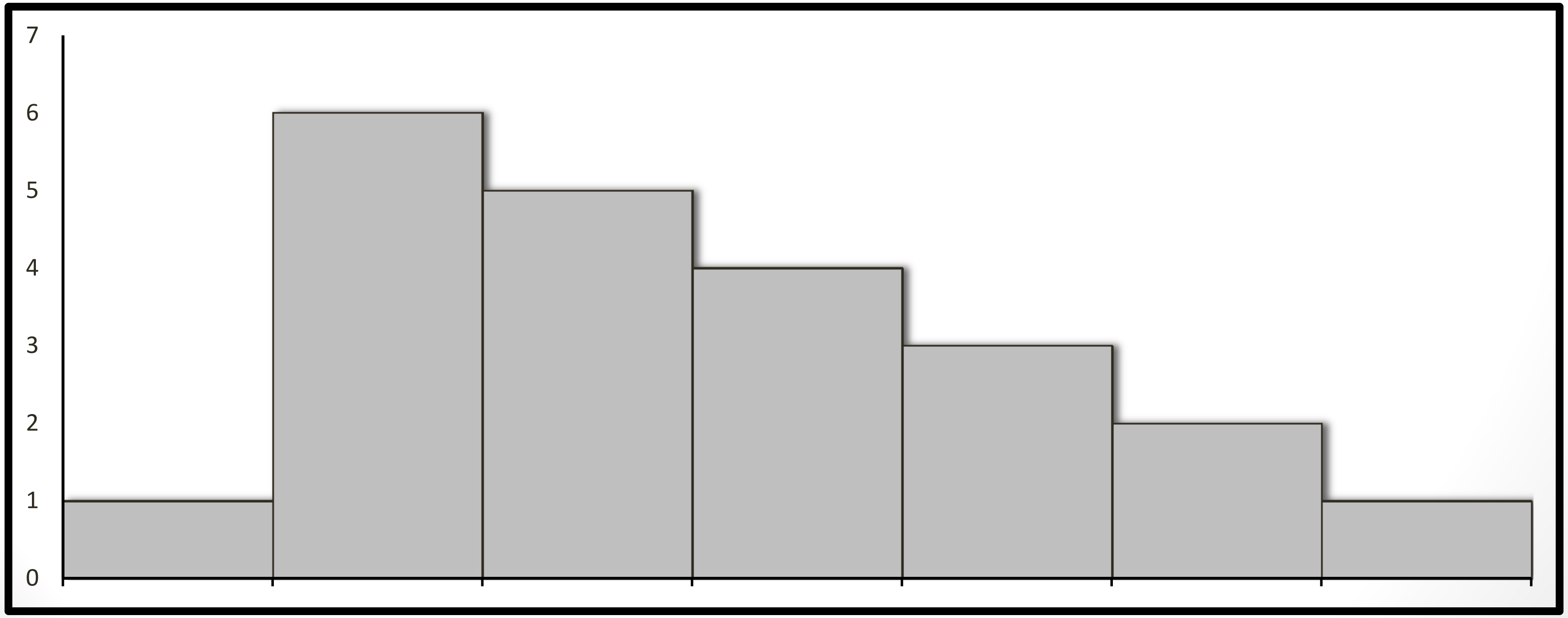
3-3 Properties of Central Tendency Measures

- The **mean** is affected by extremely high or low values and may not be the appropriate average.
- The **median** is **affected less** than the **mean** by extremely high or extremely low values.
- The **mode** can be used for categorical data, such as religious preference or gender.
- The **midrange** is affected by extremely high or low values in a data set.

3-4 Distribution Shapes

- In a **positively skewed** or **right skewed distribution**, the majority of the data values falls to the left of the mean and cluster at the lower end of the distribution.

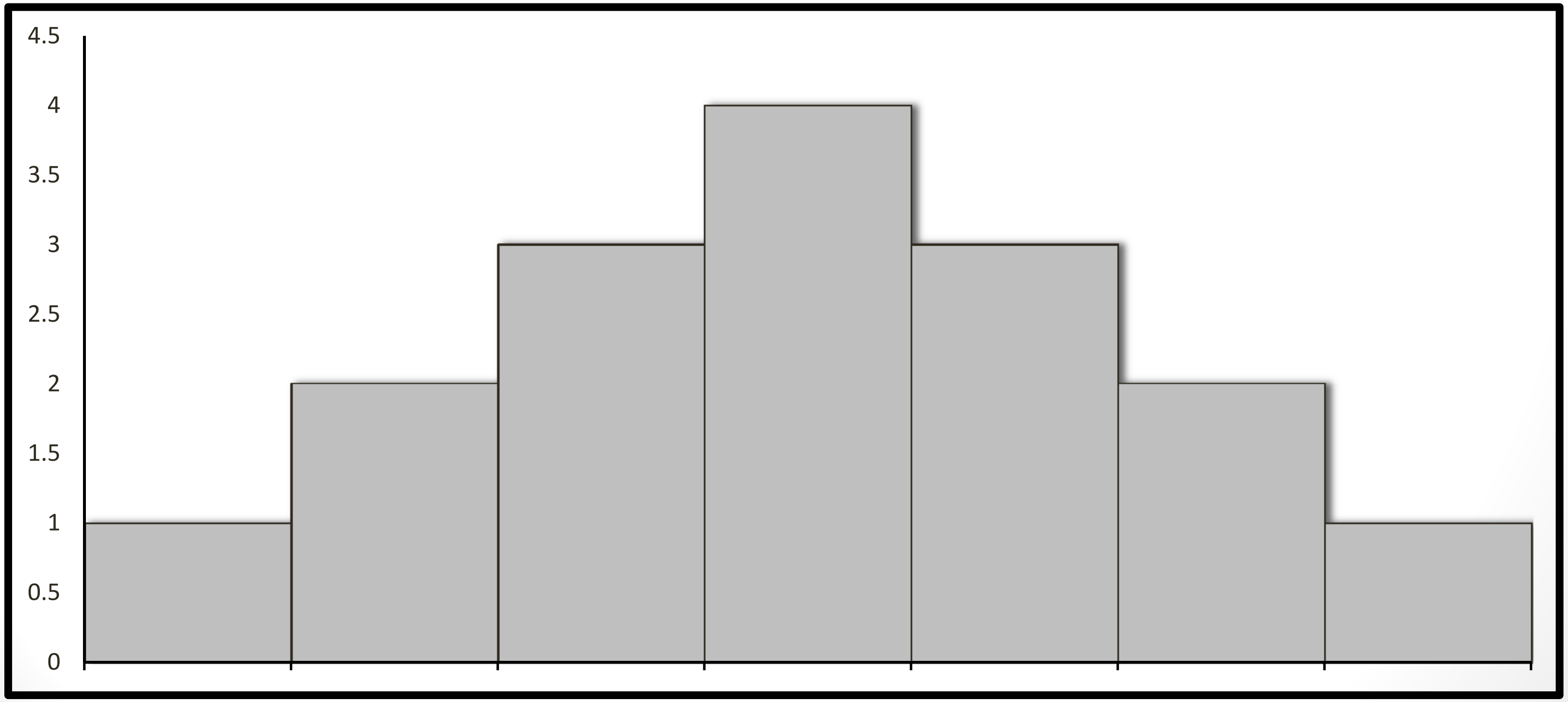
$$\boldsymbol{mode < median < mean}$$



3-4 Distribution Shapes

- In a **symmetrical distribution**, the data values are evenly distributed on both sides of the mean.

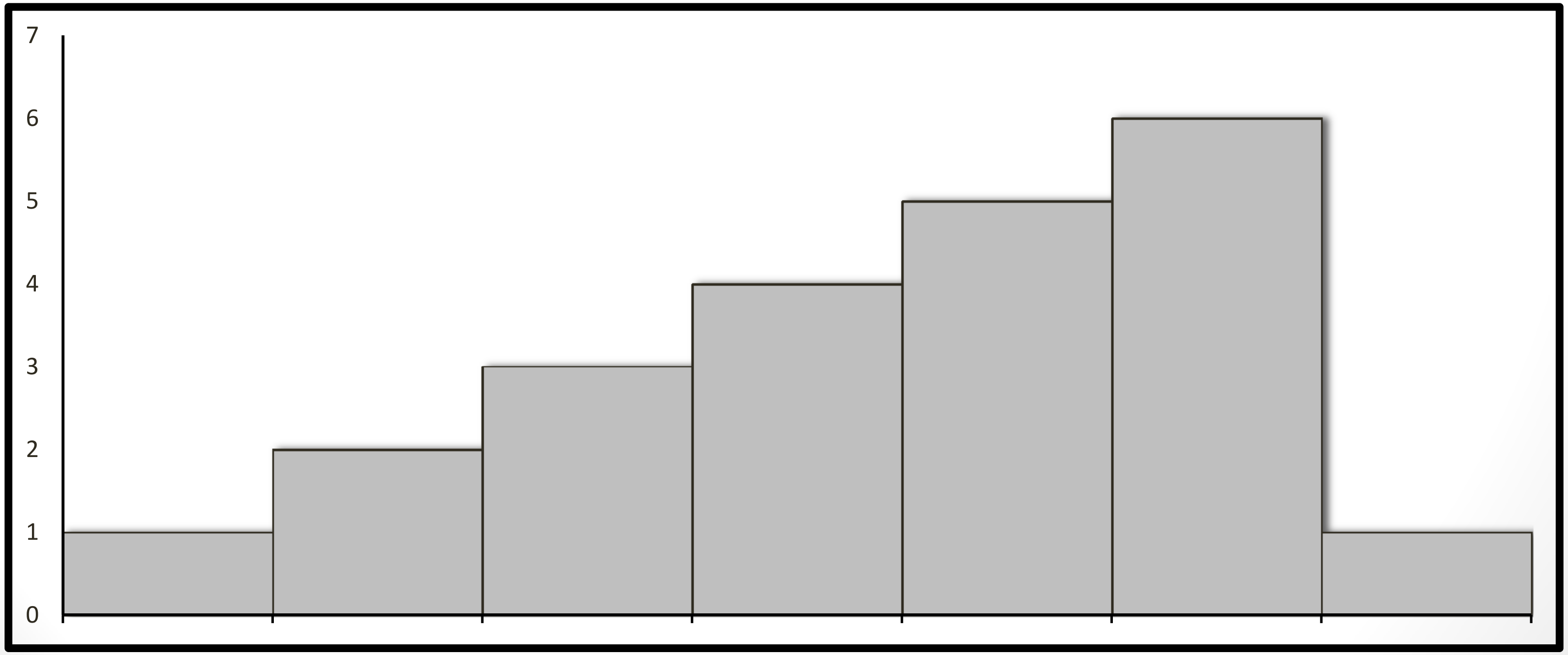
$$mean = median = mode$$



3-4 Distribution Shapes

- In a **negatively skewed** or **left skewed distribution**, the majority of the data values falls to the right of the mean and cluster at the upper end of the distribution.

$$\boldsymbol{mean < median < mode}$$



3-5 Measures of Variation

- The **range** is the highest value minus the lowest value in a data set.

$$\boldsymbol{R = Max - Min}$$

- Example:
  - Suppose that the number of burglaries reported in a specific year for nine cities are 61, 11, 1, 3, 2, 30, 18, 3, 7 Find the range.

$$\boxed{61}, 11, \boxed{1}, 3, 2, 30, 18, 3, 7$$

$$\boldsymbol{R = Max - Min = 61 - 1 = 60}$$

3-5 Measures of Variation

- The **variance** is the average of the squares of the distance each value is from the mean.

  - The symbol for the **population variance** is σ^2

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

  - The symbol for the **sample variance** is S^2

$$S^2 = \frac{\sum(x-\overline{x})^2}{n-1} = \frac{\sum x^2 - \left[\frac{(\sum x)^2}{n}\right]}{n-1}$$

3-5 Measures of Variation

- The **standard deviation** is the square root of the variance.
  - The symbol for the **population standard deviation** is $\boldsymbol{\sigma}$

$$\boldsymbol{\sigma = \sqrt{\sigma^2}}$$

  - The symbol for the **sample standard deviation** is ***S***

$$\boldsymbol{S = \sqrt{S^2}}$$

3-5 Measures of Variation

- <u>Example</u>:
  - Suppose that the number of burglaries reported in a specific year for nine cities are 61, 11, 1, 3, 2, 30, 18, 3, 7 Find the variance and the standard deviation.

x	61	11	1	3	2	30	18	3	7	$\sum x = 136$
x^2	3721	121	1	9	4	900	324	9	49	$\sum x^2 = 5138$

$$S^2 = \frac{\sum x^2 - \left[\frac{(\sum x)^2}{n}\right]}{n-1} = \frac{5138 - \left[\frac{(136)^2}{9}\right]}{9-1} = \frac{5138 - \left[\frac{18496}{9}\right]}{8}$$

$$= \frac{5138 - 2055.11}{8} = \frac{3082.89}{8} = 385.36$$

$$S = \sqrt{S^2} = \sqrt{385.36} = 19.63$$

3-5 Measures of Variation

- Variance and Standard Deviation

  - Variances and standard deviations can be used to determine the spread of the data. If the variance or standard deviation is large, the data are more dispersed.

  - The measures of variance and standard deviation are used to determine the consistency of a variable, to determine the number of data values that fall within a specified interval in a distribution and in comparing two or more data sets to determine which is more variable.

3-5 Measures of Variation

- The **coefficient of variation** is the standard deviation divided by the mean expressed as a percentage.
  - For populations

$$CVar = \frac{\sigma}{\mu} * 100\ \%$$

  - For Samples

$$CVar = \frac{s}{\bar{x}} * 100\ \%$$

- The **coefficient of variation** is used to compare standard deviations of two variables or more when the units or the values of the means are different.

- Large coefficient of variation means large variability.

3-5 Measures of Variation

- Example:
  - The average age of the employees at certain company is 30 years with a standard deviation of 5 years; the average salary of the employees is $40,000 with a standard deviation of $5000. Which one has more variation age or income?

$$\boldsymbol{CVar(age) = \frac{S}{\bar{x}} * 100 = \frac{5}{30} * 100 = 16.67\%}$$

$$\boldsymbol{CVar(income) = \frac{S}{\bar{x}} * 100 = \frac{5000}{40000} * 100 = 12.5\%}$$

Age is more variable than income.

3-6 Measure of Position

- A **<u>standard score</u>** or **<u>z score</u>** is used when direct comparison of raw scores is impossible.

- The ***<u>z score</u>*** represents the number of standard deviations a data value falls above or below the mean.

$$z = \frac{x - \bar{x}}{s}$$

3-6 Measure of Position

- Example:
  - A student scored 65 on a statistics exam that had a mean of 50 and a standard deviation of 10. Compute the *z*-score.

$$\boldsymbol{Z = \frac{x - \bar{x}}{s} = \frac{65 - 50}{10} = 1.5}$$

That is, the score of 65 is 1.5 standard deviations above the mean.

Above - since the *z*-score is positive.

3-6 Measure of Position

- <u>Example</u>:
  - Which of the following exam scores has a better relative position?

a. A score of 42 on an exam with $\overline{x} = 39$ and $S = 4$

$$Z = \frac{x - \overline{x}}{S} = \frac{42 - 39}{4} = \frac{3}{4} = 0.75$$

b. A score of 76 on an exam with $\overline{x} = 71$ and $S = 3$

$$Z = \frac{x - \overline{x}}{S} = \frac{76 - 71}{3} = \frac{5}{3} = 1.67$$

So, a score of 76 has a better relative position.

3-6 Measure of Position

- **Quartiles** divide the distribution into four groups, denoted by Q_1, Q_2, Q_3. Note that Q_1 is the same as the 25th percentile, Q_2 is the same as the 50th percentile or the median and Q_3 corresponds to the 75th percentile.

- **Quartiles** can be found as follow
  1. Arrange the data in order from lowest to highest.
  2. Find the median of the data values (Q_2).
  3. Find the median of the data values that fall bellow Q_2 (Q_1).
  4. Find the median of the data values that fall above Q_2 (Q_3).

3-6 Measure of Position

- Example:
  - Suppose that the number of burglaries reported in a specific year for nine cities are 61, 11, 1, 3, 2, 30, 18, 3, 7 Find the first, second and third quartile.

$$1,\ \boxed{2,\ 3},\ 3,\ \boxed{7},\ 11,\ \boxed{18,\ 30},\ 61$$

$$\boxed{\boldsymbol{Q_2 = 7}}$$

$$\boxed{\boldsymbol{Q_1 = 2.5}} \qquad \boxed{\boldsymbol{Q_3 = 24}}$$

3-6 Measure of Position

Outliers

- An **<u>outlier</u>** is an extremely high or an extremely low data value when compared with the rest of the data values.

- **<u>Outliers</u>** can be identified as follows:
  1. Arrange the data in order and find Q_1 and Q_3.
  2. Find the **<u>interquartile range</u>**: $IQR=Q_3-Q_1$.
  3. The values that are **<u>smaller</u>** than $Q_1-(1.5)(IQR)$ or **<u>larger</u>** than $Q_3+(1.5)(IQR)$ are called outliers.

- Outliers can be the result of measurement or observational error.

3-6 Measure of Position

- Example:
  - Suppose that the number of burglaries reported in a specific year for nine cities are 61, 11, 1, 3, 2, 30, 18, 3, 7 Find the outlier values if any.

$$1, \boxed{2, 3}, 3, \boxed{7}, 11, \boxed{18, 30}, 61$$

$$\boxed{\boldsymbol{Q_2 = 7}}$$

$$\boxed{\boldsymbol{Q_1 = 2.5}} \qquad \boxed{\boldsymbol{Q_3 = 24}}$$

$$\boldsymbol{IQR = Q_3 - Q_1 = 24 - 2.5 = 21.5}$$
$$\boldsymbol{Q_3 + (1.5)(IQR) = 24 + (1.5)(21.5) = 24 + 32.25 = 56.25}$$
$$\boldsymbol{Q_1 - (1.5)(IQR) = 2.5 - (1.5)(21.5) = 2.5 - 32.25 = -29.75}$$

$$1, 2, 3, 3, 7, 11, 18, 30, \boxed{61}$$

61 is an outlier value in this data.

3-7 Exploratory Data Analysis

- The purpose of **exploratory data analysis** is to examine data in order to find out what information can be discovered such as the center and the spread.

- **Boxplots** are graphical representations of a **five-number summary** of a data set. The five specific values that make up a five-number summary are

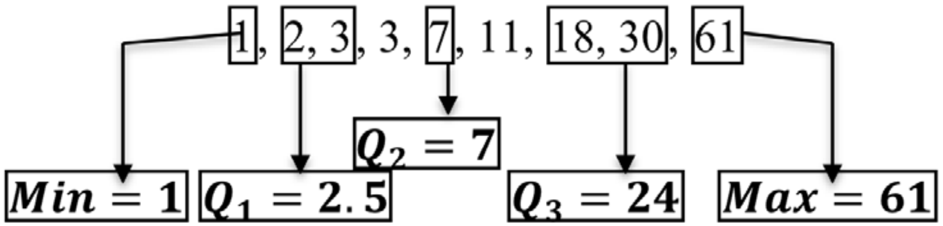
$$Min,\ Q_1,\ Q_2, Q_3,\ Max$$

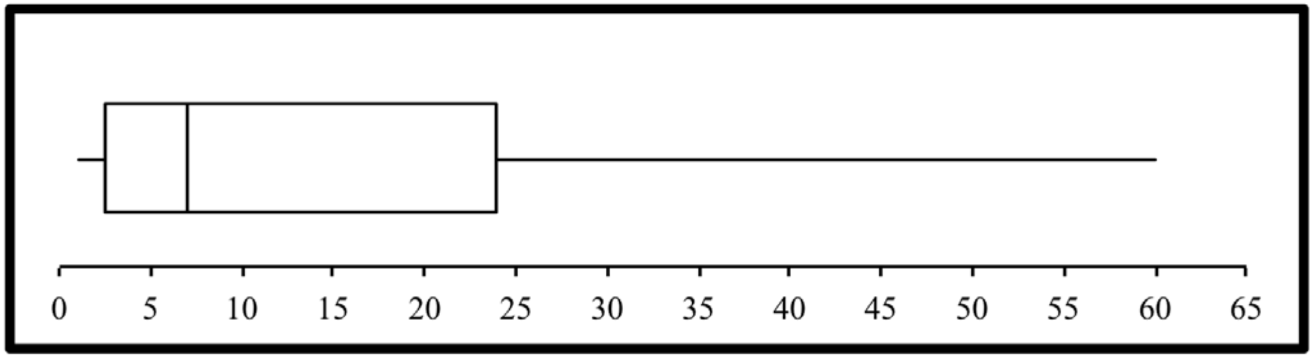
3-7 Exploratory Data Analysis

- Information obtained from a boxplot
  - Using the box:
    - If the median is near the center of the box, the distribution is approximately symmetric.
    - If the median is to the left of the box, the distribution is positively skewed.
    - If the median is to the right of the box, the distribution is negatively skewed.
  - Using the lines:
    - If the lines are about the same length, the distribution is approximately symmetric.
    - If the right line is taller, the distribution is positively skewed.
    - If the left line is taller, the distribution is negatively skewed.

3-7 Exploratory Data Analysis

- <u>Example</u>:
  - Suppose that the number of burglaries reported in a specific year for nine cities are 61, 11, 1, 3, 2, 30, 18, 3, 7 Find the five-number summary and comment on the skewness of the data.





  - The distribution is positively skewed.