Chapter 3

Data Description

Objectives

☐ Summarize data using measures of central tendency, such as the mean, median, mode, and midrange.

☐ Describe data using the measures of variation, such as the range, variance, and standard deviation.

Objectives

☐ Identify the position of a data value in a data set using various measures of position, such as standard scores and quartiles.

☐ Use the techniques of exploratory data analysis, including boxplots and five-number summaries to discover various aspects of data.

Notes

-----------------------------------------------

Dr. Saeed Alghamdi, Statistics Department, Faculty of Sciences, King Abdulaziz University
Introduction

- Statistical methods can be used to summarize data.
- Measures of average are also called *measures of central tendency* and include the mean, median, mode, and midrange.
- Measures that determine the spread of data values are called *measures of variation* or *measures of dispersion* and include the range, variance, and standard deviation.

**Notes**

- Measures of position tell where a specific data value falls within the data set or its relative position in comparison with other data values.
- The most common measures of position are *standard scores* and *quartiles*.
- The measures of central tendency, variation, and position are part of what is called *traditional statistics*.

**Notes**

- Another type of statistics is called *exploratory data analysis* which include the *box plot* and the *five-number summary*.
- A *statistic* is a characteristic or measure obtained by using the data values from a sample.
- A *parameter* is a characteristic or measure obtained by using all the data values for a specific population.
The mean is the sum of the values divided by the total number of values. The Greek letter μ (mu) is used to represent the population mean.

\[ \mu = \frac{\sum X}{N} = \frac{X_1 + X_2 + \ldots + X_N}{N} \]

The symbol \( \bar{X} \) ("x-bar") represents the sample mean.

\[ \bar{X} = \frac{\sum x}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n} \]

Example: Suppose that the number of burglaries reported for a specific year for nine communities are 61, 11, 1, 3, 2, 30, 18, 3, 7

Find the mean.

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

* See examples 3-1 and 3-2

The median is the halfway point in a data set. The symbol for the median is MD. The median is found by arranging the data in order and selecting the middle point. Example: Suppose that the number of burglaries reported for a specific year for nine communities are 61, 11, 1, 3, 2, 30, 18, 3, 7

Find the median.

\[ MD: 1, 2, 3, 3, 7, 11, 18, 30, 61 \], So \( MD = 7 \)

* See examples 3-4, 3-5, 3-6 and 3-8
The value that occurs most often in a data set is called the **mode**. A data set with one value that occurs with greatest frequency is said to be **unimodal**. A data set with two values that occur with greatest frequency is said to be **bimodal**. A data set with more than two values that occur with greatest frequency is said to be **multimodal**. When all the values in a data set occur with the same frequency is said to have **no mode**.

---

Example: Suppose that the number of burglaries reported for a specific year for nine communities are 61, 11, 1, 3, 2, 30, 18, 3, 7

Find the mode.

* **Mode = 3**

---

The **midrange** 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. The symbol is MR.

Example: Suppose that the number of burglaries reported for a specific year for nine communities are 61, 11, 1, 3, 2, 30, 18, 3, 7

Find the midrange.

* **MR = \( \frac{61 + 3}{2} = 31 \)**

---

* See examples 3-9, 3-10, 3-11 and 3-13

---

* See examples 3-15 and 3-16
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.

$$\bar{X}_w = \frac{\sum wX}{\sum w}$$

Where $w_1, w_2, \ldots, w_n$ are the weights for $X_1, X_2, \ldots, X_n$.

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.

$$\bar{X}_w = \frac{\sum wX}{\sum w} = \frac{90 \times 3 + 70 \times 3 + 80 \times 4 + 60 \times 2}{3 + 3 + 4 + 2} = \frac{766}{12} = 6.67$$

* See example 3-17

One computes the mean by using all the values of a data and used in computing other statistics.

The mean varies less than the median or mode when samples are taken from the same population and all three measures are computed for these samples.

The mean for the data set is unique, and not necessarily one of the data values.
Properties of Central Tendency Measures

- The **mean** cannot be computed for an open-ended frequency distribution.
- The **mean** is affected by extremely high or low values and may not be the appropriate average to use in these situations.
- The **median** is used when one must find the center or middle value of a data set.

Properties of Central Tendency Measures

- The **median** is used when one must determine whether the data values fall into the upper half or lower half of the distribution.
- The **median** is used to find the average of an open-ended distribution.
- The **median** is affected less than the **mean** by extremely high or extremely low values.

Properties of Central Tendency Measures

- The **mode** is used when the most typical case is desired.
- The **mode** is the easiest average to compute.
- The **mode** can be used when the data are nominal, such as religious preference or gender.
- The **mode** is not always unique. A data set can have more than one mode, or the mode may not exist for a data set.
Properties of Central Tendency Measures

- The **midrange** is easy to compute.
- The **midrange** gives the midpoint.
- The **midrange** is affected by extremely high or low values in a data set.

Distribution Shapes

- In a **positively skewed** or **right skewed distribution**, the majority of the data values fall to the left of the mean and cluster at the lower end of the distribution. 
  \[\text{mode} < \text{median} < \text{mean}\]

- In a **symmetrical distribution**, the data values are evenly distributed on both sides of the mean. 
  \[\text{mean} = \text{median} = \text{mode}\]
Distribution Shapes

- In a negatively skewed or left skewed distribution, the majority of the data values fall to the right of the mean and cluster at the upper end of the distribution. \( \text{mean} < \text{median} < \text{mode} \)

Measures of Variation

- The range is the highest value minus the lowest value in a data set. 
  \[ R = \text{highest value} - \text{lowest value} \]
- Example: Suppose that the number of burglaries reported for a specific year for nine communities are 61, 11, 3, 2, 30, 18, 3, 7. Find the range. \( R = 61 - 1 = 60 \)
  
  * See examples 3-18, 3-19 and 3-20

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 \( \sigma^2 \).
  \[ \sigma^2 = \frac{\sum (x - \mu)^2}{N} \]
- The symbol for the sample variance is \( s^2 \)
  \[ s^2 = \frac{\sum (x - \overline{x})^2}{n - 1} = \frac{\sum x^2 - \left[ \sum x \right]^2 / n}{n - 1} \]
Measures of Variation

- The **standard deviation** is the square root of the variance. The symbol for the population standard deviation is \( \sigma \).
  \[
  \sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}
  \]
- The symbol for the sample standard deviation is \( s \).
  \[
  s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}
  \]

Example: Suppose that the number of burglaries reported for a specific year for nine communities are 61, 11, 3, 2, 30, 18, 3, 7.

Find the variance and the standard deviation.

\[
\begin{align*}
\sum x^2 & = 5721 \\
\sum x & = 136 \\
\end{align*}
\]

\[
s^2 = \frac{\sum x^2 - \left( \frac{\sum x^2}{n} \right)}{n - 1} = \frac{5721 - \left( \frac{136^2}{9} \right)}{8} = 385.361 \\
s = \sqrt{s^2} = \sqrt{385.361} \approx 19.63062
\]

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 information is useful in comparing two or more data sets to determine which is more variable.
- The measures of variance and standard deviation are used to determine the consistency of a variable.
Variance and Standard Deviation

- The variance and standard deviation are used to determine the number of data values that fall within a specified interval in a distribution.
- The variance and standard deviation are used quite often in inferential statistics.

Measures of Variation

- The coefficient of variation is the standard deviation divided by the mean. The result is expressed as a percentage.
  - For populations: $CVar = \frac{s}{\mu} \times 100\%$
  - For Samples: $CVar = \frac{S}{X} \times 100\%$

Measures of Variation

- The coefficient of variation is used to compare standard deviations when the units are different for the two variables being compared.
- Large coefficient of variation means large variability.
Measures of Variation

- Example: The average age of the employees at a 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? 
  \[ CV_{\text{age}} = 16.67\% \]
  \[ CV_{\text{income}} = 12.5\% \]
- Hence, age is more variable than income.

* See examples 3-25 and 3-26

Dr. Saeed Alghamdi, Statistics Department, Faculty of Sciences, King Abdulaziz University

Measure of Position

- A standard score or \( z \) score is used when direct comparison of raw scores is impossible.
- The \( z \) score represents the number of standard deviations a data value falls above or below the mean.
  \[ z = \frac{x - \mu}{\sigma} \]

Dr. Saeed Alghamdi, Statistics Department, Faculty of Sciences, King Abdulaziz University

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.

\[ z = \frac{(65 - 50)}{10} = 1.5 \]

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

Above - since the \( z \)-score is positive.

Dr. Saeed Alghamdi, Statistics Department, Faculty of Sciences, King Abdulaziz University
Measure of Position

Example:
Which of the following exam scores has a better relative position?

a. A score of 42 on an exam with $\bar{X} = 39$ and $s = 4$
   \[ z = \frac{42 - 39}{4} = \frac{3}{4} \]

b. A score of 76 on an exam with $\bar{X} = 71$ and $s = 3$
   \[ z = \frac{76 - 71}{3} = \frac{5}{3} \]

So a score of 76 has a better relative position

* See examples 3-29 and 3-30

Dr. Saeed Alghamdi, Statistics Department, Faculty of Sciences, King Abdulaziz University

---

Measures 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.

Dr. Saeed Alghamdi, Statistics Department, Faculty of Sciences, King Abdulaziz University

---

Measures of Position

- **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 below $Q_2$ ($Q_1$).
  4. Find the median of the data values that fall above $Q_2$ ($Q_3$).

Dr. Saeed Alghamdi, Statistics Department, Faculty of Sciences, King Abdulaziz University
Measures of Position

Example: Suppose that the number of burglaries reported for a specific year for nine communities are 61, 11, 1, 3, 2, 30, 18, 3, 7

Find the first, second and third quartile.

\[
\begin{align*}
Q_1 &= \frac{1^\text{st} + 2^\text{nd}}{2} = \frac{1 + 3}{2} = 2 \\
Q_2 &= \frac{3^\text{rd} + 4^\text{th}}{2} = \frac{30 + 18}{2} = 24 \\
Q_3 &= \frac{7^\text{th} + 8^\text{th}}{2} = \frac{61 + 30}{2} = 45.5
\end{align*}
\]

* See example 3-36

Notes

3-37

Outliers

An outlier is an extremely high or an extremely low data value when compared with the rest of the data values.

Outliers can be identified as follows

1. Arrange the data in order and find Q1 and Q3.
2. Find the interquartile range: \(IQR = Q_3 - Q_1\).
3. The values that are smaller than \(Q_1 - 1.5(IQR)\) or larger than \(Q_3 + 1.5(IQR)\) are called outliers.

Outliers can be the result of measurement or observational error.

Example: Suppose that the number of burglaries reported for a specific year for nine communities are 61, 11, 1, 3, 2, 30, 18, 3, 7

Find the outlier values if any.

\[
\begin{align*}
Q_1 &= 2.5 \quad \text{and} \quad Q_3 = 24 \quad \Rightarrow \quad IQR = 24 - 2.5 = 21.5 \\
\text{So,} \quad Q_1 - 1.5(IQR) &= 24 + 1.5(21.5) = 56.25 \\
Q_3 + 1.5(IQR) &= 2.5 - 1.5(21.5) = -29.75
\end{align*}
\]

Hence 61 is an outlier value in this data.

* See example 3-37
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 minimum, $Q_1$, $Q_2$, $Q_3$, and maximum.

Example: Suppose that the number of burglaries reported for a specific year for nine communities are 61, 11, 1, 3, 2, 30, 18, 3, 7

Construct a boxplot and comment on the skewness of the data.

$Min = 1, Q_1 = 2.5, Q_2 = 7, Q_3 = 24, Max = 61$

* See example 3-39