Chapter 3 Data Description

3.1 Introduction

When we exam the graph (distribution) of the data, we usually look for its shape, center, spread, data positions, and outliers.

A distribution can be symmetric or skewed to the right or skewed to the left or in other types of shapes.

3.2 Measures of Center Tendency

Definition. 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 from a population.

Measures of Center Tendency: Mean, Median, Mode, and Midrange

Definition. If \( n \) observations are denoted by \( x_1, x_2, \ldots, x_n \), their mean is
\[
\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}
\]
or in more compact notation
\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.
\]
**Definition.** The *median* of a data set is the “middle value.” To find the median:

1. Arrange all observations in order of size, from smallest to largest.

2. If the number of observations $n$ is odd, the median $M$ is the center observation in the ordered list. Find the location of the median by counting $(n + 1)/2$ observations up from the bottom of the list.

3. If the number of observations $n$ is even, the median $M$ is the mean of the two center observations in the ordered list. The location of the median is again $(n + 1)/2$ from the bottom of the list.

**Example.** The data are arranged in increasing order as follows:

\[ 9 \ 9 \ 22 \ 32 \ 33 \ 39 \ 39 \ 42 \ 49 \ 52 \ 58 \ 70 \]

The median of this data set is 39.

If two 9’s are dropped, the median will be 40.5.

**Note.** The mean and median of a symmetric distribution are close together. In a skewed distribution, the mean is farther out in the long tail than is the median. This is because a few outliers can changed the mean, but may have no effect on the median.
Measuring Spread: The Quartiles

**Definition.** The value that occurs most often in a data set is called the mode. A data set that has only one value that occurs with the greatest frequency is said to be **unimodal**. If it has two values that are considered as the mode, then the data set is said to be bimodal. Similarly we define a **multimodal** data sets.

**Definition.** The *range* of a data set is the difference between the largest and smallest observations. The *midrange* is a rough estimate of the middle. It is defined as the average of the highest value and the lowest value of the data set ($MR = \frac{\text{Lowest value} + \text{Highest value}}{2}$). The *weighted mean* of a variable $X$ is obtained by multiplying each value by its corresponding weight and dividing the sum of products by the sum of the weights:

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 \cdots + w_n} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}.$$  

**Example.** Grade Point Average calculation.
3.3 Measures of Variation (Spread of the data)

**Definition.** The sample variance $s^2$ of a set of the sample data is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of $n$ observations $x_1, x_2, \ldots, x_n$ is

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n - 1}$$

or more compactly,

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

The **sample standard deviation** $s$ is the square root of the variance $s^2$:

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

The **population variance** $\sigma^2$ is the average of the squares of the deviations of the data values from the population mean $\mu$. In symbols, the population variance is

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}$$

or more compactly,

$$\sigma^2 = \frac{1}{n} \sum_{i} (x_i - \mu)^2.$$

The **population standard deviation** $\sigma$ is the square root of the variance $\sigma^2$:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i} (x_i - \mu)^2}.$$
Example. Consider the sample data set:

1792 1666 1362 1614 1460 1867 1439

The mean is $\bar{x} = 1600$. We can calculate the sample variance as:

<table>
<thead>
<tr>
<th>Observations</th>
<th>Deviations</th>
<th>Squared Deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i$</td>
<td>$x_i - \bar{x}$</td>
<td>$(x_i - \bar{x})^2$</td>
</tr>
<tr>
<td>1792</td>
<td>1792 - 1600 = 192</td>
<td>$192^2 = 36,864$</td>
</tr>
<tr>
<td>1666</td>
<td>1666 - 1600 = 66</td>
<td>$66^2 = 4,356$</td>
</tr>
<tr>
<td>1362</td>
<td>1362 - 1600 = -238</td>
<td>$(-238)^2 = 56,644$</td>
</tr>
<tr>
<td>1614</td>
<td>1614 - 1600 = 14</td>
<td>$14^2 = 196$</td>
</tr>
<tr>
<td>1460</td>
<td>1460 - 1600 = -140</td>
<td>$(-140)^2 = 19,600$</td>
</tr>
<tr>
<td>1867</td>
<td>1867 - 1600 = 267</td>
<td>$267^2 = 71,289$</td>
</tr>
<tr>
<td>1439</td>
<td>1439 - 1600 = -161</td>
<td>$(-161)^2 = 25,921$</td>
</tr>
</tbody>
</table>

sum = 0  sum = 214,870

So the sample variance is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{6}(214,870) = 35,811.67.$$

The sample standard deviation is

$$s = \sqrt{35,811.67} = 189.24.$$

Example 3-24 on p.127 in the book shows how to do calculation of variance and standard deviation from a grouped data.
Note. Some properties of the standard deviation are:

• $s$ measures spread about the mean and should be used only when the mean is chosen as the measure of center.

• $s = 0$ only when there is no spread. This happens only when all observations have the same value. Otherwise $s > 0$. As the observations become more spread out about their mean, $s$ gets larger.

• $s$, like the mean $\bar{x}$, is strongly influenced by extreme observations. A few outliers can make $s$ very large.

• A rough estimate of the standard deviation is $s \approx \frac{\text{range}}{4}$.

**Chebyshev’s Theorem** The proportion of values from a data set that will fall within $k$ standard deviations of the mean will be at least $1 - \frac{1}{k^2}$. 
3.4 Measures of Positions

**Definition.** A *z-score* or *standard score* for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation:

\[ z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}. \]

For samples, \( z = \frac{x - \bar{x}}{s} \). For population, \( z = \frac{x - \mu}{\sigma} \).

**Examples 3-29 and 3-30** on p.140 in the book use the \( z \)-scores to determine corresponding positions based on an exam score.

**Example.** If a student scored 74 points on a test where the mean score was 81 and the standard deviation was 5. Then what was the student’s \( z \)-score?

**Answer.** \(-1.40\).

**Example.** If the mean of a set of data is 20.00, and 14.00 has a \( z \)-score of \( 1.00 \), then what is the standard deviation?

**Answer.** 6.00.

**Example.** If a student scored 74 points on a test where the mean score was 80 and the standard deviation was 6. What was the student’s \( z \)-score?

**Answer.** \(-1.00\).
**Definition.** Percentile divides the data set into 100 equal groups. The percentile corresponding to a given value $X$ is computed by using the following formula:

$$\text{Percentile} = \frac{(\text{number of values below} X) + 0.5}{\text{total number of values}} \cdot 100\%$$

**Example.** Given the following data set, find the value that corresponds to the 75th percentile. 10, 44, 15, 23, 14, 18, 72, 56.

**Answer.** 50.

3.5 Exploratory Data

**Definition.** The first quartile $Q_1$ lies one-quarter of the way up the list, the third quartile $Q_3$ lies three-quarters of the way up the list.

**Note.** To calculate the quartiles:

1. Arrange the observations in increasing order and locate the median $M$ in the ordered list of observations.

2. The first quartile $Q_1$ is the median of the observations whose position in the ordered list is to the left of the location of the overall median.

3. The third quartile $Q_3$ is the median of the observations whose position in the ordered list is to the right of the location of the overall median.
Example. We saw above that the median of the data set:

\[9 \ 9 \ 22 \ 32 \ 33 \ 39 \ 39 \ 42 \ 49 \ 52 \ 58 \ 70\]

is 39. The first quartile is the median of the 6 observations to the left of the median, and so \(Q_1 = 27\). Similarly, \(Q_3 = 50.5\).

For the data:

\[22 \ 25 \ 34 \ 35 \ 41 \ 46 \ 46 \ 46 \ 47 \ 49 \ 54 \ 54 \ 59 \ 60\]

we have \(Q_1 = 35\) and \(Q_3 = 54\).

The Five-Number Summary and Boxplots

Definition. The five-number summary of a data set consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols the five-number summary is:

Minimum \(Q_1\) \(M\) \(Q_3\) Maximum

A boxplot is a way to graphically represent the five-number summary.

Note. The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution. Use \(\overline{x}\) and \(s\) only for reasonably symmetric distributions.

Example. Make a boxplot for the following data set. 24, 15, 34, 92, 67, 34, 78, 45, 53, 67, 83, 46.