Lecture 2

Reminder:
Home work 1 will be due on September 7.

Material Covered in This Lecture:
* Chapter 1, Section 1.2: Describing Distributions with Numbers.
* Minitab Demonstration

A brief description of a distribution should include its shape and numbers describing its center and spread.

Shape: Histogram or Stemplot.
Center ?
Spread ?

1. Measuring center

Example 1 (Example 1.12, Table 1.10): Fuel consumption (miles per gallon) of two-seater cars.

<table>
<thead>
<tr>
<th>17</th>
<th>20</th>
<th>20</th>
<th>17</th>
<th>18</th>
<th>12</th>
<th>11</th>
<th>10</th>
<th>17</th>
<th>60</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>15</td>
<td>12</td>
<td>22</td>
<td>16</td>
<td>13</td>
<td>20</td>
<td>20</td>
<td>15</td>
<td>26</td>
<td></td>
</tr>
</tbody>
</table>

(1). Mean: If the n observations are \( x_1, x_2, \ldots, x_n \), then their mean is defined as

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}
\]

Example 1 (continued): Calculate the mean of the fuel consumption data.

(2). Median: The median (M) is the midpoint of a distribution, the number such that half the observations are smaller and the other half are larger.

The steps to find the median of a data set.
   a. Arrange all observations in an increasing order.
b. If \( n \) is odd, then the median is the observation in the \((n+1)/2\)-location (Begin with the smallest number). If \( n \) is even, then the median is the average of the two observations in the \(n/2\)-th and \(n/2+1\)-th location.

*Example1 (continued):* Find the median of the fuel consumption data.

a. Arrange the data in an increasing order.

9 9 10 11 12 12 13 15 15 16 17 17 17 18 20 20 20 22 26 60

b. \( n = 21 \).

Question: How about if we remove 60 from the data set?

(3). Mean and Median

- They are the most common measures of the center of the data set.
- If the distribution is symmetric, mean and median are close together; If the distribution is skewed to the left, mean is less than median; If the distribution is skewed to the right, mean is larger than median.
- Median is robust or resistant to the outliers, mean is not robust.


**Definitions:**

*Percentile:* The \( p \)-th percentile of a distribution is the value such that \( p \) percent of the observations fall at or below it.

*First Quartile (Q1):* 25-th percentile.

*Second Quartile (M):* 50-th percentile.

*Third Quartile (Q3):* 75-th percentile.

*Interquartile Range (IQR):* Q3-Q1.

**How to find Q1 and Q3?**

a. Arrange the data in an increasing order.

b. Q1 is the median of the observations to the left of the overall median (NOT include the overall median).

c. Q3 is the median of the observations to the right of the overall median (NOT include the overall median).

*Example2 (Example 1.14, p.45).* Fuel consumption data (two-seater cars)

13 15 16 16 17 19 20 22 23 23 24 25 25 26 28 28 28 29 32 66
Find Q1, M, Q3, and IQR.

**1.5×IQR Rule for Outliers:** All observations falling more than $1.5×$IQR above Q3 and below Q1 are outliers.

*Example 2 (continued):*

**Five-Number Summary:** Minimum, Q1, M, Q3, Maximum

*Example 2 (continued):* Find the five number-summary of the fuel consumption data (two-seater cars)

**Boxplot:** A boxplot is a graph of the five-number summary.

- A central box spans the Q1 and Q3
- A line the box marks M
- Lines extend from the box out to the minimum and maximum

*Example 2 (continued):* Construct the boxplot of the fuel consumption data (two-seater cars)

**Modified Boxplot:**

Lines extend from the box out to the minimum and maximum within the interval [Q1-$1.5×$IQR, Q3+$1.5×$IQR], observations outside this interval denoted by some special symbol, e.g., star or circle.

(1). Interquartile Range (IQR): Q3-Q1

(2). Variance ($s^2$): The variance of n observations $x_1, x_2, …, x_n$ is

Formula:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + … + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

*Example 3:* Compute the variance of the following data set. 

$$4 \quad 2 \quad 3 \quad 3 \quad 6 \quad 3$$
(3). Standard Deviation (s):

\[
(s = \sqrt{s^2})
\]

*Example (continued)*: Find the standard deviation of the data set in Example 3.

Remark: IQR is robust, standard deviation is not.

Homework problems: 1.44; 1.59. (You may use Minitab or other software to draw the graphs)