3.1 Measures of Center

Measures of central location:

- Mean

Defined as

The mean is:

- easy to compute and interpret

- generally the best measure of central location

- influenced by extreme observations.
Ex. Number of days six patients of heart transplant survived:

3  15  46  64  64  623

The sample mean is

\[ \overline{x} = \frac{815}{6} = 135.8 \]

• Median

Defined as observations when arranged from smallest to largest.
Ex. Transplant data

3  15  46  64  64  623

Median =

Notice:

- If # of observations is odd, the median is the

- If # of observations is even, then the median is the
• **Mode**

  Defined as the value that

  Ex. Transplant data

  3  15  46  64  64  623

  Hence, Mode =

  **Note:**

  - If greatest frequency is 1, then the data set
  - If greatest frequency is 2 or greater, then any value that occurs with that
Ex. Transplant data.

3  15  46  64  64  623

Mean   = 135.8
Median = 55
Mode   = 64

Which of the three is a better measure of central location?

- The mode is different from the mean and median.

- The median is a

(See example 3.4 in textbook)
Relationship between mean and median

1. If the distribution is symmetrical:

2. If the distribution is skewed to the right (positively skewed):

3. If the distribution is skewed to the left (negatively skewed):
• Population and Sample Mean

The mean of a population is called the **population mean**.

The mean of a sample is called the **sample mean**.

**IMPORTANT!!**

For a particular population:
3.2 The sample mean

First, need a symbol to represent summation.

$$\sum_{j=1}^{n} x_j =$$

Ex. Let $x_1 = 2$, $x_2 = 3$ and $x_3 = 4$ then

$$\sum_{j=1}^{3} x_j =$$

The sample mean of $n$ measurements $x_1, x_2, \ldots, x_n$ is defined as:

$$\bar{x} =$$
For example above:

\[ \bar{x} = \]

Other important Sums:

Ex. \( x_1 = 2, \ x_2 = 3 \) and \( x_3 = 4 \) then

1. (Sum of squares)

\[ \sum_{i=1}^{3} x_i^2 = \]

2. (Sum of deviations from mean)

\[ \sum_{i=1}^{3} (x_i - \bar{x}) = \]
3. (Sum of squared deviations from mean)
\[ \sum_{i=1}^{3} (x_i - \bar{x})^2 = \]

3.3 Measures of Variation; The Sample Standard Deviation

Ex. We have seven comparable fast food bars sample from cities A and B. Net profit (cents) per dollar sales are:

A: 5 4 6 5 5 3 7
B: 5 9 -4 23 5 -10 7

Both samples have the same mean, median and mode (=5), yet they are very different.
Dotplot: City A, City B

Measures of variation:
• **Range** - pg106 (104)

    Range =

    **Ex. City A:**  Largest value =  Smallest value =

        Range =

    **City B:**

        Range =

    The range is:

        - easy to compute & interpret

        - provides no info about dispersion of values between smallest and largest

        - susceptible to outliers (extreme values).
Sample Standard Deviation - pg107 (105)

Measures the extent to which values differ from the mean:

\[(x_i - \bar{x}).\]

Summing over all data values:

\[\sum_{i=1}^{n} (x_i - \bar{x}) = \text{always!}\]

Instead, look at squared differences.
Find the sum of squared deviations:

and finally divide by \( n-1 \) (average of squared deviations).

**Variance of a sample** of \( n \) measurements having mean \( \bar{x} \) is:

\[
S^2 = \frac{\text{sum of squared deviations}}{n-1}
\]
Ex. City A

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$(x_i - \bar{x})$</th>
<th>$(x_i - \bar{x})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

Hence, $\bar{x} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
Units of variance:
- Not very useful!

- **Standard Deviation**

  of the variance of the measurements.

**Sample standard deviation**

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

Ex. City A: \(s_A = \)

City B: \(s_B = \)

(Units are the same as the original data.)
• Shortcut/Computational formula for sample variance

\[
s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1}
\]

Ex. For City A:

\[
\sum_{i=1}^{7} x_i^2 = \quad ,
\]

\[
\sum_{i=1}^{7} x_i = \quad ,
\]

\[
s^2 = \quad = 1.66
\]
• Computing formula for sample SD

\[ s = \sqrt{\frac{\sum x_i^2 - \left( \frac{\sum x_i}{n} \right)^2}{n-1}} \]

Note: SD is a measure of variation. The more variation it is in the data, the larger its SD.

Ex. \[ s_B = 10.41 \quad s_A = \sqrt{1.66} = 1.29 \]

Hence, data from city
3.4 The Five Number Summary; Boxplot

→ Now, descriptive measures based on percentiles (advantage: resistant).

The most commonly used percentiles are quartiles:

25th percentile

50th percentile

75th percentile
How do we calculate quartiles?

Let $n$ denote the # obs.
Arrange the data in increasing order.

- $Q_1$ is at position
- $Q_2$ is the median which is at position
- $Q_3$ is at position

If a position is not a whole number, linear interpolation is used -example later.

Note: The textbook calculates percentiles differently. Follow class notes which are consistent with software.
Ex. City A data: (ordered)

3 4 5 5 5 6 7

First quartile?

Observation at position

Choose

Third quartile?

Observation at position

Choose
Interquartile Range (IQR)

To avoid a single data value overly influencing the measure of dispersion, use the IQR (a resistant measure).

\[ \text{IQR} = \]

**Ex.**  
City A: \( Q_1 = 4, \quad Q_3 = 6 \)

\[ \text{IQR}_A = \]

City B: \( Q_1 = -4, \quad Q_3 = 9 \)

\[ \text{IQR}_B = \]

→ Data from City
The median, 1\textsuperscript{st} and 3\textsuperscript{rd} quartiles and smallest and largest observations are useful indicators of the dist’n of a data set.

→ display in a Boxplot.

Boxplot - City A
Boxplots are effective for displaying several samples for visual comparison.

Ex.
Outliers

Obs that fall

→ Use IQR to identify potential outliers.

Lower limit =
Upper limit =

Obs that lie outside the lower and upper limits are

Ex. The monthly rents -ordered- for 8 one-bedroom apartments, located in one area of the city are

525  540  570  580
585  585  625  770
Minitab output:

Descriptive Statistics: Rent

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>TrMean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rent</td>
<td>8</td>
<td>597.5</td>
<td>582.5</td>
<td>597.5</td>
<td>76.0</td>
<td>26.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Q1</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rent</td>
<td>525.0</td>
<td>770.0</td>
<td>547.5</td>
<td>615.0</td>
</tr>
</tbody>
</table>

First quartile for rent data?
(Need to use linear interpolation)

Obs. at position

Hence, \( Q_1 \) is in between the

\[ Q_1 = \text{between known as linear interpolation.} \]

Check: \( Q_3 = 615! \)
\[ \text{IQR} = 615 - 547.5 = 67.5 \]

- Lower limit = 
- Upper limit = 

\[ \uparrow \]

Useful to identify potential outliers.

Boxplot for rent data - (Modified Boxplot)
Descriptive Statistics in Excel

First you need to install the Analysis Took Pack in Excel

/Tools
/Add Ins
/Select Analysis Tool Pack

After Tool Pack is installed:

/Tools
/Select Data Analysis
/Select Descriptive Statistics
/Select Input Range (Col Row#:Col Row#)
   Output Range (Col Row#)
   Click box Summary Statistics
/Okay
Example: Find the descriptive statistics for the following data set:

3, 15, 46, 64, 126, 623

Input Range (A1:A6)
Output Range (D1)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td></td>
<td>Column1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>46</td>
<td>Mean</td>
<td>143.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>64</td>
<td>Standard Error</td>
<td>97.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>126</td>
<td>Median</td>
<td>65.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>623</td>
<td>Mode</td>
<td>#N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Standard Deviation</td>
<td>237.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Sample Variance</td>
<td>56452.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Kurtosis</td>
<td>5.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>Skewness</td>
<td>2.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>Range</td>
<td>620.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>Minimum</td>
<td>3.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>Maximum</td>
<td>623.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>Sum</td>
<td>877.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>Count</td>
<td>6.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.5 Descriptive Measures for Populations - Use of Samples.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Sample)</td>
<td>(Population)</td>
</tr>
</tbody>
</table>

- Mean
- Median
- Variance
- Standard Deviation

Note: Sample size:
Popul’n size:
Popul’n Mean:

\[ \mu = \]

Popul’n Variance:

\[ \sigma^2 = \frac{\sum x_i^2}{N} - \mu^2 = \ldots \]

Popul’n SD:

\[ \sigma = \sqrt{\frac{\sum x_i^2}{N} - \mu^2} \]
Standardized Variables and z-Scores

For a variable $x$, the variable

$$z = \frac{x - \mu}{\sigma}$$

is called the *standardized variable* corresponding to the variable $x$.

**Important!!**

A standardized variable always has
Example 3.25 - pg 136 (134)

Possible observations for the variable $x$:

$$x : \begin{array}{cccccc}
-1 & 3 & 3 & 3 & 5 & 5 \\
\end{array}$$

a. Assuming the data set is popul’n data, compute the mean and SD.

b. Using your answer in part (a), determine the standardized version of $x$.

c. Determine the observed value of $z$ corresponding to an observed value of $x$ of 5.

d. Obtain all possible observations of $z$.

e. Find the mean and SD of $z$ using the popul’n equations.

f. Obtain dotplots of the dist’ns of $x$ and $z$. 
a. Popul’n mean:

\[
\mu_x = \frac{18}{6} = 3
\]

Popul’n SD:

\[
\sigma_x = \sqrt{\frac{\sum x_i^2}{N} - \mu_x^2}
\]

\[
\sum_{i=1}^{6} x_i^2 =
\]

\[
\rightarrow \sigma_x = \sqrt{\sum x_i^2 - \frac{18}{6}^2} = \sqrt{4} = 2
\]

b. Standardized version of \( x \):

\[
Z =
\]
c. When $x = 5$, $z =$

d. When $x = -1$, $z =$, and

when $x = 3$, $z =$.

Hence, the standardized value for each of the observations are:

\[
\begin{array}{ccccccc}
x & : & -1 & 3 & 3 & 3 & 5 & 5 \\
\rightarrow & z & : & \quad & & & & \\
\end{array}
\]
e. Popul’n mean of standardized obs.:

\[ \mu_z = \]

Popul’n SD of standardized obs.:

\[ \sigma_z = \sqrt{\frac{\sum z_i^2}{6}} - \mu_z^2 \]

\[ \sum_{i=1}^{6} z_i^2 = \]

\[ \rightarrow \sigma_z = \sqrt{\frac{\sum z_i^2}{6}} = \sqrt{1} = 1 \]
f. Dotplot for $x$ and $z$ observations:

Dotplot: $x$, $z$

```
 .       .   
-2.0     0.0  2.0  4.0  6.0  8.0
```

Already know that

Standardizing **shifts** a dist’n, so that the new mean is

**Notice:**

A lot easier to look at standardized observations!