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

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

- **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

\[ 3 \quad 15 \quad 46 \quad 64 \quad 64 \quad 623 \]

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.

\[
\begin{array}{cccccc}
3 & 15 & 46 & 64 & 64 & 623 \\
\end{array}
\]

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:
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)

\[ \sum_{i=1}^{3} (x_i - \bar{x})^2 = \]

3.2 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

+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-City A

+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-City B

-14.0 -7.0 0.0 7.0 14.0 21.0

Measures of variation:
• **Range**

\[ \text{Range} = \]

Ex. **City A:**  
Largest value =  
Smallest value =  
\[ \text{Range} = \]

**City B:**  
\[ \text{Range} = \]

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**

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:

$$s^2 =$$
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>

$\bar{x} =$

$\sum_{i=1}^{n} (x_i - \bar{x})^2$

Hence, $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} =$
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

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

Ex. For City A:

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

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

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

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

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

Ex. \( s_B = 10.41 \) \( s_A = \sqrt{1.66} = 1.29 \)

Hence, data from city
3.3 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. \underline{City A}: \( Q_1 = 4, \quad Q_3 = 6 \)

\[
\text{IQR}_A =
\]

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

\[
\text{IQR}_B =
\]

→ Data from City
The median, $1^{st}$ and $3^{rd}$ 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

\[
Q_1 = \quad \uparrow
\]

known as linear interpolation.

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

Lower limit = 
Upper limit =

↑
Useful to identify potential outliers.

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

Need to install the Analysis Tool Pack in Excel

How to install ToolPak in Excel 2007:

First check if the Data Analysis Pak is installed -
1. Select the DATA tab at the top
2. Under the “Data” Tab check if “Data Analysis” is an option. If no proceed with installation of the add-in.

Steps for Installation of ToolPak:

1. Open Excel
2. Click the Microsoft Button at the top left of the window -after excel is opened-
3. Click the “Excel Options” tab at the bottom
4. On the left panel select “Add-Ins” and then in the Management Box select the Excel Add-ins to install. In this case “Analysis ToolPak” or “Analysis ToolPak BVA”
5. Click GO
6. From the add-ins box select “Analysis ToolPak” and “Analysis ToolPak BVA” and click OK
7. Installation begins and you will be taken back to the original spreadsheet. If installation worked correctly you will see the “Data Analysis” option available under the “Data” tab.
After Tool Pack is installed:

/Data
/Select Data Analysis
/Select Descriptive Statistics
/Select Input Range (Col Row#:Col Row#)
Output Range (Col Row#)
Click box Summary Statistics
/Okay

Example: Descriptive statistics - transplant data:

3, 15, 46, 64, 126, 623
So: Input Range (A1:A6)  
Output Range (D1)

OUTPUT:

<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>148.17</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>55.00</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>
3.4 Descriptive Measures for Populations - Use of Samples

<table>
<thead>
<tr>
<th>Statistics Parameters</th>
<th>(Sample)</th>
<th>(Population)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

\[ \mu = \]

Popul’n Variance:

\[ \sigma^2 = \]

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

Popul’n SD:

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

For a variable $x$, the variable

$$z =$$

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

**Important!!**

A standardized variable always has
Example

Possible observations for the variable $x$:

$x : -1 \ 3 \ 3 \ 3 \ 5 \ 5$

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 $z$. 

-63-
a. Popul’n mean:

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

Popul’n SD:

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

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

\[ \rightarrow \sigma_x = \sqrt{4} = 2 \]

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:

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

\[ \mu_Z = \]

Popul’n SD of standardized obs.:

\[ \sigma_Z = \sqrt{\frac{\sum z_i^2}{6} - \mu_Z^2} \]

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

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

Dotplot: $x$, $z$

---+---------+---------+---------+---------+---------+-x
  .          .          .
-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!