Section 1.1/1.2
Graphical and Numerical Summaries of Data

- **Shape of a Distribution**
  - Modes
  - Symmetric vs. Skewed
  - Outliers
- **Measures of the Center**
  - mean
  - median
- **Measures of Spread**
  - IQR
  - standard deviation
- **Choosing Summaries of Distributions**
- **Changing the Units of Measurement**

**Modes**

- **Question:** Does the distribution have one or several major peaks?
  - Look at histograms and stemplots.
- A distribution with one major peak is called **unimodal**. A distribution with two major peaks is called **bimodal**.
- Example of a bimodal distribution: scores on an exam
Symmetric vs. Skewed

- A distribution is **symmetric** if the values larger or smaller than the midpoint are mirror images of each other.

- A distribution is **skewed to the right** if the right tail (larger values) is much longer than the left tail (smaller values).

- A distribution is **skewed to the left** if the left tail (smaller values) is much longer than the right tail (larger values).
Outliers

Outliers – values that fall outside the overall pattern and are far from the bulk of the data

- Can be a result of natural variation.
- Or, can be evidence of a mistake (equipment failure, incorrect recording of an observation, etc.).

Removing an outlier?  Big Decision
Measures of the Center

Two different ideas for the “center” of a distribution - can be very different.

• **Mean** - “average value”

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

or,

\[
x = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

• **Median** - “middle value”

a) sort observations from smallest to largest

b) if \( n \) is odd (\( n = \) number of observations)

median = middle value of the sorted list

= \( \frac{n+1}{2} \)th observation up from the bottom of the list

c) if \( n \) is even

median = mean of the middle two observations
Mean vs. Median

- The median is a more resistant measure of the center of a distribution, i.e., the median is not as affected by extreme observations (long tails, outliers)

*Mean vs. Median Applet* - example of a dot plot
(http://bcz.whfreeman.com/ips4e/default.asp)

<table>
<thead>
<tr>
<th>Left Skewed</th>
<th>Symmetric</th>
<th>Right Skewed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean &lt; Median</td>
<td>Mean = Median</td>
<td>Mean &gt; Median</td>
</tr>
</tbody>
</table>
Example: Phyllis received 6 HW grades in her statistics class:
86  88  92  44  89  90

Her mean grade is:
\[
\frac{86 + 88 + 92 + 44 + 89 + 90}{6} = 81.5
\]

Her median grade is:
44  86  88  89  90  92
\[
\frac{88 + 89}{2} = 88.5
\]

Question: Does the mean, 81.5, give a good idea of her “typical” grade?

No, it is lower than all but one of her grades.

Question: What about the median, 88.5?

88.5 is more typical.
Measures of Spread

The $p^{th}$ percentile of a distribution is the value such that $p$ percent of the observations fall at or below it.

Most common percentiles: QUARTILES (25%, 50% (median), 75%)

- $Q_1$ (1st Quartile) - the median of the observations whose position in the ordered list is to the left of the location of the overall median.
- $Q_3$ (3rd Quartile) - the median of the observations whose position in the ordered list is to the right of the location of the overall median.

Five-Number Summary: Minimum $Q_1$ Median $Q_3$ Maximum

Boxplots

- Boxplots are graphs of five-number summaries.
  - A central box spans the quartiles $Q_1$ and $Q_3$
  - A line in the box marks the median.
  - Lines extend from the box out to the largest and smallest observations.

- Boxplots are good for side-by-side comparison of a few variables.
Measures of Spread

IQR vs. Standard Deviation

- **Inter Quartile Range (IQR)** = \( Q_3 - Q_1 \)
  - Resistant to outliers.
  - Not very useful for describing skewed distribution (as are all measures of spread).

- **1.5 X IQR criterion for outliers** - call an observation an outlier if it falls more than 1.5 X IQR above \( Q_3 \) or below \( Q_1 \).

Modified Boxplot: lines extend out from the central box only to the smallest and largest observations that are not suspected outliers.
Statistics 528 - Lecture 3

Variance (s²) - average of the squares of the deviations of the observations from their mean

\[ s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \left[ (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2 \right] \]

Standard deviation (s) - square root of the variance (has the same units as the data)

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

Properties of the Standard Deviation

- s measures the spread about the mean and should only be used when the mean is chosen as the measure of the center of a distribution.

- s = 0 only when all the observations take on the same values. Otherwise, s > 0.

- s, like the mean \( \bar{x} \), is not resistant to outliers. A few outliers can make s very large.
Choosing a Summary

- The median, IQR, or five-number summary are better than the mean and the standard deviation for describing a skewed distribution or a distribution with outliers.
- The mean and standard deviation should only be used for describing symmetric distributions with no outliers.
- Why should we ever use the mean and standard deviation?
  Answer: They completely specify a normal distribution which allows us to easily perform statistical inference.

Changing the Unit of Measurement

Linear Transformations: \( x_{\text{new}} = a + bx \)

- \( a \) (constant) shifts all of the values of \( x \) up or down by the same amount
- \( b \) (positive constant) changes the size of the unit of measurement

- A linear transformation will not change the shape of a distribution.
- Multiplying each observation by a positive constant \( b \) multiplies both measures of the center (mean and median) and measures of spread (IQR and standard deviation) by \( b \).
- Adding the same number \( a \) (either positive or negative) to each observation adds \( a \) to the measures of the center (mean and median) and to the quartiles (and other percentiles) but does not change measures of spread.