Histograms:

Shape, Outliers, Center, Spread
Frequency and Relative Histograms
Related to other types of graphical displays

Shape:
- Skewed left
- Bi-modal
- Symmetric
- Bell-shaped
- Skewed right
- Uniform:
- Normal (Be careful!)
How would you describe the shape of these distributions?

skewed right
1 peak from 1500 to 2499 employees
possible outlier at $13,000
Gap from $8500 to $9499 and $10500 to $12,499.
Most of the salaries are clustered between $500 to $4499.

Frequency Histogram and Relative Histogram

What's different?
What's the same?
What about this one?

Bar graph
Qualitative categories instead of numeric categories.

Comparing histograms and boxplots:
Your pulse rates

Use Geogebra to look at male and female data.

How do you find the center on each type of graph?
How do you determine the shape from each type of graph?
Which type of graph is easier to manipulate?
What is one thing you know from each type of graph that you don't know from the other type? Explain.
Stem and Leaf Plots

How do stem-and-leaf plots compare to histograms?

When is it better to use a histogram?

Advanced Stem Plots:

Many times the data does not lend itself to tens and units digits.

In many cases you will need to do some rounding.

*Example: Original data:*

2.234, 3.23525, 3.76447, 3.794, 4.252, 4.8886

*Rounded data: 2.2, 3.2, 3.8, 3.9, 4.3, 4.9*

Also, you may want to split the stems.

Finally, you may want to construct back-to-back stemplots in order to compare two distributions.
The stemplot below shows the number of hot dogs eaten by contestants in a recent hot dog eating contest.

80 | 1
60 | 2 7
50 | 2 5 8
40 | 0 2 5 9 9
30 | 5 7 9
20 | 7 9
10 | 1

Which of the following statements is true?

I. The range is 70.
II. The median is 46.

(a) I only
(b) II only
(c) I and II
(d) Neither is true.
(e) There is insufficient information to answer this question.

The dot plot shows the ages of people who are grandparents who live in my neighborhood. Describe the distribution of their ages.

How do dot plots compare to histograms?

When is it better to use a dot plot?
Center and Spread

How long does it take to get to work in New York State? Here are the travel times of 20 randomly chosen New York workers:

10  30  5  25  40  20  10  15  30  20  
15  20  85  15  65  15  60  60  40  45  

Make a stemplot of the data. Describe the distribution.

Find and interpret the mean and median.

Mean is __22.5______ minutes 
Median is ___31.25___ minutes.

The median is resistant to the outlier. The mean is not.

What would happen to the mean and median if the longest travel time was 200 minutes rather than 85 minutes?
Exploring Centers

- Create a data set of at least 10 observations so that the mean equals the median. Enter the data into the calculator. Describe the shape of the histogram.

- Modify your data set so that the mean is greater than the median. Describe the shape of the histogram.

- Modify your data set so that the mean is less than the median. Describe the shape of the histogram.

- Create a data set so that the data is skewed left, but has a mean that is greater than the median.

Eight members of the Ferris ASB have an average height of 5’8”, but none of them actually are 5’8”. Two of them are the same height. Determine three sets of possible heights of the eight members of the ASB, according to these criteria:

a. The mean is 5’8”, but the median is not.

b. The median is 5’8”, but the mean is not.

c. The mean and the median are both 5’8”.

For each of the following skewed distributions, estimate the location of the mean and median.
So, which is better: the mean or the median?

What are some situations you can think of where the distribution of values would be skewed to the right?

- income
- age
- bowling scores
- race time

Conclusion: It depends on the situation!

The mean annual temperature in San Francisco is 57 degrees. Springfield, Missouri has the same mean annual temperature.

*What types of clothes should I bring for my visit?*

- **S.F.**
  - Summer: 70°
  - Winter: 50°
- **Springfield**
  - Summer: 90-100°
  - Winter: 30-40°

Does the mean alone give us very much information?
Here's the data for the travel time for 20 workers in New York State.

Find and interpret the interquartile range.

<table>
<thead>
<tr>
<th>25</th>
<th>40</th>
<th>20</th>
<th>10</th>
<th>15</th>
<th>30</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>65</td>
<td>15</td>
<td>60</td>
<td>60</td>
<td>40</td>
<td>45</td>
</tr>
</tbody>
</table>

Median 22.5

$\text{IQR} = 27.5$

$LQ = 20$

$UQ = 47$
IQR is important, but there is a more common numerical description of spread:

**Standard deviation.**

Standard deviation (and variance) measure spread by looking at how far the observations are from their mean.

---

### Standard Deviation:

A group of elementary children were asked how many pets they have. Responses: 1, 3, 4, 4, 5, 7, 8, 9.

<table>
<thead>
<tr>
<th>Observations</th>
<th>Deviations from ( \bar{x} )</th>
<th>Squared deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-5=-4</td>
<td>((-4)^2=16)</td>
</tr>
<tr>
<td>3</td>
<td>3-5=-2</td>
<td>((-2)^2=4)</td>
</tr>
<tr>
<td>4</td>
<td>4-5=-1</td>
<td>((-1)^2=1)</td>
</tr>
<tr>
<td>4</td>
<td>4-5=-1</td>
<td>((-1)^2=1)</td>
</tr>
<tr>
<td>5</td>
<td>5-5=0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>7-5=2</td>
<td>(4)</td>
</tr>
<tr>
<td>8</td>
<td>8-5=3</td>
<td>(9)</td>
</tr>
<tr>
<td>9</td>
<td>9-5=4</td>
<td>(16)</td>
</tr>
</tbody>
</table>

\[ \text{Sum} = 0 \]
\[ \text{Sum} = 52 \]

The standard deviation is the "typical" distance of the data from the mean.
Formula for variance \( s_x^2 \)

\[
s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1} \sum (x_i - \bar{x})^2
\]

Standard deviation:

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

Standard deviation:

Measures the spread about the mean. (Is used only when the mean is chosen as the measure of center.)

Is always greater than or equal to 0.

When would standard deviation = 0?

What happens to value of standard deviation as observations are more spread out?

Has the same units of measurement as the original observations.

Is NOT resistant. A few outliers can make the standard deviation very large.
The heights (in inches) of five starters on a basketball team are 67, 72, 76, 76, and 84.

1. Find and interpret the mean.

2. Find the variance and standard deviation for the data. (Use a table as we did in the example.)

3. Interpret the meaning of the standard deviation in this setting.
Back to the question: Which measure of center and spread to choose:

In general, the median and IQR are better than the mean and standard deviation for describing a skewed distribution or a distribution with strong outliers. Use $\bar{x}$ and $S_x$ only for reasonably symmetric distributions that don’t have outliers.

Coming:
Using technology to find Standard deviation.
Finding Outliers
Understanding Normal Distributions
   Improved interpretation of standard deviation
Ages of the first 44 U.S. presidents when they were inaugurated:

<table>
<thead>
<tr>
<th>Age</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>40-44</td>
<td>2</td>
</tr>
<tr>
<td>45-49</td>
<td>7</td>
</tr>
<tr>
<td>50-54</td>
<td>13</td>
</tr>
<tr>
<td>55-59</td>
<td>12</td>
</tr>
<tr>
<td>60-64</td>
<td>7</td>
</tr>
<tr>
<td>65-69</td>
<td>3</td>
</tr>
</tbody>
</table>

Make and describe a relative frequency histogram of this data.

Ogive (cumulative frequency distribution)

Describe the data shown in the graph.
A consumer magazine recently reported the number of calories in a hot dog in each of 16 major brands of hot dogs. Create two graphical displays of your choice that effectively communicate the data.

Then, contrast the two displays. That is what are the advantages and disadvantages of each graph.

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>73</td>
<td>136</td>
<td>191</td>
<td>179</td>
</tr>
<tr>
<td>182</td>
<td>154</td>
<td>190</td>
<td>105</td>
</tr>
<tr>
<td>170</td>
<td>193</td>
<td>144</td>
<td>137</td>
</tr>
<tr>
<td>150</td>
<td>128</td>
<td>141</td>
<td>139</td>
</tr>
</tbody>
</table>