Measures of Center

- Three graphical displays for quantitative data: dotplot, stemplot, histogram
- Three important aspects of a distribution are shape, center, and variability.
  - Common shapes are symmetric, skewed to the right, and skewed to the left.
  - We’ll learn how to use numbers to measure the center (and later the variability) of a distribution.

**Example 4-1: Sleeping Times**
The dotplots below display the distributions of sleeping times for three sections of statistics students. Section 1 met at 7am, section 2 at 8am, and section 3 at 11am:

(a) Identify the observational units in this graph. Also identify and classify the explanatory and response variable.

Observational units:

Explanatory variable:  
Response variable: 

(b) How would you compare the centers of these three distributions of sleep times? Would you say that they are all similar, or do some differ noticeably from others?

(c) Concentrate on section 1 for the moment. What number might you choose if you were asked to select a single number to represent the center of this distribution? Briefly explain how you arrive at this choice.
The mean is the ordinary arithmetic average, found by adding up the values of the observations and dividing by the number of observations.
- The mean corresponds to the balance point of the graph.

The median is the value of the middle observation (once they are arranged in order).
- The median of an odd number of observations can be found in position \((\frac{n+1}{2})\), once the values are arranged in order, where \(n\) represents the sample size.
- The median of an even number of observations is defined to be the average (mean) of the middle two values, on either side of position \((\frac{n+1}{2})\).

(d) Calculate the mean of the 17 sleep times in section 1. [Hint: When you read the values from the dotplot, note that all values are multiples of .25 hours.] Mark the mean value with an “x” on the dotplot above.

(e) Determine the median of the sleep times in section 1. Explain how you arrive at this number.

(f) The mean sleep times in the other two sections are 7.523 hours and 7.000 hours. Which do you think is the mean for section 2, and which for section 3? Explain. [Do not bother to do the calculations.]

Section 2:    Section 3:

(g) Calculate the medians of the sleep times for section 2 and for section 3. [Hints: There are 26 students’ values in section 2 and 33 in section 3.]

Section 2:    Section 3:

Example 4-2: Rower Weights (cont.)
Recall the data on rower weights from Example 5-4, which can be found in the Minitab worksheet rowers04.

(a) Use Minitab to produce a dotplot of the weights (Graph> Dotplot). Describe the shape of the distribution of weights.

(b) Based on the dotplot, make an educated guess for the value of the average (mean) weight. Also make a prediction for how the mean and median of the weights will compare. Explain the reasoning behind your guess.
(c) Use Minitab to calculate the mean and median of the weights (Stat> Basic statistics> Display descriptive statistics). How do the mean and median compare? Explain why this makes sense.

(d) Eliminate the coxswain’s weight. What do you think will happen to the mean and median? Then use Minitab to recalculate the mean and median. How did the mean and median change?

(e) Eliminate the lightweight rowers’ weights. Predict how this will affect the mean and median. Then use Minitab to recalculate the mean and median. How did they change?

(f) Make the heaviest rower much heavier. How does this affect the mean and median?

- A measure is said to be **resistant** if it is not strongly affected by outliers.

(g) Which measure of center – mean or median – is resistant to outliers?

(h) Describe how the *shape* of the distribution relates to the relative locations of mean and median.
Example 4-3: Cancer Pamphlets
Researchers in Philadelphia investigated whether pamphlets containing information for cancer patients are written at a level that the cancer patients can comprehend. They applied tests to measure the reading levels for a sample of 63 cancer patients and also the readability levels for a sample of 30 cancer pamphlets (based on such factors as the lengths of sentences and number of polysyllabic words). These numbers correspond to grade levels, but patient reading levels of under grade 3 and above grade 12 are not determined exactly.

The counts in the following table indicate the number of patients at each reading level and the number of pamphlets at each readability level:

<table>
<thead>
<tr>
<th>Patients’ reading level</th>
<th>Under 3</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>Above 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count (n = 63)</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>2</td>
<td>17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pamphlets’ readability level</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count (n = 30)</td>
<td>3</td>
<td>3</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

a) Explain why the form of the data do not allow one to calculate the mean reading skill level of a patient.

b) Determine the median reading level of a patient and the median readability level of a pamphlet.

c) How do these medians compare? Are they fairly close?

d) Does the closeness of these medians indicate that the pamphlets are well matched to the patients’ reading levels? Explain.

e) What proportion of the patients do not have the reading skill level necessary to read even the simplest pamphlet in the study? Do you want to rethink your answer to (d) in light of this question?

- Measures of center do not tell the whole story.
- Consider the entire distribution.
- Think about the research question.
Example 4-4: Drawing Conclusions

(a) A real estate agent notes that the mean housing price for an area is $425,780 and concludes that half of the houses in the area cost more than that. Is this necessarily a valid conclusion? Explain. Is this likely a valid conclusion? (Think about the likely shape for a distribution of house prices.) What if this had been the median instead of the mean?

(b) A businesswoman calculates that the median cost of the five business trips that she took in a month is $600 and concludes that the total cost must have been $3000. Does this conclusion necessarily follow? Explain. What if this had been the mean instead of the median?

(c) An accountant reports that 90% of the company’s employees earn less than the mean salary. Is this possible? Is this plausible? Explain.

(d) I jokingly claim that when I moved from Pennsylvania to California, the average IQ dropped in both states. Is this really possible (in principle, anyway)? If so, under what circumstances will this occur?

(e) If three sections of a course have average exam scores of 90, 80, and 55, is the average exam score among students in all three sections necessarily equal to \((90 + 80 + 55) / 3 = 75\)? Explain.

(f) Create a hypothetical example of 10 exam scores (on a 100-point exam) with the property that the median is at least 20 points larger than the mean.