Measures of Variability

- Now we consider how to measure the variability of a distribution.
- We also develop another graphical display for quantitative data: boxplot

**Example 7-1: Fuel Efficiency**
The following dotplots display highway MPG ratings of new cars, classified by type of car:

(a) Based on these graphs, which type of car would you say has the highest variability for its MPG ratings? Which has the lowest?

Highest:   
Lowest:

- One measure of spread is the **standard deviation** (SD).
  - Roughly speaking, SD is almost the average of the individual values’ deviations from the mean.
  - SD is calculated by adding the squares of those deviations, dividing by one less than the number of observations, and finally taking the square root of that result: \( SD = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}} \).

(b) Calculate the standard deviation of the MPG ratings for sports cars. [First calculate the deviation between each value and the mean (28.2), recording these in the “dev” column of the table below. Then take the square of each of these deviations, recording these in the “sq dev” column. Finally, add those squared deviations, divide by 14, and take the square root].

<table>
<thead>
<tr>
<th>MPG</th>
<th>dev.</th>
<th>Sq. dev.</th>
<th>MPG</th>
<th>dev.</th>
<th>Sq. dev.</th>
<th>MPG</th>
<th>dev.</th>
<th>Sq. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td></td>
<td></td>
<td>26</td>
<td></td>
<td></td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td>27</td>
<td></td>
<td></td>
<td>31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td>28</td>
<td></td>
<td></td>
<td>31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td>29</td>
<td></td>
<td></td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td></td>
<td></td>
<td>29</td>
<td></td>
<td></td>
<td>38</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(c) The other two car types have MPG rating standard deviation of 2.52 and 1.40. Which do you think goes with which car type? Explain.
(d) What is the smallest value a standard deviation can ever have? Under what circumstances will it have this value? Explain.

- An alternative measure of variability is the **inter-quartile range** (IQR).
- IQR is calculated as the difference between the quartiles.
  - The **lower (first) quartile** (Q1) is the 25\(^{th}\) percentile.
    - Calculated as the median of the values that fall below the position of the median.
  - Similarly, the **upper (third) quartile** (Q3) is the 75\(^{th}\) percentile.
    - Calculated as the median of the values that fall above the position of the median.
  - The IQR is therefore the range of middle 50\% of the data

(e) Calculate the quartiles and then the IQR of the MPG ratings for the sports cars. (Recall that there are \( n = 15 \) sports cars).

(f) Determine the IQR of the MPG ratings for the small cars (with \( n = 25 \) observations) and for the family cars (with \( n = 28 \) observations).

(g) Is SD resistant to outliers? What about IQR? Explain.

**Example 7-2: February Temperatures**
Recall from Example 5-5 that data on daily high temperatures in February were recorded for three different locations. The data can be found in the Minitab worksheet FebTemps.

(a) Use Minitab to produce dotplots of the temperature data for the three locations on the same scale (Graph > Dotplot > Multiple Ys).

(b) Based on the dotplots, which location tends to have the highest February temperatures? Which has the lowest?

Highest:   Lowest:

(c) Based on the dotplots, which location tends to have the most *variability* in February temperatures? Which has the least variability?

Most:   Least:
(d) Use Minitab to calculate the mean and median, SD and IQR of daily high temperatures for each city (Stat > Basic statistics > Display descriptive statistics).

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lincoln</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>San Luis Obispo</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sedona</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(e) Do you want to rethink your answers to (b) or (c) based on these statistics?

**Example 7-3: Hypothetical Quiz Scores**
Consider the following four histograms, which you can think of as displaying quiz scores for 4 classes of students. The standard deviations of the values in the four classes are 1.18, 2.04, 2.66, and 4.00. Match each one to its histogram. Explain your reasoning.

**Class A:**

**Class B:**

**Class C:**

**Class D:**
Now we will encounter a fourth graphical display for quantitative data: the boxplot (also called box-and-whisker plot).
  - This plot is based on the **five-number summary**.
    - Minimum, lower quartile, median, upper quartile, maximum

**Example 7-4: Haircut Prices**

Do women pay more than men for haircuts? Is this a statistical tendency, or always true? By how much do women spend more than men, on average? How much do haircut prices vary within a gender as well as between genders?

To investigate these questions a Cal Poly professor asked students in her class to report the cost of their most recent haircut, along with their gender.

(a) Which would you consider to be the explanatory variable, and which the response?. Also classify the type (categorical or quantitative) for each variable.

Explanatory:      Type:
Response:        Type:

(b) Is this an experiment or an observational study? Explain briefly.

(c) Did the professor who collected the data make use of random sampling, random assignment, both, or neither?

The haircut prices, sorted by gender, and a dotplot are shown below:

Men \((n = 13)\):
0 0 0 14 15 15 20 20 20 22 23 60 75

Women \((n = 18)\):
0 15 15 20 25 30 35 40 45 45 50 50 60 70 75 110 120 150

(d) What shape do both distributions have? Explain why this makes sense.
(e) Which sex tends to have higher haircut prices? Which tends to have more variability in haircut prices?

(f) Determine the five-number summary of haircut prices for each group.

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Lower quartile</th>
<th>Median</th>
<th>Upper quartile</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(g) Comment on how the five-number summaries compare between the two groups.

- A **boxplot** is a visual representation of the five-number summary: a central box captures the middle 50% of the data by spanning the quartiles, a line in the box marks the median, and lines (called *whiskers*) extend out from the box to the largest and smallest observations.

(h) Produce (unmodified) boxplots to compare the two distributions of haircut prices.
• A modified boxplot: indicates outliers separately, extends “whisker” to most extreme non-outlier.
  o Outliers are values more than $1.5 \times \text{IQR}$ from nearer quartile
  ▪ An “extreme” outlier is more than $3 \times \text{IQR}$ from nearer quartile

(h) Use the $1.5 \times \text{IQR}$ rule to check for outliers in each group.

(i) Produce (modified) boxplots of haircut prices for the two groups on the same scale.

(j) Use Minitab (HaircutPrices) to calculate five-number summaries and produce boxplots. 
(Hint: Notice that the Minitab worksheet contains the data in both stacked and unstacked formats. Also be aware that Minitab uses a slightly different algorithm for calculating quartiles.)

(k) Summarize what your analysis has revealed about the distributions of haircut prices between Cal Poly men and women students.