Visualizing Quantitative Data

Box and Whisker Plots
Descriptive Statistics

● Measures of central tendency
  ○ Mean, median, mode

● Measures of spread
  ○ Range, max, min, quartiles (today)
  ○ Variance, standard deviation (coming soon)

● Measures of shape
  ○ Skew, modalities, etc.
Maximum, minimum, and range

- The **maximum** is an outcome that is greater than or equal to all others in our sample.
- The **minimum** is an outcome that is less than or equal to all others in our sample.
- The **range** is the difference between the maximum and the minimum.
- The maximum and minimum are sensitive to outliers:
  - If an outcome is added to sample that is less than the minimum, then the minimum changes.
  - If an outcome is added to our sample that is greater than the maximum, then the maximum changes.
- To determine if the maximum and the minimum in our data set are indeed outliers we can use a rule of thumb called the interquartile range rule.
Quartiles

- Quartiles within an ordered set of data are three points that divide the dataset into four equal groups, each containing a quarter of the data.
- The first quartile (Q1) is the midpoint between the minimum and the median.
- The second quartile (Q2) is the median.
- The third quartile (Q3) is the midpoint between the median and the maximum.
- The interquartile range (IQR) is the difference between Q3 and Q1 (IQR = Q3 - Q1).

![](image-source.png)
Computing quartiles

- Find the median, and use it to divide the dataset into two halves
  - If there are an odd number of data points in the dataset do not include the median in either half
  - If there are even number of data points in the dataset, split it exactly in half
- The lower quartile value is the median of the lower half of the data
- The upper quartile value is the median of the upper half of the data
Example of computing quartiles

- Computing quartiles when the sample size is odd
- Ordered sample: 3, 4, 4, 5, 6, 8, 8

<table>
<thead>
<tr>
<th>Quartile</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Quartile (Q1)</td>
<td>4</td>
</tr>
<tr>
<td>Second Quartile (Q2) (Median)</td>
<td>5</td>
</tr>
<tr>
<td>Third Quartile (Q3)</td>
<td>8</td>
</tr>
<tr>
<td>Interquartile Range (IQR = Q3 - Q1)</td>
<td>4</td>
</tr>
</tbody>
</table>
Another example of computing quartiles

- Computing quartiles when the sample size is even
- Ordered sample: 1, 3, 3, 4, 5, 6, 6, 7, 8, 8

<table>
<thead>
<tr>
<th>First Quartile (Q1)</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Second Quartile (Q2) (Median)</td>
<td>5.5</td>
</tr>
<tr>
<td>Third Quartile (Q3)</td>
<td>7</td>
</tr>
<tr>
<td>Interquartile Range (IQR = Q3 - Q1)</td>
<td>4</td>
</tr>
</tbody>
</table>
Box and whisker plots: for visualizing quartiles

- Depict minimum, first quartile, median, third quartile and maximum
- The upper whisker (from the maximum to the third quartile) represents the upper 25% of the distribution (excluding outliers)
- The interquartile range (IQR) represents the middle 50% of the data
- The lower whisker (from the first quartile to the minimum) represents the lower 25% of the distribution (excluding outliers)
Interquartile range rule for detecting outliers

1. Calculate the interquartile range (Q3-Q1).
2. Multiply the interquartile range (IQR) by 1.5.
3. Add 1.5*IQR to the third quartile. This value is called the upper fence. Values greater than this are suspected outliers.
4. Subtract 1.5*IQR from the first quartile. This value is called the lower fence. Values less than this are suspected outliers.
IQR rule for detecting outliers: visualization
Interquartile range rule example

Ordered sample: 1, 3, 4, 6, 7, 7, 8, 8, 10, 12, 17

<table>
<thead>
<tr>
<th>First Quartile (Q1)</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Second Quartile (Q2) (Median)</td>
<td>7</td>
</tr>
<tr>
<td>Third Quartile (Q3)</td>
<td>10</td>
</tr>
<tr>
<td>Interquartile Range (IQR = Q3 - Q1)</td>
<td>6</td>
</tr>
</tbody>
</table>

1.5 * 6 = 9
4 - 9 = -5 no data are less than this value
10 + 9 = 19 no data are greater than this value

Even though 17 is five more than the nearest data point, according to the interquartile range rule, it should not be considered an outlier.
Side-by-side box and whisker plots (double the fun!)

- Side-by-side box and whisker plots are a method for visualizing data when one variable is categorical (qualitative) and the other is quantitative.
- They can be used to compare the distributions associated with quantitative variable across the levels of a categorical variable.
- In this plot, the stars are outliers.
Another example: Do pets relieve stress?

- Does someone experience different level of stress when doing tasks with a pet, a good friend, or alone?
- Allen et al. had 45 people count backwards by 13s and 17s
- The people were randomly assigned to 3 different groups: pet (P), friend (F), and alone (C, for control)
- The dependent variable measured was the subject’s average heart rate during the task
Study Results

- The task was most stressful around friends and least stressful around pets
- We are comparing levels of a quantitative variable (heart rate) across levels of a categorical (qualitative) variable (treatment)