Section 2.4

Outliers, Boxplots, and Quantitative/Categorical Relationships
Outline

• One quantitative variable (continued)
  ○ Formal rule for outliers
  ○ Boxplots

• One quantitative and one categorical variable
  ○ Side-by-side boxplots
  ○ Statistics by group
  ○ Difference in means
Outliers

Outliers can be informally identified by looking at a plot, but one general rule of thumb for identifying outliers is data values more than 1.5 IQRs beyond the quartiles.

A data value is an outlier if it is

\[
\text{Smaller than } Q_1 - 1.5 \times (\text{IQR})
\]

or

\[
\text{Larger than } Q_3 + 1.5 \times (\text{IQR})
\]
Boxplot

To draw a boxplot:

• Draw a numerical scale appropriate for the data
• Draw a box stretching from $Q_1$ to $Q_3$
• Divide the box with a line at the median
• Draw a line from each quartile to the most extreme data value that is not an outlier
• Identify each outlier individually by plotting with a symbol such as an asterisk or dot
One Quantitative Variable

World gross for all 2011 Hollywood movies

HollywoodMovies2011

More graphics on profits for Hollywood movies
Boxplot

• Lines ("whiskers") extend from each quartile to the most extreme value that is not an outlier

\[
\begin{align*}
&\text{Q}_1 & \text{Median} & Q_3 \\
\end{align*}
\]

middle 50% of data
Hot Dog Eating

• How many hot dogs can you eat in 10 minutes?

• The boxplot below shows the number of hot dogs eaten by the winners of Nathan’s Famous hot dog eating contest 2002 – 2011

• Estimate the median.
• Estimate the IQR.
• Estimate the range.
Summary: One Quantitative Variable

• Summary Statistics
  ○ Center: mean, median
  ○ Spread: standard deviation, range, IQR
  ○ Percentiles
  ○ 5 number summary

• Visualization
  ○ Dotplot
  ○ Histogram
  ○ Boxplot

• Other concepts
  ○ Shape: symmetric, skewed, bell-shaped
  ○ Outliers, resistance
  ○ z-scores
Quantitative and Categorical Relationships

• Interested in a quantitative variable broken down by categorical groups
Tea and the Immune System

• Participants were randomized to drink five or six cups of either tea or coffee every day for two weeks (both drinks have caffeine but only tea has L-theanine)

• After two weeks, blood samples were exposed to an antigen, and production of interferon gamma (immune system response) was measured

• Explanatory variable: tea or coffee
• Response variable: measure of interferon gamma

Tea and the Immune System

- If the tea drinkers have significantly higher levels of interferon gamma, can we conclude that drinking tea rather than coffee caused an increase in this aspect of the immune response?

- Yes! This is a randomized experiment, so it’s possible to make conclusions about causality!
Side-by-Side Boxplots
Quantitative Statistics by a Categorical Variable

- Any of the statistics we use for a quantitative variable can be looked at separately for each level of a categorical variable

- Mean level of interferon gamma by drink:

<table>
<thead>
<tr>
<th></th>
<th>Tea</th>
<th>Coffee</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{x}_T$</td>
<td>17.70</td>
<td>$\bar{x}_C$</td>
</tr>
</tbody>
</table>
Difference in Means

- Often, when comparing a quantitative variable across two categories, and compute the *difference in means*.

<table>
<thead>
<tr>
<th></th>
<th>Tea</th>
<th>Coffee</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{x}_T )</td>
<td>17.70</td>
<td>( \bar{x}_C ) = 34.82</td>
</tr>
</tbody>
</table>

\[ \bar{x}_T - \bar{x}_C = 34.82 - 17.70 = 17.12 \]
Summary: One Quantitative and One Categorical

- **Summary Statistics**
  - Any summary statistics for quantitative variables, broken down by groups
  - Difference in means

- **Visualization**
  - Side-by-side boxplots (or dot plots, or histograms)