Exploring Data

Graphing Relationships between Variables

Administrative Notes

- Homework 1 posted last week: due in recitation this Friday, Feb. 5th
- Homework 2 posted today: due in recitation on Friday, Feb. 12th

Outline of Lecture

- Scatterplots: relationship between two continuous variables
- Side-by-side Boxplots: relationship between categorical and continuous variables
- Contingency Tables: relationship between two categorical variables

Continuous versus Categorical

- Quantitative (continuous) variables are numerical measurements
- Arithmetic operations (mean, sd, etc.) are meaningful
- Categorical variables place measurements into one of several groups
- Not all numerical variables are quantitative!
  - Workforce study - mean has no meaning here

<table>
<thead>
<tr>
<th>Agriculture</th>
<th>Maritime</th>
<th>Financial</th>
<th>Education</th>
<th>Transportation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Scatterplot of continuous variables

- Old Faithful: Eruption duration and interval between eruptions (min) from July 1995

Associations between Variables

- Positively associated if increased values of one variable tend to occur with increased values of the other
- Negatively associated if increased values of one variable occur with decreased values of the other
- Old Faithful: eruption duration is positively associated with interval between durations
- Remember that association is not proof of causation!
Another Example: US Cities

- Properties of 60 United States Cities
- Two variables of interest:
  - Mortality: Age-adjusted mortality (deaths/100,000)
  - Education: Median education (years)

<table>
<thead>
<tr>
<th>City</th>
<th>Mortality</th>
<th>Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akron, OH</td>
<td>531.37</td>
<td>11.4</td>
</tr>
<tr>
<td>Albany, NY</td>
<td>597.97</td>
<td>11.8</td>
</tr>
<tr>
<td>Allentown, PA</td>
<td>567.55</td>
<td>11.8</td>
</tr>
<tr>
<td>Altoona, PA</td>
<td>493.25</td>
<td>11.2</td>
</tr>
<tr>
<td>Atlantic City, PA</td>
<td>567.55</td>
<td>11.8</td>
</tr>
<tr>
<td>Atlanta, GA</td>
<td>521.62</td>
<td>11.4</td>
</tr>
<tr>
<td>Birmingham, AL</td>
<td>567.55</td>
<td>11.8</td>
</tr>
<tr>
<td>Binghamton, NY</td>
<td>567.55</td>
<td>11.8</td>
</tr>
<tr>
<td>Blythewood, PA</td>
<td>567.55</td>
<td>11.8</td>
</tr>
<tr>
<td>Bridgeport, CT</td>
<td>567.55</td>
<td>11.8</td>
</tr>
<tr>
<td>Buffalo, NY</td>
<td>567.55</td>
<td>11.8</td>
</tr>
</tbody>
</table>

Feb. 2, 2016  Stat 111 - Lecture 6 - Graphs  7

Scatterplot: Mortality and Education

- Negative association between mortality and education
- Potential outliers: York, PA and Lancaster, PA

Explanatory and Response Variables

- A response variable (Y-axis) measures an outcome of interest. Also called dependent
- An explanatory variable (X-axis) explains changes in response. Also called independent
- Explanatory does not mean causal: there are often several possible explanatory variables
- Example: Study of heart disease & smoking
  - Response: death due to heart disease
  - Explanatory: number of cigarettes smoked per day
- Example: City dataset
  - Response: mortality
  - Explanatory: education

Feb. 2, 2016  Stat 111 - Lecture 6 - Graphs  9

Another Example: Challenger Shuttle

- Lower temperatures associated with higher number of O-ring failures
- Temp. on day of Challenger accident: 31 degrees
- NASA only looked at number of failures through time, which doesn't show any relationship

Feb. 2, 2016  Stat 111 - Lecture 6 - Graphs  10

Another Example: Stock Market

- Scatterplot is much more effective means of seeing actual trend in data

Feb. 2, 2016  Stat 111 - Lecture 6 - Graphs  11

Linear Relationships

- Some associations are not just positive or negative, but also appear to be linear
Linear Relationships

- A perfect linear relationship is $Y = a + bX$
- Relationship will never be perfectly linear in real data
- How do we calculate $a$ and $b$? Next class!

<table>
<thead>
<tr>
<th>Positively Linear</th>
<th>Negatively Linear</th>
<th>No linear effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b &gt; 0$</td>
<td>$b &lt; 0$</td>
<td>$b = 0$</td>
</tr>
</tbody>
</table>

Labeling Points on Scatterplots

- Often we have an additional categorical variable that contributes to relationship between two continuous variables
- Add this variable to scatterplots by labeling points with different symbols
- Example: March 2002 report analyzing crack cocaine and powder cocaine penalties

Sentences for Crack versus Cocaine

- Critics claimed that sentences given for trafficking powder cocaine are much lighter than the sentences given for crack cocaine
- Justice department argued that differences weren't that big

Example: March 2002 report analyzing crack cocaine and powder cocaine penalties

- Clearly there is a large difference between the two labeled groups

Sentences for Crack versus Cocaine

- Compare sentence length for each quantity of drugs for both crack and cocaine groups

Side-By-Side Boxplots

- How does the distribution of a continuous variable change between categories?
- Divide continuous variable into each category and construct separate boxplots
- Eg. home runs for B. Ruth vs. M. McGwire

Example: Vietnam Draft Lottery

- Vietnam draft order was determined by putting 366 balls (one for each birthday) in tumbler. First birthday drawn out is drafted first

- Scatterplot seems totally random
Example: Vietnam Draft Lottery

- Instead, can use side-by-side boxplots to look at distribution of draft order by month
- Easy to see that later months have higher draft order
- Why? Balls were loaded into tumbler by month and tumbler wasn’t mixed well

Contingency Tables

- Relationship between two categorical variables examined with a contingency table
- Example: Vitamin C study (Linus Pauling, 1971)
  - Does vitamin C reduce incidence of common cold?
  - 279 people randomly given vitamin C or placebo

<table>
<thead>
<tr>
<th>Group</th>
<th>Cold</th>
<th>No Cold</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vitamin C</td>
<td>17</td>
<td>122</td>
<td>139</td>
</tr>
<tr>
<td>Placebo</td>
<td>31</td>
<td>109</td>
<td>140</td>
</tr>
<tr>
<td>Total</td>
<td>48</td>
<td>231</td>
<td>279</td>
</tr>
</tbody>
</table>

Next Class - Lecture 7

- Exploring Data: Numerical summaries of the relationship between variables
- Moore and McCabe: Section 2.2