Lecture Set 3 – Graphs used to Describe Data

Comparative Bar Charts

• Recall: we use a bar chart when presenting categorical data graphically
• Comparative bar charts can be used to visually compare two or more groups at once

Comparative Bar Charts (cont’)

Example: According to the Cal Poly Fact Book 2003-2004 the enrollment by gender and college are as follows.

<table>
<thead>
<tr>
<th>College</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>90</td>
<td>80</td>
</tr>
<tr>
<td>Science and Math</td>
<td>70</td>
<td>60</td>
</tr>
<tr>
<td>Liberal Arts</td>
<td>60</td>
<td>50</td>
</tr>
<tr>
<td>Engineering</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>Business</td>
<td>40</td>
<td>30</td>
</tr>
<tr>
<td>Architecture</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>Agriculture</td>
<td>20</td>
<td>10</td>
</tr>
</tbody>
</table>

Segmented Bar Charts

• Segmented bar charts use the entire bar to represent 100% of the category across subgroups

Pie Charts

• Categorical data can also be summarized using a pie chart
• The entire data set is represented by the “pie”, while each category is represented by a “slice”
• Useful when you have only a few categories
• Notes on how to construct (via degrees in a circle) on pages 76-77
Pie Charts (cont’)

Example: Here are the data on CP enrollment (cont’) with comparative pie charts

- Which seems more informative?

Solve example:

Stem and Leaf

- The next graphical display we will review is called a stem and leaf display.
  - Each observation is split into a stem and a leaf
  - A good place to start is to use the last digit of the observation as the leaf and the rest as the stem

Example: Hotdog data (cont’)

- How do we know that the min is 86 and not 8.6 or 0.86?

Stem and Leaf (cont’)

- Suppose you got a stem and leaf that looked like the following, what’s the difference?

Distributions for Numerical Data

- Discrete variables
  - use possible data values as categories
  - next create a frequency/relative frequency table or histogram (like a bar chart but for numerical data) using the data for each data value
  - if the variable has a large number of possible values you can create meaningful intervals

Distributions for Numerical Data (cont’)

Example: Data was collected from randomly sampled California drivers to examine trends in the typical number of speeding violations received in a 2004.

- Quantitative variables
  - need to make classes (meaningful intervals) first
  - some work needs to be done to get quantitative data into classes. One common rule of thumb is that the number of classes should be close to \( \sqrt{n} \)
  - important that classes are of equal width for accurate interpretation of data
  - Once we have our classes we can create a frequency/relative frequency table or histogram.
Distributions for Numerical Data (cont')

Example: Hotdogs (cont') Make a frequency table.
- Overall, the low is 86 calories and the high is 195 calories
  \[ \sqrt{n} = \sqrt{4} = 7.35 \times 7 \]

<table>
<thead>
<tr>
<th>Calories</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>70 - &lt;90</td>
<td>2</td>
<td>0.04</td>
</tr>
<tr>
<td>90 - &lt;110</td>
<td>7</td>
<td>0.13</td>
</tr>
<tr>
<td>110 - &lt;130</td>
<td>3</td>
<td>0.06</td>
</tr>
<tr>
<td>130 - &lt;150</td>
<td>21</td>
<td>0.39</td>
</tr>
<tr>
<td>150 - &lt;170</td>
<td>6</td>
<td>0.11</td>
</tr>
<tr>
<td>170 - &lt;190</td>
<td>10</td>
<td>0.18</td>
</tr>
<tr>
<td>190 - &lt;210</td>
<td>5</td>
<td>0.09</td>
</tr>
<tr>
<td>Total</td>
<td>54</td>
<td>1.00</td>
</tr>
</tbody>
</table>

- Seems slightly arbitrary

Distributions for Numerical Data (cont')

Example: Hotdogs (cont') Histogram using previously defined classes.

Histogram of Calories

Distributions for Numerical Data (cont')

Example: Hotdogs (cont')
- Most of the time it is easiest to just let the computer decide (ie. use the default)

Histogram of Calories

- Any difference between the two histograms?

The BIG Three

- There are three main features of data that should always be addressed in an analysis
  - Shape
  - Center
  - Spread

Shapes of Distributions

- The shape of a distribution can usually be determined by just looking at it as a histogram, dot plot or stem and leaf display.
- Definition: A distribution is unimodal if it has one mode
  - Unimodal distributions include:
    - Symmetric (aka bell, normal)
    - Positively Skewed (aka skewed right)
    - Negatively Skewed (aka skewed left)
  - Other distributions are:
    - Bimodal
    - Multimodal
    - Exponential
Shapes of Distributions (cont')

• What shape is the speeding ticket data?

Shapes of Distributions (cont')

• What seems like a logical reason for the shape of the hot dog calorie data?

Shapes of Distributions (cont')

• Classify and draw a sketch each of the following scenarios with respect to mode. Also, if unimodal, classify symmetry (symmetric, skewed right or skewed left).
  – Data collected on height of randomly sampled college students.
  – Data collected on height of randomly sampled female college students.
  – The salaries of all persons employed by a large university.
  – The amount of time spent by students on a difficult exam.
  – The grade distribution on a difficult exam.

Bivariate Data

• Bivariate data consists of measurements on two variables, called x and y
  – this data exists in the form of a pair, such as height and weight of individuals
• From this we can create a graphical display called a scatterplot
  – each pair is represented by a dot on the plot
  – we are looking for trends or patterns in the data

Bivariate Data (cont')

Example: Cereal data (cont'). In addition to gathering data on shelf location and sugar content, the researchers also collected the corresponding rating of the cereal that was calculated by Consumer Reports. Let’s examine the relationship between sugar, fiber, fat, and sodium versus the Consumer Reports rating.

Bivariate Data (cont')

• Are there any obvious trends?
• Are they positive or negative?
• Which variable seems to have the strongest relationship with rating?
**Time Series Plots**

- Time series plots are used when data is collected over a period of time.
- Similar to a scatter plot in that the data consists of x,y pairs,
  - However there is only one observation per time point.
  - Time is placed on the x-axis.
  - The data are connected with a line.

---

**Time Series Plots (cont')**

Example: Cal Poly Fact Book 2003-2004 (cont'). The fact book also reports number of application to Cal Poly by year, since 1990.

---

**Data Reporting**

- Graphical displays are a great place to start with a data analysis,
  - helps organize data and presents better than just looking at the data itself.
  - highlights main features of a data set.
- Some considerations for graphics,
  - be sure that the graph type is appropriate for the type of data.
  - include appropriate labels, scales (units), and titles.
  - keep graphics simple and honest.