MATH 112
Section 7.1: Representing and Interpreting Data

Prof. Jonathan Duncan

Walla Walla University

Autumn Quarter, 2007
In this section we statistics and how we can use statistics to answer questions. This can be broken down as follows.

**Four Basic Components**

The four components in using statistics to answer a question are:

1. **Formulating the Question**
   It is important to ask a question that is specific enough that we can collect useful data, but not so simple that the data is meaningless.

2. **Collecting Data**
   We must ensure that we collect accurate data that answers the question we formulated.

3. **Representing and Analyzing Data**
   Raw data must be organized. Our representation depends on what we want to know and may be a table, graph, or combination of both.

4. **Interpreting Our Results**
   Once we have represented the data, we interpret the graphs, tables, and numbers to turn them into conclusions which answer our question, when possible.
Refining Questions

In many instances, our initial question is not well suited for a statistical analysis.

Example

Rework the following questions so that they are well formulated for data collection.

1. What is your favorite animal?
2. How much money do teachers make?
3. How much TV do you watch?

Formulating Questions

Your questions should be focused and not too complex. You should ask yourself what your purpose is in collecting the data.
Collecting data can also be more complex than it may at first seem.

**Data Collection Issues**

Consider the following issues when collecting data to answer a question.

- Do you want to differentiate between certain groups (i.e. Men and Women)?
- Is your questioning process biased (i.e. asking people coming out of a Mexican restaurant what ethnic foods they prefer).
- Will your process illicit truthful answers (written or verbal, public or private)?

**Example**

You wish to determine the post popular ethnic food in the Walla Walla Valley. How will you collect data to answer this question?
Tables: Frequency

Once you have collected data, you must organize it in a way that makes sense.

Example

You collected the following list of favorite ethnic foods:

Mexican, Chinese, American, Mexican, Italian, Italian, French, Indian, Mexican, American, Chinese, Italian, German, Thai, Italian, Mexican, Indian, Italian, Italian, Mexican, Chinese, American, Chinese, Indian, Mexican

Frequency Tables

A frequency table can help you organize this data.

<table>
<thead>
<tr>
<th>Food</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mexican</td>
<td>6</td>
</tr>
<tr>
<td>Chinese</td>
<td>4</td>
</tr>
<tr>
<td>American</td>
<td>3</td>
</tr>
<tr>
<td>Italian</td>
<td>6</td>
</tr>
<tr>
<td>French</td>
<td>1</td>
</tr>
<tr>
<td>Indian</td>
<td>3</td>
</tr>
<tr>
<td>German</td>
<td>1</td>
</tr>
<tr>
<td>Thai</td>
<td>1</td>
</tr>
</tbody>
</table>
Tables: Grouped Frequency

Frequency tables are useful when you have a limited number of possible responses. For numerical data, we use a different table.

Example

You collect the following list of ages of professional football players.

18, 24, 21, 19, 31, 27, 26, 33, 19, 24, 28, 30, 30, 42, 25, 25, 32, 28, 29, 33, 35,
22, 26, 28, 29, 24, 31, 30, 19, 20,

Grouped Frequency Table

Counting each age gives us too many categories. Instead, we count in age groups.

<table>
<thead>
<tr>
<th>Age</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>18-22</td>
<td>7</td>
</tr>
<tr>
<td>23-27</td>
<td>8</td>
</tr>
<tr>
<td>28-32</td>
<td>11</td>
</tr>
<tr>
<td>33-37</td>
<td>3</td>
</tr>
<tr>
<td>38-42</td>
<td>1</td>
</tr>
</tbody>
</table>
Tables: Steam-and-Leaf Plots

While tables are a good way to organize the data, they don’t give us a good visualization of the data, and when we group data, we lose the original values.

Stem-and-Leaf Plots

A stem-and-leaf plot is a table in which individual data points are split into two parts: the stem (usually all but last digit) and leaf (last digit). Individual data points are then listed in ascending order by writing the leaf next to the stem.

Example

Construct a stem-and-leaf plot for the football player age data seen previously, and repeated below.

18, 24, 21, 19, 31, 27, 26, 33, 19, 24, 28, 30, 30, 42, 25, 25, 32, 28, 29, 33, 35, 22, 26, 28, 29, 24, 31, 30, 19, 20,
Graphs: Bar Graphs

The stem-and-leaf plot lets us begin to see a picture of the data. There are several other ways in which this can be done.

**Bar Graph**

A bar graph is a graphical representation of a frequency table in which a bar is created for each category and the height is determined by the frequency count.

**Example**

Below is a bar graph for our ethnic foods example.
Graphs: Circle Graphs

When the focus is on categories as parts of a whole, a circle graph can also be useful.

Circle Graphs
Each frequency is viewed as a percent of the whole. A circle is then divided into wedges of the same percent of the whole circle.

Example
Below is a circle graph for our ethnic foods example.
When working with grouped frequency tables, we use a modified bar graph called a histogram.

**Histograms**

A histogram is a bar graph in which each category is a range of consecutive values. The gaps between the bars are removed to show that values are ordered consecutively.

**Example**

Below is a histogram for our football age example.
Graphs: Line Graphs

To display data collected over time, we use a line graph.

Line Graphs

In a line graph, individual data points have two values, an $x$-value and a $y$-value. These values are plotted on a coordinate system and connected with lines.

Example

The graph below shows VCR ownership between 1978 and 1990.
While graphs and tables are useful for organizing and visualizing data, we often wish to summarize an entire collection of data with a single “center” value.

**The Mean**

To find the mean of a set of numerical data, add the values together and then divide by the number of values.

**Example**

What is the average age of a professional football player?

18, 24, 21, 19, 31, 27, 26, 33, 19, 24, 28, 30, 30, 42, 25, 25, 32, 28, 29, 33, 35, 22, 26, 28, 29, 24, 31, 30, 19, 20
Measures of Center

The Median

While the mean is a good measure of center, it is affected by outliers.

**Example**

In our list of football players ages, one is exceptionally high—the 42. If this is removed, the mean becomes 26.41.

**The Median**

To find the median, arrange data in increasing order and select the middle value if there are an odd number of values, and the mean of the two middle values if there are an even number of values.

**Example**

Find the median age of a football player.

18, 19, 19, 19, 20, 21, 22, 24, 24, 24, 25, 25, 26, 26, 27, 28, 28, 28, 29, 29, 30, 30, 30, 31, 31, 32, 33, 33, 35, 42
The Mode

If our data is not numeric, then we can not find the mean or median. However, there is still a way to measure the center.

The Mode

The mode is the value or category which appears most often. If no value appears more than once, there is no mode. If several categories are tied, each is a mode and the data set is called multi-modal.

Example

Find the mode of our ethnic food example.

Mexican, Chinese, American, Mexican, Italian, Italian, French, Indian, Mexican, American, Chinese, Italian, German, Thai, Italian, Mexican, Indian, Italian, Italian, Mexican, Chinese, American, Chinese, Indian, Mexican

Example

Find the mode of the football player ages.

18, 19, 19, 19, 20, 21, 22, 24, 24, 24, 25, 25, 26, 26, 27, 28, 28, 28, 29, 29, 30, 30, 30, 31, 32, 33, 33, 35, 42
Comparing Measures of Center

Each measure of center has its own unique advantages and disadvantages.

<table>
<thead>
<tr>
<th></th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Often easier to compute with a large set of data and gives the “center of gravity”.</td>
<td>Can be distorted by outliers and may not be a possible value.</td>
</tr>
<tr>
<td>Median</td>
<td>Not affected by outliers and gives midpoint of data.</td>
<td>Does not use every point in the data set</td>
</tr>
<tr>
<td>Mode</td>
<td>Can be used with non-numerical data and easy to find in a graph.</td>
<td>May be far from the true center and may not exist at all.</td>
</tr>
</tbody>
</table>
Our last topic is a “big picture” look at data sets. In each of the following distributions, give an example situation in which the distribution could appear and estimate the mean, median, and mode.
Another common distribution appears a lot in nature and even in test scores!
In many instances a distribution is close to normal, but skewed in one direction.

Skewed Distributions

- Skewed to the right
- Skewed to the left
Finally, a distribution may have two (or more) different modes.
### Things to Remember from Section 7.1

1. **Data Collection Components**

2. **Methods for representing data**
   - Tabular Representations
   - Steam-and-Leaf Graphs
   - Other Graphs

3. **Measures of Center: Mean, Median, and Mode**

4. **Distributions**