Lecture 1

Reminder:

Material Covered in This Lecture:
Chapter 1, Section 1.1: Displaying Distributions with Graphs.

1. Definitions (Introduction, p.4-6)
   (1). Individuals are the objects described by a set of data.
   (2). Variable: A variable is any characteristic of an individual. A variable can take different values for different individuals

   Example: A college's student data base includes data about every currently enrolled student. The students are the individuals described by the data set. For each individual, the data contain the values of variables such as date of birth, gender, choice of major, and grade point average (GPA). The following is a small part of the data set in MINITAB.

   (3). Classification of Variables
       Categorical Variable: A categorical variable places an individual into one of several groups or categories.
       Example: Gender, Major, Blood type, etc.
       Quantitative Variable: A quantitative variable takes numerical values for which arithmetic operations such as adding and averaging make sense.
       Example: Weight, Height, etc.
2. Displaying Distribution with Graphs. (Section 1.1, p.7-24)

Exploratory data analysis: Investigate the rough, basic structure of the data. Two basic strategies to organize the exploration of a set of data:
- Begin by examining each variable by itself. Then move on to study the relationships among the variables.
- Begin with a graph or graphs. Then add numerical summaries of specific aspects of the data.

(1). Describing categorical variables.

Example: Here is the distribution of the highest level of education for people aged 25 to 34 years:

<table>
<thead>
<tr>
<th>Education</th>
<th>Count(millions)</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than high school</td>
<td>4.6</td>
<td>11.8</td>
</tr>
<tr>
<td>High school graduate</td>
<td>11.6</td>
<td>30.6</td>
</tr>
<tr>
<td>Some College</td>
<td>7.4</td>
<td>19.5</td>
</tr>
<tr>
<td>Associate degree</td>
<td>3.3</td>
<td>8.8</td>
</tr>
<tr>
<td>Bachelor's degree</td>
<td>8.6</td>
<td>22.7</td>
</tr>
<tr>
<td>Advanced degree</td>
<td>2.5</td>
<td>6.6</td>
</tr>
</tbody>
</table>

a. Bar Graph.

b. Pie Chart

(2) Describing the quantitative variables.

a. Stem-and-Leaf Plot.

Example (Example 1.4, p.9): Literacy rates.

Female percent: 60 31 46 71 86 99 82 71 85 38 70 63 99 63 78 99 29
Male percent: 78 50 68 85 96 100 95 92 92 68 84 89 100 83 94 100 70
b. **Histogram.**

Steps to create histogram:
- Divide the range of the data into classes of equal width
- Count the number of individuals in each class. (Count = Frequency) (Recommendation: Construct a frequency table)
- Draw the histogram.

Example: IQ score. (Minitab Demonstration)

<table>
<thead>
<tr>
<th>TABLE 1.3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IQ test scores for 60 randomly chosen fifth-grade students</strong></td>
</tr>
<tr>
<td>145 139 126 122 125 130 96 110 118 118</td>
</tr>
<tr>
<td>101 142 134 124 112 109 134 113 81 113</td>
</tr>
<tr>
<td>123 94 100 136 109 131 117 110 127 124</td>
</tr>
<tr>
<td>106 124 115 133 116 102 127 117 109 137</td>
</tr>
<tr>
<td>117 90 103 114 139 101 122 105 97 89</td>
</tr>
<tr>
<td>102 108 110 128 114 112 114 102 82 101</td>
</tr>
</tbody>
</table>

(3). Examining distributions.

"What do I see?"
- In any graph of data, look for the overall pattern and for striking deviations from that pattern
- We can describe the overall pattern of a distribution by its shape, center, and spread.

Center: midpoint.
Spread: range (minimum, maximum)
Shape: Unimodal, bimodal, or multimodal? Symmetric, asymmetric? if asymmetric, skewed to left or right?
An important kind of deviation is an outlier, an individual value that falls outside the overall pattern.

(4). **Time plots.** A time plot of a variable plots each observation against the time at which it was measured. Always put time on the horizontal scale of your plot and the variable you are measuring on the vertical scale. Connecting the data points by lines helps emphasize any change over time.

Example: