Describing Data: Part I

John McGready
Johns Hopkins University
Lecture Topics

- Types of data
- Measures of central tendency for continuous data
- Sample versus population
- Visually describing continuous data
- Underlying shape of the “true distribution” of continuous data
Section A

Types of Data
Steps in a Research Project

1. Planning
2. Design
3. Data collection
4. Data analysis
5. Presentation
6. Interpretation
Biostatistics Issues

- Design of studies
  - Sample size
  - Role of randomization
Biostatistics Issues

- **Variability**
  - Important patterns in data are obscured by variability
  - Distinguish real patterns from random variation
Biostatistics Issues

♦ Inference
  – Draw general conclusions from limited data
  – For example: Survey

♦ Summarize
1954 Salk Polio Vaccine Trial

School Children

Randomized

Vaccinated
N=200,745

Placebo
N=201,229

Polio Cases: Vaccine 82
Placebo 162

Continued
1954 Salk Polio Vaccine Trial

Design

Features of the Polio Trial

- Comparison group
- Randomized
- Placebo controls
- Double blind
- Objective—the groups should be equivalent except for the factor (vaccine) being investigated

Continued
Design

Features of the Polio Trial

- Question
  - Could the results be due to chance?
Imbalance

There were almost twice as many polio cases in the placebo compared to the vaccine group.
Could We Get Such Great Imbalance by Chance?

- **Polio cases**
  - Vaccine—82
  - Placebo—162
  - \( p \)-value = ?

- Statistical methods tell us how to make these probability calculations
Types of Data

- Binary (dichotomous) data
  - Yes/No
  - Polio—Yes/No
  - Cure—Yes/No
  - Gender—Male/Female
Types of Data

- Categorical data
  
  *(place individuals in categories)*
  
  - Race/ethnicity  
    nominal
  
  - Country of birth  
    (no ordering)
  
  - Degree of agreement  
    ordinal
    (ordering)
Types of Data

- Continuous data
  *(finer measurements)*
  - Blood pressure
  - Weight
  - Height
  - Age
Types of Data

There are different statistical methods for different types of data
Example: Binary Data

- To compare the number of polio cases in the two treatment arms of the Salk Polio vaccine, you could use . . .
  - Fisher’s Exact Test
  - Chi-Square test
Example: Continuous Data

- To compare blood pressures in a clinical trial evaluating two blood pressure-lowering medications, you could use . . .
  - 2-sample T-test
  - Wilcoxon Rank Sum Test (nonparametric)
Section A

Practice Problems
Data Types

State the data type of each of the following:

- Homicide rate (deaths/100,000)
- High school graduate (Y/N)
- Hair color
- Hospital expenditures (yearly, in dollars)
- Smoking status (none, light, heavy)
- Coronary heart disease (Y/N)
Section A

Practice Problem Solutions
Solutions

- Homicide rate (deaths/100,000)
Solutions

- Homicide rate (deaths/100,000)
Solutions

♦ High school graduate (Y/N)
Solutions

- High school graduate (Y/N)

Dichotomous (Binary)
Solutions

- Hair color
Solutions

- Hair color

Categorical (Nominal)
Solutions

- Hospital expenditures (yearly, in dollars)
Solutions

- Hospital expenditures (yearly, in dollars)
Solutions

- Smoking status (none, light, heavy)
Solutions

- Smoking status (none, light, heavy)
Solutions

- Coronary heart disease (Y/N)
Solutions

- Coronary heart disease (Y/N)

Dichotomous
(Binary)
Section B

Measures of Central Tendency
Sample Versus Population
Summarizing and Describing Continuous Data

- Measures of the center of data
  - Mean
  - Median
Sample Mean $\bar{X}$

*The Average or Arithmetic Mean*

- Add up data, then divide by sample size ($n$)
- The sample size $n$ is the number of observations (pieces of data)
Example

- $n = 5$ Systolic blood pressures (mmHg)
  
  $X_1 = 120$
  $X_2 = 80$
  $X_3 = 90$
  $X_4 = 110$
  $X_5 = 95$
Example

\[
\bar{X} = \frac{120 + 80 + 90 + 110 + 95}{5} = 99 \text{mmHg}
\]
Notes on Sample Mean $\bar{X}$

- Formula

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$
In the formula to find the mean, we use the “summation sign” — \( \sum \).

This is just mathematical shorthand for “add up all of the observations”.

\[
\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n
\]
Notes on Sample Mean

- Also called *sample average* or *arithmetic mean*

- Sensitive to extreme values
  - One data point could make a great change in sample mean

- Why is it called the *sample* mean?
  - To distinguish it from population mean
Population Versus Sample

- **Population**—The entire group you want information about
  - For example: The blood pressure of all 18-year-old male college students in the United States
Population Versus Sample

- **Sample**—A part of the population from which we actually collect information and draw conclusions about the whole population
  - For example: Sample of blood pressures
    N=five 18-year-old male college students in the United States
Population Versus Sample

- The sample mean \( \bar{X} \) is not the population mean \( \mu \)
Population Versus Sample

Population
Population mean $\mu$

Sample
Sample mean $X$

Continued
Population Versus Sample

- We don’t know the population mean \( \mu \) but would like to know it
- We draw a sample from the population
- We calculate the sample mean \( \bar{X} \)
- How close is \( \bar{X} \) to \( \mu \)?
- Statistical theory will tell us how close \( \bar{X} \) is to \( \mu \)
Statistical Inference

- *Statistical inference* is the process of trying to draw conclusions about the population from the sample.
Sample Median

- The median is the middle number

80 90 95 110 120
Sample Median

- The sample median is not sensitive to extreme values
  - For example: If 120 became 200, the median would remain the same, but the mean would change to 115
Sample Median

- If the sample size is an even number

\[
\begin{array}{cccccc}
80 & 90 & 95 & 110 & 120 & 125 \\
\end{array}
\]

\[
\text{Median} = \frac{95 + 110}{2} = 102.5 \text{ mmHg}
\]
Section B

Practice Problems
Practice Problems

- The following data is the annual income (in $1000s of dollars) taken from nine students in the Hopkins internet-based MPH program:

37 102 34 12 111 56 72 17 33
Practice Problems

1. Calculate the sample mean income
2. Calculate the sample median income
3. What population could this sample represent?
4. Which would change by a larger amount—the mean or median—if the 34 were replaced by 17, and the 12 replaced by a 31?
Section B

Practice Problem Solutions
Solutions

1. Calculate the sample mean income

*Remember:*

\[
\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}
\]

Where \( n=9 \) and \( x_1 \) through \( x_9 \) represent the nine observed values of income.
Solutions

\[
\bar{x} = \frac{37 + 102 + 34 + 12 + 111 + 56 + 72 + 17 + 33}{9}
\]

\[
\bar{x} = \frac{474}{9} = 52.67
\]

- The mean income in our sample is $52,670
2. Calculate the sample median income

- To calculate the sample median, we must first order our data from the lowest value to the highest value
Solutions

- The ordered data set appears below

```
12 17 33 34 37 56 72 102 111
```
Solutions

- Because we have nine values (an odd number), we can just pick the middle value to calculate the median.
Solutions

3. What population could this sample represent?

- It could be representative of all Johns Hopkins Internet-based MPH students; it would need to be made clear that this is a random sample in order for it to be called representative.
Solutions

4. Which would change by a larger amount—the mean or the median—if the 34 were replaced by 17, and the 12 replaced by a 31?

- Notice that both changes do nothing to change the position of the median; therefore, the only statistic that would change is the mean
Visually Displaying Continuous Data: Histograms; The Underlying “Population Distribution”
Pictures of Data

Continuous Variables

- Histograms
  - Means and medians do not tell whole story
  - Differences in spread (variability)
  - Differences in shape of the distribution
How to Make a Histogram

- Consider the following data collected from the 1995 Statistical Abstracts of the United States
  - For each of the 50 United States, the proportion of individuals over 65 years of age has been recorded
# How to Make a Histogram

<table>
<thead>
<tr>
<th>State</th>
<th>%</th>
<th>State</th>
<th>%</th>
<th>State</th>
<th>%</th>
<th>State</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL</td>
<td>12.9</td>
<td>IN</td>
<td>12.8</td>
<td>NE</td>
<td>14.1</td>
<td>SC</td>
<td>11.9</td>
</tr>
<tr>
<td>AK</td>
<td>4.6</td>
<td>IA</td>
<td>15.4</td>
<td>NV</td>
<td>11.3</td>
<td>SD</td>
<td>14.7</td>
</tr>
<tr>
<td>AZ</td>
<td>13.4</td>
<td>KS</td>
<td>13.9</td>
<td>NH</td>
<td>11.9</td>
<td>TN</td>
<td>12.7</td>
</tr>
<tr>
<td>AR</td>
<td>14.8</td>
<td>KY</td>
<td>12.7</td>
<td>NJ</td>
<td>13.2</td>
<td>TX</td>
<td>10.2</td>
</tr>
<tr>
<td>CA</td>
<td>10.6</td>
<td>LA</td>
<td>11.4</td>
<td>NM</td>
<td>11.0</td>
<td>UT</td>
<td>8.8</td>
</tr>
<tr>
<td>CO</td>
<td>10.1</td>
<td>ME</td>
<td>13.9</td>
<td>NY</td>
<td>13.2</td>
<td>VT</td>
<td>12.1</td>
</tr>
<tr>
<td>CT</td>
<td>14.2</td>
<td>MD</td>
<td>11.2</td>
<td>NC</td>
<td>12.5</td>
<td>VA</td>
<td>11.1</td>
</tr>
<tr>
<td>DE</td>
<td>12.7</td>
<td>MA</td>
<td>14.1</td>
<td>ND</td>
<td>14.7</td>
<td>WA</td>
<td>11.6</td>
</tr>
<tr>
<td>FL</td>
<td>18.4</td>
<td>MI</td>
<td>12.4</td>
<td>OH</td>
<td>13.4</td>
<td>WV</td>
<td>15.4</td>
</tr>
<tr>
<td>GA</td>
<td>10.1</td>
<td>MN</td>
<td>12.5</td>
<td>OK</td>
<td>13.6</td>
<td>WI</td>
<td>13.4</td>
</tr>
<tr>
<td>HI</td>
<td>12.1</td>
<td>MI</td>
<td>12.5</td>
<td>OR</td>
<td>13.7</td>
<td>WY</td>
<td>11.1</td>
</tr>
<tr>
<td>ID</td>
<td>11.6</td>
<td>MO</td>
<td>14.1</td>
<td>PA</td>
<td>15.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL</td>
<td>12.6</td>
<td>MT</td>
<td>13.3</td>
<td>RI</td>
<td>15.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Source: Statistical Abstracts of the United States, 1995*
### How to Make a Histogram

<table>
<thead>
<tr>
<th>State</th>
<th>%</th>
<th>State</th>
<th>%</th>
<th>State</th>
<th>%</th>
<th>State</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL</td>
<td>12.9</td>
<td>IN</td>
<td>12.8</td>
<td>NE</td>
<td>14.1</td>
<td>SC</td>
<td>11.9</td>
</tr>
<tr>
<td>AK</td>
<td>4.6</td>
<td>IA</td>
<td>15.4</td>
<td>NV</td>
<td>11.3</td>
<td>SD</td>
<td>14.7</td>
</tr>
<tr>
<td>AZ</td>
<td>13.4</td>
<td>KS</td>
<td>13.9</td>
<td>NH</td>
<td>11.9</td>
<td>TN</td>
<td>12.7</td>
</tr>
<tr>
<td>AR</td>
<td>14.8</td>
<td>KY</td>
<td>12.7</td>
<td>NJ</td>
<td>13.2</td>
<td>TX</td>
<td>10.2</td>
</tr>
<tr>
<td>CA</td>
<td>10.6</td>
<td>LA</td>
<td>11.4</td>
<td>NM</td>
<td>11.0</td>
<td>UT</td>
<td>8.8</td>
</tr>
<tr>
<td>CO</td>
<td>10.1</td>
<td>ME</td>
<td>13.9</td>
<td>NY</td>
<td>13.2</td>
<td>VT</td>
<td>12.1</td>
</tr>
<tr>
<td>CT</td>
<td>14.2</td>
<td>MD</td>
<td>11.2</td>
<td>NC</td>
<td>12.5</td>
<td>VA</td>
<td>11.1</td>
</tr>
<tr>
<td>DE</td>
<td>12.7</td>
<td>MA</td>
<td>14.1</td>
<td>ND</td>
<td>14.7</td>
<td>WA</td>
<td>11.6</td>
</tr>
<tr>
<td>FL</td>
<td>18.4</td>
<td>MI</td>
<td>12.4</td>
<td>OH</td>
<td>13.4</td>
<td>WV</td>
<td>15.4</td>
</tr>
<tr>
<td>GA</td>
<td>10.1</td>
<td>MN</td>
<td>12.5</td>
<td>OK</td>
<td>13.6</td>
<td>WI</td>
<td>13.4</td>
</tr>
<tr>
<td>HI</td>
<td>12.1</td>
<td>MI</td>
<td>12.5</td>
<td>OR</td>
<td>13.7</td>
<td>WY</td>
<td>11.1</td>
</tr>
<tr>
<td>ID</td>
<td>11.6</td>
<td>MO</td>
<td>14.1</td>
<td>PA</td>
<td>15.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL</td>
<td>12.6</td>
<td>MT</td>
<td>13.3</td>
<td>RI</td>
<td>15.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Source: Statistical Abstracts of the United States, 1995*
### How to Make a Histogram

<table>
<thead>
<tr>
<th>State</th>
<th>%</th>
<th>State</th>
<th>%</th>
<th>State</th>
<th>%</th>
<th>State</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL</td>
<td>12.9</td>
<td>IN</td>
<td>12.8</td>
<td>NE</td>
<td>14.1</td>
<td>SC</td>
<td>11.9</td>
</tr>
<tr>
<td>AK</td>
<td>4.6</td>
<td>IA</td>
<td>15.4</td>
<td>NV</td>
<td>11.3</td>
<td>SD</td>
<td>14.7</td>
</tr>
<tr>
<td>AZ</td>
<td>13.4</td>
<td>KS</td>
<td>13.9</td>
<td>NH</td>
<td>11.9</td>
<td>TN</td>
<td>12.7</td>
</tr>
<tr>
<td>AR</td>
<td>14.8</td>
<td>KY</td>
<td>12.7</td>
<td>NJ</td>
<td>13.2</td>
<td>TX</td>
<td>10.2</td>
</tr>
<tr>
<td>CA</td>
<td>10.6</td>
<td>LA</td>
<td>11.4</td>
<td>NM</td>
<td>11.0</td>
<td>UT</td>
<td>8.8</td>
</tr>
<tr>
<td>CO</td>
<td>10.1</td>
<td>ME</td>
<td>13.9</td>
<td>NY</td>
<td>13.2</td>
<td>VT</td>
<td>12.1</td>
</tr>
<tr>
<td>CT</td>
<td>14.2</td>
<td>MD</td>
<td>11.2</td>
<td>NC</td>
<td>12.5</td>
<td>VA</td>
<td>11.1</td>
</tr>
<tr>
<td>DE</td>
<td>12.7</td>
<td>MA</td>
<td>14.1</td>
<td>ND</td>
<td>14.7</td>
<td>WA</td>
<td>11.6</td>
</tr>
<tr>
<td>FL</td>
<td>18.4</td>
<td>MI</td>
<td>12.4</td>
<td>OH</td>
<td>13.4</td>
<td>WV</td>
<td>15.4</td>
</tr>
<tr>
<td>GA</td>
<td>10.1</td>
<td>MN</td>
<td>12.5</td>
<td>OK</td>
<td>13.6</td>
<td>WI</td>
<td>13.4</td>
</tr>
<tr>
<td>HI</td>
<td>12.1</td>
<td>MI</td>
<td>12.5</td>
<td>OR</td>
<td>13.7</td>
<td>WY</td>
<td>11.1</td>
</tr>
<tr>
<td>ID</td>
<td>11.6</td>
<td>MO</td>
<td>14.1</td>
<td>PA</td>
<td>15.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL</td>
<td>12.6</td>
<td>MT</td>
<td>13.3</td>
<td>RI</td>
<td>15.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Source: Statistical Abstracts of the United States, 1995*
How to Make a Histogram

- Count the number of observations in each class
- Here are the observations:

<table>
<thead>
<tr>
<th>Class</th>
<th>Count</th>
<th>Class</th>
<th>Count</th>
<th>Class</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.0–5.0</td>
<td>1</td>
<td>9.0–10.0</td>
<td>0</td>
<td>14.0–15.0</td>
<td>7</td>
</tr>
<tr>
<td>5.0–6.0</td>
<td>0</td>
<td>10.0–11.0</td>
<td>5</td>
<td>15.0–16.0</td>
<td>4</td>
</tr>
<tr>
<td>6.0–7.0</td>
<td>0</td>
<td>11.0–12.0</td>
<td>9</td>
<td>16.0–17.0</td>
<td>0</td>
</tr>
<tr>
<td>7.0–8.0</td>
<td>0</td>
<td>12.0–13.0</td>
<td>12</td>
<td>17.0–18.0</td>
<td>0</td>
</tr>
<tr>
<td>8.0–9.0</td>
<td>1</td>
<td>13.0–14.0</td>
<td>10</td>
<td>18.0–19.0</td>
<td>1</td>
</tr>
</tbody>
</table>
How to Make a Histogram

- Divide range of data into intervals (bins) of equal width
- Count the number of observations in each class
- Draw the histogram
- Label scales
A histogram of the percent of residents older than 65 years in the 50 United States.
A histogram of the percent of residents older than 65 years in the 50 United States.
A histogram of the percent of residents older than 65 years in the 50 United States.
Recall the blood pressure data on a sample of 113 men
Histogram of the Systolic Blood Pressure for 113 men. Each bar spans a width of five mmHg on the horizontal axis. The height of each bar represents the number of individuals with SBP in that range.
Another histogram of the blood pressure of 113 men. In this graph, each bar has a width of 20 mmHg, and there are a total of only four bars making it hard to characterize the distribution of blood pressures in the sample.
Yet another histogram of the same BP information on 113 men. Here, the bin width is one mmHg, perhaps giving more detail than is necessary.
Another way to present the data in a histogram is to label the y-axis with relative frequencies as opposed to counts. The height of each bar represents the percentage of individuals in the sample with BP in that range. The bar heights should add to one.
Intervals

How many intervals (bins) should you have in a histogram?

- There is no perfect answer to this
- Depends on sample size \( n \)
- Rough rule of thumb: \( \# \text{ Intervals} \approx \sqrt{n} \)

<table>
<thead>
<tr>
<th>( n )</th>
<th>Number of Intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>About 3</td>
</tr>
<tr>
<td>50</td>
<td>About 7</td>
</tr>
<tr>
<td>100</td>
<td>About 10</td>
</tr>
</tbody>
</table>
Shapes of the Distribution

- Three common shapes of frequency distributions

A: Symmetrical and bell shaped
B: Positively skewed or skewed to the right
C: Negatively skewed or skewed to the left
Shapes of the Distribution

- Three less common shapes of frequency distributions

A (Bimodal)  
B (Reverse J-shaped)  
C (Uniform)
Distribution Characteristics

- Mode—Peak(s)
- Median—Equal areas point
- Mean—Balancing point
Shapes of Distributions

- **Right skewed** (positively skewed)
  - Long right tail
  - Mean > Median
Shapes of Distributions

- **Left skewed** (negatively skewed)
  - Long left tail
  - Mean < Median
Shapes of Distributions

- Symmetric (right and left sides are mirror images)
  - Left tail looks like right tail
  - Mean = Median = Mode
Shapes of Distributions

- **Outlier**
  - An individual observation that falls outside the overall pattern of the graph
Shapes of Distributions

Outlier

Florida

Alaska
The same histogram of BP measurements from a sample of 113 men. We are going to compare this to a histogram for a larger sample, and for the entire population.
The Histogram and the Probability Density

Histogram of blood pressure measurements, this time for a sample of 5,000 men: notice how the shape of the histogram is more defined than the previous sample of 113 men.

Continued
The Histogram and the Probability Density

The Probability Density for BP values in the entire population of men—because the population is infinite, there are no bars, and the y-axis can not have actual counts.

Continued
The Histogram and the Probability Density

- The *probability density* is a smooth idealized curve that shows the shape of the distribution in the population.
- Areas in an interval under the curve represent the percent of the population in the interval.
Section C

Practice Problems
Practice Problems

- What kind of shape do you think the distribution for the following data would have?

- Hospital length of stay for 1,000 randomly selected patients at Johns Hopkins

- Blood pressure in all women
Practice Problems

- Annual income for all JHU Internet-based MPH students (refer to results of last Practice Problems, assume a random sample)
- Number of hours of sporting events watched last week by a sample of 350 men and 350 women
Section C

Practice Problem Solutions
Solutions

- Hospital length of stay for 1,000 randomly selected patients at Johns Hopkins
Solutions

- Hospital length of stay for 1,000 randomly selected patients at Johns Hopkins

\[\text{Graph showing distribution of hospital length of stay.}\]
Solutions

- Blood pressure in all women
Solutions

- Blood pressure in all women
Solutions

- Annual income for all JHU Internet-based MPH students
Solutions

- Annual income for all JHU Internet-based MPH students

![Graph showing mean and median income distribution]

*Continued*
Solutions

- Number of hours of sporting events watched last week by a sample of 350 women and 350 men
Solutions

- Number of hours of sporting events watched last week by a sample of 350 women and 350 men
Solutions

- Number of hours of sporting events watched last week by a sample of 350 women and 350 men

Mode for Women?  Mode for Men?
Section D

Stem and Leaf Plots, Box Plots
Suppose we took another look at our random sample of 113 men and their blood pressure measurements.

One tool for “visualizing” the data is the histogram.

Sample 113 Men
Histogram: BP for 113 males
Sample 113 Men: Stem and Leaf

- Another common tool for visually displaying continuous data is the “stem and leaf” plot
- Very similar to a histogram
  - Like a “histogram on its side”
  - Allows for easier identification of individual values in the sample
Stem and Leaf: BP for 113 Males

8. | 9
9* |
9. | 9
10* | 11334
10. | 566777899
11* | 111223333344444
11. | 556666667779
12* | 00000000111223344
12. | 5566677778888999999
13* | 0011222334
13. | 5677789
14* | 0000112222
14. | 67
15* | 0122

Continued 110
## Stem and Leaf: BP for 113 Males

<table>
<thead>
<tr>
<th>&quot;Stems&quot;</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.</td>
<td>1</td>
</tr>
<tr>
<td>9*</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>1</td>
</tr>
<tr>
<td>10*</td>
<td>1</td>
</tr>
<tr>
<td>10.</td>
<td>3</td>
</tr>
<tr>
<td>11*</td>
<td>1</td>
</tr>
<tr>
<td>11.</td>
<td>6</td>
</tr>
<tr>
<td>12*</td>
<td>1</td>
</tr>
<tr>
<td>12.</td>
<td>2</td>
</tr>
<tr>
<td>13*</td>
<td>1</td>
</tr>
<tr>
<td>13.</td>
<td>2</td>
</tr>
<tr>
<td>14*</td>
<td>2</td>
</tr>
<tr>
<td>14.</td>
<td>1</td>
</tr>
<tr>
<td>15*</td>
<td>1</td>
</tr>
</tbody>
</table>

*Continued on next page*
Stem and Leaf: BP for 113 Males

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaves</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>9*</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>10*</td>
<td>11334</td>
</tr>
<tr>
<td>10</td>
<td>566777899</td>
</tr>
<tr>
<td>11*</td>
<td>111223333344444</td>
</tr>
<tr>
<td>11</td>
<td>55666667779</td>
</tr>
<tr>
<td>12*</td>
<td>00000000111223344</td>
</tr>
<tr>
<td>12</td>
<td>5566677778888999999</td>
</tr>
<tr>
<td>13*</td>
<td>000112222334</td>
</tr>
<tr>
<td>13</td>
<td>5677789</td>
</tr>
<tr>
<td>14*</td>
<td>0000112222</td>
</tr>
<tr>
<td>14</td>
<td>67</td>
</tr>
<tr>
<td>15*</td>
<td>0122</td>
</tr>
</tbody>
</table>
Stem and Leaf: BP for 113 Males

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.</td>
<td>9</td>
</tr>
<tr>
<td>9*</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>9</td>
</tr>
<tr>
<td>10*</td>
<td>11334</td>
</tr>
<tr>
<td>10.</td>
<td>566777899</td>
</tr>
<tr>
<td>11*</td>
<td>111223333344444</td>
</tr>
<tr>
<td>11.</td>
<td>556666667779</td>
</tr>
<tr>
<td>12*</td>
<td>000000000111223344</td>
</tr>
<tr>
<td>12.</td>
<td>5566677778888999999</td>
</tr>
<tr>
<td>13*</td>
<td>000112222334</td>
</tr>
<tr>
<td>13.</td>
<td>5677789</td>
</tr>
<tr>
<td>14*</td>
<td>0000112222</td>
</tr>
<tr>
<td>14.</td>
<td>67</td>
</tr>
<tr>
<td>15*</td>
<td>0122</td>
</tr>
</tbody>
</table>

Continued 113
Stem and Leaf: BP for 113 Males

8.  | 9
9*  |  
9.  | 9
10* | 11334
10. | 566777899
11* | 111223333344444
11. | 556666667779
12* | 00000000111223344
12. | 556667777888999999
13* | 00011222334
13. | 5677789
14* | 0000112222
14. | 67
15* | 0122

Continued
Sample 113 Men: Stem and Boxplot

- Another common visual display tool is the boxplot
  - Gives good insight into distribution shape in terms of skewness and outlying values
  - Very nice tool for easily comparing distribution of continuous data in multiple groups—can be plotted side by side
Boxplot: BP for 113 Males

Boxplot of Systolic Blood Pressures
Sample of 113 Men
Boxplot: BP for 113 Males

Boxplot of Systolic Blood Pressures
Sample of 113 Men

Sample Median Blood Pressure
Boxplot: BP for 113 Males

Boxplot of Systolic Blood Pressures
Sample of 113 Men

75th Percentile
25th Percentile
Boxplot: BP for 113 Males

Boxplot of Systolic Blood Pressures
Sample of 113 Men

Largest Observation

Smallest Observation
Hospital Length of Stay for 1,000 Patients

- Suppose we took a sample of discharge records from 1,000 patients discharged from a large teaching hospital
- How could we visualize this data?
Histogram: Length of Stay
Boxplot: Length of Stay

Hospital LOS (Days) for 1,000 Patients
Boxplot: Length of Stay

Hospital LOS (Days) for 1,000 Patients
Boxplot: Length of Stay

Hospital LOS (Days) for 1,000 Patients

Largest Non-Outlier

Smallest Non-Outlier

Continued
Boxplot: Length of Stay

Hospital LOS (Days) for 1,000 Patients

Large Outliers