Part 1: Quantitative and Categorical Data

Statistics: The science of data; the science of making effective use of numerical data relating to groups of individuals or experiments.

Data: Information about some group of individuals or subjects.

Individual: The person or object described by a set of data.

Variable: Any characteristic of an individual.

Two Types of Variables:

Quantitative: Has numerical values for which arithmetic operations (e.g., addition or averaging) make sense. Examples: age, height, # of AP classes, SAT score.

Categorical: Places an individual into one of several groups or categories. Examples: eye color, race, gender. May have numerical values assigned: 1=White, 2=Hispanic, 3=Asian, etc. Other numeric categorical variables include baseball jersey number or zip code.

Example: The FAA monitors airlines for safety and customer service. For each flight, the carrier must report the type of aircraft, flight number, number of passengers, and whether or not the flights departed and arrived on schedule. What variables are reported for each flight, and are they quantitative or categorical?

Answers: Variables Quantitative or Categorical

(1) Type of aircraft □ Quantitative ☑ Categorical
(2) Flight Number □ Quantitative ☑ Categorical
(3) Number of Passengers ☑ Quantitative □ Categorical
(4) Arrived/Departed on Schedule □ Quantitative ☑ Categorical

Practice Problems

Determine if the variables listed below are quantitative or categorical. Neatly print “Q” for quantitative and “C” for categorical.

_____ 1. Time it takes to get to school  ____ 8. Height
_____ 2. Number of shoes owned  ____ 9. Amount of oil spilled
_____ 3. Hair color  ____ 10. Age of Oscar winners
_____ 4. Temperature of a cup of coffee  ____ 11. Type of pain medication
_____ 5. Teacher salaries  ____ 12. Jellybean flavors
_____ 7. Facebook user  ____ 14. Type of meat
Part 2: Categorical Graphs

Graphing Data: Shows us the distribution of the variable – what values the variable takes and how often it takes these values. Graphs give us information to help us to understand the data. For all graphs, it is important to label axes for clarity.

Categorical Data Graphs:

1. Bar Graph  
   a) Bars are separated  
   b) Each bar is labeled with a category  
   c) Vertical axis shows frequency (count) or percentage  
   d) Allows quick comparisons of the frequencies of each category.

2. Pie Chart  
   a) Show what part of the whole each category represents  
   b) Must include all of the categories that make up the whole  
   c) May include a category labeled “Other”

Categorical Graphing Examples

For each problem below decide if you can display the data as a bar graph, a pie chart, or both. Then create one appropriate graph for the data. Make sure to include a title and all necessary labels.

1. Eye color of Langley AP Statistics students: Brown = 64, Blue = 23, Green = 36, Other = 5
   - Bar Graph Only  
   - Pie Chart Only  
   - Both Plots Could Represent Data
2. Average height of various animals, in cm: Walaroo = 105, Eagle = 88, Tapir = 97, Zebra = 146.

- [ ] Bar Graph Only
- [ ] Pie Chart Only
- [ ] Both Plots Could Represent Data

Note: A pie chart is not appropriate for this data because each value represents an average.

3. Percentage of Langley students with a driver’s license, by class:

Senior = 77%, Junior = 28%, Sophomore = 6%, Freshman = 2%

- [ ] Bar Graph Only
- [ ] Pie Chart Only
- [ ] Both Plots Could Represent Data

Note: A pie chart is not appropriate for this data because each percentage is not a percent of all of Langley students (percents do not sum to 100).
15. Here are data on the percent of females among people earning doctorates in 1994 in several fields of study.

<table>
<thead>
<tr>
<th>Field</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer Science</td>
<td>15.4%</td>
</tr>
<tr>
<td>Life Sciences</td>
<td>40.7%</td>
</tr>
<tr>
<td>Education</td>
<td>60.8%</td>
</tr>
<tr>
<td>Physical Sciences</td>
<td>21.7%</td>
</tr>
<tr>
<td>Engineering</td>
<td>11.1%</td>
</tr>
<tr>
<td>Psychology</td>
<td>62.2%</td>
</tr>
</tbody>
</table>

(a) Present these data in a well-labeled bar graph.

(b) Would it also be correct to use a pie chart to display these data? Is so, construct the pie chart. If not, explain why not.
Part 3: Center and Spread for Quantitative Data

**Measuring Center**

There are two ways we can measure the center of a distribution: *mean* and *median*.

**Mean:**

When our data is fairly symmetric with no big outliers, *mean* is the best measure of center. To find the mean of a set of observations, add each value and divide by the number of observations. This formula is written as:

\[ \bar{x} = \frac{1}{n} \sum x_i \]

where \( \bar{x} \) (pronounced x–bar) denotes sample mean and \( x_i \) is each observation.

Example: Find the mean of the data set:

\[ 10 \quad 18 \quad 11 \quad 28 \quad 19 \]

\[ \bar{x} = \frac{1}{n} \sum x_i = \frac{10 + 18 + 11 + 28 + 19}{5} = \frac{86}{5} = 17.2 \]

**Median:**

Median is the midpoint of a distribution—half the observations lie above the median and the other half lie below the median. Because the median is not affected by outliers, we call the median a *resistant measure*, and use the median when describing skewed distributions. To find the median of a distribution:

1. Arrange all observations in order of size, from smallest to largest.
2. If the number of observations \( n \) is odd, the median \( M \) is the center of observation in the ordered list.
3. If the number of observations \( n \) is even, the median \( M \) is the mean of the two center observations in the ordered list.

Example 1: Find the median of the data set.

ODD NUMBER

\[ 4 \quad 8 \quad 9 \quad 11 \quad 15 \]

Median is the middle number when the data is arranged in ascending order

Example 2: Find the median of the data set.

EVEN NUMBER

\[ 4 \quad 8 \quad 9 \quad 11 \quad 15 \quad 20 \]

Average the two middle numbers to get the median of an even data set:

\[ \frac{9 + 11}{2} = 10 \]
Measuring Spread

Whenever you provide a measure of center, you always want to couple it with a measure of spread. Two distributions with the same median can be very different depending on their spread. Look at data sets 1 and 2 below. Both have a median of 15 but the variability of each distribution is very different.

<table>
<thead>
<tr>
<th>Data Set 1:</th>
<th>Data Set 2:</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 10 15 20 25</td>
<td>1 2 15 90 90</td>
</tr>
</tbody>
</table>

Providing a range is one way to illustrate spread. Range is the difference between the largest and smallest observations. Data Set 1 has a range of 20 (25 – 5); Data Set 2 has a range of 89 (90 – 1). Note that range is always expressed as one number. Never express the range as an interval!

The range only uses the smallest and largest numbers in the data set to describe the variability. This can be misleading, because sometimes the smallest and/or largest number in a data set is an outlier. We can better our description of the variability by also examining the spread of the middle half of the data. To do this we find quartiles (the 25th, 50th, and 75th percentiles of the data).

Calculating Quartiles:

1. Arrange the observations in increasing order and locate the median \( M \) in the ordered list of observations.

2. The first quartile \( Q_1 \) is the median of the observations whose position in the ordered list is to the left of the location of the overall median.

3. The third quartile \( Q_3 \) is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

The five number summary of a data set consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation (min, \( Q_1 \), \( M \), \( Q_3 \), max).

The interquartile range (IQR) is the distance between the first and third quartiles, \( IQR = Q_3 - Q_1 \). IQR is also a resistant measure—it is not affected by outliers. Therefore, cases when the median is the best measure of center, IQR would be the best measure of spread for the data.

The IQR is the basis of the rule of thumb for identifying suspected outliers. Call an observation an outlier if it falls more than \( 1.5 \times IQR \) above the third quartile or below the first quartile.
Example Problems:

1. The following are the scores of 12 members of a woman’s golf team in tournament play:
   \[ 89 \quad 90 \quad 87 \quad 95 \quad 86 \quad 84 \quad 111 \quad 108 \quad 83 \quad 88 \quad 91 \quad 79 \]
   a. Find the range of the data.
   \[ \text{RANGE} = \text{MAX} - \text{MIN} = 111 - 79 = \boxed{32} \]
   
   b. Find the first quartile, median, and third quartile. Remember to put the data in ascending order.
   \[ 79 \quad 81 \quad 83 \quad 86 \quad 87 \quad 88 \quad 89 \quad 90 \quad 91 \quad 95 \quad 108 \quad 111 \]
   
   \[ 84.5 \quad 93 \quad 96.5 \quad Q_1 \quad \text{MIDIAN} \quad Q_3 \]
   
   c. Find the interquartile range.
   \[ \text{IQR} = Q_3 - Q_1 = 93 - 84.5 = \boxed{8.5} \]
   
   d. Does the data contain any outliers? If so, what are they?
   \[ \text{IQR} \times 1.5 = (8.5)(1.5) = 12.75 \]
   \[ \text{CUTOFF: LOW OUTLIERS} = Q_1 - 12.75 = 84.5 - 12.75 = 71.75 \]
   \[ \text{HIGH OUTLIERS} = Q_3 + 12.75 = 93 + 12.75 = 105.75 \]

2. Consumer Reports did a study of ice cream bars (only vanilla flavored) in their August 1989 issue. Eleven bars having a taste-test rating of at least “fair” were listed, and calories per bar was included. Calories vary quite a bit partly because bars are not of uniform size.

   \[ 439 \quad 111 \quad 201 \quad 182 \quad 197 \quad 209 \quad 147 \quad 190 \quad 151 \quad 131 \quad 151 \]
   a. Find the range of the data.
   \[ \text{RANGE} = \text{MAX} - \text{MIN} = 439 - 111 = \boxed{328} \]
   
   b. Find the first quartile, median, and third quartile.
   \[ 111 \quad 131 \quad 147 \quad 151 \quad 151 \quad 162 \quad 190 \quad 197 \quad 201 \quad 209 \quad 439 \]
   
   \[ Q_1 \quad \text{MIDIAN} \quad Q_3 \]
   
   c. Find the IQR.
   \[ \text{IQR} = Q_3 - Q_1 = 201 - 147 = \boxed{54} \]
   
   d. Are there any outliers in the data?
   \[ \text{IQR} \times 1.5 = (54)(1.5) = 81 \]
   \[ \text{LOW OUTLIERS ARE LESS THAN} \quad Q_1 - 81 = 147 - 81 = 66 \quad \text{(NONE)} \]
   \[ \text{HIGH OUTLIERS ARE MORE THAN} \quad Q_3 + 81 = 201 + 81 = 282 \]
   \[ 439 \text{ is an outlier} \]
When the mean is the best measure of center it should be coupled with the **standard deviation**. The standard deviation measures spread by looking at how far the observations are from their mean. The formula for standard deviation is:

\[ s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2} \]

The **variance** \( s^2 \) of a set of observations is the standard deviation squared, meaning the average of the squares of the deviation of the observations from their mean.

**Example:** Below is a list of test scores earned by AP Statistics students on the chapter 1 test. Find the mean and standard deviation.

88 72 96 68 81

1. Find the sample mean:

\[ \bar{x} = \frac{88 + 72 + 96 + 68 + 81}{5} = 81 \]

2.

<table>
<thead>
<tr>
<th>( x_i )</th>
<th>( x_i - \bar{x} )</th>
<th>( (x_i - \bar{x})^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>88</td>
<td>88 - 81 = 7</td>
<td>(7)^2 = 49)</td>
</tr>
<tr>
<td>72</td>
<td>72 - 81 = -9</td>
<td>(-9)^2 = 81)</td>
</tr>
<tr>
<td>96</td>
<td>96 - 81 = 15</td>
<td>(15)^2 = 225)</td>
</tr>
<tr>
<td>68</td>
<td>68 - 81 = -13</td>
<td>(-13)^2 = 169)</td>
</tr>
<tr>
<td>81</td>
<td>81 - 81 = 0</td>
<td>(0)^2 = 0)</td>
</tr>
</tbody>
</table>

1. \[ \sum (x_i - \bar{x})^2 = 49 + 81 + 225 + 169 + 0 = 524 \]

2. \[ \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{524}{5-1} = \frac{524}{4} = 131. \] This is the variance!

3. \[ \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{131} \approx 11.45. \] This is the standard deviation.
Example Problem:

4. Below is a sample of high daily temperatures (in degrees) taken in the month of August in Virginia over the past year.

\[
\begin{align*}
98 & \\
95 & \\
104 & \\
88 & \\
91 & \\
90 & 
\end{align*}
\]

a. Find the mean of the data.

\[
\frac{\sum x_i}{n} = \frac{98 + 95 + 104 + 88 + 91 + 90}{6} = \frac{560}{6} = 93.33
\]

b. Find the variance and standard deviation. Make a table to help you.

\[
\begin{array}{c|c|c}
 x_i & x_i - \bar{x} & (x_i - \bar{x})^2 \\
98 & 3.67 & 13.47 \\
95 & 0.67 & 0.45 \\
104 & 19.67 & 93.51 \\
88 & -6.33 & 40.07 \\
91 & -3.33 & 11.09 \\
90 & -4.33 & 18.75 \\
\hline
 & & 177.34 \\
\end{array}
\]

\[
\text{Variance} = \frac{177.34}{6} = 33.47
\]

\[
\text{Std. Dev.} = \sqrt{\text{Var.}} = \sqrt{33.47} = 5.79
\]

TI 83/84 Calculator Instructions:

Your calculator will compute the mean, standard deviation, and five number summary for you.

1. Press the STAT button on your calculator. #1 will be highlighted. Press ENTER.

2. Enter the data into L1.

3. Press 2nd MODE (Quit) to exit the screen.

4. Press STAT, then move the cursor to the right so CALC is highlighted. #1 should be selected. Press ENTER.

5. The screen of your calculator will say 1-Var Stats. Type in L1 (since your data is in List 1) ENTER.

6. The sample mean, \( \bar{x} \), is the first value given. \( Sx \) is the sample standard deviation. Scroll down to find the five number summary.
## Practice Problems:

<table>
<thead>
<tr>
<th>Data Set 1</th>
<th>Data Set 2</th>
<th>Data Set 3</th>
<th>Data Set 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>15</td>
<td>16</td>
</tr>
</tbody>
</table>

16. **Center**  
Set 1 | Set 2 | Set 3 | Set 4  
--- | --- | --- | ---  
Mean: | | | |  
Median: | | | |  

17. **Spread**  
Variance: | | | |  
Std Dev.: | | | |  
Range: | | | |  
IQR: | | | |  

18. **5-# Summary**  
Min: | | | |  
Q₁: | | | |  
Median: | | | |  
Q₃: | | | |  
Max: | | | |  

19. Circle the measures that are *resistant* (an extreme outlier does not affect it).  
For example, change the last value to 1000. Which measures do not change (resistant)?
Part 4: Quantitative Graphs

While Bar Graphs and Pie Charts are used to graph Categorical Data, there are many methods of graphing Quantitative Data. These include dotplots, stemplots, histograms, and boxplots.

DotPlots are one of the simplest statistical plots, and are suitable to small and moderate-sized data sets. DotPlots have the advantage of retaining the original data values (you could re-create the detailed, original data using the dotplot).

Constructing a dotplot:
1. Draw a horizontal line, and label it with the variable being graphed (in the graph below, “Weight in ounces”). Provide a descriptive title, and label the axis with relevant data values.
2. Scale the axis based on the values of the variable.
3. Mark a dot above the number on the horizontal axis corresponding to each data value. Each dot represents a single observation from the set of data.

Practice Problem:

20. In the Super Bowl, by how many points does the winning team outscore the losers? Here are the winning margins for the first 42 Super Bowl games:

<table>
<thead>
<tr>
<th>Winning Margin</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>23</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

Create a well labeled dotplot for the data above.
**Stemplots**, also known as “stem-and-leaf plots”, also allow us to plot the original data:

| Student Heart Rates after doing 1 minute of jumping jacks |
|---|---|
| **Stem** | **Leaf** |
| 6 | 1 |
| 7 |   |
| 8 | 4 6 8 |
| 9 | 2 9 9 9 |
| 10 |   |
| 11 | 3 4 |
| 12 | 1 2 4 4 7 9 |
| 13 | 8 |
| 14 | 4 7 8 |
| 15 | 1 |

**Key:** 6 | 1 = 61

Constructing a Stemplot:

1. Separate each observation into a *stem* consisting of all but the rightmost digit, and a *leaf*, the final digit. In the example above, the lowest observed heart rate of 61 beats per minute consists of a “stem” of 6 and a “leaf” of 1.

2. Write the stems vertically in increasing order from top to bottom, and draw a vertical line to the right of the stem values. Examine the data and write each leaf to the right of its stem, spacing the leaves equally. It is good practice to order the stem values from smallest to largest as you write them across the graph.

3. Title your graph, and add a *key* describing what the stems and leaves represent.

4. If the stems have a large number of leaves, it may be helpful to “split” stems (for example, the stems could go in steps of 5 instead of 10, so one stem could include values from 60-64, and another stem could include 65-69).
Practice Problems:

21. Below are the typing speeds (words per minute) for 22 secretarial applicants of an international cosmetic company:

   68  72  91  47  52  75  63  55  65  35  69
   70  84  45  58  61  69  22  46  55  66  71

a. Graph the distribution of typing speeds as a stemplot:

b. Graph the distribution of typing speeds as a dotplot:

c. What are some advantages of using a stemplot for this data instead of a dotplot?
**Histograms** allow us to graph larger sets of data by grouping values together:

The graph above represents a sample of the heights of 31 black cherry trees.

Could we re-create the original data from this graph?  

- Yes  
- No  

What is the height of the shortest measured tree?  

60 ≤ tree height < 65.

What is the height of the tallest measured tree?  

85 ≤ tree height < 90.

The graph includes the following important elements:

- Descriptive title.
- Properly scaled, labeled axes.
- Classes (“bars”) that are of equal width, and whose heights represent the frequency of observations for the values (tree heights) contained in the class. Unlike a bar graph, order of the bars is important.

Let’s build the histogram, assuming the original data was the following (sorted) heights:

60, 62, 62, 65, 67, 68, 70, 70, 71, 72, 73, 73, 74, 75, 75, 75, 76, 76, 77, 77, 79, 79, 82, 82, 82, 83, 84, 86, 88

Step 1: Determine the number of classes (k) to be used. There is no firm rule on determining this, which means that two different people could create two valid, different histograms for the same data. As a guideline, count the number of observations, \( n \), and take the square root of \( n \). Round \( \sqrt{n} \) to the nearest whole number and use that as the number of classes, \( k \).

\[
n = 31\quad \text{# of classes: } k \approx \sqrt{n} = 6 \quad \text{(nearest whole number)}
\]

Step 2: Determine width of each class. Take the range (max − min) and divide by the number of classes, \( k \). Round this number **up to the next whole number**. This is the class width.

Class width:  

\[
\frac{\text{max} − \text{min}}{k} = \frac{88 − 60}{6} = \frac{28}{6} \approx 5
\]
Step 3: Construct a *frequency table* listing the data count in each class:

<table>
<thead>
<tr>
<th>Class</th>
<th>Count</th>
<th>(Range of Values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>60 – 65</td>
<td></td>
<td>60 ≤ height &lt; 65</td>
</tr>
<tr>
<td>65 – 70</td>
<td></td>
<td>65 ≤ height &lt; 70</td>
</tr>
<tr>
<td>70 – 75</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>75 – 80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80 – 85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>85 – 90</td>
<td></td>
<td>85 ≤ height &lt; 90</td>
</tr>
</tbody>
</table>

Step 4: Draw and label axes, including the lower & upper bounds of each class on the horizontal axis and the frequency (count) of observations for each class on the vertical axis. Draw bars representing the count in each class, *with the bars touching* (no spaces in between).
Practice Problem:

22. Below are times obtained from a mail-order company's shipping records concerning time from receipt of order to delivery (in days) for items from their catalogue. Construct a histogram representing the data. Make sure to include all appropriate labels.

```
3  7  10  5  14  12  6  2  9  22  25  11  13  5
7  12  10  22  23  14  8  5  4  7  27  31  13  21
6  8  3  10  19  12  11  8
```
**Boxplots (Otherwise Known as the Box and Whisker Plot)**

The five number summary is used to create a boxplot. Because boxplots show less detail than histograms or stemplots, they are best used for side-by-side comparison of more than one distribution.

Constructing a boxplot:

1. Draw a horizontal line, and label it with the variable being graphed. Scale the axis based on the values of the variable.
2. Mark a dot where the maximum and minimum values lie (above the horizontal line).
3. Draw vertical lines where the $Q_1$, $M$, and $Q_3$ lie.
4. Connect the vertical lines to create a box around them. Draw horizontal lines from the box to connect to the maximum and minimum values.
5. Provide a descriptive title.

**Annual snow depth at Mathsville Ski Resort**

Practice Problem:

23. Students from a statistics class were asked to record their heights in inches:

   65  72  68  64  60  55  73  71  52  63  61  65
   74  69  67  74  50  44  75  67  62  66  80  64

Construct a boxplot of the data.
Modified boxplots are boxplots that show the outliers as individual points. To construct a modified boxplot, first calculate if there are any outliers (any observation(s) that is more than 1.5 x IQR outside the median). Plot any outliers as individual points. Now construct a boxplot, but the new maximum and minimum are the smallest and largest observations that are not outliers.

Example:

Below is the data on maximum daily rainfall in South Bend, Indiana over a 30-year period:

1.88  2.23  2.58  2.07  2.94  2.29  3.14  2.15  1.95  2.51
2.86  1.48  1.12  2.76  3.10  2.05  2.23  1.70  1.57  2.81
1.24  3.29  1.87  1.50  2.99  3.48  2.12  4.69  2.29  2.12

a) Compute the 5-number summary.

\[
\begin{align*}
\text{MIN} &= 1.12 \\
Q_1 &= 1.88 \\
\text{MED} &= 2.23 \\
Q_3 &= 2.86 \\
\text{MAX} &= 4.69
\end{align*}
\]

\[
\begin{align*}
\text{IQR} &= Q_3 - Q_1 = 0.98 \\
\text{IQR} \times 1.5 &= 1.47
\end{align*}
\]

OUTLIERS ARE OBSERVATIONS BELOW 0.41 (NONE) OR ABOVE 4.33 (4.69)

b) Draw a modified boxplot.
Practice Problem:

24. Students from a statistics class were asked to record their heights in inches:

<table>
<thead>
<tr>
<th>65</th>
<th>72</th>
<th>68</th>
<th>64</th>
<th>60</th>
<th>55</th>
<th>73</th>
<th>71</th>
<th>52</th>
<th>63</th>
<th>61</th>
<th>65</th>
</tr>
</thead>
<tbody>
<tr>
<td>74</td>
<td>69</td>
<td>67</td>
<td>74</td>
<td>50</td>
<td>44</td>
<td>75</td>
<td>67</td>
<td>62</td>
<td>66</td>
<td>80</td>
<td>64</td>
</tr>
</tbody>
</table>

Construct a modified boxplot of the data.

Ogives

A histogram does a good job of displaying the distribution of values of a variable but tells us little about the relative standing of an individual observation. If we want to know the percentage of values that fall at or below a value in the data set, we should construct an ogive, also called a cumulative frequency graph.

Constructing an Ogive:

1. Make a frequency table. Remember, you made a frequency table for histograms. To do this you need to first find the number of classes \((k = \sqrt{n} \text{ rounded to the nearest whole number})\) and the class width \(w = \frac{(\text{max} - \text{min})}{k} \text{ rounded to the next whole number}\). The frequency table for an ogive, however, should have three additional columns: relative frequency (this is the percentage for each count), cumulative frequency, and relative cumulative frequency.

2. Label and scale your axes and title your graph.

3. Plot a point corresponding to the relative frequency in each class interval at the left endpoint of the next class interval.

4. Connect all points with a straight line.
Let's create an ogive for the black cherry tree data.

60, 62, 62, 65, 67, 68, 70, 71, 72, 73, 73, 74, 75, 75, 75, 76, 76, 77, 79, 79, 82, 82, 82, 83, 84, 86, 88

1. a) \( n = 31 \)

b) # of classes: \( k \approx \sqrt[5.56]{31} \approx 5.56 \) so I round to 6 classes.

c) Class width: \( w = \frac{(\text{max} - \text{min})}{k} = \frac{88 - 60}{6} = 4.66 \). I always round this number up, so the class width is 5.

d) Construct the frequency table:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>60 – 65</td>
<td>60 ≤ height &lt; 65</td>
<td>3</td>
<td>3/31 ≈ 10%</td>
<td>3</td>
<td>3/31 ≈ 10%</td>
</tr>
<tr>
<td>65 – 70</td>
<td>65 ≤ height &lt; 70</td>
<td>3</td>
<td>3/31 ≈ 10%</td>
<td>6</td>
<td>6/31 ≈ 19%</td>
</tr>
<tr>
<td>70 – 75</td>
<td>...</td>
<td>8</td>
<td>8/31 ≈ 26%</td>
<td>14</td>
<td>14/31 ≈ 45%</td>
</tr>
<tr>
<td>75 – 80</td>
<td></td>
<td>10</td>
<td>10/31 ≈ 32%</td>
<td>24</td>
<td>24/31 ≈ 77%</td>
</tr>
<tr>
<td>80 – 85</td>
<td></td>
<td>5</td>
<td>5/31 ≈ 16%</td>
<td>29</td>
<td>29/31 ≈ 94%</td>
</tr>
<tr>
<td>85 – 90</td>
<td>85 ≤ height &lt; 90</td>
<td>2</td>
<td>2/31 ≈ 6%</td>
<td>31</td>
<td>31/31 ≈ 100%</td>
</tr>
</tbody>
</table>

2. An ogive always has the class numbers on the x-axis and the cumulative relative frequency on the y-axis.

Looking at an ogive we can easily locate percentiles. For example, about 70% of Black Cherry Trees are about 79 feet or less.
Practice Problems:

The graph below displays the relative cumulative frequency (ogive) of the ages of golfers who competed in the McLean Charity Golf Tournament in August.

25. Estimate the median age of the golfers.

26. What percentage of golfers were over 59 years old?

27. What percentage of the golfers who played in the tournament were 19 years old or younger?

28. What age corresponds to the 80\textsuperscript{th} percentile?

29. Plot a histogram of these data.
Part 5: Describing Quantitative Graphs

In any graph of data, look for the overall pattern and for any striking deviations from the pattern. You can describe the overall pattern of a distribution by its shape, center, and spread.

Shape: The data on the graph may resemble one of the distinct patterns below, or may show no special shape.

- Symmetric: The left and right sides are approximately mirror images.
- Skewed Left: There is a “tail” of data that extends far to the left.
- Skewed Right: There is a “tail” of data that extends far to the right.
- Uniform: All data values occur at roughly the same frequency (all “bars” are equally high).
- Unimodal: The graph has one distinct peak.
- Bimodal: The graph has two distinct peaks.

Center: You can estimate the center of a distribution by visually examining the graph, or by calculating one of the common measures of center:

- Mean: The average of all of the data values. The mean is significantly affected by extreme outliers (it is not “resistant”).
- Median: The middle value, when the data is ordered from smallest to greatest. When there is an even number of values, the middle two numbers are averaged to determine the median. The median is unaffected by extreme outliers (it is “resistant”).

Spread: Below are two sets of data that are both symmetric and have the same mean and median:

Set I: 15, 15, 15, 15, 15, 15, 15
Set II: 1, 1, 1, 1, 29, 29, 29

What distinguishes them is how widely spread the data is. There are several measures of spread in describing the distribution of a variable:

- Range: Max. Value – Min Value
- Inter-Quartile Range: More on this when we discuss boxplots.
- Variance or Standard Deviation: More on this later.

Unusual Features: These include any significant deviations from the overall pattern, including:

- Outliers: These are individual values that fall outside of the overall pattern.
- Gaps & Clusters: Parts of the distribution that contain an unusually small or large amount of data.
Comparison of Shapes of Different Graphs

<table>
<thead>
<tr>
<th>Graph</th>
<th>LEFT-SKEWED</th>
<th>SYMMETRIC</th>
<th>RIGHT-SKEWED</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Histogram</strong></td>
<td><img src="image" alt="Histogram" /></td>
<td><img src="image" alt="Histogram" /></td>
<td><img src="image" alt="Histogram" /></td>
</tr>
<tr>
<td><strong>Dotplot</strong></td>
<td><img src="image" alt="Dotplot" /></td>
<td><img src="image" alt="Dotplot" /></td>
<td><img src="image" alt="Dotplot" /></td>
</tr>
<tr>
<td><strong>Boxplot</strong></td>
<td><img src="image" alt="Boxplot" /></td>
<td><img src="image" alt="Boxplot" /></td>
<td><img src="image" alt="Boxplot" /></td>
</tr>
<tr>
<td><strong>Cumulative Frequency Chart</strong></td>
<td><img src="image" alt="Cumulative Frequency Chart" /></td>
<td><img src="image" alt="Cumulative Frequency Chart" /></td>
<td><img src="image" alt="Cumulative Frequency Chart" /></td>
</tr>
<tr>
<td><strong>Stemplot</strong></td>
<td><img src="image" alt="Stemplot" /></td>
<td><img src="image" alt="Stemplot" /></td>
<td><img src="image" alt="Stemplot" /></td>
</tr>
</tbody>
</table>
**Example:**

You are interested in studying how much time students spend using the Web each day. You study 30 students, and their times spent on the Web (in minutes) for a particular day are listed below.

<table>
<thead>
<tr>
<th>7</th>
<th>20</th>
<th>24</th>
<th>25</th>
<th>25</th>
<th>28</th>
<th>28</th>
<th>30</th>
<th>32</th>
<th>35</th>
</tr>
</thead>
<tbody>
<tr>
<td>42</td>
<td>43</td>
<td>44</td>
<td>45</td>
<td>46</td>
<td>47</td>
<td>48</td>
<td>48</td>
<td>50</td>
<td>51</td>
</tr>
<tr>
<td>72</td>
<td>75</td>
<td>77</td>
<td>78</td>
<td>79</td>
<td>83</td>
<td>87</td>
<td>88</td>
<td>135</td>
<td>151</td>
</tr>
</tbody>
</table>

a) Make a stemplot of times spent of the Web.

*Amount of Time Students Spend on the Internet (per day)*

```
<table>
<thead>
<tr>
<th>Amount</th>
<th>Times</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>14</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1</td>
</tr>
</tbody>
</table>
```

**KEY:** 0 | 7 = 7 minutes on the internet

b) Describe the distribution.

The distribution of time students spend on the internet per day is skewed right. The median time spent on the internet is 46.5 minutes and the inter-quartile range is 47 minutes. There is an outlier at 151 minutes.

Note: The median and IQR were used to describe center and spread because the data is skewed. If the data were symmetric, the mean and standard deviation would have been given instead.
Practice Problems:

30. The age of members of a local cycling club are shown below:

14, 17, 18, 18, 19, 20, 20, 21, 23, 24, 25, 25, 31, 37, 39, 53, 59, 73

a) Make a stemplot of the ages of the members of the cycling club.

b) Make a modified boxplot and list the five number summary. If there are any outliers, show supporting calculations.

c) Describe the distribution. Make sure to address shape, center, spread, and unusual features. Write in complete sentences and always include context.
Comparative Plots

Examining multiple data sets is a common practice in statistics. For example, you may be interested in determining if the distribution of grades on the Chapter 1 test in Mr. Setzer’s classes differ from the grades earned by students’ of Ms. Kreshover. When you are comparing two distributions, graph the data on one diagram then examine how the shapes, centers, spreads, and unusual features of the distributions compare.

Comparative Plots for Categorical Data

*Side-by-side bar graphs* are used when comparing multiple sets of categorical data. Below is an example of a side-by-side bar graph. Notice that by placing bars side-by-side, you can easily compare similarities and differences between the categories and between the different distributions (in this case the sales per year).
Comparative Plots for Quantitative Data

Side-by-side boxplots and back-to-back stemplots are used to compare distributions of quantitative data. Examples of each type of plot can be found on the next page.

Side-by-Side Boxplots:
Side-by-side boxplots should be drawn on one scale. This allows for easy comparison between the two distributions. Also, make sure you provide a label for each boxplot.

Analysis Example:
The distributions of Question 1 and Question 2 are both skewed left. The median of Question 2 (about 44) is greater than the median of Questions 1 (about 35); however Question 1’s IQR (about 8) is slightly more than the IQR of Questions 2 (approximately 6). Question 2 has five values that are considered outliers, while Question 1 has only one outlier.

Back-to-Back Stemplots:
Below is an example of a back-to-back stemplot. Notice that one set of data extends to the right of the stem and the other set of data extends to the left of the stem. Make sure that you provide a title for the entire graph and a key for each set of data.

Annual Monthly Rainfall in Chicago and Miami
(in inches)

<table>
<thead>
<tr>
<th>Chicago</th>
<th>Miami</th>
</tr>
</thead>
<tbody>
<tr>
<td>96</td>
<td>9</td>
</tr>
<tr>
<td>9764</td>
<td>015</td>
</tr>
<tr>
<td>76650</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>08</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>

Key: 7 9 Represents 970 feet
10 5 Represents 1050 feet
Example:

Here are the numbers of home runs that Babe Ruth hit in his 15 years with the New York Yankees, 1920 to 1934.

<table>
<thead>
<tr>
<th>54</th>
<th>59</th>
<th>35</th>
<th>41</th>
<th>46</th>
<th>25</th>
<th>47</th>
<th>60</th>
<th>54</th>
<th>46</th>
</tr>
</thead>
<tbody>
<tr>
<td>49</td>
<td>46</td>
<td>41</td>
<td>34</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Babe Ruth's home run record for a single year was broken by another Yankee, Roger Maris, who hit 61 home runs in 1961. Here are Maris's home run totals for his 10 years in the American league.

| 13 | 23 | 26 | 16 | 33 | 61 | 28 | 39 | 14 | 8  |

Compare these two distributions by constructing a back-to-back stemplot. Comment on the shape, center, spread, and unusual features of the distributions. Make sure you are using comparative language!

**HOME RUNS PER SEASON**

**RUTH**

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**MARIS**

<table>
<thead>
<tr>
<th>0</th>
<th>6</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>39 HRS</td>
</tr>
</tbody>
</table>

**Shape:** Ruth's home run distribution is fairly symmetric, while for Maris, the distribution is skewed right.

**Center:** Since the Maris distribution is skewed, we should use median to represent center. The center of the distribution of Ruth home runs is greater than that of Maris, with a median of 46 compared to 24.5 for Maris.

**Spread:** Using IQR as the measure of spread, both distributions are equally spread with an IQR of 19.

**Unusual Features:** Neither the Ruth nor the Maris distributions contain outliers (although if Maris' 61 home runs was 62, it would have been an outlier). The Maris distribution has a gap between 39 and 61, while the Ruth distribution has no gaps.
Practice Problems:

31. The scores of 18 first year college women on the Survey of Study Habits and Attitudes (this psychological test measures motivation, study habits and attitudes toward school) are given below:

154 109 137 115 152 140 154 178 101
103 126 126 137 165 165 129 200 148

The college also administered the test to 20 first-year college men. There scores are also given:

108 140 114 91 180 115 126 92 169 146
109 132 75 88 113 151 70 115 187 104

Construct side-by-side modified boxplots for the two data sets. Write a narrative comparison of the shape, center, spread, and unusual features of the two distributions.