



Understanding and Comparing Distributions

Chapter 4



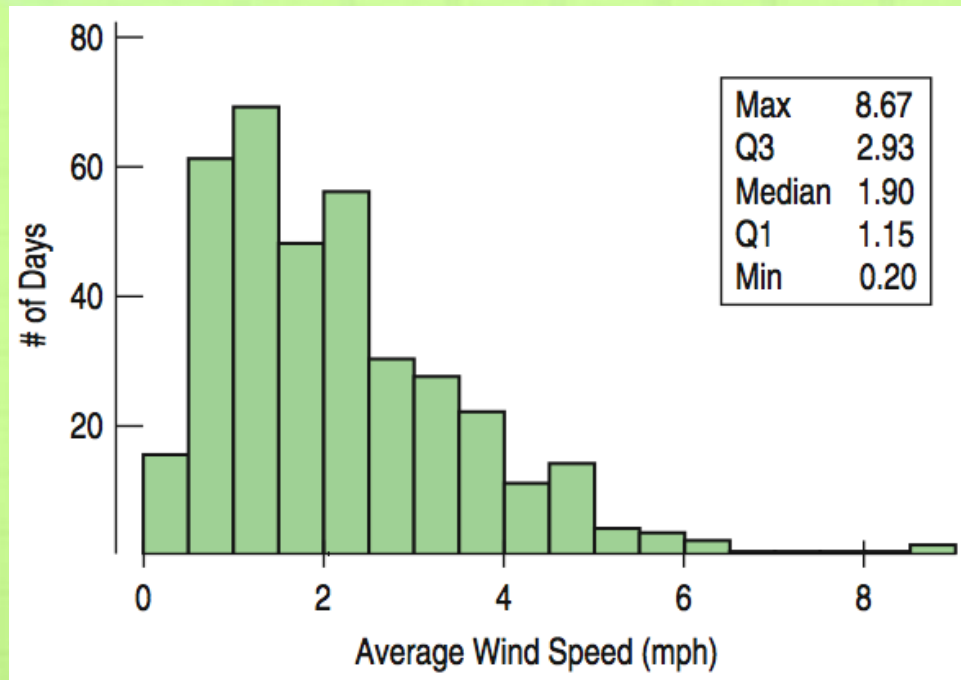
Objectives:

- Boxplot
- Calculate Outliers
- Comparing Distributions
- Timeplot



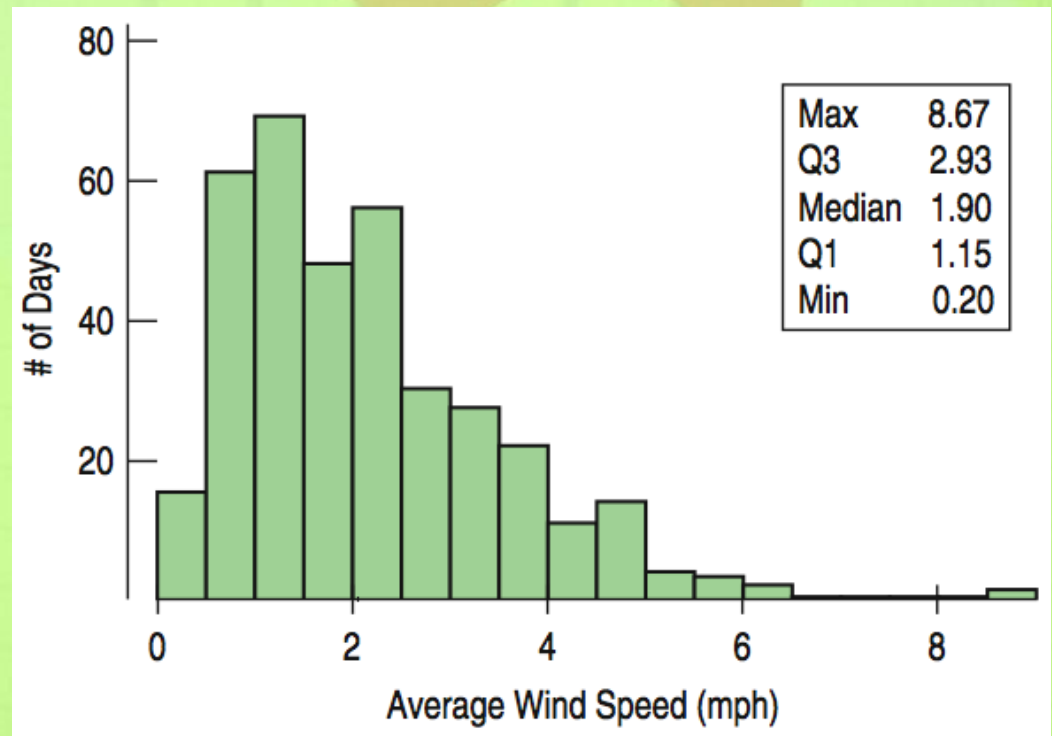
The Big Picture

- We can answer much more interesting questions about variables when we compare distributions for different groups.
- Below is a histogram of the *Average Wind Speed* for every day in 1989.



The Big Picture (cont.)

- The distribution is unimodal and skewed to the right.
 - The high value may be an outlier
- Comparing distributions can be much more interesting than just describing a single distribution.



The Five-Number Summary

- The **five-number summary** of a distribution reports its median, quartiles, and extremes (maximum and minimum).
- Example: The five-number summary for for the daily wind speed is:

Max	8.67
Q3	2.93
Median	1.90
Q1	1.15
Min	0.20



The Five-Number Summary

- Consists of the minimum value, Q_1 , the median, Q_3 , and the maximum value, listed in that order.
- Offers a reasonably complete description of the center and spread.
- Calculate on the TI-83/84 using 1-Var Stats.
- Used to construct the Boxplot.
- Example: Five-Number Summary
 - 1: 20, 27, 34, 50, 86
 - 2: 5, 10, 18.5, 29, 33



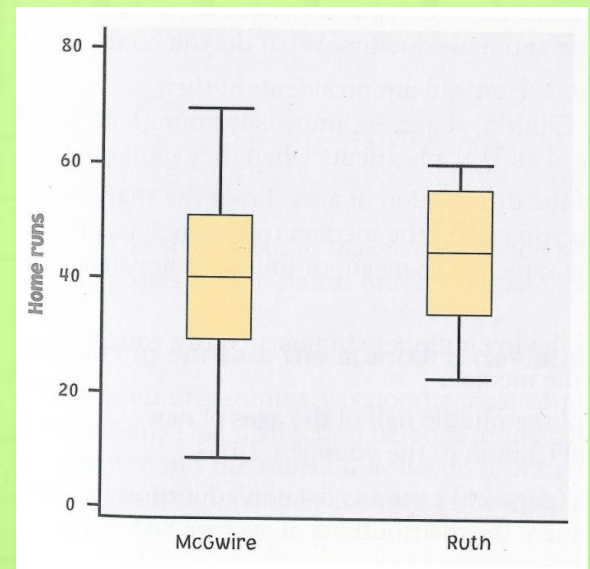
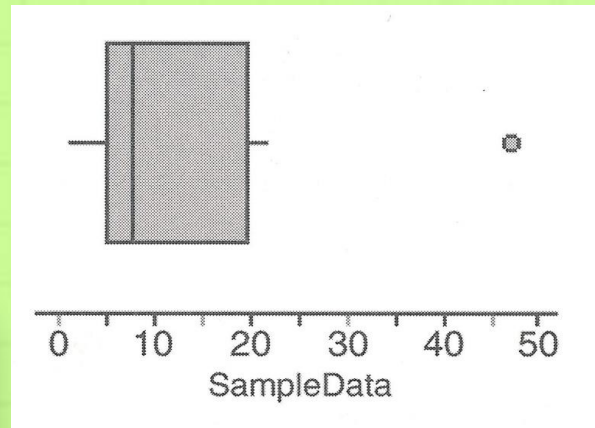
Daily Wind Speed: Making Boxplots

- A **boxplot** is a graphical display of the five-number summary.
- Boxplots are useful when comparing groups.
- Boxplots are particularly good at pointing out outliers.



Boxplot

- A graph of the Five-Number Summary.
- Can be drawn either horizontally or vertically.
- The box represents the IQR (middle 50%) of the data.
- 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 scale below the boxplot and label.
2. Draw a vertical line above the value of Q_1 , this forms the left end of the box.
3. Draw a vertical line above the value of Q_3 , this forms the right end of the box.
4. Draw a vertical line above the value of the median and complete the box.
5. Extend the “left whisker” to the minimum value.
6. Extend the “right whisker” to the maximum value.
7. Give a descriptive title to the graph.



Example Boxplot

- Data:

20, 25, 25, 27, 28, 31, 33, 34, 36, 37, 44,
50, 59, 85, 86

- Use TI-83/84



What About Outliers?

- Recall that an outlier is an extremely small or extremely large data value when compared with the rest of the data values.
- What should we do about outliers?
 - Try to understand them in the context of the data.
 - Data error
 - Special nature to the data



OUTLIERS

- If there are any clear outliers and you are reporting the mean and standard deviation, report them with the outliers present and with the outliers removed. The differences may be quite revealing.
- Note: The median and IQR are not likely to be affected by the outliers.
- The following procedure allows us to check whether a data value can be considered as an outlier.



Testing for Outliers

- IQR is used to determine if extreme values are actually outliers
- An observation is an outlier if it falls more than 1.5 times IQR below Q_1 or above Q_3 .
- To test for outliers
 1. Construct an upper and lower fence
 2. Upper Fence = $Q_3 + (1.5)IQR$
 3. Lower Fence = $Q_1 - (1.5)IQR$
 4. If an observation falls outside the fences (ie. Greater than the upper fence or less than the lower fence) than it is an outlier.

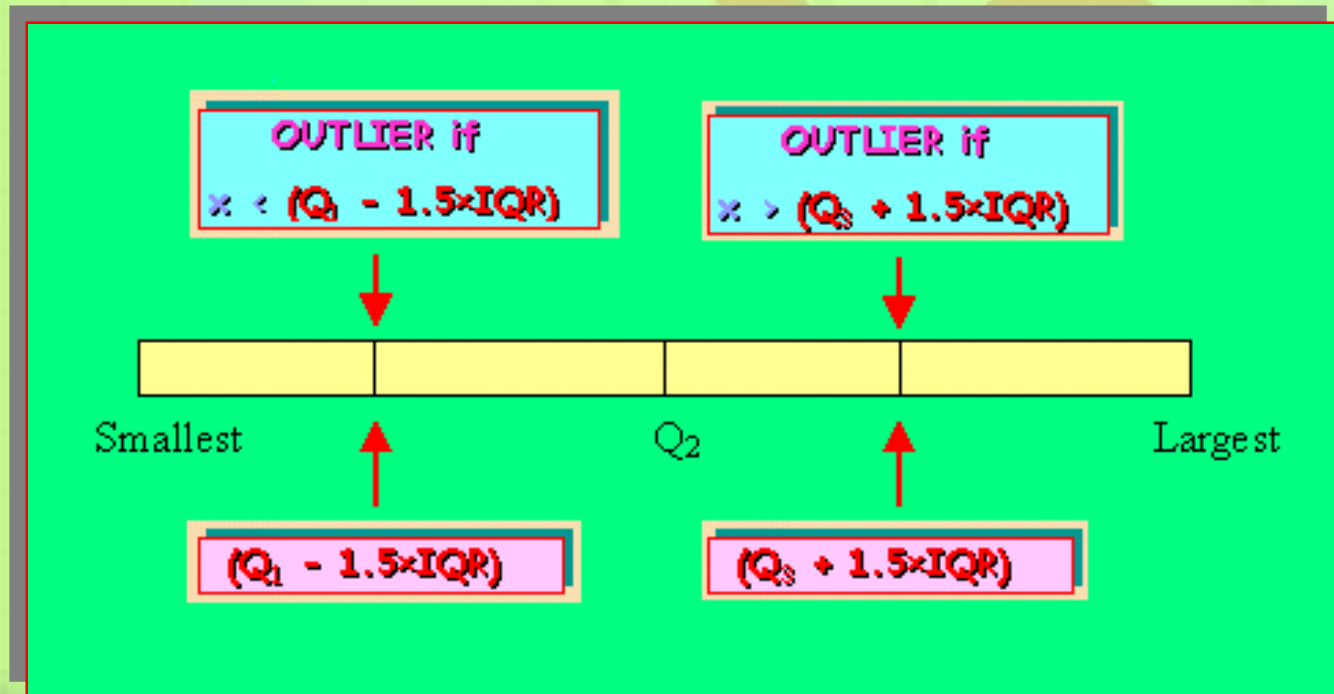


More Outliers

- Far Outlier – Data values farther than 3 IQRs from the quartiles.



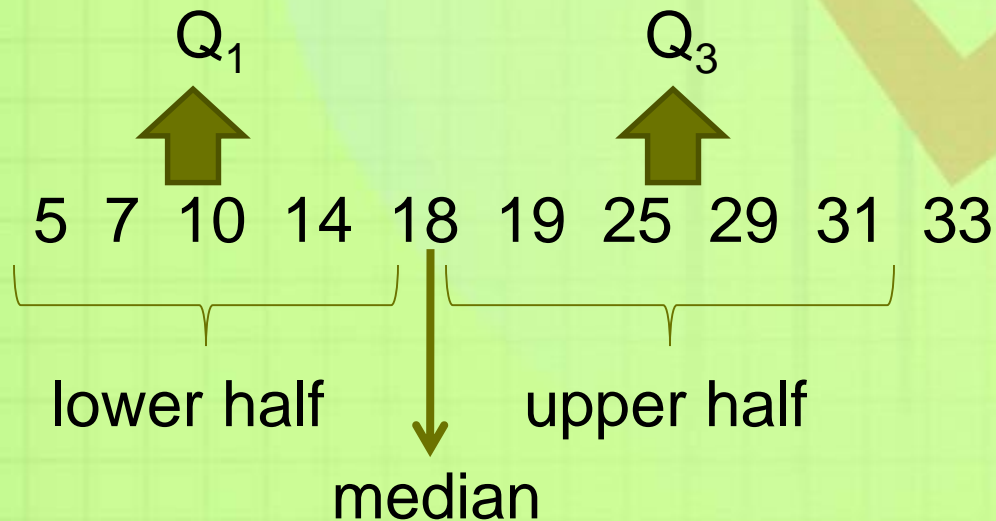
Illustration of Outliers



Example 2: Even number data set

- Data 5, 7, 10, 14, 18, 19, 25, 29, 31, 33

Find Q_1 , M , Q_3 , IQR and outliers.



- $IQR = 29 - 10 = 19$
- Upper Fence = $29 + (1.5)IQR = 29 + 28.5 = 57.5$
- Lower Fence = $10 - (1.5)IQR = 10 - 28.5 = -18.5$
- No Outliers

Your Turn: Calculate Outliers

- The data below represent the 20 countries with the largest number of total Olympic medals, including the United States, which had 101 medals for the 1996 Atlanta games. Determine whether the number of medals won by the United States is an outlier relative to the numbers for the other countries.
- Data values - 63, 65, 50, 37, 35, 41, 25, 23, 27, 21, 17, 17, 20, 19, 22, 15, 15, 15, 15, 101.



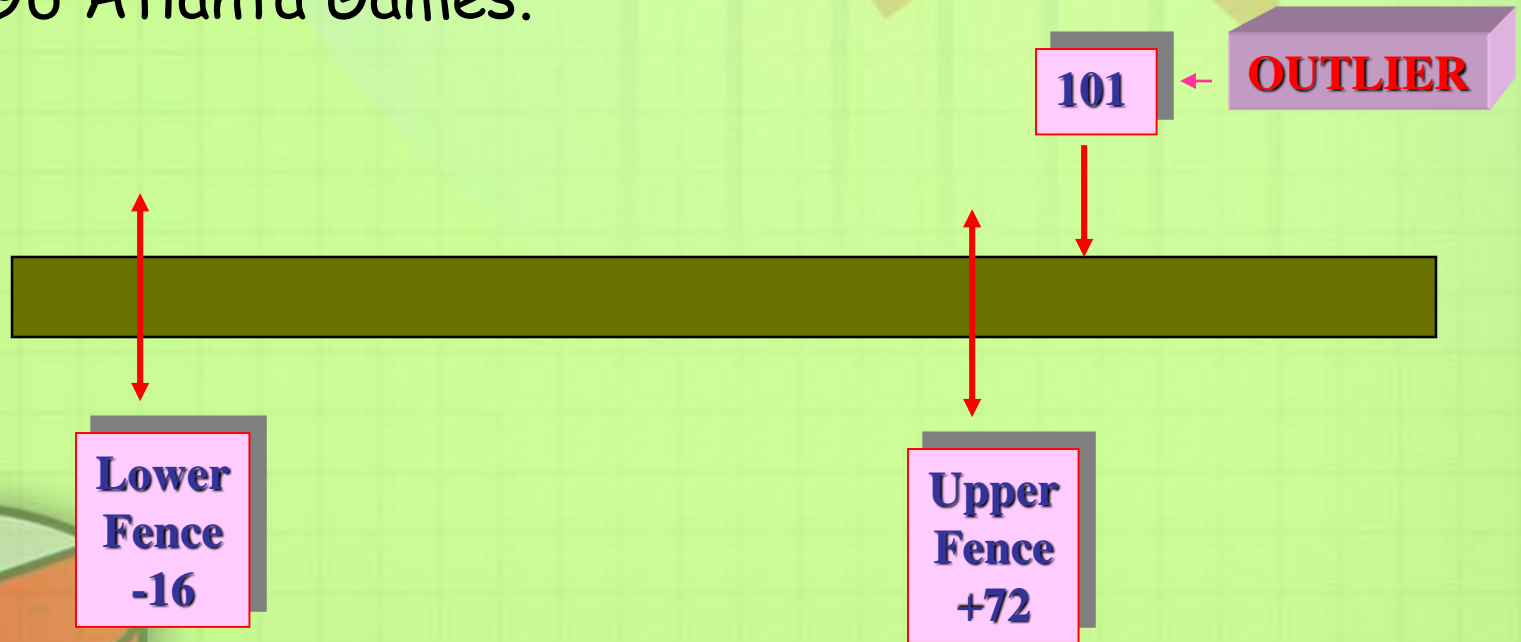
Solution

- The IQR = $39 - 17 = 22$.
- Lower Fence, $Q_1 - 1.5 \times \text{IQR} = 17 - (1.5 \times 22) = -16$.
- And Upper Fence, $Q_3 + 1.5 \times \text{IQR} = 39 + (1.5 \times 22) = 72$.
- Since, $101 > 72$, the value of 101 is an outlier relative to the rest of the values in the data set.
- That is, the number of medals won by the United States is an outlier relative to the numbers won by the other 19 countries for the 1996 Atlanta Olympic Games.



Solution (cont.)

- Pictorial Representation for the OUTLIER of the Number of Olympic Medals Won by the United States in 1996 Atlanta Games.



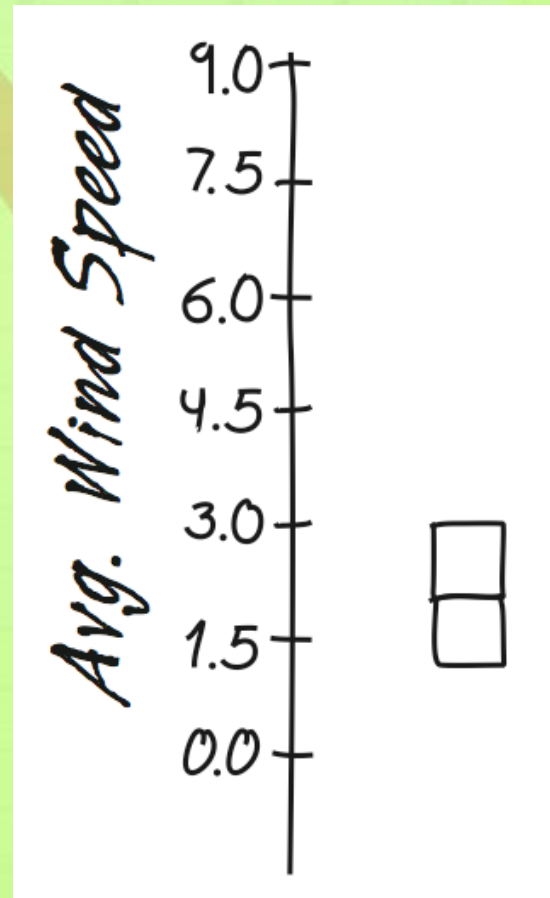
Modified Boxplot

- Plots outliers as isolated points, where regular boxplots conceal outliers.
- From now on when we say “boxplot”, we mean “modified boxplot”. The modified boxplot is more useful than the boxplot.
- Constructing a Modified Boxplot.
 1. Same as a boxplot with the exception of the “whiskers”.
 2. Extend the “left whisker” to the minimum value if there are no outliers or to the last data value less than or equal to the lower fence if there are outliers.
 3. Extend the “right whisker” to the maximum value if there are no outliers or to the last data value less than or equal to the upper fence.
 4. Outliers (either low or high) are then represented by a dot or an asterisk.



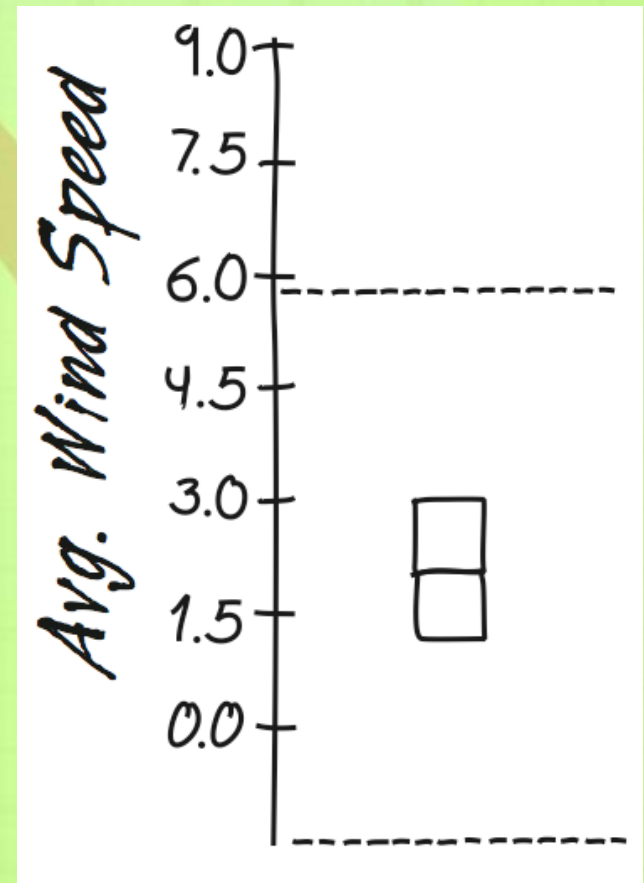
Example: Constructing Boxplots

1. Draw a single vertical axis spanning the range of the data. Draw short horizontal lines at the lower and upper quartiles and at the median. Then connect them with vertical lines to form a box.



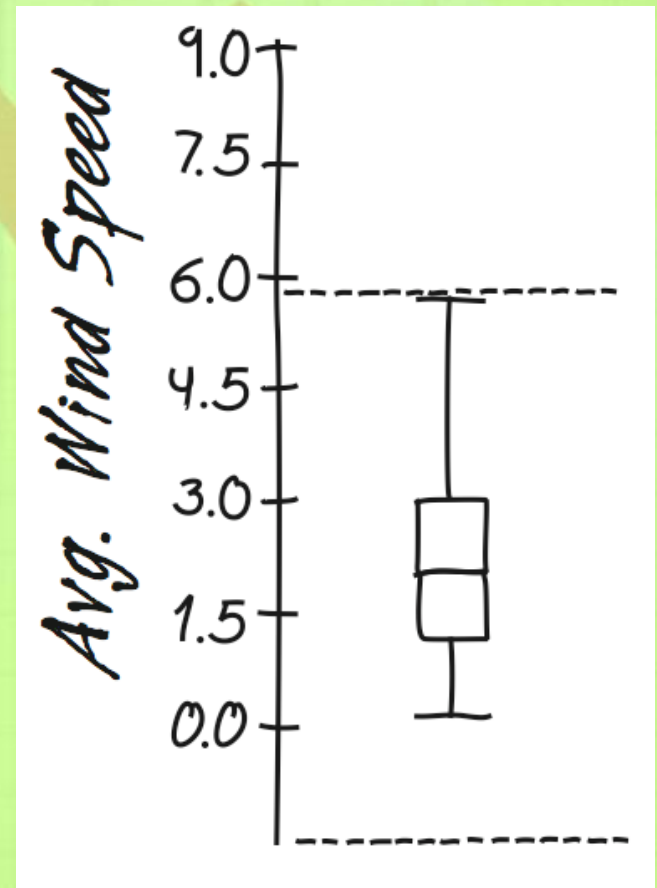
Example: Constructing Boxplots (cont.)

2. Erect “fences” around the main part of the data.
 - The upper fence is 1.5 IQRs above the upper quartile.
 - The lower fence is 1.5 IQRs below the lower quartile.
 - Note: the fences only help with constructing the boxplot and should not appear in the final display.



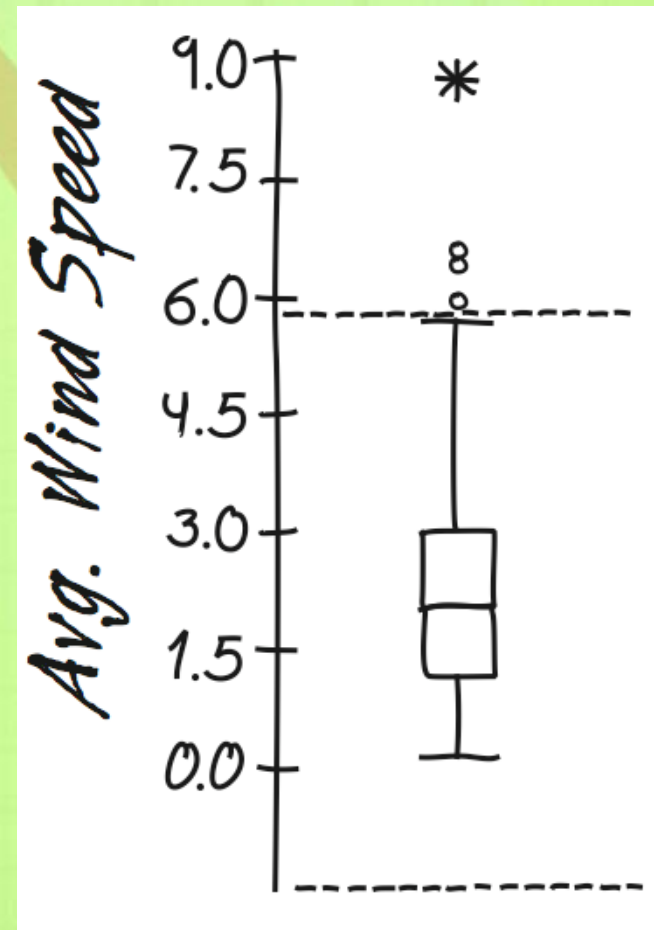
Constructing Boxplots (cont.)

- Use the fences to grow “whiskers.”
 - Draw lines from the ends of the box up and down to the *most extreme data values found within the fences*.
 - If a data value falls outside one of the fences, we do *not* connect it with a whisker.

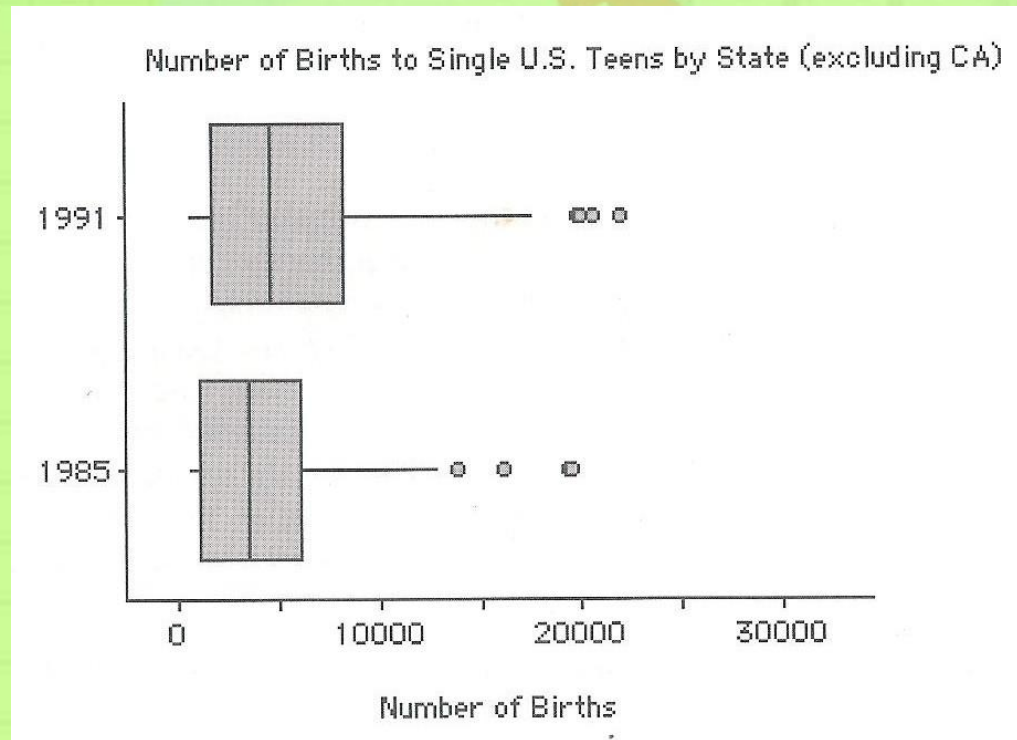


Constructing Boxplots (cont.)

4. Add the **outliers** by displaying any data values beyond the fences with special symbols.
 - We often use a different symbol for “far outliers” that are farther than 3 IQRs from the quartiles.

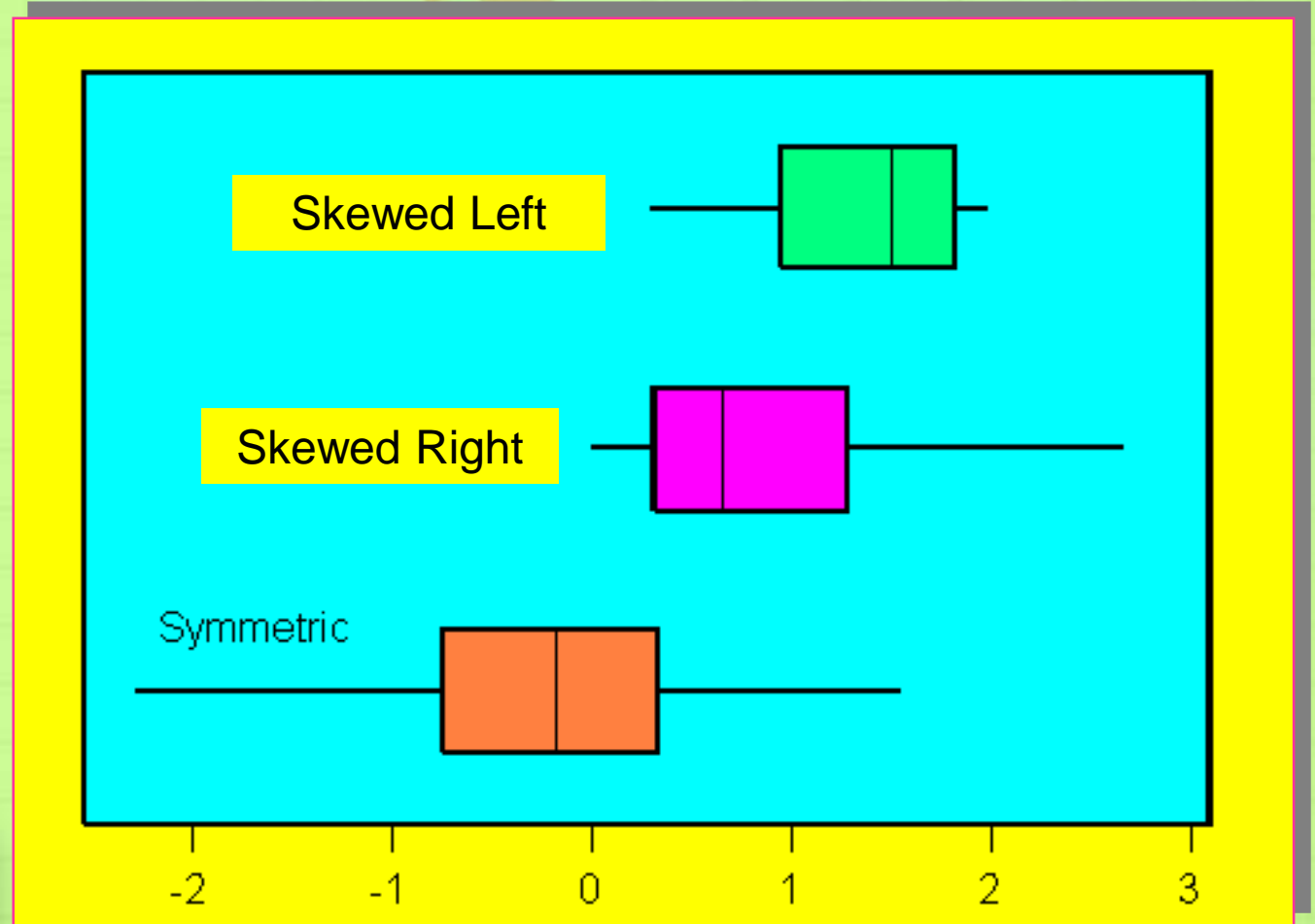


Example Modified Boxplot



- Construct Modified Boxplot using TI-83/84
- Data: 20, 25, 25, 27, 28, 31, 33, 34, 36, 37, 44, 50, 59, 85, 86

Information That Can Be Obtained From a Box Plot



Information That Can Be Obtained From a Box Plot - Looking at the Median

- If the median is close to the center of the box, the distribution of the data values will be approximately symmetrical.
- If the median is to the left of the center of the box, the distribution of the data values will be Skewed Right.
- If the median is to the right of the center of the box, the distribution of the data values will be Skewed Left.

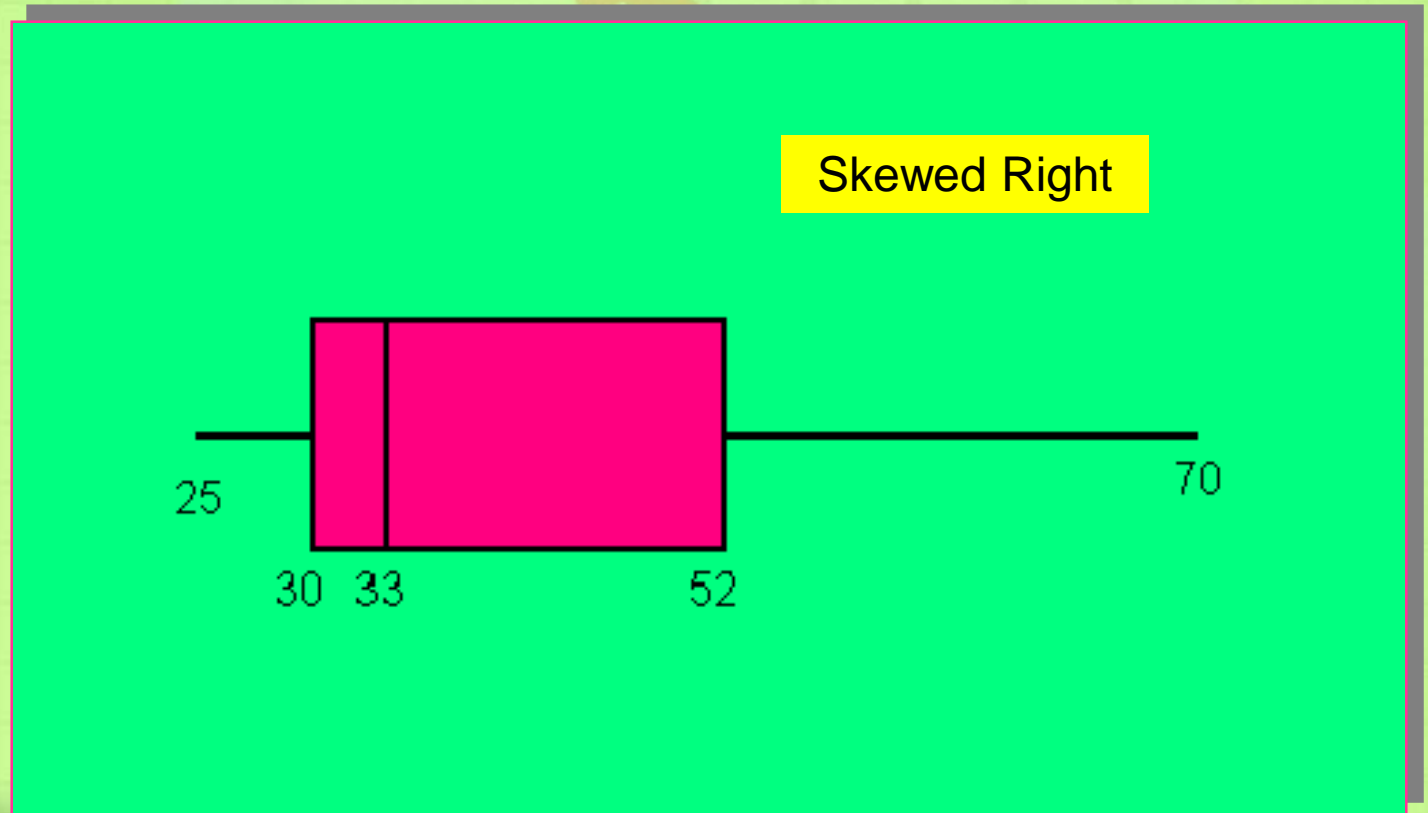


Information That Can Be Obtained From a Box Plot - Looking at the Length of the Whiskers

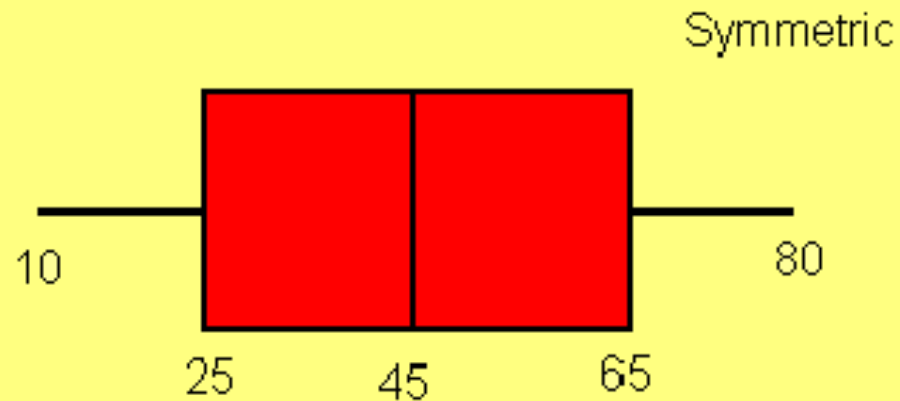
- If the whiskers are approximately the same length, the distribution of the data values will be approximately symmetrical.
- If the right whisker is longer than the left whisker, the distribution of the data values will be Skewed Right.
- If the left whisker is longer than the right whisker, the distribution of the data values will be Skewed Left.



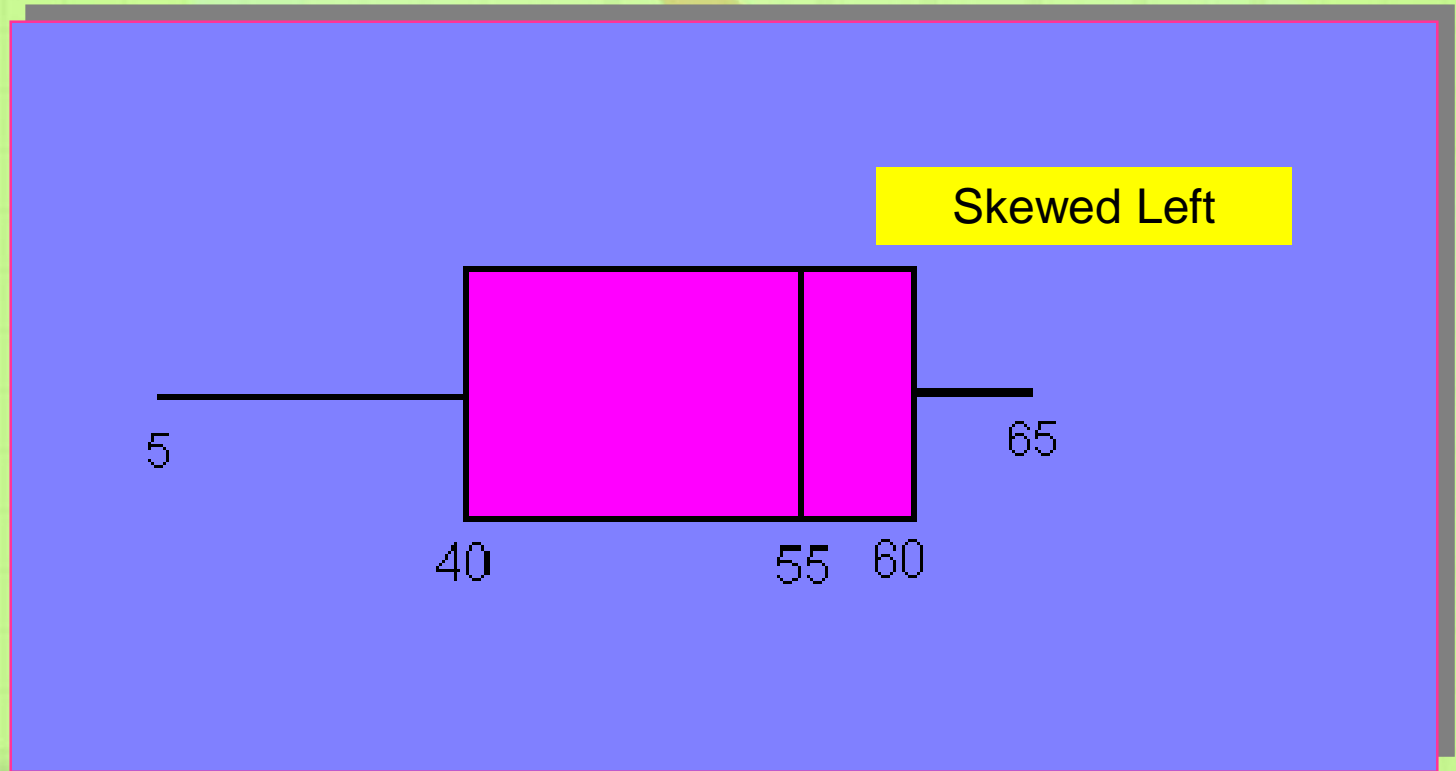
Box Plot Displaying Skewed Right Distribution



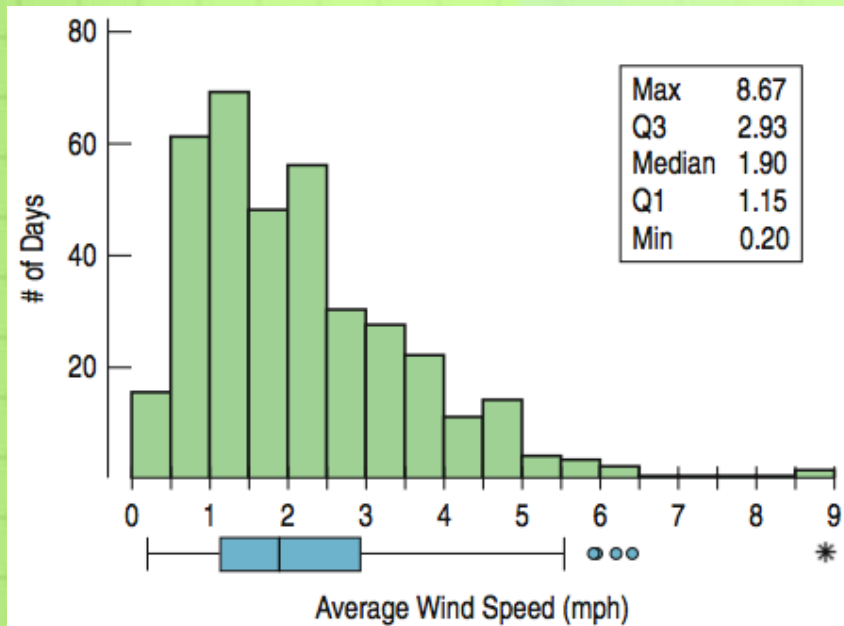
Box Plot Displaying a Symmetrical Distribution



Box Plot Displaying a Skewed Left Distribution



Comparing Distributions

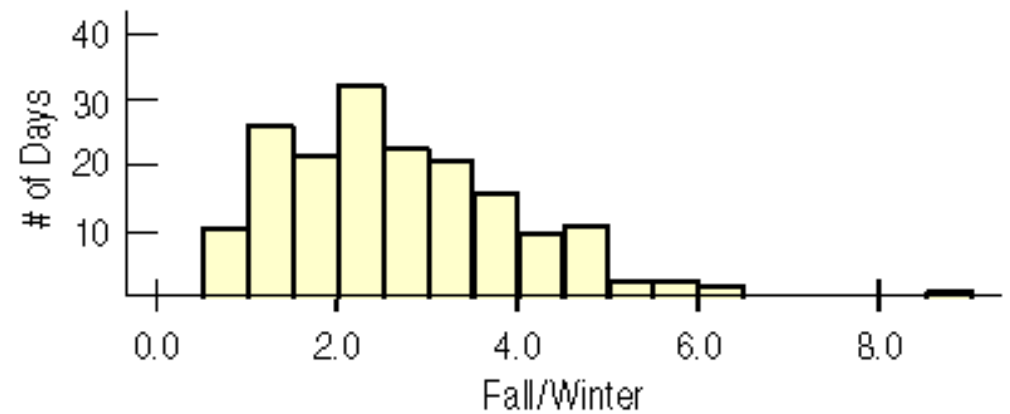
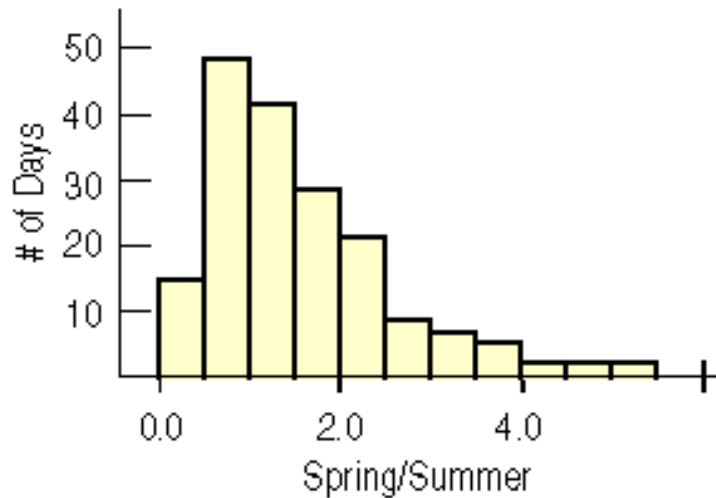


- Compare the histogram and boxplot for daily wind speeds:
- How does each display represent the distribution?
- The shape of a distribution is not always evident in a boxplot.
- Boxplots are particularly good at pointing out outliers.



Comparing Groups

- It is almost always more interesting to compare groups.

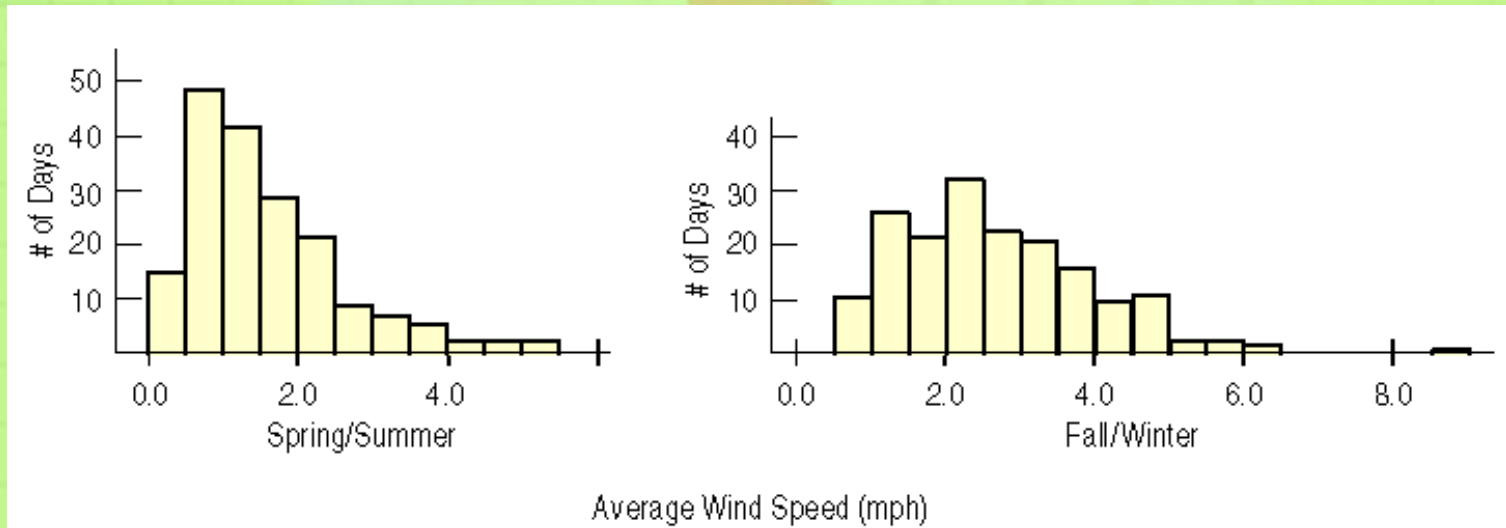


Average Wind Speed (mph)



- With histograms, note the shapes, centers, and spreads of the two distributions.
- When using histograms to compare data sets make sure to use the same scale for both sets of data.
- What does this graphical display tell you?

Comparing Groups

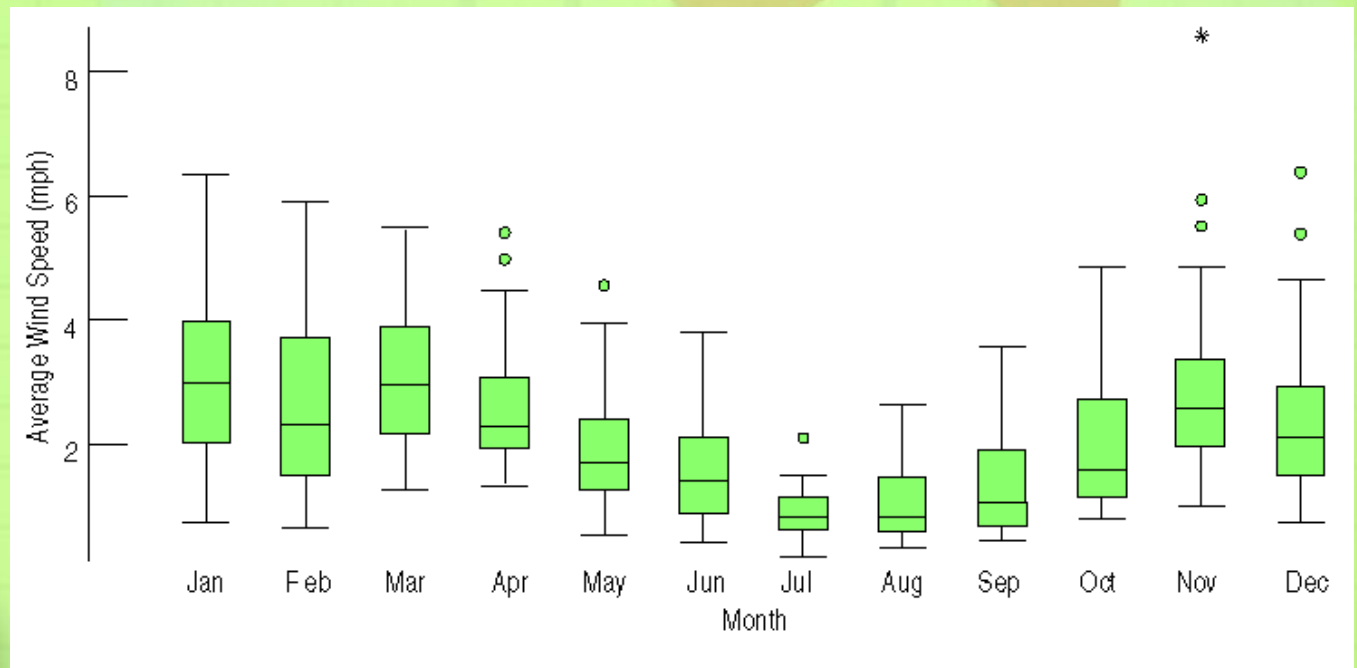


- The shapes, centers, and spreads of these two distributions are strikingly different.
- During spring and summer (histogram on the left), the distribution is skewed to the right. A typical day has an average wind speed of only 1 to 2 mph.
- In the colder months (histogram on the right), the shape is less strongly skewed and more spread out. The typical wind speed is higher, and days with average wind speeds above 3 mph are not unusual.



Comparing Groups (cont.)

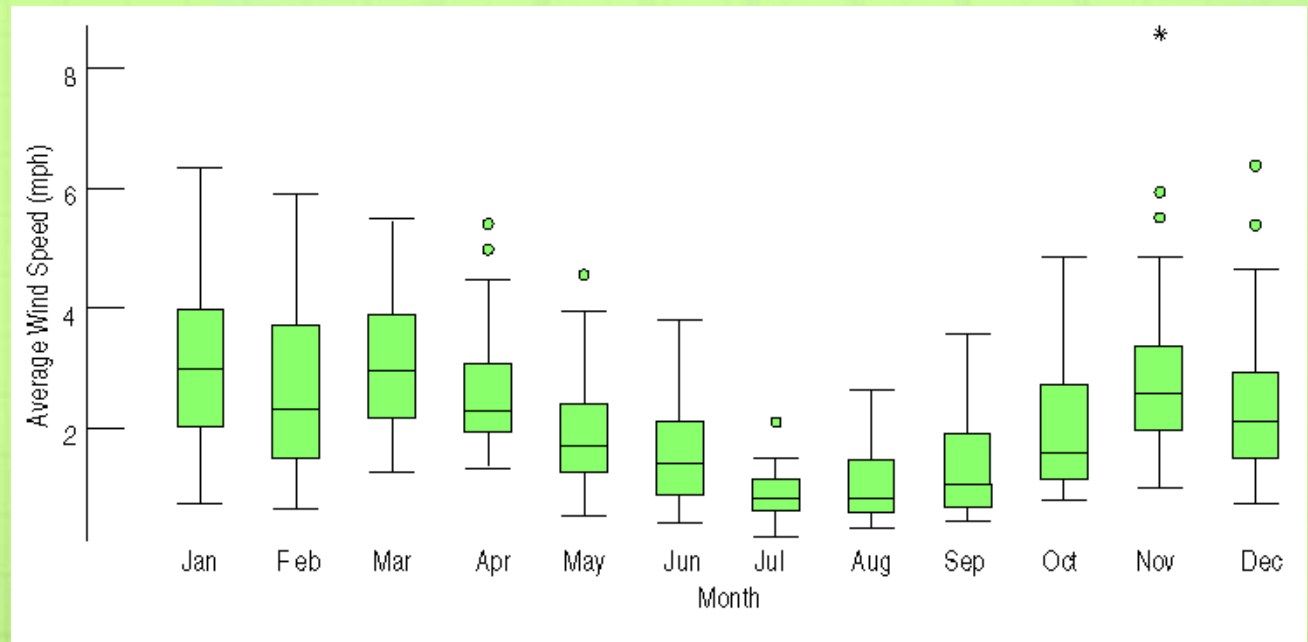
- Boxplots offer an ideal balance of information and simplicity, hiding the details while displaying the overall summary information.
- We often plot them side by side for groups or categories we wish to compare.



- What do these boxplots tell you?

Comparing Groups (cont.)

- By placing the boxplots side by side, we can easily see which groups have higher medians, which have the greater IQRs, where the middle 50% of the data is located in each group, and which have the greater overall range
- When the boxes are placed in order, we can get a general idea of patterns in both the centers and the spreads.
- Equally important, we can see past any outliers in making these comparisons because they've been displayed separately.

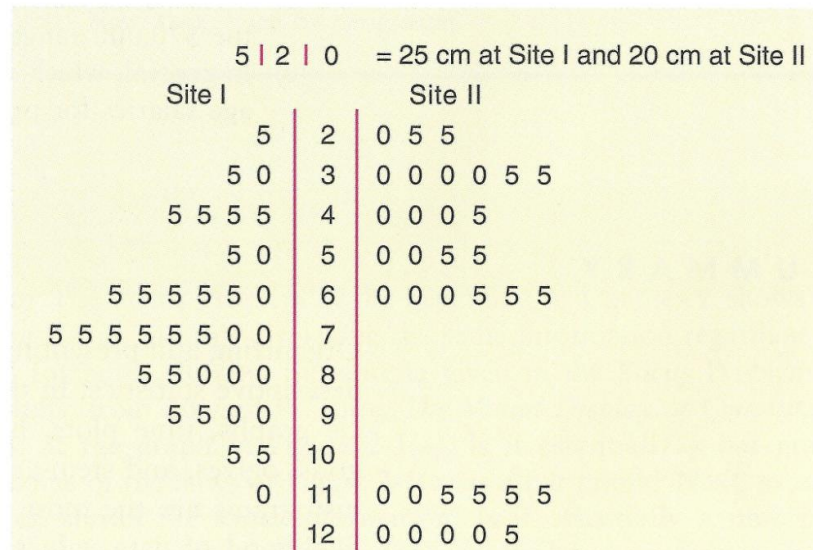


Comparing Distributions Using a StemPlot

- Variation of the Stemplot
- Back-to-Back Stemplot – Compare 2 related distributions. Common single stem with leaves on each side ordered out from the common stem.

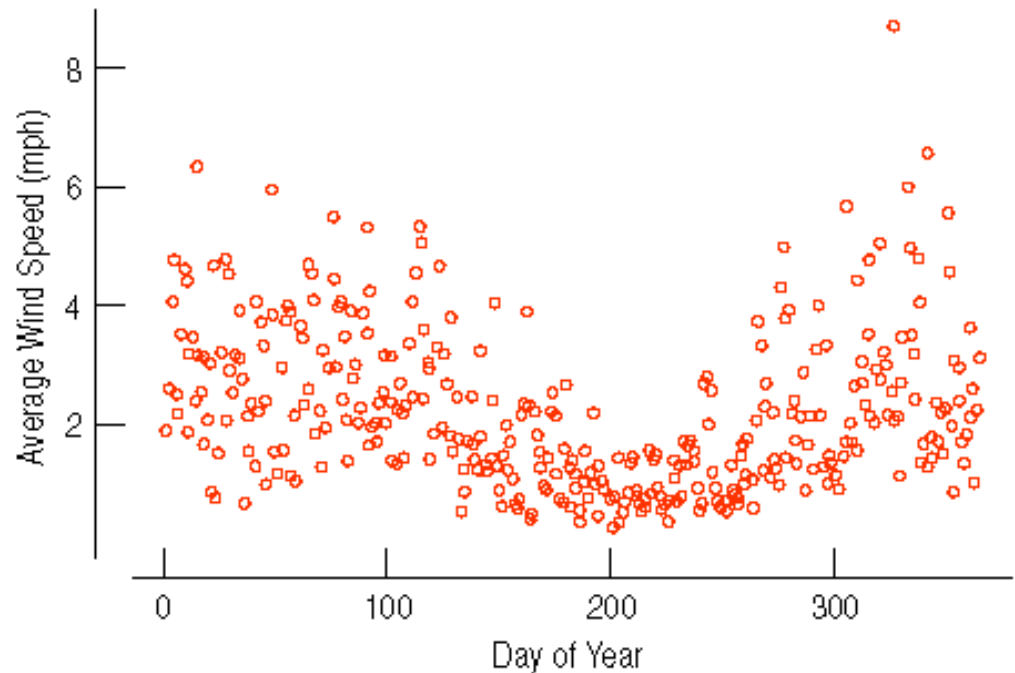
FIGURE 2-27

Depth (in cm)
of Artifact
Location



Timeplots: Order, Please!

- For some data sets, we are interested in how the data behave over time. In these cases, we construct **timeplots** of the data.



Time Plots or Line Graphs

- What is a time plot? A time plot is a plot which displays data that are observed over a given period of time.
- Note: From a time plot, one can observe and analyze the behavior of the data over time.



Time Plot

- To make a Time Plot –
 1. Put time on the horizontal scale.
 2. Put the variable being measured on the vertical scale.
 3. Connect the data points by lines.
- Interpreting Time Plots – look for
 1. An overall pattern – TREND (upward or downward)
 2. Outliers – strong deviations from the overall pattern or trend.



Time Plots -- Example

- Example: The following table gives the number of hurricanes for the years 1981 to 1990.

Year	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Number	7	2	3	5	7	4	3	5	7	8

- Display the data with a time series graph.

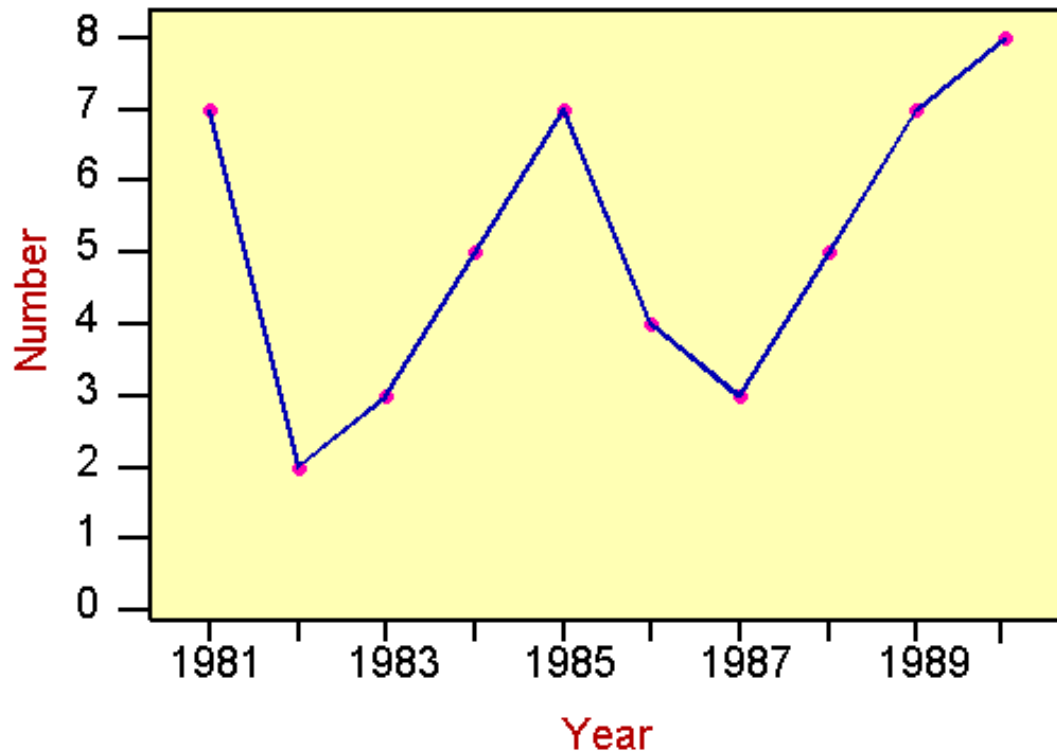


Time Plots – Example Continued

- Graph seem to display an upward trend over the years.
- Highest number was in 1990.



Time Series Plot for the Number of Hurricanes from 1981 - 1990.

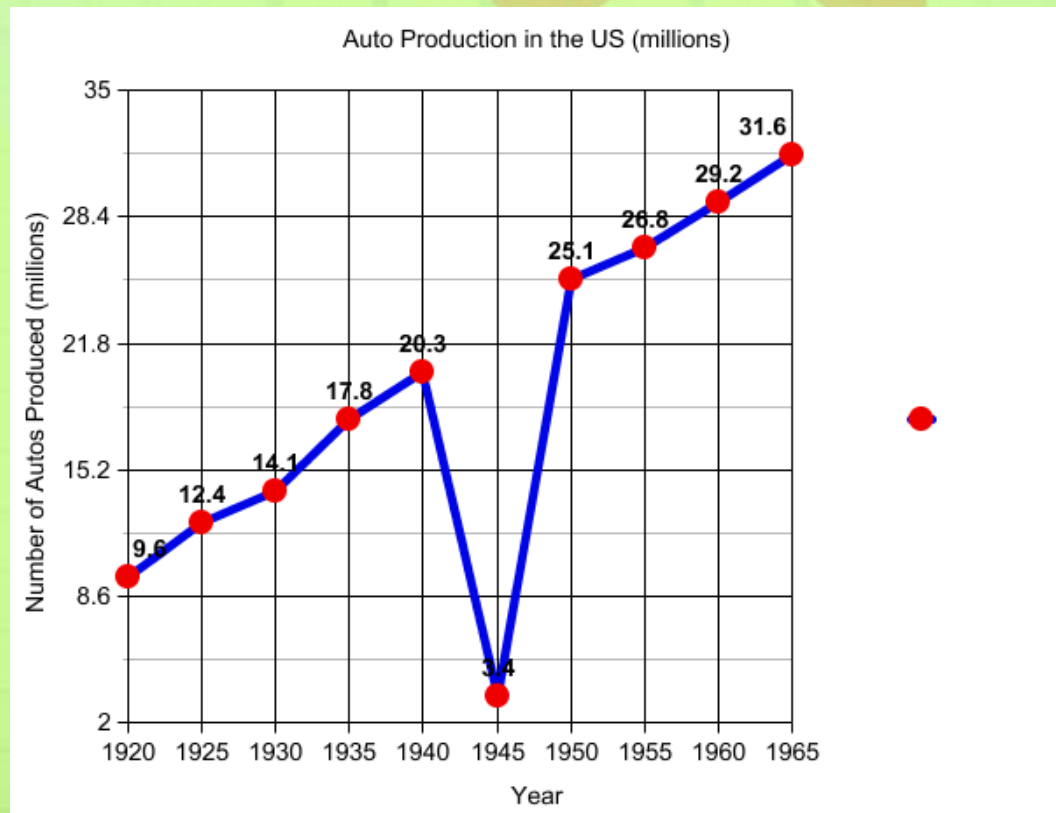


Time Plot

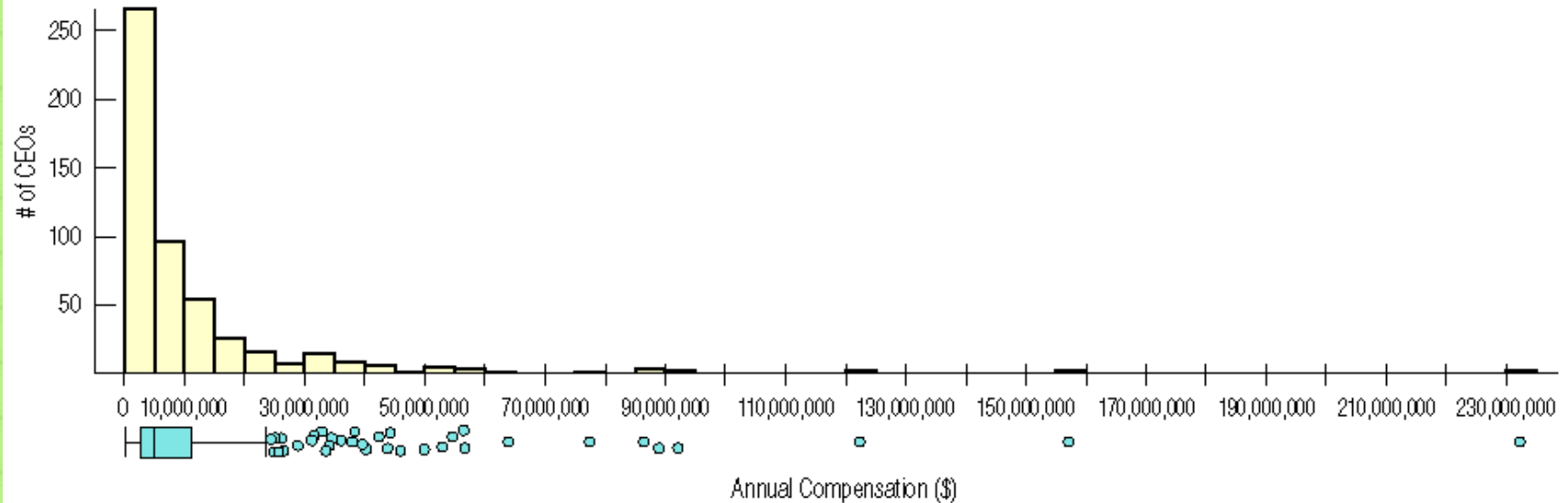
May contain outliers, deviations from the pattern.

Interpolation – to predict a value from the pattern between known values. (within the data)

Extrapolation – to predict a value by extending the pattern into the future (outside the data). Dangerous, no guarantee the pattern continues.



Re-expressing Skewed Data to Improve Symmetry

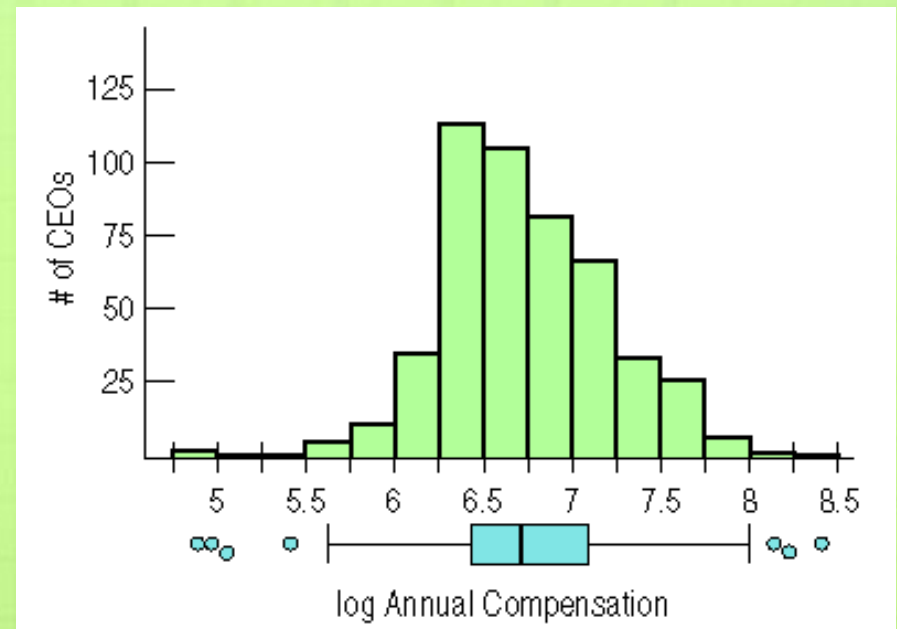
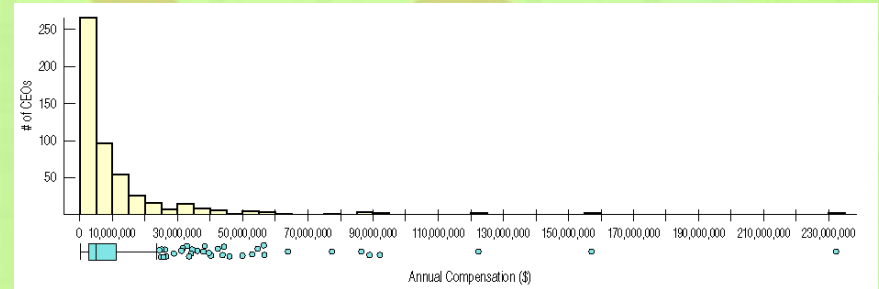


- When the data are skewed it can be hard to summarize them simply with a center and spread, and hard to decide whether the most extreme values are outliers or just part of a stretched out tail.
- How can we say anything useful about such data?
- The secret is to *re-express* the data by applying a simple function (logarithms, square roots, and reciprocals) to each value.



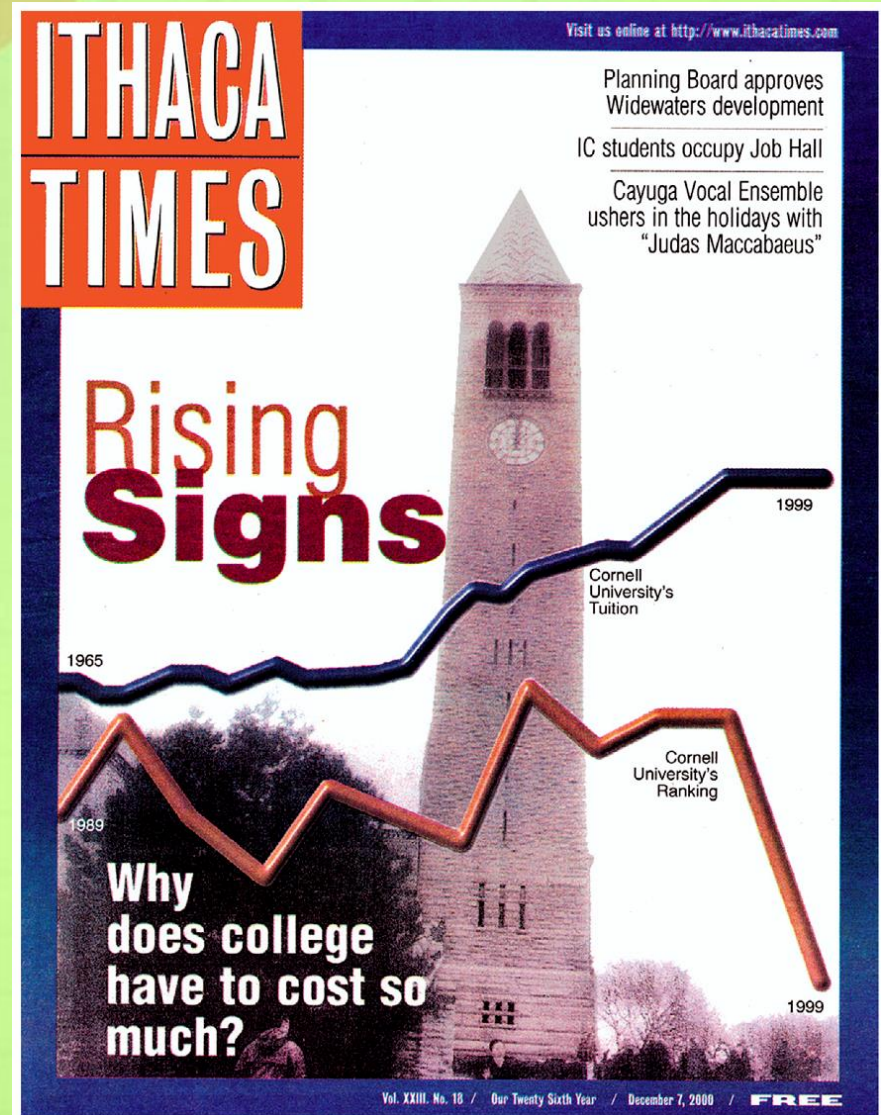
Re-expressing Skewed Data to Improve Symmetry (cont.)

- One way to make a skewed distribution more symmetric is to **re-express** or **transform** the data by applying a simple function (e.g., logarithmic function).
- Note the change in skewness from the raw data (top) to the transformed data (right):



What Can Go Wrong?

- Avoid inconsistent scales, either within the display or when comparing two displays.
- Label clearly so a reader knows what the plot displays.
 - When comparing two groups, be sure to compare them on the same scale.



What have we learned?

- We've learned the value of comparing data groups and looking for patterns among groups and over time.
- We've seen that boxplots are very effective for comparing groups graphically.
- We've experienced the value of identifying and investigating outliers.
- We've graphed data that has been measured over time against a time axis and looked for long-term trends both by eye and with a data smoother.



Assignment

- Exercises pg. 95 – 103: #6-8, 11, 12, 19-26, 28, 29, 31
- Read Ch-6, pg. 104 - 128

