Starter Ch. 5 2005 #1a

The goal of a nutritional study was to compare the caloric intake of adolescents living in rural areas of the United States with the caloric intake of adolescents living in urban areas of the United States. A random sample of ninth-grade students from one high school in a rural area was selected. Another random sample of ninth graders from one high school in an urban area was also selected. Each student in each sample kept records of all the food he or she consumed in one day.

The back-to-back stemplot below displays the number of calories of food consumed per kilogram of body weight for each student on that day.

Urban		Rural
99998876	2	
44310	3	2334
97665	3	56667
20	4	02224
	4	56889
	5	1

Stem: tens
Leaf: ones

(a) Write a few sentences comparing the distribution of the daily caloric intake of ninth-grade students in the rural high school with the distribution of the daily caloric intake of ninth grade students in the urban high school.

CW Ch. 4: Regression

L1	L2
87	88
84	86
83	73
81	67
78	83
65	80
50	78
78	?
93	?
86	?

- **Create a scatterplot**
- **Find the equation of the regression line**
- **Predict the scores**

Chapter 5:

Understanding and Comparing Distributions

Chapter Objectives

At the end of this chapter you should be able to:

- Calculate numerical summaries of quantitative data to describe center **appropriate** (median, mean, quartiles) and spread (range, interquartile range, standard deviation).
- Describe the characteristics of various numerical summaries with emphasis on the effects of outliers.
- Interpret the values of the numerical summaries for a particular data set.
- Match graphical displays of quantitative data to the values of the summary statistics.
- Explore different ways of examining the relationship between two variables when one is quantitative and the other is categorical.

Starter Chapter 5: Agility Test

"*Performance of fourth-grade students on an agility test*"

COPY THESE DATA

Boys: 22, 17, 18, 29, 22, 22, 23, 24, 23, 17, 21

Girls: 25, 20, 12, 19, 28, 24, 22, 21, 25, 26, 25, 16, 27, 22

- Enter these data in L1 (*Boys*) and L2 (*Girls*).
- Construct a side-by-side boxplot
- Write a few sentences comparing the distributions above. (Be Sure to comment on the shape, center, spread and outliers).
- How do these fourth graders compare in terms of agility?

Finding the median, quartiles and inter-quartile range.

Example 1: Find the median and quartiles for the data below.

12, 6, 4, 9, 8, 4, 9, 8, 5, 9, 8, 10

Order the data

4, 4, 5, 6, 8, 8, 8, 9, 9, 9, 10, 12

Q_1 Q_2 Q_3

Lower Quartile = 5.5

Median = 8

Upper Quartile = 9

Inter-Quartile Range = 9 – 5.5 = 3.5

Finding the median, quartiles and inter-quartile range.

Example 2: Find the median and quartiles for the data below.

6, 3, 9, 8, 4, 10, 8, 4, 15, 8, 10

Order the data

3, 4, 4, 6, 8, 8, 8, 9, 10, 10, 15

Q_1 = 4 (Lower Quartile = 4)

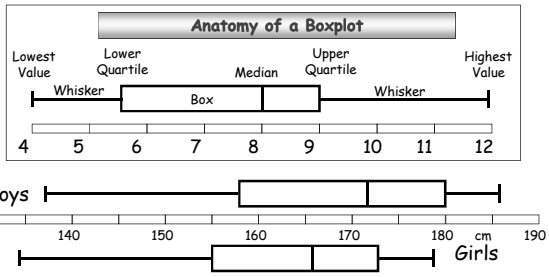
Q_2 = 8 (Median = 8)

Q_3 = 10 (Upper Quartile = 10)

Inter-Quartile Range = 10 - 4 = 6

Box-and-Whisker Plots (Boxplots)

- Boxplots are useful for comparing **two or more sets** of data like that shown below for heights of boys and girls in a class.

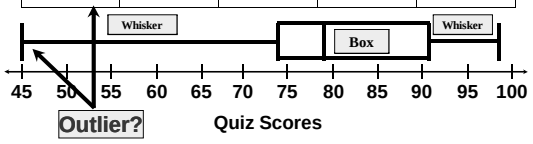


Box-and-Whisker Plots (Boxplots)

- A boxplot summarizes data using the median (Q_2), upper (Q_3) and lower quartiles (Q_1), and the extreme (least and greatest) values. This is called the **5-Number Summary.** It allows you to see important characteristics of the data at a glance.

Anatomy of a Boxplot

Min = 45	Q1 = 74	Med = 79	Q3 = 91	Max = 98



Drawing a Box Plot

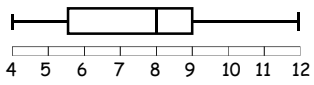
Example 1: Draw a Box plot for the data below

4, 4, 5, 6, 8, 8, 8, 9, 9, 9, 10, 12

Q_1 = 5.5 (Lower Quartile = 5.5)

Q_2 = 8 (Median = 8)

Q_3 = 9 (Upper Quartile = 9)



Drawing a Box Plot

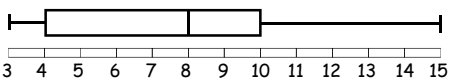
Example 2: Draw a Box plot for the data below

3, 4, 4, 6, 8, 8, 8, 9, 10, 10, 15

Q_1 = 4 (Lower Quartile = 4)

Q_2 = 8 (Median = 8)

Q_3 = 10 (Upper Quartile = 10)



Drawing a Box Plot

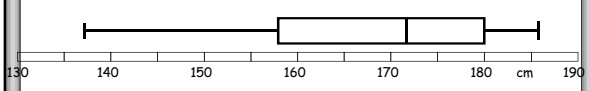
Question: Stuart recorded the heights in cm of boys in his class as shown below. Draw a box plot for this data.

137, 148, 155, 158, 165, 166, 166, 171, 171, 173, 175, 180, 184, 186, 186

Q_1 = 158 (Lower Quartile = 158)

Q_2 = 171 (Median = 171)

Q_3 = 180 (Upper Quartile = 180)



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 the daily wind speed is:

Max	8.67
Q3	2.93
Median	1.90
Q1	1.15
Min	0.20

Box-and-Whisker Plots (Boxplots)

Study your boxplot to determine what it is telling you. Make a statement about what it is saying, then support the statement with facts from your graph. You should include the following in your interpretation:

- Range or spread of the data and what it means to your graph
- Quartiles—compare them. What are they telling you about the data?
- Median- this is an important part of the graph, and should be an important part of the interpretation.
- Percentages should be used to interpret the data, where relevant

Box-and-Whisker Plots (Boxplots)

Example:

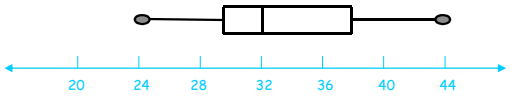
- The gas mileages in miles per gallon (mpg) of 4-cylinder manual transmission cars are in the table below.
- Find the extreme values, Q_1, Q_2, and Q_3. Interpret.

28	32	42	37
30	25	44	38
24	32	33	44
38	34	30	44
31	28	31	29
39	29	32	29

Box-and-Whisker Plots (Boxplots)

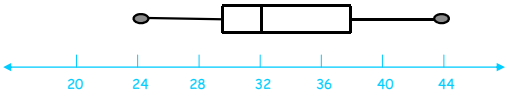
Example:

- Min = 24
- Q_1 = 29
- Q_2 = 32
- Q_3 = 38
- Max = 44



Miles per gallon (mpg)

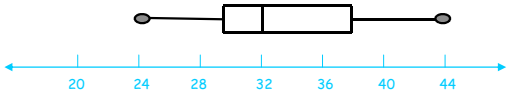
Box-and-Whisker Plots (Boxplots)



Miles per gallon (mpg)

- The boxplot clearly shows that there is a lot of different gas mileages on various 4-cylinder vehicles.
- The mileage ranged from 24 miles per gallon (mpg) to a high of 44 mpg. This is a 20 miles per gallon spread, which in car mileage is quite a bit of difference.

Box-and-Whisker Plots (Boxplots)



Miles per gallon (mpg)

- The 1st quartile reads as 29 mpg which means that 75% of the vehicles in this study got 29 mpg or more.
- The 3rd quartile tells us that 25% of these cars got 38 mpg or higher which is really good mileage.
- The median cuts the data in half. The median is 32 mpg. Therefore half the cars in the study received 32 mpg or higher.

5-Number Summary, Boxplots

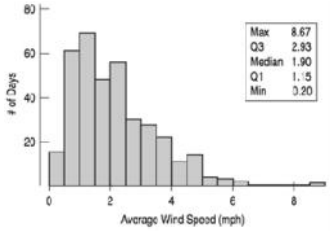
- The **5-Number Summary** provides a reasonably complete description of the center and spread of distribution

MIN	Q1	MED	Q3	MAX

- We can visualize the 5-Number Summary with a ***boxplot.***
- ***Upper fence*** = Q_3 **+ 1.5 IQR**
- ***Lower fence*** = Q_1 **- 1.5 IQR**
- **The fences are just for construction and are not part of the display. Any data beyond the fences are *outliers*.**

The Big Picture… Read p. 80

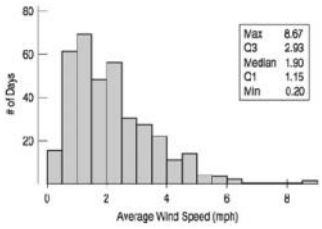
- We can answer much more interesting questions about variables when we **compare distributions for different groups.**
- Below is a histogram of the ***Average Wind Speed*** for every day in 1989.



The Big Picture…

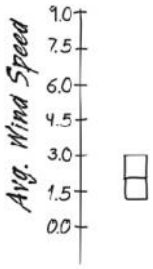
- The distribution is unimodal and skewed to the right.
- The high value may be a possible outlier.

- **Median daily wind speed is about 1.90 mph and the IQR is reported to be 1.78 mph.**
- **Can we say more?**



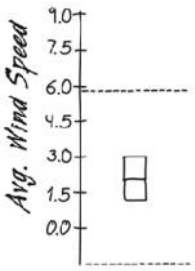
Construction Boxplots

1) Draw a single vertical (or horizontal) 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.



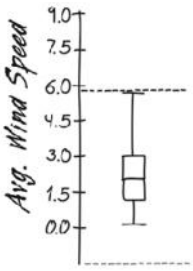
Construction Boxplots

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.**



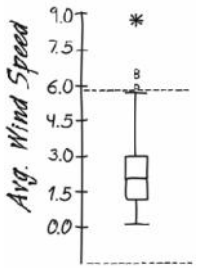
Construction Boxplots

3) 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.



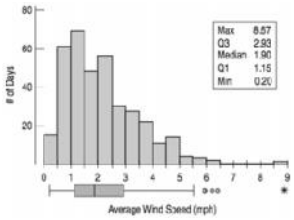
Construction Boxplots

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.



Wind Speed: Making Boxplots

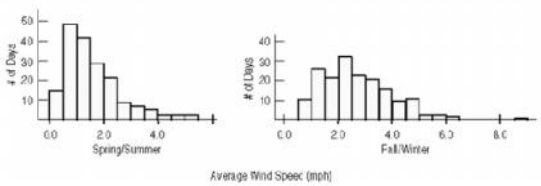
- Compare the histogram and boxplot for daily wind speeds:



- How does each display represent the distribution?

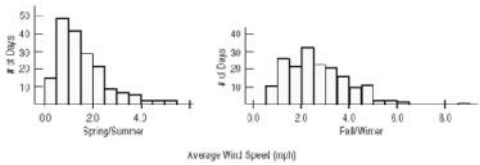
Comparing Groups

- It is almost always more interesting to compare groups.
- With histograms, note the shapes, centers, and spreads of the two distributions.



- What does this graphical display tell you?

Comparing Groups

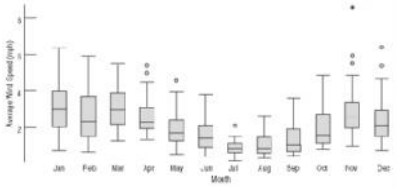


The shapes, centers, and spreads of these two distributions are strikingly different. During spring and summer, the distribution is skewed to the right. A typical day during these warm months has an average wind speed of only 1 to 2 mph, and few have average speeds above 3 mph.

In the colder months, however, 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.

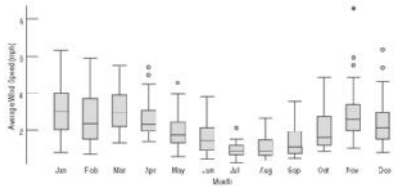
Comparing Groups

- 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



- Here we see that wind speeds tend to decrease in the summer. The months which the winds are both strongest and most variable are November through March. And there was one remarkably windy day in November.

TI-83/84: **Boxplots**

- Press STAT PLOT.
  - Select Plot1
    - Turn Plot 1 On.
    - Select the Boxplot Type.
    - Specify list L_1.
- Press ZOOM.
  - Select ZoomStat (#9) and press ENTER.

TI-83/84: **Boxplots**

- Press TRACE.
  - Use the arrow keys to see the values of the minimum, Q1, the median, Q3, and the maximum.

Anything Unusual/Outlier?

Do any unusual features stick out?

- Don't ignore outliers. Outliers can affect data summaries, but we can't just throw them out. We should call attention to them, not conceal them.
- The best policy is to make note of the outliers and try to figure out more information about them.
- If you can't identify a reason for the point, do calculations both with and without the outlier and see how much it affects the outcome.

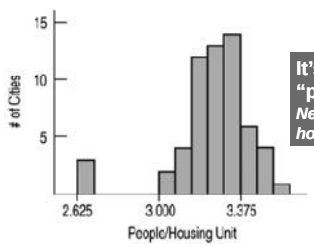
Anything Unusual/Outlier?

Do any unusual features stick out?

- Sometimes it's the unusual features that tell us something interesting or exciting about the data.
- You should always mention any stragglers, or **outliers**, that stand off away from the body of the distribution.
- Are there any **gaps** in the distribution? If so, we might have data from more than one group.

Anything Unusual/Outlier?

- The following histogram has **possible outliers**—there are three cities in the leftmost bin.



It's a good idea to say "possible" outliers.
Next time we will learn how to test for outliers.

Determining Outliers

"1.5 • IQR Rule"

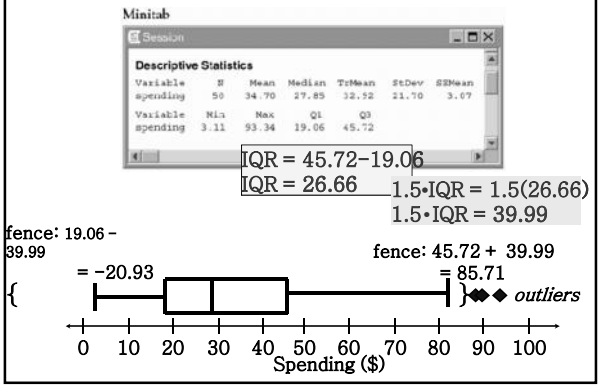
- **InterQuartile Range "IQR":** Distance between **Q1** and **Q3**. Resistant measure of spread...only measures middle 50% of data.
  - **IQR = Q3 - Q1 {width of the "box" in a boxplot}**
- **1.5 IQR Rule:** If an observation falls more than 1.5 IQRs above Q3 or below Q1, it is an **outlier**.

Why 1.5? According to John Tukey, 1 IQR seemed like too little and 2 IQRs seemed like too much...

1.5 • IQR Rule

- To determine outliers:
  - ☑ Find 5-Number Summary
  - ☑ Determine IQR
  - ☑ Multiply: 1.5 x IQR
  - ☑ Set up "***fences***" Q1- (1.5 • IQR) and Q3 + (1.5 • IQR)
  - ☑ Observations "***outside***" the fences are ***outliers***.

Outlier Example



Modified Boxplot

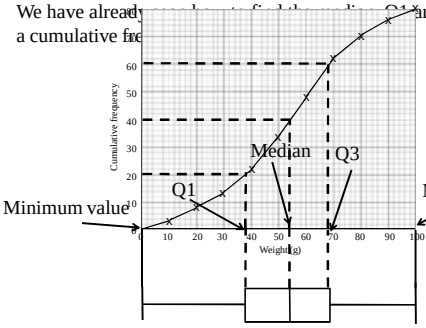
- A boxplot in which the outliers are indicated.
- Extend the whiskers from the box to the smallest and largest values that are *within* the inner fences.
- Any values that are outside the inner fences should be drawn as individual dots. These dots represent **outliers**.

Example:
Draw a modified boxplot of the sample
9, 13, 39, 40, 42, 46, 49, 54, 55, 60, 84.

Boxplots

Box plots from cumulative frequency diagrams

We have already [illegible] and Q2 from a cumulative fr[illegible]

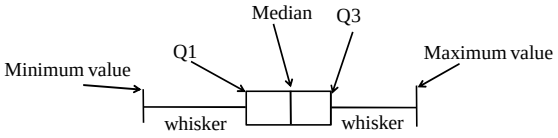


Extend the working out lines below the *x*-axis

Draw the box plot like this

Boxplots

Looking at the box plot on its own:



Boxplots

Box plots from raw data

Find the median of this set of data.
Fred keeps a record of his journey times to school each morning
These are his time to the nearest minute for 25 days:

12, 16, 16, 18, 18, 18, 19, 19, 19, 20, 21,
21, 20, 19, 20, 19, 25, 26, 17, 18, 27, 30, 30

Minimum value
Maximum value
Median
Lower Quartile
Upper Quartile

number of pieces of data, n

Median position $\frac{n+1}{2} = \frac{25+1}{2} = \frac{26}{2} = 13$

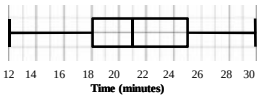
Lower Quartile position $\frac{n+1}{4} = \frac{25+1}{4} = \frac{26}{4} = 6.25$

Upper Quartile position $\frac{3(n+1)}{4} = \frac{3(25+1)}{4} = 18.75$

Boxplots

Minimum value	12
Maximum value	30
Median	21
Lower Quartile	18
Upper Quartile	25

Draw a sensible scale:



Boxplots

Now do this:

Draw a box plot for this data:

18, 8, 19, 15, 27, 13, 10, 4, 8, 31, 26, 11, 29, 28, 23

Solution:

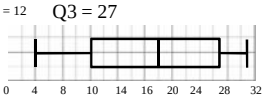
Arrange the data in order of size:

4, 8, 8, 10, 11, 13, 15, 18, 19, 23, 26, 27, 28, 29, 31

Median position $\frac{n+1}{2} = \frac{15+1}{2} = 8$ Median = 18

Q1 position $\frac{n+1}{4} = \frac{15+1}{4} = 4$ Q1 = 10

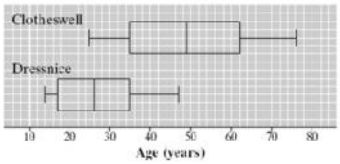
Q3 position $\frac{3(n+1)}{2} = \frac{3(15+1)}{2} = 12$ Q3 = 27



Boxplots

Interpreting Boxplots

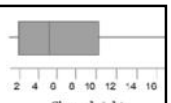
These box plots show the ages of shoppers in two clothes shops.



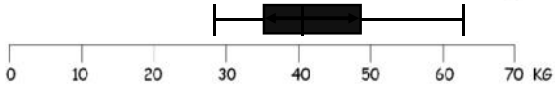
What can you say about the ages of the shoppers and what kind of shop are they?

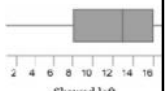
- Dressnice has a much lower median age, so the shoppers are younger.
- 75% of the Dressnice shoppers are younger than 75% of the Clotheswell shoppers.
- Clotheswell has broader appeal because the IQR is bigger.
- Dressnice is a shop for younger people, perhaps more fashionable.

Skewness



Positive skew: median closer to Q1 than Q3



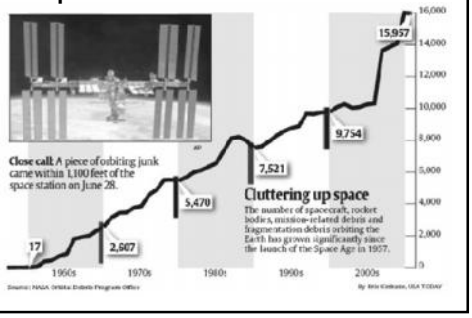


Negative skew: median closer to Q1 than Q3

Symmetrical distribution

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.



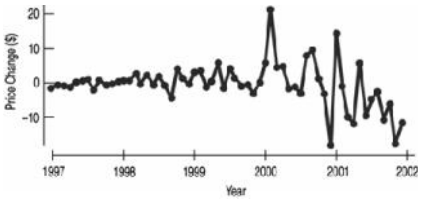
Timeplots: Order, Please!

- A timeplot of a variable plots each observation against the time a which it was measured.
- Always mark the time scale on the horizontal axis and the variable of interest on the vertical axis.
- If there are not too many points, connecting the points by lines helps show the pattern of changes over time.
- When describing a time plot, **do NOT use SOCS!!**
- Instead, describe the **TREND** you see over time!

Timeplots: Order, Please!

Look for:

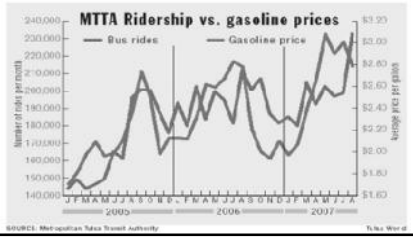
- **Trends** – overall pattern that indicates a long-term upward or downward movement over time.
- **Seasonal variation** – a pattern that repeats itself at regular time intervals.



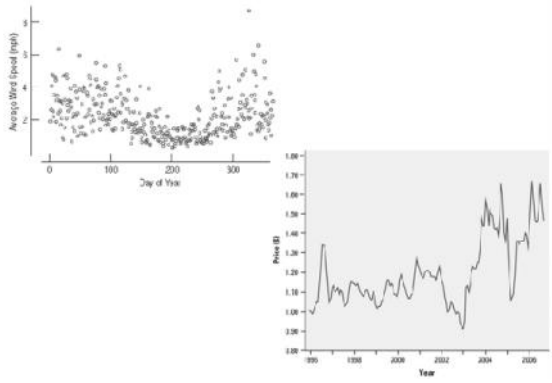
Timeplots: Order, Please!

Look for:

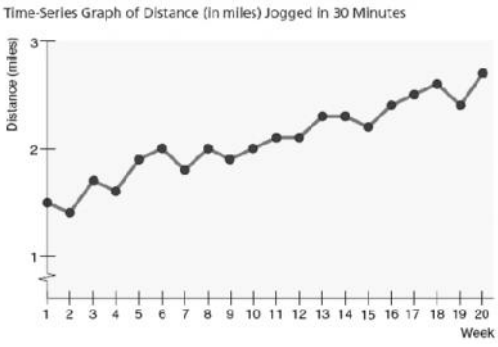
- **Trends** – overall pattern that indicates a long-term upward or downward movement over time.
- **Seasonal variation** – a pattern that repeats itself at regular time intervals.



Timeplots: Order, Please!



Timeplots: Order, Please!



Pattern of a Distribution "SOCS"

- ***Shape***
  - Modes: Major peaks in the distribution
  - Symmetric: The values smaller and larger than the midpoint are mirror images of each other
  - Skewed to the right: Right side of the graph extends much farther out than the left side.
  - Skewed to the left: Left side of the graph extends much farther out than the right side.
- ***Center (Location)***
  - Mean: The arithmetic average. Add up the numbers and divide by the number of observations.
  - Median: List the data from smallest to largest. If there is an odd number of data values, the median is the middle one in the list. If there is an even number of data values, average the middle two in the list

Pattern of a Distribution "SOCS"

- ***Spread***
  - Range: The difference in the largest and smallest value. (Max – Min)
  - Standard Deviation: Measures spread by looking at how far observations are from their mean.
    The computational formula for the standard deviation is

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

  - Interquartile Range (IQR): Distance between the first quartile ($\mathbf{Q_1}$) and the third quartile ($\mathbf{Q_3}$). **IQR = $\mathbf{Q_3 - Q_1}$**
    $\mathbf{Q_1}$ – 25% of the observations are less than Q_1 and 75% are greater than Q_1.
    $\mathbf{Q_3}$ – 75% of the observations are less than Q_3 and 25% are greater than Q_3.

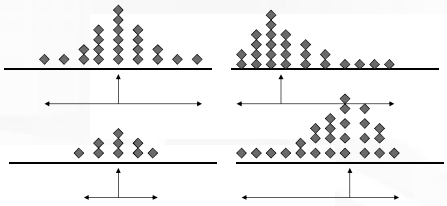
Pattern of a Distribution "SOCS"

- ***Outlier/Unusual Feature***
  - An individual value that falls outside the overall pattern.
  - Identifying an outlier is a matter of judgment. Look for points that are clearly apart from the body of the data, not just the most extreme observations in a distribution.
  - You should search for an explanation for any outlier.
  - Sometimes outliers points to errors made in recording data.
  - In other cases, the outlying observation may be caused by equipment failure or other unusual circumstances.

Rule of Thumb

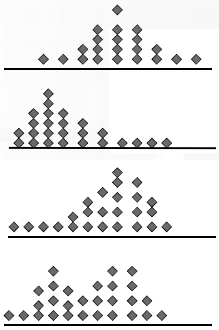
1.5 ò IQR

Interpreting Graphs: *Location and Spread*



- **Where is the data centered on the horizontal axis, and how does it spread out from the center?**

Interpreting Graphs: *Shapes*



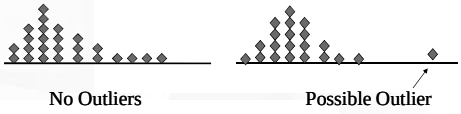
***Mound shaped* and *symmetric* (mirror images)**

***Skewed right*: a few unusually large measurements**

***Skewed left*: a few unusually small measurements**

***Bimodal*: two local peaks**

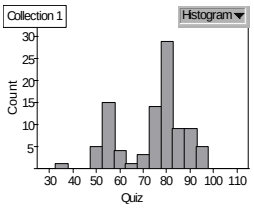
Interpreting Graphs: *Outliers*



- *Are there any strange or unusual measurements that stand out in the data set?*

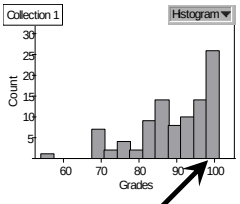
Comparing Distributions

- **Shape:** The shape is bimodal, and around each mode the shape is roughly symmetric.
- **Outlier/Unusual features:** There is a gap in the lower 40's, with a possible outlier in the mid 30's.
- **Center:** This distribution of quiz scores appears to have two modes, one at around 55, and another at around 80.
- **Spread:** The spread is from the mid-30's to the mid-90's.



Comparing Distributions

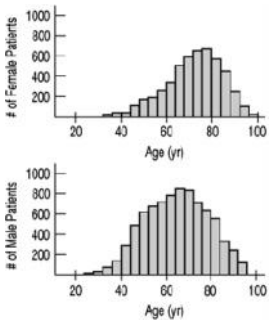
- **Shape:** The shape is unimodal and skewed to the left (to the lower grades)
- **Outlier/Unusual features:** There is a gap from the upper 50's to the upper 60's, with a possible outlier in the mid 50's.
- **Center:** This distribution of grades has a single mode at around 100.
- **Spread:** The spread is from the mid-50's to about 100.



this does NOT mean that someone had a grade of above 100. ***(more likely, a lot of 98's and/or 99's)***

Comparing Distributions

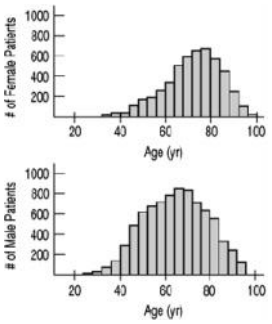
Compare the following distributions of ages for female and male heart attack patients.



Comparing Distributions

Be sure to use language of comparison.

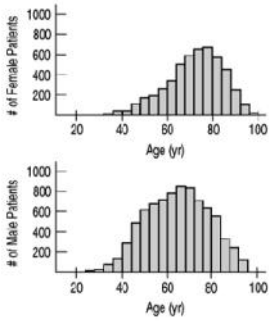
- **Center:** This distribution of ages for females has a higher center (at around 78) than the distribution for male patients (around 62).
- **Shape:** Both distributions are unimodal. The distribution for males is nearly symmetric, while the distribution for females is slightly skewed to the lower ages.



Comparing Distributions

- **Spread:** Both distributions have similar spreads: females from around 30 – 100, and males from about 24 – 96. Overall, the distribution for female ages is slightly higher than that for male ages.
- (There are no **outliers or unusual features**)

- **YOU MUST USE COMPLETE SENTENCES!!!**



Modified Boxplot

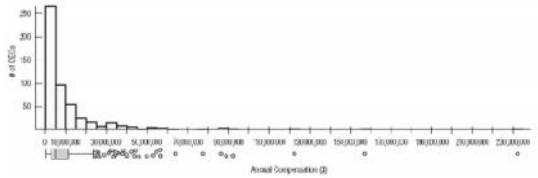
- A boxplot in which the outliers are indicated.
- Extend the whiskers from the box to the smallest and largest values that are *within* the inner fences.
- Any values that are outside the inner fences should be drawn as individual dots. These dots represent **outliers**.

Example:

Draw a modified boxplot of the sample

9, 13, 39, 40, 42, 46, 49, 54, 55, 60, 84.

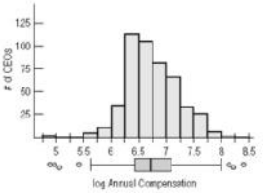
*Re-expressing/Transforming Skewed Data to Improve Symmetry



- 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?

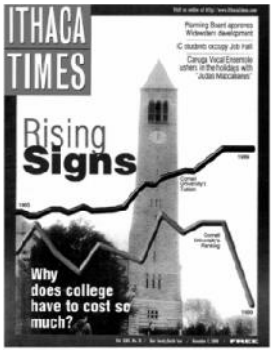
*Re-expressing/Transforming Skewed Data to Improve Symmetry (cont.)

- 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 (previous slide) to the transformed data (right):



What Can Go Wrong? (cont.)

- Avoid inconsistent scales, either within the display or when comparing two displays.
- Label clearly so a reader knows what the plot displays.
  - Good intentions, bad plot:



What Can Go Wrong? (cont.)

- Beware of outliers
- Be careful when comparing groups that have very different spreads.
  - Consider these side-by-side boxplots of cotinine levels:
  - Re-express . . .