**STCC201 Introductory Statistics**

- Instructor: Rick Gumina
- Phone: 491-5330
- Office: Room 104 Statistics
- E-mail: gumina@stat.colostate.edu

**Text and Supplies**

- Text: *Statistical Ideas and Methods* By Jessica Utts and Robert Heckard
- Audience participation clicker
- Calculator with stat functions. I recommend the TI30X IIS. However, any calculator that will handle 2 variable statistics will do.

**Grading**

- Worksheets (≤ 5) 5%
- Quizzes (≤ 5) 15%
- Assignments (≤ 15) 15%
- Labs 15%
- Exam 1 10%
- Exam 2 10%
- Exam 3 10%
- Final exam 15%
- Extra Credit 5%

**Grading Scale**

<table>
<thead>
<tr>
<th>Score Range</th>
<th>Grade</th>
<th>Score Range</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>93 - 100</td>
<td>A</td>
<td>67 - 69</td>
<td>D+</td>
</tr>
<tr>
<td>90 - 92</td>
<td>A-</td>
<td>63 - 66</td>
<td>D</td>
</tr>
<tr>
<td>87 - 89</td>
<td>B+</td>
<td>60 - 62</td>
<td>D-</td>
</tr>
<tr>
<td>83 - 86</td>
<td>B</td>
<td>&lt; 60</td>
<td>F</td>
</tr>
<tr>
<td>80 - 82</td>
<td>B-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>77 - 79</td>
<td>C+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>73 - 76</td>
<td>C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70 - 72</td>
<td>C-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Grading Policies

- **Extra credit opportunities.** You can receive up to 5% in additional credit for participation in lecture/lab as well as completing weekly assignments. Please see the “STCC201 Policies” handout for more specific information.
- **There are no curves.** The methods we use have been shown to be effective in helping learners become proficient. We’ll do our jobs. Please be sure you do yours!

What Is Statistics

- Statistics is the science of **data**
- Any task involving the collecting, processing, displaying, and interpreting **data** is part of the discipline of statistics.
- These principles and procedures help people make data based decisions when “**uncertainty**” is present.

What Is Data?

- Think of data as the result/outcome of an **experiment**.
- These responses may be numeric or labels.
- The items upon which this information is collected are called **experimental units**.

Inference

- In many studies an investigator wants to make conclusions about a large group (**population**) of items based on information collected from a smaller group (**sample**).
  - EG how much, on average, does a typical student spend on textbooks for the semester’s classes?
- These conclusions are called **inferences** and the investigator is said to “**draw inference**.”
Populations

- A population of items is the entire collection of items of interest to the investigator in his or her study.

- Sometimes all units in the population of interest do not have a chance to contribute a measurement.
  - EG Drug studies

- For this reason there are two categories of populations

Target Population

- The target population is the group of items about which the investigator ultimately wants to draw inference.

Sampled Population

- The sampled population is the group of items from which the sample is actually drawn.

- Ideally, these two groups would be the same but they don’t have to be.
  - EG drug studies

Census Data

- If measurements are obtained from every unit in the target population, then you have a census.

- In typical investigations we don’t get to observe every unit.
Sample

- A **sample** is a subset of the entire population that is captured in a study. The **variable(s)** of interest is measured on each item of the sample and **statistics** are computed based on the “values” of these variables.

Simple Random Sample

- A **Simple Random Sample (SRS)** is a sample of the population where every unit has an equal opportunity to be selected.
  - Think of drawing names from a hat
- Legitimate inferences can be drawn using data obtained via a SRS.

Self Selected Sample

- A **self selected sample** is also known as a **volunteer sample**.
  - Think of responding to one of those weird surveys in *Cosmo* or *Maxim*
- The data obtained from these types of samples are useless for drawing inference because it lacks **randomness**.

Sampling Error

- **Sampling error** arises from random fluctuations in the sample units being measured.
- E.G. Say the true average summer income for undergrads is $3500. A SRS of 50 undergrads might result in an average of $3100. A different SRS of 50 might yield a sample average of $3800.
Non-Sampling Error

- **A non-sampling error** occurs when data are incorrectly collected, recorded, or analyzed.

- Selecting a **non-representative** sample, instrument **bias**, and data input errors are examples of non-sampling error.
  - Volunteer samples are always non-representative

Bias

- **Bias** can occur if the sample isn’t representative of the population.

- **E.G.** If I want to estimate the true average age of the student body at CSU and I select a “convenience sample” of my students in STCC201

Bias

- Bias can also occur if the measurement instrument is **systematically** inaccurate

- **E.G.** A scale reports a weight that is always 1 pound more than a person’s true weight

Observational Study

- In observational studies variable values are observed and recorded. No attempt is made to modify (treat) the units being observed

- **E.G.** A survey of a hospital’s death records where the variable of interest might be “Cause of Death”
Experiment

- In an experiment some treatment is applied to the experimental unit. The investigator then proceeds to observe the effect this \( T_x \) has on the variable of interest.

- A psychologist might measure a person’s recall before and after some type of stimulus - say music.

Cross-Sectional Study (Observational)

- In a cross-sectional study, the variable(s) of interest is observed at one point in time.
- It’s a “snapshot” of the population at the time of measurement.
- E.G. Surveying voters’ attitudes about a particular candidate prior to an election.

Retrospective Study (Observational)

- In a retrospective (case control) study data are gathered from the past by reviewing old records, face-to-face interviews, etc.

- E.G. A veterinary researcher obtains information about a certain variety of cancer affecting golden retrievers by investigating clinic records from the past 5 years.

Prospective Study

Observational or Experiment

- In a prospective study, data are collected from groups who are followed forward in time.
- Prospective studies are also known as longitudinal or cohort studies.

- E.G. Following a group of Gulf War I veterans for the next 15 years to assess the prevalence of Gulf War Syndrome.
Confounding

• Variables are said to be **confounded** when the investigator can’t separate the effects of one from the other.

• E.G. A nutritionist sets up a diet plan to help a student lose weight and, at the same time, the student begins an exercise program.

Parameter

• A **parameter** is a numeric characteristic pertaining to a population. For example, the average GPA of all students attending CSU is a parameter of the population of CSU students.

**Most often parameters cannot be determined because the entire population cannot be investigated.**

Statistic

• A **statistic** is a numeric characteristic pertaining to a **sample**. For example, the average GPA of a sample of students attending CSU is a statistic.

• Generally, we’ll use sample statistics to make **inferences** about a **population**.

Uncertainty

• **Uncertainty** arises whenever a sample result is used to infer (guess) something about the entire population. This is why it is so important that the sample truly represents the entire population.

• Representation can only be achieved from randomization.
Variables

- **Variables** are the specific items of interest in a study.
- Think of them as the “information that is collected on an experimental unit”

The Lurking Variable (Confounding variable)

- A **confounding variable** is a variable that helps explain the data but is not accounted for in the study.
- For example: Did you know that the rate of sexual assault is directly correlated to the number of ice cream cones bought?

Can you think of another variable that might be related to ice cream sales and sexual assault?

Variable (Data) Types

- Variables can be either qualitative or quantitative.
- **Quantitative**: Numeric - height, weight, number of customers, blood alcohol level
- **Qualitative**: Names or categories - eye color type of car, political affiliation....

Variable Types

- Quantitative variables come in two flavors:
- **Quantitative Continuous**: Numeric variables where it makes sense to measure in fractions of units. 135.5 lb, 67.3 in, 45.05 seconds
- **Quantitative Discrete**: Numeric variables where only integer responses make sense. 5 family members, 3 pets, 4 televisions
Variable Levels

- **Nominal**: Variables (data) measured at the nominal level are non-numeric nature
- They are qualitative – Also known as categorical
- Responses that are names, labels, or categories are nominal data
- Examples: Religion, gender, major, color

Variable Levels

- **Ordinal**: Variables (data) measured at the ordinal level are also qualitative. However, they can be arranged in some order
- Examples: Letter grades, rankings, and survey responses on a Likert Scale

Data Graphics

- Pie charts, Bar charts, Multiple bar charts and Pareto charts

Pie Charts

- Pie charts can be used to summarize one qualitative variable
- Pie slices represent the proportion of observations in a class
- Sometimes frequency results are also included
Two pie charts

- Often multiple pie charts will be presented to illustrate differences between populations.
- Although you can see the differences in attitude about the death penalty for these two populations. The multiple bar chart is better for comparisons.

Bar charts

- Bar charts can be used wherever pie charts are used.
- They are the graphic of choice when there are a large number of categories.
- Number of categories is 6 or more.

Bar charts for two populations

- It’s much easier to draw conclusions about a group of populations from multiple bar charts than from a group of pie charts.
- This is because a of the bars can be drawn on one graph.

Multiple Bar charts

- Bar charts can also be used to display a comparison between two variables where both are categorical.
The Pareto Chart

- The Pareto chart is a bar chart with the categories ordered from highest frequency to lowest frequency.
- These are used to focus attention on the categories that contribute the greatest proportion to the total response.

The Pareto Chart

See the market share by ski area example overheads

The Histogram

- The histogram is used to illustrate the distribution of quantitative (numeric) data.
- The distribution of a quantitative variable is the pattern of variation that the data set exhibits.
- The histogram allows us to discern the distribution’s shape.
- It also allows us to detect any anomalous data values.
Distribution Shape

• When there are lots of low values and just a few high values the distribution is said to be **skewed to the right**.
• Another term is **positively skewed**

![Histogram of income](image1)

Distribution Shape

• When there are lots of high values and just a few low values the distribution is said to be **skewed to the left**.
• Another term is **negatively skewed**

![Histogram of home price](image2)

Distribution Shape

• When the two halves of the histogram look approximately like mirror images the distribution is said to be **symmetric**.

![Histogram of exam results](image3)

**Channel Setting for the ResponseCard® RF**

1. Press and release the "GO" button.
2. While the light is flashing red and green, enter **15** for the channel setting.
3. After the second digit is entered, the light should change to a solid green for a few seconds to indicate success.
Using your clicker: During lecture, your teacher poses a question within a PowerPoint slide in front of the classroom.

Clicker: Use your clicker to respond with your own answer. Keypad displays solid green for 3 seconds when response is received as verification for a clear, visual confirmation that your response has been received and recorded.

Dotplots

- Dotplots are similar to histograms in their use and appearance.
- They are used to display the distribution of a quantitative variable.
- The shape of the distribution as well as the presence of any possible outliers is easily discerned from the dotplot.

Dotplots

- An advantage of the dotplot over the histogram is that all individual values are visible.
- Histograms relegate all observations into classes and, thus, their individual values are lost.
- A disadvantage is that it’s difficult to compare results from multiple data sets.

Ages and annual salaries for the CEOs of the 60 top ranked small companies in America in 1993. 

Source: Forbes, Nov. 8, 1993, "America's Best Small Companies,"
Boxplots

- Boxplots are similar to histograms in their use.
- They are used to display the distribution of a quantitative variable.
- The shape of the distribution as well as the presence of any possible outliers is easily discerned from the boxplot.
- Boxplots are also great for making multiple comparisons.

Boxplot graphics

- The boxplot is sometimes called the box and whiskers plot.
- It provides a graphical representation of how numeric data is distributed.
- It also indicates the presence of outliers.

Boxplot Advantage

- One of the great things about boxplots is that they can be used to display responses from different populations easily.

Time Series

- When the order in which the data were collected is relevant, then a time series graph is used to display the data.
Time Series Data

- Time series graphics are used to display seasonal trends as well as linear trends.
- Time series models are useful in predicting events such as power consumption of a town.
- School districts use a time series model to predict enrollment from year to year. These predictions help them make staffing decisions.

Time Series Example (Trend)

A banker might be interested in the amount of money on deposit. Historical records dating to 1985 have been researched and the relevant data is tabulated here.

<table>
<thead>
<tr>
<th>Year</th>
<th>Dep</th>
<th>Year</th>
<th>Dep</th>
</tr>
</thead>
<tbody>
<tr>
<td>1985</td>
<td>2.5</td>
<td>1990</td>
<td>12.0</td>
</tr>
<tr>
<td>1986</td>
<td>4.5</td>
<td>1991</td>
<td>15.2</td>
</tr>
<tr>
<td>1987</td>
<td>5.8</td>
<td>1992</td>
<td>17.0</td>
</tr>
<tr>
<td>1988</td>
<td>7.3</td>
<td>1993</td>
<td>18.5</td>
</tr>
<tr>
<td>1989</td>
<td>9.1</td>
<td>1994</td>
<td>21.6</td>
</tr>
</tbody>
</table>

Deposits in $1 Millions

Time Series Graphs (Trend)

<table>
<thead>
<tr>
<th>Year</th>
<th>Dep</th>
<th>Year</th>
<th>Dep</th>
</tr>
</thead>
<tbody>
<tr>
<td>1985</td>
<td>2.5</td>
<td>1990</td>
<td>12.0</td>
</tr>
<tr>
<td>1986</td>
<td>4.5</td>
<td>1991</td>
<td>15.2</td>
</tr>
<tr>
<td>1987</td>
<td>5.8</td>
<td>1992</td>
<td>17.0</td>
</tr>
<tr>
<td>1988</td>
<td>7.3</td>
<td>1993</td>
<td>18.5</td>
</tr>
<tr>
<td>1989</td>
<td>9.1</td>
<td>1994</td>
<td>21.6</td>
</tr>
</tbody>
</table>

Deposits in $1 Millions

Time Series and Seasonality

- A time series graph can be used to show seasonality.
- Seasonality is the cyclic behavior of the variable of interest.
- The graph to the right is average temperature vs time.
Graphic Summaries

- Your graphics must tell a story
- They must support your data
- You want the audience to be able to look at your graphic and make the same conclusion that you came to in your analysis
- A picture is worth a thousand words........ and it’s more interesting

Data Set Characteristics

1) Location: Where is the data set “located” along the number line?

2) Spread: (variation): How spread out - disperse - are the observations?

3) Shape: What is the shape of the distribution of values in the data set?

Data Set Characteristics

4) Outliers: Are there any unusual values in the data set?

5) Time behavior: Is time an important factor in helping us understand the data?

Location Statistics
Mean, Median & Quartiles

![Histogram showing frequency distribution with location statistics M=1353, Q1=615, Q3=2350]
Sample average

- The sample average (mean) is easily computed. Simply add together all the observations and divide by the number of observations.
- The number of observations in a sample is denoted by “n”
- The sum of all the observations in a sample is denoted by $\sum x_i$

Sample Average Example

- Let’s say that a small simple random sample (SRS) of 5 electronics technicians is selected. The objective is to estimate the true average hourly wage of the entire population of electronics technicians.
- The hourly wages in the sample are: $5.75, 6.25, 6.25, 12.75, 15.00$

Sample Average Example

Hourly wages are: $5.75, 6.25, 6.25, 12.75, 15.00$

If the variable x represents the wages then $\sum x_i = 46.00$

$n = 5$

Is there any uncertainty in this estimate?

Point Estimate

- A point estimate is a numeric value derived from the sample that is used to estimate the true population parameter value.
- It is a statistic,
- The sample mean is a point estimator for the true population mean.
- In other words $\bar{x}$ estimates $\mu$
Point Estimate

- If we truly believe that the sample is accurately representative of the target population then estimates we compute from the sample should be close to the actual population parameter values.
- **This is the underlying premise of all statistical studies.**

Median

- The median is the center, or middle value in a set of “n” observations
- It is the value such that 50% of the observations lie above and 50% of the values lie below
- It does not have to be an actual observation

Steps for computing the median

1) Order the data set
2) Compute the rank of the median using the following formula:
   \[ \text{Rank} = \frac{n + 1}{2} \]
3) If “Rank” is an integer value go right to it in the sorted data set. Otherwise compute the average of the two surrounding observations
Computing the Median

- The data set to the right is already ordered. There are 19 observations.
- Rank of the median is: \( \frac{19 + 1}{2} = 10 \)
- Go to the 10'th observation. This is the sample median. 
  \( M = 96 \)

Computing the Median

- The data set to the right is already ranked. There are 20 observations.
- Rank of the median is: \( \frac{20 + 1}{2} = 10.5 \)
- Go to the 10'th and 11'th observations and find their average. This is the sample median 
  \( M = \frac{96 + 105}{2} = 100.5 \)

Location Statistics
Quartiles

- The median breaks the data set into two halves
- Quartiles break the data set into 4 quarters
- Data set must be ranked
- Formula for finding the rank of the 1’st quartile is:
  Lower Quartile: \( Q_1 = \frac{n + 1}{4} \)

Computing Quartiles

- Rank of \( Q_1 = \frac{20 + 1}{4} = 5.25 \)
- This means take the average of the 5’th and 6’th data points. 
  \( Q_1 = \frac{73 + 78}{2} = 75.5 \)

- \( Q_1 \) separates the lower 25% from the upper 75% of the data.
Computing Quartiles

• To get the rank of the 3’rd quartile you multiply the rank for Q1 by 3 so
• Rank of Q3 = 3(5.25) = 15.75
• To compute Q3 find the average of the 15’th and 16’th observations
• Q3 = (117 + 121)/2 = 119
• Q3 separates the lower 75% from the top 25% of the data.

Location Statistics

Extremes

• Any data “real life” data set will have extreme values.
• These extreme values are the minimum and the maximum.

The 5-number summary

• The 5-number summary is a group of summary statistics that can be computed for any data set.
• This group consists of the: minimum, maximum, Q1, median, and Q3
• These are all measures of location

Boxplots and the 5-number summary

• Boxplots are a good way to graphically incorporate the 5 values in a 5-number summary
Outliers

- Outliers are data points that lie outside (or far away from) the majority of the data.
- These observations should raise a flag of concern for the investigator because they are unexpected compared to the other data points.

Outlier Processing

- The outlier might be an improperly entered data value

  Go back to the data base and re-enter the point. Then reanalyze the data

Outlier Processing

- The outlier might be an improperly reported by the respondent. This might be true in the case of survey data where the question is, “What is your weight?” and the respondent reports 55 kg instead of 121 pounds.

  In this case the data entry would be correct but the original data point is in error. If possible go back to the source (respondent) and verify the accuracy of the record.

Outlier Processing

- The outlier might be a legitimate value (accurate and properly entered) but arising from an experimental unit outside the target population. This might be true in the case of survey data where the question is, “What is your weight?” and the respondent reports 72 pounds. If the investigator is interested in characterizing “normal” healthy respondents and the record comes from someone with and eating disorder then the usefulness of the information is in question.
Outlier Processing

- When the data point comes from an experimental unit that does not belong to the target population it may **bias** the conclusions.
- It’s best to just throw the data point away.

Outlier Processing

- The data point may, actually, come from a representative of the target population. It’s just an unusual observation.
- There are “normal”, healthy adults that weigh over 300 pounds or less than 100 pounds.
- In this case the data value must remain in the data set.
- If it skews the results too much then the investigator might consider collecting more observations.

Dispersion (Spread)

- Information about location is not enough to adequately summarize a data set.
- Dispersion information is also required.
- Recall the cardinal rule of statistical inference.

The Range

- The range is the easiest measure of dispersion to compute.
- It is, simply, the difference between the maximum value and the minimum value.
- You can interpret this measure as the total width of your data set.
Interquartile range

- Another measure of dispersion is the **interquartile range**
- This is also known as the **IQR**
- \( IQR = Q_3 - Q_1 \)
- In our example the IQR = 119 - 75.5 = 43.5

Variance

- The mean is an estimate of location. Specifically, it’s a measure of the center of the data.
- Never report simply a measure of location.
- Always provide a report of some measure of dispersion.
- The variance is another measure of dispersion

Sum of Squared Deviations

- To compute a variance we need a statistic called the **sum of squared deviations**
- This is often abbreviated as SS
- To get a squared deviation take an observation and subtract the mean. Then square the result
- Do this for all observations and add the results. This is SS

<table>
<thead>
<tr>
<th>( x )</th>
<th>( \bar{x} )</th>
<th>((x - \bar{x})^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.75</td>
<td>9.20</td>
<td>11.90</td>
</tr>
<tr>
<td>6.25</td>
<td>9.20</td>
<td>8.70</td>
</tr>
<tr>
<td>6.25</td>
<td>9.20</td>
<td>8.70</td>
</tr>
<tr>
<td>12.75</td>
<td>9.20</td>
<td>12.60</td>
</tr>
<tr>
<td>15.00</td>
<td>9.20</td>
<td>33.64</td>
</tr>
<tr>
<td>46.00</td>
<td>9.20</td>
<td>75.54</td>
</tr>
<tr>
<td>( \Sigma x )</td>
<td>( \bar{x} )</td>
<td>SS</td>
</tr>
</tbody>
</table>

Sample Variance

- The sample variance is denoted by the symbol \( s^2 \)
- \( s^2 = SS/(n-1) \)
- \( s^2 = 75.54/4 = 18.89 \)
- The interpretation of a variance is, “the variance is the average squared distance that a group of ‘n’ points is from the mean of the group.”
- This is not a very intuitive concept
Sample Variance

- Sample Variance = \( s^2 = \frac{SS}{n-1} \)

- Note the \((n-1)\) term in the denominator. This is called the **degrees of freedom** AKA **df**

Sample Standard Deviation

- The sample standard deviation is, simply the square root of the sample variance.

- It is given the symbol \( s \)

\[ s = \sqrt{s^2} \]

- In our example: \( s = \sqrt{18.89} = 4.35 \)

Interpret the Standard Deviation

- The standard deviation can be thought of as an average distance that a group of points lies from the group mean

- Large standard deviations mean that there is a lot of scatter in the data

- Large standard deviations mean lots of uncertainty

Comparing distributions

Say you’ve measured two different quantitative variables on each experimental unit. How do you compare the dispersion of one set of values with that of the other?

Just looking at raw standard deviation values for each data set is not appropriate because the units are not the same.
Comparing distributions

A study was conducted where the investigator collected age and income information from 35 randomly selected adults.

\[ x_{age} = 35 \text{ yr} \quad s_{age} = 7.5 \text{ yr} \]
\[ x_{inc} = \$32,000 \quad s_{inc} = \$3,250 \]

Which of these distributions is the most disperse?

The Coefficient of Variation

- Since you can’t directly compare years to dollars we need to convert the data into a proportion - expressed as a percent.
- This new statistic is called the coefficient of variation

The Coefficient of Variation

- The CV is the ratio of the sample standard deviation to the sample mean

\[ CV = \frac{s}{x} (100 \%) \]

\[ CV_{age} = \frac{7.5}{35}(100) = 21.4\% \]
\[ CV_{inc} = \frac{3250}{32000}(100) = 10.2\% \]
Interpreting the CV

- The smaller the CV the closer your data values are together.
- As the CV decreases so does the dispersion.
- In our example the income data is more tightly clustered about the mean than the age data…even though 3250 is bigger than 7.5

Z-scores

- z-scores are sometimes referred to as standardized scores
- z-scores are a measure of position. That is, they quantify the position of a data value relative to the mean of all data values in the group.
- \[ z = \frac{x - \mu}{\sigma} \]

Z-scores

- Note that the z-score formula uses the population mean and standard deviation values.
- If you’re lucky enough to have census data then there is no problem.
- If you have sample data we just replace the actual population values with their point estimates.

Z-score Example

- What is the z-score for the $15.00/hr observation?
- Recall that \( \bar{x} = 9.20 \) and \( s = 4.35 \)
- \[ z_{(15.00)} = \frac{15.00 - 9.20}{4.35} = 1.33 \]

\[
\begin{array}{ccc}
 x & \bar{x} & (x-\bar{x})^2 \\
 5.75 & 9.20 & 11.90 \\
 6.25 & 9.20 & 8.70 \\
 6.25 & 9.20 & 8.70 \\
 12.75 & 9.20 & 12.60 \\
 15.00 & 9.20 & 33.64 \\
 46.00 & 75.54 & \\
 \hline
 \Sigma x & \bar{x} & SS \\
\end{array}
\]
Interpretation of Z-scores
• The sign of the z-score indicates whether the data point lies above or below the mean.
• Positive z-scores mean that the observation is above the mean and negative z-scores mean that the point is less than the mean.
• The magnitude of the z-score indicates how far away the data point is from the mean in terms of standard deviations.
• Most z-scores lie between +3 and -3. Observations that give z-scores outside this range are often considered outliers.

Characteristics of Z-scores
• The z-score does not have any units.
• It simply expresses the distance of an observation from the mean in terms of standard deviation units.
• For example: An observation with a z-score of 1.5 lies 1.5 standard deviations above the mean.

Shapes of Distributions
• You don’t need a histogram to determine the shape of a distribution. In fact, all you need are the values for the mean and the median of your data set.

Shapes of Distributions
• What is the shape of this distribution to the right?
• Note that the mean is 86
• And the median is 92
Shapes of Distributions

• What is the shape of this distribution to the right?
• Note that the mean is 2.6
• And the median is 0.6

Mean, Median, & Shape

• If the mean is greater than the median then the distribution is skewed to the right
• If the mean is less than the median then the distribution is skewed to the left
• If the mean and median are equal then the distribution is said to be symmetric

Bell Shaped Distributions

There are lots of distribution shapes out there in the real world.

Of special interest to us is the family of **Bell Shaped** distributions
Bell Shaped Distributions

- With only information about the mean and the standard deviation we can deduce a lot about BSDs.

Bell Shaped Distributions

- These deductions are based on the Empirical “Rule”

Given that a response distribution is bell shaped $\approx 68\%$ of the observations in the data set fall b/w $\pm 1\sigma$ of the mean

Given that a response distribution is bell shaped $\approx 95\%$ of the observations in the data set fall b/w $\pm 2\sigma$ of the mean
Bell Shaped Distributions

Given that a response distribution is bell shaped $\approx 99.7\%$ of the observations in the data set fall between $\pm 3\sigma$ of the mean.

Empirical Rule for BSDs

Example: Interactive Slide – Empirical Rule 1