

Part II – Exploring Data, Chapter 1 from our textbook

Read through Chapter 1 in the *Practice of Statistics* textbook, Page 1-68.

You are responsible for understanding all of the following key words:

- **Individuals** - are the objects described by a set of data. Individuals may be people, but they may also be animals or things.
 - Ex. – A college has a database including information about every enrolled student. The students would be the individuals.
- **Variable** – a variable is any characteristic of an individual. A variable can take different values for different individuals.
 - Ex. – the data includes information such as age, major, gender, GPA, etc. Each of which is considered a variable.
- Two ways to classify variables:
 - **Categorical (aka Qualitative)** – places an individual into one of several groups or categories
 - Ex. – Gender and major simply place individuals into categories
 - Ex. – credit card numbers are also considered categorical. The first 4 numbers identify the type of card and the entire combination of numbers creates your own unique card. Although there are numbers, it wouldn't make sense to perform any mathematical operations with those numbers.
 - **Quantitative (aka Numerical)** – takes numerical values for which arithmetic operations such as adding and averaging make sense.
 - Ex. - A persons height, weight, etc.
 - Ex. – The gas mileage on various models of cars

****CAUTION****

What seems like a quantitative variable may not be: If the number isn't mathematically meaningful, but still has some type of order to it, it would be considered "other".

- Ex. –jersey numbers; not numerically meaningful. Someone with #20 isn't twice as good as #10 and your jersey number does not necessarily group you into one part of the team or another.

What seems like a quantitative variable may not be: If the word is simply naming the object, but doesn't provide the category at which it goes to, then it's considered "other."

- Ex. – A list of animals at the zoo (monkey, giraffe, snake, etc.) wouldn't be categorical unless they were put into categories based on the animal kingdom (mammals, reptiles, etc.).
- Ex. – Student last names; the last name by itself isn't categorical, but using that last name to create a new variable, such as which guidance counselor you are assigned to, is a categorical variable.

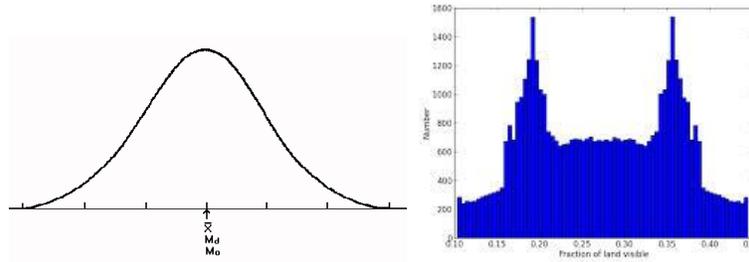
- **Distribution** – what values the variable takes and how often it takes these values
 - When describing a distribution, include the shape, center, spread, and outliers
- **Count (frequency)** – A tally of the number of individuals that fall into each category
- **Frequency Table** – A display of the counts for each category
- **Percent (relative frequency)** – A percentage of individuals that fall into each category
- **Relative Frequency Table** – A display of data that shows the percents in each category
- **Roundoff error** - the difference between the calculated approximation of a number and its exact mathematical value.
 - Happens in relative frequency tables if the percentages are rounded

- **Two-Way Tables** – Display the relationships between two categorical variables
- **Marginal Distribution** – the “total row” and “total column” are the marginal distributions.
- **Conditional Distribution** – the relative frequencies in the body of the table
- **Association** – we say that there is an association between two variables if knowing the value of one variable helps predict the value of the other.

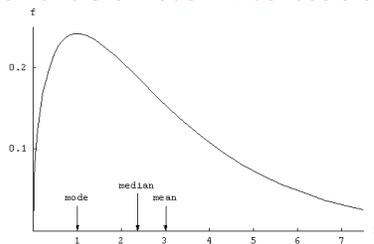
****Shape, Center, and Spread describe the overall pattern of the distribution****

Shape

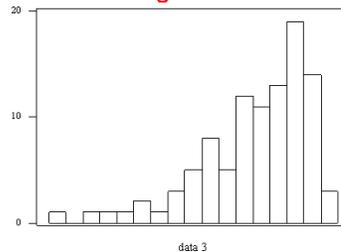
- **Symmetric** – If the right and left sides are approximately mirror images of each other



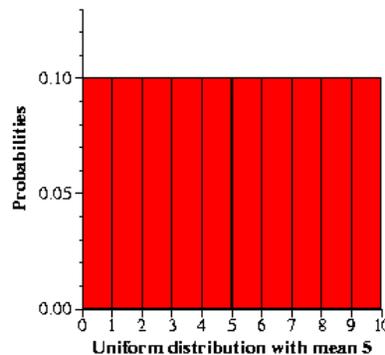
- **skewed to the right** (higher values) – when there are extreme observations that are higher than most of the data. The mean will be greater than the median and the mode will be less than the median.



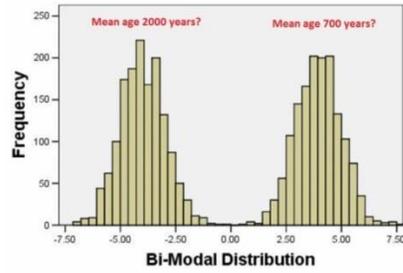
- **skewed to the left** (lower values) – when there are extreme observations that are lower than most of the data. The mean will be less than the median and the mode will be greater than the median.



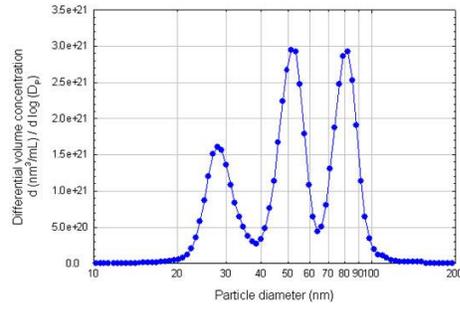
- **Rectangular (unimodal)**



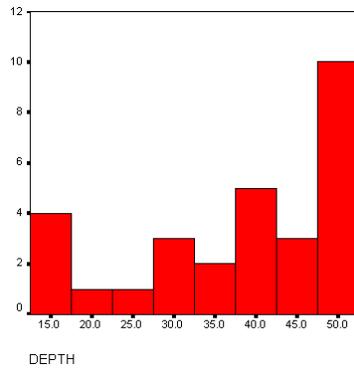
- bimodal



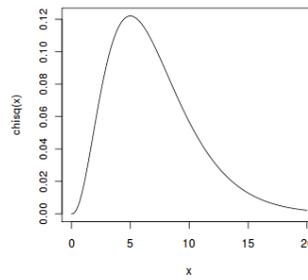
- tri-modal



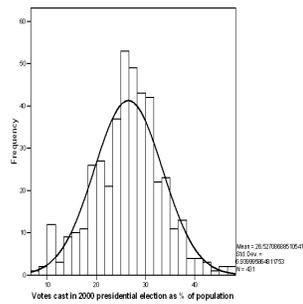
- Multimodal



- bell-shaped – not necessarily normal



- “approximately normal”



Center

- **Mean** – average (non-resistant) – cannot resist the influence of extreme observations
- **Median** – middle value – is resistant to extreme observations
- **Mode** – value or values that occur most often – is resistant to extreme observations – can have no mode, 1 mode, or more than 1 mode

Spread

- **Range** – the difference between the highest and lowest data values
- **Quartiles** – Q_1 – First quartile (median of lower half), Q_2 – Middle quartile (median of entire data set), and Q_3 – Third quartile (median of upper half)
- **interquartile range (IQR)** – The middle 50% of the data set. $IQR = Q_3 - Q_1$
- **Five-number summary** – a summary of a data set that includes the minimum value, Q_1 , Q_2 , Q_3 , and the maximum value
- **variance** – the average of the squared differences from the mean (s^2).

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

- **standard deviation (and properties on page 62)** – the square root of the variance. Measures spread by looking at how far the observations are from their mean.

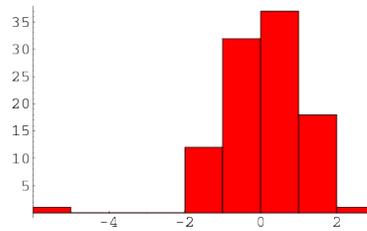
$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

Properties of standard deviation:

- s measures spread about the mean and should only be used when the mean is chosen as the measure of center (aka – only use for a normal or approx. normal distribution)
- $s = 0$ when there is no spread. This happens only when all observations have the same value. Otherwise, $s > 0$. As the observations become more spread out about their mean, s gets larger.
- s , like the mean \bar{x} , is not resistant. Strong skewness or a few outliers can make s very large.
- **Resistant measure** – When the measure cannot be affected by outliers.
 - Mean – non-resistant
 - Median – resistant
 - Mode – resistant
 - IQR - resistant
 - Standard deviation – non-resistant
- **Degrees of freedom** – The number of values in a study that are “free” to vary. Represented as $n - 1$.
 - Ex. – Suppose you need to take 10 classes and only 10 are offered. You are able to choose the order in which you take the first 9, but come time for the 10th course, you don’t have a choice. Therefore you have 9 degrees of freedom.
 - For standard deviation of a sample (s) we use $n - 1$ because the sum of deviations is always 0 and therefore the last deviation can be found once we know all other $n - 1$ deviations.

Outlier - an individual observation that falls outside the overall pattern of the graph

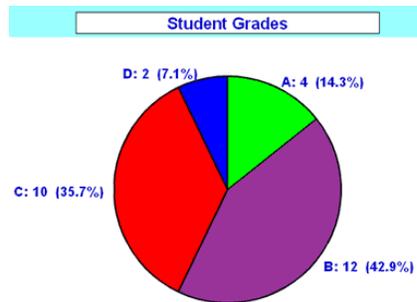
- what an outlier looks like –



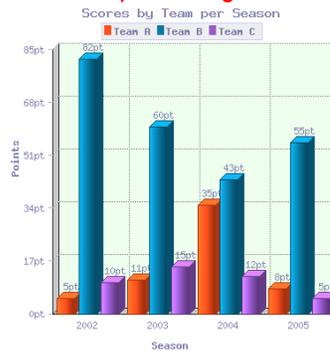
- how to calculate if something actually is one – If the value is smaller than $Q_1 - (1.5 \times IQR)$ or larger than $Q_3 + (1.5 \times IQR)$

Know how to read and construct each of the following: (Also, know when each is appropriate to use.)

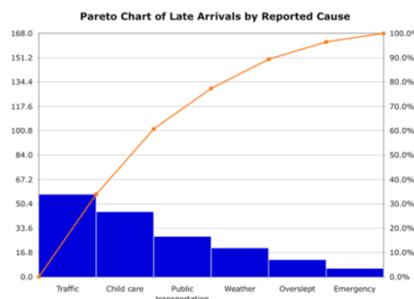
- **Pie Chart** – Use for categorical data. Pie charts are best used when trying to compare parts of a whole.



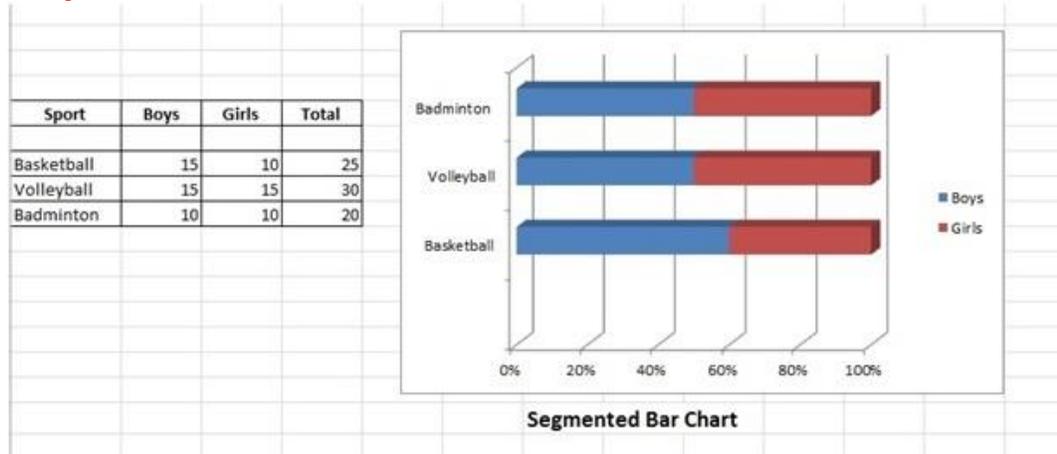
- **Bar Chart** – Used for categorical data. Used to compare things between different groups.



- **Pareto Chart** - contains both bars and a line graph, where individual values are represented in descending order by bars, and the cumulative total is represented by the line. The purpose of the Pareto chart is to highlight the most important among a (typically large) set of factors.



- **Segmented Bar Graph** – A segmented Bar chart is one kind of stacked bar chart, but each bar will show 100% of the discrete value.
 - For example, there are a total of 40 students in your classroom. Out of them, 25 students like Basketball, 30 students like Volleyball, and 20 students like Badminton. There are 25 boys and 15 girls in the class. The data along the vertical side of the box represents sports while the horizontal represents a certain percentage for each sport. Each bar will show the preference of each sport according to the number of boys and girls and the bars will be separated by stacked order, representing one group for the boys and the other for the girls.

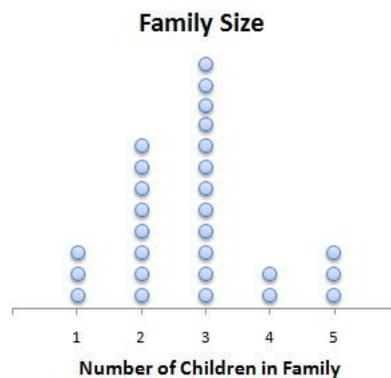


- **Frequency table** –

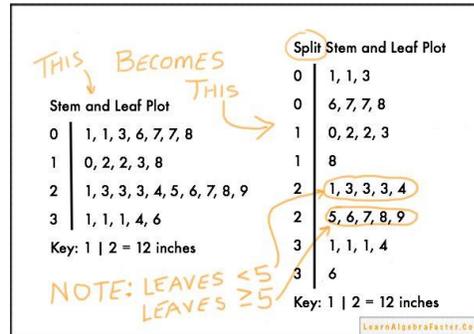
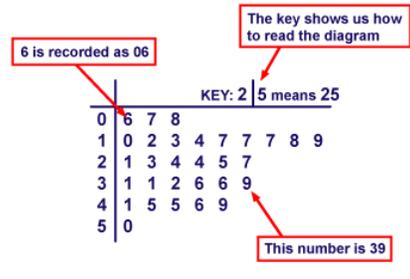
1. In the first column write down all of the data values in ascending order (or create bins)
2. In the second column tally the number of times those values occur
3. In the third column write the total number of times each value occurred

Mark	Tally	Frequency
4		2
5		2
6		4
7		5
8		4
9		2
10		1

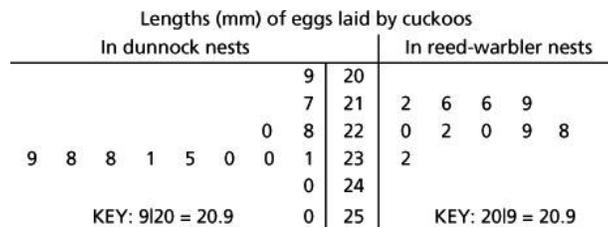
- **Dot Plot (aka Number Line Plot)** – One of the simplest ways to display quantitative data.



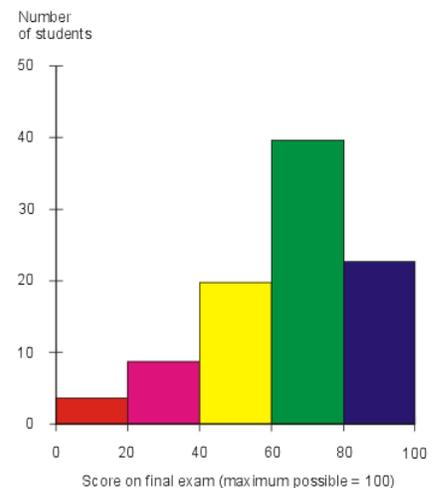
- **Stem and Leaf Plot (aka Stem Plot)** – Typically used when the values of a variable are too spread out to make a reasonable dotplot.
 - The stem consists of all but the rightmost digit (this goes in the left column)
 - The leaf is the final digit (these go in ascending order in the right column)
 - Be sure to have an equal number of possible leaf digits assigned if you split stems
 - Five stems is a good minimum



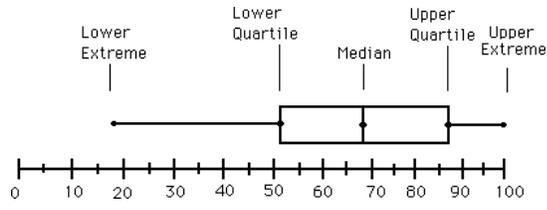
- **Back to Back Stem and Leaf Plot** – A method of constructing two sets of leaves, in the stem-and-leaf plots involving two sets of data, hanging on both sides of the same stem.



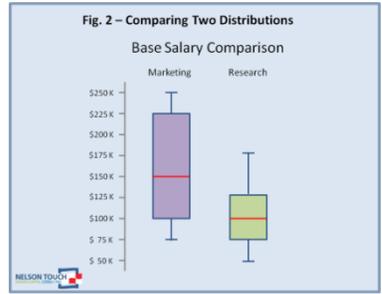
- **Histogram** (and how it differs from a bar chart) – The most common graph of the distribution of one quantitative variable. The graph is clearer if nearby values are grouped together. This is different than a bar chart because it displays quantitative data versus qualitative data. The bars are also touching in a histogram whereas they are separated in a bar chart.
 - Five classes is a good minimum
 - To group the values be sure to divide the range of the data into classes of equal width
 - Beware of letting a computer or graphing calculator choose the classes
 - The start of the class should include the number marked
 - If you are not starting at zero, put a “break” in the graph (see figure 1.7 on page 20)



- **Box and Whisker Plot (aka Boxplot)** – A graphical display based on the five-number summary. Best used for side-by-side comparison.



- **Side by Side Box and Whisker Plot**



- **Modified Boxplot (shows outliers)** - The modified box-and-whisker plot will not plot outliers as part of the box-and-whisker. The outliers are plotted as individual points beyond the whisker in an attempt to give a more accurate picture of the dispersion of the data. THE WHISKER ONLY EXTENDS TO THE NEXT HIGHEST OR LOWEST VALUE THAT IS NOT AN OUTLIER...IF THE NEXT HIGHEST VALUE THAT ISN'T AN OUTLIER HAPPENS TO ALSO BE Q3 THEN YOU WON'T HAVE A WHISKER ON THE RIGHT SIDE.

