Summary: Unit 2 & 3 "Distributions for Quantitative Data"

Topics covered in Module 2:

- ✓ How to calculate the Mean, Median, IQR
- ✓ Shapes of Histograms, Dotplots, Boxplots
- ✓ Know the difference between categorical and quantitative variables
- ✓ Analyze changes to center and spread
- ✓ Make and Analyze boxplots
- ✓ Analyze histograms
- ✓ How to calculate whether something is an outlier or not
- ✓ Write a summary analysis paragraph describing a boxplot, dotplot, or histogram with all the elements presented in Mod 2 (Shape, Center, Spread, and Outliers)

A **categorical (or qualitative) variable** names categories and answers questions about how cases fall into those categories.

- Categorical examples: sex, race, ethnicity

A **quantitative variable** is naturally measured as a number for which meaningful arithmetic operations make sense.

- Quantitative examples: income ($), height (inches), weight (pounds), score

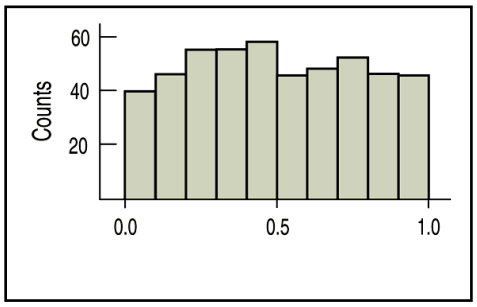
When we describe data, we say it's the **Distribution** of the data

When we describe patterns, we talk about overall pattern: **Shape, Center, Spread, and Outliers**

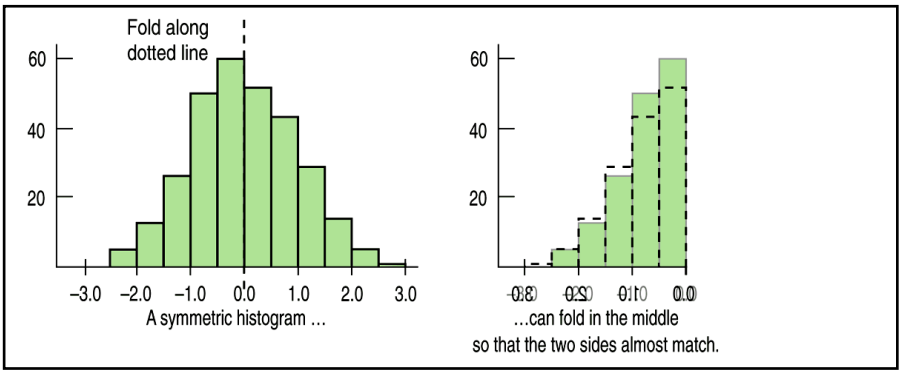
Things that don't really fit into the pattern are called **Outliers**.

SHAPE:

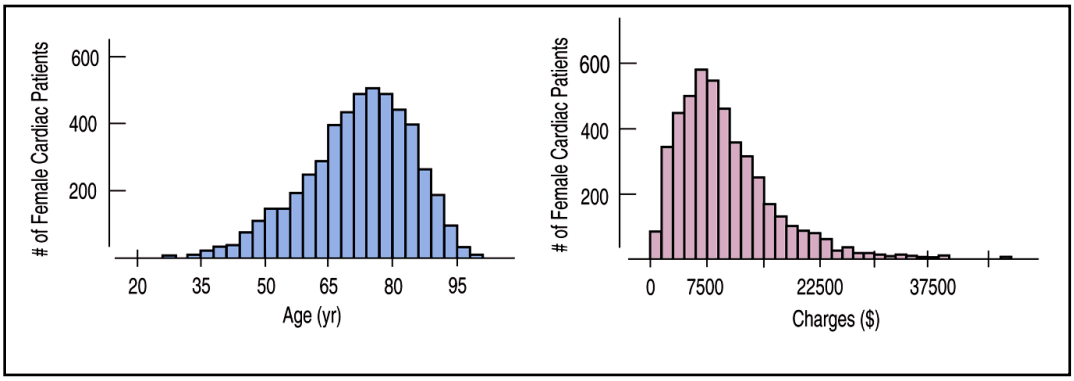
Uniform



Symmetric (Mean and Standard Deviation)



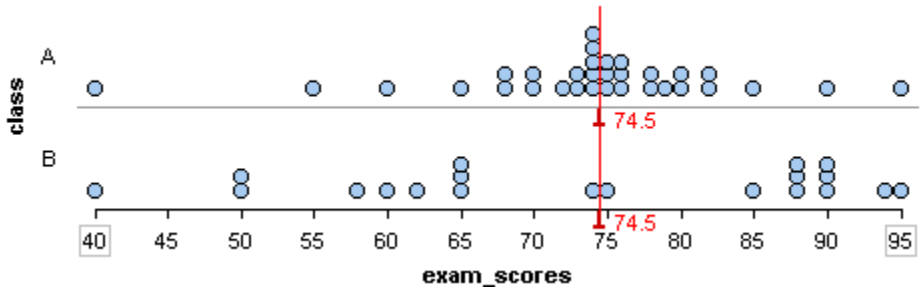
Skewed to the left/Skewed to the right (Median and IQR)



CENTER: The center of a distribution is a measurement of what a "Typical Value" would be.

We have two measurements for the center of a distribution. The **mean** and the **median**. We use the mean as a measurement for center when a distribution is symmetric/bell-shaped/normal. We use the median when the distribution is skewed.

SPREAD: The spread of a distribution is a measure of how spread out the data is.



Another word for spread is **variability.**

Initial measures of spread:

Overall Range = largest value – smallest value

To find **Typical Range**, we use the largest and smallest value of where most dots cluster.

Detailed measures of spread: We use the standard deviation and IQR (Interquartile Range) as measures of spread after we cover them. Depending on the shape, we pair center and spread accordingly.

- Symmetric – Mean and Standard Deviation
- Skewed – Median and IQR

Outliers: A data point is considered an outlier if it deviates from the overall pattern of the data.

Symmetric – A data point is considered an outlier if it is more than 2 standard deviations from the mean.

Skewed- A data point is considered an outlier if it is:

- Greater than: Q3 + 1.5(IQR)
- Less than: Q1 – 1.5(IQR)

The Mean: The mean is the balancing point in the sense that it balances the sum of the distances above and below. $\overline{x} = \frac{\sum x}{n}$

The Median: To find the median we look for the middle number. Make sure the data is in order. The median is the balance point where there is an equal number of data points on each side.

SD (Standard Deviation): know how to interpret meaning in context

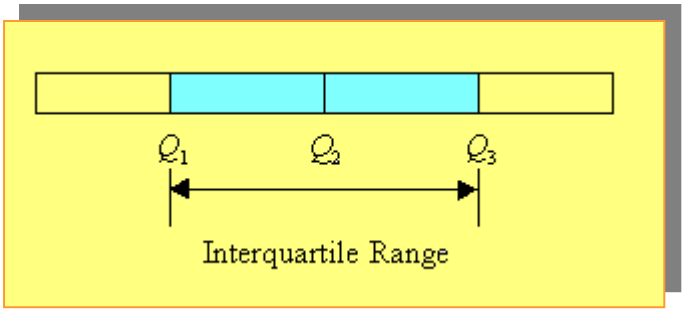
What are Quartiles? Quartile marks divide the data into____4____ subgroups with the same number of individuals (data points) in each subgroup. Each quartile contains ___25___% of the data points.

The interquartile range (IQR) measures the spread of ____the middle 50% of the data_____

To find the IQR:

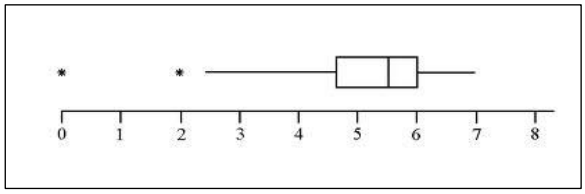
Step1: Order the data set from smallest to largest.
Step 2: Find the median for the ordered set. Denote by Q_2.
Step 3: Find the median for the first 50% of the ordered set. The median found in Step 2 is not included in this portion of the data. Denote by Q_1.
Step 4: Find the median for the second 50% of the ordered set. The median found in Step 2 is not included in this portion of the data. Denote by Q_3.

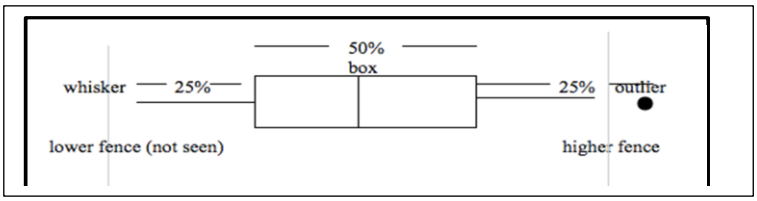
$$IQR = Q_3 - Q_1$$



Steps to create a boxplot:

- ***Step1:*** Draw a box from Q1 to Q3
- ***Step2***: Draw a line through the box at the median
- ***Step3:*** Extend a line (tail) from Q1 to the smallest value that is NOT an outlier and another line (tail) from Q3 to the largest value that is NOT an outlier.
- ***Step4:*** Indicate outliers with asterisks (*)





Writing Analysis Guide:

Shape	**Center** (This is the typical or average value that represents the data.)	**Spread** (Use the typical range.)	**Outliers** (These are the unusual data points. Remember

			that outliers are not necessarily bad.)
Left Skewed Right Skewed Uniform	Median	IQR (IQR = Q3-Q1. Approximately the middle 50% of the data in our sample is from Q1 to Q3. Also, approximately the middle 50% of the data is within the IQR of each other)	Less Than: Q1 – 1.5(IQR) More Than: Q3 + 1.5(IQR)
Symmetric	Mean	Standard Deviation (Approximately 68% or most of the data is from ($\bar{x}$-1SD) to ($\bar{x}$+1SD))	Less Than: $\bar{x}$-2SD More Than: $\bar{x}$+2SD

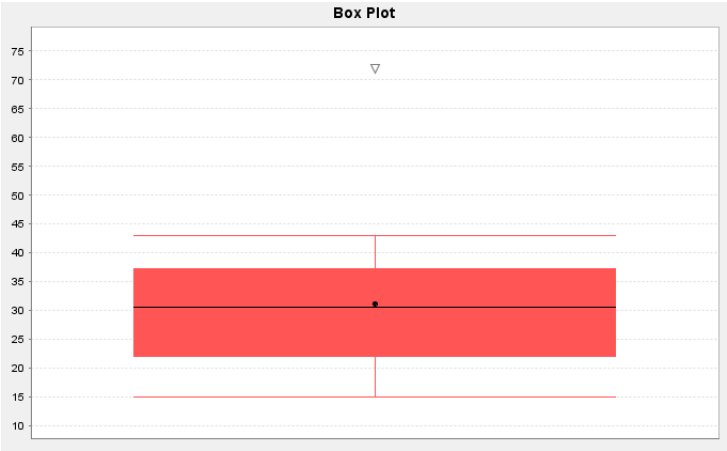
Structure:

- Introduction Sentence
- Shape
- Best measure of Center (choose one: median/mean and why) and its value including the units
- Interpret the center in the context of the data
- Best measure of Spread (choose one: IQR/SD and why) and its value including the units
- Interpret the spread in the context of the data
- Outliers (What are the outliers. Do you consider then to be invalid/valid and why?)
- Conclusion / Wrap up / Answer the posed question(s)

Notes: Be sure to use the appropriate values with the appropriate units.

Here are two examples of what an essay question would look like:

1. **Weekly TV-Viewing Times** The A. C. Nielsen Company publishes data on the TV-viewing habits of Americans by various characteristics in *Nielsen Report on Television*. A sample of 20 people yielded the weekly viewing times, in hours, is shown with the following boxplot:



Write an essay about weekly viewing times, using the boxplot. Be sure to discuss the shape, center, spread, and which measures of center and spread we should be using. Interpret the center and spread in context. How could this information be useful in the real world and to whom?

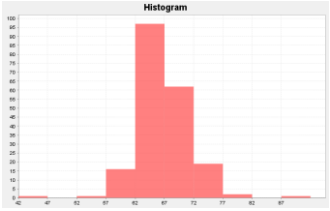
Descriptive Statistics

Mean	Standard Deviation
31.2	12.564

Q1	Median	Q3	IQR
22.0	30.5	37.25	15.25

Min	Max	Range
15.0	72.0	57.0

2. Consider the following histogram of the distribution of heights, in inches, of a random sample of 200 math pre statistic students taken in the spring semester of 2013 at College of the Canyons. Write a summary analysis of the data.



Descriptive Statistics

Variable	Mean	Standard Deviation
C1 Height in inches	65.93	4.69

Q1	Median	Q3	IQR
63.0	65.0	68.0	5.0

Min	Max	Range	
42.0	90.0	48.0	