



**Christopher Columbus High School**  
**Summer Assignment Mathematics**  
Students entering AP Statistics  
**Due Date: Tuesday, Aug 18, 2026**



**Brief Description:**

- Download worksheet to notability
- Watch Videos (1-6): Complete the assignment while watching the required videos and filling in the skeleton notes
- Sign up: [Wayground.com](https://www.wayground.com) CODE: **S856112**
- After the videos and notes complete the required multiple choice questions in Wayground.

**Resources needed:** TI-Nspire Calculator, Wayground, Notability

**Approximate time commitment during the summer:**

~6 hours (make sure to thoroughly understand the concepts as you will be tested on the material)

**Questions over the summer:** Please contact me: [maria.romero@columbushs.com](mailto:maria.romero@columbushs.com)

**Grading and mastery testing**

- The summer assignment is due on the first day of school Tuesday, Aug. 18, 2026 on Oncampus folder by **7:00am**.
- Late assignments will be accepted but will be penalized based on the number of days the assignment is late.
- The first two days of class students may ask questions and teacher will review summer assignment.
- At the end of the first week students will be expected to take a mastery test on Unit 1 which will count as their first test of the quarter.

**Required Content**

- 1: Exploring One-Variable Data
  - 1.1 Analyzing Categorical Data
    - a) Frequency tables and bar charts
    - b) Comparing distributions using bar charts
  - 1.2 Displaying Quantitative Data with Graphs
    - a) Constructing and comparing center and spread
    - b) Constructing and comparing clusters and gaps
    - c) Constructing and comparing outliers and unusual features
    - d) Constructing and comparing shape
  - 1.3 Describing Quantitative Data with Numbers
    - a. Constructing distributions with graphs.
      - i. Stem plots, dot plots, histograms, boxplots, ogives
      - ii. Graphing calculator instruction/use
    - b. Describing distributions with numbers.
      - i. Center: mean, median.
      - ii. Spread: variance, standard deviation, range, IQR
      - iii. Quartiles, percentiles
      - iv. Graphing calculator instruction/use
    - c. Interpreting distributions
      - i. SOCS: shape, center, spread & outliers.
      - ii. 4 C's: clear, concise, complete & in context

**REMEMBER:**

The CCHS Honor Code applies to this packet.  
DO NOT COPY ANSWERS FROM YOUR CLASSMATES.

**Note:**

Remember, this is an AP Course! Do not expect this to be an “easy course”. Although it may not seem as difficult computationally as calculus, it requires a great deal of outside reading and homework, and it requires a thorough understanding of many abstract concepts. This is as much a writing course as it is a math course! Explaining in complete sentences is required on this assignment and throughout the course. You cannot just write down numbers and be done, you must use numbers in **context** – what they mean to that particular problem using appropriate units. Enjoy your summer and looking forward to meeting you in August! Smiles ~ Mrs. Romero

# ASSIGNMENT

## Video1: Categorical Data: Tables and Bar Graphs

Watch video 1: <https://www.youtube.com/watch?v=qh5EyPWZtI0> Time: 16:59 min

### What are Statistics?

- Statistics is \_\_\_\_\_
- Data are observations that have been measured, recorded, collected, analyzed, and reported for use
- Data contains information about a group of \_\_\_\_\_
  - Individuals are objects described by a set of data
  - These might be people, animals, or inanimate objects
- The information is organized using \_\_\_\_\_

### Types of Variables

<ul style="list-style-type: none"><li>• Data that places individuals into specific groups</li><li>• Some examples include:<ul style="list-style-type: none"><li>○</li><li>○</li><li>○</li><li>○</li><li>○</li></ul></li></ul>	<ul style="list-style-type: none"><li>• Data that takes on numerical values, where performing arithmetic operations makes sense.</li><li>• These can be broken down into two types:<table border="1"><tr><td></td></tr><tr><td><ul style="list-style-type: none"><li>• This is numerical data where whole numbers make sense to describe the data.</li><li>• Think of variables that you would count, and decimals don't make sense to describe them</li><li>• Some examples include:<ul style="list-style-type: none"><li>○</li><li>○</li><li>○</li></ul></li></ul></td></tr><tr><td></td></tr><tr><td><ul style="list-style-type: none"><li>• This is numerical data where decimals would make sense to describe the data.</li><li>• Think of variables where measuring makes sense and a certain number of decimal places or intervals are used to report the values.</li><li>• Some examples include:<ul style="list-style-type: none"><li>○</li><li>○</li><li>○</li></ul></li></ul></td></tr></table></li></ul>		<ul style="list-style-type: none"><li>• This is numerical data where whole numbers make sense to describe the data.</li><li>• Think of variables that you would count, and decimals don't make sense to describe them</li><li>• Some examples include:<ul style="list-style-type: none"><li>○</li><li>○</li><li>○</li></ul></li></ul>		<ul style="list-style-type: none"><li>• This is numerical data where decimals would make sense to describe the data.</li><li>• Think of variables where measuring makes sense and a certain number of decimal places or intervals are used to report the values.</li><li>• Some examples include:<ul style="list-style-type: none"><li>○</li><li>○</li><li>○</li></ul></li></ul>
<ul style="list-style-type: none"><li>• This is numerical data where whole numbers make sense to describe the data.</li><li>• Think of variables that you would count, and decimals don't make sense to describe them</li><li>• Some examples include:<ul style="list-style-type: none"><li>○</li><li>○</li><li>○</li></ul></li></ul>					
<ul style="list-style-type: none"><li>• This is numerical data where decimals would make sense to describe the data.</li><li>• Think of variables where measuring makes sense and a certain number of decimal places or intervals are used to report the values.</li><li>• Some examples include:<ul style="list-style-type: none"><li>○</li><li>○</li><li>○</li></ul></li></ul>					

### Things to remember:

- Just because it is a number DOES NOT automatically make it quantitative. Can you think of an example of a number that is categorical?
  -
- The difference between discrete and continuous isn't always clear: For instance, is age continuous or discrete?
  - You use whole numbers to describe it, but it can be measured with a decimal.
- If it is not obvious if it is discrete or continuous, like age, it depends on how you use the data.
  - For age, we rarely say "I'm 16.4 years old", so we treat age like a \_\_\_\_\_ variable.

## Frequency Table

- A frequency table is a way to display how many individuals fall into each category of a categorical variable
- Frequency is also referred to as \_\_\_\_\_
- Below is the “raw data” of a sample of teacher’s favorite pizza toppings
- The term “raw data” is used when referring to data that is \_\_\_\_\_

Pepperoni	Canadian Bacon	Mushrooms	Black Olives	Pepperoni
Sausage	Pepperoni	Onions	Sausage	Canadian Bacon
Sausage	Pepperoni	Black Olives	Pepperoni	Canadian Bacon
Pepperoni	Black Olives	Mushrooms	Pineapple	Sausage
Sausage	Onions	Pepperoni	Mushrooms	Black Olives

- While the raw data gives us the information, it is difficult to find counts, patterns, or observe what the data is trying to tell us
- We will organize this information in a \_\_\_\_\_ table

<i>Category</i>	Pepperoni	Sausage	Canadian Bacon	Black Olives	Onions	Mushrooms	Pineapple
<i>Frequency</i>							

- A frequency table shows the counts in each of the categories.
- We can also make a \_\_\_\_\_ table (or relative counts) that displays the percentage in each category instead
- To find the relative frequency, you take the frequency in each category and divide it by the total. The resulting decimal is written as a percentage.
- The total number of teachers in our sample is \_\_\_\_\_

<i>Category</i>	Pepperoni	Sausage	Canadian Bacon	Black Olives	Onions	Mushrooms	Pineapple
<i>Relative Frequency</i>							

## Bar Graph

- A bar graph is used to visually show the distribution of data
- Distribution is a term used to describe the visual display of data that shows the variable and how often the data takes each value.
- To create a bar graph that displays data from a one-way frequency table or a relative frequency table:
  1. Draw your axes and label the x-axis with your categories and your y-axis with your counts or percentages.
  2. Draw a bar for each category, starting at the x-axis and going up to the corresponding value on the y-axis.
  3. Make sure the bars DO NOT TOUCH.
  4. The order of the categories does not matter, but you can put them in order of their frequency if you prefer.

Example: Create a bar graph showing the distribution of favorite pizza toppings among the teachers.

Category	Frequency (Count)
Pepperoni	7
Sausage	5
Canadian Bacon	3
Black Olives	4
Onions	2
Mushrooms	3
Pineapple	1

## Video 2: Histograms and Ogives

Watch Video 2: <https://youtu.be/8kuPjfuzJrs> Time: 36:00 min

Watch Video: How to create Bar Graph and Pie Chart on TI-Nspire

<https://www.youtube.com/watch?v=vXA9i9O-CrE>

## Displaying Quantitative Data

- Quantitative variables can be described with numerical data, where operations like averaging make sense
- Quantitative variables can be broken down into two types:
  - Discrete – \_\_\_\_\_
  - Continuous – \_\_\_\_\_
- We can use a frequency table to organize both of these data types and a \_\_\_\_\_ to display them

## Discrete Quantitative Data

- A frequency table for discrete data looks very similar to the frequency tables for categorical data.
- A histogram for this type of data is create similarly to a bar graph, with some key differences:
  - The bars will touch, to communicate that this is quantitative data
  - The x-axis must go in order from \_\_\_\_\_
  - The x-axis will be labeled with values \_\_\_\_\_ the bars
  - These values communicate what values are included in the bar, up to but not including, the upper boundary

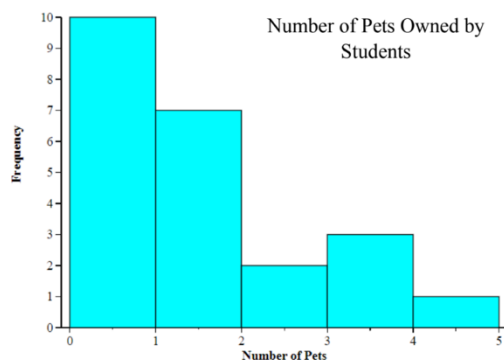
Example: A survey was given in a class. They were asked “how many pets do you own” and here are the results as a frequency table and as a histogram.

Number of Pets	Frequency
0	10
1	7
2	2
3	3
4	1

Notice how the first bar, with a frequency of 10, goes from 0 to 1.

What data is this corresponding to?

The lower boundary is always included, but the upper boundary is not.



**Example:** For your high school football team, you looked at how many points were scored in each of the games they played in the past 5 years. Here is a list of the data:

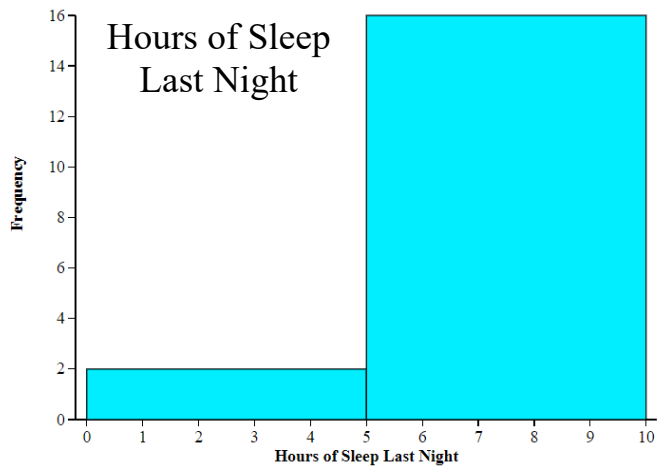
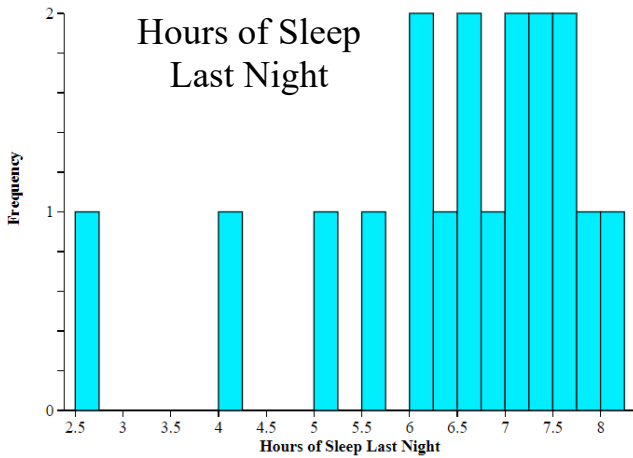
24, 27, 17, 27, 22, 10, 24, 31, 30, 34, 13, 28, 20, 24, 20, 27, 26, 7, 27, 14, 27, 22, 24, 21, 13, 17, 10, 31, 28, 19, 24, 26, 41, 20, 3, 35, 30, 27, 25, 24, 24, 27, 13, 17, 31, 36, 30, 45, 31, 24, 37, 43, 42, 37, 30, 10, 35, 28, 22, 34, 24, 34, 41, 30, 24, 40, 26, 35, 10, 21, 27, 27, 34, 23, 42, 31, 24, 11, 37, 8, 31, 20, 21, 23, 13

Fill in the frequency table below, and create a histogram with the data.

Scores	Frequency
0 to 9	
10 to 19	
20 to 29	
30 to 39	
40 to 49	

### Continuous Quantitative Data

- Constructing a frequency table for continuous data looks a little different because we need to make sure we don't have too many bars or too few:



- These graphs look silly because we do not get a good, clear visual of our data.
- To avoid having too many or too few, we try to get 5 to 7 bars with each data set

To determine the intervals needed to get the appropriate number of bars, we follow these steps:

- Calculate the range of your data set: \_\_\_\_\_
- Calculate the bar (called "class") size:
  - Divide range by desired number of bars (usually between 5 and 7)
  - If you get a decimal, round it!
- Set up each class (Note: the lowest value is \_\_\_\_\_, the highest is not)
  - 1<sup>st</sup> Class: start with the minimum value, add class size to get the next value on your x-axis
  - 2<sup>nd</sup> Class: start with that value, add the class size again to get the next value
  - Repeat for all your classes
  - Last Class: The last class does not have to end at your maximum value, but it should go over it!

**Example:** When asked how many hours of sleep students got last night, the data was reported as follows:

2.5, 4.0, 5.2, 5.5, 6.0, 6.1, 6.3, 6.5, 6.7, 6.8, 7.0, 7.2, 7.3, 7.4, 7.5, 7.6, 7.8, 8.0

Create a frequency table and a histogram to display the data. Use a class size of 6.

## Percentiles

- A percentile can be represented as a whole number percent (\_\_\_\_) or as a rank (\_\_\_\_).
- The number represents the approximate percent of data in a set that is below a single data point.

Calculating percentiles is not an exact science and there is not a universal agreement on a single procedure for calculating them. Here is an example of how we will calculate the percentiles of a small data set:



**Example:** In Disney World®, the Seven Dwarf’s Mine train is known for having a long wait time. On trill-data.com, they have the average daily wait time for every day at the park last year. The data below is the average wait time for the ride for every day in September 2024.

What is the percentile of the wait time of 66 minutes?

83	98	94	93	92	93	76	68	73	82
75	69	69	66	46	64	65	60	49	52
51	48	59	49	50	45	48	50	43	42

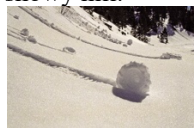
## Ogives

- Ogives (pronounced o-jives) or \_\_\_\_\_ graphs, display the percentiles of an interval of variables.
- These graphs allow us to visually see percentiles and answer questions involving them.
- To create an ogive, you first must make a relative frequency table of the quantitative data

Hours of Sleep a Night	Frequency	Relative Frequency
0 – 2	1	
2 – 4	4	
4 – 6	4	
6 – 8	7	
8 – 10	2	

“Cumulative relative frequency” ACCUMULATES the percentages as you increase your x variable.

Think of it like a snowball rolling down a snowy hill.



Hours of Sleep a Night	Cumulative Relative Frequency
0 – 2	
2 – 4	
4 – 6	
6 – 8	
8 – 10	

## Creating an Ogive

- The first interval is 0 – 2, with a cumulative relative frequency of \_\_\_\_%. Since 0 is the starting interval, that starts at 0 and \_\_\_\_% is graphed at 2.
- The next interval is 2 – 4, with a cumulative relative frequency of \_\_\_\_%. At 2, \_\_\_\_% is graphed, so at 4, the \_\_\_\_% will be graphed.
- This pattern continues until \_\_\_\_% is graphed at 10.

**Example:** Create an ogive of the number of hours of sleep a student got last night.

1) How many hours is the 30<sup>th</sup> percentile? What does this mean in the context of the problem?

2) George got 7 hours of sleep last night. What is his approximate percentile and what does it mean in the context of the problem?

### Video 3: Stem Plots and Dotplots

Watch video 2: <https://youtu.be/1R8WvlaVItA> Time: 25.16

## Stemplots

- Stemplots (sometimes called a stem and leaf plot) are an alternate way of illustrating data
- They are similar to a histogram but the individual data values are still able to be seen
- An example stemplot is shown, and it represents the ages of a random selection of teachers the building
- The “stem” is to the left of the line and represents the age digit in the tens place
- The “leaf” is to the right of the line and represents the age digit in the ones place
- Write down the 18 ages of the teachers:

Ages of Teachers

2		2	5	9	9				
3		0	4	5	7	8	8		
4		3	4	4	8				
5		1	7						
6		5	8						

Ages

KEY: 6|8 = 68

Things to note:

- The stem can be one digit or multiple digits
- Leaves are lined up by their stem and go in order from smallest to largest
- Repeat values are listed
- Between each stem, the leaves are aligned together

**Example:** The follow data represents the lowest temperature recorded each day in October of 2024 for a town in the US:

25, 26, 29, 32, 33, 36, 36, 36, 37, 38, 38, 38, 39, 40, 40, 41, 42, 42, 44, 45, 48, 51, 52, 54, 55, 55, 58, 58, 59, 61, 63

Create a stemplot of the data.

## Back-to-Back Stemplots

- You can create a back-to-back stemplot when you can separate the quantitative data into two categories.
- The stem will go in the middle of the graph, with each set of leaves branching out.
- The data should still go in order from smallest to largest, with the smaller leaves being closer to the stem.

**Example:** The table below shows the pulses of a randomly selected group of students.

Create a back-to-back stemplot of the data.

Males	46	69	61	61	68	72	55			
Females	63	88	50	68	66	76	65	70	52	90

## Split Stemplot

- Split Stemplots are helpful if you have a lot of values in a single stem.
- If you have many values, it can be helpful to “split the stem” to be able to visualize the data better.
- The first stem digit is for the leaf digits 0 to 4
- The second stem digit is for the leaf digits 5 to 9

*Regular Stemplot* of the heights of US presidents

```

16 | 3 8 8
17 | 0 0 1 3 3 3 3 3 5 7 8 8 8 9
18 | 0 0 2 2 2 3 3 3 3 3 3 5 5 8 8 8 8 9
19 | 1 2 3
    
```

Height

KEY: 19|3 = 193

*Split Stem Plot* of the heights of US presidents

## Dotplot

- A dotplot is a simple type of graph that involves plotting the data values, with dots, above the corresponding values on a number line.
- To construct a dotplot:
  1. Draw a horizontal line and label your axis with the name of your variable
  2. Scale the axis based on the values of the variable
  3. Mark a dot above the number on the horizontal axis corresponding to each data value

**Example:** Students at a high school were asked how many books they read over the summer. The data is shown in the table below. Create a dotplot to display the data.

0	2	3	8	5	2	3	4	1	2	0	1	3
7	0	1	5	1	2	1	6	1	2	4	1	0

## Choosing the Right Graph

Which graph is the best? It depends on the data! Some graphs are better suited to data sets, while some data sets can be used with any graphs! Here is a general run down of the differences.

	<i>Works Best With</i>	<i>Best Features</i>	<i>Examples</i>
<b>Histograms</b>	<ul style="list-style-type: none"> <li>• Continuous &amp; discrete data</li> <li>• Large data sets or data with a large range</li> </ul>	Shows the shape of a distribution very well	<ul style="list-style-type: none"> <li>• Heights of students</li> <li>• Daily temperatures</li> <li>• Annual rainfall amounts</li> </ul>
<b>Ogives</b>	<ul style="list-style-type: none"> <li>• Continuous data</li> <li>• Small or large data sets</li> </ul>	Easy to identify and interpret percentiles	<ul style="list-style-type: none"> <li>• Exam scores</li> <li>• Income levels</li> <li>• Age distributions</li> </ul>
<b>Stemplots</b>	<ul style="list-style-type: none"> <li>• Discrete data</li> <li>• Medium range of values; you don't want too many or too few stems</li> </ul>	Shows the shape of a distribution while maintaining the original data values	<ul style="list-style-type: none"> <li>• Test scores</li> <li>• Weights</li> </ul>
<b>Dotplots</b>	<ul style="list-style-type: none"> <li>• Discrete data</li> <li>• Small range of values</li> </ul>	Emphasizes individual data values	<ul style="list-style-type: none"> <li>• Number of books read</li> <li>• Number of goals scored</li> </ul>

### Video 4: Measures of Center and Spread

Watch video 4: <https://youtu.be/EA8ymujLEVA> Time: 27.44 min

Watch Video: How to use TI Nspire CX II to calculate mean, standard deviation, quartiles and produce a box plot <https://www.youtube.com/watch?v=FQ9EKebcp1I>

## Measures of Center

- A measure of center attempts to use a single number to summarize a set of data
- There are three measures of center in statistics are mean, median, and mode
- The word “average” is a generic term for a measure of center, but it is not a measure itself (usually when people say “average” they mean “mean”)

## Mean

- To find the mean, we add up all the values in a data set and divide by the total number of values.
- In symbols, this looks like:

Where the symbols have the following meaning:

\_\_\_\_\_ → mean of a sample (pronounced “x-bar”)

\_\_\_\_\_ → each individual value in the data set

\_\_\_\_\_ → add up (capital sigma)

\_\_\_\_\_ → total number of values in a data set

The data below represents the ages of a random sample of 24 teachers at our school.

33	25	45	31	30	26	25	29
48	36	31	29	29	49	28	30
40	39	33	38	39	24	51	32

What is the mean age of teachers in our sample?

While we use mean most often, it is not the only measure of center due to its unreliability around outliers.

## Outliers

- An outlier is a value that lies an “abnormal” distance from the other values in the data set
- We have two mathematical ways of determining if a value is an outlier, which we will discuss later
- For now, let’s look at an obvious example of an outlier:

Suppose a student took 7 exams and had these scores: 90, 92, 94, 98, 86, 88, 0

If I calculated your grade as a mean of these 7 scores, what would you get? \_\_\_\_\_

- Does this value give a good idea of how you performed on your tests?
- There is very clearly an outlier: the zero score. Sometimes, there might be a legitimate reason to eliminate an outlier from a data set. Why might that be the case with these test scores?
- Let’s say I took your reason, and eliminated the 0 from the gradebook. What would be your new mean grade?
- In general, what does an outlier do to the mean when it is in a data set?
  -
- The mean is called a \_\_\_\_\_ measure → it is strongly influenced by extreme values
- We have other measures of center that are considered to be \_\_\_\_\_ measures → they are not strongly influenced by extreme values.

## Median

- The median is a measure of center found by ordering the data from smallest to largest and then finding the middle value in that list.
- While the mean had a special symbol and formula, the median does not
- We use the symbol “Med” to label the median



Sum of the Square Deviations =

Take this value and divide it by  $n - 1$  :

- The value of 6.5 is the \_\_\_\_\_
- This represents the average squared deviation of the number of pets this group of students has
- The variance is used in some statistical methods, but for our purposes, it is not great because it does not have the same units as our original data set
- To get back to the original units, we need to take the square root of the variance

variance =  $s^2 = 6.5 \rightarrow$  \_\_\_\_\_ = \_\_\_\_\_ = \_\_\_\_\_ = \_\_\_\_\_

The standard deviation of this data set is \_\_\_\_\_ pets. This means that the number of pets typically varies from the mean, on average, by \_\_\_\_\_ pets.

We just built the formula for standard deviation:

	_____ $\rightarrow$ mean of a sample (pronounced "x-bar")
	_____ $\rightarrow$ each individual value in the data set
	_____ $\rightarrow$ add up (capital sigma)
	_____ $\rightarrow$ total number of values in a data set
	_____ $\rightarrow$ standard deviation

Important Standard Deviation Facts:

- The standard deviation measures the typical distance of the values in a distribution from the mean
- It should only be used as your measure of spread when the \_\_\_\_\_ is your chosen measure of center
- It is always \_\_\_\_\_
- It is 0 when \_\_\_\_\_
- It has the same units of measure as \_\_\_\_\_
- It is \_\_\_\_\_ to extreme values

The greater the standard deviation, the greater the \_\_\_\_\_ of the distribution

### Video 5: Boxplots and Outliers

Watch video 5: <https://youtu.be/SdsYxb9PLvg> Time: 25.55min

### The 5 Number Summary

- Percentiles measure the percent of the observations that fall below a value
- The median is the middle of a data set, where \_\_\_\_\_ of the observations fall above and \_\_\_\_\_ fall below, putting the median at the \_\_\_\_\_ percentile.

Other Important Percentiles:

- 0<sup>th</sup> percentile: \_\_\_\_\_ - lowest value in a data set.
- 25<sup>th</sup> percentile: \_\_\_\_\_ or \_\_\_\_\_ - 25% of the data is below this value.
- 50<sup>th</sup> percentile: \_\_\_\_\_ - middle value in a data set.
- 75<sup>th</sup> percentile: \_\_\_\_\_ or \_\_\_\_\_ - 75% of the data is below this value.
- 100<sup>th</sup> percentile: \_\_\_\_\_ - highest value in a data set.

- These 5 numbers make up what is called the \_\_\_\_\_
- To calculate these values by hand, place the observations in order and find the median.
- Then, find the middle value to the left and right of your median to identify your quartiles.

**Example:** Here is the data from a previous statistics class. They were asked how many hours of sleep they got before the first day of school.

2 4.5 5 5 6 6 6 6.5 7 7 7 7 7 7 7 7.5 8 8 8 8 8 8 8.5

Five Number Summary	Min	Q <sub>1</sub>	Med	Q <sub>3</sub>	Max

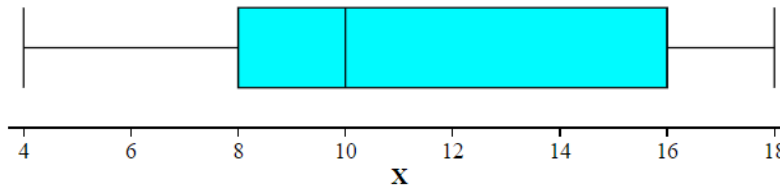
## IQR

- IQR stands for *Inter-Quartile Range* and it uses percentiles to describe the spread of the distribution
- The IQR is the range of the inter-quartiles: \_\_\_\_\_
- Just like range, this value is a SINGLE value and is not defined as an interval

**Example:** What is the IQR of the number of hours of sleep that group of AP Stat students got?

## Boxplots

- The five number summary is used to construct another graph: the boxplot



## Constructing a Boxplot

1. Set up your x-axis, and label with the correct scale.
  - a. You CANNOT only mark the values of the five-number summary. You must use a consistent scale.
2. Place points at each key value of the 5 number summary.
  - a. The Box: A box is drawn from \_\_\_\_\_ to \_\_\_\_\_
  - b. A vertical line in the box marks the \_\_\_\_\_. This is not necessarily the middle of the box, but rather where the median value is at.
  - c. The Whiskers: Lines horizontally extend from the edges of the box out to the \_\_\_\_\_ and \_\_\_\_\_ observations.
3. Connect points appropriately with boxes/whiskers.

**Example:** Here are the teacher ages from lowest to highest:

24	25	25	26	28	29	29	29
30	30	31	31	32	33	33	36
38	39	39	40	45	48	49	51

Find the 5 number summary and construct a boxplot of the data.

## Outliers

- There are two ways to determine outliers in this course
- Both of these ways attempt to determine if a point is "far enough" away from the rest of the data to be considered an outlier:
  - 1<sup>st</sup> Way: Use the IQR and Quartiles
  - 2<sup>nd</sup> Way: Use the Mean and Standard Deviation

1.5 x IQR Rule	2 Standard Deviation Rule
1. Find the 5-number summary and the IQR 2. Find lower boundary: Compute $Q1 - (1.5 * IQR)$ . Any data below that number is an outlier. 3. Find upper boundary: Compute $Q3 + (1.5 * IQR)$ . Any data above that number is an outlier.	1. Find the mean and standard deviation 2. Find the lower boundary: Compute $\bar{x} - 2s$ . Any data below that number is an outlier. 3. Find the upper boundary: Compute $\bar{x} + 2s$ . Any data above that number is an outlier.

**Example:** Here is the sleep data again:

2 4.5 5 5 6 6 6 6.5 7 7 7 7 7 7 7 7.5 8 8 8 8 8 8 8.5

Summary Statistics	Min	Q <sub>1</sub>	Med	Q <sub>3</sub>	Max	Mean	Standard Deviation
	2	6	7	8	8.5	6.64	1.5

Using each of the rules described above, determine if there are any outliers.

1.5 x IQR Rule	2 Standard Deviation Rule

## Modified Boxplot

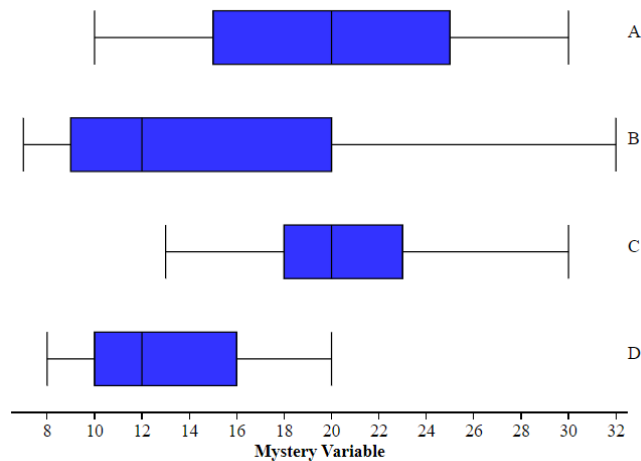
- When we have outliers in our data set, a modified boxplot helps us create a boxplot that will identify the outliers
- Any outliers identified by the  $1.5 \times \text{IQR}$  rule will be marked with a dot (or star)
- The whisker will extend out to the next observation that is within the acceptable boundary ranges (NOT the boundary ranges themselves!!)

**Example:** Create a modified boxplot for the sleep example above.

## Interpreting Data in a Boxplot

The data represented in the four boxplot measures the same variable across 4 different groups (A, B, C, and D).

Use the graphs to answer the questions below.



1) Which of the groups has the largest range?

2) Which of the groups has the largest IQR?

3) Compare the median and quartiles between A and B.

4) Compare the median and quartiles between C and D.

## Video 6: Comparing Distributions

Watch video 6: <https://youtu.be/d39VYI8hixw> Time: 23:20 min

### Shape

- When asked to describe a distribution with quantitative data, we have four items we want to comment on:
  - 
  - 
  - 
  -
- We do NOT describe categorical distributions with the above – why?
  - We cannot find the \_\_\_\_\_ of a categorical variable
  - \_\_\_\_\_ doesn't matter because we can put the data in any order we want

### Common Shapes

Once the distribution is graphed, the first thing we identify is the shape of the graph. Here are some examples of common distribution shapes you will encounter.

Unimodal	Bimodal	Uniform
Symmetric	Left Skewed	Right Skewed

## Graph Shapes

- We often use two or more of the shape descriptions above to describe what a graph looks like

Distribution	Histogram	Stemplot	Boxplot
Symmetric Uniform			
Symmetric Bimodal			
Unimodal Left Skewed			
Unimodal Right Skewed			
Multimodal  None of the main shapes above			

## Outliers

- We can comment on any \_\_\_\_\_
- Use the  $1.5 \times \text{IQR}$  rule to mathematically determine outliers
- Use the  $2 \times \text{standard deviation}$  rule to mathematically determine outliers

## Center

- When deciding what measure of center to use, it is best to consider the shape and if the distribution has extreme values (which may or may not be outliers)
- The mean is \_\_\_\_\_ to extreme values, so it is best used with symmetric, non-skewed distributions
- The median is \_\_\_\_\_ to extreme values, so it is best used with skewed distributions

In a skewed distribution (which doesn't always have outliers), the median is a good indication of center because the mean gets pulled towards the tail of the distribution.

<i>Unimodal, Symmetric</i>	<i>Left Skewed</i>	<i>Right Skewed</i>

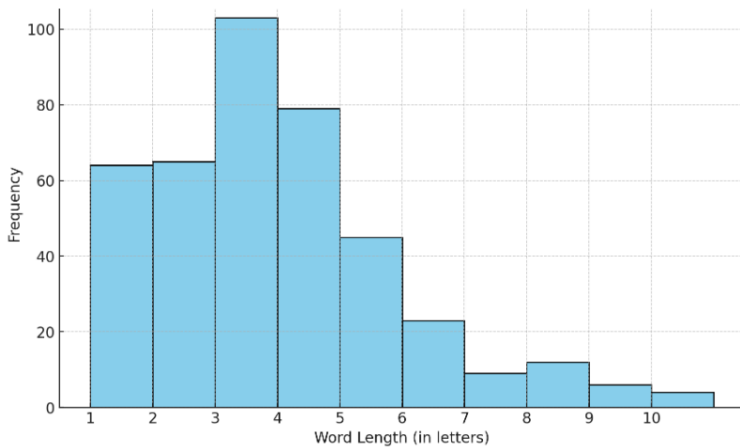
## Spread

You can describe the spread in a few different ways:

- If you have data and use the mean as the measure of center, find the \_\_\_\_\_
- If you have data and use the median as the measure of center, find the \_\_\_\_\_
- If you do not have data, report the \_\_\_\_\_ from the graph

**Example:** Describe the distribution below by commenting on the shape, center, spread, and potential outliers.

Word Length for Taylor Swift's Song, "I Can Do It with a Broken Heart"

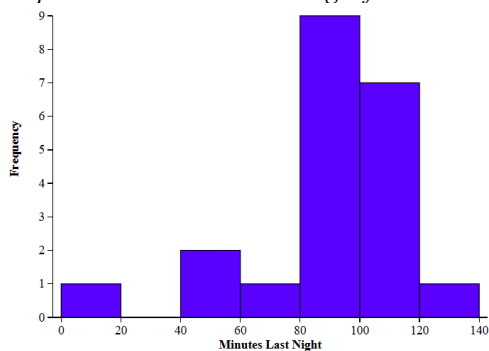


$\bar{x}$	s	min	Q <sub>1</sub>	med	Q <sub>3</sub>	max
3.5	1.94	1	2	3	4	10

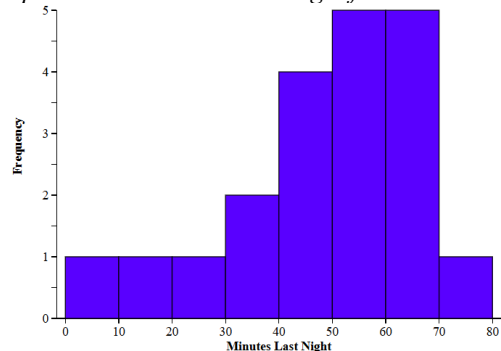
## Comparing Distributions

Do students in AP classes spend more time on homework than students in non-AP classes? A random sample of 21 "AP Students" (students enrolled in 1 or more AP classes), and 20 "non-AP students", and they were asked how long they spent on homework last night. The data is shown in the histograms below.

*Time Spent on Homework Last Night for AP Students*



*Time Spent on Homework Last Night for non-AP Students*



Summary Statistics:

n	$\bar{x}$	s	min	Q1	med	Q3	max
21	88.905	25.894	15	82.5	95	106.5	120

Summary Statistics:

n	$\bar{x}$	s	min	Q1	med	Q3	max
20	47.3	17.357	5	39	51	60	70

**Task:** Compare the two distributions.

When we compare two distributions, we want to comment on shape, outliers, center, and spread, but we want to make sure we are using comparison words and descriptions.

For Example:

- “The distribution of Group A is slightly skewed right, while Group B is roughly symmetric.”
- “Group A typically scores higher than Group B.”
- “The mean of Group A’s distribution is greater than the mean of Group B.”
- “Group B shows more variability than Group A, as seen in the larger IQR.”
- “There is an outlier in Group A that Group B does not have.”
- “The range of Group A is about the same as the range of Group B.”

Do not do...	What you should do...
The shape of the AP student homework time is left skewed with an outlier, and the shape of non-AP student homework time is left skewed with no outliers.  The mean amount of time AP students spend on homework is 88.905 minutes. The mean amount of time non-AP students spend on homework is 47.3 minutes.	Both distributions appear left skewed, while the AP student homework group appears to have a low visual outlier, the non-AP homework group does not.  The mean amount of time AP students spend on homework is 88.905 minutes, which is higher than the mean amount of time non-AP students spend on homework, which is 47.3 minutes.

**GO TO WAYGROUND TO COMPLETE THE MULTIPLE CHOICE QUESTIONS**