Unit 6: Statistic Models

Name: _____

6.1 ~ Measures of Central Tendency (Day 1)

Objective: Use statistics appropriate to the shape of the data distribution to compare center and spread of two or more data sets.

- Mean (aka Average)
  - The _____ of the data divided by the _____
- Median
  - Arrange a set of data from smallest to largest; the _____ number of the data set
    - If there is an even number of data items, then there are two middle numbers and the median is the _____ of these two numbers
      - Useful when we wish to _____
- Mode
  - A number that occurs _____ in a set of data
    - If each number occurs the same number of times, there is no mode
- Range
  - The _____ between the largest value and the smallest value in a data set

Measure	Most Useful When...
Mean	• data set has no outliers
Median	• data set has outliers • there are no big gaps in the middle of the data
Mode	• data set has many identical numbers
Range	• describing the spread of the data

Examples:

1. Consider the following data on total net income, in billions of dollars, for Starbucks Corporation for the years 2000 – 2004:

   $2.2, $2.6, $3.3, $4.1, $5.3

   What is the mean of these values?

2. A student scored the following on five tests: 78, 95, 84, 100, 82

   What is his average score?

3. Suppose five workers in a technology company manufactured the following number of computers during one day's work:

Sarah:	88	Jen:	94
Matt:	92	Mark:	91
Pat:	66		

What is the median number of computers assembled?

4. What is the median of the following set of yearly salaries?

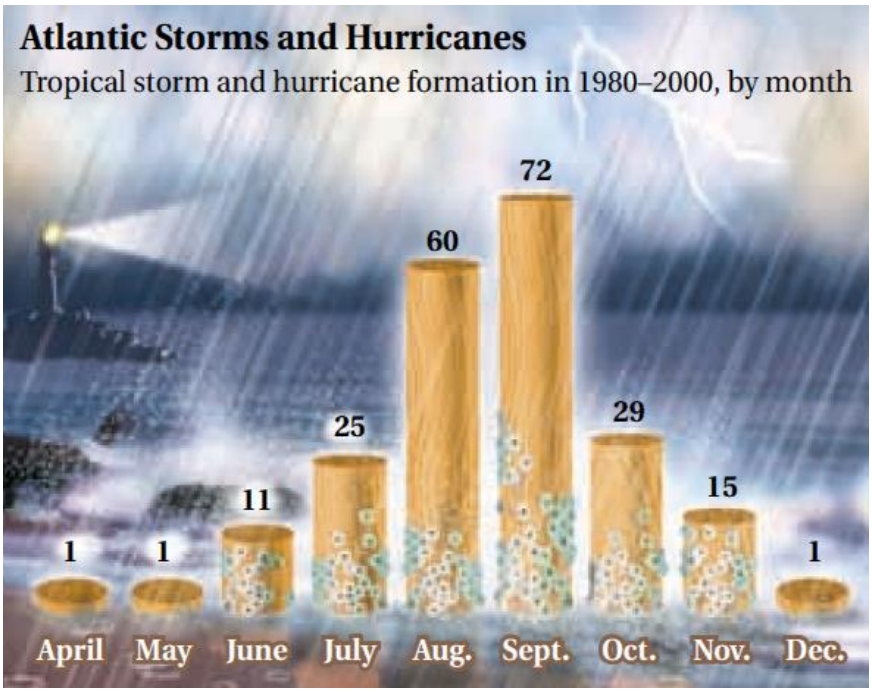
$32,500, $58,000, $87,000, $32,500, $64,800, $62,500

5. Find the mode of the following data
   a. 23, 24, 27, 18, 19, 27
   b. 83, 84, 84, 84, 85, 86, 87, 87, 87, 88, 89, 90
   c. 115, 117, 211, 213, 219

6. What is the range of the following set of data: 12, 25, 27, 29, 36, 38, 40, 43, 50, 54, 62?

7. The following prices per pound of sharp cheddar cheese were found at five supermarkets: $5.99, $6.79, $5.99, $6.99, $6.79
   a. What was the average price per pound?
   b. The median price?
   c. The mode?
   d. The range?

8. The following bar graph shows the number of Atlantic storms (hurricanes) that formed in various months from 1980 to 2000.



a. What is the average number for the nine months given?

b. The median?

c. The mode?

d. The range?

6.1 ~ Measures of Central Tendency (Day 2)

Objective: Use statistics appropriate to the shape of the data distribution to compare center and spread of two or more data sets.

❖ Mean/Median/Mode

CONCEPT SUMMARY — *Measures of Tendency*

Use	When . . .
mean	the data are spread out, and you want an average of the values
median	the data contain outliers
mode	the data are tightly clustered around one or two values

❖ Outlier

➢ A value that is very different from the other values in a data set.

Examples: Determining the Effects of Outliers

1. Identify the outlier in the data set {7, 10, 54, 9, 12, 8, 5}, and determine how the outlier affects the mean, median, mode, and range of the data.

   Outlier:

With the Outlier:	Without the Outlier:
Mean	Mean
Median	Median
Mode	Mode
Range	Range

Based on your results, are the following observations true or false?

a. An outlier can strongly affect the mean of a data set.

b. An outlier has little or no impact on the median and mode.

c. The mean is the best measure to describe a data set that contains an outlier.

d. The median or made better describes the center of a data set.

Example: Choosing a Measure of Central Tendency

2. Niles scored 70, 74, 72, 71, 73, and 96 on his six geography tests.
    a. Which measure – mean, median, or mode – gives Niles's test average? What is his test average?

    b. Which measure – mean, median, or mode – best describes Niles's typical score? Explain.

3. Would you use mean, median, mode, or range for each situation? Explain your reasoning.
    a. Jack noticed that half of the cereal brands in the store cost more than $2.00.

    b. The average score on the last test was 77%.

    c. The most common height on the basketball team is 6'1".

    d. The heights of the players on the basketball team vary by 8 inches.

    e. The most common price for a certain type of car is $25,000.

    f. Prices for tickets to the football game vary by $5.

    g. One-half of the cars at a dealership cost less than $23,000.

Analyzing Mean, Median, Mode, & Range

4. A student has six test grades. The average of those test grades is also the median of her test grades. Create a set of data that is favorable to the student, and then create a set of data that is unfavorable to the student.

5. At professional skating competitions, skaters are given scores ranging from 1 to 10 by 10 judges. The high and low scores are thrown out, and the skater receives the mean of the other eight scores. Create a set of data that is favorable to the skater and then create a set of data that is unfavorable to the skater.

6. You own a small company with eleven employees. You entire new employees by telling them that most people in your company make $100,000. Create a set of data where using the mode is unfavorable to an entry level employee.

❖ Quantitative vs. Categorical Data
- ➢ Quantitative Data
  - Data that is _____ and measures the quantity or frequency of something
- ➢ Categorical Data
  - Data that is qualitative or represents a _____ or quality of something

Quantitative
- Weight (pounds)
- Length (inches)
- Time (seconds)
- Number of questions correct on a quiz

Categorical
- Model of car
- Gender
- Yes or No
- Pass or Fail

Determine if the data is quantitative or categorical.

7. Jed's new horse: 15.2 hands high, 1250 pounds, costs $2200, age: 3 years, 4 months.

8. The tree: rough brown bark, red berries, wandering branching, small, sharply edged leaves

9. Pennsylvania cool: florescent black, earthy smell, black residue, rough texture

10. The students in the senior class at LHS High School: 578 students, 236 honor students, 150 scholarship winners, 51% male

6.2 ~ Dot Plots

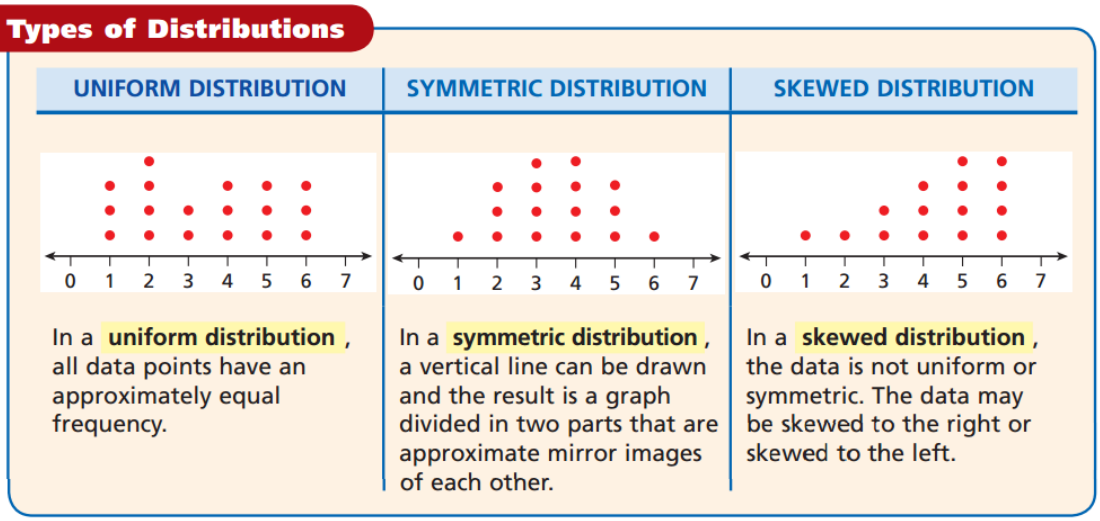
<u>Objective</u>: Represent data with plots on the real number line

- Dot Plots
  - A data representation that uses a number line and x's dots, or other symbols to show frequency.
    - Each dot can represent a ________________________________ from a set of data or a set number of observations from a set of data.
    - The dots are stacked in a column over the category so that the height of the column represents the relative or absolute frequency of observations in the category
  - Whole-Class Example:
    What is your favorite color?

Red	Orange	Yellow	Green	Blue	Purple	Black

❖ Describing Data Distributions

- Outliers: Extreme values that _____ from the rest of the data set
- Center: Graphically were _____ of the observations are on either side (_____)
- Spread: Refers to the variability
  - If it covers a wide range the spread is large
  - If observations are clustered the spread is smaller
- **Types of Distributions**



Example: An anthropology teacher at the community college is interested in analyzing the age distribution of her students. The students in her class are 21, 23, 25, 25, 24, 26, 22, 18, 19, 26, 28, 24, 22, 24, 19, 23, 24, 24, 21, 23, and 27 years old.

1. Organize the data in a dot plot.

	18	19	20	21	22	23	24	25	26	27	28	

2. Find the mean, median, mode, and range.

3. Is there an age that is much different from the rest of the data?

4. Describe the distribution of the anthropology class in as much detail as possible.

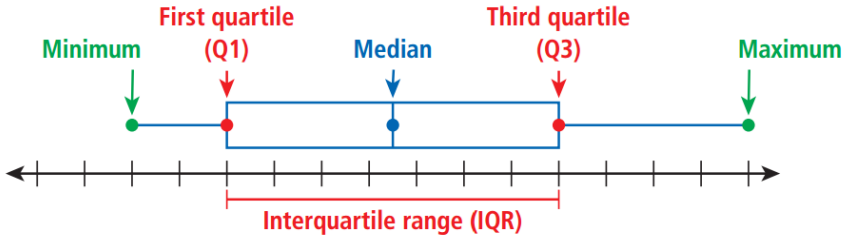
6.3 ~ Box Plots

Objective: Represent data with plots on the real number line

- Five-Number Summary
  - Minimum/Lower Extreme: the least value in the data set
  - Maximum/Upper Extreme: the greatest value in the data set
  - Median: the middle of the set of data
  - First/Lower Quartile (Q1): the median of the _____ of the data
  - Third/Upper Quartile (Q3): the median of the _____ of the data

- Box Plots – AKA Box & Whisker Plots
  - A graphical display used to display patterns of quantitative data that is split into quartiles

A *box-and-whisker plot* shows the spread of a data set. It displays 5 key points: the **minimum** and **maximum** values, the **median**, and the **first** and **third quartiles**.

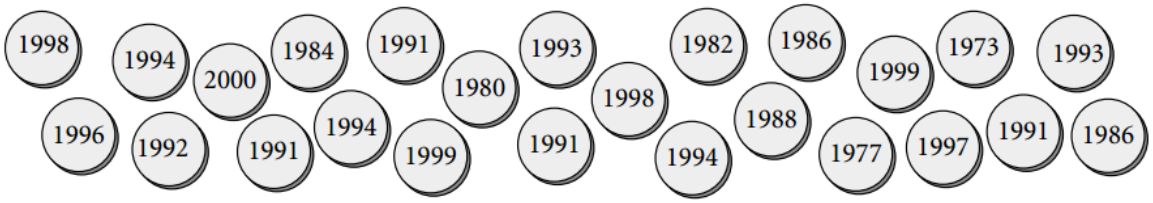


The quartiles are the medians of the lower and upper halves of the data set. If there are an odd number of data values, do not include the median in either half.

The *interquartile range*, or IQR, is the difference between the 1st and 3rd quartiles, or Q3 – Q1. It represents the middle 50% of the data.

❖ Other Number Summaries
- ➢ Inter Quartile Range (IQR): Q3 – Q1
- ➢ Outliers: Q1 – 1.5 × IQR Q3 + 1.5 × IQR

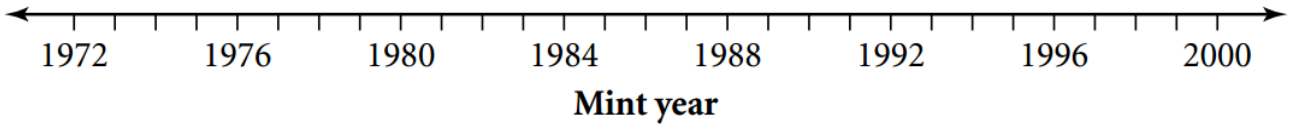
Example: Create a box plot of the mint years of the following collection of pennies.



Lower Extreme: Upper Extreme:

Lower Quartile: Upper Quartile:

Median:



Example: The ages of 9 people at a party in May are: 15, 37, 40, 87, 12, 35, 37, 39, 42.

1. Identify the…
   a. Lower extreme: b. Upper extreme:
   c. Lower quartile: d. Upper quartile:
   e. Median:

2. Create a box plot of the data.

3. What is the inter-quartile range?
4. Are there any outliers?

6.4 ~ Frequency Tables & Histograms

Objective: Represent data with plots on the real number line

- Frequency Tables
  - The _____ of a data value is the number of times it occurs.
  - A _____ shows the frequency of each data value.
    - If the data is divided into intervals, the table shows the frequency of each interval
- Histograms
  - A graphical display that shows the frequency of data divided into equal intervals

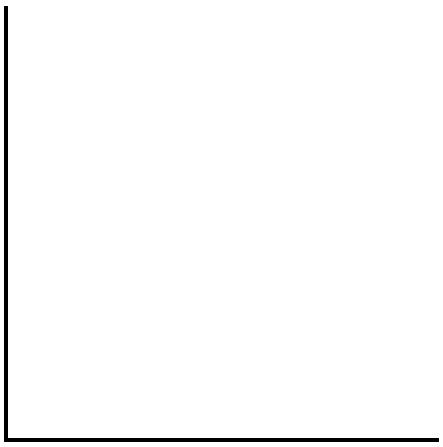
Example: The hurricane season lasts 26 weeks every year. The following data is the number of weeks into a season that a hurricane was reported from 1997 – 2000:

6, 7, 14, 12, 13, 14, 16, 16, 17, 17, 19, 21, 26, 12, 12, 13, 15, 15,
20, 20, 24, 10, 12, 15, 16, 17, 18, 20

1. Identify the least and greatest values.
2. Divide the data into equal intervals and complete the frequency table:

Interval	Tally	Frequency

3. Create a histogram based on the frequency table.



4. Describe the data in as much detail as possible. (See "Describing Data Distributions" in Lesson 6.2)

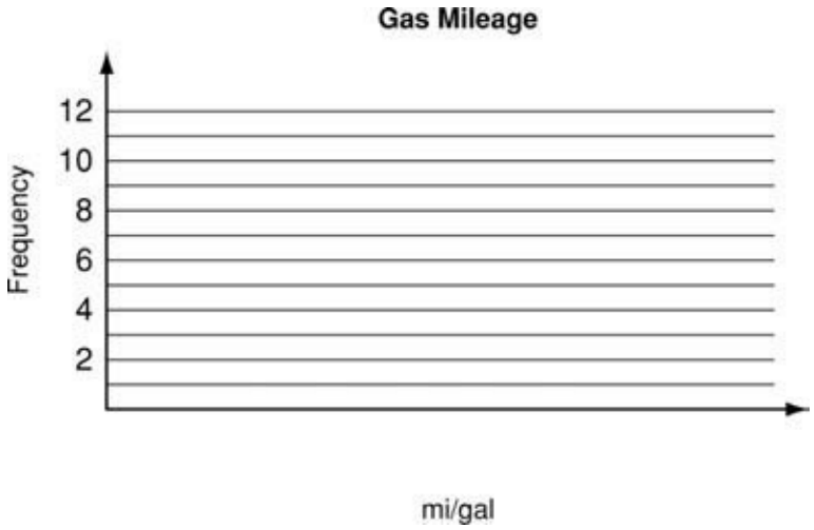
Example: The estimated miles per gallon for selected cars are shown in the table.

5. Use the data to make a frequency table with intervals.

26	28	32	33	26	15	21
35	17	18	25	29	30	26
27	30	24	25	24	32	25
19	22	32	25	31	28	23
27	23	24	20	38	44	18

Interval	Tally	Frequency

6. Make a histogram.



6.5 ~ Standard Deviation

Objective: Use and apply the mean, variance, and standard deviation of a data set

Review

1. A softball pitcher threw softballs at speeds of 55 mph, 42 mph, 62 mph, and 52 mph. Explain how you can find the *mean* speed of the softballs thrown.

❖ Variance & Standard Deviation

The data sets $\{19, 20, 21\}$ and $\{0, 20, 40\}$ have the same mean and median, but the sets are very different. The way that data are spread out from the mean or median is important in the study of statistics.

A *measure of variation* is a value that describes the spread of a data set. The most commonly used measures of variation are the *range*, the interquartile range, the *variance*, and the *standard deviation*.

The **variance**, denoted by σ^2, is the average of the squared differences from the mean. **Standard deviation**, denoted by σ, is the square root of the variance and is one of the most common and useful measures of variation.

❖ Statistical Measures

- Standard deviation is a measure of how far the numbers in a data set deviate from the mean.
  - Variance shows the same
- A measure of variation – such as range and interquartile range – describes how the data in a data set are spread out

How to find Variance and Standard Deviation

- Find the mean, $\bar{x}$, of the n values in the data set.
- Find the difference, $x_i - \bar{x}$, between each value x_i and the mean.
- Square each difference, $(x_i - \bar{x})^2$.
- Find the average (mean) of these squares. This is the variance.

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{n}$$

- Take the square root of the variance. This is the standard deviation.

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

2. The data represent the number of milligrams of a substance in a patient's blood, found on consecutive doctor visits. Find the mean and standard deviation of the data: {14, 13, 16, 9, 3, 7, 11, 12, 11, 4}

   a. Find the mean: $\overline{x}$

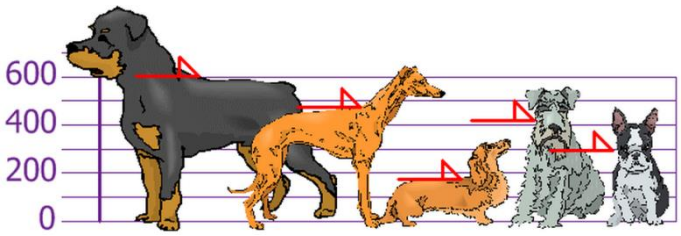
   b. Find the difference between the mean and each data value, and square it.

x_i	14	13	16	9	3	7	11	12	11	14
$x_i - \overline{x}$										
$(x_i - \overline{x})^2$										

   c. Find the variance: σ^2
   (The average of the last row of the table)

   d. Find the standard deviation: σ
   (The square root of the variance)

3. You and your friends have just measured the heights of your dogs (in millimeters). The heights (at the shoulders) are 600 mm, 470 mm, 170 mm, 430 mm, and 300 mm.



   a. Find the mean: $\overline{x}$

   b. Find the difference between the mean and each data value, and square it.

x_i	600	470	170	430	300
$x_i - \overline{x}$					
$(x_i - \overline{x})^2$					

   c. Find the variance: σ^2
   (The average of the last row of the table)

   d. Find the standard deviation: σ
   (The square root of the variance)

4. Identify the error(s) in planning the solution or solving the problem. Then write the correct solution.

What are the mean, variance, and standard deviation of these values?
62 41 54 60 49 58

$$\bar{x} = \frac{62 + 41 + 54 + 60 + 49 + 58}{6} = 54$$

$$\sigma^2 = \frac{\left(\Sigma(x - \bar{x})^2\right)^2}{n} = \frac{310^2}{6} \approx 16{,}016.67$$

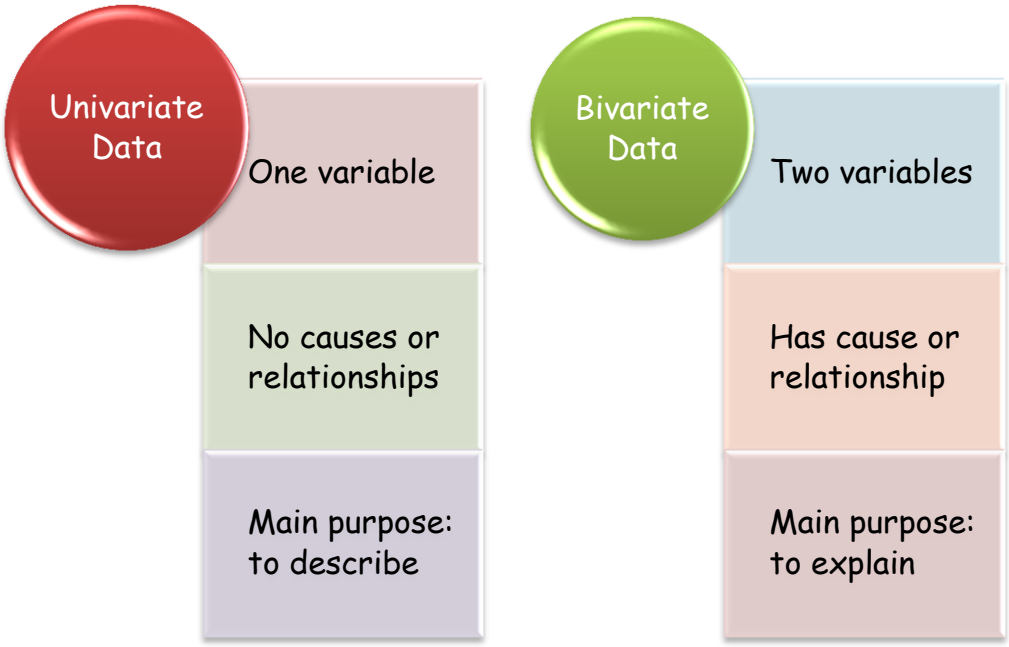
$$\sigma = \sqrt{\sigma^2} \approx \sqrt{16016.67} \approx 126.56$$

x	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
62	54	8	64
41	54	−13	169
54	54	0	0
60	54	6	36
49	54	−5	25
58	54	4	16
		Sum	310

6.6 ~ Two-Way Frequency Tables

Objective: Use two-way frequency tables to describe and interpret bivariate categorical data

- Univariate Data vs. Bivariate Data



❖ 2 Way Frequency Table
- A useful way to organize data that can be categorized by two variables

➢ Joint Relative Frequency
- The values in each category divided by the total number of values

➢ Marginal Relative Frequency
- Found either by adding the joint relative frequencies in each row or column or by taking the total of each row and column and dividing it by the total

➢ Conditional Relative Frequency
- Divide the joint relative frequency by the marginal relative frequency of the condition

Examples:

1. Jenna asked 40 randomly selected students whether they preferred dogs, cats or other pets. She also recorded the gender of each student in the table below.

Preferred Pet	Dog	Cat	Other	Total
Boys	10	5	9	24
Girls	8	7	1	16
Total	18	12	10	40

a. How many total students took the survey?

b. How many total boys took the survey?

c. How many students like dogs as pets?

d. How many boys said they like dogs as pets?

e. Calculate the Joint Relative Frequencies and The Marginal Relative Frequency for each. Use a different color for the joint and marginal frequencies.

Preferred Pet	Dog	Cat	Other	Total
Boys				
Girls				
Total				

f. Find the joint relative frequency of students surveyed who are girls that prefer dogs as pets.

g. Find the joint relative frequency of the students who are boys who prefer cats as pets.

h. Find the marginal relative frequency of students who like things other than dogs or cats as pets?

i. Find the conditional relative frequency that a student prefers cats as pets given the student is a girl.

j. Find the conditional relative frequency that a student surveyed is a girl given that she prefers cats as pets.

k. Does the table reflect a gender bias toward any particular pet?

l. What type of pet is most preferred?

2. You survey your friends about the type of party they most enjoy.

	Male	Female	Total
Bowling	6	2	8
Skating	3	11	14
Dancing	1	3	4
Total	10	16	26

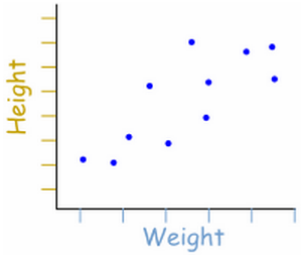
a. Calculate the Joint Relative Frequencies and The Marginal Relative Frequency for each. Use a different color for the joint and marginal frequencies.

	Male	Female	Total
Bowling			
Skating			
Dancing			
Total			

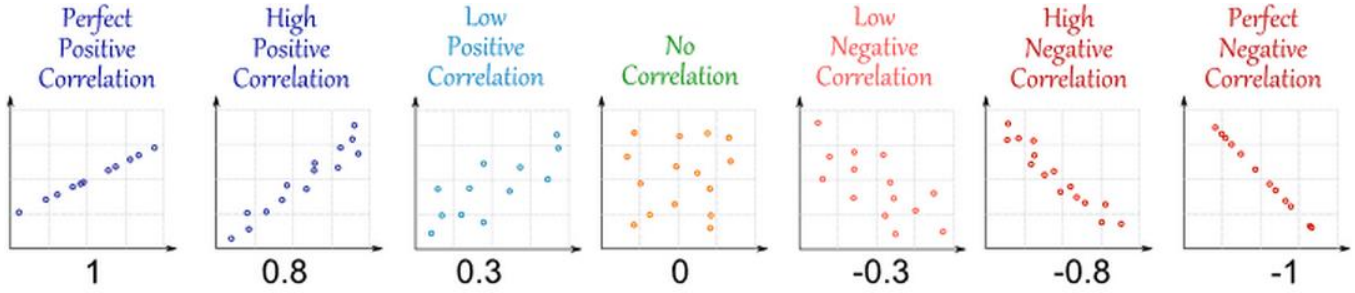
b. What type of party would you plan for your friends? Explain.

6.7 ~ Scatter Plots & Line of Best Fit

Objective: Use scatter plots to describe and interpret bivariate quantitative (numerical) data



- Scatter Plots
  - A graphical display that shows the _____ between two sets of _____ data
- Describing Scatter Plots
  - Form
    - Linear ~ the overall pattern of the data appears to be _____
    - Non-linear ~ the overall pattern of the data seems to be _____
    - None ~ the points show _____ relationship
  - Direction
    - Positive ~ the data trends _____
      - As one value _____, so does the other value
    - Negative ~ the data trends _____
      - As one value _____, the other value _____
    - Flat
      - As one value _____, the other value _____
  - Strength
    - The overall _____ of the data to a line of best first
      - If the data points are concentrated around a line the relationship is strong
- Correlation & Causation
  - Correlation: When one variable doesn't have a direct impact on the other.
    - A scatterplot may show that a relationship exists but it does not and cannot prove that one variable is causing the other.
      - There could be a third factor involved causing both, or a systemic cause, or the relationship could be a fluke.
        - i.e. Smoking correlates to alcoholism.
  - Causation: When one variable causing another variable.
    - i.e. Smoking causes lung cancer.

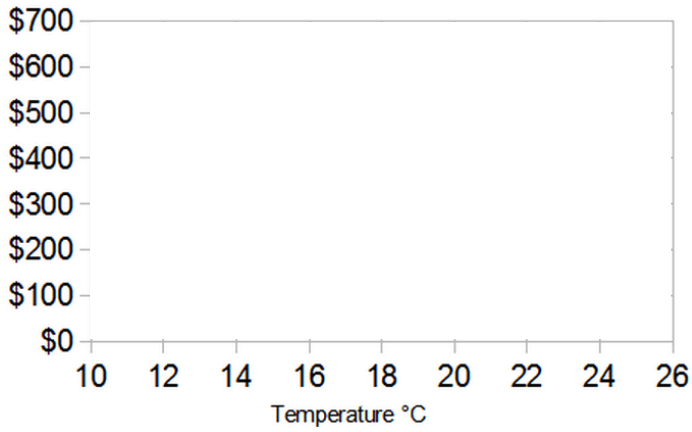


Example:

The local ice cream shop keeps track of how much ice cream they sell versus the temperature on that day. Here are their figures for the last 12 days:

Ice Cream Sales vs Temperature	
Temperature °C	**Ice Cream Sales**
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

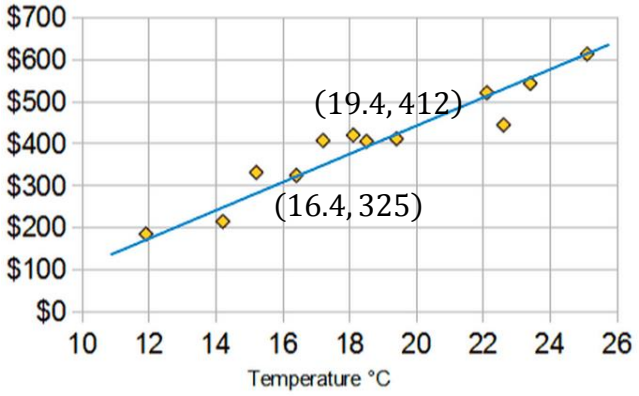
1. Create a scatter plot:



2. How are ice cream sales correlated to temperatures?

❖ Line of Best Fit

- A line that can be fit to a linear scatter plot so that about half the data is on either side of the line.
- Prediction Equation
  - The equation of a line of best fit
  - Can be used to predict one of the variables given the other variable



- To find a line of fit and a prediction equation for a set of data, select two points that appear to represent the data well, and then use the following formulas:

$$m = \frac{y_2 - y_1}{x_2 - x_1} \qquad y - y_1 = m(x - x_1)$$

3. Find an equation for the line of best fit for the scatter plot.

4. The table shows the temperature in the atmosphere at various altitudes.

Altitude (ft)	0	1000	2000	3000	4000	5000
Temp (°C)	15.0	13.0	11.0	9.1	7.1	?

Source: NASA

a. Draw a scatter plot and a line of fit, and describe the correlation.

b. Use two ordered pairs and write a prediction equation.

c. Use your prediction equation to predict the missing value.