Unit 1-Lessons

Graphical Displays of Data

MDM4U

© MARK ANDERSON, ALL RIGHTS RESERVED   WWW.ANDERTOONS.COM

"He's right! When you look at it that way, it's not so bad!"
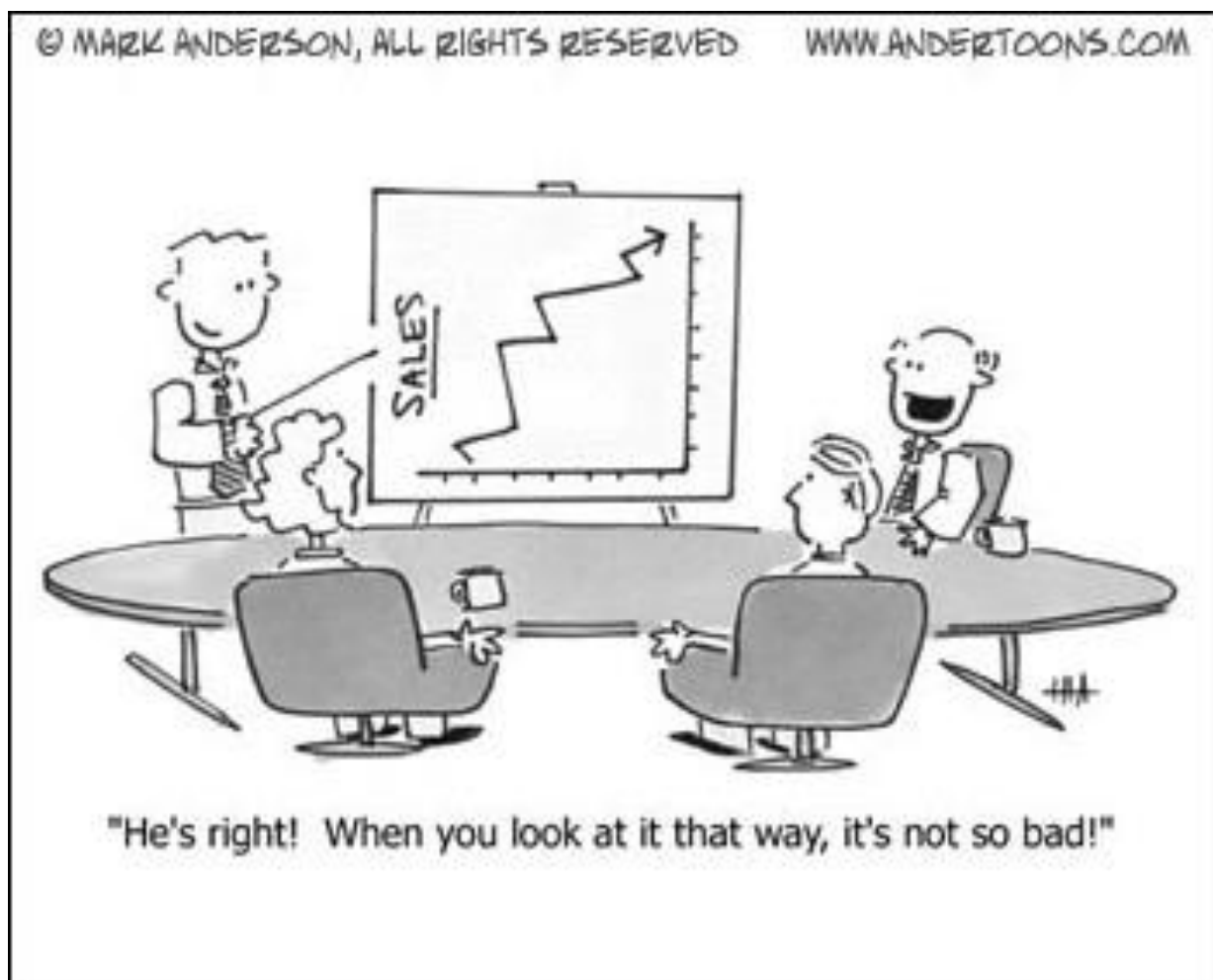
Unit Outline

Overall Unit Goal: Analyze, interpret, and draw conclusions from two-variable data using numerical, graphical, and algebraic summaries.

Section	Subject	Learning Goals	Curriculum Expectations
1.1	Intro to Statistics	- Understand what data management is	C1.1
1.2	Organizing and Displayinig Categorical Data	- understand the difference between quantitative and qualitative data - know how to display qualitative data using bar graphs, pie graphs, pictographs, contingency tables	C1.3, D1.3, D1.5, D2.1, D2.2, D2.3
1.3	Organizing and Displaying Quantitative Data	- learn how to display quantitative data using stemplots, boxplots, and histograms	B2.3, B2.4, C1.3, D1.2, D1.3, D1.5, D2.3
1.4	Scatter Plots and Correlation	- know how to create a scatterplot to show the correlation between two quantitative variables - be able to state the direction and strength of correlation based on scatterplot	D2.1, D2.2, D2.3
1.5	Correlation Using Technology	- use technology to perform a linear regression - interpret regression equation and r and r^2 values	D2.4, D2.5
1.6	Least squares line and Correlation Coefficient by Hand	- determine regression equation, r, and r^2 without technology	D2.5
1.7	Misrepresentations of Data	- understand common ways data can be misrepresented in the media	D3.1, D3.2

By the end of the unit, you will be able to:
- Create appropriate graphs for categoric and numeric data (bar, box, histogram, etc.)
- Create a scatterplot to display the correlation between two numeric variables
- Perform a linear regression algebraically and using technology to analyze the correlation between two variables

Assessments	F/A/O	Ministry Code	P/O/C	KTAC
Note Completion	A		P	
Practice Worksheet Completion	F/A		P	
Quiz – Linear Regression	F		P	
Assignment – Graphs Using Excel	O	D1.3, D2.3, D2.4, D2.5	P	K(23%), T(7%), A(43%), C(27%)
PreTest Review	F/A		P	
Test – Solving Equations	O	B2.3, B2.4, C1.3, D1.2, D1.3, D1.5, D2.1, D2.2, D2.3, D3.1, D3.2	P	K(33%), T(12%), A(25%), C(30%)

Section 1.1 – Intro to Statistics

MDM4U
Jensen

Part 1: Course Outline

Mark Breakdown:

- 5 Unit Tests – 35%
- 4 In Class Assignments – 15%
- 5 In Class Workbook Problem Sets – 5%
- Games Fair Project (LFD) – 5%
- Culminating Project – 10%
- Final Exam – 30%

Formative Assessments:

- There will be a quiz each unit
- At the end of each unit on the day of the test you will be required to submit a package that includes all completed lessons and homework

Expectations:

- Come to class ON TIME each day with unit package, graphing calculator, and pencil
- Usage of cell phones during class is not permitted
- Ask for permission to leave the class (no disappearing)
- Participate in lessons and activities
- Complete your homework every night
- Ask questions! Extra help is available Tuesday and Thursday at lunch in this room.

Part 2: Intro to Statistics

_____ are any collection of numbers, characters, images, or other items that provide information about something.

Statistics is the _____ of _____. The volume of data available to us is overwhelming. For example, astronomers work with data on tens of millions of galaxies. The checkout scanners at Walmart's 10 000 stores in 27 countries record hundreds of millions of transactions every week. Professional sports teams collect extraordinary amounts of performance data during games. In all these cases, the data are trying to tell us a story. To hear what the data are saying, we need to help them speak by _____, _____, _____, and _____. That is _____. Statistical methods enable us to look at information from a small collection of people or items and make _____about a larger collection of people or items. For instance, if we wish to estimate the proportion of people who will have a severe reaction to a flu shot without giving the shot to everyone who wants it, statistics provides appropriate methods.

To get you in a more 'statistical' mindset, read the following two stories:

1: If you have a Facebook account, you have probably noticed that the ads you see online tend to match your interests and activities. Coincidence? Hardly. According to the Wall Street Journal, much of your personal information has probably been sold to marketing or tracking companies. Why should Facebook give you a free account and let you upload as much as you want to its site? Because your data are valuable! Using your Facebook profile, a company might build a profile of your interests and activities. From Facebook's point of view, your data are a potential gold mine.

2: How dangerous is texting while driving? Researchers at the University of Utah tested drivers on simulators that could present emergency situations. They compared reaction times of sober drivers, drunk drivers, and texting drivers. The texting drivers actually responded more slowly and were more dangerous than those who were above the legal limit for alcohol.
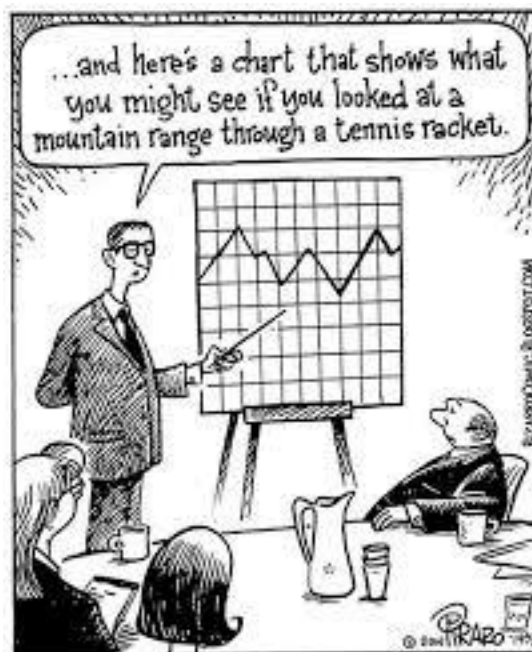
Part 3: M&M's Activity

I. Collecting the Data: Scoop out a sample of M&M's. Count the total number of M&M's in your sample. You will need exactly 25, so if you need more, randomly choose a few more to add to your sample. If you have too many, you must randomly choose M&M's to discard. *DO THIS WITH RANDOMLY!*

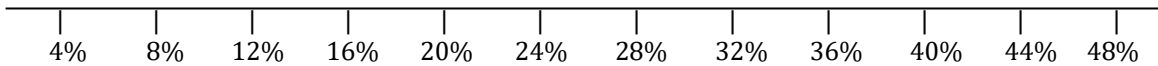Calculate the percentage of BLUE candies in your sample: _____

Record the class data using the following chart:

II. Organizing the Data: Organize the data in a meaningful way.

Title: _____

III. Displaying the Data: Display the data using a dot plot.

Title: _____

```
_____
   |     |     |     |     |     |     |     |     |     |     |     |     |
  4%    8%   12%   16%   20%   24%   28%   32%   36%   40%   44%   48%
```

IV. Analyzing the Data:

Describe some general features of the data.

What would you consider a "normal" or "typical" percentage of blue M&M's? Why?

Does our data reveal the true percentage of blue M&M's? If so, what is the true percentage? If not, what DOES it reveal about the true percentage?

Part 4: Explanation of Culminating Project

1. You will pose a significant problem whose solution would require the organization and analysis of a large amount of data.

2. You will apply the skills you learn in the course to design and carry out a study of the problem.

3. Compile a clear, well-organized, and fully justified report of the investigation and its findings.

https://www.youtube.com/watch?v=HNlgISa9Giw

http://www.youtube.com/watch?v=jbkSRLYSojo
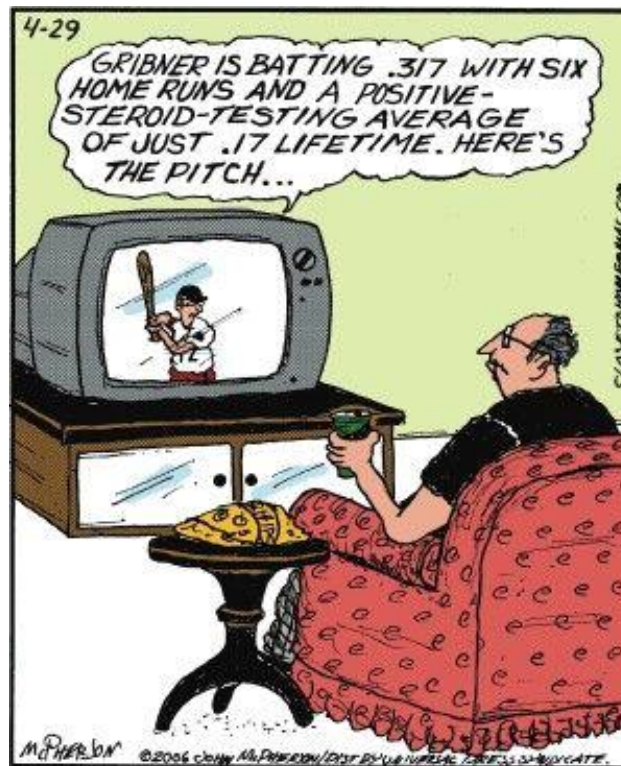
Homework Task: Explore the statistics Canada website and find at least one data table for a subject that you find interesting. Transport this data table in to a spreadsheet program (excel, numbers, etc.). Organize the data table so it is easily readable. Submit electronically to our class EDSBY page.

http://www.statcan.gc.ca/start-debut-eng.html

Part 1: Types of Variables
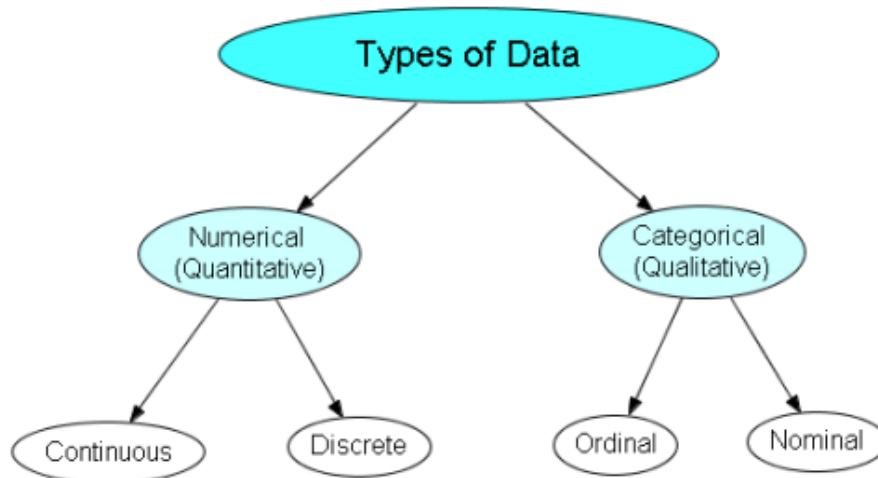
Numeric/Quantitative Variable: A variable that takes _____ values for which it makes sense to find an average. These variables can be either _____ or _____.

Continuous Variable: A numeric variable that can have an _____ number of values in a given interval. Measurable with _____.

    Examples:

Discrete Variable: A numeric variable that can take on only a _____ number of values within a given range. Usually measured with integer values only.

    Examples:

Categorical/Qualitative Variable: A variable that places an individual into one of several _____ or _____. Categorical variables may have categories that are naturally ordered (_____ variables) or have no natural order (_____variables).

Ordinal Variable: A categorical variable that has a _____of its possible values, but the distances between the values are undefined.

    Example:

Nominal Variable: Type of categorical variable that describes a name, label, or category with <u>no</u> _____.

    Example:

Part 2: Frequency Tables

To make an accurate picture of data, the first thing we have to do is make 'piles'. For categorical data, 'piling' is easy. We just count the number of cases corresponding to each category. We can organize these counts into a _____, which records the totals and category names.

Frequency tables are used to _____ data.

Example 1:

Grade 12's were asked when their spares were and these were the results:

A, B, C, D, A, D, D, B, A, C, A, C, B, B, B, A, D, C, A, A, B, D, C, A, B
B, A, C, C, D, A, B, A, B, B, B, D, D, A, D, D, C, A, D, C, D, A, B, B, A

The problem with data that is presented like this is that you can't 'see' what is going on. Organize the data in to a frequency table to better see the distribution of data.

Spare	Frequency

Counting the frequency is useful, but sometimes we want to know the _____ of data in each category, so we make a _____.

A relative-frequency table shows the frequency of a data group as a _____ or _____of the whole data set.

Spare	Frequency	Relative Frequency

Part 3: Bar Graphs

Graphs are used to _____ data. Bar graphs, segmented bar graphs, pie charts, and pictographs are appropriate types of graphs for displaying the data of _____ variables. Bar graphs can also be used for discrete numeric variables.

A bar graph displays the distribution of a categorical variable, showing the counts (frequency) for each category next to each other for easy comparison.

A bar graph consists of parallel bars of equal widths (_____) with lengths proportional to the _____ of the variables they represent.

Example 2:

The following frequency table shows the number of different U.S radio stations broken up by category based on the kind of music they broadcast.

I. Complete the relative frequency column

Genre	Frequency	Relative Frequency
Classic Rock	569	
Modern Rock	869	
Country	2066	
News/Talk	2179	
Oldies	906	
Pop	2575	
Electronic	626	
Hip-Hop/R&B	450	
Total	10240	

II. Use the table to create two bar graphs. The first showing frequencies and the second showing relative frequencies of each category.
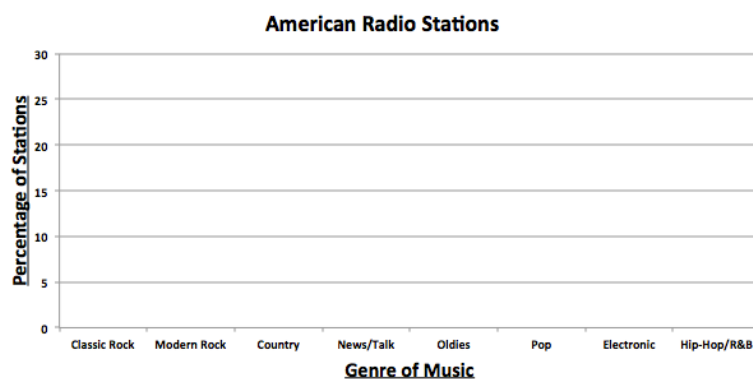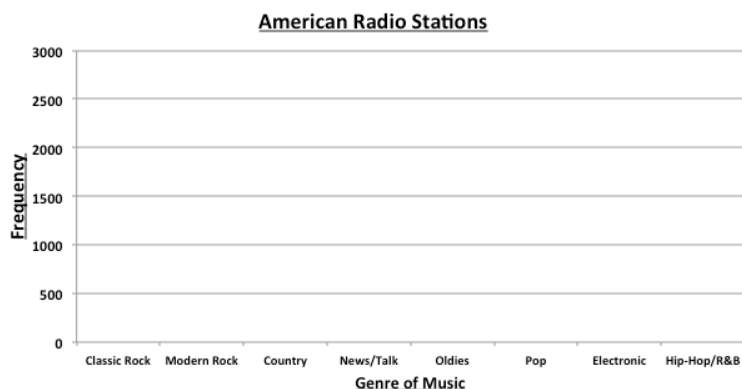
What do you notice about the shapes of the distributions?

Part 4: Pie Charts
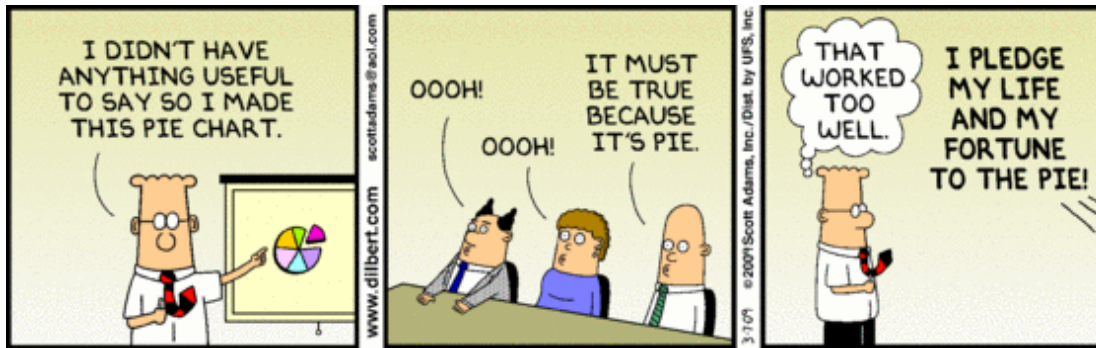
A pie chart shows the distribution of a categorical variable as a 'pie'. They slice the circle into pieces whose sizes are _____ to the fraction of the whole in each category. A pie chart is best used when trying to show a category's relation to the whole. Pie charts are awkward to make by hand, but technology will do the job for you.
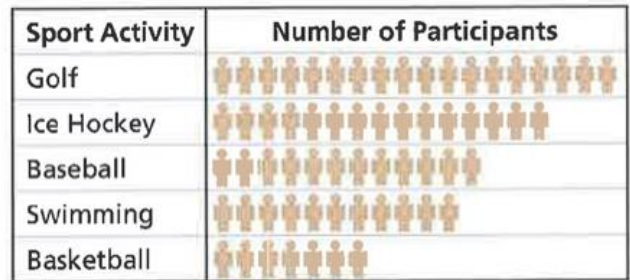
Here is a pie chart showing the data for the U.S radio stations from the previous example:
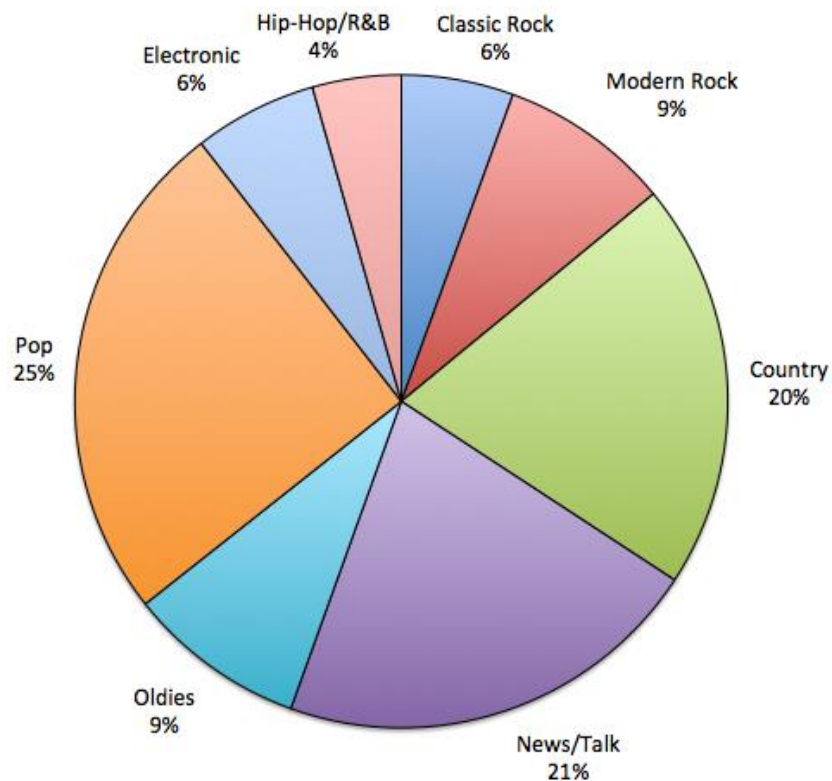
Part 5: Pictographs

A _____ is a symbolic representation of data. The following pictograph displays the number of participants, aged 15 and older, in the five most popular sports activities in Canada.

How many people aged 15 and older play hockey?

Sport Activity	Number of Participants
Golf	🧍🧍🧍🧍🧍🧍🧍🧍🧍🧍🧍🧍🧍🧍🧍🧍🧍🧍🧍
Ice Hockey	🧍🧍🧍🧍🧍🧍🧍🧍🧍🧍🧍🧍🧍🧍🧍
Baseball	🧍🧍🧍🧍🧍🧍🧍🧍🧍
Swimming	🧍🧍🧍🧍🧍🧍🧍🧍
Basketball	🧍🧍🧍🧍🧍

Legend: 🧍 represents 100 000 people

Example 3:

a) How many red delicious apples are in the store?

Varities of Apples in a food store	
Red Delicious	🍎 🍎 🍎
Golden Delicious	🍎 🍎 ◐
Red Rome	🍎 🍎 🍎 🍎
McIntosh	🍎 🍎
Jonathan	🍎 🍎 🍎 ◐
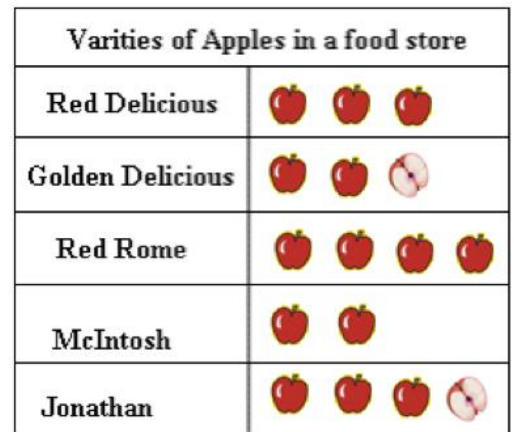
🍎 = 10 apples ◐ = 5 apples

b) How would you represent 11 apples?

Problems with Pictographs:

- Pictographs can make a graph more interesting but…

Part 6: Contingency Tables and Segmented Bar Graphs

We have learned some techniques for analyzing the distribution of a single categorical variable. If a data set involves two categorical variables, we use a _____.A two-way table of counts organizes data about two categorical variables measured from the same set of individuals.

Example 4: Only 32% of those aboard the Titanic survived. Was that survival rate the same for all ticket classes? To answer that question, we can arrange the counts for the two categorical variables, survival and ticket class, in a two-way table.

		First	Second	Third	Crew	Total
Class						
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

In this case, survival is our _____and class is our _____. The margins of the table give totals. When analyzing a contingency table, the goal is to see if the variables depend on each other. This can be done by looking at the two possible _____.

If we think that class might depend survival, then we should look at the distribution of the _____ percentages. This is the conditional distribution for class based on survival.

		First	Second	Third	Crew	Total
Class						
Survival	Alive	203	118	178	212	711
		28.6%	16.6%	25.0%	29.8%	100%
	Dead	122	167	528	673	1490
		8.2%	11.2%	35.4%	45.2%	100%

However, in this scenario it would make more sense to determine if survival depends on class. To do this, we should look at the _____ percentages. This is the conditional distribution for survival based on class.

			First	Second	Third	Crew	Total
Class							
Survival	Alive	Count	203	118	178	212	711
		% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
	Dead	Count	122	167	528	673	1490
		% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
	Total	Count	325	285	706	885	2201
			100%	100%	100%	100%	100%

Looking at how the percentages change across the row, it sure seems that class influenced whether a persons survived or not. _____of first class passengers survived while only _____of third class passengers survived.

Two-way tables are often displayed using _____.

Example: Segmented bar graph of survival based on class using frequencies
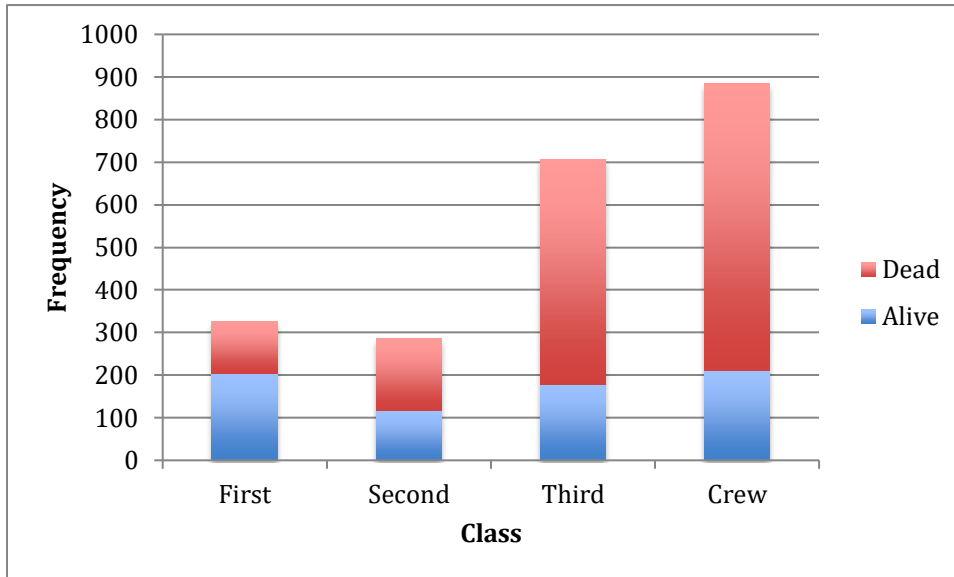
Example: Segmented bar graph of survival based on class using conditional percentages
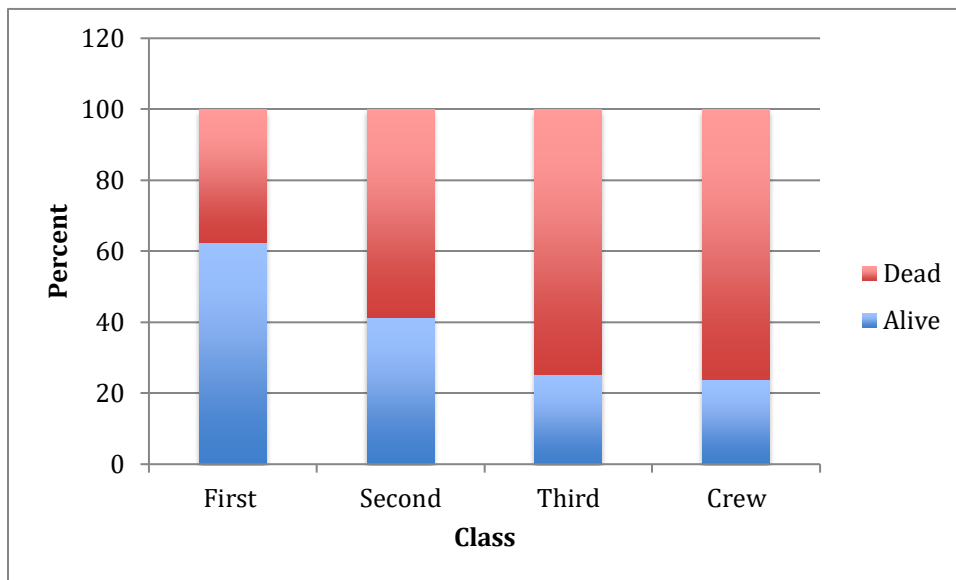
Note: The bars of each graph have the same proportions but it is easier to see in the second graph that first class passengers had the highest proportion of survivors.

Numeric Variable: A quantitative variable that takes numerical values for which it makes sense to find an average. These variables can be either continuous or discrete

Part 1: Game of Greed

Rules:

Everyone stands. Someone throws a die twice and totals the numbers. This is everybody's current score. Those that are happy with that score sit down; they have finished this round. They record their score.

For the others, the die is rolled again. Those still standing get to add the number to their total, UNLESS it is a 2. If it is a 2, the game is over and all those standing receive 0 for that round.

Keep throwing the die until a 2 comes up or everyone has sat down and recorded their score for that round. A game consists of 5 rounds.

At the end of the game, the students add their 5 scores to get their total.

Individual Results:

Round 1	Round 2	Round 3	Round 4	Round 5

My total score is: _____

Class Results:

Part 2: Stemplots

A simple graphical display for fairly small data sets of a quantitative variable is a _____(also called a stem-and-leaf plot). We made a stemplot to display the scores for the game of greed played at the beginning of this lesson.

Rules for making a stemplot:

- Each number is separated into...
  - Stem: _____
  - Leaf: _____
- Write the stems in a vertical column with the smallest at the top.
  - Do NOT skip any stems, even is there is not data value for a particular stem.
- Draw a vertical line at the right of this column
- Write each leaf, in increasing order, in the row to the right of its stem

Example 1: The points for the 30 NHL teams from the 2013 regular season are recorded below:

72, 56, 55, 49, 48, 63, 62, 57
56, 48, 57, 51, 42, 40, 36, 77
60, 56, 55, 41, 59, 55, 45, 42
39, 66, 59, 57, 51, 48

Display the data using a stemplot:

Part 3: Boxplots (5 number summary)

Let's start by watching a video introducing boxplots:

http://www.learner.org/courses/againstallodds/unitpages/unit05.html

While watching the video, fill in answers to the following five questions:

1. What variable is used to compare different brands of hot dogs?

2. What name do we give to the value for which one-quarter of the data values falls at or below it?

3. What numbers make up a five-number summary?

4. How do you calculate the interquartile range?

A boxplot is a _____that shows the distribution of a set of quantitative data.
The five numbers a boxplot displays are:

Min: smallest data value
Q_1:
Median (Q_2):
Q_3:
Max:

Note: The median of a set of data is the _____piece of data. If there are an even number of
data points in a set, the median is the _____of the two middle most pieces of data.

In a boxplot, the box contains the _____of the data and its width represents the _____of
the data.

The upper and lower limits for the box are found by finding the _____for the upper and lower half
of the data set. The median value itself is not included in the lower or upper half.

From the sides of the box, horizontal lines are drawn extending to the _____ and _____
data points that are NOT outliers.

_____ values are used to determine which pieces of data are _____.

Lower threshold =
Upper threshold =

Note: IQR stands for interquartile range which is =
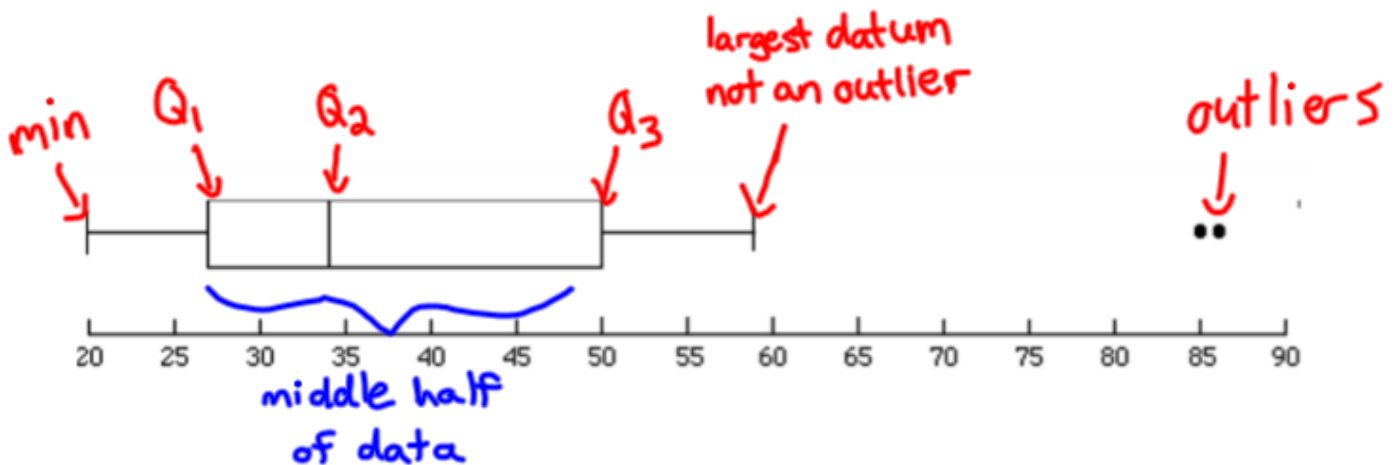
Steps to drawing a boxplot:

- Make sure the data points are in order from smallest to largest
- Find the 5-number summary (1-Var Stats), identify outliers.
- Draw a scale (number line) above which plot will be drawn—include numbers and units: can be vertical or horizontal.
- Draw rectangular box with ends at Q_1 and Q_3.
- Draw line through box at median (Q_2).
- Draw two "whiskers" from corresponding ends of box to most extreme data value that is <u>not</u> an outlier—inside thresholds. Put dots or other marks for each outlier value.

Note: Don't draw thresholds on boxplot—they are not data values. Only use them to identify outliers.

Example 2: A random survey of people at a golf course asked them how many times they had seen Happy Gilmore. The results are shown below in ascending order

$$1, 2, 2, 3, 3, 3, 4, 4, 5, 5, 6, 7, 8, 9, 10, 10, 12, 15, 26$$

a) Find the five number summary of the data (can use 1-var stats on calculator)

b) Identify outliers

c) Create a boxplot of the data
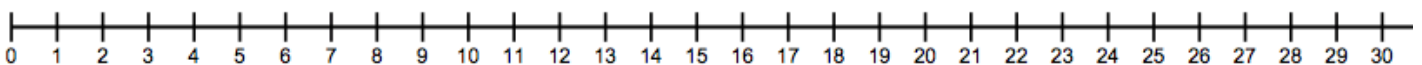
Example 3:

The times, in minutes, it took ten police officers to complete routine paperwork after their shift are given below

$$10, 32, 36, 38, 41, 43, 44, 48, 80, 89$$

Find the five-number summary all ten officers. Draw a full (modified) boxplot, showing how you identified any outliers.

Thresholds:

```
├──┼──┼──┼──┼──┼──┼──┼──┼──┼──┼──┼──┼──┼──┼──┼──┼──┼──┼──┤
0   5  10  15  20  25  30  35  40  45  50  55  60  65  70  75  80  85  90  95
```
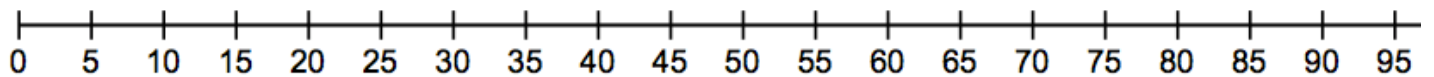
Using Ti 83/84 for Boxplots

Below shows how the previous example can be completed using the ti-83/84 calculator.

-   input data in to list L1: STAT → EDIT
-   Determine values for 5 number summary: STAT → CALC → 1-VARSTATS → List: L1 → CALCULATE

-   turn on statplot: 2nd → y= → ENTER → ENTER
-   select modified boxplot
-   view graph: GRAPH → ZOOM → ZOOMSTAT
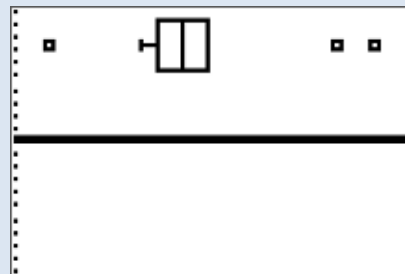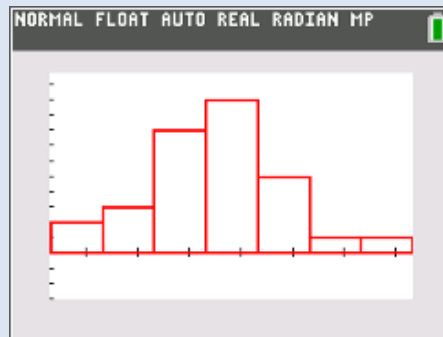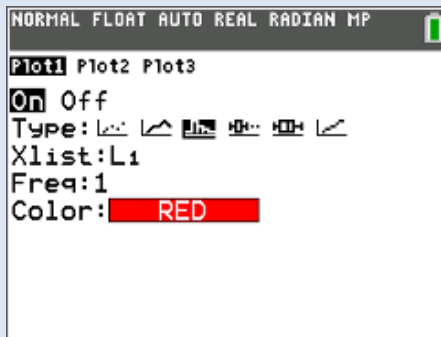
Part 4: Histograms

Quantitative variables often take many values. A graph of the distribution is clearer if nearby values are grouped together. A _____ is a frequency distribution where the horizontal access is divided into equal class _____ in to which data have been divided. The heights of the rectangles (**that have no spaces between them**) represent the frequencies associated with the corresponding intervals. It is important that each interval have the same width. Histograms are most appropriate for _____variables but you will see them for _____ variables as well.

Example 4:

a) Is the following graph a bar graph or a histogram? How do you know?

Height of High School Students

Number of Students

This shows that all other values are assumed to be included.

120 130 140 150 160 170 180 190 200
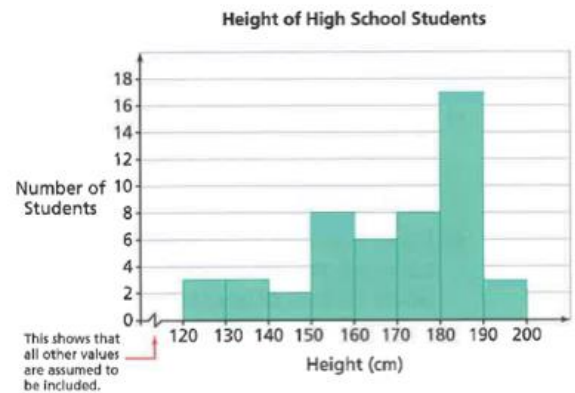
Height (cm)

b) Which height interval has the highest frequency? What is the frequency?

Steps for making a histogram:

1. Choose the number of intervals (if the question doesn't specify, choose between 5 and 10)

2. Calculate the range of your data (largest data point – smallest data point)

3. Round your range UP to a number that is easily divided by the number of intervals you chose.

4. Calculate your bin width; $\boldsymbol{bin\ width} = \dfrac{range}{number\ of\ intervals}$

5. Determine the first value for your first interval; $\boldsymbol{lowest\ value} - \dfrac{rounded\ range - actual\ range}{2}$

6. If any data points fall on the border of any of the intervals, add a decimal place to ensure that this doesn't happen.

7. Make a frequency table using the intervals you have determined.

8. Draw the histogram (no spaces between bars)

Example 5:

Here are a class' scores obtained on a data management exam:

$78, 81, 55, 60, 65, 86, 44, 90$

$77, 71, 62, 41, 80, 72, 70, 64$

$88, 73, 61, 70, 75, 98, 51, 73$

$59, 68, 65, 81, 78, 67$

a) Determine the range of the data

b) Determine an appropriate bin (interval) width that will divide the data into 6 intervals.

> Note:
>
> Round your range UP to a value that can be divided easily.

c) Determine the first value of your first interval

We added ___ to 57 when we rounded our range, therefore we should subtract _____from our smallest value _____; which makes our starting point _____.

Or use formula:

> **Note:**
>
> 1. If you have rounded your range up you should subtract half of the amount you rounded from the smallest value to evenly distribute the 'excess of your range'.
>
> 2. Make sure no data points lie on the border of two intervals. (Do this by subtracting .5 from a whole number, .05 from data with one decimal point, .005 from data with two decimal points and so on)

d) Create a frequency table using your intervals

Notice that the number one interval ends with, the next interval starts with the same number. This is because the data for a histogram is continuous!!!

Grade Interval	Frequency
39.5 - 49.5	
49.5 - 59.5	
59.5 - 69.5	
69.5 - 79.5	
79.5 - 89.5	
89.5 - 99.5	

e) Create a histogram of the data
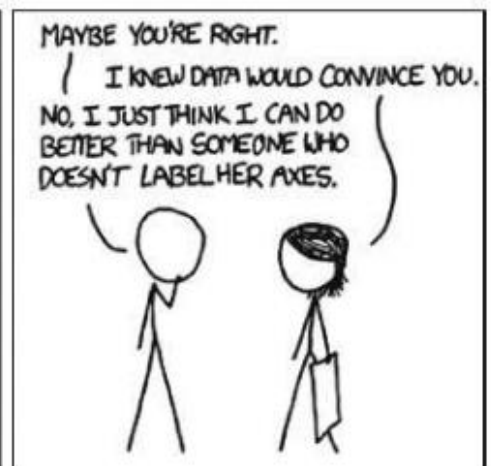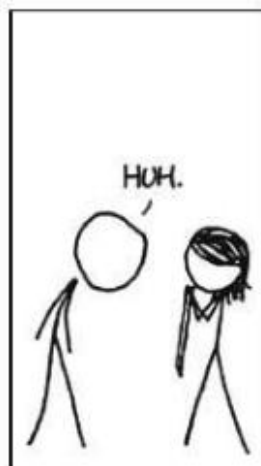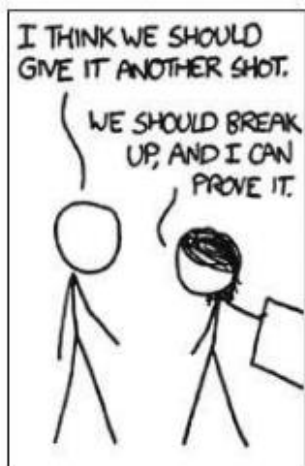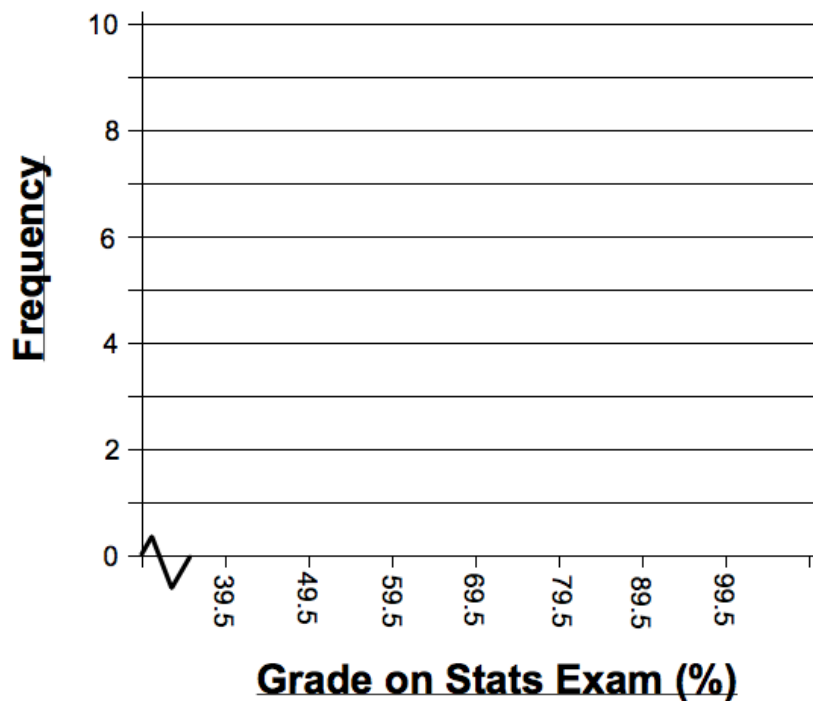
Using Ti 83/84 for Histograms

Below shows how the previous example can be completed using the ti-83/84 calculator.

- input data in to list L1: STAT → EDIT
- turn on statplot: 2nd → y= → ENTER → ENTER
- select histogram
- view graph: GRAPH → ZOOM → ZOOMSTAT

- change class intervals: WINDOW → enter values shown in picture below
- exam class intervals: TRACE → arrow left and right
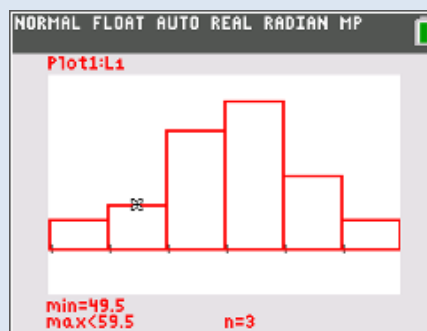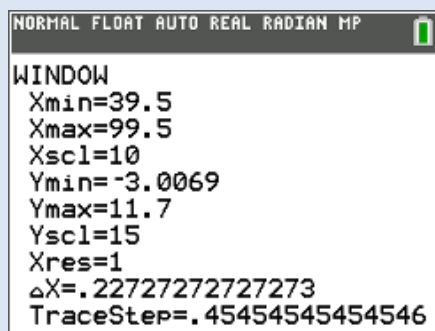
Part 1: Scatterplots Video

Let's start by watching a video on scatterplots:

http://www.learner.org/courses/againstallodds/unitpages/unit10.html

Answer the following questions while watching the movie:

1. What does a scatterplot show about the relationship between the number of powerboats registered in Florida and the number of manatees killed by powerboats?

2. Why is the number of boats plotted on the horizontal axis of this scatterplot?

3. What trend would you expect to see in a scatterplot of two variables that have a negative association?

Note: a scatterplot of manatee deaths and the number of powerboat registrations shows a positive association between the two variables. However, the fact that there is a relationship between two variables is not sufficient evidence to prove cause-and-effect linkage. A well-designed randomized experiment in which the researcher imposes some treatment on its subjects to see how they respond is THE ONLY WAY to give good evidence for cause and effect as you will learn in next unit.

Part 2: Scatterplot Basics

Remember that quantitative variables are measured using numerical values. When you have two quantitative variables (_____), you can use a scatterplot to examine the _____ (association) between the two variables.

In many cases, changes in a variable x are thought to "**ex**plain" changes in a second variable y. In such examples, x is called the _____ (or independent) variable and y is called the _____ (or dependent) variable.

A _____ is a plot of observations of quantitative variables x and y as points in the plane. The explanatory variable, if any, is always plotted on the horizontal scale (x-axis) of the scatterplot.

In the video on manatees,

The explanatory variable was:

The response variable was:
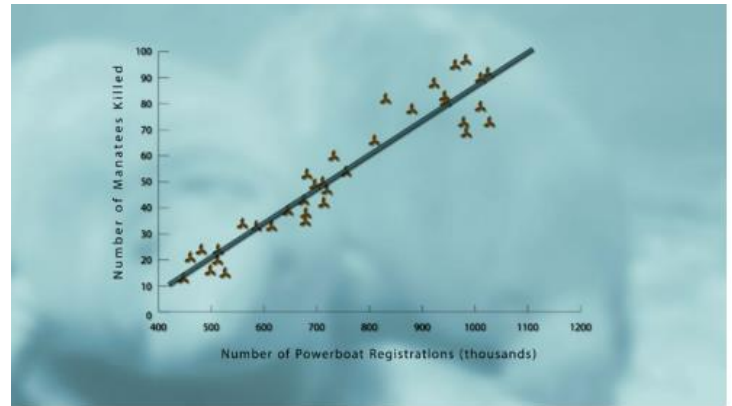
Part 3: Correlation

When analyzing a scatterplot of bivariate data, we look for:

- the overall pattern (linear, curved, random scatter)
- direction (positive, negative)
- strength of the relationship (strong, moderate, weak, no correlation)

Pattern: A scatterplot has _____form when the dots appear to be randomly scattered on either side of a straight line. However, sometimes the data form a curved pattern. In that case, we say the scatterplot has _____.

Linear Form   Nonlinear Form
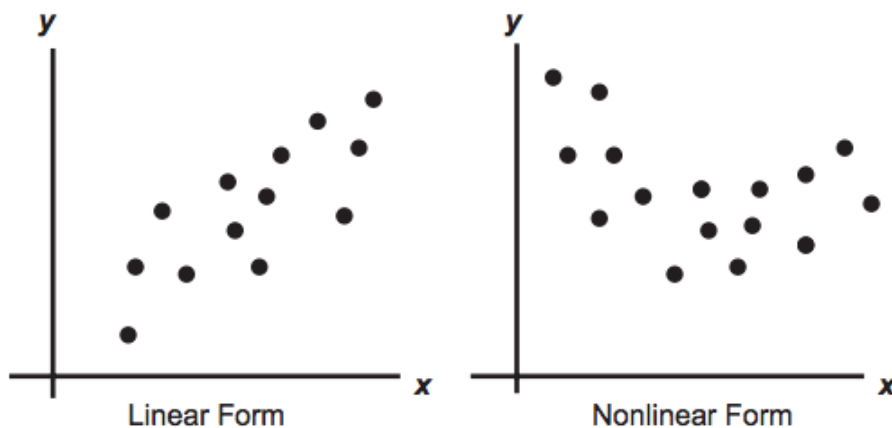
Direction:

Two variables are _____ associated (correlated) when above-average values of one tend to accompany above-average values of the other and below-average values of one tend to accompany below-average values of the other. In a scatterplot a positive association would appear as a pattern of dots in the _____ to the _____.

Two variables are _____ associated (correlated) when above-average values of one accompany below-average values of the other, and vice versa. In a scatterplot a negative association would appear as a pattern of dots in the _____ to the _____.

Positive Association   Negative Association
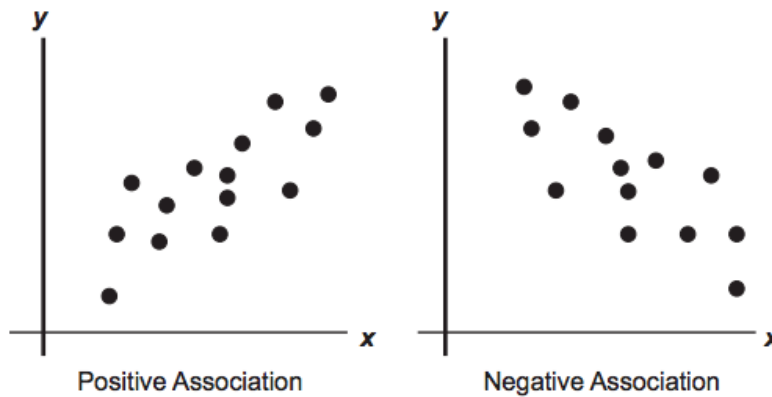
Strength:

Correlations can also be _____ or _____ depending on how close together or spread out the points on the graph are. If there seems to be no trend in the data, we say that there is

_____.

❶   ❷   ❸
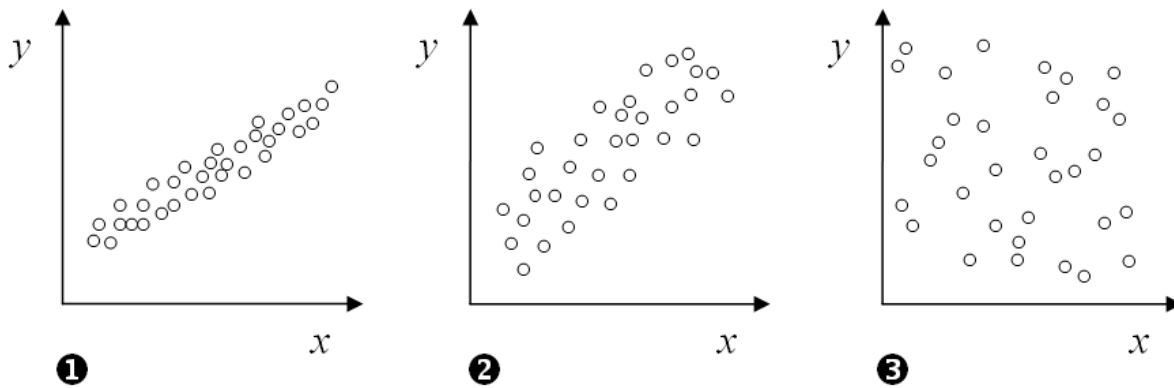
Graph 1 shows a _____ correlation than graph 2 because the points are more closely clustered together.

Graph 3 shows _____.

Time to check your understanding of what we have covered so far today.

Example 1: Circle the independent (explanatory) variable in each pair of variables

Height	**vs.**	Stride Length
Exam Score	**vs.**	Study Time
Smoking	**vs.**	Cancer Rates
Absences	**vs.**	Final Grade
Annual Income	**vs.**	Age

Example 2: Identify the type of correlation for each scatterplot. Identify the pattern, direction, and strength.
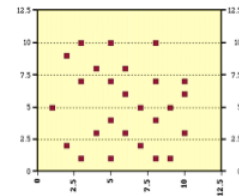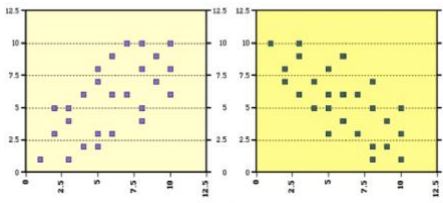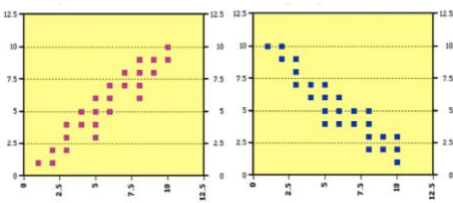
Example 3: Avengers: Infinity War had a successful opening weekend by earning more than 250 million dollars. In opening weekends, a movie's opening gross income is a way of predicting the movies eventual success. Can you predict a movie's total gross income from its opening weekend gross income?

Movie	Opening	Total Gross	Release Date
The Martian	$54,308,575	$228,433,663	10/2/15
Star Trek Beyond	$59,253,211	$158,848,340	7/22/16
LEGO Batman	$53,003,468	$175,750,384	2/10/17
Spider Man: Homecoming	$117,027,503	$334,201,140	7/7/17
Pirates of the Caribbean: Dead Men Tell no Tales	$62,983,253	$172,558,876	5/26/17
Fantastic Beasts and Where to Find Them	$74,403,387	$234,037,575	11/18/16
Coco	$50,802,605	$207,389,121	2/19/18
The Jungle Book	$103,261,464	$364,001,123	4/15/16
Frozen	$67,391,326	$400,738,009	11/27/13
Avengers: Age of Ultron	$191,271,109	$459,005,868	5/1/15
Avatar	$77,025,481	$749,766,139	12/18/09
Star Wars: The Last Jedi	$220,009,584	$618,199,339	12/15/17
Wonder Woman	$103,251,471	$412,563,408	6/2/17
Black Panther	$202,003,951	$688,796,094	2/16/18
Avengers: Infinity War	$257,698,183		4/27/18

a) Make a scatter plot of the data.
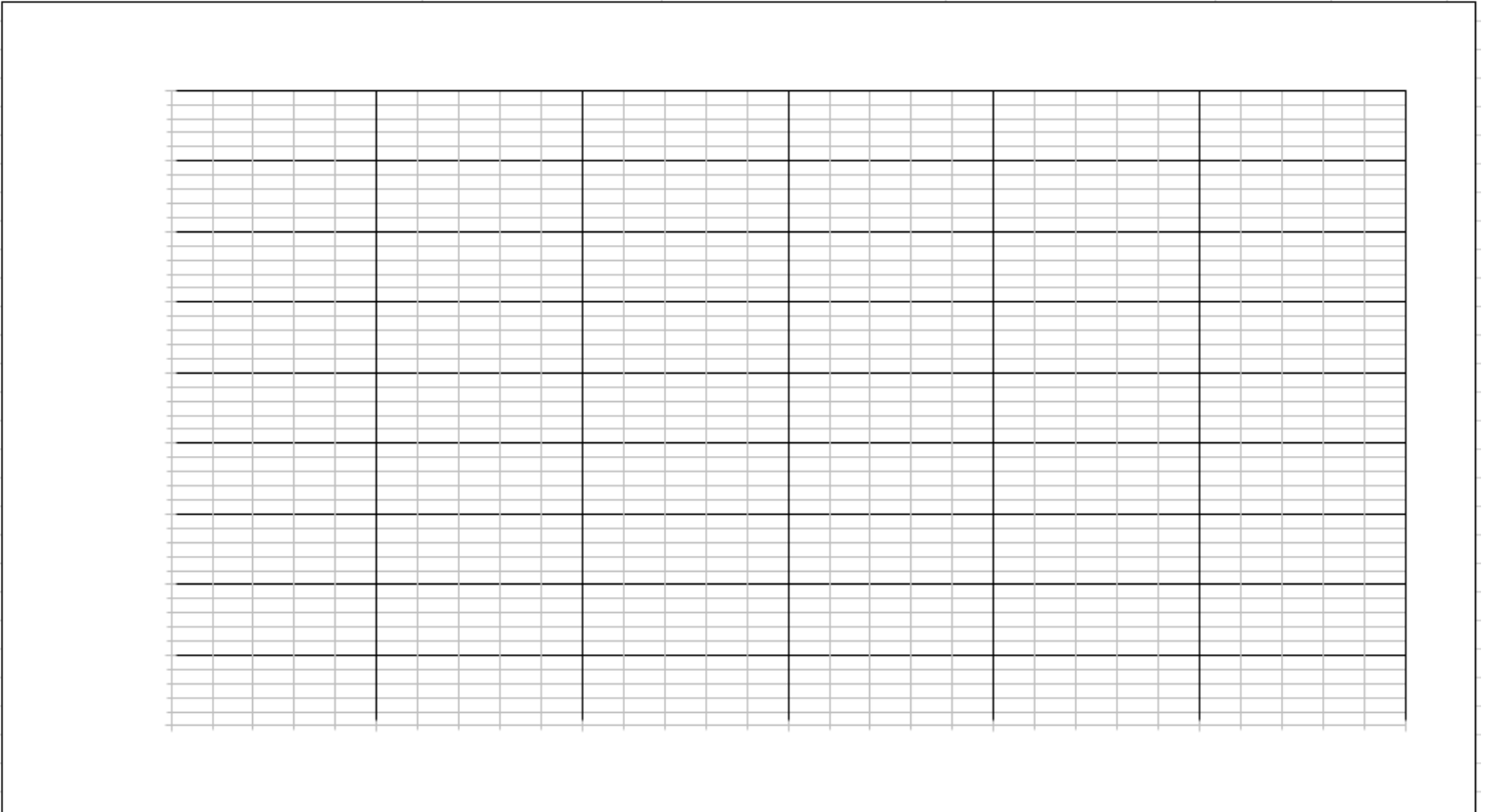
b) Describe the trend in the data

c) Use a trend line to make a prediction for the total gross revenue of Avengers: Infinity War.

The actual gross revenue so far for Avengers: Infinity War is _____
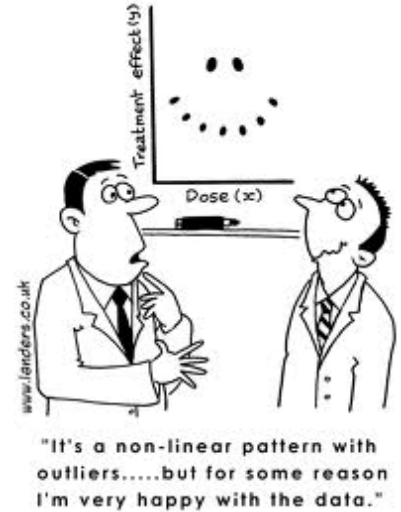
Last class, you learned that by examining a scatter plot, you can see whether the relationship between two variables is strong or weak, positive or negative, linear or non-linear.

In this lesson, you will use technology that will allow you to quantify the linear correlation between two quantitative variables. We will be looking at four main statistics to describe the correlation:

1. The correlation coefficient (r)
2. The coefficient of determination (r^2)
3. Regression line $\hat{y} = a + bx$
4. Residual values (observed y – predicted y)

"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

Part 1: The Correlation Coefficient r

The correlation coefficient, r, is a number between _____ and _____ that is an indicator of both the strength and direction of a _____ relationship between two _____ variables. A value of r = 0 indicates no correlation, while r = 1 or r = -1 indicates a perfect positive or negative correlation.

http://guessthecorrelation.com/
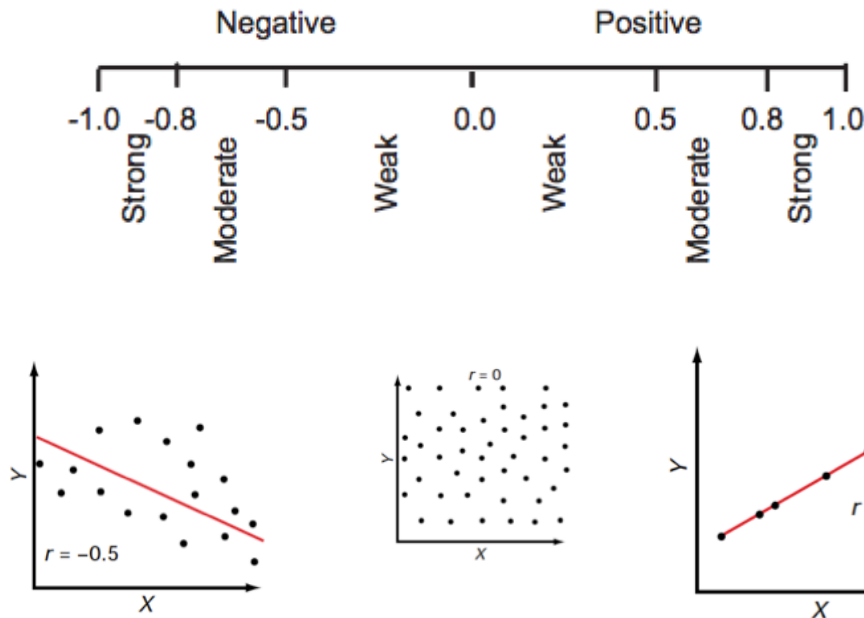
Part 2: The Coefficient of Determination r^2

The coefficient of determination r^2, is a number between _____ and _____ that is useful because it gives the _____ of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph.

The coefficient of determination is a measure of how well the regression line (line of best fit) represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The _____ the line is away from the points, the less it is able to explain.

For example, if $r = 0.922$, then $r^2 = 0.850$, which means that _____of the total variation in y can be explained by the _____between x and y (as described by the regression equation). The other _____of the total variation in y remains unexplained.

Part 3: Regression Line (Line of Best Fit)

A regression line is a line that describes how a dependent (response) variable y changes as an independent (explanatory) variable x changes. We often use a regression line to _____the value of y for a given value of x.

The equation is of the form $\hat{y} = a + bx$

In this equation,

- $\hat{y}$ is the predicted value of the dependent variable y for a given value of x

- b is the slope, the amount by which y is predicted to change when x increases by one unit

- a is the y-intercept, the predicted value of y when $x = 0$

Example 1: The equation of the regression line for the scatterplot shown to the right is $\widehat{price} = 38257 - 0.1629(miles\ driven)$. Identify the slope and y-intercept of the regression line. Interpret each value in context.
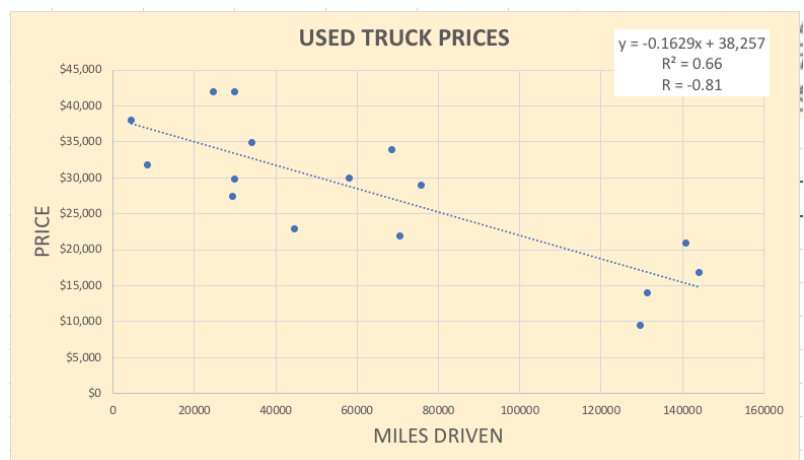
Part 4: Residual Values

A residual is the difference between an observed value of y and the value predicted by the regression line ($\hat{y}$). The residual value tells us how far off the linear regression's prediction is at a given point.

Residual = observed y – predicted y
$\qquad = y - \hat{y}$

© 2007 Thomson Higher Education
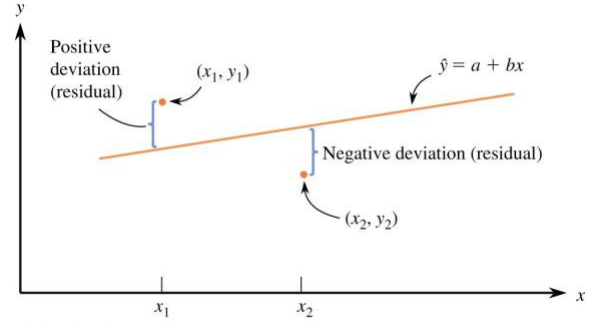
Example 2: Using the regression equation from example 1, find and interpret the residual for a truck that had 70583 miles driven and a price of $21994.

Solution:
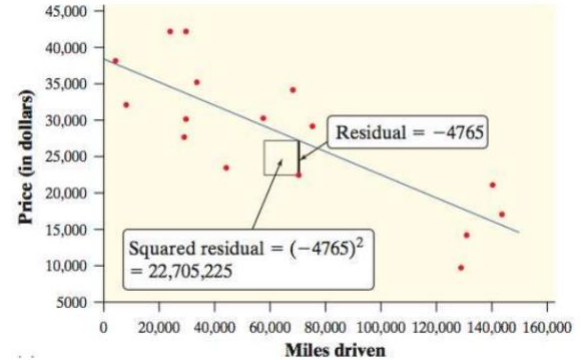
Note: If the regression model is a good fit, the residuals should be fairly _____, and there should be no noticeable pattern. _____ residuals or a _____are indicators that another model may be more appropriate.

Example 3: Sketch the residual plot for the following graph and comment about what it tells you.
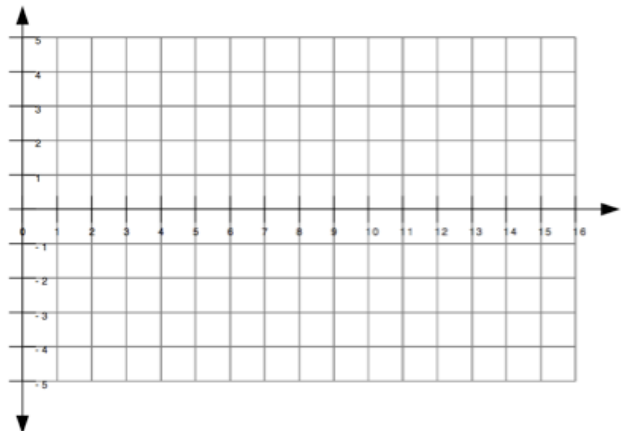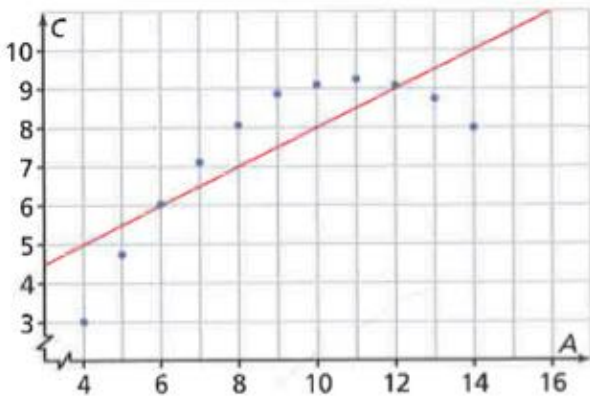
Part 5: Linear Regression Using a Ti-83/84 Calculator

Example 4: Archaeopteryx is an extinct beast having feathers like a bird but teeth and a tail like a reptile. Only six fossil specimens are known. Because these specimens differ greatly in size, some scientists think they are different species rather than individuals from the same species. If the specimens belong to the same species and differ in size because some are younger than others, there should be a positive linear relationship between the lengths of a pair of bones from all individuals. An outlier from this relationship would suggest a different species. Here are data on the lengths in centimeters of the femur and the humerus for the five specimens that preserve both bones.

Femur (x)	38	56	59	64	74
Humerus (y)	41	63	70	72	84

a) Make a scatterplot of the data

- Turn on diagnostics: 2nd → 0 → diagnosticON → ENTER
- Input data in to L1 and L2: STAT → ENTER
- Turn on statplot: 2nd → y= → ENTER → ON (make sure scatter plot is chosen)
- View graph: GRAPH → ZOOM → ZOOMSTAT

b) Find the equation of the regression line and interpret the slope and y-intercept in context.

- STAT → CALC → LinReg (a+bx) → xlist: L1 → ylist: L2 → store RegEQ: Y1 → CALCULATE

c) Find and interpret correlation coefficient, r.

d) Find the coefficient of determination, r^2. Interpret it in the context of this data.

e) Use your equation to predict the humerus length for a femur that is 85 cm.

- VARS → Y-VARS → Y1 → (85) → ENTER

```
NORMAL FLOAT AUTO REAL RADIAN MP
Y₁(85)
                            98.07692308
■
```
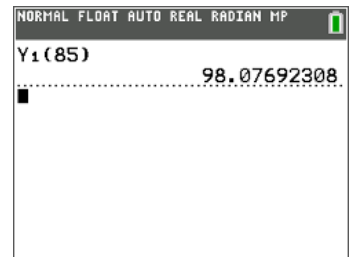
f) Calculate the residual values and analyze the residual plot

- Put residual values in to L3: STAT → ENTER → scroll to highlight L3 → ENTER → 2nd STAT → RESID → ENTER
- View residual plot: turn on plot 2 → Ylist: resid → GRAPH → ZOOMSTAT

```
NORMAL FLOAT AUTO REAL RADIAN MP
L1    L2    L3      L4     L5      3
38    41    -.8226  ------ ------
56    63    -.3668
59    70    3.0425
64    72    -.942
74    84    -.911
------ ------ ------

L₃(1)= -.822617680826
```

```
NORMAL FLOAT AUTO REAL RADIAN MP
```
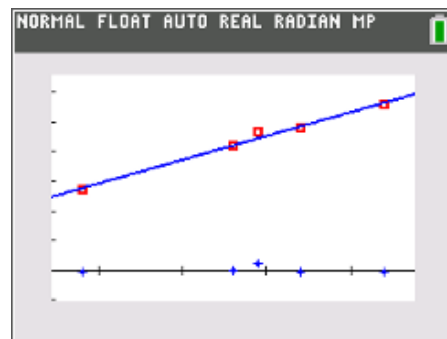
Phrases to Use in Your Answers

Underlined words/phrases or blanks indicate context is needed.

regression: interpretation, in context, of

1. r – positive or negative, weak or strong linear correlation between explanatory variable and response variable

2. r^2 – about x percent of the variation in the response variable can be explained by the approximate linear relationship with the explanatory variable.

3. **slope** – for every 1 unit increase in the explanatory variable, our model predicts an average increase of y units in the response variable.

4. **y-intercept** – at an explanatory variable value of 0 units, our model predicts a response variable value of y units.
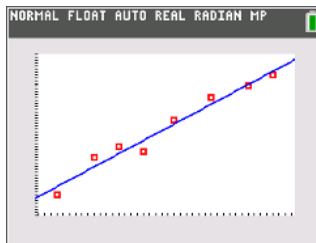
Part 1: Linear Regression Using Technology Practice

This table shows data for the full-time employees of a small company.

| Age (years)	Annual Income ($000)
33	33
25	31
19	18
44	52
50	56
54	60
38	44
29	35

a) Generate a scatterplot of the data.

b) Perform a linear regression and state the equation of the line of best fit. Explain what the slope and y-intercept mean in context.
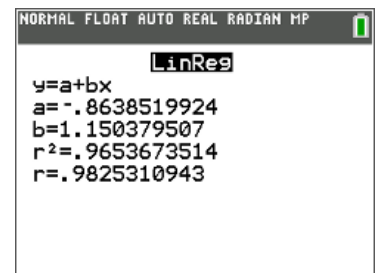
c) What is the correlation coefficient? What does this tell you about the relationship between age and annual income?

d) What is the coefficient of determination? What does it mean?

e) Use the line of best fit to predict the income for a 65 year old employee.

```
NORMAL FLOAT AUTO REAL RADIAN MP
Y₁(65)
                          73.91081594
■
```
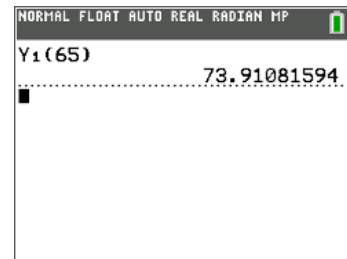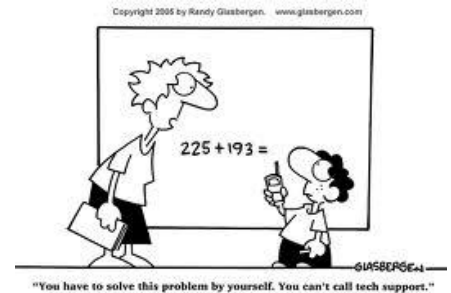
f) Find the residual values. What do they tell you about the correlation between the two variables?

```
NORMAL FLOAT AUTO REAL RADIAN MP
L1    L2    L3      L4      L5     ⌐
33    33    -4.099  ------  ------
25    31    3.1044
19    18    -2.993
44    52    2.2472
50    56    -.6551
54    60    -1.257
38    44    1.1494
29    35    2.5028
------ ------ ------
L3={-4.098671726755,3.10436
```

```
NORMAL FLOAT AUTO REAL RADIAN MP
```
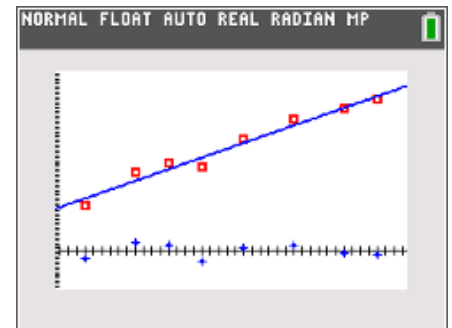

Part 2: Linear Regression by Hand

Example: The following table lists the mathematics of data management marks and grade 12 averages for a small group of students. Start by completing filling in the missing cells. You will need these values to calculate the correlation coefficient and equation of the line of best fit.

Copyright 2005 by Randy Glasbergen.   www.glasbergen.com

225 + 193 =

GLASBERGEN

"You have to solve this problem by yourself. You can't call tech support."

MDM4U Mark (x)	Grade 12 Average (y)	x^2	y^2	xy
74	77	5476		
81	87	6561	7569	7047
66	68		4624	4488
53	67			3551
92	85	8464		7820
45	55	2025	3025	
80	76		5776	
$\sum x =$	$\sum y =$	$\sum x^2 =$	$\sum y^2 =$	$\sum xy =$

a) Determine the equation of the least squares regression line (line of best fit)

b) Calculate the correlation coefficient by hand

The Media

- The media are major users of data. In addressing issues and presenting points of view, the media rely on information based on data

- One of the main purposes of the media is to inform the general public about world events in as an objective manner as possible

- However, the media may sometimes provide misleading or false impressions to sway the public

- An important reason to study statistics is to understand how information is represented or misrepresented

"I can prove it or disprove it! What do you want me to do?"

Part 1: Warm-up

Democrats say that they have won 60% of recent elections, however, Republicans say that they have won 62.5% of the most recent elections.

What is going on? Who do you think is lying?

Lets examine the real statistics.

2008 - Obama - Democrat
2004 - Bush - Republican
2000 - Bush - Republican
1996 - Clinton - Democrat
1992 - Clinton - Democrat
1988 - Bush - Republican
1984 - Reagan - Republican
1980 - Reagan - Republican

So, who was lying?

Other common ways data can be misrepresented:

1. **Data not displayed properly**
   a. Truncated y-axis
   b. Area principle violated
   c. Missing axis labels
   d. Improper scale

2. **Sample size is too small**

3. **Insufficient information**

4. **Sample is not representative of the population**

Part 2: Data not Displayed Properly

Example 1: When you purchase a new vehicle, its value drops dramatically the moment it is driven off the car dealer's lot, and then continues to drop each year thereafter. A graph is used to show this change in value over time. It is possible to communicate different messages using the same data by changing the vertical scale.

Graph A: This graph shows the car's value go from $9000 after 2 years to $1000 after 10 years.
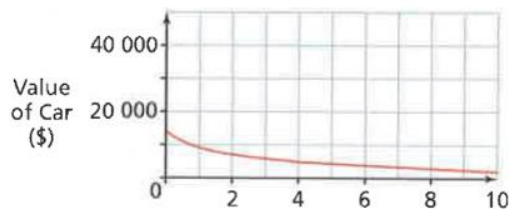
Graph B: This graph also shows the car's value go from $9000 after 2 years to $1000 after 10 years.
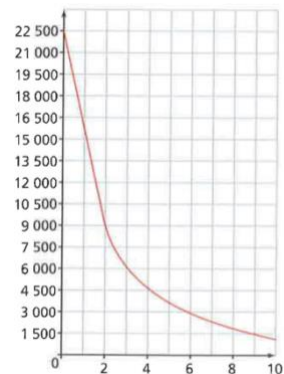
a) Look quickly at each graph. What impression does graph A give you about the change in value of the car compared to graph B?

b) Once you look more careful at both graphs, how does your impression change? What information changed your first impression of the graphs?

Example 2: How did FOX news misrepresent this data?

This is an example of a _____.

Looks like the percentage changed a lot from "now" to Jan 1, 2013. But examining closely, you can see that *the minimum point on the vertical axis is 34% instead of 0*. That's what made it misleading. Fox News exaggerated the percentage just to serve the purpose of pushing Bush's tax cut renewal. This is called "truncating the y-axis".
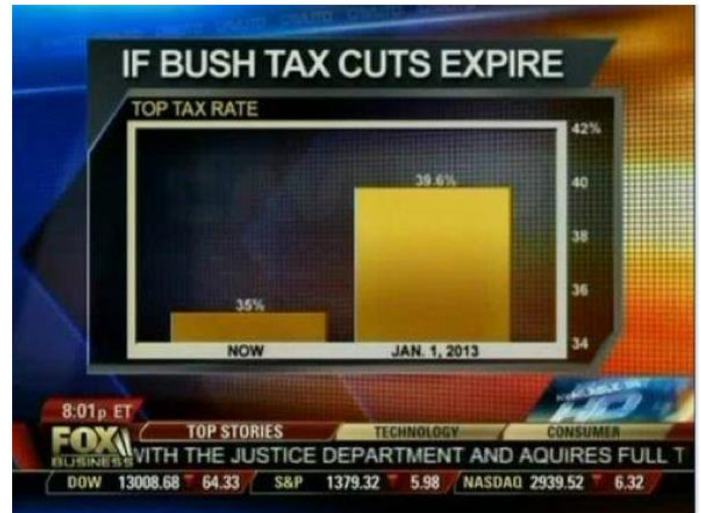

This is what the real percentage looks like:
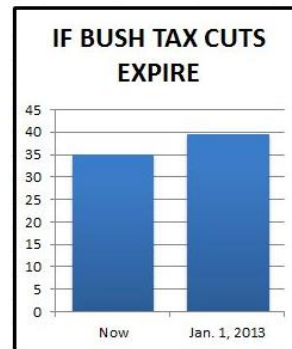
Example 3: The following graph shows the number of people on board the Titanic for each class. How does this graph misrepresent the data?

Although the lengths of the ships are accurate, our eyes respond to the _____ of the pictures. There are about _____ times as many crew members on the ship as first class passengers but the picture of the ship for crew members has an area about _____ times larger than the first class ship.

The area principle says that the _____ occupied by a part of the graph should correspond to the _____ of the value it represents.
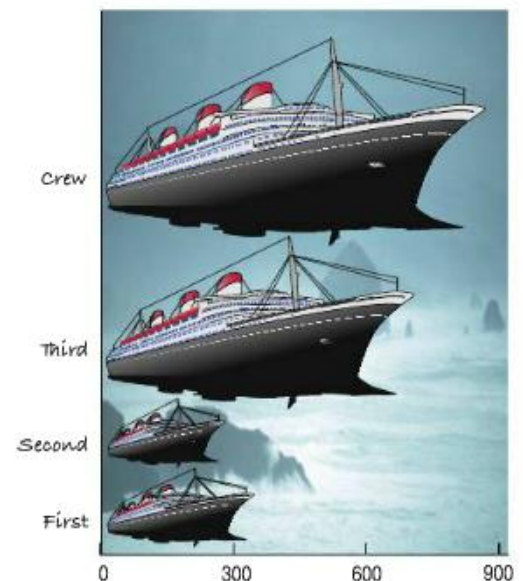
Part 3: Sample Size is Too Small

Example 4: A manager wants to know if a new aptitude test accurately predicts employee productivity. The manager has all 30 current employees write the test and then compares their scores to their productivities as measured in the most recent performance reviews. The data is ordered alphabetically by employee surname. In order to simplify the calculations, the manager selects a systematic sample using every seventh employee. Based on this sample, the manager concludes that the company should hire only applicants who do well on the aptitude test. Determine whether the manager's analysis is valid.
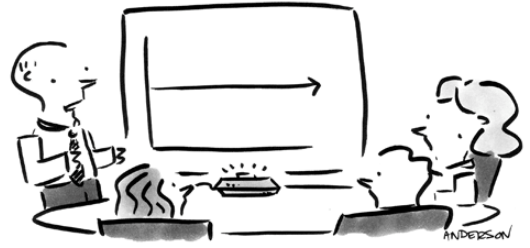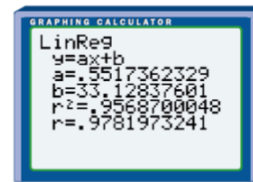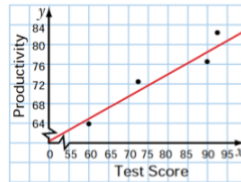
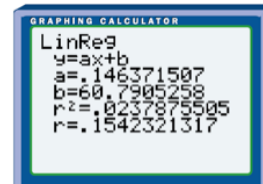| Test Score	Productivity
98	78
57	81
82	83
76	44
65	62
72	89
91	85
87	71
81	76
39	71
50	66
75	90
71	48
89	80
82	83
95	72
56	72
71	90
68	74
77	51
59	65
83	47
75	91
66	77
48	63
61	58
78	55
70	73
68	75
64	69

Based on the linear regression of the systematics sample, what would you conclude?


```
GRAPHING CALCULATOR
LinReg
 y=ax+b
 a=.5517362329
 b=33.12837601
 r²=.9568700048
 r=.9781973241
```
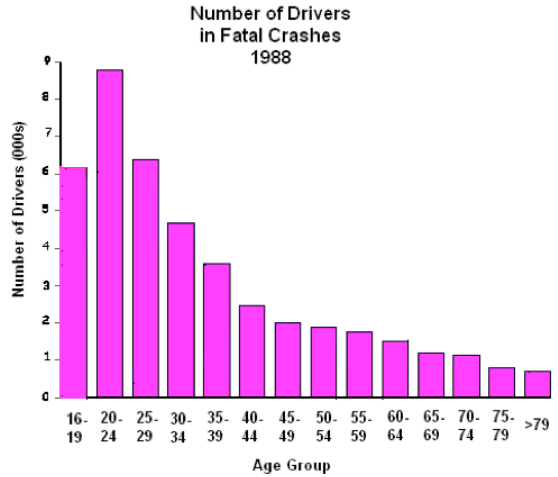
Based on the linear regression of the raw data, do you think the sample is a good representation of the population?


```
GRAPHING CALCULATOR
LinReg
 y=ax+b
 a=.146371507
 b=60.7905258
 r²=.0237875505
 r=.1542321317
```

Part 4: Insufficient Information

Example 5: What does this graph tell you about the ability of drivers as they age?


Number of Drivers
in Fatal Crashes
1988

Graph is based on data from this study: Williams, Allan F., Ph.D., and Oliver Carston, Ph.D., "Driver Age and Crash Involvement," Am J Public Health 1989; 79: 326-327.

What does this graph tell you about the ability of drivers as they age?
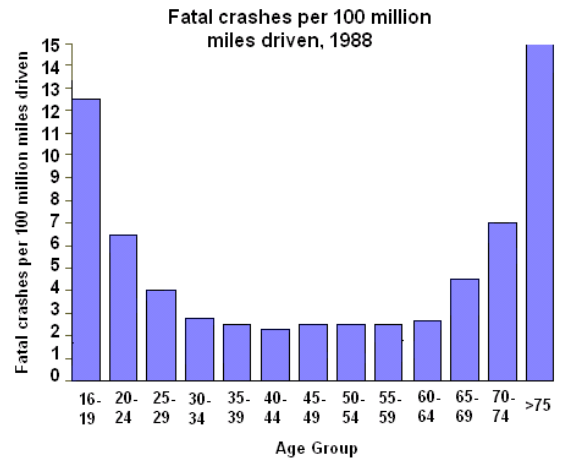
Fatal crashes per 100 million miles driven, 1988

Graph is based on data from this study: Williams, Allan F., Ph.D., and Oliver Carston, Ph.D., "Driver Age and Crash Involvement," Am J Public Health 1989; 79: 326-327.
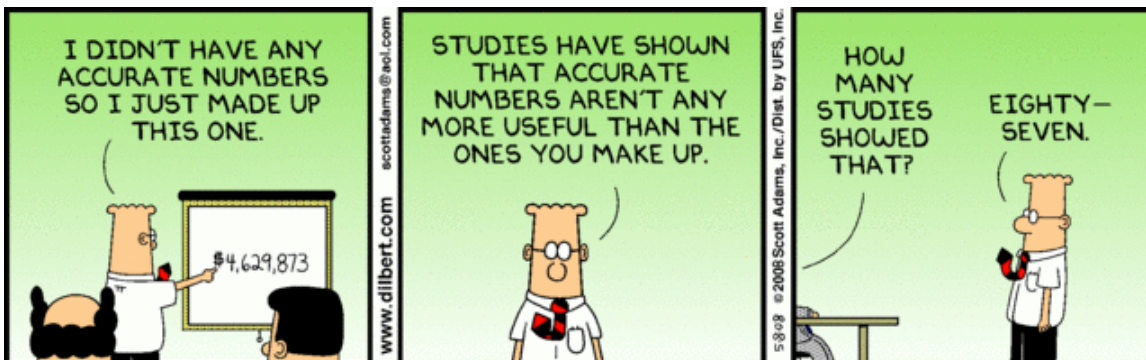
Part 4: Sample is not Representative of the Population

When reading statistics, look carefully for an indication of how the _____ was chosen. Often, companies will carefully select a sample so that they can inflate their results.