UNIVERSITY OF WATERLOO
Faculty of Mathematics
Waterloo, Ontario

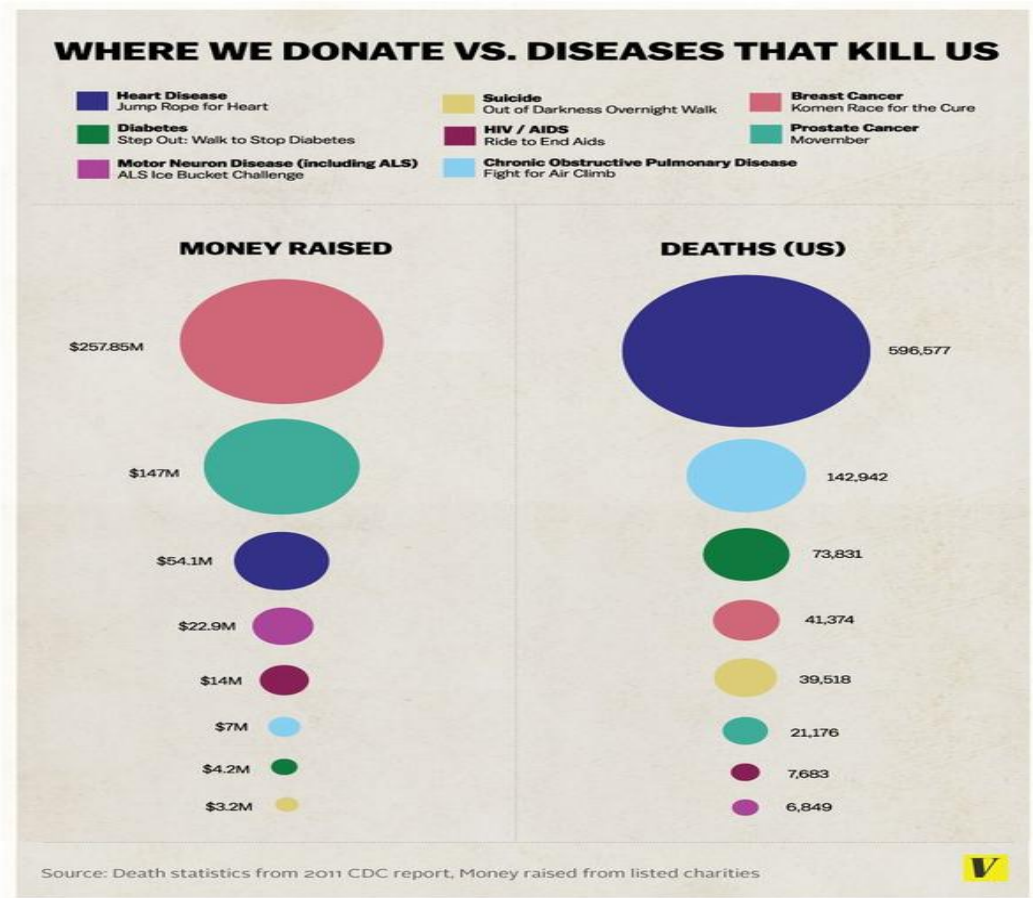
Centre for Education in
Mathematics and Computing

Grade 6 Math Circles

Fall 2014 - October 7/8

Statistics

Statistics (or Stats) is a branch of math that deals with collecting and analyzing data. Statistics can be seen almost anywhere and can be used to solve crimes, predict the outcome of an event, and has even been used to lead Major League Baseball teams to record breaking seasons. Statistics are beneficial as they also bring to the surface issues that people may not have seen before, and allow people in a position of power to make changes according to the data. Here is an interesting piece of data that is presented in an **infographic**. Infographics come in many different forms and are used to provide a clear visual representations of data.



Collecting and Organizing Data

Before any graph can be constructed, there is prior work that has to be completed. This prior work is usually conducting **surveys** to gather the data necessary that we want to represent. After this data has been collected, it is a good idea to organize it inside of a **table**.

Tables

As you can see below, tables make it easy to organize data in a neat way, so it can be referenced conveniently in the future. They also provide us with headings that people can easily identify and exact values to provide us with accurate data.

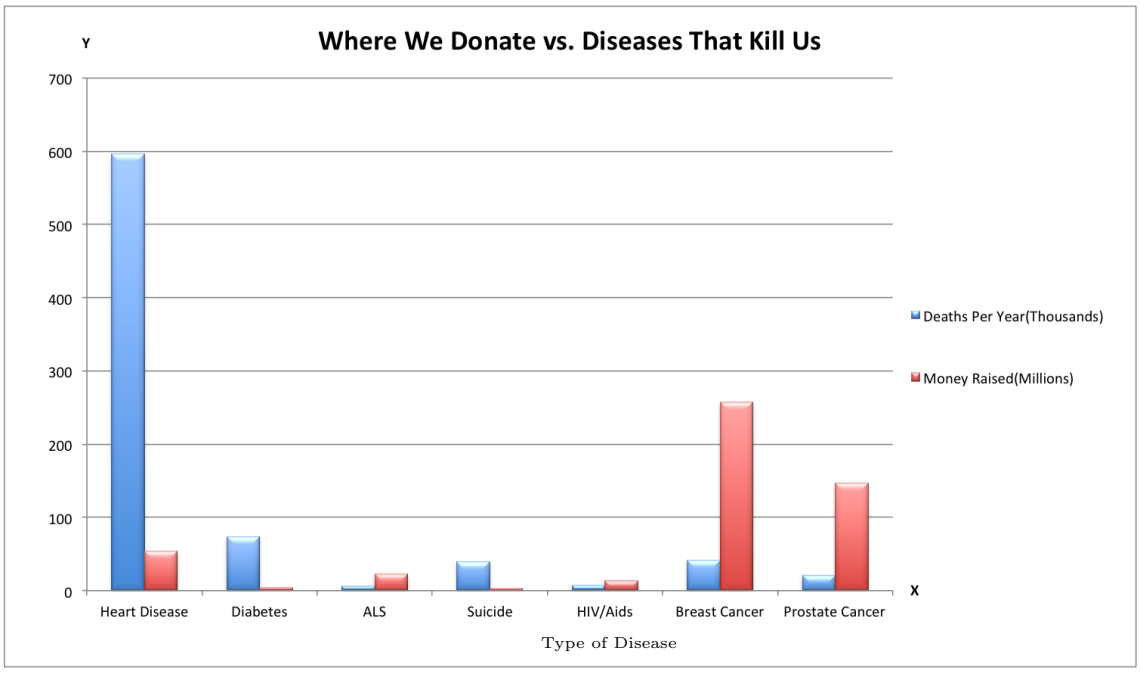
Where We Donate vs. Diseases That Kill Us

Disease	# Deaths Per Year	Money Raised Per Year($)
Heart Disease	596,577	54,100,000
Diabetes	73,831	4,200,000
Breast Cancer	41,374	257,850,000
Suicide	39,518	3,200,000
Prostate Cancer	21,176	147,000,000
HIV/Aids	7,683	14,000,000
ALS	6,849	22,900,000

Now that we have our data nicely placed in this table, we can start constructing different graphs to get a better visual representation of the data we have collected. Notice that for each disease we are given two types of data- number of deaths per year and amount of money raised per year. Let's see what this looks like on a bar graph.

Bar Graph

Bar Graphs are effective when trying to show visual comparisons between data. The relationship between the number of deaths per year and the amount of money raised per year, for each respective disease, is somewhat clear to see in this instance. Also, bar graphs are only useful when analyzing a small number of categories.



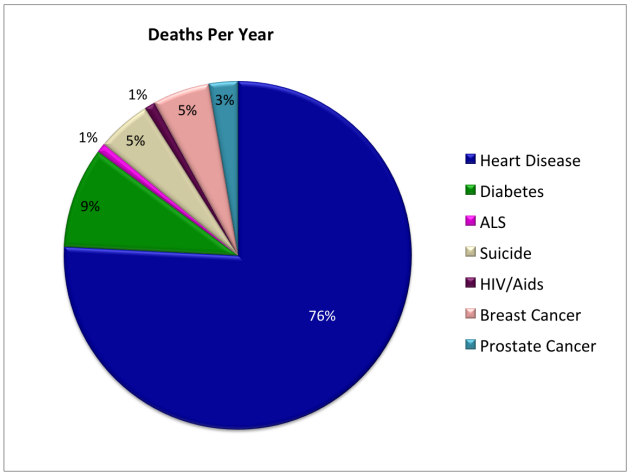
To make a bar graph:

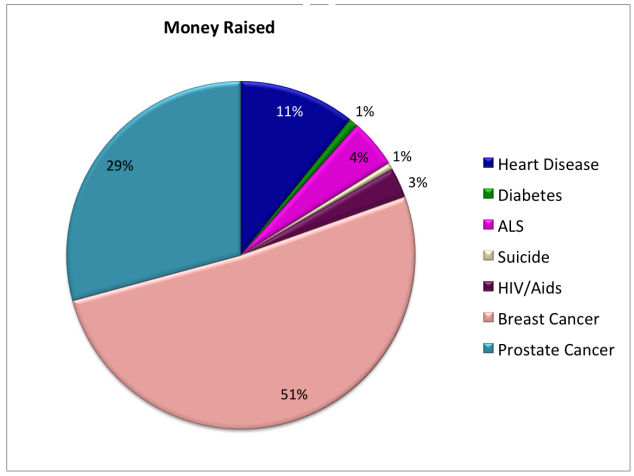
1. Draw two perpendicular lines (each called an **axis**) that connect at a point (it should look like half a rectangle).
2. Label each axis with a title.
3. Label the **vertical axis** or y**-axis** with numbers. Be sure to write the units if they are not already in the title.
4. Label the **horizontal axis** or x**-axis** with categories.
5. Draw a bar above each category. The height of the bar should be the number that corresponds with that category. The width of the bar isn't important, but should be the same for all categories.
6. Give the graph a title.

Pie Graph

Pie graphs are useful when we want to show how much of the total sum of the numerical data each data entry is responsible for. In other words, how much of the pie does each entry make up. They are most useful when there are large inequalities (unfairnesses) between data. Since it is difficult to graph two different types of data in one Pie graph, we will make two separate ones.

Where We Donate vs. Diseases That Kill Us





To draw a pie graph:

1. Start by drawing the full pie (one category of 100%).
2. Calculate the percent for each category (if not already given).

   Here is how:

   (a) Calculate the sum of all entries (add up all the values).

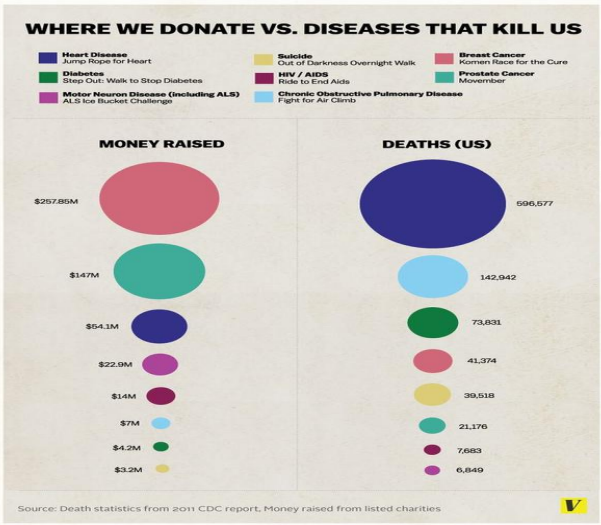
   (b) Divide the value of a category by the sum and multiply by 100. This will give you the percentage of the pie the category's value makes up.

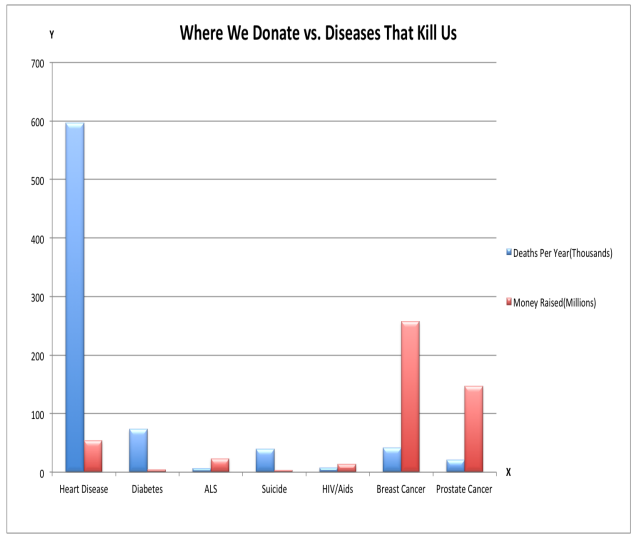
3. Divide the circle according to the calculated percents. We can do this by taking each categories percentage of 360 degrees (there are 360 degrees in a circle), and use a protractor to divide the circle accurately.

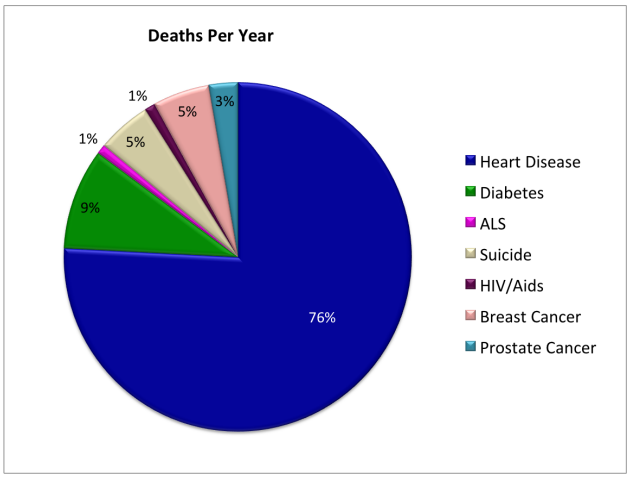
4. Label each section with the category and percent.

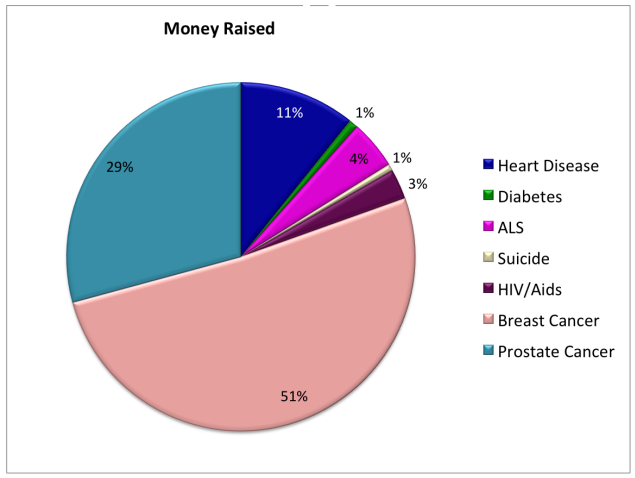
5. Give the graph a title.

Using Graphs Effectively/Manipulation









Once we have collected our data and organized it into a table, we then must use our judgement to decide which graph is going to best represent the data. This will ensure the point we are trying to get across is recognized by whoever is examining the graph. For the data we were dealing with above, which of these 3 graphical representations works best? In this example, we are trying to demonstrate that the number of deaths does not have a strong enough correlation with the amount of money being donated for each respective disease. The bar graph does this, however the difference in units of measure of the two different types of data (thousands vs. millions) throws off the accuracy of the comparison we are examining. But, the pie graph and the infographic both do a good job of showing this miscorrelation, especially with the "Heart Disease" and "Breast Cancer" cases.

Line Graph

Line graphs are similar to bar graphs in that they are drawn on the $x - y$ plane. However, they differ in that a line graph will show the exact coordinate (or (x, y) point) of each piece of data. They are most useful when trying to show trends through time. Here is line graph showing showing the stock price of Blackberry from September 5th to September 11th. Notice there is not a unit in the y-axis, because when dealing with stocks it is assumed that it is a dollar value.

BlackBerry Stock Price Short-term Trend

Data from: http://money.cnn.com/quote/quote.html?symb=BBRY

To make a line graph:

1. Draw two perpendicular lines (each called an **axis**) that connect at a point (it should look like half a rectangle).
2. Label each axis with a title.
3. Label the **vertical axis** or **y-axis** with numbers. Be sure to write the units if they are not already in the title.
4. Label the **horizontal axis** or **x-axis** with categories.
5. Plot each point on the graph using the corresponding x and y coordinates.
6. Connect each point with a line, starting from left to right
7. Give the graph a title.

Mean, Median, Mode

1. The **mean** of a set of numerical data is the average of that data. To find the average (mean) of a set of numerical data, find the sum of the numerical entries and then divide that sum by the amount of entries that made up that sum. Let's look at an example:

2. The **median** of a set of numerical data is the middle entry when the numbers are reordered from least to greatest. If there are an even number of entries, then the median is the average of the two middle numbers.

3. The **mode** of a set of numerical data is the entry that occurs most frequently. If every entry only occurs once, then there is no mode.

Let's look at an example: Find the mean, median, and mode of this set of data.

Student Test Scores

Student	Sam	Marco	Sarah	Nathan	Christine	Nadine
Test Score(%)	75	82	90	64	75	84

Mean: First we find the sum of all the numerical data, then divide this sum by the amount of students, which is 6.

$$75 + 82 + 90 + 64 + 75 + 84 = 470 \text{ and } \frac{470}{6} = 78.3\%$$

Median: First we re-order the numbers from least to greatest. Then, since there are an even amount of numbers, we find the mean of the two middle terms.

$$\{64, 75, 75, 82, 84, 90\} \text{ and } 75 + 82 = 157 \text{ and } \frac{157}{2} = 78.5\%$$

Mode: We see that the number 75 appears the most in this table of data, so the mode is 75%.

Problem Set

"*" indicates challenge question

1. The following table shows the payroll of the top two highest paying teams and the payroll of two of the top two lowest paying teams in the MLB in 2002.

2002 MLB Payroll

Team	Payroll (Millions)
NY Yankees	125
Boston Red Sox	108
Oakland Athletics	40
Tampa Bay Devil Rays	34

Source: `http://www.stevetheump.com/Payrolls.htm`

   (a) What type of graph will best describe how much of total payroll each team is responsible for?

   (b) Calculate the percentages of each team's portion of the total payroll.

   (c) Draw a graph to represent the data in the above table.

2. (a) You are conducting a survey to see how many students in your class prefer each of the four different universities: University of Waterloo, Wilfrid Laurier, University of Toronto, and McMaster. You have collected the votes, but now have to present the following data to your teacher:

   {UofW, WLU, WLU, WLU, UofT, MM, UofW, WLU, UofT, MM, MM, WLU, WLU, UofW, WLU}.

   Create a table to organize the data, then construct a bar graph to represent it.

   (b) Instead, make a pie graph. Which graph does a better job at representing this data? Explain.

3. On the map of Canada below, using the following table, create an infographic that displays the population and size of each province. Using a legend is recommended.

Province/Territory	**Population(Thousands)**	**Size(Km^2)**
Ontario	13538	1076395
Quebec	8155	1542056
British Columbia	4582	944735
Alberta	4025	661848
Manitoba	1265	647797
Saskatchewan	1108	651036
Nova Scotia	941	55284
New Brunswick	756	72908
Newfoundland and Labrador	526	405212
Prince Edward Island	145	5660
Northwest Territories	44	1346106
Yukon	37	482443
Nunavut	36	2093190

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

4. For our bar graph in the example we looked at earlier, "Where We Donate vs. Diseases That Kill Us," how could we have better represented the given data to make a more clear comparison?

5. How can we manipulate data in a pie graph to best show inequality (unfairness) between two pieces of data?

6. Given the following data, which types of graphs would you use to represent it?

   (a) Weather and climate changes for the week.

   (b) Number of votes for each political party.

   (c) Location of each building in University of Waterloo's campus.

7. In Major League Baseball, an often looked at stat is a team's "Cost Per Win." Using the table below, calculate each team's cost per win and make a new table representing the number of wins and the cost per win for each team. Round to the nearest one when calculating the cost per win. Also, comment on the new data you have come up with.

2013 MLB Payroll and Wins

Team	**Payroll ($ in Millions)**	**# Wins**
New York Yankees	229	85
Los Angeles Dodgers	217	92
Oakland Athletics	61	96
Tama Bay Rays	58	92

Source: `http://www.sportingcharts.com/mlb/stats/mlb-cost-per-win-by-season/2013/`

8. In each case, which graph, if any, is most suitable. Why?

   (a) Billy wants to show how his company's stock price has grown over the past 20 years.

   (b) Emily wants to see if hockey, soccer, or football is the most preferred sport in her classroom.

   (c) *McDonald's wants to update its "Nutritional Value" information in a more creative way.

Each Country's Share of CO2 Emissions (Top 7)

Country	Total Emissions (Millions Metric Tons of CO2)	Per Capita Emissions(Tons)
China	6,534	4.91
United States	5,833	19.18
Russia	1,729	12.29
India	1,495	1.31
Japan	1,214	9.54
Germany	829	10.06
Canada	574	17.27

Source: `http://www.ucsusa.org/global_warming/science_and_impacts/science/each-countrys-share-of-co2.html#.VDK-ccZqp_Q`

9. *The table above represents two types of data:

   (a) The total amount of carbon dioxide emissions of each country

   (b) The total amount of carbon dioxide per capita (per household) of each country.

   Draw two separate bar graphs for each type of data. How is it that one could use confuse how responsible each country is for the total amount of C02 emissions?

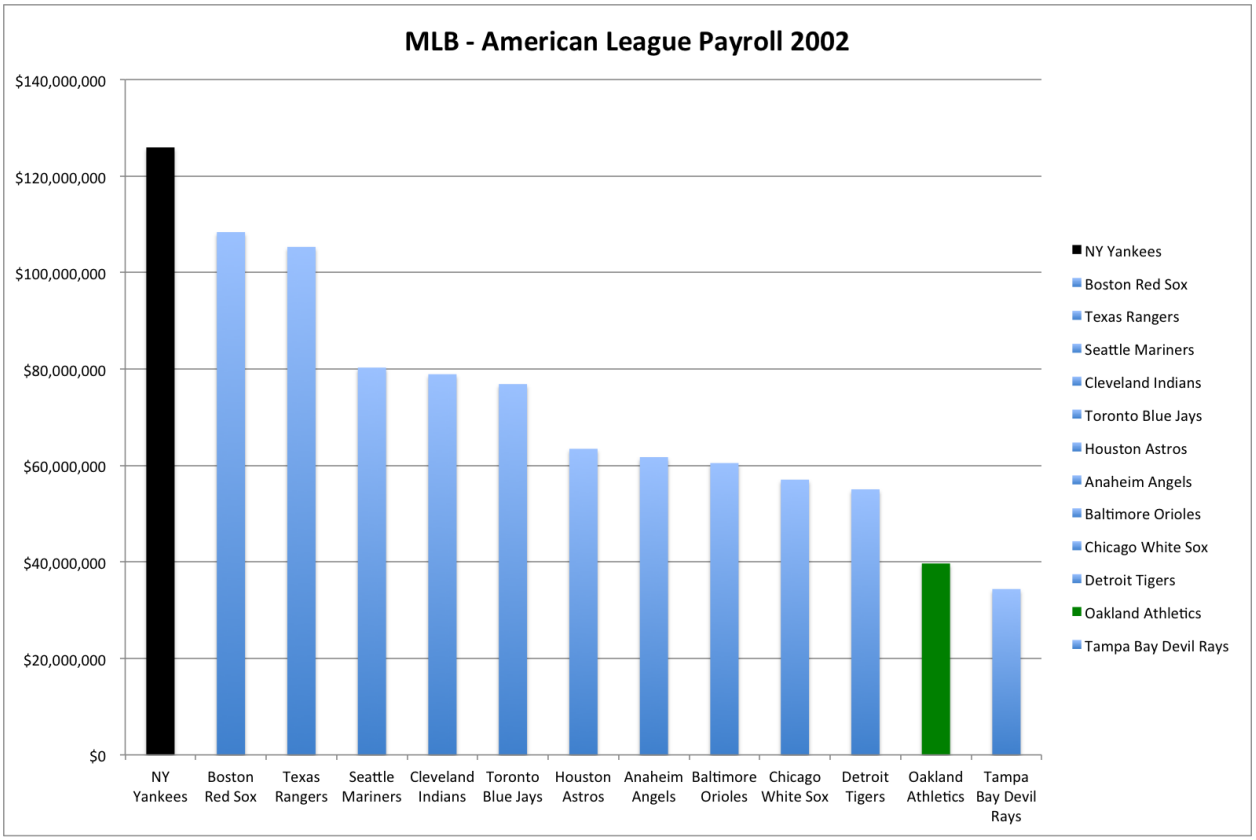
10. Find the mean, median, and mode for each set of numerical data

    (a) 4, 6, 5, 10, 14, 18, 6

    (b) 10, 20, 30, 40, 50, 60

    (c) 5, 5, 5, 5, 5, 5

11. *In the above questions, how would the mean and median change (increase/decrease) if a "10" was added to each set? Attempt to answer this question without performing calculations.



Data from: http://www.stevetheump.com/Payrolls.htm

12. *In 2002, Major League Baseball's "Oakland Athletics" placed first in their division with a record of 103-59, finishing the season with the third highest winning streak in MLB history at 20 games. Without a lot of money to form a team based on talent, the A's depended on **Sabermetrics**- a form of team development that is driven by statistical analysis. One of the key components of Sabermetrics is "OBP" or "on base percentage." Find the mean of each team's OBP using the tables below, and comment on the results while referring to the bar graph above. What does this say about success and statistical analysis in Major League Baseball?

Oakland A's OBP

Player	Hernandez	Hatterberg	Ellis	Tejada	Chavez	Justice	Long	Dye	Durham
OBP	0.313	0.374	0.359	0.354	0.348	0.376	0.298	0.333	0.350

New York Yankees OBP

Player	Posada	Giambi	Soriano	Jeter	Ventura	White	Williams	Mondesi	Johnson
OBP	0.370	0.435	0.332	0.373	0.368	0.288	0.415	0.315	0.347