University of Waterloo
Faculty of Mathematics

Centre for Education in
Mathematics and Computing

Junior Math Circles
March 24, 2010
Data Analysis

Opening Activity

Since the Winter Olympics just finished in British Columbia, here is a table with the total number of medals Canada has won in each of the Winter Olympic games.

Year	# of Medals	Year	# of Medals
1924	1	1972	1
1928	1	1976	3
1932	1	1980	3
1936	7	1984	4
1948	3	1988	5
1952	2	1992	7
1956	3	1994	13
1960	4	1998	15
1964	3	2002	17
1968	3	2006	24

Find the mean, median and mode of the number of medals Canada has won in the Winter Olympics.

Mean(average of all the data): 6

Median(middle number of the data): 3

Mode(most frequent number in the data): 3

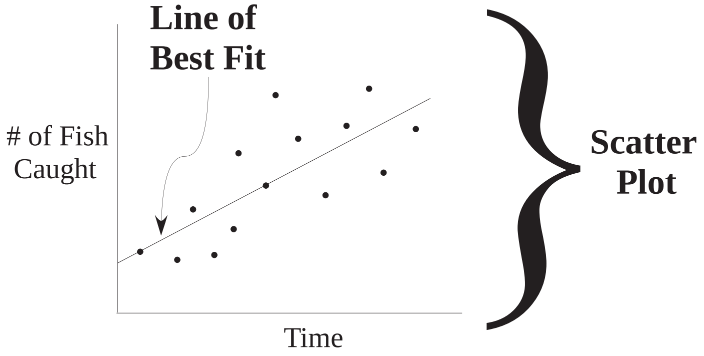
Scatter Plots & Line of Best Fit

When given a set of data, more information can be found about the data other than the mean, median and mode.

Definition: A *scatter plot* is a graph consisiting of a set of points that relate two quantities.

Definition: A *line of best fit* is a straight line that best approximates the trend shown by a group of data points.

An example of a scatter plot and line of best fit are shown below.



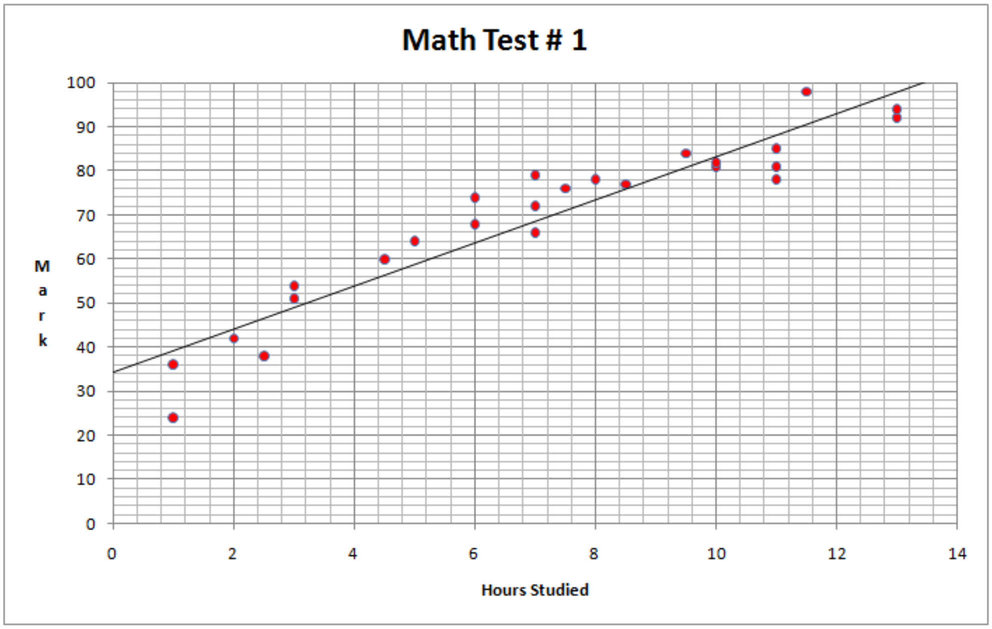
Helpful Hints: (when drawing the line of best fit)

1. Always use a ruler, a transparent one if possible.
2. Pass through as many points as possible.
3. Try to have an equal number of points above and below the line.
4. Use your best judgement.

Note: There are many possible lines of best fit when drawing them by hand. Just use the rules above to draw the best one possible.

The line of best fit in a scatter plot is used to interpolate and extrapolate data. *Interpolation* is predicting a value from a scatter plot that is inside the range of data. *Extrapolation* is predicting a value from a scatter plot that is outside the range of data. The *range* of a set of data is the difference between the greatest value and the lowest value in the data set.

Example Ms. Crescent wanted to find some information on her students' most recent test. She decided to ask each student how long they studied for the test and then graphed their hours studied against their mark on the test. Below is the scatter plot of the data she found.



a) Draw a line of best fit for this data on the scatter plot above.

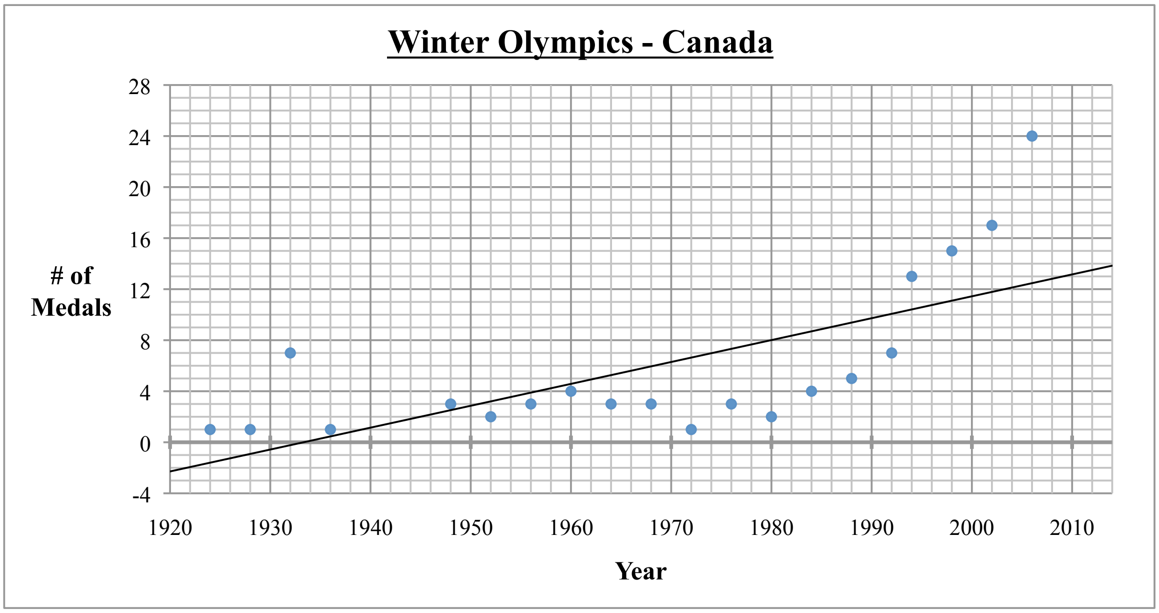
b) Using the line of best fit, estimate:
i) the mark a student could have gotten if he/she studied for 12 hours.
ii) the mark a student could have gotten if he/she didn't study at all.

To find the mark of a student had he/she studied for 12 hours, we want to *interpolate*. We can do this by drawing a vertical line perpendicular to the 12-hour mark on the horizontal axis. Then, when that line hits the best fit line, mark that point. From that point, draw a horizontal line to the vertical axis to determine an accurate mark. Doing this, we can see that a person could have gotten a mark around 93.
To find the mark of a student had he/she not studied at all, we want to *extrapolate* since this is not within our range of data. We can do this by simply extending our line until it passes the vertical axis at 0 to get an accurate measure of the mark. So, we can see that he/she would expect a mark around 34.

c) What appears to be the relationship between the number of hours studied and the mark earned?

We can see that as the number of hours studied increases, in most cases, so does a student's mark. So, we can say that there is a positive relationship - the more hours a student studies, the more likely he/she is to getting a better mark on the test.

Exercise 1 Below is a scatter plot of the data given in the Opening Activity.



a) Draw the line of best fit for the data. What do you notice about the scatter plot and the line of best fit you drew?
It does not represent the data well. A curve of best fit might be a better choice.

b) Using the line of best fit, how many medals do you predict Canada should have won in the Winter Olympics this year?
According to this line of best fit, Canada should have won 13 medals.

Note: Not all scatter plots will have a line of best fit that best represents the data as you can see with the graph above. Some graphs may have a *curve* of best fit or no line at all if the data is too randomly scattered.

Frequency

Definition: The *frequency* of an object in a data set is a count of how often it occurs in the data set. The *relative frequency* of the object is the number of times it occurs in the data set divided by the total number of objects in the data set. This may be represented as a percent, fraction or a decimal.

The mode of a data set is the value with the highest relative frequency. The sum of the relative frequencies of all the objects in the data set will be 1 or 100%.

Example Mrs. Scott asked each of the students in her class how many siblings they had. The following numbers of siblings are the answers. Calculate the relative frequency of each number of siblings.

$$\{4, 2, 2, 1, 0, 1, 1, 1, 3, 1, 1, 1, 0, 0, 0, 2, 1, 1, 6, 1, 1, 1, 2, 1, 0, 1, 1\}$$

We count and find that there are 27 answers in the data set. Of these, 0 appears five times, 1 appears fifteen times, 2 appears four times, and 3, 4 and 6 appear one time each. The relative frequency is:

# of Siblings	**Relative Frequency**	**# of Siblings**	**Relative Frequency**
0 siblings	$5 \div 27 = 0.185$	3 siblings	$1 \div 27 = 0.037$
1 sibling	$15 \div 27 = 0.556$	4 siblings	$1 \div 27 = 0.037$
2 siblings	$4 \div 27 = 0.148$	6 siblings	$1 \div 27 = 0.037$

In the past, relative frequencies of letters in the English language have been used to solve simple codes called substitution ciphers. In a substitution cipher, each letter in the message is encoded by a different one of the 26 letters, to obtain a ciphertext. People looking to crack the code would compare the letter frequencies of the ciphertext against the letter frequencies of the letters in normal English texts.
By counting the number of each letter in various English texts, the following relative frequencies (listed from highest to lowest) of each letter were found:

Letter	**Relative Frequency (%)**	**Letter**	**Relative Frequency (%)**
e	12.702	m	2.406
t	9.056	w	2.360
a	8.167	f	2.228
o	7.507	g	2.015
i	6.966	y	1.974
n	6.749	p	1.929
s	6.327	b	1.492
h	6.094	v	0.978
r	5.987	k	0.772
d	4.253	j	0.153
l	4.025	x	0.150
c	2.782	q	0.095
u	2.758	z	0.074

Exercise 2
You have found the following ciphertext:

PCSJ UJ V OJSJE OBGP JGBKPI VGX V WKOREKU BG HICRI NB AOVRJ CN, VGX C MIVOO UBSJ NIJ HBEOX.
-VERICUJXJM

1. Find the relative frequency of each letter in the ciphertext.

Letter	Relative Frequency (%)	Letter	Relative Frequency (%)
a	$1/80 = 0.0125$	n	$3/80 = 0.0375$
b	$6/80 = 0.075$	o	$7/80 = 0.0875$
c	$5/80 = 0.0625$	p	$3/80 = 0.0375$
d	0	q	0
e	$4/80 = 0.05$	r	$4/80 = 0.05$
f	0	s	$3/80 = 0.0375$
g	$5/80 = 0.0625$	t	0
h	$2/80 = 0.025$	u	$4/80 = 0.05$
i	$6/80 = 0.075$	v	$7/80 = 0.0875$
j	$10/80 = 0.125$	w	$1/80 = 0.0125$
k	$3/80 = 0.0375$	x	$4/80 = 0.05$
l	0	y	0
m	$2/80 = 0.025$	z	0

2. Using these relative frequencies, the table above and some common sense, decode the message, and write the original message in the blanks below:

GIVE ME A LEVER LONG ENOUGH AND A FULCRUM ON WHICH TO PLACE IT, AND I SHALL MOVE THE WORLD.
- ARCHIMEDES

Problem Set

1. What is the mean and median of the data set $\{11, 12, 12, 14, 15, 16\}$?
   What is the mean and median of the set if the number 102 is also included?

2. Amanda scored a mean of 32 points in her first four basketball games. After the fifth game, her mean number of points for all five games was 31. How many points did she score in her fifth game?

3. In a data set of 8 numbers, the average of 5 of the numbers is 18 and the average of the other three numbers is 25. What is the average of all 8 numbers?

4. A six-sided die was rolled 88 times. One was rolled 19 times, two was rolled 13 times, three and six were rolled 15 times each, four was rolled 16 times and five was rolled 10 times. Determine the relative frequencies of each outcome.

5. A set of five *different* positive integers has an average of 11. What is the largest possible number in the set?

6. Below is a table of the number of practices students attended and their race time in a 100-m race.

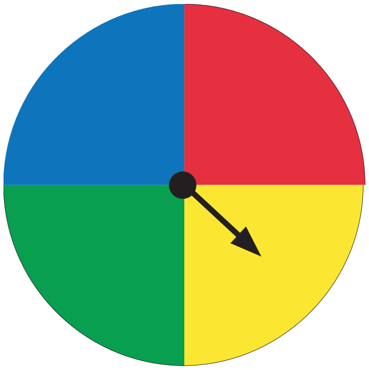
# of Practices	**Race Time**	**# of Practices**	**Race Time**
2	28.6	10	21
2	24.1	11	19.2
3	31.3	11	16.8
4	27.9	14	17.1
6	26	15	17.3
6	24.5	17	19.5
7	23.7	18	18
7	24.3	18	14.8
9	19.8	18	14.8
9	22.2	18	15.2

a) Plot the points on a graph and draw the line of best fit.
b) Using that line, estimate the time it would take if a student attended:
i) 1 practice? ii) 5 practices? iii) 15 practices?
c) Using that line, estimate how many practices a student attended if it took:
i) 25 seconds to finish? ii) 16 seconds to finish?
d) Describe any trends in the data. What kind of relationship does the data seem to have?

7. When is it not appropriate to draw a line of best fit on a scatter plot? Explain your answer.

8. What can be said about a scatter plot with a line of best fit that passes through none of the points?

9. A spinner like the one shown on the right is spun 60 times. It lands on the red section 21 times, the green section 11 times, and the blue section 13 times. What is the relative frequency of the yellow section?



10. A set of five integers has a single mode of 15, a median of 14 and a mean of 13. What are all possible sets that satisfy this condition?

11. Decode the following ciphertext:

    ENWVEX XWZMT DMJL VJN XJQ NUYY HW LJMW AUTZIIJUVEWA HX EOW EOUVBT XJQ AUAV'E AJ EOZV HX EOW JVWT XJQ AUA. TJ EOMJN JDD EOW HJNYUVWT, TZUY ZNZX DMJL EOW TZDW OZMHJM. RZERO EOW EMZAW NUVAT UV XJQM TZUYT. WSIYJMW. AMWZL. AUTRJFWM. -LZMC ENZUV

Answers

1. Mean = 13.33 , Median = 13
   If 102 included: Mean = 26, Median = 14

2. 27

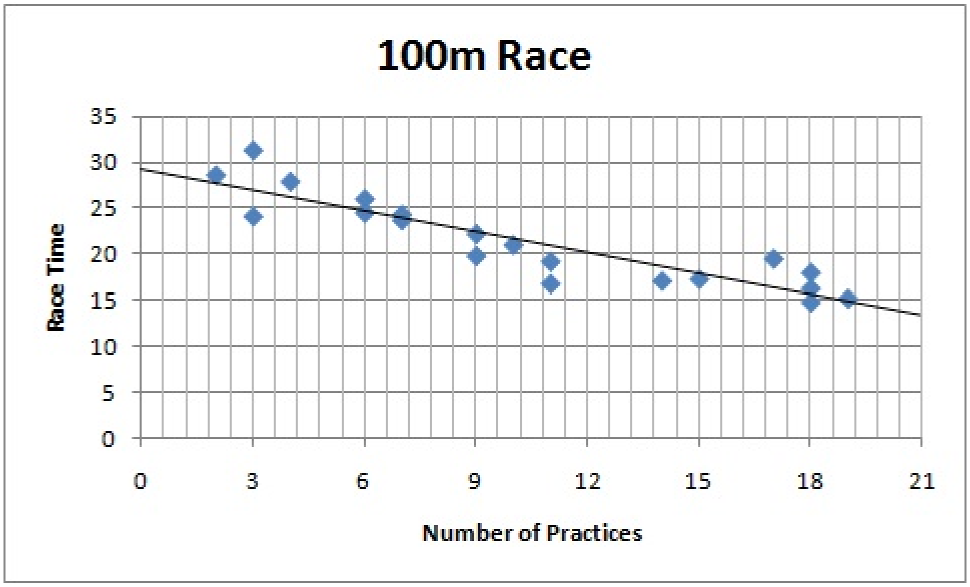
3. 20.625

4.

# on Die	# of Times Rolled	Relative Frequency
1	19	0.216
2	13	0.148
3	15	0.17
4	16	0.182
5	10	0.114
6	15	0.17

5. 45

6. (a)



   (b) i) 28 seconds
   ii) 25 seconds
   iii) 18 seconds

   (c) i) 5 practices
   ii) 17 practices

(d) According to the graph, the more practices the students attend, the better race time they will have. The graph seems to have a linear relationship where the line of best fit fits the data well.

7. It is not appropriate to draw a line of best fit when the points are too scattered and don't seem to have a relationship. Also, it is not appropriate to draw a line of best fit when the data seems to have a non-linear relationship.

8. A line of best fit may come close to all of the data points, but pass through none of them, and be a very good representation of the data. A line of best fit could also pass through some data points, but be very far away from others, and be a bad representation of the data. Therefore, whether or not a line of best fit passes through any data points is not a good indication of how well it fits the data.

9. $\frac{15}{60} = 0.25$

10. All possible sets are:
{8, 13, 14, 15, 15}
{9, 12, 14, 15, 15}
{10, 11, 14, 15, 15}

11. TWENTY YEARS FROM NOW YOU WILL BE MORE DISAPPOINTED BY THE THINGS YOU DIDN'T DO THAN BY THE ONES YOU DID. SO THROW OFF THE BOWLINES, SAIL AWAY FROM THE SAFE HARBOR. CATCH THE TRADE WINDS IN YOUR SAILS. EXPLORE. DREAM. DISCOVER.
- MARK TWAIN