2008 Excellence in Mathematics Contest
Team Project A

CHANDLER-GILBERT COMMUNITY COLLEGE

School Name:

____

Group Members:

____

____

____

____

____

____

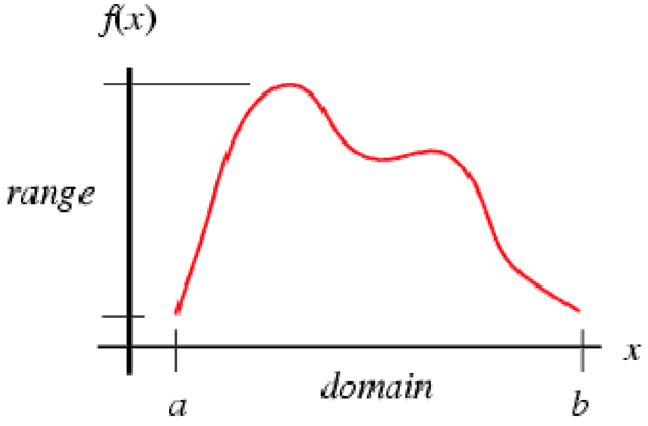
Reference Sheet

Relative Frequency is the ratio of the absolute frequency to the total number of data points in a frequency distribution.

For example, if the Arizona Diamondbacks win 96 out of 162 games, we say that they won $\frac{96}{162}$ or, equivalently, $\frac{16}{27}$ of their games. We can also say that the relative frequency is $\frac{16}{27} \approx 0.59$ or that the Diamondbacks won 59% of their games.

Domain is a term is used to describe the set of values for which a function is defined.

Range is a term used to describe the set of values that a particular function can take on over a given domain.



Coefficient of Determination, r^2 is a value that describes the strength of fit of a linear regression model to a set of data. The closer the value of r^2 is to 1, the stronger the fit.

In **statistics**, the **residual** is the observed value minus the predicted value. The bigger the residual, the poorer the model used to make the prediction.

Scrabble and Mathematical Modeling

How many of each letter should there be and how much should each letter be worth?

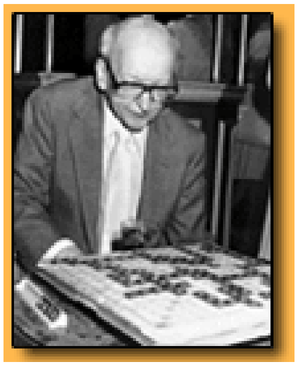
Adapted from Richardson & Gabrosek (2004 NCTM "Mathematics Teacher" Journal)

Eleven students randomly chose a starting point in a newspaper article, internet resource, or book and counted out the next 300 letters. They tallied the number of each of the letters found in their particular selection. The 11 students then pooled the results together. The results are shown in the table and will be used in Activity I.

Table 1

Letter	Total
A	434
B	133
C	175
D	169
E	563
F	147
G	192
H	292
I	307
J	7
K	63
L	186
M	146
N	300
O	326
P	118
Q	35
R	266
S	307
T	417
U	158
V	81
W	102
X	20
Y	103
Z	26
Total	5073

Background



During the Great Depression, an out-of-work architect named Alfred Mosher Butts decided to invent a board game. He did some market research and concluded that games fall into three categories: number games, such as dice and bingo; move games, such as chess and checkers; and word games, such as anagrams.

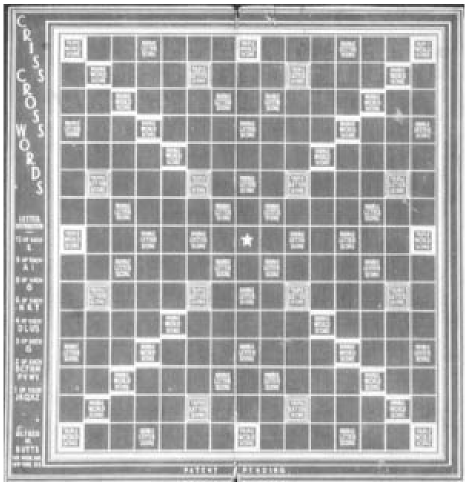
Butts wanted to create a game that combined the vocabulary skills of crossword puzzles and anagrams, with the additional element of chance. The game was originally named Lexico, but Butts eventually decided to call the game "Criss-Cross Words."

How did he do it?

Butts studied the front page of The New York Times to calculate how often each of the 26 letters of the English language was used. He discovered that vowels appear far more often than consonants, with *E* being the most frequently used vowel. After figuring out frequency of use, Butts assigned different point values to each letter and decided how many of each letter would be included in the game. The letter *S* posed a problem. While it's frequently used, Butts decided to include only four *S*'s in the game, hoping to limit the use of plurals. After all, he didn't want the game to be too easy!

Butts got it just right. His basic cryptographic analysis of our language and his original tile distribution have remained valid for almost three generations and for billions of games played.

The boards for the first Criss-Cross Words game were hand drawn with his architectural drafting equipment, reproduced by blueprinting and pasted on folding checkerboards. The tiles were similarly hand-lettered, then glued to quarter-inch balsa and cut to match the squares on the board.



Purpose of Activity 1

The purpose of this activity is to examine the relationship between a letter’s relative frequency in English and the percent of Scrabble tiles for the letter.

Use the following table to record the frequency for the letters of the alphabet in English text using the data from Table 1 on Page 3. The actual percentage of Scrabble tiles containing the letters are recorded except for the letter *L* and *W*. We wish to use this information to examine the relationship between a letter’s relative frequency and its percent of Scrabble tiles.

Percent of Scrabble Tiles

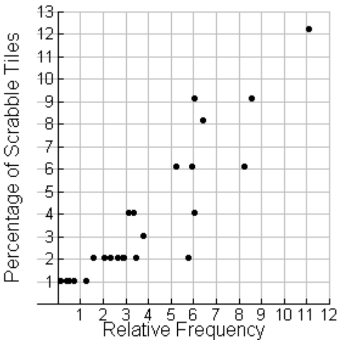
Table 2

Letter	Relative Frequency from Table 1	Percent of Scrabble Tiles	Letter	Relative Frequency from Table 1	Percent of Scrabble Tiles	Letter	Relative Frequency from Table 1	Percent of Scrabble Tiles
A	$\frac{434}{5073} \approx 8.6\%$	9.18	**J**	$\frac{7}{5073} \approx 0.1\%$	1.02	**S**	$\frac{307}{5073} \approx 6.1\%$	4.08
B	$\frac{133}{5073} \approx 2.6\%$	2.04	**K**	$\frac{63}{5073} \approx 1.2\%$	1.02	**T**	$\frac{417}{5073} \approx 8.2\%$	6.12
C	$\frac{175}{5073} \approx 3.4\%$	2.04	**L**	$\frac{186}{5073} \approx 3.7\%$	**?**	**U**	$\frac{158}{5073} \approx 3.1\%$	4.08
D	$\frac{169}{5073} \approx 3.3\%$	4.08	**M**	$\frac{146}{5073} \approx 2.9\%$	2.04	**V**	$\frac{81}{5073} \approx 1.6\%$	2.04
E	$\frac{563}{5073} \approx 11.1\%$	12.24	**N**	$\frac{300}{5073} \approx 5.9\%$	6.12	**W**	$\frac{102}{5073} \approx 2.0\%$	**?**
F	$\frac{147}{5073} \approx 2.9\%$	2.04	**O**	$\frac{326}{5073} \approx 6.4\%$	8.16	**X**	$\frac{20}{5073} \approx 0.4\%$	1.02
G	$\frac{192}{5073} \approx 3.8\%$	3.06	**P**	$\frac{118}{5073} \approx 2.3\%$	2.04	**Y**	$\frac{103}{5073} \approx 2.0\%$	2.04
H	$\frac{292}{5073} \approx 5.8\%$	2.04	**Q**	$\frac{35}{5073} \approx 0.7\%$	1.02	**Z**	$\frac{26}{5073} \approx 0.5\%$	1.02
I	$\frac{307}{5073} \approx 6.1\%$	9.18	**R**	$\frac{266}{5073} \approx 5.2\%$	6.12			

1. Is the association between the percent of scrabble tiles and the relative frequency of a letter positive or negative? Why?

The association between the percent of scrabble tiles and the relative frequency in the English language is positive. This is because the more frequent the letter is seen in the English language, the greater the percent of the Scrabble tiles found in the game. That is, the higher the relative frequency, the higher the percentage of Scrabble tiles for that letter.

2. Using your graphing calculator or Excel, make a scatterplot with the percent of Scrabble tiles for each letter on the vertical axis and the relative frequency of the letter on the horizontal axis. ***Provide a printout of the graph when you turn in this project.***



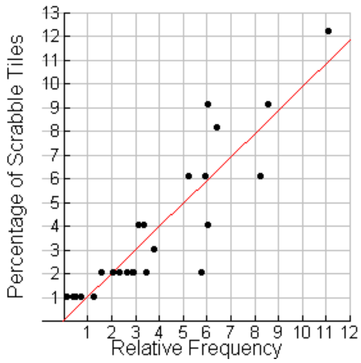
3. Are there any letters whose percent of Scrabble tiles does not follow the pattern for the majority of points? That is, do any outliers exist? Is so, which letters are they?

There do not appear to be any letters whose percent of Scrabble tiles do not follow the pattern for the majority of points. That is, there do not appear to be any outliers.

4. Use the scatterplot to describe the form, strength, and direction of the association between a letter's relative frequency and its percent of Scrabble tiles.

The association between a letter's relative frequency and its percent of Scrabble tiles is roughly linear, moderately strong, and positive.

5. Use your calculator (or Excel) to find the equation of the regression line.

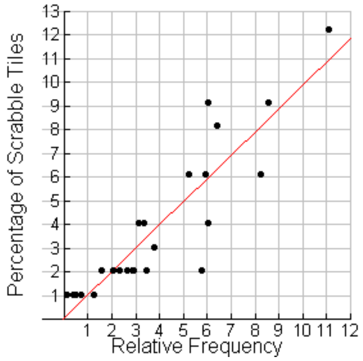


```
Linear Regression (ax+b)
regEQ(x) = .985894x + .035318
r² = .802792
```

6. Interpret the value of the coefficient of determination, r^2.

Answers will vary. We can say that approximately 80% of the error is explained by the model. Students might say that if $r^2 = 1$, the association would be perfectly linear, so with $r^2 = 0.8$, the association is strong (as predicated earlier).

7. Plot the regression line on your scatterplot.



8. Interpret the slope of the regression line in the context of this problem.

The slope of the regression line is 0.986. This says that for each percent increase in the relative frequency of a letter found in the English language, the percent of Scrabble tiles found in the game increases by about 1% (0.986% actually). Students may scale this in some way by saying "for each 10% increase in relative frequency, the percentage of Scrabble tiles increases by 9.9%."

9. Use your regression line to predict the percent of Scrabble tiles for the letters *L* and *W*. Complete the following table.

Table 3

Letter	**Relative Frequency in English Text**	**Actual Percent of Scrabble Tiles**	**Predicted Percent of Scrabble Tiles**	**Residual**
L	3.7%	4.08	3.7%	4.08 – 3.7 = 0.38
W	2.0%	2.01	2.0%	2.01 – 2.0 = 0.1

$\text{Prediction} \approx 0.986(3.7) + 0.035 \approx 3.7$

$\text{Prediction} \approx 0.986(2.0) + 0.035 \approx 2.0$

10. In Scrabble, there are some blank tiles used as “wild cards”. That is, they can represent any letter that the player chooses. According to your regression line model, what percent of the Scrabble tiles should be blank? Explain.

If we say that the relative frequency of “wild cards” in the English language is 0.0%, then the regression model predicts that $\text{Prediction} \approx 0.986(0.0) + 0.035 \approx 0.035$.

This means that 0.035% of the Scrabble tiles should be blank, wild card tiles.

11. Does question 10 provide any information that could be used in describing the domain and range of your regression line model? Explain.

The domain (practical domain, really) might be described as being [0, 9). That is, the relative frequency of a letter in the English language will fall between 0% (i.e. wild card, blank) and about 9%, based on our data.

The range can be described as being [0.035, 10). That is, the percent of Scrabble tiles will fall between 0.035% (wild card tiles) and about 10%.

Purpose of Activity 2

The purpose of this activity is to examine the relationship between a letter's relative frequency in English text and the Scrabble-tile point value of the letter.

In the following table, record the class frequency for the letters of the alphabet in English text. The corresponding Scrabble-tile point values are recorded, except for the letters *C* and *N*.

Scrabble-Tile Point Values

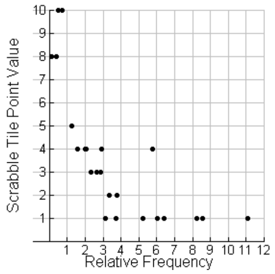
Table 4

Letter	Relative Frequency from Table 1	Scrabble Tile Points	Letter	Relative Frequency from Table 1	Scrabble Tile Points	Letter	Relative Frequency from Table 1	Scrabble Tile Points
A	$\frac{434}{5073} \approx 8.6\%$	1	**J**	$\frac{7}{5073} \approx 0.1\%$	8	**S**	$\frac{307}{5073} \approx 6.1\%$	1
B	$\frac{133}{5073} \approx 2.6\%$	3	**K**	$\frac{63}{5073} \approx 1.2\%$	5	**T**	$\frac{417}{5073} \approx 8.2\%$	1
C	$\frac{175}{5073} \approx 3.4\%$	**?**	**L**	$\frac{186}{5073} \approx 3.7\%$	1	**U**	$\frac{158}{5073} \approx 3.1\%$	1
D	$\frac{169}{5073} \approx 3.3\%$	2	**M**	$\frac{146}{5073} \approx 2.9\%$	3	**V**	$\frac{81}{5073} \approx 1.6\%$	4
E	$\frac{563}{5073} \approx 11.1\%$	1	**N**	$\frac{300}{5073} \approx 5.9\%$	**?**	**W**	$\frac{102}{5073} \approx 2.0\%$	4
F	$\frac{147}{5073} \approx 2.9\%$	4	**O**	$\frac{326}{5073} \approx 6.4\%$	1	**X**	$\frac{20}{5073} \approx 0.4\%$	8
G	$\frac{192}{5073} \approx 3.8\%$	2	**P**	$\frac{118}{5073} \approx 2.3\%$	3	**Y**	$\frac{103}{5073} \approx 2.0\%$	4
H	$\frac{292}{5073} \approx 5.8\%$	4	**Q**	$\frac{35}{5073} \approx 0.7\%$	10	**Z**	$\frac{26}{5073} \approx 0.5\%$	10
I	$\frac{307}{5073} \approx 6.1\%$	1	**R**	$\frac{266}{5073} \approx 5.2\%$	1			

1. Do you think that the association between the Scrabble-tile point value and the letter's relative frequency will be positive or negative? Why?

The association should be negative because we would expect that the more frequently a letter occurs in the English language, the fewer Scrabble points are given for using that letter in a word. For instance, the letter E is the most frequently used letter in the English language. It has a Scrabble point value of 1 because the letter E is found in many words.

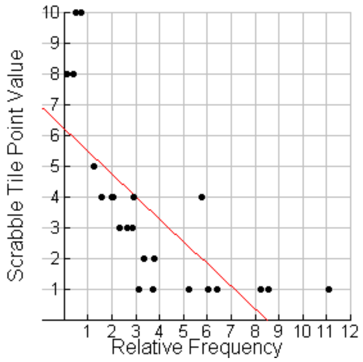
2. Using your graphing calculator or Excel, make a scatterplot with the Scrabble-tile point value for each letter on the vertical axis and the relative frequency of the letter on the horizontal axis. ***<u>Provide a printout of the graph when you turn in this project.</u>***



3. Use the scatterplot to describe the form, strength, and direction of the association between a letter's relative frequency and its Scrabble-tile point value.

A letter's relative frequency in English text and its Scrabble-tile point value have a curved association. The relationship is fairly strong. As the relative frequency of the letter increases, the Scrabble-tile point value decreases.

4. Use your calculator or Excel to fit a straight-line (regression) model to the data.



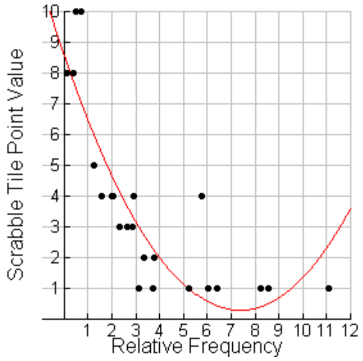
Linear Regression (ax+b)

regEQ(x) = -.734547x + 6.23307

r^2 = .53265

5. Plot the regression line on your scatterplot.

6. Use your calculator to fit the quadratic model $y = ax^2 + bx + c$ to the data. Plot your fitted-model equation on the scatterplot in question 2.

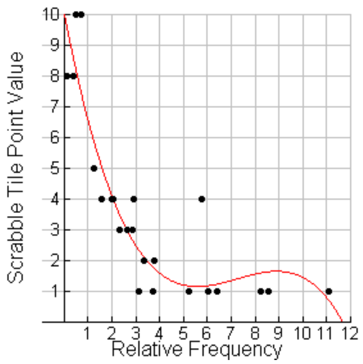


Quadratic Regression

regEQ(x) = .153366x^2 + -2.25867x + 8.60364

R^2 = .777442

7. Use your calculator to fit the cubic model $\boldsymbol{y = ax^3 + bx^2 + cx + d}$ to the data. Plot your fitted-model equation on the scatterplot in question 2.

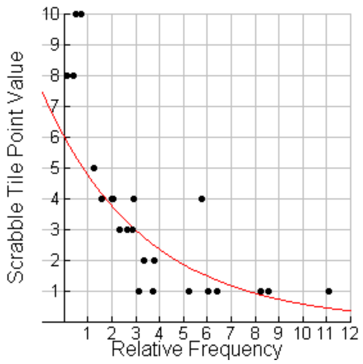


Cubic Regression

regEQ(x) = -.027235x^3 + .591082x^2 + -4.05275x + 10.1013

R^2 = .853478

8. Use your calculator to fit the exponential model $\boldsymbol{y = ab^x}$ to the data. Plot your fitted-model equation on the scatterplot in question 2.



Exponential Regression

regEQ(x) = 6.04676*.791474^x

r^2 = .645418

9. Choose the best model (linear, quadratic, cubic, or exponential) and explain why you think it is the best. Using this model, predict the Scrabble-tile point value for the letters C and N. Complete the following table.

Table 5

Letter	**Relative Frequency in English Text**	**Actual Scrabble-Tile Points**	**Predicted Scrabble-Tiles Points**	**Residual**
C	3.4%	3		
N	5.9%	1		

Answers will vary. For the letter C, linear is ~3.7, quadratic is ~2.7, cubic is ~2.1, exponential is ~2.7. For the letter N, linear is ~1.9, quadratic is ~0.6, cubic is ~1.2, exponential is ~1.5. They might decide the best model based on correlation coefficient, coefficient of determination, eyeball, residual.

10. Use the model you chose in question 9 to determine how many points the blank, "wild card" tiles should be worth. How does this compare to the actual game where the blank tiles are worth 0 points? Explain.

For linear is ~6.2, quadratic is ~8.6, cubic is ~10.1, exponential is ~6.0. Comparison is terrible. The model does not accurately predict the situation.

11. Discuss the domain and range for your chosen model.

We might say that the domain is (0, 9). That is, the relative frequency falls in the range between 0% and 9% exclusive of the endpoints.

The range is [0, 10] based on the actual Scrabble game. We get 0 points for the wild card and 10 points for Q and Z.

Scrabble - TEAM PROJECT A

2008 Excellence in Mathematics Contest

The Team Project is a group activity in which the students are presented a series of mathematical problems relating to a specific theme. The team members are to solve the problems and write a narrative about the theme which answers all the mathematical questions posed. Teams are graded on accuracy of mathematical content, clarity of explanations, and creativity in their narrative.

Scoring Sheet

The Team Project is a group activity in which the students are presented a series of mathematical problems relating to a specific theme. The team members are to solve the problems and write a narrative about the theme which answers all the mathematical questions posed. Teams are graded on accuracy of mathematical content, clarity of explanations, and creativity in their narrative. A holistic scoring approach should be used to judge the team project. For each project, assign a score to each of the major areas:

School Name: ____-

Accuracy of mathematical content: 0 1 2 3 4
Comments:

Clarity of Explanations: 0 1 2 3 4
Comments:

Creativity in Narrative: 0 1 2 3 4
Comments:

Overall Presentation: 0 1 2 3 4
Comments: