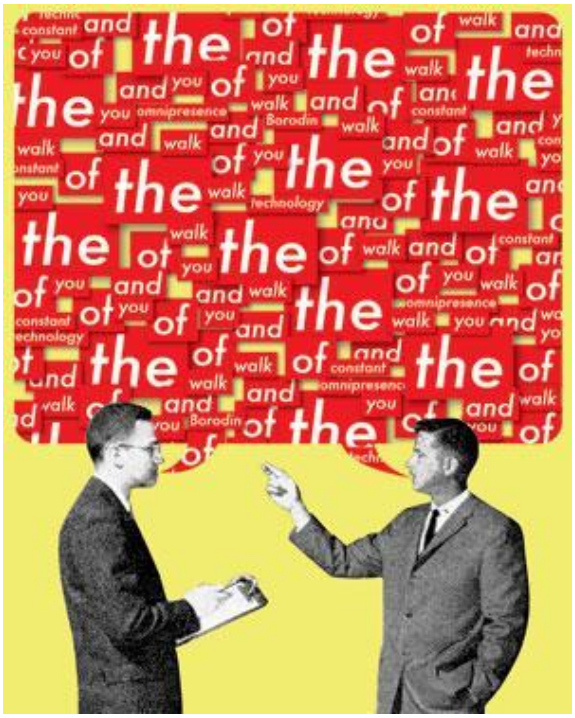
2015 Excellence in Mathematics Contest
Team Project Level II
(Below Precalculus)

CHANDLER-GILBERT COMMUNITY COLLEGE



School Name:

Group Members:

_____

_____

_____

_____

_____

_____

Reference Sheet

Formulas and Facts

You may need to use some of the following formulas and facts in working through this project. You may not need to use every formula or each fact.

$A = bh$ Area of a rectangle	$C = 2l + 2w$ Perimeter of a rectangle	$A = \pi r^2$ Area of a circle
$C = 2\pi r$ Circumference of a circle	$A = \frac{1}{2}bh$ Area of a triangle	$m = \frac{y_2 - y_1}{x_2 - x_1}$ Slope
$a^2 + b^2 = c^2$ Pythagorean Theorem	5280 feet = 1 mile	3 feet = 1 yard
P(A and B) = P(A) · P(B given A) Probability of dependent events	2.54 centimeters = 1 inch	$h = -4.9t^2 + v_0t + h_0$ $h = -16t^2 + v_0t + h_0$
1 kilogram = 2.2 pounds	1 meter = 39.3701 inches	1 gigabyte = 1000 megabytes
1 mile = 1609 meters	1 gallon = 3.8 liters	1 square mile = 640 acres
1 sq. yd. = 9 sq. ft	1 cu. ft. of water = 7.48 gallons	P(A or B) = P(A) + P(B) Probability of independent events
$V = \pi r^2 h$ Volume of cylinder	$V = (Area\ of\ Base) \cdot height$ Volume	$V = \frac{4}{3}\pi r^3$ Volume of a sphere
Lateral SA = $2\pi \cdot r \cdot h$ Lateral surface area of cylinder	$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ Quadratic Formula	$\tan\theta = \frac{\sin\theta}{\cos\theta}$

TEAM PROJECT Level II
2015 Excellence in Mathematics Contest

The Team Project is a group activity in which the students are presented an open ended, problem situation 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. We encourage the use of a graphing calculator.

Part 1: Background

In this Team Project, you will investigate and explore what is known as Zipf's Law or Zipf's Mystery. The purpose of Part 1 is to provide background about the namesake of this law – George Zipf. The following is taken from en.wikipedia.org/wiki/George_Kingsley_Zipf.

George Kingsley Zipf (/ˈzɪf/; 1902–1950), was an American linguist and philologist who studied statistical occurrences in different languages. Zipf earned his bachelors, masters, and doctoral degrees from Harvard University, although he also studied at the University of Bonn and the University of Berlin. He was Chairman of the German Department and University Lecturer (meaning he could teach any subject he chose) at Harvard University. He worked with Chinese and demographics, and much of his effort can explain properties of the Internet, distribution of income within nations, and many other collections of data.

Zipf is the eponym of Zipf's law, which states that while only a few words are used very often, many or most are used rarely,

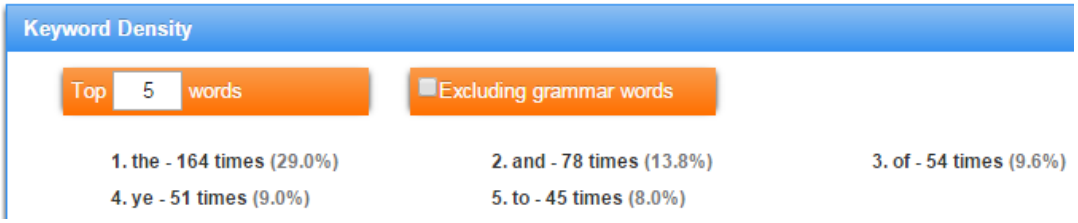
$$P_n \sim 1/n^a$$

where P_n is the frequency of a word ranked n^{th} and the exponent a is almost 1. This means that the second item occurs approximately 1/2 as often as the first, and the third item 1/3 as often as the first, and so on. Zipf's discovery of this law in 1935 was one of the first academic studies of word frequency.

Although he originally intended it as a model for linguistics, Zipf later generalized his law to other disciplines. In particular, he observed that the rank vs. frequency distribution of individual incomes in a unified nation approximates this law, and in his 1941 book, "National Unity and Disunity" he theorized that breaks in this "normal curve of income distribution" portend social pressure for change or revolution.

Part 2 – Zipf's Law

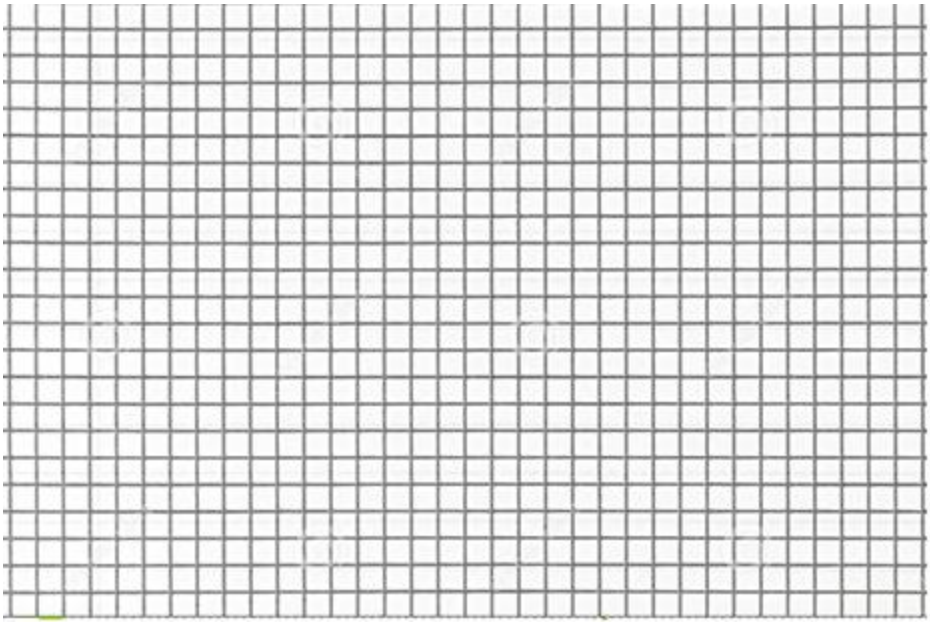
The text from Chapter 36 of the book Moby Dick was cut-and-pasted into the website wordcounttools.com. The following information was obtained.



1. Create a scatter plot showing the frequency of the word on the vertical axis and the word rank (1 through 5) on the horizontal axis. Discuss the degree to which this situation follows Zipf's Law.

2. Based on Zipf's Law, predict the frequency of the 10^{th} ranked word in Chapter 36 of Moby Dick.

3. One way to explore Zipf's Law is to consider what is known as a log-log graph. This means that rather than analyzing a graph of the raw data, we will analyze the graph of the logarithm of the data. That is, create a graph of the common logarithm of the word frequency on the vertical axis and the common logarithm of the rank on the horizontal axis. What do you notice?



4. Create a linear model for the log-log graph created in #3 above.

5. Use the log-log linear model to predict the frequency of the 10^{th} ranked word from Chapter 36 of Moby Dick. Show all your algebraic work in the space below.

Part 3 – Zipf's Law and Population

Consider a portion of an article found at /io9.com/the-mysterious-law-that-governs-the-size-of-your-city-1479244159

A mysterious law that predicts the size of the world's biggest cities

For the past century, an obscure mathematical principle called Zipf's law has predicted the size of mega-cities all over the world. And nobody knows why.

Back in 1949, the linguist George Zipf noticed something odd about how often people use words in a given language. He found that a small number of words are used all the time, while the vast majority are used very rarely. If he ranked the words in order of popularity, a striking pattern emerged. The number one ranked word was always used twice as often as the second rank word, and three times as often as the third rank. He called this a rank vs. frequency rule, and found that it could also be used to describe income distributions in any given country, with the richest person making twice as much money as the next richest, and so forth.

Later dubbed Zipf's law, the rank vs. frequency rule also works if you apply it to the sizes of cities. The city with the largest population in any country is generally twice as large as the next-biggest, and so on. Incredibly, Zipf's law for cities has held true for every country in the world, for the past century.

1. If Zipf's Law holds true for city populations, then fill out the following table. Round to the nearest whole number.

City Rank*	City	Estimated Population
1	New York City	8,491,079
2	Los Angeles	
3	Chicago	
4	Houston	
5	Philadelphia	
6	Phoenix	

*City population ranks found at: hen.wikipedia.org/wiki/List_of_United_States_cities_by_population

2. You will now compare the predicted populations from Zipf's Law to the actual populations of the top 6 ranked cities by population. Complete the table. Round the percent change to the nearest tenth of a percent.

City Rank	City	Estimated Population	Actual Population*	Percent Change
1	New York City	8,491,079	8,491,079	0%
2	Los Angeles		3,928,864	
3	Chicago		2,722,389	
4	Houston		2,239,558	
5	Philadelphia		1,526,006	
6	Phoenix		1,445,632	

*Actual city populations found at: hen.wikipedia.org/wiki/List_of_United_States_cities_by_population

3. What is the average percent change in actual city population when compared to the estimated population using Zipf's Law? Write a detailed description of what this average percent change means in the context of this situation.

Part 3 continues…

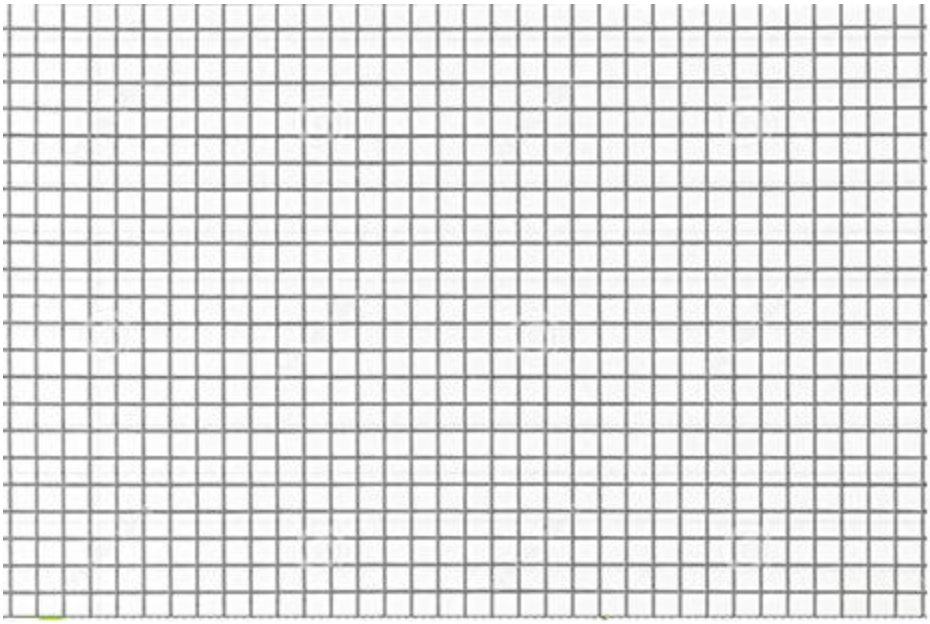
You were briefly introduced to the idea of the log-log plot in Part 2. In this portion of Part 3, you will explore the idea in greater depth. The log-log plot is often used in situations where the data being analyzed contains very large numbers. Rather than analyzing the data as is, we will analyze the logarithms of the data.

4. Complete the table.

City Rank*	**Common log of the City Rank** **(round to the nearest 0.001)**	**City**	**Estimated Population**	**Common log of the Estimated Population** **(round to the nearest 0.01)**
1		New York City	8,491,079	
2		Los Angeles		
3		Chicago		
4		Houston		
5		Philadelphia		
6		Phoenix		

*City population ranks found at: hen.wikipedia.org/wiki/List_of_United_States_cities_by_population

5. Carefully plot the Common log of the Estimated Population values against the Common log of the City Rank values below. Label the axes appropriately.



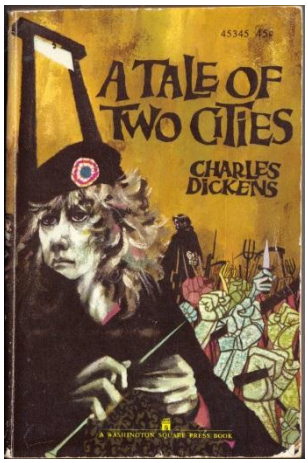
6. As you have hopefully seen, an advantage of analyzing the log-log plot is that it is nearly linear when Zipf's Law is applicable. Find a linear model for the plot you created in item #5. Write a clear explanation for the meaning of the parameters of this linear model. The parameters are the values of a and b that are output by the calculator.

7. Suppose you were sharing your linear model from item #6 with a friend. They are aware that the context is related to city populations. They become curious since the values in the linear model do not appear to be large enough to be related to city populations. Write an explanation for the connection between the vertical intercept of this linear model to the original data from which the model was determined.

Part 4 – Word Length Frequency

In this part of the project, we will examine word length rather than word count like we did previously. Consider the opening paragraph to the classic book *A Tale of Two Cities*.

> It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to heaven, we were all going direct the other way - in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.



Counting and recording the number of words of length 1 letter, 2 letters, 3 letters, etc. would be extremely tedious and uninteresting. Fortunately, there is a website that will help to do this tedious work for us (www.csgnetwork.com/documentanalystcalc.html). On the following page you will see the results from inputting the text above.

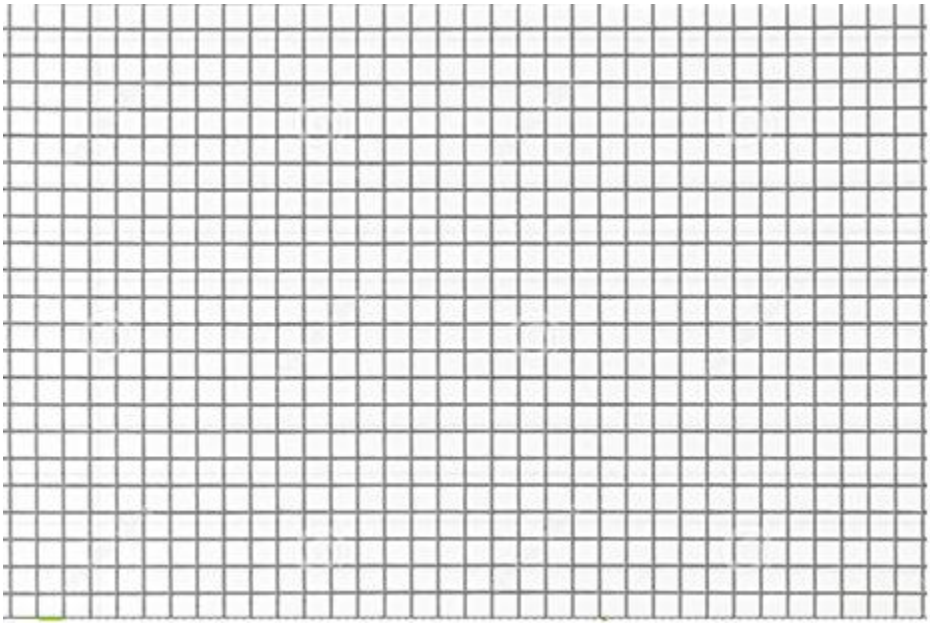
1. Analyze the data on the next page and complete the table.

Word Length (number of letters)	**Frequency** (express as a decimal to the nearest 0.001)
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	

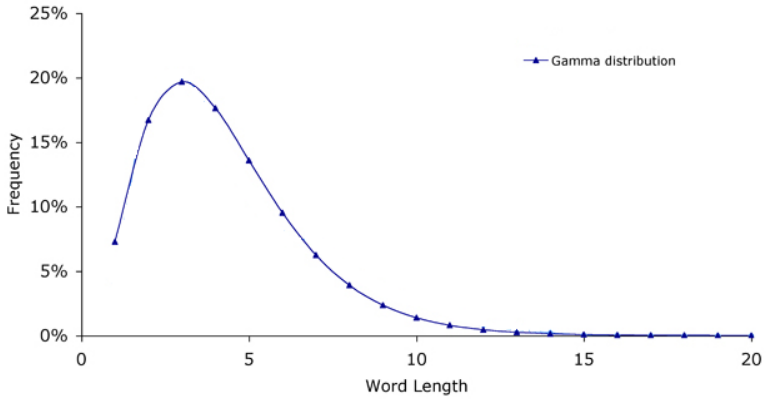
Unique words:58 Total words:120

Freq.	Word
14	THE
12	OF
11	WAS
10	IT
4	WE
2	AGE
2	EPOCH
2	SEASON
2	TIMES
2	HAD
2	BEFORE
2	US
2	WERE
2	ALL
2	GOING
2	DIRECT
2	IN
2	PERIOD
2	ITS
2	FOR
1	WISDOM
1	FOOLISHNESS
1	EVERYTHING
1	WORST
1	BELIEF
1	NOTHING
1	INCREDULITY
1	BEST
1	LIGHT
1	DARKNESS
1	TO
1	HEAVEN
1	OTHER
1	WAY
1	SPRING
1	SHORT
1	HOPE
1	SO
1	FAR
1	LIKE
1	PRESENT
1	THAT
1	SOME
1	WINTER
1	NOISIEST
1	AUTHORITIES
1	INSISTED
1	ON
1	BEING
1	RECEIVED
1	DESPAIR
1	GOOD
1	OR
1	EVIL
1	SUPERLATIVE
1	DEGREE
1	COMPARISON
1	ONLY

2. Plot the data from your table in item #1.



3. For large amounts of text, it has been shown that the frequency distribution of word lengths follow what is known as a form of the gamma distribution. The graph of this gamma distribution would look like the following for large amounts of text.



The gamma distribution function formula is of the form $f = aL^b c^L$ where f is the frequency, L is the word length and a, b, and c are constants. For large amounts of text, $a = 0.16$, $b = 2.33$, and $c = 0.49$. Use your graphing calculator to create a scatter plot of your data from item #1 and #2 above and then plot the ideal gamma distribution function, f along with the scatter plot. How well does your data fit the ideal gamma distribution?

4. According to the ideal model, from item #3, what would be the frequency of a 3 letter word? How does the ideal frequency compare to the frequency from the portion of text from *A Tale of Two Cities*? Show your work to justify your answer.

5. Suppose a word frequency of 19.22% was observed. What is the word length in this case? Show all your work to justify your response.