Cryptography:
Frequency Analysis

This activities explores frequency analysis as a means to detect patterns in messages and use that information to decrypt a message without a key. You will learn to recognize the importance of strong encryption methods and why monoalphabetic ciphers are especially weak.

Grade Level(s)
5-8

Approximate Time Required
1 hour

Materials
- Frequency Analysis Printout
- Pencil
- Computer with Internet Access for Research

NOTE: See note about Intenet safety below.

Activity Notes:
- Using an online search engine to conduct your own research is a great learning experience but should be done with constant adult monitoring. You should be informed in advance what to do if you encounter content that is not appropriate. It is suggested that a family friendly search engine like https://www.kidtopia.info/ or https://www.kiddle.co/ be used.

Introduction

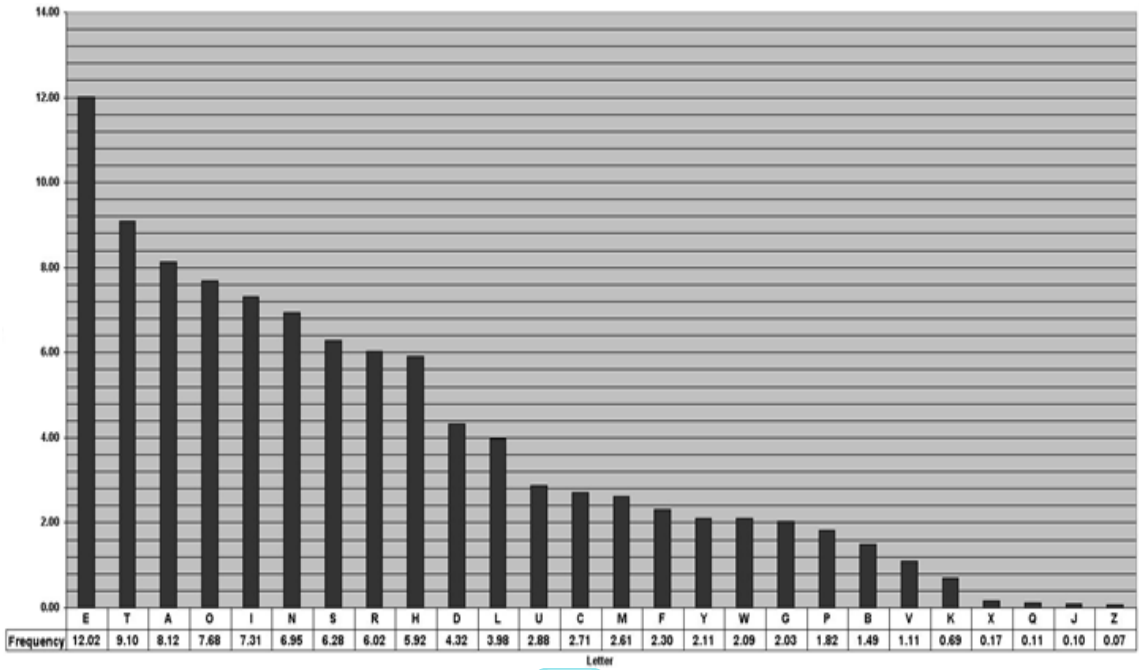
The entire purpose of encryption is to make it only readable by the intended recipient. No one else should be able to decipher the content of the message. This keeps important or critical information from falling into the wrong hands or being read by someone who shouldn't read it. Throughout time, different forms of encryption have been used and for as long as secret messages have been used, people have been trying to decrypt them.

In ancient times, some well-known and commonly used ciphers simply replaced one letter for another letter or a particular symbol. This is practice known as a simple substitution cipher. There are two primary categories of simple substitutions, monoalphabetic and polyalphabetic ciphers. With monoalphabetic (mono meaning "one" or "single") ciphers, one letter of the alphabet is substituted for another letter. In the case of pigpen ciphers, a symbol is used to represent a letter. The Caesar Cipher uses a "shift" method where one letter is represented by another one a certain distance away from it. For example, if you are using a +4 shift, the letter A would be written as the letter E which is four places to the right of the letter A. Regardless of how many spaces you move, or what symbol is used to represent a number, monoalphabetic ciphers have two major weaknesses. A polyalphabetic cipher substitutes individual letters with a number of potential substitutions at different positions in the message.

Activity 1: The English Language

The first weakness is there are only 26 letters of the alphabet. That means that for any letter there are only 25 possible options. With enough time and effort, the message can be decrypted by a process of elimination. The second weakness is the application of frequency analysis. **Frequency analysis** studies a message and compares how often each letter is used. From there, assumptions and inferences can be made about the text to decrypt the message.

In the English language, some letters are used more often than others when forming words. The graph below shows the average use of each letter in the English language. Cornell University conducted an analysis of 40,000 words in the English language then generated the graph below to display their findings. The decimals listed below each letter is a percentage of all letters used. Notice the top five letters compared to the other 21 letters. Does anything about this graph surprise you?



	E	T	A	O	I	N	S	R	H	D	L	U	C	M	F	Y	W	G	P	B	V	K	X	Q	J	Z
Frequency	12.02	9.10	8.12	7.68	7.31	6.95	6.28	6.02	5.92	4.32	3.98	2.88	2.71	2.61	2.30	2.11	2.09	2.03	1.82	1.49	1.11	0.69	0.17	0.11	0.10	0.07

To get this information, each letter of each word counted and compared to every other letter in the study. If you look at the table below, you see each letter of the alphabet and how many times it was used in the 40,000 words analyzed. The letter E was used 21,912 times followed by T, A, O, and I. This study mimics other studies and examples done by other researchers.

Table A:

E	T	A	O	I	N	S	R
21912	16587	14810	14003	13318	12666	11450	10977
12.02%	9.10%	8.12%	7.68%	7.13%	6.95%	6.28%	6.02%

H	D	L	U	C	M	F	Y
10795	7874	7253	5246	4943	4761	4200	3853
5.92%	4.32%	3.98%	2.88%	2.71%	2.61%	2.30%	2.11%

W	G	P	B	V	K	X	Q
3819	3693	3316	2715	2019	1257	315	205
2.09%	2.03%	1.82%	1.49%	1.11%	0.69%	0.17%	0.11%

J	Z
188	128
0.10%	0.07%

Using these statistics as a foundation, you can begin making inferences and decomposing an encrypted text. Frequency analysis consists of two basic steps. The first is to collect and organize the data. Basically, count the letters. The second step is to use the data in comparison to known statistics and norms to make inferences. The following activities walk you through conducting frequency analysis and decrypting a message.

Explore More: Frequency Analysis Part 1

You will need the following materials:

- Frequency Analysis Printout
- Pencil

The first step of frequency analysis is to collect and organize the data. Examine the encrypted message below. This message was encrypted using a monoalphabetic cipher but you do not know the key. Use frequency analysis to attempt decrypting the text.

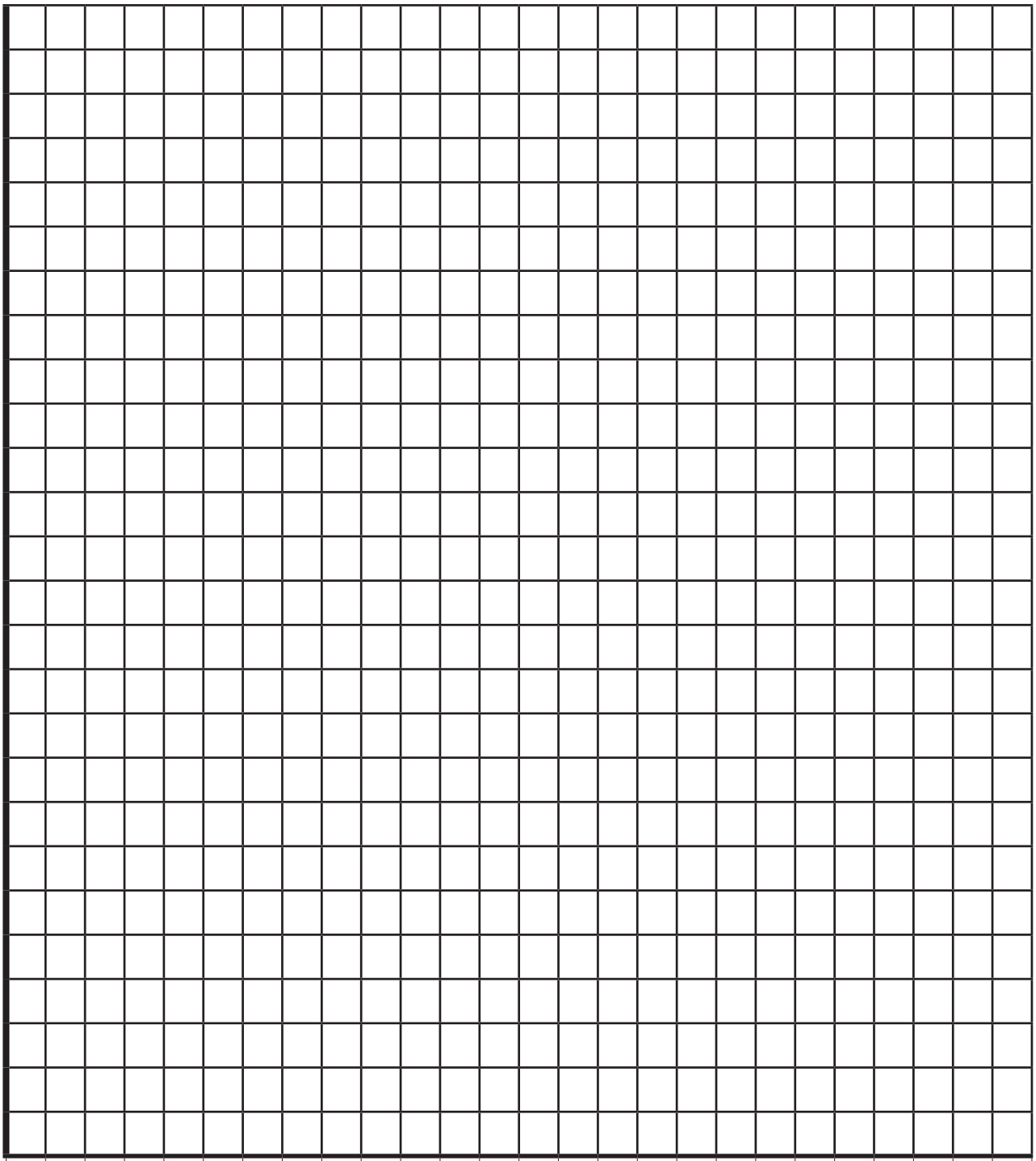
1. Count how many times each letter is used in the message and write that number in the space above that letter.
2. Once all the letters are accounted for, create a graph to display your findings.
3. Compare your chart with the one above.

Encrypted Message:

Ymj jsynwj uzwutxj tk jshwduynts nx yt rfpj ny tsqd wjfifgqj gd ymj nsyjsiji wjhnunjsy. St tsj jqxj xmtzqi gj fgqj yt ijhnumjw ymj htsyjsy tk ymj rjxxflj.

Frequency Analysis Printout

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z



a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z

Turn your data into percentages. How many letters are there total in the message?

Write the percentage of each letter below the actual number.

How do you find the percentage? Divide the ***frequency*** number by the ***total*** number of letters used. That will give you a decimal value. Multiply x 100 to get the percentage value. Round to the nearest tenths place.

For example, if your total number of letters equals 150, and the letter D appears 27 times, then your algorithm would look like this:

Example: D = 27/150 = 0.18 0.18 x 100 = 18 = 18%

Find the percentages for each of the letters used in the encrypted message.

Think: Can you make any assumptions based on comparing Table A and your graph? Do you see any patterns that might lead you to make certain inferences?

Explore More: Frequency Analysis Part 2

You will need the following materials:

- Frequency Analysis Printout
- Pencil

The second step of frequency analysis in encryption involves substituting letters, looking for patterns, and applying known English language norms.

1. Order the letters greatest to least in order of frequency. Include the frequency number (percentage optional).
2. Next, list the letters of the alphabet in order from greatest to least according to the Cornell University data (Table A).
    - Compare the data and make assumptions.
    - What letter is used most often in the text?
    - Try replacing that letter with the letter E.
    - Repeat with the other letters one at a time. Look for word or letter patterns as you work.

Analyze the encrypted message after every substitution. Are any clues to the text beginning to appear? Do you notice any patterns emerge? Once a letter is replaced, can you infer what the word might be? What do you know about three letter words that might make deciphering some letters easier?

How does frequency analysis affect larger messages? Understanding what frequency analysis is will help you to decode certain types of ciphers. However, these only work with ***monoalphabetic*** ciphers. ***Polyalphabetic*** ciphers such as the Vigenère cipher were developed to combat frequency analysis and make the practice useless.

Before computers, cryptographers would count and graph letters by hand just as you did in this activity. However, today we can use computer programs designed to analyze text and organize the data. The following website uses software to help decrypt a message.

Visit the following website. This website not only offers a review of frequency analysis but also several challenge messages to decrypt. If you get stuck, there is a hint also added.

Visit: Frequency Analysis (https://www.101computing.net/frequency-analysis/)

1. Click "Start Frequency Analysis" of the sample text.
2. Substitute a few letters based on what you know about common letters used in the English language.
3. Click "Start Substitution" .
4. Examine the text.
   - Look for clues to other letter possibilities.
   - Change potentially incorrect letters.
   - Click "Start Substitution" to update your changes.
5. Repeat until you have decrypted the message.

In your own words, describe what frequency analysis is and how letter frequency is a weakness for some types of encryption.