

WebQuest: Code Cracking and the Law of Large Numbers Part II

The Law of Large Numbers has other applications other than games of chance such as dice rolling. Suppose you had to crack a substitution cipher code, where one letter is substituted for another. We will see how the Law of Large Numbers to crack substitution codes. The article is found on the following website:

www.oreilly.com/catalog/statisticshks/chapter/hack70.pdf

Read the article, shown below, "Statistics Hacks: Tips & Tools for Measuring the World and Beating the Odds." As you are reading the article below, think about how the Law of Large Numbers is used to crack substitution codes.

Break Codes with Etaoin Shrdlu

You might have noticed that certain keys on your computer keyboard get dirty or wear out more quickly than others. That's because you hit them more often than the others. You might also notice that these letters tend to be in the middle of the keyboard or, more correctly, in small circles near where your hands are when they are centered on a keyboard. Both the wear and tear on your keys and the placement of them in a standard typewriter (a.k.a. QWERTY, for the first six letters on the top row) pattern are based on their frequency of use in English. Different letters in the alphabet are used with different frequencies in the spelling of words in a language. By applying the known frequency of these letters, along with other statistical tricks, you can quickly decode classified documents, whether they are Leonardo da Vinci's diary, a puzzle in the newspaper, or big, bright letters being turned by Vanna White on TV.

Single Substitution Ciphers

The simplest and oldest type of letter-based code is the *single substitution* format. In these codes, some message is transformed from the actual letters in the words to other letters in the alphabet. In the simplest form of this type of coding, the same letter substitutes for the same letter throughout the message. For example, a simple cipher might use the substitution pattern shown in Table 6-18, in which the letters on the top row (the *plain text*) are replaced by the letters on the bottom row (the *cipher text*).

Table 6-18. A single substitution cipher

Plain text	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
Cipher text	N A O B P C Q D R E S F T G U H V I W J X K Y L Z M

With a code like the one shown in Table 6-18, the following plain-text passage:

Tom appeared on the sidewalk with a bucket of whitewash and a long handled brush.

appears in cipher text like this:

Jut nhhpnipb ug jdp wrbpynfs yrjd n axospj uc ydrjp yhwd ngb u fugqdngbfpb aixwd.

The passage looks like nonsense, but with the key shown in Table 6-18, anyone could easily replace the nonsense letters with the original letters, causing the opening sentence of the second paragraph in Chapter Two of *Tom Sawyer* to reveal itself.

Using Probability to Decode Substitution Ciphers

Of course, the real task when deciphering ciphers is to do it without access to the code key. Real-life code breakers and winning contestants on Wheel of Fortune use the same tool to solve their problems: they apply the known distribution of letters in English language words. The advent of computers, computer analysis, and electronic copies of millions of books has made the calculation of exact probabilities for each letter of the alphabet possible, though cryptographers (code makers and breakers) have known the basics for some time. Here are some of these basics:

- The most common letter, in terms of usage in English, is E.
- The least commonly used letter is Z.
- The most common consonant is T.
- J and X are rarely used, as is Q.
- When Q is used, it is almost always followed by U.
- Only A and I are used as one-letter words in English.

With even just these basic probability facts, you could begin to tackle decoding a cipher such as our Mark Twain passage. The most commonly appearing letters in the garbled version are P and N. Because N is used as a single-letter word, it cannot be E (N is most likely A), so a good first guess for P is that it substitutes for E.

With just a little knowledge of letter distribution, we have already identified the substitutes for E and A. We can't be sure we are right, but like any good statistician, we think we are probably right. Table 6-19 shows the likely distribution for each letter of the alphabet.

Table 6-19. Frequency distribution of letters in English Letter Frequency

A 8.04 %	B 1.54 %	C 3.06 %	D 3.99 %	E 12.51 %	F 2.30 %
G 1.96 %	H 5.49 %	I 7.26 %	J 0.16 %	K 0.67 %	L 4.14 %
M 2.53 %	N 7.09 %	O 7.60 %	P 2.00 %	Q 0.11 %	R 6.12 %
S 6.54 %	T 9.25 %	U 2.71 %	V 0.99 %	W 1.92 %	X 0.19 %
Y 1.73 %	Z 0.09 %				

ETAOIN SHRDLU

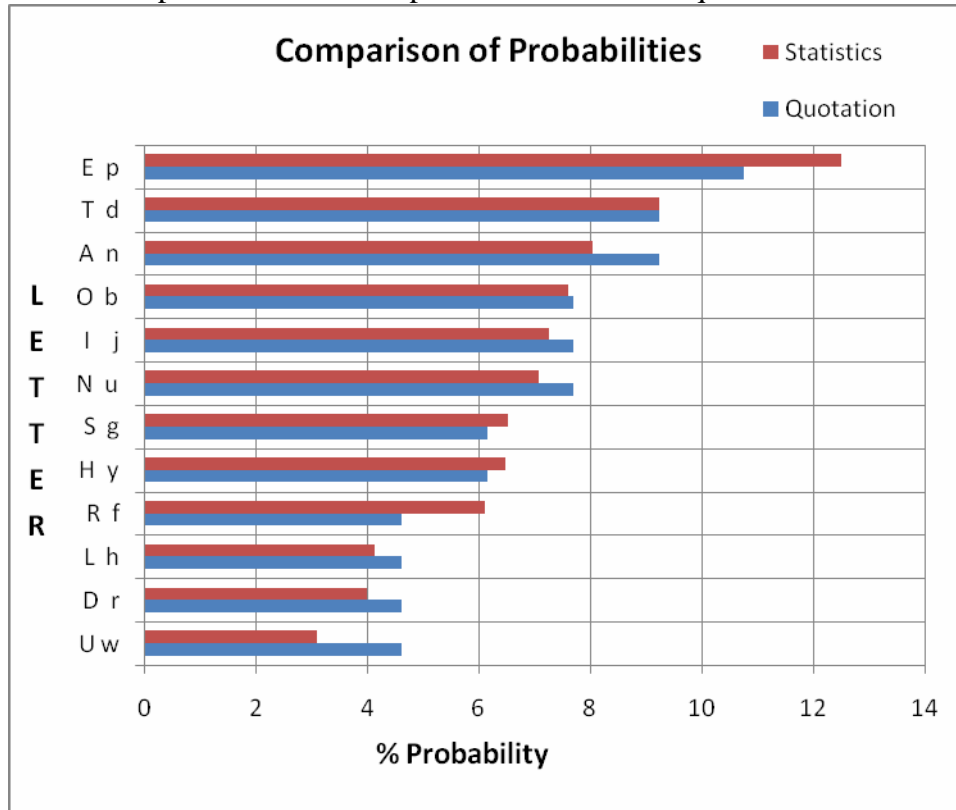
The strange phrase “ETAOIN SHRDLU” is a mnemonic device (memory tool) for remembering the most frequently occurring letters. These 12 letters account for over 80 percent of total letter frequency. You might notice that the order of letters in ETAOIN SHRDLU is not exactly the rank order of popularity shown in Table 6-19. It is close enough, though, and easier to pronounce than if it were exactly correct. Another thing to remember is that any “definitive” list of letter probability depends on the source material for the letter count. You can find many different lists of letter order and frequency, and some differ slightly from others.

For example, one organization that produced a list of statistical distributions of letters in English text relied on a computer analysis and actual count of letter occurrence in seven literary classics, such as *Jane Eyre* and *Wuthering Heights*. Two of these seven books were Tarzan novels. I’m guessing that if we were to compare that table of letter distributions with others, we would find that the proportional number of times the letter Z appeared was greater than if other sources were used. For the common letters, though—such as E, T, and A—there is wide agreement on their use as best first guesses for code breaking.

Statistical Analysis of Coded Texts

Here’s how you might use these letter statistics in real life to decode a secret message or solve a puzzle. This method works best if the coded text is lengthy, but it works surprisingly well even for shorter passages. Calculate the distribution of the coded, substitute letters (the cipher text), and then compare it to the distribution shown in Table 6-19. Figure 6-8 shows how this process might look graphically. Only the first 10 most common letters are shown, but the analysis would use all the letters. This example pretends that there is a lot of coded text and that the substitute cipher shown in Table 6-18 is being used. Because the most common substitute letters are P, followed by J, a good guess for breaking the code would be to see whether P could really be E and J could really be T. These first guesses can be made all the way down the line for each letter. By starting with the most frequently appearing letters and moving down the list, a code breaker can quickly see whether these first hypotheses are right or wrong and change guesses around until English words start to appear.

The table below compares the statistical probabilities with the quote.



Wheel of Fortune Strategy

On the TV game show Wheel of Fortune, before solving the big puzzle at the end, the producers are nice enough to provide certain letters and show whether they appear in the hangman-type phrase. They provide R, S, T, L, N, and E. These are given, of course, because they are common letters, and are in our top 12: ETAOIN SHRDLU. The player is allowed to choose three more consonants and another vowel. Using our statistical knowledge of letter frequency, a good basic strategy would be to pick A as the vowel and the three most common consonants not yet shown: H, D, and C.

Other Common Letter Patterns

Beyond just knowing the frequency of individual letters appearing, good code breakers use probability information about other patterns of letters:

Words are most likely to start with T, O, A, W, or B.

- Most words end with an E, T, D, or S.
- If two letters are doubled in a word, they are most likely to be SS, EE, TT, FF, or LL.
- Frequently appearing two-letter words include of, to, in, it, and is.
- By far, the most common three-letter words are the and and. Other common three-letter words include for, are, and but.
- Letters that tend to come in pairs include TH, HE, AN, IN, and ER.
- The most frequently used words are the, of, and, to, in, a, is, that, be, and it.

Perhaps indicating what people tend to write about, the top 100 most-used words in written texts include dollars, great, general, and public. Debts just barely failed to make the top 100, but it is surprisingly common. A good explanation of single substitution ciphers can be found under the entry for frequency analysis at http://en.wikipedia.org/wiki/Frequency_analysis.

Some of the statistics reported in this hack were found at <http://www.data-compression.com> and <http://www.scottbryce.com>.

Good information and advice for solving cryptograms and other codes using statistics can be found at those sites.

From the article you should notice three types of statistical analysis of probabilities, based on the Law Of Large Numbers, that are used:

- Frequency of letters (**ETAOIN SHRDLU**),
- Frequency of letter combinations (e.g. TH, HE, AN, IN, and ER), and
- Frequency of words (e.g. the, is, it, there).

For a long time code breaking was considered the domain of special geniuses. However, you should now see that with a computer and a good feel for statistics, you could make it into the elusive circle of code breakers.

6. Crack the following humorous coded message:

UJEJVZR QFEYGE, SV SO OFSU, JWIG FEESTGU FV VZG UJJE JC FW
FQFEVLGWV SW PZSIZ F NASVVGESWN QFEVR PFO VFYSWN QAFIG. FV
QEGISOGAR VZG OFLG LJLGVV, F XGFKVSCA XKV TFIKJKO OZJPNSEA
FEESTGU FV VZG UJJE.

CJE F LJLGVV, VZGEG PFO ZGOSVFSJW JW XJVZ OSUGO, FWU VZGW
VZG OZJPNSEA OVGQQGU XFIY VJ LFYG PFR, OFRSWN, "FNG XGCJEG
XGFKVR."

"WJV FV FAA!" OFSU UJEJVZR QFEYGE, OFSASWN VZEJKNZ. "QGFEAO
XGCJEG OPSWG!"

Make a graph or histogram of the ten most common letters in the English Language ETAOINSHRL and their frequency and compare it with the ten most common letters in the passage above.

Compare the two histograms and "crack the code." Include the histograms with your report. Explain your reasoning.

The answer is found at:

<http://mathcircle.berkeley.edu/BMC3/crypto/node2.html>

