For Whose Eyes Only?
Cryptanalysis and Frequency Analysis

David Cochran

Introduction

Cryptanalysis, the art of breaking codes, has proven to be a combat multiplier throughout history. By reading encrypted enemy message traffic, commanders often gain immediate intelligence on the enemy's disposition, strengths and intent. During World War II, many allied successes were attributed to the advantages gained by cracking both the German Enigma and Japanese PURPLE cipher systems. Further advances in the field of cryptanalysis will likewise give future American commanders a distinct advantage on the modern battlefield.

The ADFGVX Cipher

By 1917, the warring nations in World War I were facing manpower shortages due to the large-scale slaughter on the battlefields. The United States declared war on Germany and joined the Allies on April 6, 1917. On 26 June 1917, the first large contingent of American troops arrived in France. With fresh American forces arriving regularly in France, everyone realized the Americans would soon bring a numerical advantage to the Allies. They also knew there would be some time before sufficient numbers of Americans would be available to break the stalemate on the Western Front.

The success of the Bolsheviks in Russia and the resulting separate peace in December 1917 allowed the Germans to move forces from the Eastern Front to France. The change in force ratios allowed the Germans to gain a temporary numerical superiority in the West. They made a surprise attack on March 21, 1918 and made significant gains. The attack threatened to split the British and French armies and brought Paris within range of Germany's guns. More importantly, the attack depleted the previously strained manpower resources of both the British and French. On May 27, in an effort to reach Paris, the Germans attacked again between Reims and Soissons. The attacking forces penetrated the allied lines to a depth of fifteen miles and were so successful that the French government prepared to leave Paris. The Allied High Command knew the Kaiser's forces would attempt to exploit their success with yet another major attack, but only had sufficient reserves to bolster the defenses in one sector. Discovering the location and timing of the next large attack become critical. [1]

In March 1918, the Germans implemented a now famous cipher - the ADFGX cipher - to encrypt communications between their corps and division level field headquarters. This new cipher used only the letters A, D, F, G and X and resisted all attempts at cryptanalysis. This cipher contributed to the tactical

surprise achieved by the Germans during the initial spring offensive. A French cryptanalyst, Georges Panvin, attacked the new cipher using frequency analysis. By the end of April he was able to read some of the messages protected by this cipher, but he never achieved a general solution. On June 1, Panvin noticed that the Germans had slightly changed the cipher by including the letter V. Panvin used frequency analysis once more and was able to decipher some messages protected by this new cipher by June 2. He disseminated the key he uncovered to the other French cryptanalysts. On June 3, one of the French cryptanalysts, using the key provided by Panvin, deciphered an intercepted German message that provided the first hint as to the location and timing of the anticipated German offensive. When the new attack came, there was no surprise and the Germans were not as successful as in their previous attempts. Georges Panvin and the method of frequency analysis helped save France. [2]

Introduction to Cipher Systems

Koblitz defines cryptography as "the study of methods of sending messages in disguised form so that only the intended recipients can remove the disguise and read the message." [3] A cipher system consists of an enciphering map f and its inverse f^{-1} for deciphering. Often both f and f^{-1} depend on an encryption key, a parameter which may change the encryption map in some way. The enciphering map f takes the plaintext, the message we are disguising, and transforms it to the ciphertext, the disguised message. Likewise, f^{-1} takes the ciphertext and transforms it back to the plaintext.

$$Plain\ Text \xrightarrow{f} Cipher\ Text \xrightarrow{f^{-1}} Plain\ Text$$

There are two general classes of ciphering systems. The first is a symmetric ciphering system. In a symmetric ciphering system both the sender and receiver must have the same key and encryption algorithm to scramble and unscramble the message traffic. An example of a symmetric ciphering system is the simple substitution method discussed later in this paper.

The second class is a public key ciphering system. In a public key ciphering system, the sender (and everyone else) has access to the receiver's public encryption key and can use the public key to encipher the message to the receiver. The receiver alone has access to the secret deciphering key.

An example of a Public Key encryption system is the Rivest Shamir Adleman (RSA) encryption algorithm. In the RSA system, a user chooses two prime numbers, p and q, and calculates $n = pq$. Additionally, the user calculates a random number e, which has no factors in common with $(p-1)(q-1)$. The user then makes the two numbers e and n public. A person who wishes to communicate secretly with the user breaks the message in plaintext message blocks p_i and calculates the ciphertext c_i by the formula

$c_i = p_i^e \bmod n$ where e and n are the numbers previously made public to everyone. To decrypt the message the user calculates the private key d, by the relationship $d = e^{-1} \bmod((p-1)(q-1))$, and decrypts the message using the relationship $p_i = c_i^d \bmod n$. The security of the RSA system lies in the difficulty of factoring the number n into its prime factors p and q. This is a difficult mathematical problem when n is large. [4]

Example 1: Encrypt the message ATTACK using the RSA cipher system, when $p = 13$ and $q = 17$.

Solution: $n = pq = (13)(17) = 221$. Choose a random number e with no factors in common with $(p-1)\cdot(q-1) = (12)(16) = 192 = 2^6 \cdot 3$. Let $e = 25$. The public key consists of the two numbers $n = 221$ and $e = 25$. Everyone has access to these two numbers. To encrypt the message A-T-T-A-C-K, assign each of the plaintext message blocks the corresponding number for each letter in the plaintext. Then $p_1 = A = 1$, $p_2 = T = 20$, etc. The message is then encrypted as:

$$c_i = p_i^e \bmod(n)$$
$$c_1 = 1^{25} \bmod(221) = 1$$
$$c_2 = 20^{25} \bmod(221) = 150$$
$$c_3 = 20^{25} \bmod(221) = 150$$
$$c_4 = 1^{25} \bmod(221) = 1$$
$$c_5 = 3^{25} \bmod(221) = 133$$
$$c_6 = 11^{25} \bmod(221) = 193$$

The message $1-20-20-1-3-11$ is encrypted as $1-150-150-1-133-193$. To decrypt the message again, we calculate the private key $d = 25^{-1} \bmod(192), \quad d = 169$, remembering the intended recipient alone knows this number. The message is decrypted as follows:

$$p_i = c_i^d \bmod(221)$$
$$p_1 = 1^{169} \bmod(221) = 1$$
$$p_2 = 150^{169} \bmod(221) = 20$$
$$etc.$$

The security of the RSA cipher system is greatly enhanced when large prime numbers are used and the size of the message block is increased. At the current level of mathematical knowledge, a number 100 digits long may be factored in minutes, while a number 200 digits long may take years to factor. Because of

this fact, in practice this algorithm is often used with prime numbers, p and q, over 100 digits long. Prime numbers of this size will generate a number n that may take years to factor.

Substitution Ciphers

The simplest symmetric cipher system is a substitution system. In a substitution cipher system, plaintext symbols are substituted one-for-one with encrypted symbols. In the simplest example another letter represents each letter of the alphabet. For example, consider plaintext message traffic that consists only of the letters in the alphabet. An enciphering map and corresponding deciphering map, *f*, may be defined by the table:

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

Then the plain text message: ATTACK NOW is encrypted as the cipher text *MSSMTI LNW*.

Example 2: How many different enciphering maps can be created from a 26-letter alphabet? (At least one letter must be encrypted.) In general how many enciphering maps are there for *n* symbols? How many different enciphering maps are possible for *n* symbols where every symbol is mapped to a different symbol?

Solution: There are 26 symbols to choose for the first letter, 25 symbols to choose for the second symbol, 24 for the third, etc. Then there are

$$(26 \cdot 25 \cdot \cdots \cdot 2 \cdot 1) - 1 = 26! - 1$$
$$26! - 1 = 4.033 \times 10^{26}$$

ways to create a substitution enciphering map with 26 symbols. In general, there are $n! - 1$ ways to create a substitution enciphering map with *n* symbols when at least one symbol must be encrypted. If every letter must be encrypted, the problem is a derangement and there are approximately $n!/e$ ways to create different substitution alphabets.

Note: A derangement is the general problem of scrambling n elements with specified positions so that no element stays in its original position. Let D_n represent the total number of permutations of arranging n items such that no item is in its original position. Clearly if $n = 1$ then $D_1 = 0$. If $n = 2$, then $D_2 = 1$ for {1, 2} can be arranged only as {2, 1} when no element is placed in its original position. It can be shown that for $n \geq 3$, then $D_n = (n-1)(D_{n-2} + D_{n-1})$ and we can calculate $D_3 = (3-1) \cdot (0+1) = 2$ and $D_4 = (4-1) \cdot (1+2) = 9$. Additionally, it can be

shown that $e^{-1} = \frac{D_n}{n!} + (-1)^{n+1}\frac{1}{(n+1)!} + (-1)^{n+2}\frac{1}{(n+2)!} + \cdots$. The result that there are approximately $n!/e$ ways to create different substitution alphabets then follows. [5]

At first glance, the problem of finding the correct arrangement to decipher the message may seem to be a daunting challenge, but we can use clues from the message itself to simplify the task. Notice in the plaintext message from the substitution example above, the symbols T and A both occur twice. Likewise the corresponding symbols in the ciphertext, M and S both occur twice. This gives us clues that we may be able to use underlying statistical characteristics of the plaintext to find information about our ciphertext.

Frequency Analysis

As discovered above, some cipher systems may lend themselves to statistical methods of cryptanalysis. The idea of using the underlying frequency distribution of a language to aid in the attempt to read encrypted message traffic is first attributed to an Arab named Qalqashandi early in the 15th Century. [6] Since his time, frequency analysis is often the first step undertaken by cryptanalysts in the attempt to read the enemy's encrypted message traffic.

In the English language, the letter 'e' is the most frequently occurring letter. In fact, using an electronic version of Melville's Moby-Dick, prepared by Professor Eugene F. Irey at the University of Colorado [7], I calculated the following frequencies and associated probabilities:

Occurrences	Letter	Probability	Occurrences	Letter	Probability
114462	**E**	0.122748	22079	**C**	0.023677
86159	**T**	0.092396	21659	**W**	0.023227
76146	**A**	0.081658	20401	**F**	0.021878
67815	**O**	0.072724	20385	**G**	0.021861
64273	**I**	0.068926	17551	**P**	0.018822
64267	**N**	0.068919	16530	**B**	0.017727
62846	**S**	0.067396	16512	**Y**	0.017707
61469	**H**	0.065919	8457	**V**	0.009069
50912	**R**	0.054598	7903	**K**	0.008475
41969	**L**	0.045007	1512	**Q**	0.001621
37454	**D**	0.040165	1204	**X**	0.001291
26037	**U**	0.027922	1046	**J**	0.001122
22817	**M**	0.024469	630	**Z**	0.000676

Thus from the sample of the English language contained in Moby-Dick, the letter 'e' has a 12.27% probability of occurring.

Example 3. What is the expected number of occurrences of the letter 'n' in a 721-letter size sample from the novel Moby-Dick? What is the standard deviation of the number of occurrences of the letter 'n' in a 721-letter sample from Moby-Dick?

Solution: Let E_i represent the event that the randomly chosen letter is the $i\,th$ most frequently occurring letter. Then E_1 represents the event that a randomly chosen letter is the most frequently occurring letter 'e'. E_2 represents the event that the randomly chosen letter is the second most frequently occurring letter 't', etc. The events E_i are mutually exclusive and mutually exhaustive. A randomly selected letter can not be the letter 'e' and the letter 't' at the same time, nor will we encounter any other letter other than those in the alphabet. Define the random variable $\vec{X} = (x_1, x_2, \ldots, x_{26})$ such that x_i is the number of occurrences of event E_i in n trials. $\vec{X}$ has a multinomial distribution and the expected value $E(X_i)$ and variance $V(X_i)$ for event E_i in n occurrences are $E(X_i) = n \cdot p_i$ and $V(X_i) = n \cdot p_i \cdot (1 - p_i)$. The expected number of occurrences of the letter 'n' then is $E(X_7) = E('n') = (721)(0.06892) = \underline{\underline{49.69}}$ and the variance and standard deviation can be calculated as

$$V(X_7) = V('n') = 721 \cdot (0.06892) \cdot (1 - 0.06892) = \underline{\underline{46.27}}$$

$$s_{X_7} = \sqrt{V(X_7)} = \sqrt{46.27} \approx \underline{\underline{6.802}}$$

Therefore we expect 50 occurrences of the letter 'n' in a sample size of 721 with a variance of 6.802.

Example 4: Consider the following ciphertext with corresponding plaintext from the novel Moby-Dick:

GSV OZMW HVVNVW HXLIXSRMT GL SRH UVVG. DLMWVIUFOOVHG GSRMTH ZIV VEVI GSV FMNVMGRLMZYOV; WVVY NVNLIRVH BRVOW ML VYRGZYSH; GSRH HRC-RMXS XSZYGVI RH GSV HGLMVOVHH TIZEV LU YFOPRMTGLM.

Make a table of occurrences for each letter and calculate the probabilities for each letter? Which letter 'most likely' represents 'e'?

Solution: (149 total symbols)

Occurrences	Letter	Probability	Occurrences	Letter	Probability
25	**V**	0.167785	4	**T**	0.026846
13	**G**	0.087248	4	**X**	0.026846
13	**H**	0.087248	3	**F**	0.020134
12	**M**	0.080537	3	**U**	0.020134
12	**R**	0.080537	2	**E**	0.013423
10	**S**	0.067114	1	**B**	0.006711
9	**L**	0.060403	1	**C**	0.006711
7	**I**	0.046980	1	**D**	0.006711
7	**O**	0.046980	1	**P**	0.006711
6	**Y**	0.040268	0	**A**	0.000000
6	**Z**	0.040268	0	**J**	0.000000
5	**W**	0.033557	0	**K**	0.000000
4	**N**	0.026846	0	**Q**	0.000000

To answer the question which symbol represents the letter 'e' we must first develop the idea of a the probability interval. Define the random variable X as the number of occurrences of the symbol that represents the letter 'e' out of 149 trials. The symbol that represents the letter 'e' has a probability of occurring $p = 0.1227$. Then X has a binomial distribution, $X \sim BIN(149, 0.1227)$. The cumulative distribution function (c.d.f.) for this random variable is:

$$B(x; 149, 0.1227) = \sum_{i=0}^{x} \binom{149}{i} \cdot (0.1227)^{i} \cdot (1-0.1227)^{149-i} \quad .$$

If we can find an x_1 such that $B(x_1; 149, 0.1227) = 0.025$ and an x_2 such that $B(x_2; 149, 0.1227) = 0.975$, then $P(x_1 \le X \le x_2) = 0.95$ and there is a 95% probability that X will take on a value from the set {x_1, x_1+1, x_1+2,..., x_2}. Using a computer and iteration we can quickly find that the values $x_1 = 10$ and $x_2 = 26$ result in $P(10 \le X \le 26) = 0.956$.

Therefore, we are at least 95% certain that the symbol representing the letter 'e' will occur between 10 and 26 times in our sample. The symbols {V, G, H, M, R, and S} all meet this criteria, and we conclude the letter 'e' is most likely represented by one of these symbols.

Computing the 95% probability interval for the next most frequently occurring letter 't', we find that the symbol representing the letter t should occur between 6

and 21 times in the encrypted message. The symbols {G, H, M, R, S, L, I, O, Y, and Z} are all candidates to be the letter 't'.

Since there is less than a 5% chance that the letter V represents a letter other than 'e', we conclude that the letter V *most likely* represents the letter 'e'.

This method of frequency analysis can be used to find the most likely value(s) for the most frequently occurring symbols in the sample population. Note that if we are able to intelligently guess at the eight most frequently occurring letters {E, T, A, O, I, N, S, H} we will know just over 64% of the information contained in a message.

Example 5: Using the message traffic from Example 4, calculate the 95% probability interval for each of the eight most frequently occurring letters. How many deciphering maps are possible for these eight letters, if we only consider symbols that have occurrences within the 95% probability interval of each letter? How many are possible if you allow all symbols?

Solution: Using the method described above we find the following results:

	Expected Values		
Letter	**Low Value**	**High Value**	**Possible Letters**
E	10	26	V,G,H,M,R,S
T	6	21	G,H,M,R,S,L,I,O,Y,Z
A	5	19	G,H,M,R,S,L,I,O,Y,Z,W
O	4	17	G,H,M,R,S,L,I,O,Y,Z,W,N,T,X
I	4	17	G,H,M,R,S,L,I,O,Y,Z,W,N,T,X
N	4	17	G,H,M,R,S,L,I,O,Y,Z,W,N,T,X
S	4	16	G,H,M,R,S,L,I,O,Y,Z,W,N,T,X
H	3	16	G,H,M,R,S,L,I,O,Y,Z,W,N,T,X,F,U

Hence there are $6 \cdot 10 \cdot 11 \cdot 14^4 \cdot 16 = 405{,}672{,}960$ ways to represent the eight letters using the possible symbols generated by the 95% probability intervals. There are $26!/(26-8)! = 62{,}990{,}928{,}000$ ways to represent the eight letters allowing all possible symbols.

World War II

During World War II, many allied successes were attributed to the advantages gained by cracking the German Enigma and Japanese PURPLE cipher systems. Other tactical advantages were also gained by cryptanalysts working at operational and tactical levels for the allies. In North Africa, the 129th Signal Company (Radio Intelligence) discovered the Nazi's were withdrawing from Kasserine Pass. Later in North Africa the 128th gave advance warning of several attacks. In Italy, radio reconnaissance units provided "outstanding" intelligence support to VI Corps. Cryptanalysts working for General Omar

Bradley's 12th Army Group, the 849th Signal Intelligence Service (S.I.S.), read a German message at Normandy which allowed Bradley to respond to a strong counterattack against one of his vulnerable positions. Later, the 849th S.I.S. working in the Ardennes was able to decipher German message traffic disclosing the movement of armored divisions in the region of the Ardennes Forest. (Not all intelligence information is evaluated properly.) Another S.I.S. unit working for General Patton's 3rd Army, deciphered a German message that contributed to the 3rd Army inflicting heavy losses on the German 5th Parachute Division at Bastogne. Cryptanalysts continued to provide valuable support until the conclusion of the War. [8] It is certain that cryptanalysts will play a major role in future wars.

Exercises

1. Find the expected number of occurrences of the letter 'm' in a sample size of 323 from the novel Moby Dick. What is the variance for the number of occurrences of the letter m?

2. Find a 95% probability interval for the occurrences of the symbol representing the letter 's', in a sample of encrypted text of size 50.

3. Find a 90% probability interval for the occurrences of the symbol representing the letter 'o', in a sample of encrypted text of size 60.

4. Consider the cipher text of a quote from Moby Dick:

> GSVKV ZKV HLNV VMGVKIKRHVH RM DSRXS Z XZKVWFQ URHLKUVKQRMVHH RH GSV GKFV NVGSLU.

Which symbol most likely represents the letter 'e'? Which symbol(s) most likely represents the letter 'o'?

5. Using the cipher text from problem 4, develop 95% probability intervals for each of the 8 most frequently occurring letters.

6. Consider the cipher text of a quote from Moby Dick:

ACN MPNDFMFAAAFKB JAKKNP FK WCFDC DAMAAFK ACAR CAE OUFAANE ACN QAJUNI NKENPRY LS ILKELK, CAE KLA RNNK UKAAANKENE WFAC QLJN QJAII VFLINKDN AL CFQ LWK MNPQLK. CN CAE IFBCANE WFAC QUDC NKNPBY UMLK A ACWAPA LS CFQ RLAA ACAA CFQ FVLPY INB CAE PNDNFVNE A CAIS-QMIFKANPFKB QCLDH.

Which symbols most likely represent the letter 's'?

7. Would you expect the associated probabilities to change for each letter if we consider other languages besides English? Explain.

References

[1] Gilbert, Martin, The First World War, Henry Holt, Inc., 1994.

[2] Kahn, David, The Code Breakers: The Story of Secret Writing, The Macmillian Company, 1967, pp. 333-347.

[3] Kobliz, Neal, A Course in Number Theory and Cryptography, 2nd Ed., Springer-Verlag, 1994, pg. 54.

[4] Schneier, Bruce, Applied Cryptography, John Wiley and Sons, Inc., 1994, pg. 281-285.

[5] Brualdi, Richard A., Introductory Combinatorics, 2nd Ed., Elsevier Science Publishing Co., Inc., 1992, pp. 167-173.

[6] Kahn, pg. 95-98.

[7] Melville, Herman, Moby Dick, (electronic text version prepared by Prof. Eugene F. Irey at the University of Colorado from the Henricks House Edition), Internet.

[8] Kahn, pp. 507-510.