CHAPTER 2

SECURITY OF CRYPTOGRAPHIC SYSTEMS

Section I

Requirements of Military Systems

2-1. Practical Requirements

Military cryptographic systems must meet a number of practical considerations.

a. An ideal cryptographic system for military purposes is a single all-purpose system which is practical for use from the highest headquarters to the individual soldier on the battlefield. It is secure no matter how much message traffic is sent using the system. It is easy to use without special training. It presents no logistics problems in keeping the users supplied with the system's keys. It operates under all weather conditions, on all means of communication, and in the dark. Little of value is compromised if the enemy captures the system. No system exists that meets all these requirements.

b. Cryptographic system selection for military use depends on much more than its degree of security. While protecting information from unfriendly eyes, a system must still allow communications to take place rapidly, to be reliable, and to be usable by all who need to conduct communications. It must be usable under all conditions that the communications must take place. For example, a system requiring an hour of pains-taking encryption would go unused by a combat military force on the move. A system that has no tolerance for errors in its use would be inappropriate for soldiers under fire in severe weather conditions. A system that only supports a low volume of messages would be inappropriate for a major message center handling thousands of messages daily. A system that requires expensive, sophisticated equipment would be inappropriate for a military force that can barely afford to buy ammunition. No single system meets all the requirements of security, speed, reliability, flexibility, and cost. The need for security must be balanced against the practical requirements when systems are selected for use. Breakable systems are found today, despite technological advances, because of these practical requirements.

2-2. Security Requirements of Military Systems

When security must be balanced against practical considerations, how much security is enough security?

a. Almost any cryptographic system, given enough time and resources can eventually be solved. The only exception to this is a system which uses absolutely random changing keys with every character encrypted and never repeated. Such a system can be achieved under very limited conditions, but is in practice impossible on any large scale.

b. Even the most sophisticated machine or computer based cryptographic system cannot produce random, nonrepeating keys. The requirement for each communicating machine to generate the same keys prevents truly random keys. At best, a machine system can produce keys by so sophisticated a process that it appears to be random and resists efforts to recover the key generation process.

c. Given the practical considerations, a military system is expected to delay successful analysis, not prevent it. When the system is finally solved, the information obtained has lost most of its value.

2-3. Factors Affecting Cryptographic Security

As discussed above, given enough time and resources, almost any system can be solved. No nation has unlimited resources to devote to the effort. If the potential intelligence payoff is timely enough and valuable enough and the resource costs reasonable, the necessary resources will usually be devoted to the effort. A number of factors affect the vulnerability of cryptographic systems to successful cryptanalytic attack.

a. The most obvious factor is the cryptographic soundness of the system or systems in use. Systems with minimal key repetition and limited orderly usage patterns provide the most resistance.

b. The volume of traffic encoded or enciphered with a given set of keys affects system security. The longer the keys are used without change, the more chance an analyst has of finding exploitable repetition and patterns to build the attack upon.

c. The discipline of system users can play a major role in system security. A system that is very sound when used correctly can often be quickly compromised when rules are broken. An obvious example is when a user retransmits a message in the clear that has also been transmitted in encrypted form. When it is recognized, the comparison of the plaintext message with its encrypted form makes key recovery much easier. Other typical examples of undisciplined usage are–

- To mix plaintext and encrypted text in the same transmission.
- To use the same keys longer than prescribed.

- To make unauthorized changes or simplifications to the system.
- To openly discuss the contents of an encrypted message.
- To openly discuss the system or its keys.

d. The amount of collateral information available about the message sender and the situation under which the message was sent affect the security of a system. The more that is known about the sender, the more likely the contents of a message can be determined.

Section II
Cryptanalytic Attack

2-4. Role of Cryptanalysts in Communications Intelligence Operations

Communications intelligence (COMINT) operations study enemy communications for the purpose of obtaining information of intelligence value. COMINT includes the collection, processing, evaluation, and reporting of intelligence information gathered from enemy communications. When cryptanalysts is successful on a timely basis, it provides the most direct indication of the enemy's intentions. Cryptanalysis is most likely to be successful when other COMINT techniques are also productive. Collection of communications signals, transmitter location and identification, traffic analysis, and translation and analysis of cleartext transmissions all play a part in the production of COMINT.

2-5. Comparison Between Cryptanalysts and Traffic Analysis

Cryptanalysis is the study of encrypted messages. These messages, when passed as part of radio communications, or traffic, are considered the internals of the communications. Traffic analysis is the study of the externals of the communications.

a. The externals of a communications include the following:

- Call signs and call words.
- Call up procedures between operators.
- Radio frequencies.
- Times of transmissions and total volume of traffic.
- Routing information indicating where a message is to be sent.

- Chatter between radio operators.
- Serial numbers or other filing information.
- Indications of precedence or importance of the messages.
- Indicators designating what cryptographic systems or what key settings are in use.

These externals can be a rich source of information about an enemy, regardless of encrypted message recovery. The systems that communicators use to provide this external information can give substantial clues to unit type, organization, and the purpose of communications.

b. The last category of externals mentioned above, indicators of the cryptographic systems or keys in use, is of particular interest to both the traffic analyst and the cryptanalyst. For the traffic analyst, the indicators help establish patterns of usage which give clues to the enemy's organization and structure. For the cryptanalyst, the indicators help group messages into those encrypted by the same system or keys. In some cases, they may even aid directly in the solution of the system.

2-6. Steps in Cryptanalysis

The solution of nearly every cryptogram involves four basic steps–

- Determination of the language used.
- Determination of the general system used.
- Reconstruction of the specific keys to the system.
- Reconstruction of the plaintext.

a. Determination of the language used normally accompanies identification of the sender through traffic analysis or radio direction finding. If these forms of support are unavailable, or if an enemy uses several languages, the determination of the language may have to be made at a later stage of analysis.

b. Determination of the general system can come from several sources, such as–

- A detailed study of the system characteristics, aided where necessary by character frequency counts, searches for repeated patterns, and various statistical tests. The study can extend beyond single messages to searching for patterns and repetitions between different messages with similar characteristics. This single step of system determination can be the most time consuming part of the analysis.
- Past history of system usage by the sender. In most cases, the user does not change systems regularly but uses the same system or set of systems from one day to the next. The specific keys may change regularly, but the general systems remain unchanged except at longer intervals.

- System indicators included with the traffic. Whenever the user has a choice of systems or a choice of keys within the system, the choice must be made known to the receiving cryptographer. The choice is usually communicated by some form of indicators, which can appear within the text of a message or as part of the externals. When the indicators reveal the choice of system, they are called system indicators or discriminants. When they denote specific frequently changing keys to the system, they are called message indicators. Once you learn just how indicators are used from day to day, they can provide a substantial assist to cryptanalysts.

c. Reconstruction of the specific keys to the system is an important step. Although the following step of plaintext recovery produces the most intelligence information, the full key reconstruction can speed recovery of future messages. The approach used to recover keys will vary greatly from system to system.

d. Reconstruction of the plaintext, although listed as the final step, will usually proceed simultaneously with the key reconstruction. Either step can come first, depending on the system and situation. Partial recovery of one aids in the recovery of the other. The two steps often proceed alternately, with each recovery of one helping in recovery of the other until a full solution is reached.

Section III
Analytic Aids

2-7. Analytic Aids to Identification and Solution

There are a number of aids to identification and solution available to help you as a cryptanalyst. By preparing character frequency counts, performing statistical tests, and recording observed repetitions and patterns in messages, you can compare the data to established norms for various systems and languages. The appendixes to this manual include charts, lists, and tables of normal data for the English language. Similar data are available for other languages. The counting of character frequencies, performance of statistical tests, and search for repetition and patterns can be done manually or with computer assistance, where available. This section outlines the aids that apply to many types of systems. Procedures that apply to specific systems are explained in individual sections.

2-8. Language Characteristics

Each language has characteristics that aid successful cryptanalysts.

a. The individual letters of any language occur with greatly varying frequencies. Some letters are used a great deal. Others are used only a small percentage of the time. In English, the letter *E* is the most common letter used. It occurs about 13 percent of the time, or about once in every eight letters. In small samples, other letters may be more common, but in almost any sample of 1,000 letters of text or more, E will be the most frequent letter. In other languages, other letters sometimes dominate. In Russian, for example, O is the most common letter. The eight highest frequency letters in English, shown in descending order, are *E, T, N, R, O, A, I* and *S.* The eight highest frequency letters make up about 67 percent of our language. The remaining 18 letters only make up 33 percent of English text. The lowest frequency letters are J, K, Q, X, and Z. These five letters makeup only a little over 1 percent of English text. The vowels, A, E, I, O, U and Y, make up about 40 percent of English text. In many cryptographic systems, these frequency relationships show through despite the encryption. The analysis techniques explained in the following chapters make repeated use of these frequency relationships. In particular, you should remember the high frequency letters, ETNROAIS, and the low frequency letters, JKQXZ, for their repeated application. The word *SENORITA,* which includes the high frequency letters is one way to remember them. Some people prefer to remember the pronounceable ETNORIAS as a close approximation of the descending frequency order. Choose the method you prefer. The high frequency letters are referred to frequently.

b. Just as single letters have typical frequency expectations, multiple letter combinations occur with varying, but predictable frequencies, too. The most common pair of letters, or digraph, is EN. After EN, RE and ER are the most common digraphs. There are 676 different possible digraphs in English, but the most common 18 make up 25 percent of the language. Appendix A lists the expected frequencies of English language digraphs. Some cryptographic systems do not let individual letter frequencies show through the encryption, but let digraphic frequencies come through. The systems explained in Part Three of this manual show this characteristic.

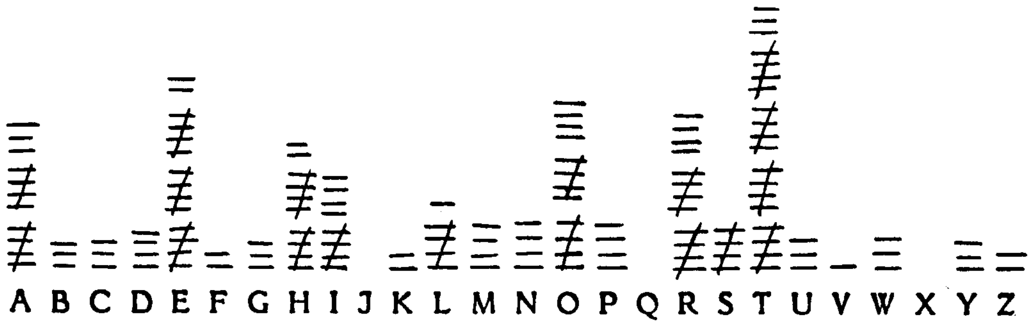
c. Appendixes B and C list frequency expectations for sets of three letters (trigraphs) and four letters (tetragraphs). Each of these can be useful when studying cryptograms in which three and four letter repeated segments of text occur.

d. Repeated segments of two to four letters will often occur because they are common letter combinations, whether or not they are complete words by themselves. Longer repeated segments readily occur when words and phrases are reused in plaintext. When words are reused in plaintext, they may or may not show up as repeated segments in ciphertext. For a word to show through as a repeat in ciphertext, the same keys must be applied to the same plaintext more than once. Even complex systems which keep changing keys will sometimes apply the same keys to the same plaintext and a repeated ciphertext segment will result. Finding such repeats gives many clues to the type of system and to the plaintext itself. The search can extend beyond single messages to all messages that you believe may have been encrypted with the same set of keys. If computer support is available to search for repeats for you, a great deal of time can be saved. If not, time spent scanning text to search for repeats will reward you for your time when you find them.

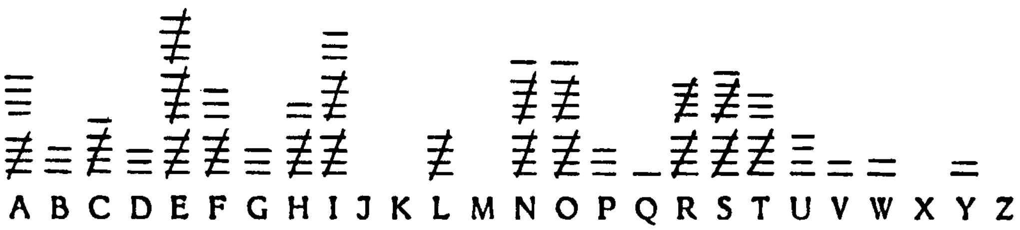
2-9. Unilateral Frequency Distribution

The most basic aid to identification and solution of cipher systems is the unilateral frequency distribution. The term unilateral means one letter at a time. A unilateral frequency distribution is a count of all the letters in selected text, taken one letter at a time.

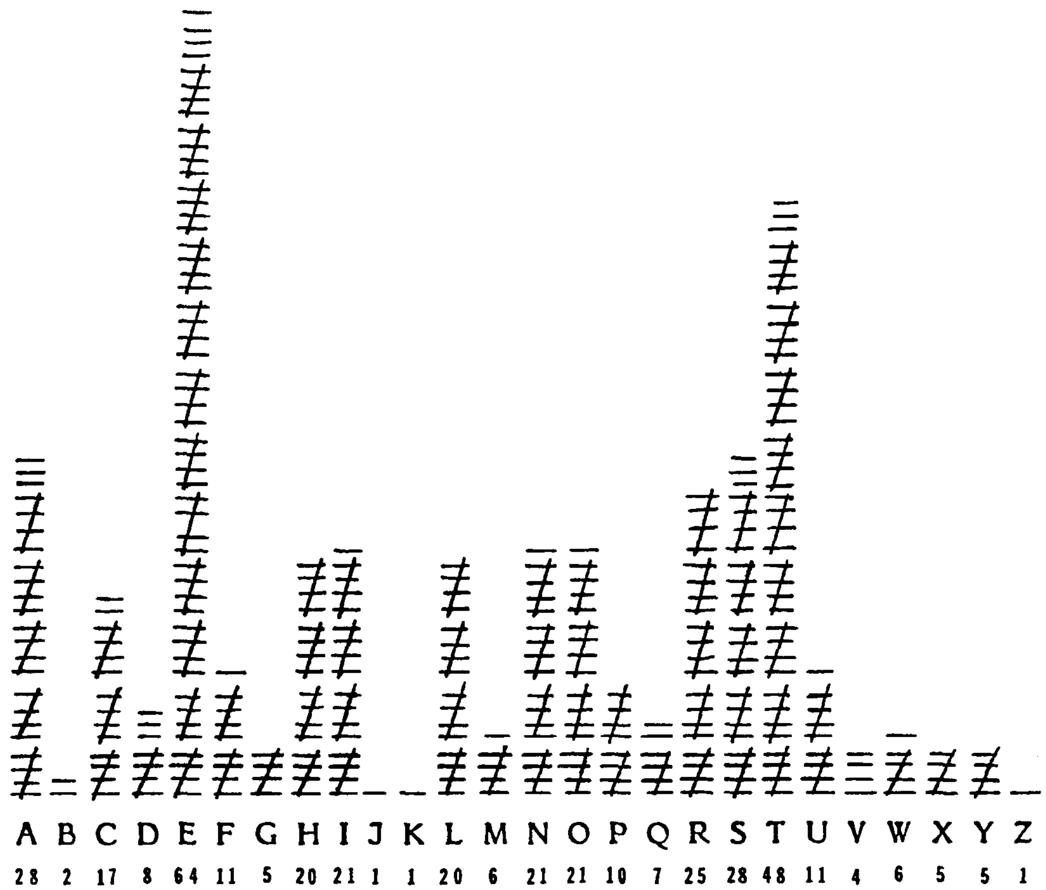
a. The customary method of taking the distribution is to write the letters A through Z horizontally and mark each letter of the cryptogram with a dash above or below the appropriate letter. Proceed through the message from the first letter to the last, marking each letter in the distribution. Avoid the alternate method of counting all the As, Bs, Cs, and so forth, which is very subject to errors. For convenience, each group of five is crossed off by a diagonal slash. The unilateral frequency distribution for the first sentence in this paragraph is shown below.



For comparison, the next example shows the frequency count for the fourth and fifth sentences in paragraph 2-9a.



b. Although individual letter frequencies differ, the pattern of high and low frequency letters is quite similar. The letters that stand above the others in each tally are, with few exceptions, the expected high frequency letters—ETNROAIS. The expected low frequency letters, JKQXZ, occur once or twice at most. Even in as small a sample as one or two sentences, expected patterns of usage start to establish themselves. Compare this to a frequency count of all letters in this paragraph.



c. When a larger sample is taken, such as the above paragraph, the letters occur much closer to the expected frequency order of ETNROAIS. As expected, E and T are the two highest frequency letters. but the next series of high frequency letters in descending order of occurrence, ASRINO, differs slightly from the expected order of NROAIS. It would take a sample thousands of letters long to produce frequencies exactly in the expected order. Even then, differences in writing style between a field manual and military message texts could produce frequency differences. For example, the word *the* is often omitted from military message traffic for the sake of brevity. More frequent use of *the* raises the expected frequency of the letter H.

2-10. Letter Frequencies in Cryptograms

As different cipher systems are explained in this manual, the ways in which letter frequencies can be used to aid identification and solution will be shown. Some basic considerations should be understood now.

a. In transposition systems, the letter frequencies of a cryptogram will be identical to that of the plaintext. A cryptogram in which the ciphertext letters occur with the expected frequency of plaintext will usually be enciphered by a transposition system.

b. In the simplest substitution systems, each plaintext letter has one ciphertext equivalent. The ciphertext letter frequencies will not be identical to the plaintext frequencies, but the same numbers will be present in the frequency count as a whole. For example, if there are 33 Es in the plaintext of a message, and if E is enciphered by the letter K, then 33 Ks will appear in the ciphertext frequency count.

c. More complex substitution cipher systems, such as the polyalphabetic systems in Part Four of this manual, will keep changing the equivalents. E might be enciphered by a K the first time it occurs and by different cipher letters each time it recurs. This will produce a very different looking frequency count.

d. To illustrate the differences in appearance of frequency counts for different types of systems, examine the four frequency counts in Figure 2-1. Each one is a frequency count of the message listed above it. The four messages are different, but each has the same plaintext. The first shows the plaintext and its frequency count. The second shows the frequencies of the same message enciphered by a transposition system. The third shows a simple substitution system encipherment. The fourth shows a polyalphabetic substitution encipherment.

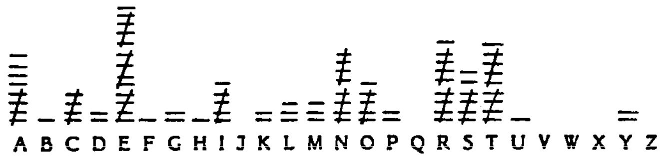
2-11. Roughness

The four examples in Figure 2-1 show another characteristic of frequency counts which is useful in system identification. The first three distributions all contain the same letter frequencies. In the first two, the plaintext and the transposition examples, there are 16 Es. In the third, where E has been replaced by W, there are 16 Ws. Where there were 9 As, there are now 9 Ls. Where there was 1 K, there is now 1 C. The first three distributions show the same wide differences between the highest frequency letters and the lowest. The fourth distribution is very different. The distribution lacks the wide differences between the highest and lowest frequency letters. Where the first three showed distinct highs and lows, or peaks and troughs, in the distributions, the fourth is relatively flat.

a. Frequency counts which show the same degree of difference between peaks and troughs as plaintext are considered to be rough distributions. Systems which suppress the peaks and troughs of plaintext letters by changing their equivalents produce flatter distributions. If letters were selected randomly from the 26 letters of the English alphabet, the resulting distribution would look very much like the fourth example. Random selection will not produce a perfectly level distribution, but it will appear quite flat in comparison to plaintext.

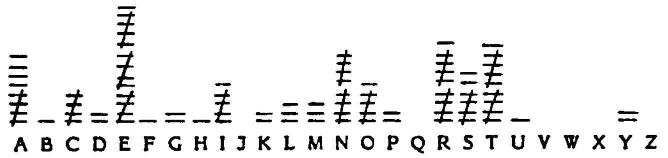
Plaintext:

AERIAL RECONNAISSANCE REPORTS ENEMY REINFORCEMENTS ESTIMATED
AT BATTALION STRENGTH ENTERING YOUR SECTOR PD CLARKE



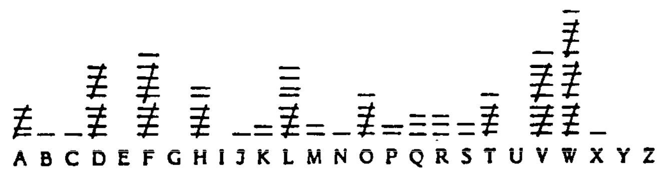
Transposition:

ANRME MTNNO ENEYM AAGGR RAPRE TLTYP IIOEN EIHOD ASRIT DOEUC
LSTNS ANNRL RASFE TSTSA ENEOS BTEER CCNRT ARRCK OEECI TEITE



Simple substitution:

LWVOL QVWAT DDLOH HLDAW VWPTV FHWDW RSVWO DNTVA WRWDF HWHFO
RLFWK LFJLF FLQOT DHFVW DMFBW DFWVO DMSTX VHWAF TVPKA QLVCW



Polyalphabetic substitution:

TARAB CZPNW TNNLL ZEFNM KLNHF OWWQM PEPVM NKRXK QNPRB FXZXE
MBXEO LFJML RWPZS GZXSS EUZYS IXWRV QZFSG FEITT HYHRW EGIKF

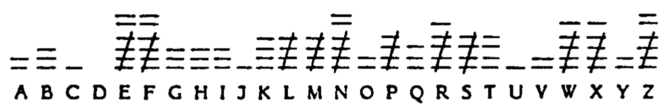


Figure 2-1. Frequency count comparison.

b. The simplest substitution systems tend to produce rough distributions. The most secure tend to produce flat distributions. Many other systems tend to fall in between. You can use the degree of roughness as one of the aids to system identification.

2-12. Coincidence Tests

Judging whether a given frequency distribution has the same degree of roughness as plaintext or random text is not easy to do by eye alone. To help you make this determination, a number of statistical tests have been developed for your use. The tests are based in probability theory, but you can use the tests whether or not you understand the underlying theories. The most common tests are called coincidence tests.

a. If you pick any two letters from a message, compare them together, and they happen to be the same letter, they are said to coincide. A comparison of the same letters, for example, two As is a coincidence. This comparison can be made of single letters or pairs of letters or longer strings of letters.

b. If you compare two single letters selected at random from the English alphabet, the probability of their being the same is 1 in 26. One divided by 26 is .0385. Expressed as a percentage, 1/26 is slightly less than 4 percent. You would expect to find a coincidence 3.85 times on the average in every 100 comparisons.

c. If you select two letters from English plaintext, however, the probability of their being the same is higher than 1 in 26. Frequency studies have shown that the probability of a coincidence in English plaintext is .0667. In other words, in every 100 comparisons, you would expect to find 6.67 coincidences in plaintext. Each language has its own probabilities, but similar traits occur in each alphabetic language.

d. Different coincidence tests use different methods of comparing letters with each other, but each rests on the probabilities of random and plaintext comparisons. The actual number of coincidences in a cryptogram can be compared with the random and plaintext probabilities to help make judgments about the cryptogram.

2-13. Index of Coincidence

A common way of expressing the results of a coincidence test is the index of coincidence (XC). The index of coincidence is the ratio of observed coincidences to the number expected in a random distribution. For plaintext, the expected index of coincidence for single letters in English is the ratio of .0667 to .0385, which is 1.73.

2-14. Monographic Phi Test

The most common coincidence test is the monographic phi test, which provides a mathematical way of measuring the roughness of a frequency count. *Monographic* is a fancy synonym for *one letter.* The term monographic distinguishes the test from the digraphic phi test, performed on two letter pairs, and other forms of the phi test. Phi is the English spelling of the Greek letter ϕ. The monographic phi test is based on the coincidence probabilities that occur when every letter in a cryptogram is compared with every other letter in the cryptogram.

a. Fortunately, the phi test can be calculated without actually comparing every letter with every other letter. Both the total number of comparisons and the total number of coincidences can be calculated from the frequency count.

b. The total number of comparisons when every letter is compared with every other letter is the total number of letters multiplied by the total number minus one. Expressed as a formula, it looks like this–

$$\text{Comparisons} = N\,(N-1).$$

c. Since one out of every 26 comparisons in a random distribution is expected to be a coincidence, the formula for the expected random value of phi is as follows:

$$\phi r = \frac{N\,(N-1)}{26}$$

or

$$\phi r = .0385\,N\,(N-1).$$

d. The expected value for plaintext coincidences is–

$$\phi p = .0667\,N\,(N-1).$$

e. Just as the total number of comparisons is N (N – 1), the total number of coincidences for each letter is f (f – 1), where f is the frequency of the individual letter. The total number of coincidences is the sum of the coincidences for all the letters. The total number of coincidences is labeled phi observed or øo, and can be expressed as either–

$$\phi o = \phi A + \phi B + \phi C + \ldots + \phi Z$$

or

$$\phi o = \Sigma f\,(f-1).$$

(The Greek letter sigma (Σ)is used to mean *sum of.)*

f. To calculate ϕo, take each letter frequency greater than 1 and multiply it times the frequency minus 1, as the formula suggests. (You can ignore letters with a frequency of 1, because they will be multiplied by 0.) Then add the results of all the multiplications.

g. The index of coincidence for the phi test is called the delta IC. The delta IC is the ratio of phi observed to phi random. It can be expressed using the Greek letter delta (Δ).

$$\Delta IC = \frac{26\ \Sigma f\ (f-1)}{N\ (N-1)}$$

h. The results of a phi test can be expressed in terms of ϕo, ϕp, and or as the ΔIC. Where computer support is available to perform the calculations, the ΔIC is the form usually shown. Where paper and pencil methods are used, either form may be used. Both methods are shown in the next example.

Letters:	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
f:	3	3	0	7	2	1	1	4	0	0	1	0	0	0	4	1	6	3	0	4	1	0	5	1	0	3
f-1:	2	2		6	1			3							3		5	2		3			4			2
f(f-1):	6	6		42	2			12							12		30	6		12			20			6

$$\begin{aligned}
\phi o &= \Sigma f\,(f-1) \\
&= 6 + 6 + 42 + 2 + 12 + 12 + 30 + 6 + 12 + 20 + 6 \\
&= 154
\end{aligned}$$

$$\begin{aligned}
\phi p &= .0667\ N\,(N-1) \\
&= .0667 \times 50 \times 49 \\
&= 163
\end{aligned}$$

$$\begin{aligned}
\phi r &= .0385\ N\,(N-1) \\
&= .0385 \times 50 \times 49 \\
&= 94
\end{aligned}$$

$$\begin{aligned}
\Delta IC &= \phi o / \phi r \\
&= 154/94 \\
&= 1.64
\end{aligned}$$

2-15. Interpreting the Phi Test

The previous example showed results close to the expected value for plaintext. This indicates the frequency count it was based on had the same approximate degree of roughness as expected for plaintext. It does not show that it was plaintext or that it was enciphered in a simple substitution system, although the latter is possible. It must be considered as just one piece of evidence in deciding what system was used.

a. In plaintext of 50 to 200 letters, the delta IC will usually fall between 1.50 and 2.00. Shorter text can vary more, and longer text will be consistently closer to 1.73. Since simple monoalphabetic systems have the same frequency distribution as plaintext, these simple systems follow the same guidelines as plaintext.

b. Random text centers around a IC of 1.00 but is subject to the same variability as plaintext. Small samples of under 50 letters vary widely. Samples in the 50 to 200 letter range will usually fall between 0.75 and 1.25. Larger samples approach 1.00 more consistently.

c. Polyalphabetic systems tend to resemble random text, and the more different alphabets that are used, the more likely the ΔIC is to approach 1.00.

d. The four frequency counts in Figure 2-1 follow these guidelines closely. Each one is 100 letters long. The first three, the plaintext, the transposed text, and the simple monoalphabetic substitution each have a ΔIC of 2.00. The fourth example, the polyalphabetic substitution example, has a ΔIC of 1.05. The system used in the example has 26 different alphabets, and the underlying plaintext frequencies have been thoroughly suppressed.