Mr. Friedman:

1. This paper used to be CONFIDENTIAL and registered in its previous edition. We sent a letter to the Navy in April requesting this paper, and we still have no answer: this is indicative of the state of chaos existing in their training section.

2. There are two footnotes (pp. 7, 12) that refer to Gaines' "Elementary Cryptanalysis". This is a sad commentary on the Navy's training resources in that they make their only references to a book available on the public market: this exaggerates the worth and importance of the Gaines book, and minimizes the resources of the Navy. Certainly a passing mention could have been made of the Army's texts on cryptanalysis, especially since the Navy has no texts of its own. . . . + theory involved-ures is adequately explained in Army texts.

3. Other than the poor terminology employed, and the plethora of mathematical eyewash that makes a simple subject difficult, this paper is potentially very good, after substantial editing and liberal re-writing.

- Capt G.

WFF

~~RESTRICTED~~

THE INDEX OF COINCIDENCE

FOREWORD

The subject of this pamphlet is coincidence.

The student may well ask, "What is coincidence and what applications has it?"

"Coincidence" as the term is used here may be defined as a recurrence of a letter in the same place, or in a corresponding place, as when two texts are lined up one under the other, letter for letter.

This mathematical evaluation assists the cryptanalyst first in preparing his material for attack, and later in the actual attack itself. It assists specifically in answering the following questions.

1) How much like random, or how different from random, is this text?

2) How similar are these texts?

3) How significant is this variation from random?

4) How significant is this similarity?

I SIMPLE MONOGRAPHIC COINCIDENCE

The test of coincidence is the evaluation of the coincidences of letters, or of digraphs, etc., between two or more messages, or within the same message.

The coincidence or "pairing" test may be consolidated into one final number or "statistic". That statistic is called the "index of coincidence" and is defined as the ratio of the actual coincidences to the coincidences to be expected from chance (coincidences in random text). For English text the expected I.C. is 1.75. For most European languages the expected I.C. is about 2.ØØ. For random text the expected I.C. is 1.ØØ.

Assume two pages of cipher text based on a complex cipher which will give a "flat" frequency table for the entire message. Select a letter at random (say the 3rd) from one page and another from the other page (say the 3rd also).

There is 1 chance in 26 of the first letter's being an "A"
There is 1 chance in 26 of the second letter's being an "A"
There is 1 chance in 676 of both letter's being "A"
There is also 1 chance in 676 of both letter's being "B"

Therefore, the chances of both letters being the same letter (in a chance selection of cipher text) are:

26 chances in 676, or 1 chance in 26, or 3.8462%.

If we select many pairs of cipher letters, the average number of identical letters to be expected "in the long run" will be 3.846% (or 1/26) of the total number of possible coincidences. We call this number the "Expected Coincidence due to Chance" (random text).

With English text it is different. Take two pages of English text. Make a chance selection from each page.

There are about 13Ø chances in 1,ØØØ of the first letter's being an "E"
There are about 13Ø chances in 1,ØØØ of the second letter's being an "E"
There are about 16,9ØØ chances in 1,ØØØ,ØØØ of both letters' being an "E"
Likewise, there are 8,464 chances in 1,ØØØ,ØØØ of both being "T"
6,4ØØ chances in 1,ØØØ,ØØØ of both being "N", etc.

(See table following).

Table

Text Letter (Telegraphic Text)	Chances in 1,000 of 1st letter's being this ltr.	Chances in 1,000 of 2nd letter's being this ltr.	Chances in 1,000,000 of both letters' being this letter
E	130	130	16,900
O	75	75	5,625
A	74	74	5,476
I	73	73	5,329
N	80	80	6,100
R	76	76	5,776
S	61	61	3,729
T	92	92	8,464
D	42	42	1,764
H	34	34	1,156
L	36	36	1,296
C	31	31	961
M	25	25	625
U	26	26	676
P	27	27	729
F	28	28	784
G	16	16	256
Y	19	19	361
B	10	10	100
V	15	15	225
W	16	16	256
K	4	4	16
J	2	2	4
Q	2	2	4
X	5	5	25
Z	1	1	1
Any letter	1,000	1,000	66,930

Finally there are 66,930 chances in 1,000,000 (the sum of the chances for the individual letters) of both letters' being the same plain text letter in a chance selection. Therefore, if we select many pairs of plain text letters, the average number of identical letters to be expected "in the long run" will be 6.693% (about 1/15) of the total number of Possible Coincidences.

We may call this number the Expected Coincidences in English Text

In actual practice we are concerned with the coincidences between our two texts, or within our alphabet, etc. The tally or count of these coincidences we call the Actual Coincidences.

To permit comparisons between results obtained from texts of varying amounts, it is most convenient to convert to an index number. We call this the Index of Coincidence and use the abbreviation I.C. or ι.

By definition $\iota = \frac{\text{Actual Coincidences}}{\text{Expected Coincidences due to Chance.}}$

The expected I.C. for English (or mono-alphabetical cipher text) is:

$$\frac{.06693}{.03846} = 1.75, \text{ approximately.}$$

The actual I.C. of unknown cipher text may take almost any value but in practice the range will generally extend from about .80 to about 2.00 (simple monographic index of coincidence).

The value of the index of coincidence for a given English text will depend on the actual distribution of letters in that text. Repetitions in short texts will increase the index of coincidence. Unrelated text (that is, text with few repetitions) will give an I.C. approaching the theoretical 1.75. As the expected number of chance coincidences is based on a flat frequency (where each cipher letter is ultimately used the same number of times) any cipher text that differs radically from such frequency distribution will have a correspondingly higher I.C. This is especially noticeable in short cipher texts where the frequency table has not had an opportunity to "flatten out".

The mono-graphic I.C. of English naval text will increase with small amounts of text to 1.80 - 2.00 (as compared with the theoretical 1.75) and small amounts of random text will give I.C.'s of 1.10 - 1.20 (as compared with the theoretical 1.00). The amount of excess attributable to the sample size will be discussed later, under "standard deviation".

For most European languages the expected I.C. is higher than in English, due to the more irregular letter distribution of their normal alphabets, namely:

Language	I.C.
Random text	1.0
English	1.7
Italian	1.9
Spanish	2.0
French	2.0
German	2.0

II POLYGRAPHIC COINCIDENCE

In addition to the simple monographic index of coincidence (ι), there are occasions when the digraphic index of coincidence (ι_2), trigraphic I.C. (ι_3), tetragraphic I.C. (ι_4), pentagraphic I.C. (ι_5), etc., can be used to advantage. They are derived from the normal digraphic (trigraphic, etc) frequency tables in the manner indicated in paragraphs 3 to 7.

Expected values for these simple polygraphic indices of coincidence are as follows:

Language	ι	ι_2	ι_3	ι_4	ι_5
Random text	1.00	1.00	1.00	1.00	1.00
English	1.75	4.75	27.89*	X	X
Italian	1.92	5.68	X	X	X
Spanish	2.02	6.29	X	X	X
French	2.02	6.29	X	X	X
German	1.98	6.57	X	X	X

Notes: X = Not computed.
* = (Computed from the only known trigraphic table. The correct index might vary widely from this estimate.)

In practice the actual polygraphic I.C.'s will usually run higher than their theoretical values, and a repeated word or two in short texts will made them sky rocket. As typical examples, we have taken the plain text of four problems in the elementary and secondary courses and computed the various I.C.'s (from ι, to ι_5, that is the monographic, digraphic, trigraphic, tetragraphic and pentagraphic indices of coincidence).

Text	ι	ι_2	ι_3	ι_4	ι_5
Expected random	1.00	1.00	1.00	1.00	1.00
Expected plain	1.75	4.75	27.89	?	?
Problem No. 1	1.80	5.23	29.11	427.	7240.
Problem No. 2	2.00	7.73	66.04	1062.	14900.
Problem No. 3	1.91	5.60	42.04	666.	12070.
Problem No. 4	1.74	4.90	31.70	456.	9190.

III THEORETICAL RECAPITULATION

Suppose we have a language for which we know the overall proportions of the letters are P_1, P_2, - - -, P_c.

$$\sum_{1}^{c} p_i = 1. \qquad (1)$$

Suppose further that we have two pieces of text from this language and line them up one above the other, and then count coincident letters. What is the expected number?

At a particular place the probability of a coincidence involving the i th letter is P_i^2. There fore, the C cases being mutually exclusive, the probability of an incidence is

$$\sum_1^C p_i^2 . \qquad (2)$$

If the length of overlap is N, then the expected number of incidences is

$$N \sum_1^C p_i^2$$

If the text is such that $p_i = p_j = 1/C$, we will refer to it as "flat", or "random". The probability of an incidence is $\sum_1^C 1/C^2 = 1/C$, and the expected number is $N \frac{1}{C}$.

The ratio of the number found in a comparison to that expected is called the "index of coincidence", ι

$$\iota = g/N/C = cg/N . \qquad (3)$$

The expected value γ of ι for our language is given by taking the expected value of $g = N \sum_1^C p_i^2$ over the expected value for flat text or

$$\gamma = \frac{N \sum_1^C p_i^2}{N/C} = C \sum_1^C p_i^2 \qquad (4)$$

Notice that the expected value of the I.C. for flat text is 1.

IV PRACTICAL APPLICATIONS

(A) TO DETERMINE WHETHER TWO MESSAGES ARE IN THE SAME KEY

During U.S. Fleet Problem V (1925) the Battle Fleet used a cipher of their own design. A total of 13 messages in this cipher were submitted to the Code and Signal Section for attack. Although a different indicator was used in each case, it was suspected that some of the messages might be in the same key. Two messages in one key (example No. 1) and two more in another key (example No. 2) were discovered. (The messages were eventually solved).

Each message was "lined up" with each other message and the coincidences were noted. (See examples No. 1 and No. 2).

Example No. 1

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

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

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

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

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

- - - - - - - - - -
Q Q U I I V H B K Q

Coincident letters are underscored. 12 coincidences in 140 pairs of letters.

Simple monographic coincidence.

Expected Coincidences $= \frac{N}{C} = \frac{14\emptyset}{26} = 5.4$ Where N = number of units examined. C = number of cells for single letter examination = 26.

$IC = \frac{12.\emptyset}{5.4} = 2.2$ (Messages are in same key).

There is one repeated trigraph, GPZ, in the messages under examination. This coincidence indicates that the keys correspond at that point, but does not necessarily indicate that the keys correspond throughout the message. To prove coincidence of the keys throughout the two messages, we must have our coincidences spread through the messages in question. (As they were in the above example). Likewise, digraphic and trigraphic coincidences may be compared and evaluated to an index of coincidence.

For example, in the above messages, 2 coincident digraphs were found (GP and PZ) (also one coincident trigraph). In this message there were 139 digraphs and 138 trigraphs in alignment with possibilities of coincidence.

$$\frac{N}{676} = \frac{139}{676} = .2\emptyset6$$

digraphs were to be expected from chance. Two were found, giving an $IC = \frac{2}{.2\emptyset6} = 9.7\emptyset$. This value, far above the normal 4.75 index of coincidence, does not necessarily indicate the messages are in the same key. All we really know is that the two keys are identical in the second group. The extremely high I.C., 9.7Ø, is due entirely to the small amount of text involved in this example. As the amount of text decreases, the variation of the I.C. from the expected will become more pronounced, until at times it is possible that small amounts of text may give entirely false indications. This effect will be discussed more fully under "standard deviation".

Example No. 2

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

P E K D A P N U T J D D U W Q Q I T N P E X T G H T K D C L
J S D N D U J K L U Z J C S Q H I O Z H U K E G X D E P W T

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

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

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

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

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

I I D T W W C C Y B X O Y U O S Q D N O B A S H H C T D D U
H N I X Y A O J B K -

14 coincidences in 22Ø pairs of letters.

$\frac{N}{C} = \frac{22\emptyset}{26} = 8.46$ coincidences expected.

$IC = \frac{14}{8.46} = 1.66$ (almost normal for English).

These two messages probably are in the same key (and actually proved to be). Note that there are no repeated digraphs or trigraphs. Note also that coincidences are well spread out.

(B) TO DETERMINE WHETHER TWO MESSAGES "OVERLAP" IN THE SAME RUNNING KEY

Copy each message on a single line, omitting all spacing. "Line up" the messages and note coincidences. Then shift one message one place to the right and note the coincidences. Repeat this process to the end.

If the index at any point is 1.75 or higher, for mono-graphic examination, the position and the fact of the "overlap" is probable. In this application the digraphic and trigraphic indices are useful adjuncts to the monographic index.

For some purposes the fundamental unit may be taken to be a set of letters, as digraphs, trigraphs, etc. Suppose we are interested in digraphic coincidences. Then the digraphic I.C., ι_2 , will be calculated as above, noting that the size C of the alphabet is larger this time.

V RELATIONS AMONG THESE STATISTICS

The monographic and digraphic I.C.'s are not independent. For if the probabilities of the various letters are P_1, P_2, - - -, P_C, then the probability γ_{ij} of the ij^{th} digraph is $p_i\ p_j$, ignoring the cohesion of the language, and for the moment treating it like newspaper which had been cut into little pieces, one letter to a piece, and then shuffled and arranged in a line. Using this estimage of $\gamma\ i\ j$ we get the I.C.

$$\gamma_2 = C^2 \sum_{i}^{C} \sum_{j=1}^{C} \gamma_{ij}^2 = C^2 \sum_{1}^{C} \sum_{1}^{C} p_i^2\ p_j^2$$

$$= C^2 \sum_{1}^{C} p_i^2 \sum_{1}^{C} p_j^2 = \left(C \sum_{1}^{C} p_i^2 = \gamma^2 \right) \qquad (5)$$

It is true however that language has cohesion, and that each letter affects the probability of occurrence of others in its vicinity. Usually then ι^2 is in excess of the estimate ι^2 above. We will sometimes calculate the ratio

$\iota_2/\iota_2 = \chi_2$ and call it the digraphic "index of cohesion". (6)

Estimates can be made in the same way for higher I.C.s. One can show that

$$\gamma_i' = \iota_{i-1}\ \iota \quad or$$

$$\gamma_i' = \iota_{i-j}\ \iota_j \qquad (7)$$

In these equations the right members are thought of as quantities already computed; while the γ's on the left are estimates or predictions of quantities which can be computed from the definition (4).

An application of these relations occurs in the study of fractionating* systems, where as a preliminary to enciphering the text is expressed as a product of two components, and each component is enciphered separately, and then the cipher text is recombined to ordinary letters. For instance, each of 25 letters may be represented by a two digit number, where the first digit comes from 0, 1, 2, 3, 4, and the second from 5, 6, 7, 8, 9. The argument for these proceeds as for digraphs, and the I.C. of the combined text is the product of those of the fractions. However, here any significant deviation from this estimate is not adequately described as cohesion, but must be due to the dependence of the two fraction streams.

VI THE ROUGHNESS OF A SINGLE SAMPLE

We have introduced the I.C. as a measure of the match between two pieces of text. We can extend this idea now to a measure of the roughness of a single sample. Suppose we have a piece of text which we duplicate on two slips of paper and then place them one under the other for the purpose of counting coincidences. There will be one position of total coincidence, which we will rule out. If we compute the I.C. for all other positions, we will have what we call the "index of coincidence of a single sample.

*Consult Gaines "Elementary Cryptanalysis," Chapter XXII

If there are M letters in our sample, we will have looked in $N = 1/2M(M-1)$ distinct places. (8)

If the text is flat we would expect $\frac{N}{C} = \frac{M(M-1)}{2C}$ coincidences. (9)

If the text is not flat but have proportions $P_1, P_2, ---, P_c$ of letters, then there are $f_i = P_i M$ occurences of the i th letter.. In the course of our counting we will compare every letter with every other, so that the ith letter will give rise to $1/2\ f_i\ (f_i - 1)$ coincidences, or (10)

$$\sum_1^c \tfrac{1}{2}\ f_i\ (f_i - 1) \tag{11}$$

in all.

Comparing this with the expected in flat text we get the I.C.

$$\delta = \frac{\sum_1^c \tfrac{1}{2}\ f_i\ (f_i - 1)}{\tfrac{1}{2} c\ M\ (M-1)} \tag{12}$$

$$\boxed{\delta = \frac{c \sum_1^c f_i\ (f_i - 1)}{M\ (M-1)}} \tag{13}$$

In theoretical context (12) is more useful than (13), but (13) is a little simpler for computational purposes.

Notice that this formula is different from (4). This is because of the omission of the perfect hit. If M is large enough then $f_i - 1$ can be replaced by $f_i = P_i M$ and M-1 by M, so that

$$\delta = \frac{c \sum_1^c f_i^2}{M^2} = c \sum_1^c p_i^2 = \gamma \tag{14}$$

is an asymptotic expression for δ
Notice that $c \geq \gamma \geq 1$

We see that

$$\delta \frac{M-1}{M} = \gamma - \frac{c}{M} \tag{15}$$

or $\boxed{\delta(M-1) = \gamma M - c}$

$$\delta = \frac{\gamma M - c}{M-1} \quad \text{or} \quad \gamma = \frac{\delta(M-1) + c}{M} . \tag{16}$$

The error in using γ (usually more convenient) in place of (17)

δ is $\boxed{\gamma - \delta = \frac{c - \gamma}{M-1} = \frac{c - \delta}{M} .}$

This error is always positive, that is, γ is an over-estimate of δ. The error is smaller for larger values of γ, or larger values of M.

Notice that γ is a measure of the shape of the distribution only, and is independent of the sample size, as is

$$\gamma = \frac{c \sum f_i^2}{M^2} = \frac{c \sum p_i^2 M_i^2}{M^2} = c \sum p_i^2 \qquad (18)$$

since $f_i = P_i M$.

But δ does depend on the sample size M, which is a desirable characteristic, since random roughness is usually present in small samples. For smaller samples

$$\delta = \gamma - \frac{c-\gamma}{M-1} \qquad (19)$$

is seen to be smaller, thus automatically compensating to some degree for small sample errors. We will usually measure the roughness of single samples by δ, using γ as an asymptotic approximation.

VII EXAMINATION OF CIPHER ALPHABETS AND CIPHER TEXTS

The indices of coincidences discussed in the previous paragraphs may be used in analyzing the internal structure of a cipher alphabet. A message of 173 letters has a frequency table as given below:

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
6	3	0	14	2	2	10	22	6	4	8	13	1	0	14	10	0	13	2	13	0	7	19	1	2	1

We make a tally count, i.e., the number of times a letter occurs with the same frequency, thus:

Tally f	Number of Tallies n	f(f-1)/2	n(x) f(f-1)/2
0	4	0	0
1	3	0	0
2	4	1	4
3	1	3	3
4	1	6	6
6	2	15	30
7	1	21	21
8	1	28	28
10	2	45	90
13	3	78	234
14	2	91	182
19	1	171	171
22	1	231	231
			Coincidences 1000

We find that there are 1000 coincidences. Although we can not count the coincidences in the examination of a single cipher text, we can evaluate the various frequency counts into actual coincidences. Having the actual coincidences from the table (1000), we obtain the I.C. by formula (13):

$$\delta = \frac{c \sum f_i (f_i - 1)}{M (M - 1)}$$

Substituting, $\delta = \frac{1000 \times 52}{173 \times 172} = 1.75$

The index of coincidence indicates that a monoalphabetic substitution was employed.

As a second example to show the results obtained from small texts, we calculate as follows from a frequency count of 36 letters assumed to be monoalphabetic.

A	3	N	2
B		O	1
C	5	P	
D		Q	
E		R	1
F	3	S	3
G	1	T	1
H	1	U	2
I	1	V	
J		W	1
K		X	4
L	3	Y	4
M		Z	

Tally or Cells (f)	n	1/2f(f-1)	T	f^{C_3}	T	f^{C_4}	T
Ø	1Ø						
1	7	Ø	Ø	Ø	Ø	Ø	Ø
2	2	1	2	Ø	Ø	Ø	Ø
3	4	3	12	1	4	Ø	Ø
4	2	6	12	4	8	1	2
5	1	1Ø	1Ø	1Ø	1Ø	5	5
			36		22		7

$$IC = \frac{36 \times 26 \times 2}{36 \times 35} = 1.49.$$

The alphabet in question was actually a monoalphabetic substitution. With a small amount of text, the simple index is somewhat indeterminate. Using the triple and quadruple indices, the results are even more so, and at times may give even false indications. It is again emphasized that sufficient text must be used to give positive indications.

As another elementary example of the application of the index of coincidence to the internal examination of a cipher text, we have, for example, a 5-letter repetition at an interval of 85. Is the cipher a polyalphabet cipher of 5 or 17 alphabets? By means of internal examination with the index of coincidence we can determine what type of cipher we have. Make a frequency count of the cipher alphabets assuming 5 and then 17 alphabets.

Calculate the index of coincidence in each case for one or more alphabets. The indices of higher value will indicate which assumption is correct. If neither assumption shows positive results (an index around 1.7) we may have a progressive cipher, running key cipher, auto key cipher, or cipher of even more complex nature.

Another elementary application is as follows. We have a cipher message which has been intercepted. The I.C. is computed and found to be σ=1.79.

This is so rough as to resemble plain text. A simple substitution has the property of leaving σ unchanged, and so has a transposition. Multi-alphabet substitutions lower the I.C.. So we are reduced to three hypotheses, that our sample is either transposed plain text, a simple substitution, or both substitution and transposition.

The digraphic I.C. is computed, σ_2=4.85. Remember that its expected value is 3.Ø5 = σ^2 in view of the known roughness. Therefore the index of coherence is K_2 = $\frac{4.85}{3.Ø5}$=1.56

Since a transposition destroys coherence we can assert that no transposition is involved. Multigraphic Indices of coincidence are preserved by a simple substitution.

VIII THE STANDARD DEVIATION

We have already several times referred to the fact that these statistics are useful only if the sample is large enough. To get an idea as to whether this is the case or not we measure our results in terms of a standard deviation, "sigma". One standard deviation is roughly one half the width of a band which when placed about the average will include two-thirds of the data. It is a measure of the dispersion. If sigma is large the data is spread out wide, and if it is small the numbers are close together. In a binomial distribution the standard deviation is $\sigma = \sqrt{N \times pq}$ where N is the number of observations and p and q are the probabilities of success and failure.

To estimate the significance of δ, we refer to (12), where the denominator is the expected number of incidences in flat text, and the numerator is the number found. Assuming a binomial distribution of the incidences we find the variance $\sigma^2 = Npq = \frac{M(M-1)(C-1)}{2C^2}$ (20)

If s is the "sigmage" or deviation of the number found divided by σ we have

$$s = \frac{\frac{1}{2}\sum f_i (f_i - 1) - \frac{M(M-1)}{2C}}{\sqrt{\frac{M(M-1)(C-1)}{2C^2}}} \tag{21}$$

$$= \frac{C\sum f_i (f_i - 1) - M(M-1)}{\sqrt{2(C-1)(M)(M-1)}} = \frac{C\sum f_i (f_i - 1)}{\frac{M(M-1)}{\sqrt{2\frac{C-1}{M(M-1)}}}} - 1 = \frac{\delta - 1}{\sqrt{2\frac{C-1}{M(M-1)}}} \tag{22}$$

(23)

$$= \frac{\delta - 1}{\sqrt{2(C-1)}} \sqrt{M(M-1)} = \boxed{\frac{\delta - 1}{\sqrt{2(C-1)}} M.}$$

For M>51 error is less than 1%.

Notice that the sigmage is a linear function of the sample size M, and also linear with the "bulge" $\delta - 1$. The denominator is relatively unimportant to the estimation of S except in shifting from code to cipher, when C can change from 500,000 to as small as 10.

The bulge $\delta - 1$ is a quantity which will recur frequently.

Formula (23) does not apply to the iota I.C. For that we have the expected number g and N/C found. Then $\sigma^2 = N/C\,(1 - 1/C)$ and the sigmage is

$$s = \frac{g - N/C}{\sqrt{N/C(1-1/C)}} = \frac{gC/N - 1}{\sqrt{\frac{C-1}{N}}} = \frac{\iota - 1}{\sqrt{C-1}} \sqrt{N} \tag{24}$$

In this case the sigmage is linear with the bulge $\iota - 1$, but varies only as the square root of the sample size.

The significance of s is given in the following table, which lists the probability of getting s or a larger result from chance.

Table I

s	prob.	s	prob.
.1	.4602	2.9	.0019
.2	.4207	3.0	.00135 = 1/800
.3	.3821	3.2	.00069
.4	.3446	3.4	.00034 = 1/3000
.5	.3085	3.6	.00016
.6	.2743	3.8	.00007
.7	.2420	4.0	.00003 = 1/33.000
.8	.2119	4.101	.0000206
.9	.1841	4.200	.0000134
1.0	.1587	4.299	.0000086
1.1	.1357	4.398	.0000055
1.2	.1151	4.497	.0000035 = 1/300,000
1.3	.0968	4.596	.0000022
1.4	.0808	4.695	.0000014 = 1/711,000
1.5	.0668	4.794	.0000008 = 1/1,250,000
1.6	.0548	4.907	.00000046164
1.7	.0446	5.006	.00000027741
1.8	.0359	5.105	.00000016513
1.9	.0287	5.209	.00000009736
2.0	.0228	5.303	.00000005686 = 1/18 million
2.1	.0179	5.402	.00000003290 =1/30 million
2.2	.0139	5.501	.00000001850
2.3	.0107	5.600	.00000001070
2.4	.0082	5.798	.00000000335
2.5	.0062	6.080	.00000000060
2.6	.0047	6.503	.00000000004
2.7	.0035	6.785	.00000000001 = 1/100,000 million
2.8	.0026		

EXAMPLE

Suppose an unknown cipher in four digit groups is being investigated and the question is whether it is reenciphered* or not. If it is the text can be expected to resemble random more than if it is not. We make a frequency count on the 10,000 groups and then determine how likely such a distribution is by chance. If it is not likely we must seek an explanation.

If 560 groups are counted and $\Sigma f_i(f_i-1) = 76$ then $\sigma = 2.41$ This gives a sigmage $s = 56$. The table shows that we would have to repeat this procedure about 120 million times on random material to get a like result. We can say that this result is unlikely by chance and that an explanation is called for. The obvious one is that there is no reencipherment, or that it is very feeble. Tests to check this further hypothesis can be quickly devised.

*See Gaines "Elementary Cryptanalysis" page 2.

IX TO DETERMINE WHETHER TWO ALPHABETS ARE IDENTICAL ALPHABETS

Assume that a complex cipher using secondary alphabets has been analyzed and reduced to 50 alphabets. There are only 26 possible secondary alphabets, so some of these 50 alphabets can be combined. Visual inspection is too inaccurate to be trusted, except within an abnormally large amount of text.

Four sample alphabets are given in the table following, A and B.

Table "A" -- Frequency Tables

	No. 1	No. 2	No. 3	No. 4
A			2	2
B	2		2	
C	1		1	1
D		1		1
E	3	3	2	
F	1			1
G			2	
H	2			
I	2		3	2
J	1	1		
K		2		
L		1		
M	2		1	
N	3	2	3	1
O				2
P		1		1
Q		1		
R		1		
S				3
T		1		
U				
V		2		
W			1	3
X		2		2
Y	2	2	1	1
Z	1		2	
	20	20	20	20

Table "B" -- Repeated Letters

	No.1-&-2	No.1-&-3	No.1-&-4	No.2-&-3	No.2-&-4	No.3-&-4
A						4
B		4				
C		1	1			1
D					1	
E	9	6		6		
F			1			
G						
H						
I		6	4			6
J						
K						
L						
M		2				
N	6	9	3	6	2	3
O						
P					1	
Q						
R						
S						
T						
U						
V						
W						3
X					4	
Y	4	2	2	2	2	1
	19	32	11	14	10	18

Total coincidences

Chance coincidences = $\frac{400}{26}$ = 15.4

Line up two alphabets at a time and cross multiply the repetitions for each letter (see table "B"). "E" occurs 3 times in No. 1 alphabet and 3 times in No. 2. There are 9 pairs of "E's" in No. 1 and No. 2. Add the coincidences noted for a pair of alphabets.

There are 20 letters in each alphabet which gives 400 pairs of letters to deal with. Chance would give 400/26, or 15.4 repeated letters in two alphabets. The index of coincidence is the sum of the actual coincidences divided by 15.4.

The indices are as follows:

No. 1-&-2	1.23	
No. 1-&-3	2.08	(above the normal 1.75)
No. 1-&-4	.71	(alphabets No.1-&-3 must be identical)
No. 2-&-3	.91	
No. 2-&-4	.65	

X THE CROSS I. C.

Suppose we have two stretches of text of which the distributions are given by $P_1, P_2, \cdots, P_c$ and $q_1, q_2, \cdots, q_c$

$$\sum_1^c p_i = 1, \qquad \sum_1^c q_i = 1.$$

If these are placed one above the other the probability of an incidence at any one position is

$$\sum_{i=1}^{c} p_i q_i \tag{25}$$

and the I.C. is

$$\boxed{\xi = c \sum p_i q_i}$$

We can show that the expected value of ξ is 1. Suppose that we apply a permutation to the q's and recompute ξ. If we do this for all C! permutations and average them we get

$$E(\xi) = \tfrac{1}{c!} \sum_{\substack{\text{All} \\ \text{permutations}}} \xi = \tfrac{1}{c!} \sum_{\substack{\text{All} \\ \text{permutations}}} c \sum_1^c p_i q_i'$$

where q_i' is one of the q's, depending on the permutation. Each q comes into a given position (C-1)! times. Thus

$$\sum_{\substack{\text{All} \\ \text{permutations}}} q_i = (c-1)! \quad \text{and}$$

$$E(\xi) = \tfrac{c}{c!} \sum_{i=1}^{c} p_i \sum_{\substack{\text{All} \\ \text{permutations}}} q_i' = \frac{c(c-1)!}{c!} \sum_1^c p_i = 1.$$

Notice that this result is independent of the roughness of either distribution.

If one of the samples of text is flat, say $q_i = \tfrac{1}{c}$, then

$$\xi = 1 \tag{26}$$

For

$$\xi = c \sum p_i q_i = \tfrac{c}{c} \sum p_i = 1 \times 1 = 1$$

If both samples are very rough, then ξ fluctuates widely as the q's are permuted. The question arises, how wide is the distribution of ξ ?

A standard measure is the "variance"

$$\sigma^2 = E(\xi^2) - [E(\xi)]^2 \tag{27}$$

We evaluate $E(\xi^2) = 1/c! \sum_{\tau} (c \sum_{1}^{c} p_i q_i)^2$ (28)

where $\sum_{\tau}$ is understood to mean the sum over all possible permutations of the q's. Then

$$E(\xi^2) = 1/c! \sum_{\tau} c^2 \sum_{i=1}^{c} \sum_{j=1}^{c} p_i q_i p_j q_j$$

$$= c/c! \sum_{i=1}^{c} \sum_{j=1}^{c} p_i p_j \sum_{\tau} q_i q_j \tag{29}$$

The term $\sum_{\tau} q_i q_j$ can be evaluated in each of the cases $i \neq j$ and $i = j$

We will use $P = c \sum_{1}^{c} p_i^2$ and $Q = c \sum_{1}^{c} q_i^2$

as the I.C.'s of the two samples.

For $i \neq j$ $\sum_{\tau} q_i q_j = (c-2)! \sum_{i \neq j}^{c} q_i q_j$

$$= (c-2)! \sum_{i=1}^{c} q_i \sum_{i \neq j}^{c} q_j = (c-2)! \sum_{i=1}^{c} q_i (1 - q_i)$$

$$= (c-2)! \sum_{i=1}^{c} (q_i - q_i^2) = (c-2)! (1 - c/c). \tag{30}$$

For $i = j$ $\sum_{\tau} q_i^2 = (c-1)! \sum_{i=1}^{c} q_i^2 = (c-1)! \; c/c$. (31)

Thus (29) becomes

$$E(\xi^2) = c^2/c! \sum_{i=1}^{c} \sum_{j \neq i}^{c} p_i p_j (c-2)! \frac{c-Q}{c} + c^2/c! \sum_{i=1}^{c} p_i^2 (c-1)! \; c/c$$

$$= c^2/c! \; (c-2)! \frac{c-Q}{c} \sum_{i=1}^{c} \sum_{j \neq i}^{c} p_i p_j + c^2/c! (c-1)! \; c/c \sum_{i=1}^{c} p_i^2$$

$$= c^2/c! \; (c-2)! \frac{c-Q}{c} \frac{c-P}{c} + c^2/c! \; (c-1)! \; c/c \; P/c = \frac{(c-Q)(c-P)}{c(c-1)} + \frac{QP}{c}$$

$$= \frac{c^2 - cP - cQ + PQ}{c(c-1)} + \frac{cQP - QP}{c(c-1)}$$

$$= \frac{c^2 - cP - cQ + cQP}{c(c-1)} = \frac{c - P - Q + PQ}{c-1} \tag{32}$$

Then since $E(\xi) = 1$, we have

$$E(\xi^2) - [E(\xi)]^2 = \frac{C-P-Q+PQ}{C-1} - 1$$

$$= \frac{C-P-Q+PQ-C+1}{C-1} = \frac{PQ-P-Q+1}{C-1}$$

$$\boxed{\sigma^2 = \frac{(P-1)(Q-1)}{C-1}} \qquad (33)$$

This says that the square of a standard deviation σ is the product of the bulges over c-1.

$$S = \frac{\xi - 1}{\sqrt{\frac{(P-1)(Q-1)}{C-1}}} = (\xi - 1)\sqrt{\frac{C-1}{(P-1)(Q-1)}}$$

is

a measure of the significance which can be judged from table I.

The function ξ can be used as a measure of correlation between two distributions.

XI THE COINCIDENCE TEST USED TO ALIGN SECONDARY ALPHABETS INTO A PRIMARY ALPHABET

We give here a special application of the I.C. statistic. An actual problem from the elementary course is used, problem 4 of assignment 6. Special frequency distribution tables were made of the sample lined up into 26 columns, i.e., lines of 26 letters each.

This problem happens to be enciphered by means of a Vigenere table, the columns being used in successions. Consequently if the cipher text is lined up 26 wide each column is enciphered by a monoalphabetic substitution. Each alphabet is a slide on that in the next column. If we knew the plain and cipher sequences the text could be decrypted. The problem is to recover these sequences. Since the sequence is not alphabetical, adjacent frequency counts as given in table C appear unrelated. But if we look at the rows they must be related by being slides on each other. If we can establish these slides will have the cipher sequence.

Table "C" -- Frequency Table of the Cryptogram

Column - / Cipher -	1	26	25	24	23	22	21	2Ø	19	18	17	16	15	14	13	12	11	1Ø	9	8	7	6	5	4	3	2	Total Letters
A	1	3		3				2		5		3	2							1		2		1	2		25
B		1	3	1		4	4	1					2	5		5		2	3	1			1		1	2	36
C		2		1	1							1		3	2			1	1			3			1	2	18
D					1		1			2	1	1		1			7				4		3		1		22
E			1	2	2				2		3			1		1	1	1				1	4	5	1		25
F				1				1	1	2						1	1		2		1		1	2		1	14
G	1	1			1	5		1	1	2	1		1		2					1	1	1		2	3	2	26
H	3		2		1	1						2		1		2			1		1	1	1			1	17
I		1			2			1		1	1			1	1		2	1				3				4	18
J			3	1	3		3					1			2		1	4			1			1	2		22
K	4	1			2			1	3			1	3	1	2	2		2				3			2		27
L			1				4				3	2				3			1		3			1	1	1	2Ø
M	1		1	3		1	1	1	2		3				5			1		1				1		1	22
N		1						3	3				4	3	1				1	2		2		2	2		24
O	3		2	1				1	1				2			2	1	1			1					1	16
P				3	3				1						1					4	4				1		17
Q				1			1				5	1						1	1	2			2			2	16
R		1	1	2		1		2	4	1	1		2		1			3	1		1		1			2	24
S	2		3		1	2				4			2			1				2		1	3				21
T		2	1		1							4	3	1			3			5			1	3		2	26
U	1			1	1	4	1	3	1		2			1				2				2					19
V	4	2			1			1	1			1		1	1	1		3	2			3	2		3	1	27
W		4	1	1		1	2				1					3	3		1	3	2			1			13
X	1				1		2			1		1		1			1		4		1		1	2	1		17
Y		1				1	2	3	1	1		2			1				3		1		2		1		19
Z	1	1	2			1				2		1		1	2		1		1		1			1			15

By use of this special table, table "C", we can build up the cipher component used by matching these frequency distributions. In selecting distributions to match we want to obtain:

(a) A maximum total number of letters involved (as the distribution will then be more reliable).

(b) A distribution with a normal count (i.e., similar to normal alphabet).

(c) A distribution without any one letter of abnormal frequency (as this gives too much weight to one letter).

When frequency distributions of two letters are properly matched, high should pair with high, low with low, blank with blank, etc. The mathematical value of each relative position is found as follows:

(a) With the two frequency distributions in question written on paper strips and slid one against the other, for any one position we cross multiply the frequencies in alignment, and then add the products of all these multiplications. This is the total number of parts or coincidences involved (see table "D").

(b) Cross-multiply the total count of the distribution of the first letter by the count of the second letter. This is the total number of possible pairs of letters. Chance would produce one coincidence in twenty-six. Therefore, divide this product by twenty-six, which gives the number of expected chance coincidences.

(c) Divide the number of the actual coincidences by the expected number of chance coincidences (that is, divide (1) by (2)). The resulting number is the Index of Coincidence.

To prove correct alignment:

(a) The index for the given relative position of two distributions must be higher than for all other positions, with no close second.

(b) The index should be 1.5Ø or higher (preferably 1.75 or higher).

(c) There must be only one acceptable alignment.

Indeterminate results will be encountered in some cases, particularly with insufficient text.

Table "E" gives the total coincidences at the various positions of one strip slid against another.

From table "C", it is seen that certain distributions have the following properties (referring to our three desired properties):

B (cipher) has a total of 36 letters, with 14 different cells involved. Its highest frequency is 5. Good. Approaches normality. Maximum text.

V (cipher) has a total of 27 letters with 15 different cells involved. Its highest frequency is 4. Not good -- too flat.

K (cipher) has a total of 27 letters with 13 different cells involved. Its highest frequency is 4. Not good - too flat.

G (cipher) has a total of 25 letters with 15 different cells involved. Its highest frequency is 5. Not good -- too flat.

D (cipher) has a total of 22 letters with only 1Ø different cells involved, but its highest frequency is 7. Not good -- too peaked.

T (cipher) has a total of 26 letters with 11 different cells involved. Its highest frequency is 5. Good. Approaches normality.

A (cipher) has a total of 25 letters with 11 different cells involved. Its highest frequency is 5. Good. Approaches normality.

N (cipher) has a total of 24 letters with 11 different cells involved. Its highest frequency is 4. Good. Approaches normality.

B, T, A and N are the best choices. Match T, A and N against B, then match A and N against T, finally match N against A. One of these combinations should give a positive index of coincidence, and thus serve as a starting point.

Table "D"--- Sliding Strips

Plain	1	2	3	4	5	6	7	8	9	1Ø	11	12	13	14	15	16	17	18	19	2Ø	21	22	23	24	25	26	"B" Master Distribution
Cipher	B																										
Freq.		1	3	1		4	4	1					2	5		5		2	3	1			1		1	2	1 3 1 4 4 1 2 5
			2	1		1							4	3	1			3			5			1	3	2	"T" set at 2 on the "B" Master Distribution.
	T																										
	1	3		3				2		5		3	2							1		2		1	2		
	A																										
		1						3	3				4	3	1				1	2		2		2	2		
	N																										

Plain	1	2	3	4	5	6	7	8	9	1Ø	11	12	13	14	15	16	17	18	19	2Ø	21	22	23	24	25	26	T-1- -- N-5- Master Distribution.
Cipher	T				N																						
Freq.		4	3		1	1						7	6	1			7	3	1	5			2	5		4	4 3 1 1 7 6 1
												1	3	1		4	4	1					2	5		5	2 3 1 1 1 2
											B																
																1	3		5				2		5	3 2 1 2 1
																A											

Table "E" -- Table of Total Coincidences

B = Master Distribution

Plain -1	2	3	4	5	6	7	8	9	1Ø	11	12	13	14	15	16	17	18	19	2Ø	21	22	23	24	25	26	Chance Coincidences
Cipher-B																										
T -X	43	49	--	39	--	--	--	--	53	--	--	41	4Ø	37	--	6Ø	--	--	--	64	--	--	--	--	39	36
A -X	--	--	38	66	--	59	--	--	--	--	--	42	--	--	--	--	--	--	--	5Ø	--	42	--	--	--	35
N -X	--	--	--	--	53	4Ø	--	--	--	--	--	--	--	--	--	--	--	--	--	43	39	63	--	--	--	33

NOTE: The numbers 43, 49, etc., represent the successive "total coincidences", i.e., the sums of products of frequencies at successive points of coincidence.

T = Master Distribution

Plain -1	2	3	4	5	6	7	8	9	1Ø	11	12	13	14	15	16	17	18	19	2Ø	21	22	23	24	25	26	
Cipher-T																										
A -X	--	32	--	--	--	--	28	29	--	42	--	--	--	31	--	51	--	43	--	--	--	--	--	--	--	25
N -X	--	--	28	5Ø	--	28	26	--	--	--	34	--	--	--	34	--	--	31	--	--	--	--	--	32	34	24

A = Master Distribution

Plain -1	2	3	4	5	6	7	8	9	1Ø	11	12	13	14	15	16	17	18	19	2Ø	21	22	23	24	25	26	
Cipher-A																										
N -X	--	26	--	--	--	--	--	--	--	--	28	26	--	43	--	34	27	--	--	--	--	27	29	--	27	23

T - 1 -- N - 5 = Master Distribution

Plain -1	2	3	4	N 5	6	7	8	9	1Ø	11	12	13	14	15	16	17	18	19	2Ø	21	22	23	24	25	26	
Cipher-T				N																						
B -X	--	--	--	X	--	1ØØ	-	--	--	123	-	8Ø	78	74	--	--	--	--	--	82	--	75	73	89	96	69
A -X	--	--	--	--	--	--	57	56	--	61	--	--	--	65	--	94	--	65	--	--	--	--	--	--	--	48

Table "F" -- Table of Coincidences

Master Distribution T-1 N-5 B-11 A-17

Plain	1	2	3	4	5	6	7	8	9	1Ø	11	12	13	14	15	16	17	18	19	2Ø	21	22	23	24	25	26	Chance Coin-cidences
Cipher	T T	V	I	D	N N	Y	U	F	P	M	B B	K	H	Q	Z	E	A A	L	G	X	R	J	W	S	C	O	
K	X	149	-	-	X	-	-	-	-	-	X	175	-	-	-	-	X	-	-	-	-	-	-	-	-	156	12Ø
V	X	181	-	-	X	-	-	-	-	-	X	X	-	-	-	-	X	-	-	-	-	143	-	-	-	158	12Ø
D	X	X	-	193	X	-	-	165	-	-	X	X	-	-	-	-	X	-	-	-	-	143	-	-	-	-	94
E	X	X	135	-	X	-	-	-	-	-	X	X	-	-	-	19Ø	X	-	-	-	148	-	-	-	-	-	1Ø2
W	X	X	-	-	X	-	-	-	-	129	X	X	-	-	-	X	X	-	-	-	-	-	2ØØ	-	-	-	1Ø2
G	X	X	-	-	X	-	-	-	-	-	X	X	-	-	143	X	X	-	193	-	131	-	X	-	-	-	1Ø7
R	X	X	-	-	X	-	-	-	133	-	X	X	-	-	-	X	X	-	X	-	158	-	X	-	-	-	1Ø2
M	X	X	133	-	X	-	-	-	-	155	X	X	-	144	-	X	X	-	X	-	X	-	X	-	-	-	94
J	X	X	-	-	X	-	-	-	-	-	X	X	-	-	-	X	X	-	X	-	X	17Ø	X	-	-	-	9Ø
L	X	X	-	-	X	144	-	-	-	-	X	X	-	-	-	X	X	154	X	-	X	X	X	-	-	-	9Ø
S	X	X	-	-	X	-	-	11Ø	-	-	X	X	-	-	1Ø9	X	X	-	X	-	X	X	X	159	-	-	9Ø
C	X	X	117	-	X	-	-	-	-	-	X	X	-	-	-	X	X	-	X	-	X	X	X	X	172	-	81
U	X	X	-	-	X	-	119	9Ø	-	-	X	X	-	-	-	X	X	-	X	-	X	X	X	X	X	-	81
Y	X	X	-	-	X	153	-	93	-	96	X	X	-	-	-	X	X	-	X	-	X	X	X	X	X	X	81
I	X	X	121	-	X	X	X	94	-	-	X	X	1Ø9	-	-	X	X	X	X	-	-	X	X	X	X	X	77
H	X	X	X	-	X	X	X	-	-	-	X	X	119	-	86	X	X	X	X	-	X	X	X	X	X	9Ø	73
P	X	X	X	-	X	X	X	-	142	-	X	X	X	-	-	X	X	X	X	116	X	X	X	X	X	-	73
X	X	X	X	-	X	X	X	11Ø	X	-	X	X	X	-	-	X	X	X	X	139	X	X	X	X	X	-	73
O	X	X	X	-	X	X	X	-	X	-	X	-	X	X	-	-	-	X	X	X	X	X	X	X	X	117	68
Q	X	X	X	-	X	X	X	-	X	81	X	X	X	1Ø8	-	X	X	X	X	X	X	X	X	X	X	X	68
Z	X	X	X	-	X	X	X	-	X	X	X	X	X	X	113	X	X	X	X	X	X	X	X	X	X	X	64
F	X	X	X	72	X	X	X	87	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	6Ø

BUILDING UP THE CIPHER COMPONENT

By utilizing the principles described in the previous sections, we can build up the cipher component. Take B (cipher) as the master distribution, since it is acceptable and contains the highest count. Copy the frequencies of B (from "C") at the bottom of a strip of paper, and repeat this sequence to the right. Over the first sequence write the numbers 1 to 26 as shown in table "D". Under No. 1 write the letters"B". The numbers represent the various unknown letters of the cipher component. Make similar master distribution strips for T and A. Next, copy the distribution of T (cipher) (from table "D") at the top of a strip of paper. Only one sequence is required for this strip, and the numbers are omitted. Indicate the space corresponding to column No. 1 (table "D") by the letter T (see table "E"). In a like manner make strips for A and N.

Note: The letter on each strip is an indicator to mark column No. 1 for that letter. When strips are properly aligned, the indicators show the relative positions of these letters in the cipher component. The student is advised to prepare strips for himself and follow these processes.

First, match T against B. As no two letters can occupy the same position in the cipher component, begin by setting the T indicator at No. 2 on the B master alphabet. Note the coincidences. Next, slide T to No. 3, and note the coincidences. Continue this process to No. 26, and record the succesive coincidences in tabular form (see table "F"). In many cases lack of good coincidences will be obvious by inspection tion and the count need not be made. In this way we discover that B and T give high indices of coincidence in two different alignments (indices computed in accordance with rule in page 23).

Index of B (1) - T (7) $\frac{64}{36} = 1.77$ (good)

Index of B (1) - T (11) $\frac{6\emptyset}{36} = 1.67$ (good)

All other alignments give such low indices that they can be at once eliminated. The above two indices, however, are both high enough to be significant, and as the second is so close to the first, it cannot be disregarded.

There can be only one acceptable point of coincidence; therefore, it is necessary to match A against B, and N against B, to see if more conclusive results can be attained.

Index of B (1) - A (5) $\frac{66}{36} = 1.89$ (excellent)

Index of B (1) - A (7) $\frac{59}{35} = 1.69$ (good)

(other alignments are eliminated)

Index of B (1) - N (21) $\frac{63}{33} = 1.91$ (excellent)

Index of B (1) - N (6) $\frac{53}{33} = 1.6\emptyset$ (fair)

(other alignments are eliminated)

Since there is no outstanding coincidence with "B" as the "master-alphabet", try "T" and "A" as the "master-alphabet":

Index of T (1) - A (17) $\frac{51}{25}$ = 2.04 (excellent)

Index of T (1) - A (19) $\frac{43}{25}$ = 1.72 (good)

Index of T (1) - A (11) $\frac{42}{25}$ = 1.68 (good)

Index of T (1) - N (5) $\frac{50}{24}$ = 2.08 (excellent)

Index of T (1) - N (12) (16) (26) $\frac{34}{24}$ = 1.42 (poor)

Index of A (1) - N (15) $\frac{43}{23}$ = 1.87 (good)

Index of A (1) - N (17) $\frac{34}{23}$ = 1.48 (poor)

The most certain combination is T (1) - N (5), and there is no doubt as to its correctness. This located "N" relative to "T" in the cipher component, and allows us to consolidate their frequencies.

For a new master distribution add the frequencies of T (at space No. 1) to those of N (at space No. 5) (see table "E"). Match "B" and "A" against this new master distribution:

Index of T (1) - N (5) - B (11) $\frac{123}{69}$ = 1.81 (good)

Index of T (1) - N (5) - B (7) $\frac{100}{69}$ = 1.45 (poor)

Index of T (1) - N (5) - B (26) $\frac{76}{69}$ = 1.40 (poor)

Index of T (1) - N (5) - A (17) $\frac{94}{48}$ = 1.96 (excellent)

Index of T (1) - N (5) - A (15) $\frac{65}{48}$ = 1.35 (very poor)

Index of T (1) - N (5) - A (19) $\frac{65}{48}$ = 1.35 (very poor)

"B" and "A" can now be consolidated with "T" and "N" for the final "master distribution" as follows:

:Plain -	1	2	3	4	5	6	7	8	9	10	11	12	13
:Cipher -	T				N								
:T &		4	3		1	1						7	6
: B		2	3	1			1		1	2		1	3
: A		3	2							1		2	

Plain -	14	15	16	17	18	19	20	21	22	23	24	25	26:
Cipher -													:
T &	1			7	3	1	5			2	5		4:
B	1		4	4	1					2	5		5:
A	1	2		1	3		3				2		5:

This locates "B" and "A" relative to "T" and "N" in the cipher component, in addition to giving the combined frequencies of all four letters.

FINAL MASTER DISTRIBUTION

Plain Comp.	-	1	2	3	4	5	6	7	8	9	10	11	12	13
Cipher Comp.	-	T				N						B		
Comb. Freq.	-		9	8	1	1	1	1		1	3		10	9

Plain Comp.	-	14	15	16	17	18	19	20	21	22	23	24	25	26
Cipher Comp.	-				A									
Comb. Freq.	-	3	2	4	12	7	1	8		4	12			14

Analyze the preceding steps. T gave two possible alignments with B, and, as we now see, the incorrect position gave the higher index. N also gave two possible alignments with B. (B was at fault due to its erratic letter distribution). However, when T and N are combined, giving twice as many letters in the master distribution, B fitted in with only one possible alignment. Adding B and A gives twice as many letters in the master distribution and this should make future results even more positive. The master distribution (of 100 letters or more) should approximate a normal frequency distribution and will give a standard to which all the other distributions can be referred. Hereafter, variations in the highest index of coincidence will be due entirely to letter distribution of the various distributions themselves.

For example:

If the highest index is 1.7 -- letter distribution is normal.

If the highest index is 2.0 -- high frequency letters predominate.

If the highest index is 1.4 -- the intermediate and low frequency letters predominate.

Therefore, when matching the remaining letters, we can accept the highest index of coincidence as establishing coincidence, unless the second highest is practically the same.

Continue the matching process and the reconstruction of the cipher component, noting that T, N, B and A are already located and thus may be deleted at once from further test. Begin with the letters of the highest frequency, as they should give the most positive results. When a letter is placed, delete this location from further test, as two letters connot occupy the same space in the cipher component.

Letters are added to the cipher component in the following order (see table "F"):

"Master alphabet" T (1) - N (5) - B (11) - A (17)

K (12) 1.46 (poor - but acceptable)

V (2) 1.51 (poor - but acceptable)

D (4) 2.05 (excellent)

D (8) 1.76 (good)(D - not certain)

E (16) 1.86 (excellent)

W (23) 1.96 (excellent)

G (19) 1.80 (good)

Note: With this many values a key-word (if any) sequence could be completed by inspection. In this case, the partially reconstructed cipher component gives no suggestion of a key-word sequence.

R (21) 1.55 (fair but acceptable)

M (1Ø) 1.65 (good)

M (14) 1.53 (M - not certain)

J (22) 1.89 (excellent)

L (18) 1.71 (good)

L (6) 1.6Ø (fair)(L - not certain)

S (24) 1.77 (good)

C (25) 2.12 (excellent)

U (7) 1.47 (poor but acceptable)

Y (6) 1.89 (excellent)

Note: This throws out L (6), but leaves L (18) as correct.

I (3) (fair)

I (13) 1.42 (poor)(I - not certain)

H (13) 1.63 (good)

Note: This throws out I (13) and leaves I (3) as correct.

P (9) 1.95 (excellent)

X (2Ø) 1.9Ø (excellent)

O (26) 1.72 (good)

Q (14) 1.59 (fair and acceptable)

Note: This throws out M (14) and leaves M (1Ø) as correct.

Z (15) 1.77 (good)

F (8) 1.45 (poor but acceptable)

F (4) 1.2Ø (very poor)

F (8) is correct and D (4) is correct

The cipher component has now been completely recovered.

Note: The process described above has actually built up the complete squared-cipher-table of a modified Vigenere table (it remains only to recover the plain component to complete the Vigenere table). We have written down the cipher component rather than the complete-squared-table merely to save time and effort.

XII THE ROUGHNESS OF MIXED TEXTS

What happens when two different distributions are mixed? As a simple case, let us suppose we mix some text of I.C. γ with flat text in the proportions R:(1 - R). Then the ith letter has probability,

$$p_i R + (1 - R)\frac{1}{c}$$

and the I.C. is $c \sum_1^c \left(p_i R + (1-R)\frac{1}{c}\right)^2$

$$= c \sum_1^c p_i^2 R^2 + 2c \sum_1^c p_i R (1-R)\frac{1}{c} + c \sum_1^c (1-R)^2 \frac{1}{c^2}$$

$$= R^2 \gamma + 2R(1-R) + (1-R)^2 \qquad (34)$$

$$= R^2 \gamma - R^2 + 1 = 1 + (\gamma - 1) R^2.$$

That is, the rough text contributes its bulge in the proportion R^2.

Now as a more complicated case suppose that we mix two rough texts in the proportions R: (1 - R). Suppose further that the distributions are P_1, P_2, ---, P_c and q_1, q_2, ---, q_c and

$$c \sum p_i^2 = P , \qquad c \sum q_i^2 = Q .$$

Then the mixture will have as probability of the ith letter $p_i R + q_i (1-R)$

and $\gamma = c \sum (p_i R + q_i (1-R))^2$

$$= c \sum p_i^2 R^2 + 2c \sum p_i q_i R (1-R) + c \sum q_i^2 (1-R)^2$$

$$= PR^2 + 2R(1-R)\, c \sum p_i q_i + Q(1-R)^2 . \qquad (35)$$

The expression $c \sum p_i q_i = \xi$ we have examined before, and seen to be a measure of the correlation of the distributions. Since the expected value of ξ is 1, the expected value of γ will be

$$E(\gamma) = PR^2 + 2R(1-R) + Q(1-R)^2$$

$$= R^2 + (P-1)R^2 + 2R - 2R^2 + (1-R)^2 + (Q-1)(1-R)^2 .$$

$$= 1 + (P-1)R^2 + (Q-1)(1-R)^2 .$$

Here again each contributes its bulge in proportion to the square of ite weight.

Now that we have seen how the argument goes, we can generalize this result to a mixture of K samples with distributions P_{i1}, P_{i2}, ---, P_{ic},

and $\gamma_i = c \sum_{j=1}^{c} p_{ij}^2$ the I.C. of the ith sample,

which we will suppose present in the proportion R_i, $\sum_{i=1}^{K} R_i = 1$.

The mixture will have the jth letter present in the proportion

$$P_j = \sum_{i=1}^{K} R_i \, P_{ij}$$

so that its I.C. is

$$\gamma = c \sum_{j=1}^{c} P_j^2 = c \sum_{j=1}^{c} \left(\sum_{i=1}^{K} R_i \cdot p_{ij} \right)^2 \qquad (36)$$

$$= c \sum_{j=1}^{c} \sum_{i=1}^{K} \sum_{L=1}^{K} R_i \, p_{ij} \, R_L \, p_{Lj}$$

$$= \sum_{i=1}^{K} \sum_{L=1}^{K} R_i \, R_L \; c \cdot \sum_{j=1}^{c} p_{ij} \, p_{Lj} .$$

The internal sum is the measure of the correlation of the ith and ℓ th samples, which we will designate by

$$\xi_{i\ell} = c \sum_{j=1}^{c} P_{ij} P_{\ell j} .$$

Notice that $\xi_{ii} = \gamma_i$ is the I.C. of the ith sample.

Now γ becomes $\gamma = \sum_{i=1}^{K} \sum_{\ell=1}^{K} R_i R_\ell \xi_{i\ell}$

$$= \sum_{i=1}^{K} \sum_{\ell \neq i}^{K} R_i R_\ell \xi_{i\ell} + \sum_{i=1}^{K} R_i^2 \gamma_i \qquad (37)$$

$$= \sum_{i=1}^{K} \sum_{\ell \neq i}^{K} R_i R_\ell \xi_{i\ell} + \sum_{i=1}^{K} \left[R_i^2 + (\gamma_i - 1) R_i^2 \right]$$

Since $E(\xi_{i\ell}) = 1$ we have

$$E(\gamma) = \sum_{i=1}^{K} \sum_{\ell=1}^{K} R_i R_\ell + \sum_{i=1}^{K} (\gamma_i - 1) R_i^2$$

$$= \sum_{i=1}^{K} R_i \sum_{\ell=1}^{K} R_\ell + \sum_{i=1}^{K} (\gamma_i - 1) R_i^2 \qquad (38)$$

$$E(\gamma) = 1 + \sum_{i=1}^{K} (\gamma_i - 1) R_i^2$$

Again each sample contributes its bulge according to the square of its presence.