Text

The Applications of Probability to Cryptography

Alan M. Turing

Copyright

Editor's Notes

Provenance

Two Second World War research papers by Alan Turing were declassified recently. The papers, *The Applications of Probability to Cryptography* and its shorter companion *Paper on Statistics of Repetitions*, are available from from the National Archives in the UK at `www.nationalarchives.gov.uk.`

The released papers give the full text, along with figures and tables, and provide a fascinating insight into the preparation of the manuscripts, as well as the style of writing at a time when typographical errors were corrected by hand, and mathematical expression handwritten into spaces left in the text.

Working with the papers in their original format provides some challenges, so they have been typeset for easier reading and access. We recommend that the typeset versions are read with a copy of the original manuscript at hand.

This document contains the text and figures for *The Applications of Probability to Cryptography*, the companion paper is also available in typeset form from arXiv at `www.arxiv.org/abs/1505.04715`. These notes apply to both documents.

Separately, a journal article by Zabell[1] provides an analysis of the papers and further background information.

The text

It is not our intent to cast Alan Turing's manuscripts into a journal style article, but more to provide clearer access to his writing and, perhaps, to answer the questions "If Turing had have had access to typesetting software, what would his papers have looked like?". Consequently no "house-style" copy-editing has been imposed. Occasional punctuation has been added to improve readability, some obvious errors corrected, and paragraph breaks added to ease the reading of long text blocks - and occasionally to give a better text flow. Turing uses typewriter underlining, single, and double quotes to indicate emphasis or style; these have been implemented using font format changes, double quotes are used as needed.

The manuscript has many typographical errors, deletions, and substitutions, all of which are indicated by over-typing, crossed out items, and handwritten pencil or ink annotations. These corrections have been implemented in this document to give the text that we presume Turing intended. Additionally, there are some hand written notes in the manuscript, which may or may not be by Turing; these are indicated by the use of footnotes.

British English spelling is used in the manuscript and this is retained, so words such as favour, neighbourhood, cancelling, etc. will be encountered. Turing appears to favour the spellings bigramme, trigramme, tetragramme, etc., although he is not always consistent; throughout this document the favoured rendering is used.

Turing's wording is unchanged to give the full flavour of his original style. This means that "That is to say that we suppose for instance that ..... " will be encountered - amongst others!

Both papers end abruptly, no summary or conclusion is offered, perhaps the papers are incomplete or pages are missing. To indicate the end of the manuscript we have marked the end of each paper with a printing sign - an infinity symbol between two horizontal bars.

[1] Zabell, S. 2012. "Commentary on Alan M.Turing: The Applications of Probability to Cryptography" *Cryptologia*, 36:191-214.

In the section on a letter subtractor problem, reference is made to other methods to be discussed later in the paper. This does not happen - perhaps another indicator of an incomplete paper or missing pages.

Finally, Turing uses some forward page references that appear in the manuscript as *see(p)*, obviously intending to return and complete the reference. This also does not happen, so these references remain unresolved.

In short, we strive to represent Turing's text as he wrote it.

Ciphertext, cleartext, etc.

In an attempt to capture the flavour of the time, ciphertext, cleartext, keys, etc. are displayed in a fixed pitch, bold, non-serif font to represent the typewriter, teletype, and telegraph machines that would have printed the original code, *viz.* `CONDITIONS`.

Mathematics

In the manuscript all mathematics is hand written in ink and pencil in spaces left between the typed text. Sometimes adequate space was left, other time not, and the handwriting spills into margins and adjacent lines, adding to the reading challenge. We have cast all mathematics into standard in-line or display formats as appropriate. We have used the `mathcal` font in places to capture the flavour of Turing's handwriting, *e.g.* "the probability p" appears as "the probability $\mathcal{P}$".

Turing uses no punctuation in his mathematics, this has been added to be consistent with modern practice[2]; he also uses letters to reference equations - numbers are used in this document. In many places we have added parentheses to give clarity to an expression, and in some places where Turing is inconsistent in his uses of parentheses for a mathematical phrase (the expression for letter probability in the Vigenère in particular) we have chosen one format and been consistent in its use.

As Turing demonstrates a love of dense mathematics the algebraic multiplication symbol × has occasionally been used for readability, so all standard forms of multiplication will be encountered, *viz.,* $ab, a \times b, a \cdot b$. Finally, convention suggests that the subject of a formula or expression sits on its own on the left hand side of the equals sign, with the subsidiary variables collected on the right hand side. Turing adheres to this convention as it suits him, his preference is retained.

In short, we strive to retain the elegance of Turing's mathematics, whilst casting it into a modern format.

Figures and tables

All figures have been included with rearrangement of some items to improve clarity or document flow. Turing uses a variety of papers, styles, inks, pen, and pencil; these have all been represented in standard figure and table format.

Contents page

Turing provides a rudimentary Contents for *The Applications of Probability to Cryptography*, this has been reworked with some additions to make it more meaningful. *Paper on Statistics of Repetitions*, being much shorter, requires no Contents.

Editors

The editor can be contacted at: *ian.taylor@maths.oxon.org.*

[2] See, for instance, Higham, Nicholas J. 1998. "Handbook of writing for the Mathematical Sciences", SIAM, Philadelphia.

Contents

CHAPTER 1

Introduction

1.1. Preamble

The theory of probability may be used in cryptography with most effect when the type of cipher used is already fully understood, and it only remains to find the actual keys. It is of rather less value when one is trying to diagnose the type of cipher, but if definite rival theories about the type of cipher are suggested it may be used to decide between them.

1.2. Meaning of probability and odds

I shall not attempt to give a systematic account of the theory of probability, but it may be worth while to define shortly *probability* and *odds*. The *probability* of an event on certain evidence is the proportion of cases in which that event may be expected to happen given that evidence. For instance if it is known the 20% of men live to the age of 70, then knowing of Hitler only *Hitler is a man* we can say that the probability of Hitler living to the age of 70 is 0.2. Suppose that we know that *Hitler is now of age 52* the probability will be quite different, say 0.5, because 50% of men of 52 live to 70.

The *odds* of an event happening is the ratio $\mathcal{P}/(1-\mathcal{P})$ where $\mathcal{P}$ is the probability of it happening. This terminology is connected with the common phraseology *odds of 5:2 on* meaning in our terminology that the odds are 5/2.

1.3. Probabilities based on part of the evidence

When the whole evidence about some event is taken into account it may be extremely difficult to estimate the probability of the event, even very approximately, and it may be better to form an estimate based on a part of the evidence, so that the probability may be more easily calculated. This happens in cryptography in a very obvious way. The whole evidence when we are trying to solve a cipher is the complete traffic, and the events in question are the different possible keys, and functions of the keys. Unless the traffic is very small indeed the theoretical answer to the problem "What are the probabilities of the various keys?" will be of the form "The key ... has a probability differing almost imperceptibly from 1 (certainty) and the other keys are virtually impossible". But a direct attempt to determine these probabilities would obviously not be a practical method.

1.4. *A priori* probabilities

The evidence concerning the possibility of an event occurring usually divides into a part about which statistics are available, or some mathematical method can be applied, and a less definite part about which one can only use one's judgement.

Suppose for example that a new kind of traffic has turned up and that only three messages are available. Each message has the letter `V` in the 17th place and `G` in the 18th place. We want to know the probability that it is a general rule that we should find `V` and `G` in these places. We first have to decide how probable it is that a cipher would have such a rule, and as regards this one can probably only guess, and my guess would be about $1/5,000,000$. This judgement is not entirely a guess; some rather insecure mathematical reasoning has gone into it, something like this:-

The chance of there being a rule that two consecutive letters somewhere after the 10th should have certain fixed values seems to be about $1/500$ (this is a complete guess). The chance of the letters being the 17th and 18th is about $1/15$ (another guess, but not quite as much in the air). The probability of a letter being `V` or `G` is $1/676$ (hardly a guess at all, but expressing a judgement that there is no special virtue in the bigramme `VG`). Hence the chance is $1/(500 \times 15 \times 676)$ or about $1/5,000,000$. This is however all so vague, that it is more usual to make the judgment "$1/5,000,000$" without explanation.

The question as to what is the chance of having a rule of this kind might of course be resolved by statistics of some kind, but there is no point in having this very accurate, and of course the experience of the cryptographer itself forms a kind of statistics.

The remainder of the problem is then solved quite mathematically. Let us consider a large number of ciphers *chosen at random.* N of them say. Of these $N/5,000,000$ of them will have the rule in question, and the remainder not. Now if we had three messages of each of the ciphers before us, we should find that for each of the ciphers with the rule, three messages have `VG` in the required place, but of the remaining $(4,999,999 \times N)/5,000,000$ only a proportion $1/676^3$ will have them. Rejecting the ciphers which have not the required characteristics we are left with $N/5,000,000$ cases where the rule holds, and $(4,999,999 \times N)/(5,000,000 \times 676^3)$ cases where it does not. This selection of ciphers is a random selection of ones which have all the known characteristics of the one in question, and therefore the odds in favour of the rule holding are:

$$\frac{N}{5,000,000} : \frac{4,999,999 \times N}{5,000,000 \times 676^3},$$
$$\textit{i.e} \quad 676^3 : 4,999,9999,$$
$$\text{or about } 60 : 1 \text{ on.}$$

It should be noticed that the whole argument is to some extent fallacious, as it is assumed that there are only two possibilities, *viz.* that either `VG` must always occur in that position, or else that the letters in the 17th and 18th positions are wholly random. There are however many other possibilities worth consideration, *e.g.*

(1) On the day in question we have `VG` in the position in question.
(2) Or on another day we have some other fixed pair of letters.
(3) Or in the positions 17, 18 we have to have one of the four combinations `VG`, `RH`, `OM`, `IL` and by chance `VG` has been chosen for all the three messages we have had.
(4) Or the cipher is a simple substitution and `VG` is the substitute of some common bigramme, say `TH`.

The possibilities are of course endless, and it is therefore always necessary to bear in mind the possibility of there being other theories not yet suggested.

The *a priori* probability sometimes has to be estimated as above by some sort of guesswork, but often the situation is more satisfactory. Suppose for example that we know that a certain cipher is a simple substitution, the keys having no specially noticeable properties. Suppose also that we have 50 letters of such a message including five occurrences of P. We want to know how probable it it that P is the substitute of E. As before we have to answer two questions.

(1) How likely is it that P would be the substitute of E neglecting the evidence of the five Es occurring in the message?
(2) How likely are we to get 5 Ps?
  (a) If P is not the substitute of E
  (b) If P is the substitute of E.

I will not attempt to answer the second question for the present. The answer to the first is simply that the probability of a letter being the substitute of E is independent of what the letter is, and is therefore always 1/26, in particular it is 1/26 for the letter P. The only guesswork here is the judgement that the keys are chosen at random.

1.5. The Factor Principle

Nearly all applications of probability to cryptography depend on the *factor principle* (or Bayes' Theorem). This principle may first be illustrated by a simple example. Suppose that one man in five dies of heart failure, and that of the men who die of heart failure two in three die in their beds, but of the men who die from other causes only one in four dies in their beds. (My facts are no doubt hopelessly inaccurate). Now suppose we know that a certain man died in his bed. What is the probability that he died of heart failure? Of all numbering N say we find that

$N\times(1/5)\times(2/3)$	*die in their beds of heart failure*
$N\times(1/5\times(1/3)$	... *elsewhere* ....
$N\times(4/5)\times(1/4)$	*die in their beds from other causes*
$N\times(4/5\times(3/4)$	... *elsewhere* ....

Now as our man died in his bed we do not need to consider the cases of men who did not die in their beds, and these consist of

$N\times(1/5)\times(2/3)$	*cases of heart failure and*
$N\times(4/5)\times(1/4)$	*from other causes*

and therefore the odds are $1\times(2/3):4\times(1/4)$in favour of heart failure. If this had been done algebraically the result would have been

$$\textit{A posteriori odds of the theory} = \textit{A priori odds of the theory} \times \frac{\textit{Probability of the data being fulfilled if the theory is true}}{\textit{Probability of the data being fulfilled if the theory is false}}.$$

In this the *theory* is that the man died of heart failure, and the *data* is that he died in his bed.

The general formula above will be described as the *factor principle*, the ratio

$$\frac{\textit{Probability of the data if the theory is true}}{\textit{Probability of the data if the theory is false}},$$

is called the factor for the theory on account of the data.

1.6. Decibanage

Usually when we are estimating the probability of a theory there will be several independent pieces of evidence *e.g.* following our last example, where we want to know whether a certain man died of heart failure or not, we may know

(1) He died in his bed
(2) His father died of heart failure
(3) His bedroom was on the ground floor

and also have statistics telling us

(a) 2/3 of men who die of heart failure die in their beds
(b) 2/5 .... have fathers who died of heart failure
(c) 1/2 .... have bedroom on the ground floor
(d) 1/4 of men who died from other causes die in their beds
(e) 1/6 .... have fathers who died of heart failure
(f) 1/20 of men who die of other cause have their bedrooms on the ground floor

Let us suppose that the three pieces of evidence are independent of one another if we know that he died of heart failure, and also if we know that he did not die of heart failure. That is to say that we suppose for instance that knowing that he slept on the ground floor does not make it any more likely that he died in his bed if we knew all along that he died of heart failure. When we make these assumptions the probability of a man who died of heart failure satisfying all three conditions is obtained simply by multiplication, and is $(2/3) \times (2/5) \times (1/2)$ and likewise for those who died from other causes the probability is $(1/4) \times (1/6) \times (1/20)$, and the factor in favour of the heart theory failure is

$$\frac{(2/3) \times (2/5) \times (1/2)}{(1/4) \times (1/6) \times (1/20)}.$$

We may regard this as the product of three factors $(2/3)/(1/4)$ and $(2/5)/(1/6)$ and $(1/2)/(1/20)$ arising from from the three independent pieces of evidence. Products like this arise very frequently, and sometimes one will get products involving thousands of factors, and large groups of these factors may be equal. We naturally therefore work in terms of the logarithms of the factors. The logarithm of the factor, taken to the base $10^{1/10}$ is called *decibanage in favour of the theory.* A *deciban* is a unit of evidence; a piece of evidence is worth a deciban if it increase the odds of the theory in the ratio $10^{1/10} : 1$. The deciban is used as a more convenient unit that the *ban.* The terminology was introduced in honor of the famous town of Banbury.

Using this terminology we might say that the fact that our man died in bed scores 4.3 decibans in favour of the heart failure theory $(10 \log(8/3) = 4.3)$. We score a further 3.8 decibans for his father dying of heart failure, and 10 for his having his bedroom on the ground floor, totalling 18.1 decibans. We then bring in the *a priori* odds 1/4 or $10^{-6/10}$ and the result is the the odds are $10^{12.1/10}$, or as we may say "12.1 deciban up on evens". This means about 16:1 on.

CHAPTER 2

Straightforward Cryptographic Problems

2.1. Vigenère

The factor principle can be applied to the solutions of a Vigenère problem with great effect. I will assume here that the period of the cipher has already been determined. Probability theory may be applied to this part of the problem also, but that is not so elementary. Suppose our cipher, written out in its correct period is[1]

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

Figure 1. Vigenère problem.
(It is only by chance that it makes a rectangular array.)

Let us try to find the key for the first column, and for the moment let us only take into account the evidence afforded by the first letter `D`. Let us first consider the key `B`. The factor principle tells us

$$\textit{Odds in favour of key } \texttt{B} = \textit{A priori odds in favour of key } \texttt{B} \times \frac{\textit{Probability of getting } \texttt{D} \textit{ in cipher if key is } \texttt{B}}{\textit{Probability of getting key } \texttt{D} \textit{ in cipher if key is not } \texttt{B}}$$

Now the *a priori* odds in favour of key `B` may be taken as 1/25. The probability of getting `D` in the cipher with the key `B` is just the probability of getting `C` in the clear which (using the count on 1000 letters in Fig 2) is 0.021. If however the key is not `B` we can have any letter other the `C` in the clear, and the probability is (1 - 0.021)/25. Using the evidence of the `D` then the odds in favour of the key `B` are

$$\frac{1}{25} \times \left(\frac{25 \times 0.021}{1 - 0.021} \right).$$

[1] Turing's statement of the ciphertext is slightly different to what he decodes. The N M at the end the first line are reversed to read `DKQHSHZMNP` in Fig 5, which gives the correct cleartext.

We may then consider the effect of the next letter in the column R which gives a further factor of (25 x 0.064)/(1 - 0.064). We are here assuming that the evidence of the R is independent of the evidence of the D. This is not quite correct, but is a useful approximation; a more accurate method of calculation will be given later. Let us write $\mathcal{P}_\alpha$ for the frequency of the letter α in plain language. Then our final estimate for the odds in favour of key B is

$$\frac{1}{25}\prod_i \frac{25\mathcal{P}_{\alpha_i-1}}{1-\mathcal{P}_{\alpha_i-1}}.$$

where $\alpha_1, \alpha_2, \ldots$ is the series of letters in the 1st column, and we use the letters and numbers interchangeably, A meaning 1, B meaning 2, ..., Z meaning 26 or 0. More generally for key β the odds are

$$\frac{1}{25}\prod_i \frac{25\mathcal{P}_{\alpha_i-\beta+1}}{1-\mathcal{P}_{\alpha_i-\beta+1}}.$$

The value of this can be calculated by having a table of the decibanage corresponding to the factors $25\mathcal{P}_\alpha/(1-\mathcal{P}_\alpha)$. One then decodes the column with the various possible keys, looks up the decibanage, and adds them up.

The most convenient form for doing this is a table of values of $20\log_{10}[25\mathcal{P}_\alpha/(1-\mathcal{P}_\alpha)]$, taken to the nearest integer, or as we may say, the values of the score in *half decibans*. One may also have columns showing multiples of these, and the table made of double height[2] (Fig 3). For the first column with key B the decoded column is CQWS••OAV,[3] and we score -5 for C, -26 for Q, -5 for W, 17 for the three letters S, 5 for O, 7 for A and -10 V, totalling -17. These calculations can be done very quickly by the use of the transparent gadget Fig 4 , in which squares are ringed in pencil to show the number of letters occurring in the column.

A	84	J	2	S	73
B	23	K	5	T	81
C	21	L	38	U	19
D	46	M	34	V	11
E	116	N	66	W	21
F	20	O	66	X	16
G	25	P	15	Y	24
H	49	Q	2	Z	3
I	76	R	64		

FIGURE 2. Count on 1000 letters.
(English text)
The value for X has been taken more of less at random as a compromise between real language & telegraphese. Also I added to each entry (see p)[4].

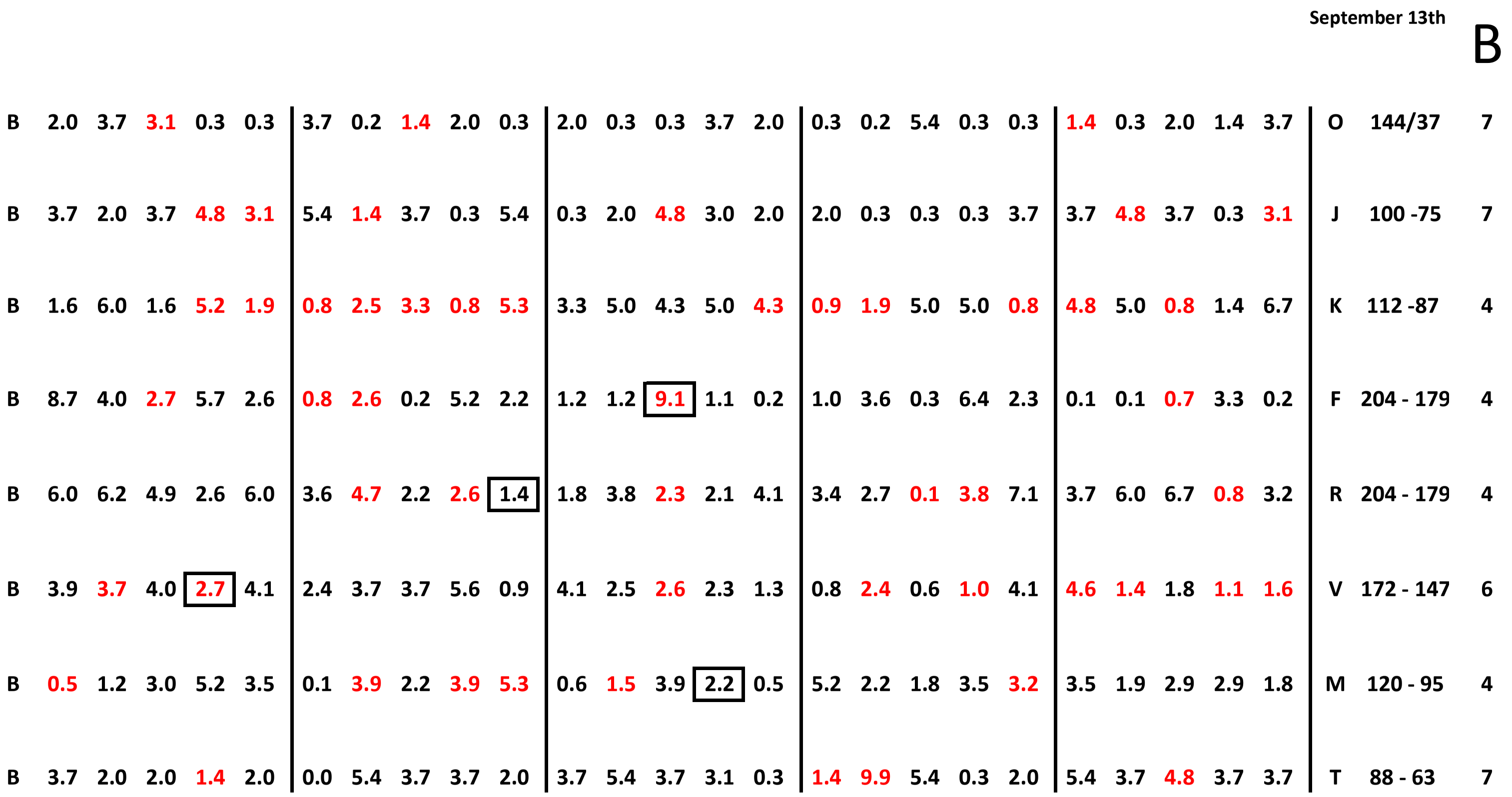
[2] Turing provides a table of double height for Fig 3 to allow the "gadget" of Figure 4 to be used with any letter of the alphabet as a decode key - hence the double alphabet. Figure 4 can be prepared as a transparency, with the original markings cleared, and markings for the new decode letter added. Fig 3 and Fig 4 are correctly proportioned in this document for this to work.

[3] S•• means SSS, for a total of three letter S, as noted in the following arithmetic. The linear decode for the example is CQWSSOAVS

[4] Forward reference left unresolved in the manuscript.

31	26	20	13	7	A
-23	-18	-14	-9	-5	B
-26	-21	-16	-10	-5	C
7	6	4	3	1	D
48	38	29	19	10	E
-28	-22	-17	-11	-6	F
-19	-15	-11	-8	-4	G
10	8	6	4	2	H
29	23	17	12	6	I
-131	-103	-77	-52	-26	J
-99	-79	-59	-40	-20	K
-2	-2	-1	-1	0	L
-6	-5	-4	-2	-1	M
23	18	14	9	5	N
23	18	14	9	5	O
-41	-33	-25	-16	-8	P
-131	-103	-77	-52	-26	Q
22	18	13	9	4	R
28	22	17	11	6	S
32	26	19	13	6	T
-31	-25	-19	-12	-6	U
-54	-43	-32	-22	-10	V
-26	-21	-16	-10	-5	W
-38	-30	-23	-15	-8	X
-20	-16	-12	-8	-4	Y
-111	-89	-67	-44	-22	Z
31	26	20	13	7	A
-23	-18	-14	-9	-5	B
-26	-21	-16	-10	-5	C
7	6	4	3	1	D
48	38	29	19	10	E
-28	-22	-17	-11	-6	F
-19	-15	-11	-8	-4	G
10	8	6	4	2	H
29	23	17	12	6	I
-131	-103	-77	-52	-26	J
-99	-79	-59	-40	-20	K
-2	-2	-1	-1	0	L
-6	-5	-4	-2	-1	M
23	18	14	9	5	N
23	18	14	9	5	O
-41	-33	-25	-16	-8	P
-131	-103	-77	-52	-26	Q
22	18	13	9	4	R
28	22	17	11	6	S
32	26	19	13	6	T
-31	-25	-19	-12	-6	U
-54	-43	-32	-22	-10	V
-26	-21	-16	-10	-5	W
-38	-30	-23	-15	-8	X
-20	-16	-12	-8	-4	Y
-111	-89	-67	-44	-22	Z

FIGURE 3. Table for scoring a Vigenère.
In units of half a deciban.



September 13th

B

B	2.0	3.7	3.1	0.3	0.3	3.7	0.2	1.4	2.0	0.3	2.0	0.3	0.3	3.7	2.0	0.3	0.2	5.4	0.3	0.3	1.4	0.3	2.0	1.4	3.7	O	144/37	7
B	3.7	2.0	3.7	4.8	3.1	5.4	1.4	3.7	0.3	5.4	0.3	2.0	4.8	3.0	2.0	2.0	0.3	0.3	0.3	3.7	3.7	4.8	3.7	0.3	3.1	J	100 -75	7
B	1.6	6.0	1.6	5.2	1.9	0.8	2.5	3.3	0.8	5.3	3.3	5.0	4.3	5.0	4.3	0.9	1.9	5.0	5.0	0.8	4.8	5.0	0.8	1.4	6.7	K	112 -87	4
B	8.7	4.0	2.7	5.7	2.6	0.8	2.6	0.2	5.2	2.2	1.2	1.2	$\boxed{9.1}$	1.1	0.2	1.0	3.6	0.3	6.4	2.3	0.1	0.1	0.7	3.3	0.2	F	204 - 179	4
B	6.0	6.2	4.9	2.6	6.0	3.6	4.7	2.2	2.6	$\boxed{1.4}$	1.8	3.8	2.3	2.1	4.1	3.4	2.7	0.1	3.8	7.1	3.7	6.0	6.7	0.8	3.2	R	204 - 179	4
B	3.9	3.7	4.0	$\boxed{2.7}$	4.1	2.4	3.7	3.7	5.6	0.9	4.1	2.5	2.6	2.3	1.3	0.8	2.4	0.6	1.0	4.1	4.6	1.4	1.8	1.1	1.6	V	172 - 147	6
B	0.5	1.2	3.0	5.2	3.5	0.1	3.9	2.2	3.9	5.3	0.6	1.5	3.9	$\boxed{2.2}$	0.5	5.2	2.2	1.8	3.5	3.2	3.5	1.9	2.9	2.9	1.8	M	120 - 95	4
B	3.7	2.0	2.0	1.4	2.0	0.0	5.4	3.7	3.7	2.0	3.7	5.4	3.7	3.1	0.3	1.4	9.9	5.4	0.3	2.0	5.4	3.7	4.8	3.7	3.7	T	88 - 63	7

Editor -This table is not referenced in the manuscript, does not have a page number, and nothing appear to indicate its purpose. It is presented here for completeness.

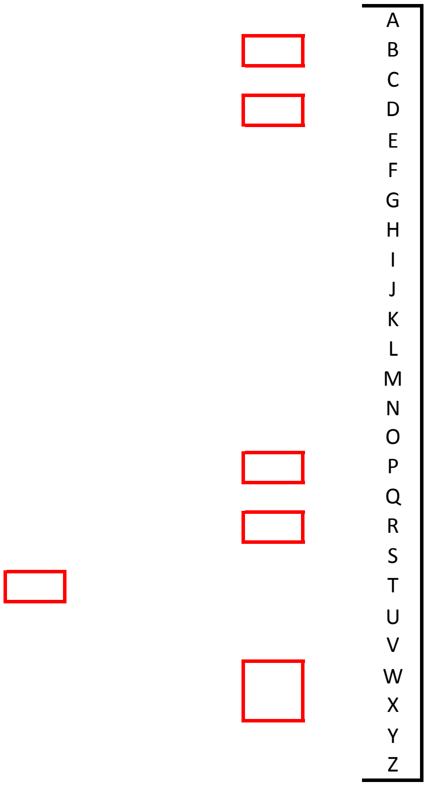


Figure 4. Apparatus for scoring a Vigenère.
Pencil marks arranged for 1st wheel of Fig. 1.

The gadget may be placed over Fig 3 in various positions corresponding to the various keys. The score is obtained by adding up the numbers showing through the various squares. In Fig 5 the alphabet has been written in a vertical below the cipher text of Fig 1, each letter representing a possible key. The score for each key has been written opposite the key, and under the relevant column. An `X` denotes a bad score, not worth adding up. Usually these will be -15 or worse. It will be seen that for the first column `P`, having a score of 43 is extremely likely to be right, especially as there is no other score better than 8. If we neglect this latter fact the odds for the key are $(1/25)10^{2.15}$ *i.e.* about 5:1 on. The effect of decoding this column with key `P` has been shown underneath.

For the second column the best key is `O`, but is by no means so certain as the first column. The decode for this column is also shown, and provides very satisfactory combinations with the first column, confirming both keys. (This confirmation could also be based on probability theory, given a table of bigramme frequencies). In the third column `I` and `C` are best although `D` would be very possible, and in the fourth column `Q` and `U` are best.

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

	A		-2		X	•	X		X	
	B	•	-17		X		X	•	-2	
	C		X	•	3		16	••	X	
	D	•	X		9		9		X	
	E		-6		X		X		X	
	F		X		X		X		X	
	G		X		X		X		-3	
	H		X	••	1		-3	•	X	
	I		X		X		17		X	
	J		X		X		X	•	13	
	K		X	•	X	••	X		-15	
Scores for	L		2		X		X		X	
possible keys	M		X		X		X		X	
	N		X		X		X		X	
	O		X	•	28		X	•	X	
	P	•	43		X	•	X		X	
	Q		X		X	•	X		22	
	R	•	X		X		X		X	
	S		X		X		-6		X	
	T	•••	8		X		-15		X	
	U		X	•	X	•	X	•	22	
	V		X		X	•	X		X	
	W	•	X	••	16	•	1		X	
	X	•	X		X		-15	•	X	
	Y		X		X		-18	•	X	
	Z		X	•	-13	•	X		X	
Best Keys			P		O		IC		QU	
			O		W		IO		RN	G
			C		O		NT		HD	I
			I		T		HN		EA	S
Possible			E		I		MW		TF	O
Decodes			E		T		OU		MI	M
			A		L		CI		YU	L
			M		A		CI		LH	I
			H		I		SY		MI	S
			E		G		RX		IE	T

FIGURE 5. Scoring and solving a Vigenère.

Writing down the possible decodes we see that the first line must read OWING and this makes the other lines read CONDI, ITHAS, EIMPO, ETOIM, ALCUL, MACHI, HISIS, EGRET. By filling in the word CONDITIONS the whole can now be decoded.[5]

[5] Solution: Keylength - 10, Key - POIUMOLQNY, Cleartext - OWINGTOWAR CONDITIONS ITHASBECOM EIMPOSSIBL ETOIMPORTC ALCULATING MACHINESXT HISISVERYR EGRETTABLE

A more accurate argument would run as follows. For the first column, instead of setting up as rival theories the two possibilities that `B` is the key and that `B` is not we can set up 26 rival theories that the key is `A` or `B` or ... `Z`, and we may apply the factor principle in the form:-

$$\frac{\textit{A posteriori probability of key A}}{\textit{A priori probability of key }\texttt{A}\times\textit{ Probability of getting the given column with key }\texttt{A}},$$

$$=\frac{\textit{A posteriori probability of key }\texttt{B}}{\textit{A priori probability of key }\texttt{B}\times\textit{ Probability of getting the given column with key }\texttt{B}},$$

$$= \textit{etc.}$$

The argument to justify this form of factor principle is really the same as for the original form. Let q_β be the *a priori* probability of key β. Then out of N cases we have Nq_β cases of key β. Let $\mathcal{P}\left(\beta, C\right)$ be the probability of getting the column C with key β, then we have rejected the cases where we get columns other than C we find that there are $Nq_\beta\mathcal{P}\left(\beta, C\right)$ cases of key β *i.e.* the *a posteriori* probability of key β is $KNq_\beta\mathcal{P}\left(\beta, C\right)$, where K is independent of β.

We have therefore to calculate the probability of getting the column C with key β and this is simply $\prod_i \mathcal{P}_{(\alpha_i-\beta+1)}$, *i.e.* the product of the frequencies of the decode letters which we get if the key is β.

Since the *a priori* probabilities of the keys are all equal we may say that the *a posteriori* probabilities are in the ratio $\prod_i \mathcal{P}_{\alpha_i-\beta+1}$ *i.e.* in the ratio $\prod_i 26\mathcal{P}_{\alpha_i-\beta+1}$ which is more convenient for calculation. The final value for the probability is then

$$\frac{\prod\limits_i 26\mathcal{P}_{\alpha_i-\beta+1}}{\sum\limits_\beta \prod\limits_i 26\mathcal{P}_{\alpha_i-\beta+1}}.$$

The calculation of the product $\prod_i 26\mathcal{P}_{\alpha_i-\beta+1}$ may be done by the method recommended before for

$$\prod_i \frac{25\mathcal{P}_{\alpha_i-\beta+1}}{1-\mathcal{P}_{\alpha_i-\beta+1}}.$$

$\big($The table in Fig 3 was in fact made up for $\prod_i 26\mathcal{P}_{\alpha_i-\beta+1}$. The differences between the two tables would of course be rather slight$\big)$. The new result is more accurate than the old because of the independence assumption in the original result.

If we only want to know the ratios of the probabilities of the various keys there is no need to calculate the denominator $\sum_\beta \prod_i 26\mathcal{P}_{\alpha_i-\beta+1}$. This denominator has however another importance: it gives us some evidence about other assumptions, such as that the cipher is Vigenère, and that the period is 10. This aspect will be dealt with later (p.)[6].

2.2. A letter subtractor problem

A substitution with the period 91 × 95 × 99 is obtained by superimposing three substitutions of periods 91, 95, and 99, each substitution being a Vigenère composed of slides of 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9.[7] The three substitutions are known in detail, but we do not know for any given message at what point in the complete substitution to begin. For many messages however we can provide a more or less probable crib. How can we test the probability of a crib before attempting to

[6] Forward reference left unresolved in the manuscript.

[7] Equivalent to keys A to J.

solve it? It may be assumed that approximately equal numbers of slides 0, 1, . . . , 9 occur in each substitution.

The principle of the calculation is that owing to the way in which the substitution is built up, not all slides are equally frequent, *e.g.* a slide of 25 can only be the sum of slides of 9, 8, and 8, or 9, 9, and 7 whilst a slide of 15 can be any of the following

9,6,0	8,7,0	7,7,1	6,6,3
9,5,1	8,6,1	7,6,2	6,5,4
9,4,2	8,5,2	7,5,3	
9,3,3	8,4,3	7,4,4	

A crib will therefore, other things being equal, be more likely if it requires a slide of 15 than if it requires a slide of 25. The problem is to make the best use of this principle, by determining the probability of the crib with reasonable accuracy, but without spending long over it.

We have to find the probability of getting a given slide. To do this we can apply several methods.

(a) We can produce a long stretch of key by addition and take a count of the resulting slides. This is obviously a very general method, and requires no special mathematical technique. It may be rather laborious, but by interpreting a small count with common sense one can probably get quite good results.

(b) There are 1000 possible combinations of slides all equally likely *viz.* 000, 001, . . . , 999. We can add up the digits in these and take the remainder on division by 26, and then count the number of combinations giving each of the possible remainders.

(c) We can make use of a trick which might appear to be rather special, but is really applicable to a multitude of problems. Consider the expression

$$f(x) = \left(1 + x + x^2 + \cdots + x^9\right)^3 .$$

For each possible way of expressing a number n as the sum of three numbers 0, . . . , 9, say $n = m_1 + m_2 + m_3$, there is a term $x^{m_1}x^{m_2}x^{m_3}$ in $f(x)$, x^{m_1} coming out of the first factor, x^{m_2} out of the second, and x^{m_3} out of the third. Hence the number of ways of expressing n in the form $n = m_1 + m_2 + m_3$, is the coefficient of x^n in $f(x)$ *i.e.* in

$$\frac{\left(1 - x^{10}\right)^3}{\left(1 - x\right)^3},$$

or in

$$\left(1 - 3x^{10} + 3x^{20} - x^{30}\right)\left(1 - x\right)^{-3} .$$

Expanding $(1 - x)^{-3}$ by the binomial theorem

$$\begin{aligned}(1 - x)^{-3} = 1 &+ 3x + 6x^2 + 10x^3 + 15x^4 + 21x^5 + 28x^6 + 36x^7 \\ &+ 45x^8 + 55x^9 + 66x^{10} + 78x^{11} + 91x^{12} + 105x^{13} \\ &+ 120x^{14} + 136x^{15} + 153x^{16} + 171x^{17} + 190x^{18} \\ &+ 210x^{19} + 231x^{20} + 253x^{21} + 276x^{22} + 300x^{23} \\ &+ 325x^{24} + 351x^{25} + 378x^{26} + 406x^{27} + 435x^{28} + \ldots .\end{aligned}$$

Now multiply by $1 - 3x^{10} + 3x^{20} - x^{30}$ and we get

$$\begin{aligned} f(x) = 1 + 3x + 6x^2 + 10x^3 + 15x^4 + 21x^5 + 28x^6 + 36x^7 \\ + 45x^8 + 55x^9 + 63x^{10} + 69x^{11} + 73x^{12} + 75x^{13} \\ + 75x^{14} + 73x^{15} + 69x^{16} + 63x^{17} + 55x^{18} \\ + 45x^{19} + 36x^{20} + 28x^{21} + 21x^{22} + 15x^{23} \\ + 10x^{24} + 6x^{25} + 3x^{26} + x^{27} \end{aligned}$$

This means to say that the chances of getting totals 0, 1, 2, ... are in the ratio 1, 3, 6, 10, ... The chances of getting remainders of 0, 1, 2, ... on division by 26 are in the ration 4, 4, 6, 10, 15, ... To get true probabilities these must be divided by their total which is conveniently 1000.

(d) There are two other methods, both connected with the last method but not relying so much on the special features of the problem. They will be discussed later.[8]

Suppose then that the probabilities have been calculated by one method or the other (as in fact we have done under (c)). We can then estimate the values of cribs. Let us suppose that a possible crib for a message beginning `MVHWUSXOWBVMMK` was `AMBASSADOR` so that the slides were 12, 9, 6, 22, 2, 0, 23, 11, 14. The slide of 12 gives us some slight evidence in favour of the crib being right for slides of 12 occur with frequency 0.073 with right cribs, whilst with wrong cribs they occur with frequency only 1/26. The factor in favour of the crib is therefore 26 × 0.073 or about 1.9. A similar calculation may be made for each of the slides, but of course the work may be greatly speeded up by having the values of the factors 26 $C_s/1000$ in half decibans tabulated: here C_s is the coefficient of x^s in the above polynomial $f(x)$. The table is given below (Fig 6)

1	0	-20
2	25	-16
3	24	-12
4	23	-8
5	22	-6
6	21	-3
7	20	-1
8	19	1
9	18	3
10	17	4
11	16	5
12	15	6
13	14	6

FIGURE 6. Scores in half decibans of the various slides.

Evaluating this crib by means of this table we score

$$6 + 3 - 3 - 6 - 16 - 20 - 8 + 5 + 6 \ (= -33 \ ),$$

i.e. the crib is worse by a factor of $10^{-33/20}$ than it was before *e.g.* if the *a priori* odds of the crib were 2:1 against it becomes 98:1 against. This crib was in fact made up at random *i.e.* the letters of the cipher text were chosen at random.

[8] No such discussion appears in the manuscript.

Now let us take one made up correctly, *i.e.* really enciphered by the method in question, but with a random chosen key.

N	Y	X	L	N	X	I	Q	H	H
A	M	B	A	S	S	A	D	O	R
13	12	22	11	21	5	8	13	19	16

(slides)

This scores 15 so that if it were originally 2:1 against, it now becomes nearly 3:1 on.

Having decided on a crib the natural way to test it is to have a catalogue of the positions in which a given series of slides is obtained if the 91 period component is omitted. We make 91 different hypotheses as to this third component, draw an inference as to what is the part of the slide arising from the components of periods 95 and 99 combined. This we look up in the catalogue. This process is fairly lengthy, and as the scoring of the crib takes only a minute it is certainly worth doing.

2.3. Theory of repeats

Suppose we have a cipher in which there are several very long series of substitutions which can be used for enciphering a message, but that one may sometimes get two messages enciphered with the same series of substitutions (or possibly, the series of substitutions for one message being those for another with some at the beginning omitted). In such a case let us say that the messages *fit*, or that they fit at such and such a distance, the distance being the number of substitutions which have to be omitted from the one series to obtain the other series. One will frequently want to know whether two messages fit or not, and we may find some evidence about this by examining the repeats between them.

By the repeats between them I mean this. One writes out the cipher texts of the two messages with the letters which are thought to have been enciphered with the same substitution under one another. One then writes under these messages a series of letters O and X, an O being written where the cipher texts differ and an X where they agree. The series of letters O and X will begin where the second message begins and end where the first to end ends. This series of letters O and X may be called the repetition figure. It may be completed by adding at the ends an indication of how many letters there are which do not overlap, and which message they belong to.

As an example:

```
GFRLIKQGVBMILAFIXMMOROGBYSKYXDAZCHMUMRKBZLDLDDOHCMVTIPRSD
        VLOVDYQCEJSOPYGBMBKYXDAZNBFIOPTFCXDOD
```

^{8}XOOOOOOOOOOXOOXXOOXXXXXXOOOOOOOOOOXOX^{11}

On the whole one expects that a fit is more likely to be right the more letters X there are in the repetition figure, and that long series of letters X are especially desirable. This is because it would not be very unusual for two fairly common words to lie directly under one another when the clear texts are written out, thus

```
THEMAINCONVOYWILLARRIVE ...
    ALLCONVOYSMUSTREPORT ...
    XOOXXXXXXOOOOOXOOOO ...
```

If the corresponding cipher texts really fit, *i.e.* if the letters in the same column are enciphered with the same substitution, then the condition for an X in the repetition figure of the cipher texts is that there be an X in the repetition figure of the corresponding clear text. Now series of several consecutive letters X can occur quite easily as above by two identical words coming under one another, or by such combinations as

```
ITISEASIERTOTEACHTHANALGEBRA ...
   THERAINWASSUCHTHATHECOULD ...
   OOOOOOOOOOOOXXXXXOOOOOOOO ...
```

if the messages really fit, but if not they can only occur by complete coincidence. One therefore tends to believe that there is a fit when one gets such series of letters X. As regards single cases of X the value of them is not so clear, but one can see that if $\mathcal{P}_\alpha$ is the frequency of letters α in plain language then the frequency of letters X as a whole in comparison of plain language with plain language is $\sum_\alpha \mathcal{P}_\alpha^2$, whilst for wrong fits of cipher text it is 1/26 which is necessarily less. Given a sufficiently long repetition figure one should therefore be able to tell whether it is a fit or not simply by counting the letters X and O.

So much is well known. The real point of this section is to show these ideas can be developed into an accurate method of estimating the probabilities of fits.

2.3.1. Simple form of theory. The complete theory takes account of the various possible lengths of repeat. As this theory is somewhat complicated it will be as well to give first two simplified forms of the theory. In both cases the simplification arises by neglecting a part of the evidence. In the first simplified form of theory we neglect all evidence except the number of letters X and the number of letters O. In the other simplified form the evidence is the number of series of (say) four consecutive letters X in a repetition figure.

When our evidence is just the number of times X occurs in the repetition figure, (n let us say) and the length of the repetition figure (N say), then the factor in favour of the fit is

$$\frac{\textit{Probability of a right repetition figure of length } N \textit{ and } n \textit{ occurrences of } \mathtt{X}}{\textit{Probability of a wrong repetition figure of length } N \textit{ having } n \textit{ occurrences of } \mathtt{X}}.$$

As an approximation we may assume that the numerator of this expression has the same value as if the right repetition figures were produced letter by letter by independent random choices, with a certain fixed probability of getting an X at each stage. This probability will have to be $\beta = \sum_\alpha \mathcal{P}_\alpha^2$. The numerator is then

$$(\textit{Number of repetition patterns with length } N \textit{ and } n \textit{ occurrences of } \mathtt{X})$$
$$\times(\textit{Probability of getting a given such repetition pattern by the process just mentioned}),$$

which we may write as $R(N,n)Q(N,n)$. Now let us denote by y_i the ith symbol of the given repetition pattern and put $\tau_x = \beta$ and $\tau_0 = 1 - \beta$. Then $Q(N,n)$, the probability of getting the repetition pattern is $\prod_{i=1}^{N} \tau_{y_i}$ which simplifies to $\beta^n(1-\beta)^{N-n}$. We may do a similar calculation for the denominator, but here we must take $\beta = 1/26$ since all letters occur equally frequently in the cipher. The denominator is then

$$R(N,n)\left(\frac{1}{26}\right)^n \left(\frac{25}{26}\right)^{N-n}.$$

In dividing to find the factor for the fit $R(N,n)$ cancels out, leaving

$$(26\beta)^n \left(\frac{26}{25}(1-\beta)\right)^{N-n}.$$

In other words we score a factor of 26β for an `X` and a factor of $(26/25)(1-\beta)$ for an `O`. More convenient is to regard it as $10\log_{10}\big[(25\beta)/(1-\beta)\big]$ decibans for an `X` and $10\log_{10}[(26/25)/(1-\beta)]$ per unit length of repetition figure (*per unit overlap*).

An alternative argument, leading to the same result, runs as follows. Having decided to neglect all evidence except the overlap and the number of repeats we pretend that nothing else matters, *i.e.* that the form of the figure is irrelevant. In this case we can regard each letter of the repetition figure as independent evidence about the fit. If we get an `X` the factor for the fit is

$$\frac{\textit{Probability of getting an } \texttt{X} \textit{ if the fit is right}}{\textit{Probability of getting an } \texttt{X} \textit{ if the fit is wrong}},$$

i.e. $\beta/(1/26)$. Similarly the factor for an `O` is $(1-\beta)/(25/26)$.

In either form of argument it is unnecessary to calculate the number $R(N,n)$. In this particular case there is no particular difficulty about about it: it is the binomial coefficient. In some similar problems this cancelling out is a great boon, as we might not be able to find any simple form for the factor which cancels. The cancelling out is a normal feature of this kind of problem, and it seems quite natural that it should happen when we think of the second form of argument in which we think of the evidence as consisting of a number of independent parts.

The device of assuming, as we have done here, that the evidence which is not available is irrelevant can often be used and usually leads to good results. It is of course not supposed that the evidence really is irrelevant, but only that the error resulting from the assumption when used in this kind of way is likely to be small.

2.3.2. Second simplified form of theory. In the second simplified form of theory we take as our evidence that a particular part of the repetition figure is `OXXXXO` (say, or alternatively `OXXXXXO` say). The factor is then

$$\frac{\textit{Frequency of } \texttt{OXXXXO} \textit{ in right repetition figures}}{\textit{Frequency of } \texttt{OXXXXO} \textit{ in wrong repetition figures}}.$$

The denominator is

$$\left(\frac{1}{26}\right)^4 \left(\frac{25}{26}\right)^2,$$

and the numerator may be estimated by taking a sample of language hexagrams and counting the number of pairs that have the repetition figure `OXXXXO`. The expectation of the number of such pairs is the sum for all pairs of the probabilities of those pairs having the desired repetition figure *i.e.* is the number of such pairs (*viz* $N(N-1)/2$ where N is the size of the sample) multiplied by the frequency of `OXXXXO` repetition figures. This frequency may therefore be obtained by division if we equate the expected number of these repetition figures to the actual number.

2.3.3. General form of theory. It is not of course possible to have statistics of every conceivable repetition figure. We must make some assumptions to reduce the variety that need to be considered. The following assumption is theoretically very convenient, and also appears to be a very good approximation.

The probability of repeats at two points known to be separated by a point where there is known to be no repeat are independent.

We may also assume that the probability of a repeat is independent of anything but the repetition figure in this neighbourhood. (We may however as a refinement produce different positions in a message). We can therefore think of repetition figures as being produced by selecting the symbols of the figure consecutively, the probability of getting an `X` at each stage being determined by the repetition figure from the point in question back as far as the last `O`. Sometimes this will take us back as far as the beginning of the message, and will include the number telling us how many more letters there are which do not repeat at all. We need in practice only distinguish two cases, where this number is 0 and when it is more. We may also neglect the question as to which message occurs first. We therefore have to distinguish the following cases

`O`	a_0	some	b_0	none	c_0
`OX`	a_1	some `X`	b_1	none `X`	c_1
`OXX`	a_2	some `XX`	b_2	none `XX`	c_2
`OXXX`	a_3	some `XXX`	b_3	none `XXX`	c_3
...		...		...	

The entries $a_0, a_1, b_0, etc.$ opposite the repetition figures are the notations we are adopting for the probability of getting another `X` following such a figure. Strictly speaking we should also bring in a notation for the probability of the message coming to an end after any given repetition figure. As the repeats at the end of a comparison do not appear to behave very differently from those in the main part of the message I shall neglect this complication by assuming that the probability of getting an `O` added to the probability of getting an `X` is 1, and that afterwards one cuts off the end of the series arbitrarily.

Let us calculate the factor for the repeat figure[9]

none	X	X	X	X	O	O	O	X
	c_0	c_1	c_2	c_3	$1-c_4$	$1-a_0$	$1-a_0$	a_0
	1/26	1/26	1/26	1/26	25/26	25/26	25/26	1/26

O	X	X	X	O	O	X	X	some
$1-a_1$	a_0	a_1	a_2	$1-a_3$	$1-a_0$	a_0	a_1	
25/26	1/26	1/26	1/26	25/26	25/26	1/26	1/26	

Underneath each symbol has been written the probability that one would get that symbol, knowing the ones which precede, both for the case of a right and of a wrong repetition figure. The factor for the fit is the product of the first row divided

[9] In the manuscript, Turing squeezes the figure into three lines by spilling into the margins and use of pen and ink. The typeset equivalent is unreadable, so the figure has been split into a left and right components.

Reassemble as: `none X X X X O | O | O | X O | X X X O | O | X X | some` by the product of the second. It is convenient to split this up as indicated by the vertical lines into the product of

$$\frac{c_0c_1c_2c_3\,(1-c_4)}{(1/26)^4\times(25/26)},$$

$$\frac{1-a_0}{(25/26)}, \qquad \text{- occurring three times,}$$

$$\frac{a_0(1-a_1)}{(1/26)\times(25/26)},$$

$$\frac{a_0a_1a_2\,(1-a_3)}{(1/26)^3\times(25/26)},$$

$$\frac{a_0a_1}{(1/26)^2},$$

and this product may be put into the form of the product of

$$\frac{c_0c_1c_2c_3\,(1-c_4)}{(1/26)^4\times(25/26)}\times\left(\frac{1-a_0}{(25/26)}\right)^{-5},$$

- which we call the factor for an initial tetragramme repeat level,

$$\frac{a_0(1-a_1)}{(1/26)\times(25/26)}\times\left(\frac{1-a_0}{(25/26)}\right)^{-2},$$

- the factor for a single repeat,

$$\frac{a_0a_1a_2\,(1-a_3)}{(1/26)^3\times(25/26)}\times\left(\frac{1-a_0}{(25/26)}\right)^{-4},$$

- the factor for a trigramme,

$$\frac{1-a_0}{1-a_2},$$

- the correction for a final bigramme,

$$\left(\frac{1-a_0}{(25/26)}\right)^{16},$$

- the factor for an overlap of 16,

$$\frac{a_0a_1\,(1-a_2)}{(1/26)^2\times(25/26)}\times\left(\frac{1-a_0}{(25/26)}\right)^{-3},$$

- the factor for a trigramme.

We shall neglect the correction for a final bigramme (or whatever it may be). It is in any case rather small, and vanishes if the repetition figure ends with 0; also with our conventions the whole question of the ends of repetition figures has been left rather in doubt.

Now let us put[10]

$$\begin{aligned}a_0a_1\dots a_r(1-a_{r+1}) &= k_r,\\ b_0b_1\dots b_r(1-b_{r+1}) &= j_r,\\ c_0c_1\dots c_r(1-c_{r+1}) &= i_r.\end{aligned}$$

The values of the i_r can be obtained as follows. We take a number of plain language messages and leave out two or three words at the beginning. Then combine the messages to form one long message; this message may be made to *eat its own tail i.e.* it may be written round a circle. If the message were compared with itself in every possible position, except level, we should expect to get repetition figures which when divided up as shown by vertical lines after each `O`, containing $(N(N-1)/2)k_r$ $(= N_r)$ parts which consist of r symbols `O`, or as we may say N_r *actual r-gramme repeats*, where h is the probability of an `O` .

The values of N_r can be calculated given the *apparent number of r-gramme repeats* M_r for each r. This apparent number of r-gramme repeats is the number of series of r consecutive symbols `X` in the repetition figures regardless of what precedes or follows the series.

By considering the ways in which an actual repeat can give rise to the apparent repeat of various lengths we see that

$$M_r = N_r + 2N_{r+1} + 3N_{r+2} + \dots,$$

and therefore

$$M_r - M_{r+1} = N_r + N_{r+1} + N_{r+2} + \dots,$$

and

$$(M_r - M_{r+1}) - (M_{r+1} - M_{r+2}) = N_r.$$

The calculation of j_r may perhaps best be done by comparing the beginners of a number of messages with the long circular message, and the values of i_r by comparing the beginners among themselves. A similar technique of actual and apparent numbers of repeats can be used. I shall not go into this in detail. The formulae required may now be assembled.

$$\begin{aligned}\mu_r &= \text{decibanage for an } r\text{-gramme repeat},\\ \gamma &= \text{negative decibanage for unit overlap},\\ S_{\beta,r} &= \text{number of occurrences in the statistics of the } r\text{-gramme } \beta,\\ N &= \text{total number of letters in the statistics.}\end{aligned}$$

Then if

$$\begin{aligned}M_r &= \sum_\beta \frac{S_{\beta,r}\,(S_{\beta,r}-1)}{2},\\ N_r &= M_r - 2M_{r+1} + M_{r+2},\\ L &= \frac{N(N-1)}{2},\\ k_r &= \frac{N_r}{Lh}.\end{aligned}$$

[10] The manuscript has a pencilled note beside k_r indicating it is to be read as k_{r+1}. We presume that this also means that j_r should be j_{r+1}, and i_r should be i_{r+1}, However, these are not indicated and no changes are made in the subsequent text. We leave the text unchanged

h may be calculated as follows. From the identity

$$(1-a_0)+a_0\,(1-a_1)+a_0a_1\,(1-a_2)+\cdots=1,$$

we get

$$\begin{aligned} k_0+k_1+k_2+\cdots&=1,\\ \therefore\qquad \frac{L-M_1}{Lh}&=1,\\ (1-a_0)=k_0=\frac{N_0}{L-M_1}&=\frac{L-2M_1+M_2}{L-M_1},\\ \mu_r=10\log_{10}\left(\frac{26^{r+1}k_r}{25}\right)&+(r+1)\,\nu,\\ \nu=-10\log_{10}\left(\frac{26(1-a_0)}{25}\right)&. \end{aligned}$$

2.4. Transposition ciphers

2.4.1. A probability problem. In making calculations about substitution ciphers we have often found it useful to treat the plain language as if it were produced by independent choices for the letters, using certain fixed frequencies with which the letters are chosen. Our method for Vigenère and one of the simplified forms of repeat theory could be based on this sort of assumption. With a transposition cipher however such an assumption would be useless or worse than useless, for it would result in the conclusion that all transpositions were equally likely. We have therefore to take a slightly less crude assumption, and the one which suggests itself is that the letters forming the plain language are chosen consecutively, the probability of getting a particular letter depending only on what the letter is and what the preceding letter was. It is easily verified the if $\mathcal{P}_{\alpha\beta}$ is the proportion of bigrammes $\alpha\beta$ in plain language and $\mathcal{P}_\alpha$ the frequency of the letter α then the probability $q_{\alpha\beta}$ of a letter β following an α is $\mathcal{P}_{\alpha\beta}/\mathcal{P}_\alpha$. The probability of a piece of plain language of length L letters saying $\alpha_1\alpha_2\ldots\alpha_L$ is then

$$\mathcal{P}_{\alpha_1}\times q_{\alpha_1\alpha_2}\times q_{\alpha_2\alpha_3}\times q_{\alpha_3\alpha_4}\times\cdots\times q_{\alpha_{(L-1)}\alpha_L},$$

which may also be written as

$$\mathcal{J}\,(\alpha_1,\ldots,\alpha_L)\,.$$

We may also calculate the probability of a given piece of plain language having certain given letters in given places, the remainder of the message being unspecified. The probability is given by

$$\sum\,(\xi_1,\ldots,\xi_L \text{ consistent with data })\,\mathcal{J}\,(\xi_1,\ldots,\xi_L)\,,$$

and if the data is that the known letters are

$$\underset{n_1\text{ dots}}{\cdots}\ \beta_1\qquad \underset{n_2\text{ dots}}{\cdots}\ \beta_2\qquad\cdots\qquad\cdots\beta_{r-1}\qquad \underset{n_r\text{ dots}}{\cdots}\ \beta_r\qquad\cdots\,,\tag{1}$$

it is approximately[11]

$$\prod_r \mathcal{P}_{\beta_r}\cdot\prod_{n_{r+1}=0}\frac{\mathcal{P}_{\beta_r\beta_{r+1}}}{\mathcal{P}_{\beta r}\mathcal{P}_{\beta_{r+1}}}.\tag{2}$$

[11] The manuscript has as the first term $\prod_r \beta_r$, a pencilled annotation indicates that the β_r is to be read as $\mathcal{P}_{\beta_r}$. This substitution has been made in the text.

A more or less rigorous deduction of this approximation from the assumptions above is given at the end of the section. For the present let us see how it can be applied. If we have two theories about the transposition of which the one requires the above pattern of letters, and the other brings the same letters in to positions in which no two of them are consecutive, then the factor in favour of the first as compared with the second is

$$\prod_{n_{r+1}=0} \frac{\mathcal{P}_{\beta_r\beta_{r+1}}}{\mathcal{P}_{\beta r}\mathcal{P}_{\beta_{r+1}}}.$$

We can apply this straightforwardly to the case of a simple transposition by columns. The following text is known to be a simple transposition of a certain type of German text with a key length of not more than 15.[12]

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

To solve this transposition, we may try comparing the first six letters of `S A T P T W` which we know form part of one column with each other series of six letters in the message, for we know that one such comparison will give entirely bigrammes occurring in the decode. We may try first

```
S F
A A
T S
P T
T A
W U
```

The factor for a transposition which brings these letters together, as compared with one which leaves them apart is

$$\frac{\mathcal{P}_{SF}}{\mathcal{P}_S\mathcal{P}_F} \times \frac{\mathcal{P}_{AA}}{\mathcal{P}_A\mathcal{P}_A} \times \cdots \times \frac{\mathcal{P}_{WU}}{\mathcal{P}_W\mathcal{P}_U}.$$

By using a table of values of

$$20\log_{10}\left(\frac{\mathcal{P}_{\alpha\beta}}{\mathcal{P}_\alpha\mathcal{P}_\beta}\right),$$

made up for the type of traffic in question, and given to the nearest integer (table of values of $\mathcal{P}_{\alpha\beta}/(\mathcal{P}_\alpha\mathcal{P}_\beta)$ expressed in half-decibans) we get the product by addition. Such a table is shown in Fig 6. The scores for this particular columns are `SF -7, AA -7, TS -2, PT -10, TA -3, WU -13`, totalling -36. If we consider this combination as *a priori* about 100:1 against (there are 95 letters in the message) it is *a posteriori* about 3000:1 against.

[12] As for the Vigenère problem above, Turing's statement of the ciphertext is slightly different from that which he scores for decryption. The second line in the ciphertext below begins `NLTS`, however, this changes to `NITS` in the scoring example in Figure 7 below. See also the notes accompanying the cleartext.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
A	-7	5	10	-4	-5	-7	3	-4	-25	-7	-8	-1	1	3	-20	-6	-2	-1	-6	-6	8	0	-11	-20	-7	-13	A
B	-7	6	-10	-9	9	-13	-6	-13	1	-11	-12	-7	-2	-18	4	-13	-2	1	-13	-15	-1	-16	-12	-18	-4	-12	B
C	-14	-3	-3	-18	-19	-20	-19	27	-21	-10	3	-12	-12	-21	-15	-13	-2	-14	-14	-21	-21	-5	-17	-17	-14	-17	C
D	5	-8	-18	-18	4	-6	-13	-16	-4	2	-9	-11	-6	-13	4	-4	-2	9	-6	-11	2	-2	-9	-10	-5	-4	D
E	-15	4	0	-8	-15	-5	-2	-5	10	-8	-14	1	-1	5	-22	-6	-3	8	-2	-6	-5	-6	-11	-13	-8	-4	E
F	-2	-11	-20	-9	-2	10	-3	-15	-12	4	-2	1	-8	-11	6	0	-3	-8	-8	-1	9	-8	-8	1	-2	-1	F
G	-1	-8	-18	-5	8	-10	-2	-13	-10	2	6	-3	-13	-11	-10	-14	-3	0	-3	-9	-2	-7	-7	1	2	-2	G
H	1	-10	-12	-8	4	-5	-11	-12	-2	5	-10	2	-3	-10	-2	-2	-2	4	-1	9	-11	-7	-4	-7	-15	-8	H
I	-14	-4	10	-10	0	-1	2	-18	-17	-6	-5	0	-1	9	-17	-7	-1	-10	-1	4	-19	-7	-16	-3	-9	-4	I
J	3	3	-4	1	-3	0	-1	2	-6	14	1	-3	1	-7	-3	7	-2	-12	1	-8	-4	5	-2	0	9	-12	J
K	-2	-9	-12	3	-3	1	-7	-9	-9	4	20	-2	1	-14	-15	-13	7	-6	-5	0	0	-3	-6	-17	-5	-8	K
L	6	0	-6	2	-4	-7	-1	-15	1	-3	-14	8	-4	-2	-2	-5	8	-18	-5	-5	2	-1	-10	3	-5	-3	L
M	6	-1	-17	-6	1	-9	-6	-5	5	1	-10	-14	15	-14	0	-2	-2	-14	-6	-5	-11	-4	-7	3	-6	2	M
N	-1	-8	-18	10	-6	3	11	-9	-8	2	-6	-11	-5	-7	-2	-9	2	-20	6	-6	4	-3	-6	3	0	-5	N
O	-10	-8	-10	-6	-3	4	-6	-13	-18	0	-1	2	6	0	1	2	-2	9	0	2	-18	3	-11	6	-6	-2	O
P	2	-7	-13	-13	4	-3	-14	-5	-8	3	-12	0	-2	-8	6	8	-2	5	-15	-10	5	-12	-12	-13	-11	-1	P
Q	-3	-2	-2	-2	-3	-3	-3	-2	-2	-1	-3	-2	-2	-3	-2	-2	3	-3	-3	-3	13	-2	-2	-2	-2	-2	Q
R	2	3	-3	5	2	0	-6	-10	-2	2	-1	-6	-2	-9	-6	1	-1	-11	-1	2	-1	-2	0	1	-1	2	R
S	-3	-1	11	-2	-4	-7	-13	-12	2	-6	1	-9	-10	-7	-5	8	-3	-19	5	7	-8	5	-9	-15	1	4	S
T	3	-3	-14	-7	5	-3	-11	-11	-4	1	-1	-4	-5	-9	-2	-7	-3	-2	-2	5	-5	-3	-4	2	-3	9	T
U	-11	-4	-3	-15	0	5	-6	-2	-26	-12	-16	10	-2	8	-11	3	2	-3	-3	-11	-2	-9	-11	-12	-9	-20	U
V	-13	-16	-15	-18	2	-15	-16	-5	12	-10	-7	-16	-15	-17	14	-12	-2	-17	-18	-8	-11	10	-14	-4	8	-8	V
W	-1	-17	-17	-17	4	-18	-18	-18	2	-10	-9	-13	-10	-20	21	-12	-2	-19	-20	-19	-13	-14	-3	-16	-6	-16	W
X	1	1	-13	6	1	0	-4	-13	-14	3	0	-12	2	-10	-12	-2	8	-15	-6	-4	-5	9	1	10	-6	3	X
Y	-11	-1	-14	7	-5	-4	-9	-9	-11	-9	0	-9	-6	-5	-4	-11	-2	-11	-6	-6	-7	9	-2	-13	25	-2	Y
Z	-13	-13	-17	-13	-4	-12	-8	-14	-7	-3	-10	-11	-17	-16	-1	-5	-2	-20	-13	-10	8	-8	27	0	-6	-2	Z
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	

FIGURE 6. Exclusive bigramme scores in half decibans, *i.e.* $20\log_{10}\left(\frac{\mathcal{P}_{\alpha\beta}}{\mathcal{P}_\alpha\mathcal{P}_\beta}\right)$, for a certain kind of German traffic.

-36	S		
X	A		
X	T		
X	P		
X	T		
X	W		X
X	S	-10	X
X	F	-36	X
-19	A	-26	X
X	S	16	X
X	T	-6	6
-10	A	18	-7
X	U	0	X
X	T	9	7
X	E	-22	-8
X	E	-24	X
X	A	-39	X
X	I	-25	-10
X	E	-27	-15
X	U	-23	X
X	F	-43	X
X	H	-27	X
X	W	-53	X
X	T	-38	-22
X	J	-60	X
X	T	-52	-25
X	D	-37	X
X	D	-44	-7
X	G	-45	X
7	C	2	X
X	N	-56	X
X	I	-22	X
X	T	-11	-15
X	S	-19	X
X	E	-18	$\boxed{27}$
	F		
	C		
	U		
	I		
	E		

X	F	-22
13	C	37
X	U	-23
X	I	-1
X	E	-2
X	B	-49
X	O	-44
X	E	-59
-22	Y	-31
X	Q	-25
X	H	8
X	G	-19
-26	T	-4
X	J	$\boxed{-4}$
X	T	4
X	E	6
X	E	-32
X	F	-38
X	I	1
X	E	-52
-11	O	-14
X	R	-46
-13	T	-19
X	A	-26
X	R	-45
X	U	-47
X	R	-87
X	N	-54
X	L	-33
X	N	-16
-13	N	-11
-5	N	17
-24	N	-36
X	A	-25
X	I	-40
	E	
	O	
	T	
	U	
	S	

X	E	-36
X	O	-55
X	T	-8
X	U	-40
X	S	-18
X	H	-44
X	L	-43
X	E	-38
X	S	-32
$\boxed{7}$	B	-32
X	F	-35
-17	B	-40
-38	R	-67
X	N	-42
	D	-58
	X	-40
	G	-20
	N	-37
	J	-25
	H	-42
	U	
	A	
	N	
	W	
	R	

FIGURE 7. Scoring the matching of columns in a simple transposition. Correct matchings noted.

Similar scoring may be done for every possible comparison of S A T P T W with six consecutive letters of the message. The comparison may be made both with S A T P T W as earlier and as later column; one may also use the last six letters of the message H U A N W R.

The results of doing this are shown in Fig 7. The message has been written out vertically. The first columns of figures after the message gives the score for S A T P T W as earlier column, entered against the first letter of the later column, *e.g.* the -36 as calculated above gets entered against the F of F A S T A U. The second column after the message consists of the scores for H U A N W R as first column [and the column before the message gives scores for H U A N W R as second column].[13] One of these columns has been worked out in detail but in the other two crosses have been put in where the scores are very bad.

The scores which eventually turned to to be right are ringed. The fourth comparison, which did not have to be done scored very badly *viz.* -27. Amongst the good scores which were wrong there was one score of 37. It was not difficult to see that this one was wrong as most of the score came from W O with requires Z to precede it, and there was no Z in the message. Apart from this fact the comparison was about evens, although if we take into account the fact that there was no better score it would be better.[14] [We have already had a case of this kind of thing in connection with Vigenère; if the various positions are *a priori* equally likely and the factors are $f_1, f_2, \ldots, f_N$ then the value $f_r / \sum f_i$ for the probability of the rth alternative is better than $(f_r/N) \,/\, (1 + f_r/N)$].

2.4.2. The Probability Formula. (Semi) Rigorous deduction of the formula (2) on page 20. (This is something of a digression).

The probability of a piece of plain language coinciding where necessary with the data (1) on page 20 is

$$\mathcal{P}_{\beta_1}\mathcal{T}_{n_2,\beta_1\beta_2}\mathcal{T}_{n_3,\beta_2\beta_3}\ldots\mathcal{T}_{n_m,\beta_{m-1}\beta_m},$$

where

$$\mathcal{T}_{n,\alpha\beta} \quad \text{is} \quad \sum_{\eta_1\eta_1\ldots\eta_n} q_{\alpha\eta_1}q_{\eta_1\eta_2}\ldots q_{\eta_n\beta},$$

since

$$\sum_{\eta_1\ldots\eta_{n_1}} \mathcal{P}_{\eta_1}q_{\eta_1\eta_2}\ldots q_{\eta_{n_1}\beta_1} = \mathcal{P}_{\beta_1}.$$

We can put

$$\mathcal{T}_{n,\alpha\beta} = \left(\mathcal{Q}^{n+1}\right)_{\alpha\beta},$$

where $\mathcal{Q}$ is the matrix whose $\alpha\beta$ coefficient is $q_{\alpha\beta}$. The formula (2) on page 20 would then be accurate if we could say that for $n > 0$,

$$\left(\mathcal{Q}^{n+1}\right)_{\alpha\beta} = \mathcal{P}_\beta.$$

This is not true, but it is true that except for very special values for $q_{\alpha\beta}$,

$$\left(\mathcal{Q}^{n}\right)_{\alpha\beta} \to \mathcal{P}_\beta, \text{ as } n \to \infty,$$

[13] *. . . and the column* to end of sentence, has the note in pencil: *I doubt it - S.W.*

[14] Using Turing's scoring recommendations and a key length of 12 with sequence 5, 11, 8, 7, 3, 10, 6, 12, 9, 4, 1, 2, a cleartext emerges: BNTO SJJ ALBA RFJ STATT IN OST B HEUTE DEN ETA RUFS PEDUNYAR NACHT FGFNQUUDNUL WICH AHTR X WIESEN WI GEN GRESFOITE TE. With Turing's original statement of the ciphertext, as noted above, GRESFOITE becomes GRESFOLTE. Turing scores for the I and not for L, although it makes no differences in the decision to align bigrams.

and the convergence is rather rapid.

To prove this I shall assume that the eigenvalues of $\mathcal{Q}$ are all different in modulus. In this case we can find a matrix $\mathcal{U}$ with unit determinant, such that $\mathcal{U}^{-1}\mathcal{Q}\,\mathcal{U}$ is in the diagonal form

$$\mathcal{M} = \mathcal{U}^{-1}\mathcal{Q}\,\mathcal{U} = \begin{pmatrix} \mu_1 & 0 & 0 & \cdots & 0 \\ 0 & \mu_2 & 0 & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & \mu_{25} & 0 \\ 0 & \cdots & 0 & 0 & \mu_{26} \end{pmatrix},$$

since $\mathcal{Q}\mathcal{U} = \mathcal{U}\mathcal{M}$ we have

$$\sum_\gamma q_{\alpha\gamma} u_{\gamma\beta} = \sum_\xi u_{\alpha\xi} m_{\xi\beta},$$

i.e.

$$\sum_\gamma q_{\alpha\gamma} u_{\gamma\beta} = \mu_\beta \mu_{\alpha\beta}.$$

That is, for each $\beta, u_{\alpha\beta}$ provides a solution of

$$\sum_\gamma q_{\alpha\gamma} l_\gamma = \mu l_\alpha, \qquad (3)$$

with $\mu = \mu_\beta$. Conversely if we have any solution of (3) then $\mu = \mu_\vartheta, l_\alpha = k u_{\alpha\vartheta}$ for some k, ϑ and all α, for as $\mathcal{U}$ is non singular we can find numbers c_γ such that

$$l_\alpha = \sum_\gamma u_{\alpha\gamma} c_\gamma \text{ for all } \alpha,$$

and then substituting in (3) we get

$$\sum_{\gamma,\delta} q_{\alpha\gamma} u_{\gamma\delta} c_\delta = \mu \sum_\delta u_{\alpha\delta} c_\delta,$$

i.e.

$$\sum_\delta \left(\mu_\delta c_\delta - \mu c_\delta\right) u_{\alpha\delta} = 0.$$

Which, since $\mathcal{U}$ is nonsingular implies $\mu = \mu_\delta$ or $c_\delta = 0$ for all δ.

As the series $\mu_1, \ldots, \mu_{26}$ are all different there is only one value ϑ of δ for which $\mu = \mu_\delta$ and so $l_\alpha = c_\vartheta u_{\alpha\vartheta}$ for all α. Now putting $l_\alpha = 1$ for all α we see that one member of the series $\mu_1, \ldots, \mu_{26}$ is 1, for (3) is certainly satisfied.

I shall prove that the remaining eigenvalues satisfy $|\mu| \leq 1$. We first prove that if $\mu \neq 1$ then $\sum p_\alpha l_\alpha = 0$. This follows by multiplying (3) on each side by $\mathcal{P}_\alpha$ and summing. Since

$$q_{\alpha\beta} = \frac{\mathcal{P}_{\alpha\beta}}{\mathcal{P}_\alpha} \text{ and } \sum_\alpha \mathcal{P}_{\alpha\beta} = \mathcal{P}_\beta,$$

we get

$$\sum_{\alpha\gamma} q_{\alpha\gamma} l_\gamma = \sum p_\gamma l_\gamma = \mu \sum p_\alpha l_\alpha,$$

which implies

$$\mu = 1 \text{ or } \sum p_\alpha l_\alpha = 0.$$

Next we show that each μ for which $|\mu| > 1$ is real and positive. Let l_α satisfy (3) with $|\mu| > 1$; then the eigenvalue for $\bar{l}_\alpha$ is $\overline{\mu}$ and so

$$\sum_\beta (\mathcal{Q}^r)_{\alpha\beta} \left(1 + \varepsilon \left(l_\beta + \bar{l}_\beta\right)\right) = 1 + 2\varepsilon \, \Re \, \mu^r l_\alpha.$$

If $\varepsilon > 0$ has been chosen so small that $\Re \varepsilon l_\beta > -1/2$ for all β then the L.H.S. is positive for the coefficients in the matrix are positive, whereas the R.H.S. is negative for suitably chosen μ, unless $l_\alpha = 0$. If now $\mu > 1$ we may take it that l_α is real for each α. As it must satisfy $\sum p_\alpha l_\alpha = 0$ it is negative for some α, but then

$$\sum_\beta (\mathcal{Q}^r)_{\alpha\beta} (1 + \varepsilon \, l_\beta) = 1 + \varepsilon \mu^r l_\alpha,$$

and if ε is chosen so that $1 + \varepsilon \, l_\beta > 0$ for all β the L.H.S is positive whereas the R.H.S is negative for sufficiently large r.

All the eigenvalues therefore satisfy $|\mu| \leq 1$ as the eigenvalues are all different in modulus this means that $|\mu| < 1$ except for one value of μ. Then as $r \to \infty, \mathcal{M}^r$ tends to a matrix which has only one element different from 0, and that a 1 on the diagonal, say in position $\sigma\sigma$.

Calling this matrix $\mathcal{X}_\sigma$ the series of matrices $\mathcal{Q}^r$ tends to the matrix $\mathcal{U}^{-1}\mathcal{X}_\sigma\mathcal{U}$. This matrix is the one and only one $\mathcal{Y}$ which satisfies $\mathcal{Y}\mathcal{Q} = \mathcal{Y}, \mathcal{Y}^2 = \mathcal{Y}, \mathcal{Y} \neq 0$ and is therefore the one whose $\alpha\beta$ coefficient is $\mathcal{P}_\beta$.

2.4.3. Another probability problem. There is another probability problem that arises in connection with simple transpositions. With a message of length L, and a key length of K what is the probability that the mth letter will be at the bottom of a column? Let D be the length of the short columns *i.e.* $D = [L/K]$, and let $E = L - DK$. Then if the mth letter is at the bottom of the wth column we must have

$$\frac{m}{D+1} \leq w \leq \frac{m}{D},$$

and there will be $(D+1)w - m$ short and $m - Dw$ long columns among these first w columns. There are[15]

$$\binom{w}{m - Dw}\binom{K - w}{E - m + Dw}$$

ways in which the short and long columns can be arranged consistently with this, and altogether $\binom{K}{E}$ ways in which the columns can be arranged, so that the probability of the m the letter being at the bottom of a column is

$$\sum_{(m/D+1) \leq w \leq (m/D)} \binom{w}{m - Dw}\binom{K - w}{E - m + Dw} \Big/ \binom{K}{E}.$$

There will normally be very few terms in the sum.

Let us take the case of the message of length 133 and consider the 45th letter, assuming the key length is between 10 and 20 (inclusive). ${L_O}^B$. $L = 133, m =$ 45.

[15] Turing is using Binomial Coefficient notation in this section;

$$\binom{n}{k} = C(n,k) = \frac{P(n,k)}{P(k,k)} = \frac{n!}{(n-k)!k!}$$

$$K = 10,\quad D = 13,\quad E = 3,\quad \frac{m}{D+1} = 3+,\quad \frac{m}{D} = 3+ \qquad \textit{no terms}$$

$$K = 11,\quad D = 12,\quad E = 1,\quad \frac{m}{D+1} = 3+,\quad \frac{m}{D} = 3+ \qquad \textit{no terms}$$

$$K = 12,\quad D = 11,\quad E = 1,\quad \frac{m}{D+1} = 3+,\quad \frac{m}{D} = 4+$$

only terms $w = 4$, $m - Dw = 1$ *probability is:* $\binom{4}{1}\binom{8}{0}\Big/\binom{12}{1} = \frac{4}{12}$

$$K = 13,\quad D = 10,\quad E = 3,\quad \frac{m}{D+1} = 4+,\quad \frac{m}{D} = 4+ \qquad \textit{no terms}$$

$$K = 14,\quad D = 9,\quad E = 7,\quad \frac{m}{D+1} = 4+,\quad \frac{m}{D} = 5$$

only terms $w = 5$, $m - Dw = 0$ *probability is:* $\binom{5}{0}\binom{9}{7}\Big/\binom{14}{7} = \frac{3}{286} = 0.0105,$

$$K = 15,\quad D = 8,\quad E = 13,\quad \frac{m}{D+1} = 5,\quad \frac{m}{D} = 5+$$

only terms $w = 5$, $m - Dw = 5$ *probability is:* $\binom{5}{5}\binom{10}{8}\Big/\binom{15}{13} = \frac{3}{7} = 0.428,$

$$K = 16,\quad D = 8,\quad E = 5,\quad \frac{m}{D+1} = 5+,\quad \frac{m}{D} = 5+$$

only terms $w = 5$, $m - Dw = 5$ *probability is:* $\binom{5}{5}\binom{11}{0}\Big/\binom{16}{5} = \frac{1}{4368} = 0.000229,$

$$K = 17,\quad D = 7,\quad E = 14,\quad \frac{m}{D+1} = 5+,\quad \frac{m}{D} = 6+$$

only terms $w = 6$, $m - Dw = 3$ *probability is:* $\binom{6}{3}\binom{11}{11}\Big/\binom{17}{14} = \frac{1}{34} = 0.0307$

($Editor - 1/34$ is 0.0294.)

$$K = 18,\quad D = 7,\quad E = 7,\quad \frac{m}{D+1} = 5+,\quad \frac{m}{D} = 6+$$

only terms $w = 6$, $m - Dw = 3$ *probability is:* $\binom{6}{3}\binom{12}{4}\Big/\binom{18}{7} = \frac{4950}{15912} = 0.311,$

$K = 19,\quad D = 7,\quad E = 0,$ *probability is:* $= 0$

$$K = 20,\quad D = 6,\quad E = 13,\quad \frac{m}{D+1} = 6+,\quad \frac{m}{D} = 7+$$

only terms $w = 7$, $m - Dw = 3$ *probability is:* $\binom{7}{3}\binom{13}{4}\Big/\binom{20}{7} = \frac{35 \times 143}{15504} = 0.323.$

∞

HW 25/37

CCR
NO. 150$^{(1)}$

CONFIDENTIAL

THE APPLICATIONS OF PROBABILITY TO CRYPTOGRAPHY

by A.M. Turing

THE APPLICATIONS OF PROBABILITY TO CRYPTOGRAPHY

The theory of probability may be used in cryptog aphy with most effect when the type of cipher used is already fully understood, and it only remains to find the actual keys. It is of rather less value when one is tryin to diagnose the type of cipher, but if definite rival theories about the type of cipher are su ested it may be used to decide between them.

Meaning of probability and odds.

I shal not attempt to give a systematic a count of the theory of probability, but it may be worth while to define shortly 'probability' and 'odds'. The probability of an event on certain evidence is the proportion of cases in which that event may be expected to happen given that evidence. For instance if it is known that 20% of men live to the age of 70, then knowing of Hitler o ly 'Hitl r is a man' we can say that the probability of Hitler living to the a e of 70 is 0.2 . Suppose however that we know that 'Hitler is now of age 52' the probability will be quite different, say ~~20%~~ 0.5, because 50% of men~~xlivextoxthexage~~ of 52 live to 70.

The 'odds' of an event happening is the ratio $p/1-p$ ~~xhe~~ where p is th e probability of it happening. This terminology is connected with the common phraseology 'odds of 5:2 on' meaning in our terminology that the odds are 5/2.

Probabilities based on part of the evidence

When the whole evidence about some event is taken into account it may be extremely difficult to estimate the probability of the event, even ~~xxtxx~~ very approximately, and it may be better to form an estimate based on a part of the evidence, so that the probability may be more easily calculated. This happens in cryptography in a very obvious way. The whole evidence when we are trying to solve a cipher is the complete traffic , and the events in question are the different possible keys, and functions of the keys. Unless the traffic is very small indeed the theoretical answer to the problem 'what are the probabilities of the various keys ?' will be of the form ' The key ... has a probability differing almost imperceptibly from 1 (certainty) and the other keys are virtually impossible'. But ~~xxxxxxxx xxxxxxxxxxxxxxxx~~ a direct attempt to determine these probabilities would obviously not be a practical method.

A priori probabilities

The evidence concerning the possibility of an event occurring usually divides into a part about which statistics are available, or some mathematical method can be applied, and a less definite part about which one can only use one's judgment. Suppose for example that a new kind of traffic has turned up and that only three messages are available. Each message has the letter V in the 17th place and G in the 18th place. We want to know the probability that it is a general rule that we should find V and G in these places. We first have to decide how probable it is that a cipher would have such a rule, and as regards this one can probably only guess, and my guess would be about 1/5,000,000 .

This judgment is not entirely ~~xaxprimrixy~~ a guess; some rather inaccurate mathematical reasoning has gone into it, something like this;-

The chance of there being a rule that two consecutive letters somewhere after the 10th should have certain fixed values seems to be about 1/500 (this is a complete guess). The chance ofthe letters being the 17th and 18th is about 1/15 (another guess , but not quite so much in the air). The probability of the letters being V and G is 1/676 (hardly a guess at all, but expressing a judgment that there is no special virtue in the bigramme VG) . Hence the chance is 1/ 500x15x676 or about 1/5,000,000 . This is however all so vague, that it is more usual to make the judgment '1/5,000,000' without explanation.

The question as to what is the chance of having a rule of this kind might of course be solved by statistics of some kind, but there is no point in having this very accurate, and of course the experience of the cryptographer itself forms a kind of statistics.

The remainder of the problem is then solved quite mathematically. Let us consider a large number of ciphers 'chosen at random' , N of them say. Of these N/5,000,000 of them will have the rule in question, and the remainder not. Now if we had three messages of each of the ciphers before us, we should find that ~~with~~ for each of the ciphers with the rule, the three messages have VG in the required place, but of the remaining 4,999,999 N/5,000,000 only a proportion $1/676^3$ will have them. Rejecting the ciphers which have not the required characteristic we are left with

N/5,000,000 cases where the rule holds, and 4,999,999 N/5,000,000 x 676^3 cases where it does not. This selection of ciphers is a random selection of ones which have ll the known characteristics of the one in question, and therefore the odds in favour of the rule holding are N/5,000,000 : 4,999,999N/5,000,000 x 676^3 i.e. 676^3 : 4,999,999 or about 60;1 on.

It should be noticed that the whole argument is to some extent fallacious, as it is assumed that there are only two possibilities, viz. that either VG must always occur in that position, or else that the letters in the 17th and 18th positions are wholly random. There are however many other possibilities worth consideration, e.g.

On the day in question we have VG in the position in question. On another day we have some other fixed pair of letters. Or

In the position 17,18 we have to have one of the four combinations VG, RH, OM, IL and by chance Vg has been chosen for all the three messages we have had. Or

The cipher is a simple substitution and VG is the substitute of some common bigramme, say TH.

The possibilities are of course endless, and it is therefore always necessary to bear in mind the possibility of there being other theories not yet suggested.

The a priori probability sometimes has to be estimated as above by some sort of guesswork,but often the situation is more satisfactory. Suppose for example that we know that a certain cipher is a simple substitution, the keys having no specially noticeable properties. Suppose also that we have 50 letters of such a message including five occurrences of P. We want to know how probable it is that P is the substitute of E. As before we have to answer two questions.How likely is it that P would be the substitute of E neglecting the evidence of the five Es occurring in the message. Secondly 'How likely are we to get 5 Ps (a) if P is not the substitute of E (b) if P is the substitute of E. I will not attempt to answer the second question for the present. The answer to the first is simply that the probability of any letter being the substitute of E is independent of what the letter is, and is therefore always 1/26, in particular it is 1/26 for the letter P. The only guesswork here is the judgment that the keys are chosen at random.

The Factor Principle.

Nearly all applications of probability to cryptography depend on the 'factor principle' (or Bayes' theorem). This principle may first be illustrated by a simple example. Suppose that one man in five dies of heart failure, and that of men who die of heart failure two in three die in their beds, but of men who die from other causes only one in four die in their beds. (My facts are no doubt hopelessly inaccurate). Now suppose we know that a certain man died in his bed. What is the probability that he died of heart failure? Of all men numbering N say, we find that

Nx	(1/5)x(2/3)	die in their beds of heart failure
Nx	(1/5)x(1/3)	... elsewhere ....
Nx	(4/5)x(1/4)	die in their beds from other causes
NX	(4/5)x(3/4)	... elsewhere ....

Now as our man died in his bed we do not need to consider the cases of men who did not die in their beds, and these consist of Nx (1/5)x(2/3) cases of heart failure and Nx (4/5)x (1/4) from other causes, and therefore the odds are 1x (2/3): 4x (1/4) in favour of heart failure. If this had been done algebraically the result would have been

A posteriori ~~probability~~ odds of the theory

= A priori ~~probability~~ odds of the theory x

$$\text{x}\ \frac{\text{Probability of the data being fulfilled if the theory is true}}{\text{Probability of the data being fulfilled if the theory is false}}$$

In this ~~theory~~ 'theory' is that the man died of heart failure, and the 'data' is that he died in his bed. The general formula above will be described as the 'factor principle', the ratio

$$\frac{\text{Probability of the data if theory true}}{\text{Probability of the data if theory false}}$$

is ~~x~~called the factor for the theory ~~if the~~ on account of the data.

Decibanage.

Usually when ~~we~~ we are estimating the probability of a theory there will be several independent pieces of evidence e.g. following our last example, where we want to know whether a certain man died of heart failure or not, we may know

a) He died in his bed

b) His father died of heart failure

c) His bedroom was on the ground floor

and also have statistics telling us

2/3 of men who die of heart failure die in their beds

2/5 . . . . have fathers who died of heart failure

1/2 . . . . have bedrooms on the ground floor

1/4 of men who died from other causes die in their beds

1/6 . . . . have fathers who died of heart failure

1/20 of men who die of other causes have their bedrooms on the ground floor

~~If we~~ Let us suppose that the three pieces of evidence are independent of one another ~~that~~ if we know that he died of heart failure, and also if we know he did not die of heart failure. That is to say we suppose ~~that~~ for instance that knowing ~~he died of heart failure, and that~~ that he slept on the ground floor does not make it any more likely that he died in his bed if we knew all along that he died of heart failure. When we make these assumptions the probability of a man who died of heart failure satisfying all three conditions is obtained simply by multiplication, and is (2/3)x(2/5)x(1/2), and likewise for those who died from other causes the probability is (1/4)x(1/6)x(1/20), and the factor in favour of the heart failure theory is

$$\frac{(2/3)x(2/5)x(1/2)}{(1/4)x(1/6)x(1/20)}$$

We may regard this as the product of the three factors (2/3)/(1/4) and (2/5)/(1/6) and (1/2)/(1/20) arising from the three independent pieces of evidence. Products like this arise very frequently, and sometimes one will get products involving thousands of factors, and large groups of these factors may be equal. We naturally therefore work in terms of the logarithms of the factors. The logarithm of the factor, taken to the base $10^{1/10}$ is called the 'decibanage in favour of the theory'. ~~The~~ A 'deciban' is a unit of evidence; a piece of evidence is worth a deciban if it increases the odds of the theory in the ratio $10^{1/10}$: 1 . The deciban is used as a more convenient unit than the 'ban'. The terminology was introduced in honour of the famous town of Banbury.

Using this terminology we might say that the fact that our man died in bed scores 4.3 decibans in favour of the heart failure theory ($10\log(8/3) = 4.3$). We score a further 3.8 decibans for his father dying of heart failure, and 10 for his having his bedroom on the ground floor, totalling 18.1 decibans. We then bring in the a priori odds 1/4 or ~~xx~~ $10^{-6/10}$ and the result is that the odds are $10^{12.1/10}$, or as we may say '12.1 deci~~x~~bans up on evens'. This means about 16;1 on.

Chapter II. Straightforward cryptographic problems.

Vigenère.

The factor principle can be applied to the solution of a Vigenère problem with great effect. I will assume here that the period of the cipher has already been determined. Probability theory may be applied to this part of the problem also, but that is not so elementary. Suppose our cipher, written out in its correct period is

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

Fig 1. Vigenère problem.

(It is only by chance that it makes a rectangular array). Let us try to find the key for the first column, and for the moment let us only take into account the evidence afforded by the first letter D. Let us first consider the key B. The factor principle tells us

Odds in favour of key B = A priori odds in favour of key B x

$$\times \frac{\text{Probability of getting D in cipher if key is B}}{\text{Probability of getting D in cipher if key is not B}}$$

Now the a priori odds in favour of key B may be taken as 1/25. The probability of getting D in the cipher with the key B is just the probability of getting C in the clear which (using the count on 1000 letters in Fig 2) is 0.021 . If however the key is not B we can have any letter other than C in the clear, and the probability is (1- 0.021)/25 . Using the evidence of the D then the odds in favour of the key B are $\frac{1}{25} \cdot \frac{25 \times 0.021}{1 - 0.021}$

We then consider the effect of the next letter in the column R which gives a further factor of 25x 0.064/(1-0.064). We are here assuming that the evidence of the R is independent of the evidence of the D. This is not quite correct, but is a useful approximation: a more accurate method of calculation will be given later. Let us write p_α for the frequency of the letter α in plain language. Then our final estimate for the odds in favour of key B is

$$\frac{1}{25}\prod_i \frac{25\, p_{\alpha_i - 1}}{1 - p_{\alpha_i - 1}}$$

where $\alpha_1\ \alpha_2 \ldots$ is the series of letters in the 1st column, and we use letters and numbers interchangeably, A meaning 1, B meaning 2,..., Z meaning 26 or 0. More generally for key β the odds are

$$\frac{1}{25}\prod_i \frac{25\, p_{\alpha_i - \beta + 1}}{1 - p_{\alpha_i - \beta + 1}}$$

The value of this can be calculated by having a table of the decibanages corresponding to the factors $25 p_\alpha / 1 - p_\alpha$. One then decodes the column with the various possible keys, looks up the decibanages, and adds them up.

The most convenient form for doing this is a table of values of $20 \log_{10} 25 p_\alpha / 1 - p_\alpha$, taken to the nearest integer (or as we may say, the values of the score in 'half decibans'). One may also have columns showing multiples of these, and the table made of double height (Fig 3). For the first column with key B the decoded column is CQWS..OAV and we score -5 for C, -26 for Q,

A	84
B	23
C	21
D	46
E	116
F	20
G	25
H	49
I	76
J	2
K	5
L	38
M	34
N	66
O	66
P	15
Q	2
R	64
S	73
T	81
U	19
V	11
W	21
X	16
Y	24
Z	3

The value for X has been taken more or less at random ~~to allow for~~ as a compromise between real language & telegraphese. Also I added 1 to each entry (see p).

Fig 2. Count on 1000 letters English text.

31	26	20	13	7		A
-23	-18	-14	-9	-5		B
-26	-21	-16	-10	-5		C
7	6	4	3	1		D
48	38	29	19	10		E
-28	-22	-17	-11	-6		F
-19	-15	-11	-8	-4		G
10	8	6	4	2		H
29	23	17	12	6		I
-131	-103	-77	-52	-26		J
-99	-79	-59	-40	-20		K
-2	-2	-1	-1	0		L
-6	-5	-4	-2	-1		M
23	18	14	9	5		N
23	18	14	9	5		O
-41	-33	-25	-16	-8		P
-131	-103	-77	-52	-26		Q
22	18	13	9	4		R
28	22	17	11	6		S
32	26	19	13	6		T
-31	-25	-19	-12	-6		U
-54	-43	-32	-22	-10		V
-26	-21	-16	-10	-5		W
-38	-30	-23	-15	-8		X
-20	-16	-12	-8	-4		Y
-111	-89	-67	-44	-22		Z
+31	26	20	13	7	→	A
-23	-18	-14	-9	-5		B
-26	-21	-16	-10	-5		C
7	6	4	3	1		D
48	38	29	19	10		E
-28	-22	-17	-11	-6		F
-19	-15	-11	-8	-4		G
10	8	6	4	2		H
29	23	17	12	6		I
-131	-103	-77	-52	-26		J
-99	-79	-59	-40	-20		K
-2	-2	-1	-1	0		L
-6	-5	-4	-2	-1		M
23	18	14	9	5		N
23	18	14	9	5		O
-41	-33	-25	-16	-8		P
-131	-103	-77	-52	-26		Q
22	18	13	9	4		R
28	22	17	11	6		S
32	26	19	13	6		T
-31	-25	-19	-12	-6		U
-54	-43	-32	-22	-10		V
-26	-21	-16	-10	-5		W
-38	-30	-23	-15	-8		X

Fig 3. Table for scoring a Vigenère. In units of half a deciban

September 6th B

3	2.0 .7 [illegible] 1 0.3 0.3	3 7 0.2 .4 2 0 0.3	2 0 0 [illegible] .3 [illegible] 0	0 3 [illegible] 4 0.3 [illegible]	1.4 0.3 2.0 1.4 .7	O [illegible] 7
3	3.7 [illegible] 7 4 3.1	5 1.4 3.7 .3 5.4	3 2 4 [illegible]	[illegible]	3.7 4 [illegible]	J. [illegible]
3	1.1 1.0 1.6 5.2 1 1	2.5 [illegible] 5.3	3.3 5.0 4.3 5.0 4.3	.7 1.1 5.0 5.0 [illegible]	4.3 5.0 .8 1.4 6.7	K 1[illegible] 4
B	1 4 0 2 7 5.7 2.6	.8 2.6 .2 5 2 2.2	1.2 1.2 (4.1) .1 .2	1.0 3 6 3 6 4 2 3	1 .1 7 3 3 .2	F 2[illegible] 4
B	6 0 24.9 2.6 6.0	3.6 4 7 2 2 2.6 (1.4)	1 8 3.8 2 2.1 4 1	5.4 3.7 .1 3 8 7 .1	3 7 6 0 6 7 3 3.2	R 264 - 179 11
B	3.9 3.7 4 0 (2.7) 4.1	2.4 3.7 3 7 5 6 9	4.1 2.5 1 2 3 1 3	.8 4.6 [illegible]	1.6 1.4 1.8 1 [illegible]	V [illegible] - 147 6
3	1 2 3.0 5 2 3.5	1 [illegible] 2 2 1 5 [illegible]	.6 1 [illegible] 3.9 (2.2) .5	5 2 2 2 1 8 3.5	3 5 1 9 2.9 9 9 1 5	M 120 - 95 4
3	3.7 2.0 2.0 2.0	3 5 4 3 7 3 7 2 0	3 7 5 4 3 7 1 .3	[illegible] 1 1 5 4 3 2.0	5 4 3 7 [illegible]	T [illegible] - 63 7

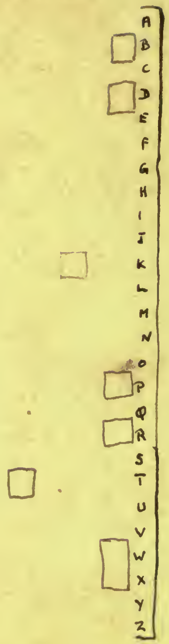


Fig 4. Apparatus for scoring in Vigenère. Pencil marks arranged for 1st column of Fig 1.

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

Scores for various keys

A	-2 -27	X	• X	X
B	• -17	X	X	• -2
C	X	• 3	16	•• X
D	• X	9	9	X
E	-6	X	X	X
F	X	X	X	X
G	X	X	X	-3
H	X	•• 1	-3	• X
I	X	X	17	X
J	X	X	X	• 13
K	X	• X	•• X	-15
L	2	X	X	X
M	X	X	X	X
N	X	X	X	X
O	X	• 28	X	• X
P	• 43	X	• X	X
Q	X	X	• X	22
R	• X	X	X	X
S	X	X	-6	X
T	⁚ 8	X	-15	X
U	X	• X	• X	• 22
V	X	X	• X	X
W	• X	• 16	•	X
X	• X	X	-15	• X
Y	X	X	-15	• X
Z	X	• -13	• X	X
	P	O	IC	QU

Best keys

Possible decodes

O	W	IO	RN	G
C	O	NT	HD	
I	T	HN	EA	S
E	I	MW	TT	O
E	T	OU	MI	M
A	L	CI	YU	L
M	A	CI	LH	I
H	*	SY	MI	S
E	G	[illegible]	[illegible]	T

EMK Scoring and solving a Vigenère.

-5 for W, 17 for the three letters S, 5 for O, 2 for A and -10 for V, totalling -17. These calculations can be done very quickly by the use of the transparent gadget Fig 4, in which squares are ringed in pencil to show the number of letters occurring in the column. The gadget may be placed over Fig 3 in various positions corresponding to the various possible keys. The score is obtained by adding up the numbers showing through the various squares. In Fig 5 the ~~possible~~ alphabet has been written in a vertical column below the cipher text of Fig 1, each letter representing a possible key. The score for each key has been written opposite the key, and under the relevent column. An X denotes a bad score, not worth adding up. Usually there will be -15 or worse. It will be seen that for the first column P, having a score of 43 is extremely likely to be right, especially as there is no other score better than 8. If we neglect this latter fact the odds for the key are (1/25) $10^{2.15}$ i.e. about 5:1 on. The effect of decoding this column with key P has been shown underneath. For the second column the best key is C, but is by no means so convincing as the first column. The decode for this column is also shown, and provides very satisfactory combinations with the first column, confirming both the keys.(This confirmation could also be based on probability theory, given a table of bigramme frequencies). In the third column I and C are best although D would be very possible, and in the fourth column Q and U are best. Writing down the possible decodes we see that the first line must read DALNG and this makes the other lines read CONDI, ITHAS, ELEFO, ETOT , ALCOL, AM ON , AISIS, LGALT. By filling in the word 'conditions' the whole can now be decoded.

A more accurate argument would run as follows. For the first column, instead of setting up as rival theories the two possibilities that B is the key and that B is not the key we can set up 26 rival theories that the key is A or B or ... or Z, and we may apply the factor principle in the form:-

$$\frac{\text{A posteriori probability of key A}}{\text{A priori probability of key A} \times \text{Probability of getting the given column with key A}}$$

$$= \frac{\text{A posteriori probability of key B}}{\text{A priori probability of key B} \times \text{Probability of getting the given column with key B}}$$

$$= \text{etc.}$$

The argument to justify this form of factor principle is really the same as for the original form. Let q_β be the a priori probability of key β. Then out of N cases we have Nq_β cases of key β. Let $P(\beta, C)$ be the probability of getting the ~~key~~ column C with key β, then when we have rejected the cases where we get columns other than C we find that there are $Nq_\beta P(\beta, C)$ cases of key β left, i.e. the a posteriori probability of key β is $K q_\beta P(\beta, C)$ where K is independent of β.

We have therefore to calculate the probability of getting the column C with key β and this is simply $\prod_i P_{\alpha_i - \beta + 1}$ i.e. the product of the frequencies of the decode letters which we get if the key is β.

Since the a priori probabilities of the keys are all equal we may say that the a posteriori probabilities are in the ratio $\prod_i p_{\alpha_i - \beta + 1}$, i.e. in the ratio $\prod_i (26\, p_{\alpha_i - \beta + 1})$ which is more convenient for calculation. The final value for the probability is then

$$\prod_i (26\, p_{\alpha_i - \beta + 1}) \Big/ \sum_\beta \prod_i (26\, p_{\alpha_i - \beta + 1})$$

The calculation of the products $\prod_i (26\, p_{\alpha_i - \beta + 1})$ may be done by the method recommended before for $\prod_i 25 p_{\alpha_i - \beta + 1} / 1 - p_{\alpha_i - \beta + 1}$ (The table in Fig 3 was in fact made up for $\prod_i (26 p_{\alpha_i - \beta + 1})$. The differences between the two tables would of course be rather slight). The new result is more accurate than the old because of the independence assumption in the original result.

If we only want to know the ratios of the probabilities of the various keys there is no need to calculate the denominator $\sum_\beta \prod_i (26\, p_{\alpha_i - \beta + 1})$. This denominator has however another importance: it gives us some evidence about our other assumptions, such as that the cipher is Vigenere, and that the period is 10. This aspect will be dealt with later (p.   ).

A letter subtractor problem

A substitution with period 91 x 95 x 99 is obtained by superimposing three substitutions of periods 91, 95, and 99, each substitution being a Vigenere composed of slides of 0,1,2,3,4,5,6,7,8, or 9. The three substitutions are known in detail, but we do not know for any given message at what point in the complete substitution to begin. For many messages however we can provide a more or less probable crib. How can we test the probability of a crib before attempting to solve it? It may be assumed that approximately equal numbers of slides 0,1,... 9 occur in each substitution.

The principle of the calculation is that owing to the way in which the substitution is built up, not all slides are equally frequent, e.g. a slide of 25 can only be the sum of slides of 9,8 and 8 or of 9,9 and 7 whilst a slide of 15 can be any of the following

9,6,0
9,5,1
9,4,2
9,3,3

8,7,0
8,6,1
8,5,2
8,4,3

7,7,1
7,6,2
7,5,3
7,4,4

6,6,3
6,5,4

A crib will therefore, other things being equal, be more likely if it requires a slide of 15 than if it requires a slide of 25. The problem is to make the best use of this principle, by determining the probability of the crib with reasonable accuracy, but without spending long over it.

We have to find out the probability of getting a given slide. To do this we can apply several methods.

(a) We can produce a long stretch of key by addition and take a count of the resulting slides. This is obviously a very general method, and requires no special mathematical technique.It may be rather laborious, but by interpreting a small count with common sense one can probably get quite good results.

(b) There are 1000 possible combinations of slides all equally likely, viz 000,001,...,999 . We can add up the digits in these and take the remainder on division by 26, and then count the number of combinations giving each of the possible remainders.

(c)We can make use of a trick which might appear to be rather special, but is really applicable to a multitude of problems. Consider the expression

$$f(x) = (1+x+x^2+\dots+x^9)^3$$

For each possible way of expressing a number n as the sum of three numbers 0,...,9, say $n = m_1 + m_2 + m_3$ there is a term $x^{m_1} x^{m_2} x^{m_3}$ in $f(x)$, x^{m_1} coming out of the first factor, x^{m_2} out of the second, and x^{m_3} out of the third. Hence the number of ways of expressing n in the form $n = m_1 + m_2 + m_3$ is the coefficient of x^n in $f(x)$ i.e. in

$$\frac{(1-x^{10})^3}{(1-x)^3}$$

or in

$$(1-3x^{10}+3x^{20}-x^{30})(1-x)^{-3}$$

Expanding $(1-x)^{-3}$ by the binomial theorem

$$(1-x)^{-3} = 1 + 3x + 6x^2 + 10x^3 + 15x^4 + 21x^5 + 28x^6 + 36x^7 + 45x^8 + 55x^9 + 66x^{10} + 78x^{11} + 91x^{12} + 105x^{13} + 120x^{14} + 136x^{15} + 153x^{16} + 171x^{17} + 190x^{18} + 210x^{19} + 231x^{20} + 253x^{21} + 276x^{22} + 300x^{23} + 325x^{24} + 351x^{25} + 378x^{26} + 406^{27} + 435x^{28} + \ldots$$

Now multiply by $1 - 3x^{10} + 3x^{20} - x^{30}$ and we get

$$f_0(x) = 1 + 3x + 6x^2 + 10x^3 + 15x^4 + 21x^5 + 28x^6 + 36x^7 + 45x^8 + 55x^9 + 63x^{10} + 69x^{11} + 73x^{12} + 75x^{13} + 75x^{14} + 73x^{15} + 69x^{16} + 63x^{17} + 55x^{18} + 45x^{19} + 36x^{20} + 28x^{21} + 21x^{22} + 15x^{23} + 10x^{24} + 6x^{25} + 3x^{26} + x^{27}$$

This means to say that the chances of getting totals of 0,1,2,... are in the ratio 1, 3, 6, 10,... The chances of getting remainders of 0,1,2,... on division by 26 are in the ratio 4, 4, 6, 10, 15, ... To get true probabilities these must be divided by their total which is conveniently 1000.

(d) There are two other methods, both connected with the last method but ~~xeqxixi~~ not relying so much on the special ~~p~~ features of the problem. They will be discussed later.

Suppose then that the probabilities have been calculated by one method or the other(as in fact ~~waxxthexxxxxxfor~~ we have done under (c)). We can then estimate the values of cribs. Let us suppose that a possible crib for a message beginning MVHWUSXOWBVMMK was ~~AMBAXXXXXXXXX~~ AMBASSADOR so that the slides were 12, 9, 6, 22, 2, 0, 23, 11, 14, The slide of 12 gives us some slight evidence in favour of the crib being right for slides of 12 occur with frequency 0.073 with right cribs, whilst with wrong cribs they occur with frequency only 1/26. The factor in favour of the crib is therefore 26x0.073 or about 1.9 . A similar calculation may be made for each of the slides, but of course the work may be greatly speeded up by having the values of the factors 26 C_s/1000 in half decibans tabulated: here C_s is the coefficient of x^s in the above polynomial $f(x)$. The table is given below (Fig 6)

1	0	-20
2	25	-16
3	24	-12
4	23	-8
5	22	-6
6	21	-3
7	20	-1
8	19	1
9	18	3
10	17	4
11	16	5
12	15	6
13	14	6

Fig 6. Scores in half decibans of the various slides.

Evaluating this crib by means of this table we ~~xxx~~ score

$$6 + 3 - 3 - 6 - 16 - 20 - 8 + 5 + 6 \quad (= -33)$$

i.e. the crib is worse by a factor of $10^{-33/20}$ than it was before e.g. if the a priori odds of the crib were 2:1 against it becomes 98:1 against. This crib was in fact made up at random i.e. the letters of the cipher text were chosen at random. Now let us take one made up correctly, i.e. really enciphered by the method in question, but with a random chosen key.

```
      N Y X L N X I   H H
      A M B A S S A D O R
     13  22  21   8  19
        12 11   5  13  16    (slides)
```

This scores 15 so that if it were originally 2:1 against it now becomes nearly 3:1 on.

Having decided on a crib the natural way to test it is to have a catalogue of the positions in which a given series of slides is obtained if the 91 period component is omitted,. We make 91 different hypotheses as to this third component, draw an inference as to what is the part of the slide arising from the components of periods 95 and 9? combined. This we look up in the catalogue . This process is fairly lengthy, and as the scoring of the crib takes only a minute it is certainly worth doing.

Theory of repeats

Suppose we have a cipher in which there are several very long series of substitutions which can be used for enciphering a message, but that one may sometimes get two messages enciphered with the same series of substitutions (or more truly, the series of substitutions for one message being those for another with some at the beginning omitted). In such a case let us say that the messages 'fit', or that they fit at such and such a distance, the distance being the number of substitutions which have to be omitted from the one series to obtain the other series. One will frequently want to know whether two messages fit or not, and we may find some evidence about this by examining the repeats between them. By the repeats between them I mean this. One writes out the cipher texts of the two messages with the letters which are thought to have been enciphered with the same substitution under one another. One then writes under these messages a series of letters o and x, an o being written where the cipher texts differ and an x where they agree. Thexx series of letters o and x will begin where the second message begins and end where the first to end ends. ~~To complete the information about the repetition figure~~ This series of letters o and x may be called the repetition figure. It may be completed by adding at the end an indication of how many letters there are which do not overlap, and which message they belong to. As an example

GFRLIKQGVBNILAFIXKMOROGBYSKYXDAZCHMUMRKBZIDLDOHCNVTIPRSD

VLOVDYQCLJSOPYGBNBKYXDAZNBFIOPTFCXDOD

^{8}xooooooooooxooxxooxxxxxxooooooooooxox^{11}

On the whole one expects that a fit is more likely to be right the more letters x there are in the repetition figure, and that long series of letters x are especially desirable. This is because it would not be very unusual for two fairly common words to lie directly under one another when the clear texts are written out, thus

```
THEMAINCONVOYWILLARRIVE . . .
     ALLCONVOYSMUSTREPORT. . .
     XOOXXXXXXOOOOOXOOOO . . .
```

If the corresponding cipher texts really fit, i.e. if the letters in the same column are enciphered with the same substitution, then the condition for an x in the repetition figure of the cipher texts is that there be an x in the repetition figure of the corresponding clear text. Now series of several consecutive letters x can occur quite easily as above by two identical words coming under one another, or by such combinations as

```
ITISEASIERTOTEACHTHANALGEBRA . . .
   THERAINWASSUCHTHATHECOULD . . .
   OOOOOOOOOOOXXXXXOOOOOOOO . . .
```

if the messages really fit, but if not they can only occur by complete coincidence. One therefore tends to believe that there is a fit when one gets such series of letters x . As regards single cases of x the value of them is not so clear, but one can see that if p_α is the frequency of letters α in plain language then the frequency of letters x as a whole in comparisons of plain language with plain language is $\sum_\alpha p_\alpha^2$, whilst for wrong fits of cipher text it is 1/26 which is necessarily less. Given a sufficiently long repetition figure one should therefore be able to tell whether it is a fit or not simply by counting the letters x and o.

So much is well known. The real point of this section is to show how these ideas can be developed into an accurate method of estimating the probabilities of fits.

Simple form of theory. The complete theory takes account of the various possible lengths of repeat. As this theory is somewhat complicated it will be as well to give first two simplified forms of the theory. In both cases the simplification arises by neglecting a part of the evidence. In the first simplified form of theory we neglect all evidence except the number of letters x and the number of letters o. In the other simplified form the evidence is the number of series of (say) four consecutive letters x in a repetition figure.

When our evidence is just the number of times x occurs in the repetition figure (n let us say), and the length of the repetition figure (N say), then the factor in favour of the fit is

$$\frac{\text{Probability of a right ~~fit~~ repetition figure of length N having n occurrences of x}}{\text{Probability of a wrong repetition figure of length N having n occurrences of x}}$$

As an approximation we may assume that the numerator of this expression has the same value as if the right repetition figures were produced letter by letter by independent random choices, with a certain fixed probability of getting an x at each stage. This probability will have to be $\beta = \sum_{\alpha} p_{\alpha}^{2}$. The numerator is then

(Number of repetition patterns with length N and n occurrences of~~x~~ x)

times (Probability of getting a given such repetition pattern by th e ~~random~~ process just mentioned)

which we may write as R(N?n) Q(N,n). Now let us denote by y_i the i th symbol of the given repetition pattern, and put $\tau_x = \beta$ and $\tau_o = 1-\beta$. Then $\varphi(N,n)$, the probability of getting the repetition pattern is $\prod_{i=1}^{N} \tau_{y_i}$ which simplifies to $\beta^n(1-\beta)^{N-n}$. We may do a similar calculation for the denominator, but here we must take $\beta = \frac{1}{26}$ since all letters occur equally frequently in the cipher. The denominator is then $R(N,n)\left(\frac{1}{26}\right)^n\left(\frac{25}{26}\right)^{N-n}$. In dividing to find the factor for the fit $R(N,n)$ cancels out, leaving $(26\beta)^n\left(\frac{26}{25}(1-\beta)\right)^{N-n}$. In other words we score a factor of 26β for an x and a factor of $\frac{26}{25}(1-\beta)$ for an o. More convenient is to regard it as ~~xxfxxtxxxofxx fxxxxxxx~~ $10\log_{10} 25\beta/1-\beta$ decibans for an x and $10\log_{10}\frac{26}{25}(1-\beta)$ decibans per unit length of repetition figure ('per unit overlap').

An alternative argument, leading to the same result, runs as follows. Having decided to neglect all evidence except the overlap and the number of repeats we pretend that nothing else matters, i.e. that the form of the figure is irrelevant. In this case we can regard each letter of the repetition figure as independent evidence about the fit. If we get an x the factor for the fit is

$$\frac{\text{Probability of getting an x if the fit is right}}{\text{Probability of getting an x if the fit is wrong}}$$

i.e. $\dfrac{\beta}{1/26}$

Similarly the factor for an o is $\dfrac{1-\beta}{25/26}$ ~~Thexfxxxx~~

In either form of argument it is unnecessary to calculate the number F(N,n). In this particular case there is no particular difficulty about it: it is the binomial coefficient In some similar problems this cancelling out is a great boon, as we might not be able to find any simple form for the factor which cancels. The cancelling out is a normal feature of this kind of problem, and it seems quite natural that it should happen when we think of the second form of argument in which we think of the evidence as consisting of a number of independent parts.

The device of assuming, as we have done here, that the evidence which is not available is irrelevant can often be used and usually leads to good results. It is of course not supposed that the evidence really is irrelevant, but only that the ~~xolxxixx fxxtxr~~ error resulting from this assumption when used in this kind of way is likely to be small.

In the second ~~fxxx~~ simplified form of theory we take as our evidence that a particular part of the repetition figure is Oxxxxo (say, or alternatively oxxxyxo say). The factor is then

$$\frac{\text{Frequency of oxxxxo in right repetition figures}}{\text{Frequency of oxxxxo in wrong repetition figures}}$$

The denominator is $\left(\frac{1}{26}\right)^4\left(\frac{25}{26}\right)^2$ and the numerator can be estimated by taking a sample of language hexagrams and counting the number of pairs that have the repetition figure oxxxxo. The expectation of the number of such pairs is the sum for all pairs of the ~~expex~~ probabilities of those pairs ~~bxinx~~ having the desired repetition figure i.e. is the number of such pairs (viz N(N-1)/2 where N is the size of the sample) multiplied by the frequency of oxxxo repetition figures. This frequency may therefore be obtained by division if we compute the expected ~~xxkxxx~~ number of these repetition figures ~~xixh~~ to the actual number.

General form of theory. It is not of course possible to have statistics of every conceivable repetition figure. We must make some assumption to reduce the variety that need be considered. The following assumption is theoretically very convenient, and also appears to be a very good approximation.

The probabilities of repeats at two points known to be separated by a point where there is ~~xx~~ known to be no repeat ~~xxxx~~ are independent.

We may also assume that the probability of a repeat is independent of anything but the repetition figure in its neighbourhood. (We may however as a refinement produce different statistics for different types of messages, and different ~~parts~~ positions in a message). We can therefore think of a repetition figure as being produced by selecting the symbols of the figure consecutively, ~~xxchxxhxxxx~~ the probability of getting an x at each stage being determined by the repetition figure from the point in question back as far as the last o. Sometimes this will take us back as far as the beginning of the message, and will include the number telling us how many more letters there are which do not repeat at all. We need in practice only distinguish two cases, where this number is ~~x~~ 0 and when it is more. We may also neglect the question as to which message comes first. We therefore have to distinguish the following cases

o	a_0	some	b_0	none	c_0
ox	a_1	some x	b_1	none x	c_1
oxx	a_2	some xx	b_2	none xx	c_2
oxxx	a_3	some xxx	b_3	none xxx	c_3
. . .		. . .		. . .	

The entries a_0, a_1, b_0 etc. opposite the repetition figures are the notations we are adopting for the probability of getting another x following such a figure. Strictly speaking we should also bring in a notation for the probability of the message coming to an end after any given repetition figure. As the repeats at the end of a comparison do not appear to behave very differently from those in the main part of the message I shall neglect this complication by assuming that the probability of getting an o added to the probability of getting an x is 1, and that afterwards one cuts off the end of the series arbitrarily.

Let us calculate the factor for the repeat figure

$$\begin{array}{cccccc|c|c|cc|cccc|c|cc|c}
\text{none} & x & x & x & x & o & o & o & x & o & x & x & x & o & o & x & x & \text{some} \\
 & c_0 & c_1 & c_2 & c_3 & 1-c_4 & 1-a_0 & 1-a_0 & a_0 & 1-a_1 & a_0 & a_1 & a_2 & 1-a_3 & 1-a_0 & a_0 & a_1 & \\
 & \frac{1}{26} & \frac{1}{26} & \frac{1}{26} & \frac{1}{26} & \frac{25}{26} & \frac{25}{26} & \frac{25}{26} & \frac{1}{26} & \frac{25}{26} & \frac{1}{26} & \frac{1}{26} & \frac{1}{26} & \frac{25}{26} & \frac{25}{20} & \frac{1}{26} & \frac{1}{26} &
\end{array}$$

Underneath each symbol has been written the probability that one would get that symbol, knowing the ones which precede, both for the case of a right and of a wrong repetition figure. The factor for the fit is the product of the first row divided by the product of the second. It is convenient to ~~divide~~ split this up, as indicated by the vertical lines into the product of

$$\frac{c_0 c_1 c_2 c_3 (1-c_4)}{\left(\frac{1}{26}\right)^4 \frac{25}{26}}$$

$$\frac{1-a_0}{\frac{25}{26}}$$ occurring three times

$$\frac{a_0(1-a_1)}{\frac{1}{26}\cdot\frac{25}{26}}$$

$$\frac{a_0\, a_1\, a_2 (1-a_3)}{\left(\frac{1}{26}\right)^3 \frac{25}{26}}$$

$$\frac{a_0\, a_1}{\left(\frac{1}{26}\right)^2}$$

and this product may be put into the form of the product of

$\dfrac{c_0 c_1 c_2 c_3 (1-c_4)}{(1/26)^4\; 25/26} \cdot \left(\dfrac{1-a_0}{25/26}\right)^{-5}$ which we call 'the factor for an initial tetragram repeat, level'

$\dfrac{a_0(1-a_1)}{\frac{1}{26}\cdot 25/26} \cdot \left(\dfrac{1-a_0}{25/26}\right)^{-2}$ the factor for a single repeat

$\dfrac{a_0 a_1 a_2 (1-a_3)}{\left(\frac{1}{26}\right)^3 \cdot 25/26} \cdot \left(\dfrac{1-a_0}{25/26}\right)^{-4}$ → the factor for a trigramme

$\dfrac{1-a_0}{1-a_2}$ → the correction for a final bigramme

$\left(\dfrac{1-a_0}{25/26}\right)^{16}$ the factor for an overlap of 16.

$\dfrac{a_0\, a_1 (1-a_2)}{\left(\frac{1}{26}\right)^2 \frac{25}{26}} \cdot \left(\dfrac{1-a_0}{25/26}\right)^{-3}$ the factor for a bigramme.

We shall neglect the correction for a final bi ramme (or whatever it may be). It is in any case rather small, and ~~ix~~ vanishes if the repetition figure ends with o: also with our conventions the whole question of the ends of repetition figures has been left rather in doubt.

Now let us put

$$a_0\, a_1 \ldots a_r\,(1-a_{r+1}) = k_r \qquad K_{r+1}$$
$$b_0\, b_1 \ldots b_r\,(1-b_{r+1}) = j_r$$
$$c_0\, c_1 \ldots c_r\,(1-c_{r+1}) = k_r$$

The values of the i_r can be obtained as follows. We take a number of plain language messages and leave out two or three words at the beginning. Then combine the messages to form one long message: this message may be made to 'eat its own tail' i.e. it may be written round a circle. If the message were compared with itself in every possible position, except level, we should expect to get repetition figures ~~including~~ which when divided up as above by vertical lines after each o, contain $h\frac{N(N-1)}{2}k_r$ $(= N_r)$ parts which consist of r symbols x followed by an o, or as we may say N_r 'actual r-gramme repeats'. ~~Thisxxxxxxx~~ The values of N_r can be calculated given ~~fxxx~~ the 'apparent number of r-gramme repeats' M_r for each r. This apparent number of r-gramme repeats is the number of series of r consecutive symbols x in the repetition figures regardless of what precedes or follows the series. By considering the ways in which an actual repeat can give rise to apparent repeats of various lengths we see that

where h is the probability of an o

$$M_r = N_r + 2N_{r+1} + 3N_{r+2} + \ldots$$

and therefore

$$M_r - M_{r+1} = N_r + N_{r+1} + N_{r+2} + \ldots$$

and

$$(M_r - M_{r+1}) - (M_{r+1} - M_{r+2}) = N_r$$

The calculation of j_r may perhaps best be done by comparing the beginners of a number of messages with the long circular message, and the values of i_r by comparing the beginners among themselves. A similar technique of actual and apparent numbersof repeats can be used. I shall not go into this in detail.

The formulae required may now be assembled.

μ_r : decibanage for an r-gramme repeat

ν : negative decibanage for unit overlap

$s_{\beta,r}$: number of occurrences in the statistics of the r- gramme β .

N = total number of letters in the statistics

Then if $$M_r = \sum_\beta s_{\beta,r}(s_{\beta,r}-1)/2$$

$$N_r = M_r - 2M_{r+1} + M_{r+2}$$

$$L = N(N-1)/2$$

$$k_r = N_r / Lh$$

h may be calculated as follows From the identity

$$(1-a_0) + a_0(1-a_1) + a_0 a_1 (1-a_2) + \dots = 1$$

we get $k_0 + k_1 + k_2 + \dots = 1$

i.e. $$\frac{L-M_1}{Lh} = 1$$

$$1 - a_0 = k_0 = N_0 / L - M_1 = \frac{L - 2M_1 + M_2}{L - M_1}$$

$$\mu_r = 10 \log_{10} \left(\frac{26^{r+1} k_r}{25} \right) + (r+1)\nu$$

$$\nu = -10 \log_{10} \frac{26(1-a_0)}{25}$$

Transpos~~k~~ition~~x~~ ciphers

In making calculations about substitution ciphers we have often found it useful to treat the plain language as if it were produced by independent choices for the letters, using certain fixed frequencies with which the letters are chosen. Our method for Vigenere and one of the simplified forms of repeat theory could be based on this sort of assumption. With a transposition cipher however such an assumption would be useless or worse than useless, for it would result in the conclusion that all transpositions were equally likely. We have therefore to make a slightly less crude assumption, and the one which suggests itself is that the letters forming the plain language are chosen consecutively, the probability of getting a particular letter depending only on what the letter is and what the preceding letter was. It is easily verified that if $p_{\alpha\beta}$ is the proportion of bigrammes $\alpha\beta$ in plain language and p_α the frequency of the letter α then the probability $q_{\alpha\beta}$ of a letter β following an α is $p_{\alpha\beta}/p_\alpha$. The probability of a piece of plain language of length L letters saying $\alpha_1\ \alpha_2 \ldots\ \alpha_L$ is then $p_{\alpha_1} q_{\alpha_1\alpha_2} q_{\alpha_2\alpha_3} q_{\alpha_3\alpha_4} \cdots q_{\alpha_{L-1}\alpha_L}$ which may also be written as $J(\alpha_1, \ldots, \alpha_L)$. We may also calculate the probability ~~f~~of a piece of plain language having certain given letters in given places, the remainder of the ~~xxxx~~ message being unspecified. The probability is given ~~by~~

by

$$\Sigma(\xi_1 \ldots \xi_L \text{ consistent with data})\ J(\xi_1, \ldots, \xi_L)$$

and if the data is that the known letters are

$$\underset{n_1 \text{ dots}}{\ldots} \beta_1 \underset{n_2 \text{ dots}}{\ldots} \beta_2 \ldots \quad \ldots \beta_{r-1} \underset{n_r \text{ dots}}{\ldots} \beta_r \ldots \qquad (D)$$

it is approximately

$$p_{\beta_r} \qquad \left(\prod_r \beta_r\right) \cdot \prod_{n_{r+1}=0} \frac{p_{\beta_r\beta_{r+1}}}{p_{\beta_r} p_{\beta_{r+1}}} \qquad (A)$$

A more or less rogorous deduction of this approximation from the assumptions above is given ~~below~~ at the end of this section. For the present let us see how it can be applied. If we have two theories ~~xxxxxxxxx~~ about the transposition of which the one requires the above pattern of letters, and the other brings the same letters in to positions in which no two of them are consecutive, then the factor in favour of the first as compared with the second is

$$\prod_{n_{r+1}=0} \frac{p_{\beta_r\beta_{r+1}}}{p_{\beta_r} p_{\beta_{r+1}}}$$

We can apply this straightforwardly to the case of ~~xxxxxx~~ simple transposition by columns. The following text is known to be a simple transposition of a certain type of German text with a key length of not more than 15.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
A	-7	5	10	-4	-5	-7	3	-4	-25	-7	-8	-1	1.	3	-20	-6	-2	-1	-6	-6	8	0	-11	-20	-7	-13	A
B	-7	6	-10	-9	9	-13	-6	-13	1	-11	-12	-7	-2	-18	4	-13	-2	1	-13	-15	-1	-16	-12	-18	-4	-12	B
C	-14	-3	-3	-18	-19	-20	-19	27	-21	-10	3	-12	-12	-21	-15	-13	-2	-14	-14	-21	-21	-5	-17	-17	-14	-17	C
D	5	-8	-18	-18	4	-6	-13	-16	-4	2	-9	-11	-6	-13	4	-4	-2	9	-6	-11	2	-2	-9	-10	-5	-4	D
E	-15	4	0	-8	-15	-5	-2	-5	10	-8	-14	1	-1	5	-22	-6	-3	8	-2	-6	-5	-6	-11	-13	-8	-4	E
F	-2	-11	-20	-9	-2	10	-3	-15	-12	4	-2	1	-8	-11	6	0	-3	-8	-8	-1	9	-8	-8	1	-2	-1	F
G	-1	-8	-18	-5	8	-10	-2	-13	-10	2	6	-3	-13	-11	-10	-14	-3	0	-3	-9	-2	-7	-7	1	2	-2	G
H	1	-10	-12	-8	4	-5	-11	-12	-2	5	-10	2	-3	-10	-2	-2	-2	4	-1	9	-11	-7	-4	-7	-15	-8	H
I	-14	-4	10	-10	0	-1	2	-18	-17	-6	-5	0	-1	9	-17	-7	-1	-10	-1	4	-19	-7	-16	-3	-9	-4	I
J	3	8	-4	1	-3	0	-1	2	-6	14	1	-3	1	-7	-3	7	-2	-12	1	-8	-4	5	-2	0	9	-12	J
K	-2	-9	-12	3	-3	1	-7	-9	-9	4	20	-2	1	-14	-15	-13	7	-6	-5	0	0	-3	-6	-17	-5	-8	K
L	6	0	-6	2	-4	-7	-1	-15	1	-3	-14	8	-4	-2	-2	-5	8	-18	-5	-5	2	-1	-10	3	-5	-3	L
M	6	-1	-17	-6	1	-9	-6	-5	5	1	-10	-14	15	-14	0	-2	-2	-14	-6	-5	-11	-4	-7	3	-6	2	M
N	-1	-8	-18	10	-6	3	11	-9	-8	2	-6	-11	-5	-7	-2	-9	2	-20	6	-6	4	-3	-6	3	0	-5	N
O	-10	-8	-10	-6	-3	4	-6	-13	-18	0	-1	2	6	0	1	2	-2	9	0	2	-18	3	-11	6	-6	-2	O
P	2	-7	-13	-13	4	-3	-14	-5	-8	3	-12	0	-2	-8	6	8	-2	5	-15	-10	5	-12	-12	-13	-11	-1	P
Q	-3	-2	-2	-2	-3	-3	-3	-2	-2	-1	-3	-2	-2	-3	-2	-2	3	-3	-3	-3	13	-2	-2	-2	-2	-2	Q
R	2	3	-3	5	2	0	-6	-10	-2	2	-1	-6	-2	-9	-6	1	-1	-11	-1	2	-1	-2	0	1	-1	2	R
S	-3	-1	11	-2	-4	-7	-13	-12	2	-6	1	-9	-10	-7	-5	8	-3	-19	5	7	-8	5	-9	-15	1	4	S
T	3	-3	-14	-7	5	-3	-11	-11	-4	1	-1	-4	-5	-9	-2	-7	-3	-2	-2	5	-5	-3	-4	2	-3	9	T
U	-11	-4	-3	-15	0	5	-6	-2	-26	-12	-16	10	-2	8	-11	3	2	-3	-3	-11	-2	-9	-11	-12	-9	-20	U
V	-13	-16	-15	-18	2	-15	-16	-5	12	-10	-7	-16	-15	-17	14	-12	-2	-17	-18	-8	-11	10	-14	-4	8	-8	V
W	-1	-17	-17	-17	4	-18	-18	-18	2	-10	-9	-13	-10	-20	21	-12	-2	-19	-20	-19	-13	-14	-3	-16	-6	-16	W
X	1	1	-13	6	1	0	-4	-13	-14	3	0	-12	2	-10	-12	-2	8	-15	-6	-4	-5	9	1	10	-6	3	X
Y	-11	-1	-14	7	-5	-4	-9	-9	-11	-9	0	-9	-6	-5	-4	-11	-2	-11	-6	-6	-7	9	-2	-13	25	-2	Y
Z	-13	-13	-17	-13	-4	-12	-8	-14	-7	-3	-10	-11	-17	-16	-1	-5	-2	-20	-13	-10	8	-8	27	0	-6	-2	Z
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	

Fig 6. Exclusive bigramme scores in half decibans, i.e. $20 \log_{10} \frac{P_{\alpha\beta}}{P_\alpha P_\beta}$ for a certain kind of German traffic.

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

To solve this transposition we may try comparing the first six letters S A T P T W which we know form part of one column with each other series of six letters in the message, for we know that one such comparison will give entirely bigrammes occurring in the decode. We may try first

S F
A A
T S
P T
T A
W U
~~B U~~

The factor for a transposition which brings these letters together, as compared with one which leaves them apart is

$$\frac{P_{SF}}{P_S P_F} \cdot \frac{P_{AA}}{P_A P_A} \cdots \frac{P_{ST}}{P_S P_T} \quad \frac{P_{WU}}{P_W P_U}$$

By using a table of values of $20 \log_{10} \frac{P_{\alpha\beta}}{P_\alpha P_\beta}$ made up for the type of traffic in question, and given to the nearest integer (table of values of $\frac{P_{\alpha\beta}}{P_\alpha P_\beta}$ expressed in half-decibans) we get the product by addition. Such a table is shown in Fig 6 . The scores for this particular column are SF -7, AA -7, TS -2, PT -10, TA -3, WU -13, ~~SU -7~~, totalling ~~-29~~ -36. If we consider this combinationas a priori about 100:1 against (there are 95 letters in the message) it is a posteriori about 3000:1 against. Similar scoring may be done for every possible comparison of S A T P T W with six consecutive letters of the message. The comparison may be made both with S A T P T W as earlier and as later column; one may also use the last six letters of the message H U A N W R . The results of doing this are shown in ~~fix~~ Fig 7 . The message has been written out vertically. The first column of figures after the message gives the scores for S A T P T W as earlier column, entered against the first letter of the later column, e.g. the -36 as calculated above gets entered against the F of F A S T A U . The second column after the message ~~is~~ consists of the scores for H U A N W R as first column [and the column before the message gives the scores for H U A N W R as second column]. I doubt it SW. One of these columns has been worked out in detail but in the other two crosses have been put in where the scores are very bad. The scores which eventually turned out to be right are ringed. The fourth comparison, which did not have to be done scored very badly viz. -27. Amongst the good scores which were wrong there was one score of 37. It was not difficult to see that this one was wrong as most of the score came from WO which requires Z to precede it, and there was no Z in the message. Apart from this fact the comparison was about evens, although if we take into account the fact that there was no better score it would be better. [We have already had a case of this kind of thing in connection with Vigenere; if the various positions are a priori equally likely and the factors are $f_1, f_2 \text{ --- } f_N$ then the value for the probability of the rth alternative is better than $\frac{f_r/N}{1+f_r/N}$ $f_r/\Sigma f_i$]

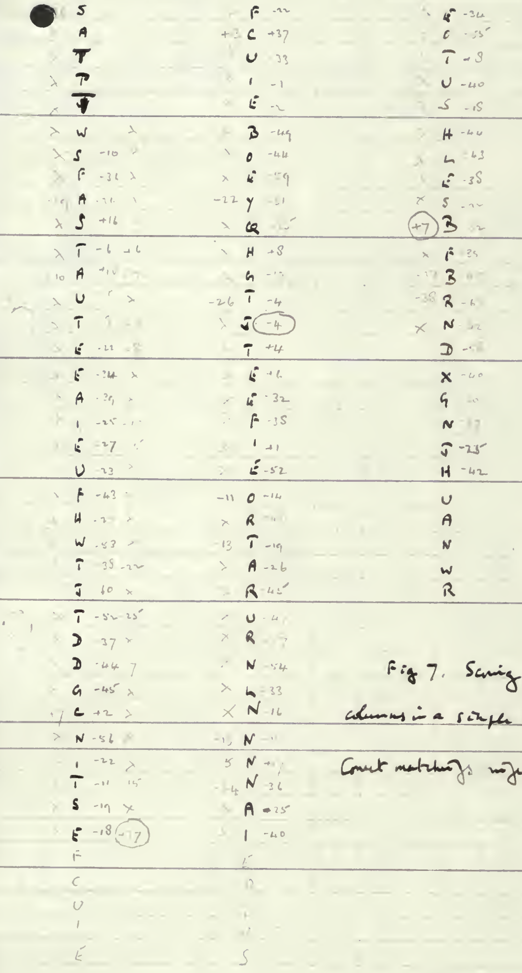


Fig 7. Scoring the matching of columns in a simple transposition.

Correct matchings noted.

(Semi-) Rigorous deduction of the formula (A). (This is something of a digression).

The probability of a piece of plain language coinciding where necessary with the data (D) is

$$p_{\beta_1}\,\tau_{n_2,\beta_1\beta_2}\,\tau_{n_3,\beta_2\beta_3}\cdots\tau_{n_M,\beta_{M-1}\beta_M}$$

where $\tau_{n,\alpha\beta}$ is

$$\sum_{\eta_1\eta_2\cdots\eta_n} q_{\alpha\eta_1}\,q_{\eta_1\eta_2}\cdots q_{\eta_n\beta}$$

since

$$\sum_{\eta_1\cdots\eta_{n_1}} p_{\eta_1}\,q_{\eta_1\eta_2}\cdots q_{\eta_{n_1}\beta_1} = p_{\beta_1}$$

we can put

$$\tau_{n,\alpha\beta} = (Q^{n+1})_{\alpha\beta}$$

where Q is the matrix whose $\alpha\beta$ coefficient is $q_{\alpha\beta}$: The formula (A) would then be accurate if we could say that for $n>0$, $(Q^{n+1})_{\alpha\beta} = p_\beta$. This is not true, but it is true that except for very special values for $q_{\alpha\beta}$, $(Q^n)_{\alpha\beta} \to p_\beta$ as $n \to \infty$, and the convergence is rather rapid. To prove this I shall assume that the eigenvalues of Q are all different in modulus. In this case we can find a matrix U with unit determinant, such that $U^{-1}QU$ is in diagonal form

$$M = U^{-1}QU = \begin{pmatrix} \mu_1 & 0 & 0 & \cdots \\ 0 & \mu_2 & 0 & \cdots \\ 0 & \ddots & \ddots & 0 \\ \vdots & & 0 & \mu_{26} \end{pmatrix}$$

since $QU = UM$ we have

$$\sum_\gamma q_{\alpha\gamma} u_{\gamma\beta} = \sum_\varepsilon u_{\alpha\varepsilon} m_{\varepsilon\beta}.$$

i.e.

$$\sum_\gamma q_{\alpha\gamma} u_{\gamma\beta} = \mu_\beta u_{\alpha\beta}$$

that is, for each β, $u_{\alpha\beta}$ provides a solution of

$$\sum_\gamma q_{\alpha\gamma} l_\gamma = \mu l_\alpha \qquad (E)$$

with $\mu = \mu_\beta$. Conversely if we have any solution of (E) $\mu = \mu_\theta$, $l_\alpha = k u_{\alpha\theta}$ then ~~[illegible]~~ for some k, θ and all α, for as U is non singular we can ~~pu~~ find numbers c_γ such that $l_\alpha = \sum_\gamma u_{\alpha\gamma} c_\gamma$ for all α, and then substituting in (E) we get

$$\sum_{\gamma,\delta} q_{\alpha\gamma} u_{\gamma\delta} c_\delta = \mu \sum_\delta u_{\alpha\delta} c_\delta$$

i.e.

$$\sum_\delta (\mu_\delta c_\delta - \mu c_\delta) u_{\alpha\delta} = 0$$

which, since U is non singular implies

$$\mu = \mu_\delta \text{ or } c_\delta = 0 \quad \text{for all } \delta.$$

As the roots $\mu_1 \ldots \mu_{26}$ are all different there is only one value θ of δ for which $\mu = \mu_\delta$ and so $l_\alpha = c_\theta u_{\alpha\theta}$ all α.

Now putting $l_\alpha = 1$ for all α we see that one member of the series $\mu_1 \ldots \mu_{2b}$ is 1, for (E) is certainly satisfied. I shall prove that the remaining eigenvalues satisfy $|\mu| \leqslant 1$. We first prove that if $\mu \neq 1$ then $\Sigma p_\alpha l_\alpha = 0$. This follows by multiplying (E) on each side by P_α and summing. Since $q_{\alpha\beta} = \frac{P_{\alpha\beta}}{P_\alpha}$ and $\sum\limits_\alpha P_{\alpha\beta} = P_\beta$ we get

$$\sum_{\alpha\gamma} q_{\alpha\gamma} l_\gamma = \Sigma p_\gamma l_\gamma = \mu \Sigma p_\alpha l_\alpha$$

which implies $\mu = 1$ or $\Sigma p_\alpha l_\alpha = 0$. now ~~Let l_α satisfy (E) with $\mu \neq 1$. Then $\Sigma p_\alpha l_\alpha = 0$ and therefore $\mathscr{R}\, l_\sigma < 0$ for some σ~~, Next we show that each μ for which $|\mu| > 1$ is real and positive. Let l_α satisfy (E) with $|\mu| > 1$; then the eigenvalue for $\bar{l}_\alpha$ is $\bar{\mu}$ and so

$$\sum_\beta (Q^r)_{\alpha\beta} \left(1 + \varepsilon(l_\beta + \bar{l}_\beta)\right) = 1 + 2\varepsilon \mathscr{R} \mu^r l_\alpha$$

If $\varepsilon > 0$ has been chosen so small that ~~thexRxHxE~~ $\mathscr{R}\,\varepsilon l_\beta > -\frac{1}{2}$ all β then the L.H.S. is positive for the coefficients in the matrix are positive, whereas the R.H.S. is negative for suitably chosen r, unless $l_\alpha = 0$. If now $\mu > 1$ we may take it that l_α is real for each α. As it must satisfy $\Sigma p_\alpha l_\alpha = 0$ it is neagtive for some α, but then

$$\sum_\beta (Q^r)_{\alpha\beta} (1 + \varepsilon l_\beta) = 1 + \varepsilon \mu^r l_\alpha$$

and if ~~thx~~ ε is chosen so that ~~thxxxt~~ $1 + \varepsilon l_\beta > 0$ all β the L.H.S. is positive whereas the R.H.S. is negative for sufficiently large r. All the eigenvalues therefore satisfy $|\mu| \leqslant 1$.

as the eigenvalues are all different in modulus. this means that ~~ixxofxform~~ $|\mu| < 1$ except for one value of μ . Then as $r \to \infty$, M^r tends to ~~thexxxxxi~~ a matrix which has only one element different from 0, and that a 1 on the diagonal, say in position $\sigma\sigma$. ~~Thexx xxndxxtoxthexlimix~~ Calling this matrix X_σ the series of matrices Q^r tends to the limit $U^{-1}X_\sigma U$. This matrix is the one and only one y which satisfies $yQ\cancel{y} = y$ $y^2 = y, y \neq 0$ and is therefore the one whose $\alpha\beta$ coefficient is p_β .

There is another probability problem that arises in connection with simple transposition. With a message of length L, and a key length of K what is the probability that the m th letter will be at the bottom of a column? Let D be the length of the short columns i.e. $D = [L/K]$, and let $E = L - DK$. Then if the m th letter is at the bottom of the ω th column we must have $\frac{m}{D+1} \leq \omega \leq m/D$, and there will be $(D+1)\omega - m$ short and $m - D\omega$ long columns amongst these first ω columns~~xxxxxx~~. There are $\binom{\omega}{m-D\omega}\binom{K-\omega}{E-m+D\omega}$ ways in which the short and long columns can be arranged consistently with this, and altogether $\binom{K}{E}$ ways in which the columns can be arranged, so that the probability of the mth letter being at the bottom of a column ~~xxx~~ is

$$\sum_{\frac{m}{D+1} \leq \omega \leq m/D} \binom{\omega}{m-D\omega}\binom{K-\omega}{E-m+D\omega} \Big/ \binom{K}{E}$$

There will normally be very few terms in the ~~xxx~~ sum. Let us take the case of a message of length 133 and consider the 45th letter, assuming the key length is between 10 and 20 (inclusive). i.e. $L = 133$, $m = 45$

$K = 10$, $D = 13$, $E = 3$, $m/D+1 = 3+$ $m/D = 3+$ no terms

$K = 11$, $D = 12$, $E = 1$, $m/D+1 = 3+$ $m/D = 3+$ no terms

$K = 12$, $D = 11$, $E = 1$, $m/D+1 = 3+$ $m/D = 4+$ ~~no terms~~.

~~only term $\omega = 4$ giving proby~~

~~hence~~, only term $\omega = 4$ giving $m - D\omega = 1$ & proby $\binom{4}{1}\binom{8}{0}/\binom{12}{1}$

i.e. $4/12$

$K=13,\ D=10,\ E=3 \quad m/D+1 = 4+,\ m/D = 4+$, no terms

$K=14,\ D=9,\ E=7 \quad m/D+1 = 4+,\ m/D = 5$ only term $\omega=5,\ m-D\omega=0$

$$\text{prob}^y \quad \binom{5}{0}\binom{9}{7}\Big/\binom{14}{7} = 3/286 = .0105$$

$K=15,\ D=8,\ E=13 \quad m/D+1 = 5,\ m/D = 5+$ only term $\omega=5,\ m-D\omega=5$

$$\text{prob}^y \quad \binom{5}{5}\binom{10}{8}\Big/\binom{15}{13} = 3/7 = .428$$

$K=16,\ D=8,\ E=5 \quad m/D+1 = 5+\ m/D = 5+$ only term $\omega=5\ m-D\omega=5$

$$\text{prob}^y \quad \binom{5}{5}\binom{11}{0}\Big/\binom{16}{5} = 1/4368 = .000229$$

$K=17,\ D=7,\ E=14 \quad m/D+1 = 5+\ m/D = 6+$ only term $\omega=6,\ m-D\omega=3$

$$\text{prob}^y = \binom{6}{3}\binom{11}{14}\Big/\binom{17}{14} = 1/34 = .0307$$

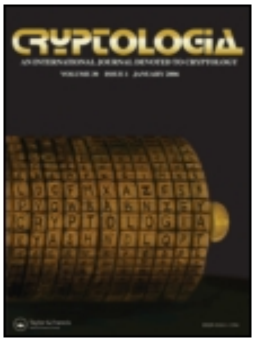
$K=18,\ D=7\ \ E=7 \quad m/D+1 = 5+\ m/D = 6+$ only term $\omega=6\ m-D\omega=3$

$$\text{prob}^y = \binom{6}{3}\binom{12}{4}\Big/\binom{18}{7} = \frac{4950}{15912} = .311$$

$K=19,\ D=7\ \ E=0 \qquad \text{prob}^y = 0$

$K=20,\ D=6,\ E=13,\ m/D+1 = 6+,\ m/D = 7+$ only term $\omega=7,\ m-D\omega=3$

$$\text{prob}^y = \binom{7}{3}\binom{13}{4}\Big/\binom{20}{7} = \frac{35\times143}{15504} = .323$$



Taylor & Francis
Taylor & Francis Group

Cryptologia

ISSN: 0161-1194 (Print) 1558-1586 (Online) Journal homepage: http://www.tandfonline.com/loi/ucry20

Commentary on Alan M. Turing: The Applications of Probability to Cryptography

Sandy Zabell

To cite this article: Sandy Zabell (2012) Commentary on Alan M. Turing: The Applications of Probability to Cryptography, Cryptologia, 36:3, 191-214, DOI: 10.1080/01611194.2012.697811

To link to this article: http://dx.doi.org/10.1080/01611194.2012.697811

Published online: 11 Jul 2012.

Submit your article to this journal

Article views: 789

View related articles

Citing articles: 4 View citing articles

Cryptologia, 36:191–214, 2012
Copyright © Taylor & Francis Group, LLC
ISSN: 0161-1194 print
DOI: 10.1080/01611194.2012.697811

Commentary on Alan M. Turing: The Applications of Probability to Cryptography

SANDY ZABELL

Abstract In April 2012, two papers written by Alan Turing during the Second World War on the use of probability in cryptanalysis were released by GCHQ. The longer of these presented an overall framework for the use of Bayes's theorem and prior probabilities, including four examples worked out in detail: the Vigenère cipher, a letter subtractor cipher, the use of repeats to find depths, and simple columnar transposition. (The other paper was an alternative version of the section on repeats.) Turing stressed the importance in practical cryptanalysis of sometimes using only part of the evidence or making simplifying assumptions and presents in each case computational shortcuts to make burdensome calculations manageable. The four examples increase roughly in their difficulty and cryptanalytic demands. After the war, Turing's approach to statistical inference was championed by his assistant in Hut 8, Jack Good, which played a role in the later resurgence of Bayesian statistics.

Keywords Alan Turing, Bayes's theorem, crib, cryptanalysis, deciban, depths, factor theorem, half-deciban, I. J. Good, index of coincidence, Jerzy Neyman, letter subtractor cipher, Markov chain, odds, prior probabilities, probability, R. A. Fisher, simple columnar transposition, theory of repeats, Vigenère cipher

On 17 April 2012, Government Communications Headquarters (GCHQ; the U.K. equivalent of the U.S. National Security Agency), released two documents on cryptanalysis written by Alan Turing during WWII. The first of these, "The Applications of Probability to Cryptography" [14], is 44 pages long; it discusses the general use of mathematical probability, specifically *Bayes's theorem*, in cryptanalysis. The second document, "Paper on the Statistics of Repetitions" [15], is much shorter (8 pages long) and derives a specific technical result extending a classical technique worked on earlier by cryptologists such as William Frederick Friedman; see, e.g., [13, pp. 68–70].

"The Applications of Probability to Cryptography" consists of five parts: an introduction, presenting Turing's favored Bayesian approach, followed by the analysis of four "Straightforward Cryptanalytic Problems" illustrating the use of this method. The first section of the paper ("Introduction") sets out Turing's basic theoretical framework. That Turing took such an approach has been known in general terms for some time. In 1979, I. J. ("Jack") Good (Turing's chief statistical assistant in Hut 8 in 1941) wrote a paper describing for the first time, albeit only in very general (and guarded) terms, how Turing used Bayesian methods of attack [6]. This was, however, only at a very early stage in the declassification of information

Address correspondence to Sandy Zabell, Department of Statistics, Northwestern University, 2006 Sheridan Rd., Evanston, IL 60208, USA. E-mail: zabell@math.northwestern.edu

relating to Allied cryptanalytic efforts during World War II, and Good's paper scrupulously avoided going into the concrete specifics of any of the cryptographic systems being attacked.

The interest of this paper, therefore, lies more in its practical examples: demonstrating how Bayesian methods can be used to attack cryptographic systems by examples involving systems of increasing complexity. It illustrates not only how such methods can be used in the cryptanalytic setting, but also something else: that the effective attack on a system requires a skillful blend of the theoretical and the practical—the awareness that sometimes the art of the cryptanalyst lies in being able to find simplifying assumptions that transform an (apparently) intractable problem into one that is feasible. Alan Turing may have been an outstanding pure mathematician who made important contributions to mathematical logic and computer science, but this paper gives us insight into a very different aspect of the man: the serious practical cryptanalyst.

A note on the commentary itself. I have largely followed Turing's notation and examples, on occasion noting when this is not the case, but have not attempted to do so in a systematic way. I have also silently changed spelling (e.g., "bigram" instead of "bigramme") and punctuation in quotations when I thought not to do so might be distracting. Because the paper itself is now readily available online (at the website of the U.K. National Archives), the reader can (and should) go back to see how Turing himself put things. One of the aims of the commentary is to facilitate this process.

1. Introduction

The first eight pages of Turing's paper give a brief synopsis of his view of probability and its use in cryptology.

1.1. Probability and Odds

First of course, there is the question of just what *is* "probability"? Turing begins by giving a brief, informal definition of the term for the purposes of the paper:

> The probability of an event on certain evidence is the proportion of cases in which that event may be expected to happen given that evidence.

Turing's definition blends elements of knowledge ("on certain evidence"), frequency ("the proportion of cases"), and belief (the "may be expected to happen"), and so places him outside the purely physical view of probability as a frequency having an objective if unknown value. This set Turing apart from the statistical mainstream of his day but was central to his approach.

Probability on this view is *conditional*, not absolute. Turing gives as an example using actuarial data to estimate the probability that Hitler will live to 70 given (a) we just know that he is a man versus (b) that he is also known to be 52. This illustrates Laplace's famous dictum: "probability is relative in part to [our] ignorance, and in part to our knowledge." The 19th century French mathematician Joseph Bertrand gave an even more piquant example of this dependence on our evidence: the king of Siam is 40; what is the probability he will live another decade? It has one value for those who have questioned his physician, yet another for the physician himself, a very different one for those conspirators who have undertaken to strangle him the next day [2, pp. 90–91]!

In modern notation, if A denotes an event of interest, and E the evidence regarding it, then $P(A|E)$ is used to denote the *probability of A given E*. For reasons that will become clear shortly, Turing often works in terms of the *odds* in favor of an event rather than its probability; if p is the value of this probability, then the odds in its favor is $p/(1-p)$.

1.2. Probabilities Based on Part of the Evidence

One of the skills that separates the successful applied mathematician from the pure theoretician is the ability to recognize the utility and power of carefully chosen simplification. Thus Turing states:

> When the whole evidence about some event is taken into account it may be extremely difficult to estimate the probability of the event, even very approximately, and it may be better to form an estimate based on a part of the evidence, so that the probability may be more easily calculated. [14, p. 2]

Turing evidently regards this as an important point, because it is the subject of an entire (if brief) subsection. He makes the interesting remark:

> Unless the traffic is very small indeed the theoretical answer to the problem 'what are the probabilities of the various keys?' will be of the form 'The key . . . has a probability differing almost imperceptibly from 1 (certainty) and the other keys are virtually impossible'. But a direct attempt to determine these probabilities would obviously not be a practical method. [14, p. 2]

This comment presumably has in mind both the computational challenge of exploiting all the information available in intercepted messages and the enormous number of possible keys for the German encryption systems, such as the Enigma. Bletchley Park's cryptanalytic counterparts in the German communications security organizations (such as Dr. Erich Hüttenhain of OKW/Chi) believed in the security of some of the German systems not because they thought they were theoretically unbreakable, but because they thought they were unbreakable in practice.

1.3. A Priori Probabilities

Effective cryptanalysis (and, more generally, any serious statistical analysis) involves the synthesis of different forms of information.

> The evidence concerning the possibility of an event occurring usually divides into a part about which statistics are available, or some mathematical method can be applied, and a less definite part about which one can only use one's judgment. [14, p. 2]

The "less definite part about which one can only use one's judgment" is where so-called "a priori probabilities" enter the picture, and Turing's willingness to use them put him apart from most statisticians at that time, who viewed them either as arbitrary or putting numbers on something that could neither be measured nor expressed in numerical form.

Contemporary distrust of prior probabilities was based in part on their use in situations where little relevant information was available beforehand; in effect, it was argued, you were pulling a rabbit out of a hat, creating something out of nothing. If you did not know something, you should just acknowledge this. Surely it was better to develop objective statistical methods based solely on the quantitative statistical data at hand; in science, the scientist always has the option of performing further experiments and generating more data.

Persuasive as this worldview was to many in the statistical profession of the time, this was a totally inappropriate paradigm for Bletchley Park. There was often a substantial amount of directly relevant prior information available, such as the type of message, its possible content, and who was sending it (so disregarding this would be wasteful); collecting more data was impossible (all you had was the message or messages in front of you); and a decision one way or the other had to be made as to whether the message should be attacked, and, if so, what the most promising next step was.

There are some cases where such *a priori* reasoning seems harmless enough. In a simple substitution cipher, if it is thought that the keys are chosen at random, it seems reasonable to say in the absence of any further information that every letter has an equal chance (i.e., 1 in 26) of being the cipher equivalent of *E*.

But in many cases, the process of assigning a prior probability can be much less clear. Turing illustrates this process with a simple example. Suppose that three messages are intercepted using a new form of encryption, and that it is observed that in each case the letter *V* is found in the 17th place and *G* in the 18th, and that one wishes to estimate the probability that this will be the case in other messages using this form of encryption. In order to do this, one has to have an estimate of how likely this would be a priori for a cipher; Turing estimates this to be about 1/5,000,000.

The estimation of the odds in favor of the rule can then be done by computing the ratio of the expected number of favorable cases versus unfavorable cases. It may be clearer if we generalize the example and consider the case of prior odds of p (where, for Turing, $p = 1/5{,}000{,}000$). Consider a large number N of *ciphers* (not messages) "chosen at random." Of these, we expect Np to obey the rule, and $N(1-p)$ not to obey. Suppose we are told that for one of the ciphers, *VG* has been observed in places 17 and 18 in three separate messages. This will be seen in all of the Np ciphers obeying the rule, but only in $(1-p)N/676^3$ of the others. (The chance of this occurring by chance in a single message is $1/26^2 = 1/676$; the chance in three messages is $1/676^3$.) Thus, the posterior odds in favor, thought of as the ratio of favorable to unfavorable cases, is

$$\frac{Np}{\frac{N(1-p)}{676^3}} = 676^3\left(\frac{p}{1-p}\right) = (308{,}915{,}776)\left(\frac{p}{1-p}\right).$$

(Note that the N has dropped out; it was merely a convenient concrete way of thinking through the argument.)

Thus, the odds in favor are the product of 676^3 and the prior odds $p/(1-p)$. What are the prior odds o in favor of the rule? This seems elusive, but Turing's point is this: *within a wide range of latitude for o, we arrive at a useful conclusion.* For example, Turing argues that a reasonable estimate for p is about 1/5,000,000, and using this value gives odds in favor of

$$(308,915,776) \cdot \frac{1}{4{,}999{,}999},$$

or about 60 to 1. Even if we took a value of p that was an order of magnitude larger (so the odds would be about 600 to 1) or an order of magnitude less (so the odds would be about 6 to 1), we would still conclude that there was some evidence in favor of the rule (although, of course, the exact strength of that evidence would depend on the prior odds).

Where did the estimate $p = 1/5{,}000{,}000$ itself come from? Here is where the mix of guesswork, experience, and mathematics meets. Turing explains:

> This judgment is not entirely a guess; some rather inaccurate mathematical reasoning has gone into it, something like this:
>
> The chance of there being a rule that two consecutive letters somewhere after the 10th should have certain fixed values seems to be about 1/500 (this is a complete guess). The chance of the letters being the 17th and 18th is about 1/15 (another guess, but not quite so much in the air). The probability of the letters being *V* and *G* is 1/676 (hardly a guess at all, but expressing a judgment that there is no special virtue in the bigramme *VG*). Hence the chance is $1/(500 \times 15 \times 676)$ or about 1/5,000,000. This is however all so vague, that it is more usual to make the judgment "1/5,000,000" without explanation. [14, p. 3]

One can well imagine why a professional statistician might be reluctant to base a theory of statistical inference on such a foundation! They would have regarded this as merely confirming their worst suspicions regarding the "arbitrary" nature of prior probabilities. But that would miss the point; for Turing, the objective is come up with some reasonable "ballpark" number:

> The question as to what is the chance of having a rule of this kind might of course be solved by statistics of some kind, but there is no point in having this very accurate, and of course the experience of the cryptographer itself forms a kind of statistics. [14, p. 3]

This point will be discussed further in the final section of the commentary.

1.4. The Factor Principle

Turing states that "[n]early all applications of probability to cryptography depend on the 'factor principle' (or Bayes's theorem)." This reflects Turing's view of the subject rather than one an amateur would find in the published literature of the time; perhaps nearly all applications of probability to cryptography at Bletchley Park either explicitly or implicitly made use of this principle.

Turing's factor principle is the so-called "odds ratio" version of Bayes's theorem. In the notation introduced earlier, if H_0 and H_1 are two "hypotheses" of interest (for example, two possible keys used in the encryption of a message), and

E represents some form of evidence or data (for example, the letters observed in an encrypted message), then the odds form of this theorem states that

$$\frac{P(H_1|E)}{P(H_0|E)} = \frac{P(E|H_1)}{P(E|H_0)} \cdot \frac{P(H_1)}{P(H_0)};$$

that is, the final or *posterior odds* for H_1 versus H_0 given E (the expression on the left) equals the *likelihood ratio* (the first ratio on the right) times the initial or *prior odds* (the second ratio on the right). Put another way, the likelihood ratio is precisely the factor that transforms, by multiplying, the initial odds into the final odds.

Let $\overline{H}$ denote the negation of H, the hypothesis "not-H." In the case $H_1 = H$ and $H_0 = \overline{H}$, Turing called the likelihood ratio the *factor in favor of the hypothesis H in virtue of the evidence E*.

How does one derive this formula? Let $A \cap B$ be shorthand for events "A and B." In mathematical probability, $P(A|B)$ is then *defined* to be $P(A \cap B)/P(B)$; using this, the odds version of Bayes's theorem may be easily derived.

For further discussion of factors and likelihood ratios from the Turing perspective, see [8, Chapter 6, especially pp. 62–66] and [6].

1.5. Decibanage

Often the evidence E consists of several independent parts $E_1, E_2, \ldots, E_n$. In this case, the overall likelihood of the E for a theory H_j is then the product of the individual likelihoods for E_j; that is, one has

$$P(E_1 \cap E_2 \cap \ldots E_n|H_j) = P(E_1|H_j) \times P(E_2|H_j) \times \cdots \times P(E_n|H_j)$$

for each competing theory H_j. (Turing uses the example of whether someone died of heart failure, and the items of evidence are that he died in his bed, his father died of heart failure, and his bedroom was on the ground floor (!); and we have statistics relating to all of these.)

It is easier to add than to multiply, so Turing introduces the concept of the *deciban*, 10 times the logarithm base 10 of the factor:

$$\begin{aligned} 10\log_{10}\frac{P(E|H_1)}{P(E|H_0)} &= 10\log_{10}\frac{P(E_1|H_1)}{P(E_1|H_0)} + \cdots + 10\log_{10}\frac{P(E_n|H_1)}{P(E_n|H_0)} \\ &= \sum_{j=1}^{n} 10\log_{10}\frac{P(E_j|H_1)}{P(E_j|H_0)}. \end{aligned}$$

The factor of 10 was included to simplify the arithmetic, dropping everything after the first decimal place. For example, in the cases $p = 0.55$ and $p = 0.9$, one has $\log_{10}(0.55/0.45) = 0.08715$ and $\log_{10}(0.9/0.10) = 0.95424$, and these would be reported in decibans as 0.9 and 9.5, respectively.

1.5.1. Half-Decibans

In 1941, there was a switch from decibans to "half-decibans," i.e., using a factor of 20 rather than 10. There is no theoretical justification for working in units of half-decibans; this was something arising from experience. This innovation was due to I. J. Good, who arrived at Bletchley Park in May 1941. Good said many years later [1, p. 9]:

> They were using decibans (weights of evidence), with one decimal point. So I thought, why don't we drop the decimal point and call the unit a centiban, thus saving a lot of writing. And then I noticed that if we used a half deciban (hdb) [as the basic unit of measurement] we would save much more time in both writing and arithmetic because most of the individual scores would then be single digits...
>
> *This must have saved half the time of the work on Banburismus*. Of course, every numerical analyst knows that you shouldn't carry more decimal places than you need, in hand calculations, and it was essentially in that spirit that I made this suggestion, but here were these highly intelligent people, who for some weeks had been using the deciban with a decimal point. [emphasis added]

Note: The "ban" in "deciban" derives from Banbury, a town in which sheets of paper were printed for finding repeats (discussed later). I. J. Good later wrote [6, p. 394]. "A deciban or half-deciban is about the smallest change in weight of evidence that is directly perceptible to human intuition." "Banburismus" was an essential part of the attack on the Naval Enigma. For further discussion of Banburismus and the impact of "half-decibans," see Good [3, pp. 206–208].

2. Straightforward Cryptographic Problems

2.1. Vigenère

In the classical Vigenère cipher, the letters of plaintext are encrypted using a sequence of different Caesar shift ciphers, repeating after a given period. Turing gives as an example the following rectangular array of ciphertext; the period is assumed to be known to be ten, hence the ten columns (Table 1). In this case, each column represents the encryption of nine plaintext letters using the same Caesar cipher, and the task of the cryptanalyst is to determine the shift used for each column.

This is, of course, a classical problem whose solution was already known in the 19th century. Besides providing a simple illustration of the Bayesian approach, the relatively short length of the columns makes it essential to use a statistically efficient

Table 1. Vigenère encrypted message, width of 10

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

method that extracts the maximum amount of information present in the sample, and this is precisely what the Bayesian method does.

Consider the first column. Let $P(X|Y)$ denote the probability of seeing letter X given the key is Y, and $P(X|\neg Y)$ the probability of seeing X given the key is not Y. Suppose the prior odds in favor of each key are 1:25. For the first letter D, the factor in favor of key B (say), if the frequency of C in plaintext is 0.21, is

$$\frac{P(D|B)}{P(D|\neg B)} = \frac{1}{25} \cdot \frac{25 \times 0.021}{1 - 0.021}.$$

As Turing explains, "The probability of getting D in the cipher with the key B is just the probability of getting C in the clear, which (using the count on 1000 letters in Fig 2) is 0.021. If however the key is not B, we can have any letter other than C in the clear, and the probability is $(1 - 0.021)/25$." [14, p. 9]

One can then proceed in similar fashion for the other letters in the alphabet. If we assume that the evidence of one letter is independent of another ("This is not quite correct, but is a useful approximation" [14, p. 10]), and p_α is the frequency of letter α in the language, then the final odds in favor of B being the key is

$$\frac{1}{25}\prod_i \frac{25 p_{\alpha_i - 1}}{1 - p_{\alpha_i - 1}}.$$

More generally (if letters and numbers are used interchangeably, the letters A, B, $C, \ldots$ correspond to the numbers 1, 2, ..., 26), the odds for letter β are

$$\frac{1}{25}\prod_i \frac{25 p_{\alpha_i - \beta + 1}}{1 - p_{\alpha_i - \beta + 1}}.$$

Given a table of empirical frequencies for letters (as in Table 2), one does the one-time work of calculating decibans corresponding for each factor $25p_\alpha/(1 - p_\alpha)$. Given a message, it is then straightforward, if tedious, to decode the column using the 26 different possible keys, look up the corresponding decibanages, and add (Table 3).

The Bayesian approach may be efficient from a statistical perspective, but it can be computationally demanding. Thus, Turing shows how the calculation can be streamlined so it is easy to use. First, one prepares a table of half-decibans based on a sample of letter frequencies. Next to each letter from A to Z, multiples of the corresponding half-decibans are computed and rounded to the nearest integer. A second copy of the resulting table is then written down immediately underneath

Table 2. Turing's Figure 2, count on 1,000 letters, English text

A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*
84	23	21	46	116	20	25	49	76	2	5	38	34
N	*O*	*P*	*Q*	*R*	*S*	*T*	*U*	*V*	*W*	*X*	*Y*	*Z*
66	66	15	2	64	73	81	19	11	21	16	24	3

Table 3. Turing's calculation of half decibans

Ciphertext	*D*	*R*	*X*	*T*	*T*	*P*	*B*	*W*	*T*
B decrypt	*C*	*Q*	*W*	*S*	*S*	*O*	*A*	*V*	*S*
Odds	0.54	0.05	0.54	1.97	1.97	1.77	2.29	0.28	1.97
Half-deciban	−5	−26	−5	6	6	5	7	−11	6

the first. Table 4 gives the first four rows of Turing's table. In all, there are 52 rows, the letters of the alphabet being listed twice. (The reason for this will become apparent shortly.) Note: the entries, given as they appear in Turing's table, are often off by one or two, for reasons noted later.

The calculation can be streamlined by preparing a transparent "gadget" consisting of a sheet of paper with the letters of the alphabet displayed in a vertical column. Holes are then punched next to each letter in the column of the message being attacked. The distance of a hole from the alphabet column depends on the number of times the corresponding letter appears in the message column. For example, since *B*, *D*, *P*, *R*, *W*, and *X* appear once in the first message column, holes are punched in a column immediately adjacent to the alphabet column. Similarly, since *K* appears twice and *T* three times, the corresponding holes are punched in the second and third column to the left of the alphabet column. For a given candidate decode letter, the gadget is placed over the table of half-decibans and shifted up the appropriate number of lines. If, for example, one wishes to test out key *B* for the first column, the apparatus would be shifted up one line. (The reason for repeating the alphabet twice in the table of decibans should now be clear.)

Turing concludes by noting that instead of viewing this as a case of just two rival hypotheses (a key letter is or is not the key), it would be more accurate to view this as a case of 26 rival hypotheses, corresponding to the 26 different possible keys *A*, *B*,..., *Z*. In that case, the "factor principle" takes the form (in modern notation, *D* denotes the data and A_i that the *i*th letter is the key)

$$\frac{P(A_1|D)}{P(A_1)P(D|A_1)} = \frac{P(A_2|D)}{P(A_2)P(D|A_2)} = \cdots.$$

If the keys are judged a priori equally likely ($P(A_j) = 1/26$ for all j), then for any pair of letters A_j, A_k, this reduces to

$$\frac{P(A_j|D)}{P(A_k|D)} = \frac{P(D|A_j)}{P(D|A_k)};$$

that is, the relative posterior odds equals the relative factor in favor of A_j versus A_k.

The problem thus reduces to one of computing the probabilities of seeing the column for each key. In terms of Turing's notation, given key β, the probability

Table 4. Table for scoring a Vigenère in units of a half a deciban

31	26	20	13	7	A
−23	−18	−14	−9	−5	B
−26	−21	−16	−10	−5	C
7	6	4	3	1	D
⋮	⋮	⋮	⋮	⋮	⋮

of seeing the message is $\prod_i (p_{\alpha_i - \beta + 1})$; and the posterior probability of key β is (inserting a factor of 26 for convenience)

$$\frac{\prod_i (26\, p_{\alpha_i - \beta + 1})}{\sum_\beta \prod_i (26\, p_{\alpha_i - \beta + 1})}.$$

This may be conveniently calculated by the method described earlier for

$$\prod_i \frac{25\, p_{\alpha_i - \beta + 1}}{1 - p_{\alpha_i - \beta + 1}}.$$

It is at this point that Turing (rather annoyingly) only now tells us that the table in his Figure 3 was computed using this method rather than the earlier one! The difference between the two tables, however, will be (as he notes) "rather slight." This is most easily seen by noting that the two quantities will in fact coincide if $p_{\alpha_i} = 1/26$ for all i (since then $25/(1 - p_{\alpha_i}) = 26$), and although the entire success of the attack depends on this not being the case (that is, letter frequencies not being flat random), it is both apparent and easily checked that the difference between $1 - p_{\alpha_i}$ and $25/26$ is too small to matter.

2.2. *A Letter Subtractor Problem*

One vulnerability of the Vigenère cipher is the fact that its period is ordinarily much shorter than the length of the message being encrypted. So it is natural to consider practical methods of constructing polyalphabetic substitutions of a much longer period. Turing's next example is of this type.

Consider three different Vigenère ciphers of (relatively prime) periods $91 (= 7 \cdot 13)$, $95 (= 5 \cdot 19)$, and $99 (= 3^2 \cdot 11)$, respectively, each using "slides" from 0 to 9 (or equivalently, keys from *A* to *J*) occurring in roughly equal proportion. If one superimposes the three resulting three substitutions (that is, applies one after the other in some order), the result is equivalent to a single substitution using slides ranging from 0 to 27, and having a period of $91 \cdot 95 \cdot 99 = 855, 855$, much longer than even (say) the longest of Hitler's rants to his generals.

One classical method of attack in this type of situation is to use a "crib," or *mot probable*—a word, phrase, or sentence thought likely or possible to occur at some point in the message. For example, one might expect to find the word AMBASSADOR at the beginning of a diplomatic message, based perhaps in part on past experience. Because there might be more than one candidate crib or the crib might not occur, and in either case the use of an incorrect crib will lead to a substantial waste of time, it is desirable to have some means of estimating the probability that the crib is correct. This is Turing's next example.

In the case of the composite letter subtractor, the vulnerability being exploited is that because of the design of the cipher, not all slides are equally likely to occur. For example, a slide of 25 can only arise as a sum of 9, 9, 7, or 9, 8, 8 (in each of three different orders), while a slide of 15 can arise in many more, viz. (apart from order):

9, 6, 0
9, 5, 1
9, 4, 2
9, 3, 3
8, 7, 0
8, 6, 1
8, 5, 2
8, 4, 3
7, 7, 1
7, 6, 2
7, 5, 3
7, 4, 4
6, 6, 3
6, 5, 4

Thus,

> A crib will therefore, other things being equal, be more likely if it requires a slide of 15 than if it requires a slide of 25. The problem is to make the best use of this principle, by determining the probability of the crib with reasonable accuracy, *but without spending long over it*. [14, p. 17, emphasis added]

Noting the phrase "without spending long over it," it is once again seen that the concern is not theoretical purity but practical utility.

To use the Bayesian approach, one needs to know the probability of the different slides. In modern notation and terminology, if X_1, X_2, and X_3 are three independent random variables having a discrete uniform distribution on the integers 0, 1,…,9, find the distribution of $S = X_1 + X_2 + X_3$; that is, compute

$$P(S = k), \quad k = 0, 1, \ldots, 27.$$

Turing mentions three methods to do this. One is what would be termed the *Monte Carlo* approach: produce a long stretch of key and tally the number of slides. Another is the *brute force* method: there are 1,000 possible slide combinations; just add, find the remainder dividing by 26, and again tally the number of slides. The third is to use *mathematics*, which is very attractive from a mathematical standpoint. Consider the polynomial

$$f(x) = (1 + x + x^2 + \cdots + x^9)^3.$$

Each partition of $n = i + j + k$ into a sum of three terms, each summand permitted to take values between 0 and 9, corresponds to a term $x^i x^j x^k$ in the expansion of $f(x)$. Since $x^i x^j x^k = x^n$, it follows that the number of different ways of partitioning n corresponds to the coefficient of x^n in $f(x)$.

There is a very simple trick for computing this coefficient. Noting that $(1 - x)(1 + x + \ldots + x^9) = (1 - x^{10})$, it follows that

$$f(x) = \frac{(1 - x^{10})^3}{(1 - x)^3} = (1 - 3x^{10} + 3x^{20} - x^{30})(1 - x)^{-3}.$$

Using the general form of the binomial theorem to expand

$$(1 - x)^{-3} = 1 + 3x + 6x^2 + 10x^3 + 15x^4 + \cdots$$

and multiplying, gives

$$f(x) = 1 + 3x + 6x^2 + 10x^3 + 15x^4 + \cdots + 3x^{26} + x^{27}.$$

The full list of coefficients is

1 3 6 10 15 21 28 36 45 55 63 69 73 75 75 73 69 63 55 45 36 28 21 15 10 6 3 1;

something that can be generated nowadays instantaneously using just a few lines of code in one's favorite computer language.

It is now a straightforward matter to compute the decibanages associated with a candidate crib. For example, consider an enciphered message beginning *MVHWUSXOWBVMMK* and the candidate crib *AMBASSADOR*. If the crib is correct, then the associated slides are 12, 9, 6, 22, 0, 23, 11, 14, and the associated deciban score is -33, a very poor fit indeed. (In fact, the cipher text letters had been chosen at random and had no relation to the crib.) If, on the other hand, the message was *NYXLNXIQHH*, the score in this case is 15, so if the initial odds were 2:1 against (say), then the final odds are almost 3:1 in favor of the crib.

2.2.1. *Using a Crib*

At the end of this section, Turing illustrates one possible use of a crib. Recall the overall slide is a sum $r+s+t$ of three components, r, s, and t having periods 91, 95, and 99, respectively. Suppose for each initial setting of the period 95 and period 99 components (that is, where in each period the initial slides s_0 and t_0 are), one constructs a catalog of the sequence of resulting sums $s+t$ (s_0+t_0, s_1+t_1, $s_2+t_2, \ldots$). One then constructs a catalog having $95 \cdot 99 = 9,405$ lines, one for each (s_0, t_0) pair, each line containing an initial segment of the sequence $s_0+t_0, s_1+t_1, s_2+t_2, \ldots$. The catalog is ordered so as to facilitate looking up the $s+t$ sequences rather than the (s_0, t_0) pairs.

If the crib is correct, then we know the total slide $r+s+t$, but not the individual components r, s, and t. For each letter in the crib, there are 91 hypotheses about the value of r, each resulting in a sequence of inferred values of $s+t$. These sequences are then looked up in the catalog. If the crib is indeed correct, then we will learn the value of r_0, s_0, and t_0 and can then generate the sequence of total slides $r+s+t$.

But why bother scoring the crib in the first place? "This process is fairly lengthy, and as the scoring of the crib takes only a minute it is certainly worth doing" (p. 21).

2.3. *Theory of Repeats*

The letter subtractor system just discussed has the advantage over the Vigenère system that the length of the key exceeds the length of the message, but the disadvantage that the sum of the slides (the numbers 0, 1,..., 25) has a nonuniform distribution. Suppose instead that one has a long series of substitutions not suffering from this later defect. In such cases, it is often useful for purposes of cryptanalysis to identify "depths," that is, two or more messages part or all of which have been enciphered using the same key stream (that is, the same series of substitutions). (Turing does not use the term "depth" but says the messages "fit.") In his next example, Turing discusses how to identify depths using the Bayesian approach.

The classical approach to this problem, using, say, Friedman's index of coincidence, exploits the fact that two such series, correctly superimposed, will exhibit a higher proportion of "repeats" in each series than otherwise expected. Thus, if the two streams are generated at random, then at any one position, there is a $(1/26)^2$ probability of finding an A in both series at that position, and similarly for B through Z, for a total probability of $26(1/26)^2 = 1/26$.

On the other hand, if the underlying plaintext is modeled as an independent sequence of letters, with the letter α occurring with frequency p_α, then the probability of a repeat is $\beta = \sum_{\alpha} p_{\alpha}^2$, which is always greater than $1/26$ (unless all $p_{\alpha} = 1/26$, which never happens in ordinary plaintext).

The problem is how to do better than this if a depth is present. Turing writes:

> One writes out the cipher texts of the two messages with the letters which are thought to have been enciphered with the same substitution under one another. One then writes under these messages a series of letters *o* and *x*, an *o* being written where the cipher texts differ and an *x* where they agree. The series of letters *o* and *x* will begin where the second message begins and end where the first to end ends. This series of letters *o* and *x* may be called the repetition figure. It may be completed by adding at the ends an indication of how many letters there are which do not overlap, and which message they belong to. [14, p. 22]

For example, the repetition figure

$$^{8}xooooooooooxooxxooxxxxxxooooooooooxox^{11}$$

indicates that the depth begins at the 9th letter of the first message, continues for the next 37 letters (during which time there are 12 repeats), and then the 2nd message continues on for 11 more letters.

2.3.1. First Simplified Form of Theory

Suppose, as in the classical Friedman index of coincidence approach, that the letters are regarded as a sequence of independent outcomes, so that "we neglect all evidence except the number of letters *x* and the number of letters *o*." Thus, suppose that in a repetition figure of length N, there are n occurrences of x and the probability of a repeat at any position is $\beta = \sum_{\alpha} p_{\alpha}^2$. In that case, the factor in favor of the fit is

$$(26\beta)^n \left[\frac{26}{25}(1-\beta)\right]^{N-n}.$$

Turing derives this result in two different ways, one of which follows.

The factor in favor of a fit is just the ratio of the probability of seeing the pattern, given a repeat rate of $p = \beta$ versus a rate of $p = 1/2$. These are each of the form $R(N, n)Q(N, n)$, where

$$R(N,n) = \frac{N!}{n!(N-n)!}$$

is a binomial coefficient that counts the number of different possible patterns having exactly n repeats out of the total of N (note $R(N, n)$ does *not* depend on p), and

$$Q(N,n) = p^n(1-p)^{N-n}$$

is the probability of seeing any *specific* pattern of n repeats.

For example, suppose the repetition figure is *oxoxx*. Then $N = 5$, $n = 3$, and

$$R(7, 5) = \frac{5!}{3!2!} = 10;$$

there are ten possible patterns with three repeats out of five. These are

ooxxx, *oxoxx*, *oxxox*, *oxxxo*, *xooxx*, *xoxox*, *xoxxo*, *xxoox*, *xxoxo*, *xxxooo*;

each has probability $Q(5,\ 3) = p^3(1-p)^2$.

Suppose the data D is that one observes a repetition figure with n repeats out of N and the repeat rate β. Then the factor in favor of a fit is

$$\frac{P(D|p = \beta)}{P(D|p = 1/26)} = \frac{R(N, n)\beta^n(1-\beta)^{N-n}}{R(N, n)\left(\frac{1}{26}\right)^n\left(\frac{25}{26}\right)^{N-n}}.$$

Turing comments:

> The device of assuming, as we have done here, that the evidence which is not available is irrelevant can often be used and usually leads to good results. It is of course not supposed that the evidence really is irrelevant, but only that the error resulting from this assumption when used in this kind of way is likely to be small. [14, p. 26]

2.3.2 Second Simplified Form of Theory

Suppose the available evidence is that there is a sequence of r contiguous repeats (such as *oxxxo* or *oxxxxxo*) in some part of the repetition figure. The point here is that such an extended sequence if at all long is very unlikely to occur by chance and therefore provides strong evidence in favor of a correct fit. For example, the word "CONVOY" might occur in both messages at the same point (or "Heil Hitler," or "Obersturmbannfuehrer," or ...).

It is instructive to compare the different competing statistical approaches here. In a classical test of significance, one computes the probability of the data given a "null hypothesis," for example, that the fit is incorrect. (The null hypothesis typically represents the skeptical position that some state of affairs does not in fact obtain.) If this "level of significance" or "P-value" is sufficiently small, this is taken as evidence against the null. Thus, one would compute the probability of six repeats,

$$\left(\frac{1}{26}\right)^6 = 0.0000000032.$$

This is obviously very small, but the problem here is that even under the competing alternative of a correct fit, the probability of a hexagram will still be quite small. For example, in the case of the Naval Enigma, the repeat rate was roughly 1 in 17, giving

$$\left(\frac{1}{17}\right)^6 = 0.000000041.$$

This too is very small!

The point is not the *absolute* magnitudes of $P(D|H_0)$ and $P(D|H_1)$, but their *relative* magnitudes or ratio:

$$L = \frac{P(D|H_1)}{P(D|H_0)} = \frac{0.0000000414}{0.0000000032} = 12.798\cdots.$$

Thus, the observed data is nearly 13 times as likely to occur given a correct fit versus an incorrect one. (This corresponds to a value of 11 decibans.)

In the classical Neyman–Pearson theory of hypothesis testing in statistics (developed in a series of papers from 1928 to 1938 by Jerzy Neyman and Egon Pearson), this problem would be viewed as one of deciding between two "simple" hypotheses, say H_0 and H_1. There are two possible errors in that situation, corresponding to accepting H_1 when H_0 holds (a "type 1 error") and accepting H_0 when H_1 holds (a "type 2" error). Each of these errors has an associated probability: $P(\text{reject}\, H_0|H_0)$ and $P(\text{reject}\, H_1|H_1)$. The so-called *Neyman–Pearson lemma* states that in such cases, for any fixed probability of type 1 error, one can minimize the occurrence of a type 2 error by choosing an appropriate cutoff c and rejecting H_0 ($=$ accept H_1) whenever the likelihood ratio exceeds that cutoff: $L > c$. (A paradigm for this might be some type of acceptance or rejection procedure in industrial quality control.)

Such an approach would be completely useless here, because *it entirely neglects the prior odds* for or against the correctness of the fit. For example, depending on the circumstances, one might have more or less reason to believe the messages were in depth and, depending on the length of the message, more or less reason to think they were properly aligned. We will return to this point later.

2.3.3. General Form of Theory

The approach just discussed (based on r-grams) illustrates the inefficiency of the classical attack based on the index of coincidence: it ignores relevant and useful information. One needs instead a more detailed statistical model, one for which the factor in favor of a fit can be both mathematically derived and readily computed. As Turing notes,

> It is not of course possible to have statistics of every conceivable repetition figure. We must make some assumption to reduce the variety that need to be considered. The following assumption is theoretically very convenient, and also appears to be a good approximation.
>
> *The probabilities of repeats at two points known to be separated by a point where there is known to be no repeat are independent.*
>
> We may also assume that the probability of a repeat is independent of anything but the repetition figure in its neighborhood.... We can therefore think of a repetition figure as being produced by selecting the symbols of the figure consecutively, the probability of getting an x at each stage being determined by the repetition figure from the point in question back as far as the last o. [14, p. 27]

In the case of the leftmost x, it may not be preceded by an o, in which case it is noted whether there is any preceding text or not. Table 5 lists the possible cases, with

Table 5. State space and probabilities for general mode

o	a_0	Some	b_0	None	c_0
ox	a_1	Some *x*	b_1	None *x*	c_1
oxx	a_2	Some *xx*	b_2	None *xx*	c_2
oxxx	a_3	Some *xxx*	b_3	None *xxx*	c_3
...		...		...	

Turing's notation for the corresponding probability of seeing an x immediately after. For example, consider the repetition figure

$$\text{none } xxxxo|o|o|xo|xxxo|o|xx| \text{ some},$$

where the vertical bars denote where a new block is computed (immediately after an o). The factor in favor of the fit may be broken down into a product of factors corresponding to each block:

$$\frac{c_0c_1c_2c_3(1-c_4)}{\left(\frac{1}{26}\right)^4\frac{25}{26}} \cdot \left[\frac{1-a_0}{\frac{25}{26}}\right]^3 \cdot \frac{a_0(1-a_1)}{\frac{1}{26}\cdot\frac{25}{26}} \cdot \frac{a_0a_1a_2(1-a_3)}{\left(\frac{1}{26}\right)^3\frac{25}{26}} \cdot \frac{a_0a_1}{\left(\frac{1}{26}\right)^2}.$$

In general, if one has a message having an r-gram of the a-type, and $k_0 = 1 - a_0$, $k_{r+1} = a_0a_1 \ldots a_r(1 - a_{r+1})$, then the appropriate decibanage is

$$\mu_r = 10\log_{10}\left(\frac{26^{r+1}k_r}{25}\right) - (r+1)10\log_{10}\frac{26(1-a_0)}{25}.$$

2.3.4. *Actual Versus Apparent r-Gram Repeats*

If h is the probability of an o and $L = N(N-1)/2$, then one has the natural estimate $k_r \approx N_r/Lh$.

The statistics N_r, the "actual" numbers of r-gram repeats, can be tedious to tally. Turing discusses how these can be computed from quantities M_r, the "apparent" number of of r-gram repeats, which require less labor. The discussion of this in [14] is only cursory, and [15] helps to understand what is happening.

Consider the case of tetragrams: there are $26^4 = 456{,}976$ of these, ranging from *AAAA*, *AAAB*, ..., to *ZZZZ*. Turing considers the example of EINS (German for "one"), the most common tetragram in many forms of German military traffic, such as that encrypted by the Enigma.

Now in the case of a genuine tetragram repeat (as opposed to a pentagram, hexagram, or in general r-gram repeat, $r \geq 5$), precisely four letters match, and the letters immediately before and the letters immediately after do not (for example, as in as QEINSR, VEINSW). Turing calls this an "actual repeat" of EINS; the grand total of all such repeats for all possible tetragrams is the "*actual* number of tetragram repeats." For any $r \geq 0$, N_r denotes the actual number of r-gram repeats.

Because N_r is not easily computed, Turing considers instead the "apparent number of repeats" M_r and shows how N_r can be obtained from M_r. An apparent r-repeat is a repeat of length r that may in fact form part of a longer s-gram repeat

($s > r$). For example, KLEINSORGE and KLEINSATZ are examples of an actual hexagram repeat of KLEINS but also an apparent tetragram repeat of EINS. In general, there is one apparent r-gram repeat for every actual r-gram repeat, two apparent r-gram repeats for every actual $r+1$-gram repeat, three apparent r-gram repeats for every actual $r+2$-gram repeat, and so on. Thus,

$$M_r = N_r + 2N_{r+1} + 3N_{r+2} + \cdots;$$

hence,

$$M_r - M_{r+1} = N_r + N_{r+1} + N_{r+2} + \cdots$$

and

$$N_r = (M_r - M_{r+1}) - (M_{r+1} - M_{r+2}) = M_r - 2M_{r+1} + M_{r+2}.$$

Thus, the computation of N_r can be reduced to that of the M_r; these can in turn be tallied in a fairly simple way:

> It is therefore sufficient to calculate only apparent numbers and to carry these two stages further than we want to go with the actual numbers. [That is, to find N_r, you need to find M_{r+2}.] In practice octagram repeats are so certain to be right that it will be sufficient to have statistics only as far as heptagrams. We therefore need statistics only as far as 9-grams. To get these numbers of apparent repeats it is sufficient to take all the 9-grams in the material (i.e., on the circle) and put them into alphabetical order. This can be done very conveniently by Hollerith [a punch-card sorting device]. The number of trigram repeats say can then be found very simply (although with a good deal of labour) by considering only the first three letters of each 9-gram. [15, p. V1].

Note: Turing's other paper, "Paper on the Statistics of Repetitions" [15], largely repeats the material in this section of [14], although some of the formulas and mathematics are different.

The mathematical theory was later developed and published in the open literature by Good [7], who describes it as a theory of "regenerative Markov chains." Good credits Turing (on p. 936) with the actual versus apparent distinction, and says a special case of the models had been invented by Turing "for the analysis of certain binary processes" (!). At Alexander's suggestion, Good derived a more general form of scoring (which Alexander humorously referred to as ROMSing, for the Resources of Modern Science [4, p. 157]).

2.4. Transposition Ciphers

In a transposition cipher the order of the letters in a message is changed, but not the letters themselves. In this case the simplifying assumption that the plaintext letters are chosen independently according to some set of frequencies "would be useless or worse than useless, for it would result in the conclusion that all transpositions were equally likely" (p. 32). That is, if H represents a possible transposition, and E the letter frequencies in ciphertext, then $P(E|H)$ is independent of H (because the implied letter frequencies in the plaintext are the same for all transpositions), and so the odds in favor of H are unchanged. It is precisely because $P(E|H)$ varies from one H to another that E furnishes evidence regarding H; if $P(E|H)$ is independent of H, then E is uninformative.

Instead, Turing assumes that successive letters form a Markov chain. In such a model the frequency $p_{\alpha\beta}$ of a bigram $\alpha\beta$ is not the product $p_\alpha p_\beta$ of the individual letters frequencies. (Indeed historically one of the earliest applications of such models, by Markov himself, was to the analysis of letter frequencies in text.) The approximate score in this case is simple, and it will simplify matters if we reverse Turing's order and consider his example first, and only then his mathematics.

2.4.1. Turing's Example

In a simple columnar transposition the plaintext is written down in rows of a given width; and the resulting columns are then permuted. The ciphertext is then written out going down successive columns from left to right. The width and permutation constitute the key. Turing gives as an example the following ciphertext encrypted using this method:

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

There are 95 letters in the message, and we are told the maximum possible key length is 15. It follows that there are between 2 and 15 columns having between 2! and 15! possible orders, for a total of 1, 401, 602, 636, 312 orders in all. Clearly a brute force attack on this message, despite the simplicity of its encryption, is out of the question.

The greater the key length, the shorter the columns. In the case of the maximum width of 15, there will be 7 rows, the last row having 5 letters; the first 5 columns will therefore have 7 letters, and the last 10 columns 6 letters. Thus the first 6 letters of the ciphertext, SATPTW, must be the first six letters of some column; and similarly the last six letters of the ciphertext, HUANWR, must be the last six letters of some column.

The statistical attack in this case compares the first 6 letters SATPTW (say), with every consecutive series of 6 letters, noting the resulting bigrams; when correctly aligned, SATPTW will be juxtaposed with the 6 letters in the column of the plaintext coming either immediately before or after it. When this happens the resulting bigrams will exhibit the roughness found in plaintext. For example, comparing SATPTW with FASTAU (starting at the 8th letter), and using SATPTW as the first column, one sees the 6 bigrams SF, AA, TS, PT, TA, WU. The natural score in half-decibans is then

$$20\log_{10}\frac{p_{SF}}{p_S p_F} + 20\log_{10}\frac{p_{AA}}{p_A p_A} + \cdots + 20\log_{10}\frac{p_{WU}}{p_W p_U}.$$

(If H is the hypothesis that the matching is correct, then $P(\alpha\beta|H) = p_{\alpha\beta}$; if $\overline{H}$ that the matching is incorrect, then $P(\alpha\beta|\bar{H}) = p_{\alpha}p_{\beta}$. Taking the sum of the individual bigram scores assumes the bigrams are at least approximately independent; the justification for this is the whole point of Turing's mathematical analysis.)

This computation is then carried out for every series of 6; and the high scoring series noted. The same process is then repeated using SATPTW as the second column in the comparison, and similarly using HUANWR for the first and second column. If an adjacent column is successfully identified this way, it can in turn be used as the basis for identifying yet another column, and so on.

Turing's Figure 6 (on p. 34) gives the bigram scores "for a certain kind of German traffic;" using these, his Figure 7 illustrates the scoring for all possible comparisons, the correct matchings being circled. Unfortunately the example itself is not particularly clean. For example, using SATPTW as first column, the correct match has a score of −4, but a number of incorrect comparisons in fact have positive scores, the highest being 36. Turing says of this highest scoring but incorrect match: "It was not difficult to see that this one was wrong as most of the score came from WO which requires Z to precede it, and there was no Z in the message". The explanation of this cryptic comment is presumably that in German military traffic ZWO (German for "two") was very common; but if so Turing's "requires" seems to overstate matters. Turing says one of the columns in Figure 7 gives the scores using HUANWR as the second column in the comparison, but in the margin there is the pencilled-in comment "I doubt it, S. W.;" and indeed, it is difficult to replicate some of the numbers given. (GCHQ in their press release suggest that "S. W." stands for Shaun Wylie, who also worked in Hut 8 in 1941 and was later Chief Mathematician in their postwar organization.)

Challenge: Turing does not give the actual underlying plaintex. Find it!

2.4.2. Markov Chain Model

Turing adopts as an alternative model to independence that the successive letters form what is today termed a Markov chain; that is, "the letters forming the plain language are chosen consecutively, the probability of getting a particular letter depending only on what the letter is and what the preceding letter was" [14, p. 32]. The subsequent mathematics are straightforward. If p_α is the frequency of the letter α, and $p_{\alpha\beta}$ is the probability of the bigram $\alpha\beta$, then (using the formula for conditional probability discussed earlier) $q_{\alpha\beta}$, the probability of seeing a β given the immediately preceding letter is α, is $q_{\alpha\beta} = p_{\alpha\beta}/p_\alpha$. In the theory of Markov chains many other quantities of interest may be computed from p_α and $q_{\alpha\beta}$, which are called the *initial distribution* and *transition matrix*, respectively, of the chain.

For example, for a stretch of plaintext of length L, say $\alpha_1, \alpha_2, \ldots, \alpha_L$, the probability of seeing the sequence of letters is

$$J(\alpha_1,\ \alpha_2, \ldots, \alpha_L) := p_{\alpha_1} \cdot q_{\alpha_1\alpha_2} \cdot q_{\alpha_2\alpha_3} \cdot q_{\alpha_3\alpha_4} \cdots q_{\alpha_{L-1}\alpha_L}.$$

By summing J over those sequences having given letters at certain places and not in others, one can compute the probability of seeing specified letters at other places. For example, if the data is that the known letters are

$$\underset{n_1\ dots}{\cdots}\ \beta_{n_1}\ \underset{n_2\ dots}{\cdots}\ \beta_{n_2}\ \cdots\ \cdots \beta_{r-1}\ \underset{n_r\ dots}{\cdots}\ \beta_r \cdots,$$

then, as Turing notes, this quantity is approximately

$$\prod_r p_{\beta_r} \cdot \prod_{n_{r+1}=0} \frac{p_{\beta_r \beta_{r+1}}}{p_{\beta_r} p_{\beta_{r+1}}}. \tag{A}$$

(The notation $n_{r+1} = 0$ means that the second product is over all r such that $n_{r+1} = 0$; that is, for all letters β_r for which the next known letter is immediately adjacent.)

If *none* of the known letters are adjacent in plaintext, this reduces to just the product $\prod_r p_{\beta_r}$, that is independence. It follows that the factor in favor of a candidate transposition being correct versus incorrect is

$$\prod_{n_{r+1}=0} \frac{p_{\beta_r \beta_{r+1}}}{p_{\beta_r} p_{\beta_{r+1}}}.$$

Turing's says his derivation of (A) is "something of a digression," but it is an interesting one.

2.4.3. Derivation

Let $\tau_{n,\,\alpha\beta}$ denote the probability, in state α, of a transition to state β after n intervening steps ($n \geq 0$). If $Q = (q_{\alpha\beta})$ is the matrix of one-step $\alpha \rightarrow \beta$ transition probabilities, and, in general, M_{ij} denotes the ijth entry of a matrix M (so that here $Q_{\alpha\beta} = q_{\alpha\beta}$), then it is a basic result in Markov chain theory that

$$\tau_{n,\,\alpha\beta} = (Q^{n+1})_{\alpha\beta};$$

that is, the probability of an $\alpha \rightarrow \beta$ transition after $(n + 1)$ steps is just the α, β entry of Q^{n+1}, the natural power of the single-step transition matrix Q.

Turing notes that (A) would hold exactly if one had $(Q^{n+1})_{\alpha\beta} = p_\beta$ for $n > 0$. Although this is not generally true, it is true ("except for very special values for $q_{\alpha\beta}$") that $(Q^{n+1})_{\alpha\beta} \rightarrow p_\beta$ as $n \rightarrow \infty$, "and this convergence is rather rapid" (p. 38). This last point is very important, ensuring the approximation should be a reasonable one.

Today this result is called the *ergodic theorem* for Markov chains; this states that under appropriate conditions, such chains converge to a unique *stationary distribution* π for the chain. That is, if Q is the matrix of one-step transition probabilities, then

$$\lim_{n\rightarrow\infty} Q^n(a, b) = \pi(b),$$

so that irrespective of where you start (the a), the probability of being in state b eventually stabilizes to a value $\pi(b)$; the resulting distribution π is "stationary" in the sense that it is the (in this case unique) probability distribution π on the states such that $\pi Q = \pi$ (which means the distribution is unchanged from one transition to the next). If you think of the successive states as describing some system in statistical mechanics, then this says that there is a equilibrium distribution π and that a system not in equilibrium converges to this equilibrium condition over time.

In the case of a finite number of states the ergodic theorem in this context is a special case of the *Perron–Frobenius theorem* for matrices. Specifically, given a matrix Q, the asserted convergence occurs if Q has nonnegative entries and if, for some n, Q^n has all positive entries. (In Turing's setting, the later corresponds to the innocuous assumption that for some n there is a positive probability of seeing any n-step $\alpha \rightarrow \beta$ transition.) For discussion of the history of the Perron–Frobenius theorem and its connection to the earlier work of Markov, see [12].

Turing proceeds to show that $\lim_{n\to\infty} Q^n(a,b) = p_\beta$, In modern terminology, Turing proves the ergodic theorem under some appropriate set of conditions. It is unclear if he was aware of this fact, or was just content to derive it for himself. (Perron's version of the theorem could certainly be found in some books of the era.)

Turing's proof, in brief outline, is as follows: assume the eigenvalues of the matrix Q have distinct moduli (absolute value). (This assumption is unnecessary; making it may be why Turing refers to his proof as "more or less rigorous.") It is then a standard result of linear algebra that one can "diagonalize"' Q: that is, find another matrix U (for "unitary") having determinant one, and such that $U^{-1}QU$ is diagonal:

$$M := U^{-1}QU = \begin{pmatrix} \mu_1 & 0 & \cdots & 0 \\ 0 & \mu_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mu_{26} \end{pmatrix}.$$

Turing shows after two pages of analysis that one of the eigenvalues μ_r equals 1 and that the rest satisfy $|\mu_r| < 1$. It follows immediately that $\lim_{n\to\infty} M^n = U^{-1} X_\sigma U$, where X_σ is a matrix "which has only one element different from 0, and that a 1 on the diagonal, say in position $\sigma\sigma$." Call this matrix Y; since it "is the one and only which satisfies" the conditions $YQ = Y$ (the stationarity condition!), $Y^2 = Y$, and $Y \neq 0$, it is necessarily the matrix whose $\alpha\beta$ coefficient is p_β.

It is apparent that the solution of a simple columnar transposition can be quite tedious. For a classical discussion of its cryptanalysis, see Sinkov [13, Chapter 5]. Sinkov states

> To assist in this step [the alignment of possible columns] we can record the frequency of every digraph [i.e., $p_{\alpha\beta}$] and assign the total of these frequencies as a score to each possibility being considered. The highest score will hopefully correspond to the correct juxtaposition. [13, p. 169]

Turing's procedure is both more efficient from a statistical perspective and more systematic that the one Sinkov described (as well as, of course, allowing one to factor in an estimate of the prior odds of correct alignment).

3. Discussion

Turing's paper is neither a general treatise on cryptology nor a detailed analysis of the cryptanalysis of a specific device, such as the Enigma. Instead, it introduces the use of the Bayesian method in cryptanalysis, illustrating it by a succession of increasingly complex examples.

It may, in fact, represent the first time Bayesian methods were discussed in the cryptological literature. The Bayesian methodology itself was of course well-known in the contemporary statistical literature of Turing's day, but sometimes a technique may be "well-known" in one field yet unknown in another. In such cases, just noting the utility of the method may be an important development.

3.1. The Contemporary Statistical Background

What is interesting and striking is that this happy resort to Bayesian methods by Turing would have been far less likely by someone having a detailed knowledge of and training in the modern statistical methods of the day. For paradoxically, the Bayesian methods that Turing found to be the perfect instrument for attacking German systems had been under sharp attack within the outside statistical profession for more than two decades.

For Turing, probability was *conditional* (relative to "certain evidence"), having both objective elements (the frequency of occurrence of an event) and subjective ones (our expectation about such frequencies). This was in contrast to the then dominant view in the statistical world, which regarded probabilities as facts about the world, manifested as frequencies in populations or repeated trials. In such a view, although judgment may enter into an analysis, it does not do so in a quantitative way.

The key issue was the nature of the prior odds used in Bayes's theorem, and more generally the nature of probability itself. Now it should be stressed that Bayes's theorem is entirely uncontroversial if the prior probabilities it requires are either frequencies or in some way objectively determined. But in the hands of those less skillful than Laplace, the method had been frequently abused, as in the notorious principle of insufficient reason.

The solution, some argued, was to recognize that probabilities were not subjective beliefs but objective facts about the world, the frequencies with which an event occurred in repeated trials. This was, to differing degrees, the view of both R. A. Fisher and Jerzy Neyman, perhaps the two most influential statisticians in England in the 1930s. (Fisher and Neyman in fact had very different views about the nature of statistical inference, but what is relevant here is that just about the only thing they did agree on was a total rejection of Bayesian methods!) For discussion of Fisher's views on inverse probability, see [17,18]; for a summary of Neyman's frequentist views, see [11].

3.2. Outside Influence and Impact

Cryptanalytic advances seldom have a technical impact on the "outside" world; signals intelligence agencies are virtually unique in being averse to advertising their successes, and one of the challenges for the historian of this area is that often key information is only released decades later. (Turing's paper is a case in point, of course, appearing some 70 years after it was written.)

But for somewhat unusual reasons, Turing's championing of Bayesian methods at Bletchley did have an almost immediate impact on the outside statistical profession. Fresh out of Bletchley Park in the fall of 1945, I. J. Good proceeded to write a book setting out the philosophy, mathematics, and application of the Bayesian approach that he had learned at Bletchley to a wide range of statistical and inferential problems. (The book, *Probability and the Weighing of Evidence* [8], was published in 1950; its cryptanalytic origins were of course discreetly omitted.) The book closely reflected Turing's views; in its preface, Turing is acknowledged as an influence, thanked for "illuminating conversations," and listed as one of three people who read and commented on the first draft of 1946. (The other two were Max Newman and Donald Mitchie of the Newmanry, where Good worked for the last two years of the war.)

During the next several decades, Good became a forceful advocate of the Bayesian viewpoint, publishing both papers in the statistical literature (some of which were elaborations of wartime statistical techniques due to Turing or Good, Turing's contribution being always carefully acknowledged) and in the philosophical literature (urging the superiority of the subjective, personalist theories of Ramsey, Savage, and de Finetti over the so-called objective frequentist theories of the time). For a summary of his philosophy four decades later, and an extensive list of references to many earlier papers, see [5].

There were other champions of the Bayesian view in the 1950s and later: initially most notably L. J. Savage at the University of Chicago and Howard Raiffa and Robert Schlaifer at the Harvard Business School. Despite their advocacy, it was not until perhaps the 1980s that Bayesian methods began to regain a considerable amount of professional respectability. Nor is this entirely surprising, given that a whole generation had been brought up viewing matters from the hostile perspective of Fisher and Neyman. Such a state of affairs must have been extraordinarily frustrating for Good, who for 30 years had to listen at conference after conference to speakers scornfully rejecting the Bayesian approach as just a lot of useless theory, having no real practical applications! (The author of this commentary once attended such a conference where Good was in attendance.) But bound as he was by his pledge of secrecy, Good had to remain silent about what can only be described as one of the outstanding statistical success stories of the 20th century. And even after 1974, when the successes of Bletchley Park became known, Good felt constrained not to reveal in detail just how central the Bayesian approach had been. (But the release of documents over the last decade, most notably the "General Report on Tunny" [9], now provides ample evidence about this.)

3.3. The Date and State of the Document

The document is undated, but internal evidence suggests a date of September 1941. The reference to Hitler being 52 on p. 1 indicates that the paper was written sometime between April 1941 and April 1942 (Hitler was born 20 April 1889); the reference to half-decibans suggests a date of no earlier than June 1941 (because they were introduced by I. J. Good, who arrived at Bletchley Park in May 1941); and the date of 12 September at the top of p. 11 further narrows this down to September of that year.

The paper may be incomplete, since it contains references to material not included. (For example, on p. 19, in the discussion of the letter subtractor problem, Turing refers to two additional methods besides the ones he has given, and says "They will be discussed later," but no such discussion appears.)

The posted manuscript has some notes and corrections, not all in Turing's hand, and it is unclear when these were made. (That is, whether made at the time Turing originally wrote the paper, or later by a reader.) For example, on p. 30, the correction k_{r+1} (in place of k_r) has inserted.

Acknowledgements

The author thanks Craig Bauer, Joe Buhler, Chris Christensen, Ralph Erskine, Al Hales, Jim Reeds, and Edward Simpson for many helpful comments on an earlier version of this article.

About the Author

Sandy Zabell is Professor of Mathematics and Statistics at Northwestern University. His research interests include mathematical probability, Bayesian statistics, legal applications of statistics (in particular forensic science), and the history of cryptology. His particular interest in Alan Turing goes back to a 1995 paper, “Alan Turing and the Central Limit Theorem” [16].

References

1. Banks, D. L. 1996. “ A Conversation with I. J. Good, ” *Statistical Science*, 11:1–19.
2. Bertrand, J. 1889. Calcul des probabilité s, (2nd ed.; reprinted 1972, Chelsea, New York). Paris: Gauthier-Villars.
3. Good, I. J. 2006. “From Hut 8 to the Newmanry.” In *Colossus: The Secrets of Bletchley Park’s Codebreaking Computers*, edited by B. J. Copeland. Oxford: Oxford University Press, pp. 204–222.
4. Good, I. J. 1994. “Enigma and Fish.” In *Codebreakers: The Inside Story of Bletchley Park* edited by F. H. Hinsley and A. Stripp. Oxford: Oxford University Press, pp. 149–166.
5. Good, I. J. 1988. “The Interface Between Statistics and Philosophy of Science,” *Statistical Science*, 3:386–412.
6. Good, I. J. 1979. “ Studies in the History of Probability and Statistics: XXXVII A. M. Turing’s Statistical Work in World War II, ” *Biometrika*, 66:393–396.
7. Good, I. J. 1973. “The Joint Probability Generating Function for Run-Lengths in Regenerative Binary Markov Chains, with Applications,” *Annals of Statistics*, 1:933–939.
8. Good, I. J. 1950. *Probability and the Weighing of Evidence*. London: Charles Griffin.
9. Good, I. J., D. Michie, and G. Timms. 1945. General Report on Tunny with Emphasis on Statistical Methods. Internal GCCS History. UK National Archives, HW 25/4 and HW 25/5
10. Lehmann, E. L. 2011. *Fisher, Neyman, and the Creation of Classical Statistics.* Springer: New York.
11. Neyman, J. 1977. “Frequentist Probability and Frequentist Statistics,” *Synthese*, 36:97–131.
12. Schneider, H. 1977. “The Concepts of Irreducibility and Full Indecomposability of a Matrix in the Works of Frobenius, König and Markov, ” *Linear Algebra and its Applications*, 18:139–162.
13. Sinkov, A. 1968. Elementary Cryptanalysis: A Mathematical Approach. In *New Mathematical Library*, Vol. 22, Random House.
14. Turing, A. M. 2012. “ The Applications of Probability to Cryptography, ” Unpublished paper, c. 1941, UK National Archives, HW 25/37.
15. Turing, A. M. 2012. Paper on the Statistics of Repetitions. Unpublished paper, c. 1941. UK National Archives, HW 25/38.
16. Zabell, S. 1995. “ Alan Turing and the Central Limit Theorem, ” *The American Mathematical Monthly*, 102:48–494.
17. Zabell, S. 1992. “R. A. Fisher and the Fiducial Argument,” *Statistical Science*, 7:369–387.
18. Zabell, S. 1989. “R. A. Fisher on the History of Inverse Probability,” *Statistical Science*, 3:247–263.