REF ID:A64688

CAUTION: THESE RECORDS WILL BE USED FOR OFFICIAL PURPOSES ONLY, DO NOT REMOVE PAPERS NOR REVEAL CONTENTS TO UNAUTHORIZED PERSON(S)

RECORDS CHARGE-OUT

10205

DATE OF REQUEST	SUSPENSE DATE
25 Jan 61	10 Feb 61

FILE OR SERIAL NUMBER AND SUBJECT	From File of Special Consultant (Friedman) Statistics for Cryptology ~~Confidential~~
TO	NAME AND EXTENSION OF PERSON REQUESTING FILE: Mr. William Friedman LI 6-8520 — ORGANIZATION, BUILDING, AND ROOM NUMBER: 310 2nd. Str., SE, Wash., D. C.
RETURN TO	Mrs. Christian, AG-24, NSA, Ft. Geo. G. Meade, Md. — DATE RET'ND.: — INITIAL HERE:
INSTRUCTIONS	WHEN TRANSFERRING FILE TO ANOTHER PERSON, COMPLETE SELF-ADDRESSED TRANSFER COUPON BELOW, DETACH, STITCH TO BLANK LETTER-SIZE PAPER AND PLACE IN OUT-GOING MAIL SERVICE.

10205

2ND TRANSFER COUPON

TO:

FILE (serial number and subject)

TRANSFERRED TO: (name and extension)

ORGANIZATION, BUILDING, AND ROOM NUMBER

DATE	(sig)	(ext.)

~~CONFIDENTIAL~~

~~Modified Handling Authorized~~

STATISTICS FOR CRYPTOLOGY

Word taken from
U.S. if it's none

~~CONFIDENTIAL~~

NATIONAL SECURITY AGENCY
WASHINGTON 25, D. C.

STATISTICS FOR CRYPTOLOGY

NATIONAL SECURITY AGENCY

WASHINGTON 25, D. C.

This edition of "Statistics for Cryptology" is published for use in training programs. Comments and suggestions for the improvement of this text are invited, and should be forwarded to the Director, National Security Agency (Attn: TNG).

PREFACE

My indebtedness to Dr. Kullback is acknowledged for his careful reading of the manuscript and his many suggestions. Credit is also due to my colleagues in NSA, too numerous to name individually, who have discovered some of this theory and have called much of the rest to my attention.

HOWARD H. CAMPAIGNE

TABLE OF CONTENTS

STATISTICS FOR CRYPTOLOGY

0. Introduction.

This treatise is a concise exposition of the mathematical statistics which has applications in cryptology. It is aimed at a hypothetical cryptanalyst who has had college mathematics and forgotten most of it, or perhaps has had none but has done fairly extensive study on his own. The object here is to give definitions and develop properties in a simple and straightforward way. This treatise is not intended to be selfcontained but is intended to be readable. Many concepts will be given without following them very far.

The theory held here is that the most formidable computation is relatively easy to understand if the reasons for it are clear. Therefore this exposition gives the basis for computational and statistical procedures, but does not dwell on their details, which are available elsewhere. The illustrative examples have been selected to have a minimum of computation, and are therefore trivial.

0, 1 Information Theory.

0, 1, 1 Coding.

Cryptology is concerned almost entirely with telecommunications, the transmission of information over long distances by means of radio or wire lines. In each method of telecommunication the information is coded in some way, such as Morse for hand sending, or amplitude or frequency modulation for speech. Of course in enciphered communications further complications are deliberately introduced, but the basic coding is unavoidable. It can usually be done in a variety of ways, depending upon whether brevity or reliability is more desirable.

0, 1, 1, 1 Efficient Coding.

The most common coding is binary, which we can represent on paper as 0's and 1's, that is, some signal can be on or off, and combinations of these on-off signal elements are used to represent elements of writing or of speech. One of the most used codes is the Baudot Code, in which each combination has five elements, and there are $2^5 = 32$ distinct combinations. These 32 combinations are used to represent the 26 letters of the alphabet and a few functions, such as carriage return, platen advance (line feed), word space, shift to upper case, etc.

By contrast the Morse code does not have a fixed number of elements for each letter, nor are the signal elements the same length. There are two conditions of the signal, on and off, and each of these is used in two versions, short (dot) and long (dash). Because of the latter, the two conditions of on and off must alternate.

Communicators encounter the need for sending messages rapidly, and ask themselves the questions: What coding conveys the most information in a given time? Which coding sends the information most reliably? Mr. Morse attempted to make his code efficient for English, for he assigned the short combinations to the frequent letters, dot for E and dash for T. He assigned the longest combinations to the least frequent letters, such as dot dash dash dash for J. There are even longer combinations for the numerals, comma, and so forth.

There is a principle of information theory which says that the more effectively a communication system is used the more it sounds like noise. Thus the Morse code will transmit a maximum of information when there are lots of E's and T's and only a few J's and Q's. The Baudot code will be at its best when the letters (including carriage returns, line feeds, and so forth) are

equally frequent. If the frequencies deviate from this then a code can be tailored to be more effective.

Making an effective code depends on knowing in advance the frequencies with which the letters (or words and phrases in the case of book codes, to which all of this applies) will occur. The effectiveness of the code will be no greater than the accuracy of the prediction of the statistics.

Given a frequency count there are some simple rules enabling us to construct an efficient code. Below is a frequency count and a binary code made to transmit rapidly sets of these symbols in these proportions.

	Frequency	*Code*	*Weighted Length*
A	.30	10	.60
B	.25	00	.50
C	.15	010	.45
D	.10	011	.30
E	.07	1100	.28
F	.05	1110	.20
G	.03	1101	.12
H	.02	11110	.10
I	.02	111110	.12
J	.01	111111	.06

The average length of combination is 2.73 binary digits; if a fixed length code were used it would take at least four signal elements per letter.

This is a Fano code, as originally proposed by Prof. R. M. Fano of the Massachusetts Institute of Technology. One way of constructing the code to fit the frequencies is to assign the first signal element for each letter, then the second for each, and so forth. Each signal element is a yes-no answer to a question, and will be efficient if the answers are equally likely. Therefore, at each step we want half of the code groups to be 0 and the other half to be 1. So we select at the first step a subset of the letters whose total frequency is as close to one-half as we can make it, assign 0 as the first signal element to these and 1 to the remaining. Then we split the subset into half and assign 0 to one-half and 1 to the other. We continue this until a letter is unique in its subset; after that there is no further need for signal elements, and the code for that letter is terminated. In this way frequent letters get short codes. A letter with n signal elements should have frequency about $\frac{1}{2^n}$.

The coding could be made for groups of letters, such as digraphs or words, instead of single letters. More efficiency can be achieved with larger units but at the expense of man-hours or equipment and delay in sending the signal. The English language is a code of this sort; frequent words tend to get shorter, as "automobile" becomes "auto", "television" becomes "TV", and so forth.

The Fano coded text is sent as a continuous stream of binary digits. It can be resolved unambiguously into its original meaning. In our example if a letter begins with 00 it is a B. If it begins with 01 then exactly one more signal element must be examined to determine whether the letter is a C or a D. If it begins with 10 it is an A. If it begins with 11 at least two more more signal elements must be examined. Thus at each signal element we know whether a letter has been determined or not, and if so then a new letter begins with the next element. Beginning from the start of a letter it can always be recognized correctly from ungarbled text.

Now if a signal element is incorrectly received at least one letter will be wrong. But worse, the beginning of the next letter may be obscured. But even so, after a few errors we will get

back in phase, and once in will stay in so long as the text received is correct. The reader can check this in the following example:

A J B A B E J F E D
10,111111,00,10,00,1100,111111,1110,1100,011

(The commas are introduced to simplify our examination of the stream; no spaces or other interruptions to the binary stream are transmitted.) If we suppose that exactly one of the first six binary digits is complemented we find that not more than three letters are garbled. The number of letters received is not necessarily the same as that sent. The number of plain letters garbled when a signal element is changed is not known in general, even in the statistical sense of knowing the expected number.

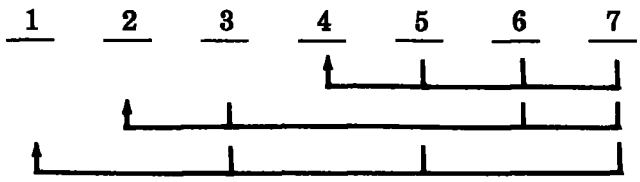
The same principles can be used in making a code even though the number of distinct signal elements is 3, 10, or 26, and the things represented by the code are digraphs, trigraphs, or words.

The efficiency of a Fano code is achieved by lowering the redundancy. This is shown in the way errors are multiplied. Later in this chapter we will discuss a way of measuring redundancy.

0, 1, 1, 2 Redundant Coding.

Instead of lowering the redundancy for efficiency we might raise it to detect errors or even correct them. A simple way to do this is the following. To the five element Baudot code adjoin a sixth element which will be 0 or 1 so that the number of 1's is even for each code combination. Now if a single element is changed in transmission the number of 1's will necessarily be odd and the occurrence of a garble will be recognizable. However, the correct value could not be reconstructed.

An error correcting code can be illustrated thus. Suppose we have a 4 binary digit code for the decimal digits, together with space, upper and lower shift, comma, period, and platen advance, 16 characters in all. Now add some more signal elements to each code combination. It will be convenient to make the new ones to be the first, second, and fourth in the group; the third, fifth, sixth, and seventh are assumed already there. The fourth is selected so that the sum of the last four signal elements is 0 mod 2. The second should be such that the sum of the 2nd, 3rd, 6th, and 7th is 0 mod 2. The first should be such that the sum of the odd numbered signal elements is 0 mod 2.



Now suppose that a single signal element has been reversed. At least one of these sums will no longer be 0. We look at each of the conditions in the order given. If the condition is satisfied put down a 0; if not put down a 1. The three binary digit number generated thus will be the position within the group of the incorrect element. For example, if the code group 1000001 arrived we look at the last four digits and add them mod 2. We get a 1, which shows that there is an error. We add the 2nd, 3rd, 6th, and 7th getting 1. We add the odd digits, getting 0. We now have generated the binary number 110 which is the number 6; the sixth digit is incorrect. The correct group was therefore 1000011. Th s kind of correctable code was first described by R. W. Hamming in the *Bell System Technical Journal*, April, 1950 (33)*.

* Numbers in parenthesis refer to the Bibliography.

The same pattern could be used to expand a four decimal digit code to make it single error correcting, but this time the sums would be 0 mod 10. As an example take the garbled group 8753708. Using the patterns given before one can find the digit to be corrected, noting only whether each sum is 0 or not 0 mod 10. The amount of correct on is easily found. In fact this correction is in this case over-determined, which shows that errors in two digits could be detected in some cases.

It is possible to make a double-error detecting or a double-error correcting code, or to make a code as reliable as one wishes by introducing enough redundancy. There is a theorem which says that by tolerating enough delay information can be sent as accurately as one chooses without decreasing the rate of sending. The delay arises because the information must be coded in large pieces.

0, 1, 2 Measuring Information.

We have seen that the efficiency can be raised by lowering the redundancy, or the reliability can be raised by increasing the redundancy. Can we measure the amount of redundancy contained in a message? We can, and in order to introduce the measure heuristically, we will give a preliminary discussion.

First, consider a yes-no question. How much information is conveyed by the answer? It clearly depends on the priori probability of the answers. Compare two questions, paraphrased from a civil service form, one of which is "Are you a male?", and the other "Do you use intoxicants to excess?" Answers to the first will be half "yes" and half "no". Answers to the second will be almost all "no". The first gives more information, of course; the second gives almost none. If the a priori probabilities are p and q, then a measure of the information will be a function H(p,q) of these probabilities. Furthermore H(½,½) has a special value, probably a maximum, and H(q,p) = H(p,q).

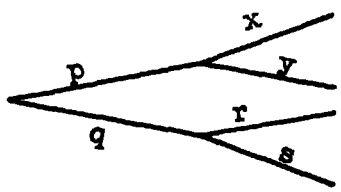
If the question has c possible values, such as a letter to be filled in on a form ("middle initial ....."; c=26), then $H(p_1,p_2, \ldots ,p_c)$ is a function of c variables. In general we would expect more information to be conveyed by more values, and in particular $H(\frac{1}{c},\frac{1}{c},\ldots,\frac{1}{c}) \geq H(\frac{1}{b},\frac{1}{b},\ldots,\frac{1}{b})$ for $c \geq b$.

We would like H to have the following properties (34):

(1) $H(\frac{1}{c},\frac{1}{c},\ldots,\frac{1}{c}) \geq H(\frac{1}{b},\frac{1}{b},\ldots,\frac{1}{b})$ for $c \geq b$.

(2) $H(p_1,p_2, \ldots , p_c)$ should be a continuous function of each p;

(3) If a decision can be made in two steps, then the information H should be the weighted sum of the information at each step. An example will make this clear. We want H(px,py,qr,qs) = H(p q) + pH(x,y) + qH(r s), where p+q=x+y=1=r+s.



The properties defined above are sufficient to uniquely define the function $H(p_1,p_2, \ldots, p_c)$, as follows:

Condition (3) implies that $H(\frac{1}{S^m}, \ldots, \frac{1}{S^m}) = mH(\frac{1}{S}, \ldots, \frac{1}{S})$. For instance, $H(\frac{1}{8}, \ldots, \frac{1}{8})$

$= H(\frac{1}{2},\frac{1}{2}) + \frac{1}{2}H(\frac{1}{4},\frac{1}{4}) + \frac{1}{2}H(\frac{1}{4},\frac{1}{4}) = \ldots = 3H(\frac{1}{2},\frac{1}{2})$. Proceeding, we set the integer $m = \log_2 t$

or $t = 2^m$. Then $H(\frac{1}{t}, \ldots, \frac{1}{t}) = mH(\frac{1}{2},\frac{1}{2}) = \log_2 t\, H(\frac{1}{2},\frac{1}{2})$. Let $H(\frac{1}{2},\frac{1}{2}) = K$.

Now consider a case where $p_i = \frac{r_i}{S}$, $S = \sum_{i=1}^{c} r_i$; that is, the probabilities are all rational. Consider this as choosing among S answers in two steps, using principle (3),

$$H(\frac{1}{S}, \ldots, \frac{1}{S}) = H(p_1, \ldots, p_c) + p_1 H(\frac{1}{r_1}, \ldots, \frac{1}{r_1}) + p_2 H(\frac{1}{r_2}, \ldots, \frac{1}{r_2}) + \ldots$$

$$+ p_c H(\frac{1}{r_c}, \ldots, \frac{1}{r_c}),$$

whence $H(p_1, \ldots, p_c) = H(\frac{1}{S}, \ldots, \frac{1}{S}) - p_1 H(\frac{1}{r_1}, \ldots, \frac{1}{r_1}) - \ldots - p_c H(\frac{1}{r_c}, \ldots, \frac{1}{r_c})$

$$= K \log_2 S - K p_1 \log_2 r_1 - K p_2 \log r_2 - \ldots - K p_c \log r_c$$

$$= (-p_1 \log \frac{r_1}{S} - p_2 \log \frac{r_2}{S} - \ldots - p_c \log \frac{r_c}{S})K.$$

If one of the r_j's $= 0 = p_j$ we agree that $p_j H(\frac{1}{r_j} \ldots,) = 0$.

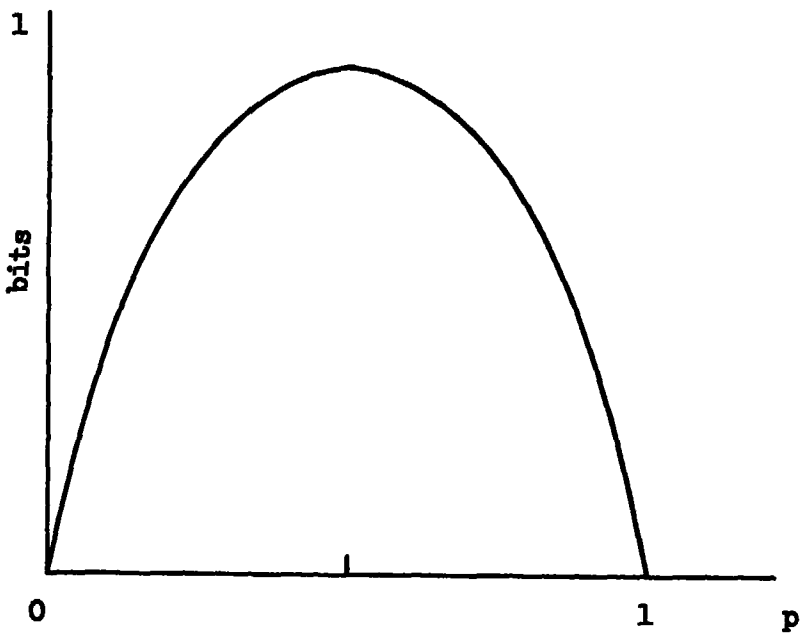
$$H(p_1,p_2, \ldots, p_c) = -(p_1 \log p_1 + p_2 \log p_2 + \ldots + p_c \log p_c)K.$$

If the p_i are irrational they can be approximated by rationals, and the assumption of continuity implies that $H(p_1, \ldots, p_c) = -K \sum_{i=1}^{c} p_i \log p_i$ still. The constant K is positive but arbitrary, a choice of unit. Let us agree to take $K = H(\frac{1}{2},\frac{1}{2}) = 1$, and call this unit a "bit" of information.

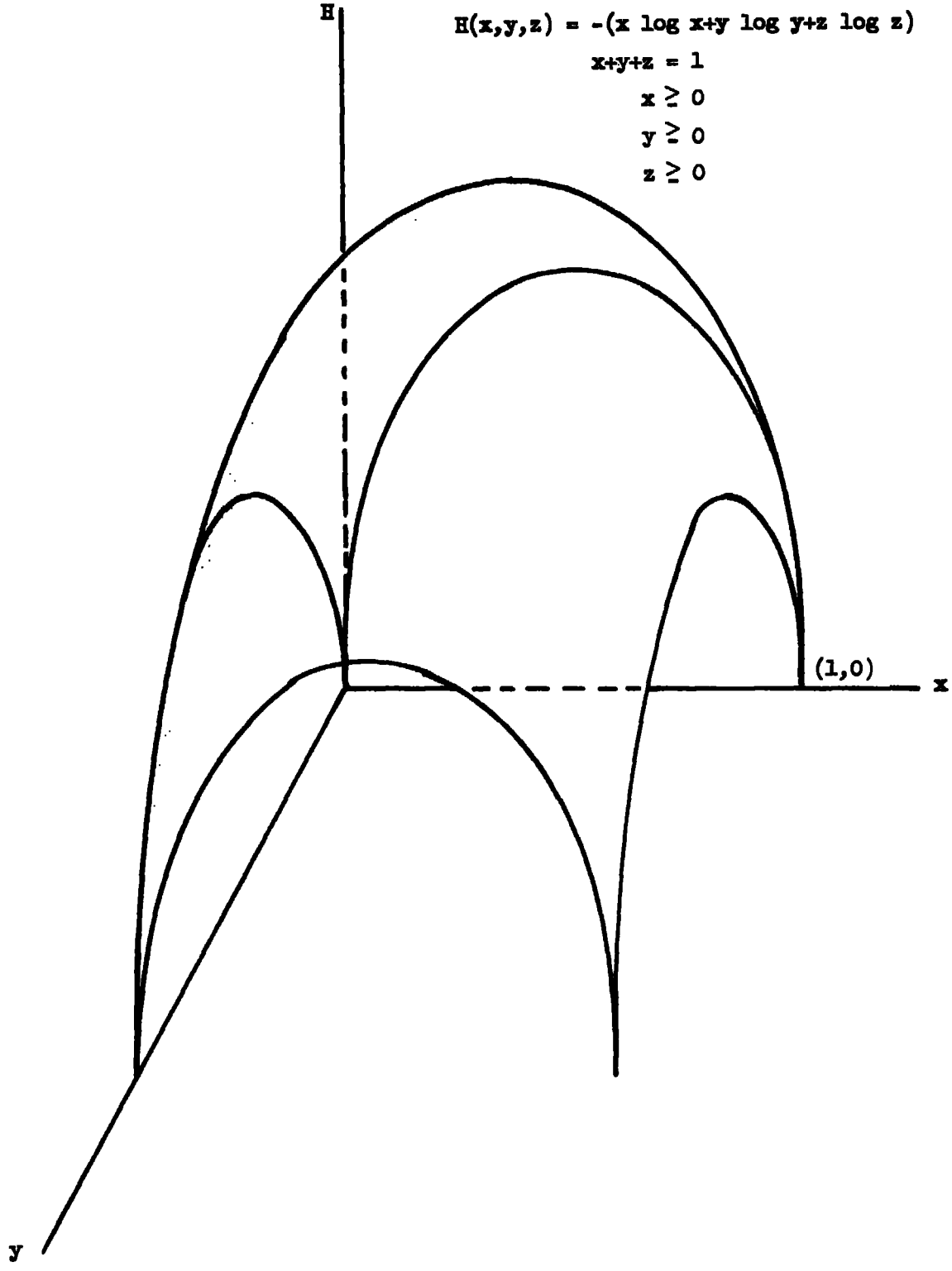
We now have a measure of the amount of information which can be transmitted under given conditions. Other measures might also exist, but this is the only one (aside from the choice of unit, the bit) which satisfies the three conditions given above. The reader is warned that this is a measure invented by a communicator, and in a sense measures the work of transmitting the data, or the capacity of a channel to carry information. It applies to a process or a channel, not to semantics. The function H gives a lower bound in binary signal elements on the abbreviation

which can be achieved by using a Fano or similar code. Purely flat random material will give a maximum value for H, even though it may have no semantic content.

The measure $H(p_1, p_2, \ldots, p_c) = -\sum_{i=1}^{c} p_i \log_2 p_i$ has been called the "entropy", since it resembles the function introduced by Boltzmann to measure the degree of disorganization in a physical system. It has many interesting properties. It is a symmetric function of its arguments $p_1, p_2, \ldots$. Its largest value occurs when the probabilities are all equal, $p_1 = p_2 = \ldots = p_c = \frac{1}{c}$, and then it is $H = \log_2 c$. It has the value 0 only if $p_1 = 1$ and $p_i = 0$, that is, the answer to the question is certain. Probabilities which are 0 can be disregarded, $H(p_1, p_2, \ldots, p_{c-1}, 0) = H(p_1, p_2, \ldots, p_{c-1})$.



Graph of $H(p,q)$, $p+q = 1$.



Shannon's Mosque

The entropy is a measure of the disorganization of a physical system, and H is a measure of the lack of pattern in a message. If we let $R = \frac{\log_2 c - H}{\log_2 c}$ we have a measure of the redundancy.

If $R = 0$ no abbreviation can be achieved by coding. If $R > 0$ then some abbreviation is possible without losing information. If $R = 1$ then $H = 0$ and no information is being transferred. An example with $R = 1$ is the meassge E E E E E E E . . . E.

The source of text can be examined digraphically by the same function R. In this case c is replaced by c^2, so that

$$R = \frac{2 \log c - H(p_{11}, p_{12}, \ldots, p_{cc})}{2 \log c}.$$

It can happen that digraphic or polygraphic examination will reveal redundancy not shown by the previous measures. For example a five digit code with a garble check will have no more information pentagraphically than it has tetragraphically.

A way of estimating the entropy of a source from a long message is the following: Consider a message of length N. It contains f_1 of the first letter, f_2 of the second, and so on. The expected value of f_i is $p_i N$. The probability q of exactly this message is $q = p_1^{f_1} p_2^{f_2} \ldots p_c^{f_c}$, or $\log q = \sum_{i=1}^{c} f_i \log p_i$. The expected value of $\log q$ is

$$E(\log q) = \sum_{i=1}^{c} p_i N \log p_i$$

$$= -NH(p_1, \ldots, p_c), \text{ or}$$

$$H = \frac{E(\log \frac{1}{q})}{N}.$$

See (34) theorem 3.

Shannon states that the redundancy for English is approximately $R = \frac{1}{2}$. One way to estimate this is to to see what proportion of letters can be eliminated (at random) from English text without concealing the meaning from a discerning reader. Following is an example with more than half the letters (and the word spaces) deleted in a pattern taken from a random number table.

Y . U . A . . . OL . - . A T . . R - . IL . IA .
. . E - . . UN . - MAN . . . . D - AN . - YOUR
. A . . - HAS . . . C . . . - V . R . . WH . . .
A . D - YE . - Y . U . INC . . S . . T . Y . . . A . .
. . . . O . R . . . A . - D . . . OU . T . INK -
. . . Y . U . - . . . . . I . - . IG . .

The symbol "-" is a word spacer. Some word spacers are present, some are absent. About half the letters are shown; find the others.

Here is another example. Again about half the letters (48%) have been suppressed, but this time it is the less frequent letters which have been suppressed.

. H E . . . . H I N . . . S . . I . S I S . . . . E . . . E . H .

N I . . . . . O . E S S I N . H I . H I . . O S S I . I . I . I E S

. . E N O . . O N S I . E . E . $_p$ H O . E . E . H . . I N . .

O . N .

No word spacers are present. The little p is a period.

E O N I S H These 6 letters have 52 occurrences.

T A R D L U C M F G B V Y W P J K Q X Z 48 occurrences suppressed

In this third example the more frequent letters are suppressed, 54% of the text.

M O D . R N . . G . B R . H . . . X P O . . D F O R . H

. F . R . . . . M . . H . F U . . V . R . . . Y . N D R

. C H N . . . O F P O . . . B . . M . . H . M . . . C . .

. Y . . M . $_p$ W . . H . . .

E T A I S L Of these 6 letters 54 occurrences are suppressed.

R N M O F G B H U V Y C W D P These 15 letters appear 45 times.

J K Q Z do not appear.

From this outline of information theory we can see two things. First, the theory is inherently statistical. The statements of information theory all deal with large numbers of elements, never with single elements. In fact most questions of statistical theory are motivated by attempts to derive information from incomplete or diluted data. Second, that cryptology is very much concerned with the same questions as is information theory.

0, 2 Statistics.

Almost all statistical theory is applicable in some way to cryptologic problems. That theory most readily applied is distinctly mathematical. We refer the reader to:

(1) "Introduction to Mathematical Probability" by Uspensky, McGraw-Hill.

(2) "Mathematical Methods of Statistics" by Harald Cramer, Princeton University Press.

(3) "Mathematical Statistics" by S. S. Wilks of Princeton.

(4) "An Introduction to Probability Theory and its Applications" by W. Feller.

One of the frequent cryptologic problems is testing hypotheses. That is, we have some data, a cryptogram say, which may have come from one of a number of causes. Which one is most likely? The available evidence in cryptanalysis can normally be digested by using Bayes' theorem. Usually a multiple of the logarithm of the Bayes factor is computed, that for each unit being called its "weight" in some appropriate unit. This is described in Chapter 2.

The cryptanalyst is often trying to draw inferences about a "universe" of cryptograms from a sample. This sample is not under his control; he has to take what is intercepted. The smallness of the samples may prejudice his inferences. The first case discussed is, given the frequency count of a sample, to estimate the distribution of the universe from which it came. A first estimate would be the proportion f_1 /N for a letter which occurred f_1 times in a sample of N. Under certain circumstances better estimates can be made. This is taken up in Chapter 3.

Because rectangular arrays are so much used in cryptanalytic statistics, a preliminary chapter is devoted to matrices. This is followed by an exposition of flagging (a technique of sorting distributions into two or more sets) with emphasis on its matrix character.

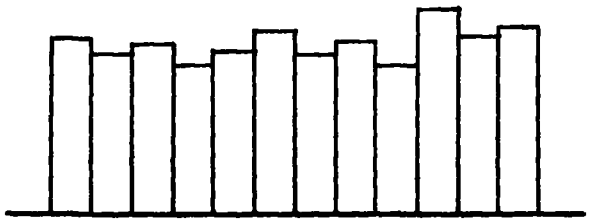
A chapter on Fourier methods is followed by one on circulices, which shows that the Fourier transform is included in matrix theory.

All the matrix and Fourier techniques are aimed at recovering periodic tendencies in a stream of cipher text or key.

1. Review of Elementary Mathematical Statistics.

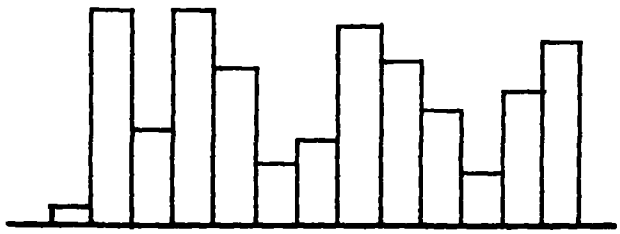
Some of the results of Kullback (6) and other sources are summarized here for ready reference.

One of the first questions arising when looking at a sample of "random" objects, such as a stream of letters, is, "Is it really 'random'?" To answer this we need to know what randomness is. It is frequently taken to be the equivalent of "flatness." This latter term can be visualized in the following way. Suppose that a histogram is drawn from a frequency count of the sample. It might look as follows:



The sample pictured here is "flat," for the deviations which occur from the average are what might occur from chance.

If the sample were from a source which is not pure chance, but has a pattern, such as letters from newspaper text or digits from a telephone directory, then the histogram might appear thus:



This sample is "rough," rather than flat.

There are several statistics invented to measure roughness. One of these is ϕ (pronounced phi). If the observed frequencies are f_1, f_2, f_3, and so forth to f_c, where c is the number of categories, then

$$\phi = \sum_{i=1}^{c} f_i(f_i-1). \tag{1, 1}$$

For a given size $N = \sum_{i=1}^{c} f_i$ of sample ϕ is larger for rougher and smaller for flatter samples. The smoothest possible sample is that in which all frequencies are the same, $f_i = \frac{N}{c} = f_j$. Then $\phi = \sum_{i=1}^{c} \frac{N}{c}(\frac{N}{c}-1) = N(\frac{N}{c}-1)$. The roughest possible frequency count is that in which one count has everything, $f_1 = N$, and all others have nothing, $f_i = 0$. Then $\phi = N(N-1)$. The values of ϕ for intermediate roughness are between these extremes. Another observation will illustrate how ϕ varies. Suppose we take a little, say 1, off a small count, say f_1, and add it to a larger count, say f_2. We now have a new count f'_i, where $f'_1 = f_1 - 1$, $f_2' = f_2 + 1$, and

$f_i' = f_i$ for $i > 2$. What has happened to ϕ? It has become:

$$\phi' = \sum_{i=1}^{c} f_i'(f_i'-1)$$

$$= f_1'(f_1'-1) + f_2'(f_2'-1) + \sum_{i=3}^{c} f_i'(f_i'-1)$$

$$= (f_1-1)(f_1-2) + (f_2+1)f_2 + \sum_{i=3}^{c} f_i(f_i-1)$$

$$= (f_1-1)f_1 + f_2^2 - f_2 + \sum_{i=3}^{c} f_1(f_i-1) - 2(f_i-1) + 2f_2$$

$$= \phi + 2(f_2-f_1) + 2.$$

Since f_2 is larger than f_1 we see that ϕ' is larger than ϕ by at least 2. The greater the a priori discrepancy f_2-f_1 between the two counts the more this operation increases ϕ.

The drawing of a sample is often pictured in the following way. Suppose we had a large barrel filled with tiny scraps of paper, on each of which is a mark, such as a letter. The number of scraps is very large, and the number of distinct marks is c. The barrel is referred to as the "universe," and the contents of the barrel as the "population." We stir up the population vigorously, and then reach in and withdraw a handful of scraps of paper; this is the "sample." The number of scraps of paper with mark i on them is f_i.

The notation E(x) is read "the expected value of x" and is defined to be the weighted mean of all the possible values of x.

Suppose the marks in the population are in proportion $p_1 : p_2 : p_3 : \ldots$.
Then the proportion f_i/N of marks i in the sample is expected to be p_i; $E(f_i/N) = p_i$, or $E(f_i) = p_iN$.

From this we can find that

$$E(\phi) = N(N-1) \sum_{i=1}^{c} p_i^2, \tag{1, 2}$$

see (6), paragraph 18.

A frequently used function of ϕ is the Index of Coincidence (abbreviated I. C.).

$$\delta = \frac{c\phi}{N(N-1)}. \tag{1, 3}$$

The expected value of δ is

$$E(\delta) = \frac{c}{N(N-1)}E(\phi) = c \sum_{i=1}^{c} p_i^2. \tag{1, 4}$$

For a flat universe, that is , one for which $p_i = 1/c$ for each i,

$$E(\delta) = c \sum_{i=1}^{c} 1/c^2 = 1,$$

which is convenient to remember and use. An exposition of the I. C. can be found in the pamphlet *The Index of Coincidence* (18).

The expression,

(1, 5) $$\gamma = \frac{c}{N^2} \sum_{i=1}^{c} f_i^2$$

approaches δ as N grows large, and is sometimes used instead. The factor $\sum_{i=1}^{c} f_i^2 = \psi$ is sometimes used also. The statistic ψ is related to ϕ by

(1, 6) $$\phi = \psi - N.$$

The variance of ϕ and ψ are derived in Kullback (6), Appendix D. It is

(1, 7) $$\sigma^2(\phi) = \sigma^2(\psi)$$

$$= 2N(N-1)\left[2(N-2)\sum_{i=1}^{c} p_i^3 + \sum_{i=1}^{c} p_i^2 - (2N-3)\left(\sum_{i=1}^{c} p_i^2\right)^2\right].$$

For the special case $p_i = 1/c$, which is used continually for comparison purposes, this reduces to

(1, 8) $$\sigma^2(\phi) = 2N(N-1)\frac{c-1}{c^2}.$$

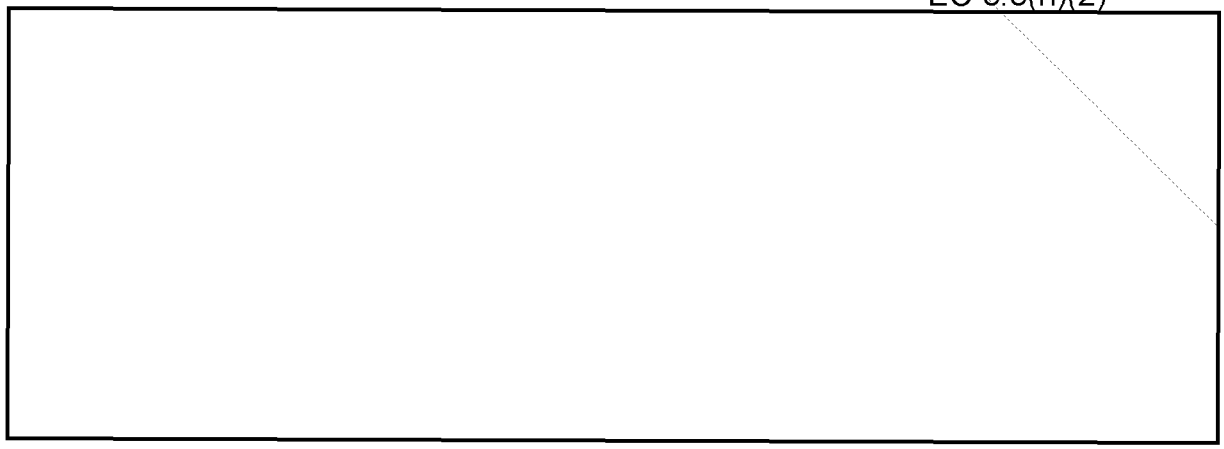
Therefore for the I. C. the variance for a flat universe is

(1, 9) $$\sigma^2(\delta) = \frac{c^2}{N^2(N-1)^2} 2N(N-1)\frac{c-1}{c^2} = 2\frac{c-1}{N(N-1)} = 2\frac{c-1}{N^2}.$$

This follows from the theorem that for a variable x and constants a and b,

(1, 10) $$\sigma^2(ax+b) = a^2\sigma^2(x)$$

PL 86-36/50 USC 3605
EO 3.3(h)(2)



If we have a quantity which takes on various values, and if these values are determined by chance, such that we can express the probabilities that $a < x < b$, then we call the variable a "statistical" or "stochastic" variable. For example, if we cut some newspapers into little pieces, one

letter on each piece, put these in a barrel, and then draw them out one by one, the letter drawn is a stochastic event. The probability for the occurrence of E is greater than that for Q, and for a large enough sample the letters will appear in proportions which can be predicted approximately. A description of the probabilities of the values is called the "distribution" of the variable.

Two events are said to be "independent" if neither affects the outcome of the other. For instance with our barrel of newspaper text, if after examining a sample we replace it in the barrel and stir before drawing another then we can be confident that the results of the two samples are independent. On the contrary if we keep out one piece with one letter on it then subsequent draws will be modified, potentially at least. The determination of whether events are independent or dependent, completely or partially, is sometimes of great importance in ascertaining the significance of those events.

1, 1 The Binomial Distribution.

If we have a stochastic event which can take on two values (such as "hit" or no "hit" between letters of two texts which have been lined up) with probabilities p and q, then if x is 1 or 0 according as one or the other event occurred, x is "binomially" distributed. Then

$$E(x) = p.1+q.0 = p, \text{ and}$$

$$\sigma^2(x) = E(x^2)-E^2(x) = p.1^2 + q.0^2-p^2 = p(1-p) = pq.$$

The sum $y = \sum_{1}^{n} x$ is the number of times the value 1 came up in n trials.

(1, 1, 1) $$E(y) = pn \text{ and}$$

(1, 1, 2) $$\sigma^2(y) = npq$$

(if the trials do not affect each other).

The probability of a specific count $y = k$ is

(1, 1, 3) $$P(y=k) = \frac{n!}{k!\,(n-k)!}\,p^k q^{n-k}.$$

PL 86-36/50 USC 3605
EO 3.3(h)(2)

The probability of $y = k$ or more is

(1, 1, 4) $$P(y \geq k) = \sum_{y=k}^{n} \frac{n!}{y!(n-y)!}\,p^y q^{n-y}.$$

Tables of these two probabilities have been compiled, (10).

1, 1, 1 Example.

1, 2 The Multinomial Distribution.

If a stochastic event can take on c values with probabilities $p_1, p_2, \ldots, p_c$, then in n trials the probability of getting a specific frequency count $f_1, f_2, \ldots, f_c$ is

$$(1, 2, 1) \qquad P(f_1, f_2, \ldots, f_c) = \frac{n!\, p_1^{f_1} p_2^{f_2} \ldots p_c^{f_c}}{f_1!\, f_2! \ldots f_c!}.$$

The probability of getting this count or a less likely one is a less easily handled question, and is more important. The usual method of handling it is to transform the question into terms of another distribution, such as the binomial or the Poisson or the normal.

1, 2, 1 Example of Multinomial Distribution.

If a barrel is full of English newspapers cut into little pieces, one letter to each piece, and then pieces are drawn haphazardly after stirring, frequency counts of samples will be multinomially distributed.

Or take a simpler example. Suppose we have a 5-letter alphabet, A, E, I, O, U, in the ratio 60:30:20:15:12 in the barrel. The probabilities are then $\frac{60}{137}$, $\frac{30}{137}$, $\frac{20}{137}$, $\frac{15}{137}$, and $\frac{12}{137}$. If we draw a sample of 46, the most probable frequency count is $f_A = 20$, $f_E = 10$, $f_I = 7$, $f_O = 5$, $f_U = 4$. The probability of getting exactly that is

$$P(20,10,7,5,4) = \frac{46!\left(\frac{60}{137}\right)^{20}\left(\frac{30}{137}\right)^{10}\left(\frac{20}{137}\right)^{7}\left(\frac{15}{137}\right)^{5}\left(\frac{12}{137}\right)^{4}}{20!\ 10!\ 7!\ 5!\ 4!} \doteqdot .001\,.$$

1, 3 The Poisson Distribution.

If we are dealing with very large total counts n it is convenient to think in terms of the expected value a. From (1, 1, 1) we have $a = pn$. Then (1, 1, 3) becomes, if we separate the factors depending on n,

$$P(y=K) = \frac{n!}{K!(n-K)!}\left(\frac{a}{n}\right)^K\left(1-\frac{a}{n}\right)^{n-K}$$

$$= \frac{a^K}{K!}\left(1-\frac{a}{n}\right)^n \frac{n(n-1)(n-2)\ \ldots\ (n-K+1)}{n^K\left(1-\frac{a}{n}\right)^K}$$

$$= \frac{a^K}{K!}\left(1-\frac{a}{n}\right)^n \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ \ldots\ \left(1-\frac{K-1}{n}\right)}{\left(1-\frac{a}{n}\right)^K}$$

$$\doteqdot \frac{a^K}{K!}e^{-a}.$$

This approximation is good if $a/n = p$ is small and n is large. Tables of this "Poisson Distribution" have been published, (13) and (14), along with the cumulative function,

$$P(y \geq K) = \sum_{y=K}^{\infty} \frac{a^y}{y!} e^{-a}.$$

Since the tables are independent of n they are short and easy to use.

1, 3, 1 Example of the Use of the Poisson Distribution.

In a list of 98,000 four digit groups the most frequent group occurs 29 times. Is this extraordinary? Here $p = .0001$, and $a = pn = .0001 \times 98{,}000 = 9.8$. Table II (13) says that when 9.8 are expected, 29 or more will occur 1 time in a million, $p = .000{,}001$, when the particular group is specified in advance. We must remember that this is the best in 10,000 tries, so that in a hundred samples of this kind there would be a million opportunities for a group to be frequent. Therefore this result would occur about once in a hundred such random experiments.

1, 4 The Normal Distribution.

If $y_1, y_2, \ldots, y_n$ is a sequence of stochastic variables all with the same distribution, where the mean is $E(y)$ and the variance $\sigma^2(y)$, then the sigmage

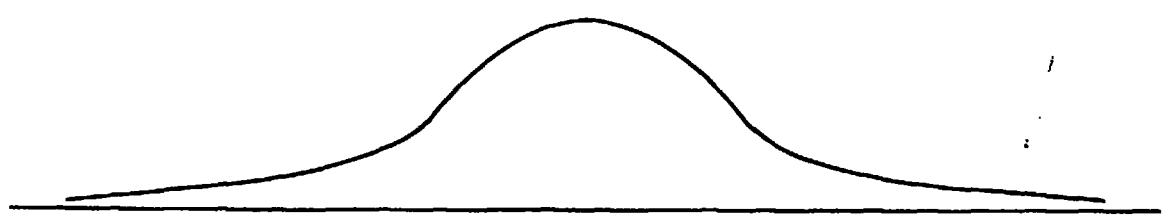
$$S = \frac{\sum_{i=1}^{n} y_i - n\,E(y)}{\sqrt{n}\,\sigma(y)} \qquad (1, 4, 1)$$

is a stochastic variable with mean $E(S) = 0$ and variance $\sigma^2(S) = 1$. If n is large, it can be shown that

$$P(S \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\,dt. \qquad (1, 4, 2)$$

This is the "normal" distribution with mean 0 and variance 1. The condition that a variable be a sum of a large number of similarly distributed quantities is frequently fulfilled, and consequently the normal distribution has wide application. For instance, Gauss assumed that errors in measurements were accumulations of smaller errors, and derived (1, 4, 2) for the distribution of errors. Sometimes this is called the "Gaussian error function," or the "bell shaped curve."

The latter refers to the appearance of the graph of $y = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$.



1, 4, 1 Example of the Use of the Normal Distribution.

In solving a cipher system a method of exhaustive trials has been devised, exactly one trial of which should give plain text. To help decide which trial was successful, each candidate for plain text is judged by a set of weights, the sum of which is a "score." The expected score for random happens to be -10, with a standard deviation of 4.5. After 10,000 trials the best score found was 8; how good is this? This is $8-(-10) = 18$ above expected, or exactly $18/4.5 = 4$ sigmas. Since the scores are sums of other stochastic variables, the normal distribution applies approximately, and 4 sigmas or better will occur about once in 33,000 trials from (16). Since we have made 10,000 trials already, we should get a score this good or better once in 3 such experiments from random material. This is a discouraging result for the cryptanalyst.

1, 5 The Chi-Squared Distribution.

Consider the distribution of the stochastic variable

(1, 5, 1) $$\chi^2 = y_1^2 + y_2^2 + \ldots + y_n^2.$$

If the y's are themselves independently and normally distributed, then it can be shown that

(1, 5, 2) $$P(\chi^2 \leqq x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\int_0^x t^{\frac{n}{2}-1} e^{-\frac{t}{2}}\, dt,$$

see (2), Section 18.1. The parameter n is the number of degrees of freedom. It is important in using formula (1, 5, 2) to have n as the number of *independent* summands in (1, 5, 1). Tables are available of (1, 5, 2), for instance (12). This distribution is important because ψ and the I. C. are asymptotically distributed this way. For an heuristic treatment of Chi-square and related statistics consult (30).

1, 6 The Zipf Distribution.

Mr. Zipf (28), in a theory he developed on the use of tools, predicted that if certain items (such as words) were ranked in the order of their use, so that f_i is the frequency of the i^{th} most frequent, then $i \cdot f_i$ would be approximately constant. That is

(1, 6, 1) $$f_i = f_1/i.$$

This is found to be reasonably accurate for codes, with the exception of the most and least frequent groups. Sometimes the assumption that the k most frequent groups have been concealed gives a better approximation. Then we have

(1, 6, 2) $$(i+k)\, f_i = (1+k)\, f_1,$$

$$\text{or } f_i = \frac{1+k}{i+k} f_1.$$

This is a distribution in the mathematical sense, and one of considerable cryptologic interest. If we write

(1, 6, 3) $$\sum_{i=1}^{c} 1/i = t$$

then from (1, 6, 1),

(1, 6, 4) $$N = \sum_{i=1}^{c} f_i = \sum_{i=1}^{c} f_1/i = tf_1.$$

Then the expected rank is

(1, 6, 5) $$E(i) = \sum_{i=1}^{c} i f_i/N = c f_1/N$$

$$= \frac{cf_1}{tf_1} = c/t$$

$$\doteq c/\log_e c$$

since $t = \sum_{i=1}^{c} 1/i \doteq \log_e c$

The variance of the rank is

(1, 6, 6) $$\sigma^2(i) = \frac{(c+1)c}{2t} - \frac{c^2}{t^2},$$

since $$E(i^2) = \frac{1}{N} \sum_{i=1}^{c} i^2 f_i$$

$$= \frac{1}{N} \sum_{i=1}^{c} i f_1$$

$$= \frac{f_1}{N} \sum_{i=1}^{c} i$$

$$= \frac{f_1}{tf_1} \frac{(c+1)c}{2} = \frac{c+1}{2t} c.$$

All the moments can be computed in terms of the sums $\sum_{i=1}^{c} i^k$, which are evaluated in Section 405 of reference (8).

1, 7 Approximate Distributions.

The normal, Poisson, and chi-squared distributions provide approximations to the finite distributions we are usually interested in, approximations which get better as the sample size increases. That is, the sum of similarly distributed variables is asymptotically normal.

There is a statistic measuring the "goodness of fit" of two counts, the "chi-squared," defined as follows: if f_i and g_i are the components of the counts, with

$$\sum_{i=1}^{c} f_i = N, \quad \sum_{i=1}^{c} g_i = M,$$

then

$$(1, 7, 1) \qquad \chi^2 = \sum_{i=1}^{c} \frac{Mf_i}{Ng_i} - N.$$

This asymmetrical version was derived from the viewpoint that the g_i represented the universe, and the f_i a sample. There are other versions. This statistic is asymptotically chi-squared distributed. One must distinguish carefully between the chi-squared statistic and the chi-squared distribution.

The cryptanalytic statistics ϕ, ψ, δ, and γ are all distributed asymptotically chi-squared. That is, these distributions can be computed from that of χ^2, given in Section 1, 5. The last two of these have been tabulated, see (17).

The index of coincidence found by counting hits has a binomial distribution (17).

The Cross I. C.,

$$(1, 7, 2) \qquad \xi = \frac{c}{MN} \sum_{i=1}^{c} f_i g_i,$$

has a distribution computable from the Incomplete Beta Function, see (17) for the tables. The distribution is as follows. Let $I_y\left(\frac{c-1}{2}, \frac{1}{2}\right)$ stand for the Incomplete Beta Function, which has been tabulated by Pearson (26). If the I.C.'s of the component distributions are $F = \frac{c}{M^2} \Sigma f_i^2$, and $G = \frac{c}{N^2} \Sigma g_i^2$, then

$$(1, 7, 3) \qquad P(\xi \geq 1 + \chi \sqrt{(F-1)(G-1)} = I_{1-\chi}\left(\frac{c-1}{2}, \frac{1}{2}\right).$$

The derivation of this is given by Gleason (27). This statistic is related to the correlation coefficient.

Since these approximations are asymptotic, and since cryptanalytic work is frequently with small samples, some experiments have been made (23) with the δ I. C. to test the accuracy of our tables. With the smallest sample tried, $c = 5$ and $N = 7$, the least accurate estimate was off by a factor of 10, where the probability of a sigmage of 6 was given as .057 by chi-square, while in fact it occurred in only .0054 of the cases. In a less extreme case, $c = 5$, $N = 15$, The probability of a sigmage of 11.5 is .00013, while a chi-squared estimate is .000025, too small by a factor of 5. The number of counts per category, N/c, is a criterion for the accuracy of the chi-square as an estimate for the distribution of the δ I. C. We see that with $N/c = 3$ it is off at most by only a factor of 5, not too bad for most cryptanalytic applications.

For many statistical questions in cryptanalysis extremes ("tails)" of a distribution are needed. For instance, from Poisson it may be required to know approximately the probability of getting 320 or more successes when 200 are expected. This is far beyond the tabulated values and the calculation is very laborious. There are special methods of getting good approximations for the extremes to some of these distributions. The standard methods for approximating near the mean are nearly useless.

A method of getting as close an approximation as one pleases to the cumulative Poisson has been given by Cramer and Gleason (21), and is presented below. The derivation uses a continued fraction expansion.

The probability of c or more successes when a are expected is

$$P(c,a) = \sum_{k=c}^{\infty} \frac{a^k e^{-a}}{K!}$$

This can be approximated in the following way.

$$P(c,a) = \frac{a^c e^{-a}}{c!} \frac{A_i}{B_i}$$

where A_i and B_i are defined recursively by

$$A_i = A_{i-1} + (-1)^{i-1} a\, a_i\, A_{i-2}$$

$$B_i = B_{i-1} + (-1)^{i-1} a\, a_i\, B_{i-2}$$

where for convenience a_i is defined as

$$a_{2m+1} = \frac{m}{(c+2m-1)\ (c+2m)}$$

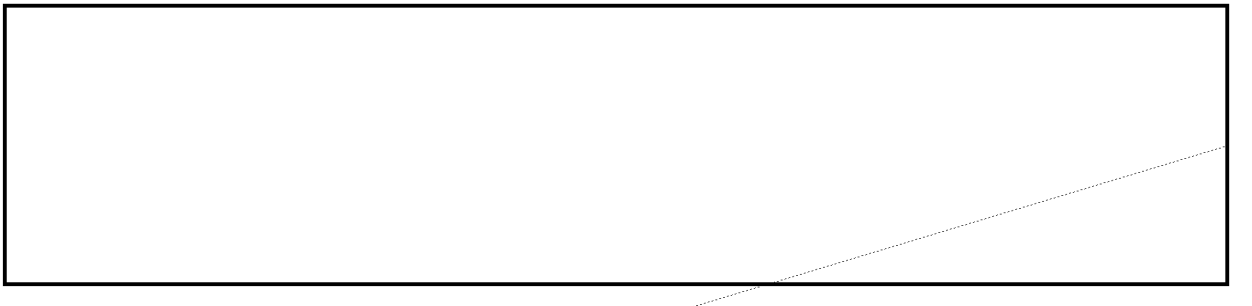
$$a_{2m} = \frac{c+m-1}{(c+2m-2)\ (c+2m-1)}$$

This gives an iterative method of approximating P(c,a). Each approximation is better than the last, and some are overestimates and some underestimates, so that the true value is boxed in.

1, 8 Regression.

A statistic may seek to measure the interrelation between the components of the data. For instance the data may be the ages of men, each with that of his spouse. In such a case the data can be plotted on a graph, one point for each datum. This graph may have to have high dimension, but theoretically this is no objection. The data will ordinarily give a cloud of points and the shape of this cloud will be of interest. One may particularly look to see if it is elongated, and if so in what direction. If it is, then there will be an axis along which it is stretched, and this axis can be found. It is called the "line of regression". It is the line which is closer to all the points, in a sense, than any other line. More precisely, it is the line such that the sum of the squares of the distances of the data from it is a minuimm. A method for finding this line will be given in section 4, 6 after some vector techniques have been discussed.

2. The Matching of Distributions.



2, 1 Goodness of Fit Defined.

If a universe U is known, the exact probability P(S) of a particular sample S of size N can be computed by using the multinomial expression of Section 1,2. Furthermore the possible samples T of size N which give a lower probability, $P(T) < P(S)$, can theoretically be enumerated, since there are only c^N samples possible. Let $M = \sum_{P(T)<P(S)} P(T) + \frac{1}{2} \sum_{P(T)=P(S)} P(T)$

Then Dawson (19) calls the relative number, $\frac{M}{c^n}$, the "goodness of fit (g.f.)" of S with U. To paraphrase then, the goodness of fit of S with U is the probability that a random sample T will give a number P(T) less than P(S). The "poorness of fit" can be defined as 1-g.f. If U is the flat universe with $p_i = \frac{1}{c}$ then the poorness of fit of S with U can be called its "roughness".

The advantage of this definition is that samples are ranked according to the probability with which they would arise. For example, consider the two samples from a 5-letter alphabet, S_1 : 10,10,10,10,0 and S_2 : 6,6,6,6,16. Here N = 40. The gamma I. C. of each is 5/4, but the roughnesses are not the same, for S_2 will occur in drawings from a flat universe over 26 times as often as S_1.

In (19) it is shown that the distribution of the goodness of fit is closely determined by the distribution of

$$\textbf{(2, 1, 1)} \qquad s = \sum_{i=1}^{c} (f_i + 1/2) \log f_i/Np_i.$$

Sometimes s itself is used as a measure of roughness. The same source (19) shows that 2s is distributed asymptotically chi-squared. That is,

$$\textbf{(2, 1,.2)} \qquad P(s \leq x) = \frac{1}{2^{\frac{c}{2}}\Gamma(\frac{c}{2})} \int_0^{2x} t^{\frac{c}{2}-1} e^{-\frac{t}{2}} dt,$$

which has been tabulated for the χ^2 statistic.

2, 2 Goodness of Fit of Two Samples.

The probability of drawing 2 samples S_1 and S_2 of size N_1 and N_2 respectively from a universe at random is the same as that of drawing a single sample S of size $N = N_1 + N_2$ and then separating S at random, getting S_1 and S_2. If the universe is unknown the likelihood of the 2 samples arising from the same source is measured by the probability of such a split. By neglecting constant factors the function

$$\textbf{(2, 2, 1)} \qquad F = \prod_{i=1}^{c} f_i! \, g_i!$$

is arrived at as a measure of the poorness of fit of

$$S_1 : \{f_i\} \text{ and } S_2 : \{g_i\}.$$

See (19), Section 10, for details. The distribution of this is not so well known.

WEIGHTING TECHNIQUES

2, 3 Bayes' Theorem.

If H_1 and H_2 are two hypotheses, and if E is an event, and if $P(E/H_i)$ is the probability of E if H_i pertains, then the probability of E and H_i is

(2, 3, 1) $$P(E,H_i) = P(H_i)P(E/H_i).$$

This can be rewritten to

(2, 3, 2) $$P(H_i)P(E/H_i) = P(H_i,E) = P(E)P(H_i/E)$$

which can be solved for

$$P(H_i/E) = \frac{P(H_i)P(E/H_i)}{P(E)}.$$

The ratio of these for i = 1,2 gives

Bayes' Theorem Stated

$$\frac{P(H_1/E)}{P(H_2/E)} = \frac{P(H_1)}{P(H_2)} \cdot \frac{P(E/H_1)}{P(E/H_2)} \text{ or}$$

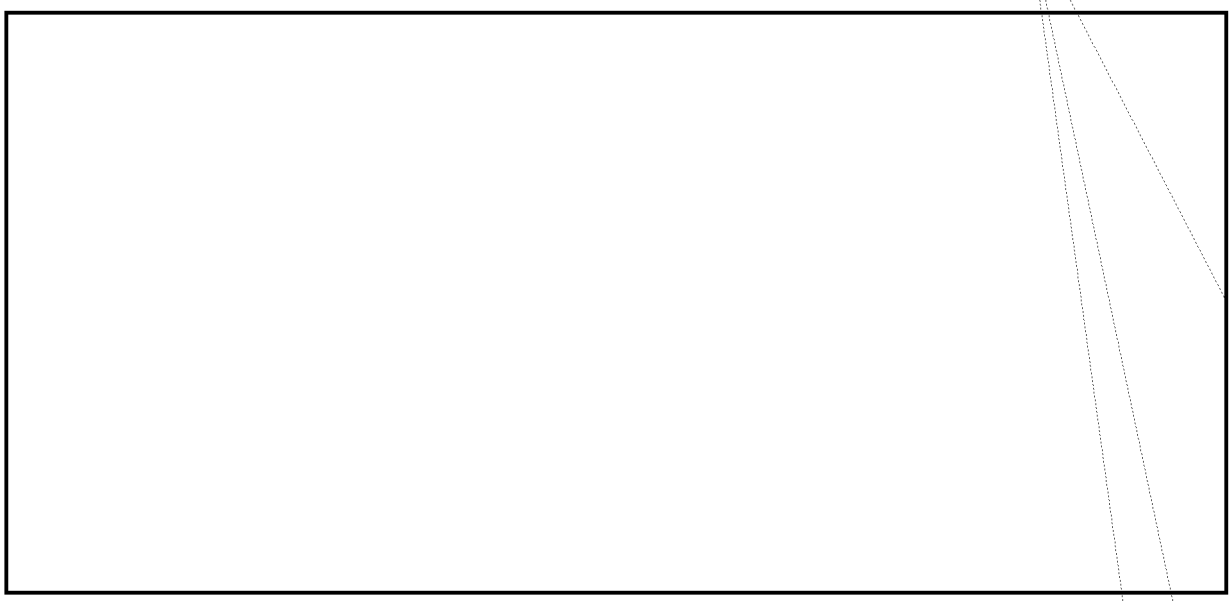
The odds on the hypothesis having observed the event } is { the odds preceding the event multiplied by a factor.

PL 86-36/50 USC 3605
EO 3.3(h)(2)

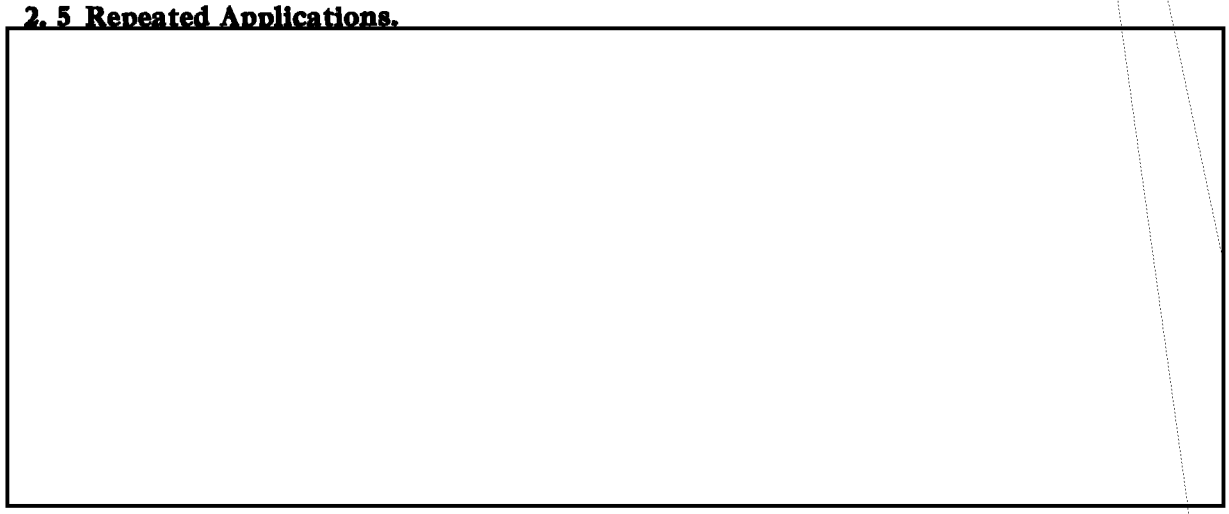
This factor is $\frac{P(E/H_1)}{P(E/H_2)}$, the "Bayes Factor".

This says that having observed an event E we can draw some conclusions about the possible causes of this event. The possible causes are designated H_1, H_2, ... The probability of H_1 is written $P(H_1)$, and the probability of H_1 after having observed the event E is written $P(H_1/E)$. The probability of the event E if H_1 is in fact the cause is written as $P(E/H_1)$. The theorem concerns the ratio of probabilities, or odds. $\frac{P(H_1)}{P(H_2)}$ is the odds in favor of H_1 against H_2. $\frac{P(H_1/E)}{P(H_2/E)}$ is the same odds after the event E has occurred. $\frac{P(E/H_1)}{P(E/H_2)}$ is not an odds, since the events are not alternatives, but is called "the Bayes factor". This theorem gives an objective way of considering circumstantial evidence.

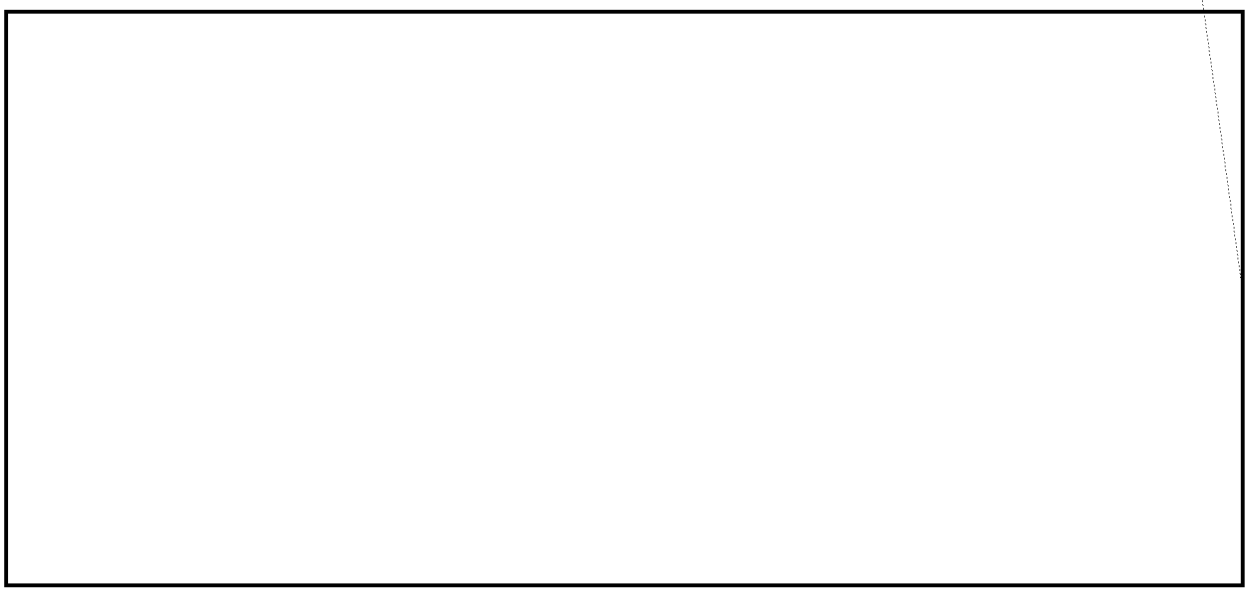
2, 4 An Example of Application.



2. 5 Repeated Applications.

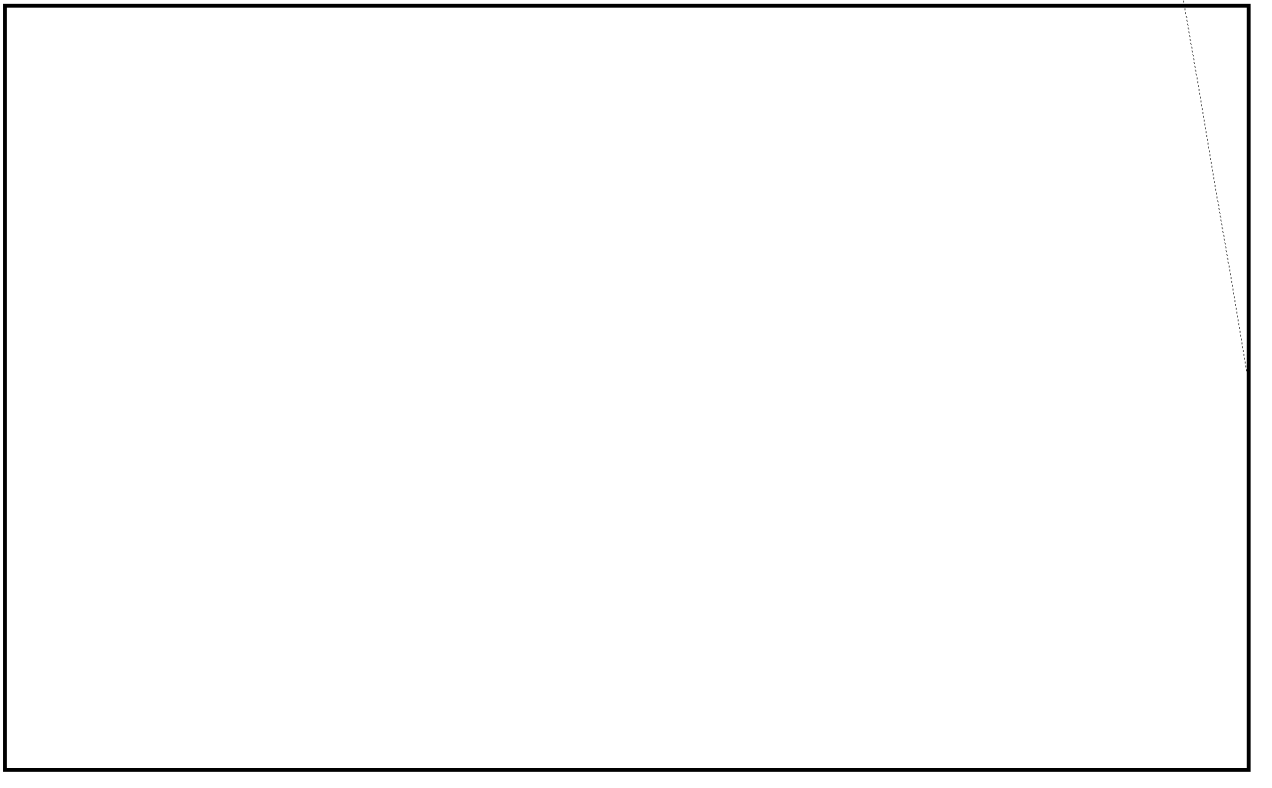


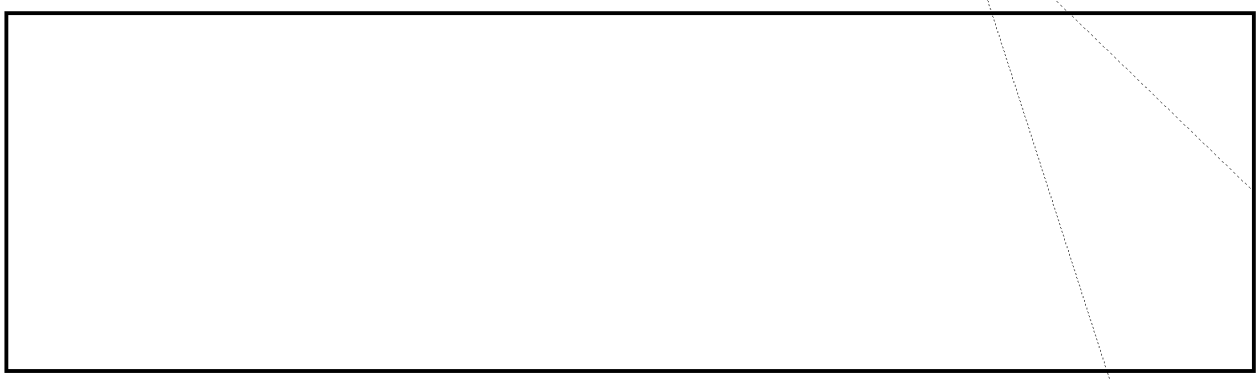
2. 6 Example of the Calculation of Bayes Factors.



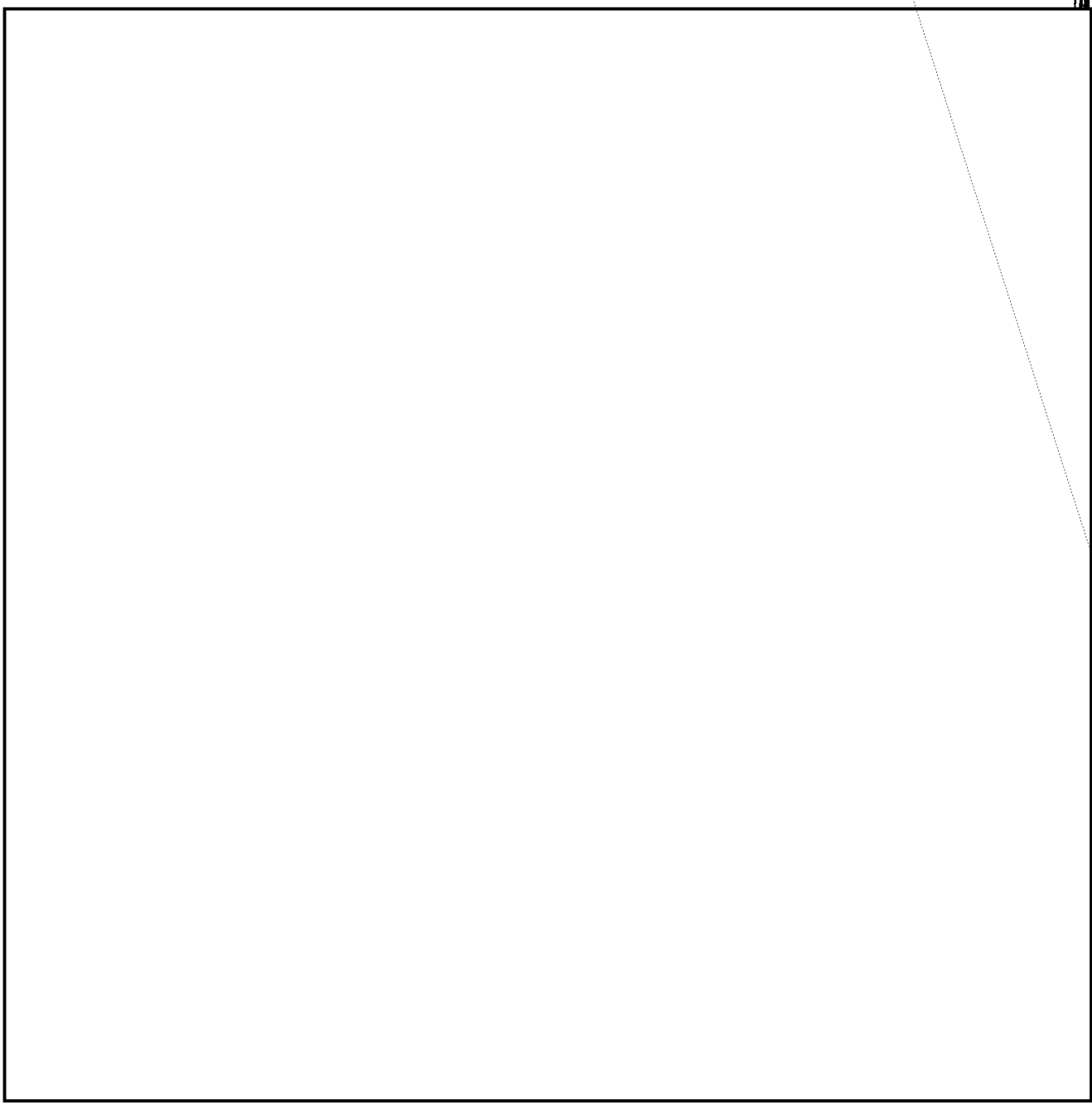
~~CONFIDENTIAL~~

2, 7 Weights.

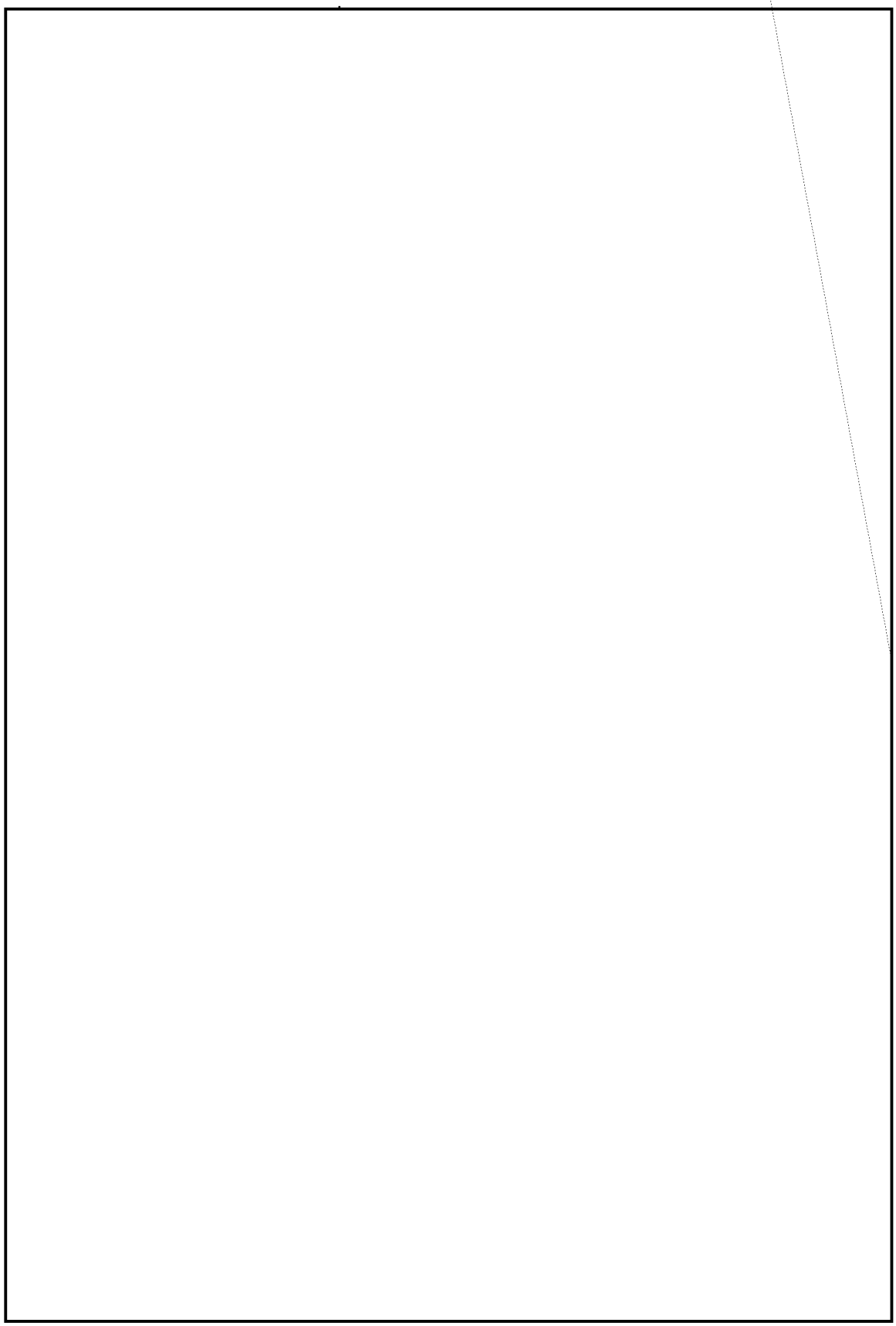




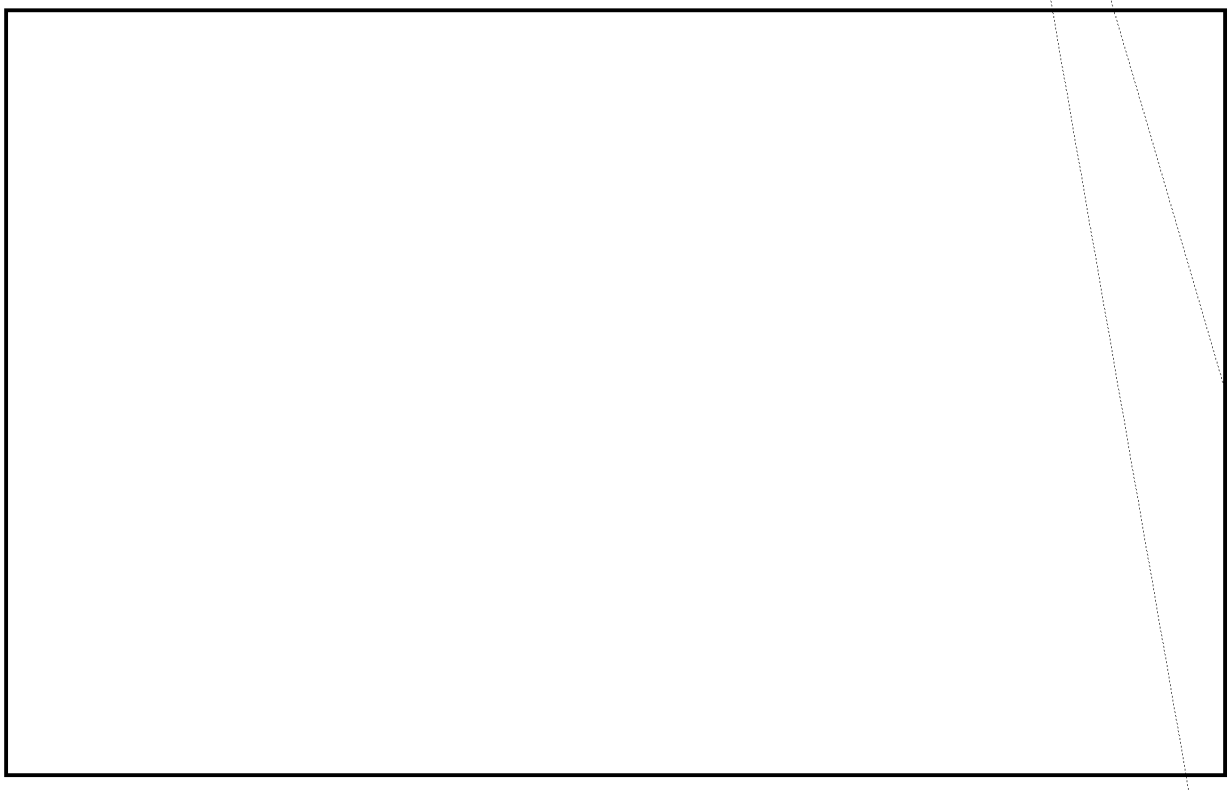
2, 8 Rounded Weights and Risk-Admission Diagrams.



~~CONFIDENTIAL~~



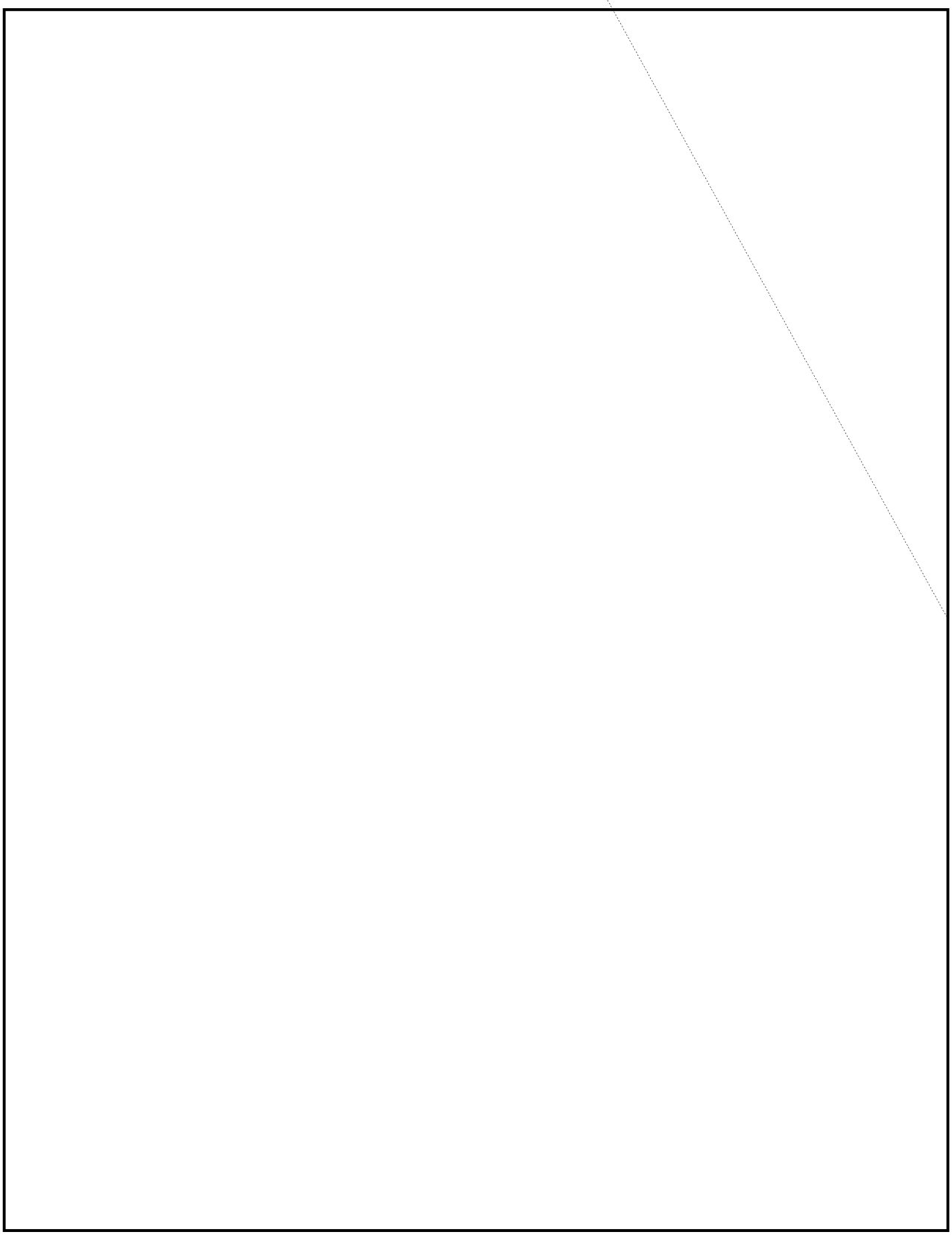
2, 9 Two-category Weights.



2, 10 Three-category Weights.

~~CONFIDENTIAL~~

2, 11 Statistics of Bayes Factors.



3. Unprejudiced Estimates of Universes.

3, 1 The Law of Succession.

To decide whether a frequency count comes from one of two universes one should know the distributions in those universes. Usually only samples from those universes are available. For example, suppose we are to devise a test for English newspaper text. We count 200 letters of newspaper text, and use this count as an estimate of the frequency distribution. But in this count the frequency of Q is zero. If we accept this as a fact, then no sample containing a Q can possibly satisfy our test for newspaper text. The sample has given us a "prejudiced" picture of the universe. If we add 1 to each count and use the result as an estimate, then no letter will be impossible. This procedure can be justified rigorously. If $\{f_i\}$ is a frequency count of such a sample, then $\{1+f_i\}$ is an unprejudiced estimate of the universe, under the hypothesis that a priori all distributions are equally likely.

The derivation is as follows: Given a process producing a c-letter alphabet, suppose a sample S of N letters has been drawn, getting f_i cases of the ith letter,

$$\sum_{1}^{c} f_i = N.$$

If one more letter is drawn, what are the odds on its being the ith letter x_i? Suppose that before looking at the frequency count $\{f_i\}$ the odds are even on all letters. Then after looking at the counts the odds are $f_1+1 : f_2+1 : \ldots : f_c+1$. A more vivid picture is as follows. Suppose that initially there are c^{N+1} hats, each with $N+1$ letters in it, and each with a different frequency count. All possible frequency counts of $N+1$ letters are available in the hats. Now a hat is selected and N letters are drawn from it. What are the odds on the next letter? The odds are $1+f_1: 1+f_2: 1+f_3: \ldots$. For the probability of drawing our sample S and then drawing x_i is the same as that of drawing a sample $S+x_i$ and then drawing x_i from the sample, that is, $P(S+x_i)\ P(x_i/S+x_i)$. We assume that $P(S+x_i) = P(S+x_j)$. Then the odds on x_i to x_j are

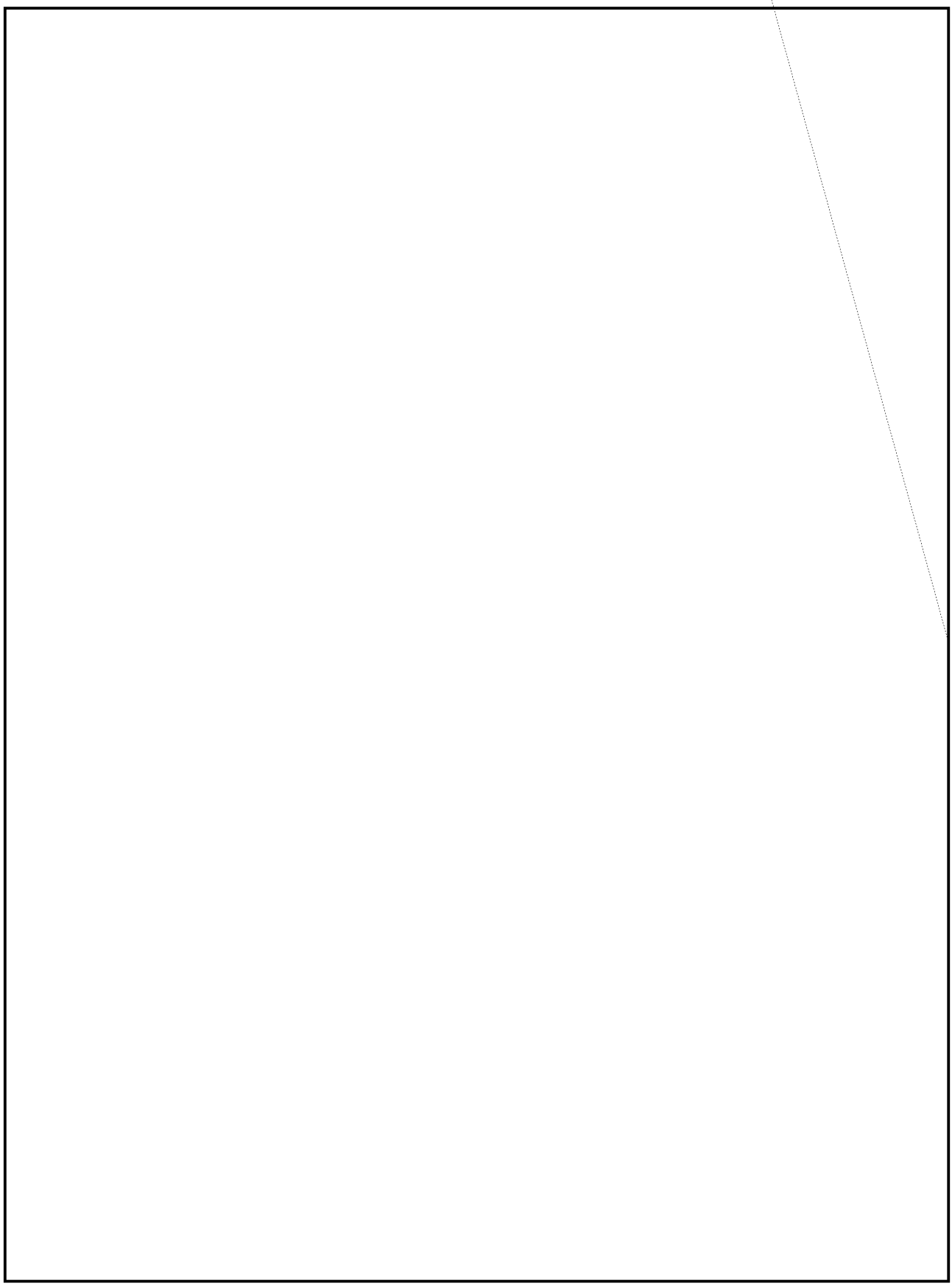
$$(3, 1) \qquad P(x_i/S+x_i) : P(x_j/S+x_j) = \frac{1+f_i}{N+1} : \frac{1+f_j}{N+1}$$

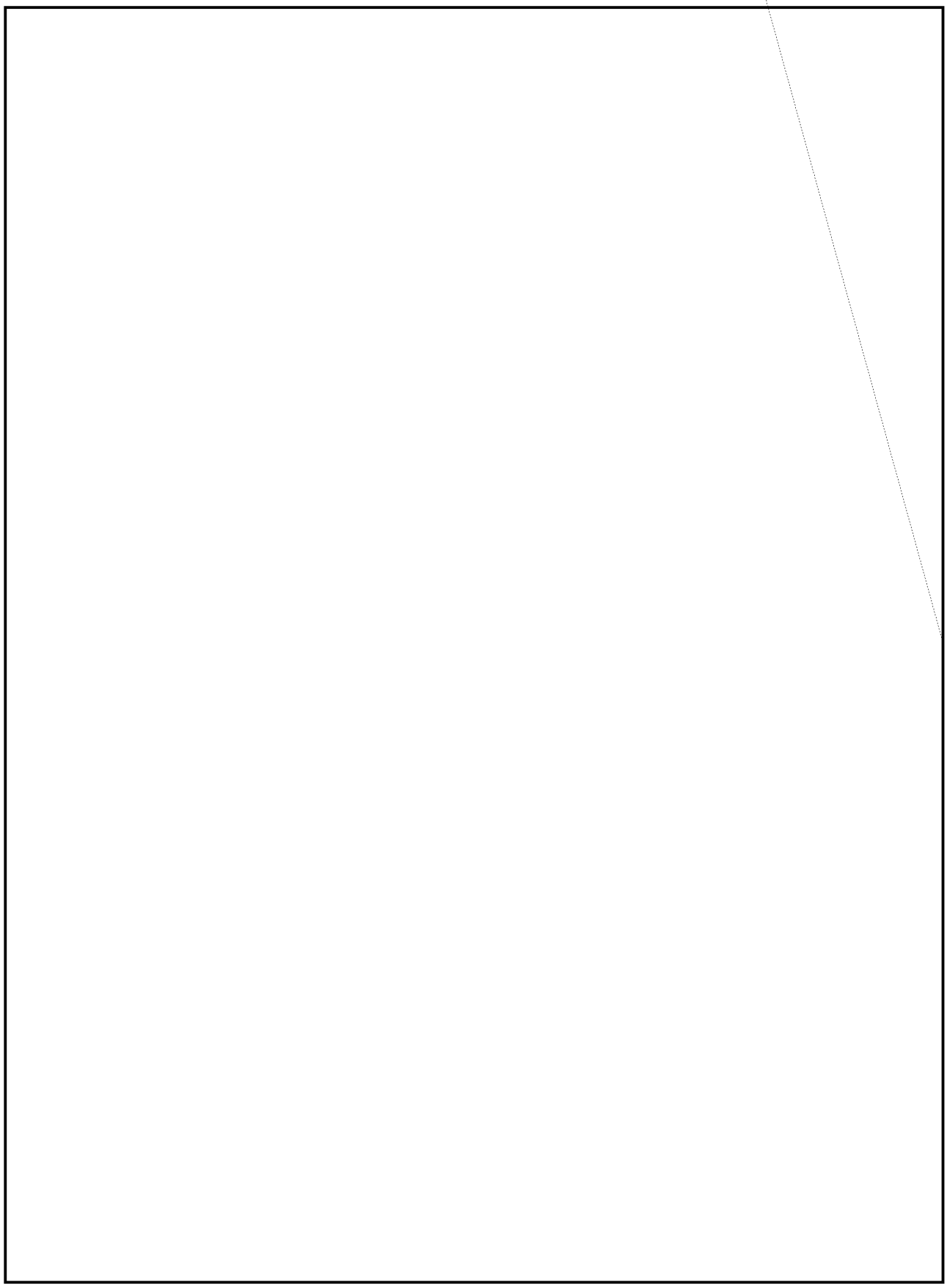
PL 86-36/50 USC 3605
EO 3.3(h)(2)

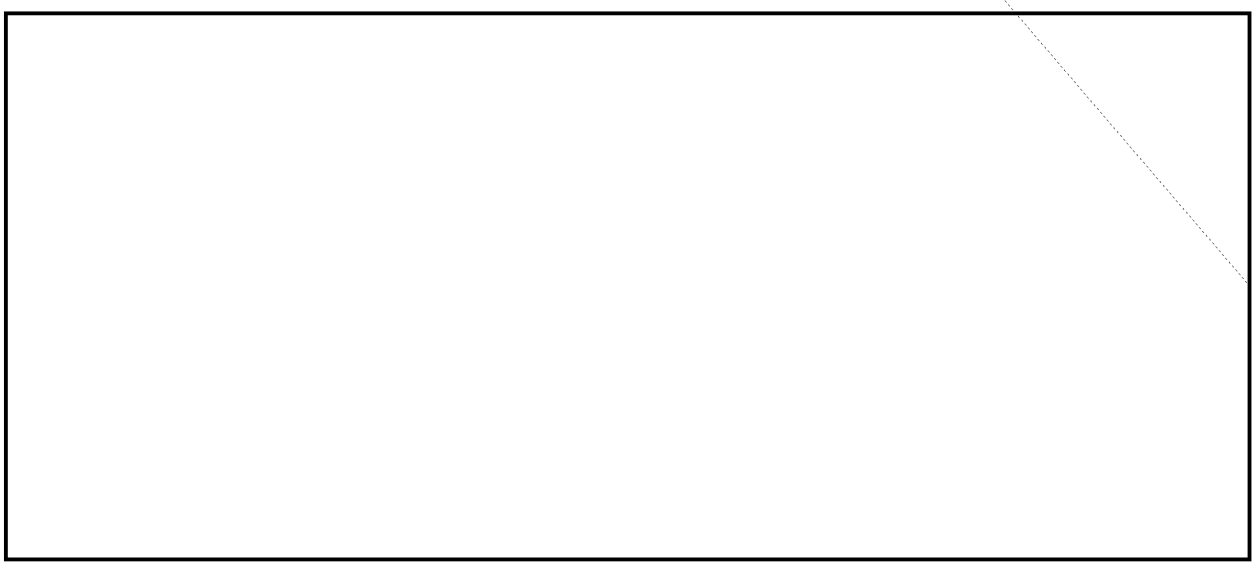
Thus the odds on the various letters are $1 + f_1: 1 + f_2: 1 + f_3: \ldots$.

The use of this estimate of the universe becomes very important if some $f_j = 0$. For in that case unreasonable log weights of ∞, or even $\infty - \infty$, may arise if the estimate f_i is used. See (29) for a longer discussion of the conditions under which the modified counts may be used.

3, 2 Code Groups.







4. Some Matrix Definitions and Properties.

A matrix is a rectangular array of numbers. Matrices are recurring items in the statistics of cryptanalysis, as will be seen in section 5.

Three ways of indicating a 3 x 2 matrix (the number of rows is always given first) are

(4, 1)
$$\begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \\ a_{20} & a_{21} \end{pmatrix} = (a_{ij}) = A.$$

4, 1 Elementary Properties of Matrices.

Two matrices of the same size can be added by adding corresponding elements,

(4, 1, 1)
$$A + B = (a_{ij}) + (b_{ij}) = (a_{ij}+b_{ij})$$

$$\begin{pmatrix} 3 & 2 \\ 0 & -1 \\ 1 & -1 \end{pmatrix} + \begin{pmatrix} -2 & 0 \\ 1 & 2 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \\ 2 & -2 \end{pmatrix}.$$

In talking about matrices as entities like numbers we need a word to describe numbers themselves. We call them "scalars". Scalar multiples of a matrix are defined by

(4, 1, 2)
$$mA = (m\, a_{ij}),$$

that is, multiply each element by the number m. For example,

$$3 \begin{pmatrix} 1/3 & 2 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 6 \\ -3 & 0 \end{pmatrix}.$$

The scalar product on the right is defined the same way,

$$A\, m = m\, A$$

The matrix of which each element is zero is represented by 0.

A combination m A + n B of two matrices A and B is called a "linear" combination. This can be extended to a linear combination of any number of matrices $\sum_{i=1}^{c} m_i A_i$. If some linear combination of a set of matrices is zero, $\sum_{i=1}^{c} m_i A_i = 0$, then the set is said to be "linearly dependent". If every linear combination (except that with 0 coefficients) of a set is different from zero the set is said to be "linearly independent".

Examples: The 1 x 2 matrices (1, 3) and (2, 6) are linearly dependent. The 1 x 3 matrices (1, 1, 5), (–2, 1, –1), and (–1, 2, 6) and linearly independent.

The product of two matrices can be defined if the number of columns of the first is the same as the number of rows of the second. It is defined thus: if A is a k x m matrix, and B an m x n, then

(4, 1, 3) $$AB = \left(\sum_{r=0}^{m-1} a_{ir} b_{rj} \right)$$

and is a k x n matrix. Each row of A is multiplied by each column of B and summed to give an element of the product. In general BA ≠ AB.

$$\begin{pmatrix} 3 & 2 \\ 0 & -1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 2 \\ 2 & 1 & -2 \end{pmatrix} = \begin{pmatrix} 7 & 5 & 2 \\ -2 & -1 & 2 \\ -1 & 0 & 4 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 1 & 2 \\ 2 & 1 & -2 \end{pmatrix} \begin{pmatrix} 3 & 2 \\ 0 & -1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 5 & -1 \\ 4 & 5 \end{pmatrix}.$$

It can be shown by straight-forward computation that

A(B+C) = AB+AC and (B+C)A = BA+CA. Also A(BC) = (AB)C.

The square matrix I = (δ_{ij}), where $\delta_{ii} = 1$ and $\delta_{ij} = 0$ for $i \neq j$, is called the "identity". I. The "identity" has 1's on the principal diagonal and 0's elsewhere. It has the property that IB = B = BI for each matrix B, when these products are defined. The matrix O with every element 0 is called "zero".

A matrix with only one column, m x 1, or with only one row, 1 x n, is called a "vector". Vectors are especially important in the theory of matrices. If v is a column vector, m x 1, and if M is a k x m matrix, then Mv = y is a k x 1 column vector. If M is square m x m then Mv has the same dimensions as v.

4, 2 Determinants.

A "determinant" is a number derived from a square matrix by applying certain rules of manipulation. These rules are; form every product possible by selecting exactly one element from each row and column, and add all these products with proper signs. The signs to be used are determined by a rule which may seem complex when first encountered. Half of the signs are positive and half negative. If the elements in a product are arranged in the order of the columns from which they come, then the rows from which they come are in a permuted order. If this permutation is odd the sign is negative. If it is even the sign is positive. An explanation

of when a permutation is odd is in order here. A permutation or rearrangement can be accomplished (sometimes in several ways) by successively interchanging pairs of elements. If the number of interchanges is odd the permutation is odd. See reference (11) for more details. If A is a square matrix, the determinant with the same elements will be written as $|A|$. For example, $|I| = 1$, and $|O| = 0$. It can be shown that $|AB| = |A| \cdot |B|$.

In general $|A+B| \neq |A|+|B|$. For instance, if $A = \begin{pmatrix}10\\00\end{pmatrix}$ and $B = \begin{pmatrix}00\\01\end{pmatrix}$, then $A+B = \begin{pmatrix}10\\01\end{pmatrix}$, and $\begin{vmatrix}10\\01\end{vmatrix} = 1$, while $|A| = 0 = |B|$.

It can be shown that a determinant vanishes if and only if its columns are linearly dependent. That is, if $|a_{ij}|$ is the c x c determinant, $|a_{ij}| = 0$ if and only if there exist c numbers, v_j, not all zero,

$$\text{such that } \sum_{j=0}^{c-1} a_{ij} v_j = 0 \text{ for each i.} \qquad (4, 2, 1)$$

Examples of the application of this last property are as follows:

$\begin{pmatrix}x & x\\y & y\end{pmatrix} = 0$, since $v_0 = 1$, $v_1 = -1$ will satisfy (4, 2, 1);

Also $\begin{pmatrix}x & y\\x & y\end{pmatrix} = 0$, since $v_0 = y$, $v_1 = -x$ will satisfy (4, 2, 1);

$$\begin{pmatrix}2 & 3 & 2\\-5 & 1 & 2\\3 & +2 & -4\end{pmatrix} = 0.$$

The coefficients v_0, v_1, v_2 can be determined by the student as an exercise.

If $A = (a_{ij})$ is an n x n matrix then (ma_{ij}) is also a matrix and $|ma_{ij}| = m^n|a_{ij}|$. For in evaluating the determinant each product has n factors of m, giving each term a factor of m^n.

If each element of a particular column of A is multiplied by m then the determinant is multiplied by m. For each product will have one factor of m, so the sum has a factor of m. Similarly, if each element of a row is multiplied by m, then the determinant is multiplied by m.

If two columns of a determinant are interchanged then the sign of the determinant is changed. To see this look back at the rule for determining the signs. The interchange of two columns puts an additional interchange into each permutation, thus changing odd to even and even to odd. Thus all the signs are changed, changing the sign of the sum. The same argument holds for the interchange of two rows.

A logical consequence of this property is that a determinant with two identical columns must be zero. For the interchange of those two columns does not change anything, and yet requires that the sign of the determinant be changed. This can only happen if it is zero. The presence of two identical rows also implies that the determinant is zero. Consequently if one column (or row) is a multiple of another the determinant is zero.

If the elements of one column of a determinant are each the sum of two numbers then the determinant can be expressed as the sum of two determinants. That is,

$$\begin{pmatrix}a_1 & b_1+x_1 & c_1\\a_2 & b_2+x_2 & c_2\\a_3 & b_3+x_3 & c_3\end{pmatrix} = \begin{pmatrix}a_1 & b_1 & c_1\\a_2 & b_2 & c_2\\a_3 & b_3 & c_3\end{pmatrix} + \begin{pmatrix}a_1 & x_1 & c_1\\a_2 & x_2 & c_2\\a_3 & x_3 & c_3\end{pmatrix}.$$

For in the expansion of the first by the definition each term would have a binomial factor which means that the term can be written as the sum of two terms, thus

$$a_i(b_j+x_j)c_k = a_ib_jc_k+a_ix_jc_k .$$

If we collect the first of each of these pairs of terms we get a determinant, and those left form a determinant. The analogous fact holds for rows.

Given a determinant, it can be modified by adding a multiple of any column to any other column without changing its value. This follows since the resultant determinant can be considered to be the sum of two determinants, one of which is the original and the other is zero. The same holds for rows. This theorem is very useful in calculating the value of determinants, using the result of the next paragraph. The direct computation from the definition is not ordinarily useful, since n! terms are involved, each a product of n factors. With n larger than 4 this is a formidable amount of work.

If a determinant has nothing but zeros above the principle diagonal it is said to be in "triangular" form. For example $\begin{vmatrix} 1 & 0 & 0 \\ 2 & 5 & 0 \\ 9 & 7 & 2 \end{vmatrix}$. The value of a triangular determinant is the product of its diagonal elements. For in the expansion by the definition each term is zero except one. Those terms using any element of the first row except the first element must be zero. Those using the first term of the first row must use the second term of the second row or be zero, and so forth.

By adding multiples of columns to other columns, or of rows to other rows, it is possible to reduce the determinant to an equal triangular determinant which is easy to evaluate.

Exercise: $\begin{vmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{vmatrix}$

Exercise: Show that this determinant is zero.

$$\begin{vmatrix} -15 & 13 & 3 & -9 \\ 6 & 8 & -6 & 7 \\ 3 & 9 & 9 & 5 \\ 7 & 8 & 5 & 8 \end{vmatrix}$$

4, 3 Inverses and Conjugate Transposes of Matrices.

If $AB = I = BA$, the matrix B is called the "inverse" of A, and is written $B = A^{-1}$. The matrices A and B must be square and the same size. Some matrices have no inverse; these are called "singular". For instance,

$$\begin{pmatrix} 0 & 1/2 \\ 3 & -3 \end{pmatrix}^{-1} = \begin{pmatrix} 2 & 1/3 \\ 2 & 0 \end{pmatrix}$$

$$\begin{pmatrix} -1 & 2 \\ 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} -1 & 2 \\ 0 & 1 \end{pmatrix}$$

The matrix $S = \begin{pmatrix} 1 & -1 \\ -2 & 2 \end{pmatrix}$ has no inverse. For $S \begin{pmatrix} x & y \\ z & w \end{pmatrix} = \begin{pmatrix} x-z & y-w \\ -2(x-z) & -2(y-w) \end{pmatrix}$ and obviously $x-z = 1$ and $-2(x-z) = 0$ are not both possible.

The matrix $(a_{ij})^* = (\bar{a}_{ji})$, is called the "conjugate transpose" of (a_{ij}). It is the result of interchanging rows with columns and taking the complex conjugate of each element. The conjugate transpose has several obvious properties.

$(A+B)^* = A^*+B^*$, and $(AB)^* = B^*A^*$. Also $(A^*)^* = A$.

$$\begin{pmatrix} -1 & 2 \\ 0 & 1 \end{pmatrix}^* = \begin{pmatrix} -1 & 0 \\ 2 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 3 & i \\ 1+i & 0 \\ 1 & 1-i \end{pmatrix}^* = \begin{pmatrix} 3 & 1-i & 1 \\ -i & 0 & 1+i \end{pmatrix}$$

If A is any matrix, then $S = A^*A$ has the special property that $S^* = S$. Any matrix with this special property is called "Hermitian".

$$\begin{pmatrix} 3 & i \\ 1+i & 0 \\ 1 & 1-i \end{pmatrix} \begin{pmatrix} 3 & 1-i & 1 \\ -i & 0 & 1+i \end{pmatrix} = \begin{pmatrix} 10 & 3-3i & 2+i \\ 3+3i & 2 & 1+i \\ 2-i & 1-i & 3 \end{pmatrix}$$

$$\begin{pmatrix} 3 & 1-i & 1 \\ -i & 0 & 1+i \end{pmatrix} \begin{pmatrix} 3 & i \\ 1+i & 0 \\ 1 & 1-i \end{pmatrix} = \begin{pmatrix} 12 & 1+2i \\ 1-2i & 3 \end{pmatrix}$$

Looking at the determinant $|A^*|$, it is clear that transposing does not affect it, while taking the conjugate of each element causes the determinant to be conjugated. Thus $|A^*| = |\bar{A}|$. For S above $|S| = |S^*| = |\bar{S}|$ whence the determinant is real.

$$|A^*A| = |\bar{A}| \cdot |A| \geq 0.$$

The inverse also has several simple properties. For one, $(AB)^{-1} = B^{-1} A^{-1}$. Another simple rule is $(A^*)^{-1} = (A^{-1})^*$. The inverse has no simple rule for sums, and in general $(A+B)^{-1} \neq A^{-1} + B^{-1}$. If A is such that $A^* = A^{-1}$, that is, $A^*A = I$, it is called "orthogonal." The matrices

$$\begin{pmatrix} 3/5 & -4/5 \\ 4/5 & 3/5 \end{pmatrix} \text{ and } \begin{pmatrix} i & \sqrt{2} \\ -\sqrt{2} & i \end{pmatrix}$$

are each orthogonal. If A has an inverse, A^{-1}, then $|A^{-1}| = 1/|A|$. For $|A| \cdot |A^{-1}| = |AA^{-1}| = |I| = 1$. This shows that if $|A| = 0$ then A has no inverse. It can be shown that if a square matrix A has no inverse then $|A| = 0$. Thus A is singular if and only if $|A| = 0$.

4, 4 Vectors.

An important special case of a matrix is the vector, which is n x 1 or 1 x n. That is, $\begin{pmatrix} 2 \\ 0 \end{pmatrix}$ and $(1, i, -1)$ are vectors. We will usually mean the vertical version. The product of a matrix by a vector, if defined, is another vector, $Mv = y$. If $M = \begin{pmatrix} 3 & 2 \\ 0 & -1 \\ 1 & -1 \end{pmatrix}$ and $v = \begin{pmatrix} a \\ b \end{pmatrix}$, then $Mv = \begin{pmatrix} 3a+2b \\ -b \\ a-b \end{pmatrix}$. If $Mu = x$ and $Mv = y$, then M takes a linear combination of u and v into the same combination of x and y.

(4, 4, 1) $\quad M(au+bv) = Mau+Mbv = aMu+bMv = ax+by.$

Any transformation $u \rightarrow x$ and $v \rightarrow y$ which has the property that

$$au+bv \rightarrow ax+by$$

is called a "linear" transformation. Any linear transformation can be represented by matrices.

It sometimes happens that for a square matrix A and vector v we have $Av = \lambda v$, where λ is a scalar. That is, A effectively only "stretches" v. The vector v is called an "eigenvector" of A, and λ is called the corresponding "eigenvalue."

If $A = \begin{pmatrix} 3 & 2 \\ 0 & -1 \end{pmatrix}$ then $v = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$ and $\lambda = -1$ are an eigenvector and associated eigenvalue of A. Another pair for A is $v = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\lambda = 3$.

A multiple of an eigenvector is also an eigenvector with the same eigenvalue. $A \cdot mv = mAv = m\lambda v = \lambda \cdot mv$. If two eigenvectors of a matrix have the same eigenvalue, then any linear combination of the two is also an eigenvector. If $Au = \lambda u$ and $Av = \lambda v$, then $A(mu+nv) = mAu+nAv = m\lambda u+n\lambda v = \lambda\,(mu+nv)$. For instance, $A = \begin{pmatrix} 2 & 0 & -1 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{pmatrix}$ has the eigenvectors $u = \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix}$ $v = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, with eigenvalue $\lambda = 2$. Then $\begin{pmatrix} 3 \\ 2 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 0 \end{pmatrix}$ is an eigenvector with eigenvalue 2, and $\begin{pmatrix} 3m \\ 2m+n \\ 0 \end{pmatrix}$ is also.

If A is a matrix and if there exists a vector $v \neq 0$ such that $Av = 0$ then A is singular. For the assumption that A^{-1} exists leads to the impossible equation,

$$v = Iv = (A^{-1}A)v = A^{-1}(Av) = A^{-1}\,0 = 0.$$

The converse can be proved, but not here. The theorem can then be stated:

A matrix A annihilates some vector if and only if A is singular.

For if $Av = 0$, where $v \neq 0$, then $\sum_{j=o}^{c-1} a_{ij}v_j = 0$, and conversely. This is a necessary and sufficient condition that $|A| = 0$, as was stated at the end of 4, 2. For instance, $A = \begin{pmatrix} 1 & -2 \\ -1 & 2 \end{pmatrix}$ annihilates $v = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$. We can now state: A matrix is singular if and only if its determinant is zero.

Now if	$Av = \lambda v$, then
	$Av - \lambda v = 0$
and	$(A-\lambda I)v = 0.$
Therefore	$\lvert A-\lambda I \rvert = 0.$
Conversely, if	$\lvert A-\lambda I \rvert = 0,$

then there exists a vector v such that $Av = \lambda v$, by the last statement of section 4, 2.

Thus a necessary and sufficient condition for λ to satisfy a relation $Av = \lambda v$ (for some v) is that λ satisfy the equation $|A-\lambda I| = 0$.

The expression $|A-\lambda I|$ can be expanded by the rules of determinants, and gives a polynomial in λ of degree c, if A is a c x c matrix. For example, the eigenvalues of $A = \begin{pmatrix} 1 & -2 \\ -1 & 2 \end{pmatrix}$ are found by solving $\begin{vmatrix} 1-\lambda & -2 \\ -1 & 2-\lambda \end{vmatrix} = \lambda^2-3\lambda = 0$. They are $\lambda = 0$ and $\lambda = 3$. By the fundamental theorem of algebra, the equation $|A-\lambda I| = 0$ has at least one root λ, and consequently A has at least one eigenvector. In general A has c eigenvalues, not necessarily all distinct. The matrix $\begin{pmatrix} 3 & -1 \\ 1 & 1 \end{pmatrix}$ has only the one eigenvalue $\lambda = 2$, instead of two as might be expected. The associated eigenvector is $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$. For each eigenvalue distinct from the others, the matrix must have an eigenvector linearly independent of the remaining. If A has c different eigenvalues then it has c independent eigenvectors.

For any vectors u and v of the same dimensions the product u*v is well-defined, and is a 1 x 1 matrix, not ordinarily distinguished from a scalar. It is frequently called the "dot product" of the vectors, u·v, and for the 3-dimensional case is the product of the lengths times the cosine of the angle between them.

For example, if $u = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$, and $v = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$, then $u^*v = (1,0,-1)\begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix} = -1$. Or if $x = \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}$, $x^*v = 0$. If $u^*v = 0$, the vectors u and v are said to be "orthogonal" or at right angles to each other. The quantity $\sqrt{v^*v}$ is called the "length" of the vector.

If x and y are vectors orthogonal to a given vector v, then any linear combination

(4, 4, 2) $$ax+by$$

is also orthogonal to v. In the example of the previous paragraph, x is orthogonal to v. The vector $y = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$ is also orthogonal to v. Therefore $ax+by = \begin{pmatrix} a-b \\ 3a+b \\ a+b \end{pmatrix}$ is also orthogonal to v. The collection of all linear combinations (4, 4, 2) is a "plane" perpendicular to v. The linear combinations of more than 3 vectors is called a "hyperplane" perpendicular to v. The collection of all vectors perpendicular to v is "the" hyperplane perpendicular to v. An application of these concepts is found in section 5, 3.

Problem: If u is a vector then uu^* is a matrix. What are its eigenvectors, and their associated eigenvalues?

Problem: If M is a matrix with the eigenvectors v_k, $Mv_k = v_k\lambda_k$, what are those of the matrix $I-M$?

Problem: Find the square of $I-uu^*$.

Problem: If M has eigenvectors v_k, what are the eigenvectors of M^2? What are the associated eigenvalues?

Problem: If M has eigenvectors v_k, and if f(x) is a polynomial, then f(M) is a matrix. What are its eigenvectors and eigenvalues?

Problem: If u and v are two vectors, what is $u^*v - v^*u$?

4, 5 Geometry.

The equation $v = u\alpha$, where α is a scalar parameter, is a parametric equation for a straight line through the origin having direction u. The equation $v = u\alpha + w$ represents a line through w parallel to u. It will be perfectly general and also convenient to take u of length 1, $u^*u = 1$.

If r and s are two vectors and d the distance between them then r-s is the vector from s to r, and $d^2 = (r-s)^*(r-s) = r^*r - 2r^*s + s^*s$.

The distance from the line $v = u\alpha$ to the point (vector) r can be found by first finding the vector s through r and perpendicular to the line, and then finding its length. Then $s = r - u\alpha$ for some α and $u^*s = 0$. To determine α take $0 = u^*s = u^*(r-u\alpha) = u^*v - u^*u\alpha$ whence $\alpha = u^*r$, using $u^*u = 1$. Now the length of s is $\sqrt{s^*s} = \sqrt{(r-uu^*r)^*(r-uu^*r)} = \sqrt{r^*(I-uu^*)^*(I-uu^*)r} = \sqrt{r^*(I-uu^*)r}$.

The line L through the point w with the direction of the vector u is $v = w + u\alpha$, where α takes on any scalar value. The distance d from L to r can be found by making the transformation $v' = v - w$, whence L becomes $v' = u$ and $r' = r - w$. Then $d^2 = (r-w)^*(I-uu^*)(r-w)$.

4, 6 The Line of Regression.

In section 1, 8 the line of regression was defined as that straight line which fitted the data best. Now that we have introduced the machinery of vectors and matrices this can be handled in the general case and a formula found for the line of regression.

Suppose the data given is the set of vectors v_k, $k = 1,2, \ldots, m$, each real and of dimension n. For example census data, each datum being the age, income, height, etc. of an individual.

Let $\bar{r} = \frac{1}{m}\sum_{k=1}^{m} v_k$. Let $M = \sum_{k=1}^{m} (r_k - \bar{r})(r_k - \bar{r})^*$, which will be needed later.

If the line of regression goes through the point w with the direction u then the square of the distance from r_k to the line is $(r_k-w)^*(I-uu^*)(r_k-w)$, and the sum of the squares of these distances is

(4, 6, 1)
$$\sum_{k=1}^{m} (r_k-w)^*(I-uu^*)(r_k-w)$$

$$= \sum_{k=1}^{m} (r_k-w)^*(r_k-w) - \sum_{k=1}^{m} (r_k-w)^* uu^*(r_k-w)$$

$$= t(M) - \sum_{k=1}^{m} u^*(r_k-w)(r_k-w)^*u$$

$$= t(M) - u^*Mu.$$

The restriction $u^*u = 1$ can be imposed on u for convenience. Now introduce the Lagrange multiplier λ and minimize

$$S(u,w) = \sum_{k=1}^{m} (r_k-w)^*(r_k-w) - u^* \sum_{k=1}^{m} (r_k-w)(r_k-w)^*u + \lambda\, u^*u.$$

A necessary condition for a minimum is that $\nabla_u S = 0$ and $\nabla_w S = 0$.

$$0 = \nabla_u S = -2 \sum_{k=1}^{m} (r_k - w)(r_k - w)^* u + 2\lambda u$$

$$0 = \nabla_w S = -2 \sum_{k=1}^{m} r_k + 2mw - \Sigma 2 u r_k^* u + 2muw^* u.$$

These reduce to

$$Mu = \lambda u$$

(4, 6, 2)

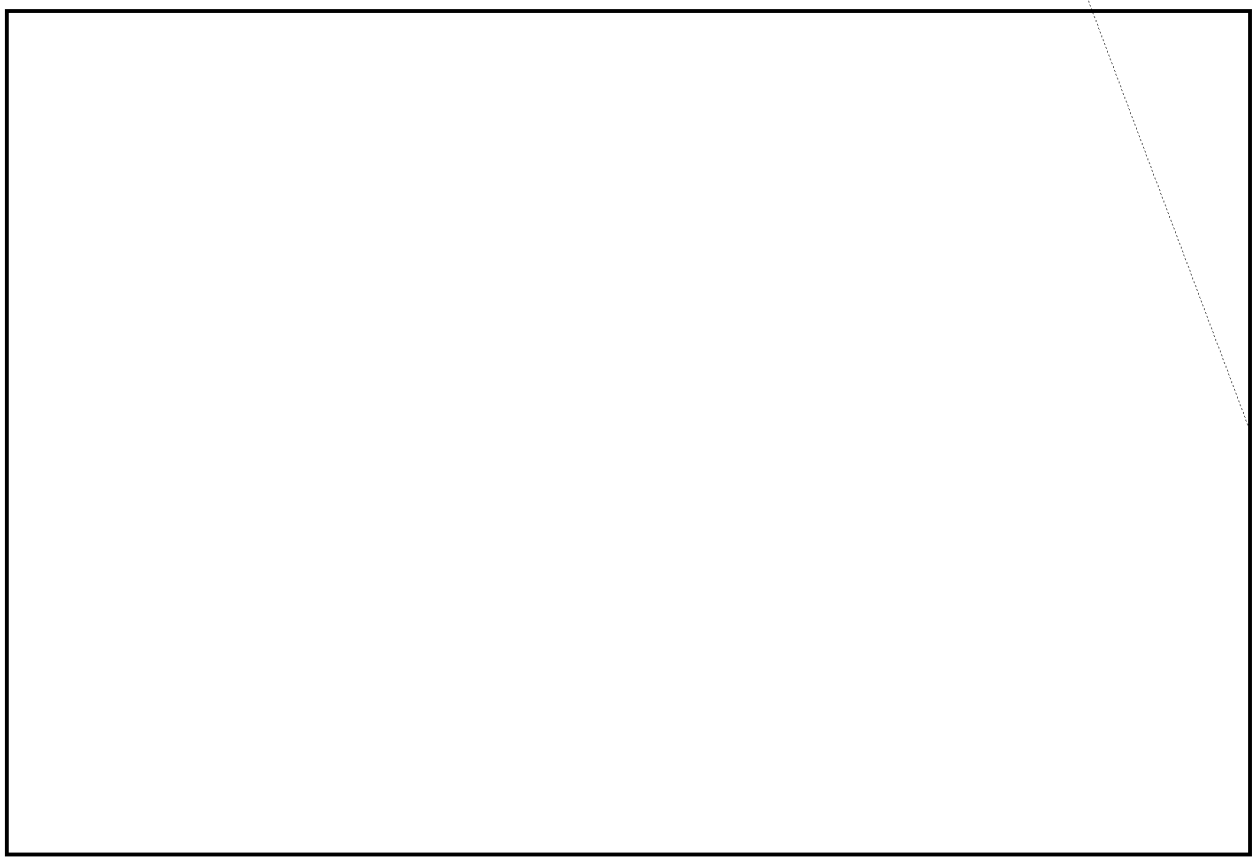
$$(I - uu^*)(\bar{r} - w) = 0.$$

The first of these says that u is an eigenvector of M and λ the corresponding eigenvalue. Multiplying both sides by u^* we get $u^*Mu = \lambda$ so that (4, 6, 1) becomes $S(u,w) = t(M) - \lambda$, from which we infer that λ is the largest eigenvalue of M.

The second equation shows that the distance of $\bar{r}$ from the line, $(\bar{r} - w)^*(I - uu^*)(\bar{r} - w)$, is 0. Thus the line goes through $\bar{r}$, the center of gravity.

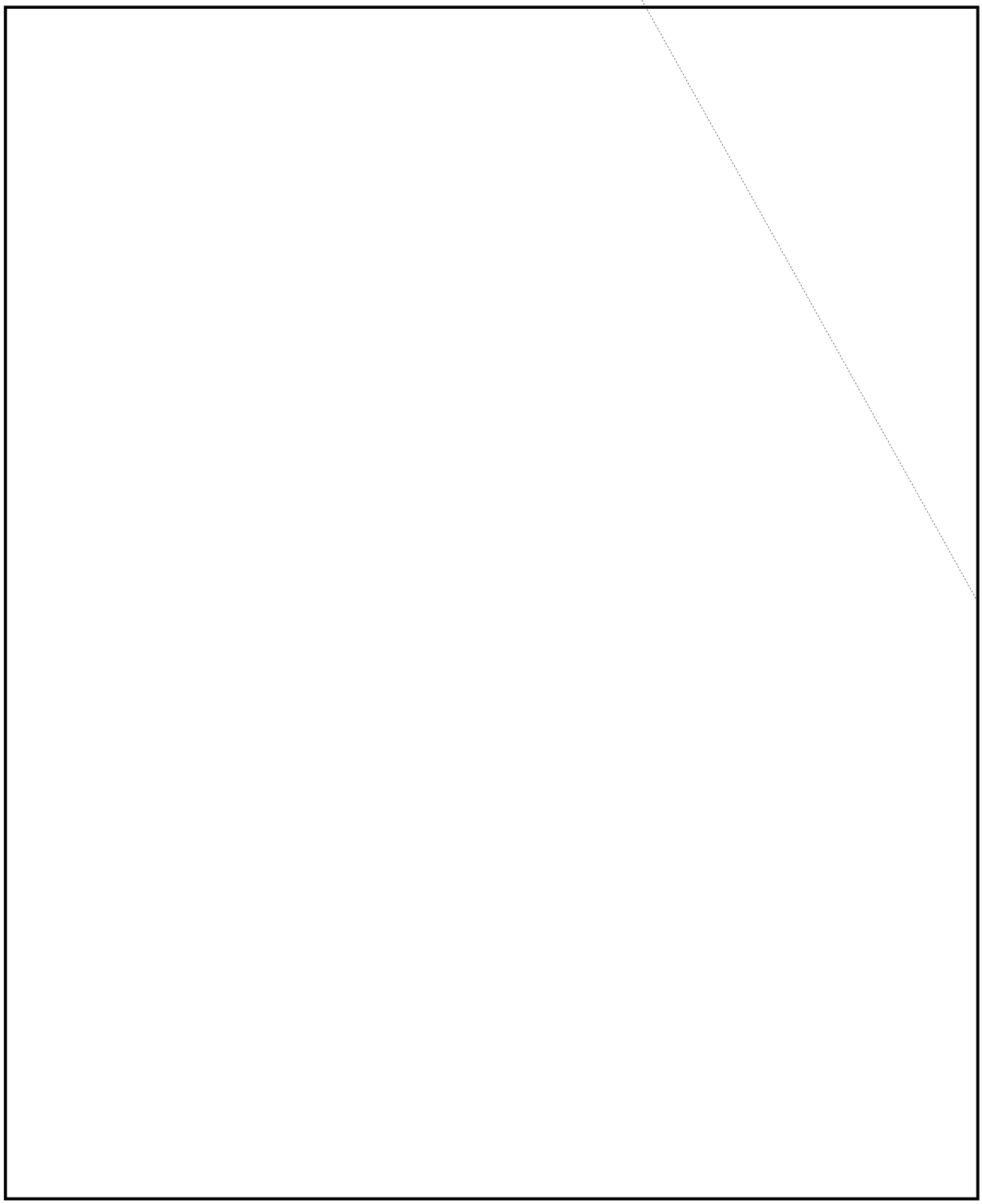
4, 7 Examples of Cryptologic Applications.

~~CONFIDENTIAL~~



~~CONFIDENTIAL~~

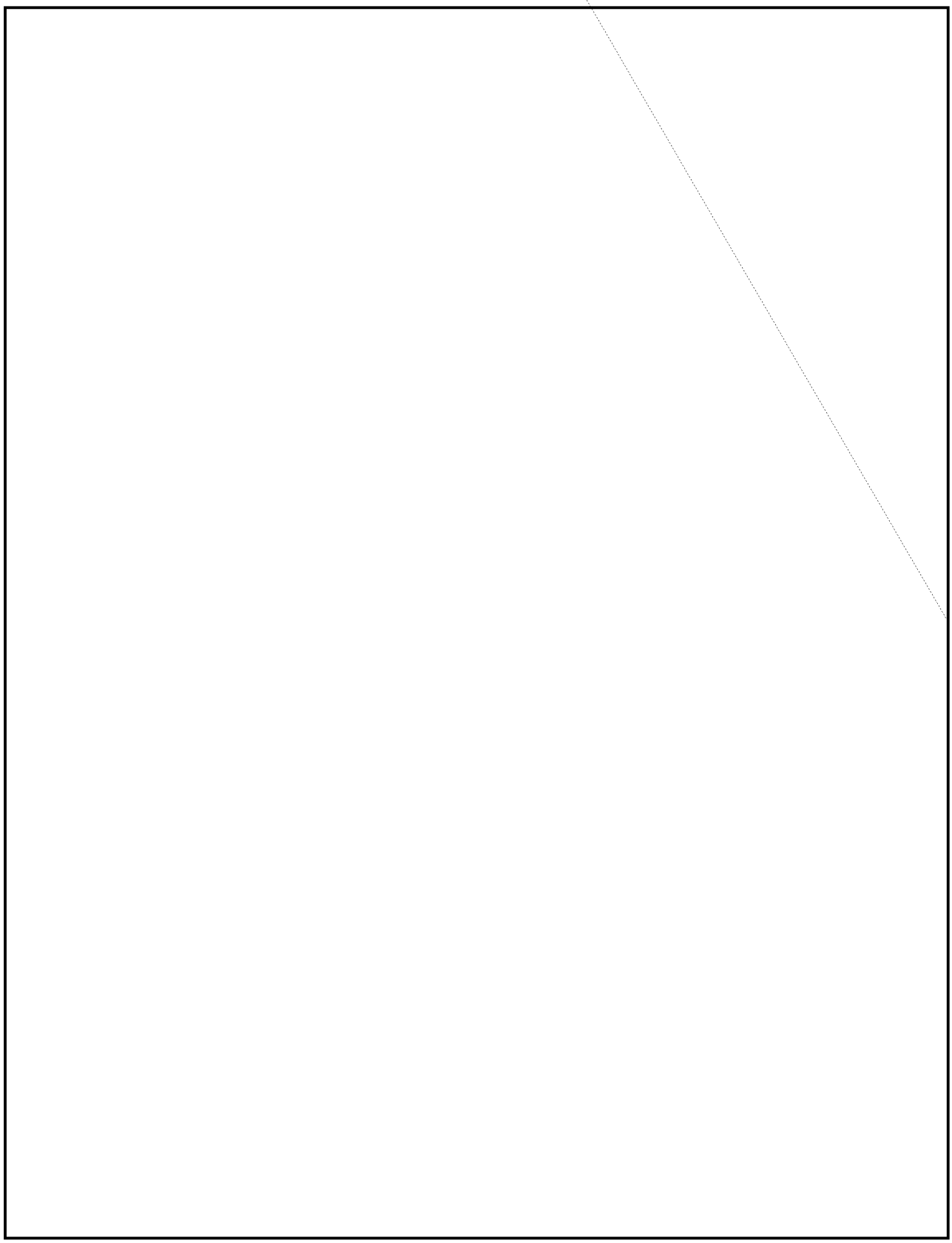
PL 86-36/50 USC 3605
EO 3.3(h)(2)



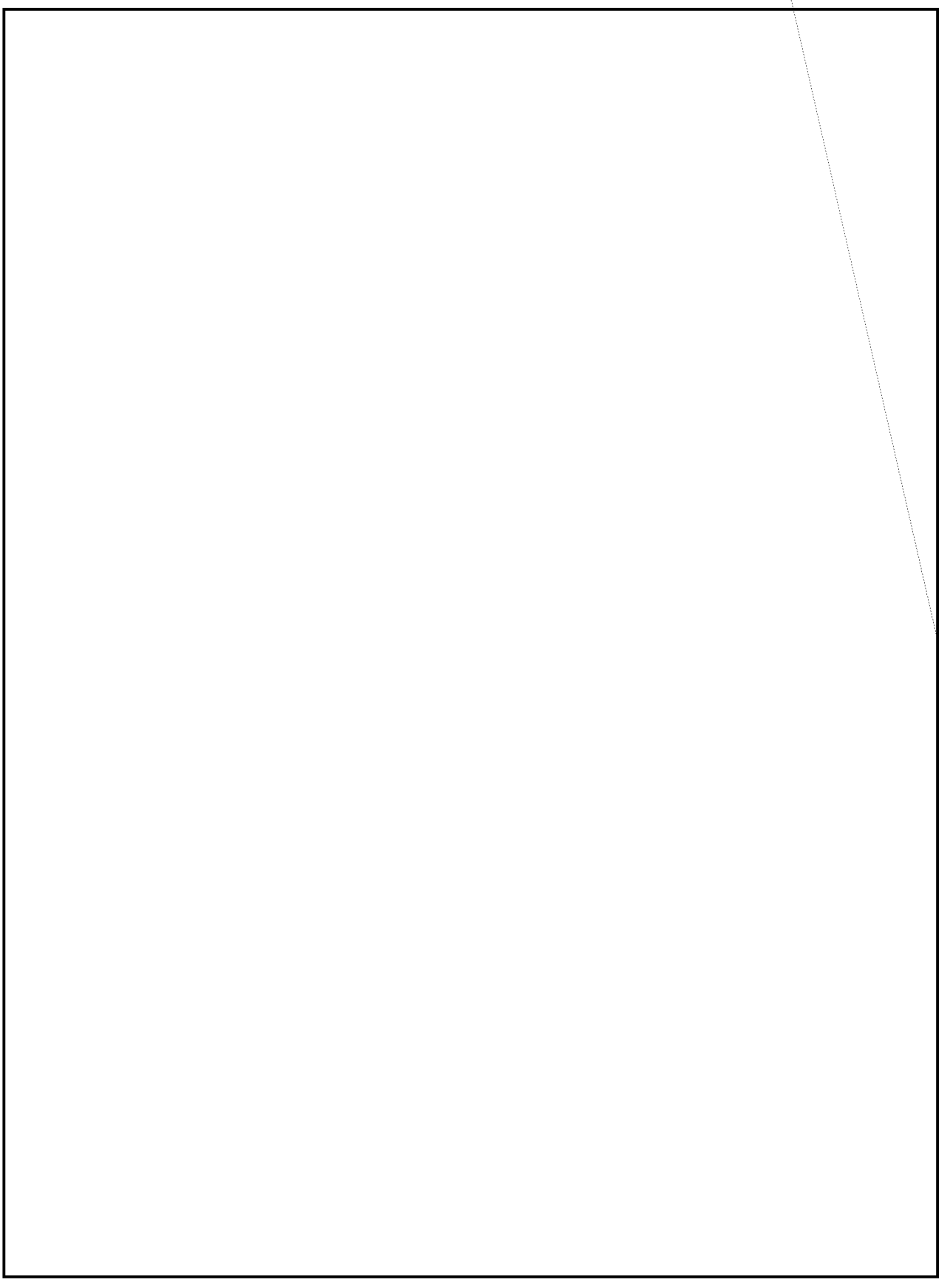
PL 86-36/50 USC 3605
EO 3.3(h)(2)

~~CONFIDENTIAL~~

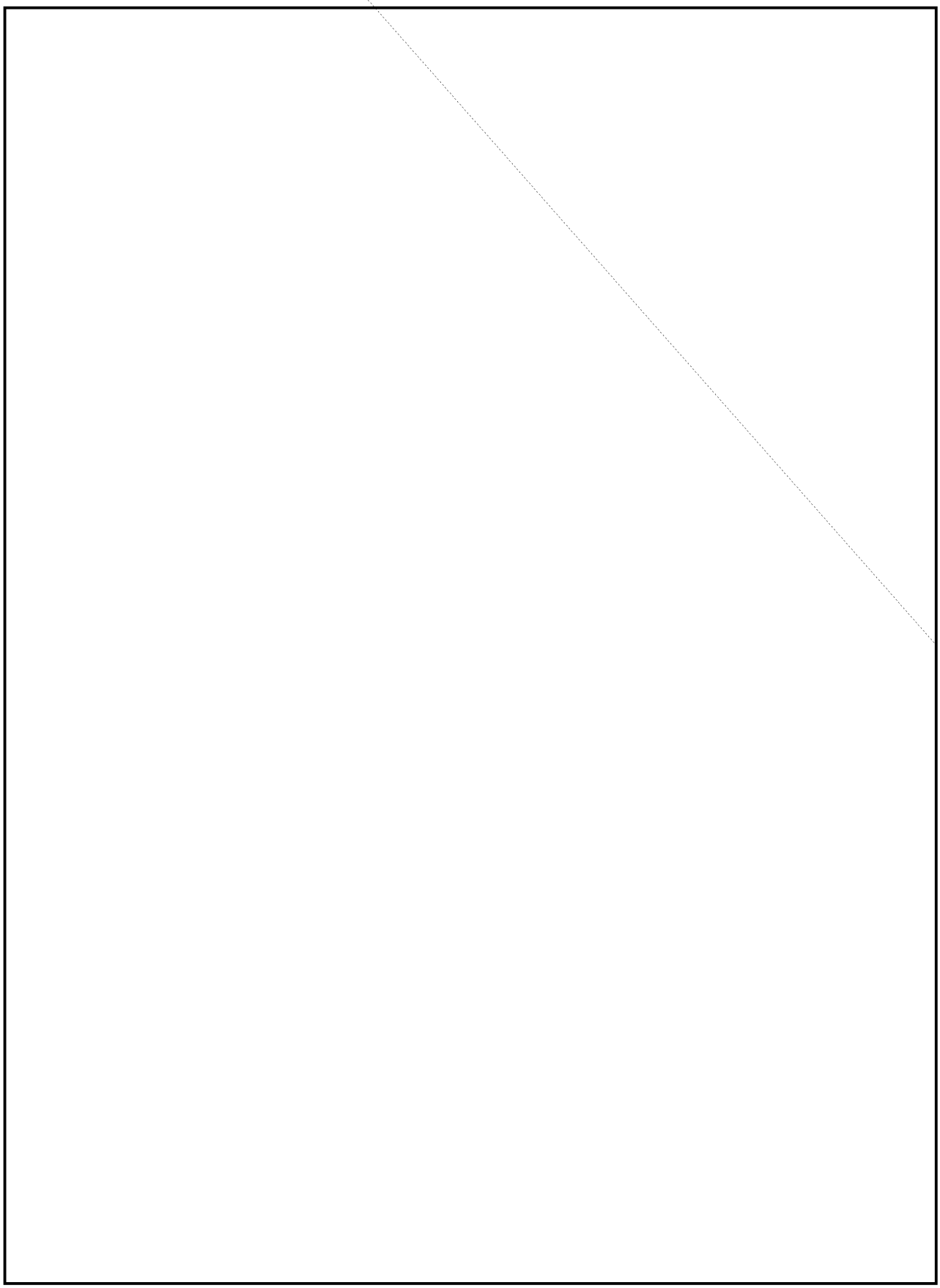
5. Flagging.



~~CONFIDENTIAL~~



~~CONFIDENTIAL~~



$$= \frac{1}{c} \sum_{j=0}^{c-1} \sum_{k=0}^{c-1} f_k e^{j(k-n)w}$$

$$= \frac{1}{c} \sum_{k=0}^{c-1} f_k \sum_{j=0}^{c-1} e^{j(k-n)w}$$

$$= f_n,$$

since

(6, 1, 4)
$$\sum_{j=0}^{c-1} e^{jmw} = \begin{cases} 0 \text{ if } m \neq 0 \\ \\ c \text{ if } m = 0. \end{cases}$$

The transformation (6, 1, 2) $F = \sum_{k=0}^{c-1} f_k e^{jkw}$ replaces the frequency distribution $\{f_i\}$ by an equivalent set of c numbers $\{F_j\}$. That it is equivalent is shown by equation (6, 1, 3), which computes $\{f\}$ in terms of $\{F\}$.

Example: $c = 4,\ f_0 = 3, f_1 = 2, f_2 = 1, f_3 = 3.$

$$w = \frac{2\pi i}{4} = \frac{\pi i}{2}.$$

$$\omega = e^w = e^{\frac{\pi i}{2}} = i.$$

Therefore $\omega^2 = -1, \omega^3 = -i$, and $\omega^4 = \omega^0 = 1.$

$$F_0 = f_0 + f_1 + f_2 + f_3 = 9$$

$$F_1 = f_0 + \omega f_1 + \omega^2 f_2 + \omega^3 f_3 = 2 - i.$$

$$F_2 = f_0 + \omega^2 f_1 + \omega^0 f_2 + \omega^2 f_3 = -1.$$

$$F_3 = f_0 + \omega^3 f_1 + \omega^2 f_2 + \omega f_3 = 2 + i.$$

(6, 1, 5)
$$\phi(t) = 1/4\,[9 + (2-i)e^{-t} - e^{-2t} + (2+i)e^{-3t}]$$

$$\phi(0) = 1/4\,[12] = 3 = f_0$$

$$\phi(w) = 1/4\,[9 - (2-i)i + 1 + (2+i)i] = 1/4\,[8] = 2 = f_1, \text{ etc.}$$

The coefficients F_i are in general complex imaginary, and satisfy the relation $F_{c-i} = \overline{F}_i$, see (6, 2, 12) below. Thus the number of degrees of freedom in the statistic is still c. Also notice that $F_0 = \sum_{i=0}^{c-1} f_i$ is the total number of letters in the count.

This "characteristic" or "Fourier transform" can be used instead of the frequency distribution. Anything which can be done with one can be done with the other, with perhaps less difficulty. There are situations in which a small subset of $\{F\}$ has most of the information of $\{f\}$ and is more easily handled. For instance, if c is even, $F_{c/2}$ is the "parity double bulge". There are favorable situations in which the parity (odd or even) alone is sufficient to betray the pattern of a cyclic component. In these situations $F_{c/2}$ is a sufficient statistic. In other situations similarly another coefficient F_i, or a set of two or three of them, may be sufficient. The reader who has used Fourier analysis for a curve fitting problem such as spectral analysis will recognize the method. The main difference is that here we are dealing with the mod 26 ring rather than a continuous range.

6, 2 Properties of the Fourier Transforms.

If $\{x\}$ and $\{y\}$ are two frequency distributions, and if $\{X\}$ and $\{Y\}$ are the corresponding Fourier transforms, then $\{f_j = x_j+y_j\}$ is a frequency distribution, and its Fourier transform $\{F\}$ is the sum of those of the components,

(6, 2, 1) $$F_k = X_k+Y_k.$$

This follows immediately from formula (6, 1, 2). As a consequence the transforms of the frequency distributions

(6, 2, 2) $$\{f^0\} = \{1,0,0,\ldots,0\}, \{f^1\} = \{0,1,0,\ldots,0\},$$

and so forth to $\{f^{c-1}\} = \{0, 0, 0, \ldots, 1\}$, can be calculated in advance, and the transform of any distribution found by adding the proper number of each.

Let $\{F^1\}$ be the corresponding transforms. Then

(6, 2, 3) $$\{F^0\} = \{1,1,1,\ldots,1\}$$
$$\{F^1\} = \{1,\omega,\omega^2,\ldots\omega^{c-1}\},$$

and so forth. In general $F_j^k=\omega^{jk}$, a complex value.

The example of 6, 1 done this way is as follows. The distribution

$\{1, 0, 0, 0\}$ has the Fourier transform $\{1, 1, 1, 1\}$

$\{0, 1, 0, 0\}$ has $\{1, i, -1, -i\}$

$\{0, 0, 1, 0\}$ has $\{1, -1, 1, -1\}$

$\{0, 0, 0, 1\}$ has $\{1, -i, -1, i\}$.

Now the transform of $\{3, 2, 1, 3\}$ is

$$3\{1, 1, 1, 1\} + 2\{1, i, -1, -i\} + \{1, -1, 1, -1\}$$
$$+ 3\{1, -i, 1, i\} = \{9, 2-i, -1, 2+i\}.$$

The expected value of F can readily be calculated.

(6, 2, 4) $$E(F_j) = \sum_{k=0}^{c-1} E(f_k)\omega^{jk},$$

since the expected value of a sum is the sum of the expected values.

Other moments can also be obtained from those of $\{f\}$, such as

$$E(F_jF_h) = \sum_{k=0}^{c-1} \sum_{m=0}^{c-1} E(f_kf_m)\omega^{jk+hm}. \tag{6, 2, 5}$$

If we assume that f is multinomially distributed (a frequent case)

then $$E(f_k) = \frac{N}{c}, \text{ where } N = \sum_{k=0}^{c-1} f_k,$$

and then

$$E(F_j) = \frac{N}{c} \sum_{k=0}^{c-1} \omega^{jk} = 0. \tag{6, 2, 6}$$

if $j \neq 0$,

$$F_0 = \sum_{k=0}^{c-1} f_k = N. \tag{6, 2, 7}$$

Since

$$E(f_kf_m) = \frac{N^2-N}{c^2}, \; k \neq m, \tag{6, 2, 8}$$

and

$$E(f_k^2) = \frac{N^2+(c-1)N}{c^2}, \tag{6, 2, 9}$$

we can calculate $E(F_jF_h)$.

$$E(F_jF_h) = \sum_{k=0}^{c-1} \sum_{m=0}^{c-1} E(f_kf_m)\omega^{jk+hm} \tag{6, 2, 10}$$

$$= \frac{N^2-N}{c^2} \sum_{m=0}^{c-1} \sum_{k \neq m} \omega^{jk+hm}$$

$$+ \frac{N^2+(c-1)N}{c^2} \sum_{k=0}^{c-1} \omega^{k(j+h)}.$$

This gives us five cases.

$$\begin{cases} E(F_0^2) = N^2 & \\ E(F_0F_h) = 0 & h \neq 0 \bmod c \\ E(F_jF_h) = N^2-N & j \neq 0, \; j+h \neq 0 \bmod c \\ E(F_jF_{-j}) = N & j \neq 0 \bmod c \\ E(F_j^2) = 0 & j \neq 0. \end{cases} \tag{6, 2, 11}$$

and

The condition $j+h \neq 0$ is a little surprising. Examination reveals that

(6, 2, 12) $$F_{c-j} = \sum_{k=0}^{c-1} f_k \omega^{jk} = \overline{F}_j,$$

the conjugate of F_j.

Now in dealing with complex statistics the definition of variance is

(6, 2, 13) $$\sigma^2(z) = E(z\bar{z}) - E(z)\overline{E(z)}.$$

Thus we have

(6, 2, 14) $$\sigma^2(F_j) = E(F_j\overline{F}_j) - E(F_j)\,\overline{E(F_j)}$$
$$= E(F_jF_{c-j}) - 0 = N \quad \text{for } j \neq 0,$$

(6, 2, 15) $$\sigma^2(F_0) = N^2 - N\,\overline{N} = 0.$$

The covariance of two complex variables is defined as

(6, 2, 16) $$\mu_{11}(x,y) = E(x\bar{y}) - E(x)\,\overline{E(y)}.$$

Thus

(6, 2, 17) $$\mu_{11}(F_jF_h) = E(F_j\overline{F}_h) - E(F_j)\,\overline{E(F_h)}$$
$$= (E(F_jF_{c-h}) - E(F_j)\,E(F_{c-h})$$
$$= \begin{cases} 0 & \text{if } jh = 0 \\ N & \text{if } j = h \neq 0 \\ N^2-N & \text{if } j \neq h \text{ and } jh \neq 0. \end{cases}$$

(6, 2, 18) $$\sum_{j=0}^{c-1} |F_j|^2 = c \sum_{k=0}^{c-1} f_k^2 = N^2\gamma.$$

Proof:
$$\sum_{j=0}^{c-1} |F_j|^2 = \sum_j F_jF_{-j}$$
$$= \sum_j \sum_k \sum_n f_kf_n\omega^{j(k-n)} = c \sum_{k=0}^{c-1} f_k^2,$$

as is seen by changing the order of summation and using (6, 1, 4).

6, 3 Real Part.

It is possible to work with real values exclusively by considering only the real value of the transform. Let R_j be the real and S_j be the imaginary part of F_j. Then

(6, 3, 1) $$F_j = R_j + iS_j.$$

By (6, 1, 1) the real part

(6, 3, 2) $$R\left(\phi\left(\frac{2n\pi i}{c}\right)\right) = \frac{1}{c} \sum_{j=0}^{c-1} R\left(F_j e - \frac{2jn\pi_i}{c}\right)$$

$$= \frac{1}{c} \sum_{j=0}^{c-1} R_j \cos j \frac{2n\pi}{c}$$

and the imaginary part is

(6, 3, 3) $$- \frac{1}{c} \sum_{j=0}^{c-1} S_j \sin j \frac{2n\pi}{c}.$$

Since by (6, 1, 2) F is additive, so are R and S.

Since $\phi(0) = f_0$ is real,

$$\frac{1}{c} \sum_{j=0}^{c-1} R_j = f_0.$$

Since $\phi(nw) = f^n$ is real, $$\frac{1}{c} \sum_{j=0}^{c-1} S_j \sin \frac{2n\pi}{c} = 0.$$

Also $$\sum_{j=0}^{c-1} S_j = - \sum_{j=0}^{c-1} \sum_{k=0}^{c-1} f_k \sin (jk \frac{2\pi}{c})$$

(6, 3, 4) $$= - \sum_{k=0}^{c-1} f_k \sum_{j=0}^{c-1} \sin (jk \frac{2\pi}{c})$$

$$= 0.$$

In the example of 6, 1 the real parts of the transforms are 9, 2, −1, and 2. The imaginary parts are 0, −1, 0, and 1.

6, 4 Absolute Value.

There is another version of the transform which is sometimes useful. This is

(6, 4, 1) $$T_j = |F_j|.$$

If one is dealing with a cipher system which applies a "slide", then this absolute value has a useful property. A slide is a known substitution with a single cycle, or a power thereof. That is, with a slide of s, the frequency distribution $\{f_k\}$ becomes $\{g_k = f_{k-s}\}$. All subscripts are taken mod c. If $|G|$ is the transform of $\{g\}$

then $$G_j = \sum_{k=0}^{c-1} g_k \omega^{jk}.$$

$$G_j = \sum_{k=0}^{c-1} g_{k+s} \omega^{j(k+s)}$$

$$G_j = \sum_{k=0}^{c-1} f_k \omega^{jk} \omega^{js}$$

(6, 4, 2) $G_j = \omega^{js} F_j$.

Then

(6, 4, 3) $T_j = |F_j| = |\omega^{-js}| \cdot |G_j| = |G_j|$.

Thus the transform T_j is invariant under a slide. Unfortunately T is not additive like F, R, and S.

In the example of 6, 1 the absolute values of the transforms are 9, $\sqrt{5}$, 1, and $\sqrt{5}$. If a slide of $s = 1$ is applied to the frequency count it becomes 3, 3, 2, 1. The Fourier transforms of these are 9, $1+2i$, 1, and $1-2i$, quite different from 9, $2-i$, -1, and $2+i$. The absolute values are, however, the same.

6, 5 Application to Minuend Systems.

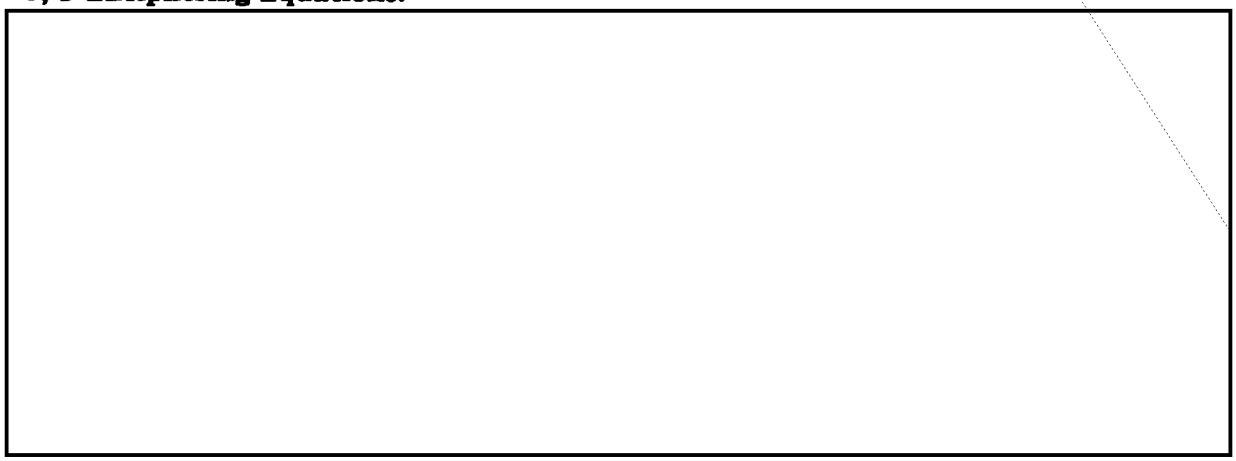
7. Theory of Circulices.

In certain algebraic and statistical procedures special matrices arise of the type

$$\begin{pmatrix} a_0 & a_1 & a_2 & \cdots & a_{c-1} \\ a_{c-1} & a_0 & a_1 & \cdots & a_{c-2} \\ & & \vdots & & \\ a_1 & a_2 & a^3 & \cdots & a_0 \end{pmatrix}.$$

That is, each row is a slide to the right of the row above. Such a matrix is called a "circulix."

7, 1 Enciphering Equations.



7, 2 Properties of Circulices.

The sum of circulices is a circulix, the zero matrix (every element 0) is a circulix, and the negative of a circulix is a circulix. Therefore they form an additive group.

The product of circulices is also a circulix, as is seen by (7, 2, 1)

(7, 2, 1) $$(a_{j-i})\,(b_{j-i}) = \left(\sum_{k=0}^{c-1} a_{k-i}\, b_{j-k}\right) = (z_{ij}).$$ All subscripts are mod c.

Now $$z_{i+s,\,j+s} = \sum_{k=0}^{c-1} a_{k-s-i}\, b_{j+s-k}$$

$$= \sum_{h=0}^{c-1} a_{h-i}\, b_{j-h}$$

$$= z_{ij},$$

where $h = k - s$. This establishes that the product is a circulix.

Circulices are permutable, for

(7, 2, 2) $$(b_{j-i})\,(a_{j-i}) = \left(\sum_{k=0}^{c-1} b_{k-i}\, a_{j-k}\right)$$

If we put $k = i+j-h \bmod c$,

$$\sum_{k=0}^{c-1} b_{k-i}\, a_{j-k} = \sum_{h=0}^{c-1} b_{j-h}\, a_{h-i},$$

which is identical with (7, 2, 1).

If a circulix has an inverse, then that inverse is a circulix. Suppose

(7, 2, 3) $$(x_{ij})\,(b_{j-i}) = (\delta_{ij}) = I.$$

Then $\sum_{k=0}^{c-1} x_{ik} b_{j-k} = \delta_{ij}$ for all i and j. Replace i by i+s, j by j+s, and let h = k − s, getting

$$\sum_{h=0}^{c-1} x_{i+s,\, h+s}\, b_{j-h} = \delta_{i+s,\, j+s} = \delta_{ij}\,. \tag{7, 2, 4}$$

Thus if (x_{ij}) is a solution of (7, 2, 3), then $(x_{i+s,\, j+s})$ is also a solution. If (b_{j-i}) has an inverse, then it is unique and $(x_{i+s,\, j+s}) = (x_{ij})$, showing that the inverse is a circulix.

If ω is a primitive cth root of 1, then $v = \begin{pmatrix} 1 \\ \omega \\ \omega^2 \\ \cdot \\ \cdot \\ \cdot \end{pmatrix}$ is a vector which is merely stretched when multiplied by a circulix A. For

$$Av = (a_{j-i})\,(\omega^i) = \begin{pmatrix} \sum_k a_{k-i}\omega^k \end{pmatrix}$$

$$= \begin{pmatrix} \omega^i \sum_{k=0}^{c-1} a_{k-i}\omega^{k-i} \end{pmatrix} \tag{7, 2, 5}$$

$$= v \sum_{k=0}^{c-1} a_{k-i}\omega^{k-i}.$$

Thus v is an "eigenvector" of A, and $F = \sum_{k=0}^{c-1} a_{k-i}\omega^{k-i}$, which is independent of i, is the corresponding "eigenvalue".

If $v_t = (\omega^{it})$, then it also is an eigenvector of A. For

$$Av_t = (a_{j-i})\,(\omega^{it}) = \begin{pmatrix} \sum_k a_{k-i}\omega^{kt} \end{pmatrix}$$

$$= \omega^{ti} \sum_{k=0}^{c-1} a_{k-i}\omega^{(k-i)t} = v_t F_t$$

(7, 2, 6) where $F_t = \sum_{k=0}^{c-1} a_{k-i}\,\omega^{(k-i)t}$ is independent of i.

The c vectors v_t are seen to be linearly independent. We can combine them into the square matrix

$$V = (\omega^{ij}).$$

Now $AV = VL$ where L is the diagonal matrix with the eigenvalues F_t as diagonal elements. Since the v_t are linearly independent V has an inverse V^{-1}, so

(7, 2, 7) $$V^{-1}AV = L.$$

The eigenmatrix V depends only on the fact that A is a circulix, and not on the a_i. Thus all circulices can be transferred to diagonal form L by the same matrix V. The diagonal elements are the eigenvalues, F_t.

If A and B are two circulices, they are each transforms of diagonal matrices, $A = VLV^{-1}$ and $B = VMV^{-1}$.

(7, 2, 8) $$A+B = VLV^{-1} + VMV^{-1} = V(L+M)V^{-1}.$$

That is, the sum of two circulices has for its eigenvalues the sum of the respective eigenvalues of the summands A and B. The order of the eigenvalues is determined uniquely by V.

(7, 2, 9) $$AB = VLV^{-1} \cdot VMV^{-1} = VLMV^{-1}.$$

The product of two circulices has for its eigenvalues the products of the respective eigenvalues of the factors.

7, 3 Fourier Transforms and Circulices.

Looking back to formula (6, 1, 2)

$$F_j = \sum_{k=0}^{c-1} f_k e^{jkw}, \text{ where } w = \frac{2\pi i}{c},$$

we see that if $\omega = e^w$ this formula is identical with (7, 2, 6)

$$F_j = \sum_{k=0}^{c-1} a_k \omega^{kj},$$

where $a_k = f_k$. Thus the eigenvalue F_j is identical with the Fourier transform F_j.

We repeat here the results from the Fourier transform theory stated in terms of eigenvalues. The eigenvalues are in general complex numbers. If the circulix is real, then the values are conjugate in pairs,

$$F_{c-j} = \overline{F}_j.$$

The first eigenvalue $F_0 = \sum_{k=0}^{c-1} a_k$. If the circulix comes from a frequency count, F_0 is the total count.

If c is even, $F_{\frac{c}{2}}$ measures the deviation from random mod 2, and has been used by itself to place cribs and set messages.

If f is a circulix of frequencies, then the sum of the diagonal elements of f*f are given by $\sum_{k=0}^{c-1} f_k^2$. Therefore the trace (the sum on the principal diagonal) of f*f is $c \sum_{k=0}^{c-1} f_k^2 = N^2\gamma$.

Since the trace is invariant under matrix transformation, therefore the trace of L*L is the same $\sum_{j=0}^{c-1} \overline{F}_j F_j = N^2$. See (6, 2, 18).

7, 4 Polynomials and Circulices.

If A is a circulix with elements a_k, we can use these elements to define a polynomial.

$$A(x) = \sum_{k=0}^{c-1} a_k x^k. \tag{7, 4, 1}$$

The variable x is an indeterminate. Any polynomial A(x) of the form of (7, 4, 1) determines a circulix.

The product of two circulices

$$AB = (a_t)\,(b_t) = \left(\sum_{k=0}^{c-1} a_k\, b_{t-k} \right)$$

(7, 4, 2) defines the polynomial $\sum_{t=0}^{c-1} \sum_{k=0}^{c-1} a_k\, b_{t-k} x^t$. The product of the polynomials

$$A(x)\,B(x) = \sum_{k=0}^{c-1} \sum_{m=0}^{c-1} a_k b_m x^{k+m} = \sum_{k=0}^{c-1} \sum_{t=k}^{c-1+k} a_k b_{t-k} x^t. \tag{7, 4, 3}$$

Comparing this with (7, 4, 2), we see it differs only in the range of t. In (7, 4, 2) the subscripts are understood to be modulo c. To impose the same convention on (7, 4, 3) would mean to interpret the exponent on x modulo c also. That is, x^{c+k} to mean x^k, and $x^{c+0} = x^0 = 1$, or $x^c - 1 = 0$. If the product A(x) B(x) of the polynomials is taken modulo $x^c - 1$ the correspondence between the circulices and the polynomials is an isomorphism.*

Reference to (7, 2, 6) shows that the eigenvalues are merely specific values of the polynomial,

$$F_j = A(\omega^j). \tag{7, 4, 4}$$

The c eigenvalues can be used to define a new polynomial,

$$E(x) = \sum_{j=0}^{c-1} F_j x^j. \tag{7, 4, 5}$$

Then the eigenvalues of E are

$$E(\omega^k) = \sum_{j=0}^{c-1} F_j \omega^{jk} = c a_k.$$

See (6, 1, 1) and (6, 1, 2). Thus the two polynomials are symmetrically related; each can be derived from the other.

*This was pointed out to me by LTJG William Blankinship.

BIBLIOGRAPHY

(1) "Introduction to Mathematical Probability", McGraw-Hill ..... J. V. Uspensky, 1937

(2) "Mathematical Methods of Statistics", Princeton ..... Harald Cramer, 1944

(3) "Mathematical Statistics", Princeton ..... S. S. Wilks, 1944

(4) "An Introduction to Probability Theory and Its Applications", Wiley ..... W. Feller, 1950

(5) "Advanced Theory of Statistics", Charles Griffin and Co. Ltd ..... Kendall, 1943

(6) "Statistical Methods in Cryptanalysis" ..... S. Kullback

(7) "Determinants and Matrices", Oliver and Boyd, Edinburgh ..... Aitken

(8) "Higher Algebra", McMillan, London ..... Hall and Knight, 1942

(9) "The Theory of Matrices", Chelsea, New York ..... C. C. MacDuffee, 1946

(10) "Tables of the Binomial Probability Distribution", Government Printing Office ..... National Bureau of Standards, 1949

(11) "Poisson's Exponential Binomial Limit", van Nostrand ..... E. C. Molina, 1943

(12) "Tables of the Incomplete Γ Functions" ..... K. Pearson, 1922, Biometrica, 1934

(13) "Table of Poisson Distribution", Cryptanalyst's Manual (Section 5–1) ..... Army Security Agency, July 1947

(14) "Poisson Difference Tables", Cryptanalyst's Manaul (Section 5–2) ..... Army Security Agency, June 1947

(15) "Expected Number of X-fold Repetitions" (Binomial Distribution), Cryptanalyst's Manual (Section 5–3) ..... Army Security Agency, January 1950

(16) "Tables of Probability Functions", Vol. II Normal ..... FWA-WPA

(17) "I. C. Distribution Tables" ..... N–31, February 1949

(18) "The Index of Coincidence", NSA ..... H. Campaigne, January 1955

(19) "Goodness of Fit", AFSA–34 ..... R. Dawson, 26 January 1950

(20) "Derivation of Analysis Weights" N–53 ..... J. J. Eachus, 12 April 1949

(21) "Estimation of the Right Tail of the Poisson Distribution", AFSA–34 ..... LCDR A. M. Gleason, G. F. Cramer

(22) "Inequalities for the Tails of the Binomial Distribution", AFSA–34 ..... C. Maple, A. H. Clifford

(23) "Distribution of Small Samples", N–31 ..... Charlotte Wootten, 17 February 1949

(24) "On Information and Sufficiency" Annals of Mathematical Statistics ..... S. Kullback, R. Leibler, March 1951

(25) "Finding the Needle in the Haystack", N–31 .. H. Campaigne, G F. Cramer, 6 February 1949

(26) "Tables of the Incomplete Beta Function", Cambridge ..... K. Pearson, 1934

(27) "Distribution of the Correlation Coefficient", N–31 ..... A. M. Gleason, 9 March 1948

(28) "Human Behavior and the Principle of Least Effort", Addison-Wesley Press ..... G. K. Zipf, 1949

(29) "Note on the Preparation of Weights from a Sample", AFSA–34 ..... R. B. Dawson, Jr., 21 April 1950

(30) "The Chi-square Test" ..... R. E. Greenwood, W. Lotz, and B. Barrett, 14 March 1952

(31) "Notes on Chi-squaring Digraph Counts" ..... Lt. P. P. Billingsley, 14 May 1954

(32) "Table of the Binomial Distribution", Army Security Agency ..... May 1946

(33) "Error Detecting and Error Correcting Codes", Bell System Technical Journal XXIX pp. 147–160 ..... R. W. Hamming, April 1950

(34) "A Mathematical Theory of Communication", Bell System Technical Journal ..... C E. Shannon, July, October 1948

(35) [redacted] ..... Wm. A. Blankinship, 3 May 1955

(36) "A Discussion of Multinomial Estimation" ..... R. B. Dawson, Jr., 7 May 1954

(37) "The Bulge of Sum or Difference Text" ..... O. S. Rothaus, 25 March 1955

(38) "The Population Frequencies of Species", Biometrika, Vol. 40 .... I. J. Good, December 1953

(39) "Estimation of the Number of Classes in a Population", Ann. Math. Stat., Vol. 20, 572–9 .... L. A. Goodman

(40) "The Method of Moments and the Distribution of Repeats in Small Samples" ..... O. S. Rothaus, 26 September 1955

(41) "The Serial Test for Sampling Numbers and Other Tests for Randomness", Proc. of the Camb. Phil. Soc., Vol. 49, 276–284 ..... I. J. Good, 1953

(42) "The Joint Distribution for the Sizes of the Generations in a Cascade Process", Proc. of the Camb. Phil. Soc., Vol. 51, 240–242 ..... I. J. Good, 1955

~~CONFIDENTIAL~~