ADVANCED GCE

MATHEMATICS (MEI) 4754B

Applications of Advanced Mathematics (C4) Paper B: Comprehension
INSERT

Friday 15 January 2010
Afternoon

Duration: Up to 1 hour

INSTRUCTIONS TO CANDIDATES

- This insert contains the text for use with the questions.

INFORMATION FOR CANDIDATES

- This document consists of **8** pages. Any blank pages are indicated.

OCR is an exempt Charity

Turn over

Cipher Systems

Introduction

Imagine you want to send a written message to a friend containing confidential information. You are keen to ensure that the information remains private but you are aware that other people might try to intercept the message before it reaches its destination. Therefore it is necessary to use some form of code, or *cipher*, known to you and your friend, so that anybody who intercepts the message will find it difficult to extract the information.

You use an encoding cipher to encode your message (called the *plaintext*) and then transmit the resulting message (called the *ciphertext*). Your friend will then decode the message using an appropriate decoding cipher. This *cipher system* is illustrated in Fig. 1.

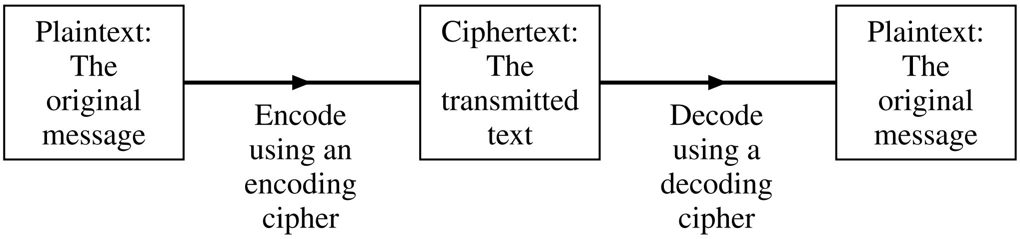
Plaintext: The original message
Encode using an encoding cipher
Ciphertext: The transmitted text
Decode using a decoding cipher
Plaintext: The original message

Fig. 1

What is the most appropriate cipher system to use so that you can send and receive messages quickly and cheaply, whilst minimising the risk that an interceptor will be able to decode the ciphertext?

This has been an important question for at least two thousand years, since messages were written on the shaved heads of messengers who would then wait until their hair grew back before setting off on journeys to the intended recipients. More sophisticated cipher systems have been developed and some of these are described in this article. The methods are not necessarily representative of those used in practice today, but highlight some of the weaknesses which need to be avoided when trying to design a secure cipher system.

Caesar cipher

This is the most basic form of cipher. In a Caesar cipher, each letter in the message is replaced by the letter a fixed number of places further on in the alphabet to give the text to transmit.

For example, Table 2 shows the Caesar cipher corresponding to a shift of the alphabet by 21 places.

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

Table 2

Using this encoding cipher, the plaintext

```
            IF I HAVE SEEN FARTHER THAN OTHER MEN,
IT IS BECAUSE I HAVE STOOD ON THE SHOULDERS OF GIANTS.
```

would be transmitted as the following ciphertext.

DA D CVQZ NZZI AVMOCZM OCVI JOCZM HZI,
DO DN WZXVPNZ D CVQZ NOJJY JI OCZ NCJPGYZMN JA BDVION.

[Warning Do not spend time in this examination checking the accuracy of this or any other ciphertext.]

To disguise the lengths of words, it is common for ciphertext to be transmitted in blocks of fixed length, with punctuation removed. For example, the ciphertext above could be transmitted as follows.

DADCV QZNZZ IAVMO CZMOC VIJOC ZMHZI DODNW ZXVPN
ZDCVQ ZNOJJ YJIOC ZNCJP GYZMN JABDV ION

Notice that the decoding cipher is a shift of the alphabet in the same direction by 5 places. This is shown in Table 3.

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

Table 3

Clearly such a cipher system is not at all secure as there are only 25 shifts of the alphabet which an interceptor would have to try to be certain of extracting the message.

Substitution cipher

If the sender and receiver agree in advance on a letter substitution, then a more secure cipher system can be used. Table 4 shows an example of a more secure encoding cipher.

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

Table 4

It is unlikely that the sender or receiver would be able to remember the 26 letter substitutions and so it would be necessary to keep a written copy of the cipher. However, by writing it down there is a risk of it being discovered by an interceptor.

An alternative is simply to use a phrase, with repeated letters removed, to form the first few letters of the cipher. The letters not appearing in the phrase are then used in alphabetical order.

For example, removing repeated letters from the phrase

GOD WROTE THE UNIVERSE IN THE LANGUAGE OF MATHEMATICS

leaves

GODWRTEHUNIVSLAFMC

and this gives the ciphertext of the first 18 letters of the alphabet. The corresponding encoding cipher is given in Table 5.

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

Table 5

If a ciphertext message based on a cipher of this type is intercepted, and the interceptor has reason to believe that letter substitution has been used, then the number of possible arrangements of the alphabet that could have been used is

$$26! = 403\,291\,461\,126\,605\,635\,584\,000\,000.$$

If the interceptor had the computing power to check 1 000 000 arrangements per second, whilst also checking the resulting messages to see if any were meaningful, then it would still take longer than the age of the universe to check all possibilities.

However, the interceptor might still have a chance of extracting the message. If the original message was in English and it was sufficiently long (about 200 letters is usually enough), then there are several techniques the interceptor might use.

- If the sender had simply encoded the words and left the gaps between the words, then it would not be difficult to discover the encoded forms of the letters A and I. This is one reason why it is normal for coded messages to be transmitted in fixed-length blocks.
- By far the most common three-letter word in English text is ‘the’ and the frequent appearance of a three-letter string in the ciphertext would probably give the encoded forms of T, H and E.
- In English, the letter Q is always followed by the letter U. If, in the ciphertext, one letter is always followed by the same other letter, this might give the interceptor the encoded forms of Q and U. This is one extreme example of a relationship between letters but other relationships also exist.
- Fig. 6 shows the percentage frequencies of the 26 letters in a large sample of English text from a wide variety of sources. In a sufficiently long passage of ciphertext, it would be sensible to look at those letters occurring most frequently and investigate if they could represent E, T and A. Once these letters are decoded, knowledge of the English language would quickly reveal many complete words.

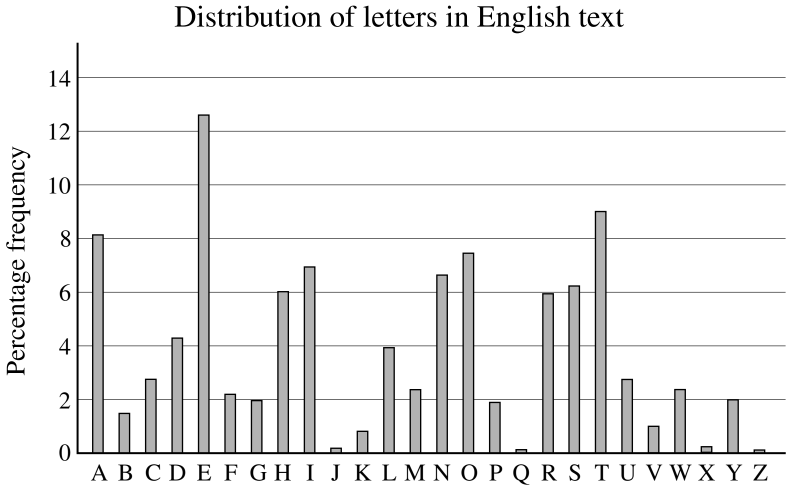

Fig. 6

Vigenère Cipher

In the ciphers mentioned above, a letter in ciphertext will always represent the same letter in the plaintext message. This is not the case with a Vigenère cipher. This cipher is illustrated using the following example.

First select a *keyword*. In the following example, the keyword is **ODE**. Rotations of the 26 letters of the alphabet are then written under the alphabet, starting with each of the letters in the keyword, as shown in Table 7.

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

Table 7

To encode a message, each of these rows is used in turn. For example, to encode the word NEWTON you would encode

- the first N using row **O**,
- the E using row **D**,
- the W using row **E**,
- the T using row **O**, and so on.

Therefore the word NEWTON would be transmitted in ciphertext as *BHAHRR*. You will notice that the two Ns in NEWTON appear as different letters in the ciphertext and the *H*s and *R*s in the ciphertext do not correspond to repeated letters in the plaintext.

This cipher is more secure than the others described so far but, for a sufficiently long passage of ciphertext, it is still vulnerable to frequency analysis, especially if the keyword is short. For example, consider the following ciphertext (in which the keyword used is not **ODE**).

NUMFU	*GPXGN*	*BBWVI*	*GMCIM*	*NRZGM*	*YQDYG*	*PXJNQ*	*GNRZL*
IEBAY	*CWXNF*	*UNMGJ*	*XVRIN*	*NVNNF*	*GPXCQ*	*MTMYQ*	*DYGPX*
WBTHO	*EAHL*___G___	***PXQBZ***	*WMZCL*	*NSQMN*	*BOX*___NU___	***MK****CAI*	*AUEUH*
HVWNM	*JIRVR*	*INNLQ*	*LNUMY*	*CEAMN*	*RAM*___NU___	***MK****YVA*	*GICMK*
GNVXH	*GXEUP*	*MBH*___GP___	***XQBZ****E*	*XSWKO*	*TTRGN*	*BAYZI*	*MCPA*

The string ***GPXQBZ*** appears twice, one a shift of 84 places from the other. It is reasonable to assume that in both cases the same six-letter string in the plaintext has been encoded. If this is indeed the case then the length of the keyword must be a factor of 84. There are also two occurrences of the string ***NUMK***, one a shift of 40 places from the other. Taken together, these two shifts suggest that the keyword has length 2 or 4.

Assume that the keyword is of length 4, since a keyword of length 2 would form a less secure cipher than one of length 4.

To carry out a frequency analysis, the ciphertext is split into four strings, S_1, S_2, S_3 and S_4, each made up of every fourth letter of the original ciphertext.

S_1 : *NUGWMNMYJNIYNMVNFCMYWOLQMNNNCUHMVNNCNNYIGHUHQXOGYC*
S_2 : *UGNVCRYGNRECFGRVGQYGBEGBZSBUAEVJRLUERUVCNGPGBSTNZP*
S_3 : *MPBIIZQPQZBWUJINPMQPTAPZCQOMIUWIIQMAAMAMVXMPZWTBIA*
S_4 : *FXBGMGDXGLAXNXNNXTDXHHXWLMXKAHNRNLYMMKGKXEBXEKRAM*

These text strings may be analysed using the information in Fig. 6. (The analysis may not always be successful with strings as short as these.)

Notice the following two characteristics of the distribution shown in Fig. 6.

- The three highest frequencies correspond to E, T and A.
- After the peak at T there is a run of six letters with low frequencies; this is the longest such run. 120

Compare this distribution with the distribution of letters in the first text string, S_1, shown in Fig. 8 below.

Distribution of letters in string S_1

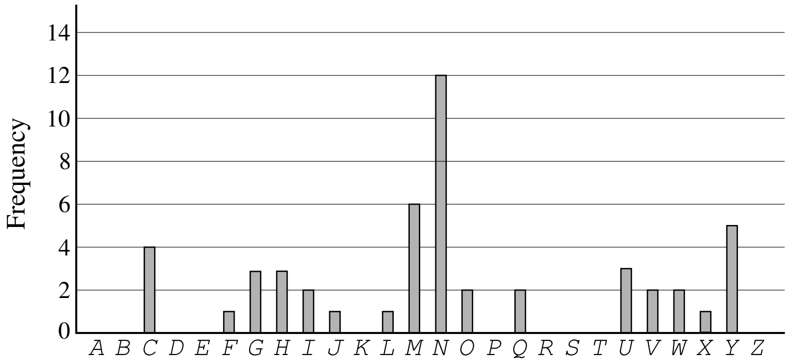

Fig. 8

There are two reasons to suspect that the letter *N* in string S_1 is the encoded form of the letter `T`.

- *N* in Fig. 8 is followed by a run of six letters with low frequency.
- The corresponding encoded forms of `A` and `E` would be *U* and *Y* respectively; both of these occur quite frequently. 125

These two reasons suggest that it is worth looking more closely at this particular shift.

Starting the distribution at *U*, as shown in Fig. 9, allows a direct comparison to be made with the distribution in Fig. 6.

Distribution of letters in string S_1 starting from *U*

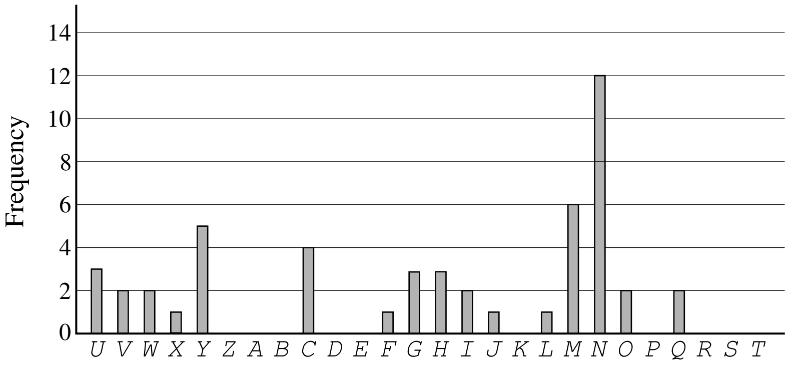

Fig. 9

The string is made up of only 50 letters so a very close match with Fig. 6 is not to be expected. 130 However the distributions are sufficiently similar to accept this shift for now and investigate the other three strings; it is only after all four strings have been decoded that we can determine if the shifts are correct.

The frequency distributions corresponding to the strings S_2, S_3 and S_4 are shown in Figs. 10, 11 and 12 respectively.

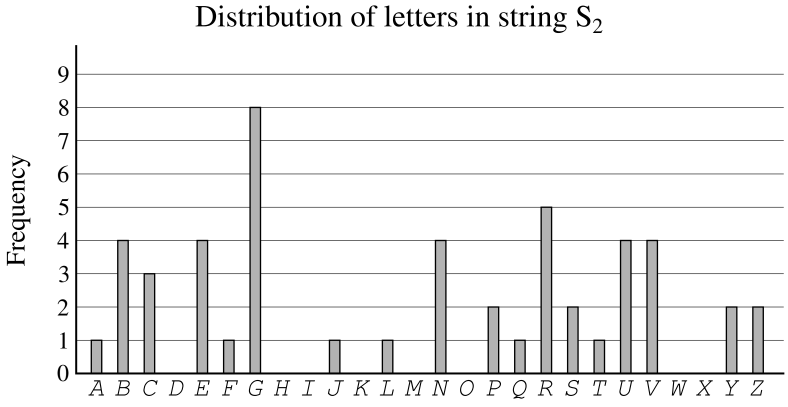

Fig. 10

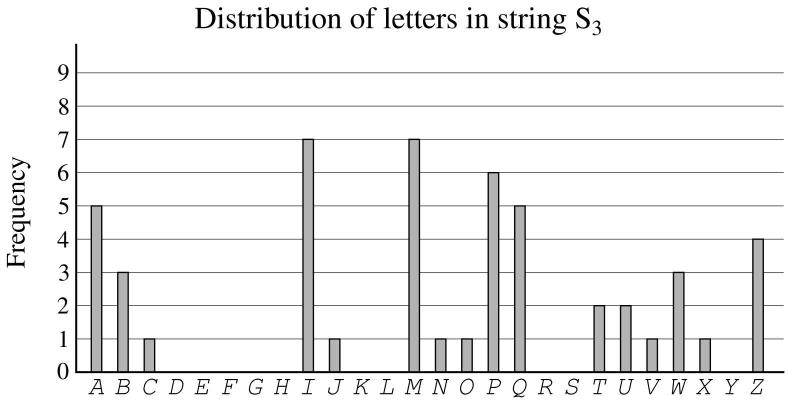

Fig. 11

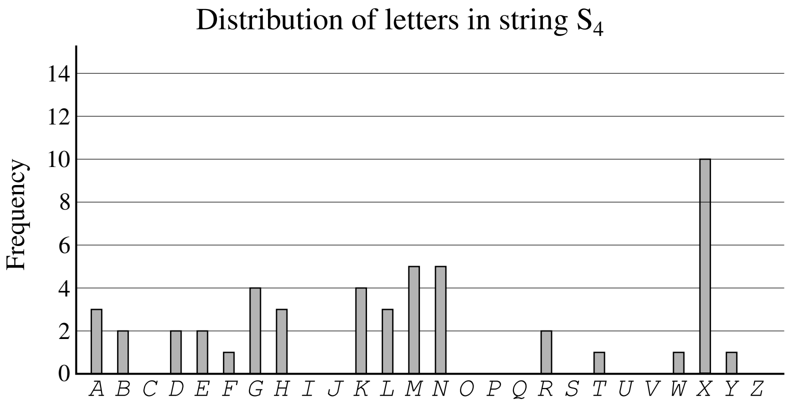

Fig. 12

By employing reasoning similar to that used in the analysis of string S_1, it is likely that

- in string S_2, the encoded form of the letter A is *N*,
- in string S_3, the encoded form of the letter A is *I*.

For string S_4 it is not quite so clear. However, it may be the case that the keyword used is a word in the English language. So it is sensible to check, in string S_4, if the encoded form of the letter A is *T* since this would make the four-letter keyword **UNIT**. With this keyword, the strings S_1, S_2, S_3 and S_4 are decoded to give the following strings.

T_1 : TAMCSTSEPTOETSBTLISECURWSTTTIANSBTTITTEOMNANWDUMEI

T_2 : HTAIPELTAERPSTEITDLTORTOMFOHNRIWEYHREHIPATCTOFGAMC

T_3 : EHTAARIHIRTOMBAFHEIHLSHRUIGEAMOAAIESSESENPEHROLTAS

T_4 : MEINTNKENSHEUEUUEAKEOOEDSTERHOUYUSFTTRNRELIELRYHT

This does indeed result in a meaningful plaintext message.

In conclusion

The methods considered have all involved substituting each letter with another letter. Variations on these are possible, such as having an encoded form of each of the 676 two-letter pairs (such as AA, AB, AC, …) or encoding a message using one cipher and then encoding the resulting ciphertext using another cipher.

The science of cryptography affects all our lives. Every time you send an email, use a cashpoint machine or make purchases on the internet, the information you transmit is encoded so that only your intended recipient can read it.

Government intelligence is heavily dependent upon the ability to transmit information securely whilst also trying to break the ciphers used by others. Indeed, it is estimated that the Second World War was shortened by two years, thereby saving many lives, thanks to the intelligence gained by the mathematicians working in cryptography at Bletchley Park.

As computing power becomes more sophisticated, more secure codes are continually being sought. The most secure codes in use today rely heavily on techniques from pure mathematics.

RECOGNISING ACHIEVEMENT