Math 135 (Summer 2006)
The Friedman and Kasiski Tests

Example: Suppose that a sample of ordinary English contains the following distribution of letters:

letter	count	letter	count
A	141	N	119
B	36	O	132
C	36	P	28
D	103	Q	1
E	188	R	95
F	37	S	64
G	34	T	182
H	102	U	59
I	123	V	13
J	4	W	55
K	18	X	3
L	56	Y	2
M	27	Z	0

Suppose that a pair of letters is selected at random from this sample. What is the probability that the two letters selected will be identical?

Solution:

If this sample is subjected to a monoalphabetic substitution, then the probability of choosing an identical pair from the resulting ciphertext is the same.

Friedman's Index of Coincidence

The **index of coincidence** (denoted by I) for a (cipher)text is the probability that two letters selected at random from it are identical.

If the text has n_0 A's, n_1 B's, n_2 C's, ..., n_{25} Z's, and

$$n = n_0 + n_1 + n_2 + \cdots + n_{25} = \sum_{i=0}^{25} n_i$$

is the number of letters in the text, then

$$I = \frac{n_0(n_0-1) + n_1(n_1-1) + \cdots + n_{25}(n_{25}-1)}{n(n-1)} = \frac{1}{n(n-1)} \sum_{i=0}^{25} n_i(n_i-1).$$

In the previous example, we found $I = 0.0656$.

Example: What is the index of coincidence for a collection of 2600 letters consisting of 100 A's, 100 B's, 100 C's, ..., 100 Z's?

Solution:

The index of coincidence of a totally random (uniformly distributed) collection of letters is about 0.0385. Vigenère ciphertexts from longer keywords have a more uniform distribution of letters. For keyword lengths closer to 1, the index of coincidence will be larger (closer to 0.0656).

Question: Can we quantify the connection between index of coincidence and keyword length?

Answer: YES!

Connection between I and Keyword Length

Suppose that the ciphertext has n letters and the Vigenère keyword has k letters. Arrange the ciphertext as follows:

Each column is a shift encipherment by an amount corresponding to that key letter. If we choose a pair at random, then either

- they come from the same column, or
- they come from two different column.

In the first case, there are k ways to choose the the column, $C(n/k, 2)$ ways to choose a pair of letters, and probability 0.065 that they are identical. Thus, the expected number of identical pairs from the same column is

$$S = k \cdot C(n/k, 2) \cdot 0.065.$$

(Remember, if there are N independent trials with probability p of an event happening on a given trial, then the expected number of events which happen is Np.)

In the second case, there are $C(k,2)$ ways to choose two different columns, n/k ways to choose a letter in one of them, n/k ways to choose a letter in the second, and probability $\frac{1}{26} \approx 0.03846$ that they are identical. Thus, the expected number of identical pairs from different columns is

$$D = C(k,2) \cdot \frac{n}{k} \cdot \frac{n}{k} \cdot 0.03846.$$

Therefore,

$$\begin{aligned} I \approx \frac{S+D}{C(n,2)} &= \frac{k \cdot C(n/k,2) \cdot 0.065 + C(k,2) \cdot \frac{n}{k} \cdot \frac{n}{k} \cdot 0.03846}{C(n,2)} \\ &= \frac{k \cdot \frac{\frac{n}{k}(\frac{n}{k}-1)}{2 \cdot 1} \cdot 0.065 + \frac{k(k-1)}{2 \cdot 1} \cdot \frac{n}{k} \cdot \frac{n}{k} \cdot 0.03846}{\frac{n(n-1)}{2 \cdot 1}} \\ &= \frac{0.065(n-k) + 0.03846n(k-1)}{k(n-1)} \end{aligned}$$

and so if we solve for k

$$k \approx \frac{0.0265n}{(0.065 - I) + n(I - 0.0385)}.$$

Example: Suppose that a polyalphabetic ciphertext has the following letter counts:

letter	count	letter	count
A	60	N	28
B	50	O	83
C	42	P	44
D	64	Q	69
E	51	R	13
F	63	S	29
G	19	T	66
H	48	U	87
I	56	V	63
J	67	W	19
K	23	X	43
L	45	Y	39
M	44	Z	67

Use Friedman's index of coincidence to estimate the length of the keyword length.

Solution:

The Kasiski Test

The following Vigenère encipherment illustrates an interesting phenomenon:

```
THEWEATHERINTHEFALLHASTHEMSWEATING
SAFETYSAFETYSAFETYSAFETYSAFETYSAFE
LHJAXYLHJVBLLHJJTJDHFWMFWMXAXYLISK

THEWEATHERINTHEFALLHASTHEMSWEATING
SAFETYSAFETYSAFETYSAFETYSAFETYSAFE
LHJAXYLHJVBLLHJJTJDHFWMFWMXAXYLISK
```

Letter groups in the ciphertext are repeated because letter groups in the plaintext line up with the keyword.

If letter groups are repeated in the ciphertext, then the keyword length may be a divisor of their separation.

Example: Determine a likely Vigenère keyword length for the following ciphertext:

```
HIBKA UQFLF SBQSX SKCFB YOAGP ALGTC RTYTL
DGBYO AGPAL OAKYB FBILY OYQTD ISVAI JJNNA
DXNLW NRQPF BVPWN IWAFB YAANR URTZE LYZLF
MEWHI BKAUQ FLALJ GTXRG VNIJP ZREQL KWZA
```

There are repeated letter groups:

```
HIBKA UQFLF SBQSX SKCFB YOAGP ALGTC RTYTL
DGBYO AGPAL OAKYB FBILY OYQTD ISVAI JJNNA
DXNLW NRQPF BVPWN IWAFB YAANR URTZE LYZLF
MEWHI BKAUQ FLALJ GTXRG VNIJP ZREQL KWZA
```

The separation (start to start) between the two occurrences of `HIBKAUQFL` is $108 = 3^3 \cdot 2^2$ letters, and the separation between those of `BYOAGPAL` is $18 = 3^2 \cdot 2$ letters. Common divisors of these two numbers are 2, 3, 4, 6, 9, 12, and 18, and so these are the most likely keyword lengths.