Index of Coincidence

The following table gives the relative frequency of the English alphabet letters in a 7834-letter sample of English writing.

<table>
<thead>
<tr>
<th>Letter</th>
<th>Relative Frequency</th>
<th>Letter</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.08399</td>
<td>N</td>
<td>0.06778</td>
</tr>
<tr>
<td>B</td>
<td>0.01442</td>
<td>O</td>
<td>0.07493</td>
</tr>
<tr>
<td>C</td>
<td>0.02527</td>
<td>P</td>
<td>0.01991</td>
</tr>
<tr>
<td>D</td>
<td>0.04800</td>
<td>Q</td>
<td>0.00077</td>
</tr>
<tr>
<td>E</td>
<td>0.12150</td>
<td>R</td>
<td>0.06063</td>
</tr>
<tr>
<td>F</td>
<td>0.02132</td>
<td>S</td>
<td>0.06319</td>
</tr>
<tr>
<td>G</td>
<td>0.02323</td>
<td>T</td>
<td>0.08999</td>
</tr>
<tr>
<td>H</td>
<td>0.06025</td>
<td>U</td>
<td>0.02783</td>
</tr>
<tr>
<td>I</td>
<td>0.06485</td>
<td>V</td>
<td>0.00996</td>
</tr>
<tr>
<td>J</td>
<td>0.00102</td>
<td>W</td>
<td>0.02464</td>
</tr>
<tr>
<td>K</td>
<td>0.00689</td>
<td>X</td>
<td>0.00204</td>
</tr>
<tr>
<td>L</td>
<td>0.04008</td>
<td>Y</td>
<td>0.02157</td>
</tr>
<tr>
<td>M</td>
<td>0.02566</td>
<td>Z</td>
<td>0.00025</td>
</tr>
</tbody>
</table>

The probability that two randomly selected letters in English are identical is given by

$$\sum_{\alpha=A}^{Z} p_{\alpha}^2 \approx 0.065$$

In a Vigenère cipher with sufficiently long keyword, the probabilities of seeing any letter in the ciphertext will converge to

$$\frac{1}{26} = 0.0385$$
Friedman Test

Definition 1 (Index of Coincidence). The index of coincidence (for a ciphertext), denoted $I$, is the probability that two randomly selected letters in the ciphertext are identical.

Remark:

- If $I \approx 0.065$ then the cipher is more likely to be mono-alphabetic substitution.
- For poly-alphabetic substitution, $0.0385 \leq I \leq 0.065$

Theorem 1. Let $n_0, n_1, n_2, \ldots, n_{24}, n_{25}$ be the respective counts of the letters $A, B, C, \ldots, Y, Z$. Let $n = \sum n_i$ be the total number of letters in the text then

$$I = \frac{1}{n(n-1)} \sum_{i=0}^{25} n_i(n_i - 1).$$

Now if an English plaintext is encrypted using a Vigenère cipher with keyword of length $k$, then

$$I \approx \frac{0.0385 \cdot n(k-1) + 0.065(n-k)}{k(n-1)}, \text{ or equivalently,}$$

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

Kasiski Test

The Kasiski Test is another way of estimating the length of the keyword for Vigenère cipher. It obtains possible keyword lengths from the gcd of the spacing between repeated letter groups in the ciphertext.
Cryptanalysis of Vigenère cipher

Remark: Both Friedman and Kasiski Tests only give the keyword length, but not the keyword itself. Furthermore, they are not very accurate when the ciphertext is small (usually less than 400 characters).

The signature of English is the graph of letter frequency distribution of English when we sort these frequencies in increasing order.

A coset are all the letters of the Vigenère ciphertext that are encrypted by the same letter of the keyword.

Remark: A coset of the Vigenère ciphertext has the same encryption as a Caesar shift cipher.
The **scrawl of English** is the graph of letter frequency distribution of English in alphabetical order.
Monty Hall Problem

Consider the following game:

Suppose there are three doors. A car is hidden behind one of the doors, and the other two have goats.

As a player, you pick one of the doors.

The host, who knows where the car is, will open another door and reveal a goat.

Now you can choose whether to swap your choice with last door or stay with original choice.

What is your best strategy?
Definition 2. The \textit{conditional probability} of an event \( B \) is the probability that this event will occur, given the knowledge that another event \( A \) has already occurred.

\[
P(B|A) = \frac{P(B \text{ and } A)}{P(A)},
\]

assuming that \( P(A) > 0 \).

Refer to the table for the frequency of character pairs in English language.