Estimating the Length of the Keyword

William Friedman's index of coincidence can also be used to estimate l the length of the keyword of a Vigenère cipher.

We will develop an approximation formula for I, the index of coincidence; this formula will contain l and n, the number of letters in the ciphertext. Then, to get an approximation for the length l, we will solve for l in terms of I and n (we know n and can calculate I).

First, assume that we know l and arrange the ciphertext into l columns. Now each column corresponds to a Caesar cipher. Although the columns might not all have the same length, we will assume that the number of letters in the ciphertext is large enough so that we can assume that they each have length $\frac{n}{l}$; i.e., we will assume that the error using this number for the length of each column is not large.

If we chose two letters from the ciphertext, what is the probability that they come from the same column and are the same letter?

One possibility is that we select two letters from the ciphertext that come from the same column and are the same letter. What is that probability?

Select a letter from the ciphertext. This selection determines a column. The probability that the next letter chosen comes from the same column is $\frac{\frac{n}{l}-1}{n-1}$. Because both letters are selected from the same Caesar cipher alphabet, the probability that both are the same is approximately the same as for standard

English 0.065. So, the probability that both letters are selected from the same column and are the same letter is approximately $\frac{\frac{n}{l}-1}{n-1} \times 0.065$.

The other possibility is that we select two letters from the ciphertext that come from different columns but are the same letter. What is that probability?

Select a letter from the ciphertext. Again, this determines a column. The probability that the next letter comes from a different column is $\frac{n-\frac{n}{l}}{n-1}$. Because the two letters are selected from different Caesar cipher alphabets, the probability that both are the same is approximately the same as for a random alphabet 0.038. So, the probability that both letters are selected from different columns and are the same letter is approximately $\frac{n-\frac{n}{l}}{n-1} \times 0.038$.

So we have two cases: the two letters are selected from the same column and are the same letter or the two letters are selected from different columns and are the same letter. To get an approximation of the index of coincidence *I*, the probability that the two letters selected are the same, we add these two probabilities:

$$I \approx \frac{\frac{n}{l}-1}{n-1} \times 0.065 + \frac{n-\frac{n}{l}}{n-1} \times 0.038.$$

Doing a bit of algebra to solve for *l*, we obtain:

$$I \approx \frac{\frac{n}{l}-1}{n-1}\times 0.065 + \frac{n-\frac{n}{l}}{n-1}\times 0.038$$

$$(n-1)I \approx \left(\frac{n}{l}-1\right)\times 0.065 + \left(n-\frac{n}{l}\right)\times 0.038$$

$$(n-1)I \approx \frac{n}{l}\times 0.065 - 0.065 + n\times 0.038 - \frac{n}{l}\times 0.038$$

$$(n-1)I + 0.065 - 0.038n \approx \frac{n}{l}\times (0.065 - 0.038)$$

$$(n-1)I + 0.065 - 0.038n \approx 0.027\frac{n}{l}$$

$$l \approx \frac{0.027n}{(n-1)I + 0.065 - 0.038n}$$

A commonly used table to estimate the length of the keyword is:

Estimated length of keyword	Index of Coincidence
1	0.0660
2	0.0520
3	0.0473
4	0.0449
5	0.0435
6	0.0426
7	0.0419
8	0.0414
9	0.0410
10	0.0407
⋮	⋮
∞	0.0388