Index Of Coincidence

We want to express a measure that would reflect how much does a given distribution differ from a perfectly flat (uniform) distribution. For a perfectly flat distribution, this value could be 0, and the more a given distribution differs from uniform, the bigger this measure gets. We could call it a ”measure of roughness”.

$$
\begin{aligned}
M.R &= \sum_i (p_i - \frac{1}{26})^2 = \sum_i p_i^2 - 2 \cdot \frac{1}{26} \cdot \underbrace{\sum_i p_i}_{1} + \sum_i \left(\frac{1}{26}\right)^2 \\
&= \sum_i p_i^2 - \frac{2}{26} + \frac{26}{26^2} = \sum_i p_i^2 - \frac{1}{26} \approx \sum_i p_i^2 - 0.038 \ \ .
\end{aligned}
$$

Since $\frac{1}{26}$ is a constant, it doesn't change and hence the only term that reflects the roughness of a distribution is $\sum\limits_i p_i^2$. Since we do not know the distribution of the plaintexts, we can't calculate $\sum\limits_i p_i^2$. However, $\sum\limits_i p_i^2$ is a probability that any two randomly selected letters have the same value. We could approximate this probability using frequencies of letters in the ciphertext.

In a set of N letters, letter A with frequency f_A. It can form

$$
\binom{f_A}{2} = \frac{f_A!}{2! \cdot (f_A - 2)!} = \frac{f_A \cdot (f_A - 1)}{2}
$$

pairs. The total number of possible pairs in a set of N letters is $\binom{N}{2}$. The probability that in a ciphertext Y any two randomly selected letters will be the same is

$$
I.C(Y) = \sum_{i=A}^{Z} \frac{\binom{f_i}{2}}{\binom{N}{2}} = \frac{f_i(f_i - 1)}{N(N-1)} \ \ .
$$

This value is called the *index of coincidence*. If instead of f_i we plug in the characteristic frequencies of letters in the English language, we will get approximately 0.066. For a flat distribution,

$$
\sum_i p_i^2 = \sum_i \left(\frac{1}{26}\right)^2 = 26 \cdot \frac{1}{26}^2 = \frac{1}{26} \ \ ,
$$

and hence the measure of roughness in this case is $M.R = \sum\limits_i p_i^2 - \frac{1}{26} = 0$.

Therefore, we can say that the index of coincidence will be in the range $0.038 \leqslant I.C(Y) \leqslant 0.066$. The bigger it is, the more it resembles English language, as opposed to a set of randomly selected letters.

We could define index of coincidence for a combination of two ciphertexts Y and Y'. For this, we calculate $I.C$ for a combined ciphertext containing $f_i + f_i'$ of letters i, and $N + N'$ letters total.

$$
I.C(Y, Y') = \sum_i \frac{(f_i + f_i')(f_i + f_i' - 1)}{(N + N')(N + N' - 1)} \ \ .
$$

The nominator can be reduced to

$$\sum_i (f_i + f'_i)(f_i + f'_i - 1) = \sum_i f_i^2 + \sum_i f_i'^2 + 2\sum_i f_i f'_i - \sum_i f_i - \sum_i f'_i \ .$$

Every expression in this sum, except for $\sum_i f_i f'_i$ is dependent on a single distribution only, and its value doesn't depend on the way in which the two distributions are positioned against one another. Therefore, to get an indication of how rough the composition of two ciphertexts is, it is sufficient to calculate $\sum_i f_i f'_i$.

A similar situation happens with denominator.

$$(N + N')(N + N' - 1) = N^2 + N'^2 + 2NN' - N - N' \ .$$

The only expression that reflects the length of the joint distribution is $2NN'$ and is the only expression that is useful to us.

Hence, the mutual index of coincidence of two ciphertexts Y and Y' is

$$I.C(Y, Y') = \sum_{i=A}^{Z} \frac{f_i f'_i}{NN'} \ .$$