## 14 Kullback's Cross-Product Sum Statistic

For a decision whether two texts $a \in \Sigma^r$, $b \in \Sigma^q$ belong to the same language we could consider $\varphi(a||b)$, the coincidence index of the concatenated string $a||b$. It should approximately equal the coincidence index of the language, or—in the negative case—be significantly smaller. This index evaluates as

$$(q+r)(q+r-1) \cdot \varphi(a||b) = \sum_{s\in\Sigma} [m_s(a) + m_s(b)]\,[m_s(a) + m_s(b) - 1]$$

$$= \sum_{s\in\Sigma} m_s(a)^2 + \sum_{s\in\Sigma} m_s(b)^2 + 2 \cdot \sum_{s\in\Sigma} m_s(a) m_s(b) - r - q$$

In this expression we consider terms depending on only one of the texts as irrelevant for the decision problem. Omitting them we are left with the "cross-product sum"

$$\sum_{s\in\Sigma} m_s(a) m_s(b)$$

From another viewpoint we could consider the "Euclidean distance" of $a$ and $b$ in the $n$-dimensional space of single letter frequencies

$$d(a,b) = \sum_{s\in\Sigma} [m_s(a) - m_s(b)]^2 = \sum_{s\in\Sigma} m_s(a)^2 + \sum_{s\in\Sigma} m_s(b)^2 - 2 \cdot \sum_{s\in\Sigma} m_s(a) m_s(b)$$

and this also motivates considering the cross-product sum. It should be large for texts from the same language, and small otherwise.

### Definition

Let $\Sigma$ be a finite alphabet. Let $a \in \Sigma^r$ and $b \in \Sigma^q$ be two texts of lengths $r, q \geq 1$. Then

$$\chi(a,b) := \frac{1}{rq} \cdot \sum_{s\in\Sigma} m_s(a) m_s(b),$$

where $m_s$ denotes the frequency of the letter $s$ in a text, is called **cross-product sum** of $a$ and $b$.

For each pair $r, q \in \mathbb{N}_1$ this defines a map

$$\chi\colon \Sigma^r \times \Sigma^q \longrightarrow \mathbb{Q}.$$

A Perl program, `chi.pl`, is in `http://www.staff.uni-mainz.de/pommeren/Cryptology/Classic/Perl/`.

Transforming $a$ and $b$ by the same monoalphabetic substitution permutes the summands of $\chi(a,b)$. Therefore $\chi$ is invariant under monoalphabetic substitution.

**Lemma 2** *Always $\chi(a,b) \leq 1$. Equality holds if and only if $a$ and $b$ consist of repetitions of the same single letter.*

*Proof.* We use the Cauchy-Schwartz inequality:

$$\begin{aligned}\chi(a,b)^2 = \left(\sum_{s\in\Sigma} \frac{m_s(a)}{r}\frac{m_s(b)}{q}\right)^2 &\leq \sum_{s\in\Sigma}\left(\frac{m_s(a)}{r}\right)^2 \cdot \sum_{s\in\Sigma}\left(\frac{m_s(b)}{q}\right)^2 \\ &\leq \sum_{s\in\Sigma}\frac{m_s(a)}{r}\cdot\sum_{s\in\Sigma}\frac{m_s(b)}{q} = 1\end{aligned}$$

Equality holds if and only if

- $m_s(a) = c \cdot m_s(b)$ for all $s \in \Sigma$ with a fixed $c \in \mathbb{R}$,
- *and* all $\frac{m_s(a)}{r}$ and $\frac{m_s(b)}{q}$ are 0 or 1.

These two conditions together are equivalent with both of $a$ and $b$ consisting of only one—the same—repeated letter. $\diamond$

Considering the quantity $\psi(a) := \chi(a,a) = \sum_s m_s(a)^2/r^2$ doesn't make much sense for Corollary 1 of the Kappa-Phi-Theorem gives a linear (more exactly: affine) relation between $\psi$ and $\varphi$:

**Lemma 3** *For all $a \in \Sigma^r$, $r \geq 2$,*

$$\varphi(a) = \frac{r}{r-1}\cdot\psi(a) - \frac{1}{r-1}$$

### Side Remark: Cohen's Kappa

In statistical texts one often encounters a related measure of coincidence between two series of observations: Cohen's kappa. It combines Friedman's kappa and Kullback's chi. Let $a = (a_1, \ldots, a_r)$, $b = (b_1, \ldots, b_r) \in \Sigma^r$ be two texts over the alphabet $\Sigma$ (or two series of observations of data of some type). Then consider the matrix of frequencies

$$m_{st}(a,b) = \#\{i \mid a_i = s, b_i = t\} \quad \text{for } s,t \in \Sigma.$$

Its row sums are

$$m_s(a) = \#\{i \mid a_i = s\} = \sum_{t\in\Sigma} m_{st}(a,b),$$

its column sums are

$$m_t(b) = \#\{i \mid b_i = t\} = \sum_{s\in\Sigma} m_{st}(a,b),$$

its diagonal sum is

$$\sum_{s\in\Sigma} m_{ss}(a,b) = \sum_{s\in\Sigma} \#\{i \mid a_i = b_i = s\} = \#\{i \mid a_i = b_i\}.$$

The intermediate values from which Cohen's kappa is calculated are

$$p_0 = \frac{1}{r} \cdot \sum_{s \in \Sigma} m_{ss}(a,b) = \kappa(a,b) \quad \text{and} \quad p_e = \frac{1}{r^2} \cdot \sum_{s \in \Sigma} m_s(a) m_s(b) = \chi(a,b)$$

Cohen's kappa is defined for $a \neq b$ by

$$\mathsf{K}(a,b) := \frac{p_0 - p_e}{1 - p_e} = \frac{\kappa(a,b) - \chi(a,b)}{1 - \chi(a,b)}$$

If $a$ and $b$ are random strings with not necessarily uniform letter probabilities $p_s$, then $\mathsf{K}$ is asymptotically normally distributed with expectation 0 and variance

$$\frac{p_0 \cdot (1 - p_0)}{r \cdot (1 - p_0)^2}$$

Therefore its use is convenient for large series of observations—or large strings—but in cryptanalysis we mostly have to deal with short strings, and considering $\kappa$ and $\chi$ separately may retain more information.

## Mean Values

For a fixed $a \in \Sigma^r$ we determine the mean value of $\kappa(a,b)$ taken over all $b \in \Sigma^q$:

$$\begin{aligned}
\frac{1}{n^q} \cdot \sum_{b \in \Sigma^q} \chi(a,b) &= \frac{1}{n^q} \cdot \sum_{b \in \Sigma^q} \left[ \frac{1}{rq} \cdot \sum_{s \in \Sigma} m_s(a) m_s(b) \right] \\
&= \frac{1}{rqn^q} \cdot \sum_{s \in \Sigma} m_s(a) \underbrace{\sum_{b \in \Sigma^q} m_s(b)}_{q \cdot n^{q-1}} \\
&= \frac{1}{rqn^q} \cdot r \cdot q \cdot n^{q-1} = \frac{1}{n}
\end{aligned}$$

where we used the corollary of Proposition 4.

In an analogous way we determine the mean value of $\chi(a, f_\sigma(b))$ for fixed $a, b \in \Sigma^r$ over all permutations $\sigma \in \mathcal{S}(\Sigma)$:

$$\frac{1}{n!} \cdot \sum_{\sigma \in \mathcal{S}(\Sigma)} \chi(a, f_\sigma(b)) = \frac{1}{rqn!} \cdot \sum_{\sigma \in \mathcal{S}(\Sigma)} \sum_{s \in \Sigma} m_s(a) m_s(f_\sigma(b))$$

As usual we interchange the order of summation, and evaluate the sum

$$\begin{aligned}
\sum_{\sigma \in \mathcal{S}(\Sigma)} m_s(f_\sigma(b)) &= \frac{1}{n} \cdot \sum_{t \in \Sigma} \sum_{\sigma \in \mathcal{S}(\Sigma)} m_t(f_\sigma(b)) \\
&= \frac{1}{n} \cdot \sum_{\sigma \in \mathcal{S}(\Sigma)} \underbrace{\sum_{t \in \Sigma} m_t(f_\sigma(b))}_{q} = \frac{1}{n} \cdot n! \cdot q = (n-1)! \cdot q
\end{aligned}$$

using the symmetry with respect to $s$. Therefore

$$\begin{aligned}\frac{1}{n!}\cdot\sum_{\sigma\in\mathcal{S}(\Sigma)}\chi(a,f_\sigma(b)) &= \frac{1}{rqn!}\cdot\sum_{s\in\Sigma}m_s(a)\cdot\sum_{\sigma\in\mathcal{S}(\Sigma)}m_s(f_\sigma(b))\\ &= \frac{1}{rqn!}\cdot r\cdot(n-1)!\cdot q=\frac{1}{n}\end{aligned}$$

Note that this conclusion also holds for $a = b$.

This derivation shows:

**Proposition 5** (i) *The mean value of $\chi(a,b)$ over all texts $b\in\Sigma^*$ of a fixed length $q$ is $\frac{1}{n}$ for all $a\in\Sigma^*$.*

(ii) *The mean value of $\chi(a,b)$ over all $a\in\Sigma^r$ and $b\in\Sigma^q$ is $\frac{1}{n}$ for all $r,q\in\mathbb{N}_1$.*

(iii) *The mean value of $\chi(a,f_\sigma(b))$ over all monoalphabetic substitutions with $\sigma\in\mathcal{S}(\Sigma)$ is $\frac{1}{n}$ for each pair $a,b\in\Sigma^*$.*

(iv) *The mean value of $\chi(f_\sigma(a),f_\tau(b))$ over all pairs of monoalphabetic substitutions, with $\sigma,\tau\in\mathcal{S}(\Sigma)$, is $\frac{1}{n}$ for each pair $a,b\in\Sigma^*$.*

## Interpretation

- For a given text $a$ and a "random" text $b$ we have $\chi(a,b)\approx\frac{1}{n}$.
- For "random" texts $a$ and $b$ we have $\chi(a,b)\approx\frac{1}{n}$.
- For given texts $a$ and $b$ and a "random" monoalphabetic substitution $f_\sigma$ we have $\chi(a,f_\sigma(b))\approx\frac{1}{n}$. This remark justifies treating a nontrivially monoalphabetically encrypted text as random with respect to $\chi$ and plaintext.
- For given texts $a$ and $b$ and two "random" monoalphabetic substitutions $f_\sigma$, $f_\tau$ we have $\chi(f_\sigma(a),f_\tau(b))\approx\frac{1}{n}$.

## Empirical Results

We collect empirical results for 2000 pairs of 100 letter texts using `chistat.pl`, from `http://www.staff.uni-mainz.de/pommeren/Cryptology/Classic/Perl/`. For English we use the book *Dr Thorndyke Short Story Omnibus* by R. Austin Freeman from Project Gutenberg. We extract a first part of 402347 letters (`Thorn1.txt`) and take the first 400000 of them for our statistic. In the same way for German we use *Die Juweleninsel* by Karl May from Karl-May-Gesellschaft (`Juwelen1.txt`, 434101 letters). For random texts we generate 400000 letters by Perl's random generator (`RndT400K.txt`). (All texts in `http://www.staff.uni-mainz.de/pommeren/Cryptology/Classic/Files/`.)

The results are in Tables 31, 32, and 33. We see that $\chi$—in contrast with the coincidence index $\kappa$—performs extremely well, in fact in our experiments it even completely separates English and German texts from random texts of length 100. It is a test with power near 100% and error probability near 0%. The $\chi$ test even distinguishes between English and German texts at the 5% error level with a power of almost 75%. For this assertion compare the 95% quantile for English with the first quartile for German.

Table 31: *Distribution of $\chi$ for 2000 English text pairs of 100 letters*

| | | | |
|---|---|---|---|
| Minimum: | 0.0500 | | |
| Median: | 0.0660 | Mean value: | 0.0663 |
| Maximum: | 0.0877 | Standard dev: | 0.0049 |
| 1st quartile: | 0.0630 | 5% quantile: | 0.0587 |
| 3rd quartile: | 0.0693 | 95% quantile: | 0.0745 |

The results for 100 letter texts encourage us to try 26 letter texts. To this end we need 104000 letters for each language. We extract the next 104009 letters from *Dr Thorndyke Short Story Omnibus* (`Thorn2.txt`), and the next 104293 letters from *Die Juweleninsel* (`Juwelen2.txt`). We construct random text by taking 104000 random numbers between 0 and 25 from `random.org` (`RndT104K.txt`). The results are in Tables 34, 35, and 36. The $\chi$-test is quite strong even for 26 letters: At the 5% error level its power is around 91% for English, 98% for German.

Table 32: *Distribution of $\chi$ for 2000 German text pairs of 100 letters*

| | | | |
|---|---|---|---|
| Minimum: | 0.0578 | | |
| Median: | 0.0792 | Mean value: | 0.0794 |
| Maximum: | 0.1149 | Standard dev: | 0.0074 |
| 1st quartile: | 0.0742 | 5% quantile: | 0.0677 |
| 3rd quartile: | 0.0840 | 95% quantile: | 0.0923 |

Table 33: *Distribution of $\chi$ for 2000 random text pairs of 100 letters*

| | | | |
|---|---|---|---|
| Minimum: | 0.0337 | | |
| Median: | 0.0400 | Mean value: | 0.0400 |
| Maximum: | 0.0475 | Standard dev: | 0.0020 |
| 1st quartile: | 0.0387 | 5% quantile: | 0.0367 |
| 3rd quartile: | 0.0413 | 95% quantile: | 0.0433 |

Table 34: *Distribution of $\chi$ for 2000 English text pairs of 26 letters*

| | | | |
|---|---|---|---|
| Minimum: | 0.0266 | | |
| Median: | 0.0666 | Mean value: | 0.0666 |
| Maximum: | 0.1169 | Standard dev: | 0.0120 |
| 1st quartile: | 0.0577 | 5% quantile: | 0.0488 |
| 3rd quartile: | 0.0740 | 95% quantile: | 0.0873 |

Table 35: *Distribution of $\chi$ for 2000 German text pairs of 26 letters*

| | | | |
|---|---|---|---|
| Minimum: | 0.0325 | | |
| Median: | 0.0784 | Mean value: | 0.0793 |
| Maximum: | 0.1538 | Standard dev: | 0.0154 |
| 1st quartile: | 0.0680 | 5% quantile: | 0.0562 |
| 3rd quartile: | 0.0888 | 95% quantile: | 0.1065 |

Table 36: *Distribution of $\chi$ for 2000 random text pairs of 26 letters*

| | | | |
|---|---|---|---|
| Minimum: | 0.0178 | | |
| Median: | 0.0385 | Mean value: | 0.0386 |
| Maximum: | 0.0680 | Standard dev: | 0.0075 |
| 1st quartile: | 0.0340 | 5% quantile: | 0.0266 |
| 3rd quartile: | 0.0429 | 95% quantile: | 0.0518 |