INDEX OF COINCIDENCE

Eleni Kontou
Student ID: 169021206

February 2020

Abstract

This paper explains thoroughly the Index of Coincidence as well as some of the basics of probability used. This topic was chosen because it is an interesting topic related to mathematics and we should all read about it. This paper shows how the formula to find the Index of Coincidence was derived, gives some applications of it and more significant results for specific languages. It also, has some examples to improve understanding for everyone.

1. Introduction

A significant number of individuals abandon engaging with math from an early stage in their lives, concluding, "I am simply not a math person, I'm bad at math". However, most of the time this is not the case since any individual can understand everything when it is explained in simple words. This paper will clarify the "Index of Coincidence" so anybody with constrained mathematical knowledge can understand it.

So, what is really meant by Index of Coincidence? Coincidence counting is the method (created by William F. Friedman in 1922) of putting two texts next to the other and checking the number of times the same letters appear in the same place across two texts. This ratio is known as the Index of Coincidence or IC for short. In other words, given a content string, the Index of Coincidence is the probability of two arbitrarily chosen letters being the same.[1]

2. Explanation of the Index of Coincidence

Suppose a particular letter appears k times among N letters. There are $\binom{N}{2} = \frac{N(N-1)}{2}$ ways we can pick two letters at random and $\binom{k}{2} = \frac{k(k-1)}{2}$ ways we can pick the designed letter, so the probability that both letters we pick are the designed letter will be:

$$\frac{\frac{k(k-1)}{2}}{\frac{N(N-1)}{2}} = \frac{k(k-1)}{N(N-1)}$$

[1] https://en.wikipedia.org/wiki/Index_of_coincidence

This stands for one letter, so it follows that for all letters the Index of Coincidence (IC) can be found by:

$$IC = \frac{\sum k_i(k_i-1)}{N(N-1)}$$

Where, k_i is the number of times the i-th symbol appears. [2]

3. Example

Consider the last part of Constantine P. Cavafy's poem "Ithaka":

"Ithaka gave you the marvelous journey.
Without her you wouldn't have set out.
She has nothing left to give you now.
And if you find her poor, Ithaka won't have fooled you.
Wise as you will have become, so full of experience,
You'll have understood by then what these Ithakas mean." [3]

The frequency count is as follows:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
17	2	2	6	27	6	3	17	11	1	3	10	3	12	25	2	0	7	10	17	14	7	7	1	9	0

Given the frequency values as shown in the table above, it is not difficult to calculate the index of coincidence of the part of the poem given above. As we can see the text has length 219 letters. So,

$$IC = \frac{17(16)+2(1)+2(1)+6(5)+27(26)+\cdots+9(8)+0(-1)}{219(218)} = 0.06292153659 \approx 0.0629$$

4. Independence

In probability, if one event affects in any way the probability of another event then these events are dependent. If on the other hand, one event does not affect another event then they are called independent. Understanding this concept can help understand the Index of Coincidence in more depth, as we recall that the Index of Coincidence is the probability to draw two equal letters from a text. Intuitively, it is clear that this probability changes if we draw two letters from a text, acknowledge that they are not equal and not place them back.

5. Application of Index of Coincidence

[2] https://pages.mtu.edu/~shene/NSF-4/Tutorial/VIG/Vig-IOC.html

[3] http://cavafis.compupress.gr/kave_17b.htm

The Index of Coincidence is being used in the analysis of common language in plaintext and in the examination of cryptanalysis. This procedure is utilized to cryptanalyze the Vigenere cipher, for example. Vigenere Cipher is a method of encrypting alphabetic text.[4] For a repeating-key polyalphabetic figure organized into a matrix, the coincidence rate inside every column will be, as a rule, the highest when the width of the matrix is a various of the key length. This reality can be utilized to decide the key length, which is initial phase in cracking the system.
Another application of Index of Coincidence is Regression, which represents an essential statistical method. In the case of data mining, it is also a significant analysis tool, used in classification applications through logical regressions as well as determined reports estimated using the least square or other methods. Non-linear data can be transformed into useful linear data and analyzed using linear regressions. Index of Coincidence is the universal test for data mining classification.[5]

6. Expected Values for Some Languages

We can find the Index of Coincidence for any language; however, this is an approximation as they depend on average frequencies of letters and these frequencies can be determined only approximately because it depends on the kind of text (scientific, historical, fiction etc.)

If you pick a letter from the English alphabet at random, at that point pick again at random, you have a 1 of every 26 probability (0.0385) that you picked the same letter both times. In the event that you pick a letter from English plaintext at arbitrary, at that point again at irregular, you have roughly a 2 out of 30 probability (0.0667) that you picked the same letter both times. This probability has been resolved through frequency contemplates. In the event that you compute the probability of coincidence for a text, at that point calculate the ratio of that probability to the probability of random coincidences in the English language to find the Index of Coincidence.

Therefore, the Index of Coincidence for English plaintext is determined by dividing the plaintext probability (0.0667 as shown above) by the random probability (0.0387 as indicated above). The resulting number is $\frac{0.0667}{0.0385} = 1.73$. Note that the Index of Coincidence of an English plaintext message is as a rule somewhere in the range of 1.50 and 2.00. The bigger the message, the closer it ought to be to 1.73.

Apply these steps for every language we can determine the index of coincidence for each language. Here are some examples:[6]

[4] https://www.geeksforgeeks.org/vigenere-cipher/

[5] https://books.google.co.uk/books?id=xeiPDwAAQBAJ&pg=PA369&lpg=PA369&dq=application+of+coincidence+index+in+data+mining&source=bl&ots=nzQeV58UEQ&sig=ACfU3U2iqUdROxU8e2-heRpqNYdd-ghKBQ&hl=en&sa=X&ved=2ahUKEwiPkunVqsznAhU8TRUIHTi6AsAQ6AEwDXoECAkQAQ#v=onepage&q=application%20of%20coincidence%20index%20in%20data%20mining&f=false

[6] http://www.crypto-it.net/eng/theory/index-of-coincidence.html

English	Russian	Spanish	Portuguese	Italian	French	German
1.73	1.76	1.94	1.94	1.94	2.02	2.05

7. General Formula

The above portrayal is just a prologue to utilization of the Index of Coincidence, which is identified with the general idea of correlation (correlation in statistics is the relationship of two random variables or bivariate data.). Different types of Index of Coincidence have been contrived; the "delta" IC measures the autocorrelation of a single distribution, whereas a "K" IC is utilized when coordinating two content strings. Although in certain applications consistent factors, for instance c and N can be disregarded, in general there is value in indexing each IC against the value to be expected for the null hypothesis, so that in each circumstance the expected value for no correlation is 1.0. Thus, any form of IC can be expressed as the ratio of the number of coincidences actually found to the number of coincidences expected, using the specific test arrangement.

The formula for K IC is:

$$IC = \frac{\sum_{j=1}^{N}[a_j = b_j]}{N/c}$$

Where N is the length of the two texts A and B, and the bracketed term is defined as 1 if the j-th letter of text A matches the j-th letter of text B, otherwise 0.

8. Conclusion

This paper explains the Index of Coincidence which is based on basic concepts of probability. We have seen how the formula was derived, and a simple example on how this formula can be applied in any text. Then, we have studied the applications of the Index of Coincidence and as a result we have shown the Index of Coincidence for some languages. After that we have looked at the general formula used in more complicated situations. It is an important part of cryptanalysis, as it makes it possible to evaluate the global distribution of letters in encrypted message for a given alphabet.

References

[1] Wikipedia 2019, Index of coincidence, viewed 30 January 2020, <https://en.wikipedia.org/wiki/Index_of_coincidence>

[2] Index of Coincidence and its applications in cryptanalysis, 1996, viewed 30 January 2020, <https://pages.mtu.edu/~shene/NSF-4/Tutorial/VIG/Vig-IOC.html>

[3] Constantine P.Cavafy, Ithaka, viewed 1 February 2020, < http://cavafis.compupress.gr/kave_17b.htm>

[4] Geeks for Geeks, Vigenere Cipher, viewed 7 February 2020 < https://www.geeksforgeeks.org/vigenere-cipher/>

[5] Adem Karahoca, Advances in Data Mining Knowledge Discovery and Applications, pg. 369, viewed 8 February 2020, <https://books.google.co.uk/books?id=xeiPDwAAQBAJ&pg=PA369&lpg=PA369&dq=application+of+coincidence+index+in+data+mining&source=bl&ots=nzQeV58UEQ&sig=ACfU3U2iqUdROxU8e2-heRpqNYdd-ghKBQ&hl=en&sa=X&ved=2ahUKEwiPkunVqsznAhU8TRUIHTi6AsAQ6AEwDXoECAkQAQ#v=onepage&q=application%20of%20coincidence%20index%20in%20data%20mining&f=false>

[6] Crypto-IT, Index of Coincidence, viewed 17 February 2020, < http://www.crypto-it.net/eng/theory/index-of-coincidence.html>