Index of Coincidence

Suppose that the ciphertext is

$$C = C_1, C_2, \ldots, C_N$$

which is known to have been encrypted by a Vigenère substitution with keyword of length p. For convenience we assume that N (the length of the ciphertext) is a multiple of p and set

$$M = \frac{N}{p}$$

If the original plaintext was

$$X_1, X_2, \ldots, X_N$$

then, for each $k = 1, 2, \ldots, M$, the k^{th} block of ciphertext

$$C_{(k-1)p+1}, C_{(k-1)p+2}, \ldots, C_{(k-1)p+p}$$

is obtained from the corresponding k^{th} block of plaintext

$$X_{(k-1)p+1}, X_{(k-1)p+2}, \ldots, X_{(k-1)p+p}$$

by p Ceasar like substitutions $\varphi_1, \varphi_2, \ldots, \varphi_p$

$$C_{(k-1)p+j} = \varphi_j(X_{(k-1)p+j})$$

To find the period p we compute the *index of coincidence* defined verbally as

$$I_C = \frac{\text{number of pairs of equal letters in the ciphertext}}{\text{total number of pairs of letters}}$$

If the numbers $N_A, N_B, \ldots N_Z$ denote respectively the total number of $A, B, \ldots, Z$ in the ciphertext, then

$$\binom{N_\alpha}{2} := \frac{N_\alpha\,(N_\alpha - 1)}{2} \qquad\qquad \alpha = A, B, \ldots, Z$$

gives the number of "$\alpha\alpha$" pairs in the ciphertext (observe that they can be very far apart). Thus

$$I_C = \frac{1}{N(N-1)} \sum_{\alpha=A}^{Z} N_\alpha(N_\alpha - 1)$$

Having computed this value, we can "estimate" the period to be

$$p \approx \frac{0.027N}{(N-1)I_C + 1 - 0.038N} \tag{1}$$

Of course when this is not an integer, the actual period is one of the closest integers.

A word of caution. Statistics carried out on too small samples may lead to grossly erroneous conclusions. The estimate in equation (1) cannot be relied upon for ciphertexts of less than 500 letters.

The reasoning that leads to expression (1) exhibits an interesting use of probabilities. The general idea is to obtain two expressions for I_C, and *solve* for p.

Let us rewrite the ciphertext in M rows, with the k^{th} row containing the k^{th} block of the ciphertext (hoping that p is the right period)

$$\begin{array}{cccc} C_1 & C_2 & \cdots & C_p \\ C_{p+1} & C_{p+2} & \cdots & C_{p+p} \\ \vdots & \vdots & \ddots & \vdots \\ C_{(k-1)p+1} & C_{(k-1)p+2} & \cdots & C_{(k-1)p+p} \\ \vdots & \vdots & \cdots & \vdots \end{array}$$

We then see that in the i^{th} column should (if p is right) consist of all the letters encrypted by the substitution φ_i. We write $M_\alpha^{(i)}$ for the number of letters equal to α (=A, B, ..., Z) in the i^{th} column. It is not difficult to see that

$$\sum_{\alpha=A}^{Z} N_\alpha \left(N_\alpha - 1\right) = \underbrace{\sum_{i=1}^{p} \sum_{\alpha=A}^{Z} M_\alpha^{(i)} (M_\alpha^{(i)} - 1)}_{(A)} + \underbrace{2 \sum_{i=1}^{p} \sum_{j=i+1}^{p} \sum_{\alpha=A}^{Z} M_\alpha^{(i)} M_\alpha^{(j)}}_{(B)} \tag{2}$$

The term (A) in (2) represents the contribution coming from pairs of letters in the same column, and the term (B) from pairs that are in different columns.

Note that for each i, we have

$$\sum_{\alpha=A}^{Z} M_\alpha^{(i)} = M$$

since this sum is the total number of letters in the column. Thus

$$(A) = \sum_{i=1}^{p} \sum_{\alpha=A}^{Z} (M_\alpha^{(i)})^2 - p\, M$$

Let us introduce the values

$$p_\alpha^{(i)} = \frac{M_\alpha^{(i)}}{M} \qquad\qquad i = 1, 2, \ldots, p$$

which give the "probability" of the letter α (=A, B, ..., Z) in the i^{th} column of the ciphertext.

Since φ_i permutes the letters of the alphabet, if p is right, then the sum

$$\sum_{\alpha=A}^{Z} (p_\alpha^{(i)})^2$$

should (for each i) be very close to

$$\sum_{\alpha=A}^{Z} (p_\alpha)^2$$

where p_α is the actual probability of α in the plaintext, and this value should be the same as those computed for standard plaintext english (if the plaintext is long enough). From our tables of english letter frequencies, we get

$$\sum_{\alpha=A}^{Z}(p_\alpha)^2 \simeq 0.065$$

hence we may conclude that

$$\begin{aligned}(A) &= M^2\sum_{i=1}^{p}\sum_{\alpha=A}^{Z}(p_\alpha^{(i)})^2 - pM\\ &\simeq 0.065pM^2 - pM\end{aligned}$$

To get an estimate of (B), we reason as follows. Note that

$$(B) = 2\,M^2\sum_{i=1}^{p}\sum_{j=i+1}^{p}\sum_{\alpha=A}^{Z}p_\alpha^{(i)}\,p_\alpha^{(j)} \tag{3}$$

For fixed i and j, let X and Y respectively denote the random variables obtained by selecting a letter at random from the i^{th} and j^{th} column. The sum

$$\sum_{\alpha=A}^{Z}p_\alpha^{(i)}\,p_\alpha^{(j)}$$

may be interpreted as the *probability* that $X = Y$, which we have denoted $P[X = Y]$. However, if we assume that the two substitutions φ_i and φ_j are unrelated, it is not unreasonable to expect that

$$P[X = Y] \simeq \frac{1}{26} \simeq 0.038$$

Substituting this estimate in (3) we obtain

$$(B) = 0.038\,p\,(p-1)\,M^2$$

Combining the approximations for (A) and (B), and recalling that $M = N/p$, we get

$$I_C = \frac{\sum_{\alpha=A}^{Z}N_\alpha\,(N_\alpha - 1)}{N\,(N-1)} = \frac{0.065\,N\,\frac{N}{p} - N + 0.038\,N\,(N - \frac{N}{p})}{N\,(N-1)}$$

Canceling N, multiplying both sides by $N - 1$ and solving for p, we get (1).

An Automated Ciphertext Only Attack of Vigenere

Suppose that the following ciphertext was encrypted using the Vigenere encipherment scheme.

```
BEJHQ DVAJS GMWUV AXTQU JSGEK EXBIE PMYHC MFLNF JNCKJ CTXFT
GWJQW TQTJT YTJXD ATXJN FHBEF UDTJX NREKJ AVHWW KMMCG KYAKG
ZNCEN EPTGL GKNGJ MXTJT YAOHS GVAJS GTWEN BKENB GETMD APWYH
GIZRU NNTQY MARIN NGLXT JTYTQ LJCWK JTJXX ETBLH VLLOX XWNOX
STUTW EKGXT KMZTG WFMQG LMGGI ETBAI PZYHG BWJWL YPQPJ RUYWO
OMMEE HSSGG YOHMM EIHAE TGJDV AFTYA JNGOJ RCGDF QKROH ZTVGK
SMGGY BGVTM GLIEU MWUEM NVGHK TJXXE GGISK MNSVA JRKZM TQYYH
GIJOR EJTQT QTGKT RVHFB QENSJ BYAPW YOKGX TKMZT GGJWI HAETG
REPMQ AABSG KMXFQ NSDCM NOPHS SWVMP TBSCK IQEUT SDQKL APBEI
PZNTU ITWGK XIPLZ CJYTR OTXTQ MMEOL MANEX EGFRO UMQIM XQYVH
JFHXH TVAJI TLFFG MDAPW MARIN NGLXP TNIEP VJIPW JEFPN LNWNC
VTYEV AFTIH AETGR EPMXL QGLEU MFBNB XHGWX HQNQD PHYBG VMAPZ
JDHHW LKZMT CGITT TSSKX STETZ SGLFN FTHCQ KIIPZ QYCEQ EZIJR
KXSCG AFTJL MEYGY HCMRA PDNNF TWEOH WEFBX PQLJD VHXUH YJRYA
NLGXA INLFR GLZFH XWADE JTJTS TQKNG JMYHG FXENO JSDRF BQENS
JBSGV AJFQK RSVHB HKVMT JXDAT XFCEN XTQFJ DDNYW JXSAN HSGVK
FIPHK ADNXE UTSDW LZRRT YIQGX PWKXU KGLIP OFRKT GLAMM EUTRE
QUOEE MJVKG HEUTI EUBLN VHWEF NHEVA JMWGI ETTGS QEZTG WJSRH
YIUFN TKLYH GBWRK ZMTKM NSVAJ ITWZT AMTTJ KTWQY KSWVM GQOJR
PFJNV TSDVH URQON DGGJW INFRF LKOTM MEKKK UVNWE UXHUT BYYUN
HHJTX BGXST JXUAV BJNVL ZFHXW APVJO HMMEU XHONH SIGLF NFLZC
JBXNQ PYHGG JCGLX IVRBH KVMCQ GXTTT NNUMM EOMTA NMJRV AJITY
TROXW SALYE OLTFI HAETG REPMY HGANS VHWYQ YYHGI WEUXS TMBSG
QYLRG TYBTB YAKGN SCANS VHWYQ YWERX FTGWN NLNWI GLFNF NXUTI
FTKHS SCEQH CONNI BSDKK JCVHG JGVYT JXJSV TGLKL MMGGY OHTSA
DLTLW MJTAK FNPRT VGKYH GLJSV TYEUM TPTHA EVANS NXYFC VYSDX
XUDFN TVXIT QTHAP WNDYH WLF
```

Using the test for monoalphabetic ciphers, we have

$$\frac{\#\text{ of equal neighbors in the ciphertext}}{n-1} = \frac{56}{1322} \approx 0.04236$$

So we can be relatively sure that this was not a simple Caesar shift. Now let us try to confirm this by approximating the length of the keyword used by calculating the index of coincidence

$$I_C = \frac{76870}{1323 \cdot 1322} \approx 0.04395$$

and thus

$$p \approx 4.046$$

Assuming that the ciphertext was encrypted using a Vigenere encipher with a 4-letter keyword, the task at hand now is to determine the 4 shifts involved. By performing a frequency analysis and drawing a histogram, we can get a *visual* clue that helps determine each shift. For example, Figure 1 suggests that the first shift sent a plaintext `A` to a ciphertext `F`.

We can also assign a number to each possible shift and let a computer decide which one is the most likely. To this end, let V_i be the vector of frequencies for the ciphertext letters

$$\{C_i, C_{i+p}, C_{i+2p}, \ldots\}$$

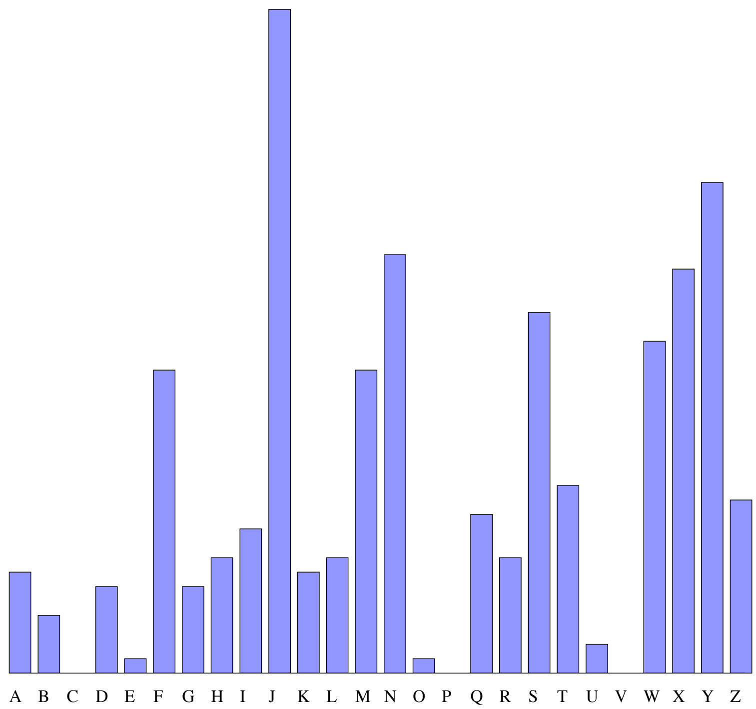

Figure 1: Histogram of 1^{st}, 5^{th}, 9^{th}, ... letters in the ciphertext.

for $1 \leq i \leq p$. For example,

$$V_1 = \{7, 4, 0, 6, 1, 21, 6, 8, 10, 46, 7, 8, 21, 29, 1, 0, 11, 8, 25, 13, 2, 0, 23, 28, 34, 12\}$$

is the vector plotted in Figure 1. Let E represent the vector of English frequencies given by

$$E = \{73, 9, 30, 44, 130, 28, 16, 35, 74, 2, 3, 35, 25, 78, 74, 27, 3, 77, 63, 93, 27, 13, 16, 5, 19, 1\}$$

and let V_i^α equal the vector of frequencies of the letters $\{C_i, C_{i+p}, C_{i+2p}, \ldots\}$ *after* being shifted by $A \to \alpha$. For example,

$$V_1^c = \{34, 12, 7, 4, 0, 6, 1, 21, 6, 8, 10, 46, 7, 8, 21, 29, 1, 0, 11, 8, 25, 13, 2, 0, 23, 28\}.$$

To each vector V_i^α, associate the number

$$V_i^\alpha \cdot E,$$

the usual dot (inner) product between V_i^α and E. For a fixed i, this value will be maximized by V_i^α if the ciphertext $\{C_i, C_{i+p}, C_{i+2p}, \ldots\}$ is decrypted by the Caesar shift $A \to \alpha$. For example,

$$V_1^c \cdot E = 34 \cdot 73 + 12 \cdot 9 + 7 \cdot 30 + \cdots + 23 \cdot 19 + 28 \cdot 1 = 11912.$$

The reason why the maximum value coincides with the decrypting Caesar shift is simple. Recall that the inner product may also be computed using the formula

$$V_i^\alpha \cdot E = |V_i^\alpha||E|\cos\theta,$$

where $|V|$ denotes the length of a vector and θ is the angle between the two vectors. Notice that for a fixed i, the length of the vector $|V_i^\alpha|$ is the same for all $\alpha = A, B, C, \ldots$. Therefore, the only difference in this value is accounted for by the factor of $\cos\theta$. This term is maximized when $\theta = 0$, or in other words, when V_i^α is a multiple of E, which is exactly what we were looking for *visually*.

The rest of the values are

$$\begin{array}{ccccc}
V_1^a \cdot E = 10552 & V_1^b \cdot E = 12153 & V_1^d \cdot E = 10433 & V_1^e \cdot E = 12960 & V_1^f \cdot E = 13861 \\
V_1^g \cdot E = 15970 & V_1^h \cdot E = 13374 & V_1^i \cdot E = 15733 & V_1^j \cdot E = 11739 & V_1^k \cdot E = 13950 \\
V_1^l \cdot E = 11059 & V_1^m \cdot E = 11520 & V_1^n \cdot E = 10356 & V_1^o \cdot E = 11217 & V_1^p \cdot E = 11385 \\
V_1^q \cdot E = 12115 & V_1^r \cdot E = 15833 & V_1^s \cdot E = 11292 & V_1^t \cdot E = 11235 & V_1^u \cdot E = 13287 \\
V_1^v \cdot E = 22296 & V_1^w \cdot E = 13760 & V_1^x \cdot E = 10356 & V_1^y \cdot E = 9377 & V_1^z \cdot E = 13275
\end{array}$$

which is clearly maximized by the shift that sends a ciphertext `A` to a `V`. Notice that this is the same as shifting a plaintext `A` to an `F`, which is the shift we came up with *visually*. The rest of the keyword, `FACT`, can be determined in a similar manner. The original plaintext is as follows:

> We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness.–That to secure these rights, Governments are instituted among Men, deriving their just powers from the consent of the governed, –That whenever any Form of Government becomes destructive of these ends, it is the Right of the People to alter or to abolish it, and to institute new Government, laying its foundation on such principles and organizing its powers in such form, as to them shall seem most likely to effect their Safety and Happiness. Prudence, indeed, will dictate that Governments long established should not be changed for light and transient causes; and accordingly all experience hath shewn, that mankind are more disposed to suffer, while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, pursuing invariably the same Object evinces a design to reduce them under absolute Despotism, it is their right, it is their duty, to throw off such Government, and to provide new Guards for their future security.–Such has been the patient sufferance of these Colonies; and such is now the necessity which constrains them to alter their former Systems of Government. The history of the present King of Great Britain is a history of repeated injuries and usurpations, all having in direct object the establishment of an absolute Tyranny over these States. To prove this, let Facts be submitted to a candid world.
>
> -Excerpt from *Declaration of Independence*

Exercises:

1. Assuming the following ciphertext was encoded using Vigenere, estimate the length of the keyword.

PLKST DRMNA EFHWB DLCGJ ILOZY DQUVX