Statistical Methods for Cryptography

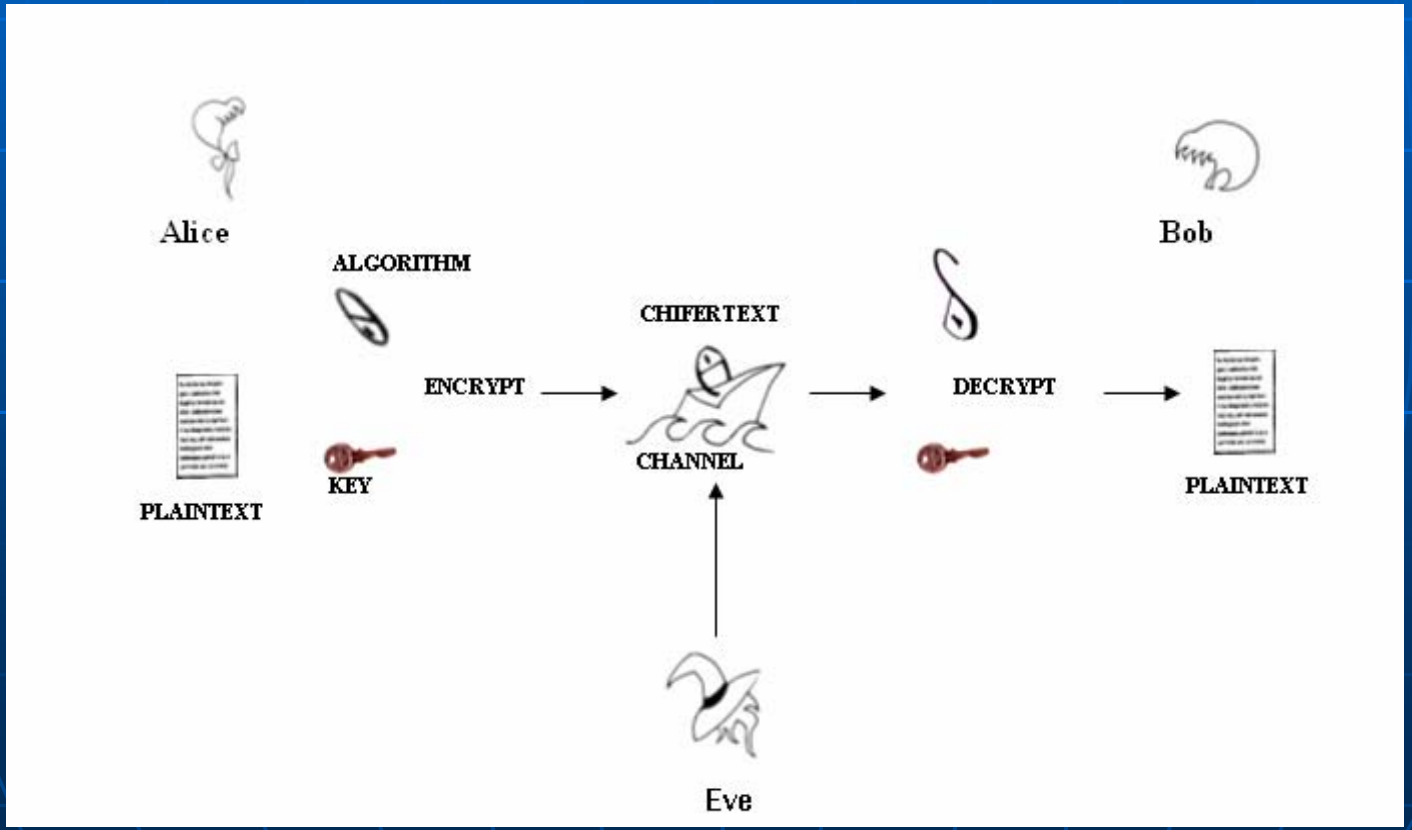
Alfredo Rizzi
Dipartimento di Statistica, Probabilità e Statistiche Applicate
Università di Roma “La Sapienza"
P.le A.Moro, 5 - 00185 Roma
alfredo.rizzi@uniroma1.it

- Cryptography
- Cryptoanalysis

- Encrypt
- Decrypt

- Linguistics, in particular Statistical Linguistics;
- Statistics, in particular the Theory of the Tests for the Analysis of Randomness and of Primality and Data Mining;
- Mathematics, in particular Discrete Mathematics;

- Bob Encrypt
- Alice Decipher
- Eve Decrypt



- Caesar Code:

  Monoalphabetic Substitution

SENATUSPOPULUSQUEROMANUS

VHQDWXVSRSXOXVTXHURPDQXV

Key=3

- Maths:

  24! = 6.2045e+23

  finite solution number

- Information Theory:

  6.2045e+14

  3.1710e+06 Years

- Vigenère:
  Multialphabetic Cipher
  Traicté des chiffres où secrètes manières d'escrire 1586

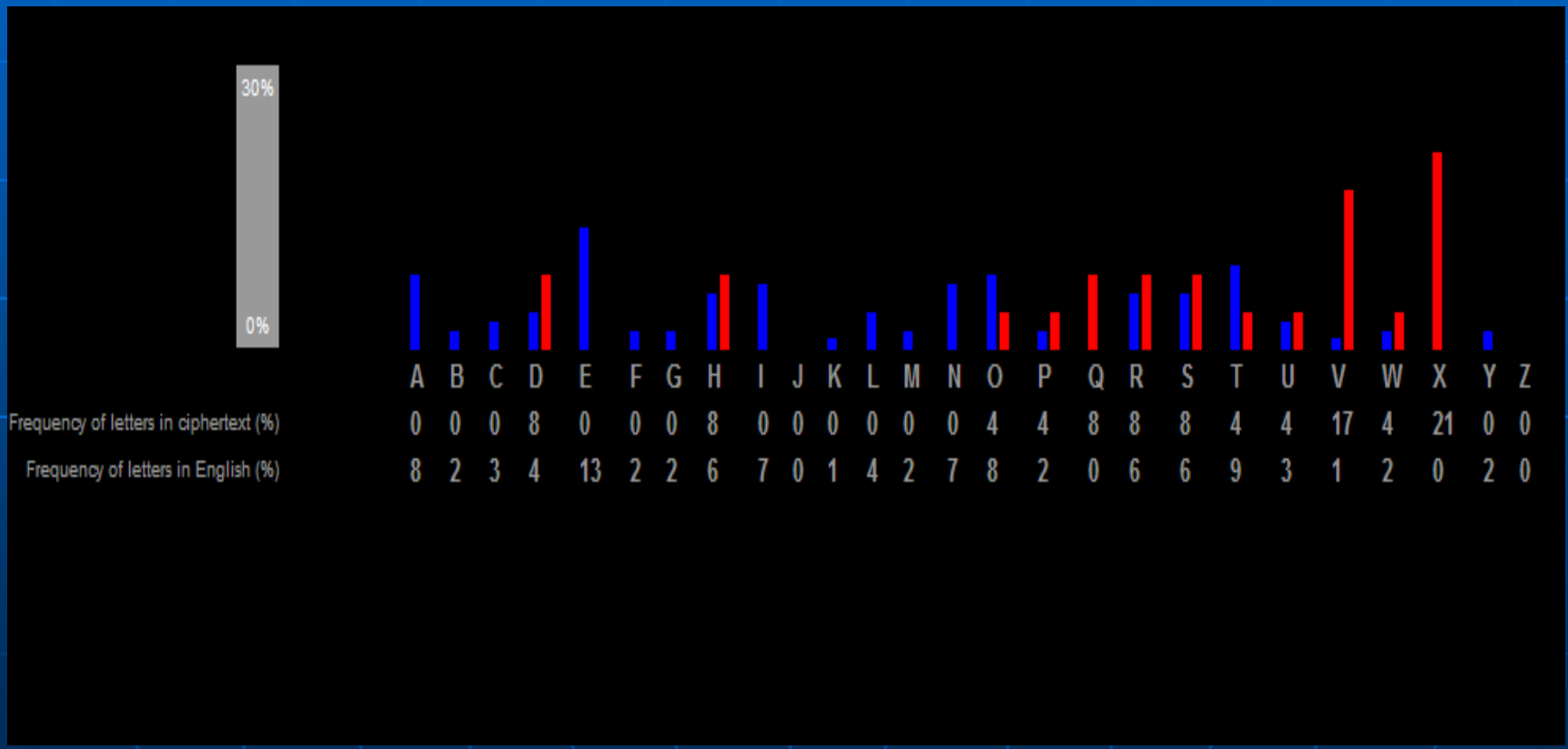
PLAINTEXT

KEY	A B C D E F G H I L M N O P Q R S T U V Z
A	A B C D E F G H I L M N O P Q R S T U V Z
B	B C D E F G H I L M N O P Q R S T U V Z A
C	C D E F G H I L M N O P Q R S T U V Z A B
D	D E F G H I L M N O P Q R S T U V Z A B C
E	E F G H I L M N O P Q R S T U V Z A B C D
F	F G H I L M N O P Q R S T U V Z A B C D E
G	G H I L M N O P Q R S T U V Z A B C D E F
H	H I L M N O P Q R S T U V Z A B C D E F G
I	I L M N O P Q R S T U V Z A B C D E F G H
L	L M N O P Q R S T U V Z A B C D E F G H I
M	M N O P Q R S T U V Z A B C D E F G H I L
N	N O P Q R S T U V Z A B C D E F G H I L M
O	O P Q R S T U V Z A B C D E F G H I L M N
P	P Q R S T U V Z A B C D E F G H I L M N O
Q	Q R S T U V Z A B C D E F G H I L M N O P
R	R S T U V Z A B C D E F G H I L M N O P Q
S	S T U V Z A B C D E F G H I L M N O P Q R
T	T U V Z A B C D E F G H I L M N O P Q R S
U	U V Z A B C D E F G H I L M N O P Q R S T
V	V Z A B C D E F G H I L M N O P Q R S T U
Z	Z A B C D E F G H I L M N O P Q R S T U V

- Lo statistico:

  Distribuzione di frequenze

  Parola Probabile



- # Index of coincidence for letter frequency (W.F. Friedman 1926)

$$I_c = \sum \frac{n_i(n_i - 1)}{N \cdot (N - 1)}$$

Per la distribuzione campionaria di Ic

$N \cdot (N-1) \cdot Ic + N = \Sigma n_i^2$.

Con semplici passaggi algebrici, sottraendo ad ambo i membri della precedente n^2/r e dividendo per r, si ha:

$1/r \cdot [N \cdot (N-1)\, Ic + N - N^2/r] = \Sigma n_i^2/r - N^2/r.$

Tenuto conto che $\Sigma n_i^2/r$ è il secondo momento e N^2/r^2 è il quadrato della media aritmetica, il secondo membro della precedente relazione è uguale alla varianza campionaria che verrà indicata con s^2. Se si suppone costante ed uguale ad 1/r la probabilità che ad una prova qualsiasi si verifichi uno qualsiasi dei simboli dell'alfabeto, per la varianza della distribuzione si avrà :

$\sigma^2 = N \cdot 1/r \cdot (1-1/r).$

Ricordando che: $\chi^2 = f \cdot s^2/\sigma^2$ ove f sono i gradi di libertà che nel nostro caso sono r-1.

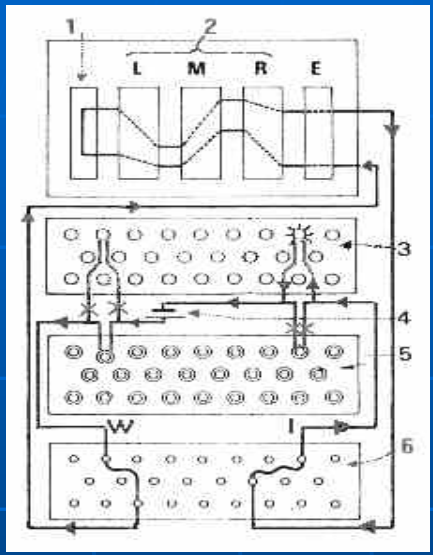
Sostituendo nella espressione precedente le due espressioni di σ^2 e di s^2 , dopo alcune semplificazioni algebriche si ha:

$\chi^2 = r \cdot (N-1) \cdot (Ic - 1)$ con r-1 gdl.

Esplicitando Ic:

$$Ic = (\chi^2)/r \cdot (N-1) + 1.$$

- ENIGMA



1)SCAMBIATORE
2)ROTORI
3)PANNELLO RETROILLUMINATO OUTPUT
4)BATTERIA
5)TASTIERA
6) PLUGBOARD O STEKER

- MACCHINA DI TURING

- Theorem 1

  Let two discrete s.v. (statistical variable) ***X***, ***Y*** assume the value: 0, 1, 2, …, m-1.

  Let ***X*** be uniformly distributed, that is it assumes the value *i* (*i* =0, 1, …,*m*-1) with probability 1/*m* and let the second s.v. ***Y*** assume the value *i* with probability *pi* .

  Then, if the two s.v. are independent, it follows that the s.v. ***Z*** obtained as a sum modulo m is uniformly distributed.

- *Proof.* If the s.v. X assumes the value *i*, then the s.v. ***Z*** can assume the values:

  i, *i*+ 1, *i*+2,…,*m*-1, 0, 1, 2, …, *i*-1

  respectively with probabilities:

  $p_0, p_1, \ldots, p_{m-1-i}, \ldots, p_{m-1}$

  according to the values

  0, 1, 2, …, *m*-*i*, *m*+1-*i*, *m*+2-*i*,…, *m*-1

  assumed by the ***Y***.

- If we let i assume the values 0, 1, 2, . . . , m-1, it follows that the s.v. Z assumes the general value h (h = 0,1, . . . ,m-1) with probability:

$$\begin{cases} \frac{1}{m} p_h & if\ X = 0 \quad Y = h \\ \frac{1}{m} p_{h-1} & if\ X = 1 \quad Y = h-1 \\ \vdots & \vdots \quad \vdots \\ \frac{1}{m} p_0 & if\ X = h \quad Y = 0 \\ \vdots & \vdots \quad \vdots \\ \frac{1}{m} p_{h-1} & if\ X = m-1 \quad Y = h+1 \end{cases}$$

It follow immediately by summation:

$$P(Z = h) = \frac{1}{m} \sum_{i=0}^{h} p_i + \frac{1}{m} \sum_{i=h+1}^{m-1} p_i = \frac{i}{m}$$

The above Theorem can be easily generalized to the sum (mod *m*) of *n* s.v., one of which be uniformly distributed.

- **Theorem 2**

  Let two independent s.v. X and Y assume the values: 0, 1, 2, …m-1 respectively with probabilities:

  $$p_0, p_1, \ldots, p_{m-1-i}, \ldots, p_{m-1}$$

  $$q_0, q_1, \ldots, q_{m-1-i}, \ldots, q_{m-1}$$

  Then, if the s.v. $Z=X+Y$ ($mod\ m$) is uniformly distributed and m is a prime number, at least one the two s. v. X and Y is uniformly distributed.

- Proof : The table of the sum (mod m) of the s.v. Xand Y is as follows:

	0	1	2	…	i	…	$m-2$	$m-1$	
0	0	1	2		i		$m-1$	$m-1$	p_0
1	1	2	3		$i+1$		$m-1$	0	p_1
2	2	3	4		$i+2$		0	1	p_2
⋮									
j	j	$j+1$			$i+j$		$j+m-2 \pmod m$	$j+m-1 \bmod m$	p_j
⋮									
$m-1$	$m-1$	0			$i-1$		$m-3$	$m-2$	p_{m-1}
	q_0	q_1			q_i		q_{m-2}	q_{m-1}	1

- As the X and Y are independent, the s. v. Z assumes the value 0 with probability:

$$p_0 q_0 + \cdots + p_2 q_{m-2} + p_1 q_{m-1}$$

Such probability, according to the hypothesis of uniform distribution of Z, must be 1/m.

By the same token, in order to compute the probabilities that the s.v. Z assumes the values 1,2,. . . ,m-1, the following system can be written:

$$\begin{cases} p_0q_0 + p_{m-1}q_1 + \cdots + p_2q_{m-2} + p_1q_{m-1} = 1/m \\ p_1q_0 + p_0q_1 + \cdots + p_3q_{m-2} + p_2q_{m-1} = 1/m \\ p_2q_0 + p_1q_1 + \cdots + p_4q_{m-2} + p_3q_{m-1} = 1/m \\ \vdots \\ p_{m-1}q_0 + p_{m-2}q_1 + \cdots + p_1q_{m-2} + p_0q_{m-1} = 1/m \end{cases}$$

- If p_i are know (the reasoning is the same if the q_i are known) the above system is a system of *m* equations in the *m* unknowns .
- It follows:

$$q_i = \frac{\begin{vmatrix} p_0 & \cdots & 1/m & \cdots & p_1 \\ p_1 & \cdots & 1/m & \cdots & p_2 \\ \vdots & & & & \\ p_{m-1} & \cdots & 1/m & \cdots & p_0 \end{vmatrix}}{\Delta}$$

- Where Δ is the determinant of the matrix of the coefficients, and it can be easily seen to be 0 if at least one. (In the opposite case the s. v. $\boldsymbol{X}$ is uniformly distributed). In fact, it is a circulating determinant.

- In order to show the theorem in general, it is sufficient to show that

$$q_i = q_0 \quad \forall\, i = 1,2,\ldots,m-1$$

- In this case, as $\Sigma q_i = 1$, $q_i \geq 0$, $q_i = 1\backslash m$.
- Then, it is sufficient to show that:

$$\begin{vmatrix} 1/m & p_{m-1} & \cdots & p_1 \\ 1/m & p_0 & \cdots & p_2 \\ \vdots & p_{m-1} & & \\ 1/m & p_{m-2} & \cdots & p_0 \end{vmatrix} = \begin{vmatrix} p_0 & p_{m-1} & \cdots & 1/m & \cdots & p_1 \\ p_1 & p_0 & \cdots & 1/m & \cdots & p_2 \\ \vdots & & & & & \\ p_{m-1} & p_{m-2} & \cdots & 1/m & \cdots & p_0 \end{vmatrix}$$

- The two determinants are equal because, in order to trasform the second into the first one, it is necessary to perform, owing to the circulating nature of the permutations of the p_i, m-2 inversions over the rows and m-2 inversions over the columns, that is $2(m-1)$ inversions over rows and columns in all. If an even number of inversions is performed, the sign of the determinant is unchanged.

- In this way, for istance, if m-1=3 we have that the numerator of q_0 is

$$\begin{vmatrix} 1/3 & p_1 & p_2 \\ 1/3 & p_2 & p_0 \\ 1/3 & p_0 & p_1 \end{vmatrix} = \begin{vmatrix} p_0 & 1/3 & p_2 \\ p_1 & 1/3 & p_0 \\ p_2 & 1/3 & p_1 \end{vmatrix}$$

SCHIFT REGISTERS:

consentono di ottenere sequenze pseudo-casuali, che *sembrano* essere casuali, anche se analisi approfondite consentono di individuare alcune regolarità.

- La pseudo-casualità va intesa nel senso che:
- Il numero degli 1 deve essere approssimativamente uguale a quello degli zeri. Analogamente la frequenza di ognuna delle quattro coppie, (00) (01) (10) (11), deve essere approssimativamente la stessa; ciò deve valere anche per le 9 terne di 3 elementi, e così via, per le 2n ennuple di n elementi.
- Sequenze di pochi 1 o di zeri devono essere più frequenti delle sequenze lunghe di 1 o di zeri.
- L'intera sequenza non deve avere autocorrelazione e la frequenza di ogni k-upla (k=1, 2,…,n), deve essere la stessa; ad esempio per k=3 la terna 000 deve avere la stessa frequenza di 001, 010,…111.

- Il funzionamento degli shift registers. Sia x_0 un vettore di n+1 componenti:
  - $x_0 \equiv \| x_{0,0}; x_{1,0}; x_{2,0}; \ldots; x_{n,0} \|$

con $x_{i,0}$ (i=0,1,…,n) cifre binarie non tutte nulle. Consideriamo, inoltre, l'evoluzione del vettore nel tempo :

- $x_t \equiv \| x_{0,t}; x_{1,t}; x_{2,t}; \ldots; x_{n,t} \|$

con:

$x_{n,t} = x_{n-1,t-1}$

$x_{n-1,t} = x_{n-2,t-1}$

$x_{n-2,t} = x_{n-3,t-1}$

…..

$x_{1,t} = x_{0,t-1}$

$x_{0,t} = f(x_{1,t}; x_{2,t}; \ldots; x_{n,t})$

- In definitiva un registro a scorrimento lineare è così caratterizzato:
- E' costituito da n+1 celle ognuna contenente una cifra binaria;
- Evolve dallo stato t-1 allo stato t, t=1,…r,…. riempiendo tutte le n celle da 1 ad n con il contenuto della cella immediatamente a sinistra nello stato t-1.
- Allo stato iniziale il contenuto di almeno una cella è 1.
- Il contenuto della cella 0 al tempo t è calcolato come funzione lineare delle celle da 1 a n dello stesso stato t.
- La lunghezza massima del periodo è 2n -1
- Il periodo della sequenza è massimo qualora i coefficienti ci, (i=1,2,…,n), siano quelli di un polinomio primitivo

- TEOREMA 1.

  La lunghezza massima del periodo di un registro a scorrimento lineare di dimensione n è $2n - 1$.
- TEOREMA 2
- Se il registro a scorrimento ha periodo massimo, allora i coefficienti ci (1=1,2,…,n) sono coefficienti di un polinomio irriducibile di grado n in GF(2).
- Si noti che l'inverso del teorema non è valido come si verifica nell'esempio 3.
- Ricordiamo che: In GF(p) un polinomio intero di grado n è primitivo se, oltre ad essere irriducibile, ossia non essere il prodotto di almeno due polinomi di grado ≥1, divide $x^m - 1$ per m non inferiore a $p^n - 1$.

Fermat's little theorem (1637)

- Let a, p integer | $(a, p)=1$. If p is prime

  then:

$$a^{p-1} \equiv 1 \bmod p$$

- The composite p for which Fermat's little theorem is true for every a are called Carmichael numbers.

  the first three are 561, 1105, 1729.

  Although these number are very rare, it is proved that there are infinitely many of them.

- P we can find the solution in polynomial time O(f(n))
- NP (Non-deterministic Polynomial) we can just check the solution in polynomial time

M.Agrawal, N.Kayal and N. Saxena

- deterministic test based on the following:
- Theorem 4:

  p is prime if and only if

  $(x-a)^p \equiv (x^p-a) \pmod{p}$

  where a is a number prime with p.

- Agrawal, Kayal and Saxena

  algorithm would run in at most $O((\log n)^{12} f(\log \log n))$ time where f is a polynomial.