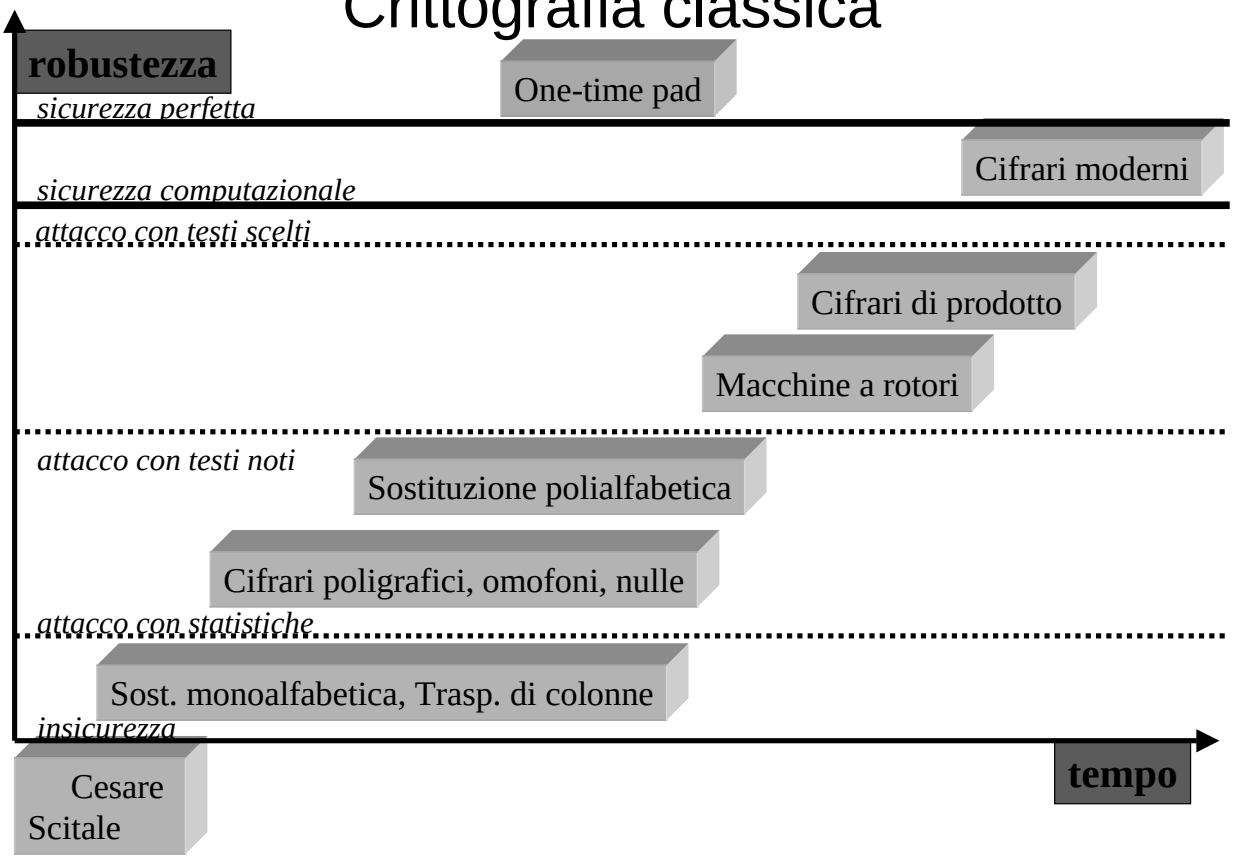
Crittanalisi dei cifrari simmetrici

- Sostituzione monoalfabetica di caratteri
- Sostituzione polialfabetica di caratteri
- One-time pad
- Cifrario perfetto
- Cifrario composto

Crittografia classica



Crittografia classica: la sostituzione monoalfabetica

regola di sostituzione (o chiave)

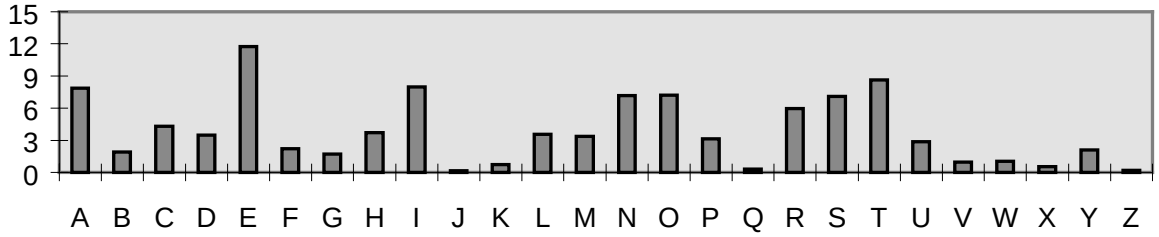
A B C D E F G H I L M N O P Q R S T U V Z
| | | | | | | | | | | | | | | | | | | | |
Q E M R F Z T B L U P O N H A S C G V D I

testo in chiaro:**CRITTOGRAFIA**
testo cifrato: **MSLGGNTSQZLQ**

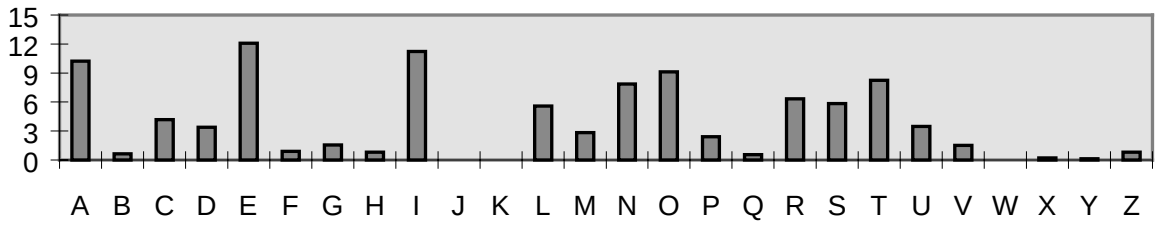
$$26! = 4 \times 10^{26}$$

Statistiche dei caratteri

Frequenze di occorrenza (%) nella lingua Inglese



Frequenze di occorrenza (%) nella lingua Italiana



Probabilità di occorrenza

Statistiche di digrammi e trigrammi

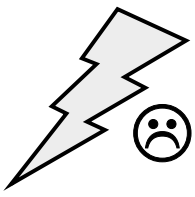
Lingua inglese

TH	3,16%,
IN	1,54%
ER	1,33%
RE	1,3%
ecc.	
THE	4,72
ING	1,42
ecc.	

- un linguaggio naturale è ridondante
- la probabilità di occorrenza di stringhe corte è indipendente dal testo
- in un testo lungo le frequenze di occorrenza approssimano le probabilità

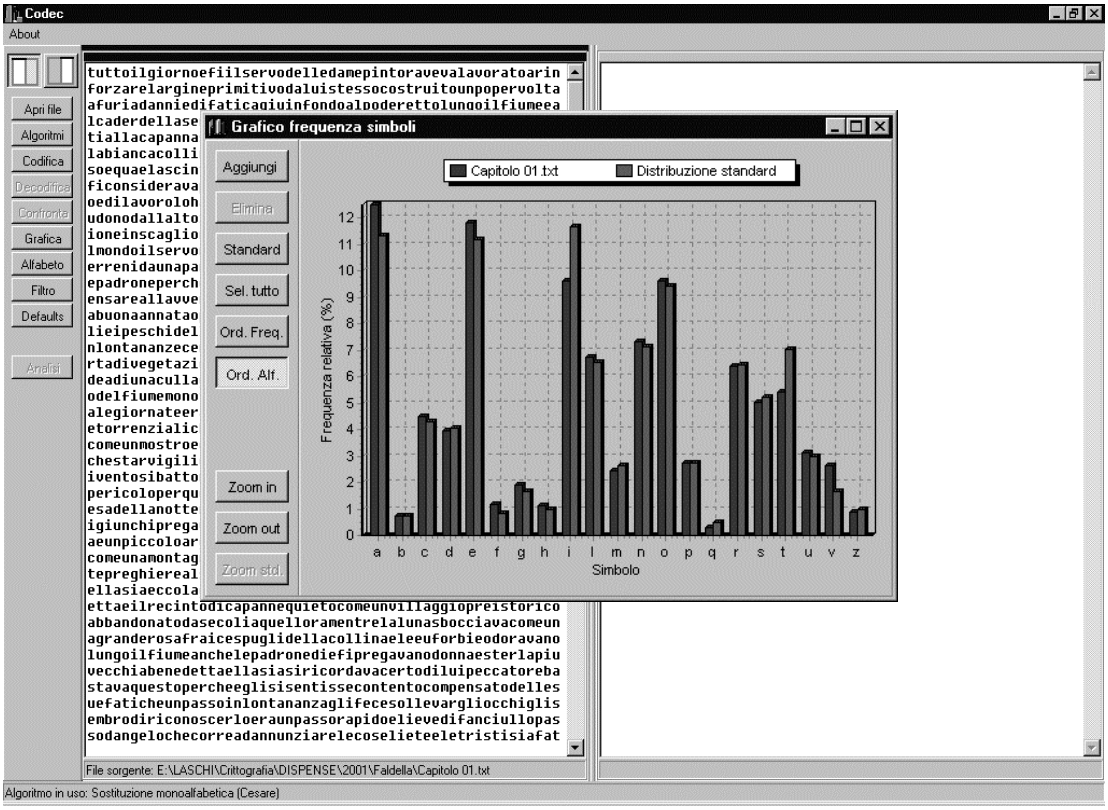
Il punto debole della monoalfabetica

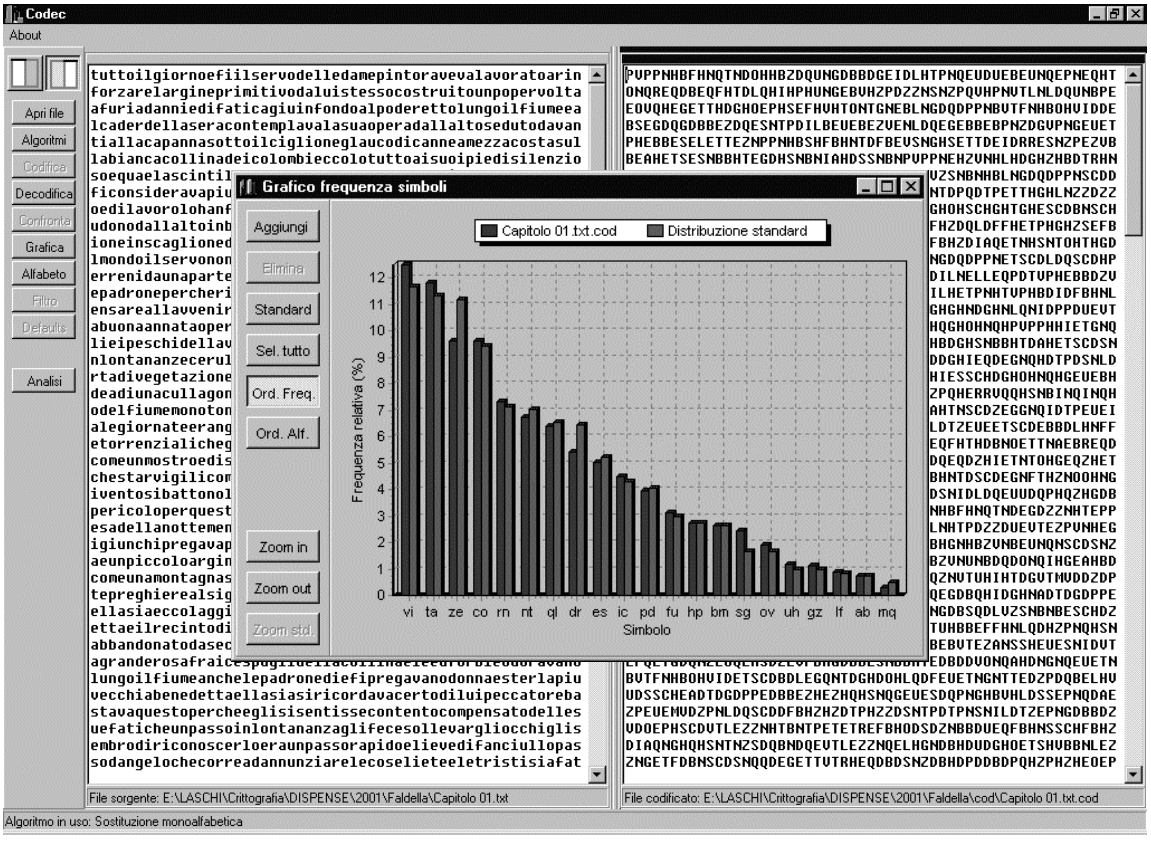
Le proprietà statistiche di ogni carattere del testo in chiaro vengono trasferite immutate sul carattere che lo sostituisce nel testo cifrato



Un grande spazio delle chiavi può non servire a nulla!

Verifiche con Codec





Rottura di una monoalfabetica

1) testo inglese cifrato con una sostituzione monoalfabetica
UNUFT OST, SII QNUF RBU GQFIO, HQDWXRUF KURGQFVL SKO BQLR ...
2) analisi frequenziale dei caratteri del testo cifrato:
U 15,3%, R 9,8% , S 7,8%
UF, FU e RB 3,3%,
RBU 3,5%, USV 2%.
S spesso da sola e 13% come prima lettera di una parola
3) ipotesi: U ↔ ***E***, R ↔ ***T***; conferma: RBU ↔ ***THE*** ; conseguenza: B ↔ ***H***.
4) sostituzioni:
EN***E***FT OST, SII QN***E***F ***THE*** GQFIO, HQDWX***TE***F K***ET***GQFVL SKO ***H***QL***T*** .....
5) nuove ipotesi: S ↔ ***A*** , UF ↔ ***ER***, FU ↔ ***RE*** ; conseguenza: F ↔ ***R***.
6) sostituzioni:
EN***ER***T O***A***T, ***A***II QN***ER THE*** GQ***R***IO, HQDWX***TER*** K***ET***GQ***R***VL ***A***KO ***H***QL***T***..
N.B. 18 caratteri su un totale di 46, cioè circa il 40%.
7) Statistiche non ancora prese in considerazione e significati probabili :
EVERY DAY, ALL OVER THE WORLD, COMPUTER NETWORKS AND HOSTS…

La sostituzione polialfabetica (Vigenere)

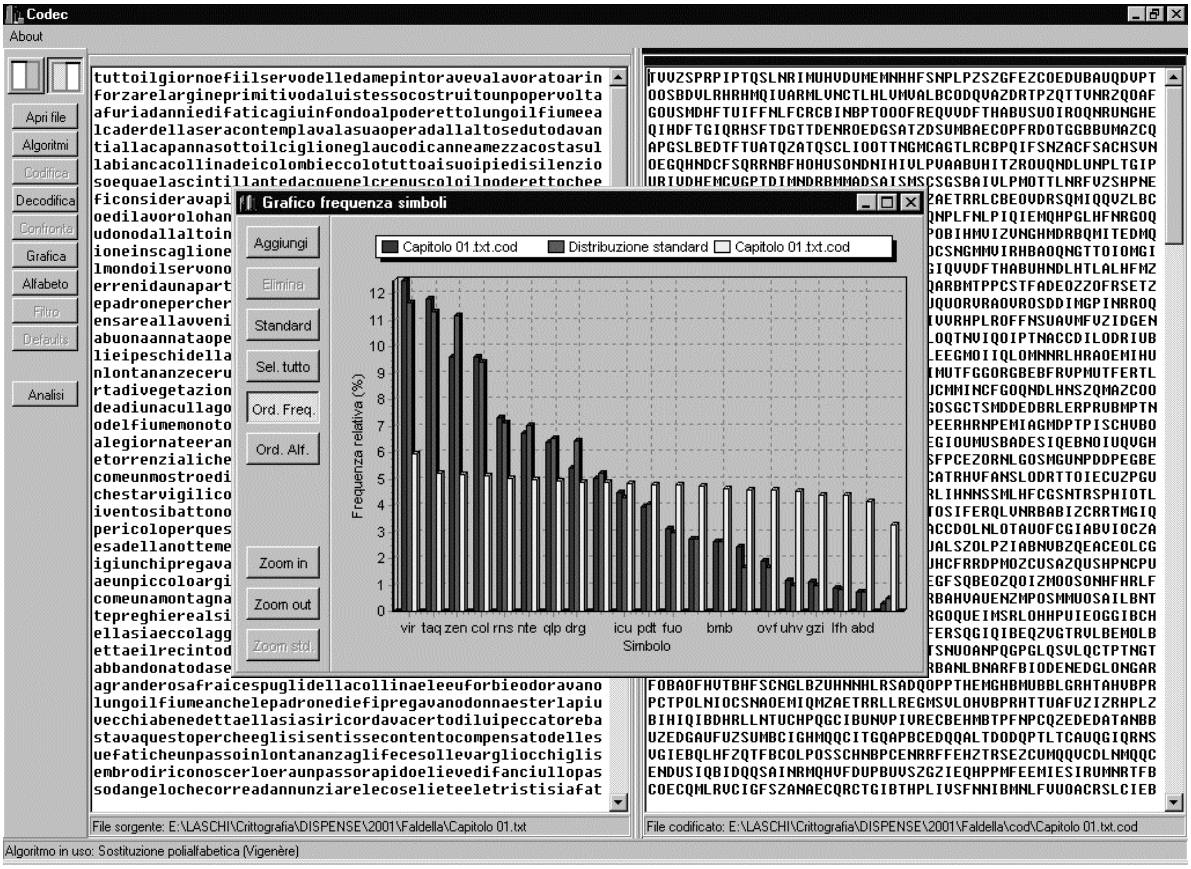
testo

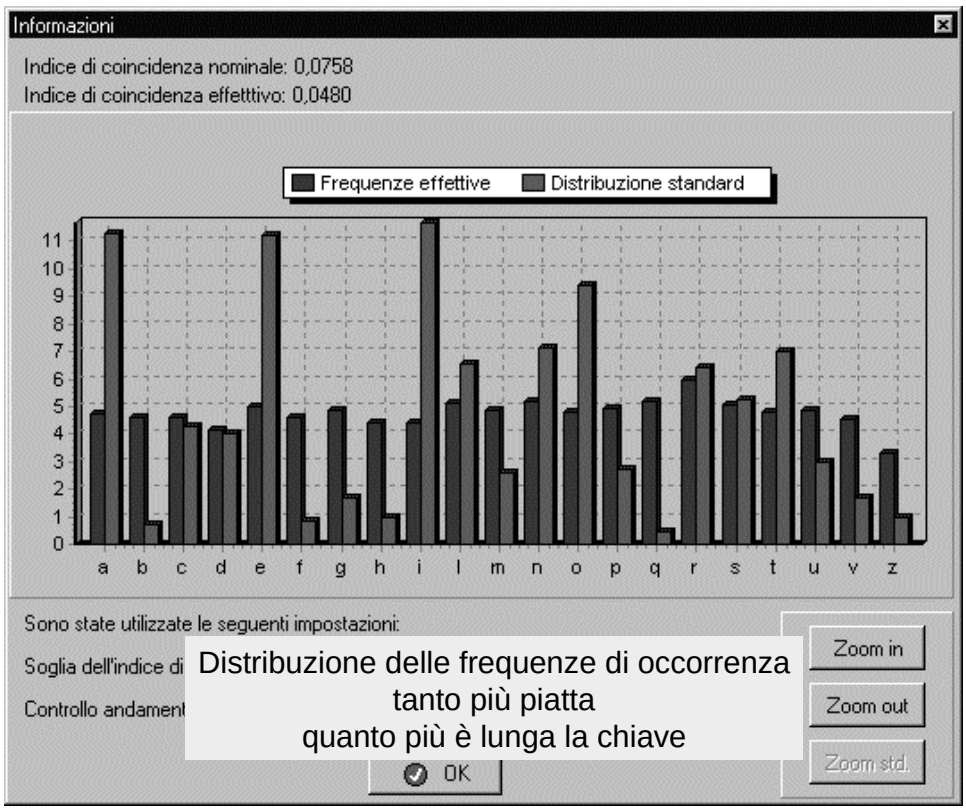
chiave →

ABCDEFGHIJKLMNOPQRSTUVWXYZ
BCDEFGHIJKLMNOPQRSTUVWXYZA
CDEFGHIJKLMNOPQRSTUVWXYZAB
DEFGHIJKLMNOPQRSTUVWXYZABC
.....
ZABCDEFGHIJKLMNOPQRSTUVWXY

Chiave: CIAO
testo in chiaro : DOMANI NON POSSO
Cifratura:

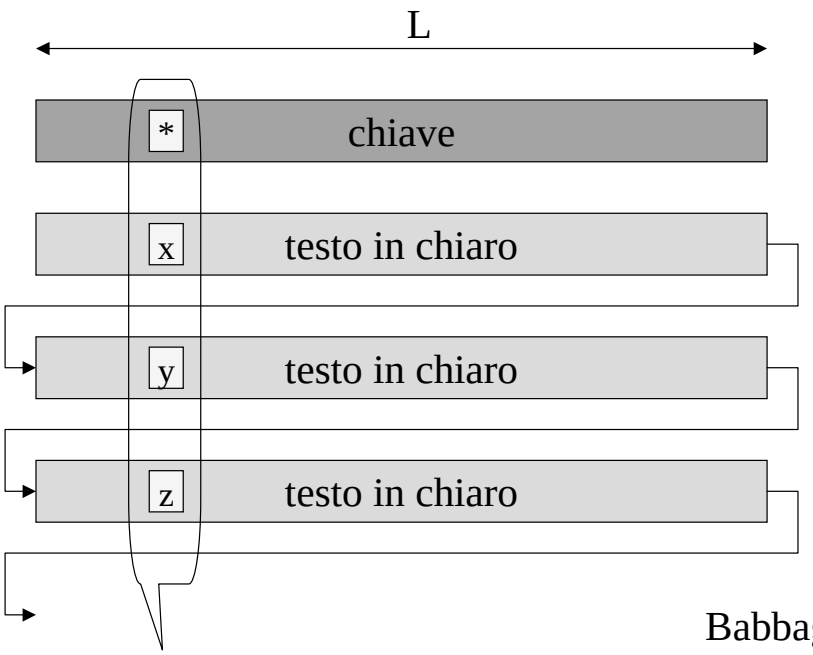
C I A O C I A O C I A O C I
D O M A N I N O N P O S S O
F Z M O P S N C P A O H U Z





Distribuzione delle frequenze di occorrenza
tanto più piatta
quanto più è lunga la chiave

Il punto debole della polialfabetica



Sostituzione monoalfabetica

Babbage (1854?)
Kasiski (1863)
Friedman (1925)

1-Indice di coincidenza

DATO: stringa lunga n formata da simboli di A = {1,2, .., 26}

DEFINIZIONE: *l'***indice di coincidenza I** *della stringa fornisce la probabilità di trovare due simboli identici scegliendo a caso la loro posizione all'interno della stringa .*

Stima di I

- **Occorrenze dei simboli: n(1), n(2), .., n(26)** con Σ_i **n(i) = n**
- **N° di coppie di simboli in diversa posizione: n (n-1) /2**
- **Coppie formate da due "i": n(i) (n(i)-1) /2**
- **Coppie formate da due simboli identici: ½** Σ_i **n(i) (n(i)-1)**

$$I = \frac{\sum_{i=1}^{26} n(i)\left[n(i)-1\right]}{n(n-1)}$$

2-Indice di coincidenza di un linguaggio naturale

- **Probabilità dei caratteri: p(1), p(2), .., p(26)**
- **Probabilità di due "i": $p(i)^2$**
- **Probabilità di due caratteri identici: $\Sigma_i\, p(i)^2$**

$$I_l = \sum_{i=1}^{26} p(i)^2$$

IPOTESI
•testi lunghi
•argomenti diversi

$$I_l = \sum_{i=1}^{26} f(i)^2$$

Inglese:0,0667
Italiano:0,0738
Russo: 0,0529

3-Indice di coincidenza di una stringa di caratteri scelti a caso

- **Probabilità dei caratteri: p(1) = p(2) = .. = p(26) = 1/26**

$$I_c = \sum_{i=1}^{26} 1/26^2 = 0{,}0384$$

$$I_c = \sum_{i=1}^{21} 1/21^2 = 0{,}0476$$

4-Indice di coincidenza di un testo cifrato (monoalf., trasp., polialf.)

IPOTESI: simboli di chiave scelti a caso
I*: valore atteso dell'indice di coincidenza di un testo cifrato

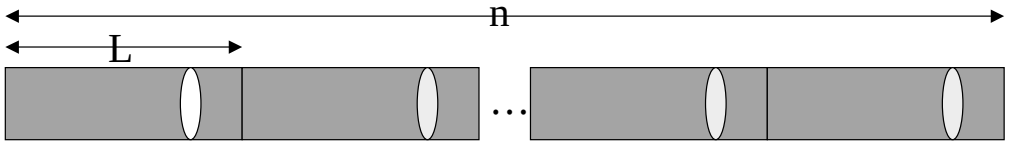
Sostituzione monoalfabetica: $I^* = I_l$
Trasposizione: $I^* = I_l$
Sostituzione polialfabetica: $I_l > I^* > I_c$

Soglia: 0,85 I_l

5-Indice di coincidenza di un testo cifrato con una polialfabetica

IPOTESI: chiave lunga **L**

- **Probabilità che due simboli della stringa, scelti a caso, distino k×L**



$$p = [(n/L)-1)]/n-1$$

- **Indice di coincidenza per caratteri a distanza k×L: I_l**
- **Indice di coincidenza per ogni altra distanza: I_c**
- **Valore atteso dell'indice di coincidenza per una polialfabetica:**

$$E\,(I^*) = p \times I_l + (1-p) \times I_c$$

6 - La formula di Friedman

$$p = [(n/L)-1]/n-1$$

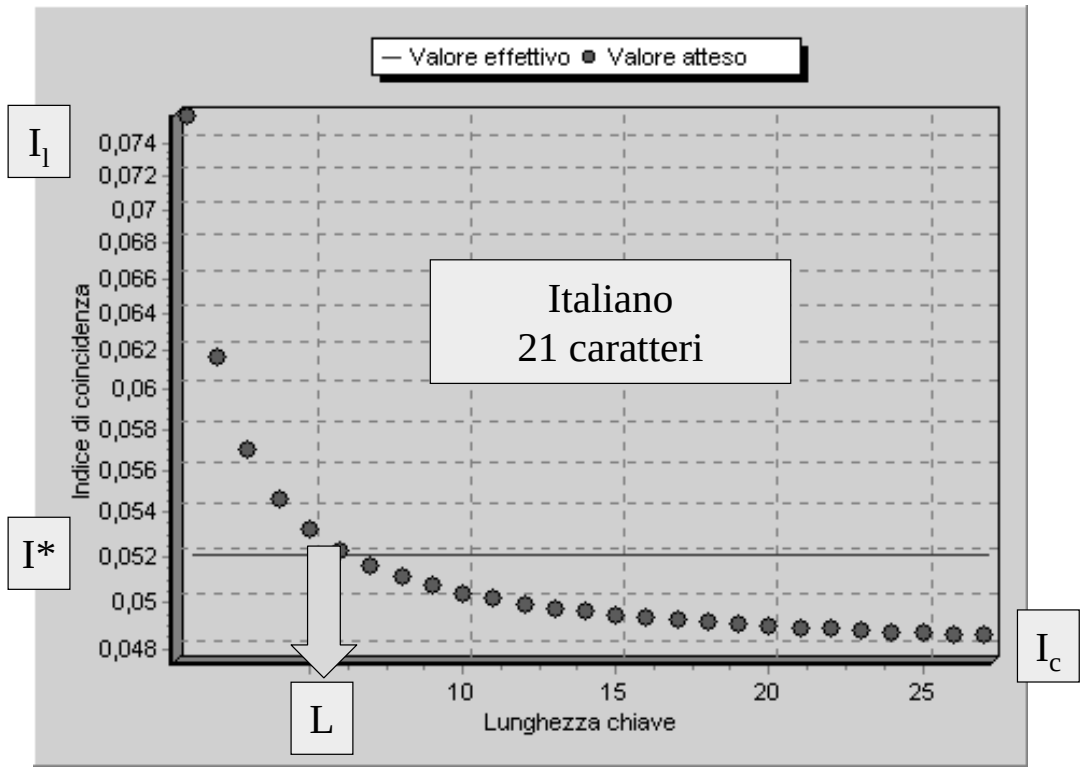
$$E\,(I^*) = p \times I_l + (1-p) \times I_c$$

$$E\,(I^*) = I^* = F\,(n, I_l, I_c, L)$$

$$L = \frac{I_l - I_c}{I^* - I_c + (I_l - I_c)/n}$$

N.B. formula valida solo per L “piccolo”

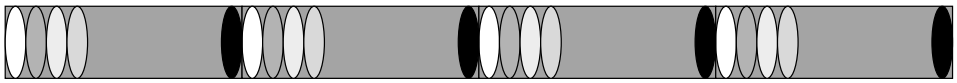
$F(L)$ e I_l, I_c, I^*

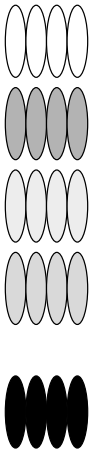


L grande: test di Friedman

Sia L un'ipotesi di lunghezza di chiave

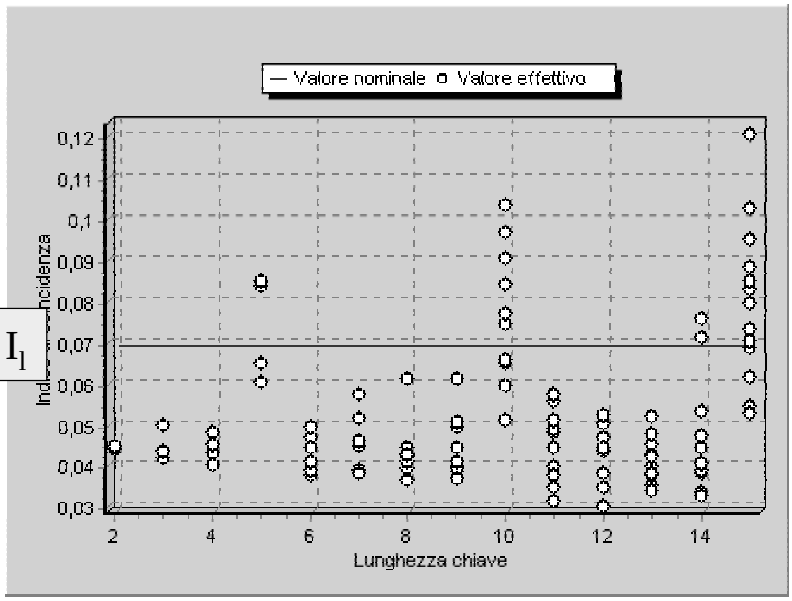
Si suddivide **c** in **L** sottoinsiemi contenenti ciascuno **n/L** simboli posti ad una distanza multipla di **L** uno dall'altro





Si calcola l'indice di coincidenza I di ciascun sottoinsieme.
Se L è la vera lunghezza, ogni sottoinsieme contiene solo simboli ottenuti con una monoalfabetica:
il valore del loro indice di coincidenza
deve dunque essere molto prossimo a I_l

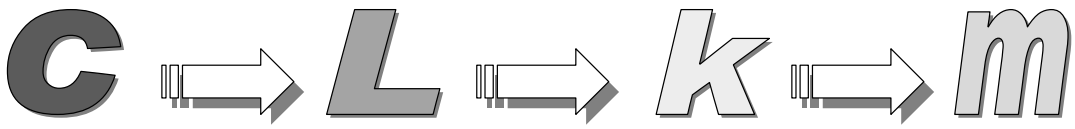
Esempio: testo in chiaro cifrato con una chiave di lunghezza 5



I_l

I_i centrati attorno a **I_l** per L multiplo intero di 5
.. ma anche sempre più sparpagliati: il numero di simboli su cui calcolare il valore dell'indice di coincidenza cala sempre più
Si cerca quello che ha il minimo scarto quadratico da I_l

Attacco con solo testo cifrato



Rappresentazione matematica della polialfabetica

k\m

	A	B	C	D	E	F	G	H	I	L	M	N	O	P	Q	R	S	T	U	V	Z
A	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
B	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	0
C	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	0	1
D	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	0	1	2
..	.	.	.	.																	

$$c = (m + k) \bmod 21$$

La rappresentazione numerica del carattere di testo cifrato è data dal resto della divisione per 21 della somma delle rappresentazioni numeriche del carattere di testo e del carattere di chiave

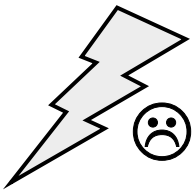
Attacco con testo in chiaro noto o probabile

• L lunghezza della chiave
•k(1), k(2), ... k(L) interi corrispondenti ai caratteri di chiave (0=A, 1=B, ecc.).
•m(1), m(2), ... m(L) interi corrispondenti ai caratteri di testo in chiaro

Cifratura: $c(i) = (m(i) + k(i)) \bmod 26$ per $1 \leq i \leq L$
Decifrazione: $m(i) = (c(i) + 26 - k(i)) \bmod 26$ per $1 \leq i \leq L$.

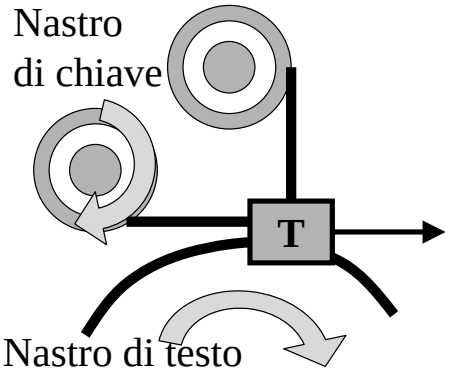
$k(i) = (c(i) + 26 - m(i)) \bmod 26.$

Attacco con testo in chiaro noto
Attacco con parola probabile



Il cifrario Vernam
One time pad

Il Cifrario di Vernam (1917)



Telegrafo di Vernam

- *codifica binaria (5 bit)*
- *chiave lunga quanto il testo*

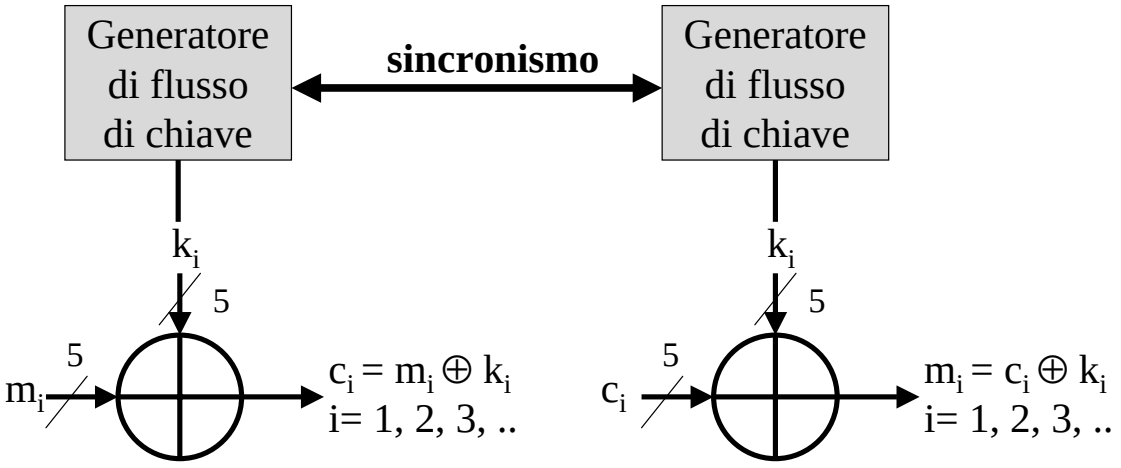
	Chiave							
Testo	000	001	010	011	100	101	110	111
000	000	001	010	011	100	101	110	111
001	001	000	011	010	101	100	111	110
010	010	011	000	001	110	111	100	101
011	011	010	001	000	111	110	101	100
100	100	101	110	111	000	001	010	011
101	101	100	110	111	001	000	011	010
110	110	111	100	101	010	011	000	001
111	111	110	101	100	011	010	001	000

8 righe: 8 permutazioni di {0,1,..,7}

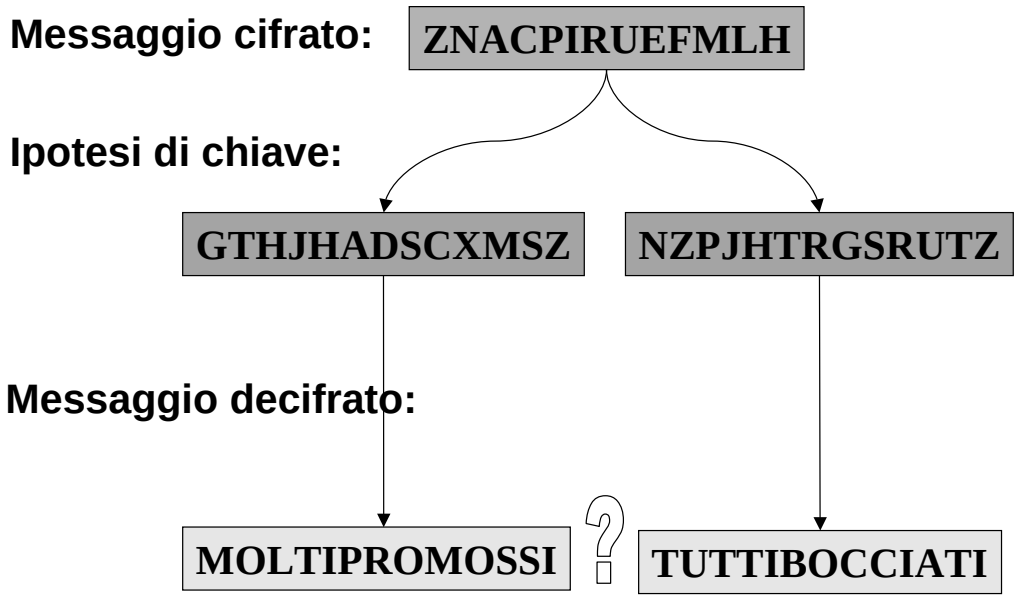
Mauborgne: *chiave scelta a caso e usata una sola volta*

Polialfabetica con *running key*

L'operazione di somma modulo due



One-time pad: inviolabile con attacco passivo



Per trasmettere un messaggio riservato su un canale insicuro bisogna concordare una chiave altrettanto lunga su un canale sicuro

Problemi di one-time pad

- Accordo riservato su molte chiavi molto lunghe
- Uguale probabilità di occorrenza dei simboli di chiave
- Ricezione di tutto il testo cifrato in ordine

Bletchley Park | Spie russe | Telefono rosso

Attacco attivo

- Impiego di meccanismi di autenticazione (H, S)

Il problema dell'intercettazione

Quanta informazione
sul testo in chiaro e sulla chiave di cifratura
può essere dedotta
da un testo in chiaro intercettato?

NOTA BENE
Il **messaggio** è una stringa di **simboli** che la sorgente sceglie all'interno di un certo **alfabeto**.
Dopo la **codifica binaria** dei simboli, il messaggio è una **stringa di bit**.
Una stringa di bit può essere interpretata come un **numero reale**.

Sicurezza di un Cifrario (C. Shannon)

SEGRETEZZA PERFETTA

Un Cifrario è detto **perfetto**, o **assolutamente sicuro**, se, dopo aver intercettato un certo testo cifrato **c**, l'incertezza *a posteriori* sul testo in chiaro **m** corrispondente è uguale all'incertezza che si aveva *a priori* , cioè prima dell'intercettazione.

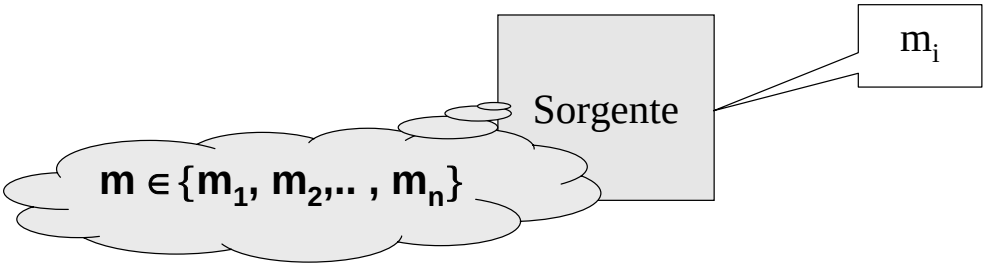
SICUREZZA

Un Cifrario è detto **sicuro** se dato un qualsiasi testo cifrato **c**, il trovare un **m** tale che $\mathbf{E_k(m) = c}$ è impossibile per chi non conosce $\mathbf{E_k^{-1}}$ e quindi **k**.

SICUREZZA COMPUTAZIONALE

Un Cifrario è detto **computazionalmente sicuro** se il calcolare **m** da un **c** è possibile, ma richiede una potenza di elaborazione superiore a quella a disposizione dell'attaccante.

Il messaggio come variabile aleatoria



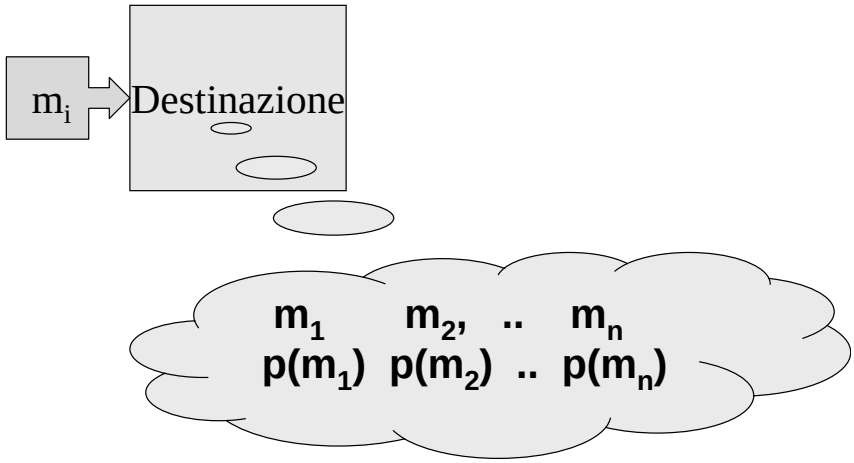
- **Ipotesi: la sorgente è priva di "memoria"**
- Ogni messaggio ha una sua prefissata probabilità di essere scelto:

Variabile aleatoria M

$\mathbf{P(M = m_i) = p(m_i)}$, con $\mathbf{0 \le p(m_i) \le 1}$

$$\sum_{i=1}^{n} p(m_i) = 1$$

L’informazione come variabile aleatoria



L’informazione **I(m)** fornita dall’arrivo di un messaggio **mi** è tanto più grande quanto più piccola è la sua probabilità.

Variabile aleatoria **I(m)**: il valore **$I(m_i)$** ha probabilità **$p(m_i)$**

Misura dell’informazione

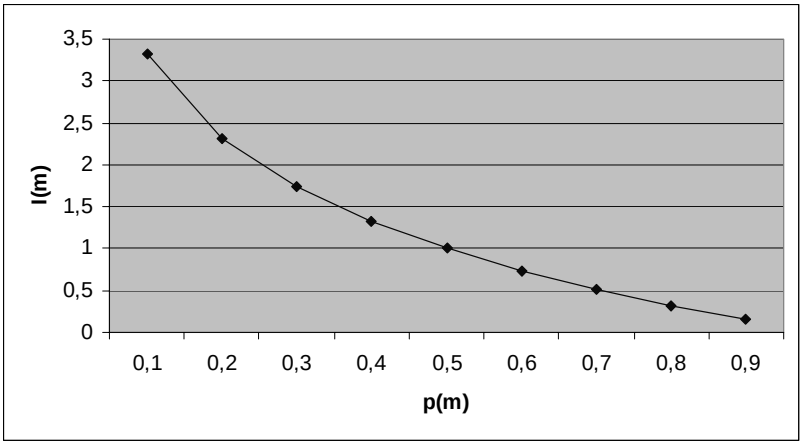
$$I(m) = \{I(m_1), I(m_2), \ldots, I(m_n)\}$$

Per misurare l’informazione è stata scelta la **funzione logaritmo**

DEFINIZIONE: l’informazione trasportata da **m** è indicata con **I(m)** ed espressa da:

$$I(m)=\log_2(1/p(m))=-\log_2 p(m)$$

L’unità di misura era il *bit* ed oggi è il *shannon (sh)*



Entropia come valor medio dell'informazione

DEFINIZIONE – L'**entropia** della variabile aleatoria **M** è il valor medio dell'informazione trasportata da un messaggio m

H(M) = *E*(I(m)) = $\sum_{i=1}^{n} p(m_i) \times I(m_i) = -\sum_{i=1}^{n} p(m_i) \times \log p(m_i)$

L'entropia è anche una misura dell'**incertezza** sull'informazione trasportata da un messaggio, prima di riceverlo.

L'entropia è inoltre il numero atteso di bit necessari per la codifica

Un messaggio ha probabilità 1 e tutti gli altri 0

$0 \leq H(M) \leq \log_2 n$

Tutti i messaggi hanno la stessa probabilità di occorrenza

Entropia condizionata

1. **Entropia** della variabile aleatoria **X**, **condizionata** dalla osservazione di un certo valore **y** della variabile aleatoria **Y**:

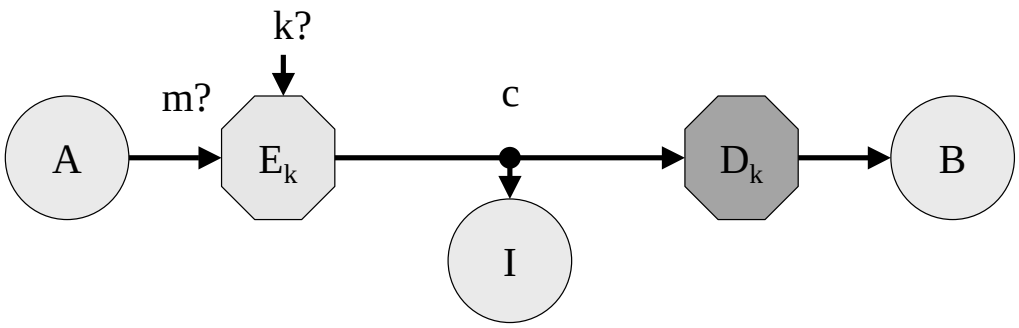
H(X|Y=y) = $-\sum_{x} p(x \mid y) \times \log p(x \mid y)$

in cui **p(x|y)** è la probabilità condizionata di **x**, dato **y**

2. **Entropia** della variabile aleatoria **X**, **condizionata** dalla osservazione di un qualsiasi valore della variabile aleatoria **Y**:

H(X|Y) = $-\sum_{y} p(y) \sum_{x} p(x \mid y) \times \log p(x \mid y)$

L'attacco con solo testo cifrato e con testo in chiaro noto



Incertezza su M condizionata dalla conoscenza di C : **H(M | C)**

Incertezza su K condizionata dalla conoscenza di C : **H(K | C)**

Incertezza su K condizionata dalla conoscenza di M,C: **H(K | M,C)**

In un Cifrario robusto le tre **equivocazioni** devono avere valori alti

Entropia congiunta: incertezza su una coppia

H(X,Y) = $-\sum_{y}\sum_{x} p(x,y)\times\log p(x\mid y)$

Dalla definizione segue:

H(X,Y) ≤ H(X) + H(Y)
= H(X) + H(Y) se X,Y indipendenti

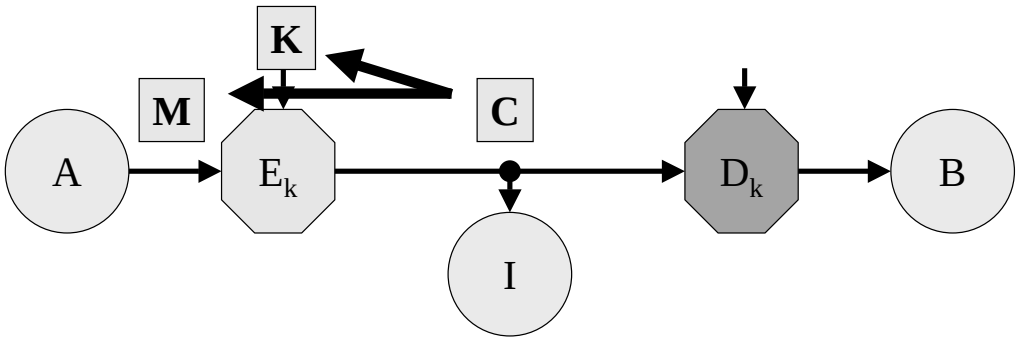
Proprietà notevole:

H(X,Y) = H(X) + H(Y|X)
= H(Y) + H(X|Y)

Data una terza variabile aleatoria Z si ha anche:

H(X,Y|Z) = H(X|Z) + H(Y|X,Z)
= H(Y|Z) + H(X|Y,Z)

Incertezza su K,M condizionata da C



H(K,M|C) = H(K|C) + H(M|K,C) = H(M|C) + H(K|M,C)

Incertezza sulla chiave

Incertezza sul testo in chiaro

Attacco con testo in chiaro noto

Il Cifrario perfetto

H(K|C) = H(M|C) + H(K|M,C)
H(K|C) ≥ H(M|C)

Noto **c,** l'individuazione di **k** è più difficile di quella di **m**

Se il Cifrario ha segretezza perfetta, deve essere:
p(m|c) = p(m) per ogni **m** e per ogni **c**.

incertezza a posteriori sul testo in chiaro

incertezza a priori sul testo in chiaro

Condizione necessaria per la segretezza perfetta:
H(M|C) = H(M)

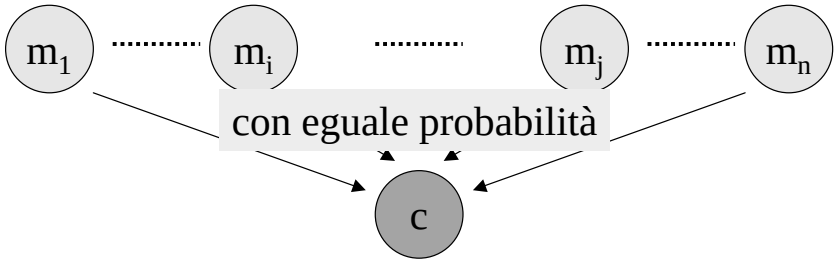
H(K) ≥ H(M)
k casuale e più lunga

Non esistono correlazioni
A nulla serve il disporre di un calcolatore con potenza illimitata
Si può solo "tirare ad indovinare" ma non si avranno conferme

Come si costruisce un Cifrario perfetto?
Per rispondere è sufficiente ricordare i teoremi di Bayes:

$$p(m) \times p(c|m) = p(c) \times p(m|c)$$

T1: "Condizione necessaria e sufficiente per la segretezza perfetta è **p(c|m) = p(c)** per ogni **m** e **c**"
N.B. p(c|m) deve dunque essere indipendente da **m**



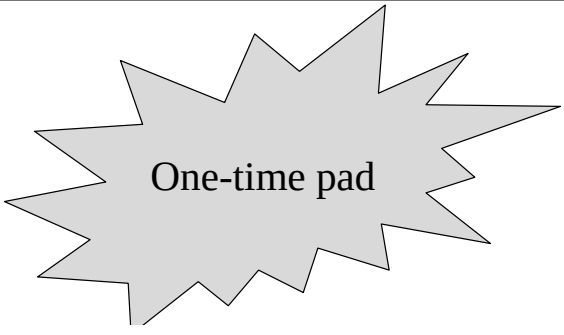
- la probabilità totale di tutte le chiavi che trasformano un $\mathbf{m_i}$ in un certo **c** è uguale a quella di tutte le chiavi che trasformano $\mathbf{m_j}$ nello stesso **c**, per ogni $\mathbf{m_i}$, $\mathbf{m_j}$ e **c**
- **il numero di chiavi deve almeno essere pari al numero di messaggi**

T2: "Tra le cardinalità degli spazi **M**, **K** di un Cifrario perfetto deve valere la diseguaglianza

$$|K| \geq |M|$$"

Il caso più semplice si ha quando il numero di chiavi è esattamente uguale al numero di messaggi.

T3: "un Cifrario con **|M| = |K| = |C|** è perfetto
se e solo se c'è esattamente una chiave
che trasforma ciascun messaggio **m** in ciascun crittogramma **c**
e se tutte le chiavi **k** sono **equiprobabili**"



Insieme di messaggi: $\{m_1, m_2, m_3, m_4, m_5\}$,
Insieme di crittogrammi: $\{c_1, c_2, c_3, c_4, c_5\}$
Insieme di chiavi **equiprobabili** $\{k_1, k_2, k_3, k_4, k_5\}$

Un Cifrario perfetto deve trasformare ogni **m** in ogni **c**

	c_1	c_2	c_3	c_4	c_5
m_1	k_1	k_2	k_3	k_4	k_5
m_2	k_5	k_1	k_2	k_3	k_4
m_3	k_4	k_5	k_1	k_2	k_3
m_4	k_3	k_4	k_5	k_1	k_2
m_5	k_2	k_3	k_4	k_5	k_1

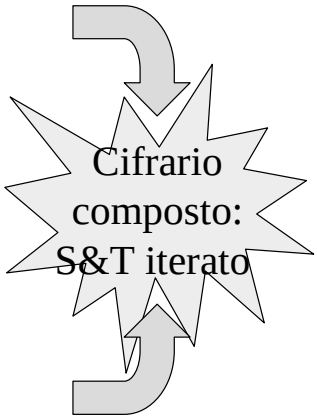
$$c_i = E(m_j, k_k)$$

"quadrato latino": $i = (j + k - 1) \bmod 5$

Si noti che essendo $p(k) = 1/5$ si ha anche $p(m|c) = p(c) = 1/5$.

Confusione & Diffusione (C. Shannon)

La **confusione** nasconde la relazione esistente tra testo in chiaro e testo cifrato e rende poco efficace lo studio del secondo basato su statistiche e ridondanze del primo.
La **sostituzione** è il mezzo più semplice ed efficace per creare confusione.



La **diffusione** nasconde la ridondanza del testo in chiaro spargendola all'interno del testo cifrato.
La **trasposizione** è il mezzo più semplice ed efficace per ottenere diffusione

Il cifrario composto