2.6. Joint principles of the above

$\mathcal{X}, \mathcal{Y}$: alphabets = finite set of characters

$\mathcal{X} = \{x_1, \dots, x_m\}$, $\mathcal{Y} = \{y_1, \dots, y_n\}$

$\mathcal{X}^\ell, \mathcal{Y}^\ell$: words of length $\ell \in \mathbb{N}_0$ over $\mathcal{X}, \mathcal{Y}$

$\mathcal{M} \subseteq \bigcup_{\ell=0}^{\infty} \mathcal{X}^\ell$: set of possible plaintexts, messages

$\mathcal{C} \subseteq \bigcup_{\ell=0}^{\infty} \mathcal{Y}^\ell$: set of possible ciphertexts

$M \in \mathcal{M}$ is called message or plaintext

$C \in \mathcal{C}$ is called ciphertext or cryptogram

$\mathcal{K}$: set of possible keys, the key space

$K \in \mathcal{K}$: is called key.

Encryption is described by a function

$$e : \mathcal{M} \times \mathcal{K} \to \mathcal{C} : (M, K) \mapsto C,$$

decryption by a function

$$d : \mathcal{C} \times \mathcal{K} \to \mathcal{M} : (C, K) \mapsto M$$

Def. 2.8. A cryptosystem is a five-tuple $(\mathcal{M}, \mathcal{K}, \mathcal{C}, e, d)$ with $\mathcal{M}, \mathcal{K}, \mathcal{C}$ as above and e, d functions with

$$e: \mathcal{M} \times \mathcal{K} \to \mathcal{C}, \quad d: \mathcal{C} \times \mathcal{K} \to \mathcal{M}$$

such that $d(e(M,K),K) = M$ for all $(M,K) \in \mathcal{M} \times \mathcal{K}$. ⊥

(2.6.1) Cryptanalysis

General assumption: O/E knows the cryptosystem being used. Kerckhoff's principle

Further information, side information:
language, context, statistical frequencies, etc.

Objective: determine the key

Different levels of attacks.

a) Ciphertext only (a string of ciphertext known)
b) Known plaintext (string of ciphertext and corr. plaint.)
c) Chosen plaintext (access to the enc. machinery)
d) Chosen ciphertext (access to the decr. machinery)

b) is minimal, c) and d) are hardest.
Classical systems would fail.

3. Cryptanalysis of Classical Systems

3.1 Frequency analysis

Monoalphabetic ciphers retain the frequencies of characters. In English, $\{E, T, A, O, I, N\}$ combine 51.75% of all frequencies.

Avoid this attack by:
enlarge the alphabet, e.g., DES with $\mathcal{X} = \{0,1\}^{64}$
non-natural languages, compression,

3.2. Friedman-Test

Objective: Decide whether a cipher is mono- or polyalphab.
Alphabet: $\mathcal{Y} = \{1, \dots, m\}$
Ciphertext: $C = (C_1, \dots, C_n)$ modeled by i.i.d. r.v. $C_1 \dots C_n$
with $P(C_i = \ell) = q_\ell$, $\ell = 1, \dots, m$

Def. 3.1.

$$I_C = I(C_1, \dots, C_n) = \frac{|\{(i,j) \mid C_i = C_j,\ 1 \le i < j \le n\}|}{\binom{n}{2}}$$

is called index of coincidence.

$$I_C = \frac{\text{no. of pairs with identical entries}}{\text{no of all pairs}}$$

Obviously: $I_C = 1 \iff C_1 = \dots = C_n$
$I_C = 0 \iff$ all C_i are different

Different representation of I_c

Let $N_\ell = |\{i \mid C_i = \ell\}|$, $\ell = 1, \dots, m$

Then

$$I_c = \frac{1}{n(n-1)} \sum_{\ell=1}^{m} N_\ell (N_\ell - 1)$$

By the strong law of large numbers

$$\frac{N_\ell}{n} \to q_\ell \quad (n \to \infty) \ (a.e.) \quad \forall \ell = 1, \dots, m$$

Hence

$$I_c = \sum_{\ell=1}^{m} \frac{N_\ell}{n} \frac{(N_\ell - 1)}{n-1} \to \sum_{\ell=1}^{m} q_\ell^2 = K_c \quad (n \to \infty) \ (a.e.)$$

Another representation of I_c:

Let $Y_{ij} = \begin{cases} 1, & C_i = C_j \\ 0, & C_i \neq C_j \end{cases}$, $1 \le i < j \le n$

Then $$I_c = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} Y_{ij}$$

Lemma 3.3. $E(I_c) = \sum_{\ell=1}^{m} q_\ell^2 = K_c$. ∎

Proof. $$E(Y_{ij}) = P(C_i = C_j) = \sum_{\ell=1}^{m} P(C_i = \ell, C_j = \ell)$$
$$= \sum_{\ell=1}^{m} q_\ell^2 = K_c$$

Hence, $$E(I_c) = \frac{1}{\binom{n}{2}} \sum_{i<j} K_c = K_c$$ ∎

Summary: I_c is an unbiased, strongly consistent estimator of K_c.

Cauchy-Schwarz inequality:

$$\Big(\underbrace{\sum_{\ell=1}^{m} q_\ell}_{1}\Big)^2 \le m \sum_{\ell=1}^{m} q_\ell^2 \iff \sum_{\ell=1}^{m} q_\ell^2 \ge \frac{1}{m}$$

with equality iff $q_\ell = \frac{1}{m} \quad \forall \ell = 1, \dots, m$.

If $q_\ell = \frac{1}{26}$ (uniform distribution), then

$$\kappa_U = \sum_{\ell=1}^{26} \frac{1}{26^2} = 0.0385$$

For German language: $\kappa_G = 0.0762$

$I_C \sim 0.0762 \rightarrow$ monoalphabetic cipher

$I_C \sim 0.0385 \rightarrow$ polyalphatic

Table of κ-values:

	English	French	Russian	Arabic
κ	0.066895	0.074604	0.056074	0.075889

3.3. Vigenère Cipher, estimate key length

Stochastic model:

$\mathcal{X} = \{0, \dots, m-1\}$ Alphabet

k keyword length, n message length, $k \ll n$

$M = (M_1, \dots, M_k, M_{k+1}, \dots, M_{2k}, M_{2k+1}, \dots, M_n)$

$K = (K_1, \dots, K_k, K_1, \dots, K_k, \dots \dots K_k)$

$\oplus$ $C = (C_1, \dots, C_k, C_{k+1}, \dots, C_{2k}, \dots \dots C_n)$

Plaintext, M_i i.i.d, $P(M_i = \ell) = p_\ell$ (known)

K_i i.i.d., $P(K_i = \ell) = \frac{1}{m}$

$$I_C = \frac{1}{\binom{n}{2}} \sum_{i<j} Y_{ij}, \quad Y_{ij} = \begin{cases} 1, & C_i = C_j \\ 0, & \text{otherwise} \end{cases}, \quad i<j$$

$$\kappa_M = \sum_{\ell=0}^{m-1} p_\ell^2$$

Lemma 3.5.

$$E(I_C) = \frac{1}{k(n-1)} \left[(n-k)\kappa_M + n(k-1)\frac{1}{m} \right] \quad (*)$$

(Outline of proof → see lecture notes)

We are interested in k

Resolve for k

$$k = \frac{n\left(\kappa_M - \frac{1}{n}\right)}{(n-1)E(I_c) + \kappa_M - \frac{n}{m}}$$

Estimate $E(I_c) = \kappa_c$ by I_c

$$I_c \to E(I_c) = \kappa_c \quad \text{a.e.} \quad (n \to \infty)$$

In German: $\kappa_M = 0.0762$, $m = 26$

Hence:

$$\hat{k} = \frac{0.0377\, n}{(n-1)\, I_c - 0.0385\, n + 0.0762}$$

is an estimator of k.

If k is known, write C as follows

$$\tilde{C} = \begin{pmatrix} c_1, \dots, c_k \\ c_{k+1}, \dots, c_{2k} \\ \vdots \qquad \vdots \\ c_{sk+1}, \dots, c_n \end{pmatrix}$$

The columns are monoalphabetic, apply a frequency analysis.