# CRYPTOGRAPHY AND STATISTICS: A DIDACTICAL PROJECT

**Massimo BORELLI, Anna FIORETTO,**

**Andrea SGARRO, Luciana ZUCCHERI**

DSM (Department of Mathematical Sciences)

University of Trieste, 34100 Trieste (Italy)

{borelli, sgarro, zuccheri}@units.it anna.fioretto@adriacom.it

## ABSTRACT

Cryptography is a stimulating way to introduce and consolidate ideas in statistics, computational linguistics, combinatorics and modular arithmetic. Two of the authors have been carrying out didactical experiences starting back in 1989 at a primary school level, without any special technology. A game is set up which involves cryptographers and cryptanalysts. Simple substitution ciphers are broken by building letter frequency hystograms by parallel work, so as to achieve what is being felt as statistical significance. Pupils quickly discover Markov models and the slight non-stationarity of the linguistic process. We have initiated a new round of experiments at a different level of age, 14-16, and technology. We take advantage of computer software to deepen our analysis of cipher systems and Markov models. The friendly (and cheap) technology of graphing calculators is used to analyse perfect and pseudo-perfect ciphers and to discuss the elusive notion of randomness.

*Keywords*: cryptography, statistics, mathematics education, maximum likelihood, Markov processes, randomness.

## 1. Why cryptography?

A blunt answer to the question posed in the title might be: because cryptography is a charming and rewarding way to introduce into the classroom subjects of traditional or less traditional mathematics, algebra, modular arithmetic, computational linguistics, combinatorics, algorithms, and more specific to our point, statistical estimations and statistical tests. It is not just a matter of gratifying one's audience for fun's sake: a solid reason to use cryptography is its being an effective response to the decrease of logical abilities which has been observed in students entering university. Many ways out have been suggested, based e.g. on algebra or Euclidean geometry (cf Mammana and Villani 1998): the approach taken is then the traditional approach of axiomatic deductive theories, which, unfortunately, is not always appropriate or exciting from the point of view of pre-college students, especially when they are too young, or when they are more technical-oriented. Instead, cryptography stimulates the problem-solving skills of the pupils and enhances their argumentative abilities, in a way which is directly linked to the "soft" logic of natural languages; cryptography is (perceived as) a *game*, but it is a motivating and sophisticated logical game! Actually, our experimentation has shown that cryptography may be introduced in classrooms at a very early stage, even at a primary school level (cf Section 3; cf also Zuccheri 1992, Leder et al. 2001); kids spontaneously formulate conjectures and develop arguments to prove or disprove them. An additional point in favour of cryptography (cf Section 4) is that it is ideally suited to make clear the advantage of an empirical approach to mathematics, pursued in a *math laboratory*, where one can use both hand calculators and full scale computers, according to the case; actually the computations involved can be quite lengthy with paper and pencil only, or even infeasible. Last, we think that, nowadays, cryptography by itself should be part of everybody's culture. In the age of the Internet and of the dramatic privacy and security problems it poses, one should understand the difference between trivial tactical aids like passwords, and professional strategic systems, as are DES (Data Encryption Standard) and RSA (so called from the names of its inventors, Rivest, Shamir and Adleman). Security is no longer a prerogative of the secret services.

Our team includes two persons active in cryptographic research and in mathematics education research (A.Sgarro and L. Zuccheri, respectively) and two teachers in charge of the class project of Section 4 (M. Borelli and A. Fioretto).

## 2. Ideas of cryptography, from the Bible to the web

This instant history of cryptography is used to introduce some of its basic notion; observe, however, that historical hints can be presented in the classroom to add some flavour to the technical material of Sections 3 and 4 (the standard reference to the history of cryptography is still Kahn 1967; cf also Sgarro 1989; as a reference to modern cryptography we suggest e.g. Schneier

1994).

Cryptography (i.e. *secret writing*, in old Greek) is nowadays felt as a part of computer science, and also as a part of our daily life, used as it is to protect the privacy of on-line transactions: and yet, it has always been with us. The simplest and possibly the oldest type of cipher, called a *simple substitution cipher*, appears already in the Bible. When such a cipher system is used, a permutation of the alphabet is chosen to be the *key* of the cipher; in practice one has two matched orderings of the alphabet, the normal ordering and a permuted ordering. The *clear text* is enciphered by substituting each letter as specified by the key. Breaking such a system is quite easy when the cipher text (the *cryptogram*) is long enough: in practice 25 to 30 letters will do. How to accomplish all this is masterly explained in *The golden bug*, a tale written by Edgar Allan Poe. In a natural language letters have a typical occurrence frequency (E appears 12% of an English clear text, say); this typical frequency is "inherited" by the corresponding cipher text letter, and so, after a few trials and some semantic aid, the cryptogram can be broken. The underlying method, called *maximum likelihood*, is typical of statistics, and was well known in the old Arab world: to this end the Aristotelian philosopher Al Kindi had prepared an accurate statistical description of the Arab language, obtained by sampling part of the Qur'an. Long forgotten in Europe, cryptography was re-born in Italy during Renaissance, but the lessons of the Arabs had been learnt, and it was well understood that a good cipher system should be able to "cheat" statistics. One of the ways out which were adopted was *polyalphabetic ciphers*. Initially supported by theoreticians rather than practitioners, these cipher systems took the lead in the 19th century, their implementation being now obtained by use of mechanical devices, so as to get rid of the synchronisation problems which had marred polyalphabetic ciphers in the age of paper-and-pencil cryptography. In a polyalphabetic cipher *several* permutations are selected (their number is called the *period of the cipher*), and they are used *in turns* according to a fixed scheme. As a rule, each permutation is very simply a *rotation* of the alphabet, and so it is completely specified by the substitute of clear letter A (if A is substituted by D, say, then B is substituted by E, C by F, and finally Z by C; observe that one is simply making sum modulo 3, as soon as one thinks of the letters as numbers: A=0, B=1, C=2, D=3, etc; sum is performed letter by letter, with no carry-over). This way, the very same clear text letter is enciphered by different permutations, and so has different substitutes in the cipher text. In the Renaissance, two concentric wheels were used to implement a polyalphabetic cipher; sometimes the cipher alphabet was a fancy one. Later, electro-mechanical machines, based on a cute system of rotors, allowed one to obtain extremely long periods, so safeguarding the cipher text from the cryptanalytic techniques which were developed at the end of the 19th century by Friedrich Kasiski, a German officer. Such machines were still in use during the second world war: an example is the notorious Enigma, adopted by the Germans and broken by the allied secret services. Substitution and polyalphabetic ciphers still exist nowadays in "extreme" forms. As for substitution ciphers, they are part of *composed* ciphers, as is DES, the *Data Encryption Standard*

widely used in commercial cryptography. In a composed cipher the clear text is enciphered many times in series, and in different ways: in DES one alternates substitutions and *transpositions*, i.e. anagrams. The "asymptotic" version of a polyalphabetic cipher is called a *one-time pad*: in it the key is a totally random and potentially infinite sequence which is summed to the clear text. Usually, the random sequence is binary and the sum is bit-by-bit sum modulo 2 (1+1=0, no carry-over), the clear text being itself binary, because it has been preliminarily encoded by means of ASCII, say. (Note that ASCII is *not* a cipher, but simply a transcription code, widely used by computer people to convert information to binary). Already the late Claude Shannon had shown that *the one-time pad is perfect,* i.e. provably unbreakable. In practice, the one-time pad needs too much key material, and so genuinely random sequences are replaced by more convenient pseudo-random sequences (cf Section 4). Unfortunately, a "pseudo-perfect" cipher is no longer unbreakable; actually, this is possibly the only example when the standard software used to generate random digits has proved to be sorely insufficient. Nowadays, besides commercial ciphers as DES, or sophisticated pseudo-random ciphers as used by the militaries, *public-key cryptography* has entered the lists. This is a revolutionary approach, which is based on the theory of algorithmic complexity; for example, the intolerable complexity of factoring integers is made good use of by a cipher system called RSA, which is widely used to safely transmit DES keys along the web.

## 3. Cryptography in primary schools: an exciting didactical experience

Our experimentation has been carried out during several years. It began back in 1989, in schools of North-East Italy, with approximately 300 pupils aged 7 to 10, and has continued up to the present date, due to the enthusiastic response of pupils and school teachers alike; cryptography is experienced as an exciting *game* (cf Zuccheri 1992 and Leder et al. 2001). Work in the classrooms consists of *cryptographic* and *cryptanalytic* activities (building and breaking ciphers, respectively), based on secret messages sent by encryptors (cryptographers) and intercepted by decryptors (cryptanalysts). We use substitution ciphers; cf Section 2; cryptanalysts end up "re-inventing" statistical inference, and in particular the principle of *maximum likelihood*; the basic underlying notion is *relative frequency*. The tools used are limited to zero-level technology, i.e. to paper and pencil. Initially, we use rotation ciphers only, and keep for simplicity the spacing between words; pupils readily find out cute “tricks”, semantic rather than statistical, to guess the key (*rotation ciphers* are the simplest form of substitution ciphers, since the only permutations allowed are those obtained by rotating the alphabet; cf Section 2). Encryptors soon perceive that they should make life harder to decryptors. One moves to general substitution ciphers. Permutations easily memorised are based on a secret motto: one writes down the motto by dropping repeated letters, and adds all the lacking letters in the reversed order (so SMALL IS BEAUTIFUL becomes SMALIBEUTFZWV…DC); however, pupils prefer to use special

alphabets for the encrypted text, e.g. the numbers from 1 to 26, and so the permutation has to be written down (an unwise policy, actually…). After a short training, we take out word-spacing; pupils work on encrypted texts of approximately 300 characters each. To break the cryptogram, they begin by counting the letter frequencies of a clear text of approximately 1000 words and build up hystograms by "parallel work", so as to achieve what is being felt as "statistical significance". Pupils compare their results with standard tables of frequencies. Now they are ready to successfully apply maximum likelihood, aided by their semantic competence. Pupils go as far as discovering some basics of Markov models (e.g., in Italian, letter Q is always followed by letter U, except in the unruly word SOQQUADRO, which incidentally means, disorder, "unruliness"); they quickly realise that the linguistic stochastic process is a slightly non-stationary, especially at the incipits. Polialphabetic or homophonic substitutions can be pointed out as clever tricks to "cheat" statistical cryptanalysis (in a *homophonic cipher* the cipher text alphabet is made up of many fancy letters, 50, say - fancy letters, incidentally, can be fun in themselves - and each clear letter is given *many* possible substitutes; this way the frequency of each clear-text letter is "spread" in the cryptogram among its possible substitutes, homophonic ciphers were in use up to the age of Napoleon). Pupils construct their own enciphering devices, rotating wheels and sliding rules (the latter are quite easy to make out of cardboard paper: one writes once the alphabet on the fixed strip, and twice on the sliding strip; in a way, one "linearises" the rotating wheels). On the way, the teacher has a chance to illustrate notions as one-to-one mappings, inverse mappings and modular arithmetic.

## 4. From paper and pencil to calculators and computers.

In school year 2001-2002 we have extended our experimentation to a different range of age, 14 to 16-year old students attending a technical school. Two classrooms have been involved; in the first, we are simply extending and deepening the material of Section 3.

- Substitution ciphers

This part of the project makes use of full-scale computers provided with standard software to re-take the ciphers of Section 3. Hystograms as in Section 3 can be now be built in a more sophisticated way; statistical significance and converging of relative frequencies to their "asymptotic values", i.e. to *probabilities*, can be made quite explicit by the support of graphics. Actually, one can construct typical frequency tables also for couples and triples, sampling large texts already available in the computer memory (our tables are arrays for single letters, matrices for couples, and dynamic lists for triples: actually, most triples are never encountered in a natural language text). These tables can be used to simulate "statistical" Italian (or English) of the 1st, 2nd and 3rd order. In the latter case one produces a sequence as ALLESTRORAMIA…; even such a short chunk contains genuine Italian words, as ESTRO (= gad-fly, and also: inventive whim), and ORA (= hour). These "texts" are meaningless, but one can soon discriminate between English and

Italian, say. Application of the principle of maximum likelihood by itself leads to phoney Italian (or English) of this type, the final touch pertaining to semantics. On the way, the teacher has the chance to introduce some combinatorics; e.g. the number of keys which are available in a simple substitution cipher (the number of ways one can permute the natural alphabet) is a nice way to introduce factorials and the factorial growth.

The experimentation in the second classroom is more taxing. This is done in co-operation with the Association for the Didactic with Technology, the Italian branch of $T^3$, Teachers Teaching with Technologies: the friendly technology of graphing calculators can help the teacher to set up a sophisticated *math laboratory* in the classroom, for a wide range of school levels up to university, in a cheap and handy way. Such math laboratories have been introduced at an undergraduate level (cf Invernizzi et al. 2000). In particular, the program covers Monte Carlo methods and simulations by means of random digits: this is directly linked to the present project, in which cryptography is used to teach and consolidate statistical notions as are randomness and testing; we take advantage of the powerful tools for manipulating data lists, which are available in graphing calculators.

- (Pseudo)-random digits

The idea is to simulate a binary one-time pad; cf Section 2. Cryptographic theory teaches us the following: if the binary key sequence is genuinely random - is obtained by tossing a fair coin - so is the cipher text sequence, and, what is more surprising, the resulting cipher cannot be broken: the latter statement is a rigorous theorem, not just wishful thinking! Unfortunately, generating long random sequences is extremely inconvenient, and so one is tempted to resort to convenient *pseudo*-random sequences, generated by the calculator (or by the computer), as normally done in similar cases. Since graphing calculators essentially perform operations on numbers, it is better to use a "numerical alphabet", rather than using the natural one: so, the clear-text must be preliminarily encoded, e.g. by ASCII, a standard code which, we stress it, has nothing to do with secrecy. To this end we have developed computer software which converts normal texts to a numerical form, and which can be used by the students to feed the encoded (but not yet enciphered) text to the graphing calculator, so as to form a clear text list. Random binary digits or, rather, *pseudo*-random binary digits, can be generated by the RANDOM function of the graphing calculator (suitably modified), so as to form a further list, which will contain the bits of the key sequence. The students encipher the message by summing the two lists; one uses bit-by-bit sum modulo 2 (without carry-over), i.e. *xor* logical sum. The output sequence (i.e. the cryptogram) is itself random-looking, like the key sequence; however, this "feeling" should be put to test.

- Statistical tests

One has to find a way for testing randomness. This can be done in a naive fashion by checking the occurrences of 0's and 1's in the list, or the occurrences of couples (00, 01, 10, 11), or the occurrences of triples (000, ... , 111). All this is easily accomplished on the calculator, by running a suitable cycle over the tested sequence. At a more sophisticated level, one can use the $\chi^2$ test

(goodness-of-fit), which is available on the calculators we are using. This way, one shows that the key sequence and the cipher-text sequence are indistinguishable from genuine coin-flipping sequences, at least from the point of view of statistical tests (these sort of statistical checks for randomness are generally considered to be enough in the general context of simulations, cf Knuth 1981). This concludes the technical work in the classroom; however, the teacher provides a "historical" addendum, to show that cryptography is special indeed: in cryptography one should never overlook the difference between a genuine coin-flipping sequence and a random-looking sequence generated by a cute *deterministic* algorithm like the one implemented in the calculator, even when this algorithm is considered to be quite good in the general context of simulation, since it has proved to be able to "cheat" standard randomness tests. Actually, cryptographers have proved that ciphers like ours, which rely on standard deterministic algorithms to generate the random-looking key, are quite insecure, at least from the very severe point of view of *strategic* cryptography. More specifically, they are extremely weak against attacks of a special type, when the cryptanalyst gets hold of some clear text matched with the corresponding cipher text (the clear text might be his own, e.g. because he was permitted to operate the enciphering machine for a short while); this is enough to reconstruct the key-generating function, and so to impersonate the legitimate user indefinitely; cf Schneier 1994, or Sgarro 1986. Good pseudo-random ciphers require generation programs which are extremely sophisticated, and are sometimes classified military material.

## REFERENCES


-Invernizzi S., Rinaldi M. and Sgarro A., 2000, *Moduli di Matematica e Statistica*, (Bologna: Zanichelli)

-Kahn D.,1967, *The Codebreakers*, (New York: Macmillan)

-Knuth D.E., 1981, Seminumerical Algorithms, vol. 2 of *The Art of Computing*, (Reading, Ma: Addison-Wesley)

-Leder D., Scheriani C. and Zuccheri L., 2001, The mathematics of the boys/girls: exchange of experience among boys/girls of the same age, in *Proceedings of CERME2*, Marianske Lazné, Czech Republic

-Mammana C. and Villani V. (eds.), 1998*, Perspectives on the teaching of geometry for the 21st century. An ICMI study,* (Dordrecht: Kluwer)

-Schneier B., 1994, *Applied Cryptography*, (New York: J. Wiley)

-Sgarro A., 1993, *Crittografia,* (Padova: Muzzio)

-Sgarro A., 1989, *Codici segreti,* (Milano: Mondadori); also: 1991, *Geheimschriften* (Augsburg: Weltbild)

-Zuccheri L., 1992, Crittografia e Statistica nella Scuola Elementare, in *L'Insegnamento della Matematica e delle Scienze Integrate*, vol. 15 n.1, pp 19-38