Statistical Methods for Cryptography

Alfredo Rizzi

Abstract In this note, after recalling certain results regarding prime numbers, we will present the following theorem of interest to cryptography: Let two discrete s.v.'s (statistical variable) X, Y assume the value: $0, 1, 2, \ldots, m-1$. Let X be uniformly distributed, that is, it assumes the value $i (i = 0, 1, \ldots, m-1)$ with probability $1/m$ and let the second s.v. Y assume the value i with probability $(p_i : \sum_{i=1}^{m-1} p_i = 1, p_i \geq 0)$. If the s.v. $Z = X + Y \pmod m$ is uniformly distributed and m is a prime number, at least one of the two s. v. X and Y is uniformly distributed.

1 Introduction

In today's world the need to protect vocal and written communication between individuals, institutions, entities and commercial agencies is ever present and growing. Digital communication has, in part, been integrated into our social life. For many, the day begins with the perusal of e-mail and the tedious task of eliminating spam and other messages we do not consider worthy of our attention. We turn to the internet to read newspaper articles, to see what's on at the cinema, to check flight arrivals, the telephone book, the state of our checking account and stock holdings, to send and receive money transfers, to shop on line, for students' research and for many other reasons. But the digital society must adequately protect communication from intruders, whether persons or institutions which attack our privacy. Cryptography (from $\kappa\rho\upsilon\pi\tau o\varsigma$, hidden), the study and creation of secret writing systems in numbers or codes, is essential to the development of digital communication which is absolutely private insofar as being impossible to be read by anyone to whom it is not addressed. Cryptography seeks to study and create systems for ciphering and to verify and authenticate the integrity of data. One must make the distinction between

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

F. Palumbo et al. (eds.), *Data Analysis and Classification*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-03739-9_2, © Springer-Verlag Berlin Heidelberg 2010

cryptoanalysis, the research of methods an "enemy" might use to read the messages of others and cryptography. Cryptography and cryptoanalysis are what make up cryptology.

Until the 1950s cryptography was essentially used only for military and diplomatic communication. The decryption of German messages by the English and of Japanese messages by the Americans played a very important role in the outcome of the Second World War. The great mathematician Alan Touring made an essential contribution to the war effort with his decryption of the famous Enigma machine which was considered absolutely secure by the Germans. It was the Poles, however, who had laid the basis for finding its weak link. Cryptography also played a vital role in the Pacific at the battle of Midway Regarding Italy, the naval battles of Punta Stilo and of Capo Matapan were strongly influenced by the interception and decryption of messages.

1.1 Different disciplines in cryptography

There are four disciplines which have important roles in cryptography:

1. Linguistics, in particular Statistical Linguistics
2. Statistics, in particular the Theory of the Tests for the Analysis of Randomness and of Primality and Data Mining
3. Mathematics, in particular Discrete Mathematics
4. The Theory of Information

The technique of Data Mining seems to be of more use in the analysis of a great number of data which are exchanged on a daily basis such as satellite data. Technical developments are largely inter-disciplinary. This suggests that new applications will be found which will, in turn, lead to new queries and problems for the scholars of Number Theory, Modular Arithmetic, Polynomial Algebra, Information Theory and Statistics to apply to cryptography.

Until the 1950s the decryption of messages was based exclusively on statistical methods and specific techniques of cryptography. In substance, the working instruments of cryptography, both for the planning of coding systems and for reading messages which the sender intended remain secret, were statistical methods applied to linguistics. The key to decoding systems using poly-alphabetic substitution and simple and double transposition has always been the analysis of the statistical distribution of graphemes (letters, figures, punctuation marks, etc.). Mathematics was not fundamental to the work of the cryptoanalyst.

Today, with the advent of data processing technology, coding of messages is done by coding machines. The structure of reference is the algebra of Galois (GF(q)). The search for prime numbers, in particular tests of primality, are of notable interest to modern cryptology.

In this note, after recalling certain results regarding prime numbers, we will present a theorem of interest to cryptography.

2 Prime Numbers

The questions regarding prime numbers have interested many scholars since the dawn of mathematics. We need only recall Euclid in ancient times and Fermat, Eulero, Legendre, Gauss and Hilbert in the last four hundred years. Gauss, in 1801, in *Disquisitiones Arithmeticae*, stated that the problem of distinguishing prime numbers from composite numbers and that of the factorization of these composite numbers were among the most important and useful in arithmetics. Moreover, he added, the very dignity of science itself seemed to require that such an elegant problem be explored from every angle which might help clarify it.

The great calculation resources which are today available to scholars all over the world have led many to deal with questions relative to primes and some to try and *falsify* certain conjectures. Numerous are the web sites devoted to these numbers. The most noteworthy fact of this situation is that information arrives on the web in real time, not only in print and these are among the most frequented sites. This leads many to confront questions regarding primes which are of limited importance. A form of emulation is stimulated in which we see many universities in the entire world, but particularly the United States, make great efforts to find a new prime and so become the "leader of the pack", if only for a short while as with setting a record in a sport. This happened, and is happening in the efforts to find the largest known prime to which some universities devote massive calculation resources for many years as occurred with the confirmation of the famous theorem of four colors in postal zones at the University of Illinois and the proof that the 23rd Mersenne number is prime.

When speaking of research in prime numbers reference is often made to possible applications in cryptography and in particular cryptographic systems with an RSA public key. The RSA system is based on the choice of two primes of sufficient size and on the relations introduced by Eulero in 1700. This is the source of interest in basic research in prime numbers which could, in some way, have operative results in various coding systems.

2.1 Tests of primality

The theoretical basis for the tests of primality, whether deterministic or probabilistic, has its origin in the research of the Swiss mathematician Leonardo Eulero (1707–1783) and the Frenchman Pierre de Fermat (1601–1665). Let $\mathbf{Z_n}$, be the set $[1, 2, \ldots, n]$. Let $\mathbf{Z}_n^*$ be the set of the integers prime with n. The cardinality of $\mathbf{Z}_n$ is indicated by $\phi(n)$. This is known as Eulero's function.

Theorem 1. *The number of primes with n is equal to:*

$$\phi(n) = n \prod \left(1 - \frac{1}{p_j}\right),$$

where p_j varies in all the primes which are divisors of n (including n if it is prime).

This demonstration can be seen in texts of the Theories of Numbers. If n is a prime number Eulero's function $\phi(n)$ is reduced to:

$$\phi(n) = n(1 - \frac{1}{n}) = n - 1.$$

If n is a composite number it is reduced to: $\phi(n) < n - 1$.

Theorem 2 (Eulero's). *For any* $n > 2$ *and* $a : (a, n) = 1$

$$a^{\phi(n)} \equiv 1(\mathrm{mod}n)\ \forall a \in \mathbf{Z_n}.$$

With Fermat's so-called Little Theorem one is able to consider a particular case as Eulero did whenever n is prime. In essence Fermat, had formulated a concept which was completely demonstrated to be a particular case of the preceding theorem. This was also demonstrated in various ways during the eighteenth century.

Theorem 3 (Fermat's). *If n is prime then:*

$$a^{n-1} \equiv 1(\mathrm{mod}n)\ \forall a \in \mathbf{Z_n^*}.$$

2.2 Deterministic tests

Those procedures which allow the determination of prime numbers through the application of a certain algorithm are called *deterministic tests*. The theory of complexity, an important branch of Computer Science, allows one to quantify the computational difficulty of a specific procedure. In general, complexity is measured by the processing resources necessary for the implementation of the algorithms in terms of memory capacity used, time taken for their execution, etc. For the problem of determining the primality of an integer it is enough to refer to the time taken for the execution of the algorithm. The simplest deterministic test of primality for a number n is based on the successive division of n by all primes inferior to the square root of n. Naturally this test is not applicable to very large integers. There are many valid deterministic tests of primality in numbers smaller than a particular n. For example (Pomerance et al. 1980):

1. If $n < 1.373.653$ and satisfies Fermat's relation (par. 2.1) for base 2 and 3, then n is prime.
2. If $n < 25.326.001$ and satisfies Fermat's relation (par. 2.1) for base 2, 3 and 5 then n is prime.
3. If $n < 2.152.3002.898.747$ and satisfies Fermat's relation (par. 2.1) for base 2, 3, 5, 7 and 11 then n is prime.
4. If $n < 341.550.071.728.321$ and satisfies Fermat's relation (par. 2.1) for base 2, 3, 5, 7, 11, and 13 then n is prime.

2.3 Some deterministic tests

The important results of M. Agrawal, N. Kayal and N. Saxena appear in "Annals of Mathematics", where they have proposed a deterministic test based on the following:

Theorem 4. *p is prime if and only if*

$$(x-a)^p \equiv (x^p - a)(\mathrm{mod} p),$$

where a is a number prime with p.

The simple demonstration is based on the fact that, if i comes between 0 and p the coefficients $\binom{n}{p}$ calculated on *modulo p*, in the binomial development of the first member of the preceding relation are null and, furthermore, $a^p \equiv a(\mathrm{mod} p)$.

Vice-versa, if p is not prime one of its factors does not divide $\binom{n}{p}$ mod *(p)* and therefore, the indicated relation is not valid. The algorithm supplied by the authors, carried out in only 13 lines, allows one to discover whether a number is prime or composite.

The result of greatest theoretic interest, demonstrated by the authors in the work cited, is the following:

Theorem 5. *The asymptotic complexity of the algorithm is* $O^{\sim}(\log^{21/2} n)$ *where the symbol* $O^{\sim}(f(n))$ *is for* $O^{\sim}(f(n)\ poly(\log f(n))$*, where f(n) is any function of n.*

In practice, however, the authors recall that in many cases this is faster than indicated. Therefore one deals with an algorithm P, or actually an algorithm in which the time of execution is the function of a polynomial which depends on n. The other algorithms for the analysis of primality noted in literature are *NP*, or rather, their execution in a *polynomial* time depends on non-deterministic procedures. In 1986 Goldwasser and Kilian proposed a randomized algorithm, based on an elliptical curve which works in very wide hypotheses, in polynomial time for *quasi* all inputs. The algorithm certifies primality. The probabilistic tests of primality verify the null hypothesis $\mathbf{H_0}$: n is a prime number. If the hypothesis is not verified the number is surely composite. This is a statistical test in which the probability of errors of the second type, or rather of accepting a false hypothesis, is a number other than zero. Very little attention has been paid by scientific literature to these, very particular statistical tests.

The most noted statistical test of primality is that of Miller and Rabin, proposed in 1976. We define as *witness* a number which meets the requirements of Fermat's so-called Little Theorem to be a composite number. The test in question is based on the following:

Theorem 6. *If n is an odd composite number then the number of witnesses of which it is composed will be at least:* $(n-1)/2$.

Theorem 7. *Considering an odd integer and an integer s, the probability that a composite number is considered to be prime is less than* 2^s.

Empirical evidence shows that the probability that a composite number is found to be prime is actually, in the long term, less than that indicated. There have been shown to exist only 21, 253 composite numbers in base 2 which are inferior to 25 billion and which satisfy Fermat's Little Theorem. These are called pseudo-primes. There is, therefore a probability of about 8×10^{-6} that a composite number n will satisfy the relation $2^{n-1} \equiv 1(\mathrm{mod} n)$.

The problems of factorizing a number and of determining if a number is prime are by their nature diverse. In many processing procedures, however, these are treated together. In every case it is easier to determine whether a number is prime than to find all of its factors. Today, even with the super computers available and the improved algorithms which are known, it is not possible to factorize numbers having more than a few hundred digits.

3 The Sum Modulo *m* of Statistical Variables

The deterministic and non-deterministic methods co-exist, at times in the same procedure. Algorithms are being found which are always more efficient and easier to use. But there is no doubt that probabilistic tests of primality are the only ones applicable when n is *particularly* high and one hasn't the fortune to find oneself in a very particular situation. For instance, if the number is composite and divisible by one of the initial primes. Deterministic tests of primality can be applied, relatively quickly, to numbers having a very large amount of digits. There is, however, a limit on the number of digits as we learn from the *theory of complexity*. Probabilistic tests of primality furnish results which are completely acceptable in situations which are very general. They require negligible time to process and are those applicable in research situations Rizzi (1990), Scozzafava (1991).

Theorem 8. *Let two discrete s.v. (statistical variable) X, Y assume the value:* $0, 1, 2, \ldots, m-1$. *Let X be uniformly distributed, that is, it assumes the value* $i\,(i = 0, 1, \ldots, m-1)$ *with probability* $1/m$, *and let the second s.v. Y assume the value i with probability* $(p_i, \sum_{i=1}^{m-1} p_i = 1, p_i \geq 0)$. *Then, if the two s.v. are independent, it follows that the s.v. Z obtained as a sum modulo m:* $Z = X + Y(\mathrm{mod}\ m)$ *is uniformly distributed.*

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

$$i, i+1, i+2, \ldots, m-1, 0, 1, 2, \ldots, i-1$$

respectively with probabilities:

$$p_0, p_1, p_2, \ldots, p_{m-1} - i, \ldots, p_{m-1}$$

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

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

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 will be uniformly distributed.

Theorem 9. *Let two independent s.v. X and Y assume the values:* $0, 1, 2, \breve{E} m-1$ *respectively with probabilities*

$$p_0, p_1, \ldots, p_{m-1} \left(p_i \geq 0, \sum_{i=1}^{m-1} p_i = 1 \right)$$
$$q_0, q_1, \ldots, q_{m-1} \left(q_i \geq 0, \sum_{i=0}^{m-1} q_i = 1 \right)$$

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

Proof. The table of the sum *(mod m)* of the s.v. X and Y is as follows: As the X and Y are independent, the s. v. Z assumes the value 0 with probability

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

	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	$j+1$			$i+j$		$j+m-2 \pmod m$	$j+m-1 \pmod m$	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

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 $1, 2, \ldots, m-1$, 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 the p_i are known (the reasoning is the same if the q_i are known) the above system is a system of m equations in the m unknowns q_i. It follows:

$$q_i = \frac{1}{\Delta} \begin{vmatrix} p_0 p_{m-1} & \ldots 1/m \ldots & p_1 \\ p_1 p_0 & \ldots 1/m \ldots & p_2 \\ \vdots & & \\ p_{m-1} p_{m-2} & \ldots 1/m \ldots & p_0 \end{vmatrix},$$

where Δ is the determinant of the p_i coefficients, and it can be easily seen to be $\neq 0$ if at least one $p_i \neq 1/m (i = 0, 1, \ldots, m-1)$. (In the opposite case the s.v. X is uniformly distributed and the theorem is proved). 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 $\sum_{i=o}^{m-1} q_i \geq 0$ we have $q_i = 1/m$. Then, it is enough to show that

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

The two determinants are equal because, in order to transform the second into the first one, it is necessary to perform, due to the circulating nature of the permutation of the p_i, $m-2$ rows' inversions and $m-2$ inversions over the columns, that is $2(m-2)$ 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 instance, if $m - 1 = 3$ we have that the number of q_0 is

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

References

Agrawal, M., Kayal, N., & Saxena, N. (2004). Primes in p. *Annals of Mathematics, 160*, 781–793.

Goldwasser, S., & Kilian, J. (1986) Almost all primes can be quickly certified. In *Proceedings of the eighteenth annual ACM symposium on Theory of Computing* (pp. 316–329). New York: ACM Press.

Pomerance, C., Selfridge, J. L., & Wagstaff, Jr., S. S. (1980). The pseudoprimes to 25×10^9. *Mathematics of Computation, 35*, 1003–1026.

Rizzi, A. (1990). Some theorem on the sum modulo m of two independent random variables. *Metron, 48*, 149–160.

Scozzafava, P. (1991). Sum and difference modulo m between two independent random variables. *Metron, 49*, 495–511.

http://www.springer.com/978-3-642-03738-2

Data Analysis and Classification
Proceedings of the 6th Conference of the Classification and Data Analysis Group of the Società Italiana di Statistica
Palumbo, F.; Lauro, C.N.; Greenacre, M. (Eds.)
2010, XXII, 482 p. 109 illus., Softcover
ISBN: 978-3-642-03738-2