Random Numbers

AK Bhateja

Random Numbers

- Why random numbers?
- Random number generators
- Golomb's Randomness Postulates for Binary Sequences
- Statistical Tests for Randomness

One-time Pad Cipher System

- The simplest and most secure of all cryptosystems is the one-time pad.
- Plain text and key are added modulo 10/26/2.
- The key is truly random.
- Suppose the message is given as a string of bits i.e. elements of F_2. A long random string of bits is formed; i.e. the key which is known to sender and receiver.
- The key string must be at least as long as the message string and is used only once.
- This is a perfect, unbreakable cipher.
- The major disadvantage of this cryptosystem is that it requires as much key as there is data to be sent.

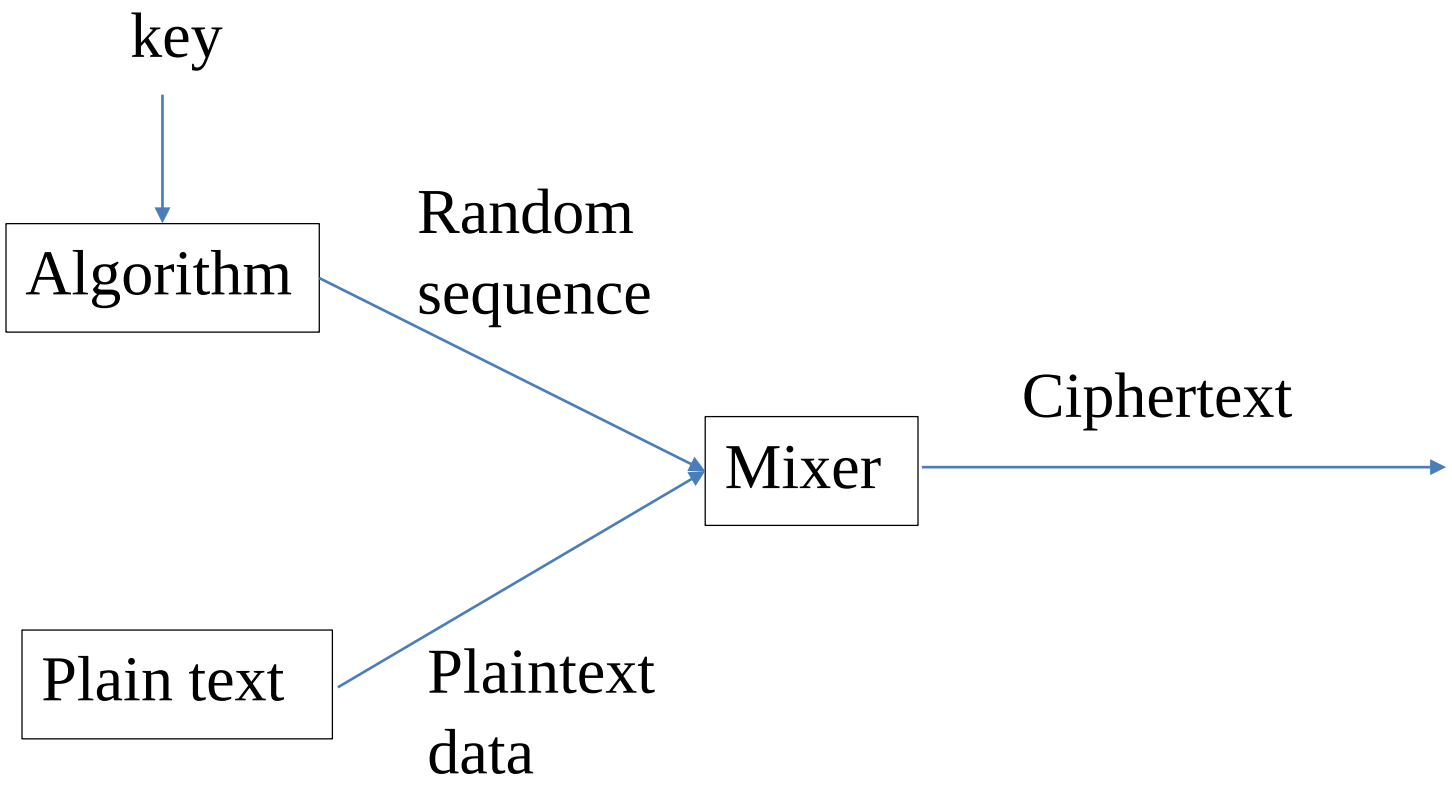
Stream cipher

- A stream cipher is an encryption algorithm that encrypts bits/bytes of plaintext with pseudorandom sequences, usually with XOR. e.g. One-time pad, Shift registers, RC4.

Plaintext:	1 0 0 1 0 1 1 1 0 0
Key stream	0 0 1 1 1 0 0 1 0 1
Ciphertext	1 0 1 0 1 1 1 0 0 1

- Some general properties of a stream cipher:
  - The number of possible keys must be large enough so that an exhaustive search for the key is not feasible.
  - The key stream have a guaranteed minimum length for their periods which exceeds the length of the message strings.
  - The ciphertext must appear to be random.

A Stream Cipher



Random numbers

- Random Numbers
  - True random numbers
  - Pseudorandom numbers
- True random numbers
  - flipping a coin
  - rolling a dice
- Computer generated true random numbers
  - the radioactive decay of an atom

    According to quantum theory, there's no way to know for sure when radioactive decay will occur, so this is essentially "pure randomness" from the universe.
  - atmospheric noise
  - use the exact time you press keys on your keyboard

Pseudorandom number/bit Generator

- A pseudorandom generator (PRG) is a deterministic algorithm which, gives a sequence appears to be random. The input to the PRG is called the seed, while the output of the PRG is called a pseudorandom bit sequence.
- A random bit generator can be used to generate (uniformly distributed) random numbers.
- Properties of a Good Generator
  - It should be efficiently computable.
  - The period should be large.
  - The successive values should be independent and uniformly distributed

Types of Random-number Generators

- Linear congruential generators
- Tausworthe generators
- Combined generators

Linear-Congruential Generators

- Discovered by D. H. Lehmer in 1951
- A linear congruential generator produces a pseudorandom sequence of numbers $x_1, x_2, x_3, \ldots$ according to the linear recurrence

  $$x_{i+1} = a\,x_i + c \bmod m, \quad i = 0, 1, 2, \ldots,$$

  a (multiplier), c (increment), and m (modulus) are parameters which characterize the generator, while $x_0 < m$ is the (secret) seed.
- The length of a cycle is called the period of a generator.
  It denoted by LCG(x_0, a, c, m).
- Selection of LCG Parameters
  - a, c, and m affect the period
  - The modulus m should be large.
  - The period can never be more than m.
  - For mod m computation to be efficient, m should be a power of 2.

Linear-Congruential Generators

- When $c = 0$, the form is known as the Multiplicative Linear Congruential Generator.
- When $c \neq 0$, the form is called the Mixed Linear Congruential Generator.
- $x_i \equiv a\, x_{i\text{-}1} \pmod{m}$ has full period (i.e. $m - 1$) iff
  1. m is a prime number
  2. a is a primitive root modulo m.
- For example, we choose $m = 7$ and $a = 3$, the above conditions satisfy. Period is 6.

Mixed Linear Congruential Generator

When $c \neq 0$, correctly chosen parameters allow a period equal to m (maximum period), for all seed values. This will occur if and only if (Hull–Dobell Theorem)

- c is relatively prime to m;
- $a \equiv 1 \pmod{p}$ if p is a prime factor of m;
- $a \equiv 1 \pmod{4}$ if 4 is a factor of m.

- All of these conditions are met if $m = 2^k$, $a = 4b + 1$, and c is odd. Here, b, c, and k are positive integers.
- To obtain random numbers on [0, 1], we let $u_i = x_i / m$
- An unfavorable choice of the parameters m, a, and c, may result in a very short length of the period.

  e.g. $m = 10$, $a = c = x_0 = 5$, sequence generated 5, 0, 5, 0, …

  gives period = 2.

Fact: If $x_{i+1} = a\,x_i + c \bmod m, \quad i = 0, 1, 2, \ldots$, then for all $k \in \{1, 2, \ldots, n\}$

$$x_k = \left(a^k x_0 + c\,\frac{a^k - 1}{a - 1}\right) \bmod m$$

Proof: $x_{i+1} = a\,x_i + c \bmod m, \quad i = 0, 1, 2, \ldots$

$x_1 = (a\,x_0 + c) \bmod m$

$x_2 = (a\,x_1 + c) \bmod m = a^2\,x_0 + c\,(a + 1) \bmod m$

……..

$x_k = a^k\,x_0 + c\,(a^k + \ldots + a + 1) \bmod m$

or $x_k = \left(a^k x_0 + c\,\frac{a^k - 1}{a - 1}\right) \bmod m$

Theorem: Let $m = p_1^{e_1} p_2^{e_2} \ldots p_k^{e_k}$ be the decomposition of a + ve integer m into distinct prime factors. Let α be the period LCG(x_0, a, c, m), α_i period of LCG$\left(x_0, a, c, p_i^{e_i}\right)$ for i = 1, … k, then α = lcm(α_1, α_2, …, α_k)

Proof: Let's prove it for $m = m_1 \cdot m_2$, where $m_1 = p_1^{e_1}$ and $m_2 = p_2^{e_2}$

Let $y_i = x_i \bmod m_1$ and $z_i = x_i \bmod m_2$ be the two sequences, based on LCG(x_0, a, c, m), for $m = m_1$ & m_2 respectively.

Let α, α_1 and α_2, be the periods of $\{x_i\}$, $\{y_i\}$ and $\{z_i\}$ respectively.

$\therefore$ $x_\alpha = x_0 \bmod m$ and $x_n \neq x_0 \bmod m$ for all $0 < n < \alpha$.

Since $x_\alpha = x_0 \bmod m \Rightarrow m$ divides $(x_\alpha - x_0)$

Since m_1 divides m, then m_1 divides $(x_\alpha - x_0)$,

$\therefore$ $x_\alpha = x_0 \bmod m_1$ hence $y_\alpha = y_0 \bmod m_1$

Since α_1 is period of $\{y_i\}$, therefore α is divisible by α_1.

Similarly α is divisible by α_2.

Hence α = lcm(α_1, α_2). It can be extended i.e. α = lcm(α_1, α_2, …, α_k)

Let us assume that LCG(x_0, a, c, p^e) generates the full sequence and m. We will prove all the three conditions of LCG. There is no loss of generality in considering $x_0 = 0$

$$x_j = c + ac + a^2c + \cdots + a^{j-1}c = \frac{a^j - 1}{a - 1} c \bmod p^e$$

One of these must be equal to 1. $\frac{a^j-1}{a-1} \cdot c \equiv 1 \bmod p^e$

$\Rightarrow$ neither $\frac{a^j-1}{a-1}$ nor c is divisible by p. Hence c is coprime to p^e.

Also, since $x_m = 0 \bmod m$ and let $m = p^e$ then $\frac{a^{p^e}-1}{a-1} \cdot c \bmod p^e = 0$

$\Longrightarrow a^{(p^e)} \equiv 1 \bmod p^e \Longrightarrow a^{(p^e)} \equiv 1 \bmod p$

because when a number is divisible by p^e it must be divisible by p.

Since $a^p \equiv a \bmod p$ (by Fermat's theorem)

$$\therefore \quad a^{p.p.p \ldots p} = \left(\left(\left((a^p)^p\right)^p\right) \ldots\right)^p = a \bmod p \Rightarrow a^{(p^e)} \equiv a \bmod p$$

Therefore $a \equiv 1 \bmod p$

Now consider when $p = 2$ and $e > 1$, we have to prove $a \equiv 1 \bmod 4$

Let $a \equiv 3 \bmod 4$

$x_{i+1} = a\,x_i + c \bmod m, \quad i = 0, 1, 2, \ldots$

$x_2 = (a + 1)c$

$x_4 = a^2\,x_2 + (a + 1)c$

…

$x_{2j} = a^2\,x_{2j-2} + (a + 1)c.$

i.e. each x_{2j} is divisible by $a + 1 = 4$.

This it cannot have a cycle longer than $m/4$.

Hence $\{x_j\}$ sequence cannot have period longer than $m/2$.

Hence $a \equiv 1 \bmod 4$

Now, we prove sufficiency of these conditions.

i.e. let us assume

- c is relatively prime to m;
- $a \equiv 1 \pmod{p}$ if p is a prime factor of m;
- $a \equiv 1 \pmod{4}$ if 4 is a factor of m.

we have to prove that the period of LCG(x_0, a, c, m) is maximum.

We use induction (assume sufficiency for $m = p^{k-1}$, and extend it to $m = p^k$. We also take $x_0 = 0$, without a loss of generality

Case 1: When $k = 1$, a must equal to 1, and then $x_i = i \cdot c$,

When $c \neq 0$, generates the full sequence

because c, $2c$, $3c$, …, $(p-1)c$ must consists of $p-1$ distinct numbers (since $sc = rc \bmod p$ implies $s = r$).

Case 2: When $k > 1$.

Since $a \equiv 1 \bmod p^e \Rightarrow x = 1 + qp^e$ for some $q \in Z$ and not a multiple of p.

By the binomial theorem

$$a^p = (1 + qp^e)^p = 1 + \binom{p}{1} qp^e + \binom{p}{2} q^2 p^{2e} + \cdots + \binom{p}{p} q^p p^{pe}$$

$$= 1 + qp^{e+1}\left(1 + \frac{1}{p}\binom{p}{2} qp^e + \frac{1}{p}\binom{p}{3} q^2 p^{2e} + \cdots + \frac{1}{p}\binom{p}{p} q^{p-1} p^{(p-1)e}\right)$$

$$or \quad a^p = 1 + s \cdot q \cdot p^{e+1}$$

Where $s \equiv 1 \bmod p$. Now

$$x_p = \frac{a^p - 1}{a - 1} c = \frac{sqp^{e+1}}{qp^e}\ c = spc \bmod p^k$$

which implies that not only x_p, but consequently all

$$x_{ip} = a^p x_{(i-1)p} + \frac{a^p - 1}{a - 1} c \ \bmod p^k$$

are divisible by p.

Define $y_i = \frac{x_{ip}}{p}$, the corresponding sequence is generated by

$$y_i = a^p y_{(i-1)} + \frac{a^p - 1}{a - 1} \cdot \frac{c}{p}$$

$$= a^p y_{(i-1)} + s \cdot c \ \ mod\ p^{k-1}$$

where neither *s* nor *c* is divisible by *p*.

By the induction assumption, the $\{y_i\}$ sequence is of the full period p^{k-1}

$$\therefore x_{p^k} = p \cdot y_{p^{k-1}} = 0 \bmod p^k \quad \& \quad x_{p^{k-1}} = p \cdot y_{p^{k-2}}$$
$$\neq 0 \bmod p^k$$

$$x_{p^{k-i}} = p \cdot y_{p^{k-(i+1)}} \neq 0 \bmod p^k \qquad \forall\, i = 1, 2, \ldots$$

Hence the period of the $\{x_j\}$ sequence must be equal to p^k.

Tausworthe Generators

- Need long random numbers for cryptographic applications
- Generate random sequence of binary digits (0 or 1)
- Divide the sequence into strings of desired length
- Proposed by Tausworthe (1965)

  $x_{t+k} = (a_0 x_t + a_1 x_{t+1} + a_2 x_{t+2} + \ldots + a_{k-1} x_{t+k-1}) \bmod 2$

  where $a_i , x_i \in \{0, 1\}$
- This generator can have a maximum period of $2^k - 1$.

Tausworthe Generators

- Let $f(x) = x^k + a_1 x^{k-1} + a_2 x^{k-2} + \ldots + a_k$ be a polynomials in $Z_p[x]$.
- This polynomial generate a sequence of numbers by

$$x_m = (a_1 x_{m-1} + a_2 x_{m-2} + \ldots + a_k x_{m-k}) \bmod p \,, \quad m \geq k$$

for any initial vector $(x_0, x_1, x_2, \ldots, x_{k-1}) \in Z_p^k - \{0\}$

- To achieve the maximum period $p^k - 1$, a necessary and sufficient condition is that the polynomial $f(x) \in Z_p[x]$ is primitive polynomial.
- A polynomial of degree k in $Z_p[x]$ is called primitive polynomial if it has a root that is a primitive element of the finite field F_{p^k}.

Combined Generators

- Longer period generator is needed because of the increasing complexity of the systems.
- Let $y_{i,1}, y_{i,2}, \ldots, y_{i,k}$ be the i^{th} outputs from k different multiplicative LCG
- The combined linear congruential generator is

$$x_i = \left(\sum_{j=1}^{k} (-1)^{j-1}\, y_{i,j}\right) \bmod (m_1 - 1)$$

where $y_{i,j}$ is the i^{th} input from the j^{th} LCG and x_i is the i^{th} random generated value.

- The maximum possible period: $\frac{(m_1-1)(m_2-1)\ldots(m_k-1)}{2^{k-1}}$. This bound is attained only when all values of $(m_j - 1)/2$ are relatively prime.

Ex: Combining $k = 2$ generators with $m_1 = 2147483563$, $m_2 = 2147483399$.

$$x_i = (y_{i,1} - y_{i,2}) \bmod (m_1 - 1) = (y_{i,1} - y_{i,2}) \bmod 2147483562$$

Combined generator has period: $(m_1 - 1)(m_2 - 1)/2 \approx 2 \times 10^{18}$

Concept of Randomness

- No periodic sequence is truly random.
- In cryptography we require unpredictability rather than randomness.
- Cryptographer wants that if a cryptanalyst intercepts part of the sequence, should not be able to predict what comes next.
- Security may be ensured if a segment of ciphertext which is considerably shorter than the period is intercepted.
- Any deterministic sequence satisfying these properties is normally called a pseudo-random sequence.

- Consider a binary sequence (s_t) of period p.
- Definition. Run means a string of consecutive identical sequence elements which is neither preceded nor succeeded by the same symbol.
- A run of 0s is called a gap and a run of 1s is a block.

  e.g. 0111001 it begins with a run of one 0, contains a run of three 1s and a run two 0s and then ends with a run of one 1.
- Autocorrelation function:

$$C(\tau) = \frac{A - D}{p} \qquad 0 \leq \tau < p$$

Where A is the number of positions in which the sequences (s_t) & ($s_{t+\tau}$) agree and D is the number of positions in which they disagree.

When $\tau = 0$, it is in-phase autocorrelation. $C(0) = 1$.

When $\tau \neq 0$ it is out-of-phase autocorrelation.

Golomb's Three Randomness Postulates for Binary Sequences

1. In every period, the number of zeroes is nearly equal to the number of ones
2. In every period, half the runs have length 1, one-fourth have length 2, one-eighth have length 3 etc. Moreover, for each of these lengths, there are equally many gaps and blocks.
3. The out-of-phase autocorrelation is a constant.

A sequence satisfying these postulates is called G-random or pn sequence.

Example: Let (s_t) = 0111001, p = 7

(s_t) = 0111001

(s_{t+1}) = 1110010

$$C(1) = \frac{3-4}{7} = -1/7$$

(s_{t+2}) = 1100101

$$C(2) = \frac{3-4}{7} = -1/7\ ,\ \ldots$$

i.e. out-of-phase autocorrelation is constant

Statistical Tests for Local Randomness

- Let $s = s_0, s_1, \ldots, s_{n-1}$ be a binary sequence
- Statistical tests to determine whether the binary sequence possesses specific characteristics that a truly random sequence is likely to have.
- Five Basic Tests
  - Frequency test
  - Serial test
  - Poker test
  - Autocorrelation test
  - Run test

The Frequency Test

- To ensure same number of 0s and 1s.

$$\chi^2 = \frac{n_0 - n_1}{n}$$

- From the table of χ^2 distribution, value of χ^2 for one degree of freedom & for 5% significance level is 3.84.
- Reject any sequence for which the calculated value is greater than 3.84.

The Serial Test

- to ensure that the transition probabilities are reasonable i.e. the probability of consecutive entries being equal or different is about the same.
- It gives some level of confidence that each bit is independent of its predecessor.
- Let 01 occurs n_{01} times, 10 occurs n_{10} times, 00 occurs n_{00} times and 11 occurs n_{11} times.
- $n_{01} + n_{10} + n_{00} + n_{11} = (n - 1)$ since in a sequence of length n there are only $n - 1$ transitions and the subsequences are allowed to overlap.
- The purpose of this test is to determine whether the number of occurrences of 00, 01, 10, and 11 as subsequences of s are approximately the same. Ideally $n_{01} = n_{10} = n_{00} = n_{11} = \frac{n-1}{4}$.

The Serial Test

- Good has shown that (Reference: Good, I.J., “On the serial test for random sequence”, The Annals of Mathematical Statistic, 1957)

$$\frac{4}{n-1}\sum_{i=0}^{1}\sum_{j=0}^{1}(n_{ij})^2 - \frac{2}{n}\sum_{i=0}^{1}(n_i)^2 + 1$$

is approximately distributed as χ^2 with two degrees of freedom, which is 5.99.

- Reject any sequence for which the calculated value is greater than 5.99.

The Poker Test

- For any integer *m* there are 2^m different possibilities for a section of length *m* of a binary sequence.
- Treats numbers grouped together as a poker hand. Then the hands obtained are compared to what is expected using the chi-square test.
- Partition the sequence into blocks of size *m*.
- Count the frequencies of each type of section of length *m* in the sequence.

$$\text{If the frequencies are } f_0, f_1, \ldots, f_{2^m-1}, \text{then } \sum_{i=0}^{2^m-1} f_i = F = \left\lfloor \frac{n}{m} \right\rfloor$$

- Evaluate

$$\chi^2 = \frac{2^m}{F}\left(\sum_{i=0}^{2^m-1} (f_i)^2\right) - F$$

- Compare this value with the table value for χ^2 having $2^m - 1$ degrees of freedom.
- This test can be applied for different values of *m*.
- Setting $m = 1$ in the poker test yields the frequency test

The Autocorrelation Test

- The purpose of this test is to check for correlations between the sequence *s* and (non-cyclic) shifted versions of it.
- Let *d* be a fixed integer, $1 \leq d \leq \lfloor n/2 \rfloor$. The number of bits in *s* not equal to their *d*-shifts is

$$A(d) = \sum_{i=0}^{n-d-1} s_i \oplus s_{i+d}$$

- Evaluate

$$2\left(A(d) - \frac{n-d}{2}\right)/\sqrt{n-d}$$

which approximately follows an $N(0, 1)$ distribution if $n - d \geq 10$.

The Run Test

- The purpose of the runs test is to determine whether the number of runs (of either zeros or ones); of various lengths in the sequence *s* is as expected for a random sequence.
- Let r_{0i} be the number of gaps of length *i* and r_{1i} be the number of blocks of length *i*.
- If r_0 and r_1 are the no. of gaps and blocks respectively then

$$r_0 = \sum_{i=1}^{n} r_{0i} \text{ and } r_1 = \sum_{i=1}^{n} r_{1i}$$

$n_{01} = r_0$ or r_0 -1, $n_{10} = r_1$ or r_1 -1, $n_{00} = n_0 - r_0$ and $n_{11} = n_1 - r_1$.

- This should not be applied if the sequence had not already passed the serial test.
- We expect about half the gaps (or blocks) to have length 1, a quarter to have length 2 and so on.

(Reference: Mood, A.M., "The distribution theory of runs", The Annals of Mathematical Statistics, 1940)

References

- D. E. Knuth, "The Art of Computer Programming, Volume 2: Seminumerical Algorithms", Addison-Wesley, Longman Publishing, Boston, Mass, USA.
- Alfred J. Menezes, Pall C. van Oorschot, Scott A. Vanstone, "Handbook of Applied Cryptography", CRC Press