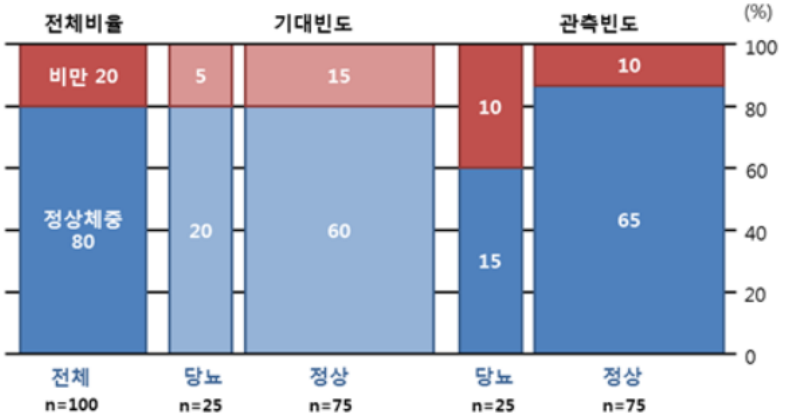
Pseudo Random Generator

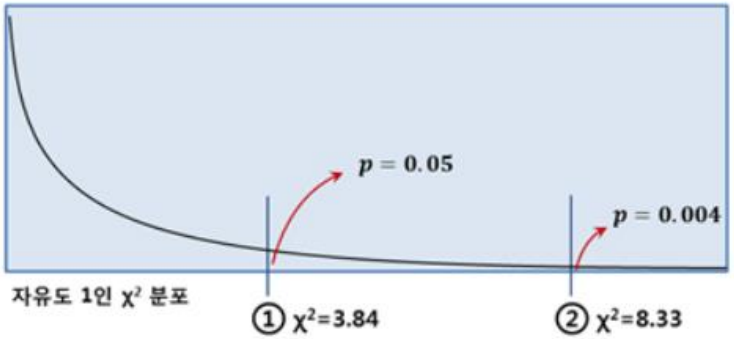
이리나 석사과정

2017.11.6

카이제곱검정 가설 검정

- 귀무가설 : 두 변수는 연관성이 없다
- 대립가설 : 두 변수는 연관성이 있다.
- 두 변수 사이의 연관성이 전혀 없다는 귀무가설 하에서 검정통계량의 값 (카이제곱값) 이 클수록 이러한 현상이 관찰될 가능성(p-value)는 적어지게 된다.





$$\chi^2 = \sum \frac{(\text{관측빈도} - \text{기대빈도})^2}{\text{기대빈도}} = \frac{(+5)^2}{5} + \frac{(-5)^2}{20} + \frac{(-5)^2}{15} + \frac{(+5)^2}{60} = 8.33$$

- 자유도 1인 카이제곱 분포에서 카이제곱값 3.84인 현상이 관찰될 가능성이 5%이며(p value= 0.05) 카이제곱값이 8.33인 현상이 관찰될 확률은 0.4%이다. (p-value=0.004)
- 따라서 귀무가설을 기각하고 당뇨와 비만 사이에 연관성이 있다는 대립가설을 채택할 수 있다.

http://dermabae.tistory.com/?page=16

Five Basic Test?

Let $s = s_0, s_1, s_2, \ldots, s_{n-1}$ be a binary sequence of length n. This subsection presents five statistical tests that are commonly used for determining whether the binary sequence s possesses some specific characteristics that a truly random sequence would be likely to exhibit. It is emphasized again that the outcome of each test is not definite, but rather probabilistic. If a sequence passes all five tests, there is no guarantee that it was indeed produced by a random bit generator (cf. Example 5.4).

- 무작위 비트 생성기(random number generator) 의 성질을 조사하기 위한 테스트를 수행.
- 생성된 표본 출력 수열(Sample out sequence)에 몇 가지 테스트를 적용하여 이루어진다.
- 각각의 통계적 테스트는 무작위 수열이 가져야 하는 속성을 얼마나 만족하는지를 결정적이기보다 확률로 나타낸다.

Five Basic Test

The purpose of this test is to determine whether the number of 0's and 1's in s are approximately the same, as would be expected for a random sequence. Let n_0, n_1 denote the number of 0's and 1's in s, respectively. The statistic used is

$$X_1 = \frac{(n_0 - n_1)^2}{n} \tag{5.1}$$

which approximately follows a χ^2 distribution with 1 degree of freedom if $n \geq 10$. [7]

- 0과 1의 개수가 같다는 가정에 근거한 테스트.
- 0과 1의 개수를 각각 이라고 하면, n>=10인 경우

$X_1 = \frac{(n_0 - n_1)^2}{n}$ 은 자유도(degree of freedom)가 1인 분포를 따른다.

- 이 때 $n_0 + n_1 = n$ 을 만족한다.

Serial test

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, as would be expected for a random sequence. Let n_0, n_1 denote the number of 0's and 1's in s, respectively, and let n_{00}, n_{01}, n_{10}, n_{11} denote the number of occurrences of 00, 01, 10, 11 in s, respectively. Note that $n_{00} + n_{01} + n_{10} + n_{11} = (n-1)$ since the subsequences are allowed to overlap. The statistic used is

$$X_2 = \frac{4}{n-1}\left(n_{00}^2 + n_{01}^2 + n_{10}^2 + n_{11}^2\right) - \frac{2}{n}\left(n_0^2 + n_1^2\right) + 1 \qquad (5.2)$$

which approximately follows a χ^2 distribution with 2 degrees of freedom if $n \geq 21$.

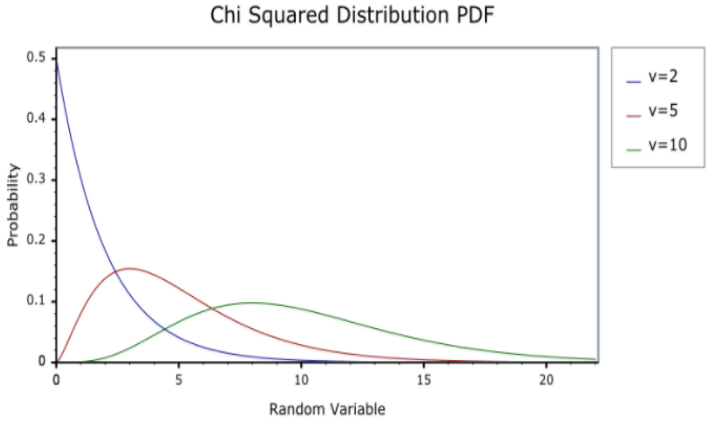
– 연속된 2비트 00, 01, 10, 11의 개수가 같다는 가정에 근거한 테스트.

– 00, 01, 10, 11의 개수를 각각 이라고 하면, $n_{00}, n_{01}, n_{10}, n_{11}$ 인 경우

확률변수 X는 $X_2 = \frac{4}{n-1}\left(n_{00}^2 + n_{01}^2 + n_{10}^2 + n_{11}^2\right) - \frac{2}{n}\left(n_0^2 + n_1^2\right) + 1$

로 정의 되고, 이 X는 자유도 2인 카이 제곱 분포를 따른다.

(이 때 $n_{00}+n_{01}+n_{10}+n_{11}=n-1$, $n_0+n_1=n$ 을 만족함.)



Poker test

Let m be a positive integer such that $\lfloor \frac{n}{m} \rfloor \geq 5 \cdot (2^m)$, and let $k = \lfloor \frac{n}{m} \rfloor$. Divide the sequence s into k non-overlapping parts each of length m, and let n_i be the number of occurrences of the i^{th} type of sequence of length m, $1 \leq i \leq 2^m$. The poker test determines whether the sequences of length m each appear approximately the same number of times in s, as would be expected for a random sequence. The statistic used is

$$X_3 = \frac{2^m}{k} \left(\sum_{i=1}^{2^m} n_i^2 \right) - k \qquad (5.3)$$

– 양의 정수 m이 $\lfloor \frac{n}{m} \rfloor \geq 5 \cdot (2^m)$ 을 만족하면, 양의 정수 $k = \lfloor \frac{n}{m} \rfloor$ 에 대해 수열 s를 각 m비트인 k개의 부분수열로 둘 수 있고, 각각의 부분수열 a는 m비트 수임.

– ni는 i번째 type의 sequence가 나타난 횟수

– 이 때 $X_3 = \frac{2^m}{k} \left(\sum_{i=1}^{2^m} n_i^2 \right) - k$ 는 자유도 $2^m - 1$ 의 카이제곱 분포를 따른다.

Poker test

$$X_3 = \frac{2^m}{k}\left(\sum_{i=1}^{2^m} n_i^2\right) - k \qquad (5.3)$$

- $$\chi^2 = \Sigma\frac{(O-E)^2}{E}$$
- O : Observed Frequency(실제 빈도)
- E : Expecetd Frequency(기대 빈도) (두 변수가 독립, 즉 연관성 없는 경우)

표 15.2 간호 학생의 교실에서 세 인종 집단의 관찰빈도와 기대빈도

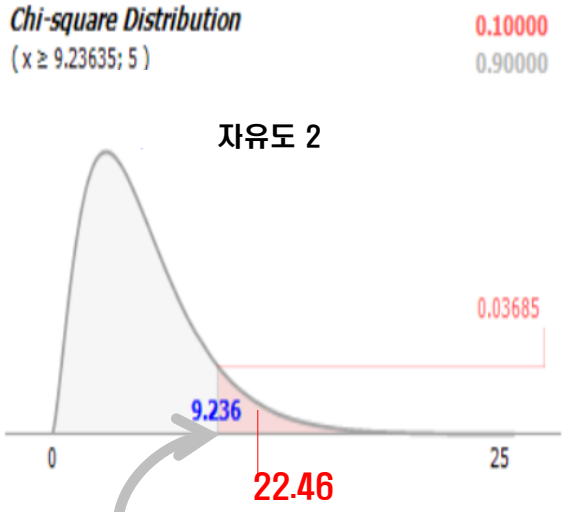
빈도	백인(UK와 EU)	캐리비안인	인도계인	계
관찰	34	62	28	124
기대	59.5	41	23.5	124

$$\chi^2 = \Sigma\frac{(O-E)^2}{E} = \frac{(34-59.5)^2}{59.5} + \frac{(62-41)^2}{41} + \frac{(28-23.5)^2}{23.5}$$

$$= \frac{(-25.5)^2}{59.5} + \frac{(21)^2}{41} + \frac{(4.5)^2}{23.5}$$

$$= \frac{650.25}{59.5} + \frac{441}{41} + \frac{20.25}{23.5}$$

$$= 10.9 + 10.7 + 0.86 = 22.46$$



계산한 카이제곱 값이 특정 자유도와 p값 (ex.p=0.05) 하에서의 카이제곱 값(9.236)보다 크다->귀무가설을 가설을 만족한다->유의수준 0.05 하에서 연관성이 있다.

Poker test

Example (*basic statistical tests*) Consider the (non-random) sequence s of length $n = 160$ obtained by replicating the following sequence four times:

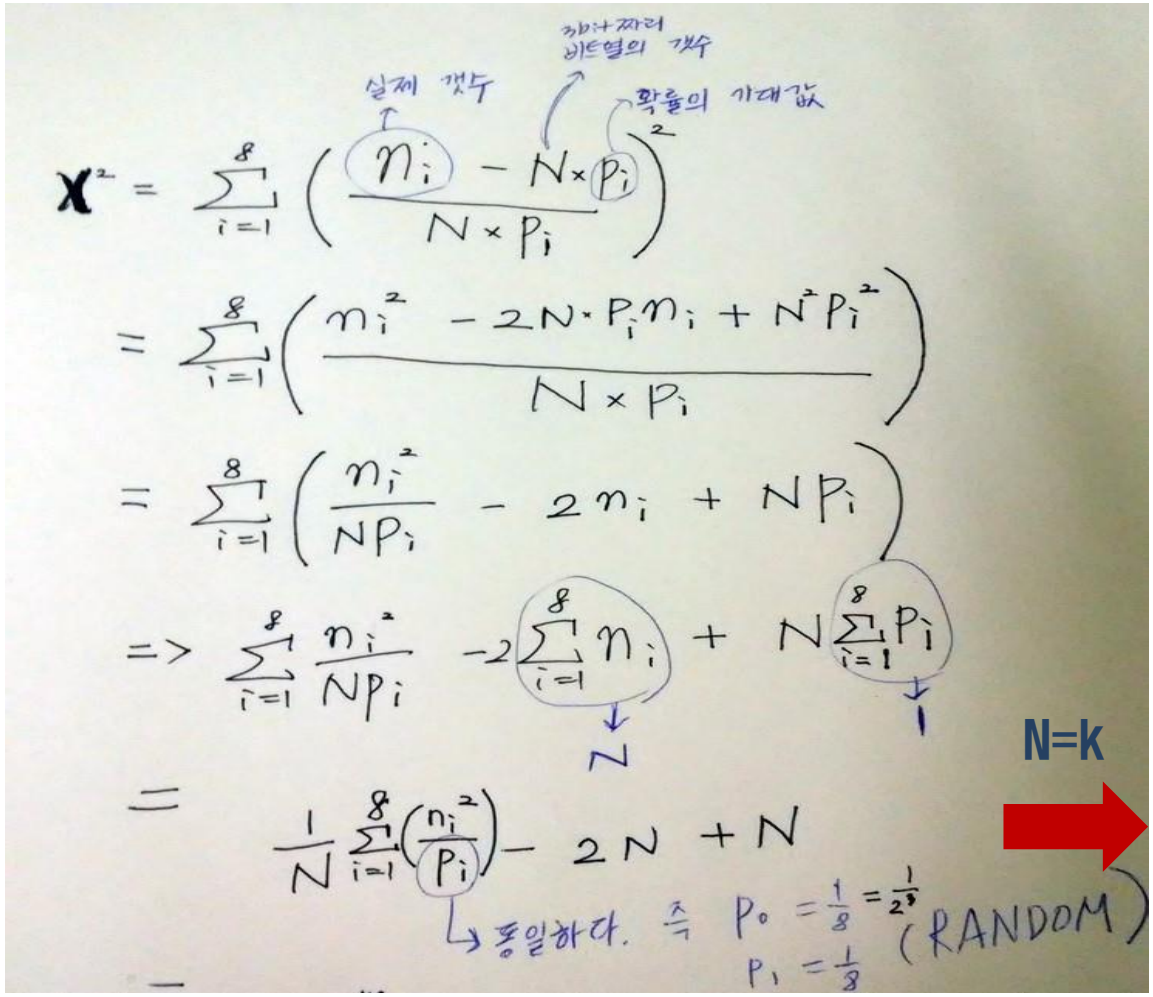
(*poker test*) Here $m = 3$ and $k = 53$. The blocks 000, 001, 010, 011, 100, 101, 110, 111 appear 5, 10, 6, 4, 12, 3, 6, and 7 times, respectively, and the value of the statistic X_3 is 9.6415.

- 각각의 3bit짜리 000, 001, 011의 실제 개수를 n1, n2, ..으로 이름붙임
- n1 = 5, n2 = 10, n3=6 …
- 각각이 등장할 확률의 기대값을 p1, p2, p3…으로 명명
- p1, p2, p3.. 즉 각각의 수열이 나타낼 확률은 동일해야 가장 random함. (3비트의 경우는 1/8)

$$X_3 = \frac{2^m}{k}\left(\sum_{i=1}^{2^m} n_i^2\right) - k$$

- (공식 설명은 다음 페이지)

Poker test



$$X_3 = \frac{2^m}{k}\left(\sum_{i=1}^{2^m} n_i^2\right) - k$$

Runs test

The purpose of the runs test is to determine whether the number of runs (of either zeros or ones; see Definition 5.26) of various lengths in the sequence s is as expected for a random sequence. The expected number of gaps (or blocks) of length i in a random sequence of length n is $e_i = (n-i+3)/2^{i+2}$. Let k be equal to the largest integer i for which $e_i \geq 5$. Let B_i, G_i be the number of blocks and gaps, respectively, of length i in s for each i, $1 \leq i \leq k$. The statistic used is

$$X_4 = \sum_{i=1}^{k} \frac{(B_i - e_i)^2}{e_i} + \sum_{i=1}^{k} \frac{(G_i - e_i)^2}{e_i} \tag{5.4}$$

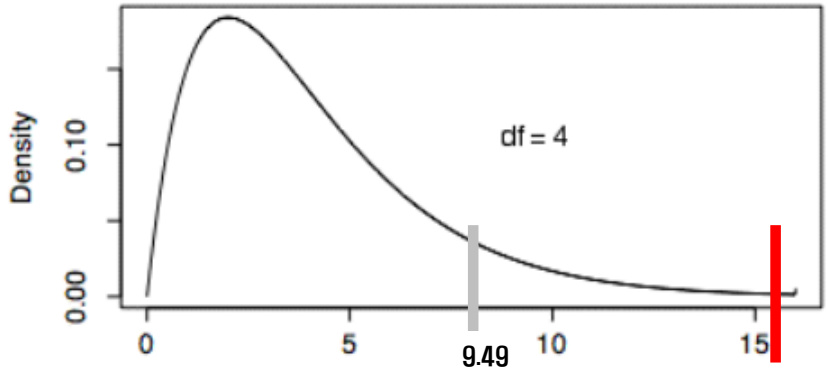
which approximately follows a χ^2 distribution with $2k - 2$ degrees of freedom.

- 0 또는 1이 연속하여 나타나는 정도가 추정치에 근접한지 확인함.
- n비트의 수열에서 0 또는 1이 i개 연속하여 나타날 확률은 $e_i = (n-i+3)/2^{i+2}$ 이다.
- Block(1이 연속된 것) Gap(0이 연속된 것) 의 개수를 각각 B_i, G_i라한다.
- $\sum_{i=1}^{k} \frac{(B_i - e_i)^2}{e_i}$ 은 block의 카이제곱 값. $\sum_{i=1}^{k} \frac{(G_i - e_i)^2}{e_i}$ 은 Gap의 카이제곱 값.
- Block의 카이제곱값과 gap의 카이제곱값을 합하여 run test를 검증한다.

Runs test

example

(*runs test*) Here $e_1 = 20.25$, $e_2 = 10.0625$, $e_3 = 5$, and $k = 3$. There are 25, 4, 5 blocks of lengths 1, 2, 3, respectively, and 8, 20, 12 gaps of lengths 1, 2, 3, respectively. The value of the statistic X_4 is 31.7913.



0,1	0,05	0,025	0,01	0,005
2.71	3.84	5.02	6.63	7.88
4.61	5.99	7.38	9.21	10.60
6.25	7.81	9.35	11.34	12.84
7.78	9.49	11.14	13.28	14.86

두 변수(실제 비트열의 분포와 random한 비트열의 분포)는 연관성이 있다.

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 $A(d) = \sum_{i=0}^{n-d-1} s_i \oplus s_{i+d}$, where $\oplus$ denotes the XOR operator. The statistic used is

$$X_5 = 2\left(A(d) - \frac{n-d}{2}\right) / \sqrt{n-d} \qquad (5.5)$$

which approximately follows an $N(0,1)$ distribution if $n - d \geq 10$. Since small values of $A(d)$ are as unexpected as large values of $A(d)$, a two-sided test should be used.

- 일정한 간격 차이의 두 비트들 간의 관계(Si와 d만큼 shift된)를 테스트함.
- n/2 미만의 양의 정수 d에 대해 $A(d) = \sum_{i=0}^{n-d-1} s_i \oplus s_{i+d}$ 즉 S_i와 S_{i+d} 의 XOR 값은
- $X_5 = 2\left(A(d) - \frac{n-d}{2}\right) / \sqrt{n-d}$ 은 정규분포 N(0, 1)을 따른다.

Maurer's universal statistical test

- The basic idea behind Maurer's universal statistical test is that it should not be possible to significantly compress (without loss of information) the output sequence of a random bit generator. Thus, if a sample output sequence s of a bit generator can be significantly compressed, the generator should be rejected as being defective. Instead of actually compressing the sequence s, the universal statistical test computes a quantity that is related to the length of the compressed sequence.
- 기본 아이디어는 주어진 비트 수열이 무작위라면 그 수열을 압축하였을 때 그 압축률이 너무 크지는 않을 것이라는 사실이다. 즉 어떤 수열을 압축하였을 때 큰 압축률을 보인다면 그 수열을 거부한다. 이 테스트는 실제로 주어진 수열을 압축하는 대신 압축률에 관련된 값만을 계산하므로 그 실행속도는 비교적 빠르다. 그러나 매우 큰 비트 수열을 필요로 하기 때문에 비트생성기가 표준출력수열을 생성하는데 비교적 많은 시간이 소요된다.

Maurer's universal statistical test

Algorithm Computing the statistic X_u for Maurer's universal statistical test

INPUT: a binary sequence $s = s_0, s_1, \ldots, s_{n-1}$ of length n, and parameters L, Q, K.
OUTPUT: the value of the statistic X_u for the sequence s.

1. *Zero the table* T. For j from 0 to $2^L - 1$ do the following: $T[j] \leftarrow 0$.
2. *Initialize the table* T. For i from 1 to Q do the following: $T[b_i] \leftarrow i$.
3. sum$\leftarrow$0.
4. For i from $Q+1$ to $Q+K$ do the following:
   4.1 sum$\leftarrow$sum $+ \lg(i - T[b_i])$.
   4.2 $T[b_i] \leftarrow i$.
5. $X_u \leftarrow$ sum$/K$.
6. Return(X_u).

- 주어진 2진 수열 s를 Q+K개의 L비트 블록으로 분리
- T[j]는 가장 최근에 j값이 나타낸 블록의 위치
- $A_i = i - T[b_i]$, $Q + 1 \le i \le Q + K$ 라고 하면 Ai는 bi값이 몇 블록 이전에 나타났는지를 의미
- $$X_u = \frac{1}{K} \sum_{i=Q+1}^{Q+K} \lg A_i.$$ 는 u, o에 대해 평균 u와 분산 σ^2 을 갖는 정규 분포를 따른다.

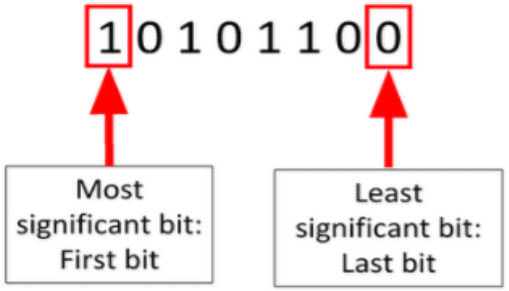
RSA pseudorandom bit generator

Algorithm RSA pseudorandom bit generator

SUMMARY: a pseudorandom bit sequence $z_1, z_2, \ldots, z_l$ of length l is generated.

1. *Setup*. Generate two secret RSA-like primes p and q (cf. Note 8.8), and compute $n = pq$ and $\phi = (p-1)(q-1)$. Select a random integer e, $1 < e < \phi$, such that $\gcd(e, \phi) = 1$.
2. Select a random integer x_0 (the *seed*) in the interval $[1, n-1]$.
3. For i from 1 to l do the following:
   3.1 $x_i \leftarrow x_{i-1}^e \bmod n$.
   3.2 $z_i \leftarrow$ the least significant bit of x_i.
4. The output sequence is $z_1, z_2, \ldots, z_l$.

1. P와 q라는 두 소수를 생성
2. n=pq 계산. 파이값(p-1)(q-1) 계산
3. 파이 값과 서로 소인 random한 정수 e를 정함.
4. 1과 n-1 사이의 임의의 정수 x0(seed값)을 선택함
5. x_i는 x_{i-1}mod n
6. i번째 z값은 x의 least significant bit를 셋



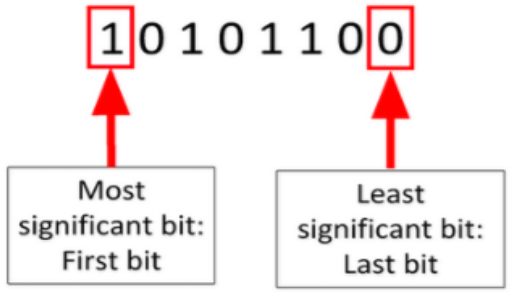
Micali-Schnorr pseudorandom bit generator

Algorithm Micali-Schnorr pseudorandom bit generator

SUMMARY: a pseudorandom bit sequence is generated.

1. *Setup*. Generate two secret RSA-like primes p and q (cf. Note 8.8), and compute $n = pq$ and $\phi = (p-1)(q-1)$. Let $N = \lfloor \lg n \rfloor + 1$ (the bitlength of n). Select an integer e, $1 < e < \phi$, such that $\gcd(e, \phi) = 1$ and $80e \leq N$. Let $k = \lfloor N(1 - \frac{2}{e}) \rfloor$ and $r = N - k$.
2. Select a random sequence x_0 (the *seed*) of bitlength r.
3. *Generate a pseudorandom sequence of length* $k \cdot l$. For i from 1 to l do the following:
   - 3.1 $y_i \leftarrow x_{i-1}^e \bmod n$.
   - 3.2 $x_i \leftarrow$ the r most significant bits of y_i.
   - 3.3 $z_i \leftarrow$ the k least significant bits of y_i.
4. The output sequence is $z_1 \| z_2 \| \cdots \| z_l$, where $\|$ denotes concatenation.

1. p와 q라는 두 소수를 생성
2. n=pq 계산. 파이값=(p-1)(q-1) 계산
3. N = [logn]+1 (n의 비트 길이)
4. 정수 e를 선택, gcd() = 1 파이 값과 서로 소인 random
4. 1과 n-1 사이의 임의의 정수 x0(seed값)을 선택함
5. x값으로 y를 생성. y값으로 x와 z를 생성.

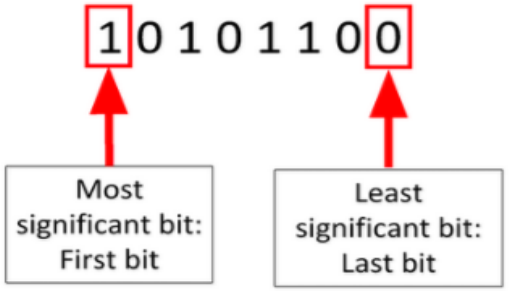


Blum Blum Shub 생성기

SUMMARY: a pseudorandom bit sequence $z_1, z_2, \ldots, z_l$ of length l is generated.

1. *Setup*. Generate two large secret random (and distinct) primes p and q (cf. Note 8.8), each congruent to 3 modulo 4, and compute $n = pq$.
2. Select a random integer s (the *seed*) in the interval $[1, n-1]$ such that $\gcd(s, n) = 1$, and compute $x_0 \leftarrow s^2 \bmod n$.
3. For i from 1 to l do the following:
   3.1 $x_i \leftarrow x_{i-1}^2 \bmod n$.
   3.2 $z_i \leftarrow$ the least significant bit of x_i.
4. The output sequence is $z_1, z_2, \ldots, z_l$.

- 임의의 소수 p와 q를 선택(아주 큰 숫자여야 소인수분해가 어렵다.)
- 1과 n-1 사이의 임의의 정수 s(seed값)를 선택함. s와 n은 서로 소여야 한다.
- x_0을 셋팅 후 xi값을 제곱 후 mod n값으로 순차적으로 계산함
- 생성된 z값은 least significant bit으로 사용



감사합니다

Q & A