通信情報保護學會誌
第3卷 第4號, 1993. 12

랜덤 치환의 안전성과 통계적 검정

Statistical Tests for the Random Permutations

이 경 현*

요 약

본 논문에서는 n개의 원소를 임의로 나열하는 대표적인 치환 발생 알고리즘을 소개하고 이들 중 발생 치환의 랜덤성이 우수하다고 알려진 Knuth의 알고리즘을 근간으로 설계된 랜덤 치환 발생기의 암호학적 안전성을 분석하고 발생 치환들에 대한 난수 발생 수열 관점의 통계적 임의성과 연속 발생 치환끼리의 통계적 독립성 적용을 위한 통계량 소개 및 각 관점에서 통계적 검정을 통과함을 시뮬레이션 결과를 통하여 보인다.

1. 서 론

n개의 원소를 임의로 나열하는 치환(Permutation) 발생 알고리즘은 지난 20여년간 여러 가지 방법으로 제안되어져 왔다.[3~7] 이러한 알고리즘들은 대부분 인접된 원소끼리의 교환(Interchange)에 그 근간을 두고 있으며 각각은 실제적인 영역에서의 사용목적에 따라 서로 다른 발생 알고리즘으로 모든 n!개의 치환을 발생하여 이용되어왔다. 한편 이러한 치환 발생 알고리즘들은 전자 계산학에 기본적 개념이 되는 카운팅(Counting)문제, 재귀적 및 반복적 문제들간의 관계를 잘 나타내어주므로 다양한 연구가 진전되어져 오고 있다. 본 논문에서는 이러한 각종 치환 발생 알고리즘중 랜덤성이 우수하며 고속으로 실시간 발생 가능한 램덤 치환 발생기[1,10]에 의해 발생된 치환들이 통계적으로 우수함을 이진 비트들에 대한 난수의 통계적 검정법[2,9]과 연속적으로 발생되는 치환끼리의 독립성 검정[6]을 통해 보이고자 한다.

제2장에서는 각종 치환 발생 알고리즘들중 대표적인 것을 몇개 알아보고 3장에서는 치환발생 알고리즘들중 단 방향성(Trapdoor-Oneway)과 암호학적 안정성(Cryptographically Secure)이 뛰어난 치환 발생법을 Knuth의 랜덤 치환 발생법을 근간으로, 임의 정수의 계승수 체계(Factorial Number System)와 랜덤 치환과의 관계를 이용한 발생법에 대해 간략히 알아본다.[1,5,10] 4,5장에서는 3장에서 발생환 치환들의 안전성 평가를 위한 도구로써 이진 수열의 임의성 검정법 및 연속된 치환끼리의 독립성 검정법에 대해 소개하고 검정에 의한 컴퓨터 시뮬레이션 결과를 기술한다.

* 부산수산대학교 전자계산학과

2. 각종 치환 발생 알고리즘

n개 원소로 구성된 임의의 치환의 집합을 S_n이라고 하면 S_n은 차수 n의 대칭군을 이루며 S_n의 개수는 n!개가 된다. 또한 표현의 편의를 위하여 임의의 치환 $\alpha \in S_n$에 대해 $\alpha=(\alpha_1, \alpha_2, ..., \alpha_n)$으로 나타내고 이것은 1이 α_1로 2가 α_2로, ..., n이 α_n으로 자리 바꿈함을 의미하는 것으로 약속하자. 한편 수의 표현에 있어서 일반적으로 많이 사용되고 있는 진법은 흔히 우리들이 사용하고 있는 십진법 체계이다. 그러나 컴퓨터를 사용한 수의 표현에는 이진법과 십육진법(Hexadecimal)이 가장 많이 쓰이고 있으며 이때 진법 표현의 기본이 되는 수를 기저(Base) 또는 기(Radix)라고 하고 이러한 기저로 이루어진 수의 표현 방법을 기 체계(Radix System)라고 부른다. 한편 기 체계를 일반화하여 각 자리수마다 한개의 기가 아닌 여러개의 기들로 구성될 때 이를 혼합 기 체계(Mixed Radix System)라고 하며 이러한 혼합 기 체계의 특수한 경우(즉, 각 자리수의 기가 계승인 경우)가 바로 계승 수 체계(Factorial Number System)이다. 즉, 임의의 $m \in (0, n!-1)$에 대해 $m=(d_1, d_2, ..., d_{n-1})$의 표현은 $m=(d_1 1! + d_2 2! + ... + d_{n-1}(n-1)!(0 \le d_i \le i,\ m \neq 0)$임을 의미한다. 또한 0과 n!−1 사이의 임의의 수의 계승 수 체계와 n개 원소로 구성된 임의의 치환 사이에는 일대일 대응관계가 있음이 알려져 있다.[3,5] 따라서 이러한 대응관계에 따라 각종 치환 발생 알고리즘이 고려될 수 있으며 대표적인 것들은 다음과 같다.[3,5]

2.1. Nested Cycles (Tompkins-Paige's Algorithm)

치환들중에 가장 간단한 종류는 단 한개의 사이클로 구성된 것으로써 이러한 치환들은 치환의 구성 원소를 오른쪽으로 순환시키거나 몇몇 위치의 원소들만을 회전시킴으로써 발생시킬 수 있다. 특히 S_n의 임의의 치환이 차수(Order) k이고 정도(Degree) d인 순환 치환이라는 것은 주어진 치환의 왼쪽 k개 원소들이 오른쪽으로 d 위치만큼 순환하고 나머지 (n−k)개의 원소들은 제자리에 고정되어 있다는 것을 나타낸다. 이 경우 S_n의 임의의 치환 α는 차수 i의 순환 치환 P_i의 연속된 곱, 즉 $\alpha=P_n, P_{n-1}, ..., P_2 P_1$과 같은 성질을 만족하는 치환 P_i들의 곱으로 유일하게 표시될 수 있다. 단, 여기서 치환의 곱이란 각 치환 원소끼리의 왼쪽에서 오른쪽으로의 합성(Composition)을 의미하며 P_1은 모든 원소가 자리 바꿈 하지 않은 항등 치환을 나타낸다. 따라서 P_n, P_{n-1}, ..., P_2 등에 대해 모든 가능성을 체계적으로 나열함으로써 n개의 원소에 대한 모든 치환들을 발생할 수 있으며, 이러한 대응은 임의의 계승 수 체계에 대해 차수 i+1이고 정도 $d_i (0 \le d_i \le i,\ 1 \le i \le n-1)$인 n−1개의 단순 치환들의 합성으로 나타낼 수 있다.

2.2. 사전식 순서 (Lehmer's Algorithm)

n개 원소로 구성된 치환 수열의 사전식 순서(Lexicographic Order)는 n개 원소가 증가되는 순서대로 치환을 배열하는 것을 의미한다. 예를 들면 1,2,3의 3개의 원소로 구성된 치환 수열의 사전식 순서는 (123), (132), (213), (231), (312), (321)이다. 이때 계승 수 체계에 의한 임의의 정수 $m \in (0, n!-1)$에 대응되는 치환은 전체 n!개 치환중 사전식 순서에서 m번째에 해당하는 치환이 되며, 먼저 d_i의 각 원소를 1 증가시켜서 $d_i+1 (i=2, 3, ,,,, n)$을 얻고 1을 쓰고 d_i보다 큰 모든 수를 1 증가시키면서 d_i를 왼쪽편에 써내려감으로써 생성시킬 수 있다.

2.3. 역 벡터 (Hall's Algorithm)

임의의 치환 $\alpha=(\alpha_1, \alpha_2, ..., \alpha_n) \in S_n$에 대해 $i<j$이고 $\alpha_i > \alpha_j$ 일때 한쌍의 원소(α_i, α_j)를 α의 역(Inversion)이라 하고 (α_i, α_j)가 역이 되는 계수를 d_i로 두면 d_j는 바로 α_j 보다 크면서 α_j의 왼쪽에 있는 원소의 개수가 된다. 따라서 $0 \le d_j < j$ 이고$(d_1=0)$, d_j 수열은 바로 계승 수 체계에 해당된다. 역으로 계승 수 체계 $\{d_j\}$에 대응되는 치환 α도 역 벡터를 이용하여 쉽게 찾을 수 있으며 주어진 치환내에서 역 벡터의 합은 그 치환의 재배열의 정도(일종의 거리)를 나타내므로 이것을 치환의 통계적 검정의

통계량으로 사용할 수 있다.

2.4. 인접한 원소간의 교환 (Well's Algorithm)

한개의 주어진 임의의 치환에서 특정한 두 곳(Two Marks)의 원소를 서로 교환함으로써 연속적으로 그 뒤에 따르는 모든 치환들을 발생할 수 있다. 이때 특정한 Marks를 결정하는 함수 h=h(m)(0≤m<n!)은 각 m에 대해 계승 수 체계의 d_i 즉 i을 만족하는 가장 작은 인수(Least Subscript)로 두면 m번째 치환에서 (m+1)번째 치환을 다음과 같은 방법으로 발생시킬 수 있다.

1) h가 홀수이거나 h가 짝수이고, $d_{h+1}<2$이면 Marks가 h와 h-1이 되어 서로 교환한다.

2) 그 외의 경우는 Marks가 h와 $h-d_{h+1}$이 되어 서로 교환한다. 단 여기서 각 Mark의 위치는 0부터 n-1로 순서가 매겨지고 음수 Mark는 0으로 해석된다.

2.5. Random 발생법 (Knuth's Algorithm)

n개원소로 구성된 랜덤 치환은 임의의 치환 $\alpha=(\alpha_1,\alpha_2,...,\alpha_n)$으로부터 출발하여(보통 원소의 크기 순서로 나열된 항등 치환(1, 2, ..., n)을 이용) α_n을 $\alpha_1, \alpha_2, ..., \alpha_{n-1}$ 중의 한개와 교환하고 또 α_{n-1}을 $\alpha_1, \alpha_2, ..., \alpha_{n-2}$ 중의 한개와 교환하는 방식으로해서 (n-1)번의 원소끼리의 교환을 반복함으로써 발생한다.

Algorithm : 임의의 초기 치환 $\alpha=(\alpha_1,\alpha_2,...,\alpha_n)$에 대해
For i=n to 2 by -1 do
Interchange $\alpha_i \longleftrightarrow \alpha_k$
단 k=rand(1,i) : 1과 i사이의 난수

한편 위의 방식으로 발생된 n!개의 치환 각각이 확률적으로 항등적(Equally Uniformly)인 것은 n에 대한 수학적 귀납법으로 쉽게 보일 수 있다.

3. 단 방향 치환 (Trapdoor-Oneway Permutation) 발생

3.1. 치환 발생 알고리즘

앞절의 치환 발생 알고리즘들은 주어진 원소끼리의 교환(Interchange)에 그 근간을 두고 있으며 이러한 알고리즘들 중 Random성이 우수하며 주어진 허용 처리 시간내에서 고속으로 또한 연속적으로 치환 발생이 가능한 알고리즘을 채택하여 통계적 검정법에 적용하고자 한다.[1,10] 특히, 발생 치환의 안전성 관점에서 최종 출력 치환으로부터 구성 성분의 치환을 추출해내는 알고리즘의 구조적 취약성을 없애기 위해 치환끼리의 결합(Composition)을 이용함으로써 발생 치환에 단 방향성(Trapdoor-Oneway)을 부여해 왔다.

한편, 이러한 Random 치환 발생의 근간은 앞절에서 언급했듯이 Knuth의 알고리즘에 두고 있으며 이는 n개의 원소를 크기 순서대로 배열한 후 (즉 P=(1,2,...,n)),0과 n!-1 사이의 임의의 난수를 선형 합동법 $X_{m+1}=A*X_m+C$ Mod n! 이 보장하는 주기성을 이용하여 발생하였으며 본 논문에서는 발생의 편의 및 통계적 검정을 위해서 Akl-Meijer[1]가 제안한 방법(상수 A=1인 경우)을 사용하였다. 사용 알고리즘의 절차는 다음과 같다.

Step 1 : $Z_{n!}$={0, 1, ..., n!-1}내의 임의의 수 x^0와 GCD(C, n!)=1이 되는 C를 (0, n!-1) 내에서 초기값으로 잡는다.

Step 2 : Step 1에서 발생된 x^i를 n개 원소의 Vector 형태로 배열한 계승 수 체계를 이용하여 다음의 방법으로 치환 P1을 발생한다.

$x^i=(x^i(0), x^i(1), ..., x^i(n-1))$
항등치환 P1=(0, 1, ..., n-1) (즉, P1(j)=j)
j = 0, 1, ..., n-1에 대해
$P1(j) \longleftrightarrow P1(x^i(j))$를 n-1번 교환한다.
같은 방법을 반복하여 치환 P2를 발생한다.
$x^{i+1}=(x^{i+1}(0), x^{i+1}(1), ..., x^{i+1}(n-1))$
j = 0, 1, ..., n-1에 대해
$P2(j) \longleftrightarrow P2(x^i(j))$를 n-1번 교환한다.

Step 3 : Step 2에서 발생한 치환 P1과 P2를 결합하여 단 방향 치환 P를 발생한다.

j = 0, 1, ..., n-1에 대해
P(j)=P2(P1(j))

필요에 따라서는 Step 2와 Step 3을 k번까지 반복 수행하여 최종적인 단 방향 치환 P를 발생할 수 있다.

Step 4 : Step 3에서 발생된 치환에 대해 원소의 불변 요인(Unmoved Factor)에 대한 제약 조건을 적용하여 발생치환에 제약을 가한다.

j = 0, 1, ..., n-1에 대해

또한 |j-P(j)| <2 이면 Go to Step 1

3.2. Good Permutation 발생 조건

위에서 발생된 치환들 중에서 치환의 사용 및 응용성격에 따라 실시간 발생 가능한 범위 내에서 치환 발생에 제약 조건을 가할 수 있다. 즉, 발생 치환들에 적절한 Check 조건을 적용하여 Good Permutation을 선택한다. 보통 고려되는 제약 조건은 치환하기 이전 각 원소 본래의 위치에서 치환 후 얼마만큼 위치를 옮겼는가의 척도가 되는 천이 요인(Shift Factor), 치환의 어떠한 원소도 그것의 이웃에 연속되어 있는 원소에 대해 치환 후 다시 연속적으로 이웃해 남게 되는 경우를 배제하기 위한 연속 요인(Consecutive Factor) 및 어떤 원소가 치환 후 본래의 위치에 그대로 남게 될 경우를 배제시키는 불변 요인(Unmoved Factor) 등이 있고 이들 중 필요에 따라 적절한 범위내에서 제약조건을 치환 발생 알고리즘에 적용할 수 있다. 본 논문에서는 위의 조건들 중 치환의 랜덤 특성상 가장 일반적인 조건인 불변 요인만을 고려하여 이를 알고리즘 발생에 적용시켰으며 이 경우 발생되는 치환의 개수는 근사적으로 다음과 같이 계산되어질 수 있다.

만약 n개 원소로 구성된 치환에서 한개 원소라도 자리바꿈하지 않은 치환들을 제외한 치환의 개수를 UM(n)이라 할 때

$UM(n) = Per(D_n)$: Permanent of D_n

단 $D_n = J - I_n$

J는 원소가 모두 1로 구성되어 있는 nxn 행렬

I_n는 nxn 항등행렬(Identity Marix)

이때 $Per(D_n)$은 다음과 같이 Recursive하게 주어진다.

$$Per(D_n) = (n-1)\cdot(Per(D_{n-1}) + Per(D_{n-2}))$$

또한 정확한 $Per(D_n)$의 값은

$Per(D_n) = n\cdot Per(D_{n-1}) + (-1)_n$, $n>2$이며,

수학적 귀납법에 의해

$$Per(D_n) = n!(1-1/1!+1/2!-1/3! + \ldots + (-1)_n 1/n!)$$ 이고

충분히 큰 n에 대해 근사적으로 $Per(D_n) = n!\cdot e^{-1}$

단, 위에서 사용된 Permanent 함수는 다음과 같다.

주어진 $A=(a_{ij})nxn$; nxn 행렬에서

$$Per(A) = \sum_{\sigma} \alpha_{1\sigma(1)}\, \alpha_{2\sigma(2)} \ldots \alpha_{n\sigma(n)}$$

단 σ는 대칭군 S_n 내의 모든 치환을 취한다.

따라서 위의 수식에서 $e^{-1}=0.367879....$ 이므로 불변 요인에 의해서는 약 10% 정도가 이용가능한 치환수에서 제한되어진다고 볼 수 있다. 따라서 발생 치환에 제약 조건을 가하더라도 사용가능한 치환의 총 개수는 충분하다는 것을 보장할 수 있다. 랜덤 치환 발생법에서 설명했듯이 계승 수 체계의 각 d_i는 앞세우는 원소의 개수를 의미하게 된다.

또한 위의 알고리즘은 d_i 배열에 대해 한번의 Scanning으로 치환을 발생할 수 있다는 점과 치환의 분석적인 측면에서 치환에서 계승 수 체계를 추출해내기가 어렵다는 두가지 장점이 있다.

2.3. 치환발생 알고리즘의 분석

2-3-1. 치환의 통계적 특성 및 임의성

알고리즘에 의해 발생되는 치환 각각이 사용 목적에 맞는 치환 효과(예를 들면 음성의 스크램블링 효과)를 낼 수 없는 약점을 가진다든지 또한 연속적으로 발생된 치환끼리의 통계적 종속성(Statistical Dependence)이 존재한다면 암호 해독가는 이러한 치환의 통계적 특성 및 약점을 최대한 이용할 것이므로 이러한 점을 고려하여 치환 발생 알고리즘을 설계하여야 한다.

2-3-2. 치환 발생 알고리즘의 강인성

일반적으로 암호 해독가는 충분한 양의 치환을 얻을 수 있고 얻은 치환에서부터 역으로 키이를 얻고자 노력할 것이다. 따라서 많은 양의 치환에서 실제 치환을 발생시키는 키이에 대한 유추가 어렵게 알

고리즘이 설계되어야 한다. 앞서 설명된 랜덤 치환 발생 알고리즘에 대한 이러한 강인성(Robustness)은 다음과 같이 설명할 수 있다.

치환발생 부분에서 설명되었듯이, S_n은 Symmetric Group of Order n이고 $|S_n|$은 S_n의 원소 갯수이다.

°, *를 2개의 S_n위에서의 연산(Operations)이라고 할 때 임의의 치환 수열 $\{P_i\}=P_0, P_1, P_2, \ldots$는 다음과 같이 발생된다.

Step 0 : x^0를 S_n위의 임의의 원소이고,

Step 1 : $x^i = x^{i-1} * C, i>0$이고 적당한 $C \in S_n$에 대해

Step 2 : $p^i = x^{ki} \circ x^{ki+1} \circ \ldots \circ x^{ki+k-1}$, 임의의 $i>0$과 적당한 양의 정수 k에 대해.

이때 S_n, C, k와 두 연산 *와 °는 $\{P^i\}$가 암호학적으로 안전하도록(Cryptographically Secure) 선택되어야 한다. 발생되는 치환 수열 $\{P^i\}$는 다음에 의해 안정성(Security)이 보장된다.

(1) C는 $\{x^i\}$의 주기가 최대가 되도록 선택되어진다.

실제 수열 $\{x^i\}$는 최대의 경우 $|S_n|$ 만큼의 주기를 가진다. 이때 이에 대응되는 치환 수열$\{P^i\}$의 주기는 $|S_n| / GCD(k, |S_n|)$으로 주어진다.

(2) 두 연산 *와 °는 분배법칙(Distributive Law)을 만족하지 않는다.

알려지지 않은 $\{x^i\}$와 C에 의해 수열 $\{P^i\}$의 원소를 구하는 공식을 유도하든지 간단히 하는 것은 어렵다. 실제 랜덤 알고리즘에서 *연산은 +mod n!로 주어졌고, °는 치환끼리의 합성(Composition)으로 주어짐으로써 위의 성질을 만족한다.

(3) 집합 S_n은 연산 °에 대해 군(Group)을 이룬다.

즉, 임의의 π, τ에 대해 $\pi=\tau \circ x$를 만족하는 해 $z \in S_n$가 존재한다. 만약 치환 수열 $\{P^i\}$의 한 원소 P^i가 알려지더라도 각 연산 °(Composition)에 대해 해(Solution)가 존재하므로 $x^{kj}, x^{kj+1}, \ldots, x^{kj+k-1}$ 값들을 알아내기는 거의 힘들다.

4. 통계적 치환 검정법

4.1. 임의성 검정

랜덤 치환 발생기를 이용하여 이진 난수를 발생하는 기법으로서는 특정한 비트 패턴(예를 들면, 1과 0이 교대로 나타나는 101010...1010, 또는 전체 치환 크기의 반은 0이고 나머지 반은 1인 000...0111...11의 형태)을 랜덤 치환의 입력으로 사용하여 비트의 치환 출력을 직접적인 이진 난수로 사용할 수 있다. 이 경우 연속적인 n-1개 비트를 이용하여 n번째 비트를 완전히 예측할 수 있는 약점이 있지만 이러한 약점은 쉽게 보완될 수 있으며, 본 절의 임의성 검정은 이러한 치환의 이진 난수 출력을 기존의 알려진 이진 난수의 검정법[10]을 이용한다. 적용한 이진 난수 검정법 3가지는 비트열을 m비트 단위로 나누었을 때 m차원에서 균등하게 분포되었는가를 검정하는 도수-m 검정(Poker 검정), 0이 0이나 1로 천이되어 가는 과정이 독립인가를 검정하는 계열 검정, 0-runs과 1-runs의 길이를 이용한 연 검정 등이다.

4.2. 독립성 검정

일반적으로 두 개의 확률변수(Random Variable) X, Y에 대한 독립성 검정(Independence Test)을 위해서는 두개의 서로 다른 등급 사이의 관계를 측정하는 방법이 이용되며 대표적인 것으로 Kendall's 샘플 계수 T와 Spearman's 등급 상관 계수 R이 통계치(Statistics)로 사용된다. 본 절에서는 임의로 발생된 치환에 대해 본래의 순서(항등치환, 즉 $\sigma=(1, 2, \ldots, n)$)와의 무관련성, 즉 독립성(Indenpendence)을 보장하는 독립성 검정(Independence Test)을 치환 알고리즘에 적용하였다.

1) Kendall의 샘플 계수 T (Kendall's Sample Tau Coefficient)

(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)를 2변량 모집단(Bivariate Population)에 의한 샘플이라 할 때 임의의 쌍 (X_i, Y_i)와 (X_j, Y_j)에 대해 만약 $Y_i < Y_j$ 일때마다 $X_i < X_j$ 이거나, $Y_i > Y_j$ 일때마다 $X_i > X_j$ 일때 두쌍은 완전한 일치의 관계(The Relation of Perfect Concordance)라 하고, 만약 $Y_i < Y_j$ 일때마다 $X_i >$

X_j 이거나, $Y_i > Y_j$ 일때마다 $X_i < X_j$ 일때 두쌍은 완전히 불일치의 관계(The Relation of Perfect Discordance)라 부른다. 이때 완전한 일치의 확률, 즉, $Pr\{(X_j - X_i)(Y_j - Y_i) > 0\} = \pi_c$라고 두고, 완전한 불일치의 확률, $Pr\{(X_j - X_i)(Y_j - Y_i) < 0\}$을 π_d라 둘 때, 두 확률변수 $X = (X_1, X_2, \ldots, X_n)$와 Y 사이의 관계에 대한 측도(Measure)를 $\tau = \pi_c - \pi_d$라 정의하고 τ를 Kendall's Tau라고 부른다. 독립인 연속 확률변수인 경우에는 $\tau = 0$이나 일반적으로 역은 성립하지 않는다. 보통 $\tau = 0$이면 두 확률변수 X, Y간의 상관계수(Correlation Coefficient) $\rho = COV(X, Y) / \sigma_X \sigma_Y$ 또한 0이 된다(단, σ_X, σ_Y는 X와 Y의 표준편차이고 COV(X, Y)는 X, Y의 Covariance이다). 따라서, $\tau = 0$으로서 X, Y 두 확률변수간의 독립성을 판단할 수 있으며 이러한 τ를 독립성 검정을 사용하기 위해 표본으로부터 τ에 대한 적절한 추정량(Estimator)를 찾아야 한다. $sgn(X_j - X_i)sgn(Y_j - Y_i)$를 A_{ij}라 두면(이때 sgn u는 sign함수로서 $u > 0$이면 1, $u < 0$이면 -1, $u = 0$면 0의 값을 갖는다.)

A_{ij}가 다음의 a_{ij} 값을 가짐에 따라

$$a_{ij} = \begin{cases} 1, & \text{만약, 두쌍 } (X_i, Y_i) \text{와 } (X_j, Y_j) \text{가 일치할때} \\ 0, & \text{만약, 두쌍 } (X_i, Y_i) \text{와 } (X_j, Y_j) \text{가 일치하지도 불일치 하지도 않을때} \\ -1, & \text{만약, 두쌍 } (X_i, Y_i) \text{와 } (X_j, Y_j) \text{가 불일치 할때} \end{cases}$$

$$Pr(A_{ij} = a_{ij}) = \begin{cases} \pi_c & \text{만약 } a_{ij} = 1 \\ \pi_d & \text{만약 } a_{ij} = -1 \\ 1 - \pi_c - \pi_d & \text{만약 } a_{ij} = 0 \end{cases}$$

가 되고 또한 $E(A_{ij}) = \pi_c - \pi_d = \tau$이다. 따라서 A_{ij}는 τ의 불편 추정량(Unbiased Estimator)이다. 또한 $a_{ij} = a_{ji}$이고 $a_{ii} = 0$으로 두면 $T = \Sigma\Sigma_{i<j} ({}_nC_2)^{-1} A_{ij}$ 도 τ의 불편추정량이 된다. 이 통계치 T를 Kendall의 샘플 Tau계수라 하고 사용의 편의를 위해 다음과 같이 형태를 변경할 수 있다. 즉, P를 양인 A_{ij}의 개수라 두고 N을 음인 A_{ij}의 개수라 두면 $(1 \le i < j \le n)$,

$$T = (P - N) / {}_nC_2 \quad \text{단, } {}_nC_k = n!/(k!(n-k)!)$$

만약 두개의 관찰치 X와 Y내에 비등(tie)이 없다면 $A_{ij} \neq 0 (i \neq j)$이고 $P + N = {}_nC_2$, 따라서

$$T = (2P/{}_nC_2) - 1 = 1 - (2N/{}_nC_2) = 1 - 4N/(n(n-1)), \quad -1 \le T \le 1$$

이 T를 X와 Y의 독립성 Test에 적용한다. 즉,

Null Hypothesis H_0 : X와 Y는 독립
Alternative Hypothesis H_1 : X와 Y는 종속

에서 만약 T가 크면 H_0를 Reject한다. 또한 H_0하에서는 $\tau = 0$이고 따라서 T의 Null Distribution은 0을 중심으로 대칭이다.

결론적으로 T의 관찰치(Observed Value), t가 $|t| > t_{\alpha/2}$을 만족시킬때(단, $Pr\{|T| > t_{\alpha/2} \mid H_0\} = \alpha$) 유의수준 α에서 H_0를 기각(Reject)시킨다. 확률변수 X, Y의 순서를 재배열함으로써(즉, 예를들어 X를 크기가 증가하는 순서로 배열한다.) P를 $1 \le i < j \le n$에서 $Y_i - Y_j > 0$인 (Y_i, Y_j)쌍의 개수로 둘 수 있고 이것을 T의 관찰치로 받아들인다. 한편 n이 작은 값에 대해서는 Null Distribution을 쉽게 계산할 수 있으나, n이 큰 값일때는 (보통 $n \ge 8$) 통계치 T는 근사적으로 정규분포를 가진다. 즉,

$$Pr(T \le t) \longrightarrow N(0, 2(2n+5)/(9n(n-1))) \quad \text{as } n \longrightarrow \infty.$$

따라서 확률론의 중심극한 정리(Central Limit Theorem ; CLT)에 의하여 충분히 큰 n에 대해 통계치 $Z = 3\sqrt{n(n-1)}\,T / \sqrt{2n(2n-5)}$는 N(0, 1)(표준정규 분포)로 취급된다.

표준 정규 분포에 대한 확률분포 $Pr\{Z > z_\alpha\} = \alpha$는 표준정규분포표를 이용하여 찾을 수 있고 이때 α값은 유의 수준이 되는데 보통 0.05로 두고 검정한다.

2) Spearman의 등급 산관 계수 R (Spearman's Coefficient)

'1)'의 검정법에서와 같은 조건하에서 두 확률변수 X, Y의 표본상관계수(Sample Correlation Coefficient) R은 다음과 같이 정의된다.

$$R = \{\Sigma (X_i - X')(Y_i - Y')\} / \{\Sigma (X_i - Y')^2 \Sigma (Y_i - Y')^2\}^{1/2},$$
$$X' = \Sigma X_i / n, \quad Y' = \Sigma Y_i / n.$$

만약 $X_1, X_2, \ldots, X_n$과 $Y_1, Y_2, \ldots, Y_n$ 표본값이 각각

1부터 n의 증가하는 순서대로 등급(rank)이 매겨지고 $R_i = \text{rank}(X_i)$, $S_i = \text{rank}(Y_i)$로 두면,

$R_i, S_i \in \{1, 2, \ldots, n\}$, $\Sigma R_i = \Sigma S_i = n(n+1)/2$,
$R' = \Sigma R_i/n = (n+1)/2$. $S' = \Sigma S_i/n = (n+1)/2$,
$\Sigma(R_i - R')^2 = \Sigma(S_i - S')^2 = n(n^2-1)/12$,

따라서 $$R = 12\Sigma(R_i - R')(S_i - S')^2/(n^3-n)$$

한편, D_i를 $(R_i - R') - (S_i - S')$로 두면

$$\Sigma D_i^2 = \Sigma(R_i - R')^2 + \Sigma(S_i - S')^2 - 2\Sigma(R_i - R')(S_i - S')$$
$$= (1/6)n(n^2-1) - 2\Sigma(R_i - R')(S_i - S')$$

따라서 $R = 1 - \{6\Sigma D_i^2/(n(n^2-1))\}$을 얻고 이 통계량(Statistics) R을 Spearman의 등급 상관 계수라 한다(단 Σ은 i=1부터 n까지의 합을 의미한다). 한편 R의 평균값 ER은

$ER = (12/n^2-1)\ E(R_i S_i) - 3(n+1)/(n-1)$이고
H_0하에서는 X와 Y가 독립이므로 rank R_i와 S_i도 독립이고

$$E_{H0}R = 12(n+1/2)^2/(n^2-1) - 3(n+1)/(n-1) = 0$$

따라서 '1)'의 검정법에서와 마찬가지 방법으로 만약 $|R| > R_\alpha$면 H_0를 기각한다(단 $P_H\{|R| > R_\alpha\} \le \alpha$, α : 유의 수준)

한편 실제적인 R값을 구하기 위해 R의 Null Distribution을 알아야 한다. 일반성을 잃지않고 $R_i = i$이라고 가정하면($i=1, 2, \ldots, n$) $D_i = i - S_i$가 되고 H_0하에서는 X와 Y가 독립이므로 (i, S_i)의 n!개의 등급쌍(Rank Pair)은 Equally Likey되어서 다음이 성립한다.

$$P_{H0}\{R=r\} = (n!)^{-1} \times \{R=r\text{이 성립하는 등급쌍의 개수}\} = n_r/n!$$

한편 상관 계수 R은 $-1 \le R \le 1$을 만족하고, 만약 모든 $i=1, 2, \ldots, n$에 대해 $R_i = S_i$이면 $R=1$이고 $R_i = n+1-S_i$인 경우 $R=-1$이 된다.

$n \le 10$인 경우 Kendall에 의해 ΣD_i^2의 정확한 분포가 구해질수 있으나 n이 큰 값 일때는 통계치 R은 근사적 정규 분포를 따른다. 즉 충분히 큰 n에 대해 통계치 $Z = R\sqrt{n-1}$은 근사적으로 N(0, 1) (표준 정규분포)로 취급된다(중심극한정리).

한편 이러한 통계량 R을 Kendall의 방법과 마찬가지로 유의수준 $\alpha = 0.05$에 대해서 검정을 수행한다.

5. 시뮬레이션 결과 및 응용

앞절에서 설명된 통계적 검정법을 치환의 크기 n=64, 최종 치환 출력을 위한 발생 치환 결합의 크기 k=2인 경우에 적용시켜 보았다. 시뮬레이션은 PC386에서 서로 다른 5개의 초기값에 대해 각각 20,000개의 임의 치환을 발생시켜 적용하였으며 검정법 통과 개수의 평균값을 유의 수준 $\alpha = 0.05$에서 조사한 결과를 표 5-1에 수록하였다.

표 5-1. 검정 통과 개수 평균값

통계적 검정	통과 개수 평균값	임의성/독립성
도수-2 검정	18939	임의성
계열 검정	19001	임의성
연 검정	18963	임의성
Kendall 검정	19034	독립성
Spearman 검정	19025	독립성

검정 결과표에서 알 수 있듯이 발생된 치환들은 이진 난수 관점의 임의성 관점이나 연속 치환끼리의 독립성 관점 모두에 대해 통계적으로 우수함을 알 수 있다. 한편 이러한 랜덤 치환 발생 알고리즘은 앞에서 언급했듯이 여러가지 목적으로 응용될 수 있다. 특히 음성이나 영상의 시간 영역에서의 스크램블링 기법에서는 실시간적인 랜덤 치환 발생이 필수적으로 요구되어진다. 또한 몬테 카르로 시뮬레이션용의 난수 발생 알고리즘으로 흔히 이용되는 프로그램 내장 난수 알고리즘인 경우, 난수의 출현 빈도 관점에서만 랜덤성이 만족되므로 통신에서 이용되는 BER(Bit Error Rate)용 등으로 적용하기에는 현실성이 결여되어 있다. 따라서 랜덤 치환 발생 알고리즘을 이용하여 요구되는 BER 비트 패턴의 발생은 물론이고 특정한 분포를 가지는 난수 발생 알고리즘도 쉽게 구현할 수 있다.

참 고 문 헌

[1] S.G. Akl & H. Meijer, "A Fast Pseudo Random Permutation Generator with Application to Cryptology", Advanced in Cryptology : Proceedings of CRYPTO 84, pp.269-275, 1985.

[2] H.J. Beker & F.C. Piper, Secure Speech Communications, Academic Press, 1984.

[3] T.C. Hu, Combinatorial Algorithms, Addison-Wesely, 1982.

[4] G.D. Knott, "A Numbering System for Permutations of Combinations", Comm. of ACM, June Vol.19, No.6, pp.355-356, 1976.

[5] D.E. Knuth, The Art of Computer Programming, Vol.2 : Seminumerical Algorithms, Addison-Wesely, 1981.

[6] V.K. Rohatgi, An Introduction to Probability Theory and Mathematical Statistics, John Siley & Sons, 1976.

[7] N.J.A. Sloane, "Encrypting by Random Rotations", Advances in Cryptology, Proceedings of CRYPTO 82, pp.71-128, 1983.

[8] 현대 암호학, 한국전자통신연구소편, 1991.

[9] Randomness 특성 분석에 관한 연구, 한국전자통신연구소편, 1991.

[10] 고승철, 이경현, "랜덤 치환 고속 발생기" Proceedings of 1st Workshop in Applied Mathematics, pp.379-383, 1993.

□ 著者紹介

李 京 炫 (正會員)

1982. 2. 慶北大學校 師範大學 數學科 卒業(理學士)
1985. 2. 韓國科學技術院 應用數學科 卒業(理學碩士)
1992. 8. 韓國科學技術院 數學科 卒業(理學博士)
1985～1993.2. 韓國電子通信研究所 研究員, 先任研究員
1993.3.～現在 釜山水産大學校 電子計算學科 專任講師
관심분야 : 네트워크 성능분석, 암호이론, 암호알고리즘 설계