通信情報保護學會論文誌
第 8 卷 第 3 號, 1998. 9

유니버설 통계적 검정에서 표본 수열의 길이에 대한 분석

강 주성*, 박 상우*, 박 춘식*

On the Length of Sample Sequence in Universal Statistical Test

Ju-sung Kang*, Sang-woo Park*, Choon-sik Park*

요 약

Maurer가 제안한 유니버설 통계적 검정을 소개하고 검정에 사용된 통계량의 의미를 분석한다. 기존 검정법을 포괄하는 이 검정법은 보다 넓은 의미의 통계적 결점들을 탐지해낼 수 있다. 또한, 검정에 사용되는 통계량은 엔트로피와 밀접한 연관이 있으며 암호학적 응용에서 시스템의 안전성에 영향을 미치는 요소들을 탐지해 낸다. 이러한 특징과 함께 기존 검정법 보다 상당히 긴 표본 수열을 필요로 한다는 사실이 유니버설 검정의 단점으로 지적되어 왔다. 그러나 본 논문에서는 빈도(freequency) 검정법과의 비교를 통해서 미세한 편의(bias)를 탐지해내기 위한 도구로서는 유니버설 검정이 오히려 더 효율적이라는 사실을 보였다.

Abstract

We survey Maurer's universal statistical test and analyze the meaning of statistics which used in it. This test is able to detect any one of a general class of statistical defects and measures per-bit entropy which is related to cryptographic significance of a defect. A drawback of universal statistical test is that it requires a much longer sample sequence. But we shou that drawback is not a concern to detect a minute bias and universal test is efficient rather than frequency test.

Key words : Universal statistical test, Entropy, Ergodic stationary soruce, Length of sample

1. 서 론

난수 발생기는 확률적으로 서로 독립이면서 대칭적인 이진 확률 변수들의 수열을 생성하기 위한 장비이다. 이상적인 이러한 장비를 *BSS*(binary symmetric source)라고 부른다. 반

* 한국전자통신연구원

면, 우리가 사용하는 대부분의 난수 발생기는 *BSS*에 의해서 생성된 것처럼 보이는 이진 수열을 결정된 어떤 규칙에 의해서 발생시키는 의사 난수 발생기(pseudorandom bit generator)이다. 어떤 의사 난수 발생기가 생성하는 난수열은 완전한 의미의 랜덤 수열은 아니기 때문에 통계적 검정을 필요로 한다. 완전히 랜덤한 이상적인 수열들이 가지는 통계적 특성들을 조사하여 적절한 기준을 정한 다음에 의사 난수 발생기로부터 출력되는 수열에 대해서 이 기준을 적용하여 검정하는 것이다. 다양한 통계적 검정법이 있지만 지금까지는 빈도(frequency), 시리얼(serial), 포커(poker), 런(run), 자기상관관계(autocorrelation) 검정법 등이 통계적 검정에 주로 이용되어 왔다[2, 10].

본 논문에서는 기존 통계적 검정 방법을 포괄하면서 엔트로피와 암호학적 관점에서도 어떤 의미가 있는 것으로 알려진 새로운 통계적 검정 방법에 대한 분석을 실시하고자 한다. 유니버설 통계적 검정으로 불리는 이 검정 방법은 Maurer[9]가 1992년에 제안한 것으로서 그동안 널리 사용되어온 기존의 방법들에 비해서 다음과 같은 두가지 큰 장점을 지니고 있다. 첫째, 이 검정법은 매우 일반적인 의미의 통계적인 결점을 발견해낼 수 있다. 여기에서 일반적인 통계적 결점이란 것은 BSS를 포괄하는 난수 발생기인 정류(stationary) 에르고드(ergodic) 난수 생성기에 의해서 모델화 가능한 것으로서 다섯가지 기본 검정에 의해서 발견될 수 있는 모든 통계적 결점들을 포괄하는 것이다. 둘째, 유니버설 검정은 암호 시스템의 안전성에 영향을 미치는 실질적인 요소를 측정한다. 만일 검정 대상인 난수 발생기가 키 생성을 위해 사용된다면 이 검정 방법은 효율적인 키의 크기를 측정해 낼 수 있다.

위와 같은 장점과 함께 유니버설 검정은 기존 다섯가지 기본 검정법 보다 훨씬 긴 표본 출력 수열을 필요로 한다는 장애 요인을 안고 있다. 이렇게 검정에 요구되는 표본 수열의 길이가 길다는 것은 실제 응용상에서 상당한 불편함으로 작용하는 경우가 많다. 그렇지만 우리는 통계적 결점이 쉽게 보이지 않는 미세한 것일 때에는 Maurer의 유니버설 검정이 보다 더 효율적인 방법이 된다는 사실을 보인다. 각 출력 비트들이 어떤 편의(bias)를 갖을때 이 통계적 결점을 탐지하기 위해서 필요로 하는 표본 수열의 길이를 빈도 검정과 비교함으로써 이를 분석한 결과가 3절에 나타나 있다.

간결한 표현을 위해서 난수 발생기들을 다음과 같이 분류하자. 서로 독립이면서 동일 분포를 갖는 이진 수열을 생성하는 난수 발생기를 BMS(binary memoryless source), 서로 독립이고, 1을 발생시킬 확률이 p, 0을 발생시킬 확률이 $1-p$로 주어지는 난수 발생기를 BMS_p로 표시하자. 그리고 0과 1은 같은 확률로 발생시키면서 서로 다른 두 비트를 생성할 확률이 p, 서로 같은 두 비트를 발생시킬 확률은 $1-p$인 전이 확률을 갖는 난수 발생기를 ST_p(stationary source with p)라 하자.

본 논문은 모두 4개의 절로 구성된다. 2절에서는 Maurer의 유니버설 검정을 분석하고, 3절에서 일종의 통계적 결점인 편의를 탐지하기 위한 검정을 수행할때 필요로 하는 표본 수열의 길이에 관한 결과를 기술하며, 4절은 결론부이다.

2. Maurer의 유니버설 통계적 검정

Maurer의 유니버설 통계적 검정은 세개의 양의 정수 L, Q, K를 모수(parameter)로 갖는다. 먼저 표본 수열을 길이가 L인 서로 겹치지 않는 블럭들로 나눈다.

표본 수열 s^N의 전체 길이는 $N=(Q+K)L$이다. 이때 Q는 초기화 과정에 필요한 블럭들의 갯수이고, K는 실제로 검정에 사용되는 블럭

들의 갯수이다. $1 \le n \le Q+K$인 n에 대하여 n번째 블럭을

$$\beta_n(s^N) = \{s_{L(n-1)+1}, \cdots, s_{Ln}\}$$

으로 표시하기로 하자. 그리고 $n = Q+1, \cdots, Q+K$에 대하여

$\beta_n(s^N) = \beta_{n-i}(s^N)$인 $i(\le n)$가 존재하면

$$\nu_n(s^N) = min\{i \le n : \beta(s^N) = \beta_{n-i}(s^N)\}$$

으로 놓고, 1부터 n까지의 모든 정수 i에 대하여 $\beta_n(s^N) \neq \beta_{n-i}(s^N)$일때는 $\nu_n(s^N) = n$으로 정의하자. Maurer의 유니버설 통계적 검정에서 사용하는 검정 함수는

$$Mu(s^N) = \frac{1}{K} \sum_{n=Q+1}^{Q+K} log_2 \nu_n(s^N)$$

이다. $\nu_n(s^N)$는 n번째 블럭 $\beta_n(s^N)$의 가장 최근의 출력 위치를 탐지해내는 양이고, 검정 함수 Mu는 $\nu_{Q+1}(s^N), \cdots, \nu_{Q+K}(s^N)$ 로그 값들의 산술평균을 구한 것이다. $\nu_n(s^N)$의 계산은 $\beta_{n-1}(s^N)$, $\beta_{n-2}(s^N)$, … 을 모두 조사할 필요 없이 각각의 L-비트 블럭에 대하여 가장 최근의 발생 시점을 저장해 둠으로써 쉽게 해결할 수 있다.

검정 함수 Mu를 이용하여 통계적 랜덤성을 검정하기 위해서 생성기가 BSS일때를 고려하자. R^N을 BSS로부터 생성된 길이 N인 수열을 나타내는 확률 변수라고 할때, 통계량(statistics) $Mu(R^N)$의 평균과 분산은 각각

$$\mu = E[Mu(R^N)] = E[log_2 \nu_n(R^N)],$$
$$\sigma^2 = Var[Mu(R^N)] = c(L,K)^2 \frac{Var[log_2 \nu_n(R^N)]}{K}$$

으로 표현할 수 있다. $K \ge 2^L$에 대해서

$$c(L, K) \approx 0.7 - \frac{0.8}{L} + \left(1.6 + \frac{12.8}{L}\right) K^{-\frac{4}{L}}$$

이고, $L \ge 3$일때 $c(L, 2^L)$은 0.8에 매우 가깝고, $K \gg 2^L$일때 $c(L, K)$는 $L=4$인 경우 0.5에, $L=8$인 경우 0.6에, 그리고 $L=12$인 경우 0.65에 가까운 값이다. 확률론의 중심 극한 정리(central limit theorem)에 의하면 $N \to \infty$일때

$$\frac{Mu(R^N) - \mu}{\sigma} \xrightarrow{d} N(0,1)$$

이므로 이사실을 이용하여 주어진 유의수준하에서 정규(normal) 검정을 실시할 수 있다.

검정을 실행하기 위해서 Maurer는 $6 \le L \le 16$, $Q \ge 10 \cdot 2^L$, 그리고 K는 가능한한 크게 할 것을 권고한다.

예를들면 $K = 1000 \cdot 2^L$ 정도가 적당하다. Q에 대한 조건은 처음 Q개의 블럭들에서 높은 확률로 각각의 L-비트 패턴이 적어도 한번은 나오도록 보장하는 것인데 이 확률은

$$1 - \left(\frac{2^L - 1}{2^L}\right)^Q \qquad (2.1)$$

이 된다. $Q = 10 \cdot 2^L$일때 (2.1)의 값은 $L=4$인 경우 0.999967, $L=8$인 경우 0.999955로 거의 1에 가까운 값이다. $1 \le L \le 16$인 L에 대해서 $Mu(R^N)$의 기대값과 $log_2 \nu_n(R^N)$의 분산을 $Q \to \infty$라는 가정하에 근사적으로 구한 값들이 표1에 나타나 있다.

표 1: $Mu(R^N)$의 기대값 및 $log_2 \nu_n(R^N)$의 분산

L	$E[Mu(R^N)]$	$Var[log_2 \nu_n(R^N)]$
1	0.7326495	0.690
2	1.5374383	1.338
3	2.4016068	1.901
4	3.3112247	2.358
5	4.2534266	2.705
6	5.2177052	2.954
7	6.1962507	3.125
8	7.1836656	3.238
9	8.1764248	3.311
10	9.1723243	3.356
11	10.170032	3.384
12	11.168765	3.401
13	12.168070	3.410
14	13.167693	3.416
15	14.167488	3.419
16	15.167379	3.421

한편, 생성기 BMS_p의 길이 N인 출력 수열을 나타내는 확률 변수를 U^N_{BMSp}로 나타내기로 하면

$$\lim_{L\to\infty}(E[M_u(U^N_{BMSp})]-LH(p))=C=-0.83276$$

이 성립하여 엔트로피와 통계량 $M_u(U^N_{BMSp})$ 사이에는

$$E[M_u(U^N_{BMSp})]\approx L\cdot H(p)+C \qquad (2.2)$$

라는 관계식이 만족된다. 여기에서 엔트로피 함수 $H(x)$는 $0<x<1$인 x에 대해서

$$H(x)=-x\log_2 x-(1-x)\log_2(1-x)$$

와 $H(0)=H(1)=0$으로 정의된다. 식 (2.2)는 Maurer의 유니버설 검정이 BMS의 엔트로피를 측정한다는 사실을 나타내 주고 있다.

$L=8$과 $L=16$일때 $E[M_u(U^N_{BMSp})]$, $LH(p)+C$, $Var[\log_2 v_n(U^N_{BMSp})]$를 몇가지 p값에 대해서 계산한 결과를 나타내면 표 2와 같다.

표 2: L=8, 16일때 $E[M_u(U^N_{BMSp})]$, $LH(p)+C$, $Var[\log_2 v_n(U^N_{BMSp})]$

L	p	$E[M_u(U^N_{BMSp})]$	$LH(p)+C$	$Var[\log_2 v_n(U^N_{BMSp})]$
8	0.50	7.18367	7.16725	3.239
8	0.45	7.12687	7.10945	3.393
8	0.40	6.95559	6.93486	3.844
8	0.35	6.66713	6.63980	4.561
8	0.30	6.25683	6.21758	5.482
16	0.50	15.16738	15.16725	3.421
16	0.45	15.05179	15.05165	3.753
16	0.40	14.70268	14.70246	4.733
16	0.35	14.11275	14.11234	6.319
16	0.30	13.26886	13.26791	8.425

Willems[11]가 사용했던 방법과 비슷한 논리를 이용하면 ST_p를 일반화한 임의의 정류 에르고드 생성기(stationary ergodic generator) G에 대해서

$$\lim_{L\to\infty}\frac{E[M_u(U^N_G)]}{L}=H_G \qquad (2.3)$$

임을 보일 수 있다. 여기에서 U^N_G는 G에 의해서 생성되는 길이 N인 이진 수열을 나타내는 확률 변수이고, H_G는 $H(p)$와 비슷하게 정의된 생성기 G의 비트당 엔트로피를 나타낸다. 정류 에르고드 마르코프 연쇄의 엔트로피에 대한 정의는[1]을 참조하면 된다. 식 (2.3)은 유니버설 검정에 사용하는 검정함수 M_u가 정류 에르고드 생성기에 대해서도 비트당 엔트로피를 측정해낼 수 있음을 보여주는 것이다. (2.2)와 (2.3)에서 보면 $L\to\infty$라는 조건이 있어서 이 관계식들은 L이 상당히 큰 경우에만 성립하는 것으로 오해할 수 있다. 그러나 표 2에서 엿볼 수 있는 것처럼 $L=8$만 되어도 그 수렴 오차는 아주 작게 된다. 그러므로 Maurer[9]가 추천한 $6\le L\le 16$에 대해서 위의 극한식들은 모두 동치식으로 받아들여도 큰 오차는 발생하지 않는다.

지금까지 우리는 유니버설 검정에서 사용하는 통계량이 엔트로피와 밀접한 관계가 있다는 사실에 대해서 논했다. 이제, 엔트로피와 어떤 암호 시스템의 안전성과의 관계를 규명

함으로써 궁극적으로 유니버설 검정함수 M_U가 암호 시스템의 안전성에 영향을 미치는 요소를 측정해낼 수 있음을 살펴보자.

좋은 암호 시스템이란 전수 조사(exhaustive key search)보다 본질적으로 빠른 공격 방법이 없도록 설계된 것이다. 그러나 키 생성기가 가능한 모든 키값을 똑같은 확률로 생성하지 않는다면 공격자는 확률이 높은 키값들부터 조사함으로써 좀 더 효율적인 공격을 할 수 있을 것이다.

이러한 관점에서 통계적으로 어떤 결함이 있는 키 생성기를 소유한 암호 시스템의 효율적인 키의 크기에 관해서 논해보자.

KY를 n-비트 크기의 랜덤한 비밀키라 하고,

$$P(KY=z_1)\geq P(KY=z_2)\geq\cdots\geq P(KY=z_{2^n})$$

을 만족하는 $z_1, z_2, \cdots, z_{2^n}$을 키의 목록이라고 하자. 주어진 난수 발생기 G에 대하여 공격자가 가장 효율적인 방법으로 키를 찾는다고 가정할 때, 적어도 δ의 확률로 공격을 성공시키기 위해서 시도해야만 하는 최소의 키 갯수 $\mu_G(n, \delta)$를 다음과 같이 정의한다.

$$\mu_G(n, \delta) = min\left\{k: \sum_{i=1}^{k} P(KY=z_i)\geq\delta\right\},\ 0\leq\delta\leq 1.$$

키생성기를 G로 사용하는 암호 시스템의 효율적인 키의 크기는

$$log_2\ \mu_G(n, 1/2)$$

로 정의하는데, 이것은 적어도 50%의 확률로 올바른 키를 찾고자할 때 공격자가 알고있어야할 비트수를 의미한다. G가 BSS라면 효율적인 키의 크기는 n-1이 됨을 쉽게 알 수 있다.

BMS_P가 주어졌을때 새로운 난수 발생기의 n번째 비트를 BMS_P에서 생성되는 처음 n개의 비트들을 모듈러 2로 모두 더한 값으로 정의하면 이 새로운 난수 발생기는 근사적으로 ST_P가 된다. 즉, ST_P는 항상 BMS_P에 의해 만들어질 수 있다. 그러므로 두 발생기는 키 확률들의 집합이 동일하여 효율적인 키의 크기 역시 같게 된다. 실제로 Maurer[9]는 BMS_P를 키 생성기로 갖는 암호 시스템에 대해서 $0<\delta<1$일때,

$$\lim_{n\to\infty}\frac{log_2\mu_{BMSp}(n, \delta)}{n} = H(p) \qquad (2.4)$$

라는 사실을 보였으며, ST_P를 일반화한 에르고드 정류 생성기 G를 키 생성기로 사용하는 암호 시스템에 있어서도 비슷한 관계식

$$\lim_{n\to\infty}\frac{log_2\mu_G(n, \delta)}{n} = H_G \qquad (2.5)$$

가 성립함을 보였다. 식 (2.4)와 (2.5)는 효율적인 키의 크기가 엔트로피와 밀접한 관계가 있다는 사실을 잘 보여주고 있다. 결과적으로 유니버설 검정 함수 M_U는 (2.2)와 (2.3)에 의해서 엔트로피와 관계가 있고, 엔트로피는 (2.4)와 (2.5)를 보면 효율적인 키의 크기와 연관되어 있으므로 유니버설 검정은 암호 시스템의 안전성에 영향을 끼치는 요소와 엔트로피를 동시에 측정해낼 수 있음을 알 수 있다.

3. 표본 수열의 길이

유니버설 검정의 단점중 하나는 다른 통계적 검정법에 비해서 상대적으로 긴 표본 수열을 필요로 한다는 것이다. 2절에서 본 바와 같이 L=8인 경우 표본수열의 길이는 약 200만 비트가 되어야 한다. 유의수준이 다르기는 하지만 FIPS 140-1[10]에서 적용하는 2만 비트와 비교하면 약 100배 정도의 표본 수열이 더 필요하다는 뜻이므로 상당한 차이라고 볼 수 있다. 그러나 똑같은 유의수준 하에서같은 종류의 통계적 결점을 탐지해내기 위해 필요로 하는 표본 수열의 길이에 대해서도 유니버설 검정이 항상 더 긴것을 요구한다고 결론 지을 수는 없다. 본절에서는 기본 검정법 중에서 가장 단순한 빈도 검정과 유니버설 검정을 비교해 봄으로써 필요로 하는 표본 수열의 길이를 정량적으로 나타내고자 한다.

표본 수열의 길이를 N이라 하고, 검정의 유의 수준을 α라 하자. Z가 표준 정규 분포를 따르는 확률 변수라 할때 문턱치(threshold) t_α를 $P(|Z|>t_\alpha)=\alpha$를 만족하는 값으로 정의하자. Maurer의 유니버설 검정은 $0.001\le\alpha\le0.01$인 유의수준 α를 추천하고 있으며, $t_{0.01}=2.58$이고 $t_{0.001}=3.27$이다. 빈도 검정에 정규(normal) 검정을 적용할때 사용되는 통계량은

$$X=\frac{2n_1-N}{\sqrt{N}}$$

이다. 여기에서 n_1은 표본 수열 내의 1의 갯수이다. 표본 수열 내의 각 비트가 1이 될 확률을 p라 하고, $p=\frac{1}{2}\pm\gamma(0<\gamma<\frac{1}{2})$일때 유의수준 α로 통계적 결점 γ를 탐지하는데 필요한 표본 수열의 길이를 조사하자. 이 경우 $n_1=Np=N(\frac{1}{2}\pm\gamma)$이기 때문에

$$|X|=2\gamma\sqrt{N}\ge t_\alpha$$

일때 γ를 탐지해낼 수 있으므로 표본 수열의 길이 N은

$$N\ge\frac{t_\alpha^2}{4\gamma^2} \qquad (3.1)$$

을 만족해야만 한다.

한편, Maurer의 유니버설 검정에서 통계량 $M_U(R^N)$의 평균 μ와 분산 σ^2은 표 1을 참조하여 얻을 수 있다. $\sigma^2=Var[\log_2 v_n(R^N)]$일때

$$\sigma^2=c(L,K)^2\cdot\frac{\sigma^2}{K}$$

이다. 표 2를 참조하면 $L=8$, $p=0.45$인 경우에 $E[M_U(U^N_{BMSp})]=7.12687$이고 $\sigma^2=3.238$이므로 $c(8,K)=0.6$으로 놓으면

$$\frac{|E[M_U(U^N_{BMSp})]-\mu|}{\sigma}\approx0.05261\cdot\sqrt{K} \qquad (3.2)$$

이 된다. 그러므로 $\gamma=0.5-0.45=0.05$를 탐지하기 위해서는

$$K\ge\left(\frac{t_\alpha}{0.05261}\right)^2 \qquad (3.3)$$

을 만족해야 한다. 유의수준 $\alpha=0.01$인 경우에 식 (3.3)에 의하면 $K\ge2405$이다. Maurer의 유니버설 검정에서 $N=(Q+K)L$이지만 Q는 검정 대상에 포함되지 않는 초기 블럭들의 갯수이므로 K에 비해서 상당히 작은 것으로 생각하면 이 경우에 $\gamma=0.05$를 탐지하기 위한 최소의 표본 수열의 길이는 $8\cdot2405=19240$이 된다. 한편, 식 (3.1)에 $\gamma=0.05$와 $\alpha=0.01$을 대입하면 $N\ge666$이다. 빈도 검정으로 편의(bias) $\gamma=0.05$를 탐지하기 위해서는 최소한 666 비트 길이의 표본 수열이 필요하다는 뜻이다. 위의 결과를 종합하면 유의 수준 0.01로 편의 0.05를 탐지하는데 필요로 하는 이론적인 표본 수열의 최소 길이는 유니버설 검정이 빈도(frequency) 검정을 적용할때보다 29배 정도 더 길어야 한다는 것이다. 그러나 편의 γ가 아주 작은 경우에도 표본 수열의 길이에 있어서 유니버설 검정이 빈도 검정보다 항상 비효율적인 것만은 아니다.

유니버설 검정인 경우에는 식 (2.2)에서 보는 바와 같이 엔트로피 함수 $H(p)$를 계산함으로써 $M_U(U^N_{BMSp})$의 기대값을 근사적으로 구할 수 있다.

표 2를 참조하면 $E[M_U(U^N_{BMSp})]$와 $LH(p)+C$ 사이에 $L=8$인 경우 약 0.002의 오차가 발생한다. 그러므로

$$E[M_U(U^N_{BMSp})]=8H(p)-0.852746 \qquad (3.4)$$

으로 놓고 통계량 $M_U(U^N_{BMSp})$의 기대값을 구하는 것이 타당하다. 편의 γ가 0.1, 0.05인 경우에는 표 2를 참조하고, $\gamma=0.01$, 0.005, 0.001인 경우에는 (3.4)에 의해서 통계량의 기대값을 구함으로써 검정에 필요로 하는 블럭의 갯수 K를 (3.2)와 (3.3)을 얻어낸 과정과 똑같은 과정을 통해서 구할 수 있다. $L=8$인 경우 $Q\ge10\cdot2^8=2560$이 요구되므로 이 블럭들의 갯수도 표본 수열의 길이에 포함시켜야 한다.

위와 같은 요소들을 모두 고려하여 편의

$0.001 \le \gamma \le 0.1$인 몇몇 값들에 대해서 유의수준 $\alpha = 0.01$ 하에서 각 편의들을 탐지하기 위해서 필요로 하는 표본 수열의 최소 길이를 구한 결과가 표 3에 나타나 있다. 빈도 검정은 식 (3.1)에 의해서 얻은 결과이고, 유니버설 검정의 경우에는 (3.4)에 의해서 먼저 통계량의 기대값을 구하고 (3.3)을 얻어낸 과정과 동일한 방법에 의해서 K를 계산한 후에 $N = 8(Q+K)$로 표본 수열의 길이를 구한 것이다. 여기에서 $Q = 10 \cdot 2^8 = 2560$으로 놓았다.

표 3:편의 (bias) γ를 탐지하기 위한 표본 수열의 길이

편의(bias) γ	표본 수열의 길이	
	빈도(frequency)	Maurer의 유니버설 검정
0.10	167	21,673
0.05	666	39,720
0.01	16,641	61,866
0.005	66,564	65,842
0.001	1,664,100	67,227

표 3의 결과를 잘 관찰해 보면 γ가 큰 값일 때는 빈도 검정이 유니버설 검정보다 효율적이지만 γ가 0.005보다 작으면 유니버설 검정이 더 효율적이라는 사실을 알 수 있다. 이런 현상이 나타나는 이유는 빈도 검정에서 필요로 하는 표본 수열의 길이는 (3.1)에 의해서 γ의 제곱 값에 반비례하여 γ가 작아지면 길이 N은 급격히 증가하지만 유니버설 검정의 경우 N은 엔트로피 $H(p)$에 종속되고, $H(p)$값은 $p = 0.5$ 근처에서 완만하게 변하므로 표본 수열의 길이 역시 급격히 변하지는 않기 때문이다. 그러므로 미세한 편의를 탐지하는데는 표본 수열의 길이 관점에서 Maurer의 유니버설 검정이 보다 더 좋은 검정 도구라고 볼 수 있다.

4. 결 론

통계적 검정의 수행은 그 검정의 기초가 되는 통계적 모델에 대단히 많이 종속되므로 모델이 일반적이라면 탐지할 수 있는 결점의 영역도 넓어지게 된다. 한편, 모델이 제한되면 될수록 이 모델에 기초를 둔 검정은 어떤 의미에서 더 편리해진다고 볼 수 있다. 즉, 통계적 결점을 탐지해 내기 위해서 보다 더 짧은 길이의 표본 수열이 필요하게 되는 측면이 있다. 이와 같은 이유로 통계적 검정을 디자인할 때 적절한 모델을 사용하는 것은 매우 중요한 일이다.

Maurer가 제안한 유니버설 통계적 검정은 2절에서 살펴본 것처럼 다음과 같은 두가지 큰 장점을 갖고 있다. 첫째, 유니버설 통계적 검정은 매우 일반적인 영역의 통계적 결점들을 탐지할 수 있게 만들어진 것이다. 이 일반적인 영역의 통계적 결점들이란 BMS_p와 ST_p를 포괄하는 에르고드 정류 생성기에 의해서 모델화될 수 있는 것이므로 이전의 다섯가지 기본 검정에 의해서 탐지될 수 있는 모든 결점들을 포함한다. 둘째, 유니버설 검정은 암호 시스템의 안전성에 영향을 끼치는 실질적인 양을 측정한다. 만일 검정 대상인 생성기 G가 키 생성기로 사용된다면 이 검정은 효율적인 키의 크기 $\mu_G(n, 1/2)$을 측정한다.

난수 발생기가 $p = 0.45$인 BMS_p에 의해서 모델화된 것이라면 비트당 엔트로피 $H(0.45) = 0.9928$가 되어 거의 1에 가깝기 때문에 이경우 빈도 검정과 똑같은 탐지 확률로 비랜덤성을 가려내기 위해서 유니버설 검정은 훨씬 긴 표본 수열이 필요하다. 예를들어 우리는 $L = 8$인 경우 표본 수열의 길이는 29배 더 길어야 한다는 것을 3절에서 볼 수 있었다. 그러나 p가 거의 0.5에 가까워서 편의가 미세한 양일때는 엔트로피의 변화가 작아서 유니버설 검정의 경우 필요한 표본 수열의 길이가 급격히 늘어나지 않는 반면, 빈도 검정에서는 편의가 작아지면 필요로 하는 표본 수열의 길이가 급격히 증가하게 된다.

실제로 우리는 3절에서 편의가 0.005 이하일

때는 유니버설 검정이 더 효율적이라는 사실을 보였다.

결론적으로 Maurer가 제안한 유니버설 검정은 기본적으로 요구하는 표본 수열이 상당히 길다는 단점에도 불구하고 의사 난수 발생기의 통계적 특성을 심층적으로 평가하는 도구로서는 여러가지 측면에서 우수한 것으로 생각된다. 특히, 엔트로피와 암호학적인 관점에서 통계적 결점의 의미를 파악할 수 있다는 점과 실제 응용시 구현이 용이하고 계산이 효율적인 단 하나의 통계량을 고려하면 된다는 사실은 유니버설 검정의 우수성을 대표하는 것이라고 하겠다.

참고문헌

[1] N. Abramson, *information theory and coding*, New York, McGraw-Hill, 1963.

[2] H. Beker and F.Piper, Cipher systems, London: Northwood Books, 1982.

[3] P. Elias, Interval and recency rank source coding:Two on-line adaptive variable-length schemes, *IEEE Transactions on Information Theory*, vol. 33, Jan. 1987, pp. 3-10.

[4] W. Feller, An Introduction to Probability Theory and Its Applications, 3rd edn, vol. 1, New York, Wiley, 1968.

[5] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*, Oxford, Clarendon Press, 1982.

[6] D. E. Knuth, *The Art of Computer Programing*, vol. 2, 2nd edn., Reading, MA, Addison Wesley, 1981.

[7] A. N. Kolmogrov, Three approaches to the quantitive definition of information, *Problemy Peredachi Informatsii*, vol. 1, no. 1. 1965, pp 3-11

[8] J. L. Massey, An introduction to conremporary cryptology, *Proceedings of the IEEE*, vol. 76, no. 5, 1988, pp 533-549.

[9] U. M. Maurer, A universal statistical test for random bit generators, *J. Cryptology*, vol 5, 1992. pp. 89-105

[10] A. J. Menezes. P. C. Oorschot and S. A. Vanston, *Handbool of Applied Cryptography*, CRC Press, 1997

[11] F. M. Willems, Universal data compression and repetition times, *IEEE Transactions on Information Theory*, vol. 35, Jan. 1989, pp. 54-58

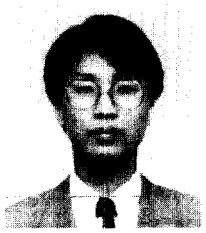
[12] J. M. Wzencraft and B. Reiffen, Sequential Decoding, Cambridge, MA, Technical Press of the MIT, 1960.

[13] J. Ziv and A. Lempel, A universal algorithm for sequential data compression, *IEEE Transactions on Inforamtion Theory*, vol. 23. May 1977, pp. 337-343.

□ 著者紹介

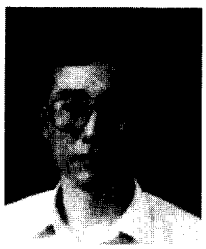
강 주 성

1989년 2월 고려대학교 이과대학 수학과(학사)
1991년 2월 고려대학교 대학원 수학과(이학석사)
1996년 2월 고려대학교 대학원 수학과(이학박사)
1997년 12월 - 현재 한국전자통신연구원 선임연구원



박 상 우

1989년 2월 고려대학교 사범대학 수학교육학과(이학사)
1991년 2월 고려대학교 대학원 수학과(이학석사:응용수학 및 확률론)
1995년 2월 - 현재 한국전자통신연구원 연구원



박 춘 식

광운대학교 전자통신과 졸업(학사)
한양대학교 대학원 전자통신과 졸업(석사)
일본 동경공업대학 전기전자공학 졸업(암호학 전공, 공학박사)
1989년 10월 ~ 1990년 9월 일본 동경공업대학 객원 연구원
1982년 - 현재 한국전자통신연구원 책임연구원
1997년 한국통신정보보호학회 편집이사, 종신회원

※ 주관심분야 : 암호이론, 정보이론, 통신이론