通信情報保護學會誌
第2卷 第4號, 1992.12

난수의 다차원 간격검정

최봉대*, 신양우**, 이경현***

1. 서 론

간격검정(gap test)은 주어진 의사 난수열의 독립성을 검정하기 위하여 많이 사용되는 검정법중의 하나이다. [0,1]수열 (U_n)에 대한 일차원 간격검정에 대하여 간략하게 기술해 보면 다음과 같다.[4),5),7)] 임의의 주어진 구간 $(\alpha,\beta), 0 \leq \alpha < \beta \leq 1$에 대하여 연속적인 부분수열 U_j, U_{j+1}, …, U_{j+r}로써 $U_{j-1} \in [\alpha,\beta)$, $U_{j+r} \in [\alpha,\beta)$이고 다른 U_i들은 $[\alpha,\beta)$에 속하지 않을 때 이 연속적 부분수열의 간격의 길이를 r이라고 한다. 만일 (U_i)가 서로 독립이고 [0,1] 상에서 일양 분포를 따른다면 간격의 길이의 분포는 모수가 $P(\alpha < U_1 < \beta) = \beta - \alpha$인 기하분포를 따른다. 수열 (U_i)가 서로 독립이라면 연속적으로 정의된 간격의 길이들도 서로 독립이므로 경험적 분포와 기하분포를 비교하여 χ^2-검정을 행하는 것이 1차원 간격검정이다. $\alpha=0$, $\beta=\frac{1}{2}$ 또는 $\alpha=\frac{1}{2}$, $\beta=1$인 경우는 보통 runs above the mean 또는 runs below the mean이라고 알려져있다. 이진수열에 대한 간격검정은 "0" run의 길이와 "1" run의 길이를 이용하여 시행하며 run검정으로 알려져 있기도 하다.[1)] 일반적으로 난수열을 이용한 암호계에서 암호화할 때 수열의 block으로 이루어진 vector열을 사용하는 경우가 많다. 그러므로 다차원 공간상에서의 수열의 randomness는 매우 중요한 문제라 하겠다. 이 논문에서는 d-진수열에 대한 다차원 간격검정법을 개발하고 이를 이용하여 [0,1]-수열의 다차원 간격검정을 수행하는 방법을 제시한다. 2절에서는 d-진수열의 다차원 검정법을 제시하고 3절에서는 [0,1]-수열에 대한 다차원 간격검정을 제시한다. 끝으로 4절에서는 기존의 몇몇 난수생성자에 대한 모의실험을 행한다.

2. d-진 수열의 다차원 검정법

2.1. 하나의 벡터의 출현시간 간격을 이용한 검정법

2.1.1. 기본이론

$X_1, X_2, \cdots$는 $S=\{0,1,2,\cdots d-1\}$상에서 균등분포를 따르는 $i.i.d.$인 확률변수열이라 하자. 즉 $\{X_n\}$의 분포는 $P(X_n=i)=\frac{1}{d}$, $i=0,1,2,\cdots d-1$이다.

먼저 $\{X_n\}$으로 구성된 연속벡터열 $Z_n=(X_n, X_{n+1}, \cdots, X_{n+k-1})$, $n=1,2,\cdots$이 벡터 $U=(u_1,u_2,\cdots,u_k)$, $(u_i \in \{0,1,\cdots,d-1\})$에 도달하는 시간간격 $S_i(X,U)$의 분포를 알아보자. 여기서 $S_i(X,U)$은

* 한국과학기술원 수학과
** 창원대학교 통계학과
*** 한국전자통신연구소

은 다음과 같이 정의한다.

$$S_0=0, \tag{2.1}$$

$S_j(X,U)=\min\{n|Z_{S_{j-i+n}}=U\}+k-1$, $j=1,2,\cdots$.

예를 들어 X_i가 연속해서 주사위를 던질때 i번째 나오는 눈이라하자. 그러면 X_i가 취할 수 있는 값은 {1,2,3,4,5,6}중에 하나이다. $U=(1,3,4,2)$가 주어진 벡터라고 할때 U가 나올 때 까지 주사위를 몇번 던져야 하겠는가? 만일 계속해서 주사위를 던졌을 때 그 결과가

6324134252432413426113

이었다면 $S_1(X,U)=8$, $S_2(X,U)=10$이다.

$S_j(X,U)$의 정의로 부터 우리는 다음의 결과를 쉽게 알 수 있다.

(1) $S_j(X,U)\geq k$

(2) $S_j(X,U)$, $j=1,2,\cdots$, 는 독립이고 같은 분포를 따른다.

$S_1(X,U)$의 확률질량함수 $p_i=P(S_1(X,U)=i)$, $i=1,2,\cdots$를 구하기 위하여 다음의 몇가지 용어를 도입하자.

정의 2.1. (Leading Numbers)

$$\varepsilon_j=\begin{cases}1, & \text{if } u_{k-j+i}=u_i,\ i=1,2,\cdots,j,\\ 0, & \text{otherwise.}\end{cases}$$

Leading number $\varepsilon_j(U)$는 벡터 U의 끝부분과 U의 앞부분이 j만큼 겹쳐있는가를 나타낸다. 예를 들어 이진수로 이루어진 벡터 $U=(00100)$에서 $\varepsilon_1(U)=1$, $\varepsilon_2(U)=1$, $\varepsilon_3(U)=0$, $\varepsilon_4(U)=0$, $\varepsilon_5(U)=1$이다. 정의 2.2를 이용하여 다음의 기호를 정의하자.

$$q_r(U)=\varepsilon_{k-r}(U)d^{-r},\ r=0,1,\cdots,k-1.$$

$q_r(U)$는 U로 시작하는 난수열이 주어져있다는 가정하에서 $r+1$과 $r+k$번째 사이에 U가 있을 조건부 확률이다.

정리 2.1. $S_1(X,U)$의 확률질량함수 $p_i=P(S_1(X,U)=i)$, $i=1,2,\cdots$은 다음과 같이 반복적인 형태로 얻어진다.

$$p_i=\begin{cases}0 & i=1,2,\cdots,k-1.\\ d^{-k}-d^{-k}\sum_{j=k}^{i-k}p_j-\sum_{j=i-k+1}^{i-k}p_jq_{i-j}(U), & i=k,k+1,\cdots.\end{cases} \tag{2.2}$$

증명. 벡터 U가 정확하게 i번째 시행에서 일어날 사건 E를 생각하자. 이 때 i번째 시행에서 처음으로 U가 일어날 필요는 없다. 이 사건의 확률은 d^{-k}이고 다음 세가지의 배반적인 사건 E_1, E_2, E_3로 나누어진다.

사건 E_1. U가 i번째만에 처음으로 일어난다.

사건 E_2. U가 j번째만에 처음으로 일어나고 그 뒤로 i번째도 일어나며, 두번의 U가 일어나는 시간간격은 k이상이다. 이 경우 두개의 U는 서로 독립적으로 일어난다. 그러므로 이 경우의 확률은 p_jd^{-k}가 된다.

사건 E_3. U가 j번째만에 처음으로 일어나고 그 뒤로 i번째도 일어나며, 두번의 U가 일어나는 시간간격은 k이하이다. 그러므로 이 경우의 확률은 $p_jq_{i-j}(U)$가 된다.

세사건 E_1, E_2, E_3는 서로 소이고, $E=E_1\cup E_2\cup E_3$ 이므로

$$d^{-k}=p_i+d^{-k}\sum_{j=k}^{i-k}p_j+\sum_{j=i-k+1}^{i-k}p_jq_{i-j}(U). \quad \square$$

예를 들어 정상적인 동전을 던진다고 할때 연속해서 세번의 앞부분(HHH)이 나올 때까지 시행할 횟수의 확률 p_i, $i=3,4,\cdots,11$는 각각 $\frac{1}{8}$, $\frac{1}{16}$, $\frac{1}{16}$, $\frac{1}{16}$, $\frac{7}{128}$, $\frac{13}{256}$, $\frac{12}{256}$, $\frac{11}{256}$, $\frac{81}{2048}$이 된다.

정리 2.2. $S_1(X,U)$의 확률생성함수는 다음과 같다.

$$\phi(t)=\frac{1}{1+(1-t)\eta\left(\frac{d}{t}\right)}$$

이다. 여기서 $\eta(x)=\Sigma_{i=1}^{k}\varepsilon_i(U)x^i$은 $\varepsilon_i(U)$의 생성함수이다.

증명. 식(2.2)를 다시 정리하면 다음과 같다.

$$d^k p_i=\sum_{j=i-k+1}^{\infty}p_j-d^k\sum_{j=i-k+1}^{i-1}p_jq_{i-j}(U).$$

위 식의 양 변에 t^i를 곱하고 나서 i에 대하여 합하면

$$d^k\phi(t)=\sum_{j=1}^{\infty}\sum_{i=k}^{k+j-1}p_jt^i-d^k\sum_{i=k}^{\infty}\sum_{j=i-k+1}^{i-1}p_jq_{i-j}(U)t^i$$

이 된다. 이것은 다시 다음과 같은 형태로 쓸 수 있다.

$$d^k\phi(t)=\frac{t^k}{1-t}\sum_{j=1}^{\infty}p_j(1-t^j)-d^k\sum_{j=k}^{\infty}t^jp_j\sum_{s=1}^{k-1}t^sq_s(U)$$
$$=\frac{t^k}{1-t}(1-\phi(t))-\phi(t)d^k\sum_{i=1}^{k-1}t^i\varepsilon_{k-i}(U)d^{-i}.$$

위 식을 $\phi(t)$에 대하여 풀면 정리의 결과를 얻는다. □

따름정리 2.3. $S_1(X,U)$의 평균과 분산은 다음과 같다.

$$E(S_1(X,U))=\eta(d),$$
$$\mathrm{Var}(S_1(X,U))=\eta(d)^2+\eta(d)-2d\eta'(d).$$

여기서 $\eta'(x)=\Sigma_{i=1}^{k}i\varepsilon_i(U)x^{i-1}$이다.

2.1.2. 검정수행 방법

이 절에서는 앞절에서 얻어진 결과를 이용하여 주어진 수열에 대한 간격검정을 수행하는 방법을 제시한다. $\{x_n\}$을 길이가 N인 d-진 수열이라하고 $U=(u_1,u_2,\cdots,u_k)$, $u_j\in\{0,1,\cdots,d-1\}$를 고정된 벡터라하자. 이때 검정을 수행하는 과정은 다음과 같다.

① 간격의 길이에 대한 자료를 수집한다. 즉,

$$g_n(U)=\begin{cases}\Sigma_j 1(S_j(x,U)=n), & n=k,k+1,\cdots,k+t-1,\\ \Sigma_j 1((x,U)\geq k+t), & n=k+t.\end{cases}$$

여기서 $1(D)$는 D의 지표함수(indicator function)이다.

② 간격의 총 개수 $G(U)$를 계산한다.

$$G(U)=\sum_{n=k}^{k+t}g_n(U).$$

③ $g_n(U)$의 기대값 $G_n(U)$를 계산한다.

$$G_n(U)=\begin{cases}G(U)P(S_1(X,U)=n), & n=k,k+1,\\ & \cdots,k+t-1,\\ G(U)P(S_1(X,U)\geq k+t), & n=k+t.\end{cases}$$

④ 검정통계량을 계산한다.

$$T_1(U)=\sum_{n=k}^{t+k}\frac{(g_n(U)-G_n(U))^2}{G_n(U)},$$
$$T_2(U)=\frac{\bar{S}(x,U)-E(S_1(X,U))}{\sqrt{\frac{\mathrm{Var}(S_1(X,U))}{G(U)}}}. \qquad (2.3)$$

여기서 $\bar{S}(x,U)=\Sigma_{j=1}S_j(x,U)/G(U)$.

$T_1(U)$는 근사적으로 자유도가 t인 χ^2분포를 따른다. $\chi^2_{\alpha,t}$를 자유도가 t인 χ^2분포의 상위 $100\alpha\%$ 백분위수라 하자. $T_1(U)$를 이용하여 검정을 할 때 유의수준 α의 기각역은 $T_1(U)>\chi^2_{\alpha,t}$이다. 또한 $T_2(U)$는 근사적으로 표준정규분포를 따르므로 $T_2(U)$를 이용하여 검정을 할 때 유의수 α의 기각역은 $|T_2(U)|>z_{\alpha/2}$이다. 여기서 z_α는 표준정규분포의 상위 $100\alpha\%$ 백분위수이다.

2.2. 벡터집합의 출현시간 간격을 이용한 검정법

2.2.1. 기본이론

$X_1,X_2,\cdots$는 $S=\{0,1,2,\cdots,d-1\}$에서 값을 갖는 $i.i.d.$인 확률변수열이라 하자. 이 절에서는 X_1의 분포를 균등분포로 한정하지 않고 $P(X_1=i)=p(i)$라고 하자. 벡터열 $Z_n=(X_n,X_{n+1},\cdots,X_{n+k-1})$, $n=1,2,\cdots$이 집합 $A=\{U_1,U_2,\cdots,U_n\}$에 도달하는 시간간격은 다음과 같다. 이때 $U_i,(i=1,2,\cdots,n)$는 k digits로 이루어진 유한수열이라 하자.

$$T_0(X,A)=0, \qquad (2.4)$$
$$T_j(X,A)=\min\{n|Z_{T_{j-i}+n}\in A\}+k-1,\ j=1,2,\cdots.$$

예를 들어 X_i가 연속해서 주사위를 던질때 i번째 나오는 눈이라 하자. $U_1=(1,3,4,2)$, $U_2(1,2,3,4)$가 주어진 벡터들이고 $A=\{U_1,U_2\}$이라하자. 만일 계속해서 주사위를 던졌을 때 그 결과가

6324134252432413426112345632 16

이었다면 $T_1(X,A)=8$, $T_2(X,A)=10$, $T_3(X,A)=6$ 이다. $T_j(X,A)$의 정의로부터 우리는 다음의 결과를 쉽게 알 수 있다.

(1) $T_j(X,A)\geq k$.

(2) $T_j(X,A)$, $j=1,2,\cdots,$는 독립이고 같은 분포를 따른다.

(3) $T_j(X,A)=\min\{S_j(X,U_i)|i=1,2,\cdots,n\}$.

$T_1(X,A)$의 분포를 구하기 위하여 다음과 같은 용어를 도입하자. 벡터 $U\in S^k$의 마지막 j개의 성분과 $V\in S^k$의 처음 j개의 성분이 일치할때 U와 V는 j만큼 겹친다고 한다. 이와 같은 겹침을 설명하기 위하여 다음의 용어를 도입한다.

정의 2.4. (Leading Numbers) $V=(v_1,v_2,\cdots,v_k)$상에서 $U=(u_1,u_2,\cdots,u_k)$의 leading numbers는 다음과 같이 정의한다.

$$\varepsilon_j(U,V)=\begin{cases}1, & \text{if } u_{k-j+i}=v_i,\ i=1,2,\cdots,j,\\ 0, & \text{otherwise.}\end{cases}$$

$T_1(X,A)$의 확률 생성함수를 다음과 같이 정의한다.

$$\pi_i(z)=\sum_{m=1}^{\infty}P(T_1(X,A)=T_1(X,U_i)=m)z^m,\quad i=1,2,\cdots,n,$$

$$\pi(z)=\sum_{m=1}^{\infty}P(T_1(X,A)=m)z^m.$$

확률 생성함수들을 계산하기 전에 먼저 다음의 기호를 도입하자.

$$U*V(z)=\sum_{j=1}^{l}\left\{\frac{z^{-j}}{p(v_1)\cdots p(v_j)}\,|\,1\leq j\leq l,\varepsilon_j(U,V)=1\right\}.$$

여기서 $z>0$, $U=(u_1,u_2,\cdots,u_k)\in S^k$, $V\in(v_1,v_2,\cdots,v_k)\in S^k$ 이다.

정리 2.5. ([3]) 확률생성함수 $\pi(z)$와 $\pi_i(z)$, $i=1,2,\cdots,n$은 다음의 연립방정식을 만족한다.

$$\begin{pmatrix}-1 & 1 & 1 & \cdots & 1\\ \frac{1}{1-z} & U_1*U_1(z) & U_2*U_1(z) & \cdots & U_n*U_1(z)\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \frac{1}{1-z} & U_1*U_n(z) & U_2*U_n(z) & \cdots & U_n*U_n(z)\end{pmatrix}\begin{pmatrix}\pi(z)\\ \pi_1(z)\\ \vdots\\ \pi_n(z)\end{pmatrix}=\begin{pmatrix}0\\ \frac{1}{1-z}\\ \vdots\\ \frac{1}{1-z}\end{pmatrix}$$

따름정리 2.6. ([6]) A에 있는 (n−1)개의 다른 벡터들보다 U_i가 제일 먼저 일어날 확률 $\pi_i=\pi_i(1)$과 $T_1(X,A)$의 평균 $E(T_1(X,A))$는 다음의 연립방정식을 만족한다.

$$\begin{pmatrix}0 & 1 & 1 & \cdots & 1\\ -1 & U_1*U_1(1) & U_2*U_1(1) & \cdots & U_n*U_1(1)\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ -1 & U_1*U_n(1) & U_2*U_n(1) & \cdots & U_n*U_n(1)\end{pmatrix}\begin{pmatrix}E(T_1(X,A))\\ \pi_1\\ \vdots\\ \pi_n\end{pmatrix}=\begin{pmatrix}1\\ 0\\ \vdots\\ 0\end{pmatrix}$$

정리 2.7. ([2]) X_1이 균등분포를 따를 경우 즉, $P(X_1=i)=\frac{1}{d}$, $i=0,1,\cdots,d-1$, 일때 $\pi_{im}=P(T_1(X,A)=T_1(X,U_i)=m)$, $i=1,2,\cdots,n$은 다음과 같은 반복적인 형태로 연어진다. $i=1,2,\cdots,n$과 모든 m에 대해서

$$\pi_{im}=d^{-k}-d^{-k}\sum_{j=k}^{m-k}\sum_{l=1}^{n}\pi_{jl}-\sum_{j=m-k+1}^{m-1}\sum_{l=1}^{n}\pi_{jl}\,q_{i-j}(l,m).$$

여기서

$$q_r(l,m)=\varepsilon_{k-r}(U_l,U_m)d^{-r},\ r=0,1,\cdots,k-1.$$

또한 $\pi_i(z)$, $i=1,2,\cdots,n$은 다음의 관계를 만족한다.

$$\sum_{i=1}^{n}\pi_i(z)(1+(1-z)\eta_{ij}(\frac{d}{z}))=1,\quad j=1,2,\cdots,n \qquad (2.5)$$

여기서 $\eta_{ij}(x)=\Sigma_{r=1}^{n}\varepsilon_r(U_i,U_j)x^r$은 $\varepsilon_r(U_i,U_j)$의 생성함수이다.

정리 2.8. ([2]) X_1이 균등분포를 따를때 $T_1(X,A)$의 평균과 분산 그리고 π_i, $i=1,2,\cdots,n$은 다음과 같이 주어진다.

$$E(T_1(X,A))=\frac{1}{\Sigma_{i=1}^{n}x_i},$$

$$\pi_j=\frac{x_j}{\Sigma_{i=1}^{n}x_i},\ j=1,2,\cdots,n,$$

$$\mathrm{Var}(T_1(X,A))=[E(T_1(X,A))]^2+E(T_1(X,A))-2dE(T_1(X,A))\sum_{i=1}^{n}x_i\eta'_{ij}(d).$$

여기서 $x_1,x_2,\cdots,x_n$은 다음의 연립방정식을 만족한다.

$$\sum_{i=1}^{n}x_i\eta_{ij}(d)=1,\ j=1,2,\cdots,n.$$

증명. (2.5)의 양변을 $z=1$에서 미분하면

$$E(T_1(X,A))=\sum_{i=1}^{n}\pi_i'(1)=\sum_{i=1}^{n}\pi_i\eta'_{ij}(d)$$

이 된다. $\pi_i=Cx_i$(여기서 C는 임의상수)로 두면

$$E(T_1(X,A))=C\sum_{i=1}^{n}x_i\eta_{ij}(d).$$

$C=E(T_1(X,A))$로 두면 x_i, $i=1,2,\cdots,n$은 다음의 연립방정식을 만족한다.

$$\sum_{i=1}^{n}x_i\eta_{ij}(d)=1,\ j=1,2,\cdots,n. \qquad (2.6)$$

$1=\pi(1)=\Sigma_{i=1}^{n}\pi_i=C\Sigma_{i=1}^{n}x_i=E(T_1(X,A))\Sigma_{i=1}^{n}x_i$ 이므로 다음의 결과를 얻는다.

$$E(T_1(X,A))=\frac{1}{\Sigma_{i=1}^{n}x_i},$$

$$\pi_i=E(T_1(X,A))\,x_i.$$

$T_1(X,A)$의 분산을 구하기 위하여 (2.5)의 양변을 $z=1$에서 두번 미분하면 다음과 같이 된다.

$$\sum_{i=1}^{n}\pi_i''(1)+\sum_{i=1}^{n}\pi_i 2d'_{ij}(d)=2\sum_{i=1}^{n}\pi_i'\eta_{ij}(d). \qquad (2.7)$$

$\pi_i'(1)=C_1y_i$(여기서 C_1은 임의상수)로 두자.

$$2C_1=E(T_1(X,A)^2)-E(T_1(X,A))+2d\sum_{i=1}^{n}\pi_i\eta'_{ij}(d) \qquad (2.8)$$

라고 두면

$$\pi''(1)=\sum_{i=1}^{n}\pi_i''(1)=E(T_1(X,A)^2)-E(T_1(X,A))$$

이므로 식(2.7)로부터 다음의 연립방정식을 얻는다.

$$\sum_{i=0}^{n}y_i\eta_{ij}(d)=1,\ j=1,2,\cdots,n. \qquad (2.9)$$

연립방정식 (2.9)는 (2.6)과 동일함을 알수 있다. 그러므로

$$y_i=x_i,\ i=1,2,\cdots,n.$$

또한

$$E(T_1(X,A))=\sum_{i=1}^{n}\pi_i'(1)=C_1\eta'_{ij}(d)\sum_{i=1}^{n}x_i=\frac{C_1}{E(T_1(X,A))}$$

이므로

$$C_1=(E(T_1(X,A)))^2. \qquad (2.10)$$

(2.8)과 (2.10)으로부터 다음을 얻는다.

$$E(T_1(X,A)^2)=2(E(T_1(X,A)))^2+$$
$$E(T_1(X,A))-2d\sum_{i=1}^{n}\pi_i\eta'_{ij}(d).$$

그러므로 분산 $\mathrm{Var}(T_1(X,A))$는 다음과 같이 주어진다.

$$\mathrm{Var}(T_1(X,A))=[E(T_1(X,A))^2)+E(T_1(X,A))$$
$$-2d\,E(T_1(X,A))\sum_{i=1}^{n}x_i\eta'_{ij}(d).\ \square$$

예를 들어 위의 결과를 설명하여 보자. 이진수 벡터 $U_1=(0000)$, $U_2=(1000)$이 주어지고 $A=\{U_1, U_2\}$라 하자. 이때 각각의 $r=1,2,3,4$에 대한 leading numbers는 다음과 같다.

$$\varepsilon_r(U_1,U_1)=1,1,1,1,\qquad \varepsilon_r(U_1,U_2)=0,0,0,0,$$
$$\varepsilon_r(U_2,U_1)=1,1,1,0,\qquad \varepsilon_r(U_2,U_2)=0,0,0,1,$$

또한 $\eta_{11}(2)=30$, $\eta_{12}(2)=14$, $\eta_{21}(2)=0$, $\eta_{22}(2)=16$이다. 그러므로 연립방정식은 다음과 같다.

$$30x_1+14x_2=1,$$
$$16x_2=1.$$

이 연립방정식의 해는 $x_1=\frac{1}{240}$, $x_2=\frac{1}{16}$이므로

$$E(T_1(X,A))=\left(\frac{1}{240}+\frac{1}{16}\right)^{-1}=15,$$

$$\pi_1=\frac{1}{16},\ \pi_2=\frac{15}{16}.$$

다른 한 예를 들어보자, 한 남자가 해적에게 잡히게 되었다. 그런데 해적의 선장은 게임을 매우 좋아해서 그 남자에게 다음과 같은 제안을 했다. 주사위의 각 면에 L, I, V, E, K를 쓰고 나머지 한면은 비워두었다. 주사위를 LIVE나 KILL이 나올때까지 계속 던져서 나온 결과에 따라 그 남자를 처리하기로 하였다. 그 남자는 게임을 시작하기 전에 해적 선장에게 KILL을 DEAD로 바꾸어 줄 것을 요청하였다. 해적 선장은 그의 요청을 받아들여 주사위에다가 L, I, V, E, A, D를 썼다. 이렇게 함으로써 그 남자가 살아남을 확률은 $\frac{124}{240}\approx 0.4980$에서 $\frac{217}{433}\approx 0.5012$로 높아지게 되었다.

2.2.2. 검정수행 방법

이 절에서는 앞절에서 얻어진 결과를 이용하여 주어진 수열에 대한 간격검정을 수행하는 방법을 제시한다. $\{x_n\}$을 크기가 N인 d-진수열이라고 하고 $A=\{U_1, U_2, \cdots U_n\}$을 k차원 d-진수열들의 집합이라 하자. 이때 검정을 수행하는 과정은 다음과 같다.

① 간격의 길이에 대한 자료를 수집한다. 즉,

$$g_n(A)=\begin{cases}\Sigma_j\,1(T_j(x,A)=n),\\ \qquad\qquad \text{if } n=k,k+1,\cdots,k+t-1,\\ \Sigma_j\,1(T_j(x,A)\geq k+t),\ \text{if } n=k+t.\end{cases}$$

여기서 $1(D)$는 D의 지표함수(indicator funcion)이다.

② 간격의 총 개수 $G(A)$를 계산한다.

$$G(A)=\sum_{n=k}^{k+t}g_n(A).$$

③ $g_n(A)$의 기대값 $G_n(A)$를 계산한다.

$$G_n(A)=\begin{cases}G(A)P(T_1(X,A)=n), & n=k,k+1,\\ & \cdots,k+t-1,\\ G(A)P(T_1(X,A)\geq k+t), & n=k+t.\end{cases}$$

④ 검정통계량을 계산한다.

$$\tau_1(A)=\sum_{n=k}^{t+k}\frac{(g_n(A)-G_n(A))^2}{G_n(A)},$$

$$\tau_2(A)=\frac{\bar{T}(x,A)-E(T_1(X,A))}{\sqrt{\frac{\mathrm{Var}(T_1(X,A))}{G(A)}}}.$$

여기서 $\bar{T}(x,A)=\Sigma_j\frac{T_j(x,A)}{G(A)}$는 수열 $\{x_n\}$이 A에 도달하는 간격의 평균길이이다.

$\tau_1(A)$는 근사적으로 자유도가 t인 χ^2분포를 따른다. 그러므로 $\tau_1(A)$를 이용하여 검정을 할때는 2.1.2절의 $T_1(U)$를 사용할때와 같은 방법으로 한다.

또한 $\tau_2(A)$는 근사적으로 표준정규분포를 따르므로 $\tau_2(A)$를 이용하여 검정을 할때는 *2.1.2*절의 $T_2(U)$와 같은 방법으로 한다.

3. [0,1]-수열에 대한 k-차원 간격검정

구간 [0,1]내에서 값을 갖는 수열 $\{u_n\}$을 [0,1]-수열이라 한다. 이 절에서는 [0,1]-수열에 대한 k-차원 간격검정법을 제시한다. 이의 한 방법은 $x_n=\lfloor du_n \rfloor, n=1,2,\cdots$이라고 하면 $\{x_n\}$은 $\{0,1,\cdots, d-1\}$에서 값을 갖는 d-진열이 된다. 그러므로 $\{x_n\}$에 대하여 2절에서 제시한 k차원 간격검정을 수행하면 된다.

다른 한가지 검정법은 다음과 같다. $R_i, i=1,2,\cdots$을 [0,1]상에서 균등분포를 따르는 $i.i.d.$ 확률변수열이라 하자. 그리고 $Z_n=(R_n,R_{n+1},\cdots,R_{n-k+1}), n=1,2,\cdots$이라하자. $I_1,I_2,\cdots,I_k$가 [0,1]의 부분구간(subintervals)들 이라고 할때 $I=I_1\times I_2\times\cdots\times I_k$로 두자. k차원 간격검정은 다음과 같은 간격의 분포를 이용한다.

$$D_0(R,I)=0,$$
$$D_j(R,I)=\min\{n|Z_{D_{j-i}+n}\}+k-1, j\geq 1.$$

이론 전개를 단순화 하기 위하여 $k=2$인 경우에 대해서만 언급한다. $I_1\subset[0,1]$, $I_2\subset[0,1]$에 대하여 $|I_i|$를 I_i의 길이(Lebesgue측도)라 하자. R_n을 이용하여 $\{1,2,3,4\}$내에서 값을 갖는 확률변수열 X_n을 다음과 같이 정의하자.

$$X_n=\begin{cases}1, & \text{if } R_n\in I_1\setminus I_2,\\ 2, & \text{if } R_n\in I_1\cap I_2,\\ 3, & \text{if } R_n\in I_1\setminus I_1,\\ 4, & \text{if } R_n\notin I_1\cup I_2.\end{cases}$$

그러면 $\{x_n\}$의 분포는 다음과 같이 된다.

$$p(1)=p(X_1=1)=|I_1\setminus I_2|,$$
$$p(2)=p(X_1=2)=|I_1\cap I_2|,$$
$$p(3)=p(X_1=3)=|I_2\setminus I_1|,$$
$$p(4)=p(X_1=4)=1-|I_1\cup I_2|.$$

$A_1=\{j|p(j)>0, j=1,2\}$, $A_2=\{j|p(j)>0, j=2,3\}$, $A=A_1\times A_2$라고 하자. $\{R_n\in I_1\}=\{X_n=1\}\cup\{X_n=2\}$이고 $\{R_n\in I_2\}=\{X_n=2\}\cup\{X_n=3\}$이므로 $\{(R_n,R_{n+1})\in I_1\times I_2\}=\{(X_n,X_{n+1})\in A\}$이다. 그러므로 $D_j(R, I_1\times I_2)=T_j(X,A), j=0,1,\cdots$가 된다. 따라서 $T_j(X,A)$에 대한 2.2절의 결과를 이용하여 검정을 행한다.

4. 모의실험 결과 및 토의

이 절에서 우리는 다음과 같은 난수 생성자에 의해서 만들어지는 이진수열에 대한 k차원($k=1,3,5$) 간격검정을 시행하였다. 여기서 사용한 난수생성자에 대한 자세한 설명은 현대암호학([8])을 참조하기 바란다.

(1) 선형합동법(Linear congruential generator)

$$x_i=16807x_{i-1} \bmod (2^{31}-1),$$
$$x_0=123456789,$$
$$b_i=\begin{cases}0 & x_i\leq(2^{31}-1)/2,\\ 1 & x_i>(2^{31}-1)/2,\end{cases}$$

(2) 선형귀환 쉬프트 레지스터(Linear feedfack shift register)

$$b_i=b_{i-p}\oplus b_{i-(p-q)}\ (p=33,\ q=13).$$

여기서 $\oplus$는 modulo 2 sum을 의미한다.

(3) 승산(multiplication)

$$c_i=a_i\times b_i.$$

여기서

$$a_i=a_{i-p_1}\oplus a_{i-(p_1-q_1)}\ (p_1=31,\ q=13),$$
$$b_i=b_{i-p_2}\oplus b_{i-(p_2-q_2)}\ (p_2=33,\ q_2=13).$$

(4) J-K플립플롭(J-K Flip-Flop)

$$c_i=((a_i+b_i+1)c_{i-1}+a_i) \bmod 2,$$
$$c_0=0.$$

여기서 a_i, b_i는 (3)의 승산시스템에서 사용한 것과 동일한 것이다.

(5) Geffe 시스템

$$g_i = a_i b_i \oplus c_i b_i \oplus c_i .$$

여기서 a_i, b_i는 승산시스템에서 사용한 것과 동일한 것이며,

$$c_i = c_{i-p} \oplus c_{i-(p-q)} \quad (p=28,\ q=9)$$

이다.

(6) 상호대칭 시스템

$$s_i = a_i b_i \oplus b_i c_i \oplus c_i a_i .$$

여기서 a_i, b_i, c_i는 Geffe시스템에서 사용한 것과 동일한 것이다.

모의실험을 위해 사용한 수열의 길이는 $2^{20} = 1,048,576$이다. 표 1, 표 2, 표 3은 통계량 $T_1(U)$를 이용하여 얻었고 표 4, 표 5, 표 6은 통계량 $T_2(U)$의 결과이다. 표에서 U의 자리에 들어가 있는 정수는 k비트 이진수를 정수로 고쳐놓은 것이다. 예를 들어 표 2에서 4는 "100", 2는 "010"에 대응된다. 각 표에서 난수생성자 번호는 다음과 같다.

1 : 선형합동법　2 : 선형귀환 쉬프트트레지스터
3 : 승산시스템　4 : J-K플립플롭
5 : Geffe시스템　6 : 상호대칭 시스템

또한 표 2와 표 3에서 자유도(df)가 서로 다른 것은 비트벡트 U에 따라서 간격의 길이가 서로 다르고

표 1. $T_1(U)$의 값 (k=1, 자유도=14)

발생기 U	1	2	3	4	5	6
0	5.7344	17.0243	224053.8000	10.3265	13.0407	22.0256
1	15.8517	20.9551	2155600.0000	8.5275	29.0164	20.7803

유의수준 1%에서 검정 통과영역 : $T_1(U) < 29.14$

표 2. $T_1(U)$의 값 (k=3)

생성자	U	d.f.	$T_1(U)$	U	d.f.	$T_1(U)$	U	d.f.	$T_1(U)$	U	d.f.	$T_1(U)$
1	0	36	40.0831	1	16	10.0167	2	24	24.6511	3	16	18.2181
	4	16	19.8886	5	24	25.4656	6	16	11.2839	7	36	48.8401
2	0	36	483.9248	1	16	22.3747	2	24	32.5576	3	16	16.9935
	4	16	20.6629	5	24	22.3534	6	16	11.5310	7	36	177.8256
3	0	36	5854497.0	1	16	209135.5	2	24	178654.1	3	16	143274.5
	4	16	209185.8	5	24	90289.1	6	16	143707.4	7	36	86133.1
4	0	36	316.1055	1	16	13.9334	2	24	23.8341	3	16	13.4985
	4	16	15.9572	5	24	19.0300	6	16	22.3402	7	36	120.2355
5	0	36	47.2363	1	16	8.1392	2	24	15.1399	3	16	9.7331
	4	16	11.2153	5	24	20.6269	6	16	12.8939	7	36	76.8158
6	0	36	55.3531	1	16	17.0534	2	24	17.7716	3	16	19.4786
	4	16	14.1669	5	24	10.7813	6	16	11.4032	7	36	65.7266

유의수준 1%의 각 자유도에 대한 $x^2(d, 0.01)$ 값

d.f.	16	24	36
$x^2(d, 0.01)$	32.00	42.98	58.619

검정의 통과 영역 $T_1(U) < x^2(d, 0.01)$

표3. $T_1(U)$의 값 ($k=5$)

생성자	U	d.f.	$T_1(U)$	U	d.f.	$T_1(U)$	U	d.f.	$T_1(U)$	U	d.f.	$T_1(U)$
1	0	73	88.6587	1	53	45.4121	2	55	47.9158	3	53	47.3374
	4	58	59.4655	5	53	79.8442	6	55	48.8153	7	53	36.7117
	8	55	53.9745	9	57	57.4100	10	62	79.9385	11	53	46.4792
	12	55	49.5260	13	57	55.0704	14	55	45.2394	15	53	54.7066
	16	53	56.1292	17	55	56.5239	18	57	40.9331	19	55	78.0708
	20	53	50.7186	21	62	47.5564	22	57	46.2941	23	55	55.5283
	24	53	58.1314	25	55	46.6203	26	53	41.3720	27	58	62.4222
	28	53	49.3960	29	55	67.5637	30	53	44.8531	31	73	68.9446
2	0	73	506.9178	1	53	845.8359	2	55	196.3260	3	53	160.8862
	4	58	119.7249	5	53	145.6003	6	55	261.6571	7	53	132.8497
	8	55	183.5773	9	57	107.2506	10	62	207.5249	11	53	107.5047
	12	55	231.6223	13	57	192.0410	14	55	84.4735	15	53	173.8793
	16	53	843.0416	17	55	173.6612	18	57	139.1526	19	55	125.4950
	20	53	110.9050	21	62	157.5863	22	57	149.4852	23	55	130.3932
	24	53	161.6243	25	55	129.5050	26	53	152.7405	27	58	216.8324
	28	53	87.0464	29	55	156.8287	30	53	197.9715	31	73	100.3007
3	0	73	221694.50	1	53	79633.04	2	55	62715.29	3	53	1105.48
	4	58	50895.60	5	53	1239.87	6	55	751.25	7	53	33263.26
	8	55	62641.14	9	57	1204.37	10	62	357.37	11	53	33950.44
	12	55	754.99	13	57	27983.61	14	55	31071.52	15	53	32501.46
	16	53	79598.10	17	55	1560.72	18	57	1180.12	19	55	31355.75
	20	53	1180.59	21	62	23732.32	22	57	27956.04	23	55	32892.63
	24	53	1218.04	25	55	31880.31	26	53	33509.16	27	58	26120.74
	28	53	33992.80	29	55	31689.71	30	53	32411.17	31	73	10968.64
4	0	73	307.4007	1	53	517.7211	2	55	89.3189	3	53	60.9204
	4	58	62.8671	5	53	61.6673	6	55	71.8559	7	53	88.1998
	8	55	67.7232	9	57	71.4274	10	62	64.1195	11	53	51.9442
	12	55	51.5765	13	57	75.9196	14	55	42.2095	15	53	430.7101
	16	53	393.6660	17	55	41.9577	18	57	56.0929	19	55	73.5474
	20	53	82.6803	21	62	62.8153	22	57	54.7003	23	55	66.9211
	24	53	57.1059	25	55	49.7284	26	53	41.0474	27	58	74.3716
	28	53	78.3723	29	55	70.2033	30	53	508.8640	31	73	308.2416
5	0	73	79.3044	1	53	80.6234	2	55	52.8744	3	53	39.5662
	4	58	63.2304	5	53	54.4863	6	55	57.3286	7	53	55.7667
	8	55	59.1052	9	57	59.7782	10	62	52.7243	11	53	58.9638
	12	55	63.8453	13	57	62.9922	14	55	40.1620	15	53	74.8740
	16	53	98.1454	17	55	69.7759	18	57	53.6145	19	55	43.4424
	20	53	53.5725	21	62	75.6294	22	57	51.8182	23	55	50.8111
	24	53	56.1560	25	55	55.1802	26	53	47.1059	27	58	58.7164
	28	53	46.6709	29	55	70.2818	30	53	66.0879	31	73	70.5127
6	0	73	76.4078	1	53	70.3573	2	55	89.0810	3	53	66.2532
	4	58	94.9373	5	53	54.2781	6	55	50.6969	7	53	41.6111
	8	55	69.1891	9	57	42.3768	10	62	60.3112	11	53	57.7473
	12	55	56.6758	13	57	43.3259	14	55	61.9942	15	53	77.0152
	16	53	67.4449	17	55	45.9229	18	57	51.1590	19	55	40.0114
	20	53	51.9486	21	62	58.8793	22	57	59.4152	23	55	79.5814
	24	53	77.2267	25	55	49.9013	26	53	59.0388	27	58	66.9186
	28	53	49.9707	29	55	56.6656	30	53	83.1911	31	73	86.5660

유의수준 1%의 각 자유도에 대한 x^2(d. 0.01) 값

표3. (계 속)

df	53	55	57	57	62	73
x^2(d. 0.01)	79.843	82.292	84.733	87.166	90.789	103.996

유의수준 1%의 검정의 통과 영역 $T_1(U) < x^2(d.\ 0.01)$

표4. $T_2(U)$의 값 (k=1)

생성자 U	1	2	3	4	5	6
0	−0.9351	2.0372	−417.4928	−0.5994	1.1686	2.5080
1	0.9123	−1.9821	1219.7030	0.5780	−1.1398	−2.4365

유의수준 1%에서 검정 통과영역 : $|T_2(U)| < Z_{0.005} = 2.576$

표5. $T_2(U)$의 값 (k=3)

생성자	U	$T_2(U)$	U	$T_2(U)$	U	$T_2(U)$	U	$T_2(U)$
1	0	−.1524	1	−1.3629	2	−1.0514	3	.4828
	4	−1.3634	5	−.0345	6	.4862	7	1.4490
2	0	17.9197	1	24.6231	2	62.0508	3	−22.3083
	4	24.6225	5	15.2017	6	−22.3094	7	−60.9820
3	0	−312.3767	1	−70.0799	2	−71.7835	3	600.3456
	4	−70.0820	5	409.4594	6	600.3331	7	656.5037
4	0	−.5955	1	−1.0084	2	−1.5733	3	1.8958
	4	−1.0084	5	.7078	6	1.8980	7	.0626
5	0	.6704	1	1.9914	2	.1399	3	.1714
	4	1.9880	5	−1.1323	6	.1680	7	−.0310
6	0	2.0885	1	1.4711	2	1.9284	3	−.7167
	4	1.4689	5	.1849	6	−.7207	7	−2.3598

유의수준 1%에서 검정 통과영역 : $|T_2(U)| < Z_{0.005} = 2.576$

간격의 길이를 여러개의 범주로 나눌때 각 범주의 확률이 같아지도록 나누었기 때문이다. 표1, 표2, 표3에서 알 수 있듯이 유의수준 1%에서 승산시스템을 제외한 모든 난수생성자의 출력 수열이 k=1일 때는 통과하였다. 그러나 선형쉬프트레지스터와 J-K플립플롭의 출력수열은 k=3일때 U=000, U=111인 경우에 통과하지 못하였다. 표4, 표5, 표6은 k=1,3일때 승산시스템을 제외한 모든 난수생성자의 출력수열이 $T_2(U)$를 이용한 검정을 통과하지만 k=5일때는 선형합동법, J-K플립플롭, Geffe시스템에서 출력된 수열은 검정을 통과하고 상호대칭 시스템에서는 U=0010, U=1111인 경우를 제외하고는 검정을 통과하는 것을 보여준다. 표 6에서는 $T_2(U)$를 이용한 검정법이 $T_1(U)$를 이용한 검정법보다 더욱 민감하다는 것을 알 수 있다.

표 6. $T_2(U)$의 값 (k=5)

생성자	U	$T_2(U)$	U	$T_2(U)$	U	$T_2(U)$	U	$T_2(U)$
1	0	.7572	1	.2655	2	−1.6395	3	.9388
	4	−.5033	5	−2.0067	6	.2990	7	.8081
	8	−.2441	9	−.6691	10	−.4357	11	−1.0490
	12	−1.4807	13	−.0863	14	.2653	15	.7429
	16	.2653	17	−1.1823	18	−.6839	19	−.1816
	20	−.4660	21	.3564	22	−.6058	23	.1330
	24	−.5489	25	−.3986	26	.4964	27	.7332
	28	.1033	29	.9924	30	.7435	31	.3765
2	0	10.3719	1	15.4604	2	−21.0284	3	44.9214
	4	67.9376	5	−4.9852	6	−3.9004	7	−25.6342
	8	91.0356	9	−6.1108	10	−11.7218	11	−3.6240
	12	93.8030	13	−4.6294	14	−79.1534	15	−21.2401
	16	15.4600	17	1.6993	18	91.7486	19	−61.6586
	20	14.6052	21	−10.5982	22	93.2881	23	−75.8959
	24	−49.4184	25	5.5244	26	16.0983	27	−14.7012
	28	−98.7330	29	3.7698	30	−21.2454	31	−46.3136
3	0	−240.2013	1	−205.5680	2	−186.5631	3	35.5985
	4	−170.3413	5	36.5476	6	28.7192	7	284.7231
	8	−186.5748	9	35.3909	10	17.9437	11	290.8354
	12	28.8942	13	246.3207	14	258.5434	15	622.6193
	16	−205.5704	17	41.5516	18	35.8504	19	265.4561
	20	36.6257	21	224.3664	22	246.3376	23	574.5276
	24	35.6687	25	263.6413	26	291.0394	27	500.3738
	28	283.1589	29	581.0095	30	662.4999	31	612.9183
1	0	.2579	1	.0959	2	−.2887	3	−.2286
	4	−1.3367	5	−.7375	6	.2870	7	−.0941
	8	−2.2226	9	−.2993	10	.0159	11	.3760
	12	1.0708	13	1.2829	14	1.1475	15	−.8213
	16	.1100	17	−.4268	18	−6368.	19	.9638
	20	−.8012	21	−0941.	22	1.2423	23	.6073
	24	1.6145	25	−.9429	26	.3281	27	.9390
	28	−.4959	29	.7897	30	−.8209	31	.1342
5	0	.5160	1	−.8134	2	−.4702	3	−.4652
	4	1.5301	5	−.9582	6	.7891	7	1.0372
	8	−1.1152	9	2.2985	10	−1.7195	11	−.8629
	12	.6922	13	−1.0867	14	1.8857	15	−1.7045
	16	−8136.	17	.0925	18	1.0216	19	1.9517
	20	.0436	21	−1.8473	22	−.4821	23	−.7561
	24	.0383	25	.3077	26	.1399	27	−.4916
	28	−.1899	29	.5332	30	−1.7107	31	−1.2385
6	0	.7922	1	.8942	2	2.6184	3	−0.212
	4	1.8123	5	.8095	6	.9117	7	−.8306
	8	1.2767	9	.1231	10	−.6608	11	.7607
	12	1.2889	13	.2488	14	−.7087	15	−1.6596
	16	.8938	17	2.0810	18	−.7048	19	−.2054
	20	−.0116	21	−1.1637	22	.6908	23	−1.2981
	24	1.7722	25	−1.3204	26	−.0560	27	−1.0445
	28	−.6011	29	−1.3189	30	−1.6659	31	−3.0362

유의수준 1%에서 검정 통과영역 : $|T_2(U)| < Z_{0.005} = 2.576$

참 고 문 헌

1. H.Beker and F.Piper, *Ciper Systems, The Protection of Communications*, John Wiley & Sons New York, 1982.

2. G. Blom and D. Thorburn, *How many random digits are required until given sequences are obtained?*, Journal of Applied Probability 19 (1982), 518-531.

3. H.V. Gerber and S.Y. Robert Li, *The occurrence of sequence patterns in repeated experiments and hitting times in a Markov chain*, Stochastic Processes and their Applications 11(1981), 101-108.

4. D.E. Knuth, *The Art of Computer Programming, vol.2, Seminumerical Algorithms, 2nd ed*, Addison-Wesley Publishing Company, 1981.

5. B.D. Ripley, *Stochastic Simulation*, John Wiley & Sons, New york, 1987.

6. S.Y. Robert Li, *A martingale approach to the study of occurence of a sequence patterns in repeated experiments*, The Annals of Probability 8(6)(1980), 1171-1176.

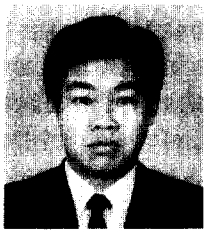
7. 최봉대, *Randomness* 특성 분석에 관한 연구, 한국전자통신연구소 연구보고서, 1991.

8. 현대암호학(*Modern Cryptology*), 한국전자통신연구소.

□ 著者紹介

崔 鳳 大(정회원)

慶北大學校 師範大學 數學科(理學士)
慶北大學校 大學院 數學科(理學碩士)
Ohio State University 大學院 數學科(理學博士)
University of North Carolina 訪問教授
慶北大學校 數學科, 專任講師, 助教授
韓國科學技術院 數學科, 助教授, 副教授, 教授
현재 : 韓國科學技術院 數學科 教授
大韓數學會 總務理事



신 양 우(정회원)

慶北大學校 自然科學大學 數學科(理學士)
韓國科學技術院 應用數學科(理學碩士)
韓國科學技術院 數學科(理學博士)
韓國科學技術院 應用數學科 助教
현재 : 昌原大學校 自然科學大學 統計學科 助教授

이 경 현(정회원)

慶北大學校 師範大學 數學教育科(理學士)
韓國科學技術院 應用數學科(理學碩士)
韓國科學技術院 數學科(理學博士)
현재 : 韓國電子通信研究所 先任研究院