The Korean Communications
in Statistics Vol. 9, No. 1, 2002
pp. 75-86

# Change-point Estimation based on Log Scores[1)]

**Jaehee Kim[2)] and Hyunjoo Seo[3)]**

## Abstract

We consider the problem of estimating the change-point in mean change model with one change-point. Gombay and Huskova(1998) derived a class of change-point estimators with the score function of rank. A change-point estimator with the log score function of rank is suggested and is shown to be involved in the class of Gombay and Huskova(1988). The simulation results show that the proposed estimator has smaller mse, larger proportion of matching the true change-point than the other estimators considered in the experiment when the change-point occurs in the middle of the sample.

*Keywords* : Change-point, Rank, Score function, Wiener process.

## 1. 서론

현대는 정보화 사회인 까닭에, 신문, 잡지 등을 통하여 다량의 자료를 접하게 되며 자료들로부터 올바른 추론을 하는 것은 쉽지 않다. 통계학은 올바른 예측을 얻기 위한 도구학문으로 볼 수 있으며 여러 분야에서 통계적 방법이 필요하게 된다.

현대의 모든 학문분야에 있어서 어떠한 이론에 대한 실증적 연구와 과학적 분석기법이 중요시됨에 따라 사회과학분야이든 자연과학분야이든 심지어 예술분야이든 통계학은 널리 응용되고 있다. 따라서 수많은 자료들 중에서 의미있는 자료를 선택하여 분석하고, 자료 내에 변화가 있는 경우 변화의 흐름을 파악하고 그 변화의 시점을 찾아내어 변화원인에 대한 분석과 대안을 제시할 필요가 현대 통계학에서는 요구되어지고 있다. 즉, 방대한 자료들을 압축하고 여과시켜 단순한 자료의 분석이 아닌 자료 속에서 의미 있는 정보를 찾아내는 방법론의 개발이 필요하다.

변화점이란 관측값이 연속적인 시간에 의해 발생하는 경우 혹은 어떤 일정한 패턴에서 순차적으로 관측된 경우, 자료 내에 변화가 발생하는 시점을 의미한다. 변화점 추정(change-point estimation)은 이러한 변화를 감지하고 변화가 발생하는 시점을 찾아내는 방법으로, 변화점 추정을

1) This research was supported by the Research Fund 2001 from Duksung Women's University and partially supported by the Promoting Program for the basis of Women's University from Korea Institute of S&T Evaluation and Planning.

2) Associate Professor, Department of Statistics, Duksung Women's University, Ssangmun-Dong, Tobong-Ku, Seoul, S.KOREA
E-mail : jaehee@duksung.ac.kr
3) Graduate student in Master program, Department of Statistics, Duksung Women's University, Seoul, S.Korea

통해서 우리는 자료를 균일한 부분으로 나눌 수 있게된다.

본 논문에서는 비모수적 방법을 이용하여, 위치모수의 변화가 있는 경우 변화점 추정 통계량을 제안하고자 한다. 비모수적 방법에서 흔히 사용하는 모집단에 대한 가정은 "연속성"이며, 모형에 따라서 "대칭성"을 추가로 가정하는 경우가 많다. 전통적인 비모수적 방법에서 흔히 사용되는 도구는 관측값의 부호(sign)와 순위(rank) 또는 순위에 기초한 점수(score)를 사용하여 정보의 손실을 줄일 수 있다. 특히 Gombay 와 Huskova(1998) 의 조건에 맞는 점수함수(score function)와 순위점수(rank score)를 이용하여 변형시킨 변화점 추정통계량을 제안하고, 모의실험을 통하여 변화점 추정통계량들의 움직임을 비교하고자 한다.

본 논문에서는 기존의 변화점 추정통계량들에 대한 이론을 2장에서 소개하고, log 점수함수와 순위를 이용한 변화점 추정량을 제안하고 4장에서는 2장과 3장에서 소개된 변화점 추정통계량을 S-plus 를 이용한 모의실험을 통해 비교하고 분석한다. 마지막 5장에서는 본 논문의 결론 및 제안을 한다.

## 2. 기존 변화점 추정통계량

변화점(change-point)란 연속적인 시간에 의해 발생된 확률변수가 자료에서 변화가 발생될 때의 시점을 나타내며 위치모수의 변화, 분산모수의 변화 등이 일어나는 시점이다. 변화점 추정연구에서 고려되어지는 문제로는 변화시간의 추정과 발생된 변화의 크기 추정으로 나눌 수 있으며 본 연구에서는 변화시간의 추정, 즉 변화점 추정 문제를 다루고자 한다.

$X_1, X_2, \cdots, X_n$ 는 연속분포를 따르는 독립변수로 다음의 모형을 만족한다.

$$\begin{aligned} &X_1, X_2, \ldots, X_\tau \quad \sim iid\ F(x), \\ &X_{\tau+1}, X_{\tau+2}, \ldots, X_n \sim iid\ G(x), \quad \tau \in \{1, \ldots, n-1\} \\ &F(x) \neq G(x). \end{aligned}$$

여기서, 자료내 실제 변화점 $\tau$는 분포 $F$에서 $G$로 변화가 시작하는 시점이 된다. Hinkley(1970)는 최대가능도법(maximum likelihood method)을 이용하여 변화점 추정통계량을 제안했다. 동일한 모수족에서 평균치의 차이가 존재하는 분포함수 F와 G를 얻어 그 가운데 존재하는 변화점을 최대가능도방법을 이용하여 추정한다. Hinkley(1970)의 모형을 살펴보면, 연속된 확률변수 $\{X_1, X_2, \cdots, X_n\}$에 1개의 변화점이 존재하는 경우

$$X_t = \begin{cases} \theta_0 + \varepsilon_t, & t = 1, 2, \cdots, \tau \\ \theta_1 + \varepsilon_t, & t = \tau+1, \cdots, n \end{cases}$$

여기서 오차항 $\varepsilon_t$는 독립이고 동일한 분포 $N(0, \sigma^2)$을 따른다. $\theta$는 평균이고 변화점 $\tau$는 모르는 시간 상수이다. Hinkley(1970)는 오차항이 정규분포를 따를 때 우도함수와 $\theta_0$

$$Z_t^2 = \frac{t(n-t)(\overline{x_t} - \overline{x_t^*})^2}{n}, \quad t = 1, \cdots, n-1$$

를 얻고 변화점 추정량으로

$$T_{Hink} = \arg\max_{1<K<n} \{Z_t^2\}$$

을 제안했다. 즉, 우도함수(likelihood function)를 최대화하는 시점이 주어진 모형 내에서 발생하는 변화점으로 추정된다.

Darkhovsh(1976)은 맨-휘트니 통계량(Mann-Whitney statistic)에 기초를 둔 변화점 통계량을 제안했고, Bhattachryya and Johnson(1978)은 변화점의 유무에 관한 비모수적 검정법을 제안하였다.

Pettit(1979)는 평균이 변화하는 모형에서 두 표본일 경우 맨-휘트니 통계량(Mann-Whitney statistic)을 기초로 하여 비모수적 변화점 추정량들을 제안하였다. Pettit(1979)의 변화점 추정량은 다음과 같다.

$$U_{(t,n-t)}=\frac{1}{2}\left\{\sum_{i=1}^{t}\sum_{k=t+1}^{n} sgn(X_i-X_k)+t(n-t)\right\}$$

여기서, 부호함수의 정의는 다음과 같다.

$$sgn(x)=\begin{cases}1, & x>0\\ 0, & x\le 0\end{cases}$$

이다. 여기서

$$W_t=2U_{(t,n-t)}-t(n-t),\quad t=1,2,\cdots,n-1$$

를 이용하여 Pettit(1979)은 변화점 추정통계량으로

$$T_p=\arg\max_{1<t<n}\{W_t\}$$

을 제안하였다.

Schechtman(1982)은 $U_{(t,n-t)}$이 $(X_1, X_2, \cdots, X_j)$와 $(X_{j+1}, \cdots, X_n)$와 같이 두 표본인 경우, 맨-휘트니 윌콕슨 통계량(Mann-Whitney-Wilcoxon statistic)으로부터 맨-휘트니(Mann-Whitney) 형태를 갖는다고 가정하였다. 이로부터 다음과 같은 형태를 갖게된다.

$$U_{(t,n-t)}=\frac{1}{2}\left\{\sum_{i=1}^{t}\sum_{k=t+1}^{n} sgn(X_i-X_k)+t(n-t)\right\}$$

이다. Schechtman(1982)의 평균 변화 모형에서 변화점 추정량은 다음과 같다.

$$V_t=\frac{\left[\frac{U_{(t,n-t)}}{t(n-t)}-0.5\right]}{\left[\frac{(n+1)}{12t(n-t)}\right]^{0.5}},\quad t=1,2,\cdots n-1.$$

변화점 추정통계량으로

$$T_{Sche}=\arg\max_{1<t<n}\{V_t\}$$

을 제안하였다.

Hawkins(1977, 1986)는 최소제곱추정법(LSE: least square estimation method)을 이용하여 평균의 차이가 존재하는 임의의 표본에 대한 변화점 추정방법을 제안하였다. Hawkins(1977, 1986)는 다음의 모형으로부터 나온 자료로부터 최소제곱방법을 이용하여 변화점 $\tau$ 의 추정을 제안하였다.

$$X_i=\mu+\delta I_{(i\ge\tau)}+\varepsilon_i,\quad i=1,2,\cdots,n$$

여기서 $\mu, \delta, \tau$는 모르는 상수이고 $\varepsilon_i$는 $E[\varepsilon_i]=0$이고 $\sigma^2=E[\varepsilon_i^2]<\infty$, $1\le i\le n$인 독립이고 동일한 분포를 따르는 오차항이다. $\overline{X_t}=\sum_{i=1}^{t}X_i/t$, $t=1,2,\cdots,n$ 이고

$$Q_t=\frac{nt}{n-t}\left[\overline{X_t}-\overline{X_n}\right]^2/\sigma^2$$

일 때 Hawkins(1986)의 변화점 추정통계량은

$$T_{Hawk} = \arg\max_{1<t<n}\{Q_t\}$$

으로 제안하였다. 특히 오차 항이 정규분포를 따를 경우 Hinkley(1970)의 변화점 추정량은 Hawkins(1986)의 변화점 추정통계량과 같게된다.

Lombard(1987)은 독립적인 관측치로부터 하나 또는 그 이상의 변화점존재여부를 검정하기 위해 순위 통계량에 기초를 둔 변화점 통계량을 제안하였다. 데이터를 순위함수로 바꿈으로써 변화가 없는 귀무가설 하에서 분포무관 검정통계량을 얻을 수 있고 이상점의 영향을 적게 받게 된다. 서로 독립인 확률변수 $X_1, X_2, \cdots, X_n$이 연속인 분포함수 $F(x, \theta_1), \cdots, F(x, \theta_n)$을 갖는다고 하면, $\theta_1 = \cdots = \theta_\tau = \theta$, $\theta_{\tau+1} = \cdots = \theta_n = \theta^*$인 경우 시점 $\tau$를 변화점이라 한다. Lombard(1987)은 평활변화모형(smooth change model)과 가파른 변화모형(abrupt change model) 2가지를 모두 고려하였다.

단순 가파른 변화(single abrupt change)는 다음과 같다.

$$\theta_i = \begin{cases} \xi_1, & 1 \le i \le \tau \\ \xi_2, & \tau < i \le n \end{cases}$$

점진적 변화를 고려한 모형으로 평활변화모형(smooth change model)은 다음과 같다.

$$\theta_i = \begin{cases} \xi_1, & i \le \tau \\ \xi_1 + (i-\tau_1)(\xi_2-\xi_1)/(\tau_2-\tau_1), & \tau_1 < i \le \tau_2 \\ \xi_2, & i > \tau_2 \end{cases}$$

여기서 $\theta_i$는 위치모수일 필요는 없고 $\xi_1, \xi_2, \tau_1, \tau_2$는 모르는 모수이다. $\tau_2 = \tau_1 + 1$인 경우에는 평활변화모형이 단순 가파른 변화모형이 됨을 알 수 있다. 평활변화모형(smooth change model)에서 Lombard(1987)의 변화점 추정통계량은 다음과 같다. rank $(x_i) = r_i$라고 할 때, 점수함수(score function) $\phi$는 $0 < \int_0^1 \phi^2(u)du < \infty$을 만족한다. 순위점수(rank score)를 s $(r_i)$라고 하면,

$$s(r_i) = \left[\phi\left(\frac{r_i}{n+1}\right) - \bar{\phi}\right]/A$$

여기서

$$\bar{\phi} = \frac{1}{n}\sum_{i=1}^{n}\phi\left(\frac{i}{n+1}\right),$$

$$A^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[\phi\left(\frac{i}{n+1}\right) - \bar{\phi}\right]^2$$

이다. 순위통계량으로

$$v_{t_1,t_2} = \sum_{j=t_1+1}^{t_2}\sum_{i=1}^{t_1} s(r_i)$$

를 계산하고 표준화하여

$$\tilde{v}_{t_1,t_2} = v_{t_1,t_2}/\sigma(t_1/n, t_2/n)$$

를 얻는다. 여기서 $\sigma^2(u,v) = (1-u)^3(1+3u)/12 - (1-v)^3(1+3v)/12 - (1-v)^2(v^2-u^2)/2$이다.

Lombard(1987)는 변화점 추정 통계량으로 $T_L = \arg\max_{t_1,t_2}\{|\tilde{v}_{t_1,t_2}|\}$ 을 제안하였다. 특히 변화점이 1 개일 경우,

$$T_L = \arg\max_{1<t<n}\{|\tilde{v}_{t,t+1}|\}$$

로 표현된다.

Carlstein(1988)은 한 시점을 중심으로 두 분포간의 거리를 최대로 하는 시점을 변화점으로 추정하였다. 주어진 독립변수들은 다음 두 분포를 따른다고 모형을 가정한다.

$$X_1, X_2, \cdots X_\tau \sim iid\ F(x)$$
$$X_{\tau+1}, X_{\tau+2}, \cdots X_n \sim iid\ G(x)$$

여기서, $F(x)$와 $G(x)$에 대한 분포함수의 특별한 가정 없이 단지 $\psi = \{x \in R: |F(x) - G(x)| > 0\}$ 에서 $\int_\psi dF(x) > 0$ 또는 $\int_\psi dF(x) > 0$ 가정을 기본 조건으로 한다. $t \in \Lambda = \{i/n: 1 \le i \le n-1\}$에 대해 t시점 이전의 경험누적함수(pre-t empirical cdf) ${}_t h(x)$와 t시점 이후의 경험누적함수(post-t empirical cdf) $h_t(x)$는 다음과 같다.

$${}_t h(x) = \sum_{i=1}^{nt} I\{X_{i \le x}\}/nt$$

$$h_t(x) = \sum_{i=nt+1}^{n} I\{X_{i \le x}\}/n(1-t)$$

여기서 지시함수(indicator function)

$$I(X \le a) = \begin{cases} 1, & x \le a \\ 0, & x > a \end{cases}$$

이다. t시점 이전, 이후의 경험누적 함수를 이용하여 Carlstein(1988)은 3가지의 거리 기준을 고려하여 다음의 변화점 통계량을 제안하였다.

(1) 차이의 절대값의 합을 이용한 통계량

$$D_1(t) = t^{0.5}(1-t)^{0.5} n^{-1} \sum_{i=1}^{n} |\ {}_t h(x_i) - h_t(x_i)|$$

변화점 추정통계량은 $T_{Carl1} = \arg\max_{1<t<n}\{D_1(t)\}$ 이다.

(2) 차이의 제곱 합을 이용한 통계량

$$D_2(t) = t^{0.5}(1-t)^{0.5}\left[n^{-1}\sum_{i=1}^{n}\{\ {}_t h(x_i) - h_t(x_i)\}^2\right]^{0.5}$$

변화점 추정통계량은 $T_{Carl2} = \arg\max_{1<t<n}\{D_2(t)\}$ 이다.

(3) 차이의 최대값을 이용한 통계량

$$D_3(t) = t^{0.5}(1-t)^{0.5}\ SUP_{1 \le i \le n}|\ {}_t h(x_i) - h_t(x_i)|$$

변화점 추정통계량은 $T_{Carl3} = \arg\max_{1<t<n}\{D_3(t)\}$ 이다.

Carlstein(1988)은 Carlstein의 통계량 $\hat{\tau}$에 대하여, $\delta \in [0, 1/2)$에 고정되어 있다고 하면, $|\hat{\tau} - \tau| n^\delta$는 $n$이 커짐에 따라 0에 수렴함을 증명하고 또한 $\varepsilon > 0$에 대하여 $c_1 > 0$ 이고 $c_2 > 0$ 인 상

수와 모든 $n \geq n(\varepsilon)$에 대하여 $P\{|\hat{\tau} - \tau| > \varepsilon\} \leq c_1 n \exp\{-c_2 \varepsilon^2 n\}$이 성립함을 보였다.

## 3. 점수함수를 이용한 변화점 추정통계량 제안

독립적인 확률변수 $\{X_1, X_2, \cdots, X_n\}$의 평균의 변화가 한 번 발생하는 경우 다음의 변화점 모형을 갖는다.

$$X_i = \begin{cases} \mu_1 + \varepsilon_i, & i = 1, 2, \cdots, \tau \\ \mu_2 + \varepsilon_i, & i = \tau + 1, \cdots, n \end{cases}$$

여기서, $\delta = \mu_2 - \mu_1$ 으로 0이 아닌 상수이고( $\mu_1 \neq \mu_2$), 오차항 $\varepsilon_i$는 서로 독립이며 평균이 0, 분산이 $\sigma^2$인 동일한 연속 분포를 따르며 $\tau$ 는 평균의 변화가 일어나는 변화점이다.

$\{X_1, \ldots, X_n\}$ 은 순위 $\{r_1, \ldots, r_n\}$ 을 가지며 $rank(X_i) = r_i$ 라고 할 때 순위를 이용한 점수함수(score function)로서 다음의 로그함수 $\phi(t) = \log(t+1)$, $t = r_i/(n+1)$ 이용을 제안하고자하며 log 함수를 통해 점수(score) 간의 차이를 줄이게 된다. $a(r_i) = \phi(r_i/(n+1))$ 로 놓으면 $\boldsymbol{a} = (a(r_1), \ldots, a(r_n))$ 은 순위를 이용한 점수벡터(a vector of scores)이며 부분합(partial sum)은

$$S_k = \sum_{j=1}^{k} (a(r_i) - \overline{a_n}), \qquad k = 1, \ldots, n$$

으로 표현하며 여기서 $\overline{a_n} = \sum_{j=1}^{n} a(j)/n$이다.

본 논문에서는 순위에 기초하여 log 점수함수를 이용한 변화점 추정통계량으로

$$T = \arg\max_{1 < k < n} |S_k|$$

를 제안하고자한다.

Gombay 와 Huskova(1998)는 순위에 근거한 변화점 추정량의 형태를 제안하고 극한분포를 규명하였다. 본 연구에서 제안하는 변화점 추정통계량 $T$ 은 Gombay 와 Huskova(1998)가 제안한 추정량집합에 속하며 필요한 가정을 만족하므로 Gombay와 Huskova(1998)의 결과들을 이용할 수 있다. 다음에서는 가정이 만족됨을 보이고 $T$ 의 극한분포에 관한 정리를 보이고자한다.

**가정1.** $\tau = [n\theta]$ 일 때, $\theta \in (0,1)$는 존재한다. 여기서 $[a]$는 정수를 나타낸다.

만족 : $\theta = \tau/n$, $\tau \in (1, n)$ 이므로 $\theta \in (0,1)$ 가 존재한다.

**가정2.** $n \to \infty$ 일 때, 양의 정수 c와 D가 존재한다.

$$\sigma_n^2(\boldsymbol{a}) \to c$$

그리고

$$\frac{1}{n} \sum_{i=1}^{n} (a(i) - \overline{a_n})^4 \leq D \quad , \quad n \geq 1$$

여기서 $\sigma_n^2(\boldsymbol{a})=\frac{1}{n-1}\sum_{i=1}^{n}(a(i)-\bar{a}_n)^2$ 이다.

만족 :

$$\begin{aligned}0\le\sigma_n^2(\boldsymbol{a}) &= \frac{1}{n-1}\sum_{i=1}^{n}(a(i)-\bar{a}_n)^2\\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\log\left(1+\frac{i}{n+1}\right)-\bar{a}_n\right)^2\\ &\le \frac{1}{n-1}\sum_{i=1}^{n}\left(\log\left(1+\frac{i}{n+1}\right)\right)^2\\ &\approx \int_0^1\log(1+x)^2dx\le\int_0^1(1+x)^2dx=\frac{7}{3}<\infty\end{aligned}$$

그러므로 integral test에 의해 어떤 양의 상수 $c$ 가 존재해서 $\sigma_n^2(\boldsymbol{a})\to c$ 이다.

**가정3.** $n\to\infty$ 일 때, $\frac{1}{\tau}\sum_{i=1}^{\tau}(a(r_i)-\bar{a}_n)^2-\sigma_n^2(\boldsymbol{a})\to^p 0$ 이다.

만족: $E(a(r_i)-\bar{a}_n)^2=\sigma_1^2,\quad i=1,\ldots,n$ 이므로

$$\begin{aligned}\frac{1}{\tau}\sum_{i=1}^{\tau}(a(r_i)-\bar{a}_n)^2-\sigma_n^2(\boldsymbol{a}) &= \frac{n-\tau}{\tau n}\sum_{i=1}^{\tau}(a(r_i)-\bar{a}_n)^2-\frac{1}{n}\sum_{i=\tau+1}^{n}(a(r_i)-(\bar{a}_n))^2.\\ &\to^p 0\end{aligned}$$

**가정4.** $n\to\infty$ 일 때, 다음의 조건을 만족하는 일련의 정수 $\{d_n\}$이 존재한다.

$$|d_n|\to 0,\quad |d_n|\sqrt{\frac{n}{\log\log n}}\to\infty$$

이고

$$\frac{n}{\tau(n-\tau)d_n}S_\tau\to b,\quad b\neq 0.$$

만족 : $d_n=n^{-a},\quad a<1/2$ 로 택하면 가정의 조건을 만족하게 된다.

이와같이 Gombay 와 Huskova(1998)의 가정들을 만족하므로 Gombay 와 Huskova (1998)에서 증명한 다음의 정리가 성립한다.

**정리1.** $(X_1,\cdots,X_\tau)$과 $(X_{\tau+1},\cdots,X_n)$는 각각 독립변수로 분포가 $F$와 $G$인 연속분포함수를 갖는다고 하자. $n\to\infty$ 때 가정1-가정4 에 의해서 다음을 얻는다.

$$\frac{b^2d_n^2}{\sigma_n^2(\boldsymbol{a})}(T-\tau)\ \to\ \min\{\ z\in R^1\ :\ \max\{W(t)-|t|g(t)\ ,\ t\in R^1\}=W(z)-|z|g(z)\ \}$$

여기서

$$g_1(t)=\begin{cases}1-\theta, & t<0\\ \theta, & t>0\end{cases}$$

$$g_2(t)=\frac{1}{2},\quad t\in R^1$$

$$W(t)=\begin{cases}W_1(-t) & ,\ t<0\\ W_2(t) & ,\ t>0\end{cases}$$

이고 $\{W_1(t),\ t>0\}$과 $\{W_2(t),\ t>0\}$ 는 서로 독립인 Wiener process 이다.

**정리2.** $X_1,\cdots,X_n$ 은 임의의 변수로 분포가 $F$인 연속분포함수를 갖는다. 그리고 그 점수(score)를 $a(1),\cdots,a(n)$이고 $n\to\infty$ 일 때 가정2를 만족하면

$$\frac{T}{n}\to \min\{t\in(0,1)\ ;\ |B(t)|=\max_{0\le v\le 1}|B(v)|\},$$

이고, $\{B(v)\ ,\ 0\le v\le 1\}$은 Brownian bridge를 나타낸다.

## 4. 모의실험

변화점 추정능력을 비교하기 위하여 S-plus를 이용한 모의실험을 행하고자한다. 모의실험에서 변화점이 1 개 존재하는 다음의 변화점 모형을 사용한다.

$$X_i=\begin{cases}\mu_1+\varepsilon_i\ , & i=1,2,\cdots,\tau\\ \mu_2+\varepsilon_i\ , & i=\tau+1,\cdots,n\end{cases}$$

여기서 $\mu_1=0$ 에 대해 평균의 변화량을 고려하여 $\mu_2=1,2,3,4$ 의 경우에 대해 실험한다. 오차항 $\varepsilon_i$ 는 평균이 0, 분산이 1인 정규분포, 이중지수분포, 균일분포를 따르며, 균일분포의 경우에는 (-1.7, 1.7) 구간에서 동일한 확률값을 갖는 경우가 된다. 표본의 크기 $n=100$ 에 대해 변화점 $\tau=50,30$ 인 경우, $\mu_1=0$, $\mu_2=1,2,3,4$ 에 대하여 1,000번의 반복실험으로 변화점 추정통계량의 움직임을 관찰하고자한다. 변화점 추정시 자료의 매우 앞부분이나 매우 뒷부분에서 변화점에 대한 발생가능성을 제외하고자 $\tau=50$ 인 경우에는 $t=20,\ldots,80$ 범위에서, $\tau=30$ 인 경우에는 $t=10,\ldots,90$ 범위에서의 변화점추정을 고려하였다. 변화점 추정능력을 비교하기위한 통계량으로 반복실험에서의 평균, 평균제곱오차(MSE), 변화점 추정비율, 변화점에 대한 95% 신뢰구간을 계산한다. 여기서 변화점 추정비율은 주어진 변화점을 정확히 추정한 비율을 계산한 것이다.

표 1, 2, 3에서 보면 $\tau=50$ 일 때, 즉 변화점이 자료의 중간에서 발생할 때, 제안하는 변화점 통계량 $T$ 의 평균제곱오차가 다른 통계량에 비해 적고, 변화점 추정비율이 더 높은 것으로 나타나 기존추정량인 Hinkley(1970), Schechtman(1982), Carlstein(1988), Lombard(1987)의 추정량 보다 추정능력이 우수함을 알 수 있다. 그러나 $\tau=30$ 일 때, 즉 변화점이 자료의 중간보다 앞부분에서 발생할 때, 제안하는 변화점통계량 $T$ 의 평균제곱오차가 $T_{Carlstein}$ 보다는 우수하지만 그 외의 다른 추정통계량보다는 추정능력이 약간 떨어지는 것을 볼 수 있다.

## 5. 결론

변화점 추정문제는 최근 여러 분야에서 대두되고 있는 문제로, 데이터가 동일한 분포에서 나오지 않을 경우 분포의 변화가 일어나는 시점을 추정하고자한다. 이번 연구에서는 자료내에 위치모수의 변화가 한번 일어날 경우 순위와 log 점수함수를 이용한 비모수적 변화점 추정통계량을 제안하였다. 특히 Gombay 와 Huskova(1998)가 규명한 극한 분포로 근사될 수 있어 제안하는 추정

량의 극한분포를 밝힐 수 있었으며 Gombay 와 Huskova(1998)가 제시하는 가정에 맞는 다른 추정량의 모색이 가능함을 보여주었다. 또한 모의실험을 통해 제안하는 통계량의 변화점추정능력이 우수함을 보였으며 새로운 접근방법에 위한 변화점 추정량을 기대한다.

## 참고문헌


[1] Bhattcharyya, G. K. and Johnson, R. A. (1968). Non-parametric Test Shift at an Unknown Time Point. *Annals of Mathematical Statistics.* 39, 1731-1743.

[2] Carlstein, E. (1988). Nonparametric Change-point Estimation. *Annals of Statistics,* 16, 188-197.

[3] Darkhovsh, B. S. (1976). A Non-parametric Method for the a Posteriori Detection of the "Disorder" Time of a Sequence of Independent Random Variables. *Theory of Probability and Its Applications* 21, 178-183.

[4] Gombay, E. and Huskova, M. (1998). Rank based Estimators of the Change-point. *Journal of Statistical Planning and Inference.* 67, 137-154.

[5] Hawkins, D. L. (1986). A Simple Least Squares Method for Estimating a Change in Mean. *Communications in Statistics.* 15, 655-679.

[6] Hinkley, D. V. (1970). Inference about the Change-point in a Sequence of Random Variables. *Biometrika.* 57, 1-16.

[7] Kim, Jaehee and Jang Heeyoon (1999). Change-point Estimators using Rank Average in Location Change Model, 「한국통계학회논문집」 6, 467-477.

[8] Lombard, F (1987). Rank Test for Changepoint Problems. *Biometrika.* 74, 615-624.

[9] Pettit, A. N. (1979). A Nonparametic Approach to the Change Problem. *Applied Statistics.* 28, 126-135.

[10] Schechtman, E. (1982). A Nonparametic Test for Detecting Changes in Location. *Communications in Statistics-Theory and Methods.* 11, 1475-1482.


[ 2001년 9월 접수, 2002년 2월 채택 ]

**표1.** n=100, 변화점 $\tau=50$, $\tau=30$ 일 때 오차항이 정규분포 N(0,1) 따르는 경우

| 변화점 | | $\tau=50$ | | | | $\tau=30$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 평균 | 평균 제곱 오차 | 추정 비율 | 95% 신뢰구간 | 평균 | 평균 제곱 오차 | 추정 비율 | 95% 신뢰구간 |
| $\mu_0=0$ $\mu_1=1$ | $T_{Hink}$ | 50.027 | 30.975 | 0.473 | (38, 62) | 30.874 | 60.766 | 0.463 | (18, 52) |
| | $T_{Sche}$ | 50.006 | 31.842 | 0.463 | (37, 62) | 31.169 | 57.607 | 0.456 | (19, 51) |
| | $T_{Carl1}$ | 49.077 | 37.513 | 0.466 | (33, 60) | 29.300 | 52.706 | 0.456 | (13, 44) |
| | $T_{Carl2}$ | 49.049 | 41.317 | 0.450 | (31, 61) | 29.567 | 59.409 | 0.453 | (13, 46) |
| | $T_{Carl3}$ | 49.204 | 155.750 | 0.248 | (20, 77) | 31.903 | 255.981 | 0.262 | (10, 83) |
| | $T_L$ | 48.869 | 32.375 | 0.411 | (36, 60) | 29.963 | 58.299 | 0.423 | (18, 48) |
| | $T$ | 49.808 | 15.434 | 0.503 | (40, 58) | 33.401 | 60.231 | 0.466 | (26, 54) |
| $\mu_0=0$ $\mu_1=2$ | $T_{Hink}$ | 50.057 | 1.769 | 0.844 | (47, 53) | 30.124 | 1.922 | 0.853 | (27, 33) |
| | $T_{Sche}$ | 50.054 | 1.684 | 0.847 | (47, 53) | 30.371 | 2.535 | 0.931 | (28, 35) |
| | $T_{Carl1}$ | 49.938 | 1.760 | 0.847 | (47, 53) | 30.182 | 2.408 | 0.830 | (27, 34) |
| | $T_{Carl2}$ | 49.969 | 2.089 | 0.834 | (47, 53) | 30.158 | 2.266 | 0.844 | (27, 33) |
| | $T_{Carl3}$ | 51.484 | 88.938 | 0.516 | (20, 73) | 31.314 | 114.866 | 0.538 | (10, 62) |
| | $T_L$ | 49.045 | 2.587 | 0.794 | (46, 52) | 29.350 | 2.738 | 0.818 | (27, 33) |
| | $T$ | 49.880 | 1.462 | 0.861 | (47, 52) | 31.097 | 6.803 | 0.763 | (29, 38) |
| $\mu_0=0$ $\mu_1=3$ | $T_{Hink}$ | 50.018 | 0.278 | 0.969 | (49, 51) | 30.003 | 0.297 | 0.969 | (29, 31) |
| | $T_{Sche}$ | 50.013 | 0.257 | 0.968 | (49, 51) | 30.253 | 0.617 | 0.967 | (29, 32) |
| | $T_{Carl1}$ | 49.977 | 0.339 | 0.959 | (48, 51) | 30.167 | 0.483 | 0.954 | (29, 32) |
| | $T_{Carl2}$ | 49.994 | 0.326 | 0.963 | (49, 51) | 30.126 | 0.396 | 0.960 | (29, 32) |
| | $T_{Carl3}$ | 51.456 | 78.832 | 0.596 | (20, 74) | 31.315 | 100.257 | 0.609 | (10, 58) |
| | $T_L$ | 49.011 | 1.256 | 0.930 | (48, 50) | 29.231 | 1.145 | 0.949 | (28, 31) |
| | $T$ | 49.927 | 0.297 | 0.963 | (48, 51) | 30.608 | 1.946 | 0.857 | (30, 34) |
| $\mu_0=0$ $\mu_1=4$ | $T_{Hink}$ | 49.992 | 0.078 | 0.994 | (49, 51) | 29.995 | 0.063 | 0.977 | (29, 30) |
| | $T_{Sche}$ | 49.992 | 0.066 | 0.995 | (49, 50) | 30.253 | 0.467 | 0.952 | (30, 32) |
| | $T_{Carl1}$ | 49.968 | 0.080 | 0.993 | (49, 50) | 30.179 | 0.289 | 0.974 | (30, 32) |
| | $T_{Carl2}$ | 49.977 | 0.083 | 0.994 | (49, 50) | 30.141 | 0.233 | 0.981 | (30, 31) |
| | $T_{Carl3}$ | 50.573 | 85.819 | 0.619 | (20, 68) | 32.184 | 110.642 | 0.634 | (10, 64) |
| | $T_L$ | 48.980 | 1.120 | 0.964 | (49, 50) | 29.237 | 0.943 | 0.988 | (29, 31) |
| | $T$ | 49.928 | 0.114 | 0.991 | (49, 50) | 30.518 | 1.360 | 0.887 | (30, 33) |

표2. n=100 변화점 $\tau=50$, $\tau=30$ 일 때 오차항이 평균 0, 분산 1 인 이중지수분포 따르는 경우

| 변화점 | | $\tau=50$ | | | | $\tau=30$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 평균 | 평균 제곱 오차 | 추정 비율 | 95% 신뢰구간 | 평균 | 평균 제곱 오차 | 추정 비율 | 95% 신뢰구간 |
| $\mu_0=0$ $\mu_1=1$ | $T_{Hink}$ | 49.927 | 32.721 | 0.524 | (37, 63) | 30.723 | 49.799 | 0.497 | (18, 45) |
| | $T_{Sche}$ | 50.153 | 22.243 | 0.561 | (40, 61) | 30.755 | 29.999 | 0.530 | (22, 43) |
| | $T_{Carl1}$ | 49.328 | 26.690 | 0.557 | (36, 58) | 29.594 | 25.168 | 0.521 | (17, 40) |
| | $T_{Carl2}$ | 49.503 | 22.969 | 0.565 | (38, 58) | 29.847 | 20.805 | 0.545 | (19, 39) |
| | $T_{Carl3}$ | 49.361 | 157.905 | 0.317 | (20, 80) | 32.871 | 285.287 | 0.304 | (10, 89) |
| | $T_L$ | 49.018 | 23.280 | 0.530 | (38, 59) | 29.625 | 27.481 | 0.496 | (20, 42) |
| | $T$ | 49.879 | 11.854 | 0.594 | (42, 57) | 32.699 | 33.885 | 0.483 | (26, 46) |
| $\mu_0=0$ $\mu_1=2$ | $T_{Hink}$ | 50.021 | 1.889 | 0.857 | (47, 53) | 30.042 | 1.724 | 0.867 | (27, 33) |
| | $T_{Sche}$ | 50.044 | 1.728 | 0.856 | (47, 53) | 30.028 | 1.700 | 0.933 | (28, 33) |
| | $T_{Carl1}$ | 49.986 | 1.790 | 0.862 | (47, 53) | 30.044 | 1.436 | 0.854 | (27, 33) |
| | $T_{Carl2}$ | 50.015 | 1.553 | 0.871 | (47, 53) | 30.071 | 1.279 | 0.868 | (28, 33) |
| | $T_{Carl3}$ | 50.841 | 108.947 | 0.525 | (20, 75) | 32.271 | 131.569 | 0.509 | (10, 70) |
| | $T_L$ | 49.030 | 2.680 | 0.911 | (46, 52) | 29.224 | 2.174 | 0.847 | (27, 32) |
| | $T$ | 49.920 | 1.322 | 0.943 | (47, 52) | 31.029 | 5.703 | 0.734 | (29, 37) |
| $\mu_0=0$ $\mu_1=3$ | $T_{Hink}$ | 50.010 | 0.296 | 0.968 | (49, 51) | 29.982 | 0.428 | 0.965 | (29, 31) |
| | $T_{Sche}$ | 50.012 | 0.318 | 0.968 | (49, 51) | 30.247 | 0.829 | 0.928 | (29, 33) |
| | $T_{Carl1}$ | 49.972 | 0.312 | 0.968 | (49, 51) | 30.131 | 0.663 | 0.943 | (29, 33) |
| | $T_{Carl2}$ | 49.984 | 0.272 | 0.971 | (49, 51) | 30.106 | 0.550 | 0.952 | (29, 32) |
| | $T_{Carl3}$ | 51.231 | 81.445 | 0.605 | (20, 72) | 31.443 | 109.161 | 0.586 | (10, 58) |
| | $T_L$ | 49.012 | 1.294 | 0.939 | (48, 50) | 29.225 | 1.337 | 0.945 | (28, 31) |
| | $T$ | 49.938 | 0.336 | 0.963 | (49, 51) | 30.685 | 2.465 | 0.832 | (30, 35) |
| $\mu_0=0$ $\mu_1=4$ | $T_{Hink}$ | 49.996 | 0.100 | 0.990 | (49, 51) | 30.011 | 0.123 | 0.988 | (29, 31) |
| | $T_{Sche}$ | 50.007 | 0.107 | 0.990 | (49, 51) | 30.211 | 0.417 | 0.951 | (30, 32) |
| | $T_{Carl1}$ | 49.983 | 0.115 | 0.988 | (49, 51) | 30.155 | 0.321 | 0.963 | (30, 32) |
| | $T_{Carl2}$ | 49.994 | 0.118 | 0.989 | (49, 51) | 30.132 | 0.266 | 0.970 | (30, 32) |
| | $T_{Carl3}$ | 51.009 | 87.455 | 0.602 | (20, 70) | 32.129 | 112.749 | 0.617 | (10, 64) |
| | $T_L$ | 49.001 | 1.109 | 0.961 | (48, 50) | 29.205 | 1.013 | 0.978 | (29, 31) |
| | $T$ | 49.931 | 0.131 | 0.988 | (49, 51) | 30.593 | 2.019 | 0.687 | (30, 31) |

**표3.** n=100, 변화점 $\tau=50$, $\tau=30$ 일 때 오차항이 균일분포 U(-1.7,1.7) 따르는 경우

| 변화점 | | $\tau=50$ | | | | $\tau=30$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 평균 | 평균 제곱 오차 | 추정 비율 | 95% 신뢰구간 | 평균 | 평균 제곱 오차 | 추정 비율 | 95% 신뢰구간 |
| $\mu_0=0$ $\mu_1=1$ | $T_{Hink}$ | 50.000 | 33.914 | 0.481 | (36, 64) | 31.033 | 58.581 | 0.465 | (19, 48) |
| | $T_{Sche}$ | 49.947 | 43.497 | 0.457 | (35, 65) | 31.853 | 82.997 | 0.430 | (18, 57) |
| | $T_{Carl}$ | 48.745 | 55.003 | 0.439 | (26, 61) | 29.551 | 81.111 | 0.411 | (12, 48) |
| | $T_{Carl2}$ | 48.822 | 61.284 | 0.427 | (25, 64) | 29.910 | 91.300 | 0.394 | (12, 52) |
| | $T_{Carb}$ | 49.255 | 158.521 | 0.247 | (20, 77) | 31.860 | 238.300 | 0.250 | (10, 83) |
| | $T_L$ | 48.834 | 44.804 | 0.406 | (33, 63) | 30.555 | 80.431 | 0.409 | (15, 56) |
| | $T$ | 49.588 | 20.450 | 0.487 | (39, 59) | 33.729 | 69.823 | 0.411 | (24, 54) |
| $\mu_0=0$ $\mu_1=2$ | $T_{Hink}$ | 49.995 | 1.815 | 0.843 | (47, 53) | 30.031 | 1.455 | 0.848 | (27, 33) |
| | $T_{Sche}$ | 49.988 | 2.014 | 0.847 | (47, 53) | 30.322 | 2.420 | 0.825 | (28, 34) |
| | $T_{Carl}$ | 49.920 | 1.944 | 0.846 | (46, 53) | 30.160 | 1.988 | 0.828 | (27, 34) |
| | $T_{Carl2}$ | 49.929 | 2.185 | 0.842 | (46, 53) | 30.126 | 2.210 | 0.822 | (27, 34) |
| | $T_{Carb}$ | 50.918 | 83.588 | 0.537 | (20, 72) | 31.894 | 110.854 | 0.510 | (10, 60) |
| | $T_L$ | 48.965 | 3.109 | 0.760 | (45, 52) | 29.296 | 2.664 | 0.774 | (27, 33) |
| | $T$ | 49.828 | 1.728 | 0.856 | (47, 52) | 31.053 | 6.371 | 0.737 | (29, 37) |
| $\mu_0=0$ $\mu_1=3$ | $T_{Hink}$ | 50.026 | 0.188 | 0.986 | (49, 51) | 29.968 | 0.188 | 0.985 | (29, 31) |
| | $T_{Sche}$ | 50.005 | 0.163 | 0.986 | (49, 51) | 30.201 | 0.383 | 0.967 | (30, 32) |
| | $T_{Carl}$ | 49.970 | 0.156 | 0.987 | (49, 51) | 30.150 | 0.258 | 0.973 | (30, 32) |
| | $T_{Carl2}$ | 49.994 | 0.172 | 0.986 | (49, 51) | 30.115 | 0.207 | 0.981 | (30, 31) |
| | $T_{Carb}$ | 51.450 | 81.900 | 0.612 | (20, 72) | 31.903 | 116.073 | 0.625 | (10, 59) |
| | $T_L$ | 49.001 | 1.163 | 0.939 | (48, 50) | 29.191 | 0.977 | 0.989 | (29, 31) |
| | $T$ | 49.939 | 0.139 | 0.992 | (49, 51) | 30.483 | 1.331 | 0.894 | (30, 34) |