The Korean Communications
in Statistics Vol. 10 No. 2, 2003
pp. 353-359

Tests for Normal Mean Change with the Mean Difference[1)]

Jaehee Kim[2)] and PilKyoung Yun[3)]

Abstract

This paper deals with the problem of testing mean change with one change-point with the normal random variables. We propose a test with the mean difference for change in a location parameter. A power comparison study of various change-point test statistics is performed via Monte Carlo simulation with S-plus software.

Keywords : test, location parameter, change-point, mean difference, power

1. 서론

품질관리분야에서 시작된 변화점 문제는 자료의 변화에 관심이 있는 여러 분야에서 다루어지고 있으며 변화에 대한 통계적 분석의 필요성이 요구되고 있다. 특히, 자료에서 변화에 대한 추론, 즉 변화존재여부에 대한 검정, 변화점 추정에 대한 방법은 계속 연구되고 새로운 방법이 제안되고 있다. 변화검정후 자료 내에서 변화가 있는 경우, 변화의 흐름을 파악하고 그 변화시점을 찾아내어 변화원인에 대해 분석하고 그에 대한 대안을 제시하는 것은 매우 중요한 작업이다.

변화점(change-point)이란 시간의 흐름에 따라 일정하게 관측하여 기록된 자료에서 변화가 일어나는 시점을 말한다. 예를 들어 매일 증권 거래 마감시간에 종합주가지수를 일정기간 동안 관측하여 기록할 때 평균 종합주가지수의 변화여부와 변화가 있을 경우 변화점 추정은 큰 관심 문제가 된다.

본 연구에서는 자료의 위치모수에 변화점이 1개 있는 경우 평균차와 그 합을 이용한 검정통계량을 제안하고자한다. 본 논문의 구성은 다음과 같다. 2장에서는 기존에 제안된 변화점 검정통계량들을 소개하고 3장은 두 부분의 평균의 차의 제곱합을 이용한 새로운 검정통계량을 제안하고 그 특징들에 대해 설명한다. 그리고 4장에서는 S-plus를 이용한 모의실험을 통해 변화점 검정통계량의 검정력을 비교한다. 마지막 5장에서는 본 논문의 결론 및 제안을 한다.

2. 통계적 모형과 기존 변화점 검정통계량

연속적인 시간에 의해 관측 기록된 자료를 확률변수라 할 때 변화가 일어나는 시점 즉 위치모

1) This research was supported by R04-2001-000-00135-0(2002) KOSEF.
2) Associate Professor, Department of Statistics, Duksung Women's University, Ssangmun-Dong 419 Tobong-Gu, Seoul, Korea. jaehee@duksung.ac.kr
3) Master in Statistics, Department of Statistics, Duksung Women's University, Seoul, Korea pil-suni@hanmail.net

수의 변화, 분산모수의 변화 또는 분포의 변화가 발생하는 시점을 가리킨다.

변화점 연구에서 고려되어지는 방법으로는, 매 시점에서 자료를 두 부분으로 나누었을 때 두 부분의 비균일성을 최대로 하는 측도를 이용하여 변화 검정통계량이 개발되었으며, 이 절에서 기존 변화점 통계량을 소개하고자한다.

확률표본 $X_1, X_2, \ldots, X_n$ 은 $E(X_i)=\mu_i$이고 $Var(X_i)=\sigma^2$이고 서로 독립인 정규분포를 따른다. 변화점 검정에 대한 가설은

$$\begin{aligned} H_0 &: \mu_1=\mu_2=\cdots=\mu_n=\mu \\ H_1 &: \mu_1=\mu_2=\cdots=\mu_\tau \neq \mu_{\tau+1}=\cdots=\mu_n, \quad \tau\in\{1,\ldots,n-1\} \end{aligned} \tag{1}$$

이다.

Chernoff 와 Zacks(1964)는 μ_1 이 $N(0,a^2)$ 의 사전분포를 따를 때 Bayes 검정을 이용하여

$$T_{CZ}=\begin{cases} \sum_{i=1}^{n-1}(i+1)X_i & \mu_1 \text{ 알려진 경우} \\ \sum_{i=1}^{n-1}(i+1)(X_i-\overline{X_n}), & \mu_1 \text{ 모르는 경우} \end{cases} \tag{2}$$

의 검정통계량을 제안하였다.

Gardner(1969)는 균일 사전분포를 따를 때 다음의 검정통계량을 유도하였다:

$$T_{Ga}=\frac{1}{n^2}\sum_{i=1}^{n-1}\left[\sum_{j=i}^{n-1}(X_{j+1}-\overline{X_n})\right]^2. \tag{3}$$

Brown(1975)은 다음과 같은 재귀적 잔차(recursive residual)

$$\begin{aligned} z_i&=\left\{\frac{n}{(n+1)}\right\}^{\frac{1}{2}}(X_{i+1}-\overline{X_i}), \\ \overline{X_i}&=\frac{1}{i}\sum_{j=1}^{i}X_j \qquad i=1,\ldots,n \end{aligned} \tag{4}$$

과 누적합 $\tilde{S}_i=z_1+\cdots+z_i$ 으로 재귀적 잔차통계량을 이용한 검정통계량을 제안하였다:

$$T_B=\max_{n_0\le i\le n}\frac{|\tilde{S}_{n-1}-\tilde{S}_{n-i}|}{\sqrt{i-1}}. \tag{5}$$

Sen 과 Srivastava(1975)가 최대우도비함수를 이용하여 다음의 검정통계량을 제안하였다 :

$$T_S=\max_{1\le i\le n-1}\frac{\left\{\frac{S_n-S_i}{(n-i)}-\frac{S_i}{i}\right\}}{\sqrt{\frac{n}{i(n-i)}}}. \tag{6}$$

Hawkins(1977)는 μ_1과 μ_n의 최대우도추정량을 구하고 분산분석을 고려하여 검정통계량을 제안하였다:

$$T_H=\frac{n}{i(n-i)}\left\{\sum_{j=1}^{i}(X_j-\overline{X_n})\right)^2\Big\}. \tag{7}$$

Pettitt(1980)는 점수함수의 누적합을 이용한 검정통계량을 제안하였다:

$$T_P=\max_{1\le i\le n}\left|i\frac{S_n}{n}-S_i\right|. \tag{8}$$

James(1987)는 Brown(1975)의 재귀적 잔차통계량을 이용한 검정통계량을 변형하여 검정통계량

$$T_J = \max_{n_0 \le k < n} \frac{(\widetilde{S_{n-1}} - \widetilde{S_{n-k}})}{(k-1)^{1/2}} \tag{9}$$

을 제안하였다.

Gombay(1990)는 우도비 검정 통계량으로부터 함수를 이용하여 다시 표현할 수 있다는 점에 착안하여 검정통계량을 제안하였다. 함수의 형태 $g(t) = t^2/2$ 일 때는 검정통계량으로

$$T_{GO} = \max_{n_0 \le i \le n_1} \frac{(nS_i - iS_n)^2}{ni(n-i)} \tag{10}$$

를 얻게되고 $g(t) = \exp(t)$일 경우에는 검정통계량으로

$$T_{GO2} = \max_{n_0 \le i \le n_1} \frac{\{i\exp(\overline{X_i}) + i^*\exp(\overline{X_i^*}) - n\exp(\overline{X_n})\}}{\exp(\overline{X_i})}$$

가 얻어진다. 귀무가설하에서, T_{GO}의 점근분포는

$$\sup_{\lambda_1 \le t \le 1-\lambda_2} \frac{\{W(t) - tW(t)\}^2}{t(1-t)}$$

이며, 여기서 $W(t)$는 Wiener process이다.

3. 평균차 제곱을 이용한 검정통계량

확률표본 $X_1, X_2, \ldots, X_n$ 은 $E(X_i) = \mu_i$이고 $Var(X_i) = \sigma^2$이고 서로 독립인 정규분포를 따른다. 변화점 검정에 대한 가설은 (1)에 나타나 있으며, 평균변화가 한 번 일어나는 분포를 따를 때 통계적 모형은 다음과 같다:

$$X_i = \begin{cases} \mu_0 + \varepsilon_i, & 1 \le i \le \tau \\ \mu_0 + \delta + \varepsilon_i, & \tau + 1 \le i \le n. \end{cases} \tag{11}$$

여기서 $\delta \neq 0$, τ는 변화점이다. 오차항 ε_i는 평균이 0, 분산이 σ^2 인 서로 독립인 정규분포를 따른다.

자료가 평균이 다른 두 부분으로 나뉠 때 변화점에서 두 부분의 평균의 차가 최대가 되는 것과 매 시점의 차이를 이용하여 검정통계량을 제안하고자 한다. 매 시점에서 두 부분의 평균을 구하여 두 평균의 차를 제곱한 값을 구하고 이 값에 분산을 이용한 가중치를 곱하고 모든 시점에서의 합으로 변화점 존재 여부에 대한 새로운 변화점 검정통계량으로 제안한다:

$$T_n = \frac{1}{n-1} \sum_{i=1}^{n-1} (\overline{X_i} - \overline{X_i^*})^2 \frac{i(n-i)}{n}.$$

여기서 n 은 데이터의 수, $\overline{X_i} = (X_1 + \ldots + X_i)/i$, $\overline{X_i^*} = (X_{i+1} + \ldots + X_n)/(n-i)$ 이다.

제안하는 변화점 검정통계량 T_n에서 성분 $(\overline{X_i} - \overline{X_i^*})$은 매 시점에서 시점의 앞부분 자료와 시점의 뒷부분 자료의 평균차가 된다. 귀무가설하에서 $(\overline{X_i} - \overline{X_i^*})$의 기대값과 분산은

$$E(\overline{X_i} - \overline{X_i^*}) = E\left[\frac{1}{i}\sum_{j=1}^{i} X_j - \frac{1}{(n-i)}\sum_{j=i+1}^{n} X_j\right] = 0,$$

$$V(\overline{X_i}-\overline{X_i^*}) = V\left[\frac{1}{i}\sum_{j=1}^{i}X_j-\frac{1}{(n-i)}\sum_{j=i+1}^{n}X_j\right]=\frac{1}{i}\sigma^2+\frac{1}{(n-i)}\sigma^2=\frac{n}{i(n-i)}\sigma^2$$

이다. 표준화 과정을 고려하여 매 시점에서 평균차의 제곱값에 평균차의 분산을 나누어주고 모든 시점에서의 합을 구한다. 이 과정을 통해서 제안하는 변화점 검정통계량 T_n을 얻게 된다.

귀무가설하에서 제안하는 변화점 검정통계량 T_n의 기댓값을 구하면

$$E(T_n) = E\left[\frac{1}{n-1}\sum_{i=1}^{n-1}(\overline{X_i}-\overline{X_i^*})^2\frac{i(n-i)}{n}\right]$$
$$=\frac{1}{n-1}\sum_{i=1}^{n-1}E\left[\left\{\frac{1}{i^2}\sum_{j=1}^{i}X_j^2+\frac{1}{(n-i)^2}\sum_{j=i+1}^{n}X_j^2\right\}\frac{i(n-i)}{n}\right]=\sigma^2$$

이 되어 표본의 크기에 의존하지 않으며 일정한 값을 가짐을 알 수 있다.

제안하는 변화점 검정통계량 T_n의 주요성분인 d_i와 d_i의 누적합 cd_j

$$d_i=(\overline{X_i}-\overline{X_i^*})^2\frac{i(n-i)}{n}, \qquad cd_j=\sum_{i=1}^{j}d_i, \quad j=1,\ldots,n$$

의 움직임을 그래프를 통해 알아보고자 한다. [그림 1]은 귀무가설하에서의 d_i와 cd_j의 움직임을 나타내는 그래프이다. 평균의 변화가 없는 변화점 검정통계량 T_n의 주요성분 d_i와 cd_j의 움직임에는 별다른 경향성이 없어 보이며 뚜렷하게 큰 값이 나타나지 않는다. [그림 2]는 $n=100$일 때 변화점 $\tau=50$에서 평균의 변화가 있는 검정통계량 T_n의 주요성분 d_i와 cd_j의 움직임을 나타내는 그래프이다. 변화점 $\tau=50$에서 d_i가 큰 값을 갖는다. 변화점 $\tau=50$ 근처에서 cd_j의 기울기가 커져 변화점 가능성을 시사한다.

4. 모의실험

본 절에서는 변화점 검정능력을 비교하기 위하여 S-plus를 이용한 모의실험을 통하여 검정통계량의 검정력(power)을 비교한다.

변화점 검정에 대한 가설에 대해 평균변화가 한 번 존재하는 분포를 따를 때 다음의 통계적 모형을 고려한다:

$$X_i=\begin{cases}\mu_0+\varepsilon_i, & 1\le i\le\tau\\ \mu_0+\delta+\varepsilon_i, & \tau+1\le i\le n.\end{cases}$$

여기서 δ는 0이 아닌 상수, τ는 변화점이다. 오차항 ε_i는 $E(\varepsilon_i)=0$, $Var(\varepsilon_i)=1$이고 서로 독립인 정규분포를 따른다.

우선 귀무가설하에서의 모의실험을 통해 $\alpha=0.05$ 에 해당하는 경험적 기각치를 구한다. 대립가설 상황으로는 표본의 크기는 $n=50$ 인 경우 변화점이 $\tau=10,20,25,35,40$ 에서 발생한 경우, $n=100$ 인 경우에는 변화점 $\tau=20,40,50,60,80$ 에서 발생했을 때, 변화량으로는 $\mu_0=0$ 에 대하여 위치모수의 차이 $\delta=0.5,1.0,1.5$ 로 변화시켜가면서 1,000번의 반복실험을 통해 변화점 검정통계량들의 검정력(power)을 구한다.

표 1을 보면 유의 수준 $\alpha=0.05$하에서, 표본의 크기가 50인 경우, 변화점이 앞부분 $\tau=10$

에서 일어날 때 T_{G2}의 검정력이 우수하고 변화점이 자료의 중간 ($\tau=25$)에서 발생했을 때는 T_{GA}, T_P, T_n의 검정력이 우수하다. 변화점이 자료의 뒷부분($\tau=40$)에서 일어날 때는 T_B의 검정력이 우수하다. 표본의 크기가 100인 경우의 검정력을 보여주는 표 2에서도 비슷한 경향을 보여준다. 표본의 크기가 클수록, 변화점이 자료의 앞뒤부분에서 발생했을 때는 T_{G2}를 제외한 상황에서는 약간 우수함을 볼 수 있으며 자료의 중간에서 변화가 발생했을 때 검정력이 기존 통계량에 비해 크게 떨어지지 않음을 알 수 있다. 그리고 기존의 통계량과 마찬가지로 자료의 중간에서 변화점이 발생했을 때 높은 검정력을 보여준다.

제 5 장 결론 및 제안

본 연구에서는 정규분포를 따르는 자료 내에서 위치모수의 변화가 있는지에 대한 검정에 대하여 평균차의 제곱에 분산을 이용한 가중치를 곱한 새로운 검정통계량을 제안하였고 모의실험을 통해 기존 통계량과 비교하였다. 여기서 제안한 통계량은 비모수적 방법으로의 확장이 가능하며 이에 대한 연구를 기대한다.

참고문헌

[1] 장희윤(1999), 위치모수의 변화가 있는 경우 비모수적 변화점 추정통계량에 관한 연구, 덕성여자대학교 석사논문.

[2] 서현주(2002), 점수함수를 이용한 비모수적 변화점 추정, 덕성여자대학교 석사논문.

[3] Brown, R. L., Durbin, J. and Evans J. M.(1975), Techniques for testing the constancy of regression relationships over time, *Journal of Royal Statistical Society*. B 37, 149-92.

[4] Chernoff, H. and Zacks, S.(1964), Estimating the current mean of a normal distribution which is subjected to changes in time, *The Annals of Mathematical Statistics*.35, 999-1028.

[5] Gardner, L. A.(1969), On detecting changes in the mean of normal variates, *The Annals of Mathematical Statistics* 40, 1, 116-126.

[6] Gombay, E.(1990), Asymptotic distributions of maximum likelihood tests for change in the mean, *Biometrika,* 77, 2, 411-414.

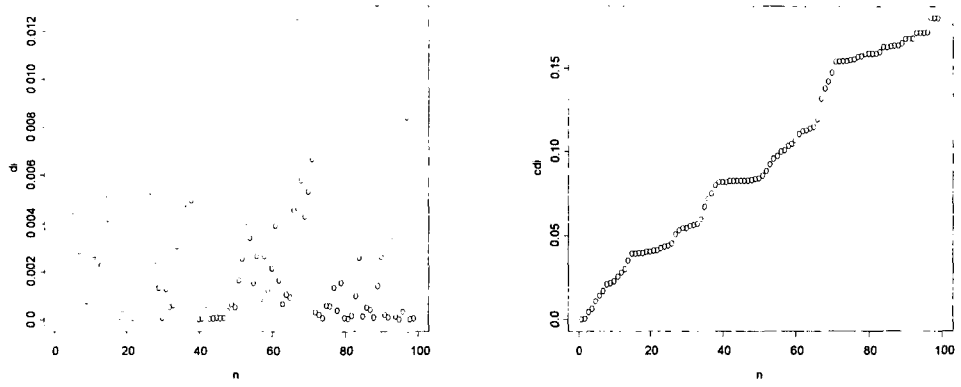
[7] Hawkins, D. M.(1977), Testing a sequence of observations for a shift in location, *Journal of the American Statistical Association,* 15, 655-679.

[8] James, B., James, K. L. and Siegmund, D.(1987), Tests for a Change-point, *Biometrika,* 74, 1, 71-83.

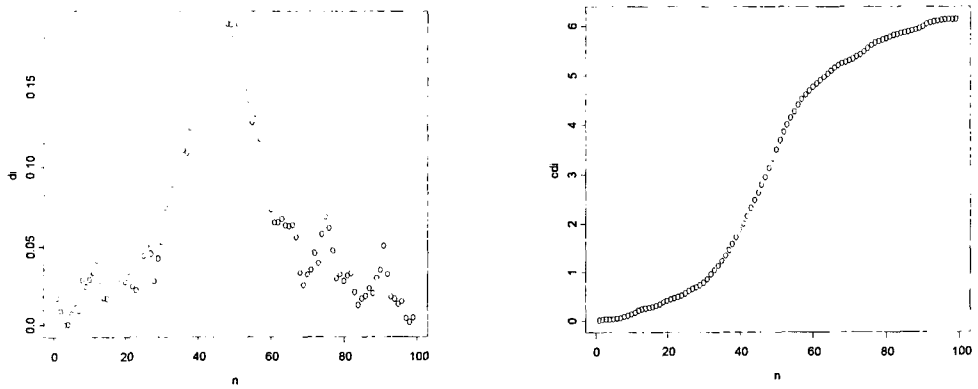
[9] Pettitt, A. N.(1980), A simple cumulative sum type statistic for the change-point problem with zero-one observation, *Biometrika,* 67, 1, 79-84.

[10] Sen, A. and Srivastava, M. S.(1975), On tests for detecting change in mean, *The Annals of Statistics*. 3, 98-108

[2003년 3월 접수, 2003년 6월 채택]



[그림 1] 귀무가설하에서(변화점이 없을 때), $n=100$인 경우 통계량 d_i와 cd_j의 움직임 그래프



[그림 2] $n=100, \tau=50, \delta=1.0$일 때 통계량 d_i와 cd_j의 움직임 그래프

표 1. $n=50$ $\alpha=0.05$일 때 검정력 비교

δ		0.5			1.0			1.5		
변화점(τ)		10	25	40	10	25	40	10	25	40
$n=50$ $\alpha=.05$	T_{CZ}	0.170	0.335	0.173	0.508	0.847	0.521	0.847	0997	0.859
	T_{GA}	0.176	0.353	0.176	0.546	0.871	0.566	0.890	0.999	0.915
	T_B	0.142	0.275	0.212	0.487	0.819	0.679	0.883	0.997	0.962
	T_S	0.192	0.252	0.180	0.606	0.802	0.628	0.951	0.996	0.951
	T_H	0.164	0.256	0.187	0.617	0.846	0.621	0.941	0.998	0.944
	T_P	0.166	0.342	0.164	0.546	0.880	0.568	0.922	0.999	0.940
	T_J	0.192	0.252	0.180	0.606	0.802	0.628	0.951	0.996	0.951
	T_{GO}	0.192	0.252	0.180	0.606	0.802	0.628	0.951	0.996	0.951
	T_{GOL}	0.261	0.284	0.208	0.693	0.808	0.639	0.959	0.996	0.953
	T_n	0.184	0.328	0.198	0.585	0.853	0.604	0.922	0.997	0.933

표 2. $n=100$, $\alpha=0.05$일 때 검정력 비교

δ		0.5			1.0			1.5		
변화점(τ)		20	50	80	20	50	80	20	50	80
$n=100$ $\alpha=.05$	T_{CZ}	0.268	0.583	0.309	0.815	0.989	0.788	0.991	0.999	0.979
	T_{GA}	0.283	0.624	0.306	0.854	0.993	0.847	0.997	0.999	0.997
	T_B	0.218	0.504	0.368	0.835	0.985	0.914	0.997	0.999	0.999
	T_S	0.304	0.501	0.321	0.906	0.985	0.895	0.999	0.999	0.99
	T_H	0.303	0.452	0.330	0.912	0.992	0.992	0.999	0.999	0.999
	T_P	0.266	0.636	0.330	0.877	0.992	0.870	0.999	0.999	0.999
	T_J	0.304	0.501	0.321	0.906	0.985	0.895	0.999	0.999	0.999
	T_{GO}	0.304	0.501	0.321	0.906	0.985	0.895	0.999	0.999	0.999
	T_{GOL}	0.354	0.484	0.314	0.928	0.979	0.881	0.999	0.999	0.999
	T_n	0.310	0.583	0.331	0.884	0.990	0.876	0.997	0.999	0.998