變化時點 問題에 대한 統計的 檢定과 事例 分析

金 善 佑*

時間이 흐름에 따라서 順次的으로 추출한 標本들에서 變化가 발생하는 경우 그 變化時點을 알고 있다면 단순한 두 標本集團의 問題가 되어 이는 두 標本 比較問題로 귀착된다. 그러나 變化時點을 모르고 있다는 假定은 새로운 檢定問題를 제기하게 되는데 이러한 問題를 變化時點 問題(change point problem)라 지칭한다. 기존의 많은 分析들은 단순한 數値上의 差異 등으로써 時間의 흐름에 따르는 變化를 서술하는데 그쳐왔다. 따라서 먼저 分析對象 資料들의 變化 有無를 檢定해야 하며 만약 變化가 발생했다면 그 變化時點을 推定할 필요가 있다. 본 논문에서는 이에 대한 統計的 方法을 서술하고 變化時點 問題에 대한 事例들을 分析하였다.

I. 緖 論

독립 확률 변수 x_1, x_2, …, x_N은 時間이 흐름에 따라서 順次的으로 발생되는 변수이며, x_1, x_2, …, x_r은 분포함수 F(x)를 x_{r+1}, x_{r+2}, …, x_N은 분포함수 $F(x-\Delta)$, $-\infty<\Delta<\infty$을 따른다고 하자. 여기서 정수 r은 未知의 變化時點(change point)이라 하고 未知의 母數 Δ는 變化時點에서의 變化의 크기를 나타낸다. 일반적으로 變化時點이 여러 개가 있을 수 있으나 본 논문에서는 變化時點이 기껏해야 한 개인 경우만을 다루기로 한다. 變化의 有無 檢定을 위한 여러가지 母數的 方法과 非母數的 方法이 있으나 본 논문에서는 Sen과 Srivastava(Sen and Srivastava, 1975 : 593-602)가 제안한 母數的 檢定方法과 Pettitt(Pettitt, 1979 : 126-135)가 제안한 非母數的 檢定方法을 소개한다. 母數的 方法을 이용하려면 변수들의 분포 함수를 알거나 분포 함수에 대한 가정을 할 수 있어야 하나 非母數的 方法은 분포함수에 대한 정보가 불필요하며 變化의 有無를 檢定하고 대략의 變化時點을 찾는데 계산이 쉽고 간편하다. 이러한 變化時點 模型에 실제 적용된 자료들을 보면 평균 신장의 변화, 주가의 변동, 환자들의 병상 이용률의 변화, 일조 시간의 변화 등을 볼 수 있다. 특히 가족 계획 사업 또는 의료보험 실시 등으로 인한 변화를 분석하는데 이러한 變化時點 問題를 適用시켜 볼 수 있다.

* 本院 責任研究員

Ⅱ. 變化時點 問題의 統計的 檢定方法

1. Sen과 Srivastava의 母數的 檢定方法

$x_1, x_2, \cdots, x_N$이 평균$(\mu_1, \mu_2, \cdots, \mu_N)$, 공통분산 σ^2를 갖는 일련의 독립 정규 확률 변수일때 귀무가설 $H_0 : \mu_1=\mu_2=\cdots=\mu_N=\mu$ 對 대립가설 $H_1 : \mu=\mu_1=\cdots=\mu_r<\mu_{r+1}=\cdots\mu_N=\mu+\Delta(\Delta>0)$을 검정한다고 하자. 단, μ, r, Δ은 未知의 값이다. 대부분의 경우 σ를 모르기 때문에 σ를 모를 경우 위의 가설을 검정하기 위한 檢定 統計量을 서술하면 (1), (2), (3)과 같다.

$$S_1=U/V^{1/2},$$

$$(U=\sum_{i=1}^{N-1} i(x_{i+1}-\bar{x}),\ \bar{x}=\sum_{i=1}^{N} x_i/N,$$

$$V=\sum_{i=1}^{N} (x_i-\bar{x})^2/(N-1)) \qquad (1)$$

$$S_2=U/V_1^{1/2},\ (V_1=\sum_{i=1}^{N-1} (x_{i+1}-x_i)^2/2(N-1)) \qquad (2)$$

$$S_3=\sup_{1\le r<N} \{(\overline{x_{N-r}}-\overline{x_r})/[W(1/r+1/(N-r))]^{1/2}\},$$

$$(W=\{\sum_{i=1}^{r} (x_i-\overline{x_r})^2+\sum_{i=r+1}^{N} (x_i-\overline{x_{N-r}})^2\}/(N-2),$$

$$\overline{x_r}=\sum_{i=1}^{r} x_i/r,\ \overline{x_{N-r}}=\sum_{i=r+1}^{N} x_i/(N-r)) \qquad (3)$$

여기서 U는 각 관측치와 표본평균과의 차이값 $(x_{i+1}-\bar{x})$에 가중치(i)를 곱한 값들의 합(i=1, …, N−1)이다. 따라서 귀무가설하에서 U는 작은 값을 갖게 되나 대립가설하에서는 U값이 커지게 된다. 가장 간단한 檢定 統計量인 S_1은 U를 공통분산 σ^2의 한 推定量인 표본분산 V의 제곱근으로 나누어 구해진다. 그런데 대립가설하에서 변화량이 클 때 이 변화량 때문에 V값이 매우 커지게 된다. 한 시점과 그 다음 시점의 관측치들의 차이의 제곱의 합을 이용하여 최소한 변화량 Δ이 큰 값을 갖더라도 이에 덜 민감하도록 만들어진 V_1을 사용한 檢定 統計量이 S_2이다. 그리고 檢定 統計量 S_3는 대립가설하에서 유도된 最大尤度 統計量(maximum likelihood statistic)이며 이는 두 표본 t-統計量의 수정된 형태이다. 더욱이 統計量 S_3를 이용하여 變化有無의 檢定뿐만 아니라 變化時點을 推定할 수가 있다.

標本의 크기가 20 이하이고 變化量 Δ이 작을 때는 統計量 S_1, S_2가 S_3보다 우월한 檢定力을 가지고 있고 標本의 크기가 20 이상이 되거나 變化量 Δ이 클 때는 變化時點이 順次的 標本의 중앙 부분에 위치할 때는 S_1, S_2이 그리고 變化時點이 양 끝 부분에 위치할 때는 S_3의 檢定力이 더 우월하다. N=10, 20, 30인 경우 統計量 S_1, S_2, S_3에 대한 유의수준 5%하에서의 기각치는 다음과 같다.

	S_1	S_2	S_3
N=10	15.8257	19.5288	3.326
N=20	43.4568	45.7087	2.965
N=30	83.3528	84.8233	2.920

2. Pettitt의 非母數的 檢定方法

變化量 Δ에 관한 無變化 對 有機化를 檢定하기 위해서 變化時點을 알고 있는 경우의 非母數的 方法인 Mann-Whitney의 檢定 統計量을 적절히 변형하여 變化時點을 모르는 경우의 檢定 統計量이 제안되었고 이에 대한 近似 有意確率(approximate significance probability)이 유도되었다. 귀무가설 $H_0 : \Delta=0$(無變化) 對 대립가설 $H_1 : \Delta \neq 0$(有變化)를 檢定하기 위한 양측 검정 통계량은 다음과 같다.

$$P=\max_{1\le r<N} |U_{r,N}|$$

$$(U_{r,N}=\sum_{i=1}^{r}\sum_{j=r+1}^{N} D_{ij},\ D_{ij}=\mathrm{sgn}(x_i-x_j),$$

$$\mathrm{sgn}(x_i - x_j) = \begin{cases} 1 ; x_i > x_j, \\ 0 ; x_i = x_j, \\ -1 ; x_i < x_j \end{cases} \quad (4)$$

단측 검정 $H_0 : \Delta = 0$ 對 $H_1 : \Delta > 0$과 $H_0 : \Delta = 0$ 對 $H_1 : \Delta < 0$을 검정하기 위한 단측 검정 통계량은 각각 (5)와 (6)이다.

$$P^+ = -\min_{1 \le r < N} U_{r,N} \quad (5)$$

$$P^- = \max_{1 \le r < N} U_{r,N} \quad (6)$$

검정의 판단을 위해 Pettitt가 유도한 近似 有意確率은 다음과 같다.

檢定 統計量	近似 有意確率
P	$2\exp\{-6P^2/(N^3+N^2)\}$
P^+	$\exp\{-6P^{+2}/(N^3+N^2)\}$
P^-	$\exp\{-6P^{-2}/(N^3+N^2)\}$

Ⅲ. 事例 分析

실제로 키, 몸무게, I.Q.와 같은 자료들은 정규 분포를 따른다고 볼 수가 있어서 이러한 자료들은 母數的인 變化時點 問題로 취급하여 자료 분석을 할 수가 있다. 이에 대한 適用 事例로는 1975년부터 1984년까지의 우리나라 중·고등학생의 평균 신장에 관한 분석이다. 母數的인 檢定 統計量을 이용하여 檢定한 결과 남, 녀 모두 변화가 있었으며 推定된 變化時點은 남자 중·고등학생은 1980년부터 평균 신장이 변화하였으며 여자 중·고등학생의 경우 變化時點은 각각 1979년, 1982년으로 나타났다(강창완, 1987). 그러나 우리 주변의 정규 분포를 따르지 않는 많은 자료들은 非母數的인 變化時點 問題로 취급하여 분석하는 것이 타당하다. 이에 대한 適用 事例로는 1970년부터 1985년까지의 우리나라 병원의 환자들의 병상 이용률 변화이다. 非母數的 檢定方法을 이용하여 분석한 결과 이 기간내에 변화가 있었으며 變化時點은 1977년으로 나타났다. 이는 의료보험 실시로 인하여 1977년부터 병상 이용률이 증가했을 것으로 생각될 수 있다(백구환, 1987).

본 논문에서 다루고자 하는 事例는 1969년부터 1988년까지 우리나라 병원의 1일 평균 외래 환자수의 변화와 1981년부터 1990년까지 우리나라에서 당뇨병으로 인한 사망률의 변화, 그리고 1983년부터 1990년까지 우리나라 여자의 자궁경부암으로 인한 사망률의 변화이다. 이들 자료들은 時間이 흐름에 따라서 順次的으로 뽑은 서로 독립인 標本이고 여기서 變化가 일어났다면 變化가 한번 일어날 것으로 가정한다. 이들 자료들을 정리하면 〈표 1〉, 〈표 2〉 그리고 〈표 3〉과 같다.

〈표 1〉로부터 1일 평균 외래 환자 수는 증가 추세에 있음을 볼 수 있으므로 $H_0 : \Delta = 0$ 對 $H_1 : \Delta > 0$와 같은 가설을 검정하기로 한다. 더욱이 1일 외래 환자 수는 정규 분포를 따른다고 가정할 수 있고 標本의 크기가 20이고 變化量이 크므로 母數的 檢定 統計量 S_2와 S_3를 이용하기로 한다.

Table 1. Out-Patients Average/day
1일 평균 외래 환자 수

Year	1969	1970	1971	1972	1973
Patients	18,602	19,537	18,866	15,296	21,827
Year	1974	1975	1976	1977	1978
Patients	21,298	24,834	27,423	32,363	40,861
Year	1979	1980	1981	1982	1983
Patients	54,955	59,864	62,382	76,704	94,298
Year	1984	1985	1986	1987	1988
Patients	106,115	120,081	107,993	92,586	134,922

Source : National Bureau of Statistics Economic Planning Board, *Korea Statistical Yearbook*, 1970, 1979, 1989.

Table 2. Death rate from Diabetes Mellitus(per 100,000 persons)
당뇨병으로 인한 사망률(10만당)

Year	1981	1982	1983	1984	1985
Death rate	3.43	4.58	4.30	5.43	6.87
Year	1986	1987	1988	1989	1990
Death rate	7.65	7.72	7.42	9.38	11.83

Table 3. Death rate from Malignant Neoplasm of Uterus(per 100,000 persons)
자궁경부암으로 인한 사망률(10만당)

Year	1983	1984	1985	1986	1987
Death rate	7.09	6.97	7.67	7.94	7.57
Year	1988	1989	1990		
Death rate	7.55	8.38	7.91		

Note : Cause of specific death rate is prorated according to the total number of registered deaths.

Source : National Bureau of Statistics Economic Planning Board, "Deaths from causes(Special list of 50 causes)", *Causes of death statistics,* 1981, 1982.
National Bureau of Statistics Economic Planning Board, "Deaths from causes(The 124-abridged tabulation list of mortality for Korea)", *Causes of death statistics,* 1983, 1984.
National Bureau of Statistics Economic Planning Board, "Deaths from causes(The 124-abridged tabulation list of mortality for Korea)", *Annal report on the cause of death statistics,* 1985-1989.
National Statistical Office, "Deaths from causes(The 124-abridged tabulation list of mortality for Korea)", *Annual report on the cause of death statistics,* 1990.

S_2 값을 구해보면 71.4850이고 N=20일 때 유의수준 5%하에서 기각치는 45.7087이므로 귀무가설을 기각하게 된다. 따라서 1일 평균 외래 환자수에 있어서 변화가 있었다고 볼 수 있다. 덧붙여 統計量 S_3 값을 구할 수 있는데 그 값은 8.7822로 계산된다. 이 경우도 마찬가지로 유의수준 5%에서 기각치가 2.965이기 때문에 귀무가설을 기각하게 된다. 더욱이 統計量 S_3을 통하여 變化時點을 推定할 수 있는데 變化時點이 13번째일 때 S_3의 값이 8.7822이므로 1981년부터 1일 평균 외래 환자수가 증가하였다고 말할 수 있다.

당뇨병의 경우 사망률이 증가 추세에 있어서 $H_0 : \Delta=0$ 對 $H_1 : \Delta>0$와 같은 가설을 세우고 자궁경부암의 경우 사망률의 변동이 일정한 추세를 보이고 있지 않아 變化의 有無를 檢定하는데 있어서 $H_0 : \Delta=0$ 對 $H_1 : \Delta \neq 0$와 같은 가설을 설정하게 된다. 이 두가지 사망 원인에 의한 사망률의 變化 有無 檢定方法은 母數的 檢定方法을 적용할 수 없으므로 非母數的 統計量을 사용하여 檢定하고 變化가 일어났다면 變化時點을 推定해 보고자 한다. Pettitt의 統計量 (4), (5), (6)을 그대로 사용하면 계산이 복잡하다. 그래서 $U_{r,N}=U_{r-1,N}+V_{r,N}$($V_{r,N}=\sum_{j=1}^{N}$ sgn (x_r-x_j))를 이용하면 계산이 편리하므로 이것을 이용하여 統計量의 값을 구하기로 하자. 〈표 2〉와 〈표 3〉의 자료를 사용하여 統計量을 구하기 위하여 먼저 구한 $U_{r,N}$의 값은 〈표 4〉와 〈표 5〉와 같다.

당뇨병의 경우 이용되는 統計量은 $P^+=-\min_{1\le r<N} U_{r,N}$이고 〈표 4〉로부터 구해지는 P^+값과 이에 대한 近似 有意確率은 각각 25와 0.0331이 되어 유의수준을 5%로 택할 때 변화가 없다는 귀무가설을 기각하게 된다. 推定되는 變化時點은 $P^+=25$일때 r의 값이므로 1985년이 된다. 자궁경부암의 경우 대립가설이 $H_1 : \Delta \neq 0$이므로 統

Table 4. The values of $U_{r,N}$(Diebetes Mellitus)
$U_{r,N}$의 값(당뇨병)

r	1	2	3	4	5	6	7	8	9	10
$U_{r,N}$	−9	−14	−21	−24	−25	−22	−17	−16	−9	0

Table 5. The values of $U_{r,N}$(Malignant Neoplasm of Uterus)
$U_{r,N}$의 값(자궁경부암)

r	1	2	3	4	5	6	7	8
$U_{r,N}$	−5	−12	−11	−6	−7	−10	−3	0

計量 $P=\max_{1\le r<N} |U_{r,N}|$을 이용하여야 한다. 〈표 5〉로부터 P값은 12로 구해지고 이에 대한 近似 有意確率은 0.4463이 되어 1983년부터 1990년 사이에 자궁경부암으로 인한 사망률에는 변화가 없다는 귀무가설을 받아들이게 된다.

Ⅳ. 結　論

본 논문에서는 變化時點을 모르고 變化時點이 기껏해야 한 개인 경우의 變化時點 問題에 관해서 대표적인 母數的 檢定 統計量과 非母數的 檢定 統計量을 소개하였다. 관측치가 정규 분포를 따르거나 또는 정규 분포를 따른다고 가정할 수 있을 때는 母數的 檢定方法을 사용하는 것이 좋다. 왜냐하면 非母數的 檢定 統計量은 관측치들의 크기를 비교함으로써 계산되는 반면 母數的 檢定 統計量은 관측치들의 값들을 그대로 이용하여 구해지므로 그만큼 자료에 대한 정보의 손실이 적기 때문이다. 실제로 관측치가 정규 분포를 따른다는 가정하에서는 母數的 檢定 統計量의 檢定力이 非母數的 檢定 統計量의 檢定力보다 우월하다. 그러나 非母數的 檢定方法은 母數的 檢定方法에 비해 계산이 간편하고 쉬우며 더욱이 정규 분포의 가정이 어려울 때는 非母數的 檢定方法이 사용된다.

본 논문에서는 變化時點 問題에 대한 事例로 1969년부터 1988년까지 우리나라 병원의 1일 평균 외래 환자 수의 변화(母數的 檢定方法 適用)와 1981년부터 1990년까지 우리나라에서 당뇨병으로 인한 사망률의 변화, 그리고 1983년부터 1990년까지 우리나라 여자의 자궁경부암으로 인한 사망률의 변화 문제(以上 非母數的 檢定方法 適用)가 適用되었다. 時間의 흐름에 따른 관측치들의 대체적인 變化를 살펴보면 1일 평균 외래 환자 수와 당뇨병으로 인한 사망률은 감소한 時點도 있었으나 대체로 증가해왔다고 말할 수 있다. 그러나 이는 관측치들의 數値上의 比較를 통한 서술이며 특히 이를 통해 變化時點 年度를 찾아내는 것은 쉽지 않다. 더욱이 자궁경부암의 경우는 사망률의 변화 여부를 단순히 결론짓기는 어렵기 때문에 변화 여부를 판정하고 變化時點을 찾아내는 보다 정밀한 分析技法이 요구된다. 變化時點에 관심이 있고 기껏해야 한 개의 變化時點이 있다고 가정할 때 위의 세가지 事例를 통하여 관측치들의 時間의 흐름에 따른 변화 여부가 제시된 統計的 檢定方法을 통하여 分析되었고 이를 통하여 변화가 있었다고 판정되어진 경우 그에 대한 變化時點이 推定되었다.

또한 變化時點이 두 개 이상인 變化時點 問題인 경우에도 이에 대한 統計的 檢定方法이 있으나 본 논문에서는 생략하기로 한다.

參 考 文 獻

Pettitt, A. N., "A nonparametric approach to the change-point problem", *Applied Statistics*, 28, 1979, pp. 126−135.

Sen., A. and Srivastava, M., "On tests for detecting change in the mean when variance is unknown", *Ann. Inst. Statist. Math.*, 27, 1975, pp. 593−602.

강창완, "변화 시점 문제에 관한 모수적 검정 고찰 및 사례 연구", 고려대학교 대학원 통계학과 석사학위 논문, 1987.

백구환, "변화 시점 문제에 관한 비모수적 고찰과 사례 분석", 고려대학교 대학원 통계학과 석사학위 논문, 1987.

〈Summary〉

Statistical Testing and Analysis of Case Studies for the Change point Problem

Seonwoo Kim*

Consider a specific problem where independent observations are generated sequentially over time and the distribution of this random sequence is subjected to change at several possible points during data collection. Then it is of interest to elicit information from the observations concerning the possibility of such change points in this random sequence. In most of these cases, the numeric differences among observations have been used to describe the change over time. Accordingly, we need to test whether the observations change over time, and if so, to estimate the change points. There may be several change points, but, here we are concerned with at most one. If the possible change point is known, it becomes a traditional two sample comparison problem. Otherwise, it introduces a new testing problem, which is called the change point problem.

Let $x_1, x_2, \cdots, x_N$ be independent random variables that are successively drawn from a continuous population $F(\cdot)$ such that $x_1, x_2, \cdots, x_r \sim F(x)$, $x_{r+1}, x_{r+2}, \cdots, x_N \sim F(x-\Delta)$, $-\infty<\Delta<\infty$, where r is the change point, and Δ is the location-shift parameter which is the magnitude of the postulated change after the change point. Two statistical testing methods for this change point problem have been introduced. We are then able to use the parametric testing procedure proposed by Sen and Srivastava if we can assume that F (x) is a normal distribution function. For many cases when we can not make such an assumption, we use the nonparametric testing procedure by Pettitt.

Three case examples for the changes point problem have been studied. With the parametric testing method by Sen and Srivastava, it can be concluded that there was a change in the out-patient average / day between 1969 and 1988, and the change point is estimated to be 1981. Whether there was a change in the death rate from diabetes mellitus between 1981 and 1990 was also tested, as was the death rate from malignant neoplasm of the uterus between 1983 and 1990. Using the nonparametric testing method by Pettitt, it is concluded that there was an increase in the death rate from diabetes mellitus in 1985, and there was no change for the death rate from malignant neoplasm of the uterus.

* Senior Researcher, Korea Institute for Health and Social Affairs(KIHASA)