The Korean Communications
in Statistics Vol. 7, No. 2, 2000
pp. 447-456

Outlier Tests in Sample Surveys[1)]

Pyong Namkung[2)], Joon Suk Lee[3)]

Abstract

In this paper, we considered three methods for outlier identification in sample surveys. First, we studied method of handling and adjusting outliers in normal population. Second, we studied existing methods using mean, maximum and minimum and proposed a test using of median which well reflects characteristic of data regardless of sampling distribution. Finally, we showed our test using median works better than Dixon and mean test through simulation.

Keywords : outlier, Dixon test statistics, mean test statistic, median test statistic

1. 서 론

표본자료로부터 추정된 추정량의 정도를 향상시키고 편의를 감소시키기 위해 무응답 오차의 대체방법과 비표본 오차의 최소화 방법에 대한 연구가 활발히 진행되고 있다. 그러나 조사 결과의 정도를 감소시키고 편의를 심각하게 발생시키는 요인으로 특이값(outlier)의 영향을 고려해 볼 필요가 있다.

특이값의 처리 문제는 일변량 및 다변량 표본, 회귀분석, 실험계획 그리고 시계열분야 등에서 많이 다루어지고 있으나, 표본조사 분야에서는 특이값에 대한 탐색이나 검정을 심도 있게 다루지 않고 있다.

물론, 관측된 표본자료에 대해서 임의의 가정된 분포 하에서 관측자료들의 특이값을 판정하고 있지만, 가정된 분포에 위배되는 자료에 대해서는 특이값의 판단이 모호해 지거나 잘못 판정됨으로써 여전히 조사결과의 정도를 떨어뜨리고 편의가 존재하게 될 것이다. 이와 같이 표본자료에 특이값이 존재한다면 모수 추정이나 자료분석 결과에 심각한 편의를 제공하므로 표본조사에서 관측된 자료에 대한 특이값을 판단하는 근본적인 검토가 필요하다.

Barnett(1992)는 무한 모집단으로부터 추출한 표본조사에서의 특이값 검정에 대한 방법을 제안하였고, Barnett와 Roberts(1993)는 무한 모집단에서의 특이값 검정을 유한 모집단에 적용함으로써 나타나는 문제점들을 검토하였다. 그러나 이들의 연구에서도 관측자료들이 정규분포를 따르거나 균일분포, 지수분포와 같은 기본적인 분포 하에서만 특이값 검정 방법에 대해 다루고 있다. 또

1) This paper was supported by 63 Research Fund, Sungkyunkwan University, 1998.
2) Professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea
E-mail : namkung@yurim.skku.ac.kr
3) Lecturer, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea.

한 특이값의 영향에 대한 현재까지의 주된 연구 방향도 회귀모형과 같은 선형모형 하에서 특이값을 탐색하거나 영향에 대한 많은 연구가 진행되고 있는 실정이다. 이와 같이 표본자료에 대한 분포가 알려져 있지 않거나 잘못 알려져 있다면, 특이값의 판단은 부정확하게 될 것이며 편의가 발생된 결과를 제공하게 될 것이다.

따라서 표본조사에서 특이값을 판단하고 검정하는 보편적이고 일반화된 기준에 대한 연구와 통계전문가가 아니더라도 특이값 탐색을 보다 쉽게 적용하고 활용할 수 있는 방법에 대한 연구가 요구된다.

본 연구에서는 첫째, 실증 자료를 이용한 기존의 방법들의 활용 여부를 검토해 보고 둘째, 특이값 검정방법으로 평균이나 최대값 그리고 최소값을 이용하는 기존의 방법과는 달리 표본자료 분포의 인지여부에 관계없이 관측자료의 특성을 잘 반영하고 있는 중위수를 이용하는 새로운 방법을 제안하고 셋째, 모의실험을 통하여 새로 제안한 방법과 기존의 방법을 비교 검토하여 직관적이고 보다 쉽게 특이값을 판단할 수 있는 특이값 검정방법을 제안한다.

2. 정규표본에서의 특이값

2.1 정규표본에서 특이값의 처리문제

정규분포에서 특이값의 처리방법은 단순 제거나 가중값을 적용하였으나 Dixon(1950)과 Grubbs(1950) 등이 혼합모형을 제안하였는데 Dixon의 두 가지 혼합모형은 다음과 같다.

(1) 모형 A : 평균 이동모형(mean-shift model)

$$N(\mu\ ,\sigma^2)\ \Rightarrow N(\mu+\lambda\ ,\sigma^2)$$

(2) 모형 B : 분산 팽창모형(variance-inflation model)

$$N(\mu\ ,\sigma^2)\ \Rightarrow N(\mu, a^2\sigma^2)\ ,a^2>1$$

2.2 특이값의 조정

(1) L-추정량

순서통계량 $X_{(1)},X_{(2)},\cdots,X_{(n)}$에 가중값을 도입하여 특이값을 조정한다.

$$T=\sum_{i=1}^{n} a_i X_{(i)}\ , \tag{2.1}$$

여기서 a_i는 적절한 가중값

이 경우에 Tukey(1960)는 꼬리가 긴 경우 평균보다 중위수가 보다 효율적임을 보였으나 Hodges와 Lehmann(1963)은 다음과 같은 새로운 통계량을 제시하였다.

$$T = median(\frac{1}{2}(X_i + X_j) \ , 1 \leq i \leq j \leq n) \tag{2.2}$$

(2) M-추정량

잔차에 손실함수를 가중값으로 하여 특이값을 조정하는 방법으로 Huber(1964)가 로버스트 추정 이론의 새로운 접근을 하였다.

$$T = \sum \rho(X_i - T) \ , \tag{2.3}$$

여기서 ρ는 손실함수.

(3) 추정에 의한 제거기법

Dixon(1953)은 여러 가지 형태의 평균과 중위수의 평균제곱오차를 비교하여 특이값 제거 전과 제거 후의 평균제곱오차를 비교하였으며, Guttman(1967)과 Smith(1969, 1971)등이 분산이 알려져 있고 표본내에 한 개의 특이값이 존재하는 경우에 대한 연구를 하였으며 Guttman(1973)은 표본내에 두 개 이상의 특이값이 있는 경우로 확장하였다.

3. 기존의 특이값 검정방법

특이값 검정에 대하여 Mckay(1935)는 모표준편차를 알고 있는 경우에 극단값과 표본평균의 차이 $(X_{(n)} - \overline{X})/\sigma$의 분포를 이용하였으며, Thompson(1935)은 표준화잔차(studentized residual) $\tau_i = (X_i - \overline{X})/\hat{\sigma}$ 의 분포를 유도하였는데 그는 $(n-2)^{1/2}\tau_i/(n-1-\tau_i^2)^{1/2}$ 가 자유도 $n-1$의 t 분포를 따름을 보였다.

한편, Pearson 과 Chandra Sekar(1936)가 Thompson의 기준이 특이값을 발견하는데 매우 유용함을 보였으며, Grubbs(1950)는 τ의 정확한 분포를 유도하여 검정통계량으로 다음을 제시하였다.

$$S_n^2/S^2 = 1 - \tau^2/(n-1) \tag{3.1}$$

여기서 $\tau = \max(\tau_i)$, $S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2$, $S_n^2 = \sum_{i=1}^{n-1}(X_{(i)} - \overline{X}_n)^2$, $\overline{X}_n = \sum_{i=1}^{n-1} X_{(i)}/(n-1)$.

또한 Dixon은 한 개 혹은 두 개의 특이값 검정을 위해 범위의 비를 제시하였다. 즉,

(1) $r_{10}=\frac{X_{(2)}-X_{(1)}}{X_{(n)}-X_{(1)}}$; 한 개의 특이값 X_1

(2) $r_{20}=\frac{X_{(3)}-X_{(1)}}{X_{(n)}-X_{(1)}}$; 두 개의 특이값 X_1 , X_2

한편, Barnett(1992), Barnett와 Roberts(1993)는 특이값 검정방법으로 다음과 같은 방법을 제안하였다. 즉, 모집단 F로부터 표본크기가 n인 단순임의표본을 추출한 관측자료에서 특이값으로 최대값 $x_{(n)}$를 가정하고, 모집단 분포로는 정규분포, 균일분포, 지수분포를 가정하였다. 그리고 무한모집단에서의 특이값 검정방법으로 가정된 분포에 따라 평균을 이용하거나 자료의 최대값, 최소값 등을 이용하는 방법을 사용하고 있다.

3.1 모집단 F가 $N(\mu, \sigma^2)$인 경우

$$t_{N1}=\frac{x_{(n)}-\bar{x}}{s} \quad , \text{(극단적 표준화 거리)} \tag{3.2}$$

$$t_{N2}=\frac{x_{(n)}-x_{(n-1)}}{x_{(n)}-x_{(1)}} \quad , \text{(Dixon 통계량을 이용하는 경우)} \tag{3.3}$$

$$t_{N3}=\frac{x_{(n)}-x_{(n-2)}}{x_{(n)}-x_{(1)}} \quad , (x_{(n-1)}\text{에서 masking될 때 Dixon통계량을 이용하는 경우}) \tag{3.4}$$

3.2 모집단 F가 $U(a, b)$인 경우

$$t_{U1}=\frac{x_{(n)}-x_{(n-1)}}{x_{(n)}-x_{(1)}} \quad , \text{(Dixon 통계량을 이용하는 경우)} \tag{3.5}$$

$$t_{U2}=\frac{x_{(n)}-x_{(n-2)}}{x_{(n)}-x_{(1)}} \quad , (x_{(n-1)}\text{에서 masking될 때 Dixon통계량을 이용하는 경우}) \tag{3.6}$$

$$t_{U3}=\frac{x_{(n)}-x_{(n-2)}}{x_{(n)}-a} \quad , \text{(최소값 } a\text{가 알려진 경우)} \tag{3.7}$$

3.3 모집단 F가 $\mathrm{Exp}(\lambda)$인 경우

$$t_{E1}=\frac{x_{(n)}}{\sum x_i} \tag{3.8}$$

$$t_{E2} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}} \text{ , (Dixon 통계량을 이용하는 경우)} \quad (3.9)$$

Dixon 통계량을 이용하는 경우 식(3.3), 식(3.5) 그리고 식(3.9)는 같은 형태이나 각각 다른 분포와 다른 임계값을 갖는다.

4. 중위수를 이용한 특이값 검정방법

특이값 검정에 있어 정규 모집단을 가정하더라도 선택된 표본이 양 혹은 음의 방향으로 심하게 치우친 자료를 포함하고 있는 경우, 검정통계량으로 평균을 사용한다면 모집단 분포에 심각한 왜곡을 초래할 수 있다. 이러한 경우 평균 대신 중위수를 사용한다면 모집단 분포에 대한 왜곡을 방지할 수 있고 보다 안정된 검정통계량을 구할 수 있다.

새로운 특이값 검정방법은 자료의 최대값 $x_{(n)}$에서 중위수 $\tilde{x}$를 뺀 차를 Tippett(1925)의 정의에 의한 s로 나눈 검정통계량으로 다음과 같다.

$$t_M = \frac{x_{(n)} - \tilde{x}}{s} \text{ , (중위수를 이용한 극단적 표준화 거리)} \quad (4.1)$$

여기서 $s=(x_{(n)}-x_{(1)})/4$.

제안된 중위수를 이용한 특이값 검정통계량을 기존 결과와 비교하기 위하여 동일한 조건에서 모의실험을 한 후, 실증 자료를 이용한 기존의 방법과 비교 검토하여 보고자 한다.

5. 모의 실험

Dixon의 검정 통계량, 평균을 이용한 검정 통계량 그리고 중위수를 이용한 검정통계량을 구하기 위해 서울시내 33개 유명 체인점에서 한 달간 판매된 특정 상품의 판매액 자료 800개 중 표본크기를 30으로 하여 750회 모의실험을 하였다.

표본크기 $n=30$일 때 Dixon의 검정통계량, 평균을 이용한 검정통계량 그리고 중위수를 이용한 검정통계량은 다음과 같다.

(1) Dixon의 검정 통계량

$$t_D = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}} = 0.374485$$

(2) 평균을 이용한 검정 통계량

$$t=\frac{x_{(n)}-\bar{x}}{s} = 3.096545$$

(3) 중위수를 이용한 검정 통계량

$$t_M=\frac{x_{(n)}-\tilde{x}}{s} = 3.808455$$

여기서 $s=(x_{(n)}-x_{(1)})/4$.

또한, 기존의 Dixon 검정과 평균 검정을 위한 임계값과 비교하기 위하여 유의수준 5%와 1%에서 중위수 검정을 위한 임계값을 표본크기 $n=3$부터 $n=30$까지 각각 1,000회의 모의실험을 실시하여 구한 결과가 표 5.1과 같다.

표 5.1 Dixon 통계량과 평균, 중위수를 이용한 임계값

n	Dixon 통계량		평 균		중 위 수	
	5%	1%	5%	1%	5%	1%
3	0.941	0.988	1.15	1.15	3.69	3.93
4	0.765	0.889	1.46	1.49	3.40	3.82
5	0.642	0.780	1.67	1.75	3.37	3.71
6	0.560	0.698	1.82	1.94	3.20	3.51
7	0.507	0.637	1.94	2.10	3.11	3.49
8	0.468	0.590	2.03	2.22	3.01	3.39
9	0.437	0.555	2.11	2.32	3.00	3.27
10	0.412	0.527	2.18	2.41	2.88	3.21
12	0.376	0.482	2.29	2.55	2.84	3.15
14	0.349	0.450	2.37	2.66	2.79	3.10
15	0.329	0.426	2.41	2.71	2.78	3.08
16	0.313	0.407	2.44	2.75	2.73	3.06
18	0.300	0.391	2.50	2.82	2.71	2.94
20	0.277	0.362	2.56	2.88	2.67	2.92
30	0.260	0.341	2.74	3.10	2.60	2.87

위의 결과를 살펴보면 극단적인 특이값 $x_{(n)}$이 포함된 자료에서 구한 검정통계량 중 Dixon의 통계량과 중위수를 이용한 검정통계량은 $x_{(n)}$이 특이값 임을 보여주고 있으나, 평균을 이용한 검

정통계량의 경우 $\alpha=0.01$에서 $x_{(n)}$이 특이값 임을 보여주지 못함을 알 수 있다. 이는 극단적인 특이값이 평균에 상당한 영향을 미치고 있어 평균이 모집단의 특성을 잘 반영하지 못하고 있는 것으로 판단된다. 따라서 이러한 경우에는 평균을 이용한 검정통계량 보다는 Dixon, 혹은 중위수를 이용한 검정통계량의 사용이 바람직할 것이다. 결과적으로 모집단 분포가 어느 한쪽으로 심하게 치우쳐있는 경우 중위수를 사용하는 것이 모집단 특성을 잘 반영하면서 특이값을 판단할 수 있을 것이다.

그림 5.1은 극단적인 특이값이 포함된 자료의 히스토그램이다.

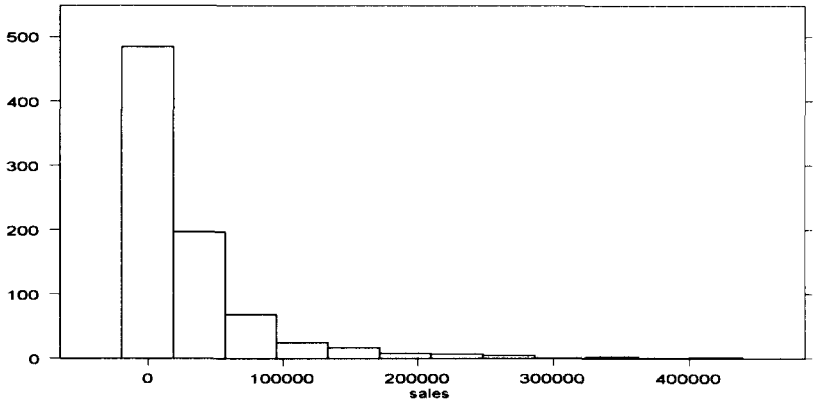

그림 5.1 특이값을 포함한 자료의 히스토그램

그림 5.2에서 보면 Dixon 검정통계량은 0.2와 0.5에 집중되어 나타났으며, 그림 5.3에서는 평균을 이용한 검정통계량이 2.5에서 4.5 사이에 분포되어 있고 그림 5.4의 중위수를 이용한 검정통계량은 3에서 4 사이에 분포되어 있음을 히스토그램을 통해서 볼 수 있다.

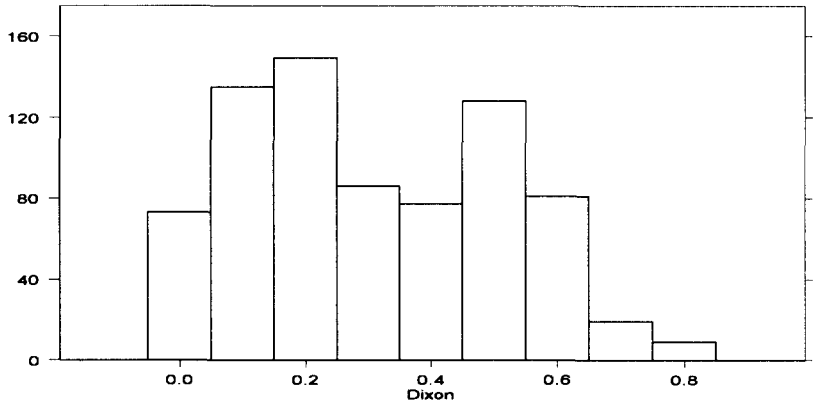

그림 5.2 Dixon 검정통계량의 히스토그램

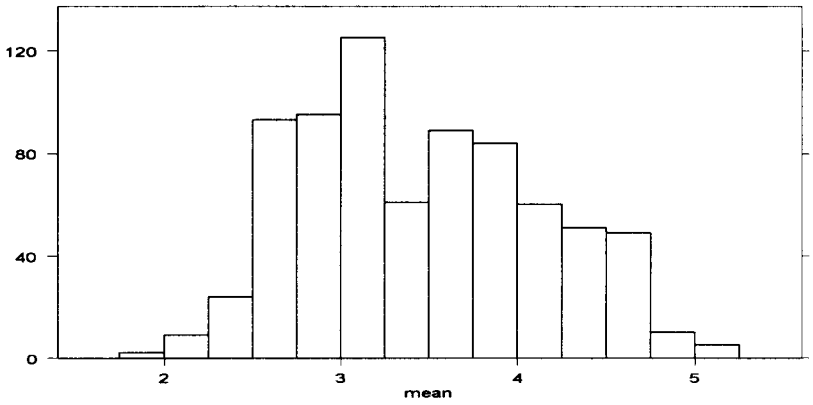

그림 5.3 평균을 이용한 검정통계량의 히스토그램

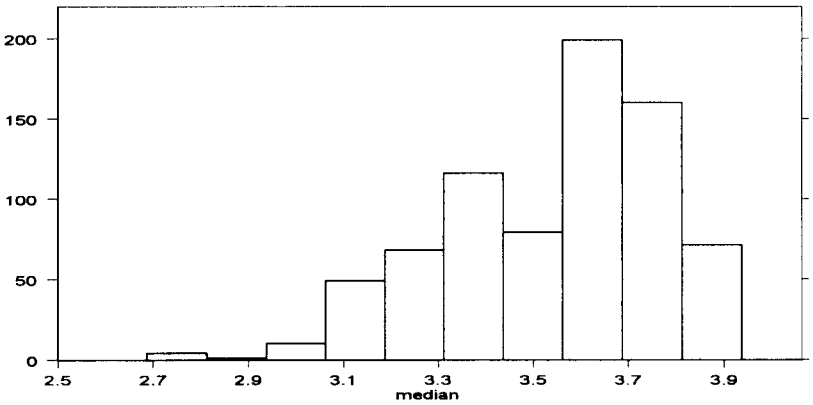

그림 5.4 중위수를 이용한 검정통계량의 히스토그램

6. 결 론

선형모형 하에서 특이값을 식별하고 특이값의 영향에 대한 연구가 최근까지 꾸준하게 진행되고 있으나, 표본조사에서의 특이값 검정방법에 대한 연구는 활발하지 않은 실정이다. 따라서 표본조사의 필요성과 중요성에 비추어 볼 때 특이값 검정방법에 대한 연구가 활발해질 것으로 기대된다.

첫째, 중위수를 이용하는 특이값 검정방법은 각종 표본조사에서 모수를 추론하기 전에 특이값을 탐색함으로써 추정량의 정도가 향상 될 것이며, 신뢰성이 있는 결과를 제공하게 될 것이다.

둘째, 평균이나 최대값, 최소값을 이용하는 기존의 검정 방법에 비해 중위수를 이용함으로써 표본자료의 분포에 관계없이 직관적으로 쉽게 이용할 수 있는 방법을 제공함으로써 표본조사에서 특이값을 보다 쉽게 식별할 수 있게 될 것이다.

셋째, 실증 자료를 이용한 특이값 검정 방법을 비교함으로써 여러 유형의 표본조사에 대하여 간편하고 편리한 검정 방법을 제공할 것으로 기대된다.

본 연구 결과를 이용할 경우 표본조사에서는 보다 효율적으로 특이값을 식별하게 될 것이며, 보다 신뢰성있는 표본자료로부터 모수를 추정할 수 있을 것이다. 또한 표본조사를 수행하는 조사기관에서는 특이값 검정의 효율적인 방법을 활용할 수 있을 것으로 기대한다.

참고문헌

[1] Barnett, V.(1992). Outliers in Sample Surveys. *Presented at 7th International Conference on Multivariate Analysis*, New Delhi, December 1992.

[2] Barnett, V. and Lewis, T.(1998). *Outliers in Statistical Data, 3rd Edition.* John Wiley & Sons.

[3] Barnett, V. and Roberts, D.(1993). The Problem of Outlier Tests in Sample Surveys. *Commun. Statist.-Theory Meth.*, Vol. 22, 2703-2721.

[4] Davies, L. and Gather, U.(1993). The Identification of multiple outliers. *Journal of the American Statistical Association.*, Vol. 88, 782-792.

[5] Dixon, W. J. (1950). Analysis of extreme values. *Ann. Math. Statist.*, 21, 488-506.

[6] Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *Ann. Math. Statist.*, 21. 27-58.

[7] Guttman, I., and Smith, D. E. (1969). Investigation of rules for dealing with outliers in small samples from the normal distribution I : Estimation of the mean. *Technometrics*, 13, 101-111.

[8] Hadi, A. S.(1994). A Modification of a Method for the Detection of Outliers in Multivariate Samples. *Journal of Royal Statistical Society,* B, 56, 393-396.

[9] Hadi, A. S.(1992). Identifying Multiple Outliers in Multivariate Data. *Journal of Royal Statistical Society,* B, 54, 761-771.

[10] Hadi, A. S. and Simonoff, J. S.(1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, 88, 1264-1272.

[11] Hawkins, D. M.(1980). *Identification of Outliers.* London. Chapman & Hall.

[12] Hodges, J. L. Jr., and Lehmann, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.*, 34, 598-611.

[13] Huber, P. J.(1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, 35, 73-101.

[14] McKay, A. T.(1935). The distribution of the difference between the extreme observation and the sample mean in samples of n from a normal universe. *Biometrika*, 27, 466-471.

[15] Pearson, E. S., and Chandra Sekar, C.(1936). The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, 28, 308-320.

[16] Simonoff, J. S.(1984). The Calculation of Outlier Detection Statistics. *Commun. Statist. Theory Meth.*, 13, 275-285.

[17] Thompson, W. R.(1935). On a criterion for the rejection of observations and the distribution of the ratio of the deviation to the sample standard deviation. *Ann. Math. Statist.*, 6, 214-219.

[18] Tukey, J. W.(1960). A survey of sampling from contaminated distributions. In Olkin, (ED.)(1960). *Contributions to Probability and Statistics*. University Press, Stanford, California.