응용통계연구 제 7권 2호
1994년 9월 pp. 1-7

윈저화 χ^2의 양태에 대하여[1)]

성내경[2)]

요약

몬테칼로 모의실험 기법을 사용하여 모집단이 정규분포를 따를 때 g-g 대칭 윈저화 제곱합에 기초를 둔 윈저화 카이제곱 통계량의 경험적 분포가 자유도 (n-3g-1)의 통상적인 카이제곱 분포에 만족할만하게 근사되어짐을 보였다. 여기서 n은 표본 크기, g는 한쪽 꼬리 부분에서 윈저화가 적용되는 양이다. 산출된 경험적 분포의 일부를 수록하였다. 윈저화 카이제곱 통계량의 적용 사례의 한 예로써 단일 표본에서 분산 검증을 다룬다. 이 연구는 Tukey와 McLaughlin (1963), Yuen과 Dixon (1973) 등의 연구 성과를 확대하는 것으로, 긴꼬리 분포에서 도출되는 자료 해석을 단순화하는 실용성을 위주로 한다.

1. 서론

실험계획, 회귀분석, 분산분석 등에서 널리 사용되는 많은 선형 통계모형들은 오차항이 서로 독립이며 평균이 0, 분산이 미지 상수 σ^2을 따르는 정규 확률변수라는 기본 가정하에 발전되었다. 그러나 Tukey와 McLaughlin (1963)이 지적한대로 실제 자료분석에서 유추할 수 있는 오차의 대표적인 분포형은 대칭이기는 하지만 정규분포보다 더 긴 꼬리를 가짐이 잘 알려져있다. 또한, 정규 가정하에서 최적인 통계 절차들을 긴꼬리 분포에 적용하면 검정력 등 여러 면에서 대체로 나쁜 결과가 나타남은 주지의 사실이다.

단일 표본의 경우에 비정규분포 상황에서 통상의 스튜던트 t 통계량의 취약성은 Tukey와 McLaughlin (1963)의 절단화 t나, Dixon과 Tukey (1968)의 윈저화 t와 같은 로버스트 통계량이 나타나는 계기가 되었다. 이 통계량들은 특히 정규성에서 크게 벗어나는 현상, 즉, 이상점의 악영향을 막는 목적으로 고안되었다.

Tukey와 McLaughlin의 절단화 t 통계량은 분자가 절단평균, 분모는 윈저화 표준편차인 형태로서, 몬테칼로 모의실험을 통하여 결정된 적당한 자유도의 t 분포로 놓아도 실용적인 면에서 부족함이 없다.

Tukey와 McLaughlin의 절단화 t 통계량의 개념을 확장하여 Yuen과 Dixon (1973)은 이표본 절단화 t 통계량을 제안하였는데, 이 경우에도 분자는 절단평균, 분모는 합동 윈저화 표준편차로 설정된다.

이러한 개연성에 착안하여 Stigler (1973)는 대표본에서 절단평균의 근사분포가 정규분포로서 특히 분산이 표본 크기에 의존하는 상수를 곱한 형태의 윈저화 분산이 됨을 증명하였다. 따라

1) 이 연구는 이화여자대학교 한국생활과학연구원 연구비와, 부분적으로 1993년도 한국학술진흥재단의 자유공모과제 연구비에 의하여 수행되었음.

2) (120-750) 서울시 서대문구 대현동 11-1, 이화여자대학교 통계학과.

서 절단화 t 통계량은 Stigler 정리의 유한 표본형으로 간주될 수 있다.

이러한 관점에서 절단화 평균과 윈저화 분산의 상호 관계는 셋 이상의 모평균들의 비교, 즉, 분산분석 문제에 즉각 적용할 수 있다고 판단된다. 즉, 통상의 분산분석은 평균과 카이제곱분포를 따르는 제곱합을 근간으로 만들어지듯, 로버스트 분산분석의 한 가지 형태가 절단평균과 윈저화 제곱합을 기초로 형성될 수 있다고 예상된다. 그러나 이 경우에 대두되는 한 가지 문제는 윈저화 제곱합의 정확한 분포를 알지 못한다는 점이다.

이 소고에서는 Tukey와 McLaughlin, Dixon과 Tukey, 그리고 Yuen과 Dixon 등에 의하여 수행된 기존 몬테칼로 모의실험의 형식을 빌려, 이들의 연구 성과를 분산분석에 확장시켜보려는 기초 단계로서 윈저화 제곱합의 근사 행태를 조사한 결과를 보고한다. 특히 윈저화 정규 표본에서 산출된 제곱합 항들이 적당한 자유도의 카이제곱 분포로 만족할만하게 근사됨을 보인다. 그리고 이런 결과를 실제 문제에 적용하는 한 가지 예로 단일 표본에서 분산에 대한 가설 검증 문제를 다룬다.

2. 윈저화 카이제곱의 근사 행태

우리의 주된 관심사는 정규분포 하에서 윈저화 자료로부터 형성된 유사 카이제곱 통계량의 분포 행태를 탐색하는데 있다. 이 유사 카이제곱 통계량의 행태에 대한 탐색은 후에 정의될 윈저화 평균을 기초로 계산된 수정 제곱합 S_w와 모분산 σ^2 간의 비의 분포 행태에 대한 탐색과 동등하다. S_w/σ^2의 정확한 분포는 분석적으로 얻을 수 없기 때문에 몬테칼로 모의실험 기법으로 경험적 분포함수를 추정한다.

$y_1 \le y_2 \le \cdots \le y_n$을 정규분포 $N(\mu, \sigma^2)$에서 추출된 n 개 관측을 순서화한 표본이라 하자. 이 자료에 대한 g-g ($g \ge 1$) 대칭 윈저화 자료를 다음과 같이 표기한다.

$$
\begin{aligned}
z_1 = z_2 = \cdots = z_g &= y_{g+1} \\
z_{g+i} = y_{g+i}, \quad & 1 \le i \le h = n-2g \\
z_n = z_{n-1} = \cdots = z_{n-g+1} &= y_{n-g}.
\end{aligned}
$$

여기서 n=g+h+g (순서화를 강조하는 방편으로 이런 표현이 n=2g+h보다 선호된다)이며, 양쪽 꼬리 부분에서 g 개의 관측들의 값이 수정되고 가운데의 h 개 관측들은 윈저화의 영향을 받지 않는다.

g-g 윈저화 평균 $\bar{y}_w$와 윈저화 평균 기준의 수정 제곱합 S_w를 다음과 같이 정의한다.

$$
\bar{y}_w = \sum_{i=1}^{n} z_i / n ,
$$

$$
S_w = \sum_{i=1}^{n} (z_i - \bar{y}_w)^2 ,
$$

일반성의 상실없이 $\mu=0$와 $\sigma^2=1$을 가정하자. SAS 6.04 패키지에 내장된 정규 난수 생성기인 RANNOR를 이용하여 표본 크기 n의 랜덤 표본 1,000 개를 10조 생성한다. 고려된 표본 크기는 5부터 30까지로 1 단위씩 증가시켰다. 각 표본 크기마다 생성된 관측들을 순서화하고 윈저화를 적용한다. 윈저화의 양 g는 표본 크기에 따라 달라지지만, Gastwirth와 Cohen (1970)의 연구 결과를 따라 20%까지의 윈저화를 적용하였다. 한 조의 1,000개 표본에 대하여, 각 윈저화 자료마다 S_w를 계산하고 선택된 분포함수값들에 대하여 S_w의 경험적 분위수를 산출한다.

여기서 S_w는 윈저화 제곱합이며 S_w/σ^2으로 간주해도 무방하기 때문에 앞으로는 이것을 윈저화 카이제곱이라 부른다. 윈저화 카이제곱의 경험적 누적분포함수를 구한 후 이 함수값들을 통상의 카이제곱 분포와 비교하여, 만족할만한 카이제곱 근사가 이루어지는 최적의 자유도를 탐색하였다. 즉, $S_w/\sigma^2 \approx \chi^2_{df}$ 로 놓을 수 있는 정수 df 를 탐색한다.

최대 관심 영역은 일반적으로 임계값이 위치하는 위꼬리 부분이므로, 이 영역에서 자유도를 변화시키면서 얻어지는 진짜 카이제곱 분포의 분위수와 윈저화 카이제곱의 경험적 분위수 간의 비를 취하여 가장 만족스러운 근사가 되는 자유도를 탐색하였다. 이러한 탐색 방법은 Dixon과 Tukey의 탐색법을 원용한 것으로, 이 결과 자유도 df 가 (n-3g-1)로 주어질 때 가장 적절한 카이제곱 근사가 이루어짐을 발견하였다. 자유도가 (n-3g-1)이라는 단순한 패턴으로 주어질 수 있음은 상당히 놀라운 결과라 하겠다.

표 1에 몇 가지 선택된 확률값에 대한 S_w의 경험적 분위수, 표준오차, 그리고 경험적 분위수와 도출된 χ^2_{df} 간의 비가 수록되어 있다. 표 1을 보면 일반적으로 S_w의 분포는 우리가 근사시킨 카이제곱 분포에 비하여 낮은 백분위수에서 다소 높은 꼬리를 가지며, 또 높은 백분위수에서 약간 낮은 꼬리를 갖는다. 그러나 우리가 주로 위꼬리 부분에만 관심이 있다는 점을 감안하면 낮은 백분위수들에서 높아보이는 꼬리 부분은 전혀 문제가 되지 않는다. 대부분의 경우 윈저화 제곱합의 경험적 분포는 자유도 (n-3g-1)의 카이제곱 분포에 비하여 미세하나마 다소 높은 꼬리 부분을 갖고 있다. 따라서 S_w를 직접 χ^2_{n-3g-1}로 간주하는 것은 실제 상황에 적용을 고려한 약간 보수적인 선택이라 하겠다.

표 1. 수록된 값들은 몇 개의 선택된 확률값들에 대한 대칭 윈저화 제곱합 분포의 경험적 분위수와 그들의 표준오차, 그리고 경험적 분위수 대 χ^2_{df} 분위수의 비이다. n은 표본 크기, g 는 각 꼬리에서 대칭 윈저화의 양이며 df는 (n-3g-1)으로 주어진다. 이 표에 나와있는 값들은, $n \leq 30$과 $g \geq 1$의 모든 조합에 대하여 산추된 완전한 경험적 분위수들 중 일부이다.

n	g	χ^2_{df}	0.05	0.1	0.5	0.9	0.95	0.975	0.99
5	1	1	0.051	0.112	0.860	3.300	4.486	5.634	7.125
			0.002	0.002	0.016	0.048	0.069	0.132	0.203
			12.894	7.067	1.889	1.220	1.168	1.121	1.074

표 1. (계속됨)

n	g	χ^2_{df}	0.05	0.1	0.5	0.9	0.95	0.975	0.99
6	1	2	0.227	0.372	1.673	4.858	6.118	7.450	9.150
			0.004	0.006	0.024	0.046	0.096	0.080	0.116
			2.210	1.766	1.207	1.055	1.021	1.010	0.993
7	1	3	0.483	0.745	2.525	6.364	7.992	9.466	11.197
			0.011	0.016	0.017	0.076	0.136	0.198	0.109
			1.372	1.275	1.067	1.018	1.023	1.013	0.987
8	1	4	0.812	1.178	3.417	7.684	9.160	10.738	12.595
			0.015	0.013	0.020	0.054	0.078	0.147	0.214
			1.143	1.108	1.018	0.988	0.965	0.964	0.949
9	1	5	1.248	1.717	4.385	9.256	11.082	12.725	14.947
			0.018	0.023	0.015	0.057	0.103	0.170	0.179
			1.090	1.066	1.008	1.002	1.001	0.992	0.991
10	1	6	1.686	2.302	5.315	10.505	12.384	14.172	16.585
			0.029	0.027	0.038	0.097	0.137	0.171	0.258
			1.031	1.044	0.994	0.987	0.984	0.981	0.986
10	2	3	0.663	0.952	2.913	6.716	8.240	9.617	11.517
			0.020	0.016	0.032	0.070	0.124	0.153	0.247
			1.885	1.629	1.231	1.074	1.054	1.029	1.015
20	1	16	7.875	9.087	15.090	22.966	25.685	28.283	31.233
			0.084	0.068	0.071	0.123	0.172	0.298	0.390
			0.989	0.976	0.984	0.976	0.977	0.981	0.976
20	2	13	5.848	7.048	12.158	19.379	22.038	24.393	27.370
			0.071	0.061	0.071	0.138	0.214	0.191	0.204
			0.992	1.001	0.985	0.978	0.985	0.986	0.989
20	3	10	4.123	5.051	9.490	15.870	17.932	19.905	22.523
			0.041	0.055	0.055	0.087	0.154	0.282	0.347
			1.046	1.038	1.016	0.993	0.980	0.972	0.970
30	1	26	15.084	16.983	24.946	35.141	38.531	41.717	45.479
			0.119	0.099	0.089	0.102	0.179	0.234	0.332
			0.981	0.982	0.985	0.988	0.991	0.995	0.996
30	2	23	12.654	14.428	21.724	31.188	34.539	37.330	41.099
			0.065	0.089	0.084	0.159	0.248	0.305	0.353
			0.967	0.972	0.973	0.974	0.982	0.980	0.987
30	3	20	10.857	12.358	18.872	27.619	30.400	33.037	36.221
			0.053	0.060	0.086	0.100	0.171	0.187	0.328
			1.001	0.993	0.976	0.972	0.968	0.967	0.964
30	4	17	8.737	10.051	16.188	24.220	26.910	29.686	32.081
			0.059	0.053	0.081	0.123	0.194	0.227	0.294
			1.007	0.997	0.991	0.978	0.975	0.983	0.960

3. 응용

제안된 윈저화 카이제곱 통계량을 적용하는 하나의 예로써, 단일 표본에서 분산에 대한 가설 검증을 고려한다. 분산에 관해서는 일면 가설이 보편적이므로 $H_0 : \sigma^2 \leq \sigma_0^2$와 $H_1 : \sigma^2 > \sigma_0^2$라 하자. 데이타가 다음과 같은 오염된 정규분포를 따른다고 가정하자:$(1-p)\times N(\mu, \sigma^2)+p\times N(\mu, c_2\sigma^2)$. 여기서 c는 척도상수이다. 오염률 p는 0.2로 고정하였다. 편의상 유의수준은 0.05로 놓는다.

이 예에서는 몇 가지 선택된 n, g, c들의 값에 대하여 모의실험을 통하여 귀무가설이 옳을 때 귀무가설을 기각하는 경험적 확률을 산출한다. 우선 생성된 원시 자료로부터 통상의 카이제곱 통계량을 사용하여 귀무가설하에서 귀무가설을 기각할 경험적 확률을 계산한다. 그 다음, 원시 자료에 윈저화를 적용하고, 윈저화 자료에서 동일한 경험적 확률을 근사 자유도 (n-3g-1)을 따르는 윈저화 제곱합을 기초로 계산한다.

모든 계산은 각 표본 크기마다 오염된 정규분포로부터 생성한 10조의 100개 랜덤 표본들을 기초로 수행되었다. 아래에 보인 결과는 귀무가설이 옳을 때 귀무가설을 기각하는 경험적 확률 값들과 표준오차들이다.

이 결과에서 윈저화 카이제곱 통계량에 비록 평소보다 더 작은 자유도가 부여되었을지라도 귀무가설 하에서 통상의 카이제곱 통계량보다 오염에 더 둔감함을 알 수 있다. 최대 이득은 척도 상수 c와 윈저화 양 g가 증가할 때 나타난다.

이는 오염된 정규분포의 경우에 대한 간단한 예로써 윈저화 카이제곱 통계량의 로버스트성을 보인 것이지만, 다른 긴꼬리 분포에서도 동일한 현상이 나타날 것임은 자명하다.

			χ^2		윈저화 χ^2	
c	*n*	*g*	확률	*se*	확률	*se*
3	5	1	.370	.012	.146	.007
	10	1	.555	.012	.255	.014
		2	.574	.016	.188	.011
	20	1	.782	.012	.522	.019
		2	.766	.014	.324	.012
5	5	1	.541	.013	.157	.013
	10	1	.806	.013	.389	.014
		2	.797	.009	.204	.012
	20	1	.961	.007	.796	.009
		2	.952	.006	.524	.014
7	10	1	.881	.011	.442	.018
		2	.914	.008	.240	.015
	20	1	.976	.006	.876	.009
		2	.985	.003	.629	.021

감사의 말

많은 시간을 소모하는 경험적 분포함수의 계산을 도와준 이화여자대학교 통계학과의 이서원 양에게 사의를 표한다. 또한 더 나은 논문이 되게끔 건설적 비판을 제기한 두 심사위원께 감사 드린다.

참고문헌

[1] Andrews, D. F., Bickel, W. H., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust estimates of location,* Princeton University Press, Princeton, New Jersey.

[2] Dixon, W. J. and Tukey, J. W. (1968). Approximate behavior of the distribution of Winsorized t (Trimming/Winsorization 2), *Technometrics,* Vol. 10, 83-98.

[3] Gastwirth, J. L. and Cohen, M. L. (1970). Small sample behavior of some robust linear estimators of location, *Journal of the American Statistical Association,* Vol. 65, 946-973.

[4] Stigler, S. M. (1973). The asymptotic distribution of the trimmed mean, *Annals of Statistics* Vol. 1, 472-477.

[5] Tukey, J. W. and McLaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/ Winsorization 1, *Sankhya,* Series A, Vol. 25: 331-352.

[6] Yuen, K. K. and Dixon, W. J. (1973). The approximate behaviour and performance of the two-sample trimmed t, *Biometrika,* Vol. 60, 369-374.

On the behavior of Winsorized χ^2[1)]

Nae Kyung Sung[2)]

Abstract

Using a Monte-Carlo simulation technique we evaluate the empirical distribution of a pseudo-chi-square statistic based on symmetrically Winsorized sum of squares when the population is normally distributed, and search for a chi-square distribution with appropriate degrees of freedom which can be referred to an approximate distribution for Winsorized chi-square.

1) Research was supported by Korean Research Institute for Better Living, Ewha Womans University, and in part by '93 Non-directed Research Fund, Korea Research Foundation.

2) Statistics Department, Ewha University, Seoul, 120-750, KOREA.