응용통계연구 제 **16**권 *1*호, *2003*년, *pp. 181-193*

표본크기 결정을 위한 IQR의 활용방법

홍종선[1] 김현태[2] 윤상호[3] 정민정[4]

요 약

표본크기를 결정할 때 모표준편차 σ의 추정량으로 표본표준편차를 구할 수 없는 경우 범위(R) 또는 사분위간 범위(IQR)를 이용하여 σ의 추정량으로 사용할 수 있다. R과 IQR의 함수로 나타난 추정값은 최소한 95% 이상의 확률로 σ보다 크거나 같아야 과소추정됨을 피할 수 있다. 다양한 확률분포로부터 추출된 여러 표본의 범위와 사분위간 범위에 대하여 Browne(2001)이 연구한 추정량 $R/4$과 본 연구에서 제시한 추정량 IQR이 σ이상일 확률에 대하여 비교 분석을 하였다. 그리고 표본의 범위와 사분위간 범위를 상수로 나누었을 때 σ이상일 확률을 가질 수 있는 대안적인 분모를 각각 구하여 비교 연구하였다.

주요용어 : σ의 검정력 보존 추정량, 범위, 사분위간 범위, 왜도, 표준평균범위, 첨도.

1. 서론

두 개이상의 모평균을 비교하는 연구에 필요한 적절한 표본크기(N)를 결정하는 공식은 잘 알려져 있다(Cohen 1969). 이 공식은 제1종과 제2종의 오류(α, β), 모평균 차이(Δ) 그리고 모표준편차(σ)의 값을 필요로 하는데, 처음 세 개의 값은 쉽게 구할 수 있으나 실제 σ값을 구하기에는 어려운 경우가 많다. 이러한 경우 σ의 추정량으로 유사한 연구 또는 사전조사의 결과에서 얻은 표본표준편차를 이용할 수도 있다.(N을 계산하기 위한 σ의 소표본 추정량을 이용하는 최적의 방법은 Browne(1995)의 연구에 상술되었다.)

σ의 추정량은 전체 자료로부터 표본표준편차를 계산하는 것이 일반적인 방법이지만 결측값이 많이 존재하거나 자료전체를 공개하지 않는 경우 표본표준편차를 쉽게 구할 수 없다. 어떤 연구보고서에서는 조사수집된 자료의 평균, 표본크기(n_s) 그리고 최소값과 최대값을 언급하였으나 표본표준편차는 제공하지 않는 경우도 있다.(본 연구에서 n_s는 사전조사 및 연구에 의해 알고있는 표본크기이고, N은 계산에 의해 결정되는 사후연구의 표본크기이다.) 연구보고서에서 제공된 자료의 두 극한값인 최소값과 최대값의 차이로부터 범위(R)를 구할 수 있으며, 이러한 경우 σ의 추정량으로 R을 이용하는 세 가지 방법은 다음과 같다 :

1) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 경제학부 통계학전공, 교수
E-malil: cshong@skku.ac.kr
2) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 대학원
E-malil: neopit@hanmail.net
3) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 대학원
E-malil: yunsangho@korea.com
4) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 대학원
E-malil: prograde@korea.com

1. $\sigma \approx R/4$ (Mendenhall, Ott와 Scheaffer 1971)
2. $\sigma \approx R/6$ (Daniel과 Terrell 1979)
3. S $\approx R/$표준평균범위 (Rhiel 1986, 1989).

σ의 첫 번째 추정량인 $R/4$은 정규분포를 따르는 자료 중에서 중심의 95.4%에 해당하는 자료가 $\mu \pm 2\sigma$ 구간에 포함된다는 사실에 기초하며, 두 번째 추정량인 $R/6$은 정규분포 중심의 99.7%를 구간 $\mu \pm 3\sigma$에 포함되기 때문에 σ의 적절한 추정량으로 간주한다. 표준평균범위(standardized mean range)는 d_n으로 표시하고, R/d_n의 기대값이 σ가 되도록 d_n을 설정하는데 이것은 자료에 적합한 확률분포의 영향을 받으며 n_s의 함수로 표현된다.(예를 들어 정규분포의 경우 n_s=20 일 때 d_n=3.714이다.)

σ의 정확한 추정값을 찾는 것도 중요하지만 더욱 중요한 연구 목적은 σ의 추정량에 관련된 표본크기 N을 결정하는 것이다. σ의 추정량은 통계량이기 때문에 모표준편차 σ보다 크거나 작은 값으로 나타날 수 있다. 만약 추정값이 σ보다 크면, N은 불필요하게 커지나 원하는 검정력을 충분히 확보할 수 있다. 하지만 추정값이 σ보다 작으면, N은 작아지며 과소 검정력(under-powered)을 가진 연구로 판명된다. N은 σ의 단조증가 함수이므로, 충분히 큰 N값을 갖기 위해서 σ의 적절한 통계량값이 필요하다. 즉 σ보다 크거나 같은 값을 가진 σ의 추정량이 필요한데, 이것은 주어진 검정력을 갖기 위한 N 값을 결정하는 것이다.

Browne(2001)은 R로부터 σ를 추정하기 위해 위에서 언급한 일반적인 방법을 고려하였으며, 이 방법보다 더욱 만족스러운 다른 방법도 고려하였다. Browne(2001)은 추정값이 σ이상일 가능성이 적어도 95%인 추정량을 'σ의 검정력 보존 추정량(power-preserving estimator of σ)' 이라고 정의하였으며, 앞에서 언급한 추정량 $R/4$, $R/6$, R/d_n이 σ의 검정력 보존 추정량 인지를 연구하였다. 그리고 대안적인 σ의 검정력 보존 추정량을 구하기 위하여 R을 나누는 적절한 분모를 탐색적으로 선택하는 연구를 하였다.

Browne(2001)은 σ의 추정량으로 표본범위(R)를 활용하여 σ의 검정력 보존 추정량을 연구하였지만, 본 연구에서는 범위 대신 특이값에 민감하지 않은 사분위간 범위(interquar-tile range; IQR)를 이용하여 σ의 검정력 보존 추정량을 연구하고자 한다. 즉, 범위를 알수 있는 최대값과 최소값 대신 특이값에 민감하지 않고 로버스트(robust)한 상사분위수 (upper quartile, Q3)와 하사분위수(lower quartile, Q1)가 제공되는 어떤 연구보고서를 토대로 표본크기 N을 결정하는 경우에 사분위간 범위를 σ의 추정량으로 이용하고자 한다. 예를 들어 "말기 암환자의 생존기간은 평균 3달이며 짧게는 1달에서 길게는 2년까지 사는 경우도 있다."라는 통보를 받은 환자의 절박한 심정을 상기해보자. 의사가 제시한 최대값은 물론이고 최소값에 환자는 당황할 뿐만 아니라 절망감에 빠지기 쉽다. 이런 경우에는 최대값과 최소값보다는 로버스트(robust)한 상사분위수와 하사분위수를 이용하여 "말기 암환자의 생존기간은 평균 3개월이며 하사분위수는 2개월, 상사분위수는 8개월입니다."라고 언급한다면, 환자나 가족은 더욱 안정적으로 그리고 긍정적으로 인정할 것이다(R=23개월, IQR=Q3 - Q1=6개월). 따라서 본 연구에서는 표본크기를 결정하고자 할 때, σ의 추정량으로 범위(R) 대신에 사분위간 범위(IQR)를 이용하고자 한다.

Browne(2001)의 연구에서는 다양한 형태의 확률분포와 여러 표본크기를 응용하여 Monte Carlo방법을 이용하였는데, 본 연구에서는 Browne(2001)의 방법을 이용하여 결과를 비교

하고 토론하였다. σ의 검정력 보존 추정량에 관한 연구결과는 확률분포의 형태에 따라 영향을 많이 받는다는 것을 알 수 있었는데, 특히 확률분포의 왜도와 첨도의 크기에 의존한다는 것을 파악하였다. 따라서 Browne(2001)의 연구를 확장시켜 다양한 왜도와 첨도값을 가진 확률분포를 생성한 후 이 확률분포에 의존하는 σ의 검정력 보존 추정량에 관한 연구로 확대하였다.

2. σ의 검정력 보존 추정량에 관한 연구

모표준편차 σ가 알려져 있는 분포가 주어졌을 때 표본크기가 n_s인 확률표본 10,000개를 독립적으로 생성하고, 각각의 자료에서 10,000개의 범위 R과 사분위간 범위 IQR을 계산한다. 표본크기 n_s의 값은 4부터 10,000까지 하나씩 증가하도록 설정하였다.

앞에서 언급한 3가지 방법($R/4$, $R/6$, R/d_n)에 의해 σ값을 추정하는데 각각의 10,000개의 추정값들 중 실제 σ보다 크거나 같은 추정값의 비율을 계산한다. 주어진 표본크기에서의 실제 σ보다 크거나 같은 추정값의 비율이 95%를 넘는 최소표본크기를 구한다.(95% 이상의 확률로 추정값이 σ 이상일 때, 이를 σ의 검정력 보존 추정값이라 정의하였다.) Browne(2001)의 연구에서 R/d_n은 고려된 어떠한 확률분포에서도 σ의 검정력 보존 추정량이 아님을 보였기 때문에 본 연구에서는 R/d_n에 대하여는 생략하였다.

σ의 검정력 보존 추정방법에 대하여 IQR의 연구는 다음과 같다. R은 표본 전체에 해당하는 통계량이며 IQR은 표본자료 중 가운데에 위치한 50%에 대응하는 통계량이므로 IQR을 나눈 분모는 3이하의 정수로 생각할 수 있다. 연구 방법은 위와 동일한 방법으로 ($IQR/3$, $IQR/2$, $IQR/1$)에 의해 σ의 추정값을 계산하였다. 그러나 $IQR/3$과 $IQR/2$인 경우 σ 이상일 가능성이 적어도 95%인 σ의 검정력 보존 추정값을 구할 수 없었기 때문에 IQR을 나눈 분모는 1로 고정시킨 IQR자체에 대한 σ의 검정력 보존 추정값을 구하였다.

표 2.1: σ의 검정력 보존 추정량이 되기 위한 $R/4$, $R/6$, IQR의 최소표본크기 n_s

	추정방법		
확률분포	$R/4$	$R/6$	IQR
표준정규	87	1750	54
카이제곱(1)	215	1452	N/A
카이제곱(2)	142	1783	421
카이제곱(6)	111	1882	92
베타(0.5,0.5)	N/A	N/A	15
베타(1,1)	N/A	N/A	20
베타(2,2)	263	N/A	31
베타(4,4)	126	N/A	39
베타(10,10)	98	4714	46
베타(2,4)	161	N/A	38

N/A = Not Achievable

Browne(2001)이 선정한 10개의 확률분포에 대하여 $R/4$, $R/6$, IQR의 추정값들이 σ 이상일 비율을 10,000개의 표본 중에서 계산하고 그 비율이 95%이상을 만족할 때의 최소표본크기를 계산한 결과는 표 2.1과 같다.

3. 결과 I

표 2.1의 추정량 $R/4$을 이용하는 경우를 살펴보면, 정규분포일 때의 최소표본크기가 제일 작다. 이것은 정규분포에서 표본크기 n_s가 87보다 작을 때에는 본 논문에서 연구된 여러 확률분포에서 추정값이 σ 이상일 확률이 95%에 미치지 못한다는 것을 의미한다. 또한 균일분포와 동일한 베타분포(1,1)과 베타분포(0.5, 0.5)일 때 $R/4$값은 σ보다 항상 작다. 균일분포와 같이 하한과 상한이 정해진 유계한 구간으로 나타나는 정의역을 가진 확률분포 또는 정의역의 상한과 하한 근처에 많은 확률이 몰려있는 확률분포로부터 생성된 자료일 때, $R/4$값은 과소 추정되기 때문에 $R/4$을 σ의 추정량으로 이용하지 않는다. n_s가 263보다 큰 경우 $R/4$은 베타분포(0.5,0.5)과 균일분포를 제외한 연구된 분포를 모두 고려하여도 σ 이상일 확률이 적어도 95%보다 크거나 같다. 이것은 $R/4$ 추정량을 이용하는 경우 특이값이 발생할 수 있는 확률분포로부터 생성된 자료에서 $R/4$값이 σ보다 크거나 같을 확률이 높다는 것을 의미한다.

정규분포에서 σ의 검정력 보존 추정량이기 위한 $R/6$은 n_s가 1,750보다 커야함을 알 수 있다. 자유도가 1이상인 카이제곱분포에서는 검정력 보존 추정량이기 위한 $R/6$은 n_s가 1,452보다 작은 경우 검정력 보존 추정량이 될 수 없고 카이제곱분포의 자유도가 커질수록 n_s도 커진다. 베타분포(0.5,0.5), 균일분포, 베타분포(2,2), 베타분포(4,4), 베타분포(2,4)일 때의 자료에서 $R/6$은 σ보다 항상 작다. 따라서 확률변수의 정의역이 유계한 구간인 확률분포 자료에서 $R/6$은 σ의 적절한 추정량이 될 수 없으며, 베타분포(10,10)인 경우 σ의 검정력 보존 추정량이기 위한 n_s는 4,714보다 커야한다. 그러므로 $R/6$은 구간이 정해진 분포로부터 자료를 이용할 때 σ를 추정하는 방법으로서 좋지 않다.

IQR을 정규분포에서 이용하는 경우에 σ의 검정력 보존 추정량이 되기 위한 최소표본크기는 54이므로 $R/4$과 $R/6$과 같이 범위 R을 이용한 경우보다 더 작은 표본을 필요로 하는 것을 알 수 있다. 범위를 이용하는 방법에 의해서 σ의 검정력 보존 추정량을 구할 수 없었던 베타분포(0.5,0.5)와 균일분포와 동일한 베타분포(1,1)의 경우 IQR을 이용함으로써 최소표본크기를 구할 수 있으며, 모든 베타분포에서 $R/4$을 이용하는 경우보다 IQR을 이용하는 경우에 더 작은 표본크기로 σ의 검정력 보존 추정량이 된다. 그러나 자유도가 크지 않은 카이제곱분포에서는 σ의 검정력 보존 추정량이기 위한 최소표본크기를 계산하기 위해 IQR보다는 $R/4$을 이용하는 것이 더 좋은 결과를 나타낸다. 그러므로 자유도가 1 또는 2의 카이제곱분포와 같이 왜도값이 큰 확률분포에서는 $R/4$을 이용하는 것이 더 좋고, 정규분포나 베타분포와 같이 첨도값이 0보다 작거나 같은 분포에서는 IQR을 이용하는 경우보다 작은 표본크기로 σ의 검정력 보존 추정량을 얻을 수 있다는 것을 유도하였다. 이에 관한 연구를 더욱 상세히 하기 위하여 다음 절에서는 다양한 왜도와 첨도값을 가지는 확률분포를 생성하고, 생성된 분포로부터 표본을 독립적으로 추출하여 σ의 검정력 보존 추정량에

관한 연구를 하여보자.

4. 결과 II

σ의 검정력 보존 추정량이기 위한 추정값이 모수보다 크거나 같은 비율의 변화를 살펴보기 위하여 다양한 첨도와 왜도값을 갖는 확률분포로부터 표본을 생성하였다. 우선 다양한 첨도값을 나타내는 확률분포는 다음과 같이 설정하였다. 여기서 첨도값이 -0.274인 확률분포는 Rhiel(1986)이 제안한 분포이다.

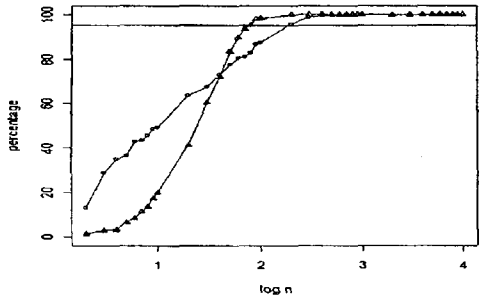
표 4.1: 첨도값에 대응하는 확률분포

첨도	확률분포	
3.0	자유도 6의 t-분포	
1.5	자유도 8의 t-분포	
1.0	자유도 10의 t-분포	
0	표준정규분포	
-0.274	$f(x) = x/20$	$(0.2 < x < 1.0)$
	$= 0.1525x$–0.1025	$(1.0 < x < 3.0)$
	$= 0.3350$	$(3.0 < x < 3.4)$
	$= 0.8735$–$0.1525x$	$(3.4 < x < 5.4)$
	$= 0.3200$–$(1/20)x$	$(5.4 < x < 6.2)$
-1.2	균일분포 ; U(0,1)	

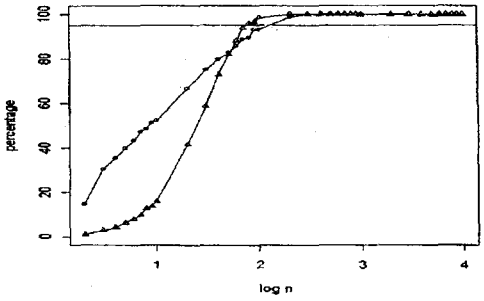
그림 4.1은 각 첨도값에 대응하는 확률분포에서 표본크기가 증가함에 따라 $R/4$과 IQR에 대한 σ의 검정력 보존 추정값의 비율을 나타낸 그림이다. 그리고 표 4.2은 그림 4.1의 결과에서 그 비율이 95%일 때에 해당하는 최소표본크기를 정리한 것이다. 그림 4.1과 표 4.2을 살펴보면 첨도값이 작을수록 $R/4$을 이용한 최소표본크기는 증가하지만 IQR을 이용한 최소표본크기는 감소한다. 첨도값이 1인 자유도 10의 t-분포에서 최소표본크기가 반전되어 IQR을 이용하는 경우 $R/4$을 이용하는 경우보다 적은 표본에서도 σ의 검정력 보존 추정량을 구할 수 있음을 확인하였다. 그러므로 첨도값이 1보다 큰 경우에는 $R/4$을 이용하는 것이 적은 표본에서 σ의 검정력 보존 추정량을 구할 수 있고, 첨도값이 1이하인 경우에는 IQR을 이용할 것을 추천한다.

표 4.2: 첨도에 따른 σ의 검정력 보존 추정량이 되기 위한 $R/4$과 IQR의 최소표본크기n_s

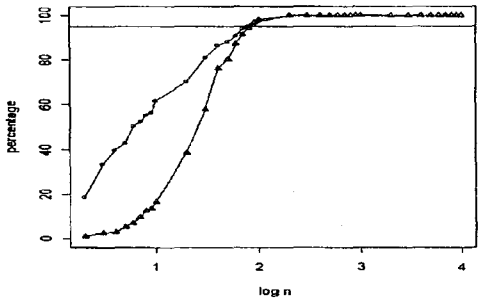
첨도	$R/4$	IQR	첨도	$R/4$	IQR
3.0	77	182	0	87	54
1.5	79	116	-0.274	92	52
1.0	82	80	-1.2	N/A	19



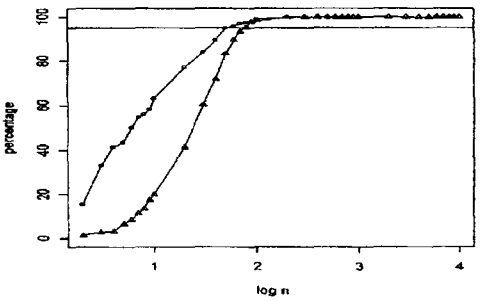
Ⅰ. 첨도 : 3



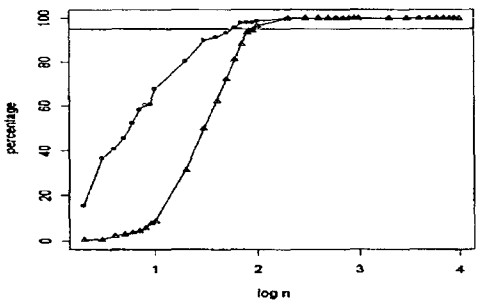
Ⅱ. 첨도 : 1.5



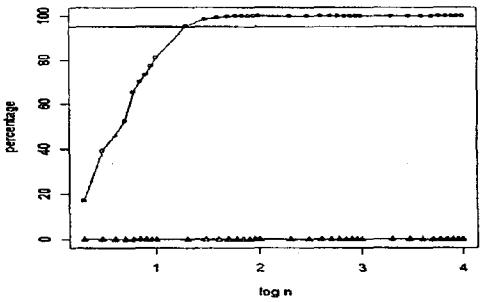
Ⅲ. 첨도 : 1



Ⅳ. 첨도 : 0



Ⅴ. 첨도 : -0.274



Ⅵ. 첨도 : -1.2

범례 : IQR : ○○○○○ , $R/4$: △△△△△

그림 4.1: 각 분포의 첨도에 따른 σ의 검정력 보존 추정량을 위한 표본의 크기에 따른 $R/4$과 IQR을 이용할 때의 비율

다음으로는 $R/4$과 IQR을 이용하여 왜도값의 크기에 따라 σ의 검정력 보존 추정량이 기 되기 위한 표본크기 n_s의 변화를 살펴보자. 다양한 평균과 분산값을 갖고있는 정규분포를 선형 결합(linear combination)하여 0부터 1.46까지의 왜도값을 갖는 6개의 확률분포를 생성하였다.(홍종선과 그 외 1994)

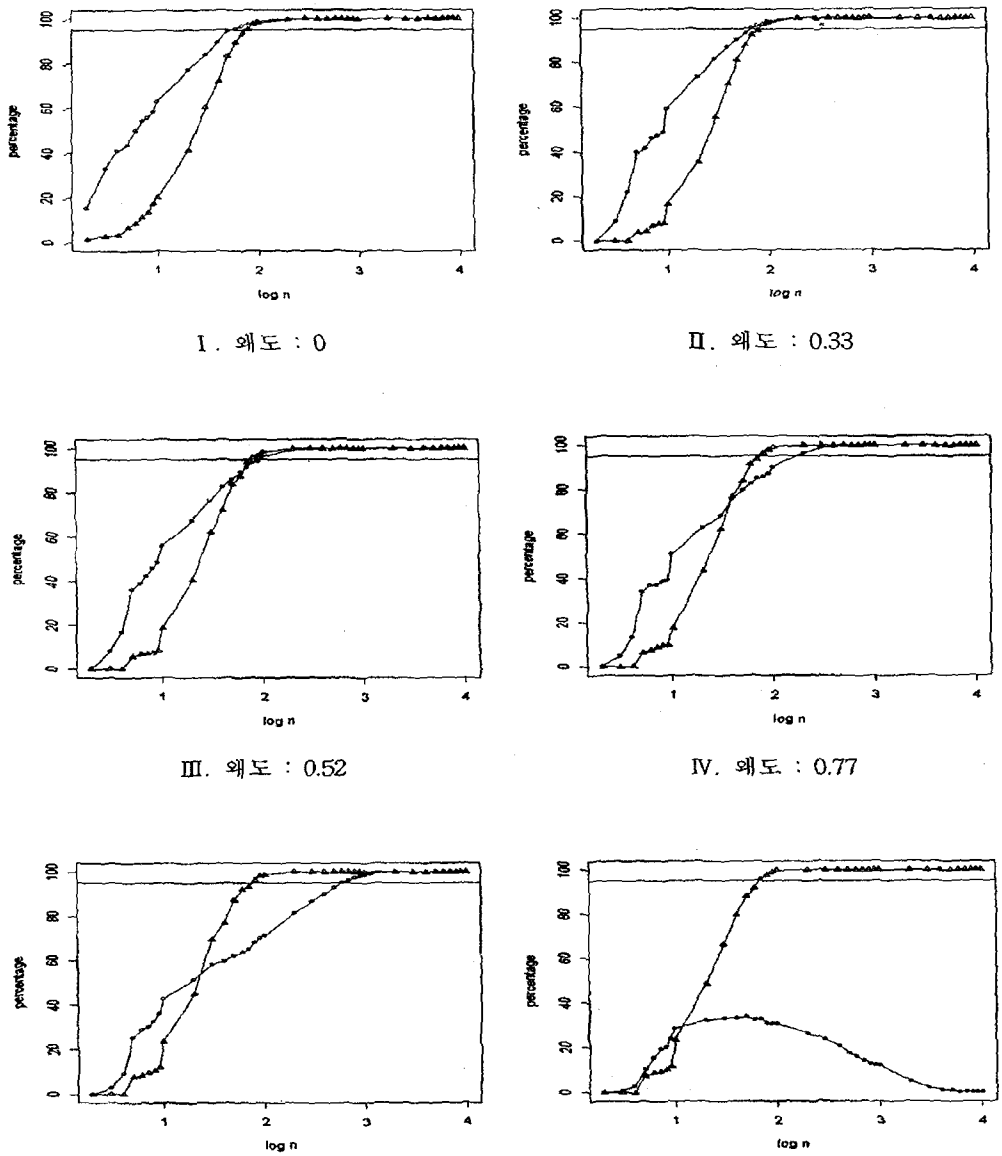
표 4.3: 왜도값에 대응하는 확률분포

왜도	확률분포
0.0	$N(0,1)$
0.33	$0.8 \times N(0,64) + 0.2 \times N(0,100)$
0.52	$0.8 \times N(0,49) + 0.2 \times N(0,100)$
0.77	$0.8 \times N(0,36) + 0.2 \times N(0,100)$
1.09	$0.8 \times N(0,25) + 0.2 \times N(0,100)$
1.46	$0.8 \times N(0,16) + 0.2 \times N(0,100)$

이 확률분포로부터 임의의 크기를 가진 독립적인 표본을 10,000개 생성하였다. 생성된 확률분포에 대응하는 왜도값과 표본의 증가에 따라서 $R/4$과 IQR에 대응하는 σ의 검정력 보존 추정값의 비율을 구하여 그림 4.2에서 나타내었으며, 표 4.4는 각 왜도에 따른 σ의 검정력 보존 추정량을 위한 $R/4$과 IQR의 최소표본크기를 구한 결과를 나타낸 것이다. 그림 4.2와 표 4.4를 살펴보면 왜도값이 커질수록 $R/4$을 이용한 최소표본크기는 약간씩 감소하지만 IQR을 이용한 최소표본크기는 크게 증가함을 알 수 있다. 따라서 왜도값이 0 근처일 경우 $R/4$보다는 IQR을 이용하면, 더 작은 표본으로 σ의 검정력 보존 추정량이 된다. 그러나 왜도값이 0.5이상의 큰 경우에는 IQR보다는 $R/4$을 이용하는 것이 더 작은 표본으로 σ의 검정력 보존 추정량이 될 수 있다는 것을 확인하였다.

표 4.4: 왜도에 따른 σ의 검정력 보존 추정량이 되기 위한 $R/4$과 IQR의 최소표본크기n_s

왜도	$R/4$	IQR	왜도	$R/4$	IQR
0	87	54	0.77	75	155
0.33	81	68	1.09	72	600
0.52	79	92	1.46	72	N/A



Ⅰ. 왜도 : 0 Ⅱ. 왜도 : 0.33

Ⅲ. 왜도 : 0.52 Ⅳ. 왜도 : 0.77

Ⅴ. 왜도 : 1.09 Ⅵ. 왜도 : 1.46

그림 4.2: 각 분포의 왜도에 따른 σ의 검정력 보존 추정량을 위한 표본의 크기에 따른 $R/4$과 IQR을 이용할 때의 비율

그러므로 우리는 정규분포, 균일분포 그리고 베타분포와 같이 첨도값이 작거나 음수인 경우와 왜도값이 0에 가까운 경우에는 IQR을 이용하면 보다 적은 표본크기에서 σ의 검정력 보존 추정량이 되며, 첨도값이 크고 꼬리가 두터운 분포와 왜도값의 절대값이 커서 양 끝으로 치우쳐진 분포에서는 $R/4$을 이용하는 것이 더욱 작은 표본크기로 σ의 검정력 보존 추정량을 구할 수 있다는 사실을 발견하였다.

5. σ의 검정력 보존 추정량을 위한 IQR의 분모에 관한 연구 결과

여러 확률분포에서 추출한 표본에 대하여 추정량이 모수 σ이상일 확률이 95%이상일 때 즉, σ의 검정력 보존 추정량일 때의 R/K값의 적절한 분모 K와 IQR/K의 적절한 분모인 K를 탐색적인 방법으로 연구해보자. 2절에서 이용한 Monte Carlo 방법을 이용하는데 여기에서는 R과 IQR을 실제 σ값으로 나누어 계산($K=R/\sigma$ 또는 IQR/σ)한 뒤에 순차 정렬된 K의 5번째 백분위수($K_{0.05}$)를 10,000개의 K값에서 구한다. 각각의 결과를 표 5.1에 함께 나타내었다.

표 5.1을 통하여 Browne(2001)이 연구한 R의 분모와 본 논문에서 제안한 IQR의 분모와를 비교하면서 토론하여 보자. 다양한 확률분포에서 R의 분모는 4를 중심으로 큰 변동을 갖으며, IQR의 분모는 1을 중심으로 작은 값을 나타내고 있으며 그 변동도 매우 작은 것을 파악할 수 있다. 특히 표본의 크기 n_s가 작은 경우와 왜도값이 큰 자유도 1의 카이제곱분포인 경우에는 IQR의 분모가 1보다 작다. 이것으로부터 왜도값이 0에서 멀리 떨어진 비대칭적 분포로부터 추출한 표본에서는 사분위간 범위 IQR이 모표준편차를 과소 추정되고 있다는 사실을 알 수 있다. 이 경우에는 사분위간 범위 IQR 보다는 범위 $R/4$을 이용하여 모분산을 추정하는 것이 바람직하다는 것으로 확인된다.

자유도가 3 이상인 카이제곱분포와 정규분포에서는 범위 R과 사분위간 범위 IQR을 이용하는 각각의 경우에 동일한 K값을 갖는다. 이것은 모의실험 결과에 의해 확률변수의 구간이 유한하지 않으며 왜도값이 0에 가까운 단일봉형태의 확률분포는 정규분포의 범위와 사분위간 범위의 분모와 동일하게 간주한다는 것을 의미하며, 모표준편차를 추정하기 위하여 범위 R보다 사분위간 범위 IQR을 이용하는 것을 제안한다.

예를 들어 정규분포로 추출된 표본의 범위가 100이며 사분위간 범위는 25인 경우를 가정하자. n_s가 10과 100이면, σ의 검정력 보존 추정량일 때의 R/K값의 K는 각각 1.86과 4.12이다. 따라서 Δ=20, α=0.05, 1-β=0.80인 두 그룹의 비교를 가정하면, n_s=10이면 각 그룹의 표본크기 N=115개가 추천되고, n_s=100이면 N=25개가 추천된다. 또한 n_s가 10과 100이면, σ의 검정력 보존 추정량일 때의 IQR/K값의 K는 각각 0.64와 1.12이다. 동일한 Δ=20, α=0.05, 1-β=0.80 상황에서, n_s=10이면 각 그룹의 N=61개가 추천되고, n_s=100 이면 N=21개가 추천된다(Cohen 1969). 위의 예의 경우에서는 IQR에 의한 추정량이 R을 이용하는 경우보다 더 작은 표본크기 N을 제안한다.

Rousseeuw와 Leroy(1987)과 David(1998)은 로버스트한 모표준편차의 추정량으로 IQR / (2×0.6745) 를 제안하였으며 가중최소제곱법에 응용하여 많은 연구를 하였다. 본 연구의 3절과 4절에서는 모표준편차 σ의 검정력 보존 추정량으로 IQR을 나눈 분모를 1로 설정

하였고 5절에서는 1을 중심으로 IQR의 적절한 분모를 설정하기 위해서 연구하였는데, 표 5.1에서와 같이 정규분포에서 표본크기가 10,000인 경우 IQR의 분모는 1.32이다. 이 결과는 Rousseeuw와 Leroy(1987)과 David(1998)의 결과와 유사함을 확인할 수 있다.

표 5.1: 각 분포에서 σ의 검정력 보존 추정량이 되기 위한 R과 IQR의 분모 K

표본 크기	정규와 카이제곱 (3이상)		카이제곱 (1)		카이제곱 (2)		베타 (0.5,0.5)		균일		베타 (2,2)	
n_s	R	IQR	R	IQR	R	IQR	R	IQR	R	IQR	R	IQR
5	1.04	0.27	0.35	0.11	0.63	0.14	1.15	0.30	1.17	0.36	1.14	0.33
10	1.86	0.64	0.91	0.28	1.27	0.38	2.16	0.86	2.07	0.77	2.03	0.72
20	2.62	0.81	1.56	0.42	1.93	0.57	2.61	1.26	2.71	1.03	2.72	0.93
30	3.04	0.92	2.06	0.45	2.33	0.64	2.72	1.39	2.95	1.18	3.04	1.08
40	3.30	0.99	2.37	0.54	2.59	0.68	2.77	1.49	3.07	1.25	3.23	1.09
50	3.50	1.01	2.54	0.54	2.82	0.73	2.79	1.56	3.14	1.32	3.37	1.17
60	3.68	1.02	2.79	0.56	2.99	0.78	2.80	1.62	3.20	1.37	3.47	1.21
70	3.79	1.05	2.98	0.58	3.17	0.79	2.81	1.64	3.24	1.40	3.54	1.22
80	3.91	1.06	3.14	0.62	3.29	0.80	2.81	1.70	3.26	1.43	3.60	1.23
90	4.03	1.09	3.30	0.64	3.42	0.82	2.82	1.70	3.28	1.44	3.66	1.27
100	4.12	1.12	3.45	0.65	3.52	0.83	2.83	1.73	3.30	1.46	3.70	1.29
250	4.81	1.19	4.55	0.73	4.41	0.92	2.83	1.81	3.40	1.54	3.99	1.40
500	5.29	1.24	5.45	0.76	5.12	0.97	2.83	1.89	3.43	1.61	4.13	1.44
1000	5.75	1.27	6.27	0.78	5.80	1.01	2.83	1.91	3.45	1.64	4.23	1.47
10000	7.09	1.32	9.31	0.83	8.06	1.07	2.83	1.97	3.45	1.70	4.40	1.52

6. 결론

σ의 추정량으로 범위 R과 사분위간 범위 IQR에 대해 비교 연구하여 얻은 두 가지 결과는 다음과 같다. 첫 번째로 다양한 자료의 확률분포로부터 추출한 표본에서 $R/4$, $R/6$, R/d_n 중 가장 효과적인 $R/4$와 IQR, $IQR/2$, $IQR/3$중 가장 효과적인 IQR이 σ보다 크거나 같은 값을 갖는 추정량을 제공하는지 비교 연구하였다. 왜도값이 큰 카이제곱분포에서는 $R/4$을 이용하는 것이 더 좋고, 정규분포나 베타분포와 같이 첨도값이 0보다 작거나 같은 분포에서는 IQR을 이용하는 경우 더욱 작은 표본크기에서 σ의 검정력 보존 추정량이 된다는 것을 연구하였다. 다양한 왜도와 첨도값을 가지는 확률분포를 생성하여 연구를 확장한 결과 첨도값이 작은 정규분포와 베타분포의 경우와 왜도값이 0에 가까운 경우에는 σ의 검정력 보존 추정량을 위해 IQR을 이용하는 것이 작은 표본크기를 요구하며, 첨도값이 큰 분포와 왜도값의 절대값이 커서 양끝으로 치우쳐진 분포에서는 $R/4$을 이용하는 것이 더욱

작은 표본크기로 σ의 검정력 보존 추정량을 구할 수 있다는 것을 추가적으로 연구하였다.

두 번째는 다양한 표본크기에 대해서 범위 R과 사분위간 범위 IQR의 어떤 분모가 σ의 검정력 보존 추정량으로서 적절한가를 비교 연구하였다. 다양한 확률분포에서 IQR의 분모는 R의 분모보다 작은 1을 중심으로 적은 변동을 나타내고 있다는 것을 탐색하였다. 특히 표본크기 n_s가 작은 경우와 왜도값이 큰 자유도 1의 카이제곱분포인 경우 IQR의 분모가 1보다 작다. 이것으로부터 알 수 있는 것은 왜도값이 0에서 멀리 떨어진 분포로부터 추출된 표본에서는 사분위간 범위가 모표준편차를 과소 추정하고 있다는 것이다. 그러므로 이러한 경우에는 사분위간 범위 IQR보다는 범위 $R/4$을 이용하여 모표준편차을 추정하는 것이 바람직하다는 것을 알 수 있다. 반면에 왜도값이 0에 가까운 베타분포인 경우에는 모표준편차을 추정하기 위해서 범위 $R/4$을 이용하는 것보다 사분위간 범위 IQR을 이용하는 것이 더욱 작은 표본크기로 σ의 검정력 보존 추정량을 구할 수 있다는 사실을 유도하였다.

참고문헌

[1] 홍종선, 최병수, 엄종석. (1994). 정규분포의 혼합성 판단기준, <*응용통계연구*>, 제7권 1호, 131-140.

[2] Browne, R. H. (1995). On the Use of a Pilot Sample for Sample Size Determination, *Statistics in Medicine*, Vol. 14, 1933-1940.

[3] Browne, R. H. (2001). Using the Sample Range as a Basis for Calculating Sample Size in Power Calculations, *The American Statistician*, Vol. 55, No. 4, 293-298.

[4] Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*, New York: Academic Press.

[5] Daniel, W. W., and Terrell, V. C. (1979). *Business Statistics: Basic Concepts and Methodology*, Boston, MA: Houghton Mifflin Company.

[6] David, H. A. (1998). Early Sample Measures of Variability, *Statistical Science*, Vol. 13, No. 4, 368-377.

[7] Mendenhall, W., Ott, L., and Scheaffer, R. L. (1971). *Elementary Survey Sampling, Belmont*, CA: Duxberry Press.

[8] Rousseeuw, P. J., Leroy, A. (1987). *Robust Regression and Outlier Detection*, New York, Wiley.

[9] Rhiel, G. S. (1986). The Effect of Non-normality on the Use of the Range in Estimating the Population Standard Deviation, *Journal of Statistical Computation and Simulation*, Vol. 24, 71-82.

[10] Rhiel, G. S. (1989). An Improved Range Estimator of Sigma for Determining Sample Sizes, *Communication in Statistics-Simulation*, Vol. 18, 1295-1309.

[2002년 7월 접수, 2002년 12월 채택]

Using the Sample IQR for Calculating Sample Size

C. S. Hong[1] H. T. Kim[2] S. H. Yun[3] M. J. Jung[4]

ABSTRACT

Without a sample standard deviation for an estimator of the population standard deviation σ in a sample size computations, we often use some functions of a sample range (R) or interquartile range (IQR) by an estimator of σ. In order to avoid under-powered studies, these estimates must have a high probability of being greater than or equal to σ. In this paper, these probabilities of being greater than or equal to σ are estimated for IQR for various parents distributions, and are compared with the probabilities for $R/4$ (Browne 2001). Alternative divisors (K) are explored and discussed for which the probabilities of R/K and IQR/K being greater than or equal to σ is at least 95%.

Keywords: : Interquartile range; Kurtosis; Range; Power-preserving estimator of σ; Sample size; Skewness; Standardized mean range.

1) Professor, Department of Statistics, Sungkyunkwan University.
E-malil: cshong@skku.ac.kr
2) Graduate Student, Department of Statistics, Sungkyunkwan University.
E-malil: neopit@hanmail.net
3) Graduate Student, Department of Statistics, Sungkyunkwan University.
E-malil: yunsangho@korea.com
4) Graduate Student, Department of Statistics, Sungkyunkwan University.
E-malil: prograde@korea.com