

원점수 순위와 표준점수 순위 비교를 통한 표준화 방법비교[†]

민대기¹ · 정지현²

¹ 덕성여자대학교 정보통계학과 · ² 덕성여자대학교 통계학과
접수 2011년 12월 5일, 수정 2011년 12월 28일, 게재확정 2012년 1월 5일

요약

수시입학이나 면접에서 여러 조가 동시에 면접을 진행하고, 그 점수로 합격 여부를 결정할 때 면접관에 따른 편차를 조정하기 위한 점수의 표준화는 필수적인 과정이다. 그리고 대부분은 표준편차를 이용하여 표준화를 취한다. 그러나 이 과정에서 자칫 표준편차가 큰 그룹에 속한 후보자는 표준편차가 작은 그룹에 속한 후보자에 비하여 불이익을 당할 우려가 있다. 본 논문에서는 표준편차를 이용한 표준화가 최상의 선택이며, 또한 위험도는 어느 정도 되는지를 알아보기 위하여 다른 표준화 방법과 비교 연구하였다.

주요용어: 사분위, 중위수, 표준편차, 표준화.

1. 연구배경 및 목적

대학입시의 수시나 회사 면접 시, 많은 지원자가 있을 때 여러 조가 동시에 면접을 진행하는 것은 필수적인 일이다. 이 과정에서 면접관에 따른 편차가 내재하여 표준화를 통하여 산출된 점수로 합격 여부를 결정한다. 일반적으로 많은 경우에 표준편차를 통하여 표준화를 하는 것으로 알고 있다. 그러나 표준편차는 극한값에 민감하고 이로 말미암아 합격선에 있는 학생이 불행하게도 불합격할 수 있는 위험이 있다. 본 연구에서는 어떤 표준화 방법이 가장 안정적이며, 각각의 표준화 방법은 어느 정도 위험이 발생하는지를 모의실험을 통하여 비교 연구하였다. 모의실험은 중심변동이 비교적 작고 산포도가 다른 경우를 가정하여 정규분포와 포아송분포를 이용하여 원점수 순위와 표준점수 순위를 각각 50번 비교하였다. 또한, 일반적으로 중심과 산포도가 변동이 심한 경우를 가정하여 정규분포와 감마분포를 가정하여 각각의 경우에 50개씩 5그룹을 생성하여 50회 실험을 시행하였다. 표준화 방법으로는 가장 많이 알려진 STD (Standard Deviation)와 Range, MAD (Median Absolute Deviation from Median), IQR (Interquartile Range) 등을 이용하였다. 각 표준화방법은 사용하는 스케일과 중심 값에 따른 장, 단점을 갖고 있다. 표준화를 실행한 후에, 원점수순위와 표준점수순위를 비교하여 방법에 대한 안정성 평가는 정 분류비율과 피어슨 상관도를 이용 하였다.

2. 연구내용

그림 2.1은 수시면접에서 1개 그룹이 25명씩, 125명이 5개로 그룹으로 나뉘어 각각 다른 면접관에 의한 면접 후, 원점수와 STD를 이용한 표준점수의 산포도를 비교한 것이다. 극한값에 취약한 STD를 이

[†] 위 연구는 2010 덕성여자대학교 교내연구비를 지원받았음.

¹ 교신저자: (132-714) 서울특별시 도봉구 근화교길19, 덕성여자대학교 정보통계학과, 교수.

E-mail: dkmin@duksung.ac.kr

² (132-714) 서울특별시 도봉구 근화교길19, 덕성여자대학교 통계학과 대학원, 석사.

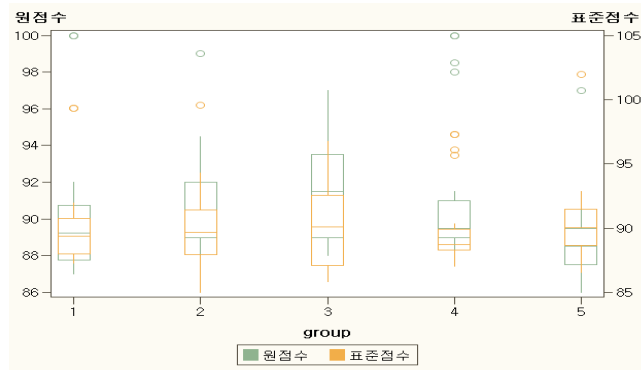


그림 2.1 원점수와 표준점수 비교

용한 표준점수 산출이후 각 그룹에서 원점수와 표준점수의 분포를 살펴보면, 산포도크기가 타 그룹에 비하여 비교적 다른 그룹3과 그룹5에 많은 변화가 있음을 알 수 있다. 그림에서 보는 바와 같이 그룹3은 산포도가 다른 그룹에 비하여 크기 때문에 원점수와 표준점수에 대한 상자도표를 비교하여 보았을 때 표준점수가 하향 조정되었다. 반면 그룹5는 산포도가 비교적 작아서 표준점수가 전체적으로 상향 조정되었다. 전체 125명 중 상위 40등까지를 합격선이라 가정하여 원점수 순위와 표준점수 순위를 비교하여 보았을 때, 그룹5의 원점수 88점이 그룹3의 원점수 92보다 표준점수가 높아 합격선에 속하는 것을 볼 수 있었다. 만약 원점수 순위가 실제 실력에 근거하였다면 표준화 과정에서 순위가 바뀌는 정도를 위험의 발생이라 할 수 있다. 본 연구에서는 어느 경우에서 위험이 얼마나 발생하는지를 연구하기 위하여 중심의 변동이 작고 산포도만 다른 경우와 중심과 산포도가 다른 일반적인 경우를 가정하여 실험하였다. 그림 2.2와 같이 두 경우에 대하여 모의실험을 통하여 5그룹에 각 50개의 자료를 생성하여 원점수 순위를 정하고, 표준화를 한 후에 다시 그 점수를 기준으로 표준점수 순위를 정하여, 원점수 순위와 표준점수 순위가 어느 정도 변동이 있는지를 측정하였다. 원점수가 실제 성적을 반영한 것이라 가정할 때 표준화 점수에 근거한 순위와 원점수 순위의 차이를 위험의 정도로 생각하여 표준화 방법에 따른 안정성을 비교해 보았다.

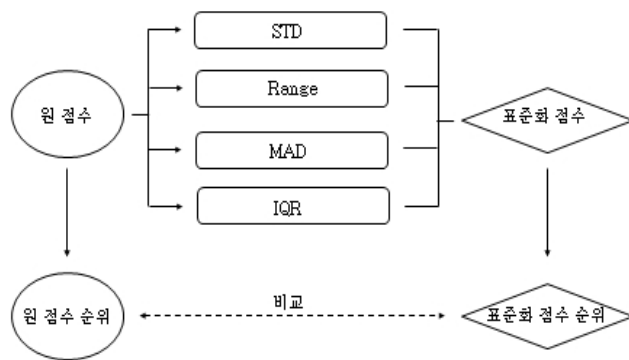


그림 2.2 연구진행절차

2.1. 표준화 방법

자료의 중심과 퍼진 정도를 어떤 값으로 측정하느냐에 따라 표준화 방법의 중심과 산포도 값은 달라진다. 본 논문에서는 그룹 자료의 퍼진 정도를 분산으로 표현하지 않고 산포도로 표현하고자 한다. 표준화 방법으로 가장 일반적으로 많이 쓰이는 평균과 표준편차를 이용한 STD, 최댓값, 최소값 만을 활용하는 Range, 중앙값을 이용한 MAD, 75%와 25% 값을 이용한 IQR 등이 있다. STD는 자료의 중심이 평균이고 표준편차를 산포도로 하여 표준화를 취하는 방법으로 극한값에 민감한 약점이 있다. 극한값에 민감하지 않은 로버스트 추정치에 대한 연구는 여러 분야에서 진행되고 있다. 국외에서는 Goodall (1983)과 Iglewicz (1983) 등이 하였으며 국내에서는 군집분석에서는 황진수 (2009) 등이 극한값에 민감하지 않은 표준화된 거리를 근거로 군집을 분류하는 방법에 연구하였다. 회귀분석에서는 Hwang (2006) 등이 새로운 접근방법으로 연구하였다. 표준편차가 크다는 것은 각 관측 값이 평균으로부터 그만큼 멀리 떨어져 있다는 것을 의미한다. Range는 그룹의 중심이 최소값이고 산포도는 Range이다. 그러나 Range는 오직 그룹의 최댓값과 최소값을 근거하여 그 차이로 결정되므로 그룹의 크기가 충분히 크지 않을 때 산포도에 대한 측정값으로 매우 취약하다. MAD는 그룹에 대한 중심은 중앙값이고 산포도는 중앙값과 각 관측치의 절댓값에 대한 중앙값으로 결정되므로, 극한값에 민감하지 않고 비교적 안정적으로 그룹의 퍼진 정도를 나타낸다. 중앙값의 신뢰구간에 대해서는 Park (2004)이 연구하였다. Peter (1993)가 연구한 바와 같이 특히 범위가 없는 cauch 분포와 같은 그룹의 산포도를 나타내는데 적절하며 인구통계에서 중요하게 사용된다. IQR은 그룹의 중심은 중앙값이고 산포도는 그룹의 75%에 해당하는 값과 25%에 해당하는 값의 차이로 결정된다. Range에 비하여 상위, 하위 25% 값을 취하지 않기 때문에, 그룹의 산포도를 나타내는 데 안정적이다. 분포가 대칭인 때는 MAD는 IQR의 반값이다.

2.2. 모의실험 및 비교방법

표준화하는 과정은 SAS/STAT (1988)의 Proc Stdize를 이용하였다. 표준화를 취한 후 산출되는 점수는 산포도를 나타내는 스케일과 위치를 나타내는 중심 값에 의해서 결정되므로 두 가지 경우를 고려하여 실험을 계획하였다. 첫 번째는 실제 입시상황과 비슷한 그룹의 중심변동이 비교적 작고 산포도만 다른 경우, 두 번째는 제약조건을 없애고 일반화하여 중심변동이 심하고 산포도 역시 아주 다른 경우이다.

2.2.1. 그룹의 중심변동이 비교적 작고 산포도가 다른 그룹

실제상황과 비슷한 중심변동이 작고 산포도가 다른 그룹에 대해 모의실험을 하기 위하여 정규분포와 포아송분포를 각각 가정하고 1그룹에 50개씩 5그룹을 만드는 과정을 SAS의 ranpoi와 rannor 함수를 이용하여 각각 250회 실시하였다. 각 5개 그룹에 대하여 정규분포의 평균은 80으로 하고, 표준편차는 8, 10, 12, 14, 16 등으로 모두 다르게 설정하였다. 포아송분포는 그 자체의 분포가 오른쪽으로 치우침이 있기 때문에 5개 그룹에 같은 평균값만 설정하였다. 아래의 그림은 포아송분포 함수를 이용하여 만든 모의 그룹자료의 한 예이다. 그림 2.2에서 보는 바와 같이 중심의 변동은 비교적 작고 각 그룹의 산포도는 다르게 설정되어 있다.

2.2.2. 그룹의 중심과 산포도가 아주 다른 그룹

제약조건이 없는 상황을 가정하여 정규분포와 감마분포 함수를 이용하여 모의실험자료를 만들었다. 분포가 대칭인 경우와 그렇지 않은 경우를 모두 고려하여 정규분포와 감마분포를 이용하였다. 감마분포는 스케일이 작은 경우 오른쪽으로 심하게 치우치고 분산이 큰 분포이어서 선택하였다. 각각의 분포를 가정하여 5개의 그룹에 각각 다른 값의 평균과 분산을 다르게 설정하여 각 그룹마다 50개의 난수를

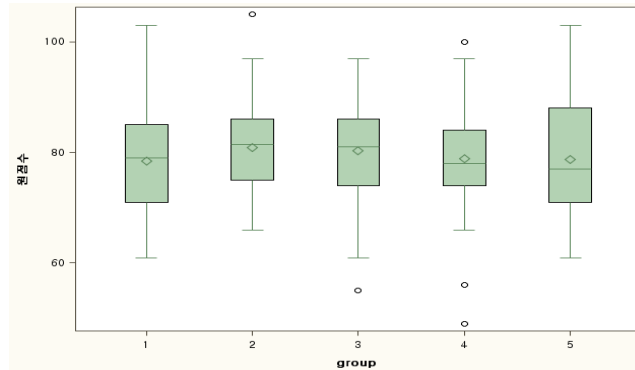


그림 2.3 포아송분포를 이용한 그룹

만들었으며 이 실험을 50회 실시하였다. 감마분포는 SAS의 rangam 함수를 사용하였다. 감마함수의 shape 모수는 첫 번째 그룹부터 5, 6, 7, 8, 9로 스케일 모수는 3, 4, 5, 6, 7로 설정하였다. 그림 2.3에서 보는 바와 같이 평균과 분산이 아주 다르게 그룹별로 설정된 것을 알 수 있다. 정규분포는 감마분포의 평균과 분산 값에 대응되게 SAS의 rannor 함수를 이용하여 5그룹에 각각 50개의 난수를 만들었으며 역시 50회 실험을 하였다. 아래 그림은 감마분포를 이용하여 만든 그룹의 예이다.

2.2.3. 비교방법

모의실험에 대한 비교평가는 상위 1등부터 일정 등수까지를 합격선이라 가정했을 때 합격권에 속하는 비율, 즉 합격권에 속하면 1, 그렇지 않으면 0으로 분류하여 원 점수 순위와 표준점수 순위가 일치하는 정 분류비율로 하였다. 이 비율을 상위 20% 50명, 30% 75명, 40% 100명에 대하여 적용하여 비율에 대한 일관성도 살펴보았다. 또한, 전체적인 순위의 일치 도를 비교하기 위하여 피어슨 상관계수를 구하였다. 두 순위가 많이 일치하는 때의 상관계수가 그렇지 않은 경우에 비하여 높으므로 전체 순위에 대한 일치 도를 평가하기 위한 적절한 방법이다. 순위에 대한 상관도이기 때문에 Spearman의 상관도는 피어슨 상관도와 같은 결과를 얻었다.

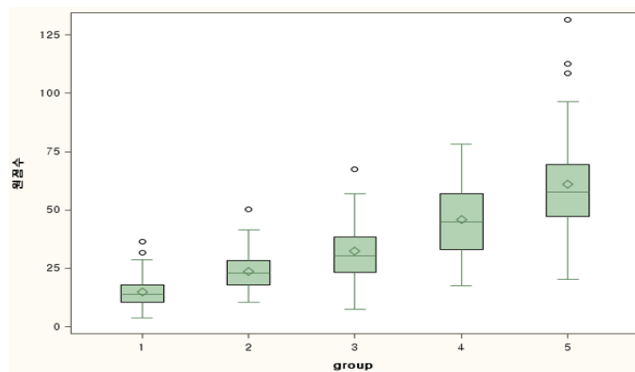


그림 2.4 감마분포를 이용한 그룹

3. 3 연구결과

3.1. 그룹의 중심변동이 비교적 작고 산포도가 다른 그룹

아래의 표를 살펴보면 평균이 같고 분산이 다른 정규분포의 경우나 포아송분포의 경우 같은 결과를 보여주는 것을 알 수 있다. 정규분포는 Range를 이용한 표준화가 상위 20%, 30%, 40%까지 가장 불안정 결과를 보여주고 표준편차를 스케일로 이용한 STD가 근소하나마 다른 두 표준화 방법에 비하여 좋은 결과를 보여주고 있다. 중심을 중앙값으로 이용한 MAD나 IQR는 STD와 비슷한 결과를 보여 주고 있다. 두 분포에서 Range를 이용한 방법이 다른 3 방법에 비하여 결과가 좋지 않았다. 상관계수의 평균을 비교한 결과를 보면 정규분포에서나 포아송분포에서 STD, MAD, IQR과 Range는 유의수준 0.05에서 차이가 있었다. STD방법을 이용한 상관계수가 가장 높고 Range가 가장 낮은 결과를 보여 주었다. 이는 전체적으로 STD방법이 Range를 이용한 방법보다 전체적인 순위가 일치하고 위험도가 낮다는 것을 의미한다. 50회 상관계수 평균에 대한 표준편차도 Range가 다른 표준화 방법에 비하여 높게 나타났다. 이는 Range의 스케일이 최대, 최소값으로 결정되기 때문에 다른 방법에 비하여 안정적이지 못하다는 것을 보여주고 있다.

표 3.1 정 분류비율 비교

기준	정규분포			포아송		
	50	75	100	50	75	100
STD	17.256	27.368	38.064	19.536	29.592	39.536
Range	16.160	25.672	35.808	17.496	27.232	36.784
MAD	17.096	26.944	37.688	19.056	28.840	38.816
IQR	16.976	26.952	37.704	19.040	28.936	39.064

표 3.2 상관계수 비교

	정규분포 상관계수 (표준편차)		포아송분포 상관계수 (표준편차)	
	상관계수	표준편차	상관계수	표준편차
STD	0.980	(0.006)	0.983	(0.012)
Range	0.949	(0.034)	0.946	(0.059)
MAD	0.974	(0.013)	0.981	(0.018)
IQR	0.973	(0.013)	0.981	(0.018)

3.2. 그룹의 중심과 산포도가 아주 다른 그룹

표 3.3과 3.4의 결과에서 보듯, 중심변동이 작은 경우와 달리 정 비율이나 상관계수가 매우 낮게 나타났다. 이는 원점수 순위와 표준점수 순위에 많은 변동이 있다는 것을 의미한다. 상위 50, 75, 100명까지의 정 비율에서도 큰 차이를 보였으며, 상관계수 평균은 STD는 0.98에서 0.58로 떨어졌다. 그러나 표준화 방법 4가지 결과는 앞의 경우와 달리 정규분포나 감마분포에서 정 비율이나 상관계수에서 차이가 없었다. 각 방법에 따른 상관계수평균에 대한 표준편차를 살펴보면 앞의 경우와 같이 STD의 경우가 가장 안정된 것을 알 수 있고 Range의 결과가 가장 불안정하다는 것을 알 수 있다. 이는 Range의 스케일이 최대값과 최소값에 의해서만 결정되기 때문에 상관계수가 높을 때와 낮을 때가 차이가 크다는 것을 다시 한 번 알 수 있었다. 예상과 달리 STD는 극한값에 많은 영향을 받음에도 MAD나 IQR와 비슷한 결과를 보여 주었다.

표 3.3 정 분류비율 비교

기준	정규분포			감마분포		
	50	75	100	50	75	100
STD	8.95	15.63	24.38	9.90	16.45	24.70
Range	8.66	15.46	23.74	9.86	16.74	24.81
MAD	9.17	15.77	24.41	9.65	16.26	24.47
IQR	9.18	15.76	24.41	9.84	16.38	24.52

표 3.4 상관계수 비교

	정규분포 상관계수 (표준편차)	감마분포 상관계수 (표준편차)
STD	0.576 (0.036)	0.562 (0.031)
Range	0.54 (0.125)	0.577 (0.119)
MAD	0.572 (0.047)	0.553 (0.048)
IQR	0.572 (0.047)	0.552 (0.049)

3.3. 상관계수와 산포도비율의 관계

아래의 표는 산포도비율과 상관계수와의 상관도이다. 산포도비율은 각 분포에 따라 만들어진 5개 그룹을 표준화할 때 표준화 방법에 따라 만들어진 5개 스케일 중 최대스케일을 최소스케일로 나눈 값이다. STD를 예를 들면 5개 그룹에서 최대 표준편차를 최소 표준편차로 나눈 비율이다. 포이송분포는 5개 그룹에 값은 평균값을 설정하였기에 산포도비율이 낮았고, 감마분포는 형상모수, 척도모수를 다르게 설정하였기에 이 비율이 높았다. 각 분포에 대한 상관계수와 산포도비율에 대한 관계는 아래의 표3.5 에서 보는 바와 같이 유의하게 나타났다. 즉 표준화 과정에서 산포도비율이 높을수록 원점수 순위와 표준점수 순위의 변동 가능성이 있다는 것이다. 그러므로 특정 경우에서 위험도가 작은 표준화 방법을 택하려면, 산포도비율이 작은 표준화 방법을 고려하여 택하면 순위의 변동을 줄일 가능성이 있다.

표 3.5 상관계수와 산포도비의 관계

	정규분포 (중심 고정)	포이송분포	정규분포	감마분포
상관계수	-0.228	-0.221	-0.184	-0.349
P-값	0.0012	0.0017	0.0009	0.001
산포도비율	2.06-2.17	1.19-1.36	3.22-3.35	3.24-3.47

표 3.6은 실제 입시자료에 4가지 표준화 방법을 적용한 결과이다. 보는 바와 같이 Range에 대한 산포도비율이 작아서 Range를 이용한 결과가 정 분류비율이나 상관계수에서 다른 방법에 비하여 가장 좋게 나왔다. 다음으로 MAD나 IQR의 산포도비율이 STD보다 높지만 상관계수는 높게 나왔다. 이 자료만을 고려한다면 Range를 이용하여 표준화점수를 산출 하는 것이 바람직하다. 하지만 앞으로 이와 비슷한 입시자료에 지속적으로 같은 방법을 적용할 것이기 때문에, 일시적으로 위험도가 낮은 표준화 결과를 만든 Range보다 안정적인 STD를 선택하여 표준화 점수를 산출하는 것이 바람직할 것이다.

표 3.6 입시성적에 대한 표준화 결과비교

	STD	Range	MAD	IQR
산포도비율	1.71	1.44	2.25	2.55
50까지 정 분류비율	15.25	16.94	11.86	13.56
75까지 정 분류비율	22.03	23.72	22.03	22.03
100까지 정 분류비율	27.97	31.36	28.81	28.81
상관계수	0.85	0.92	0.89	0.88

4. 결론

본 연구를 통해서 그룹의 중심이 작고 산포도가 다른 경우나 중심과 산포도가 다른 경우나 표준화를 취할 때 STD 방법이 가장 안정적이고 좋은 결과를 보여주고 Range 방법을 취하는 것이 불안정하고 위험이 크다는 것을 알 수 있었다. MAD나 IQR은 극한값에 대한 장점이 있음에도 STD보다 좋은 결과를 보여 주지 못했다. 또한, 산포도비율은 상관계수와 유의하지만 약한 관계가 있다. 그러므로 표준화 방법을 택할 때, 산포도비율을 고려할 수 있지만 절대적으로 영향을 주지 않는다는 것을 알 수 있었다. 입시면접이나 취업면접 등 일시적으로 특수한 경우에서 위험도가 낮은 표준화 방법을 찾는다면 산포도비율을 보고 방법을 고려 할 수도 있지만, 지속적으로 적용할 안정적인 방법을 찾는다면 4가지 방법 중에서 STD 방법이 적절한 표준화 방법이 될 것이다.

참고문헌

- 황진수, 김지연 (2009). 마이크로어레이 자료에서 서포트벡터머신과 데이터 템플을 이용한 분류방법의 비교. <한국데이터정보과학회지>, **20**, 311-319.
- Goodall, C. (1983). M-estimators of location: An outline of theory. In *Understanding robust and exploratory data analysis*, edited by H. Mosteller, John Wiley & Sons, Inc., New York, 339-403.
- Iglewicz, B. (1983). Robust scale estimators and confidence intervals for location. In *Understanding robust and exploratory data analysis*, edited by H. Mosteller, John Wiley & Sons, Inc., New York, 404-431.
- Hwang, C. (2006). Switching regression analysis via fuzzy LS-SVM. *Journal of the Korean Data & Information Science Society*, **17**, 609-617.
- Park, S. (2004) A improved method for constructing confidence interval of median: Small sample case. *Journal of the Korean Data & Information Science Society*, **15**, 973-980.
- Peter, J. R. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, **88**, 1273-1283.
- SAS/STAT User's guide (1988). *The STDIZE procedure*, vol. 3, 3117-3152.

The study of comparisons of standardization methods[†]

Dae Kee Min¹ · Ji Hyeon Jung²

¹Department of Information & Statistics, Duksung Women's University

²Department of Statistics, Duksung Women's University

Received 5 December 2011, revised 28 December 2011, accepted 5 January 2012

Abstract

When we evaluate prospective students in the interview process, we have to implement a system in which each student can be fairly judged. This process, the standardization of the scores which the interviewers have produced based on a student's performance, is implemented to ensure that each student receives a score that objectively translates one's performance. Although we don't know exactly how effective the standardization is in many different cases, we have researched which standardization methods are most stable and have minimum risks among the four methods such as STD, Range, MAD and IQR. These methods use scales such as standard deviation, range, maximum median and interquartile range.

Keywords: Interquartile, median, range, standardization.

[†] This paper was supported by Duksung Women's University in 2010.

¹ Corresponding author: Professor, Department of Information & Statistics, Duksung Women's University, 419 Ssangmoon_dong Dobong-gu, Seoul 132-714, Korea. E-mail: dkmin@duksung.ac.kr

² Master, Department of Statistics, Duksung Women's University, 419 Ssangmoon_dong Dobong-gu, Seoul 132-714, Korea.