Journal of the Korean Data & Information Science Society
2011, **22**(2), 189–196

한국데이터정보과학회지

면접점수 표준화 방법 모의실험 비교

박철용[1]

[1]계명대학교 통계학과

접수 2011년 1월 23일, 수정 2011년 2월 21일, 게재확정 2011년 2월 25일

요 약

이 연구에서는 면접점수 표준화 방법으로 흔히 사용되고 있는 절사평균 방법, 순위평균 방법 및 z-점수평균 방법을 모의실험을 통해 비교하고자 한다. 모의실험 기법은 피면접자의 참값 점수와 이것과 독립적인 잡음 변수가 심사자의 전문성에 의해 가중평균 형태로 심사자의 평가점수에 영향을 미친다고 가정한다. 다시 말해 심사자의 전문성이 커지면 개인의 참값 점수에 가까운 심사자의 점수가 관측되고, 심사자의 전문성이 작아지면 참값 점수 대신에 잡음 변수에 더 가까운 심사자의 점수가 관측된다. 여기에 심사자의 성향편의가 더해져 심사자의 최종 평가점수가 관측된다고 가정한다. 이 모의실험에서는 각 표준화 방법에 의한 심사자의 평균점수와 참값의 순위상관 값을 계산하여 이 값이 큰 방법을 좋은 방법으로 평가하였다. 그 결과 참값의 분포가 정규분포이면 z-점수평균이 가장 좋은 성능을 보였으며, 라플라스분포이면 전체면접에서는 z-점수평균이 순위평균보다 다소 성능이 좋았으나 반분면접에서는 순위평균이 z-점수평균보다 다소 성능이 좋았다. 절사평균은 일반적으로 성능이 가장 낮게 나타났다.

주요용어: 순위상관, 순위평균, 절사평균, z-점수평균.

1. 머리말

한국에서 대학입시는 관련되어 있는 모든 국민들이 몰두하는 아주 중요한 과제로 인식되어 있다. 그런데 송필준과 김종태 (2010)와 윤용화와 김종태 (2010)의 연구에서도 알 수 있듯이 향후 입학자원의 감소가 두드러지게 전개될 것이기 때문에 대학에서는 홍보전략 수립과 우수학생 유치를 위한 다양한 연구와 활동이 전개하고 있는 실정이다 (최승배 등, 2009; 최승배 등, 2011).

대학입시에서 면접이 차지하는 비중은 점차 확대되어 왔다. 이전에는 주로 수능 점수와 내신과 같은 지필고사를 통해 평가하였으나 지필고사를 통해 평가하기 어려운 인성 및 적성 등을 평가하기 위해서 면접의 중요성이 강조되고 있기 때문이다. 그런데 면접의 강화는 수험생에게 상당한 부담을 안겨준 것이 사실이지만, 각 대학에서도 유능한 학생들을 자기 학교로 유치하기 위해서 객관적인 면접점수 표준화 방법을 도입하여야 하는 부담을 안게 되었다.

대학입시에서 객관성을 보장하기 위한 연구들은 수능시험에서 표준점수 제도가 도입되면서 본격적으로 진행되어 왔다. 허명회 (1994)는 표준점수제에서의 교육측정의 기본 이론을 다루고 있으며, 박성현 등 (2000)은 대학입시에서의 선택과목점수 표준화에 있어 등분위수 등화가 선택과목 점수의 변별력을 증가시키는 효과가 있다고 하였다. 또한 황형태 (2005)는 선택과목별 가산점수제를 통해 표준점수제의 보완을 제안하였다.

[1] (704-701) 대구광역시 달서구 신당동 1000번지, 계명대학교 통계학과, 교수. E-mail: cypark1@kmu.ac.kr

황형태 등 (2004)은 각 심사자 별 순위 (rank)에 근거한 면접점수 표준화 방법을 제안하였다. 이 방법은 기본적으로 각 심사자의 점수에 순위를 매기고 그 순위에 해당되는 표준정규분포에 근거한 점수를 부여하는 방식이다. 특히 상위 5%와 하위 5%에는 동일한 점수를 부여하여 강건성 (robustness)을 증대시켰다.

표준점수제와 순위에 근거한 표준화 방법과 더불어 체조와 다이빙과 같이 여러 명의 심판이 심사하는 스포츠 종목에서는 절사평균 (trimmed mean)을 이용한 표준화 방법이 많이 사용되고 있다. 예를 들어 5명의 심판이 심사하는 경우 최하와 최상의 점수를 버리고 중간 3명의 심판의 점수의 평균을 사용하는 것이다.

이 연구에서는 모의실험을 통해 세 가지 표준화 방법, 즉 절사평균, 순위평균 및 z-점수 (z score)평균 방법을 비교하고자 한다. 모의실험의 기법은 피면접자의 참값 점수와 이것과 독립적인 잡음 변수가 심사자의 전문성에 의해 가중평균 형태로 심사자의 평가점수로 관측된다고 가정한다. 다시 말해 심사자의 전문성이 커지면 피면접자의 참값 점수에 가깝게 심사점수가 관찰되고, 심사자의 전문성이 작아지게 되면 피면접자의 참값 대신에 잡음 변수에 가깝게 심사자의 점수가 관측된다.

황형태 (2005)에서도 언급되었듯이 z-점수를 과목별 표준점수로 변환하기 위해 수능시험에서 사용하고 있는 선형변환에도 문제가 발생할 수 있기 때문에, 이 연구에서는 z-점수 자체를 사용하였다. 즉 심사자들의 평균 z-점수를 z-점수평균 방법에 의한 최종 면접점수로 간주하였다. 또한 심사자들의 순위평균을 순위평균 방법에 의한 최종 면접점수로 간주하였다.

요약하면 세 가지 방법에 의해 관측되는 심사자의 점수는 절사평균을 제외하고는 모두 피면접자의 참값과 다른 척도를 가지게 된다. 따라서 이 연구에서의 모의실험에서는 순위의 중요성만 따지는 Kendall (1938)의 순위상관계수를 이용하여 세 가지 방법이 얼마나 충실히 피면접자의 참값의 순서를 유지하는지 비교하고자 한다. 순위상관계수를 사용함으로 해서 z-점수평균과 순위평균에 어떠한 단조증가함수 (monotone increasing function)에 의한 변환을 하여 최종 면접점수를 부과하더라도 동일한 상관계수 값을 얻을 수 있게 된다.

황형태 등 (2004)에서 사용된 방법은 각 심사자의 피면접자들 점수에 순위를 매긴 후 그 순위에 해당되는 정규분포 근거 점수를 부여하고 심사자의 평균을 최종 면접점수로 간주하였으나, 이 연구에서는 단순히 각 심사자별 순위평균을 사용하였다. 순위상관계수 관점에서 보자면 황형태 등 (2004)에서 상위 5%와 하위 5%에는 동일한 점수를 부여하는 부분만 없다면 거의 동일한 결과를 얻을 수 있을 것이다. 다만 심사자별 평균을 최종 면접점수로 하기 때문에 최종 면접점수에는 약간의 차이가 발생할 수 있을 것이다.

이 논문은 다음과 같이 구성되어 있다. 2절에서는 이 논문에서 제안하고 있는 모의실험 방법과 모의실험 결과를 제시한다. 3절에서는 이 연구의 결론과 한계점 및 추후연구과제에 대해서 논의한다.

2. 모의실험

이 절에서는 먼저 심사자가 피면접자의 면접점수를 부여하는 간단한 모형을 제시하고자 한다. 이 모형에서는 심사자의 면접점수가 피면접자의 참값과 이와 독립인 잡음 변수의 가중평균으로 얻어지게 되며, 그 가중값은 심사자의 전문성에 의해 결정된다고 가정한다. 여기에 심사자의 성향편의가 더해져 최종 평가점수가 구해진다고 가정한다. 그 다음에 이 모형의 여러 모수의 변화에 따른 모의실험 결과를 토대로 세 가지 표준화 방법을 비교하고자 한다. 심사자가 모든 피면접자를 심사하느냐에 따라 전체면접과 반분면접으로 나누어 고려하였으며, 피면접자를 평가하는 심사자의 수는 절사평균에 적합한 3명과 5명인 경우를 고려하였다. 또한 피면접자의 참값의 분포로는 표준정규분포와 라플라스분포를 고려하였으며 심사자의 전문성의 분포는 균일분포를 고려하였다. 추가로 심사자의 성향편의가 존재한다고 가정

하였으며 또한 피면접자 개개인에 따라 심사자의 관측오차가 존재한다고 가정하였다. 성향편의와 관측오차의 분포는 평균이 0이되 참값의 분산보다 작은 분산을 가지는 정규분포를 사용하였다.

전문성을 θ, 피면접자의 참값을 X, 잡음 변수를 W, 심사자의 성향편의를 B이라고 표시하고 피면접자 관측오차를 ϵ이라 표기하자. 전문성 θ, 피면접자의 참값 X, 잡음 변수 W 및 성향편의 B가 주어졌을 때 심사자의 평가점수 Y는 다음과 같이 관측된다고 가정한다.

$$Y = \theta X + (1-\theta)W + B + \epsilon.$$

X의 분포는 표준정규분포와 이보다 꼬리가 두꺼운 라플라스분포를 고려한다. 일반적으로 z-점수는 정규분포에서 좋은 성능을 발휘하고 순위는 정규분포에서 벗어날 때 좋은 성능을 보이기 때문에 이 두 분포를 고려한 것이다. 그리고 θ의 분포는 균일분포를 사용하였다. 심사자의 전문성에 따라 각각 0.9 ± 0.1와 0.75 ± 0.1 사이의 균일분포에서 난수를 생성하여 전문가 집단, 비전문가 집단이라 명명하였다. 보통의 상식을 가진 심사자라면 피면접자의 참값에 3/4, 잡음 변수에 1/4 정도의 비중을 가지고 관찰하며, 전문가라면 참값에 대한 비중이 0.9 정도는 되어야 한다고 생각하였기 때문이다. 잡음 변수 W의 분포는 X의 분포와 같이 하는 것도 고려하였지만, X의 분포를 안다고 가정하여 사용하는 것보다 이와 상관없는 균일분포를 사용하는 것이 현실적이라 생각하였다. 심사자의 성향편의 B는 후하게 (혹은 인색하게) 점수를 매기는 성향이 있는 사람은 모든 피면접자들에게 동일하게 후하게 (혹은 인색하게) 점수를 매기는 점을 고려하여 포함시켰는데 평균이 0인 정규분포를 사용하였다. 심사자의 성향편의는 심사자가 모든 피면접자를 심사하지 않는 면접에서는 흔히 발생하는 문제이기 때문에 이 모형에 포함시켰다. 피면접자를 대하는 심사자의 관측오차 ϵ는 통상적으로 사용되는 평균이 0인 정규분포를 사용하였다. 전문성에 의한 가중평균을 사용하기 때문에 X와 W의 분산은 1로 고정시켰으며 B와 ϵ의 표준편차는 각각 1/2과 1/3로 잡아 체계적인 변동부분의 1/2, 1/3에 해당되도록 하였다.

심사자 $i = 1, 2, \cdots, p$가 피면접자 $j = 1, 2, \cdots, m$의 평가점수를 얻는 과정을 간략하게 정리하면 다음과 같다. 참값 X_j를 가진 피면접자가 θ_i $(i = 1, \cdots, p)$의 전문성과 B_i $(i = 1, \cdots, p)$의 성향편의를 가지는 심사자 그룹에 들어오면 심사자들은 $Y_{ij} = \theta X_j + (1-\theta_i)W_j + B_i + \epsilon_{ij}$로 심사점수를 부여한다. 여기서 W_j는 잡음 변수로 피면접자와의 참값 X_i와 무관하며 심지어 균일분포에서 난수를 생성하여 사용해도 상관없다. 이 심사점수는 평균적으로 $\theta X_j + (1-\theta_i)W_j + B_i$이며 ϵ_{ij} 만큼의 관측오차를 동반하게 된다. 안정적인 모의실험 결과를 얻기 위해서는 m이 커지면 좋겠지만, 현실적으로 한 심사자가 면접할 수 있는 피면접자의 수에는 한계가 있기 때문에 $m = 50$ (반분면접) 혹은 $m = 100$ (전체면접)으로 고정하였으며, 결과적으로 전체 피면접자의 수는 $n = 100$으로 고정되었다.

구체적인 모의실험 설계는 다음과 같다. 세 가지 표준화 방법을 비교하기 위해 제일 먼저 심사자 모두가 피면접자 모두를 면접하는 전체면접과 심사자가 두 그룹으로 나뉘어 피면접자의 절반만 면접하는 반분면접으로 나누어 모의실험한다. 따라서 전체면접 시 필요한 심사자는 p명이고 100명을 모두 면접하게 되며, 반분면접 시는 $2p$명의 심사자가 두 그룹으로 나뉘어 각 50명을 면접하게 된다. 다음으로 면접자의 수 p가 3 혹은 5가 될 때 세 가지 방법의 성능이 어떻게 달라지는지 비교한다. 이는 절사평균이 사용될 수 있는 현실적인 심사자의 인원이라 생각되어 사용되었다. 또한 θ가 0.9 ± 0.1 사이의 균일분포를 따르는 전문가 집단과 0.75 ± 0.1 사이의 비전문가 집단에 따른 성능비교와 함께 참값 X의 분포가 표준정규분포와 라플라스분포로 변동할 때의 성능비교를 시도한다.

세 가지 방법에서 최종 면접점수를 얻는 과정을 간략히 살펴 보면 다음과 같다. 먼저 절사평균 방법은 $Y_{ij},\ i = 1, 2, \cdots, p$ 중 최소값과 최대값을 제외한 표본평균으로 최종 면접접수를 정한다. 순위평균 방법은 $Y_{ij},\ j = 1, 2, \cdots, m$ 중의 순위 R_{ij}를 구한 후 심사자들의 순위평균 $\sum_i R_{ij}/p$로서 최종 면접점수를 정한다. 마지막으로 z-점수평균 방법은 $Y_{ij},\ j = 1, 2, \cdots, m$의 z-점수 Z_{ij}를 구한 후 심사자들의 z-점수평균 $\sum_i Zij/p$로서 최종 면접점수를 정한다.

세 가지 방법의 성능을 비교하는 기준은 서론에서도 언급되었듯이 Kendall (1938)의 순위상관계수이다. 다시 말해 참값과 세 가지 방법의 최종 면접점수 간의 순위상관계수를 계산하여 이 값이 가장 큰 방법을 최고의 성능을 가지는 방법으로 인정하는 것이다.

1000번의 반복실험을 통해 순위상관계수의 표본평균과 표준오차를 계산하여 정리한 것이 표 2.1에 주어져 있다.

표 2.1 참값과 세 가지 방법의 최종 면접점수 간 순위상관계수의 표본평균(표준오차)

				표준화 방법		
면접유형	심사자수	참값 분포	심사자집단	절사평균	순위평균	z-점수평균
전체면접	3	표준정규	전문가	.8077(.0309)	.8461(.0235)	.8472(.0233)
			비전문가	.7106(.0487)	.7405(.0413)	.7411(.0412)
		라플라스	전문가	.7691(.0393)	.8128(.0318)	.8142(.0318)
			비전문가	.6667(.0543)	.6976(.0491)	.6980(.0489)
	5	표준정규	전문가	.8601(.0217)	.8704(.0208)	.8714(.0207)
			비전문가	.7499(.0370)	.7557(.0352)	.7564(.0353)
		라플라스	전문가	.8297(.0284)	.8414(.0273)	.8424(.0271)
			비전문가	.7091(.0453)	.7161(.0442)	.7161(.0442)
반분면접	3	표준정규	전문가	.7668(.0566)	.8304(.0254)	.8338(.0256)
			비전문가	.6731(.0596)	.7303(.0352)	.7318(.0352)
		라플라스	전문가	.7261(.0613)	.7986(.0317)	.7975(.0334)
			비전문가	.6279(.0657)	.6913(.0429)	.6905(.0433)
	5	표준정규	전문가	.8221(.0467)	.8514(.0247)	.8546(.0250)
			비전문가	.7254(.0452)	.7466(.0326)	.7477(.0324)
		라플라스	전문가	.7892(.0552)	.8253(.0295)	.8228(.0330)
			비전문가	.6803(.0564)	.7075(.0409)	.7062(.0417)

이 표에서 다음과 같은 일반적인 경향을 읽을 수 있다. 반분면접보다 전체면접에서, 심사자 3인보다 심사자 5인에서, 라플라스분포보다 표준정규분포에서 또한 비전문가보다 전문가 심사자 집단에서 각각 더 높은 순위상관계수 값을 얻을 수 있다. 이는 심사가가 상대하는 피면접자의 수가 많은 경우, 심사자가 많은 경우, 꼬리가 얇아 이상값이 적게 나오는 분포인 경우 및 전문가인 경우 각각 훨씬 신뢰성이 높은 심사점수를 얻을 수 있다는 상식을 반영하고 있다.

절사평균은 다른 두 가지 방법에 비해 성능이 떨어지는 것을 관측할 수 있다. 비록 미세한 차이처럼 보이지만 표준정규분포인 경우에는 z-점수평균이 순위평균보다 더 좋은 성능을 보이고 있으며, 라플라스분포인 경우에는 전체면접에서는 z-점수평균이, 반분면접에서는 순위평균이 더 좋은 성능을 보였다.

z-점수평균과 순위평균 간의 차이가 미세해 보이는 것 같아 통계적으로 의미가 있는 차이인지 알아보기 위해 가능한 모든 두 집단 간 차이를 비교하는 쌍체 t검정 (paired t-test)을 시도하였다. 여기서 쌍체 t검정을 이용한 이유는 세 개의 평균 사이에는 양의 상관성이 강하게 나타나기 때문이다. 쌍체 t검정의 결과는 표 2.2에 주어져 있다.

이 표에는 각 쌍체 t검정통계량 값의 유의성을 손쉽게 알 수 있도록 별표를 사용하였다. 구체적으로 별표가 하나이면 0.05 유의수준에서 유의적인 차이가 있는 경우이며, 별표가 두 개이면 0.01 유의수준에서 유의적인 차이가 있는 경우이다. 이 유의성 검정의 결과 전체면접, 심사자가 5명, 라플라스분포이며 비전문가 심사자 집단의 조건을 모두 만족하는 경우에 z-점수평균과 순위평균 간에 유의적인 차이가 없을 뿐 모든 짝에 대해서 유의적인 차이가 존재하는 것으로 나타났다. 물론 다중비교에 따른 실제 유의확률의 변화를 생각하더라도 t검정통계량 값의 절대값이 상당히 커서 소수 몇 개의 짝을 제외하고는 대부분 유의적인 차이가 있는 것으로 판정될 것이라 생각한다.

표 2.1에서는 미세한 차이처럼 느껴졌는데 왜 쌍체 t검정에서는 유의적인 차이가 발생한 것일까? 가

표 2.2 세 가지 방법 간 순위상관계수 값의 차에 대한 쌍체 t검정통계량

면접유형	심사자수	참값 분포	심사자집단	표준화 방법		
				z-점수, 순위 평균차	z-점수, 절사 평균차	순위, 절사 평균차
전체면접	3	표준정규	전문가	7.6656**	57.4963**	56.1015**
			비전문가	4.4135**	36.2571**	35.4722**
		라플라스	전문가	8.8801**	54.2658**	52.7160**
			비전문가	2.7993**	34.5798**	33.8589**
	5	표준정규	전문가	8.0981**	34.6540**	30.9756**
			비전문가	4.9357**	15.7687**	13.8138**
		라플라스	전문가	7.2551**	32.8083**	29.4210**
			비전문가	-0.0031	15.2688**	15.2275**
반분면접	3	표준정규	전문가	13.1697**	37.9734**	35.7754**
			비전문가	6.8557**	35.8227**	34.7184**
		라플라스	전문가	-3.0909**	39.3591**	40.9835**
			비전문가	-2.5270*	37.4957**	38.0871**
	5	표준정규	전문가	11.8966**	21.5885**	19.3570**
			비전문가	5.5199**	19.1163**	18.0992**
		라플라스	전문가	-6.3992**	19.3487**	21.4832**
			비전문가	-4.7486**	19.1467**	20.3781**

*. 쌍체 t-검정통계량이 0.05 유의수준(양쪽)에서 유의하다.
**. 쌍체 t-검정통계량이 0.01 유의수준(양쪽)에서 유의하다.

장 중요한 원인은 z-점수평균과 순위평균 방법 사이에는 상당히 큰 상관관계를 가진다는 것이다. 실제로 두 방법 간의 피어슨 상관계수를 구하여 보았더니 가장 작은 것이 0.926이며 가장 큰 것이 0.994에 이르렀다. 상관계수가 크기 때문에 차의 분산은 근사적으로 다음의 관계를 만족한다.

$$s_d^2 \approx (s_1 - s_2)^2$$

여기서 s_1, s_2은 각각 z-점수평균과 순위평균의 표준편차이다. 표 2.1에서 이 두 값의 차이가 상당히 작기 때문에 차의 분산은 더욱 작아지게 되는 것이다.

큰 상관계수와 더불어 반복실험 횟수가 1000이라는 점도 유의적인 차이가 발생하는 데 중요하게 작용하였다. 실제로 차의 변동계수 (coefficient of variation)가 $\bar{d}/s_d = 0.1$에 불과하더라도 t검정통계량 값은 $0.1\sqrt{1000} = 3.16$에 이르게 되어 유의적인 차이가 있는 것으로 나타나게 되는 것이다. 이러한 두 가지 이유로 미약한 차이로 보였던 z-점수평균과 순위평균 간의 차이가 대부분 유의적으로 의미 있는 차이로 나타나게 된 것이다.

3. 결론과 논의

이 연구에서는 면접점수 표준화 방법으로 흔히 사용되고 있는 절사평균 방법, 순위평균 방법 및 z-점수평균 방법을 모의실험을 통해 비교하였다. 모의실험 기법은 피면접자의 참값 점수와 이것과 독립적인 잡음 변수가 심사자의 전문성에 의해 가중평균 형태로 심사자의 평가점수에 영향을 미친다고 가정하였다. 다시 말해 심사자의 전문성이 커지면 개인의 참값 점수에 가깝게 심사자의 점수가 관찰되고, 심사자의 전문성이 작아지면 참값 점수 대신에 잡음 변수에 더 가깝게 심사자의 점수가 관찰되는 것이다. 여기에 심사자의 성향편의가 더해져 심사자의 최종 평가점수가 관측된다고 가정한다.

절사평균 방법은 3인 혹은 5인 심사자 중에서 최소와 최대 점수를 제외하고 피면접자의 최종 면접점수로 정하는 방법이다. 순위평균 방법은 각 심사자별로 순위를 매겨 심사자들의 평균으로 피면접자의 최종 면접점수를 정하는 방법이다. 마지막으로 z-점수평균 방법은 각 심사자별로 z-점수를 구하여 심

사자들의 평균으로 피면접자의 최종 면접점수를 정하는 방법이다. 이 모의실험에서는 각 표준화 방법에 의한 심사자의 평균점수와 참값의 순위상관 값을 계산하여 이 값이 큰 방법을 좋은 방법으로 평가하였다. 순위상관을 사용함으로써 각 방법의 평균점수에 어떠한 단순증가함수 (monotone increasing function)를 적용하여 면접점수가 결정되는 시스템이라도 동일한 결과를 얻을 수 있다. 따라서 이 모의실험 결과는 절사평균, 순위평균 및 z-점수평균의 어떠한 단순증가함수에 의해 면접점수를 결정하는 시스템에도 적용될 수 있는 것이다.

모의실험에서는 여러 가지 모수의 변화를 시도하였다. 먼저 심사자가 모든 피면접자를 평가하는 전체면접과 심사자를 반분하여 각각 피면접자 절반을 면접하는 반분면접을 고려하였다. 또한 심사자 수를 절사평균이 잘 적용될 수 있으며 현실적인 3인과 5인을 고려하였다. 또한 피면접자의 참값의 분포로 표준정규분포와 라플라스분포를 고려하였으며, 심사자 집단으로는 전문성 지수가 평균적으로 좋은 전문가 집단과 전문성이 다소 떨어지는 비전문가 집단을 고려하였다. 모의실험 결과 참값의 분포가 정규분포이면 z-점수평균이 가장 좋은 성능을 보였으며, 라플라스분포이면 전체면접에서는 z점수평균이 순위평균보다 다소 성능이 좋았으나 반분면접에서는 순위평균이 z-점수평균보다 다소 성능이 좋았다. 절사평균은 일반적으로 성능이 가장 낮게 나타났다. 이는 z-점수평균이 정규분포에서 좋은 성능을 나타내지만, 꼬리가 두꺼운 분포에서는 순위평균의 성능이 좋아져 전체면접에서 성능의 차이가 줄어 들고 반분면접에서는 마침내 z-점수를 능가한다는 결론이다.

이 모의실험에서 여러 가지 모수 변화를 시도하였으나 이 모의실험 결과를 일반적으로 확대해석하는 것은 곤란할 것이라 생각한다. 여기서는 그 결과를 제시하지 않았지만 실제로 심사자의 성향편의가 존재하지 않는 경우의 모의실험에서는 심사자 5인의 반분면접에서 절사평균이 가장 좋은 성능을 발휘하였다. 표준화 방법이 결국은 심사자의 성향편의를 극복하는 방법 중의 하나이기 때문에, 성향편의가 존재하지 않는다면 굳이 순위나 z-점수와 같은 변환을 시도하는 것보다 심사자들의 원점수를 사용하는 것이 좋은 성능을 보일 것이다. 이러한 이유로 심사자 5인인 반분면접에서는 최소와 최대를 제외한 3인의 원점수 평균을 사용하는 것이 반분의 피면접자를 대상으로 순위나 z-점수를 구하여 두 개를 결합하는 것보다 더 좋은 성능을 보인 것이라 생각된다. 그러나 심사자가 3인인 반분면접에서는 최소와 최대를 제외하면 1인의 원점수만 사용할 수 있기 때문에 순위평균과 z-점수평균보다 좋은 성능을 유지하지 못하고 결국 세 방법 중 가장 나쁜 성능을 보였다.

이 연구의 다음 연구과제로 생각할 수 있는 것은 심사자의 평가점수 모형을 참값과 잡음 변수의 가중평균이 아니라 좀 더 현실적인 모형으로 일반화하는 것이다. 그 첫 번째 방향은 참값과 독립적인 잡음 변수가 아니라 참값과 연관성이 있을 수 있는 새로운 변수를 도입하는 것이라 생각한다. 예를 들어 피면접자의 외모는 평가점수 결정에 영향을 미치며 참값과 연관되어 있다고 볼 수도 있기 때문에 이런 변수를 장애 변수라는 이름으로 모형에 포함시킬 필요가 있는 것이다. 다른 연구 방향은 참값과 잡음 혹은 장애 변수를, 가중평균이 아니라 쉽게 이해할 수 있으면서 현실적으로 적용 가능한 다른 함수를 사용하여 결합하는 시도가 될 수 있을 것이다.

참고문헌

박성현, 김춘원, 박준오 (2000). 대학입시에서의 선택과목 점수 표준화에 관한 연구. <품질경영학회지>, **28**, 124-132.

송필준, 김종태 (2010). 로지스틱함수모형과 비례이동평균모형에 의한 학생 수 추계와 분석. <한국데이터정보과학회지>, **21**, 503-511.

윤용화, 김종태 (2010). 수도권 지역의 고3학생 수 예측과 대학입학정원수와 분석. <한국데이터정보과학회지>, **21**, 523-534.

최승배, 강창완, 조장식 (2009). 웹 로그데이터를 이용한 대학입시 지원자 행태 분석. <한국데이터정보과학회지>, **20**, 493-504.

최승배, 강창완, 조장식 (2011). 학과 홈페이지 평가지수 개발에 관한 연구. <한국데이터정보과학회지>, **22**, 출간예정.
허명회 (1994). 새 대학입시의 통계적 계획과 분석 - 문항분석과 선택과목 등화(표준점수제)를 중심으로. <한국통계학회논문집>, **1**, 215-225.
황형태 (2005). 대학수학능력시험에서 표준점수제의 개선방안에 대한 연구. <응용통계연구>, **18**, 521-532.
황형태, 이강섭, 이장택 (2004). 대학입시에서의 면접점수 표준화에 관한 연구. <대한수학교육학회지 시리즈 A: 수학교육>, **43**, 309-314.
Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika,* **30**, 81-93.

Journal of the Korean Data & Information Science Society
2011, **22**(2), 189–196

한국데이터정보과학회지

Simulation comparison of standardization methods for interview scores

Cheolyong Park[1]

[1]Department of Statistics, Keimyung University

Received 23 January 2011, revised 21 February 2011, accepted 25 February 2011

Abstract

In this study, we perform a simulation study to compare frequently used standardization methods for interview scores based on trimmed mean, rank mean, and z-score mean. In this simulation study we assume that interviewer's score is influenced by a weighted average of true interviewee's true score and independent noise whose weight is determined by the professionality of the interviewer. In other words, as interviewer's professionality increases, the observed score becomes closer to the true score and if interviewer's professionality decreases, the observed score becomes closer to the noise instead of the true score. By adding interviewer's tendency bias to the weighed average, final interviewee's score is assumed to be observed. In this simulation, the interviewers' scores for each method are computed and then the method is considered best whose rank correlation between the method's scores and the true scores is highest. Simulation results show that when the true score is from normal distributions, z-score mean is best in general and when the true score is from Laplace distributions, z-score mean is better than rank mean in full interview system, where all interviewers meet all interviewees, and rank mean is better than z-score mean in half split interview system, where the interviewers meet only half of the interviewees. Trimmed mean is worst in general.

Keywords: Rank correlation, rank mean, trimmed mean, z-score mean.

[1] Professor, Department of Statistics, Keimyung University, Daegu 704-701, Korea.
E-mail: cypark1@kmu.ac.kr