응용통계연구 제10권 2호
1997년 9월 pp. 275-291

대표성분점수화법의 제안과 이의 타당성 및 활용성에 관한 연구[1)]

이 광 진[2)]

요 약

직접측정이 불가능한 단일 상위목표개념을 직접측정이 가능한 관련 하위개념들을 이용하여 간접적으로 측정한 경우, 많은 연구자들이 비록 모호하나마 연구초기에 이해했던 목표개념과 관련개념들간의 관계를 자료분석의 단계에서 구체화하지 못하고, 목표개념의 점수화를 편의성에 의존한다거나 통계분석자에게 일임해 버리는 경향이 있다. 그러나 사용된 점수화 방법에 대한 객관성이 결여되면 그릇된 정보가 도출되거나 연구결과가 독자들로부터 수용되어지지 못할 수도 있다.

본 연구에서는 이러한 상황에서 측정이론에 충실하면서 목표개념의 조작적 정의를 어느 정도 객관화할 수 있는 한 방법을 제안하고, 이에 바탕을 둔 대표성분점수화법의 타당성 및 활용성에 관해 연구하였다.

1. 서론

단일 목표개념을 염두에 두고 의도적으로 선정된 일단의 변수들이 연구집단의 각 대상들에 대해 측정되었을 때, 측정값들을 적절히 결합하여 목표개념에 대한 각 대상들의 점수를 구해야 하는 문제는 계량분석적 연구에서 흔히 발생하고 있다. 예를 들면, 여러 과목의 시험점수들을 결합하여 각 학생들의 학업성취도를 평가한다든가, 설문조사에서 동류의 문항점수들을 결합하여 각 응답자들의 대표점수를 구하고자 할 때 등이다. 이런 경우 분석의 초기단계에서 직면하게 되는 가장 큰 어려움 중의 하나가 바로 일변량 점수화 방법이다. 즉 측정된 변수들을 결합하여 각 대상들의 하나의 대표값을 정하는 가장 좋은 혹은 가장 적절한 결합방법이 무엇이냐 하는 것이다. 그러나 불행하게도 모든 면에서 보아 최상인 결합방법은 제시된 적도 없고 또한 있을 수도 없다. 사실, 이는 상정된 목표개념의 정의에 관련된 문제이기 때문에 그 목표개념의 객관적 정의가 존재하는 경우에는 아무런 문제가 없다. 하지만 그렇지 않은 경우 특별한 대안을 찾을 수 없다는 이유 때문에 기존에 편의적으로 사용되어져 온 단순평균법, 제1주성분점수법, 제1인자점수법 등이 무비판적으로 사용되어지고 있는 경우가 많다.

본 연구에서는 관련된 모든 변수들이 하나의 어떤 목표개념을 염두에 두고 의도적으로 선정된 경우에 있어서 문제가 되는 일변량 점수화방법에만 국한하여, 전통적으로 사용되어져온 기존의 점수화 방법들을 비교·고찰하고(2절), 측정이론에 충실하면서 목표개념의 조작적 정의를

1) 이 논문은 1994년도 교육부지원 한국학술진흥재단의 신진교수과제 학술연구조성비에 의해 연구되었음.
2) (301-729) 대전시 중구 목동 목원대학교 사회과학대학 응용통계학과 조교수.

어느 정도 객관화할 수 있는 한 방법을 제안하면서(3절), 이에 바탕을 둔 점수화 방법인 대표성 분점수화법의 여러 성질들을 유도함(4절)과 아울러 가상예제와 실제사례를 통해 기존의 방법들과 비교하고 몇가지 관련된 사항들을 토의한다(5절).

2. 대표점수화에 관한 기존방법들

2.1 산술평균법

설문응답의 분석과정에서 서로 관련되어 있는 문항들을 결합하여 대표점수를 구할 때, 또는 개별 시험점수들을 결합하여 각 학생들의 학업성취도 등을 구하고자 할 때 전통적으로 산술평균법이 가장 자주 사용되어지고 있다. 왜냐하면 이 방법은 개념상 단순하여 통계학에 대한 큰 지식이 없더라도 쉽게 이용할 수 있기 때문이다.

이 방법은 독립표본에서 모평균의 추정량으로 표본평균을 사용하고 있는 바를 변수들의 관계에 원용한 것이다. 그러나 본 연구에서 고려하는 바와 같이 어떤 한 개념을 염두에 두고 의도적으로 선정된 여러 변수들에 대한 대표점수를 구하고자 할 경우에 있어서는 이 방법은 변수들 간에 존재하는 일반적으로 서로 다른 상관관계의 정도를 전혀 고려하고 있지 않다는 점, 문항선정의 과정을 통해 악용 또는 왜곡될 가능성이 있다는 점, 그리고 변수들의 측정단위가 다를 경우 그 의미가 완전히 사라져 버린다는 점 등에서 비판의 소지가 있어 무분별하게 사용될 성질의 점수화방법은 결코 아니다.

2.2 제1주성분점수법

서로 상관되어 있는 변수들을 일변량화하는 방법으로 제1주성분점수법도 많이 사용되고 있다. 여기서 제1주성분이란 선정된 변수들의 선형결합으로서 원자료에 내재하는 전체변이 중 가장 큰 부분을 보유하는 하나의 인공변수를 말한다. 이는 정보의 손실을 최소화하는 차원축약의 의미가 큰 점수화법이라고 할 수 있다..

만약 대표점수화의 목적이 개체들의 대표점수들을 최대한 벌려놓아 그들 간의 분산이 가장 크도록 하는데 있다면 제1주성분점수는 대표점수로서의 의미를 지닌다고 하겠다. 그러나 이 방법은 상관계수가 큰 변수들의 집단에 너무 높은 가중값을 부여하게 된다는 점과 이를 악용할 소지가 있다는 점에서 각 개체의 대표점수로 이용하기에는 신중을 기할 필요가 있다.

가상예를 하나 들어 본다. 학생들의 능력을 평가하기 위해서 A,B,C 세 과목의 시험을 치루었다. 그런데 이들의 표본상관행렬이 다음과 같이 나타났다고 한다면 이들 과목들에 어떤 가중값을 부여하는 것이 상식적으로 타당할까?

$$\begin{pmatrix} 1 & \approx 1.0 & \approx 0.0 \\ \approx 1.0 & 1 & \approx 0.0 \\ \approx 0.0 & \approx 0.0 & 1 \end{pmatrix}$$

이런 자료에 대해서 제1주성분점수법을 적용하면 거의 1/2 : 1/2 : 0의 비율로 가중값을 부여하게 된다. 그러나 이 상관행렬을 잘 살펴보면 A, B 두 과목은 거의 완전 양의 상관관계가 있고, A와 C, B와 C는 거의 독립인 것으로 보아 학생들의 능력을 평가하는데 있어서 A와 B

두 과목은 거의 중복평가된 것으로 볼 수 있다. 따라서 이런 경우라면 A, B, C 세 과목에 1/4 : 1/4 : 1/2 정도의 비율로 가중값을 부여하는 것이 더 타당할 것으로 여겨진다.

2.3 제1인자점수법

개체들의 대표점수를 구하고자 할 때 제1주성분점수와 마찬가지로 하나의 인자만을 고려하는 단일 인자모형에 근거한 인자점수도 자주 사용된다. 여기서의 인자란 상호연관된 다변량 확률변수들간의 내부적 의존관계를 재현하고 있는, 변수들의 저변에 내재하는 가설적인 개념인 공통성분을 의미한다. 이와 같이 관찰할 수 없는 가설적인 인자는 하나의 확률변수로 생각할 수 있으며, 인자점수는 모형에 따른 확률적 인자에 대한 추측값의 의미를 갖는다. 이와 같은 인자점수를 얻기 위한 방법으로는 대표적으로 Thomson의 회귀적방법과 Bartlett의 가중회귀방법등이 있지만, Thomson의 방법에 의해 구해진 인자점수가 주로 사용되어지고 있다.

그러나 어떤 방법에 의해 구하여졌든 인자점수라는 것은 인자적재행렬의 비유일성에 의해 일반적으로 유일하게 존재하지 않으며, 2.2절에 주어진 표본상관행렬의 경우 제1주성분점수와 마찬가지로 인자점수법도 인자분석의 속성상 상관관계가 높은 변수들의 집단에 높은 가중값을 부여한다. 즉 인자점수도 세 과목에 각각 1/2 : 1/2 : 0 정도의 비율로 가중값을 준다. 만약 우리의 목적이 서로 상관된 점수들을 결합하여 전체적인 의미에서의 각 개체들의 대표값을 구하는데 있다면 이렇듯 한 쪽으로만 치우쳐진 가중값을 이용한다는 것은 불합리한 일일 것이다. 그렇다면 이런 경우 합당한 가중값을 부여하는 점수화 방법을 유도할 수는 없을까? 이에 대한 하나의 해답으로 '대표성분점수화법'이라는 새로운 대표점수화 방법을 제안하고자 한다.

3. 대표성분점수법

3.1 측정개념의 정의와 객관화의 중요성

전통적으로 사회과학이나 행동과학의 여러 학문분야 뿐만 아니라 일반사회에서도 측정과 관련하여 측정도구로서 설문지 또는 시험지를 많이 활용하고 있다. 그러나 이 측정방법은 일반연구자들이 생각하는 것 이상의 난해한 여러 문제점들을 내포하고 있다. 하지만 측정이론 또는 조사방법론에서는 설문지와 시험지의 작성법, 문항의 선정 및 평가등에 관련하여 주로 타당성과 신뢰성 등의 측면에 초점을 맞추고 있을 뿐, 서로 상관되어 있는 문항점수들을 결합하여 하나의 값으로 대표점수화하는 방법에 대해서는 연구의 큰 비중을 두고 있지 않다. 이는 아마 점수화라는 것이 연구자가 의도한 목표개념의 정의와 직접 관련된 문제이므로 이에 대한 일반적 연구가 이루어질 수 없다는 그 이유 때문이라고 생각된다. 그러나 사용된 점수화의 방법이 객관성을 지니지 못한다면 이 점수를 이용하여 얻어지는 여러 통계분석의 결과 또한 객관성을 지니지 못하게 됨으로서 결과적으로 그 연구결과는 독자들에게 받아들여질 수 없게 되는 것이다. 경우에 따라서는 독자들에게 그릇된 정보를 심어주기도 한다.

우선 하나의 예를 들어본다. 한 연구자가 '갑'과 '을' 두 회사 사원들의 '회사몰입도' 정도를 비교하기 위해 회사몰입도를 잴 수 있을 것으로 판단되는 아래의 다섯 문항으로 이루어진 설문지를 구성하였다.(각 문항은 5점 척도로 이루어져 있음)

(1) 나는 주위 사람들에게 우리회사는 정말 한번 일해 볼 만한 직장이라고 이야기하곤 한다.
(2) 나는 우리 회사의 발전을 위해 많은 노력을 기울인다.
(3) 나는 우리 회사에 대해 충성심을 느낀다.
(4) 나는 우리 회사의 경영이념과 나의 가치관이 매우 비슷하다고 생각한다.
(5) 나는 우리 회사의 장래에 대해 진정으로 걱정하고 있다.

각 문항들은 회사몰입도와 관련이 있는 변수들의 집합인 변수공간에서 연구자가 의도적으로 선정한 변수들로 볼 수 있는데, 이 다섯 개의 문항들을 가지고 '회사몰입도'라는 상위목표개념을 어떻게 객관적으로 정의할 수 있을까? 많은 성급한 연구자들이 이 문제에 관해서 깊은 사려도 없이 단순히 다섯 문항의 점수들의 합으로, 아니면 제1인자점수 또는 제1주성분점수로 회사몰입도를 정의한다. 그러나 상위목표개념의 정의를 어떻게 하느냐에 따라 분석결과가 완전히 달라질 수도 있는 사실에 유념할 필요가 있다.

이 설문지의 응답결과를 분석하는 단계에서 연구자는 회사몰입도의 정도를 다섯 문항의 점수합으로 정의하고(이 점수화방법은 통상의 연구에서 흔히 이용되고 있다는 사실에 유념할 필요가 있다), 이 점수를 이용하여 t-검정을 한 결과 "'갑'회사의 사원들보다 '을'회사의 사원들이 회사몰입도가 더 높다"라는 결론을 이끌어냈다. 한편 한 통계분석자가 회사몰입도의 정도를 이 다섯 문항의 제1인자점수로 하고(이 점수화방법도 통상의 연구에서 흔히 이용되고 있다), 마찬가지로 t-검정을 하여보았더니 위의 결론이 역전되었다. 이 두가지 서로 상반된 결론에 대해 연구자는 어떻게 설명할 것이며, 독자들은 위의 결론들을 어떻게 이해해야 할 것인가?

이 경우 '회사몰입도'라는 개념이 이미 다른 연구자들에 의해 사회적 또는 학문적 통념의 바탕하에서 합리적이고도 객관성 있게 조작되어 있어 연구자가 그 방식을 그대로 따른 상황이라면 아무런 문제가 없다. 그러나 그렇지 못한 경우라면 연구자는 아마 다음과 같이 주장할지도 모른다. "이는 순전히 '회사몰입도'라는 개념의 정의에 대한 시각의 차이에서 오는 것이다. 내가 염두해 둔 '회사몰입도'란 위의 다섯 문항의 점수합이지 제1인자점수는 아니다. 그러므로 나의 정의에 동의하지 않는다면 나의 연구결과를 받아들이지 않으면 된다." 이 답변은 그럴듯해 보이지만 만일 다수의 독자들이 연구자의 정의에 동의하지 않는다면, 많은 시간과 연구비를 들인 이 연구자의 연구결과는 일반성을 지니지 못하고 단지 자기의 정의에 동의하는 소수의 독자에게만 의미있는 결과로 전락하게 된다. 이 예를 통하여 우리는 측정개념을 정의함에 있어서 편의성보다는 객관성과 합리성의 유지가 얼마나 중요한가를 느끼게 될 것이다.

과학적 연구에서 일단 연구문제가 제기되고 나면, 문제제기에 관련되었던 개념(들)의 정의문제에 부딪히게 된다. 만일 개념이 명확히 정의되어 있지 않는다고 하면 위의 예와 같이 연구의 초점이 흐려져 많은 혼란을 초래하게 될 뿐만 아니라, 그 해석의 난맥은 연구의 전과정에 있어 많은 문제의 제기를 불가피하게 한다.

자연과학에서는 대부분의 경우 측정개념들이 명확히 정의되어 있어-예를 들어 길이, 무게, 온도 등과 같이-직접측정이 가능할 뿐만 아니라 측정도구로 인한 편의와 측정오차를 잘 통제하기만 하면 연구의 객관성유지에는 큰 어려움이 없다. 하지만, 사회과학이나 행동과학에서 다루어지는 대부분의 개념들은-예를 들어 인지도, 국방력, 사회성숙도 등과 같이-일반적으로 그 자체가 아주 추상적이거나 관념적이어서 합리적이고 객관적인 정의를 내리기가 힘들다.

사회과학이나 행동과학에서의 실제 연구과정을 통해 볼 때 정의작성은 결코 용의한 것은 아

니다. 구성적 정의나 조작적 정의를 막론하고 그 정의수립은 일시에 손쉽게 만들어지는 것은 아니다. 특히 조작적 정의를 수립할 때의 그 고충은 참으로 지대하다고 하지 않을 수 없다. 과학적 연구에 있어 조작적 정의는 이론, 가설, 관찰 또는 측정을 가교하는 필요불가결의 요인이다. 이는 경험적 관찰없는 과학적 연구가 있을 수 없는 것과 같이, 무엇을 어떻게 경험적으로 관찰하여 측정할 것인가를 보여주는 조작적 정의없는 과학적 연구 또한 있을 수 없기 때문이다.

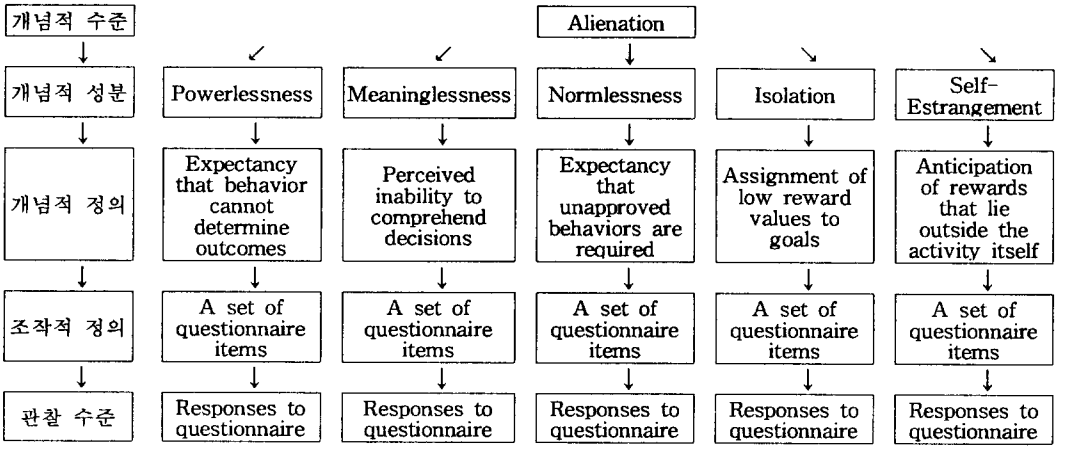
어떤 한 상위개념을 경험적으로 측정하려고 할 때는 이 상위개념이 뜻하는 경험적인 현상 또는 지표들을 기준으로 연구자 혹은 연구자그룹이 조작하지 않으면 안되는 것이다. 이렇게 조작을 하려고 할 때, 다행히 조작하려는 개념이 이미 다른 연구자들에 의하여 합리적이고도 객관성 있게 조작되어 있다고 하면 그대로 표준화된 조작적 정의를 따르면 되는 것이다. 그러나 그렇지 않고 생소한 개념을 새로 측정하기 위해 조작적 정의를 수립하고자 할 때는, 그들을 합리적으로 측정하기 위한 온갖 노력이 경주되지 않을 수 없다. 하지만 사회과학이나 행동과학도 '과학'이라는 인식을 얻어내기 위해서는 자연과학의 방법론을 원용하여 연구대상인 그 개념 자체에 대한 합리적이고도 객관적인 정의를 내릴 필요가 있다. 이를 위해서는 연구자 혹은 연구자그룹은 우선 사회적 또는 학문적 통념에 바탕을 둔 합의를 이끌어내어야만 하고 이를 구체화할 수 있어야 한다. 이는 앞에서 지적되었듯이 결코 쉬운 작업은 아니다.

하나의 상위목표개념을 간접적으로 측정할 때 그 바탕에 깔린 일반적인 절차를 간략히 정리하면 다음과 같다. 연구자가 측정하고자하는 상위목표개념을 일단 상정한 경우 우선 측정가능하면서 타당성있는 관련 하위개념들을 선정하는 작업일 것이다. 이 선정작업은 마치 상위개념과 관련이 있는 하위개념들로 이루어진 무한공간에서 의도적으로 소수의 하위개념을 표본추출하는 것과 흡사하다. 이때 연구자는 선정된 하위개념이 원래 측정하고자 했던 상위목표개념과 어떤 관련성을 지니고 있는지, 그리고 이 하위개념들을 가지고 어떻게 상위개념을 정의할 수 있는지를 미리 객관적으로 규정해 두어야 할 것이다. 다음으로 선정된 관련하위개념들을 측정할 수 있는 적절한 측정도구 또는 문항의 개발작업이 뒤따른다. 마지막으로 실제측정작업을 하고, 이의 결과인 측정값들을 정해진 규정에 따라 결합시켜 얻어진 값을 원래 측정하고자 했던 상위목표개념의 측정값으로 여긴다.

이 과정을 Nachmias(1981)에 있는 한 예를 통하여 살펴본다. 이 예에서는 'alienation' 이라는 매우 추상적이고 복합적인 개념이 어떻게 경험적으로 연구되는지를 보여주고 있다(<그림1> 참조). Melvin Seeman은 'alienation'이라는 단어가 사전에 "a sense of the splitting asunder of what was once held together, the breaking of the seamless mold in which values, behavior, and expectations were once cast into interlocking forms."라고 개념화되어 있다고 주장하면서, 이러한 사전적 개념화는 'alienation'이라는 상위개념에 대해 다섯가지의 의미(powerlessness, meaninglessness, normlessness, isolation, self-estrangement)를 부여하게 된다라고 주장하였다. 물론 이들 각 용어에 대한 개념적 정의가 요구됨은 당연하다. 이후 여러 연구에서 이 다섯가지의 개념들, 즉 alienation의 차원들, 은 각 차원에 대한 설문문항들을 구성함으로서 조작적으로 정의되었다. 여기에서 한 가지 주지하고 있어야 할 점은 개념적 성분의 조작적 정의에 사용된 문항들은 속성상 그 개념적 성분과 밀접한 상관이 있으며, 그 분야의 전문가들이라면 그 상관의 정도에 관한 나름대로의 느낌들을 모두 지니고 있다는 점이다.

문제는 관찰수준에서 얻어진 설문의 응답결과를 가지고 alienation의 각 개념적 성분들의 값

을 어떻게 구할 것이며, 더 나아가 이들 각 개념적 성분들의 값을 가지고 alienation의 값을 또한 어떻게 얻을 것인가 하는 점이다. 즉 점수화방법이 무엇이냐는 것이다.



<그림1> alienation의 경우 개념적 수준에서 관찰수준까지의 전의과정

위의 상황과 같은 경우 측정개념의 점수화와 관련하여 연구자가 취하게 되는 일반적 접근방법으로는 사전규정방법, 사후규정방법, 그리고 일부사전규정-일부사후보정의 방법의 세 가지가 있을 수 있다. 그러나 사전규정법이든 사후규정법이든 점수화의 규정은 조작화의 과정처럼 학문적 또는 사회적 통념을 지닌 전문가들의 의견에 바탕을 두어야 함은 당연하지만, 전문가들의 의견일치의 방법이 계량적 잣대에 의하지 않는 한 점수화라는 것이 구체적이고 객관적이지 못하여 비판을 받을 소지가 항상 있게 된다. 뿐만 아니라 그 연구결과의 해석상에도 많은 문제의 재기가 불가피하게 한다. 따라서 이를 구현할 수 있는 한 방법으로서 3.2절에서 '대표성분점수화법'이라는 것을 제안한다.

3.2 대표성분점수화법과 대표성분점수의 정의

우선 직접측정이 불가능한 단일 목표개념(ξ)을 조작화하여 직접측정이 가능한 관련변수들을 이용하여 간접적으로 측정한 자료의 전형적인 구조(<표1>)를 제시한다.

Non-Bayesian계열의 통계학에서 각 변수들의 모평균(μi)들을 미지의 모수로 인식하듯이 직접측정이 불가능한 목표개념(ξ)도 Guttman(1954)에 의하면 미지의 모변수로 볼 수 있다. 또한 미지의 모평균을 추정하기 위하여 모평균에 관한 정보를 지닌 개체들의 모집단(개체모집단)에서 표본개체들을 추출하여 측정값을 얻어 이들의 어떤 대표값을 이용하듯이, 우리가 알지 못하는 직접측정이 불가능한 하나의 목표개념(모변수)을 추정하고자 할 때 통상의 연구자들은 그 목표개념과 관련이 있는 측정가능한 변수들의 모집단(변수모집단)에서 표본변수들을 추출하여 추출된 각 변수의 측정값을 얻고 이들을 적절히 결합하여 얻어진 일변량화한 어떤 대표값을 이용하는 과정을 거친다.

<표1> 단일 목표개념 측정을 간접적으로 측정한 자료의 전형적인 구조

변수 / 개체	X_1	X_2	…	X_p
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{21}	…	x_{2p}
⋮	⋮	⋮	⋮	⋮
N	x_{N1}	x_{N2}	…	x_{Np}
↓ 추정				
모평균	μ_1	μ_2	…	μ_p

→ 추정

모변수(ξ)
ξ_1
ξ_2
⋮
ξ_N

여기서 주목할 점은 Guttman(1954)의 생각에 의하면 연구자가 선택한 변수들은 목표개념과 관련이 있는 변수들로만 이루어진 변수공간에서 의도적으로 뽑힌 표본변수로 볼 수 있다는 것이다. 따라서 이들은 속성상 서로 밀접한 양의 상관관계를 지니게 되며, 이들을 결합하여 연구자가 의도했던 목표개념의 측정값을 구하는 일은 일반연구에서 많이 요구되어 진다.

한편 개체추출과 변수추출 간의 차이점을 요약하면 다음과 같다.

첫째, 개체추출의 경우는 1934년 Neyman의 논문이 발표된 이후 거의 임의추출법에 의존하고 있지만, 변수추출의 경우는 상위개념에 관련이 있을 것으로 여겨지는 즉 타당성을 지닌 변수들만을 추출하는 의도적 추출과정(purposive selection process)을 거친다.

둘째, 미지의 모평균은 객관적으로 정의되지만, 상위개념은 속성상 일반적으로 객관적 정의가 이루어지기 힘들다. 자연과학에서의 개념들-예를 들어 길이, 온도, 질량등-과는 달리 사회과학에서 나오는 대부분의 개념들-예를 들면 인지도, 소비성향, 회사만족감등-은 매우 추상적인 것으로 사람마다 시간과 공간에 따라 그 느낌의 정도가 달라서 객관적으로 정의되기 어렵기 때문이다. 그러나 이에 대한 조작적 정의가 구체적으로 이루어지지 않은 경우 대부분의 연구자들은 자료분석의 과정에서 나름대로 이를 객관화하는 과정을 거친다.

셋째, N개의 표본개체가 추출되었다고 하더라도 실제 측정이 이루어지기 전에는 i번째 표본개체의 측정값이 모평균에 어느 정도 가깝게 나타날지를 전혀 알 수 없다. 이와는 달리 p개의 변수들이 추출된 경우에는 의도적 추출이라는 특성 때문에 실제 측정이 이루어지기 전에라도 전문가들은 j번째 변수의 측정값들이 상위목표개념과 상대적으로 얼마나 관련성이 있을 것인지에 관하여 어느 정도는 미리 감을 가질 수 있다.

개체모집단의 모평균을 제곱손실함수의 기대값을 최소화하는 값으로 정의하는 것처럼, 계량분석을 위해서는 상위개념 즉 모변수도 객관적으로 정의할 필요가 있다. 따라서 이제부터 상위목표개념의 정의(점수화)를 객관화 할 수 있는 새로운 방법을 소개하고자 한다.

Ω를 상위목표개념(ξ)과 관련이 있는 직접측정이 가능한 모든 변수들의 집합이라고 하자. 일반적으로, Ω는 무한집합의 성격을 지니면서 그 원소들 간에는 양의 상관관계를 지닌다. 한편 연구자를 포함한 전문가들이 "Ω의 각 원소들이 ξ와 가지는 상대적 관련성의 정도를 상관계수의 잣대로 보면 다음의 식(3.2.1)과 같을 것이다"라고 의견일치를 내릴 수 있다면, 그리고 ξ는 Ω의 원소들의 선형결합으로 구현될 수 있다는 가정을 받아들일 수 있다면, 이는 바로 상위목표개념을 고정화시키는 것 즉 객관적으로 정의하는 것이 된다.

$$corr(X_1, \xi) : corr(X_2, \xi) : \cdots : corr(X_t, \xi) = k_1 : k_2 : \cdots : k_t \quad (3.2.1)$$

여기서, $k_1, k_2, \cdots, k_t$는 상수들이고, t는 ∞일 수도 있다.

그러나 이는 현실적으로 불가능하다. 그 주된 이유는 아무리 전문가라 할지라도 상위목표개념과 관련이 있는 모든 변수들을 완벽하게 다 생각해 낼 수 없기 때문이다. 따라서 상위목표개념의 조작적 정의 과정에서 연구자는 중요한 즉 관련성이 높을 것으로 여겨지는 변수들만을 의도적으로 선정하게 된다.

이제, 연구자가 식(3.2.1)을 만족하는 ξ의 추정값(Y), 즉 조작화된 상위목표개념의 측정값, 을 구하기 위해서 Ω로부터 p개의 표본변수들 $X_1, X_2, \cdots, X_p$를 의도적으로 추출하였다고 하자. 그리고 위에서와 마찬가지로, 연구자를 포함한 전문가들이 "각 X_i들이 Y와 가지는 상대적 관련성의 정도를 상관계수의 잣대로 보면 다음의 식(3.2.2)과 같다"라고 의견일치를 내렸다고 하자

$$corr(X_1, Y) : corr(X_2, Y) : \cdots : corr(X_p, Y) = c_1 : c_2 : \cdots : c_p \quad (3.2.2)$$

단, $c_1, c_2, \cdots, c_p$는 상수들이다.

물론, 경우에 따라서는 전문가들에 따라 c_i값들의 크기에 대해 견해가 다를 수 있다는 점에서 이 과정도 쉽지는 않다. 그러나 c_i값들에 대한 의견불일치가 발생하는 경우 중도적 입장에서 그리고 의도적 추출이라는 속성을 고려한다면, 각 변수들이 같은 정도의 중요성을 갖고 상위개념과 관련되어 있다고 가정하는 것이 합당할 것이다. 따라서 이러한 경우에 식(3.2.2)는 다음의 식(3.2.3)으로 바뀌어 진다.

$$(corr(X_1, Y) \quad corr(X_1, Y) \quad \cdots \quad corr(X_1, Y))' \propto \mathbf{1}_p \text{ , 단 } \mathbf{1}_p = (1\ 1\ \cdots\ 1)' \quad (3.2.3)$$

이제 X_j들의 선형결합으로서 위의 식(3.2.3)을 만족하는 Y를 구해본다.

표준화된 확률벡터 $\underline{X} = (X_1\ X_2\ \cdots\ X_p)'$가 공분산행렬 $R = (r_{ij})$을 가진다고 하자. 이 때 상수벡터 $\underline{a} = (a_1\ a_2 \cdots\ a_p)'$에 대해 다음 조건을 만족하는 변수 X_j ($j = 1, 2, \cdots, p$)들의 선형결합인 Y를 고려한다.

$$Y = \underline{a}'\underline{X}, \qquad \text{단, } \quad \underline{a}'\mathbf{1}_p = 1. \quad (3.2.4)$$

식(3.2.3)의 가정이 합당하다고 할 때, $\underline{X}$와 Y의 표본공분산은 식(3.2.5)와 같이 나타난다.

$$cov(\underline{X}, \underline{X}'\underline{a}) = R\underline{a} = c\mathbf{1}_p, \quad \text{단 c는 상수이다.} \quad (3.2.5)$$

따라서 식(3.2.5)를 만족하는 $\underline{a}$와 Y는 각각 식(3.2.6), 식(3.2.7)과 같이 주어진다.

$$\underline{a} = c R^{-1}\mathbf{1}_p, \qquad \text{단, } c = 1 / \mathbf{1}_p' R^{-1}\mathbf{1}_p. \quad (3.2.6)$$

$$Y = \mathbf{1}_p' R^{-1}\underline{X} / \mathbf{1}_p' R^{-1}\mathbf{1}_p \quad (3.2.7)$$

본 연구에서는 식(3.2.7)과 같이 얻어진 Y를 서로 상관된 p개의 표본변수들 $X_1, X_2, \cdots, X_p$의 '대표성분'이라고 정의한다.

앞의 <표1>과 같은 자료구조, 즉 N개의 독립개체들 각각에 대해서 p개의 변수들에 관한 관찰값을 얻은 경우의 자료행렬의 각 관찰벡터 $x_i = (x_{i1}\ x_{i2} \cdots\ x_{ip})'$를 식(3.2.7)의 $\underline{X}$ 대신 대입하면 N개의 값들을 얻어진다. 이를 표현하면 다음과 같다.

$$y_i = \mathbf{1}_p' R^{-1} x_i \;/\; \mathbf{1}_p' R^{-1} \mathbf{1}_p , \qquad i = 1, 2, \cdots, N \tag{3.2.8}$$

이에 의해 얻어진 y_i를 i번째 개체에 대한 '대표성분점수', 그리고 이를 벡터로 나타낸 $y = (y_1\ y_2 \cdots y_N)'$을 '대표성분점수벡터'라고 정의한다.

3.3 기하적 의미

이광진(1992)에 의하면 대표성분점수벡터가 가지는 기하적 의미는 다음과 같다. 표준화된 자료행렬 $X_{N\times p} = (\ x_1\ \ x_2 \cdots x_p\ )$이 있고, 이의 j번째 열벡터 $x_j = (x_{1j}\ x_{2j} \cdots x_{Nj})'$를 j번째 변수의 관찰벡터라고 하자. 그리고 모든 변수의 관찰벡터들의 종점(end point)들을 지나는 초평면을 $L(X)$라고 표현하자. 그러면 초평면 $L(X)$상에는 원점으로부터의 유클리드 거리가 최소인 한 벡터가 존재한다. 이 벡터가 바로 3.2절에서 정의한 대표성분점수 벡터가 된다. 이를 증명해 보자.

$L(X)$상의 임의의 한 벡터 x^*는 다음과 같이 표현될 수 있다.

$$x^* = c_1 x_1 + c_2 x_2 + \cdots + c_p x_p = X\underline{c} \tag{3.3.1}$$

여기서 $\underline{c} = (c_1\ c_2 \cdots c_p)'$는 $\underline{c}'\mathbf{1}_p = 1$을 만족하는 상수벡터이다. 이때 벡터 x^*의 길이의 제곱은

$$\| x^* \|^2 = x^{*\prime} x^* = \underline{c}' X' X \underline{c} \tag{3.3.2}$$

가 되는데, 제약조건 $\underline{c}'\mathbf{1}_p = 1$ 하에서 이를 최소화하는 $\underline{c}$를 라그랑지 승수법을 이용하여 구하는 과정은 다음과 같다.

$$\begin{aligned} f(\underline{c}, \lambda) &= \underline{c}' X' X \underline{c} - \lambda(\underline{c}'\mathbf{1}_p - 1) \\ \partial f / \partial \underline{c} &= 2X'X\underline{c} - \lambda \mathbf{1}_p = \underline{0}_p \\ \partial f / \partial \lambda &= -\underline{c}'\mathbf{1}_p + 1 = 0 \end{aligned} \tag{3.3.3}$$

따라서 식(3.3.3)으로부터

$$\begin{aligned} \lambda &= 2 \;/\; \mathbf{1}_p'(X'X)^{-1}\mathbf{1}_p \\ \underline{c} &= (X'X)^{-1}\mathbf{1}_p / \mathbf{1}_p'(X'X)^{-1}\mathbf{1}_p \end{aligned} \tag{3.3.4}$$

이다. 그런데 자료행렬 X가 표준화된 자료행렬이므로, $\underline{c}$는 식(3.3.5)처럼 표현이 되며

$$\underline{c} = R^{-1}\mathbf{1}_p \;/\; \mathbf{1}_p' R^{-1}\mathbf{1}_p \tag{3.3.5}$$

벡터 x^*는 식(3.3.1)에 의해 식(3.3.6)가 같이 표현되며 이는 식(3.2.8)과 일치한다.

$$x^* = X\underline{c} = XR^{-1}\mathbf{1}_p \;/\; \mathbf{1}_p' R^{-1}\mathbf{1}_p \tag{3.3.6}$$

3.4 대표성분점수의 성질

3.4.1 산술평균점수법과의 관계

통계학의 추정문제에서 여러 추정방법들 중에서 가장 상식적이고 많이 사용되는 것이 일반화 최소제곱법(generalized least square method)이다. 이 방법을 간단히 묘사하면 다음과 같다.

$$\sum (\text{개별추정자료에 대한 가중값}) * \| \text{개별추정자료} - \text{추정대상후보} \|^2$$

을 최소화하는 추정대상후보를 추정대상의 추정값으로 사용하는 방법.

이 추정방법을 사용함에 있어서, 어떤 한 개별추정자료가 다른 개별추정자료들에 비해 추정대상에 더 가까울 것이라는 정보가 전혀 없는 상황에서는 모든 개별추정자료에 대해 동일한 가중값을 부여하는 것 또한 지극히 상식적이다.

이런 입장에서 보면, 일변량 독립표본의 경우 추정전에는 어떤 표본이 다른 표본보다 모평균에 더 가깝게 나타날 지에 관한 정보가 전혀 없는 상황이므로 개별표본들에 대한 가중값들을 모두 동일하게 둘 수 있으며, 이를 일반화된 최소제곱법에 적용하면 모평균의 추정값으로 표본평균(즉 산술평균)이 얻어진다. 따라서 일변량 독립표본에서 모평균의 추정값으로 표본평균을 사용하는 것은 확률적인 의미보다는 상식에 바탕을 둔 것이라 할 수 있다. 참고로 여기에서는 추정대상후보에 대해 '표본들의 선형결합'이어야 한다는 자격조건을 부여하지는 않았다.

한편 이 추정방법을 p개 표본변수들에 대한 N명의 측정값들을 통하여 N명의 모변수의 값들을 추정하려는 우리의 경우에 적용하여 보자. 우선 추출된 어떤 한 표본변수가 다른 표본변수보다 모변수에 더 가까운 지에 관해서 특별히 말할 수 없는 상황에서 가중값들을 모두 동일하게 둔다는 것은 위의 논리에 의해 받아들일 만하다. 그리고 개별추정자료는 각 표본변수들에 대한 N명의 관찰벡터(p개)가 되고. 추정대상은 N명의 모변수의 값들로 구성되는 벡터이다. 여기에서 추정대상후보에 대해 '표본변수들의 선형결합으로 가중값의 합이 1 이어야 한다'는 자격조건을 부여하여 일반화된 최소제곱법에 적용하면 3.3절의 기하적 의미로부터 쉽게 N명의 모변수의 추정값들로 구성되는 벡터가 바로 대표성분점수벡터가 된다라는 사실이 얻어진다. 이를 정리하면, 대표성분점수법이란 일변량 독립표본에서 모평균의 추정값으로 표본평균을 이용한 그 착상을 표본변수들의 일변량화에 그대로 적용한 것임을 알 수 있다.

그러면 일반적으로 많은 사람들이 추출된 변수들과 모변수와의 상관성의 정도에 관한 정보가 전혀 없는 상황에서 사용되어질 수 있는 점수화법으로 알고 있는 산술평균점수법이 실제로 일반화된 최소제곱법의 입장에서 어떤 의미를 파악해 볼 필요가 있다. 이를 위해 위에서 밝혔듯이 실제로 대표성분점수법이 추출된 변수들과 모변수와의 상관성의 정도에 관한 정보가 전혀 없는 상황을 반영한 점수화법에서의 가중값들의 벡터 $(R^{-1}\mathbf{1} \,/\, \mathbf{1}\,R^{-1}\mathbf{1})$와 산술평균점수법에서의 가중값들의 벡터 $(1/p\ \mathbf{1})$가 어떤 상황에서 같게 되는지를 살펴본다.

$$R^{-1}\mathbf{1} \propto \mathbf{1} \qquad (3.4.1.1)$$

이 식(3.4.1.1)이 의미하는 바는 R이 등예측상관구조(equipredictability correlation structure)를 지닌다는 점이다. 이는 어떤 한 변수가 나머지 변수들과 가지는 다중상관계수들이 모두 같다는, 즉 한 변수를 제외한 나머지 변수들이 제외된 그 한 변수를 예측할 수 있는 능력이 모두 동일하다는, 성질을 가지고 있는 것이다. 이런 의미에서 산술평균점수법은 추출된 표본변수들과

모변수와의 상관성의 정도에 관한 정보가 전혀없는 상황을 구현한 것은 아니라, 실제로는 선정된 표본변수들간의 상관성만을 반영하는 상관행렬이 등예측상관구조를 지닌다라는 사실을 반영한 점수화법이라고 할 수 있다. 또한 표본변수들의 상관행렬이 등예측상관구조를 가지는 경우 산술평균점수법과 대표성분점수법은 동일한 점수화를 제공하기 때문에 대표성분점수법의 의미가 더 한층 부각된다고 할 수 있다.

3.4.2 제1주성분점수와의 관계

제1주성분점수가 개체들의 전체변이를 최대화하려는데 목적을 둔 점수라면, 어느 한쪽으로도 치우치지 않고 전체적으로 대표할 수 있는 하나의 값을 구하고자 할 때 대표성분점수는 그 의미가 있다고 할 수 있겠다. 이처럼 두 점수는 개체들의 다른 측면을 묘사하여 자료를 축약하고 있다고 할 수 있으며 따라서 일반적으로 다른 결과를 준다. 그러나 다음과 같이 표본상관행렬 R이 특별한 패턴을 가지는 경우 제1주성분점수와 대표성분점수는 동일한 값을 가지게 된다.

경우1 등상관구조(equicorrelation pattern)의 경우

p개의 변수 $X_1, X_2, \cdots, X_p$가 있고 이 변수들의 표본상관행렬 $R=(r_{ij})$이 다음과 같이 등상관구조, 즉 $r_{ij}=r,\ \forall\ i\neq j$ 를 만족한다고 가정하자.

$$R = (r_{ij}) = \begin{pmatrix} 1 & r & \cdots & r \\ r & 1 & \cdots & r \\ \cdot & \cdot & \ddots & \cdot \\ r & r & \cdots & 1 \end{pmatrix} \tag{3.4.2.1}$$

이 때 행렬 R의 첫번째 고유값은 $\lambda_1=1+(p-1)r$ 이고, 이에 대응하는 $p\times 1$ 고유벡터는 $\underline{e_1}=(1/\sqrt{p}\ \ 1/\sqrt{p}\ \cdots 1/\sqrt{p})'$이다. 따라서 i번째 관찰벡터 $\underline{x_i}=(x_{i1}\ x_{i2}\ \cdots\ x_{ip})'$에 대한 제1주성분점수는

$$pc1_i = \underline{e_1}'\underline{x_i} = 1/\sqrt{p}\ \sum_{j=1}^{p} x_{ij}\ ,\quad i=1,2,\cdots,N \tag{3.4.2.2}$$

이다. 이 경우 제1주성분점수는 p개의 변수 각각에 똑같은 가중값을 주고 있다.

한편 식(3.4.2.1)의 표본상관행렬 R은 다음과 같이 표현할 수 있다.

$$R = (1-r)\ I + r\ \mathbf{1}\ \mathbf{1}' \tag{3.4.2.3}$$

따라서,

$$R^{-1} = 1/(1-r)\ [I - \{r/(1+(p-1)r)\}\mathbf{1}\ \mathbf{1}'\ ] \tag{3.4.2.4}$$

로 부터

$$R^{-1}\mathbf{1} = 1/(1+(p-1)r)\ \mathbf{1} \tag{3.4.2.5}$$

이고, 대표성분점수 y_i는 식(3.2.8)에 의하면 다음과 같다

$$y_i = \mathbf{1}'R^{-1}\underline{x_i}\ /\ \mathbf{1}'R^{-1}\mathbf{1} = 1/p\ \sum_{j=1}^{p} x_{ij}\ ,\quad i=1,2,\cdots,N \tag{3.4.2.6}$$

이처럼 대표성분점수도 제1주성분점수와 마찬가지로 각각의 변수에 같은 가중값을 준다. 즉 등상관 구조를 갖고 있는 변수들의 경우에 제1주성분점수나 대표성분점수는 같은 의미를 갖는다. 이는 변수들 간에 똑같은 상관관계가 있을 경우 대표성분이라는 관점에서 각 변수들이 동일한 수준의 공헌도를 가짐을 나타낸다.

경우2 등상관예측구조(equipredictability correlation pattern)의 경우

Morrison(1990)에서 언급하고 있는 다음과 같은 패턴의 상관행렬을 고려해 보자.

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & r_{14} \\ r_{12} & 1 & r_{14} & r_{13} \\ r_{13} & r_{14} & 1 & r_{12} \\ r_{14} & r_{13} & r_{12} & 1 \end{pmatrix} \qquad (3.4.2.7)$$

식(3.4.2.7)과 같은패턴의 행렬은 등예측상관구조행렬(equipredictability correration pattern matrix)로 알려져 있는데(Bargmann(1957)), 이는 어떤 한 변수가 나머지 변수들과 가지는 다중상관계수(mutiple correlration coefficient)들이 모두 같다는 성질을 가지고 있다

식(3.4.2.7)과 같은 행렬의 첫 고유값에 해당하는 고유벡터인 $\underline{e_1}$은 (1/2 1/2 1/2 1/2)′이므로 제1주성분점수는 각 변수에 동일한 가중값을 부여하여 $\sum x_j$에 비례하게 됨을 알 수 있다. 대표성분점수도 식(3.4.2.8)에서 보듯이 변수들의 합에 비례함을 알 수 있다. 즉, 식(3.4.2.7)과 같은 상관구조에서는 제1주성분점수와 대표성분점수는 같은 의미를 갖는다.

$$\begin{aligned} y &= 1'R^{-1}x \,/\, 1'R^{-1}1 \\ &= 1/4 \sum_{j=1}^{4} x_j \qquad \text{단,} \quad x=(x_1\ x_2\ x_3\ x_4)' \end{aligned} \qquad (3.4.2.8)$$

이 처럼 j번째 변수 X_j와 이를 제외한 나머지 변수들간의 다중상관계수가 모두 동일하다는 의미는 대표성분점수의 의미에서 보면 각 변수들이 동일한 수준의 공헌을 함을 뜻한다.

3.4.3 제1인자점수와의 관계

Spearman의 단일인자모형(one-factor-model)에 근거한 인자점수를 생각한다. i번째 개체의 관찰벡터 $\underline{x_i}=(x_{i1}\ x_{i2}\ \cdots\ x_{ip})'$에 대하여 Thomson(1951)에 의해 제안된 회귀적 방법에 의한 i번째 인자점수 f_i는 $f_i=\hat{\lambda}'R^{-1}x_i$와 같이 계산할 수 있다. 이 때 $\hat{\lambda}$는 인자적재벡타의 추정벡터로서 $\hat{\lambda} = \hat{\lambda}1$이라고 가정하자. 이는 p개의 변수들 각각에 대해 동일한 인자적재 $\hat{\lambda}$를 부여한 것으로, 곧 $corr(X_j, F)=\lambda_j=\lambda$, $j=1,2,\cdots,p$ 를 의미한다. 따라서 f_i는 아래와 같이 쓸 수 있다.

$$f_i = \hat{\lambda}1'R^{-1}x_i \qquad (3.4.3.1)$$

이 결과는 식(3.2.8)에서의 대표성분점수와 동일한 의미를 가진다. 따라서 단일인자모형에서 각 변수들에 똑같은 인자적재를 줄 경우의 인자점수는 대표성분점수와 같은 결과를 가짐을 알 수 있다. 이는 단일 공통인자 F를 앞에서 언급한 상위목표개념(ξ)의 추정변수로 간주한다면 합당한 결과라고 할 수 있겠다.

또한 제1주성분점수와 마찬가지로 인자점수도 등상관구조와 등예측상관구조를 가지는 변수들의 경우 각 변수들에 같은 가중치를 줌으로써 대표성분점수와 동일한 의미를 가진다.

4. 사례분석

다음의 <표2>와 <표3>은 XX초등학교 6학년 X반의 41명 학생들의 사회, 산수, 자연, 체육 4개 과목에 대한 1학기 성적자료와 이들 과목에 대한 상관행렬이다. 여기서 우리는 이 네 과목의 점수를 적절히 결합하여 각 학생의 대표점수를 얻고 싶다고 하자. 흔히 생각할 수 있는 것은 산술평균이다. 그러나 이것은 각 과목간의 상관관계를 전혀 고려하지 않는다는 단점이 있다. 따라서 각 과목에 적절한 가중값을 부여하는 다른 대표점수를 생각할 필요가 있다. 여기서는 자주 사용되고 있는 제1주성분점수, 단일인자모형 하에서의 제1인자점수, 그리고 본 논문에서 제안하고 있는 대표성분점수를 서로 비교해 보도록 하겠다.

우선 사회, 산수, 자연, 체육 네 과목 점수의 표본상관행렬은 다음과 같다.

위의 <표3>을 보면 사회, 산수, 자연 세 과목은 서로 상관관계가 꽤 높은 편이나 이 세 과목과 체육은 상관관계가 비교적 낮음을 알 수 있다. 따라서 사회, 산수, 자연 점수는 학생의 여러가지 능력중에 하나의 거의 동일한 능력 즉 수리·탐구에 관련된 지적 능력을 재고 있다고 생각할 수 있으며, 체육 점수는 학생의 또 다른 능력인 신체적 능력을 나타낸다고 할 수 있을 것이다. 그러므로 우리가 어떤 학생의 능력을 평가하고자 할 때 특별한 하나의 능력에 치우치지 않고, 그 학생이 가지고 있는 모든 능력을 골고루 나타낼 수 있는 대표값을 얻고자 한다면, 위와 같은 경우에 네 과목 각각에 1/6, 1/6, 1/6, 1/2에 가까운 가중값을 주는 것은 합당한 일일지도 모른다.

<표2> XX초등학교 학생들의 성적자료

번호	사회	산수	자연	체육	번호	사회	산수	자연	체육
1	77	81	82	89	22	74	68	79	83
2	84	91	88	77	23	80	79	77	71
3	91	80	89	85	25	92	74	92	85
4	93	86	90	92	26	65	63	79	72
5	78	55	78	84	27	64	42	70	68
6	66	69	79	88	28	75	66	79	82
7	92	87	93	90	29	69	72	74	92
8	78	53	68	84	30	87	86	93	80
9	75	72	74	82	31	87	85	92	93
10	95	79	86	91	32	91	93	73	79
11	87	60	80	91	33	98	93	93	86
12	89	83	89	83	34	60	52	57	75
13	64	70	74	79	35	92	92	94	70
14	74	74	80	93	36	83	74	75	86
15	89	77	83	87	37	79	84	87	93
16	60	45	66	89	38	87	58	87	81
17	60	61	81	78	39	84	90	88	91
18	77	64	76	83	40	93	88	98	97
19	54	54	62	90	41	60	38	63	90
20	42	50	48	81	42	78	80	82	91
21	96	82	94	73					

<표3> 표본상관행렬

	사회	산수	자연	체육
사회	1	0.735	0.846	0.146
산수	0.735	1	0.770	0.150
자연	0.846	0.770	1	0.157
체육	0.146	0.150	0.157	1

<표2>의 자료에 대하여 제1주성분점수와 제1인자점수에서의 각 변수별 가중계수와 본 연구에서 제안하고 있는 대표성분점수에서의 가중값들을 비교해 보면 다음 <표4>와 같다.

<표4> 과목별 가중계수

	제1주성분	제1인자	대표성분
사회	0.575	0.909	0.190
산수	0.556	0.830	0.224
자연	0.579	0.930	0.132
체육	0.159	0.170	0.454

<표4>에 의하면 각 학생들의 제1주성분점수, 제1인자점수, 대표성분점수는 다음과 같은 식들에 의해 구해진다.

제1주성분점수 = 0.575*사회점수 + 0.556*산수점수 + 0.579*자연점수 + 0.159*체육점수
제1인자점수 = 0.909*사회점수 + 0.830*산수점수 + 0.930*자연점수 + 0.170*체육점수
대표성분점수 = 0.190*사회점수 + 0.224*산수점수 + 0.132*자연점수 + 0.454*체육점수

위의 제1주성분점수에서의 각 변수별 가중계수들은 <표3>으로 주어진 표본상관행렬의 첫 고유값 2.6226에 해당하는 고유벡터의 원소들이며, 제1인자점수에서 네 과목의 가중값으로 사용된 계수값들은 Thomson에 의해 제안된 회귀적 방법에 의하여 계산한 값들이다. 이 때 인자적재들의 추정값을 구하기 위하여 주축인자법을 사용하였다. 그리고 대표성분점수에서의 가중값들은 앞의 3.2절에 있는 식(3.2.6)을 만족하는 벡터 $\underline{\alpha}$의 원소들이다. <표4>를 보면 제1주성분점수는 상관계수가 큰 사회, 산수, 자연 과목에 너무 높은 가중치를 주고 있음을 볼 수 있다. 제1인자점수의 경우도 사회와 자연 과목에 높은 가중값을 주고 있으며 산수의 경우 제1주성분점수에 비해 상대적으로 낮은 가중값을 부여하고 있음은 제1주성분점수와 마찬가지이다. 만일 제1주성분점수법과 제1인자법을 사용해야 한다면 이 경우에는 두 개의 주성분과 인자 - 즉 지적 능력을 나타내고 있는 제1주성분(인자)과 신체적 능력을 나타내는 제2주성분(인자)- 를 고려해야 합당할 것이다. 그러나 지적 능력과 신체적 능력을 포괄하면서도 한 쪽 능력에 치우침 없이 각 학생의 능력을 하나의 대표점수로 나타내고자 한다면, 네 과목 각각에 1/6, 1/6, 1/6, 1/2에 가까운 가중값을 주고 있는 대표성분점수가 더 합당한 의미를 가지는 것으로 생각할 수 있다.

다음 <표5>는 각 학생들에 대한 산술평균, 대표성분점수, 제1주성분점수, 제1인자점수이다. 단 여기서 제1주성분점수와 제1인자점수는 표준화된 점수이며, 이 점수들과의 비교를 위해 표준화된 대표성분점수(대표성분점수*)도 구해 보았다. <표5>에 의하면 19번 학생의 경우 산술평

균은 65점이고 대표성분점수는 산술평균보다 높은 71점이다. <표2>의 자료를 볼 때 19번 학생의 체육점수는 91점으로 반학생들의 평균점수 84점보다 매우 높은 반면 다른 세과목의 점수는 각각 54점, 54점, 62점으로 반 평균 78점, 71점, 80점에 미치지 못함을 알 수 있다. 따라서 학생들의 지적 능력 못지 않게 신체적 능력 또한 그 학생의 중요한 능력으로 간주하는 취지에서의 대표성분점수는 앞의 <표3>과 같은 상관관계가 존재함을 간과하고 각 과목에 똑같은 가중값을 부여하여 결국은 지적 능력을 재는 사회, 산수, 자연 과목에 높은 비중을 두고 신체적 능력을 나타내는 체육 과목에 낮은 비중을 두는 산술평균보다 높은 값을 주게 된다. 반대로 2번 학생처럼 사회, 산수, 자연 과목의 점수는 반 평균점수보다 높고 체육점수는 반 평균점수보다 낮은 경우 대표성분점수는 산술평균보다 낮은 값을 나타낸다. 이는 제1주성분점수와 제1인자점수, 그리고 대표성분점수*의 비교에서도 그대로 적용될 수 있다. 즉 19번 학생의 경우 제1주성분점수 -2.557과 제1인자점수 -1.670에 비해 대표성분점수 -0.473으로 높고, 2번 학생의 경우는 제1주성분점수가 1.186, 제1인자점수가 0.708로 대표성분점수 -0.002보다 높은 값을 보인다.

5. 결론

지금까지 서로 상관된 변수들을 결합하여 하나의 대표점수를 구하는 방법에 관하여 살펴보았다. 전통적으로 많이 사용되어지고 있는 산술평균점수법은 선정된 변수들간의 상관성만을 반영하는 상관행렬이 등예측상관구조를 지니는 특수한 경우에 사용되어질 수 있는 점수화법이라는 사실과 이 경우 본 연구에서 새로이 제안하는 대표성분점수법과 같은 의미를 지니고 있음을 보았다. 또한 제1주성분점수법과 제1인자점수법은 상관관계가 큰 변수들의 집단에 너무 높은 가중치를 부여한다는 단점을 지니고 있음을 보았고, 이런 경우 대표성분점수법이 타당한 결과를 나타냄을 알 수 있었다. 물론 모든 면에서 최상인 방법은 있을 수 없다. 그러나 차원축소를 위하여 소수 몇개의 주성분들이나 인자들로 변수간에 내재하는 구조를 재현하고자 함이 아니고, 단지 선정된 여러 변수들에 치우침 없이 전체적으로 대표할 수 있는 하나의 값을 찾고자 할 때 대표성분점수화법은 의미를 가진다.

따라서 본 논문의 결과는 사회과학 및 행동과학의 많은 계량적 연구에서 뿐만 아니라 사회, 경제, 경영의 일반사회 분야에서 요구되는 지수개발에 있어서도 일조할 수 있으리라 여겨진다. 그러나 대표성분점수의 유도과정에서 일단의 변수들을 의도적으로 추출할 경우 상위목표개념의 측정에 심각한 왜곡이나 한쪽으로 치우쳐진 변수들이 섞여 있을 수도 있음을 무시하고 있는데, 이처럼 상위개념을 흐리게 하지 않는 변수들의 선정문제에 관한 또 다른 연구가 요구되어 진다고 하겠다.

<표5> 각 학생들에 대한 산술평균, 대표성분점수, 제1주성분점수, 제1인자점수

번호	산술 평균	대표성 분점수	제1주성 분점수	제1인자 점수	대표성분 점수*	번호	산술 평균	대표성 분점수	제1주성 분점수	제1인자 점수	대표성분 점수*
1	82.75	84.01	0.464	0.155	0.434	22	76.00	77.41	-0.437	-0.231	-0.217
2	58.00	82.92	1.186	0.708	-0.002	23	76.75	75.29	-0.140	-0.037	-0.749
3	86.25	85.55	1.312	0.818	0.450	24	85.75	84.79	1.290	0.899	0.411
4	90.25	90.58	1.828	1.005	1.022	25	69.75	69.59	-1.255	-0.560	-1.114
5	73.75	75.58	-0.775	-0.327	-0.303	26	61.00	61.69	-2.636	-1.247	-1.802
6	75.50	78.39	-0.635	-0.427	-0.001	27	75.50	76.70	-0.490	-0.231	-0.296
7	90.50	90.10	1.935	1.115	0.932	28	76.75	80.79	-0.5690	-0.515	0.279
8	70.75	73.80	-1.374	-0.782	-0.452	29	86.50	84.39	1.460	0.947	0.215
9	75.75	77.38	-0.530	-0.370	-0.266	30	89.25	89.94	1.658	0.916	1.008
10	87.75	88.41	1.423	0.797	0.835	31	84.00	83.61	0.821	0.283	0.073
11	79.50	81.44	0.059	0.081	0.366	32	92.50	90.77	2.328	1.347	0.855
12	86.00	84.93	1.292	0.798	0.340	33	61.00	64.63	-2.966	-1.776	-1.424
13	71.75	73.48	-1.146	-0.701	-0.642	34	87.00	82.27	1.730	1.184	-0.244
14	80.25	83.43	0.059	-0.093	0.515	35	79.50	81.29	0.031	-0.076	0.142
15	84.00	84.61	0.844	0.471	0.431	36	85.75	87.54	1.013	0.470	0.820
16	65.00	70.61	-2.443	-1.451	-0.537	37	78.25	77.78	0.131	0.338	-0.211
17	70.00	71.18	-1.307	-0.625	-0.813	38	88.25	89.05	1.459	0.722	0.866
18	75.00	76.68	-0.612	-0.329	-0.270	39	94.00	94.36	2.432	1.383	1.463
19	65.00	71.42	-2.557	-1.670	-0.473	40	62.75	69.10	-2.838	-1.666	-0.614
20	55.25	62.31	-4.159	-2.668	-1.440	41	82.75	84.89	0.515	0.174	0.560
21	86.25	82.15	1.599	1.172	-0.146						

참고문헌

[1] 고홍화, 김현수, 백영승(1988). 「사회 · 행동과학 연구방법의 기초」, 성원사, 서울.

[2] 김기영, 전명식(1989). 「SAS 주성분분석」, 자유아카데미, 서울.

[3] 김기영, 전명식(1990). 「SAS 인자분석」, 자유아카데미, 서울.

[4] 김해동(1988). 「조사방법론 -이론과 기법」, 법문사, 서울.

[5] 이관우(1983). 「신조사방법론」, 형설출판사, 서울.

[6] 이광진(1992). 상관구조의 기하적 고찰과 고유변환에 관한 연구. 박사학위논문, 고려대학교.

[7] 허명회(1991). 설문지 시험지 문항의 신뢰성 분석, 「응용통계연구」, 제4권 1호, 93-105.

[8] Bargmann, R. E.(1957). A study of independence and dependence in multivariate normal analysis, University of north Carolina Institute of Statistics Mimeographed Series, No 186, Chapel Hill.

[9] Guttman, L. A. (1954). A new approach to factor analysis: The radix. In P.F.Lazarsfeld (Ed.): Mathematical thinking in the social sciences. New York: Columbia University Press.

[10] Mardia, K, V. and Kent, J. T. and Bibby, J. M. (1979). Multivariate Analysis, Academic Press.

[11] Morrison, D. F. (1990). Multivariate Statistical Method, McGraw Hill.

[12] Nachmias. D. & Nachmias. C. (1981). Research Methods in the Social Sciences. 2 ed., St.Martin's Press, Inc.

Representative Component Scoring System and Its Validity and Applicability[1)]

Kwang Jin Lee[2)]

Abstract

In the case that an abstract concept was measured indirectly by using its indicators, many researcher have obtained its score by using the simple mean, the first principal component, or the first factor, etc. In this paper, an scoring method named as the representative component scoring system was suggested as an alternative and its validity and applicability were studied.

1) This research was supported by the Korean Research Foundation, 1994.

2) Department of Applied Statistics, Mokwon University, 24 Mokdong, Chung-gu, Taejon, 301-729, Korea