개념

- 범주형 변수 범주의 유사성 표현
- 교차표(분할표)로 나타내어지는 자료의 행과 열 범주를 저차원 공간상(2차원)의 좌표로 표현하여 관계를 탐구하려는 탐색적 자료 분석 기법

기원

- 대응분석의 수리적인 기원은 1930년대 Hirshfeld의 논문『상관관계와 분할표의 연관성』
- 대응분석의 기하적인 면은 1960년대 프랑스에서 Jean-Paul Benzecri에 의해서 발전되었다.
- 일본: 1950년대 Chikio Hayashi에 의해서 수량화 제3방법으로 개발되어 발전
- 프랑스: 1960년대 Jean-Paul Benzecri가 이끄는 자료분석 모임이 다양한 분야로부터 수집된 자료를 분석하는데 대응분석 기법을 응용하고 발전

$$\text{검정통계량} = \frac{\sum_{i=1}^{k}(O_i - E_i)^2}{E_i} \sim \chi^2(k-1)$$

RXC 분할표

X \ Y	1	2	…	C	Total
1	π_{11}	π_{12}	…	π_{1c}	π_{1+}
2	π_{21}	π_{22}	…	π_{2c}	π_{2+}
…	…	…	…	…	…
R	π_{r1}	π_{r2}	…	π_{rc}	π_{r+}
Total	π_{+1}	π_{+2}	…	π_{+c}	π_{++}

- π_{ij}: (X, Y) 결합밀도함수
- π_{i+}: (X) 주변밀도함수
- π_{+j}: (Y) 주변밀도함수

Homogeneity (동질성)

$$H_0 : \pi_{ij} = \pi_{kj} \text{ for } j = 1,2,\ldots,c \text{ and } k \neq i$$

- 각 행에 대해 열의 분포가 동일한가?

Independence (독립성)

- (X, Y)는 서로 독립인가?

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$$

결과 해석

- χ^2검정 결과 p-value(유의확률) 0.05보다 적으면 두 변수 상관관계 존재
- 관계 해석: 행 퍼센트 혹은 열 퍼센트에 의한 차이 해석
- R×C 셀이 많아지면 퍼센트에 의한 해석이 복잡해지고 신뢰도가 떨어짐
- 행 범주, 열 범주의 유사성 정도를 표현하지 못함

- Notation
  - n_{ij}: (i, j) 셀 관측빈도
  - n_{i+}: (i)번째 행의 관측빈도 합
  - n_{+j}: (j)번째 열의 관측빈도 합

- 방법
  - (i, j) 셀의 빈도 n_{ij} (≥ 0)의 i번째 행 (n_{i1}, …, n_{ic})은 총빈도가 $n_{i+} = n_{i1} + \ldots + n_{ic}$ 이고 C개 범주를 갖는 다항 분포
  - Multinomial 분포의 대응 확률은 상대빈도 $f_{ij} = n_{ij} / n$: 이것을 행 프로파일 (row profile)이라 정의
  - 각 행의 상대적 빈도 $f_{i+} = f_{i1} + \ldots + f_{ic}$ 를 선형계수로 (주성분 분석과 유사) 하여 좌표 계산
  - $r_i = (f_{i1} / f_{i+}, f_{i2} / f_{i+}, \ldots, f_{ic} / f_{i+})$는 C 차원 가중 Euclid 공간의 좌표
  - 가중(weighted) Euclid 공간이란 두 개의 좌표 r_a, r_b 사이의 거리가 다음과 같이 정의
    - $f_{+j} = f_{1j} + \ldots + f_{rj}$
  - 같은 방식으로 열 프
  - 행, 열 프로파일을 각각 2차원 공간에 표현하거나 동시에 표현

$$d(r_a, r_b) = \sqrt{\sum_j \left(\frac{f_{aj}}{f_{a+}} - \frac{f_{bj}}{f_{b+}}\right)^2 / f_{+j}}$$

X \ Y	1	2	…	C	Total
1	n_{11}	n_{12}	…	n_{1c}	n_{1+}
2	n_{21}	n_{22}	…	n_{2c}	n_{2+}
…	…	…	…	…	…
R	n_{r1}	n_{r2}	…	n_{rc}	n_{r+}
Total	n_{+1}	n_{+2}	…	n_{+c}	n_{++}

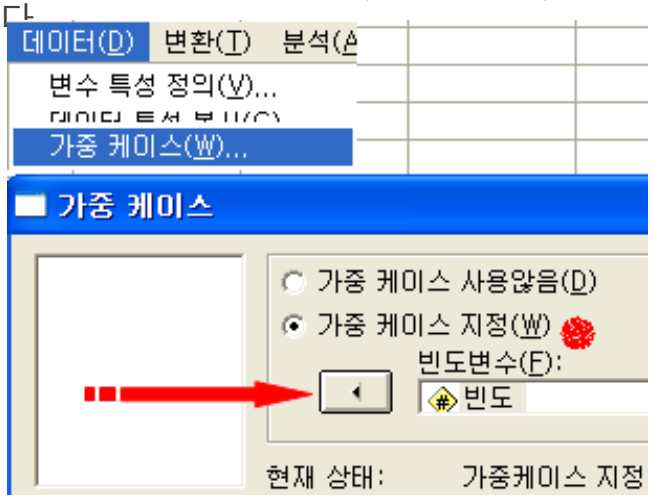
- 예제 데이터
  - 학과별 영어등급 조사한 데이터이다.

영어 등급 \ 학과	경영	경제	통계
A	78	65	68
B	22	8	30
C	20	2	7

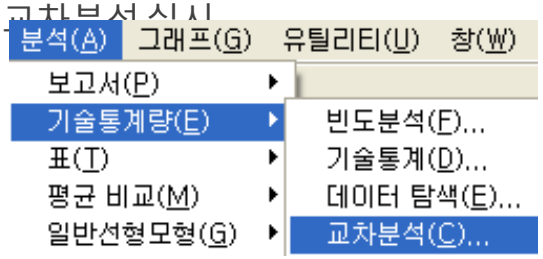
	전공	영어등급	빈도
1	경영	A	78
2	경영	B	22
3	경영	C	20
4	경제	A	65
5	경제	B	8
6	경제	C	2
7	통계	A	68
8	통계	B	30
9	통계	C	7

http://wolfpack.hannam.ac.kr
한남대학교 통계학과 권세혁 교수

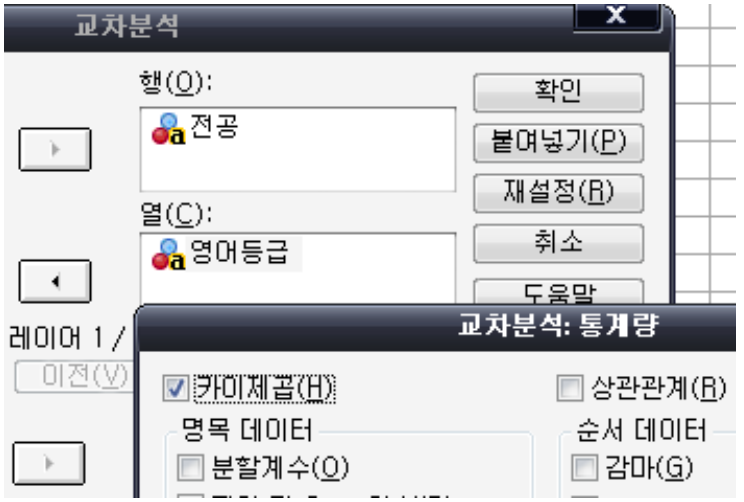
- 우선 전공별 학점의 차이가 있는지 알아보자.
  - 교차분석
    - 우선 빈도를 가중케이스(관측치 반복)로 만들어야 한다.



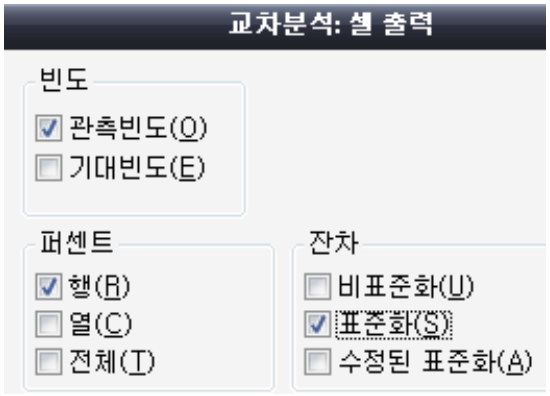
    - 교차분석 실시



- 행 변수에 영향(설명변수)을 주는 변수 설정
- 검정통계량 chi-square 값 출력, 행과 열 모두 순서형이면 상관계수 보는 것이 필요.



- 결과 해석 위하여
  - 행 퍼센트
  - 표준화 잔차 필요



교차분석 결과

전공 * 영어등급 교차표

			영어등급			전체
			A	B	C	
전공	경영	빈도	78	22	20	120
		전공의 %	65.0%	18.3%	16.7%	100.0%
		표준화 잔차	-.7	-.4	2.5	
	경제	빈도	65	8	2	75
		전공의 %	86.7%	10.7%	2.7%	100.0%
		표준화 잔차	1.7	-1.8	-1.9	
	통계	빈도	68	30	7	105
		전공의 %	64.8%	28.6%	6.7%	100.0%
		표준화 잔차	-.7	2.0	-1.0	
전체		빈도	211	60	29	300
		전공의 %	70.3%	20.0%	9.7%	100.0%

카이제곱 검정

	값	자유도	점근 유의확률 (양측검정)
Pearson 카이제곱	21.946[a]	4	.000
우도비	22.571	4	.000
유효 케이스 수	300		

a. 0 셀 (.0%)은(는) 5보다 작은 기대 빈도를 가지는 셀입니다. 최소 기대빈도는 7.25입니다.

해석

검정통계량 활용

- 유의확률이 <0.001이므로 귀무가설 기각, 전공별 영어성적의 차이는 있다.

행 퍼센트 이용

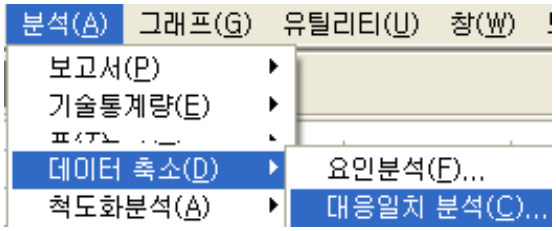
- A학점의 비율이 다른 학과에 비해 높은 경제학과 학생의 영어 성적이 높다고 할 수 있다.
- C학점의 비율 면에서 경영학과가 다른 학과에 비해 높으므로 영어 성적이 낮은 학과이다.

표준화 잔차

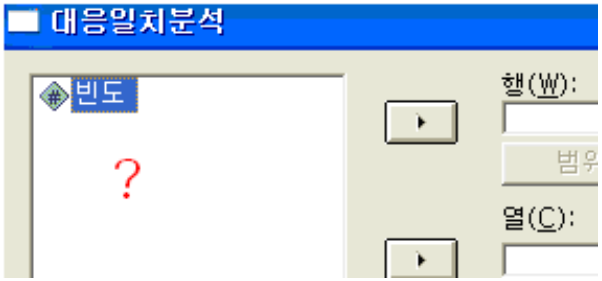
- A 등급: 경제전공 비율 높다.
- B등급: 통계 전공 높고 경제전공 낮다
- C등급: 경영전공 높고 경제전공 낮다.

그러므로 경제>통계>경영순이다.

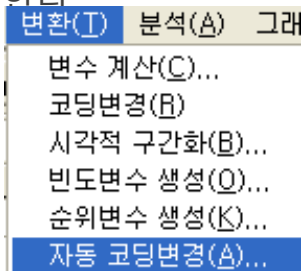
- 대응분석 실시

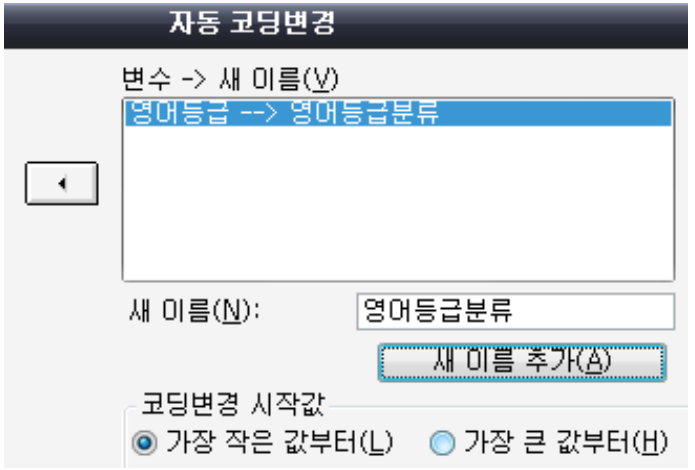


  - 변수명이 나타나지 않는다.



  - 대응분석을 하려면 행과 열 변수도 숫자로 입력되어 있어야 한다. 그러므로 열과 행 변수를 숫자형 변수로 변환하여야 한다.

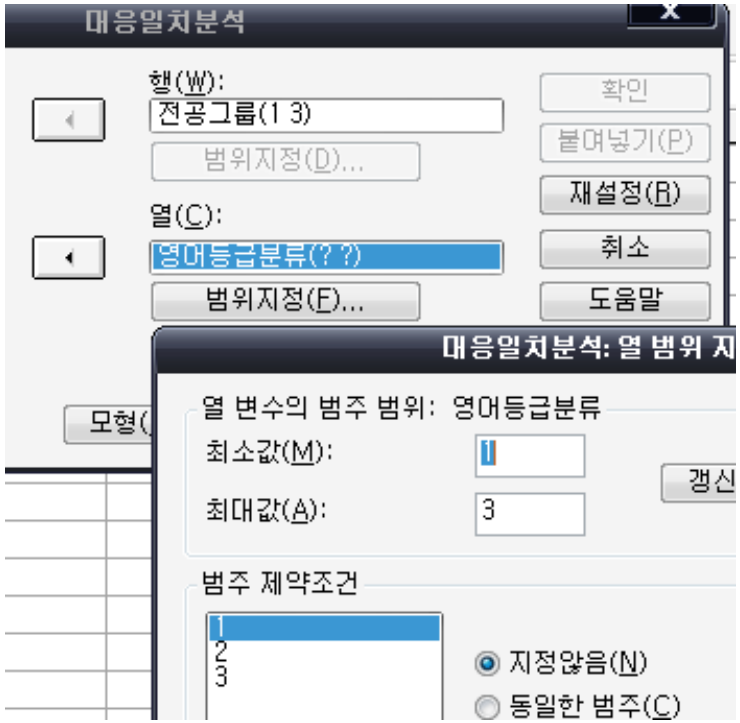




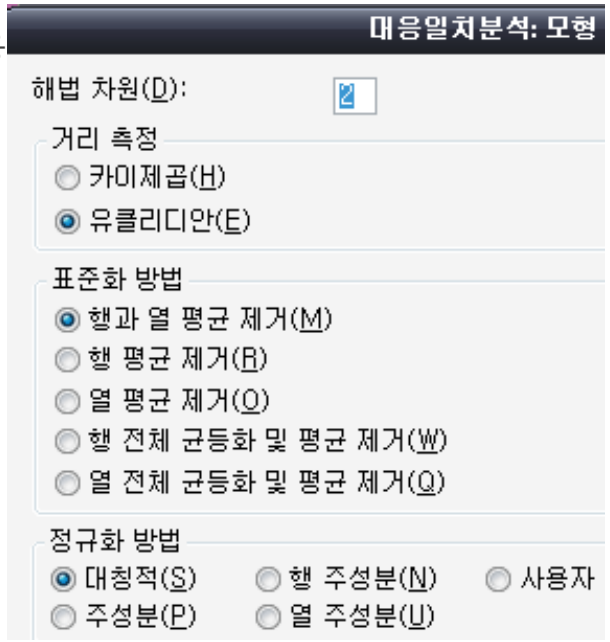
	전공	영어등급	빈도	영어등급분류	전공그룹
1	경영	A	78	1	1
2	경영	B	22	2	1
3	경영	C	20	3	1
4	경제	A	65	1	2
5	경제	B	8	2	2
6	경제	C	2	3	2
7	통계	A	68	1	3
8	통계	B	30	2	3
9	통계	C	7	3	3

http://wolfpack.hannam.ac.kr
한남대학교 통계학과 권세혁 교수

- 행에 설명변수, 열에 종속변수를 지정하고 각 변수의 범주를 지정한다.

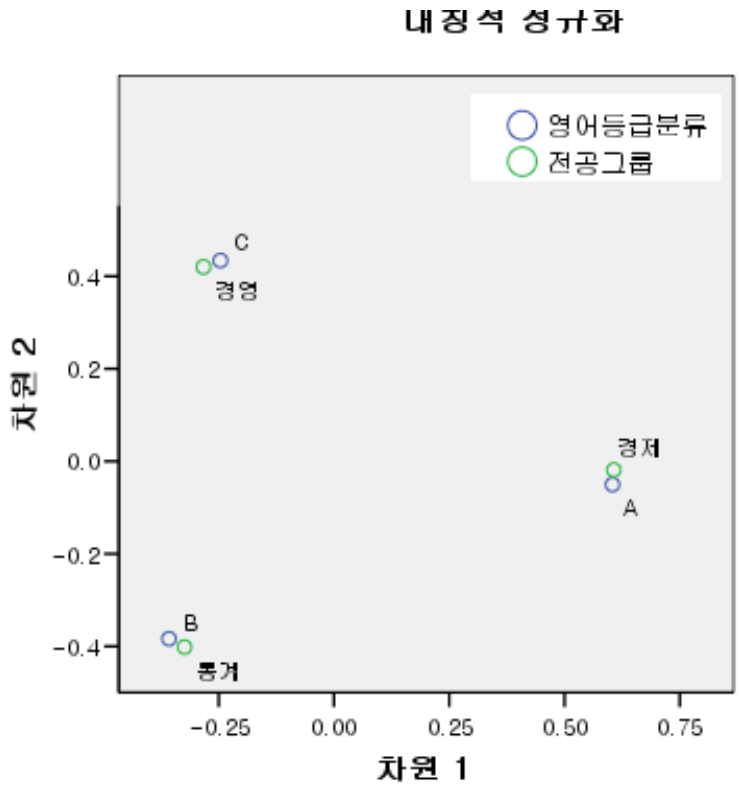


- 옵션 설정
  - 이 설정 사용



- 범주가 동일 산점도에 표현되는 Bi-plot만 보자.

- 결과 해석



- (경제, A)가 유사, 경제학과 학생의 영어 등급이 높다.
- (경영, C), (통계, B)이므로 경제>통계>경영 순이다.

- HOMEWORK due 12.5.2007
  - 다음 분할표(15대 대통령 선거 지역별 득표 현황)에 대해 후보에 따른 지역별 득표 차이가 있는지 교차분석하고 대응분석을 실시하시오. VOTE.xls

지역 \ 후보	이회창	김대중	이인제	권영길	허경영	김한식	신정일
서울	2,394,311	2,627,309	747,846	65,663	5,430	8,975	5,230
부산	1,117,069	320,178	623,756	25,581	2,252	2,211	3,359
대구	965,607	166,576	173,649	16,258	1,661	1,229	4,108
인천	470,560	497,839	297,739	20,340	1,915	2,356	1,862
광주	13,294	754,159	5,181	1,478	154	660	273
대전	199,266	307,493	164,374	8,444	1,028	1,352	936
울산	268,657	80,671	139,615	32,135	625	427	988
경기	1,612,108	1,781,577	1,071,704	47,608	7,077	8,035	7,618
강원	358,921	197,438	257,138	8,231	3,201	1,853	4,161
충북	243,210	295,666	232,254	10,232	2,784	2,313	3,357
충남	235,457	483,093	261,802	9,604	3,011	4,109	4,122
전북	53,114	1,078,957	25,037	4,189	943	4,981	1,973
전남	41,534	1,231,726	18,305	2,199	1,027	4,790	2,255
경북	953,360	210,403	335,087	22,382	4,177	2,476	11,723
경남	908,808	182,102	515,869	27,823	3,215	2,150	8,047
제주	100,103	111,009	56,014	3,856	551	799	1,245