CHAPTER 8

교차 분석

8.1. 교차 분석 (cross-TABULATION) 개요

8.1.1. 교차 분석 개념

두 분류형(범주형) 문항(변수)간의 연관 관계(association)를 볼 때 교차표(분할표)를 작성하여 변수들간 관계를 분석하게 된다. 이를 교차 분석 혹은 χ^2 (chi-SQUARE)검정이라 한다. 교차 분석의 의미는 두 변수의 빈도 표를 교차시켰다는 의미이며 교차 분석에 사용되는 검정 통계량이 χ^2-분포를 (물론 근사 통계량이지만) 따르기 때문에 χ^2-검정이라 한다.

교차표(cross-tabulation: 분할표: contingency table)는 각 분류형 변수에 대한 빈도표를 행과 열로 결합시켜 놓은 형태이다. 일반적으로 행에는 설명 변수에 해당되는 변수를 열에는 반응 변수(종속변수)를 놓으면 된다. 원인이 되는 변수를 독립 변수 또는 설명 변수라 하고 결과 변수를 종속 변수 또는 반응 변수라 한다.

성별(남녀)과 전공 선택(중국, 경제, 정보통계) 문항 간 관계를 알아보고자 한다. 두 문항(변수) 모두 분류형 변수이므로 빈도표를 교차시켜 놓으면 된다. 이 때 성별에 따른 전공 선택의 차이라고 재해석 할 수 있으니 성별이 설명 변수, 전공 선택은 종속 변수가 된다. 성별을 행으로 전공 선택을 열로 해서 교차표를 작성하면 된다.

	중국	경제	정보 통계
남자	n_{11}	n_{12}	n_{13}
여자	n_{21}	n_{22}	n_{23}

8.1.2. RxC 교차표와 검정 통계량

다음은 행(row) 변수의 범주가 R 개, 열(column) 변수의 범주가 C 개일 때 교차표이다. 교차표 작성시 행은 설명 변수(영향을 미치는 혹은 ~따라서), 열은 종속 변수로 하는 것이 일반적이다.

종속 변수 / 설명변수	1	2	…	C	행 총합
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}	$n_{r.}$
열 총합	$n_{.1}$	$n_{.2}$	…	$n_{.c}$	n

$$n_{i.} = \sum_j n_{ij}$$

$$n_{.j} = \sum_i n_{ij}$$

$$n = \sum_i \sum_j n_{ij}$$

- n_{11}=1 번 행, 1 번 열의 빈도수 (위의 예제) 남자이고 중국 전공을 선택한 학생
- n_{23}=2 번 행, 3 번 열의 빈도수 (위의 예제) 여자이고 정보통계 전공을 선택한 학생

두 변수가 관계가 없다, 혹은 설명 변수가 종속 변수에 영향을 미치지 않는다 (예를 들어 성별에 따른 전공 선택의 차이가 없다.) 의미는 두 변수(문항)가 서로 독립이라는 의미이다. 두 변수가 서로 독립이라면 확률 이론에 의해 $P(AB) = P(A)P(B)$ 이 성립한다. (예) P(남자∩경제)=P(남자)P(경제전공)

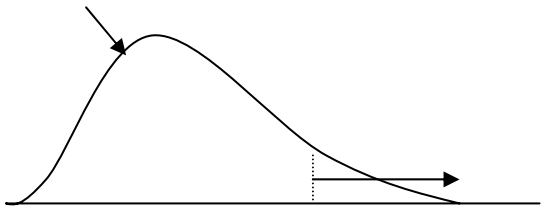
교차표에서 두 변수가 서로 독립이라면 셀 (i, j) 의 확률(비율=빈도/총 개수) P_{ij} 는 $P_{i.} \times P_{.j}$ 으로 나타낼 수 있다. $P_{i.}$ 은 i-번째 행의 확률이고 $P_{.j}$ 은 j-번째 열의 확률이다. 교차표의 빈도 기호로 다시 표시하면

$P_{ij} = \frac{n_{ij}}{n}$, $P_{i.} = \frac{n_{i.}}{n}$, $P_{.j} = \frac{n_{.j}}{n}$ 이고 독립이라면 $P_{ij} = \frac{n_{ij}}{n} = \frac{n_{i.}}{n} \times \frac{n_{.j}}{n}$ 이 성립한다.

두 변수가 독립이라는 가정 하에 i-행, j-열 셀의 예상 빈도는 $\frac{n_{i.}}{n} \times \frac{n_{.j}}{n}$ 이다. 이를 기대 빈도라 하고 E_{ij} 으로 나타낸다. 표본으로부터 계산된(관측된) 빈도를 관측 빈도라 하고 O_{ij} 라 한다. 이 사실을 이용하여 귀무가설(두 변수는 서로 독립이다)을 검정하는데 다음 통계량을 생각할 수 있을 것이다.

$$T = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

위의 검정 통계량의 의미는? 만약 두 변수가 독립이라면 $(O_{ij} = E_{ij})$ 이고 T 값은 0 일 것이다. 즉 T 가 0 에 가까우면 두 변수는 관계가 없다고 결론 내릴 수 있는 것이다. 또한 이 검정 통계량은 $\chi^2(df = (R-1)(C-1))$ 에 근사함이 밝혀져 있다.



기각역 (critical region)

8.1.3. 기대 빈도 5 미만 문제

교차 분석에 사용되는 $\chi^2(df = (R-1)(C-1))$ 검정 통계량은 근사적으로 χ^2-분포에 따른다. 근사 조건으로는 각 셀의 기대 빈도(관측 빈도가 아니다)가 5 이상이어야 한다. Cochran 은 자유도 2 이상인 경우 기대 빈도 5 이상인 셀이 전체 20%만 넘으면 교차표에서 구한 검정 통계량은 Chi-square 분포에 근사 한다고 했다. 일반적으로 Cochran 의 이론을 받아들인다. 통계 소프트웨어는 기대 빈도가 5 미만인 셀의 비율을 출력하여 사용자에게 경고 메시지를 준다. (통계 소프트웨어 이용 방법에서 논의)

기대 빈도가 5 미만인 셀의 비율이 20%를 넘으면 계산된 검정 통계량은 χ^2- 분포에 근사하지 않는다. 이런 경우 해결책은 무엇인가? ①표본의 크기 n 을 늘리면 되지만 이미 설문이 끝난 상태이므로 해결책이 되지 못한다. ②독립성 검정의 경우 변수의 수준을 합쳐 셀의 수를 줄이는 방법이다. 위의 예에서 변수 X 의 수준 중 0 과 1 을 합쳐 하나의 수준으로 하

면 이 문제는 해결된다. 셀을 합칠 경우 그룹으로 할 수 있는 것을 합친다. 예를 들어 수준이 (상, 중, 하)인 경우 '상'과 '하'를 합치는 것은 정말 어리석은 일이다. 수준의 의미가 상실되기 때문이다. (3)동질성 검정의 경우 Exact test 를 시행하는 것이다. 물론 이 방법은 독립성 검정에도 적용될 수 있다. 이는 근사 통계량을 이용하는 것이 아니다. 처음 이 방법을 제안한 사람은 Fisher 인데 그는 2x2 분할표의 경우 제안하였고 후에 RxC 분할표로 확대되었다.

8.1.4. 수작업

A 학부 1 학생 230 명을 대상으로 남녀별 전공 선택(3 개 전공)의 차이가 있는지 알아보고자 하여 자료를 조사하여 다음 교차표를 얻었다고 하자.

성별	전공			
	A 전공	B 전공	C 전공	Total
남자	75	46	23	144
여자	30	32	24	86
Total	105	78	47	230

(1)가설

①귀무가설

두 분류형 변수간의 관계가 없다. 두 변수는 서로 독립이다. 남녀별 전공 선택의 차이는 없다.

②대립가설

관계가 있다. 서로 독립이 아니다. 남녀별 전공 선택의 차이는 있다

(2)검정통계량: 만약 두 변수가 서로 독립이라면 P(표본 n 명 중 ij 셀에 속하는 빈도) = $E_{ij} = (\frac{n_{i.}}{n..})(\frac{n_{.j}}{n..})n_{..}$ 위의 예에서 두 변수가 독립이라면 P(남자∩A 전공) = P(남자) P(A 전공) 이다. 그러므로

▪1 행 1 열의 기대 빈도는 $E_{11} = (\frac{n_{1.}}{n_{..}})(\frac{n_{.2}}{n_{..}})n_{..} = 105 \times 144 / 230 = 65.7$

▪다른 셀도 같은 방법... 2 행 3 열 기대 빈도 $E_{23} = 86 \times 47 / 230 = 17.6$

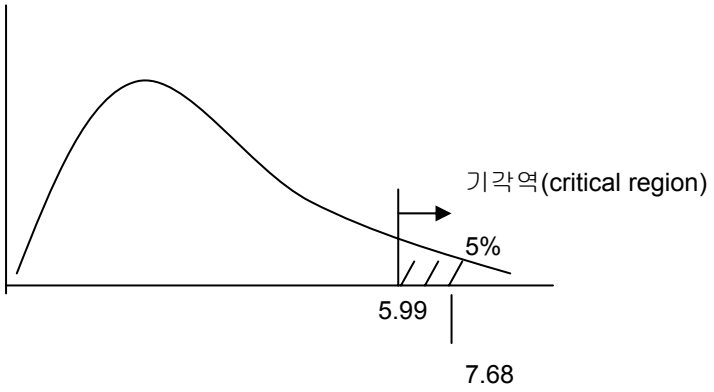
만약 귀무가설이 성립하면 (독립이라면) 각 셀의 $(O_{ij} - E_{ij})$ 는 0 에 가까운 값일 것이다.

검정통계량 $T = \sum\sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ 는 자유도 $(r-1)(c-1)$ χ^2-분포에 근사(approximate) 한다.

$$T = \frac{(75-65.7)^2}{65.7} + ... + \frac{(24-17.6)^2}{17.6} = 7.68 \sim \chi^2(df = (2-1)(3-1) = 2)$$

(3)결론

계산된 검정통계량 값이 χ^2-분포표로부터 구한 기각역에 속하면 귀무가설을 기각하고 그렇지 않으면 귀무가설을 채택한다.



계산된 검정 통계량이 7.68 이므로 기각역에 $\{>5.99\}$ 속하므로 귀무가설이 기각되고 성별과 전공 선택에는 관계가 있다고 할 수 있다. 그러면 어떤 관계가 있는가? 이것에 대한 대답은 행 퍼센트를 참고하면 된다. 설명 변수가 행에 있으므로... SAS 출력을 미리 살펴보면 남자는 A 전공을, 여자는 B 전공을 선호한다고 말할 수 있다. 여자는 각 전공을 골고루 택하지만 남자는 A 전공을 선호하고 있다.

대립 가설이 양측 검정의 형태인데 기각역이 왜 한 쪽 방향만 고려되느냐고 묻고 싶은 사람이 있다면 묻기 전에 생각해 보라. 검정 통계량의 값이 어떻게 계산되었는지를... 그래도 이해가 되지 않으면 머리를 벽에다 세 번 박으면 알게 될 것이다.

8.1.5. 교차표가 주어진 경우

교차표가 주어진 경우에도 SAS 를 이용하여 χ^2-검정 통계량과 검정 결과를 얻을 수 있다.

▪교차표가 만들어져 있는 경우 각 셀의 빈도를 가중치(WEIGHT)로 사용하면 된다.

▪EXPECTED 옵션은 기대 빈도 출력한다.

▪빈도, 백분율(비율), 행 비율, 열(컬럼) 비율 출력된다.

```
DATA ONE;
    INPUT GENDER $ MAJOR $ F @@;
    CARDS;
MALE A 75 MALE B 46 MALE C 23
FEMALE A 30 FEMALE B 32 FEMALE C 24
;
PROC FREQ DATA=ONE;
    TABLE GENDER*MAJOR/CHISQ EXPECTED;
    WEIGHT F;
RUN;
```

GENDER MAJOR

빈도 기대도수 백분율 행 백분율 칼럼 백분율	A	B	C	총합
FEMALE	30 39.261 13.04 34.88 28.57	32 29.165 13.91 37.21 41.03	24 17.574 10.43 27.91 51.06	86 37.39
MALE	75 65.739 32.61 52.08 71.43	46 48.835 20.00 31.94 58.97	23 29.426 10.00 15.97 48.94	144 62.61
총합	105 45.65	78 33.91	47 20.43	230 100.00

GENDER * MAJOR 테이블에 대한 통계량

통계량	자유도	값	확률값
카이제곱	2	7.6823	0.0215
우도비 카이제곱	2	7.6869	0.0214
Mantel-Haenszel 카이제곱	1	7.6186	0.0058
파이 계수		0.1828	
분할 계수		0.1798	
크래머의 V		0.1828	

카이 제곱, 우도비 카이 제곱은 같은 개념의 검정 방법이므로 독립성 검정을 위해서는 카이 제곱(χ^2) 검정을 이용하면 된다. Mantel-Haenszel 카이제곱, 파이 계수, 분할 계수, 크래머의 V 통계량은 순서형(리커드 척도 문항이나 우선 순위 문항, 상/중/하 등이 이에 해당) 문항(변수) 간 상관 정도를 분석할 때 사용한다. 유의 수준은 Mantel-Haenszel 카이제곱(순서형 변수의 상관 계수로 생각하면 된다)만 주어져 있으므로 이를 사용하면 된다. 귀무가설은 "상관 관계가 없다" 이다.

8.2. 설문 분석에 교차 분석 적용

①인구학적 변인(일반적으로 분류형 문항, 즉 보기 문항)에 따른 본 문항(이 문항 역시 보기 문항, 즉 폐쇄형 문항으로 리커드 척도 문항 아님) 선택의 차이는 있는지 ②본 문항의 폐쇄형 문항간 응답자의 선택에 차이는 있는지 알아보는데 교차 분석이 사용된다. 예제 설문의 경우 다음을 알기 위해서는 교차 분석을 실시하면 된다.

- ▪성별(Q1)에 따른 전공 선택(Q27) 차이는 있는가?
- ▪성별(Q1)에 따른 전공 선택 시 취업의 우선 순위 (Q26_1)의 차이는 있는가?
- ▪대학원서 접수 때 원하는 전공선택 여부(Q28)에 따른 전공 선택(Q27)의 차이는 있는가?

교차 분석은 두 문항 모두 분류형 변수이어야 하고 두 문항과 관계(이것이 연구 가설)를 보기 위하여 실시하는 것이다. 물론 리커드 척도 문항도 폐쇄형 문항으로 간주하여 인구학적 변인과 교차 분석이 가능하지만 이미 척도라는 개념에 의해 분산분석을 실시하는 것이 일반적이다. 이 부분에 대해서는 8.4 절에서 좀 더 다루기로 한다.

우선 순위 문항도 순위가 5 개 정도 되면 점수로 간주하여 기초 통계량 분석을 할 수 있으나 3 개 정도면 (물론 무리해서 평균, 표준 편차를 구할 수 있으나) 각 우선 순위를 범주형 (폐쇄형) 변수로 간주하여 교차 분석을 실시하는 것이 더 옳은 방법이다.

8.3. 통계 소프트웨어 이용

다음 연구 주제에 대해 분석하여 보자.

①성별(Q1)에 따른 전공 선택(Q27)의 차이는 있는가?

②출신 지역(Q3)에 따른 전공 결정 여부(Q28)의 차이는 있는가?

③성별(Q1)에 따른 취업 전망 우선 순위 선택(Q26_1)의 차이는 있는가?

④대학 원서 접수 때 원하는 전공 선택 여부(Q28)에 따른 전공 선택(Q27)의 차이는 있는가?

8.3.1. SAS

▣성별(Q1)에 따른 전공 선택(Q27)의 차이는 있는가?

```
PROC FREQ DATA=SURVEY;
    TABLE (Q1 Q28)*Q27/NOCOL NOPERCENT CHISQ;
RUN;
```

교차표를 작성할 때 행은 설명 변수(~따른, ~의해)를 적어야 한다. NOCOL 옵션은 열 퍼센트를 출력하지 말라는 옵션이고 NOPERCENT 는 셀 퍼센트(%) 출력하지 말라는 옵션이다. CHISQ 는 검정 통계량을 출력하는 옵션이다. 교차 분석이 이 옵션을 모두 사용하자. TABLE (Q1 Q28)*Q27의 의미는 (Q1*Q27), (Q28)*(Q27) 두 개의 교차표를 출력하라는 의미이다.

Q1 * Q27 교차표

Q1	Q27			
빈도 행 백분율	1	2	3	총합
1 여자	81 91.01	4 4.49	4 4.49	89
2 남자	27 71.05	5 13.16	6 15.79	38
총합	108	9	10	127

각 셀에는 빈도와 행 백분율만 나타나 있다. 남녀별 차이를 보기 위해서는 각 행에서 비율이 가장 큰 셀 혹은 가장 낮은 셀에 표시하자. 남녀 모두 중국 전공을 선호하고 있으며

남자의 경우 그 다음 전공으로 정보 통계를 생각하고 있다. 물론 이런 해석도 카이-제곱(χ^2) 검정 결과 유의해야 가능하다. 그러나 꼭 통계적 유의성이 필요한가? 사실 우리의 관심은 남자, 여자의 전공 선택 비율의 순서에만 있다고 한다면 통계적 유의성은 학문 연구에서만 중요하지 않을까?

Q1 * Q27 테이블에 대한 통계량

통계량	자유도	값	확률값
카이제곱	2	8.3826	0.0151
우도비 카이제곱	2	7.6998	0.0213
Mantel-Haenszel 카이제곱	1	7.7340	0.0054
파이 계수		0.2569	
분할 계수		0.2488	
크래머의 V		0.2569	

경고: 셀들의 33%가 5보다 작은 기대도수를 가지고 있습니다.
카이제곱 검정은 올바르지 않을 수 있습니다.

유의 확률(p-값)이 0.0151 이므로 남녀별 전공 선택의 차이는 있다. 앞의 해석이 유효하다. 문제가 발생했다. 경고에 보면 기대 빈도가 5 미만인 셀이 전체 33%이다. 그러므로 카이제곱(χ^2) 검정을 사용할 수 없다. 이런 경우 문항(변수)의 범주를 합쳐 수를 줄이거나 Fisher 가 제안한 Exact 검정을 하면 된다.

①기대 빈도 출력

우선 기대 빈도를 출력하여 어느 부분이 문제인지 살펴보자.

```
PROC FREQ DATA=SURVEY;
    TABLE Q1*Q27/NOCOL NOPERCENT CHISQ EXPECTED;
RUN;
```

Q1 * Q27 교차표

Q1	Q27			
빈도 기대도수 행 백분율	1	2	3	총합
1	81 75.685 91.01	4 6.3071 4.49	4 7.0079 4.49	89
2	27 32.315 71.05	5 2.6929 13.16	6 2.9921 15.79	38
총합	108	9	10	127

빨간 박스 안에 2개 셀이 기대 빈도 5미만이다. 그러므로 2/6=33(%)이다.

②범주 합치기

셀의 수를 줄이려면 각 문항의 셀을 합쳐야 한다. 그러나 성별은 범주가 2 개이므로 합치는 것은 적절하지 않고 전공도 합치기에는 문제가 있다. 이런 경우 Fisher 의 Exact (정확) 검정을 사용하시오. **다음은 전공을 중국 경제를 합칠 수 있다고 가정하고 실시한 분석이다.**

```
DATA SURVEY3;
    SET SURVEY;
        IF (Q27=1) OR (Q27=2) THEN Q27R=1;
        IF (Q27=3) THEN Q27R=2;
RUN;

PROC FREQ DATA=SURVEY3;
    TABLE Q1*Q27R/NOCOL NOPERCENT CHISQ;
RUN;
```

Q1 Q27R

빈도 행 백분율	1	2	총합
1 여자	85 95.51	4 4.49	89
2 남자	32 84.21	6 15.79	38
총합	117	10	127

Q1 * Q27R 테이블에 대한 통계량

통계량	자유도	값	확률값
카이제곱	1	4.6835	0.0305
우도비 카이제곱	1	4.2386	0.0395
연속성 수정 카이제곱	1	3.2558	0.0712
Mantel-Haenszel 카이제곱	1	4.6466	0.0311
파이 계수		0.1920	
분할 계수		0.1886	
크래머의 V		0.1920	

경고: 셀들의 25%가 5보다 작은 기대도수를 가지고 있

5 미만인 셀의 비율이 20% 이상이므로 χ^2 검정을 사용할 수 없다. 그러므로 Fisher 의 정확 검정을 사용해야 한다. 만약 이런 경고가 나오지 않았다면... 유의 확률이 0.0305 이므로 남녀별 전공 선택의 차이는 있다. 그리고 해석은 "남녀 모두 중국/경제 전공을 선호하지만 남자는 여자에 비해 중국/경제 선호 정도가 낮다."라 하면 된다.

③Fisher 의 Exact 검정

```
PROC FREQ DATA=SURVEY;
    TABLE Q1*Q27/NOCOL NOPERCENT CHISQ EXACT;
RUN;
```

EXACT 옵션을 사용하면 된다. Fisher 의 정확 검정은 초기하 분포에 기초한다.

Fisher의 정확 검정

테이블 확률 (P)	0.0016
Pr <= P	0.0115

유효한 표본 크기 = 127
결측값의 개수 = 3

유의 확률이 0.0115이므로 남녀별 차이가 있다. 남녀 모두 중국 전공을 선호하고 있으며 남자의 경우 그 다음 전공으로 정보 통계를 생각하고 있다.

▣출신 지역 (Q3)에 따른 전공 선택(Q27)의 차이는 있는가?

```
PROC FREQ DATA=SURVEY;
    TABLE Q3*Q28/NOCOL NOPERCENT CHISQ;
RUN;
```

Q3 을 행으로 Q28 을 열로 하여 교차표를 작성한다.

Q3 Q28

빈도 / 행 백분율	1	2	총합
1	49 57.65	36 42.35	85
2	11 55.00	9 45.00	20
3	0 0.00	2 100.00	2
4	0 0.00	5 100.00	5
5	11 64.71	6 35.29	17
총합	71	58	129

통계량	자유도	값	확률값
카이제곱	4	9.4447	0.0509
우도비 카이제곱	4	12.0807	0.0168
Mantel-Haenszel 카이제곱	1	0.2370	0.6264
파이 계수		0.2706	
분할 계수		0.2612	
크래머의 V		0.2706	

경고: 셀들의 40%가 5보다 작은 기대도수를 가지고 있

셀의 기대 빈도가 미만인 셀의 비율이 40%로 20%를 넘으므로 χ^2-검정을 사용할 수 없다. 셀을 합치거나 Fisher Exact 검정을 하면 된다. Fisher exact 검정 방법은 위의 예제

(1)를 보기 바란다. 셀 합치기를 다시 한 번 살펴 보기로 하자. 대전+충남, 그외 지역으로 나누어 보자.

```
DATA SURVEY3;
    SET SURVEY;
        IF (Q3<=2) THEN Q3R=1;
        IF (Q3=>3) THEN Q3R=2;
RUN;
```

이런 식으로 하면 결측치가 2보다 적은 수로 인식되어 1로 변환된다.

```
PROC FREQ DATA=SURVEY3;
    TABLE Q3R*Q28/NOCOL NOPERCENT CHISQ;
RUN;
```

Q3R Q28

빈도 / 행 백분율	1	2	총합
1	60 57.14	45 42.86	105
2	11 45.83	13 54.17	24
총합	71	58	129

결측값의 개수 = 1

Q3R * Q28 테이블에 대한 통계량

통계량	자유도	값	확률값
카이제곱	1	1.0097	0.3150
우도비 카이제곱	1	1.0048	0.3162

대전/충남 지역 학생은 원서 접수 전에 전공을 정하였으나 기타 지역 출신자들은 지원 시 성적에 맞는 전공을 선택하기 위하여 정하는 않은 비율이 높았다. 물론 이것은 통계적으로 유의하지 않았지만 (유의 확률=0.315) 중요한가? 비율이 다르다는 것이 사실이 중요하지 않은가?

▣성별(Q1)에 따른 취업 전망 우선 순위 선택(Q26_1) 차이는 있나?

```
PROC FREQ DATA=SURVEY;
    TABLE Q1*Q26_1/NOCOL NOPERCENT CHISQ;
RUN;
```

Q1 Q26_1

빈도 행 백분율	1	2	3	4	5	총합
여자 1	47 55.29	33 38.82	4 4.71	0 0.00	1 1.18	85
남자 2	20 62.50	11 34.38	0 0.00	1 3.13	0 0.00	32
총합	67	44	4	1	1	117

기대 빈도 경고로 인하여 χ^2-검정을 사용할 수 없다. 우선 순위의 수가 5 개인 경우에는 집단간 우선 순위 점수 평균의 차이를 검정(분산 분석 혹은 t-검정)하는 것이 좋다.

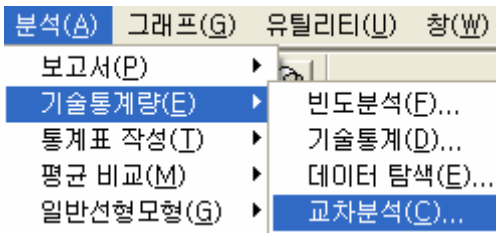
우선 순위가 3 개인 경우에는 교차 분석을 하는 것이 적당하다. 만약 우선 순위 개수가 보기 문항의 수보다 적어 설문 데이터에 입력된 값이 우선 순위 점수가 문항 보기 번호인 경우 다음과 같이 프로그램 하면 된다. Q26_1 에는 1 순위로 선택된 문항 보기가 있다. 프로그램은 동일하지만 이제 열은 더 이상 우선 순위가 아니라 문항 보기이다.

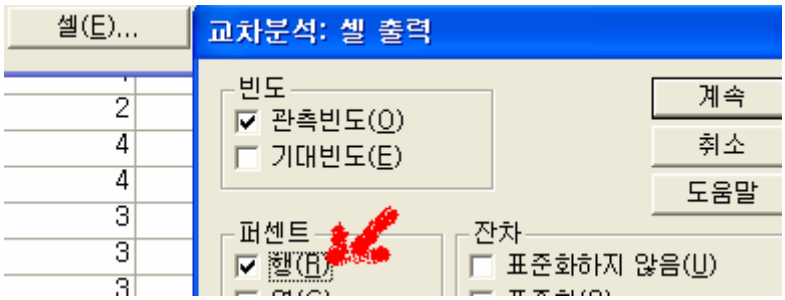
```
PROC FREQ DATA=SURVEY;
    TABLE Q1*Q26_1/NOCOL NOPERCENT CHISQ;
RUN;
```

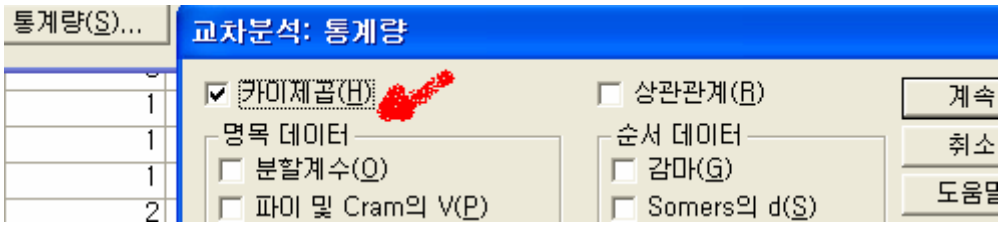
빈도 행 백분율	취업 1	학문 2	적성 3	교수 4	선후배5	총합
여자 1	1 1.23	1 1.23	22 27.16	44 54.32	13 16.05	81
남자 2	0 0.00	0 0.00	8 27.59	13 44.83	8 27.59	29
총합	1	1	30	57	21	110

8.3.2. SPSS

▣성별(V1)에 따른 전공 선택(V31)의 차이는 있는가?







V1 * V31 교차표

			V31			전체
			1	2	3	
V1	1	빈도	81	4	4	89
		V1의 %	91.0%	4.5%	4.5%	100.0%
	2	빈도	27	5	6	38
		V1의 %	71.1%	13.2%	15.8%	100.0%
전체		빈도	108	9	10	127
		V1의 %	85.0%	7.1%	7.9%	100.0%

카이제곱 검정

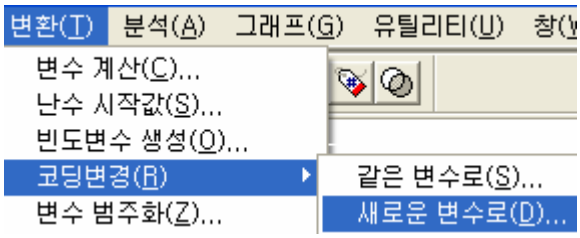
	값	자유도	점근 유의확률 (양쪽검정)
Pearson 카이제곱	8.383[a]	2	.015
우도비	7.700	2	.021
선형 대 선형결합	7.734	1	.005
유효 케이스 수	127		

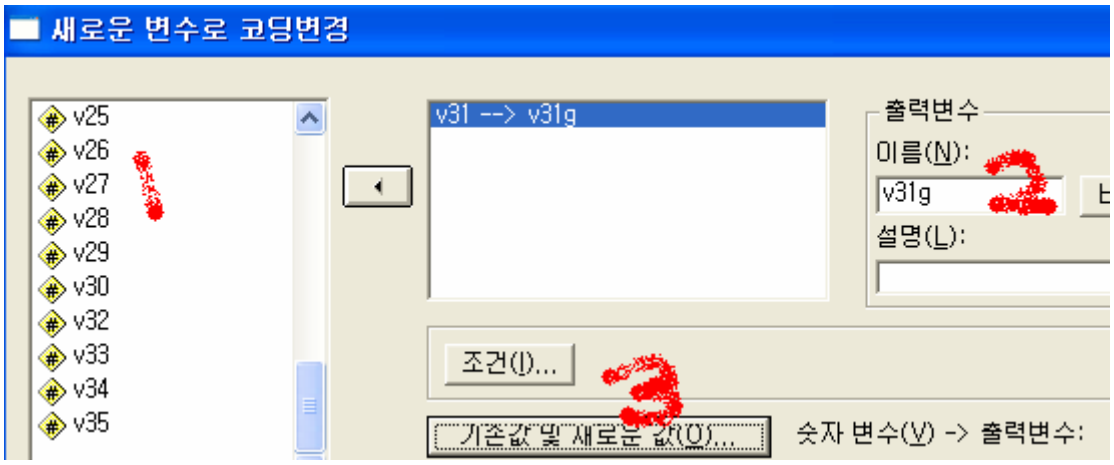
a. 2 셀 (33.3%)은(는) 5보다 작은 기대 빈도를 가지는 셀입니다. 최소 기대빈도는 2.69입니다.

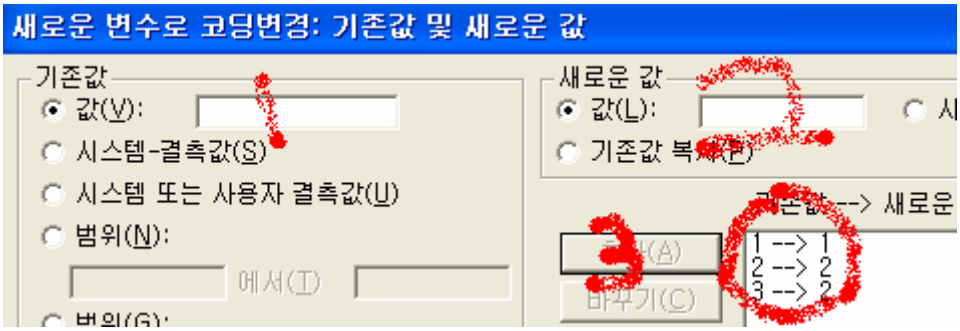
기대 빈도가 5 미만인 셀의 수가 33%로 χ^2-검정을 사용할 수 없다. 셀을 합치거나 Fisher의 Exact 검정을 사용하면 되지만 SPSS 는 (2X2) 교차표 경우에만 사용 결과가 출력된다. 셀 합치는 방법을 사용해 보자.

①셀 합치기

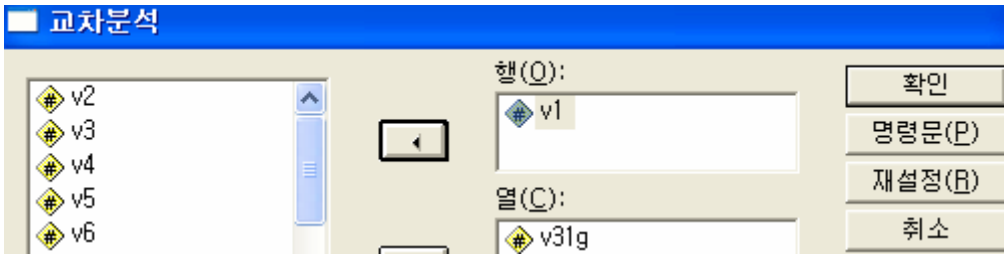
V31 에서 경제와 통계를 하나의 범주로 합하여 문제를 해결하려고 한다고 가정하자.







V1 와 V31G 변수간 교차 분석을 실시하면 된다.



아래 결과와 같이 2X2 교차표에서만 Fisher Exact 검정 결과가 나타난다.

V1 * V31G 교차표

			V31G		
			1.00	2.00	전체
V1	1	빈도	81	8	89
		V1의 %	91.0%	9.0%	100.0%
	2	빈도	27	11	38
		V1의 %	71.1%	28.9%	100.0%
전체		빈도	108	19	127
		V1의 %	85.0%	15.0%	100.0%

카이제곱 검정

	값	자유도	점근 유의확률 (양쪽검정)	정확한 유의확률 (양쪽검정)	정확한 유의확률 (한쪽검정)
Pearson 카이제곱	8.338[b]	1	.004		
연속수정[a]	6.843	1	.009		
우도비	7.661	1	.006		
Fisher의 정확한 검정				.006	.006
선형 대 선형결합	8.272	1	.004		
유효 케이스 수	127				

a. 2x2 표에 대해서만 계산됨

8.4. 보고서 작성

교차 분석 결과는 교차표와 바 차트로 정리하면 된다. 출신 지역을 대전(1), 충남(2), 그외 지역(3)으로 범주를 나누었다고 가정하자. 출신 지역에 따른 전공 선택의 차이가 있는지 알아 보았다.

```
DATA SURVEY3;
     SET SURVEY;
          IF (Q3=1) THEN Q3R=1;
          IF (Q3=2) THEN Q3R=2;
          IF (Q3>=3) THEN Q3R=3;
RUN;

PROC FREQ DATA=SURVEY3;
     TABLE Q3R*Q27/NOCOL NOPERCENT CHISQ;
RUN;
```

빈도
행 백분율

Q3R * Q27 교차표

Q3R	Q27 1	Q27 2	Q27 3	총합
1	75 89.29	5 5.95	4 4.76	84
2	16 80.00	1 5.00	3 15.00	20
3	18 75.00	3 12.50	3 12.50	24
총합	109	9	10	128

결측값의 개수 = 2

다음은 SAS 의 웹 결과를 엑셀에 복사한 후 정리한 후 워드 문서로 가져온 것이다.

출신지	전공 선택			총합
	중국	경제	정보통계	
대전	75 89.29	5 5.95	4 4.76	84
충남	16 80.0	1 5.0	3 15.0	20
기타	18 75.0	3 12.5	3 12.5	24
총합	109	9	10	128

$\chi^2 = 4.79, p = 0.3088$

출신 지역에 따른 전공 선택의 차이는 유의하지 않았다. 출신 지역에 관계없이 중국 전공을 선호하였으나 충남 지역 출신 학생들은 2 순위로 정보 통계 전공을 선호하였다. 다른 지역에서는 경제나 정보 통계 전공 선호에는 차이가 없었다.

교차 분석 절차는 다음과 같다.

①설명 변수 문항을 행, 종속 변수 문항을 열로 하여 교차 분석을 실시한다.

②기대 빈도 5 미만인 셀에 대한 경고 메시지를 확인한다.

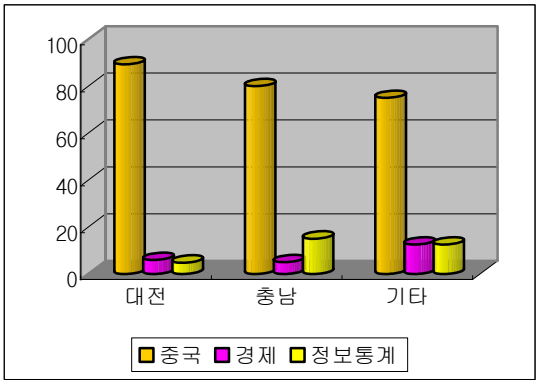
③기대 빈도 5 미만인 셀이 20% 이상이면 셀 합치기(권장)를 하거나 Fisher 의 Exact 검정을 실시한다.

④ χ^2-검정 통계량의 유의성을 확인한다.

⑤통계적으로 유의 하다면 설명 변수 문항(행)의 각 보기(범주)에 대해 행 퍼센트가 높은 순으로 정리한다. 행 퍼센트 비율 차이나 순서 차이를 살펴 해석한다.

적절한 그래프를 그리려면 우선 엑셀에서 표를 다음과 같이 정리할 필요가 있다. 물론 이 표는 위의 표를 수정한 것이 아니라(가능하면 남겨 주는 것이 좋다) 그 아래 복사한 후 수정한 것이다.

	중국	경제	정보통계
대전	89.29	5.95	4.76
충남	80.0	5.0	15.0
기타	75.0	12.5	12.5



인구학적 변인에 따른 리커드 척도 문항 응답의 차이(예를 들면 성별(Q1)에 따른 입학한 것에 대한 만족도의 차이(Q24)는 있는가?)에 대한 적절한 분석 방법은 교차 분석이 아니라 분산 분석이다. 왜냐하면 리커드 척도 문항은 분류형 보기 문항을 척도 개념을 이용하여 측정형 변수로 변환하였기 때문이다. 그럼에도 불구하고 교차 분석을 하려면 다음과 같이 하면 된다.

선택하는 전공(Q27)에 따른 입학한 것에 따른 만족도(Q24)의 차이는 있는지를 알아보고자 한다. 만족도 점수가 7점 척도이므로 흔히 사용되는 5점 척도인 경우로 데이터를 변환하자. 6점과 7점은 5점, 5점은 4점, 4점은 3점, 3점은 2점, 1점과 2점은 1점으로 변환하자. 5점 척도인 경우에는 다음 프로그램이 필요 없다.

```
DATA SURVEY1;
    SET SURVEY;
        IF (Q24=1) OR (Q24=2) THEN Q24R=1;
        IF (Q24=3) THEN Q24R=2;
        IF (Q24=4) THEN Q24R=3;
        IF (Q24=5) THEN Q24R=4;
        IF (Q24=6) OR (Q24=7) THEN Q24R=5;
RUN;
```

```
PROC FREQ DATA=SURVEY1;
    TABLE Q27*Q24R/NOCOL NOPERCENT CHISQ;
RUN;
```
[교차 분석]

```
PROC MEANS DATA=SURVEY1 MEAN STD;
    CLASS Q27;
    VAR Q24R;
RUN;
```

[기초 통계량 계산]

기대 빈도가 5이하인 셀의 비율이 67%로 χ^2-검정 방법을 사용할 수 없으나 예제이므로 이를 무시하자.

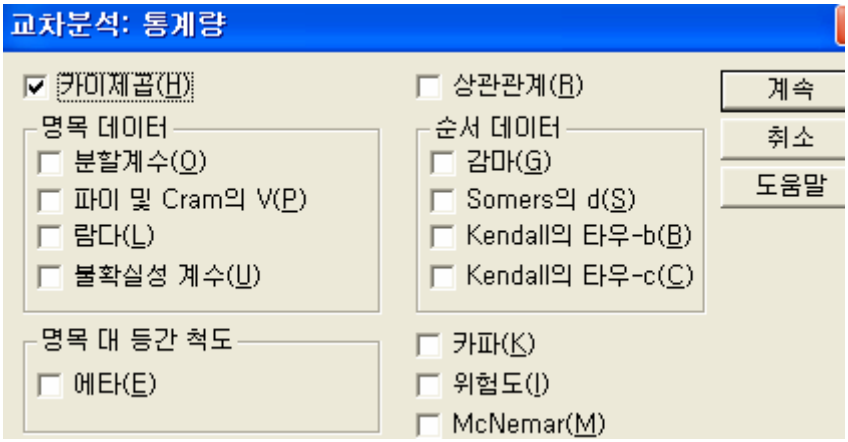
	입학에 대한 만족도					분산분석	
	매우 불만족	불만족	보통	만족	매우 만족	평균	표준편차
중국	19 17.27	21 19.09	39 35.45	21 19.09	10 9.09	2.83[A]	1.19
경제	0 0	3 33.33	3 33.33	0 0	3 33.33	3.33[A]	1.32
정보 통계	2 20	1 10	4 40	3 30	0 0	2.8[A]	1.14
검정통계량=11.37, 유의확률=0.1812						유의확률=0.48	

분산 분석 결과 얻는 방법과 해석 방법은 9장에서 다루기로 한다. 교차분석, 분산 분석 결과 모두 전공 선택에 따른 입학 만족의 차이는 없다고 결론 내릴 수 있다. 교차 분석은 각 행의 수준에 대해 5개 비율이 나타나므로 행 변수에 의한 차이를 보기 위하여 행 퍼센트 순위 차이를 보거나(중국 전공 선택 학생은 보통>불만족. 만족>매우 불만족>매우 만족 순이고 경제 전공은 불만족, 만족, 매우 만족비율이 같다) 각 행에서 행 퍼센트가 가장 큰 셀을 참조하여(중국과 정보 통계는 보통, 경제는 불만족, 보통, 매우 만족) 해석한다. 다소 복잡하다.

분산 분석은 평균 하나의 값에 의해 차이가 해석되므로 입학 만족도가 가장 높은 집단은 경제 전공 선택 학생이고 정보 통계 선택 학생이 만족도가 가장 낮다고 할 수 있다. 물론 유의 확률이 0.48이므로 통계적으로 유의한 차이는 없다.

8.5. McNemar 검정(optional)

SPSS 교차 분석 통계량 옵션을 보면 McNemar 검정이 나온다. 이에 대한 살펴보기로 하자.



수준이 짝을 이루었다는 것은 무슨 뜻인가? 예를 들어 보자. 새로운 이슈가 발생한 경우 A 대통령 후보에 대한 지지 여부가 바뀌었는지 알아보거나 (물론 이 경우 동일 응답자, 즉 패널(panel) 구성), 안전 벨트 착용 거부여부가 교육 전후에 바뀌었는지 알아보고자 할 때 사용되는 방법이다. 전후 사이에 응답 대상이나 실험 대상이 바뀌는 경우나, 전현 다른 것을 측정하는 경우(즉 수준이 달라지는 경우)는 McNemar 방법을 사용할 수 없다.

다음 교차표는 McNemar 검정 방법을 사용할 경우 교차표의 형태이다.

		After		
		Yes	No	Total
Before	Yes	**A**	**B**	A+B
	No	**C**	**D**	C+D
	Total	A+C	B+D	N

반드시 동일 실험 대상이 전후에 사용되어야 하고 같은 개념을 묻거나 실험해야 한다. McNemar 는 이 방법을 수준이 2 개(Yes, No)인 경우만 제안했으나 Bennett & Underwood 가 3 개 이상인 경우로 확대하였다. 편의를 위하여 수준이 2 개인 경우를 가설 검정 순서를 살펴보기로 하자.

(1)가설 (hypothesis)

①귀무가설: $p_1 = p_2$ (실험 전의 yes 비율과 실험 후의 yes 비율이 같다)

②대립가설: $p_1 \neq p_2$ (양측 검정) $p_1 > p_2$ 혹은 $p_1 < p_2$ (단측 검정)

(2)검정 통계량 (test statistic)

표본 추정치: $\hat{p}_1 = \frac{A+B}{N}$, $\hat{p}_2 = \frac{A+C}{N}$ 표본 추정치 차이: $\hat{p}_1 - \hat{p}_2 = \frac{B-C}{N}$

귀무가설이 맞다면 $(B-C)/N = 0$ 이므로 이를 이용하여 McNemar 는 검정 통계량으로 다음을 제안하였고 이가 성립하기 위해서는 (B+C)가 적어도 10 이상이어야 한다.

$$z = \frac{B-C}{\sqrt{B+C}} \sim Normal(0,1)$$

EXAMPLE 안전벨트 교육 효과를 알아보기 위하여 85 명을 임의로 선택하여 교육 전 벨트 작용 여부와 교육 후 벨트 작용 여부를 조사하여 아래 표를 만들었다. 교육 효과가 있는지 검정하시오 (유의수준=0.05)

		교육 후		
		Yes	No	Total
교육 전	Yes	**7**	**37**	41
	No	**26**	**15**	44
	Total	33	52	85

(1) 가설 (hypothesis)

①귀무가설: $p_1 = p_2$ (교육 전후 벨트 착용 비율의 같다.)

②대립가설: $p_1 < p_2$ (교육 후 벨트 착용 비율이 높아졌다.)

(2)검정 통계량 (test statistic): $z = \frac{37-26}{\sqrt{37+26}} = 1.385$

유의 확률 p-value= $pr(z \geq 1.38)$ *2=0.084*2=0.1658 이므로 귀무가설을 기각하지 못한다.

```
DATA ABC;
    INPUT PRE $ POST $ F @@;
    CARDS;
Y Y 7 Y N 37 N Y 26 N N 15
;
RUN;

PROC FREQ DATA=ABC;
    WEIGHT F;
    TABLE PRE*POST/NOCOL NOPERCENT;
    EXACT MCNEM;
RUN;
```

PRE * POST 교차표

PRE	POST		
빈도 | 행 백분율	N	Y	총합
N	15 36.59	26 63.41	41
Y	37 84.09	7 15.91	44
총합	52	33	85

PRE * POST 테이블에 대한 통계량

McNemar 검정

통계량 (S)	1.9206
자유도	1
근사적인 Pr > S	0.1658
정확한 Pr >= S	0.2074

SAS 출력 검정 통계량은 정규분포 $z=\frac{B-C}{\sqrt{B+C}}$ 가 아니라 $\chi^2=\frac{(B-C)^2}{B+C}$ 통계량 값이다.

[연습문제]

(1)성별(Q1)에 따른 ○○대학교에 입학한 것에 대한 만족도(Q24) 차이가 있는지 교차 분석하시오. 7 점 척도이므로 1, 2 점은 불만족, 3, 4, 5 점은 보통, 6, 7 점은 만족으로 하여 분석하시오.

(2)(1)에서 성별을 출신 지역에 바꾸어 교차 분석하시오.

(3)팀 프로젝트 설문에서 인구학적 문항과 본 문항 중 교차 분석이 가능한 것에 대해 분석하고 보고서 작성하시오.