교차분석과 카이제곱 검정

# 1. 교차분석의 통계량 선택 방법

‘카이제곱’을 선택하는 이유는 교차분석의 유의성을 판단하기 위함이다. 이때 설문의 유형에 따라 선택하는 통계량이 다르다.

* 명목척도로 이루어진 경우 : 명목의 ‘람다’에 🗹
* 서열척도(순서척도)로 이루어진 경우 : 순서형의 ‘감마’ 또는 ‘Somers의 d’에 🗹
* 명목척도와 등간척도로 이루어진 경우 : 명목대 구간의 ‘에타’에 🗹

이러한 분석을 통해 유의성이 판단된 후의 결합 정도를 확인할 수 있다.

우리가 사용한 유형은 지역으로, 명목척도 이므로 명목의 ‘람다’에만 🗹를 하여 분석을 진행한다.

# 2. 카이제곱 검정의 전체와 기대빈도의 관계

카이제곱 검정 결과를 적용하여 유의성을 판단하려면 “기대빈도가 5 미만인 셀이 전체의 20%를 넘지 않아야 한다.”는 전체를 만족해야 한다. 왜냐하면 이 전제를 충족하지 못하면 카이제곱 분포에서 벗어나기 때문이다.

앞서 결과창[지역\*구매의사 교차표]를 토대로 관측빈도, 기대빈도를 살펴보고, 카이제곱 부석의 전제를 만족하는지 살펴보자

|  |  |  |  |
| --- | --- | --- | --- |
| 구분 | 구매의사 있음 | 구매의사 없음 | 전체 |
| 지역1 | 154(95.7%) | 52(31.7%) | 206(63.4%) |
| 지역2 | 7(4.3%) | 112(68.3%) | 119(36.6%) |
| 전체 | 161(100%) | 164(100%) | 325(100%) |

▲ 지역\*구매의사 교차표

이 표는 구매의사가 있는 사람 161명과 구매의사가 없는 사람 164명을 기준으로 지역별 차이가 있는지를 조사한 자료이다. 측정도구(설문지)를 이용하여 직접 측정한 값으로, 이를 관측빈도라 한다. 즉 지역 1에서 구매의사가 있음이 95.7%, 구매의사가 없음이 31.7%이고, 전체표본(325명) 중 63.4%가 지역 1에 해당된다.

구매의사가 있음과 없음이 무상관(상관관계=0)이라는 가정하에 지역1의 관측빈도는 63.4%이므로 ‘구매의사 있음” 에 해당하는 빈도의 합인 161의 63.4%(102)를 지역1의 기대값으로 예측할 수 있고, ‘구매의사 없음’에 해당하는 빈도의 합인 164의 36.6%(59)를 지역2의 기대값으로 예측할 수 있다. 이와 같이 관측빈도를 기준으로 예상한 빈도가 기대빈도이다. 이를 계산하면 다음과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
| 구분 | 의사 있음 | 의사 없음 | 전체 |
| 지역1 | 관측빈도 = 154(95.7%)  기대빈도=161\*63.4%=102 | 관측빈도 = 52(31.7%)  기대빈도 = 164\*63.4%=104 | 206(63.4%) |
| 지역2 | 7(4.3%)  기대빈도=161\*36.6%=59 | 관측빈도 = 112(68.3%)  기대빈도 = 164\*36.6%=60 | 119(36.6%) |
| 전체 | 161(100%) | 164(100%) | 325(100%) |

기대빈도는 관측빈도로부터 계산되는데, 그 계산법을 외울 필요는 없다. 그보다는 위에서 설명한 ”기대빈도가 5미만인 셀이 전체의 20%넘지 않아야 한다.” 를 기억하도록 하자. SPSS Statistics는 위와 같이 관측빈도와 거래빈도를 계산해준다.

이를 바탕으로 셀의 개수(표의 행과 열을 곱한 값)를 구해보자. 위의[지역&구매의사 교차표]에서는 2개의 행(‘지역1’, ’지역2’)과 2개의 열(‘구매의사 있음’ , ‘구매의사 없음’)이 있으므로, 셀의 개수는 총 4개(2x2)이다. 한 셀이라도 기대빈도가 5 미만이면 전체 셀의 25%이므로 기준이 되는 20%를 넘어간다. 다행히도 [지역\*구매의사 교차표]에서는 기대빈도가 카이제곱 분포를 벗어나는 경우가 없다. 만약 그 조건들이 성립하지 않는다면 Fisher의 정확한 검정(Fisher’s exact test)으로 추가분석을 해야 한다.