교차분석 및 카이제곱 검정

- 교차법을 통한 자료요약 방법 익히기
- 카이제곱 검정의 개념 파악
- 카이제곱 검정 실습

부산가톨릭대학교
배화숙

교차분석 위해 개념 다시 정리!

- 기술통계 / 추리통계
- 모수 통계
  - 모수에 대하여 정상분포, 분산의 동일성과 같은 분포에 관한 가정을 필요로 하는 통계방법
- 비모수 통계
  - 모수통계 방법을 사용할 수 없을 때 가설검정 또는 의사결정을 위해 사용하는 통계방법
  - 정상분포에 대한 가정 필요 없음
  - 명목 및 서열 변수에도 사용가능
  - 비모수 통계에서 가장 많이 사용되는 방법 χ^2 검증

개념 정리 계속.....

- 일원 통계
  - 1개 변수에 대한 분석
  - 빈도, 평균, 최빈값, 중앙값, 분산 등
- 이원 통계
  - 두 개의 변수에 대한 분석을 동시에 실시하는 통계
  - 성별과 교육정도/ 성별과 경제활동 상태
  - 연령에 따른 학력의 정도
- 그렇다면 교차분석은?

I.교차분석 cross tabulation, crosstabs

- 두 변수의 각 범주가 지니는 값을 교차시켜
  ➔두 변수 간의 관계를 나타내는 분석

- 기술통계로서 교차분석
  - 명목변수와 명목변수를 동시에 분석 통계방법
  - 주로 각 셀의 빈도와 백분율을 설명함

- 추리통계로서 교차분석
  - 변수들 간의 관계에 대한 설명을 위한 분석
  - 두 변수 간 관계 파악하는데 카이제곱검정 사용

2. 교차분석 원리

- 빈도 및 %의 분포가 독립변수의 범주에 따라 통계적으로 유의미한 차이가 있는가 → χ^2 검정

- 동질성 검정
  - 둘 또는 그 이상의 표본이 어떤 분류 기준에 대해
  - 모집단으로부터 동질적으로 추출되었는지의 여부

  - 예) 남 23/45, 여 49/62 합격 ➔ 합격률이 동일한가?

    - Ho : 성별과 합격여부는 관계가 없다

교차표

- 교차분석표 crosstabs
  - 범주 변수를 교차시켜 해당되는 셀에 빈도를 기록한 표
  - 각 변수가 갖는 값의 범주에 따라
    - 행 行 row
    - 열 列 column
  - r ×c 교차분석표
    - 셀에 포함된 사례 수 = 관찰 빈도

	남	여	계
찬성	70	70	140
반대	20	40	60
계	90	110	200

카이제곱 통계량 계산

- 교차분석 목적
  - 변수간 서로 독립적이라는 가설검정 위해
    - 독립적? '두 변수간 아무런 관련이 없다'는 의미
      - 예) '성별과 찬반 의견이 서로 독립적이다' 의 의미
        - ➔ 남학생 찬반 비율과 여학생 찬반비율 동일하다 (≒성별과 찬반은 상관 없다)
    - 즉, 독립변수 집단 간... 종속변수 범주 분포에 의미 있는 차이가 있는지 검정
      - 두 변수 간의 통계적 독립성(statistical independence) 확인

카이제곱 통계량 계산

- 기대빈도
  - 영가설이 참일때 기대되는 빈도

$$\text{기대빈도} = \frac{\text{해당 칸이 속한 열의 총빈도} \times \text{해당칸이 속한 행의 총빈도}}{\text{전체 빈도}}$$

  - 관찰빈도와의 기대빈도 차이가 얼마나 큰가에 따라 영가설 기각 or 수용

➔ 이것을 판단하는 기준 : χ^2통계량

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$\chi^2 = \frac{(70-63)^2}{63} + \frac{(70-77)^2}{77} + \frac{(20-27)^2}{27} + \frac{(40-33)^2}{33}$$

카이제곱 검정에서 자유도

- χ^2 값이 얼마가 되어야 영가설을 기각할까?
  - 자유도에 따라 그 기준 값이 결정됨

A	B	A+ b
c	D	C+ d
A+ c	B+d	A+ b+ c+ d

자유도 df = (r-1)(C-1)

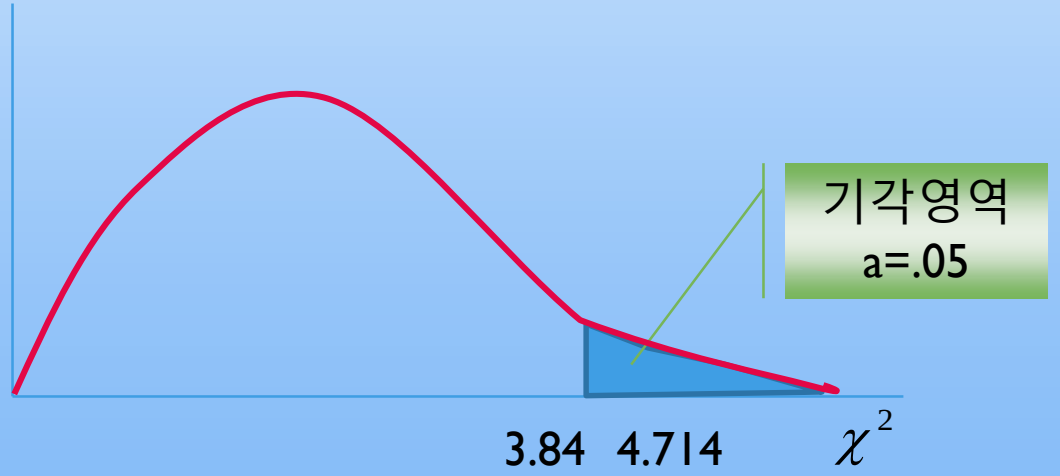
하나의 모수치 추정에 사용되는 독립적 정보의 수 =독립적 정보가 될 수 없는 수를 뺀 정보의 수

3. chi-square 값의 특성

- 양의 값 0 ~ ∞
- 자유도에 영향을 받음
  - 자유도에 따라 카이제곱 값과 분포는 달라짐
- 관찰치의 백분율이 매우 적다면, 관찰빈도는 이론적 분포에 따른 빈도와 일치하지 않는다
- 추리통계학에서 유의도 검증의 근거로 사용하나 관계의 강도를 측정하는 통계량은 아님

3) 카이제곱 검정 → 영가설의 수용과 기각

영가설 전제한 이론적 분포



- 관찰분포와 기대분포의 차이가 영가설이 옳다는 기준을 고려한 것보다 크다($p<.05$)
- ➔영가설 기각
- ➔찬성반대에 대해 남녀간 차이가 남 (우연히 일어난 차이가 아니라는 의미)

4. two-way chi-square test

- 가정
  1. 두 변수가 범주형 변수여야 함
  2. 전체 표본 크기가 최소 30이상 되어야 함
     - 기대빈도가 5보다 작은 셀이 전체 셀의 20%이하여야 함
     - 경험과 통계에 의하면 각 셀의 빈도가 5이상이 되는 경우에 이론적 카이제곱 값에 접근성 확보됨
  3. 각 칸에 떨어져 있는 빈도는 각각 독립적이어야 함

교차분석의 의미(연습)

성별에 따른 흡연 여부

흡연	남	여	계
피움	30	15	45
안피움	20	35	55
계	50	50	100

자유도 df = (r-1)(c-1)

$$\text{기대빈도} = \frac{\text{해당칸이속한열의총빈도} \times \text{해당칸이속한행의총빈도}}{\text{전체빈도}}$$

가설검정절차

1) 각 셀별 기대빈도 구하기
2) 카이제곱 값 구하기
3) 자유도 확인
4) 기준 카이제곱 값과 비교
   - 예) 95% 신뢰수준, 유의 수준 0.05에서의 임계치 확인
5) 영가설 기각 여부 결정
6) 보고서 작성 및 해석