비율비교와 독립성 검정

숙명여자대학교 경영학부 오중산

비율비교 소개

- 비율비교의 정의
  - 두 집단 간에 모비율이 동일한지 여부를 파악하기 위한 통계분석방법
- 표본비율과 표본비율 차이의 표본분포
  - n이 증가하여 np & $n(1\text{-}p) \geq 5$이면 $\bar{p}$는 정규분포로 수렴 $\bar{p} \sim \mathrm{N}(p, \sqrt{\frac{p(1-p)}{n}}^2)$
  - $n_i p_i$와 $n_i(1\text{-}p_i) \geq 5$ $(i=1, 2)$ 이면

$$(\bar{p}_1 - \bar{p}_2) \sim \mathrm{N}(p_1 - p_2,\ \sigma^2{}_{\bar{p}_1 - \bar{p}_2}) \quad SE = \sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}$$

$$\bar{p}(\text{공동추정량}) = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$$

모비율 차이에 대한 가설검정

- 모비율 차이 검정
  - 독립변수와 종속변수
    - 독립변수: 집단을 구분하기 위해 범주형 척도로 측정된 변수
    - 종속변수: 어떤 사건의 발생 여부(두 가지 경우)에 대해 범주형 척도로 측정된 변수
  - 두 가지 가설
    - H_0: $p_1 - p_2 = 0$
    - H_a: $p_1 - p_2 \neq 0$

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - 0}{\sigma_{\bar{p}_1 - \bar{p}_2}}$$

    - 예시1: 20~30대(1)와 40~50대(2) 간의 맥주광고 선호율 차이 검정
    - 예시2: 판촉행사 후 20대(1)의 장바구니-구매 전환률과 30대(2)의 전환률 차이 검정

모비율 차이에 대한 가설검정

- 모비율 차이 검정
  - 예시2에 대한 가설검정
    - H_0: $p_{20} - p_{30} = 0$
    - H_a: $p_{20} - p_{30} \neq 0$
    - 독립변수(customer): 20대와 30대를 구분하기 위한 변수
    - 종속변수(trans): 장바구니에 있던 품목을 실제 구매했는지 여부와 관련된 변수
  - 가설검정 절차
    - 데이터 불러와서 matrix 함수와 rownames & colnames 함수를 이용하여 빈도수 교차표(cross table) 만들기
    - prop.table 함수를 이용하여 비율에 대한 교차표(cross table) 만들기
    - prop.test 함수를 이용하여 양측검정 실시

적합성 검정

- 다항모집단의 정의
  - 모집단이 두 차례 이상의 서로 다른 실험에 의해 두 개 이상의 부분 집합으로 분리될 경우 다항 모집단이라고 함
    - 개별 사례는 반드시 하나의 부분집합에만 속해야 함
    - 부분집합 개수는 '실험 횟수 + 1'과 동일함
    - 예: 숙명여대 학생들을 학년에 따라 네 개 부분집합으로 구분
      - 세 차례 실험(1학년인지 묻는 실험 / 2학년인지 묻는 실험 / 3학년인지 묻는 실험)에 따라 네 개 부분집합으로 구분할 수 있음

적합성 검정

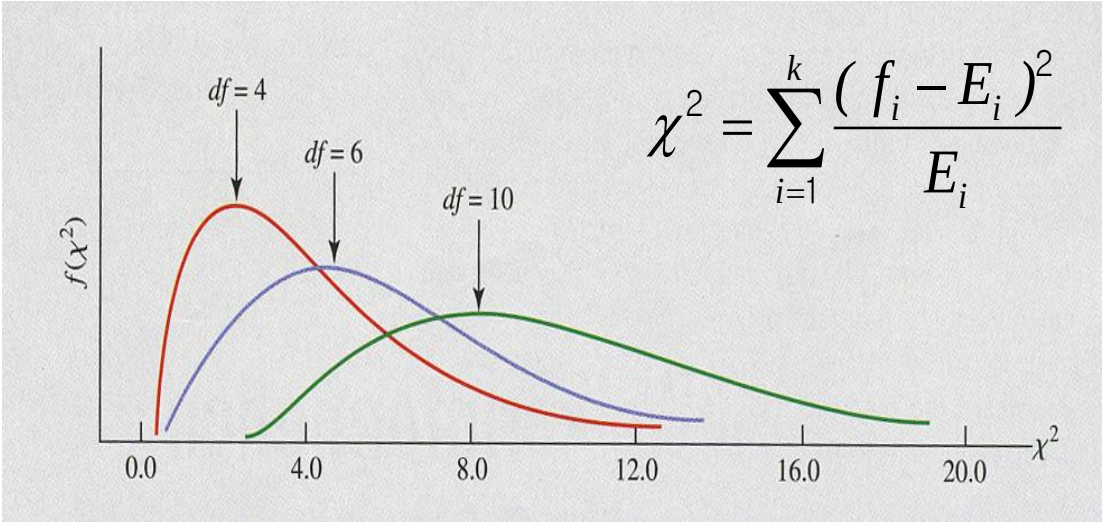
- 다항모집단에 대한 적합성 검정
  - 데이터: telecom.csv
    - 과거(past)와 현재(current)에 대해 300명 사례를 대상으로 이동통신 3사(SKT, KT, LGU+) 가입 여부 조사
    - 과거 시점에 300명을 가입 통신사에 따라 120명, 100명, 80명으로 구분한 다항모집단 존재
    - 현재 시점에 300명을 가입 통신사에 따라 149명, 85명, 66명으로 구분한 다항모집단 존재
  - 다항모집단에 대한 적합성 검정을 위한 세 가지 질문
    - 첫째, 과거에 이동통신 3사의 시장점유율은 동일했는가?
    - 둘째, 현재 이동통신 3사의 시장점유율은 동일한가?
    - 셋째, 과거 이동통신 3사 시장점유율과 현재 이동통신 3사 시장점유율은 동일한가?

적합성 검정

- 다항모집단에 대한 적합성 검정
  - 세 가지 질문과 관련된 가설
    - H_0: (과거에 혹은 현재) 이동통신 3사의 시장점유율은 서로 같다.
    - H_a: (과거에 혹은 현재) 적어도 하나의 이동통신사 시장점유율은 다른 이동통신사 시장점유율과는 다를 것이다.
    - H_0: 과거와 현재 이동통신 3사 각각의 시장점유율은 변화가 없다.
    - H_a: 과거와 현재 이동통신 3사 각각의 시장점유율에 변화가 있다.
      - 만약 H_a가 채택되면 이동통신 3사 각각에 대해 과거-현재 시장점유율 변화에 대해 추가 분석해야 함

적합성 검정

- 다항모집단에 대한 적합성 검정
  - χ^2 검정을 통한 가설검정
    - χ^2 분포는 표준정규분포를 그리는 서로 독립인 k개 변수 제곱의 합
      - 자유도(= 부분집합 개수(k) – 1)가 커지면 정규분포에 수렴
      - 기대값 = 자유도 = 0.5 × 분산
      - F 분포는 χ^2 분포의 비율
    - χ^2값 바깥 쪽 넓이(p-value)가 유의수준 보다 작으면 대립가설 채택
      - 관찰빈도(f_i 혹은 O_i)와 기대빈도(E_i) 차이가 클수록 χ^2은 커지고, 대립가설 채택 가능성도 커짐



적합성 검정

- 다항모집단에 대한 적합성 검정
  - 첫 번째(과거)와 두 번째(현재) 가설에 대한 검정 절차
    - Step1: 귀무가설과 대립가설 수립
    - Step2: chisq.test 함수를 이용하여 가설검정
  - 세 번째(과거와 현재 비교) 가설에 대한 검정 절차
    - Step1: 귀무가설과 대립가설 수립
    - Step2: chisq.test 함수를 이용하여 가설검정
    - Step3: 대립가설이 채택되면 각 집단별로 모비율 비교하는 사후분석 실시

독립성 검정

- 독립성의 정의
  - 두 변수가 서로 영향을 주고 받지 않는, 즉 관련성이 없음을 의미
    - 예: 연령대는 선호하는 정당에 영향을 미치지 않음
      - 연령대는 독립변수이고, 선호하는 정당이 종속변수
      - 연령대에 따라 집단은 r개이고, 선호하는 정당 개수는 k개
      - 연령대별로 선호하는 정당에 대해 k개의 부분집합으로 구분
        - ✓ 연령대별로 다항모집단을 k개의 부분집합으로 나누는 실험 진행
      - 두 개 변수 측정값을 이용하여 다음과 같은 교차표를 작성해야 함

	자유당	보수당	진보정당	합계
18~19세와 20대	40(43)	40(43)	20(15)	100
30대	50(47)	35(47)	25(16)	110
40대	55(49)	40(49)	20(17)	115
50대	50(47)	50(47)	10(16)	110
60대 이상	35(45)	65(45)	5(16)	105
합계	230	230	80	540

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(f_{ij} - E_{ij})^2}{E_{ij}}$$

독립성 검정

- 독립성 검정
  - 두 개의 가설
    - H_0: 연령대는 선호하는 정당(정당 선호도)에 영향을 미치지 않는다.
    - H_a: 연령대는 선호하는 정당(정당 선호도)에 영향을 미친다.
  - 가설에 대한 검정 절차
    - Step1: 귀무가설과 대립가설 수립
    - Step2: chisq.test 함수를 이용하여 가설검정
    - Step3: 대립가설이 채택되면 각 집단별로 모비율 비교하는 사후분석 실시

독립성 검정

- 독립성 검정 실습
  - 데이터: tennis.csv
    - 세계 3대 남자 테니스 선수(Novak Djokovic, Rafael Nadal, Roger Federer) 경기 결과
  - 두 가지 가설
    - 일반적으로 Nadal을 clay court의 강자라고 칭함
      - 프랑스오픈에서 2017~2020년 4연패를 포함 13회 우승
    - 문제제기: Rafael Nadal은 다른 코트에 비해 클레이에서 승률이 더 좋을까?
    - H_0: 코트 유형은 경기결과(승패 혹은 승률)에 영향을 미치지 않는다.
    - H_a: 코트 유형은 경기결과(승패 혹은 승률)에 영향을 미친다.