02강. 이차원 분할표(2)

■ 주요용어

용어	해설
카이제곱분포	카이제곱분포는 자유도 df에 의해서 분포모양이 결정되며 자유도 df가 커짐에 따라 좌우대칭인 형태로 접근해 간다. 자유도가 k 인 카이제곱분포는 서로 독립인 표준정규분포를 따르는 k 개의 확률변수를 제곱하여 합한 분포이다.
우도비 검정	귀무가설이 참인 가정에서 구한 최대우도값과 모수에 대한 제한이 없는 상황에서 구한 최대우도값을 비교하여 가설검정하는 방법이다.
수정잔차	이차원 분할표에서 일반적으로 μ_{ij}가 커짐에 따라 잔차인 $n_{ij}-\hat{\mu}_{ij}$도 커지는 경향이 있다. 이러한 단점을 보완한 수정잔차는 $r_{ij}=\frac{n_{ij}-\hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-p_{i+})(1-p_{+j})}}$로 구한다.
중간순위	범주에 대해서 점수를 부여하는 방법으로 해당 범주에 속한 모든 개체에 대해서 범주에 속한 개체들의 순위의 평균값을 부여하는 방법이다.
적률상관계수	두 변수 X, Y에 대한 적률상관계수는 두 변수 사이 선형관계의 방향과 강도를 살펴볼 수 있는 측도로 다음과 같이 정의된다. $r=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

정리하기

1. 피어슨 통계량과 카이제곱분포
   - 가설 형태

     H_0 : 두 변수 X와 Y가 서로 독립

     H_1 : 두 변수 X와 Y가 서로 연관됨
   - 귀무가설 H_0가 성립하면 다음 사항이 성립함
   - $P(X=i, Y=j) = P(X=i)P(Y=j) \Leftrightarrow \pi_{ij} = \pi_{i+} \cdot \pi_{+j}$
   - $\mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j} \Rightarrow \hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = n\frac{n_{i+}}{n}\frac{n_{+j}}{n}$
   - 카이제곱 검정통계량
   - $X^2 = \sum_{all\ cells} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$
   - X^2은 표본크기가 클 때 자유도 $df = (I-1)(J-1)$인 카이제곱분포를 따름
   - X^2 통계량 값이 클수록 귀무가설 H_0가 옳지 않다는 증거가 뚜렷해짐

     $P-$값 $= P(X^2 \geqq X^2\ observed)$

2. 가능도비 검정(Likelihood Ratio Test, 우도비 검정)
   - 가능도비 검정의 일반형태

     $$\Lambda = \frac{\text{모수가 } H_0\text{를 만족할 때의 최대우도값}}{\text{모수에 대한 제한조건이 없는 상황에서의 최대우도값}}$$
   - 가능도비 Λ는 1을 초과할 수 없고, 모수가 H_0를 만족하지 않을 때 가능도비 Λ는 1보다 훨씬 작아질 것임. 이는 H_0에 대한 강한 반증을 의미함.
   - 가능도비 검정통계량은 $-2\log(\Lambda)$와 동치임
   - 이차원 분할표에 대한 가능도비 검정
   - $G^2 = 2\sum n_{ij}\log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right)$
   - G^2는 표본크기가 클 때 자유도 $df = (I-1)(J-1)$인 카이제곱분포를 따름
   - G^2 값이 클수록 H_0에 대한 강한 반증이 됨

3. 잔차
   - 각 칸별로 관측값과 기댓값을 비교하면 검증결과를 더 잘 이해할 수 있음
   - 잔차 $n_{ij}-\hat{\mu}_{ij}$만을 고려하는 것은 충분한 정보를 주지 못함
     (기대도수가 크게 되는 칸에서는 $n_{ij}-\hat{\mu}_{ij}$도 커지는 경향이 있음)
   - 수정잔차(adjusted residual)
     - $r_{ij}=\dfrac{n_{ij}-\hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-P_{i+})(1-P_{+j})}}$
     - 귀무가설 하에서 r_{ij}는 표준정규분포를 따름
     - $|r_{ij}|>2$ or 3이면 H_0를 따른다고 하는 것이 적합하지 않음을 의미함

4. 카이제곱의 분할
   - $2\times J$ 분할표에서 독립성 검정을 위한 G^2 통계량의 분할
     - 검정통계량의 자유도인 $df=J-1$개로 분할 가능
     - 분할표의 첫 번째 두 열을 비교하는 G^2 값, 첫 번째 두 열을 합한 결과와 세 번째 열과 비교하는 G^2 값, 같은 방법으로 $J-1$번째 열을 합한 결과와 마지막 J번째 열과 비교하는 G^2 값으로 분할
   - 각 G^2는 자유도가 1이고, 합하면 $2\times J$ 분할표에서 G^2 값과 일치함

5. 독립성에 대한 선형추세 대립가설
   - 행변수 X와 열변수 Y가 순서형일 경우의 추세 연관성 분석
     - X 수준이 높아질 때 Y 반응수준이 높아지거나 낮아지는 경향이 있는 경우의 일반적인 분석방법: 범주수준에 점수(score)를 부여하여 선형추세나 상관관계를 측정함
     - 두 변수 X, Y에 대하여 행 점수와 열 점수를 부여하여 피어슨 적률 상관계수를 계산하여 검정함
   - 범주 점수의 선택
     - 대개 범주점수 부여방법에 따른 분석결과에 미치는 영향은 크지 않음
     - 범주점수의 부여 방법은 주관적인 범주점수 부여 방법이나 중간순위 부여 방법이 있음
     - 연구자의 판단에 따라 범주간의 거리를 반영하는 점수를 선택하는 것이 바람직함

6. 코크란-아미티지 추세검정(Cochran-Armitage trend test)
   - 행변수 X가 설명변수이고, 열변수 Y가 반응변수인 경우에 $I\times 2$ 분할표 분석에 관심이 있음

- 순서형 변수 X에 임의의 단조점수를 부여하고 Y에 대해서도 임의의 점수를 부여한 후 $M^2=(n-1)r^2$을 계산하여 X 수준 변화에 따라 특정 범주 비율의 선형추세 여부를 파악함
- M^2 값이 클수록 선형추세의 기울기가 0이 아님을 나타냄

7. 소표본에 대한 정확 추론
   - 카이제곱통계량은 표본크기가 증가함에 따라 근사적으로 카이제곱분포를 따르게 된다는 사실을 이용하여 가설검정 수행
   - 피셔의 정확 추론은 표본크기가 작은 경우의 독립성 검정방법으로 초기하 분포를 이용하여 P-값을 계산함

과제하기

구분	내용
과제 주제	1. 박태성 & 이승연 (2020) 교재 p82의 문제 2.21 2. 박태성 & 이승연 (2020) 교재 p83의 문제 2.23
목적	2주차 강의 내용을 복습하고, 카이제곱통계량의 분할과 순서형 자료 분석에 대한 심층적인 이해를 목적으로 함.
제출 기간	2주차 강의 후 1주 후 토요일 밤 12시까지
참고 자료	범주형 자료분석에서 R 사용방법 관련 자료는 참고문헌에 제시한 자료를 참고하기 바람
기타 유의사항	