12장 교차검증과 카이제곱

Sangkon Han(sangkon@pusan.ac.kr)

2023-03-27

REVIEW

카이제곱 검정

- 교차분석처럼 "범주형" 데이터를 대상으로 범주별 차이를 관련성을 분석
  - 적합도 검정은 확률 모형이 데이터를 얼마나 잘 설명하는지를 검정
  - 독립성 검정은 두 변인의 관계가 독립적인지를 검정
    - H_0 두 사건은 관련성이 없음
  - 동질성 검정은 각 범주간의 비율이 동일한지 검정
    - H_0 모든 표본들의 비율은 동일
- 일원 카이제곱 검정, 이원 카이제곱 검정으로 분류
  - 일원은 한개의 범주 -> 적합도 검정
  - 이원은 두개 이상의 범주 -> 독립성 / 동질성

자유도

일반적으로 카이제곱 분포는 자유도에 따라 달라지기 때문에, 수학자들이 '카이제곱분포표'와 같은 것을 제공합니다. 일반적으로 통계학에서 '자유도(degree of freedom)'란 개념은 매우 모호하고 이해하기 힘듭니다(하지만 계산이 쉬운 경우도 있습니다. 예를 들어서, 다변량정규분포를 따르는 확률벡터의 이차형식(quadratic form)으로 카이제곱분포를 나타내는 경우 행렬의 계수(rank)를 구함으로써 자유도를 구할 수 있는데, 멱등행렬에서는 대각합(trace)와 계수가 같다는 성질을 이용).

'자유도'를 아주 간단한 예를 들어 설명하자면, 5개의 숫자 평균이 3이라고 가정합시다. 이 때 5개 중 4개의 숫자는 마음대로 고를 수 있지만, 마지막 한 개의 숫자는 정해져 있습니다. 제가 1, 1, 1, 1 숫자를 선택했다면 평균 3이 되기 위해서 마지막 숫자는 반드시 11이 되어야 합니다. 이때 숫자 4개는 제가 마음대로 골랐기 때문에 이 경우에 자유도는 4가 됩니다.

12장 연습문제 풀이

1. 교육수준(education)과 흡연율(smoking) 간의 관련성을 분석하기 위한 연구가설을 수립하고, 각 단계별로 가설을 검정하시오. [독립성 검정]

먼저 데이터를 가져와서, 해당 데이터를 확인 후에 원하는 가설과 결론에 사용될 데이터 형태를 파악하세요. 범주형 데이터로 되어있는지 확인하세요.

```r
smoke <- read.csv("./data/smoke.csv", header=T)
head(smoke)
```

```
##   education smoking
## 1         1       1
## 2         1       1
## 3         1       1
## 4         1       1
## 5         1       1
## 6         1       1
```

가능하다면 해당 데이터를 이해하기 쉽도록 정리하세요.

```r
smoke$education2[smoke$education==1] <- "1.대졸"
smoke$education2[smoke$education==2] <- "2.고졸"
smoke$education2[smoke$education==3] <- "3.중졸"
smoke$smoking2[smoke$smoking==1] <- "1.과대흡연"
smoke$smoking2[smoke$smoking==2] <- "2.보통흡연"
smoke$smoking2[smoke$smoking==3] <- "3.비흡연"
head(smoke)
```

```
##   education smoking education2   smoking2
## 1         1       1     1.대졸 1.과대흡연
## 2         1       1     1.대졸 1.과대흡연
## 3         1       1     1.대졸 1.과대흡연
## 4         1       1     1.대졸 1.과대흡연
## 5         1       1     1.대졸 1.과대흡연
## 6         1       1     1.대졸 1.과대흡연
```

교차 분할표를 만들어서 가설을 확인하세요.

```r
table(smoke$education2, smoke$smoking2)
```

```
##
##          1.과대흡연 2.보통흡연 3.비흡연
##   1.대졸         51         92       68
##   2.고졸         22         21        9
##   3.중졸         43         28       21
```

CrossTable을 활용해서 해당 가설의 독립성을 검증하세요.

```r
CrossTable(smoke$education2, smoke$smoking2, chisq = TRUE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
```

```
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  355
##
##
##                  | smoke$smoking2
## smoke$education2 | 1.과대흡연 | 2.보통흡연 |   3.비흡연 | Row Total |
## -----------------|-----------|-----------|-----------|-----------|
##           1.대졸 |        51 |        92 |        68 |       211 |
##                  |     4.671 |     0.801 |     1.633 |           |
##                  |     0.242 |     0.436 |     0.322 |     0.594 |
##                  |     0.440 |     0.652 |     0.694 |           |
##                  |     0.144 |     0.259 |     0.192 |           |
## -----------------|-----------|-----------|-----------|-----------|
##           2.고졸 |        22 |        21 |         9 |        52 |
##                  |     1.476 |     0.006 |     1.998 |           |
##                  |     0.423 |     0.404 |     0.173 |     0.146 |
##                  |     0.190 |     0.149 |     0.092 |           |
##                  |     0.062 |     0.059 |     0.025 |           |
## -----------------|-----------|-----------|-----------|-----------|
##           3.중졸 |        43 |        28 |        21 |        92 |
##                  |     5.568 |     1.996 |     0.761 |           |
##                  |     0.467 |     0.304 |     0.228 |     0.259 |
##                  |     0.371 |     0.199 |     0.214 |           |
##                  |     0.121 |     0.079 |     0.059 |           |
## -----------------|-----------|-----------|-----------|-----------|
##     Column Total |       116 |       141 |        98 |       355 |
##                  |     0.327 |     0.397 |     0.276 |           |
## -----------------|-----------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  18.91092     d.f. =  4     p =  0.0008182573
##
##
##
```

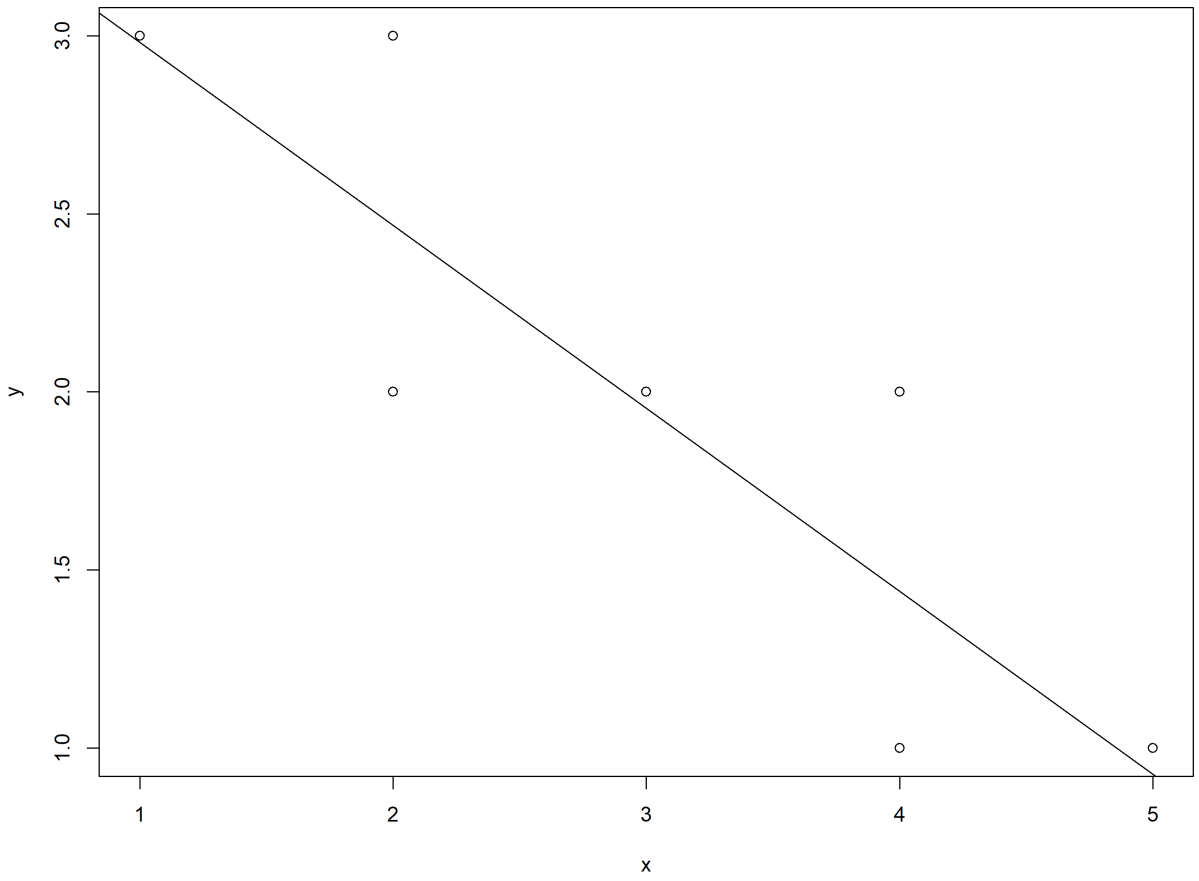
검정 결과의 p-value를 확인하세요. 해당 연습문제의 p-value는 0.0008이므로 유의미한 수준에서 '교육수준과 흡연율 간의 관련성이 있다'라고 볼 수 있습니다.

2. 나이(age3)와 직위(position) 간의 관련성을 단계별로 분석하시오. [독립성 검정]

```
data <- read.csv("./data/cleanData_part3.csv", header=T, fileEncoding = "euc-kr")
head(data)
```

```
##   resident gender job age position price survey   age2 resident2 gender2 age3
## 1        1      1   1  26        4   5.1      5 청년층     특별시     남자    1
## 2        2      1   2  54        1   4.2      4 장년층     광역시     남자    3
## 3        4      2  NA  45        2   3.5      4 중년층     광역시     여자    2
## 4        5      1   3  62        1   5.0      5 장년층     시구군     남자    3
## 5        3      1   2  57       NA   5.4      4 장년층     광역시     남자    3
## 6        2      2   1  36        3   4.1      2 중년층     광역시     여자    2
```

```r
x <- data$position
y <- data$age3
plot(x,y,abline(lm(y~x),main="나이와 직위에 대한 산점도"))
```




```r
CrossTable(x,y, chisq = TRUE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
```

```
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  208
##
##
##              | y
##            x |         1 |         2 |         3 | Row Total |
## -------------|-----------|-----------|-----------|-----------|
##            1 |         0 |         0 |        60 |        60 |
##              |    17.019 |    18.462 |    51.343 |           |
##              |     0.000 |     0.000 |     1.000 |     0.288 |
##              |     0.000 |     0.000 |     0.706 |           |
##              |     0.000 |     0.000 |     0.288 |           |
## -------------|-----------|-----------|-----------|-----------|
##            2 |         0 |        30 |        25 |        55 |
##              |    15.601 |    10.105 |     0.283 |           |
##              |     0.000 |     0.545 |     0.455 |     0.264 |
##              |     0.000 |     0.469 |     0.294 |           |
##              |     0.000 |     0.144 |     0.120 |           |
## -------------|-----------|-----------|-----------|-----------|
##            3 |         0 |        23 |         0 |        23 |
##              |     6.524 |    35.827 |     9.399 |           |
##              |     0.000 |     1.000 |     0.000 |     0.111 |
##              |     0.000 |     0.359 |     0.000 |           |
##              |     0.000 |     0.111 |     0.000 |           |
## -------------|-----------|-----------|-----------|-----------|
##            4 |        23 |        11 |         0 |        34 |
##              |    18.496 |     0.028 |    13.894 |           |
##              |     0.676 |     0.324 |     0.000 |     0.163 |
##              |     0.390 |     0.172 |     0.000 |           |
##              |     0.111 |     0.053 |     0.000 |           |
## -------------|-----------|-----------|-----------|-----------|
##            5 |        36 |         0 |         0 |        36 |
##              |    65.127 |    11.077 |    14.712 |           |
##              |     1.000 |     0.000 |     0.000 |     0.173 |
##              |     0.610 |     0.000 |     0.000 |           |
##              |     0.173 |     0.000 |     0.000 |           |
## -------------|-----------|-----------|-----------|-----------|
## Column Total |        59 |        64 |        85 |       208 |
##              |     0.284 |     0.308 |     0.409 |           |
## -------------|-----------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  287.8957     d.f. =  8     p =  1.548058e-57
```

```
##
##
##
```

'나이와 직위는 관련성이 있다.'를 분석하기 위해서 A회사 223명을 표본으로 추출한 후 설문조사하여 교차분석과 카이제곱검정을 실시하였다. 분석결과를 살펴보면 나이와 직위의 관련성은 유의미한 수준에서 차이가 있는 것으로 나타났다.(X^2=309.369, p<0.05) 따라서 관련 가설을 채택하고 세부적인 내용을 살펴보도록 하자.

3. 직업유형에 따른 응답정도에 차이가 있는가를 단계별로 검정하시오.[동질성 검정]

```
result <- read.csv("./data/response.csv", header=T, fileEncoding = "euc-kr")
head(result)
```

```
##   job response
## 1   1        1
## 2   1        1
## 3   1        1
## 4   1        1
## 5   1        1
## 6   1        1
```

```
result$job2[result$job==1] <- "1.학생"
result$job2[result$job==2] <- "2.직장인"
result$job2[result$job==3] <- "3.주부"
result$response2[result$response==1] <- "1.무응답"
result$response2[result$response==2] <- "2.낮음"
result$response2[result$response==3] <- "3.높음"
```

```
table(result$job2, result$response2)
```

```
##
##            1.무응답 2.낮음 3.높음
##   1.학생         25     37      8
##   2.직장인       10     62     53
##   3.주부          5     41     59
```

```
str(result)
```

```
## 'data.frame':    300 obs. of  4 variables:
##  $ job      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ response : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ job2     : chr  "1.학생" "1.학생" "1.학생" "1.학생" ...
##  $ response2: chr  "1.무응답" "1.무응답" "1.무응답" "1.무응답" ...
```

```
chisq.test(result$job, result$response)
```

```
##
##  Pearson's Chi-squared test
##
## data:  result$job and result$response
## X-squared = 58.208, df = 4, p-value = 6.901e-12
```

세 집단의 응답률이 6.901e-12로 응답간 차이가 없음을 확인할 수 있습니다.