컴퓨터응용통계
범주형 자료분석

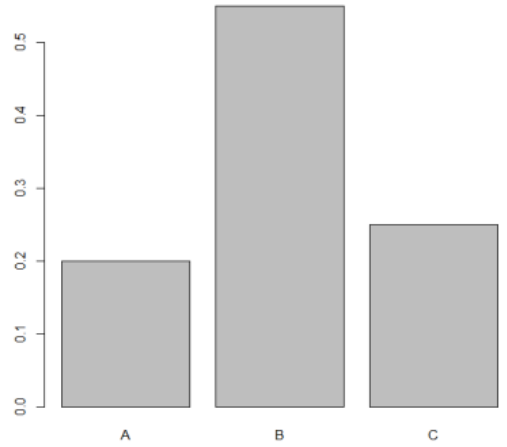
최경미

예제 11.3 적합도 검정

- 휴대전화 브랜드 A, B, C의 시장점유율이 1:2:1인지 조사하기 위해서 100명을 무작위로 뽑아 휴대전화 보유현황을 조사하여, A 20개, B 55개, C 25개를 얻었다고 가정하자.
- 이 빈도를 100으로 나누어 계산한 상대빈도를 막대의 높이로 나타낸 막대그래프(barplot)은 아래와 같다.

브랜드	A	B	C	합
빈도 (상대빈도 %)	20 (20%)	55 (55%)	25 (25%)	100 (100%)

$$p_A = \frac{1}{4},\ p_B = \frac{1}{2},\ p_C = \frac{1}{4}\ ?$$



범주형 자료(categorical data)

- 질병여부, 후보나 브랜드는 집단(group)을 나타낸다.
- 빈도표(frequency table), 상대빈도표(relative frequency table), 막대그래프(bar plot), 모자이크 그래프 등을 사용한다.
- 적합도 검정 (Goodness-of-fit test): 이들 비율(proportion)이 가정된 분포에 적합한지 검정한다.

11.3 적합도 검정

- 가설

H_0: $p_1 = p_{10}^{1/4}, p_2 = p_{20}^{1/2}, \ \dots, p_K = p_{K0}^{1/4}$ $\quad H_1$: H_0이 아니다.

- 표본크기 n
- 집단 i의 관측빈도(Observed frequency)= $O_i = n_i$
- 귀무가설 H_0이 참일 때 기대빈도(Expected frequency)= $E_i = np_{i0}$
- 둘의 차이가 작으면 귀무가설 H_0이 적합하고,

  둘의 차이가 크면 귀무가설 H_0이 적합하지 않다.
- 피어슨(Pearson) χ^2 –검정통계량

$$\chi^2 = \sum_{\text{모든셀 } i} \frac{(O_i - E_i)^2}{E_i} = \sum_{\text{모든셀 } i} \frac{(n_i - np_{i0})^2}{np_{i0}}$$

- 귀무가설 H_0이 참일 때, 검정통계량의 분포는 근사적으로 $\chi^2 \sim \chi^2(k-1)$ 을 따른다.
- $\chi^2 \geq \chi^2_\alpha(\mathrm{k}-1)$이면, H_0을 기각한다.
- 자유도 = 추정하는 확률 p_i의 개수=k-1

관측빈도 (O), 표본

집단(group)	1	2	…	K	합
관측빈도 (counts)	n_1 O_1	n_2 O_2	…	n_K O_K	n

기대빈도 (E), H_0 에서 가정하는 모집단

집단(group)	1	2	…	K	합
확률(prob)	p_1	p_2	…	p_K	1
기대빈도	$E_1 = np_1$	$E_2 = np_2$	…	$E_K = np_K$	

예제 11.3 적합도 검정

휴대전화 브랜드 A, B, C의 시장점유율이 1:2:1인지 조사하기 위해서 100명을 무작위로 뽑아 휴대전화 보유현황을 조사하여, A 20개, B 55개, C 25개를 얻었다고 가정하자.

$$H_0: p_A = \frac{1}{4},\ p_B = \frac{1}{2}, p_C = \frac{1}{4} \quad H_1: H_0\text{이 아니다}$$

브랜드 집단	A	B	C
O_i	20	55	25
E_i	$E_1 = 100 \times \frac{1}{4} = 25$	$E_2 = 100 \times \frac{1}{2} = 50$	$E_3 = 100 \times \frac{1}{4} = 25$
$\frac{(O_i - E_i)^2}{E_i}$	$\frac{(20-25)^2}{25}$	$\frac{(55-50)^2}{50}$	$\frac{(25-25)^2}{25}$

$$\chi^2 = \frac{(20-25)^2}{25} + \frac{(55-50)^2}{50} + \frac{(25-25)^2}{25} = 1.5 < \chi^2_{0.05}(2) = \text{qchisq(0.95,2)=5.99}$$

유의수준 0.05에서 귀무가설 $H_0: p_A = \frac{1}{4},\ p_B = \frac{1}{2}, p_C = \frac{1}{4}$ 을 기각하지 않는다.

휴대전화의 브랜드 A, B, C별 시장점유율이 1:2:1이라고 볼 수 있다.

R

```
# 표 11.1, 표 11.2
> x <- c(A = 20, B = 55, C = 25)
> x <- as.table(x)
> p <- c(1/4, 1/2,1/4)
> X2 <- chisq.test(x, p=p)
> X2
Chi-squared test for given probabilities
data:  x
X-squared = 1.5, df = 2, p-value = 0.4724
> X2$observed                 # 관측빈도
> X2$expected                 # 기대빈도
```

> 0.05. H_0을 기각한다.

11.4 독립성 검정

- 흡연과 폐암의 관계를 알아보자.
- 설명변수 X는 흡연/비흡연, 반응변수 Y는 폐암/비폐암이다.

설명변수 X = 조건 노출 여부 (Exposed or not)

반응변수 Y = 사건 발생 여부 (Event or not)

- i,j 셀의 빈도 n_{ij}
- 행의 합 $n_{i\bullet} = n_{i1} + n_{i2},\ (\mathrm{i} = 1,2)$
- 열의 합 $n_{\bullet j} = n_{1j} + n_{2j}, (\mathrm{j} = 1,2)$
- i,j 셀의 확률 p_{ij}
- 행의 합 $p_{i\bullet} = p_{i1} + p_{i2}\ (\mathrm{i} = 1,2)$
- 열의 합 $p_{\bullet j} = p_{1j} + p_{2j}\ (\mathrm{j} = 1,2)$

Y=0 Y=1

표본 (O)	Y=1 (No Event)	Y=2 (Event)	total
X=0 X=1 (Not Exposed)	n_{11}	n_{12}	$n_{1\bullet}$
X=1 X=2 (Exposed)	n_{21}	n_{22}	$n_{2\bullet}$
total	$n_{\bullet 1}$	$n_{\bullet 2}$	n

H_0모집단	Y=1 (No Event)	Y=0 (Event)	total
X=1 (Not Exposed)	p_{11}	p_{12}	$p_{1\bullet}$
X=2 (Exposed)	p_{21}	p_{22}	$p_{2\bullet}$
total	$p_{\bullet 1}$	$p_{\bullet 2}$	1

11.4 독립성 검정

- H_0: 조건에 대한 노출 여부와 사건발생은 독립이다. H_1: H_0이 아니다.
- 관측빈도 $O_{ij} = n_{ij}$

- 귀무가설이 참일 때,

  기대빈도 $E_{ij} = n\widehat{p_{ij}} = n\widehat{p_{i\bullet}p_{\bullet j}} = n\,\hat{p}_{i\bullet}\hat{p}_{\bullet j} = \mathrm{n}\,\frac{n_{i\bullet}}{\mathrm{n}}\,\frac{n_{\bullet j}}{\mathrm{n}}$, i, j = 1,2 (독립)

- 피어슨 검정통계량은

$$\chi^2 = \sum_{\text{모든 셀}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- 귀무가설 H_0이 참일 때, $\chi^2 \sim \chi^2(1)$
- 검정통계량 $\chi^2 \geq \chi^2_\alpha(1)$이면, H_0을 기각하고, X와 Y가 독립이 아니다.
- 자유도= (행의 수-1)× (열의 수-1) = (2-1)(2-1)=1
- 설명변수 X가 r개의 범주 (집단), 반응변수 Y가 c개의 범주 (집단)을 나타내면, r × c 교차표에서 피어슨 검정통계량의 분포는 근사적으로 다음과 같다.

$$\chi^2 \sim \chi^2((r-1)(c-1))$$

표본 (O)	Y=1 (No Even)	Y=2 (Event)	total
X=1 (Not Exposed)	n_{11}	n_{12}	$n_{1\bullet}$
X=2 (Exposed)	n_{21}	n_{22}	$n_{2\bullet}$
total	$n_{\bullet 1}$	$n_{\bullet 2}$	n

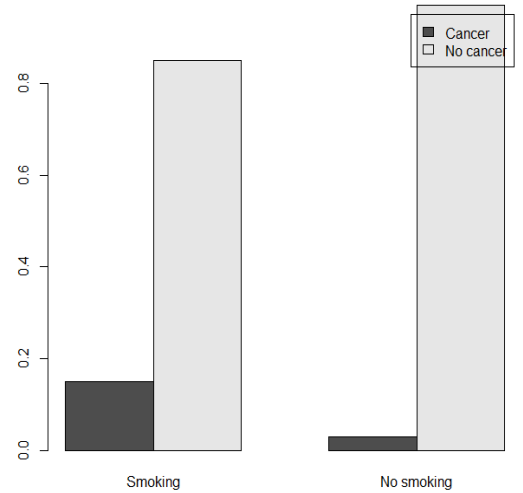
모집단	Y=1 (No Event)	Y=0 (Event)	total
X=1 (Not Exposed)	p_{11}	p_{12}	$p_{1\bullet}$
X=2 (Exposed)	p_{21}	p_{22}	$p_{2\bullet}$
total	$p_{\bullet 1}$	$p_{\bullet 2}$	1

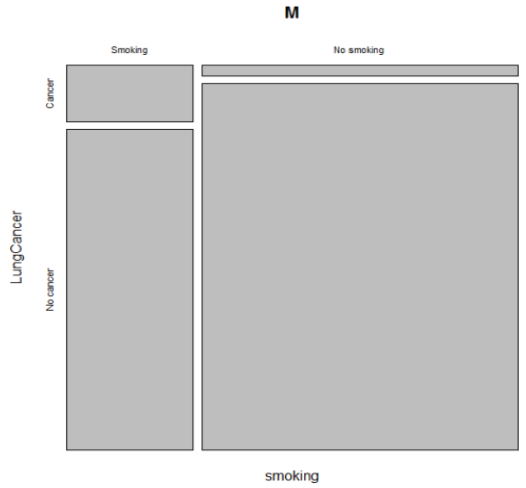
A,B 독립 $\Leftrightarrow P(A \cap B) = P(A)P(B)$

$p_{11} = p_{1\bullet}\, p_{\bullet 1}$
$p_{12} = p_{1\bullet}\, p_{\bullet 2}$
$p_{21} = p_{2\bullet}\, p_{\bullet 1}$
$p_{22} = p_{2\bullet}\, p_{\bullet 2}$

예제 11.4 (가짜 폐암 자료)

행비율 열비율 전체비율	폐암에 걸림	폐암에 걸리지 않음	행합
흡연	6 0.150 0.667 0.043	34 0.850 0.260 0.243	40 0.286
비흡연	3 0.030 0.333 0.021	97 0.970 0.740 0.693	100 0.714
열합	9 0.064	131 0.936	140





H_0: 흡연과 폐암은 독립이다.

H_1: 흡연과 폐암은 독립이 아니다.

O_{ij} (E_{ij})	폐암 = 1	폐암 = 0	행합
흡연 = 1	6 (2.57)	34 (37.43)	40
흡연 = 0	3 (6.43)	97 (93.57)	100
열합	9	131	140

- $E_{11} = 140\ \frac{40}{140}\ \frac{9}{140} = 2.57$
- $E_{12} = 140\ \frac{40}{140}\ \frac{131}{140} = 37.43$
- $E_{21} = 140\ \frac{100}{140}\ \frac{9}{140} = 6.43$
- $E_{22} = 140\ \frac{100}{140}\ \frac{131}{140} = 93.57$
- $\chi^2 = \frac{(6-2.57)^2}{2.57} + \frac{(34-37.43)^2}{37.43} + \frac{(3-6.43)^2}{6.43} + \frac{(97-93.57)^2}{93.57}$

  $= 4.9902$

  $> \chi^2_{0.05}((2-1)(2-1)) = 3.84$
- 유의수준 0.05에서 귀무가설 "H_0: 흡연과 폐암은 독립이다"를 기각한다.
- 유의수준 0.05에서 흡연이 폐암에 영향을 미친다고 볼 수 있다 (p = 0.02549 < 0.05)
- 흡연자와 비흡자 집단에서 폐암의 비율이 다르다.

R

H_0: 흡연과 폐암은 독립이다.

H_1: 흡연과 폐암은 독립이 아니다.

```
# 표 11.5
> r1<-c(6,34)
> r2<-c(3,97)
> M<-as.table(rbind(r1,r2))
> dimnames(M) <- list(smoking = c("Smoking", "No smoking"), LungCancer = c("Cancer", "No cancer") )  # 행과 열의 이름
> library(gmodels)
> CrossTable(mytable)
```

6	34
3	97

```
> X2<-chisq.test(M)
경고메시지(들):
In chisq.test(M) : 카이제곱 approximation은 정확하지 않을 수도 있습니다
> X2
        Pearson's Chi-squared test with Yates' continuity correction
data:  M
X-squared = 4.9902, df = 1, p-value = 0.02549

> X2$observed                                  # 관측빈도
             LungCancer
smoking       Cancer No cancer
  Smoking          6         34
  No smoking       3         97
> X2$expected                                  # 기대빈도
             LungCancer
smoking       Cancer No cancer
  Smoking    2.571429  37.42857
  No smoking 6.428571  93.57143
```

연습문제

2. R의 타이타닉 (Titanic)의 남자 어른 자료에서 객실등급과 생존여부에 대한 교차표 <표1>과 같이 얻고,이에 대한 피어슨 통계량을 아래와 같이 얻었다. 다음 설명 중 틀린 것은 어느 것인가? ②

<표1> 타이타닉의 객실 등급과 생존에 대한 교차표

남자(Male) 어른(Adult)	생존여부 (survived)	
객실등급(class)	No	Yes
1 등실	118	57
2 등실	154	14
3 등실	387	75

```
> mytable <- Titanic[1:3,"Male","Adult",]
> mytable
        Survived
Class     No Yes
  1st    118  57
  2nd    154  14
  3rd    387  75
> chisq.test(mytable)
        Pearson's Chi-squared test
data:  mytable
X-squared = 36.56, df = 2, p-value = 1.151e-08
```

< 0.05

① H_0: 객실 등급이 생존에 영향을 미치지 않는다.

② 검정통계량 χ^2은 36.56이며, 자유도는 4이다.

③ 유의수준 0.05에서 귀무가설을 기각한다.

④ 유의수준 0.05에서 객실등급이 생존여부에 영향을 미친다.

⑤ 위 보기 중 답 없음

H_0: 객실등급과 생존여부는 독립이다.

H_1: 객실등급과 생존여부는 독립이 아니다.

H_0: 객실등급 별 생존율이 동일하다.

H_1: 객실등급 별 생존율이 동일하지 않다.