

## 문제1

1991년 일반사회조사에서 조사된 다음의 4X4 분할표는 흑인들에 대해서 직업만족도와 수입에 대해서 조사한 결과를 보여준다.

수입	직업만족도 매우 불만족	직업만족도 약간 만족	직업만족도 적절히 만족	직업만족도 매우 만족
<5000	2	4	13	3
5000 ~ 15000	2	6	22	4
15000 ~ 25000	0	1	15	8
>25000	0	3	13	8

a.  $\chi^2$  검정을 이용하여 직업만족도와 수입 간의 독립성검정을 수행하라. 검정 결과를 해석하고 이 자료에 대해서 이 검정이 갖고 있는 문제점을 설명하라. 표준화잔차를 구하라. 어떠한 패턴을 보이고 있는가?

```
In [62]: # 데이터 입력
data = matrix(c(2, 4, 13, 3,
                2, 6, 22, 4,
                0, 1, 15, 8,
                0, 3, 13, 8),
              nrow=4, byrow=TRUE)

colnames(data) = c("매우불만족", "약간만족", "적절히만족", "매우만족")
rownames(data) = c("<5000", "5000~15000", "15000~25000", ">25000")
```

```
In [20]: # 카이제곱 검정을 이용하여 직업만족도와 수입 간의 독립성검정 수행
result = chisq.test(data)
result
```

```
Warning message in chisq.test(data):
"Chi-squared approximation may be incorrect"
Pearson's Chi-squared test
```

```
data: data
X-squared = 11.524, df = 9, p-value = 0.2415
```

< 카이제곱 검정 결과 > chisq.test 함수를 통해 카이제곱검정을 실시한 결과, 검정 통계량이 11.524, 자유도는 9, p값은 0.2415 이다.

자유도 df = (I-1)(J-1) 이므로 (4-1)×(4-1) = 9 이다.

p값은 1-pchisq(11.524, df=9) 명령어를 통해 0.2415임을 확인할 수 있다.

```
In [21]: p_value = 1-pchisq(11.524, df=9)
p_value
```

0.241492497242428

< 검정결과에 대한 해석 >

귀무가설은 "수입(독립변수)과 직업만족도(반응변수)는 서로 독립이다"이며,  
대립가설은 "수입(독립변수)과 직업만족도(반응변수)는 서로 독립적이지 않다." 이다.

검정 결과 p값이 0.2415로서 0.05보다 크므로 귀무가설을 기각할 수 없는 것으로 해석할 수 있다.

< 이 검정이 갖고 있는 문제점 >

다만, 행과 열의 각각 4개씩 존재하는 분할표에서 전체 관측치가 104개라면 평균적으로 하나의 셀에 6.5개의 관측이 존재할 수 있는데, 각 셀마다 관측치 수의 편차가 크다는 점을 고려하면 검정결과가 왜곡될 우려가 있다고 판단된다.

```
In [22]: # 표준화잔차 구하기  
result$stdres
```

A matrix: 4 × 4 of type dbl

	매우불만족	약간만족	적절히만족	매우만족
<5000	1.4406207	0.7305280	-0.1606257	-1.079180
5000~15000	0.7525419	0.8715747	0.6005081	-1.772568
15000~25000	-1.1171392	-1.5211385	0.2198088	1.509794
>25000	-1.1171392	-0.1573592	-0.7326960	1.509794

< 표준화잔차의 패턴 > 표준화잔차의 패턴을 살펴보면

첫째, 수입이 낮으면 매우불만족, 약간만족이 기대값보다 많이 나타나고, 수입이 높으면 매우불만족, 약간만족이 기대값보다 적게 나타난다.

둘째, 수입이 낮으면 적절히만족, 매우만족이 기대값보다 대체로 적게 나타나고, 수입이 높으면 적절히만족, 매우만족이 기대값보다 대체로 많이 나타난다.

b. 수입에 대해서는 (3, 10, 20, 35) 점수를 사용하고 직업만족도에 대해서는 (1, 3, 4, 5) 점수를 사용하여 변수들을 질적인 형태로 간주한 후에 검정을 수행하라. 왜 이 검정 결과가 (a)와 다른지 설명하라.



FREQ 프로시저

빈도  
백분율  
행 백분율  
열 백분율

테이블 income * good					
income	good				
	1	3	4	5	합계
3	2 1.92 9.09 50.00	4 3.85 18.18 28.57	13 12.50 59.09 20.63	3 2.88 13.64 13.04	22 21.15
10	2 1.92 5.88 50.00	6 5.77 17.65 42.86	22 21.15 64.71 34.92	4 3.85 11.76 17.39	34 32.69
20	0 0.00 0.00 0.00	1 0.96 4.17 7.14	15 14.42 62.50 23.81	8 7.69 33.33 34.78	24 23.08
35	0 0.00 0.00 0.00	3 2.88 12.50 21.43	13 12.50 54.17 20.63	8 7.69 33.33 34.78	24 23.08
합계	4 3.85	14 13.46	63 60.58	23 22.12	104 100.00

income \* good 테이블에 대한 통계량

통계량	자유도	값	Prob
카이제곱	9	11.5243	0.2415
우도비 카이제곱	9	13.4673	0.1426
Mantel-Haenszel 카이제곱	1	7.0449	0.0079
파이 계수		0.3329	
우발성 계수		0.3158	
크래머의 V		0.1922	
WARNING: 56%개의 셀이 5보다 적은 기대빈도를 가지고 있습니다. 카이제곱 검정은 올바르지 않을 수 있습니다.			

표본 크기 = 104

테이블 income \* good에 대한 요약 통계량

Cochran-Mantel-Haenszel 통계량 (테이블 스코어에 기반)				
통계량	대립가설	자유도	값	Prob
1	영(0)이 아닌 상관계수	1	7.0449	0.0079

전체 표본 크기 = 104

< 점수를 반영하여 질적인 형태로 간주한 변수의 검정 결과 > (a)에서 실시한 카이제곱 검정에 따른 p값은 0.2415 였는데, (b)에서 점수를 부여한 Mantel-Haenszel 카이제곱 검정에 따른 p 값은 0.0079이다. 즉, Mantel-Haenszel 카이제곱 검정에 따르면 귀무가설을 기각할 수 있으며, 수입과 만족도 사이에는 연관성이 존재한다.

< 검정 결과가 a와 다른 이유 > 데이터에 점수를 반영하면 데이터는 단순히 범주형이 아니라 순서형 정보를 포함하게 된다. 점수를 반영한 데이터는 변수의 순서와 강도를 추가로 반영하게 되며, 이는 검정통계량과 p값에 큰 영향을 미치며, 이에 따라 변수간의 관계에도 영향을 미치는 것으로 보인다.

## 문제2

<표 2.15>은 후두암을 치료하는 데 있어 수술과 방사선 치료를 비교하는 연구 결과이다. R 수행한 결과가 다음과 같다.

```
> fisher.test(matrix(c(21,2,15,3), ncol=2, byrow=TRUE),  
alternative="two.sided")  
p-value = 0.6384  
> fisher.test(matrix(c(21,2,15,3), ncol=2, byrow=TRUE), alternative="greater")  
p-value = 0.3808  
> fisher.test(matrix(c(21,2,15,3), ncol=2, byrow=TRUE), alternative="less")  
p-value = 0.8947
```

<표 2.15>

	후두암 치료됨	후두암 치료안됨
수술	21	2
방사선 치료	15	3

a. (i)  $H_0: \theta=1$  과 (ii)  $H_1: \theta \neq 1$  를 검정하기 위해 Fisher의 정확검정 p-값을 구하고 그 결과를 해석하라.

```
In [1]: # 빈도표 정의  
data_matrix <- matrix(c(21, 2, 15, 3), nrow = 2, byrow = TRUE)  
# Fisher의 정확검정 수행  
result <- fisher.test(data_matrix)  
# 결과 출력  
result$p.value
```

0.638425776423209

<  $H_0: \theta = 1$   $H_1: \theta \neq 1$  을 검정하기 위한 Fisher의 정확검정 P-값 >

대립가설이  $H_1: \theta \neq 1$  인 경우 귀무가설은 수술과 방사선 치료가 후두암 치료에 같은 효과를 가진다는 것이고,

대립가설은 수술과 방사선 치료가 후두암 치료에 보이는 효과가 다르다는 것이다.

주어진 바에 따르면 이 경우 P-값은 0.6384 이며, 귀무가설은 기각할 수 없다.

b.  $H_a: \theta \neq 1$  검정을 위한 중앙 p-값을 구하고 그 결과를 해석하고 이 값에 대응되는 신뢰구간을 구하라. 이 중앙 p-값이 기존의 p-값에 비해 어떤 장점이 있는지 설명하라.

```
In [3]: #install.packages('exact2x2')  
library(exact2x2)  
  
exact2x2(data_matrix, tsmethod="central", midp=TRUE)
```

### Central Fisher's Exact Test (mid-p version)

```
data: data_matrix
p-value = 0.4862
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2752008 19.1427855
sample estimates:
odds ratio
 2.061731
```

< 중앙 p-값과 그 결과의 해석 > 중앙 p-값은 0.4862로 기존 p-값인 0.6384보다 작지만 귀무가설을 기각하기에는 여전히 큰 값이다.

95% 신뢰구간은 [0.275~19.14] 이다.

< 중앙 p-값의 장점 > 이산확률분포에서 일반적인 Fisher의 정확검정 P-값은 때때로 과도하게 보수적일 수 있다. 다시 말하면 이산형분포에 대해서는 일반적인 p값이 커진다는 뜻이다. 중앙 P-값은 (더 극단적인 결과의 확률 + 그 밖에 관측된 결과의 확률의 절반)으로 구한다. 이러한 특징 때문에 중앙 P-값은 기존의 P-값보다 덜 보수적이게 되며, 작은 표본에서 보다 신뢰할 수 있는 검정결과를 제공하고, 극단적인 데이터에 대해 더 균형잡힌 평가를 제공한다.