2. 범주형 데이터와 관련된 통계적 추론

독립성 검정

내용

1. 2×2 분할표의 연관성 측도: 오즈비(odds ratio)
2. 2차원 분할표에 대한 독립성 검정

- 2차원 I×J 분할표의 구조(관찰값 분할표)

		Y				
		1	2	$\cdots$	J	
	1	n_{11}	n_{12}	$\cdots$	n_{1J}	n_{1+}
X	2	n_{21}	n_{22}	$\cdots$	n_{2J}	n_{2+}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	I	n_{I1}	n_{I2}	$\cdots$	n_{1J}	n_{I+}
		n_{+1}	n_{+2}	$\cdots$	n_{+J}	n

n_{ij} : i번째 행, j번째 열의 관찰값 빈도 수

n_{i+} : i번째 행의 빈도 수

n_{+j} : j번째 열의 빈도 수

n : 총 빈도 수

1) 2×2 분할표의 연관성 측도: Odds ratio

- 이항변수: 두 개의 범주를 갖는 범주형 변수

- 두 이항변수의 연관성 측도: 오즈비(Odds ratio)

- 오즈(Odds): 어떤 사건이 일어날 확률을 일어나지 않을 확률로 나눈 값

$$odds = \frac{P(A)}{1 - P(A)}$$

● 2×2 분할표에서의 오즈비(Odds ratio)

2×2 관찰 분할표

X	Y	
	Success	Failure
1	n_{11}	n_{12}
2	n_{21}	n_{22}

- X=1인 경우, P(Y=Success)= π_1 X=2인 경우, P(Y=Success)= π_2
- X=1인 경우 Y의 Success odds: $odd1 = \pi_1/(1-\pi_1)$

  X=2인 경우 Y의 Success odds: $odd2 = \pi_2/(1-\pi_2)$
- 두 odds의 비율인 odds ratio:

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

● 오즈비의 특성

X	Y	
	Success	Failure
1	n_{11}	n_{12}
2	n_{21}	n_{22}

$$\theta = \frac{\pi_1/1-\pi_1}{\pi_2/1-\pi_2}$$

- $0 < \theta < \infty$
- 두 변수 X , Y 가 서로 독립이면, $\pi_1 = \pi_2$ → $\theta = 1$
  또한 $\theta = 1$ 이면 두 변수 X , Y는 서로 독립이다.
- 만일 $\theta > 1$, $\pi_1 > \pi_2$ → X=1에서의 성공 가능성이 더 높다.
- 만일 $\theta < 1$, $\pi_1 < \pi_2$ → X=1에서의 성공 가능성이 더 낮다.

- odds ratio θ 와 역수 $1/\theta$ 는 두 변수 사이의 같은 정도의 연관성을 보이나, 방향은 반대

- $\theta = 0.5$: 첫 행의 odds가 둘째 행 odds의 0.5배

  → 둘째 행의 odds가 첫 행 odds의 $1/0.5 = 2$배

- Odds ratio θ 의 추정량

$$\hat{\theta} = \frac{p_1/1-p_1}{p_2/1-p_2} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

X	Y	
	Success	Failure
1	n_{11}	n_{12}
2	n_{21}	n_{22}

$$p_1 = \frac{n_{11}}{n_{11}+n_{12}} \qquad p_2 = \frac{n_{21}}{n_{21}+n_{22}}$$

- Odds ratio 추정량 $\hat{\theta}$의 분포: 오른쪽으로 심하게 치우쳐진 형태
  → $(0, 1)$의 구간과 $(1, \infty)$의 구간이 실질적으로 동일함.

- 효과적인 추론을 위해 추정량의 로그변환이 필요한 상황

- 로그 오즈비 추정량의 점근적인 분포

$$\log\hat{\theta} \;\approx\; N\left(\log\theta\,, \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)$$

- log θ에 대한 100×(1-α)% CI

$$\log\hat{\theta} \;\pm\; z_{1-\alpha/2}\; SE\left(\log\hat{\theta}\right)$$

- 오즈비 θ에 대한 신뢰구간:
  - log θ 신뢰구간의 하한과 상한에 지수 역변환을 적용하여 계산
  - 두 이항변수의 독립성 검정($H_0 : \theta = 1$, $H_1 : \theta \neq 1$)에 사용

예제: Aspirin 복용 여부가 Heart Attack에 미치는 영향 분석

Group	Heart Attack		Total
	Yes	No	
Placebo	189	10845	11034
Aspirin	104	10933	11037

- Placebo 그룹의 odds: 189/10845 = 0.0174
- Aspirin 그룹의 odds : 104/10933 = 0.0095
- odds ratio 추정값 : 0.0174/0.0095 = 1.83
- 로그 odds ratio의
  95% 신뢰구간 : $0.605 \pm 1.96 \times 0.123 = (0.365,\ 0.846)$
- odds ratio의
  95% 신뢰구간 : (exp(0.365), exp(0.846) = (1.44, 2.33)

- R에서 odds ratio 계산

  - 패키지 vcd의 함수 oddsratio( ) 이용

  - 사용 방법:

    oddsratio(x , log = TRUE)

    x : 2×2 행렬 혹은 table 객체

    log = TRUE : 로그 오즈비 계산(디폴트)

    log = FALSE : 오즈비 계산

  - 두 이항변수의 독립성 검정:
    - 함수 oddsratio( )로 생성된 객체에 함수 summary( ) 또는 confint( )를 적용

예제: Aspirin 자료

- 자료 입력

```
> aspirin <- matrix(c(189,104,10845,10933), ncol=2,
            dimnames=list(Group=c("Placebo","Aspirin"),
                          HeartAttack=c("Yes","No")))
> aspirin
         HeartAttack
Group      Yes    No
  Placebo 189 10845
  Aspirin 104 10933
```

- $\log\theta$의 추론

```
> library(vcd)
> my_odd1 <- oddsratio(aspirin)
> summary(my_odd1)

z test of coefficients:

                        Estimate Std. Error z value  Pr(>|z|)
Placebo:Aspirin/Yes:No  0.60544    0.12284  4.9286 8.282e-07
```

$H_0 : \log\theta = 0$

- θ의 추론(95% 신뢰구간)

```
> my_odd2 <- oddsratio(aspirin, log=FALSE)
> confint(my_odd2)
                              2.5 %  97.5 %
Placebo:Aspirin/Yes:No 1.440042 2.33078
```

2) 2차원 분할표에 대한 독립성 검정

- 두 범주형 변수의 독립성 검정
  - Pearson 카이제곱 검정(대표본의 경우)
  - Fisher의 정확검정(소표본의 경우)

- 두 범주형 변수의 분포

  - 결합분포(Joint distribution)

$$\pi_{ij} = P(X = i, Y = j)$$

  - 한계분포(Marginal distribution)

$$\pi_{i+} = P(X = i), \qquad \pi_{+j} = P(Y = j)$$

		Y				
		1	2	$\cdots$	J	
	1	π_{11}	π_{12}	$\cdots$	π_{1J}	π_{1+}
X	2	π_{21}	π_{22}	$\cdots$	π_{2J}	π_{2+}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	I	π_{I1}	π_{I2}	$\cdots$	π_{1J}	π_{I+}
		π_{+1}	π_{+2}	$\cdots$	π_{+J}	1

- 독립성

  - 사건의 독립성: $P(A \cap B) = P(A)\,P(B)$

  - 확률변수의 독립성: $P(X=x, Y=y) = P(X=x)\,P(Y=y)$

- 두 범주형 변수 X와 Y의 독립성:

  → $\pi_{ij} = \pi_{i+}\pi_{+j}\ ,\ i = 1, \cdots, I,\ j = 1, \cdots, J$

2.1) Pearson 카이제곱 독립성 검정

H_0: 두 범주형 변수는 서로 독립 H_1: 두 범주형 변수는 독립이 아님

H_0: $\pi_{ij} = \pi_{i+}\pi_{+j}$ H_1: $\pi_{ij} \neq \pi_{i+}\pi_{+j}$

- 관찰 빈도수: n_{ij}
- 귀무가설에서의 기대 빈도수: $\mu_{ij} = n\pi_{ij}$
- 귀무가설이 사실인 경우 $n_{ij} - \mu_{ij} \approx 0$

- 검정 통계량

$$\chi^2 = \sum \frac{\left(n_{ij} - \hat{\mu}_{ij}\right)^2}{\hat{\mu}_{ij}}$$

- $\hat{\mu}_{ij} = np_{i+}p_{+j} = \frac{n_{i+}n_{+j}}{n}$
- 귀무가설에서 검정통계량의 점근분포 : $\chi^2(df),\ \ df = (I-1)(J-1)$
- 카이제곱 분포를 사용하기 위해서는 대표본이 필수적
  → $\mu_{ij} \geq 5$의 만족이 필요함

- R에서의 Pearson 카이제곱 검정

  - 함수 chisq.test( )의 사용법

    chisq.test(x , y=NULL , simulate.p.value=FALSE)

    - x , y : 두 범주형 변수를 나타내는 벡터
      만일 x가 행렬 혹은 table 객체이면 y는 무시됨

    - simulate.p.value=FALSE : 검정통계량의 근사분포로
      카이제곱 분포를 사용하여 p값 계산

    - simulate.p.value=TRUE : 모의실험을 통하여 p값 계산.
      소규모의 표본에 적합

예제: Aspirin 자료

```
> aspirin
           HeartAttack
Group        Yes    No
  Placebo 189 10845
  Aspirin 104 10933

> chisq.test(aspirin)

        Pearson's Chi-squared test with Yates'
continuity correction

data:  aspirin
X-squared = 24.429, df = 1, p-value = 7.71e-07
```

Yates' continuity correction

- 2×2 분할표에서만 적용
- 이산형인 이항분포를 연속형인 카이제곱 분포로 근사(approximation) 할 때의 오류 감소 효과
- 표본 수가 너무 작은 경우에는 생략(correct=FALSE)

예제: vcd::Arthritis의 Treatment와 Improved의 독립성 검정

- 분할표에 의한 카이제곱 검정

```
> library(vcd)

> (my_table1=with(Arthritis,table(Treatment,Improved)))
          Improved
Treatment None Some Marked
  Placebo   29    7      7
  Treated   13    7     21

> chisq.test(my_table1)

        Pearson's Chi-squared test

data:  my_table1
X-squared = 13.055, df = 2, p-value = 0.001463
```

- 두 범주형 변수의 입력에 의한 카이제곱 검정

```
> with(Arthritis, chisq.test(Treatment, Improved))

        Pearson's Chi-squared test

data:  Treatment and Improved
X-squared = 13.055, df = 2, p-value = 0.001463
```

2.2) Fisher의 정확검정

- Pearson 카이제곱 검정은 표본크기가 충분히 큰 경우 적용 가능한 방법
- 표본크기가 작은 경우 근사분포를 사용하지 않는 방법이 필요

- R. A. Fisher와 관련된 일화

  어떤 영국 부인이 milk tea를 만들 때 찻잔에 차를 먼저 붓고 우유를 나중에 부었는지 아니면 우유를 먼저 붓고 차를 나중에 부었는지 맛으로 구분할 수 있다고 주장하였다. 이 주장을 검정하기 위하여 두 가지 방법으로 각각 4잔의 차를 만들고 맛으로 보게 하여 다음의 결과를 얻었다.

Guess	Truth	
	Milk	Tea
Milk	3	1
Tea	1	3

8잔 중 6잔을 옳게 구분

- Fisher의 검정 절차

Guess	Truth		합계
	Milk	Tea	
Milk	3	1	n_{1+}
Tea	1	3	n_{2+}
합계	n_{+1}	n_{+2}	n

- 2×2 분할표의 열과 행 합계는 모두 고정
- n_{11}만 결정되면 나머지 칸 모두 결정
- n_{11}이 가질 수 있는 값은 0,1,2,3,4

→ 각 값을 가질 확률은 초기하 분포로 결정

- n개의 잔 중 Milk가 먼저 들어간 n_{+1}개의 잔을 선택하는 경우의 수에서
  - Milk Guess n_{1+} 중 n_{11} 이 실제 Milk
  - 따라서 Tea Guess n_{2+} 중 n_{+1}-n_{11} 이 실제 Milk일 확률은

$$\mathrm{P}(n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}}$$

- 부인의 주장을 검정하기 위한 가설

  H_0: 맛으로 구분할 수 없다(θ = 1)
  H_1: 맛으로 구분할 수 있다(θ > 1)

- 위 가설에 대한 p-값: 실험 결과 얻어진 n_{11}의 값 보다 대립가설에서 설정된 방향으로 더 극단적인 값을 취하게 될 확률

  초기하 분포에서 계산

  → p값 = $P(n_{11} = 3) + P(n_{11} = 4)$

- p-값 계산

  1) dhyper( ) 이용: m=4, n=4의 바구니에서 k=4의 공을 꺼내는 경우
  x=3, 4의 확률 계산

```
> dhyper(x=3,m=4,n=4,k=4) + dhyper(x=4,m=4,n=4,k=4)
[1] 0.2428571
```

  2) fisher.test( ) 이용

- 함수 fisher.test( )의 사용법

fisher.test(x , y = NULL , or=1, alternative = "two.sided" ,
conf.int=TRUE, simulate.p.value=FALSE)

- x : 요인 객체 혹은 행렬, table 객체
- y : 요인 객체. x가 행렬이면 무시
- simulate.p.value : 분할표가 2×2보다 큰 경우 p값을 모의실험을 통해 계산할 것인지 여부

나머지 옵션은 2×2 분할표에만 적용
- or=1 : 귀무가설에서 설정되는 odds ratio 값
- alternative= : 대립 가설. 디폴트 값 외에 "less", "greater" 가능
- conf.int : odds ratio에 대한 신뢰구간

```
> TeaTaste <- matrix(c(3,1,1,3),ncol=2,
                  dimnames=list(Guess=c("Milk","Tea"),
                                Truth=c("Milk","Tea")))
> TeaTaste
      Truth
Guess   Milk Tea
  Milk    3   1
  Tea     1   3
```

```
> fisher.test(TeaTaste, alternative="greater")

        Fisher's Exact Test for Count Data

data:  TeaTaste
p-value = 0.2429
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.3135693       Inf
sample estimates:
odds ratio
  6.408309
```

odds ratio의 계산 방식이 앞에서
정의된 것과 다름. 무시하기 바람.

- p-값은 0.2429로 계산
- 두 변수 Guess와 Truth 사이의 통계적 양의 연관성을 확립할 수 없음
- 비록 8잔 중 6잔을 옳게 구분하였으나 부인의 주장이 통계적으로는 입증되지 않음

- 예제: 직업 만족도와 수입의 연관성

Income	Satisfaction			
	VeryD	LittleD	ModerateS	VeryS
< 15K	1	3	10	6
15 – 25K	2	3	10	7
25 – 40K	1	6	14	12
> 40K	0	1	9	11

- 자료입력

```
> Job <- matrix(c(1,2,1,0, 3,3,6,1, 10,10,14,9, 6,7,12,11), ncol=4,
         dimnames = list(income=c("<15k","15-25k","25-40k",">40k"),
             satisfaction=c("VeryD","LittleD","ModerateS","VeryS")))
```

```
> Job
            satisfaction
income    VeryD LittleD ModerateS VeryS
  <15k        1       3        10     6
  15-25k      2       3        10     7
  25-40k      1       6        14    12
  >40k        0       1         9    11
```

- 주어진 분할표가 2×2를 초과. odds ratio의 검정은 불가능
- Pearson 카이제곱 검정과 Fisher의 정확검정으로 독립성여부 확인
- 두 변수의 범주 개수에 비하여 표본 수가 매우 적은 경우
  - 카이제곱 검정에 문제가 발생할 수 있음

    → 해결방법?

- 카이제곱 검정 실시

```
> (Job.t=chisq.test(Job))

        Pearson's Chi-squared test

data:  Job
X-squared = 5.9655, df = 9, p-value = 0.7434

Warning message:
In chisq.test(Job) : 카이제곱 approximation은 정확하지 않을수도 있
습니다
```

- 기대 빈도수 확인

```
> Job.t$expected
              satisfaction
income            VeryD    LittleD   ModerateS    VeryS
    <15k     0.8333333   2.708333    8.958333    7.500
  15-25k     0.9166667   2.979167    9.854167    8.250
  25-40k     1.3750000   4.468750   14.781250   12.375
     >40k    0.8750000   2.843750    9.406250    7.875
```

- 분할표의 전체 칸 중 50% 칸의 기대빈도수가 5미만
  → 카이제곱 분포를 사용하는데 문제가 있음

- 대안 1) p-값을 카이제곱 분포가 아닌 모의실험을 통해 계산
  2) Fisher의 정확검정 적용
  3) 두 범주형 변수의 범주 개수 축소하여 카이제곱 검정 적용

대안 1) 모의실험에 의한 p값 계산

```
> chisq.test(Job, simulate.p.value=TRUE)

        Pearson's Chi-squared test with simulated p-value
(based on 2000 replicates)

data:  Job
X-squared = 5.9655, df = NA, p-value = 0.7746
```

모의실험에 의한 것이기 때문에 실행마다 p값에 약간의
차이가 날 수 있음

대안 2) Fisher의 정확검정 적용

```
> fisher.test(Job)

        Fisher's Exact Test for Count Data

data:  Job
p-value = 0.7827
alternative hypothesis: two.sided
```

대안 3) 범주 개수 축소

- 범주 개수를 축소하여 분할표 전체 칸 수를 줄이면 각 칸의 기대도수가 증가하여 카이제곱 분포 근사의 부정확성 문제 해결

- 변수 income 범주 2개로 축소
  - <15k + 15-25k → <25k
  - 25-40k + >40k → >25k
- 변수 satisfaction 범주 2개로 축소
  - VeryD + LittleD → D
  - ModerateS + VeryS → S

- 패키지 vcdExtra에 있는 함수 collapse.table( ) 사용

- 원래의 분할표

```
> Job
        satisfaction
income    VeryD LittleD ModerateS VeryS
  <15k        1       3        10     6
  15-25k      2       3        10     7
  25-40k      1       6        14    12
  >40k        0       1         9    11
```

- 함수 collapse.table( )에 의한 조정

```
> library(vcdExtra)
> Job.r <- collapse.table(as.table(Job),
                     income=c("<25k","<25k",">25k",">25k"),
                     satisfaction=c("D","D","S","S"))
> Job.r
      satisfaction
income  D  S
  <25k  9 33
  >25k  8 46
```

4×4 분할표를 2×2 분할표로 변환

- 변환된 분할표를 대상으로 독립성 검정

```
> chisq.test(Job.r)

        Pearson's Chi-squared test with Yates'
continuity correction

data:  Job.r
X-squared = 0.3279, df = 1, p-value = 0.5669
```

- 카이제곱 분포의 근사 부정확성 문제는 해결
- 범주의 개수 축소: 몇 개의 범주를 결합시켜 새로운 범주 만드는 작업

  → 범주의 특성을 그대로 유지할 수 있도록 하는 것이 중요