R을 활용한 연관성 분석

호서대학교 빅데이터경영공학부 연규필

질적자료간 연관성 분석

1. 분할표에 대한 교차분석
2. 오즈비
3. 범주형 변수들간의 연관성 지표
4. 심프슨의 패러독스
5. 참고 : 피셔의정확검정, 맥니머검정

1. 분할표에 대한 교차분석

1.1 동일성검정/독립성검정

■ 분할표

두 개의 범주형 변수 A와 B에 대한 교차표(cross table) 혹은 분할표(contingency table)에서, 변수 A가 r개의 범주, 변수 B가 c개의 범주를 가지고 있다면 $r \times c$ 교차표는 다음과 같은 형태를 가진다.

변수 A	변수 B					행 합계
	1	$\cdots$	j	$\cdots$	c	
1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1c}	$n_{1.}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ic}	$n_{i.}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
r	n_{r1}	$\cdots$	n_{rj}	$\cdots$	n_{rc}	$n_{r.}$
열 합계	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.c}$	n

1.1 동일성검정/독립성검정

'청량음료' 데이터

연령대	청량음료				행 합계
	coke	pepsi	fanta	others	
20대	10	14	4	12	40
30대	13	9	10	8	40
40대	12	8	10	10	40

'교육수준과 소득수준' 데이터

교육수준	소득수준			행 합계
	상	중	하	
대졸	255	105	81	441
고졸	110	92	66	268
중졸	90	113	88	291
열 합계	455	310	235	1,000

- 분포에 대한 동일성 검정

$$H_0 : (p_{11}, p_{12}, \cdots, p_{1c}) = \cdots = (p_{r1}, p_{r2}, \cdots, p_{rc})$$

- 두 변수의 독립성 검정

$$H_0 : p_{ij} = p_{i.}p_{.j}, \;\; i = 1, \cdots, r; \;\; j = 1, \cdots, c$$

1.1 동일성검정/독립성검정

■ 관찰도수와 기대도수

관찰도수

n_{ij}

연령/상품	A	B	C	전체
30세 이하	20 20%	20 20%	60 60%	100
30세 이상	70 35%	100 50%	30 15%	200
전체	90 30%	120 40%	90 30%	300

기대도수

e_{ij}

$$e_{ij} = \frac{n_{i.}\, n_{.j}}{n}$$

연령/상품	A	B	C	전체
30세 이하	30 30%	40 40%	30 30%	100
30세 이상	60 30%	80 40%	60 30%	200
전체	90 30%	120 40%	90 30%	300

1.1 동일성검정/독립성검정

■ **카이제곱 검정**

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - E_{ij})^2}{E_{ij}} : \text{Pearson's chi-square}$$

IxJ 분할표에서 카이제곱에 대한 준거분포는 자유도가 (I-1)x(J-1)인 카이제곱 분포 (대표본 이론)

예) (20-30)^2/30 + (20-40)^2/40 + ... + (30-60)^2/60 = 65

p-값 = 0.001, $\chi^2(2)$ = 5.99

⇒ 연령과 상품이 독립이라는 귀무가설을 기각

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c}\log\left(\frac{n_{ij}}{e_{ij}}\right)$$

※ 카이제곱 검증을 위한 충분조건(Cochran의 기준) :

80% 이상의 칸들에서 기대빈도 $E_{ij} \geq 5$.

※ 이 조건을 만족시키지 못하는 경우에는 인접(유사)행 또는 열의 병합 후 카이제곱 검정을 적용하거나 Fisher’s Exact Test를 수행해야 함.

1.1 동일성검정/독립성검정

■ 카이제곱 통계량에 근거한 측도들

- 파이 계수 (Phi Coefficient)

$$\phi = \frac{n_{11}\, n_{22} - n_{12}\, n_{21}}{\sqrt{n_{1.}\, n_{2.}\, n_{.1}\, n_{.2}}} \quad \text{for } 2 \times 2 \text{ tables} \quad \boxed{-1 \le \phi \le 1}$$

$$\phi = \sqrt{Q_P/n} \quad \text{otherwise} \quad \boxed{0 \le \phi \le min(\sqrt{R-1}, \sqrt{C-1})}$$

- 크래머 V (Cramer's V)

$$V = \phi \quad \text{for } 2 \times 2 \text{ tables} \quad \boxed{-1 \le V \le 1}$$

$$V = \sqrt{\frac{Q_P/n}{\min(R-1, C-1)}} \quad \text{otherwise} \quad \boxed{0 \le V \le 1}$$

- 분할계수 (Contingency Coefficient)

$$P = \sqrt{\frac{Q_P}{Q_P + n}} \quad \boxed{0 \le P \le \sqrt{(m-1)/m}\text{, where } m = min(R, C)}$$

1.1 동일성검정/독립성검정

■ 사례 : Prefer 데이터

Prefer 데이터 (첫 10개, $n = 300$)

ID	Agegroup	Product	ID	Agegroup	Product
1	30<	B	6	30<	C
2	30>=	B	7	30<	C
3	30<	B	8	30>=	B
4	30<	A	9	30<	C
5	30>=	B	10	30>=	A

```
Prefer = read.csv("data/Prefer.csv",header=TRUE)
Prefer.table = xtabs(~Agegroup+Product,data=Prefer)
addmargins(Prefer.table,margin=2)   # 관측빈도
addmargins(prop.table(Prefer.table,margin=1),margin=2)
```

```
        Product
Agegroup   A   B   C Sum
    30<   20  20  60 100
    30>=  70 100  30 200
```

```
        Product
Agegroup    A    B    C  Sum
    30<  0.20 0.20 0.60 1.00
    30>= 0.35 0.50 0.15 1.00
```

1.1 동일성검정/독립성검정

■ 사례 : Prefer 데이터

```
fisher.test(Prefer.table)
chisq.test(Prefer.table)
Prefer.chisq = chisq.test(Prefer.table)
addmargins(Prefer.chisq$expected,margin=2)   # 기대빈도
```

```
> fisher.test(Prefer.table)

        Fisher's Exact Test for Count Data

data:  Prefer.table
p-value = 1.703e-14
alternative hypothesis: two.sided
```

```
> chisq.test(Prefer.table)

        Pearson's Chi-squared test

data:  Prefer.table
X-squared = 65, df = 2, p-value = 7.681e-15
```

```
         Product
Agegroup   A   B   C Sum
    30<   30  40  30 100
    30>=  60  80  60 200
```

1.1 동일성검정/독립성검정

■ 사례 : Softdrink 데이터

Softdrink(청량음료) 데이터

Agegroup	Drink	Count	Agegroup	Drink	Count	Agegroup	Drink	Count
1	1	10	2	1	13	3	1	12
1	2	14	2	2	9	3	2	8
1	3	4	2	3	10	3	3	10
1	4	12	2	4	8	3	4	10

```
Softdrink = read.csv("data/Softdrink.csv",header=TRUE)
Softdrink$Agegroup = factor(Softdrink$Agegroup,labels=c("20대","30대","40대"))
Softdrink$Drink = factor(Softdrink$Drink,labels=c("coke","pepsi","fanta","others"))
Softdrink.table = xtabs(Count~Agegroup+Drink,data=Softdrink)
addmargins(Softdrink.table,margin=2)
addmargins(prop.table(Softdrink.table,margin=1)*100,margin=2)
```

```
        Drink
Agegroup coke pepsi fanta others Sum
    20대   10    14     4     12  40
    30대   13     9    10      8  40
    40대   12     8    10     10  40
```

```
        Drink
Agegroup  coke pepsi fanta others   Sum
    20대  25.0  35.0  10.0   30.0 100.0
    30대  32.5  22.5  25.0   20.0 100.0
    40대  30.0  20.0  25.0   25.0 100.0
```

1.1 동일성검정/독립성검정

■ 사례 : Softdrink 데이터

```
fisher.test(Softdrink.table)
chisq.test(Softdrink.table)
```

```
> fisher.test(Softdrink.table)

        Fisher's Exact Test for Count Data

data:  Softdrink.table
p-value = 0.3922
alternative hypothesis: two.sided
```

```
> chisq.test(Softdrink.table)

        Pearson's Chi-squared test

data:  Softdrink.table
X-squared = 6.2, df = 6, p-value = 0.4012
```

1.2 적합도검정

'완두콩' 데이터

완두콩의 형태(Type)	1	2	3	4	합계
관측빈도	315	108	101	32	556
관측비율	56.7%	19.4%	18.2%	5.8%	100%

$$H_0 : p_1 = p_{10}, p_2 = p_{20}, \cdots, p_c = p_{c0}$$

$$\chi^2 = \sum_{i=1}^{c} \frac{(n_i - e_i)^2}{e_i} = \sum_{i=1}^{c} \frac{(n_i - n\,p_{i0})^2}{n\,p_{i0}}$$

≪예≫ p_1=9/16, p_2=3/16, p_3=3/16, p_4=1/16

완두콩 데이터에 대한 적합도 검정

```
> Pea = c(315,108,101,32)
> P0 = c(9/16,3/16,3/16,1/16)
> chisq.test(Pea,p=P0)
#---------------------------
Chi-squared test for given probabilities
data:  Pea
X-squared = 0.47002, df = 3, p-value = 0.9254
#------------------------------------------------------------------------
> Pea.chiaq = chisq.test(Pea,p=P0)
> Pea.chiaq$expected
#------------------------------------------------------------------------
[1] 312.75 104.25 104.25  34.75
```

$$\chi^2 = \frac{(315-312.75)^2}{312.75} + \frac{(108-104.25)^2}{104.25} + \frac{(101-104.25)^2}{104.25} + \frac{(32-34.75)^2}{34.75} = 0.470$$

2. 오즈비

2.1 오즈비

■ 코호트 연구(Cohort study)

- 코호트: 같은 특성을 갖는 개체(사람)들의 모임
- 코호트연구는 전향적 연구 (연구 진행 방향이 시간의 흐름과 일치)의 형태를 가짐
- 추적조사를 해야 하므로 시간이 많이 걸리고 방대한 비용이 들며 연구관리가 어려움

■ 연관성 측도

	결과=0	결과=1	합
그룹0	n_{00}	n_{01}	n_0
그룹1	n_{10}	n_{11}	n_1

	비폐암 (Y=0)	폐암 (Y=1)
흡연그룹 (그룹 0)	99.30% $P_{00}=n_{00}/n_0$	0.70% $p_{01}=n_{01}/n_0$
비흡연그룹 (그룹 1)	99.95% $p_{10}=n_{10}/n_1$	0.05% $p_{11}=n_{11}/n_1$

- 비율의 차 (difference in rate)

$$\Delta = p_{11}-p_{01} = 0.70\%-0.05\% = 0.65\%$$

- 상대비 (relative rate)

$$\Delta = p_{11}/p_{01} = 0.70\%/0.05\% = 14.0$$

- 오즈비 (odds ratio)

$$\psi = (p_{11}/p_{10})/(p_{01}/p_{00})$$
$$= (0.70/99.30)/(0.05/99.95) = 14.09$$

✓ 반응(Y=1)의 출현비율이 아주 작은 경우 오즈비는 상대비율과 거의 동일한 값을 갖는다.

2.1 오즈비

■ 사례-대조 연구(Case-Control study)

- 후향적 연구 형태를 띔
- 시간과 비용 측면에서 경제적임
- 관측편의 및 선택편의가 개입될 소지가 있음

■ 연관성 측도

	비폐암 대조군	폐암 사례군
흡연그룹 (그룹 0)	200 n_{00}	780 n_{01}
비흡연그룹 (그룹 1)	800 n_{10}	220 n_{11}
합계	1000	1000

- 오즈비 (odds ratio)

$$\psi = (n_{01} \times n_{10})/(n_{00} \times n_{11})$$

$$= (780 \times 800)/(200 \times 220) = 14.2$$

$$= \frac{\text{흡연그룹의 폐암 대 비폐암의 비율}}{\text{비흡연그룹의 폐암 대 비폐암의 비율}}$$

✓ 사례-대조 연구에서는 p_{ij}를 알 수 없으므로 비율의 차나 상대비율을 계산할 수 없다.

✓ 이 경우 오즈비가 상대비율의 추정치로 사용될 수 있다.

2.1 오즈비

■ 오즈비 (Odds Ratio)

	성공	실패	합
범주1	a	b	a+b
범주2	c	d	c+d
합	a+c	b+d	a+b+c+d

- 이차원 분할표에서 두 변수간 연관성을 나타내는 하나의 측도.
- 범주1의 오즈 = a/b : 범주1에서 성공 대 실패의 비
- 범주2의 오즈 = c/d : 범주2에서 성공 대 실패의 비
- 오즈비 = 범주1의 오즈/범주2의 오즈 = (a/b)/(c/d) = (ad)/(bc)

- 오즈비는 0에서 무한대의 값을 가짐.
- 1일 때 두 변수의 연관관계가 없다고 해석함.
- 0이나 무한대 값으로 갈수록 두 변수가 밀접한 연관관계 성립.
- 변수 내에서 범주의 순서를 바꾸어도 오즈비의 의미는 변하지 않는다.

2.2 R을 활용한 오즈비의 계산

■ 오즈비의 계산과 독립성 검정

```
Smoking = data.frame(Y=c(1,1,2,2),X=c(1,2,1,2),Count=c(780,220,200,800))
Smoking
Smoking.table = xtabs(Count~X+Y,data=Smoking)
addmargins(Smoking.table,margin=1)
fisher.test(Smoking.table,alternative="greater")
```

```
> addmargins(Smoking.table,margin=1)
     Y
X        1    2
  1    780  200
  2    220  800
  Sum 1000 1000
```

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

        Fisher's Exact Test for Count Data

data:  Smoking.table
p-value < 2.2e-16
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 11.76064      Inf
sample estimates:
odds ratio
  14.16288
```

$$H_0 : OR \le 1 \quad vs \quad H_1 : OR > 1$$

3. 범주 변수들간의 연관성 지표

3.1 카이제곱 통계량에 근거한 척도

- 파이 계수 (Phi Coefficient)

$$\phi = \frac{n_{11}\, n_{22} - n_{12}\, n_{21}}{\sqrt{n_{1.}\, n_{2.}\, n_{.1}\, n_{.2}}} \qquad \text{for } 2 \times 2 \text{ tables} \qquad \boxed{-1 \le \phi \le 1}$$

$$\phi = \sqrt{Q_P / n} \qquad \text{otherwise} \qquad \boxed{0 \le \phi \le min(\sqrt{R-1}, \sqrt{C-1})}$$

- 크래머 V (Cramer's V)

$$V = \phi \qquad \text{for } 2 \times 2 \text{ tables} \qquad \boxed{-1 \le V \le 1}$$

$$V = \sqrt{\frac{Q_P / n}{\min(R-1, C-1)}} \qquad \text{otherwise} \qquad \boxed{0 \le V \le 1}$$

- 분할계수 (Contingency Coefficient)

$$P = \sqrt{\frac{Q_P}{Q_P + n}} \qquad \boxed{0 \le P \le \sqrt{(m-1)/m}\ \text{, where } m = min(R, C)}$$

3.2 람다

✓ 2차원 분할표에서 행과 열의 결합정도를 수량화

- λ(C|R): 명목형 행과 열에 적용.
- 행 수준의 유무에 따라 열에 대한 "예측오류감소"계산.

	A	B	C	합계
30<	20 (20%)	20 (20%)	60 (60%)	100 (100%)
30≥	70 (35%)	100 (50%)	30 (15%)	200 (100%)
전체	90 (30%)	120 (40%)	90 (30%)	300 (100%)

$$\epsilon_1 = 90 + 0 + 90 = 180$$

$$\epsilon_2 = \{20 + 20 + 0\} + \{70 + 0 + 30\} = 140$$

$$\lambda = 1 - \frac{\epsilon_2}{\epsilon_1} = 1 - \frac{140}{180} = 0.222$$

3.2 람다

■ 람다

- 분할표에서 두 명목형 변수 사이의 연관도를 재는 측도. ($0 \le \lambda \le 1$)

$$\text{람다} = 1 - \frac{\text{행정보를 이용할 때 열정보 예측에서의 오류수}}{\text{행정보를 무시할 때 열정보 예측에서의 오류수}}$$

- $\lambda=1$
  - ✓ 행의 정보를 이용할 때 열의 정보를 예측함에 전혀 오류가 없다.
  - ✓ 행변수는 열 변수를 예측하는데 중요한 요인이다.
  - ✓ 두 변수간 큰 연관관계가 있다.

- $\lambda=0$
  - ✓ 열에 대한 정보를 예측하는데 행의 정보가 아무런 역할이 없다.
  - ✓ 두 변수간에 연관관계가 없다.

3.2 람다

■ **사례**

```
library(DescTools)
options(scipen=100)
Desc(Prefer.table,plotit=TRUE)
```

```
> Desc(Prefer.table,plotit=TRUE)
------------------------------------------------------------
Prefer.table (xtabs, table)

Summary:
n: 300, rows: 2, columns: 3

Pearson's Chi-squared test:
  X-squared = 65, df = 2, p-value = 0.000000000000007681
Likelihood Ratio:
  X-squared = 63.854, df = 2, p-value = 0.00000000000001362
Mantel-Haenszel Chi-squared:
  X-squared = 39.867, df = 1, p-value = 0.000000002719

Phi-Coefficient         0.465
Contingency Coeff.      0.422
Cramer's V              0.465
```

3.2 람다

■ 사례

```
Assocs(Prefer.table,conf.level=0.95)
```

```
> Assocs(Prefer.table,conf.level=0.95)
                            estimate  lwr.ci  upr.ci
Phi Coeff.                    0.4655       -       -
Contingency Coeff.            0.4220       -       -
Cramer V                      0.4655  0.3482  0.5754
Goodman Kruskal Gamma        -0.5676 -0.7223 -0.4128
Kendall Tau-b                -0.3447 -0.4502 -0.2391
Stuart Tau-c                 -0.3733 -0.4908 -0.2558
Somers D C|R                 -0.4200 -0.5478 -0.2922
Somers D R|C                 -0.2828 -0.3754 -0.1902
Pearson Correlation          -0.3651 -0.4594 -0.2628
Spearman Correlation         -0.3651 -0.4594 -0.2628
Lambda C|R                    0.2222  0.1363  0.3081
Lambda R|C                    0.3000  0.1444  0.4556
Lambda sym                    0.2500  0.1456  0.3544
Uncertainty Coeff. C|R        0.0977  0.0516  0.1439
Uncertainty Coeff. R|C        0.1672  0.0889  0.2455
Uncertainty Coeff. sym        0.1234  0.0654  0.1813
Mutual Information            0.1535       -       -
```

3.3 감마

γ : 순서형 행과 열에 적용.
행과 열 결합의 일치성·비일치성의 경우 수를 비교.

(예) 지위에 따른 업무 만족도

지위	만족도		합계
	낮음	높음	
하	20	10	30
상	5	25	30
합계	25	35	60

지위와 만족도 사이의
일치쌍(concordant pair)의 수 = 20x25 = 500
비일치쌍(discordant pair)의 수 = 5x10 = 50

$$\text{감마} = \frac{\text{일치쌍의 수 - 비일치쌍의 수}}{\text{일치쌍의 수 + 비일치쌍의 수}} = 450/550 = 0.82$$

3.3 감마

■ 감마(Gamma)

- 분할표에서 두 순서형 변수 사이의 연관도를 재는 측도. ($-1 \le \gamma \le 1$)

- 일치쌍 (concordant pair)

각 변수에 대한 관측값이 크기순서에서 같은 방향에 있는 한 쌍의 관측개체. (P)

- 불일치쌍 (discordant pair)

각 변수에 대한 관측값이 크기순서에서 반대방향에 있는 한 쌍의 관측개체. (Q)

$$\gamma = \frac{P-Q}{P+Q}$$

✓ 감마가 0에 가까울수록 두 범주간 연관관계가 없고, 1에 가까울수록 양(+)의 연관관계, -1에 가까울 수록 음(-)의 연관관계를 가진다.

✓ 2×2 분할표의 경우 오즈비(OR)와 양의 단조관계에 있음. $\gamma=(OR-1)/(OR+1)$

3.4 타우, D

■ **켄달의 타우-b**

- 같은 순위를 가지는 쌍의 경우에 대하여 감마를 수정한 것 (-1≤타우b≤1)

■ **스튜어트의 타우-c**

- 같은 순위를 가지는 쌍의 경우와 분할표의 크기를 이용하여 감마를 수정한 것 (-1≤타우c≤1)

■ **Somer's D**

- 켄달의 타우b를 비대칭적 관계를 고려하여 수정한 것

■ **Mantel-Haenszel 카이제곱 통계량** $\boxed{\mathrm{MH}\ \chi^2 = (n-1)\, r^2}$

- N은 관측개체수, r은 행과 열 변수의 각 수준에 적절한 숫자값을 부여하고 계산한 상관계수

3.5 사례

■ 사례 : Economic 데이터

Economic 데이터 (첫 10개, $n = 1,000$)

ID	Education	Income	ID	Education	Income
1	2	1	6	1	1
2	3	3	7	2	3
3	3	2	8	2	3
4	1	1	9	3	1
5	2	1	10	1	1

```
Economic = read.csv("data/Economic.csv",header=TRUE)
Economic$Education = factor(Economic$Education,labels=c("상","중","하"))
Economic$Income = factor(Economic$Income,labels=c("상","중","하"))
Economic.table = xtabs(~Education+Income,data=Economic)
Desc(Economic.table,plotit=TRUE)
```

3.5 사례

■ **사례 : Economic 데이터**

```
Pearson's Chi-squared test:
  X-squared = 54.254, df = 4, p-value = 0.0000000004656
Likelihood Ratio:
  X-squared = 54.972, df = 4, p-value = 0.0000000003293
Mantel-Haenszel Chi-squared:
  X-squared = 42.401, df = 1, p-value = 0.0000000007434

Phi-Coefficient           0.233
Contingency Coeff.        0.227
Cramer's V                0.165
```

3.5 사례

■ **사례 : Economic 데이터**

```
Assocs(Economic.table,conf.level=0.95)
```

```
> Assocs(Economic.table,conf.level=0.95)
                           estimate  lwr.ci  upr.ci
Phi Coeff.                   0.2329       -       -
Contingency Coeff.           0.2269       -       -
Cramer V                     0.1647  0.1154  0.2044
Goodman Kruskal Gamma        0.2926  0.2134  0.3718
Kendall Tau-b                0.1935  0.1395  0.2476
Stuart Tau-c                 0.1873  0.1351  0.2396
Somers D C|R                 0.1924  0.1387  0.2462
Somers D R|C                 0.1946  0.1403  0.2489
Pearson Correlation          0.2060  0.1459  0.2646
Spearman Correlation         0.2158  0.1559  0.2741
Lambda C|R                   0.0422  0.0000  0.0923
Lambda R|C                   0.0268  0.0000  0.0949
Lambda sym                   0.0344  0.0000  0.0844
Uncertainty Coeff. C|R       0.0259  0.0124  0.0394
Uncertainty Coeff. R|C       0.0256  0.0123  0.0389
Uncertainty Coeff. sym       0.0257  0.0124  0.0391
Mutual Information           0.0397       -       -
```

4. 심프슨의 패러독스

4.1 심프슨의 패러독스(Simpson's paradox)

- 분할표 분석에 있어 전체분석결과와 세부분석의 결과가 모순되는 현상

(예) Berkeley Admission Data, 1973

	합격	불합격	지원자 계
남자	1400 **(52%)**	1291 (48%)	2691 (100%)
여자	772 **(42%)**	1063 (58%)	1835 (100%)
전체	2172 (48%)	2354 (52%)	4526 (100%)

성차별 주장. 그러나….

	남		여	
분야	지원자	합격률	지원자	합격률
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	60%	341	70%

4.1 심프슨의 패러독스(Simpson's paradox)

약품	성공	실패	합계
A	20 (50%)	20 (50%)	40 (100%)
B	16 (40%)	24 (60%)	40 (100%)

A가 B보다 치료율에 있어 10% 포인트 우세한 것으로 보임.

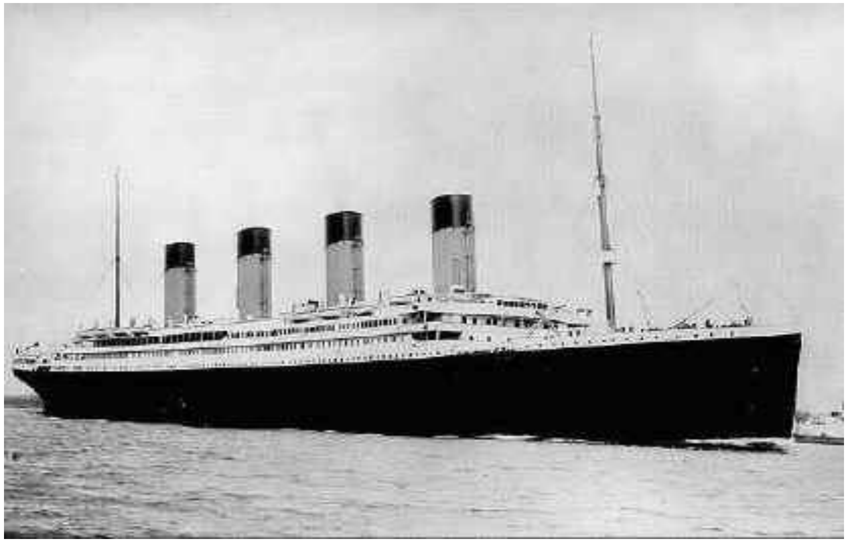
그러나 전체 사례를 중증도에 따라 분류하여 보면,

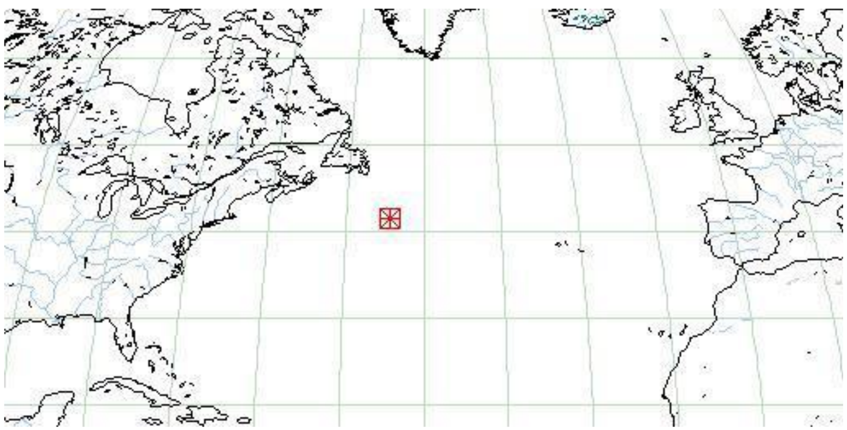
결론은 그 정반대로 나온다.

경증	성공	실패	합계
A	18 (60%)	12 (40%)	30 (100%)
B	7 (70%)	3 (30%)	10 (100%)

중증	성공	실패	합계
A	2 (20%)	8 (80%)	10 (100%)
B	9 (30%)	21 (70%)	30 (100%)

4.2 타이타닉 사례





"All the News That's Fit to Print."

The New York Times.

THE WEATHER.

TITANIC SINKS FOUR HOURS AFTER HITTING ICEBERG;
866 RESCUED BY CARPATHIA, PROBABLY 1250 PERISH;
ISMAY SAFE, MRS. ASTOR MAYBE, NOTED NAMES MISSING

Col. Astor and Bride, Isidor Straus and Wife, and Maj. Butt Aboard.

"RULE OF SEA" FOLLOWED

Women and Children Put Over in Lifeboats and Are Supposed to be Safe on Carpathia.

PICKED UP AFTER 8 HOURS

Vincent Astor Calls at White Star Office for News of His Father and Leaves Weeping.

FRANKLIN HOPEFUL ALL DAY

Manager of the Line Insisted Titanic Was Unsinkable Even After She Had Gone Down.

HEAD OF THE LINE ABOARD

Biggest Liner Plunges to the Bottom at 2:20 A. M.

RESCUERS THERE TOO LATE

Except to Pick Up the Few Hundreds Who Took to the Lifeboats.

WOMEN AND CHILDREN FIRST

Cunarder Carpathia Rushing to New York with the Survivors.

SEA SEARCH FOR OTHERS

The California Stands By on Chance of Picking Up Other Boats or Rafts.

OLYMPIC SENDS THE NEWS

The Lost Titanic Being Towed Out of Belfast Harbor.

PARTIAL LIST OF THE SAVED.

CAPT. E. J. SMITH.

4.2 타이타닉 사례

	생존자	사망자	생존율
1등실 성인 승객	197	122	61.7%
2등실 성인 승객	94	167	36.0%
3등실 성인 승객	151	476	24.1%
승무원	212	673	24.0%

4.2 타이타닉 사례

	생존자	사망자	생존율
1등실 성인 승객	197	122	61.7%
2등실 성인 승객	94	167	36.0%
3등실 성인 승객	151	476	24.1%
승무원	212	673	24.0%

	어린이				성인				생존율		여성 비율 %
	남자		여자		남자		여자		남자 %	여자 %	
	생존	사망	생존	사망	생존	사망	생존	사망			
1등실	5	0	1	0	57	118	140	4	32.6	97.2	45.0
2등실	11	0	13	0	14	154	80	13	8.3	86.0	35.0
3등실	13	35	14	17	75	387	76	89	16.2	46.1	26.3
승무원	0	0	0	0	192	670	20	3	22.3	86.9	2.6

5. 참고

참고 A. 피셔의 정확검정법

■ 피셔의 정확 검정법 (Fisher's Exact Test)

- **예제 : 환자 10명의 처리와 반응간 관계**

10명의 환자들이 각각 처리 1과 처리 2를 받고 난 후, 다음 표와 같이 반응 또는 무반응의 결과를 보였다. 처리와 반응간에 유의한 연관성이 있는지 알아보자

[환자 10명의 처리와 반응간 관계]

	반응	무반응	합
처리 1	1	4	5
처리 2	3	2	5
합	4	6	10

고정

	반응	무반응	합
처리 1	0	5	5
처리 2	4	1	5
합	4	6	10

고정

귀무가설 : 처리와 반응 간에는 아무런 연관이 없다.
대립가설 : 처리2에서의 반응율이 처리1에서의 반응율 보다 높다.

참고 A. 피셔의 정확검정법

■ 피셔의 정확 검정법 (Fisher's Exact Test)

▪ 예제 : 환자 10명의 처리와 반응간 관계

- 처리 1의 반응에 대한 도수만 정해지면 나머지 도수는 자동적으로 정해짐
- 귀무가설 하에서 처리 1의 반응에 대한 도수가 1이 되는 확률

$$P\{n_{11} = 1\} = \frac{\binom{4}{1}\binom{6}{4}}{\binom{10}{5}} = 0.238$$

- 보다 극단적인 분할표

$$P\{n_{11} = 0\} = \frac{\binom{4}{0}\binom{6}{5}}{\binom{10}{5}} = 0.024$$

- 정확한 p-값 = 0.238 + 0.024 = 0.262 가 됨

=> 유의수준 0.05에서 귀무가설을 기각할 수 없음

참고 A. 피셔의 정확검정법

■ 피셔의 정확 검정법 (Fisher's Exact Test)

- 예제 : 환자 10명의 처리와 반응간 관계

```
trt<-factor(c(1,1,2,2))
response<-factor(c("Y","N","Y","N"), levels=c("Y","N"))
count <-c(1,4,3,2)
mytable=xtabs(count~trt+response)
fisher.test(mytable, alternative="less")
```

```
        Fisher's Exact Test for Count Data

data:  mytable
p-value = 0.2619
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.00000 3.17237
sample estimates:
odds ratio
 0.2033268
```

참고 B. 맥니머 검정(McNemar Test)

■ 맥니머 검정 (McNemar's Test)

- 연속형 변수의 분석에서 짝지은 t검정이 있었던 것처럼, 도수를 구한 데이터에서도 짝지은 표본 (대응표본)에 대해서 분석방법을 달리 고려
- 한 환자에 대해서 처리 전후의 반응을 보는 것임
- 이는 한 개체에 대해 처리 전,후를 비교하는 문제이기에 앞의 카이제곱 검정과는 다름 (독립이 아니므로)

[짝지은 범주형 자료의 형태]

조건 1	조건 2		합
	예	아니오	
예	n_{11}	n_{12}	n_{1+}
아니오	n_{21}	n_{22}	n_{2+}
합	n_{+1}	n_{+2}	n

참고 B. 맥니머 검정(McNemar Test)

■ 맥니머 검정 (McNemar's Test)

- 조건1과 조건2에서의 반응률이 같은지를 알고 싶음
- 가설 $H_0 : p_1 = p_2 \quad vs. \quad H_1 : p_1 \neq p_2$

  여기서 조건 1에서의 반응률 p_1, 조건 2에서의 반응률 p_2 각각의 추정량은

$$\hat{p}_1 = \frac{n_{1+}}{n}, \quad \hat{p}_2 = \frac{n_{+1}}{n}$$

- 맥니마 검정은 두 조건에서 일치한 쌍의 개수는 고려하지 않고 일치하지 않은 쌍의 개수에 근거하여 통계량을 구함. 즉, $n_{12} + n_{21}$ 이 고정되어 있을 때(일정할 때), n_{12} 는

  이항분포 $B(n_{12} + n_{21}, \frac{1}{2})$ 를 따름을 이용.

- 검정통계량(exact): $n_{12} \sim B(n_{12} + n_{21}, \frac{1}{2})$
- 검정통계량(approximate): 맥니머 카이제곱 검정

$$Q_M = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})} \approx \chi^2(1) \qquad Q_M = \frac{(|n_{12} - n_{21}| - 1)^2}{(n_{12} + n_{21})} \quad \text{(연속성보정)}$$

참고 B. 맥니머 검정(McNemar Test)

■ 맥니머 검정 (McNemar's Test)

▪ **예제 : 결혼 전후 경제생활 만족도 관계**

다음 표의 자료는 결혼한 지 3년이 된 남성 60명을 대상으로 결혼 전후의 경제생활의 만족도를 조사한 자료이다. 이 자료에 대해 결혼 전과 결혼 후의 경제생활의 만족도에 있어서 차이가 있는지 알아보자.

[결혼 전후 경제생활 만족도 자료]

결혼 전	결혼 후		합
	만족	불만족	
만족	23	7	30
불만족	18	12	30
합	41	19	60

$$Q_M = \frac{(n_{12}-n_{21})^2}{(n_{12}+n_{21})} = \frac{(7-18)^2}{7+18} = 4.84 > \chi^2(1, 0.05) = 3.84$$

- 유의수준 0.05에서 귀무가설을 기각함

참고 B. 맥니머 검정(McNemar Test)

■ 맥니머 검정 (McNemar's Test)

- **예제: 결혼 전후 경제생활 만족도 관계**

```
pre <-factor(c("Y", "Y", "N", "N"), levels=c("Y","N"))
post <-factor(c("Y","N","Y","N"), levels=c("Y","N"))
count <-c(23,7,18,12)
mytable=xtabs(count~pre+post)
mcnemar.test(mytable)
```

```
        McNemar's Chi-squared test with continuity correction

data:  mytable
McNemar's chi-squared = 4, df = 1, p-value = 0.0455
```

```
        Fisher's Exact Test for Count Data

data:  mytable
p-value = 0.2668
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.6301791 7.9380709
sample estimates:
odds ratio
  2.161641
```