범주형 자료분석

장원철

서울대학교 통계연구소

이번 강좌에서 다룰 내용

- $r \times c$ 분할표
  - χ^2 검정
  - Cochran-Armitage 검정
- 2×2 분할표
  - 역학설계연구
  - Fisher's exact test
  - McNemar test for paired data
- 층화 분석
  - Mantel-Haenszel (combined) OR estimate
  - Mantel-Haenszel test for association
  - Breslow-Day test for Homegeneity

시작하기 전에...

- 수업시간에 사용할 R script와 자료를 통계연구소 강좌자료실에서 download 받고, working directory 설정하기.
- R 패키지 `tibble`, `ISwR`, `DescTools`, `ggplot2`를 설치하자.

```
install.packages("tibble") # install R pacakge tibble for data input
install.packages("ISwR") # install R package ISwR for a data set
install.packages("DescTools") # for Descriptive Statistics and simple test proce
install.packages("ggplot2") # Graphing supplement

library(tibble)
library(DescTools)
library(ISwR)
library(ggplot2)
```

Windows에서 한글 깨짐 현상 해결방법

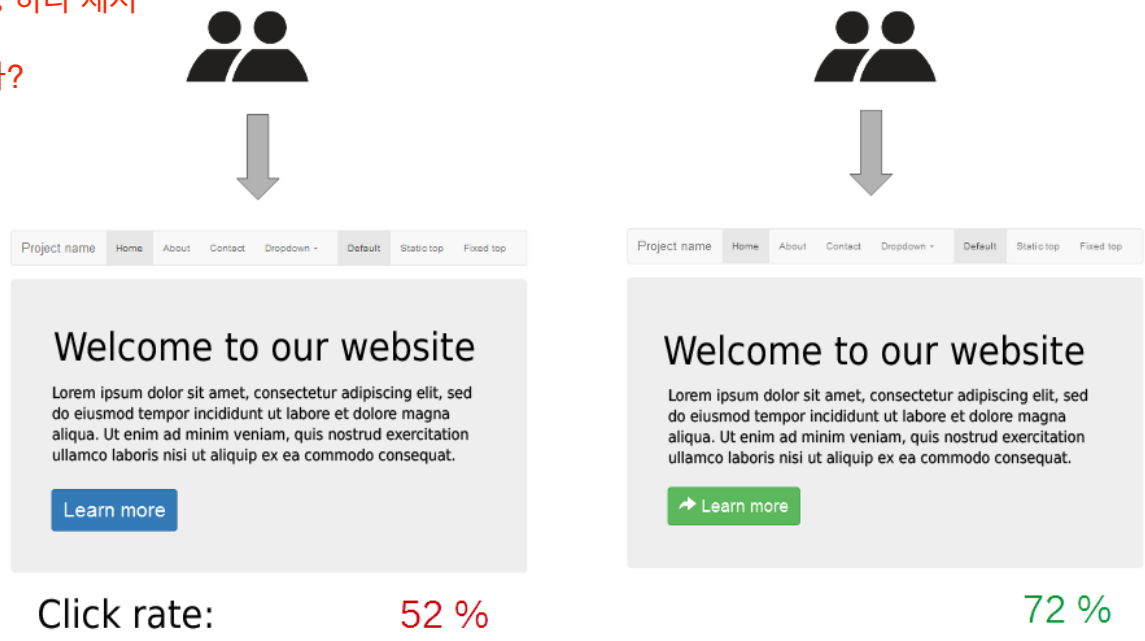
RStudio에서

1. Tools > Global Options > Code > Saving > Default text encoding을 UTF-8으로 변경하고 Apply
2. CDAs.R 파일을 RStudio에 불러온 후에 File > Save with Encoding > UTF-8을 선택하고 아래 Set as default encoding for source files를 체크
3. RStudio에서 CDAs.R를 닫고 CDAs.R 를 다시 불러오기

A/B Testing

- Website에 접근하는 사람에게 A, B 두 포맷 중 임의의 선택을 보여주고 클릭을 하는 비율을 비교한다.

랜덤하게 두 개의 디자인 중 하나 제시
-> 뭐가 더 관심을 끄나?
클릭한 비율이 뭐가 더 높나?



source:

https://commons.wikimedia.org/wiki/File:A-B_testing_simple_example.png

질문: B가 A보다 나은가? 또는 $P($ 클릭 $|A)$ 와 $P($ 클릭 $|B)$ 이 같은가, 다른가?

A/B Testing의 자료(dataset)

- A/B testing을 100 명에게 진행한 결과 다음과 같은 구조의 자료를 생성할 수 있다.

Individual	Format	Click
id1	A	Yes
id2	A	No
id3	B	No
id4	A	Yes
⋮	⋮	⋮

2 * 2 Table

- 각 행은 사람(individual)에 해당하며, 각 열은 변수에 해당한다.
- 변수 Format과 Click은 자료의 형태가 범주형(Categorical)이다.

타이타닉 자료

- 여기에 소개하는 자료는 호화여객선 타이타닉에 탑승한 1309 명들에 대한 기록으로 개인별로 14개의 변수가 있다.
- 개인별 변수는 pclass(좌석등급), survived (1: 생존, 0:사망), name (이름), sex(성별), age (나이), sibsp (동승한 형제자매/배우자 수), parch (동승한 부모/자녀 숫자), ticket (티켓번호), fare(티켓가격), cabin (방번호), embarked (승선항), boat(구명보트 번호), body (사망자 인식번호), home.dest (목적지) 로 구성되어 있다.
- 다음과 같이 자료를 살펴본다.

```
titan <-read.csv("titanic.csv")
View(titan)
```

분할표 작성

- 범주형 자료는 일반적으로 다음 3가지 경우 중 하나의 형태를 이용하여 주어진다
  - 개인별 raw data가 주어졌지만 (대부분의) 변수들이 범주형 자료인 경우
  - 자료자체가 분할표로 주어진 경우
  - 데이타 테이블의 각 행이 분할표 각 Cell의 빈도수로 이루어진 경우
- 어떤 형태로 자료가 주어지던지 분할표를 작성할 수 있으며, 분할표를 이용해 원 자료를 다시 구할 수 있다.

분할표 작성: 개인별 자료 → 분할표

- 타이타닉 자료의 성별(sex)과 생존여부(survived)를 분할표로 만들자
- xtabs command 사용하여 table 형태의 자료로 정리
- 문법:

```
xtabs( ~  행 변수 + 열 변수, data = data.frame)
```

```
# Use xtabs ("cross-tabulation") to creat a contingency table
xtabs(~ sex + survived, data=titan)

##          survived
## sex         0   1
##   female  127 339
##   male    682 161
```

분할표 직접입력

- 자료가 분할표로 직접주어진 경우를 고려해보자.
- 앞장 슬라이드의 타이타닉 자료 분할표를 직접 입력한다면

```
# A direct, through error-prone, method of manually creating a contingency table
tab_3 <- matrix(c(127,682,339,161),nrow=2)
dimnames(tab_3) <- list(sex=c("Female","Male"), survived=c("No","Yes"))
tab_3

##          survived
## sex        No Yes
##   Female  127 339
##   Male    682 161
```

빈도수 자료 직접입력 후 분할표 변환

- `tibble::tribble` 함수를 이용하여 간단한 자료를 직접 입력할 수 있다.

```
titan_2 <- tribble(
    ~sex, ~survived, ~Freq,
    # --- / -- / ---
    "Female", "No", 127,
    "Female", "Yes", 339,
    "Male", "No", 682,
    "Male", "Yes", 161,
)
```

- 이 때, `xtabs`의 문법은

```
xtabs(빈도 ~ 행 변수 + 열 변수, data =  빈도수자료 )
```

```
(tab_2 <- xtabs(Freq ~ sex + survived, data = titan_2))

##          survived
## sex        No Yes
##   Female 127 339
##   Male   682 161
```

자료 변환

- 개인별 자료 → 분할표

```
tab_1<-xtabs(~ sex + survived, data=titan)
```

- 분할표 → 개인별 자료

```
## Using DescTools::Untable ; use head() to show first few lines
head( Untable(tab_1) )
```

```
##      sex survived
## 1 female        0
## 2 female        0
## 3 female        0
## 4 female        0
## 5 female        0
## 6 female        0
```

- 분할표 → 빈도 수 자료

```
data.frame(tab_1)
```

```
##      sex survived Freq
## 1 female        0  127
## 2   male        0  682
## 3 female        1  339
## 4   male        1  161
```

연습문제 1

1. 심근경색과 혈압과의 관계를 알아 보고자 한다. 다음 파일 (`MIdata.csv`)에서 자료를 읽고 혈압(SBPgt140: 수축기 혈압 140 이상) 과 심근경색 (MI) 유무의 결과는 분할표로 작성하라.
2. 1에서 작성한 table을 `matrix()` 함수를 이용하여 직접 입력하라.
3. 빈도 수 자료를 `tribble()` 함수를 이용하여 직접 입력 후 분할표로 변환하라

연습문제 1

```
MI <- read.csv("MIdata.csv")
View(MI)
MI.tab1 <-xtabs(~SBPgt140+MI, data=MI)
MI.tab1

##          MI
## SBPgt140    0    1
##        0 1244   27
##        1  711   29

MI.tab2 <-matrix(c(1244,711,27,29), nrow=2)
dimnames(MI.tab2)<-list(SBPgt140 = c("저혈압" , "고혈압"),
                        MI =c("건강" , "심장마비" ) )
MI.tab2

##          MI
## SBPgt140 건강 심장마비
##   저혈압 1244       27
##   고혈압  711       29
```

연습문제 1

```
Freq_dat <- tribble(
   ~SBPgt140, ~MI, ~Freq,
   "저혈압",   "건강",  1244,
   "저혈압",   "심장마비",    27,
   "고혈압",   "건강",   711,
   "고혈압",   "심장마비",    29)
MI.tab3 <- xtabs(Freq ~ SBPgt140 + MI, data = Freq_dat)
MI.tab3

##          MI
## SBPgt140 건강 심장마비
##   고혈압  711       29
##   저혈압 1244       27
```

분할표의 기술통계량

- 타이타닉 선실등급과 생존유무

```
tab <- xtabs( ~ survived + pclass,  data = titan)
tab
```

```
##          pclass
## survived 1st 2nd 3rd
##        0 123 158 528
##        1 200 119 181
```

관심: 분할표에 기록된 각 선실 등급별 인원수는? (주변합과 주변분포)
관심: 선실 등급별 생존률은? (상대빈도)

분할표의 기술통계량: 주변합

- 주변합: `margin.table()`로 분할표의 주변 (margin) 여백을 채운다.

```
margin.table(tab, margin = 2)
```

```
## pclass
## 1st 2nd 3rd
## 323 277 709
```

- 행별(row-wise) 주변합은 `margin = 1`,
- 열별(column-wise) 주변합은 `margin = 2`.
- 전체 주변 여백을 채운 분할표는 `addmargins()`

```
addmargins(tab)
```

```
##          pclass
## survived  1st  2nd  3rd  Sum
##      0    123  158  528  809
##      1    200  119  181  500
##      Sum  323  277  709 1309
```

분할표의 기술통계량: 상대빈도표

- 기준이 되는 변수(행 또는 열)를 지정한다
- 열(margin = 2)에 있는 변수인 선실 등급(pclass)별로 생존여부(survived)의 상대빈도를 구하면:

```
prop.table(tab, margin = 2)
```

```
##          pclass
## survived    1st    2nd    3rd
##        0 0.3808 0.5704 0.7447
##        1 0.6192 0.4296 0.2553
```

- 전체 탑승인원수를 기준으로 상대빈도를 구하면:

```
prop.table(tab)
```

```
##          pclass
## survived     1st     2nd     3rd
##        0 0.09396 0.12070 0.40336
##        1 0.15279 0.09091 0.13827
```

$r \times c$ 분할표 분석

- 동일성과 독립성 검정
- Trend 검정

예제: 교육수준과 임상실험

HIV 예방접종의 참여도와 교육수준의 관련성에 대해 알고자 한다. 4850명을 대상으로 "만약 내일 HIV 백신연구가 시작된다면 연구에 참여하겠는가?"라는 질문을 한 후 교육수준별 응답을 다음과 같이 정리하였다.

	절대 안함	아마도 안함	아마도 함	반드시 함	합계
고교 중퇴	52	79	342	226	699
	7.4%	11.3 %	48.9%	32.3%	
고졸	62	153	417	262	894
	6.9%	17.1%	46.6%	29.3%	
대학 중퇴	53	213	629	375	1270
	4.2%	16.8%	49.5%	29.5%	
대졸	54	231	571	244	1100
	4.9%	21.0%	51.9%	22.2%	
대학원 중퇴	18	46	139	74	277
	6.5%	16.6%	50.2%	26.7%	
대학원 졸	25	139	330	116	610
	4.1%	22.8%	54.1%	19.0 %	
합계	264	861	2428	1297	4850
	5.4%	17.8%	50.1%	26.7%	

관심: 교육수준과 참여도가 서로 독립적인가, 아니면 관련이 있는가? (독립성 검정)

예제: 폐암과 흡연 (Doll and Hill, 1952)

영국의사들을 상대로 한 후행적(restropective) 연구를 통해 흡연과 폐암과의 관련성에 대해 알고자 한다.

	일일 흡연량						
	0	< 5	$5-14$	$15-24$	$25-49$	$50+$	합계
폐암환자	7	55	489	475	293	38	1357
	0.5%	4.1%	36.0 %	35.0%	21.6%	2.8 %	
대조군	61	129	570	431	154	12	1357
	4.5%	9.5%	42.0%	31.8%	11.3%	0.9%	
합계	68	184	1059	906	447	50	2714

질문: *이 자료*에 대해서 일반적인 독립성 검정이 바람직하지 않은 이유는?

Table 형태 자료입력

```r
# 예제: 교육수준과 임상실험
hiv.edu <-matrix(
  c(52,62,53,54,18,25,79,153,213,231,46,139,342,417,629,
     571,139,330,226,262,375,244,74,116),
  nrow=6
  )
dimnames(hiv.edu) <-list(
  education=c("고교중퇴","고졸","대학중퇴",
              "대졸","대학원중퇴","대학원졸"),
           참여도 =c("절대안함", "아마도 안함", "아마도 함","반드시 함")
)

# 예제 : 폐암과 흡연
lung.cancer <-matrix(c(7,61,55,129,489,570,475,431,293,154,38,12),
                     nrow=2)
dimnames(lung.cancer)<-list(
      group=c("case","control"),
      smoking=c("0","<5","5-14","15-24","25-49","50+"))
```

기댓값의 계산

기댓값: 귀무가설이 참이라 가정하고, 단순 확률 곱으로 구함

- O_{ij}와 E_{ij}는 i번째 열과 j번째 행의 관측치와 기댓값을 나타낸다고 하자.
- 첫번째 예제에서 귀무가설은 "교육수준과 임상실험의 참여도가 관련이 없다"
- 즉 귀무가설은 교육수준과 참여도가 서로 독립이다.
- 만약 귀무가설이 참이라면 E_{ij}는 어떻게 계산할까?
- $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$ if A and B are indep.
- 따라서

$$E_{ij} = \text{전체자료의 수} \times \Pr(\text{교육수준이 } i\text{번째이고 참여여부가 } j\text{번째일때의 경우})$$

로 생각할 수 있고 귀무가설하에서

$$\Pr(\text{교육수준이 } i\text{번째이고 참여여부가 } j\text{번째일때의 경우}) = \Pr(\text{교육수준이 } i\text{번째}) \cdot \Pr(\text{참여여부가 } j\text{번째})$$

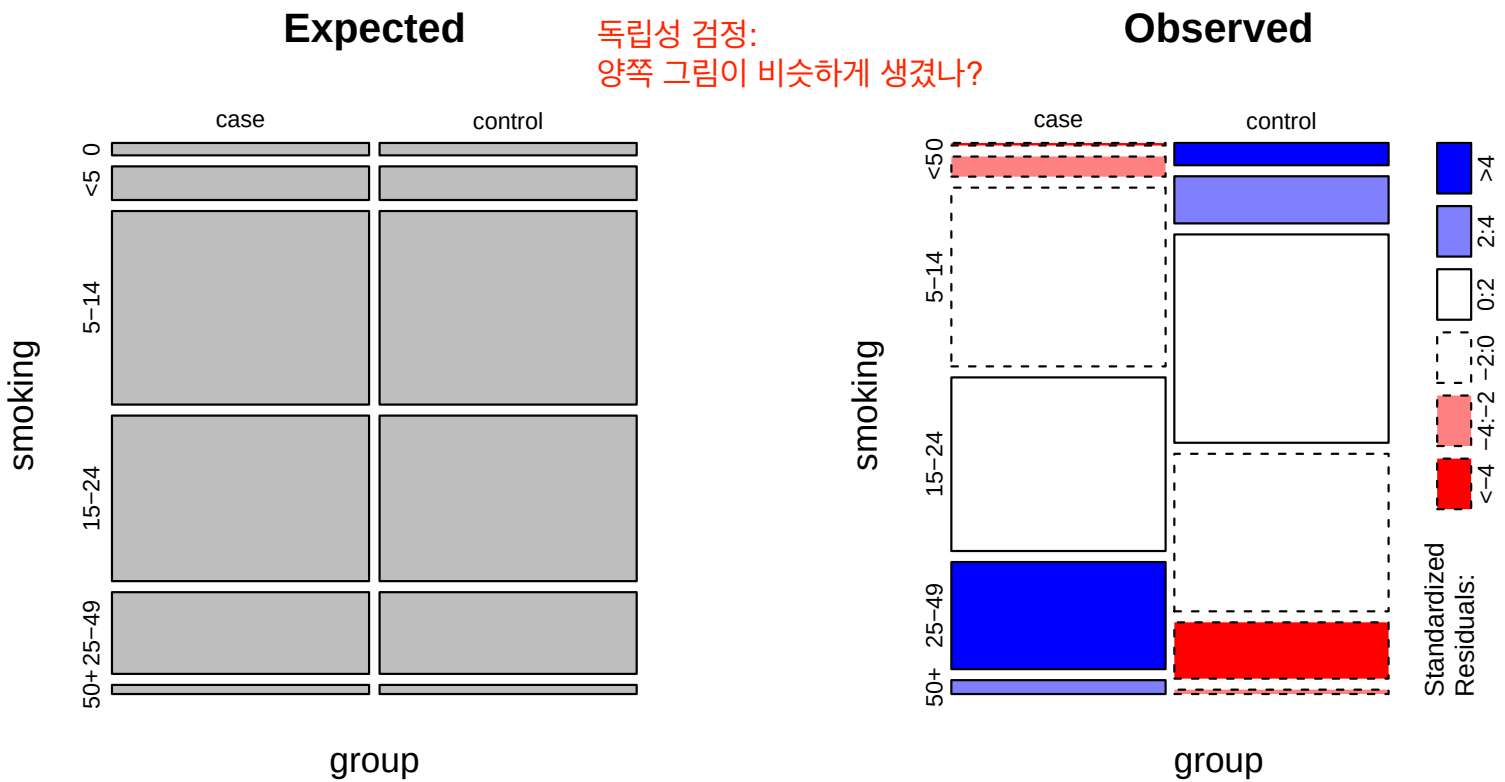
기댓값의 계산

	일일 흡연량						
	0	< 5	$5-14$	$15-24$	$25-49$	50+	합계
폐암환자	7	55	489	475	293	38	1357
	0.5%	4.1%	36.0 %	35.0%	21.6%	2.8 %	
대조군	61	129	570	431	154	12	1357
	4.5%	9.5%	42.0%	31.8%	11.3%	0.9%	
합계	68	184	1059	906	447	50	2714

$$
\begin{aligned}
E_{13} &= 2714 \times \Pr(\text{일일흡연량이 5-14개비인 폐암환자}) \\
&= 2714 \times \Pr(\text{폐암환자}) \times \Pr(\text{일일흡연량이 5-14개비 }) \\
&= 2714 \times \frac{1357}{2714} \times \frac{1059}{2714} = 529.5
\end{aligned}
$$

기댓값의 이해

```r
lung.cancer.e <- chisq.test(lung.cancer)$expected # 기댓값 계산
par(mfrow=c(1,2))
mosaicplot(lung.cancer.e, main = "Expected")
mosaicplot(lung.cancer, shade = TRUE, main = "Observed")
```



χ^2 검정

표준화

$r \times c$ 분할표에서 두 변주형 자료의 독립성/동일성 검정

- 귀무가설이 참이라면 O_{ij}와 E_{ij}의 값이 비슷할 것으로 짐작할 수 있으므로 검정통계량은 실제 관측치와 기댓값의 차이에 기반해야 함을 알 수 있다.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2\left((r-1)\times(c-1)\right)$$

- 백신연구의 경우 참여도와 교육수준과 관련이 있는지 여부에 관한 검정 (독립성)
- 폐암연구의 경우 환자군과 대조군에서 흡연량의 분포가 같은지 여부에 관한 검정 (동일성)

동일성과 독립성 검정 결과

```
chisq.test(lung.cancer)
```

```
##
##  Pearson's Chi-squared test
##
## data:  lung.cancer
## X-squared = 137.7, df = 5, p-value < 2.2e-16
```

폐암연구에서 환자군과 대조군의 흡연량의 분포는 다르다고 할 수 있다. (p-value $< 2.2 \cdot 10^{-16}$)

```
chisq.test(hiv.edu)
```

```
##
##  Pearson's Chi-squared test
##
## data:  hiv.edu
## X-squared = 89.72, df = 15, p-value = 1.117e-12
```

결과해석: 백신연구에서 참여도와 교육수준과는 관련이 있어 보인다. (p-value $= 1.117 \cdot 10^{-12}$)

순서가 있는 범주형 자료

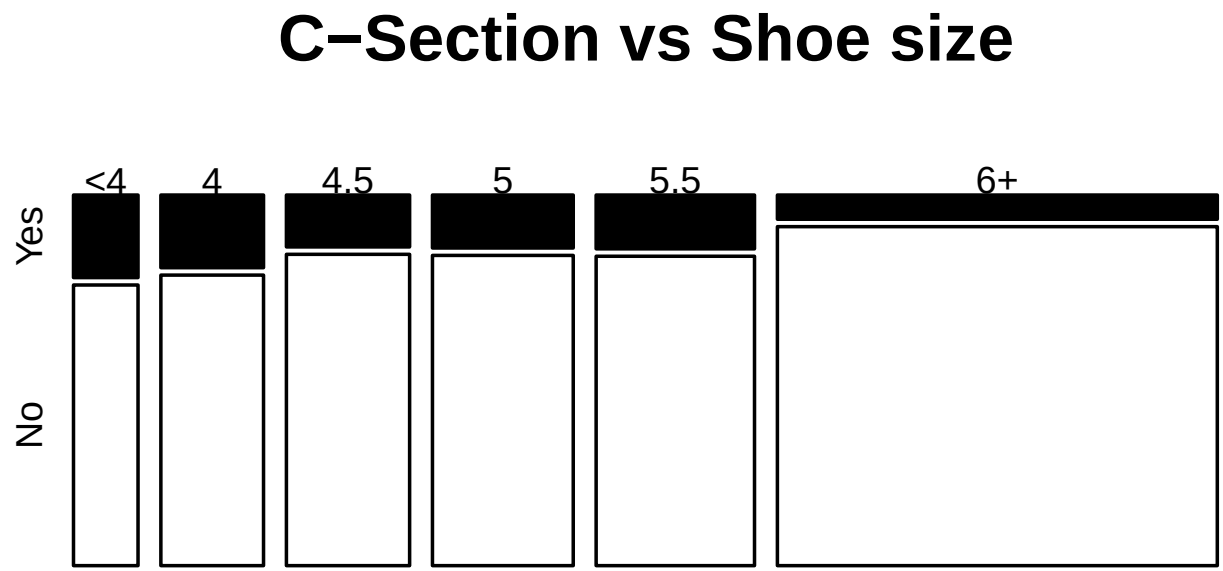
- 첫번째 예제의 교육수준과 두번째 예제의 하루 흡연량은 순서가 있는 범주형 자료이다.
- 이런 경우 우리가 알고 싶은 결론은 조금 더 구체적일 수 있다. 예를 들면 "담배를 많이 필수록 폐암에 걸릴 가능성이 높다".
- 대립가설이 이와 같이 보다 구체적일 경우 실제로 test의 검정력 (power)가 더 좋아질 수 있다.

예제: 신발 사이즈와 제왕절개수술 여부

임신부의 발크기가 제왕절개수술 여부와 관련이 있는지 알고자 한다. (Altman, 1991, p.229)

Table: 신발사이즈 vs 제왕절개여부

C-section	< 4	4	4.5	5	5.5	6+
Yes	5	7	6	7	8	10
No	17	28	36	41	46	140



Cochran-Armitage 검정

- $2 \times k$ 분할표에서 Trend가 있는지 여부를 검정하고 싶을 경우
- Test statistic: Cochran-Armitage test statistic
- 귀무가설하에서 $\chi^2(1)$을 따른다!

예제: 신발 사이즈와 제왕절개수술 여부

```
ISwR::caesar.shoe
```

```
##     <4  4 4.5  5 5.5  6+
## Yes  5  7   6  7   8  10
## No  17 28  36 41  46 140
```

```
(caesar.shoe.yes <- caesar.shoe["Yes", ])
```

```
##  <4   4 4.5   5 5.5  6+
##   5   7   6   7   8  10
```

```
(caesar.shoe.total <- margin.table(caesar.shoe, 2))
```

```
##  <4   4 4.5   5 5.5  6+
##  22  35  42  48  54 150
```

Cochran-Armitage 검정

χ^2를 사용할 경우 여기서 귀무가설과 대립가설은 다음과 같다.

$$H_0 : p_1 = p_2 = p_3 = p_3 = p_4 = p_5 = p_6$$
$$H_a : \text{모든 집단에서 비율들이 같은 것은 아니다}$$

```
chisq.test(caesar.shoe)

## Warning in chisq.test(caesar.shoe): Chi-squared approximation may be
incorrect

##
##  Pearson's Chi-squared test
##
## data:  caesar.shoe
## X-squared = 9.3, df = 5, p-value = 0.1
```

기댓값이 낮은 셀이 많이 존재할 경우, 카이 사용 권장 X

Cochran-Armitage 트렌드 검정

만약 다음과 같은 대립가설를 가정한다면 좀더 Powerful한 검정을 할 수 있다.

자유도가 낮기 때문에,, p 밸류가 낮아진다.

$$H_a : p_1 \leq p_2 \leq p_3 \leq \ldots \leq p_k \quad (or \geq)$$

```
prop.trend.test(caesar.shoe.yes, caesar.shoe.total)

##
##  Chi-squared Test for Trend in Proportions
##
## data:  caesar.shoe.yes out of caesar.shoe.total ,
##  using scores: 1 2 3 4 5 6
## X-squared = 8, df = 1, p-value = 0.005
```

발이 클수록 제왕절개 수술비율이 낮다 (p.value=0.005)

Cochran-Armitage 검정

- 일반적 χ^2 검정의 경우 자유도가 5이기때문에 p-value가 0.10 이었다.
- Cochran-Armitage test의 경우 대립가설을 보다 구체적으로 함으로써 검정통계량은 약간의 차이가 생기지만 자유도는 항상 1이므로 p-value의 값이 낮아진다.
- Cochran-Armitage test를 사용할 경우 자료의 입력이 일반적인 table의 형태가 아니라는 점을 기억하자.

분할표 정돈하기

- 가끔 분할표가 원하는 순서로 정돈되어 있지 않을 수가 있다.
- 예를 들어 다음 타이타닉 자료의 embarked변수의 값은 알파벳 순으로 자동 정렬되어 있다

```
(tab <- xtabs( ~ survived + embarked, data = titan))
```

```
##          embarked
## survived     Cherbourg Queenstown Southampton
##        0   0       120         79         610
##        1   2       150         44         304
```

분할표 정돈하기

- 먼저 빈 값을 지운다.

```
(tab_mod1 <- tab[,2:4]) # 행은 전부(빈칸으로 놔둠), 열은 2-4번째만
```

```
##          embarked
## survived Cherbourg Queenstown Southampton
##        0       120          79         610
##        1       150          44         304
```

- 탑승 순서(Southampton, Cherbourg, Queenstown)대로 정렬하자.

```
(tab_mod2 <- tab_mod1[,c(3,1,2)]) #3열이 첫번째, 1열이 두번째, 2열이 세번째 열이 된다
```

```
##          embarked
## survived Southampton Cherbourg Queenstown
##        0         610       120          79
##        1         304       150          44
```

연습문제 2

흡연여부와 본인이 생각하는 건강상태와의 관련을 알고자 한다.

Table: Self Report Quality of Health

Smoke	Poor	Fair	Good	V. Good	Exc.	Total
No	11	27	42	53	11	144
Yes	7	15	16	13	1	52

1. `Exercise2dat.csv`에서 위의 자료를 읽고,
2. χ^2검정을 이용하여 흡연과 본인의 건강상태인지와 관련이 있는지 검정해보아라.
3. Cochran-Armitage 검정을 이용하여 위의 귀무가설을 검정해보아라.
4. `mosaicplot()` 을 이용하여 분할표를 시각화하라.

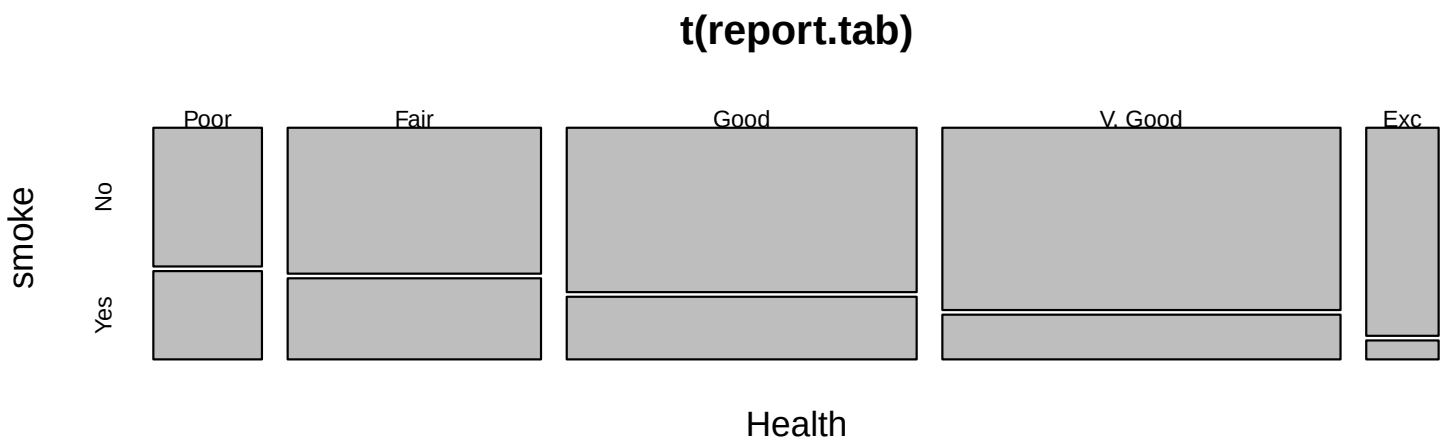
연습문제 2

```
dat<-read.csv("Exercise2dat.csv")
report.tab <- xtabs(~ smoke + Health, data = dat)
report.tab <- report.tab[,c(4,2,3,5,1)]

chisq.test(report.tab)

##
##  Pearson's Chi-squared test
##
## data:  report.tab
## X-squared = 6.9, df = 4, p-value = 0.1

report.smoke <-report.tab["Yes",]
report.total <-margin.table(report.tab,2)
prop.trend.test(report.smoke, report.total)

##
##  Chi-squared Test for Trend in Proportions
##
## data:  report.smoke out of report.total ,
##  using scores: 1 2 3 4 5
## X-squared = 6.7, df = 1, p-value = 0.01
```

연습문제 2

자신의 건강이 좋다고 생각할 수록 흡연의 비율이 낮은 경향이 있다. (p.value=0.01)

```
mosaicplot(t(report.tab))
```



2×2 분할표

- 역학연구설계
  - 코호트 연구 현재 시점 데이터 모아 -> 앞으로 나가는...
  - 사례-대조군 연구
  - 단면조사 연구
- Association measures in 2×2 분할표
  - Odds Ratio (OR)
  - Relative Risk (RR)
  - Risk Difference (RD)
- Fisher's exact test
- McNemar test for paired data

2×2 분할표; 코호트 연구 (Pauling, 1971)

텍스트

환자들을 임의로 두 그룹으로 나누어 비타민 C과 위약 (placebo) 을 주고 각 그룹에서 얼마나 많은 감기환자가 발생했는지 살펴본다.

Cold	Vitamin C	Placebo	Total
Yes	17	31	48
No	122	109	231
Total	139	140	279

관심: 비타민 C가 감기예방에 도움이 되는가?

역학연구 설계: 코호트 (cohort) 연구

비타민 복용 시 감기 걸릴 위험 ~ 복용하지 않았을 때 감기 걸릴 위험

특정위험요인에 노출된 집단과 노출되지 않은 집단의 질병 발생률을 비교 분석하여 위험요인과 질병사이의 관련성을 조사

- 위험도 (Risk):
  $p_1 = P(\text{ 감기 } | NoVitC)$ (요인에 노출된 집단의 발병)
  $p_2 = P(\text{ 감기 } | VitC)$ (요인에 노출되지 않은 집단의 발병율)
- 오즈 (Odds): $O_1 = p_1/(1-p_1)$, $O_2 = p_2/(1-p_2)$
- 상대위험도 (Relative Risk): $RR = p_1/p_2$
- 오즈비 (Odds Ratio) $OR = O_1/O_2$
- 만약 $OR > 1$ 또는 $RR > 1$이면, 비타민 C를 복용하지 않을 때 감기에 걸릴 위험이 높다고 본다.

귀무가설: $p_1 = p_2$, $RR = 1$, 또는 $OR = 1$. **관심**: 귀무가설이 기각된다면, 비타민 C가 위험도를 얼마나 낮춰주는가 (RR의 추정)

2×2 분할표; 사례-대조군 연구 (Keller, 1965)

구강암과 흡연의 관계를 알기 위해 사례군(구강암환자)을 과대표집하였다.

	Smoker	Nonsmoker	Total
Case	484	27	511
Control	385	90	475
Total	869	117	986

질문: 흡연이 구강암과 관련이 있는가? 그렇다면 흡연자가 구강암에 걸릴 가능성을 비흡연자와 비교한다면 얼마나 높은가?

역학연구 설계: 사례-대조군 (case-control) 연구

환자와 대조군 각각에서 과거에 위험요인에 노출된 정도를 비교하여 질병과 위험요인과의 관련성을 알고자 한다.

- 사례군이 과대표집되어 위험도 추정이 불가!
  $p_1 = P(\text{ 암 } | Smoker) \neq 484/869$
  $p_2 = P(\text{ 암 } | NonSmoker) \neq 27/117$
- 여기서 계산된 오즈비(Odds Ratio)만이 코호트 연구에서 계산되는 오즈비와 수학적으로 동일하다.
- 위험도와 상대위험도를 계산할 수 없으므로 오즈비만 유일한 관련성 측도이다.
- 희귀한 질병에 대해서 오즈비와 상대위험도는 비슷하다.

관심: 흡연이 암에 걸릴 오즈를 얼마나 높이는가? (OR의 추정)

오즈비와 상대위험은 엄연히 다름

2×2 분할표; 단면조사 (Norusis, 1988)

1984년 미국 성인을 대상으로 한 임금과 직업 만족도에 관한 설문조사

	Dissatisfied	Satisfied	Total
$\leq 15,000$	104	391	495
$\geq 15,000$	66	340	406
Total	170	731	901

질문: 연봉과 직업 만족도와 관련이 있는가? 그렇다면 그 영향력은 어느 정도인가?

역학연구 설계: 단면조사 (cross-section) 연구

- 질병의 유병상태와 위험요인의 노출간의 관련성을 특정 시점 또는 기간 동안에 조사하는 것을 말한다.
- 위험도, 상대위험도, 오즈비 모두 계산할 수 있다.

상대위험도와 오즈비의 해석

다음과 같은 형태의 분할표에서 상대위험도와 오즈비를 해석한다.

조건	결과 위험	Okay
A	10	90
B	40	60

- p_A = 조건 A일 때의 위험도 = 10%, A의 오즈 = 1/9
- p_B = 조건 B일 때의 위험도 = 40%. B의 오즈 = 4/6
- 상대위험도 = 조건 A일 때의 위험도 / 조건 B일 때의 위험도 = 1/4 " 조건 A가 B에 비해 위험도를 4배 낮춘다"
- 오즈비 = 조건 A일 때의 오즈 / 조건 B일 때의 오즈 = (1/9)/(4/6) = 1/6. " 조건 A가 오즈를 6배 낮춘다."

희귀한 질병같은 경우 상-위, 오즈비가 비삿해짐

상대위험도와 오즈비의 해석

다음과 같은 형태의 분할표에서 상대위험도와 오즈비를 해석한다.

```
    결과
조건 위험 Okay
   B   40   60
   A   10   90
```

(조건 A, B의 순서가 바뀌었다.)

- p_B = 조건 B일 때의 위험도 = 40%. B의 오즈 = 4/6
- p_A = 조건 A일 때의 위험도 = 10%, A의 오즈 = 1/9
- 상대위험도 = 조건 B일 때의 위험도 / 조건 A일 때의 위험도 = 4 " 조건 B가 A에 비해 위험도를 4배 높인다"
- 오즈비 = 조건 B일 때의 오즈 / 조건 A일 때의 오즈 = (4/6)/(1/9) = 6. " 조건 B가 오즈를 6배 높인다."

R을 이용한 2×2 분할표 자료 분석

- 사례-대조군 분석시 `DescTools::OddsRatio()` 함수를 이용해 오즈비 추정
- 코호트, 단면 조사의 경우, `DescTools::OddsRatio()`, `DescTools::RelRisk()`를 이용해 오즈비와 상대위험도를 모두 추정
- 다음과 같은 형태의 분할표를 가정할 때,

	결과	
조건	위험	Okay
A	10	90
B	40	60

  - 상대위험도 = 조건 A일 때의 위험도 / 조건 B일 때의 위험도
  - 오즈비 = 조건 A일 때의 오즈 / 조건 B일 때의 오즈

  를 추정한다.

R을 이용한 2×2 분할표 자료 분석

```
ex1

##         조건
## 결과     A  B
##   Okay 90 60
##   위험 10 40

(ex2 <- t(ex1)) # 행과 열을 교환한다

##      결과
## 조건 Okay 위험
##    A   90   10
##    B   60   40

(ex3 <- Rev(ex2, margin = 2)) # 열의 변수값 순서를 바꾼다

##      결과
## 조건 위험 Okay
##    A   10   90
##    B   40   60
```

R을 이용한 2×2 분할표 자료 분석

```
RelRisk(ex3)

## [1] 0.25

OddsRatio(ex3)

## [1] 0.1667

RelRisk(ex3, conf.level = 0.95) #95% 신뢰구간 추정

## rel. risk    lwr.ci    upr.ci
##     0.2500    0.1324    0.4602

OddsRatio(ex3, conf.level = 0.95)

## odds ratio     lwr.ci     upr.ci
##    0.16667    0.07747    0.35856
```

비타민 C와 감기 자료 분석 (Pauling, 1971)

```
pauling <-matrix(c(17,122,31,109),nrow=2)
dimnames(pauling)<-list(cold=c("Yes", "No"), treatment=c("Vitamin C","Placebo"))
(pauling <- Rev( t(pauling), margin = 1 ))
```

```
##             cold
## treatment    Yes  No
##   Placebo     31 109
##   Vitamin C   17 122
```

```
chisq.test(pauling)$p.value
```

```
## [1] 0.04186
```

```
RelRisk(pauling, conf.level = 0.95)
```

```
## rel. risk    lwr.ci    upr.ci
##     1.811     1.063     3.110
```

결과해석:
1. 비타민 C복용이 감기발병율에 영향을 미친다. (χ^2검정)
2. 비타민 C복용을 하지 않으면 감기에 걸릴 위험도가 1.8 배 높아진다 (상대위험도 추정)

흡연과 구강암 자료 분석 (Keller, 1965)

```
keller <-matrix(c(484,385,27,90),nrow=2)
dimnames(keller)<-list(group=c("case", "control"), smoking=c("Yes", "No"))
(keller.new <-t(keller))

##         group
## smoking case control
##     Yes  484     385
##     No    27      90

OddsRatio(keller.new , conf.level = 0.95)

## odds ratio     lwr.ci     upr.ci
##      4.190      2.671      6.575
```

결과해석:
1. 흡연그룹에서 구강암의 오즈 추정값은 비흡연그룹에서 구강암의 오즈 추정값의 4.19배 (신뢰구간 2.67-6.57)이다. (오즈비의 해석)
2. 구강암이 희귀한 질병이므로, 흡연자의 구강암 위험도가 비흡연자에 비해 대략 4.19배 높다고 볼 수 있다.

비슷하다...

임금과 직업 만족도 자료 분석 (Norusis, 1988)

```
norusis.tab1 <-matrix(c(104,66,391,340),nrow=2)
dimnames(norusis.tab1)<-list(wage=c("Low","High"), satisfaction=c("Low", "High"
norusis.tab1

##       satisfaction
## wage   Low High
##   Low  104  391
##   High  66  340

RelRisk(norusis.tab1 , conf.level = 0.95)

## rel. risk    lwr.ci    upr.ci
##    1.2924    0.9799    1.7097

chisq.test(norusis.tab1)$p.value

## [1] 0.08379
```

결과해석: 임금이 낮은 그룹에서 직업에 불만인 비율은 21%이고 높은 그룹에서 불만족자의 비율은 16% 이며 그 비는 1.29 (신뢰구간 0.98-1.71) 이지만 1과의 차이는 통계적으로 유의하지 않다.

연습문제 3: 마약 정맥주사와 HIV 감염여부

뉴욕주 감옥에 수감되어 있는 475명의 여죄수를 대상으로 마약정맥주사 (Intravenous drug use)를 사용한지 여부와 HIV 감염여부에 대해 조사해 보았다.

		HIV+	HIV-	Total
IVDU	Yes	59	77	136
IVDU	No	29	310	339
	Total	88	387	475

1. 마약정맥주사와 HIV 감염여부가 관련이 있는지 통계적 검정을 실시하라.
2. 마약정맥주사를 사용한 그룹과 그렇지 않은 그룹간의 HIV 감염여부에 대한 relative risk를 구하고 이를 해석하라.
3. 마약정맥주사를 사용한 그룹과 그렇지 않은 그룹간의 HIV 감염여부에 대한 odds ratio를 구하고 이를 해석하라.

연습문제 3: 마약정맥주사와 HIV 감염여부

```
prison <- tribble(
      ~Drug, ~HIV, ~Freq,
      "Yes", "Positive", 59,
      "Yes", "Negative", 77,
      "No",  "Positive", 29,
      "No", "Negative" , 310
)
prison.tab <- xtabs(Freq~Drug + HIV, data = prison)
( prison.tab <- Rev(prison.tab, margin = c(1,2)) )

##      HIV
## Drug  Positive Negative
##   Yes       59       77
##   No        29      310

## 1. Chi-square test of independence
chisq.test(prison.tab)$p.value

## [1] 3.287e-18
```

유의확률 (p.value $\approx$ 0.000)이 작으므로 관련이 있다고 볼 수 있다.

연습문제 3: 마약정맥주사와 HIV 감염여부

```
## 2. Relative risk estimation
RelRisk(prison.tab , conf.level = 0.95)

## rel. risk    lwr.ci    upr.ci
##     5.071     3.419     7.536
```

마약정맥주사를 하지 않은 그룹에서 HIV 감염 비율은 8.6% 이며
마약정맥주사를 한 그룹에서 HIV 감염 비율은 43.4%이다.
상대위험도는 5.07(신뢰구간 3.41-7.55), 즉 대략 다섯 배가 높다.

```
## 3. Relative risk estimation
OddsRatio(prison.tab , conf.level = 0.95)

## odds ratio     lwr.ci     upr.ci
##      8.191      4.920     13.637
```

마약정맥주사를 사용한 그룹의 HIV감염에 대한 오즈 추정값은
마약정맥주사를 사용하지 않은 그룹의 HIV 감염 오즈 추정값의
8.19배 (신뢰구간 4.92-13.64)이다.

2×2 분할표; 쌍을 이룬 자료 (HIVNET, 1995)

- 백신임상실험에 관한 지식 설문조사를 2번에 걸쳐서 실시함.
- 처음 설문조사 후 6개월이 지난 후 동의서 절차를 거친 후에 두 번째 설문조사를 실시한다.

	Month 6		
Baseline	Incorrect	Correct	Total
Incorrect	251	178	429
Correct	68	98	166
Total	319	276	595

관심: 동의서 처리절차가 백신임상실험에 관한 이해도를 높이는데 기여하였는가?

쌍을 이룬 자료

- Pre-tratment vs Post-treatment
- Matched-Case-Control: 사례군과 대조군을 성별/나이등에 따라 쌍을 지음
- Paired observation: 왼눈 vs 오른눈

	Month 6		
Baseline	Incorrect	Correct	Total
Incorrect	251	178	429
Correct	68	98	166
Total	319	276	595

- 첫 번째 설문조사에서 오답율: $429/595 = 0.72$
- 두 번째 설문조사에서 오답율: $319/595 = 0.54$

두 변수의 독립성이 아니라, 두 오답율이 같은지 다른지가 중요하다.

McNemar 검정

		Controls	
		E=1	E=0
Cases	E=1	π_{11}	π_{10}
	E=0	π_{01}	π_{00}

McNemar 검정을 이용하여 아래 귀무가설을 검정한다.

$$H_0 : \pi_{10} = \pi_{01}$$

즉,

$$P(\text{Case}|E = 1) = \pi_{11} + \pi_{10} = \pi_{11} + \pi_{01} = P(\text{Control}|E = 1)$$

McNemar 검정

```
hivnet <- matrix(c(251, 68, 178, 98), nrow = 2)
dimnames(hivnet) <-list(baseline =c("Incorrect",  "correct"),
                        post = c("Incorrect", "correct"))
mcnemar.test(hivnet)

##
##  McNemar's Chi-squared test with continuity
##  correction
##
## data:  hivnet
## McNemar's chi-squared = 48, df = 1, p-value
## = 4e-12
```

결과해석: 동의서 처리결과가 백신 임상실험에 관한 이해도를 변경시켰다는 것이 통계적으로 유의하다.

예제: 소금 섭취량와 심장혈관 질환

소금 섭취량과 심장혈관 질환과 관련이 있을까?

	High Salt	Low Salt	Total
CVD	5	30	35
Non-CVD	2	23	25
Total	7	53	60

이 경우 OR $= 1.90$이지만 χ^2검정의 경우 기대값이 5이하인 cell이 전체 cell의 25% 이상이면 그 결과를 신뢰하기 힘들다. 이러한 경우 FIsher's exact test를 사용한다.

Fisher's Exact Test

행과 열의 합이 주어졌다고 가정한다면, 귀무가설 (소금 섭취량과 심장병이 독립) 아래에서 우리가 관측할 수 있는 분할표들은 다음과 같다.

0		35
		25
7	53	60

1		35
		25
7	53	60

2		35
		25
7	53	60

3		35
		25
7	53	60

4		35
		25
7	53	60

5		35
		25
7	53	60

6		35
		25
7	53	60

7		35
		25
7	53	60

Fisher's Exact Test

초기하 분포를 사용한다면 각각의 분할표를 관측할 확률을 구할 수 있다.

0	35	35
7	18	25
7	53	60

.001

1	34	35
6	19	25
7	53	60

.016

2	33	35
5	20	25
7	53	60

.082

3	32	35
4	21	25
7	53	60

.214

4	31	35
3	22	25
7	53	60

.312

5	30	35
2	23	25
7	53	60

.252

6	29	35
1	24	25
7	53	60

.105

7	28	35
0	25	25
7	53	60

.017

즉 주어진 자료나 그보다 더 극단적인 결과를 관측할 확률 (p-value)은 0.252+0.105+0.017=0.374이다.

Fisher's Exact Test

```
CVD.salt.tab <- matrix (c(5,2,30,23), nrow=2)
dimnames(CVD.salt.tab) <-list(CVD=c("Yes", "No"), Salt=c("High", "Low"))
fisher.test(CVD.salt.tab, alternative = "greater")

##
##  Fisher's Exact Test for Count Data
##
## data:  CVD.salt.tab
## p-value = 0.4
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  0.3573    Inf
## sample estimates:
## odds ratio
##      1.897
```

결과해석: 소금 섭취량과 심장혈관 질환과 관련여부는 통계적으로 유의하지 않다.

연습문제 4: Tea Tasting Lady

R. A. Fisher의 동료 Dr. Bristol은 컵에 차와 밀크를 컵에 넣는 순서를 차의 맛을 보면 알아 낼 수 있다고 주장하였다. 그녀의 주장을 검증하기 위해 Fisher는 다음과 같은 실험을 실시하였다. Dr. Bristol에게 8개의 차가 제공되었는데 4개는 우유를 먼저, 나머지는 차를 먼저 부었다. 제공된 차의 순서는 임의로 정해졌고 실험결과는 다음과 같다.

실제로	추측한 것		
먼저 부은것	우유	차	합계
우유	3	1	4
차	1	3	4
합계	4	4	8

Dr. Bristol의 주장을 검증하라.

연습문제 4: Tea Tasting Lady

```
TeaTasting <-matrix(c(3, 1, 1, 3),nrow = 2)
dimnames(TeaTasting) <- list(Guess = c("Milk", "Tea"),
                      Truth = c("Milk", "Tea"))
fisher.test(TeaTasting, alternative = "greater")

##
##  Fisher's Exact Test for Count Data
##
## data:  TeaTasting
## p-value = 0.2
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  0.3136    Inf
## sample estimates:
## odds ratio
##      6.408
```

$OR = 6.41$이지만 주장이 사실인지 여부 (OR > 1) 에 대한 검정결과 p.vaule=0.2429로 귀무가설을 기각할 수 없다. 즉 Dr. Bristol의 주장이 사실이라고 단정할 수 없다.

층화분석

- Confounding Variables
- 층화분석
  - Mantel-Haenszel OR estimate
  - Mantel-Haenszel test for assoication
  - Breslow-Day test for Homogeneity

Simpson의 역설

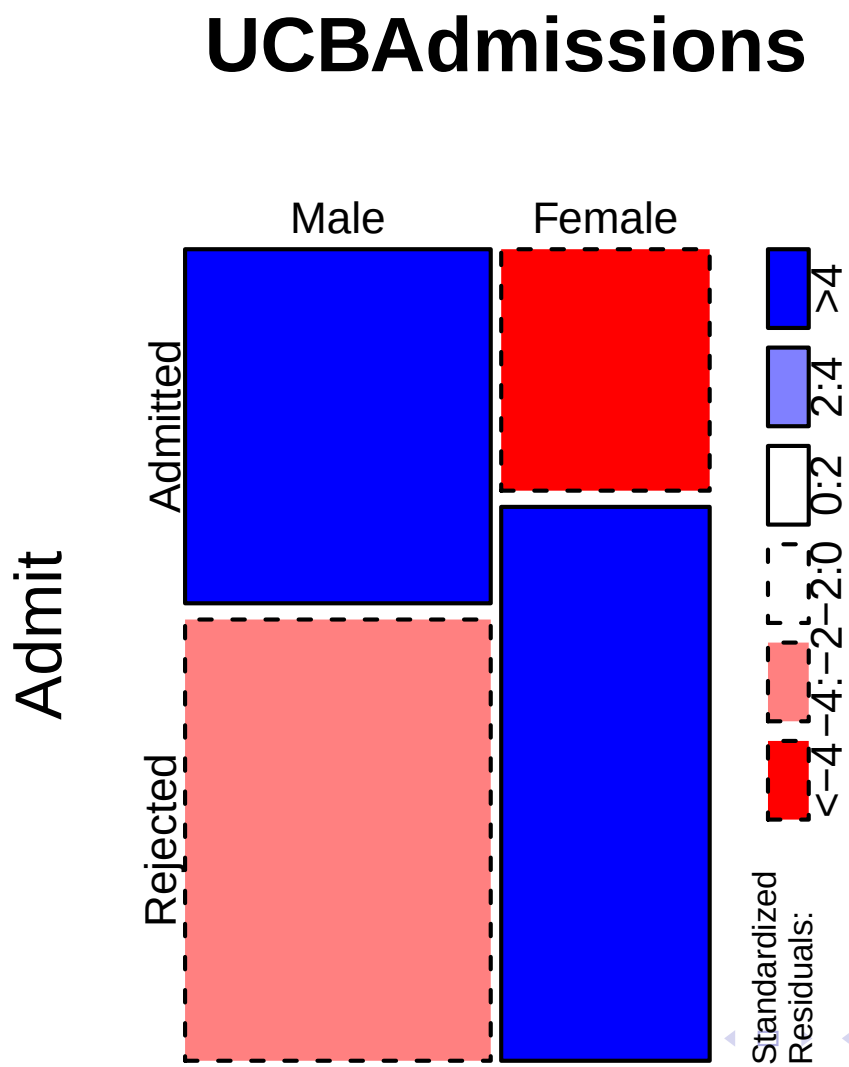
1970년대 초반 대학원 입시에 성차별이 있었다는 이유로 버클리 대학이 고소를 당한다. 아래는 버클리 대학의 1973년 6개 대학원 입시결과이다.

	남자	여자	Total
합격	1198	557	1755
불합격	1493	1278	2771
합계	2691	1835	3526

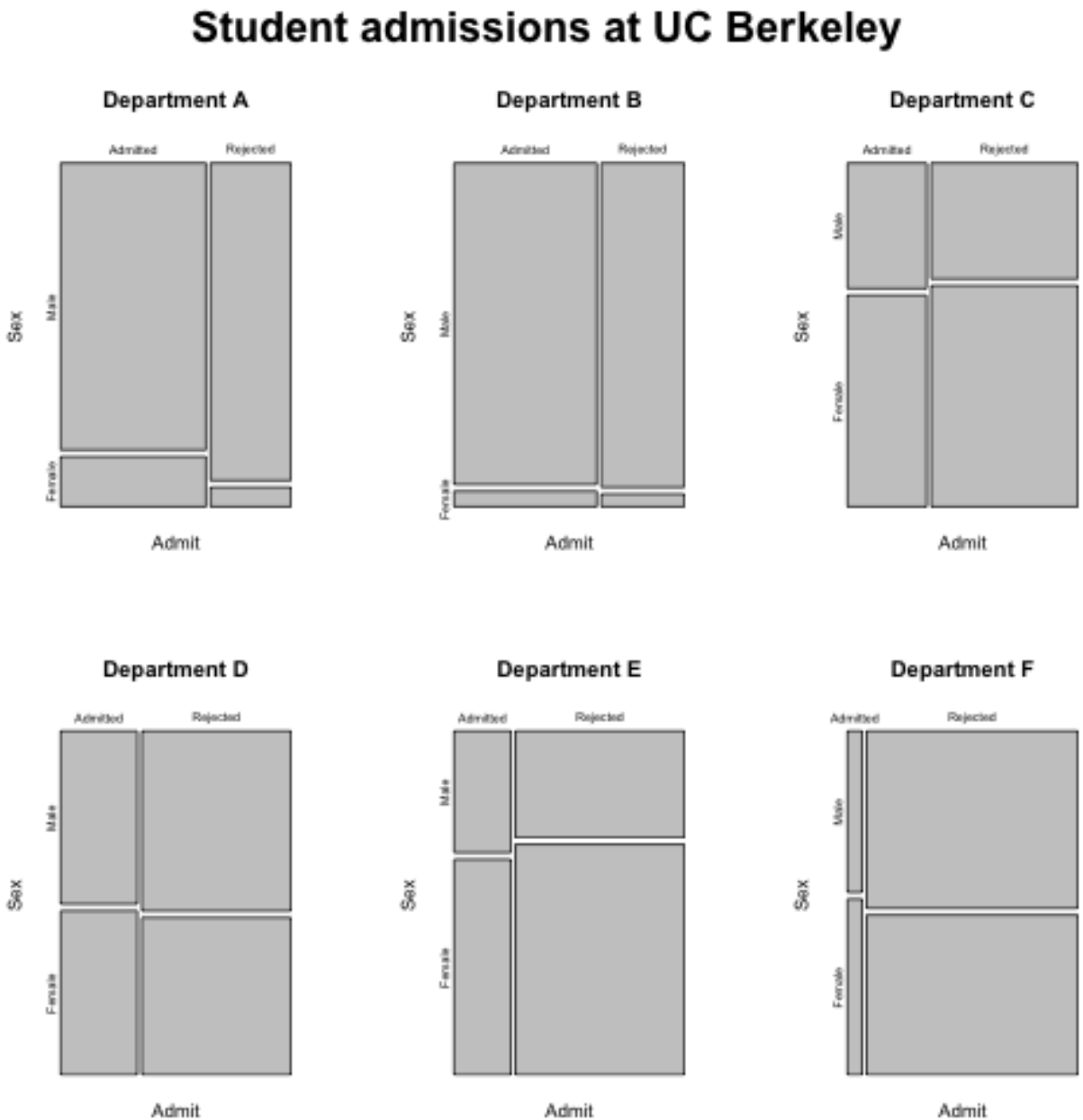
하지만 입시결과를 대학원별로 나누어서 분석한 결과 대부분의 대학원에서 오히려 여학생들의 합격율이 높다는게 밝혀졌다. 즉 대학원이 중첩변수(confounding variable)이다.

Mosaic Plot

```
mosaicplot(~ Gender + Admit, UCBAdmissions, shade = TRUE)
```

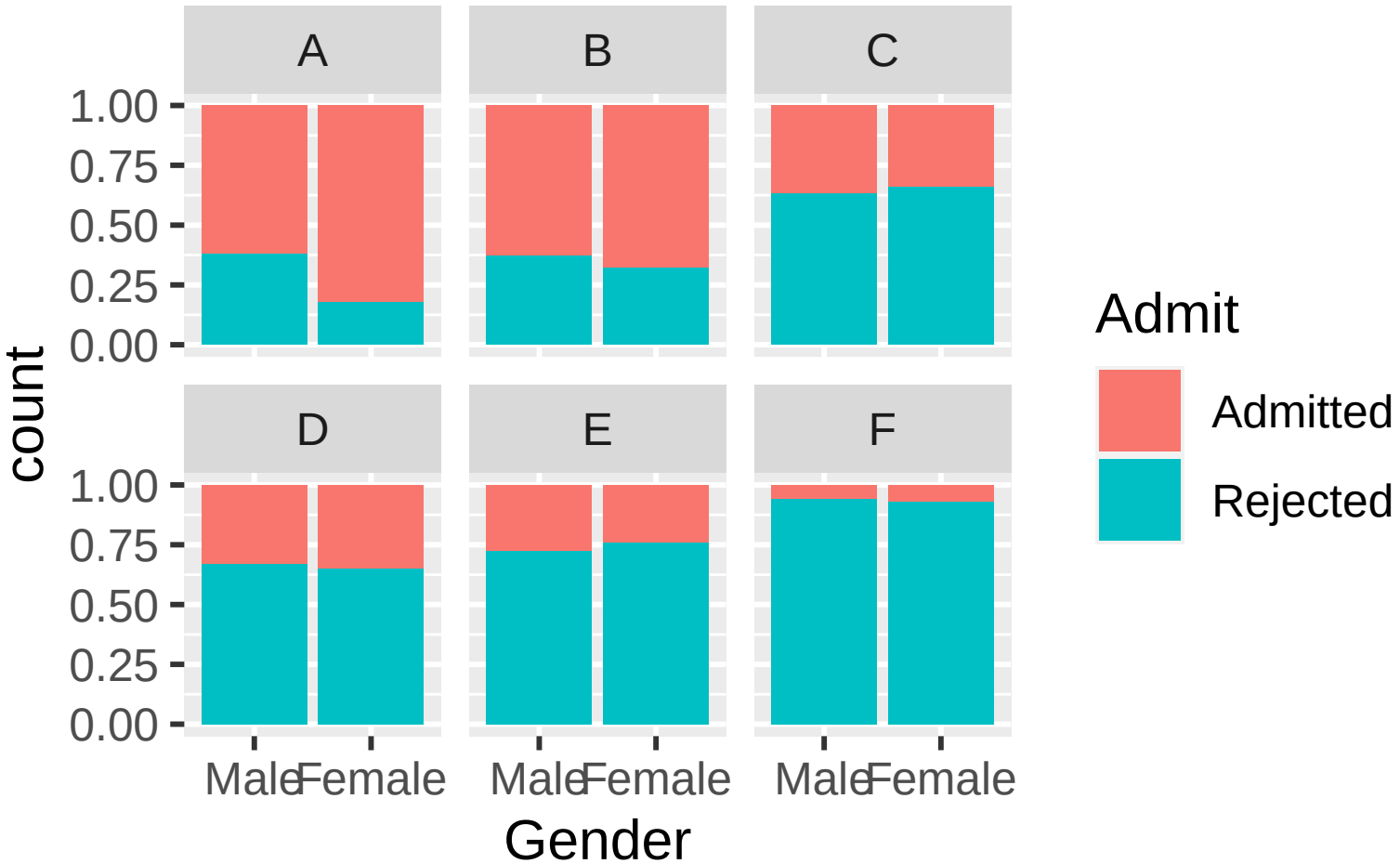


Mosaic Plot


Student admissions at UC Berkeley
Department A
Department B
Department C
Department D
Department E
Department F
Admitted
Rejected
Male
Female
Sex
Admit

층별 비율 비교

```
library(ggplot2)
ggplot( data = Untable(UCBAdmissions), aes(x = Gender, fill = Admit)) +
geom_bar(position = "fill")  +facet_wrap(~Dept)
```



심슨의 역설: 농구슛 성공률

- 3점슛 성공률

	시도	성공	성공률
하승진	50	20	40%
양동근	54	22	41%

- 2점슛 성공률

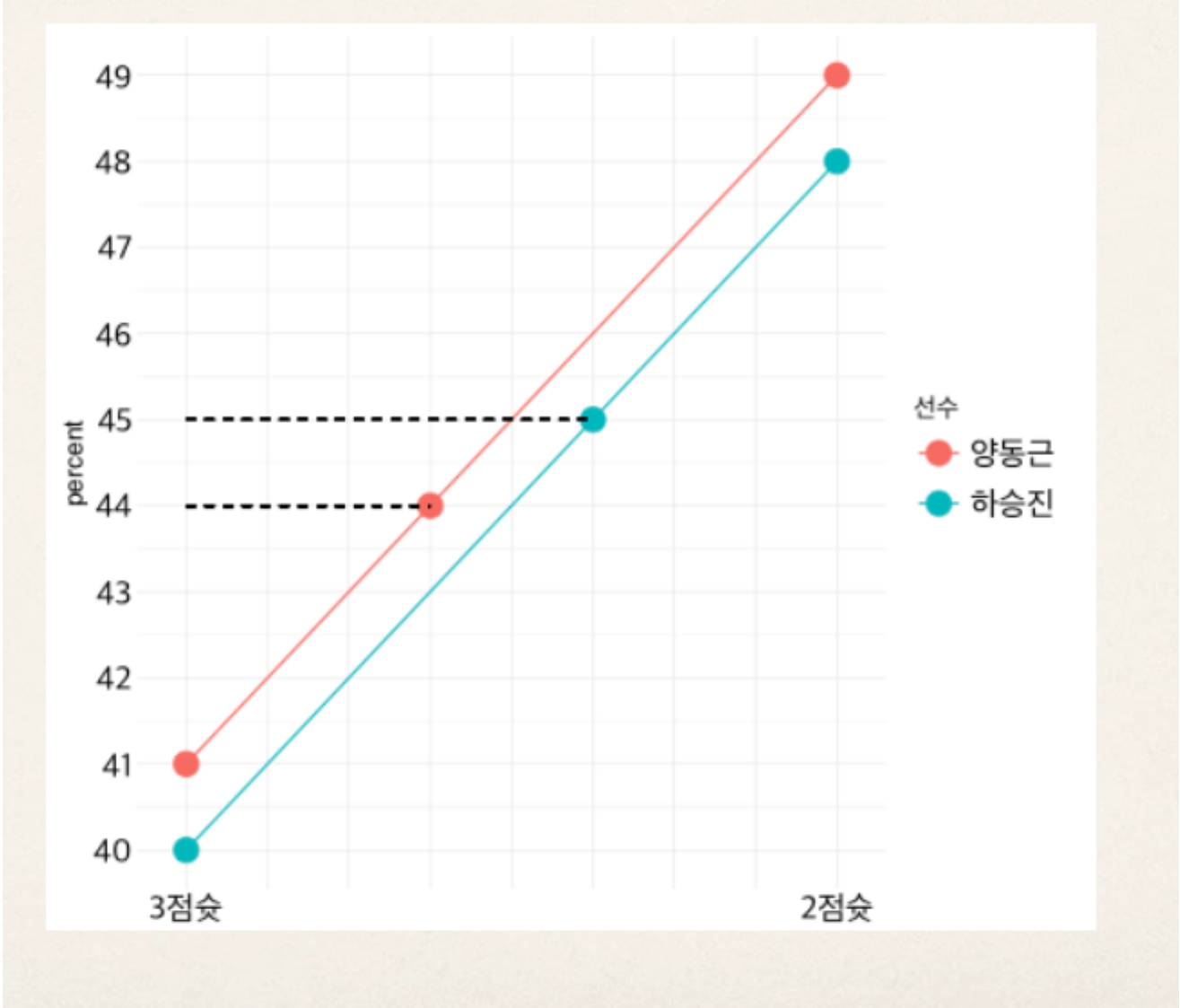
	시도	성공	성공률
하승진	64	31	48%
양동근	39	19	49%

심슨의 역설: 농구슛 성공률

- 전체 슛 성공률은 하승진이 앞선다

	시도	성공	성공률
하승진	114	51	45%
양동근	93	41	44%

심슨의 역설: 농구슛 성공률



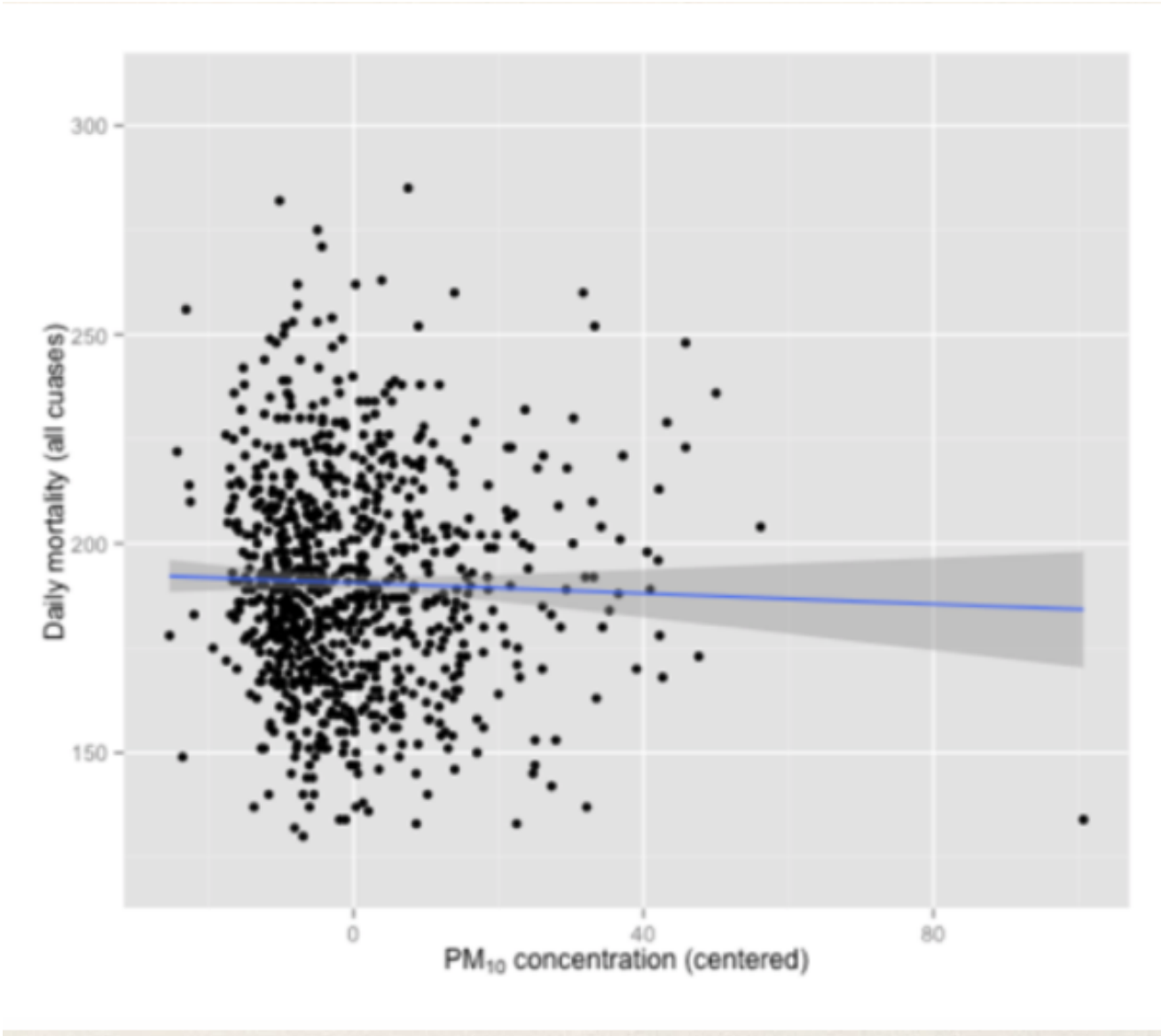
심슨의 역설: 농구슛 성공률

왜 그럴까?

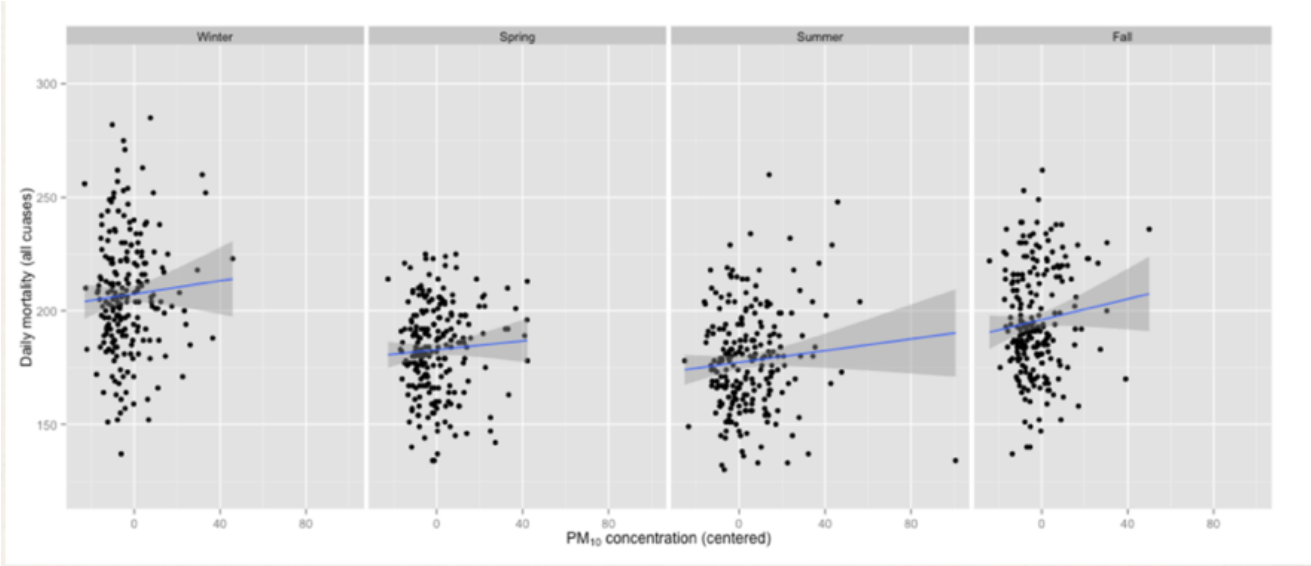
- 성공률이 높은 2점슛을 하승진이 상대적으로 더 많이 시도하므로, 하승진의 성공률 평균을 내면 2점슛의 성공횟수가 3점슛에 비해 과대하게 대표된다. (하-45% vs 양-44%)
- 슛의 종류별 성공률의 평균:
  - 하승진: (40% + 48%) / 2 = 44 %
  - 양동근: (41% + 49%) / 2 = 45 %
- 오즈비를 득점의 종류별로 계산하면 비슷하다
  - 오즈비(2점슛) = 하승진의 2점슛성공오즈 / 양동근의 2점슛성공오즈 = 0.99
  - 오즈비(3점슛) = 하승진의 3점슛성공오즈 / 양동근의 3점슛성공오즈 = 0.97
- 오즈비의 평균? 0.98?

대안: 2점슛과 3점슛 전체 개수가 다른 경우 가중평균을 해야 한다 (M-H combined OR)

심슨의 역설: 미세먼지


300
250
200
150
Daily mortality (all cuases)
0
40
80
PM10 concentration (centered)

심슨의 역설: 미세먼지


Winter
Spring
Summer
Fall
Daily mortality (all cuases)
300
250
200
150
0
40
80
PM10 concentration (centered)

Kahn and Sempos (1989)

나이와 수축기혈압이 심장마비에 어떤 영향을 미치는지 알고자 한다.

	heart attack		
	present	absent	Total
SBP$\geq$ 140	29	711	740
SBP $<$ 140	27	1244	1271
Total	56	1955	2011

이 경우 OR $=$ 1.88 (95% CI:1.10-3.20)이며 p.value=0.03이다.
(수축기 혈압이 높으면 심장마비가 일어날 위험이 크다.)

연령대 분석

Age $\geq$ 60	MI cases	MI negative	
SBP$\geq$ 140	9	115	124
SBP $<$ 140	6	73	79
Total	15	188	203

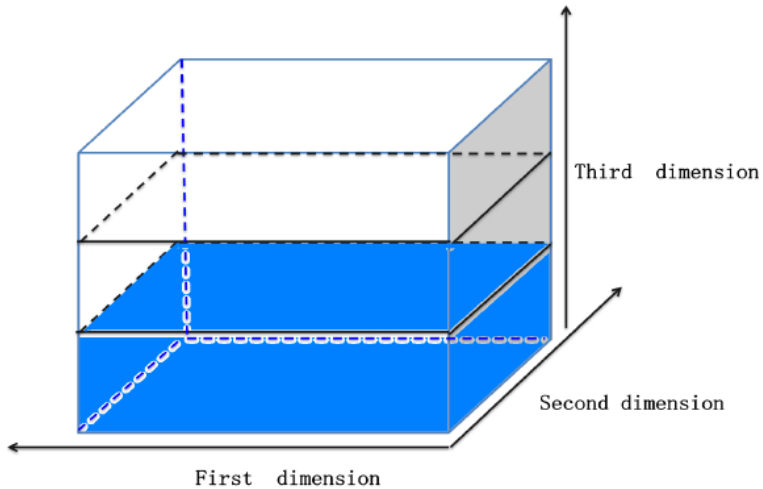
60세 이상에서는 심장마비와 수축기혈압의 오즈비는 =0.95 (95% CI: 0.33-2.79)이다. (60세 이상에서는 수축기 혈압과 심장마비가 관계가 있다고 단정 불가)

Age $<$ 60	MI cases	MI negative	
SBP$\geq$ 140	20	596	616
SBP $<$ 140	21	1171	1192
Total	41	1767	1808

60세 미만에서는 심장마비와 수축기혈압의 오즈비는 =1.87 (95% CI: 1.01-3.48)이다. (60세 미만에서는 수축기 혈압이 높으면 심장마비가 일어날 위험이 크다.)

3차원 분할표

- 3차원 분할표는 table형태로 주어지며 1번째 argument가 설명변수, 2번째가 outcome, 3번째가 층화변수이다.



3차원 분할표

나이와 심근경색 자료에서 나이를 confounding variable로 간주하고 혈압과 심근경색의 관계를 알고자 한다면

```
MI.raw <-read.csv("MI2.csv")
( MIdata <- Rev( xtabs(~ SBP + MI + Age, MI.raw), margin = 1) )

## , , Age = <60
##
##        MI
## SBP     case negative
##   >140   20      596
##   <140   21     1171
##
## , , Age = >60
##
##        MI
## SBP     case negative
##   >140    9      115
##   <140    6       73
```

층화분석

3차원 분할표의 층화분석은 다음과 같은 분석을 진행할 수 있다.

- 중첩변수(confounding variable)의 층별로 분석을 실시하여 추정치를 구한다.
- 각 층별 오즈비가 모두 같은지 여부를 Breslow-Day test를 이용하여 검정 (교호작용 여부 검정과 동일)
- 교호작용이 있을 경우 각 층별로 오즈비의 추정치를 보고 한다,
- 교호작용이 없을 경우 각 층별 오즈비 추정치의 가중평균을 계산하여 하나의 통일된 오즈비 (M-H Combined OR) 추정치 계산한다. 이 경우 오즈비는 중첩변수를 고려하여 계산한 오즈비이다.
- 여기서 추정된 M-H Combined OR가 1인지 여부를 Mantel-Haenszel test를 이용하여 검정한다,

1. 층별 오즈비, 상대위험도 추정

다음의 R Code를 이용하여 층별 오즈비와 상대위험도를 추정한다.

```
apply(MIdata, 3,
      function(x) list(rbind(
   "Case-control (odds ratio)" = OddsRatio(x, conf.level = 0.95),
   "Cohort (col1 risk)" = RelRisk(x, conf.level = 0.95),
   "Cohort (col2 risk)" = RelRisk(Rev(x, 1), conf.level = 0.95))))
```

```
## $`<60`
## $`<60`[[1]]
##                            odds ratio lwr.ci
## Case-control (odds ratio)     1.8712 1.0063
## Cohort (col1 risk)            1.8429 1.0150
## Cohort (col2 risk)            0.5426 0.2993
##                            upr.ci
## Case-control (odds ratio) 3.4793
## Cohort (col1 risk)        3.3414
## Cohort (col2 risk)        0.9852
##
##
## $`>60`
## $`>60`[[1]]
##                            odds ratio lwr.ci
## Case-control (odds ratio)     0.9522 0.3254
## Cohort (col1 risk)            0.9556 0.3695
## Cohort (col2 risk)            1.0464 0.3993
##                            upr.ci
## Case-control (odds ratio)  2.787
## Cohort (col1 risk)         2.504
## Cohort (col2 risk)         2.706
```

2. Breslow-Day test

`DescTools::BreslowDayTest()` 를 이용하여 층별 오즈비가 모두 같은지 여부를 검정한다. 이 경우, 고연령에서의 혈압과 심장마비의 오즈비 (OR_1)와 저연령에서의 혈압과 심장마비의 오즈비 (OR_2)가 같다가 귀무가설이다.

$$H_0 : OR_1 = OR_2$$

```
BreslowDayTest(MIdata)

##
##  Breslow-Day test on Homogeneity of Odds
##  Ratios
##
## data:  MIdata
## X-squared = 1.2, df = 1, p-value = 0.3
```

결과분석: 심장마비와 수축기혈압의 오즈비는 연령별로 다르다고 할 수 없다 (Homegeneity test statistic = 1.2, p.value=0.3). 만약 연령별로 다르다면 (이 경우는 아니다) 오즈비를 연령별로 따로 보고한다. 끝.

3. 나이 효과를 조정한 통합 오즈비 추정 및 검정

만약 Breslow-Day test의 결과 오즈비가 연령별로 차이가 없다면,

1. 나이의 효과를 조정한 통합 오즈비 (Mantel-Haenszel Combined Odds Ratio)를 추정한다.
2. 통합 오즈비가 1인지를 검정한다.

```
mantelhaen.test(MIdata)
```

3. 나이 효과를 조정한 통합 오즈비 추정 및 검정

```
##
##  Mantel-Haenszel chi-squared test with
##  continuity correction
##
## data:  MIdata
## Mantel-Haenszel X-squared = 2.3, df = 1,
## p-value = 0.1
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.9134 2.6937
## sample estimates:
## common odds ratio
##             1.569
```

결과해석:

- 연령을 고려한 심장마비과 수축기혈압은 관련이 있다고 할 수 없다. (Mantel-Haenzel test statistic =2.3, p.value=0.1).
- 나이를 고려하여 계산한 심장마비와 수축기혈압의 오즈비 (M-H Combined OR) 는 1.57 이며 나이를 고려하지 않은 경우 (crude OR =1.88) 보다 낮아졌다. 때문에 OR = 1 이 아닌 증거가 충분치 않다.

연습문제 5: 안전벨트 착용과 교통사고 사망율

스피드를 고려하여 안전벨트 착용과 교통사고 사망율의 관계를 분석하라.

	Impact Speed			
	< 40 mph		≥ 40 mph	
Driver	seat belt		seat belt	
	worn	not	worn	not
dead	31	22	73	185
alive	273	184	137	129

연습문제 5: 안전벨트 착용과 교통사고 사망율

```
accident <-as.table(array(c(31,22,273,184,73,185,137,129),
                          dim = c(2, 2, 2)))
dimnames(accident) <-list(Seat_belt  = c("worn","not"),
                          Drive = c("dead","alive"),
                          Impact_Speed = c("<40mph", ">=40mph"))
accident <- Rev(accident, 1)


# 0. Visualize
mosaicplot(~ Seat_belt + Drive, data = accident)
mosaicplot(~ Impact_Speed + Seat_belt + Drive, data = accident)
ggplot( data = Untable(accident), aes(x = Seat_belt, fill = Drive)) +
geom_bar(position = "fill")  + facet_wrap(~Impact_Speed)
```

연습문제 5: 안전벨트 착용과 교통사고 사망율

```r
# 1. 속도별 오즈비 추정
apply(accident, 3,
      function(x) list(rbind(
   "Case-control (odds ratio)" = OddsRatio(x, conf.level = 0.95),
   "Cohort (col1 risk)" = RelRisk(x, conf.level = 0.95),
   "Cohort (col2 risk)" = RelRisk(Rev(x, 1), conf.level = 0.95))))

# 2. 오즈비가 층별로 다른지 검정
BreslowDayTest(accident)

# 3. 통합 오즈비 추정 및 검정
mantelhaen.test(accident)
```

정리

- 범주형 자료와 분할표 `xtabs()`, `tribble()`
- 범주형 변수들의 동일성(독립성) 검정 `chisq.test()`, `fisher.test()`
- 역학연구와 2×2 분할표
- 상대위험도와 오즈비 `RelRisk()`, `OddsRatio()`
- 쌍을 이룬 자료 `mcnemar.test()`
- 중첩변수의 효과를 제거하는 층화분석 `BreslowDayTest()`, `mantelhaen.test()`

일단 분할표로 만들 것. 혹은 이미 분할표만 주어졌다면?
3가지로 나뉘는 역학연구..

참고자료

- Dalgaard, P. (2008) *Introductory Statistics with R*. 2nd edition. Springer
- McGready, J. Lecture notes of *Statistical Reasoning II*. `http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/StatisticalReasoning2/coursePage/index/`
- Verzani, J. (2005) *Using R for Introductory Statistics*. Chapman & Hall.
- Signorell, J (2015) Tables in R – A quick practical overview, Manuscript.