03 강

데이터분석방법론2

삼차원 분할표

통계·데이터과학과 이기재 교수

학습목차

학습개요 및 목표

이번 강의는 3차원 분할표에 대한 분석 방법에 대해서 학습하겠습니다.
부분연관성의 개념, 코크란 - 맨틀 - 핸첼 검정,
브레슬로 - 데이 검정 방법에 대해서 공부하겠습니다.

1. 삼차원 분할표의 부분연관성 개념을 설명할 수 있다.
2. 동질연관성의 개념을 설명할 수 있다.
3. 코크란-맨틀-핸첼 검정, 브레슬로-데이 검정을 적용할 수 있다.

제 3장. 삼차원 분할표

1 개요

2 부분연관성

3 코크란-맨틀-핸첼 방법

01

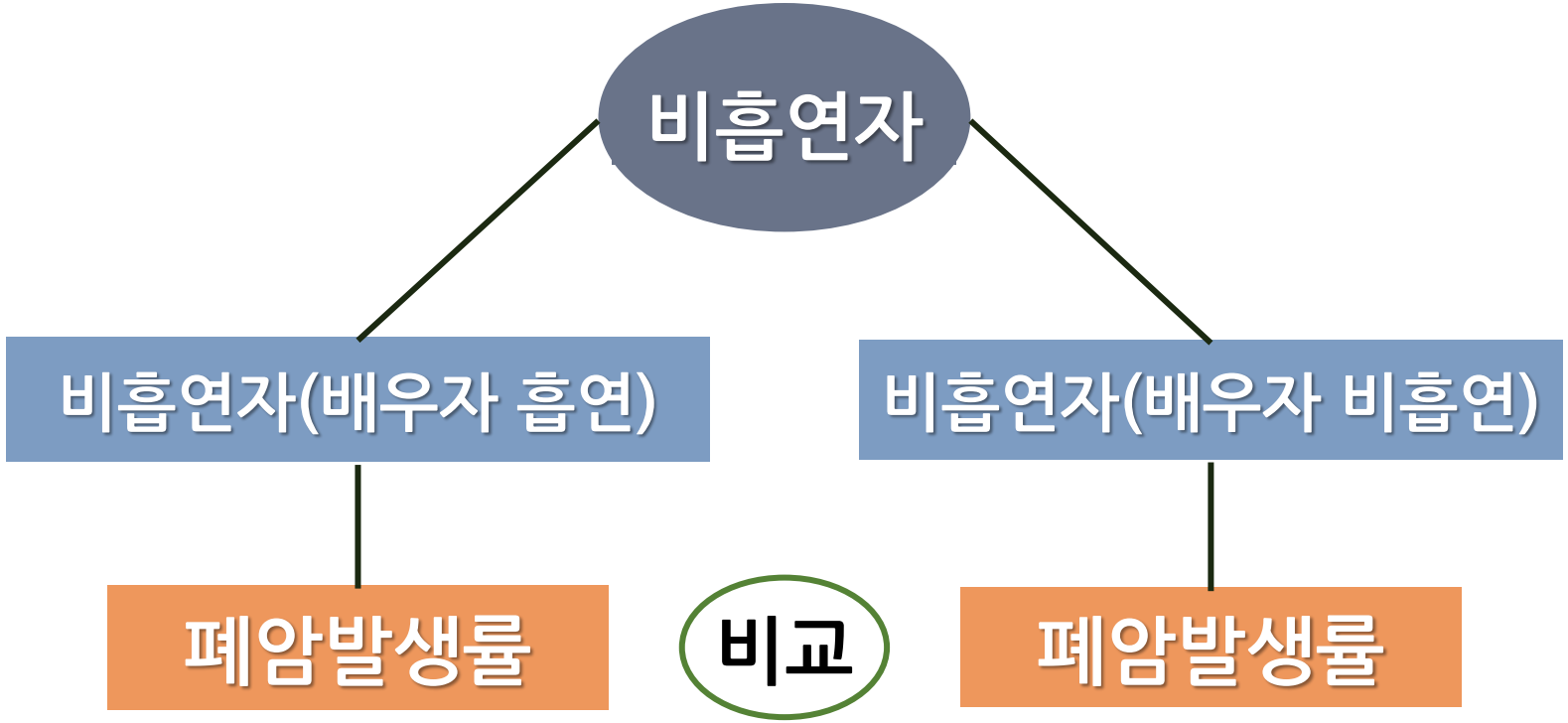
제 3장. 삼차원 분할표

개요

1. 교락변수(Confounding Variable)의 통제 필요성

- 흡연자와 함께 사는 비흡연자에게
  간접 흡연이 미치는 영향을 분석하고자 하는 경우

- 간접흡연과 폐암 발생간의 연관성을
  횡단면 연구(Cross-sectional Study)를 통해서 파악하고자 함



1. 교락변수(Confounding Variable)의 통제 필요성

- 연구 참여자의 나이, 사회·경제적 지위 등은 배우자의 흡연 여부와 연구 참여자의 폐암 발생여부에도 영향을 미칠 수 있음

예

배우자가 비흡연자인 사람들이 상대적으로 젊다면
나이 변수를 통제하지 않을 경우
폐암 발생률의 단순한 비교는 의미가 없음

- 교락변수(Confounding Variable) Z의 효과를 통제하면서 범주형 변수 X와 Y의 연관성을 분석

- 고정된 Z의 수준에 대하여 X와 Y의 연관성을 살펴봄
  ➔ 3차원 분할표 분석

- 여러 교락변수를 동시에 통제할 수 있는 모형과 일반적인 방법에 대한 학습(교재 4장부터의 내용)

02

제 3장. 삼차원 분할표

부분연관성

1. 부분분할표 (Partial Tables)

예제 : FL 사형선고 판결 사례

피해자의 인종	피고의 인종	사형선고		“예”의 비율
		예	아니오	
백인	백인	53	414	11.3%
	흑인	11	37	22.9%
흑인	백인	0	16	0.0%
	흑인	4	139	2.8%
총계	백인	53	430	11.0%
	흑인	15	176	7.9%

Y = 사형선고 여부(반응변수)
X = 피고의 인종(설명변수)
Z = 피해자의 인종(통제변수)

1. 부분분할표 (Partial Tables)

“Z = 피해자의 인종”을 통제하는 경우

① 피해자 = 백인

피고	사형선고	
	예	아니오
백인	53	414
흑인	11	37

② 피해자 = 흑인

피고	사형선고	
	예	아니오
백인	0	16
흑인	4	139

1. 부분분할표 (Partial Tables)

X-Y 주변분할표 (Marginal Table)

- 모든 부분분할표를 결합해서 얻은 이차원 분할표
- 주변분할표는 변수 Z를 통제하지 않고 통합하여 작성함

피고	사형선고		
	예	아니오	"예" 비율
백인	53	430	11.0%
흑인	15	176	7.9%

1. 부분분할표 (Partial Tables)

FL 사형선고판결 사례분석

피해자의 인종	피고의 인종	사형선고		"예"의 비율
		예	아니오	
백인	백인	53	414	11.3%
	흑인	11	37	22.9%
흑인	백인	0	16	0.0%
	흑인	4	139	2.8%
총계	백인	53	430	11.0%
	흑인	15	176	7.9%

Y = 사형선고 여부(반응변수)
X = 피고의 인종(설명변수)
Z = 피해자의 인종(통제변수)

1. 부분분할표 (Partial Tables)

- 피해자의 인종이 「백인」인 경우
  : 흑인에 대한 사형선고 비율이 11.6% 높음

- 피해자의 인종이 「흑인」인 경우
  : 흑인에 대한 사형선고 비율이 2.8% 높음

- 피해자의 인종을 무시한 주변표에서는
  반대로 백인의 사형선고 비율이 흑인에 비해 3.1% 높음

➔ 부분분할표와 주변분할표는
반대 양상의 연관성을 보여줌

1. 부분분할표 (Partial Tables)

❑ 피해자 인종의 통제 여부에 따른 두 변수 연관성 차이 발생 이유

① 피해자와 피고 인종 간의 연관성이 매우 강함

백인은 흑인을 살해하기보다 같은 백인을 살해할 오즈가 87.0[=(467 X 143)/(48 X 16)]로 대단히 큼

피고	피해자	
	백인	흑인
백인	467	16
흑인	48	143

② 피고의 인종에 관계없이 백인이 피해자인 경우는 흑인 피해자의 경우에 비해 사형판결 비율이 더 높음

결과적으로 백인은 백인을 살해하는 경향이 높고, 백인을 살해한 경우 사형판결 가능성 또한 높기 때문에 주변분할표에서 백인 피고의 사형선고 비율이 더 높음

피해자	사형 여부	
	예	아니오
백인	64	451
흑인	4	155

1. 부분분할표 (Partial Tables)

심프슨의 역설 (Simpson's Paradox)

- 조건부연관성과 주변연관성이
  서로 다른 방향으로 나타나는 현상을 말하며,
  이것은 범주형 변수뿐만 아니라 양적 변수에서도 발생함

- 교락변수 Z를 무시하고,
  주변분할표를 이용하여 분석할 경우에는
  잘못된 분석 결과를 얻을 수 있다는 점에 유의해야 함

2. 조건부 오즈비와 주변 오즈비

조건부 오즈비

$$Z=\text{백인}: \widehat{\theta_{XY(1)}}=\frac{53\times 37}{414\times 11}=0.43$$

$$Z=\text{흑인}: \widehat{\theta_{XY(2)}}=0.00$$

(각 cell에 0.5를 더한 후 계산한 오즈비는 0.94임)

피해자의 인종(Z)을 통제할 때 사형선고를 받을 오즈는
흑인 피고에 비해서 백인 피고의 경우가 낮음

2. 조건부 오즈비와 주변 오즈비

주변분할표에 대한 오즈비

- 부분분할표를 결합하여 XY 주변분할표를 작성

피고	사형선고	
	예	아니오
백인	53	430
흑인	15	176

$$\widehat{\theta_{XY}} = \frac{53 \times 176}{15 \times 430} = 1.45$$

피해자의 인종(Z)를 무시하면
사형선고를 받을 오즈는 백인 피고의 경우가 더 높음

➔ 심프슨의 역설 발생

3. 주변독립성, 조건부 독립성

- 정의 : Z가 주어졌을 때 X와 Y가 조건부 독립
  (Conditionally Independent)

- 2 x 2 x K 분할표에서 X와 Y의 조건부 독립

$$\Leftrightarrow \quad \theta_{XY(1)} = \cdots = \theta_{XY(K)} = 1.0$$

3. 주변독립성, 조건부 독립성

❑ 사례 : 병원별 처리방법에 따른 반응

병원	처리	반응	
		성공	실패
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32
총계	A	20	20
	B	20	40

① $\theta_{XY(1)} = \frac{18 \times 8}{12 \times 12} = 1.0$ ② $\theta_{XY(2)} = \frac{8 \times 8}{2 \times 32} = 1.0$

"각 병원에서 반응과 처리는 조건부독립이다."

3. 주변독립성, 조건부 독립성

▪ 사례 : 병원별 처리방법에 따른 반응

③ $\theta_{XY} = \frac{20 \times 40}{20 \times 20} = 2.0$

"반응과 처리는 주변독립이 아니다."

Note

- 조건부독립이 성립한다고 주변부독립이 성립하는 것은 아니다.
- 주변부 독립이 성립한다고 조건부 독립이 성립하는 것도 아니다.

Note

주변분할표만을 이용해서 분석하면 처리 A가
처리 B에 비해서 성공률이 높다는 잘못된 결론을 내릴 수 있다.

3. 주변독립성, 조건부 독립성

동질적 연관성의 정의

- K개의 2 x 2 부분분할표에서 오즈비 $\theta_{XY(k)}$에 대해 $\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$이 성립할 때 「2 x 2 x K 분할표에서 X-Y동질적 연관성이 있다.」고 함

- K 조건부독립은 X와 Y의 조건부오즈비가 특별히 1.0인 경우임

- I x J x K 분할표의 경우 Z의 각 수준에서 X와 Y 각각의 임의의 두 수준에 의해 결정되는 조건부 오즈비가 동일할 때 **동질적 X-Y조건부 연관성이 있다**고 말함

03

제 3장. 삼차원 분할표

코크란-맨틀-핸첼 방법

1. 중국 도시별 흡연과 폐암자료(8개 도시 대상)

도시	흡연	폐암		오즈비	μ_{11k}	$Var(n_{11k})$
		예	아니오			
베이징	흡연자	126	100	2.20	113.0	16.9
	비흡연자	35	61			
상하이	흡연자	908	688	2.14	773.2	179.3
	비흡연자	497	807			
선 양	흡연자	913	747	2.18	799.3	149.3
	비흡연자	336	598			
난 징	흡연자	235	172	2.85	203.5	31.1
	비흡연자	58	121			
하 빈	흡연자	402	308	2.32	355.0	57.1
	비흡연자	121	215			
쩡조우	흡연자	182	156	1.59	169.0	28.3
	비흡연자	72	98			
타이윈	흡연자	60	99	2.37	53.0	9.0
	비흡연자	11	43			
난 창	흡연자	104	89	2.00	96.5	11.0
	비흡연자	21	36			

1. 중국 도시별 흡연과 폐암자료(8개 도시 대상)

- 2 x 2 x K 분할표의 K개의 조건부 오즈비에 대한
  조건부 독립성 검정과 동질적 연관성 검정을 살펴보고자 함

- K개의 부분분할표로부터
  계산된 표본오즈비를 결합하여 부분연관성에 대한
  하나의 요약된 측도를 구하고자 함

- X : 흡연여부(설명변수)
  Y : 폐암 발병 여부(반응 변수)
  Z : 도시(통제변수)

2. 코크란-맨틀-핸첼 검정

- 2 x 2 x K 분할표에서 Z가 주어졌을 때
  X와 Y가 조건부 독립이라는 귀무가설을 검정하고자 함

- “X와 Y가 조건부 독립성”
  ⇔ “X와 Y간의 조건부 오즈비 $\theta_{XY(k)}$ 가
  모든 부분분할표에서 1이 됨”

- **표본추출모형**
  : 각 칸 도수에 대해 ① 독립 포아송분포,
  ② 전체 표본크기가 고정된 다항분포,
  ③ 각 부분분할표가 정해진
  표본크기를 갖고 서로 독립인 다항분포,
  ④ 각 부분분할표가 행의 합이 고정되어 있고
  서로 독립인 이항분포 등으로 간주할 수 있음

2. 코크란-맨틀-핸첼 검정

K번째 부분분할표

흡연		폐암		합계
		예	아니오	
k	흡연자	n_{11k}	n_{12k}	n_{1+k}
	비흡연자	n_{21k}	n_{22k}	n_{2+k}
합계		n_{+1k}	n_{+2k}	n_{++k}

- 행합계 $\{n_{1+k}, n_{2+k}\}$, 열합계 $\{n_{+1k}, n_{+2k}\}$가 주어지면 n_{11k}는 다른 모든 칸 도수를 결정함
- n_{11k}는 초기하 분포를 따름

2. 코크란-맨틀-핸첼 검정

- 귀무가설(X와 Y가 조건부 독립성) 하에서

$$\mu_{11k} = E(n_{11k}) = \frac{n_{1+k} n_{+1k}}{n_{++k}}$$

$$Var(n_{11k}) = \frac{n_{1+k} n_{2+k} n_{+1k} n_{+2k}}{n_{++k}^2 (n_{++k} - 1)}$$

- k번째 부분분할표에서 오즈비 $\theta_{XY(k)}$ 가 1.0보다 크면 $(n_{11k} - \mu_{11k}) > 0$ 이 기대됨

➔ $\sum_k (n_{11k} - \mu_{11k})$ 값이 클수록

귀무가설이 옳지 않다는 증거

2. 코크란-맨틀-핸첼 검정

검정통계량

- $CMH = \frac{[\sum_k (n_{11k} - \mu_{11k})]^2}{\sum_k Var(n_{11k})}$: 코크란-맨틀-핸첼 통계량
- 표본크기가 클 때 $CMH \sim \chi^2(1)$
- 각 부분분할표의 연관성이 유사하게 나타날 때 CMH방법 은 각 각 부분분할표에 대한 개별적인 검정보다 우수함
- 단순하게 모든 부분분할표를 합하여 2 x 2 분할표를 만들어 검정하는 것은 부적절한 방법일 수 있음 (심프슨의 역설 참고)

3. 폐암에 대한 메타분석 예제

사례

- 흡연과 폐암에 관해
  중국의 8개 도시에서 행해진 사례-대조연구

- 각 도시별로 폐암환자와
  그렇지 않은 환자를 대응시켜
  과거의 흡연 여부를 조사한 것

- 흡연과 폐암간의 조건부 독립성,
  즉, 각 도시의 오즈비가 1.0이라는
  가설을 검정하고자 함

3. 폐암에 대한 메타분석 예제

SAS프로그램

```
DATA cmh ;
INPUT center smoke cancer count @@ ;
CARDS;
1 1 1 126  1 1 2 100  1 2 1  35   1 2 2   61
2 1 1 908  2 1 2 688  2 2 1 497  2 2 2 807
3 1 1 913  3 1 2 747  3 2 1 336  3 2 2 598
4 1 1 235  4 1 2 172  4 2 1  58   4 2 2 121
5 1 1 402  5 1 2 308  5 2 1 121  5 2 2 215
6 1 1 182  6 1 2 156  6 2 1  72   6 2 2  98
7 1 1  60  7 1 2  99   7 2 1  11   7 2 2  43
8 1 1 104  8 1 2  89   8 2 1  21   8 2 2  36
run;
PROC FREQ ;
  WEIGHT count ;
  TABLES center*smoke*cancer / cmh ;
Run;
```

3. 폐암에 대한 메타분석 예제

SAS 분석결과

Cochran-Mantel-Haenszel 통계량(테이블 스코어에 기반한)				
통계량	대립가설	자유도	값	확률
1	영(0)이 아닌 상관계수	1	280.1375	<.0001
2	행 평균 스코어 차이	1	280.1375	<.0001
3	일반 연관성	1	280.1375	<.0001

- 조건부 독립성의 가설을 매우 뚜렷하게 기각함
- 도시별 자료를 결합하여 얻은
  매우 큰 표본크기 n=8,419인 경우
  매우 유의한 결과를 얻게 됨

3. 폐암에 대한 메타분석 예제

메타분석 (Meta Analysis)

- 여러 연구 결과로부터 얻은 정보를 결합하여 분석하는 통계분석
- 제시한 사례는 각각의 분할표에서 보여주는 것보다 더 강한 연관성의 증거를 보여줌

4. 공통오즈비의 추정

공통오즈비 추정의 필요성

- 단순히 연관성에 대한 가설을 검정하는 것보다
  연관성의 강도를 추정하면 더 많은 정보를 얻을 수 있음

- 모든 부분분할표에서 연관성이 유사하게 나타나면
  K개 오즈비들의 공통값을 추정할 수 있음

4. 공통오즈비의 추정

맨틀 - 핸첼 공통오즈비의 추정값

- 맨틀-핸첼 공통오즈비의 추정값이란?
  : 2 x 2 x K분할표에서 $\theta_{XY(1)} = \cdots = \theta_{XY(K)}$ 일 때

- $$\widehat{\theta_{MN}} = \frac{\sum_K (n_{11K} n_{22K} / n_{++K})}{\sum_K (n_{12K} n_{21K} / n_{++K})}$$

- $\widehat{\log\theta_{MN}}$ 의 표준오차는 복잡함
  (SAS FREQ 절차문 이용)

4. 공통오즈비의 추정

SAS 분석결과

상대 리스크의 추정값(행1/행2)				
통계량	방법	값	95% 신뢰한계	
오즈비	Mantel-Haenszel	2.1745	1.9840	2.3832
	로짓	2.1734	1.9829	2.3823
상대 리스크(칼럼 1)	Mantel-Haenszel	1.5192	1.4417	1.6008
	로짓	1.5132	1.4362	1.5942
상대 리스크(칼럼 2)	Mantel-Haenszel	0.6999	0.6721	0.7290
	로짓	0.7011	0.6734	0.7300

4. 공통오즈비의 추정

- $\log\widehat{\theta_{MN}}$, 공통오즈비 θ에 대한 95% 신뢰구간(1.98, 2.38)
  ➔ 흡연자가 폐암에 걸릴 오즈는 비흡연자의 약 2배임

- 부분분할표에서 오즈비가 크게 다른 경우가 아니라면
  $\widehat{\theta_{MN}}$는 K개 조건부 연관성을 요약하는 효과적인 방법임

5. 공통오즈비에 대한 브레슬로 - 데이 검정

귀무가설

- 변수 Z의 모든 수준에서 X-Y 오즈비는 동일함

$\Leftrightarrow H_0 = \theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$

5. 공통오즈비에 대한 브레슬로 - 데이 검정

- $\{\widehat{\mu_{11k}}, \widehat{\mu_{12k}}, \widehat{\mu_{21k}}, \widehat{\mu_{22k}}\}$ k번째 부분분할표에서의 추정 기대도수
  - 관측된 자료와 동일한 행 및 열의 주변합 분포를 갖고, 공통오즈비 $\widehat{\theta_{MN}}$ 를 갖는다는 조건에서 추정된 기대도수임

5. 공통오즈비에 대한 브레슬로 - 데이 검정

Breslow-Day 통계량

- Breslow-Day 통계량 = $\sum_{i,j,k} \frac{(n_{ijk} - \hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}}$
- n_{ijk} 가 $\hat{\mu}_{ijk}$ 에 가까울수록,
  이 통계량 값은 작아지고, H_0를 반증하는 증거가 약해짐
- 근사적으로 df=K-1인 카이제곱분포를 따름
- SAS FREQ문을 통해서 계산

5. 공통오즈비에 대한 브레슬로 - 데이 검정

SAS 분석결과

오즈비의 동질성에 대한 Breslow-Day 검정	
카이제곱	5.1997
자유도	7
Pr > ChiSq	0.6356

- 부분분할표의 오즈비가 동일한지를 검정하는 브레슬로 - 데이검정 결과
- P-값이 0.6356이므로 오즈비는 동일하다고 볼 수 있음

5. 공통오즈비에 대한 브레슬로 - 데이 검정

R을 이용한 CMH 검정

- R에서 CMH 검정을 하기 위해서는

  mantelhaen.test 이용 :

  Cochran-Mantel-Haenszel Chi-Squared Test

  for Count Data

5. 공통오즈비에 대한 브레슬로 - 데이 검정

R을 이용한 CMH 검정

- Breslow-Day 검정

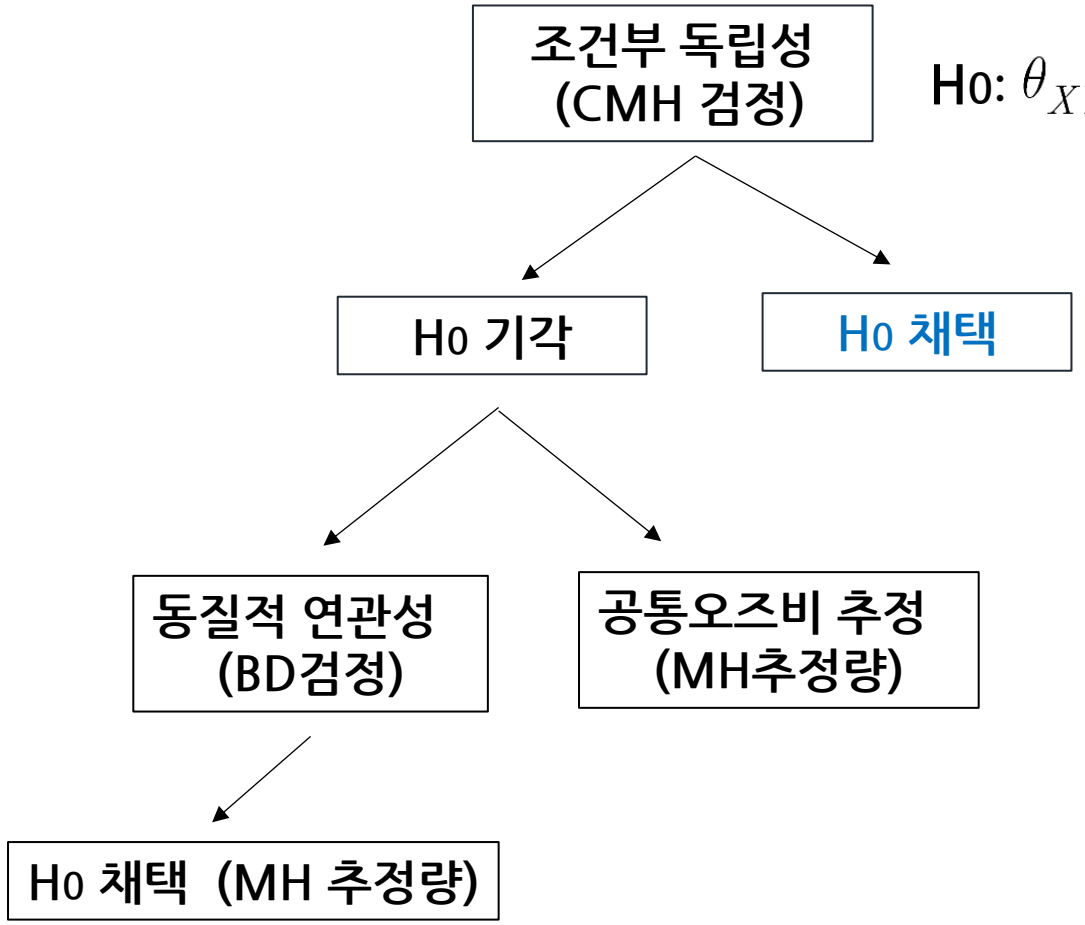
➔ use BreslowDayTest( ) from the DescTools package.

Requires a 2 x 2 x k table as its main argument.

〈참고〉 xtabs( ) : To create 2 x 2 x k tables from a data-set

6. Summary

요약


조건부 독립성
(CMH 검정)
H0: $\theta_{XY(1)} = \cdots = \theta_{XY(K)} = 1.0$
H0 기각
H0 채택
동질적 연관성
(BD검정)
공통오즈비 추정
(MH추정량)
H0 채택 (MH 추정량)

04 강

다음시간안내

일반화선형모형

수고하셨습니다.