01 강

데이터분석방법론2

이차원 분할표(1)

통계·데이터과학과 이기재 교수

한국방송통신대학교 대학원

학습목차

학습개요 및 목표

이번 강의는 범주형 자료분석의 개념과 이차원 분할표에 대한 분석방법에 대하여 학습하도록 하겠습니다. 범주형 자료의 확률분포와 비율 추론에 대해서 살펴보고, 이차원 분할표에 대한 분석방법을 살펴보겠습니다.

1. 범주형 자료의 확률분포를 설명할 수 있다.
2. 이차원 분할표에 대한 확률구조를 설명할 수 있다.
3. 오즈비의 개념을 설명할 수 있다.

제 1장. 서론

1 범주형 데이터와 분석 개요

2 표본추출모형

3 비율에 대한 추론

01

제 1장. 서론

범주형 데이터와 분석 개요

1. 범주형 자료분석

- t-검정, ANOVA, 선형회귀모형 등은
  모두 반응변수(종속변수)가 연속형(정규분포를 따름) 변수를 가정함

- 비모수적 분석법도 종속변수가 연속형(이산형 포함)이거나
  적어도 순서형 척도로 측정된 변수에 적용됨

- 범주형 데이터는 종속변수가 범주형인 경우를 말하며,
  이 과목에서 주로 다루게 됨

2. 측정수준에 따른 변수 유형

- 범주형 (Categorical)
  - binary
  - nominal
  - ordinal
- 양적 변수 (Quantitative)
  - discrete
  - continuous

이산형 확률변수

2개 범주 +

여러 개 범주 +

순서 중요(order matters) +

수치형(numerical) +

연속형

3. 검정방법 개요

독립변수(설명변수) = predictor

종속변수(반응변수) = outcome

예

BMI= 체중, 나이, 폐경여부(1, 0)

Continuous outcome

Continuous predictors

Binary predictor

4. 분석 대상 변수 유형과 분석 방법

독립변수(설명변수)	종속변수(반응변수)	분석법 또는 연관성 척도
이분형(Dichotomous)	연속형	T-검정
범주형	연속형	ANOVA
연속형(일변량)	연속형	단순선형회귀분석
다변량	연속형	다중선형회귀분석
이분형(Dichotomous)	이분형(Dichotomous)	오즈비, 상대위험도, 차이 검정 등
범주형	범주형	카이제곱검정
다변량	이분형(Dichotomous)	로지스틱회귀분석
범주형	Time-to-event	Kaplan-Meier curve/log-rank test
다변량	Time-to-event	Cox-proportional hazards model

02

제 1장. 서론

표본추출모형

1. 분석에서 분포의 역할

연속형 자료에 대한 회귀분석이나 분산분석에서 정규분포가 중요한 역할을 하듯 범주형 자료에 대한 분석에서는 포아송분포와 이항분포(다항분포)가 중요한 역할

- 포아송 분포(Poisson Distribution)
- 이항분포(Binomial Distribution)
- 다항분포(Multinomial Distribution)

2. 분포의 유형

포아송 분포(Poisson Distribution)

- 어떤 정해진 기간 동안에 발생하는 희귀한 사건의 발생 건수에 대한 확률분포

- Y= 어떤 정해진 기간 동안 관심 사건의 발생 건수
- $P(Y=y)=\dfrac{e^{-\mu}\mu^{y}}{y!},\quad y=0,\ 1,\ 2,\ 3,\ \cdots$
- $E(Y)=\mu,\quad Var(Y)=\mu$

- 평균이 증가함에 따라 분산도 함께 증가함

2. 분포의 유형

이항 분포(Binomial Distribution)

- n 회의 서로 독립인 베르누이 시행
  (각 시행의 결과는 성공 또는 실패임)
- 각 시행에서 성공확률은 동일하며,
  $\pi = P(성공),\ \ 1-\pi = P(실패)$
- $Y=$ " n 회 베르누이 시행에서의 총 성공횟수"
- $P(Y=y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}, \quad y=0,\ 1,\ 2,\ \cdots,\ n$

2. 분포의 유형

이항 분포(Binomial Distribution)

- $E(Y) = n\pi, \quad Var(Y) = n\pi(1-\pi)$

  $p = \hat{\pi} = \frac{Y}{n}$: 성공확률 추정량

  $E(p) = E\left(\frac{Y}{n}\right) = \pi, \quad \sigma(p) = \sqrt{\frac{\pi(1-\pi)}{n}}$

- 각 시행에서 3개 이상의 가능한 결과를 가질 때 여러 개 범주에 대한 발생 건수의 분포는 다항분포(Multinomial Distribution)를 따름
  ➔ 이항분포는 다항분포의 특수한 형태

03

제 1장. 서론

비율에 대한 추론

1. 최대가능도 추정 (최대우도 추정)

가능도 함수 (Likelihood Function, 우도 함수)

- 미지의 모수(Parameter)의 함수로 표현된 관측자료(Observed Data)의 확률

예제

- 이항분포 $Y \sim B(n,\ \pi)$

$$P(Y=y) = \binom{n}{y}\pi^{y}(1-\pi)^{n-y} = l(\pi)$$

이항분포 $n=2,\ \ y=1$ 인 경우

$$p(1) = \frac{2!}{1!1!}\pi^{1}(1-\pi)^{1} = 2\pi(1-\pi) = l(\pi)$$

1. 최대가능도 추정 (최대우도 추정)

최대가능도추정 (Maximum Likelihood Estimator)

- 가능도함수가 최대값을 갖게 하는 모수(Parameter) 값을 추정값으로 정의함

예제

- 이항분포 $n=2,\ \ y=1$ 인 경우

$$p(1)=\frac{2!}{1!1!}\pi^{1}(1-\pi)^{1}=2\pi(1-\pi)=l(\pi)$$

$l(\pi)=2\pi(1-\pi)$ 은 $\hat{\pi}=0.5$ 에서 최대값을 가짐

➔ π 의 ML 추정값은 $\hat{\pi}=0.5$

2. 최대가능도 추정량의 성질

- 이항분포인 경우에 성공확률 π 에 대한 ML 추정량 :

$\hat{\pi} = \dfrac{y}{n}$ (표본 성공비율)

- $y_1, \; y_2, \; \cdots, \; y_n$ 이 정규분포(or 포아송분포)로부터 랜덤표본일 때, μ 에 대한 ML 추정량은 $\hat{\mu} = \bar{y}$ 임

- 표본크기가 클 때 ML추정량은 최적(optimal)의 추정량이고, 근사적으로 정규분포를 따름

3. 비율에 대한 검정

가설

$$H_0 : \pi = \pi_0 \quad \text{VS} \quad H_1 : \pi \neq \pi_0$$

검정통계량

$$z = \frac{p - \pi_0}{\sigma(p)} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} \simeq N(0,1)$$

4. 비율에 대한 신뢰구간

- **모수 θ에 대한 대표본 신뢰구간 :** $\hat{\theta} \pm z_{\alpha/2}(SE)$

- **모비율 추정**

$$\theta = \pi, \quad \hat{\theta} = \hat{\pi} = p$$

$$\sigma(p) = \sqrt{\frac{\pi(1-\pi)}{n}} \leftarrow SE = \sqrt{\frac{p(1-p)}{n}}$$

➔ 95% 신뢰구간 : $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$

4. 비율에 대한 신뢰구간

➔ 95% 신뢰구간 : $p \pm 1.96\sqrt{\dfrac{p(1-p)}{n}}$

" $\pi < 0.2$ 이거나 $\pi > 0.8$ 때는 표본크기가
상당히 크더라도 실제 포함확률(Coverage Probability)이
0.95에 가깝게 나오지 않을 수도 있다."

4. 비율에 대한 신뢰구간

- 유의성 검정으로부터 신뢰구간을 만드는 방법

95% 신뢰구간 :
유의수준 0.05에서 귀무가설을 “기각하지 않은 모든 π_0값을 포함하는 구간”

$\Rightarrow \dfrac{|p-\pi_0|}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}}$ = 1.96을 만족하는 π_0값을 구함

- 대략적으로 $np>5,\ n(1-\pi)>5$ 을 만족할 때
앞서 구한 두 가지 신뢰구간은 정확한 결과를 줌

제 2장. 이차원 분할표(1)

01

제 2장. 이차원 분할표(1)

분할표의 확률 구조

1. 분할표 예제

성별	사후세계에 대한 믿음	
	예	아니오 또는 불확실
여성	435	147
남성	375	134

분할표(Contingency Table)

- I개 행과 J개 열로 이루어진 이차원 분할표를 I x J 분할표라고 함
- 두 개의 범주형 변수 X, Y에 대해서 X(I개 수준)을 행에, Y(J개 수준)을 열에 표시하면 I x J 분할표를 얻게 됨

2. 결합확률, 주변확률, 조건부 확률

- $\{\pi_{ij}\}$= $\{P(X=i,\ Y=j)\}$: 확률변수 X와 Y의 결합분포

	π_{11}	π_{12}	π_{1+}
	π_{21}	π_{22}	π_{2+}
	π_{+1}	π_{+2}	1.0

- 주변분포(marginal distribution)
  - $\{\pi_{i+}\}$, $\{\pi_{j+}\}$로 표현

2. 결합확률, 주변확률, 조건부 확률

- $\{n_{ij}\}$: 각 칸 도수 (cell counts)
- $\{p_{ij}\}$: 칸 비율 (cell proportions)

$$p_{ij} = \frac{n_{ij}}{n} \quad , \quad n = \sum_i \sum_j n_{ij}$$

- 조건부분포 : X가 x값으로 고정되었을 때 Y의 분포
- 사후세계 자료

성별	사후세계에 대한 믿음		합계
	예	아니오 또는 불확실	
여성	$n_{11} = 435$	$n_{12} = 147$	$n_{1+} = 582$
남성	$n_{21} = 375$	$n_{22} = 134$	$n_{2+} = 509$
합계	$n_{+1} = 810$	$n_{+2} = 281$	$n = 1091$

2. 결합확률, 주변확률, 조건부 확률

예제 : Diagnostic disease test

- Y = 진단결과 : 1 = 양성(Positive)   2 = 음성(Negative)
  X = 실제 : 1 = diseased   2 = not diseased

		Y(진단결과) 1	2	
X (실제)	1			
	2			

- 민감도(Sensitivity) = P(Y=1|X=1)
  특이도(Specificity) = P(Y=2|X=2)

"만약 진단결과로 양성이라면 P(X=1| Y=1)에 관심이 있음"

3. 독립성

- X와 Y는 통계적 독립

⇔ Y의 조건부 확률이 X의 각각의 수준에서 동일

⇔ $\pi_{ij} = \pi_{i+} \cdot \pi_{+j}$ 모든 i,j

		Y(진단결과)		
		0	1	
X (실제)	0	0.28	0.42	0.7
	1	0.12	0.18	0.3
		0.4	0.6	1.0

4. 포아송분포, 이항분포, 다항분포

- **독립이항 표본추출 (Independent Binomial Sampling)**
  - 각 행의 표본이 서로 독립이고,
    각 행의 표본에 대해 이항분포를 가정할 수 있는 경우

- **다항표본추출 (Multinomial Sampling)**
  - 분할표에서 전체 표본크기만 고정되어 있는 경우

- **각 표본추출 모형에 대한**
  주요 추론 방법들의 결과는 동일함

02

제 2장. 이차원 분할표(1)

2 X 2 분할표에서 비율 비교

1. 비율의 차이

이항변수 (Binary Variable)

- 두 개의 범주를 갖는 반응변수

		Y	
		S	F
X	1	π_1	$1-\pi_1$
	2	π_2	$1-\pi_2$

- $\pi_1 - \pi_2$ 의 추정

: $\hat{\pi}_1 - \hat{\pi}_2 = p_1 - p_2$

: $SE(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

2. 아스피린과 심장마비의 예제

- 미국 하버드 의과대학 내과의사 연구그룹의 연구결과
- 눈가림법에 의한 5년 연구

	심근경색		총계
	예	아니오	
위약	189	10,845	11,034
아스피린	104	10,933	11,037

2. 아스피린과 심장마비의 예제

- p_1 = 0.017, p_2 = 0.009, $p_1 - p_2$ = 0.008
- $SE = \sqrt{\dfrac{0.017 \times 0.983}{11.034} + \dfrac{0.009 \times 0.991}{11.037}}$
- $\pi_1 - \pi_2$에 대한 95% 신뢰구간 :
  $0.008 \pm 1.96(0.0015) = (0.005,\ 0.011)$
  $\Leftrightarrow \pi_1 > \pi_2$

아스피린 복용이 심장혈관질환의 위험을
감소시킨다고 볼 수 있음

3. 상대 위험도 (Relative Risk)

- **상대 위험도(Relative risk) =** $\dfrac{\pi_1}{\pi_2}$

		Y	
		S	F
X	1	π_1	$1-\pi_1$
	2	π_2	$1-\pi_2$

- 아스피린과 심장마비 예제

$$\frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{p_1}{p_2} = \frac{0.017}{0.009} = 1.83$$

3. 상대 위험도 (Relative Risk)

"위약 복용집단에서 심근경색을 일으키는 비율이 아스피린 복용집단에 비해서 약 83%만큼 더 높다."

- 두 비율이 모두 0에 가까울 때 비율의 차이만으로 두 집단을 비교하는 것은 잘못된 결론을 가져올 수 있음

03

제 2장. 이차원 분할표(1)

오즈비

1. 오즈비 (Odds ratio) 란?

	S	F
1	π_1	$1-\pi_1$
2	π_2	$1-\pi_2$

- $odds_1 = \dfrac{\pi_1}{1-\pi_1}$ (첫째 행에서 성공의 오즈)
- $odds_2 = \dfrac{\pi_2}{1-\pi_2}$ (둘째 행에서 성공의 오즈)
- 오즈비(Odds ratio) $\theta = \dfrac{\pi_1 / (1-\pi_1)}{\pi_2 / (1-\pi_2)}$

1. 오즈비 (Odds ratio) 란?

- 상대 위험도는 두 확률의 비인 반면 오즈비 θ는 오즈의 비(比)임
- 아스피린과 심장마비 예제

$$odds_1 = \frac{0.0171}{0.9829} = 0.0174 \quad \text{(위약)}$$

$$odds_2 = \frac{0.0094}{0.9906} = 0.0095 \quad \text{(아스피린)} \qquad \hat{\theta} = \frac{0.0174}{0.0095} = 1.83$$

"위약집단에서 심근경색을 일으킬 오즈는
아스피린 복용 집단의 1.83배로 추정된다."

2. 오즈비의 성질

- 각 $odds \geq 0, \quad \theta \geq 0$
- 두 변수 X와 Y가 서로 독립이면 $\pi_1 = \pi_2$, $odds_1 = odds_2$
  $\Rightarrow \theta = 1$
- 오즈비 θ가 1로부터 멀리 떨어질수록 더 강한 연관성을 나타냄
- 행의 순서가 바뀌거나 열의 순서가 바뀌면 오즈비는 역수가 됨

$$\theta \quad \rightarrow \quad \frac{1}{\theta}$$

$\theta = 3, \ \theta = \frac{1}{3}$ 은 같은 강도의 연관성을 나타내며
행과 열의 배열 방법에 따라 다른 값을 갖게 됨

2. 오즈비의 성질

- 분할표에서 행이나 열을 서로 바꾸더라도 오즈비는 변하지 않음
(열을 반응변수로 행을 설명변수로 다루거나,
행을 반응변수로 열을 설명변수로 다루더라도 같은 오즈비를 갖게 됨)

	S	F
1	n_{11}	n_{12}
2	n_{21}	n_{22}

$$\Rightarrow \hat{\theta} = \frac{n_{11} \,/\, n_{12}}{n_{21} \,/\, n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

교차적비 (Cross-Product Ratio)

2. 오즈비의 성질

- $\theta = 1 \Leftrightarrow \log\theta = 0$
  로그오즈비는 0에 대하여 대칭임
  $\theta = 2 \Rightarrow \log\theta = 0.7$
  $\theta = 1/2 \Rightarrow \log\theta = -0.7$

- $\hat{\theta}$의 표본분포(Sample Distribution)는 오른쪽으로 기울어져 n이 큰 경우에만 근사적으로 정규분포를 따름

2. 오즈비의 성질

- $\log\hat{\theta}$ 의 표집분포는 정규분포에 더 근사됨

$\log\hat{\theta}$ 의 점근적 표준오차

$$ASE(\log\hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

$\log\hat{\theta}$ 의 신뢰구간

$$\log\hat{\theta} \pm z_{\alpha/2} \times ASE(\log\hat{\theta}) \Leftrightarrow (L, U)$$

θ 의 신뢰구간

$$(e^{L},\ \ e^{U})$$

2. 오즈비의 성질

- 예제

$$\hat{\theta} = \frac{189 \times 10933}{104 \times 10845} = 1.83$$

$$\log\hat{\theta} = 0.605$$

$$ASE(\log\hat{\theta}) = \sqrt{\frac{1}{189} + \frac{1}{10933} + \frac{1}{104} + \frac{1}{10845}} = 0.123$$

θ의 95% 신뢰구간

$$(e^{0.365},\ \ e^{0.846}) = (1.44,\ 2.33)$$

apparently $\theta > 1$

2. 오즈비의 성질

- $\hat{\theta}$은 θ에 대한 신뢰구간의 중간지점이 아님
  (∵ $\hat{\theta}$ 의 표집분포가 오른쪽으로 길게 기울어짐)

- 어떤 $n_{ij} = 0$이면 $\{n_{ij}\}$ 대신에 $\{n_{ij} + 0.5\}$ 를
  사용하여 추정값과 표준오차 추정값을 계산하는 것이 바람직함

3. 오즈비와 상대위험의 관계

- 오즈비 = $\dfrac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$ = 상대위험도 X $\dfrac{(1-\pi_2)}{(1-\pi_1)}$

- π_1 과 π_2가 모두 0에 가까우면

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

4. 사례: 사례대조 연구 (Case-control study)

이탈리아 북부지방에서 심근경색으로 치료를 받은 262명의 69세 이하 중년여성 환자들(사례집단)과 이들 각각에 대해서 같은 병원에 다른 질병으로 입원한 두 명씩의 환자들(대조집단)을 대응시켜 조사대상 선정

4. 사례: 사례대조 연구 (Case-control study)

	심근경색	
	예	아니오
예	172	173
아니오	90	346
합계	262	519

후향적 설계 (Retrospective Design)

- 독립 이항표본추출 모형 적용 사례
- 심근경색 발병여부를 반응변수, 흡연상태는 설명변수로 간주

4. 사례: 사례대조 연구 (Case-control study)

- 실험설계의 특성상

$$P(Y=y|x),\ \pi_1-\pi_2=P(Y=yes|X=yes)-P(Y=yes|X=no),$$

π_1/π_2 등을 추정할 수 없음

- $P(X|Y)$만 추정 가능

$$\hat{\theta}=\frac{\hat{P}(X=yes|Y=yes)/\hat{P}(X=no|Y=yes)}{\hat{P}(X=yes|Y=no)/\hat{P}(X=no|Y=no)}$$

$$=\frac{(172/262)/(90/262)}{(173/519)/(346/519)}$$

$$=\frac{172\times 346}{173\times 90}=3.82$$

4. 사례: 사례대조 연구 (Case-control study)

- $P(Y = yes|X) \cong 0$

(즉, 중년여성의 심근경색 발생 확률은 흡연여부와 무관하게 낮음)

$\Rightarrow \theta \approx \pi_1 / \pi_2$

" 흡연 여성은 비흡연 여성에 비해 심근경색 위험이 약 4배 높다"

5. 연구방법의 구분

- **관측연구 (Observation Studies)**
  - 누가 어떤 그룹에 속하는 지와 어떤 반응 결과를 가지는가를 관측
- **실험연구 (Experimental Studies)**
  - 각 개체를 어떤 그룹(처리)에 포함시킬 것인지를 연구자가 결정하여 실험
    (예 : 어떤 사람에게 아스피린이나 위약
    어느 것을 투여할지 연구자가 결정)

실험연구는 확률화 원리를 기초로 하기 때문에
잠재적인 함정에 빠질 위험이 거의 없지만,
의학연구나 사회과학 분야에서는
관측연구가 널리 사용되고 있음

6. 연구 진행 유형에 따른 구분

전향적 연구 (Prospective study)

- 연구의 진행방향이 시간의 흐름에 따라 진행함
- 코호트(Cohort) 연구는 전향적 연구의 대표적 사례임

횡단면적 연구 (Cross-Sectional Study)

- 조사대상을 표본추출하여
  동시에 설명변수와 반응변수에 따라 분류하여 분석
- 대표적인 사례로는 표본조사가 있음

후향적 연구 (Retrospective Study)

- 사례-대조(Case-control)연구가 대표적 유형임
- 조사대상을 사례와 대조집단으로 구분하고
  시간을 거슬러 연구 진행

04

제 2장. 이차원 분할표(1)

데이터 분석 실습

데이터 분석 실습

1. Data

	심근경색		총계
	예	아니오	
위약	189	10,845	11,034
아스피린	104	10,933	11,037

데이터 분석 실습

2. SAS

```
Data aspirin;
Input group mi count @@;
CARDS;
 1    1   189    1   2   10845
 2    1   104    2   2   10933
;
RUN;
PROC FREQ order=data;
        WEIGHT count;
        TABLES group*mi/measures nocol nopercent;
RUN;
```

데이터 분석 실습

3. 분석결과

테이블 : group * mi			
group	mi		
	1	2	총합
1	189 1.71	10,845 98.29	11,034
2	104 0.94	10,933 99.06	11,037
총합	293	21,778	22,071

오즈비 및 상대 리스크			
통계량	값	95% 신뢰한계	
오즈비	1.8321	1.4400	2.3308
상대 리스크 (칼럼1)	1.8178	1.4330	2.3059
상대 리스크 (칼럼2)	0.9922	0.9892	0.9953

표본 크기 = 22,071

정리

제 1강. 이차원 분할표(1)

Summary & Learning Content

정리하기

- 최대가능도 추정법 (Maximum Likelihood Estimator)
- 두 변수의 독립성
- 상대위험도 (Relative Risk)
- 오즈비 (Odds Ratio)
- 오즈비와 상대위험의 관계
- 연구방법의 구분
  -관측연구 (Observational Studies)
  -실험연구 (Experimental Studies)
  -전향적 연구 (Prospective Study)
  -횡단면 연구 (Cross-Sectional Study)
  -후향적 연구 (Restrospective Study)

02 강

다음시간안내

이차원 분할표(2)

수고하셨습니다.