02 강

데이터분석방법론2

이차원 분할표(2)

통계·데이터과학과 이기재 교수

학습목차

학습개요 및 목표

범주형 자료는 이차원 분할표로 요약하여 나타낼 수 있습니다.
오늘 강의는 I × J 이차원 분할표에 대한 독립성 검정방법,
설명변수와 반응변수가 순서형 변수인 경우의 독립성 검정법,
소표본 범주형 자료에 대한 정확 추론 방법을 학습하겠습니다.

1. 이차원 분할표에 대한 카이제곱 검정법을 적용할 수 있다.
2. 순서형 자료에 대한 독립성 검정법을 설명할 수 있다.
3. 이차원 분할표에 대한 정확 추론 방법을 설명할 수 있다.

제 2장. 이차원 분할표(2)

1 독립성에 대한 카이제곱 검정

2 순서자료에 대한 독립성 검정

3 소표본에 대한 정확 추론

01

제 2장. 이차원 분할표(2)

독립성에 대한 카이제곱 검정

예제

▪ 성별에 따른 지지정당의 분할표

성별(X)	지지정당 (Y)		
	민주당	공화당	무소속
여성	495	272	590
남성	330	265	498
합계	825	537	1088

Data : General Social Survey (2016)

1. 피어슨 통계량과 카이제곱 분포

- 가설 형태

H0 : X(성별)과 Y(지지정당)은 서로 독립
H1 : X(성별)과 Y(지지정당)은 서로 연관 됨

1. 피어슨 통계량과 카이제곱 분포

- H_0가 성립한다는 가정

① $P(X=i, Y=j) = P(X=i)P(Y=j)$

$\Leftrightarrow \pi_{ij} = \pi_{i+}\pi_{+j}$

② 기대도수(Expected Frequency)

$$\mu_{ij} = n\pi_{ij}$$

$$= n\pi_{i+}\pi_{+j} \quad \text{under } H_0$$

③ 기대도수 추정값

$$\widehat{\mu_{ij}} = n\widehat{\pi_{i+}}\widehat{\pi_{+j}}$$

$$= n\left(\frac{n_{i+}}{n}\right)\left(\frac{n_{+j}}{n}\right) = \frac{n_{i+}n_{+j}}{n}$$

1. 피어슨 통계량과 카이제곱 분포

카이제곱 검정 통계량

$$X^2 = \sum_{all\,cells} \frac{(n_{ij} - \widehat{\mu_{ij}})^2}{\widehat{\mu_{ij}}}$$

: Chi-squared statistics
(Karl Pearson, 1900)

- X^2은 표본크기가 클 때 자유도 $df = (I-1)(J-1)$ 인 카이제곱분포를 따름
- $\{n_{ij}\}$ 와 $\{\widehat{\mu_{ij}}\}$간의 차이가 클수록 귀무가설 H_0 가 옳지 않다는 증거가 뚜렷해짐

$$P-값 = P(X^2 \geq X^2\ observed)$$

1. 피어슨 통계량과 카이제곱 분포

▪ 성별과 지지정당 관계 예제

성별	지지정당			합계
	민주당	공화당	무소속	
여성	495 (456.9)	272 (297.4)	590 (602.6)	1357
남성	330 (368.1)	265 (239.6)	498 (485.4)	1093
합계	825	537	1088	2450

1. 피어슨 통계량과 카이제곱 분포

- 괄호 안의 수치는 기대도수임

예 $\widehat{\mu_{11}} = 2450 \times \frac{495}{825} \times \frac{495}{1357} = 456.9$

- $X^2 = \sum \frac{(n_{ij} - \widehat{\mu_{ij}})^2}{\widehat{\mu_{ij}}} = 12.57$

$df = (I-1)(J-1) = 1 \times 2 = 2$

$P-\text{값} = P(X^2 \geq 12.57) = 0.002$

⟺ 성별과 정당선호도는 서로 독립이 아님

1. 피어슨 통계량과 카이제곱 분포

카이제곱 분포

- 카이제곱 분포 개형

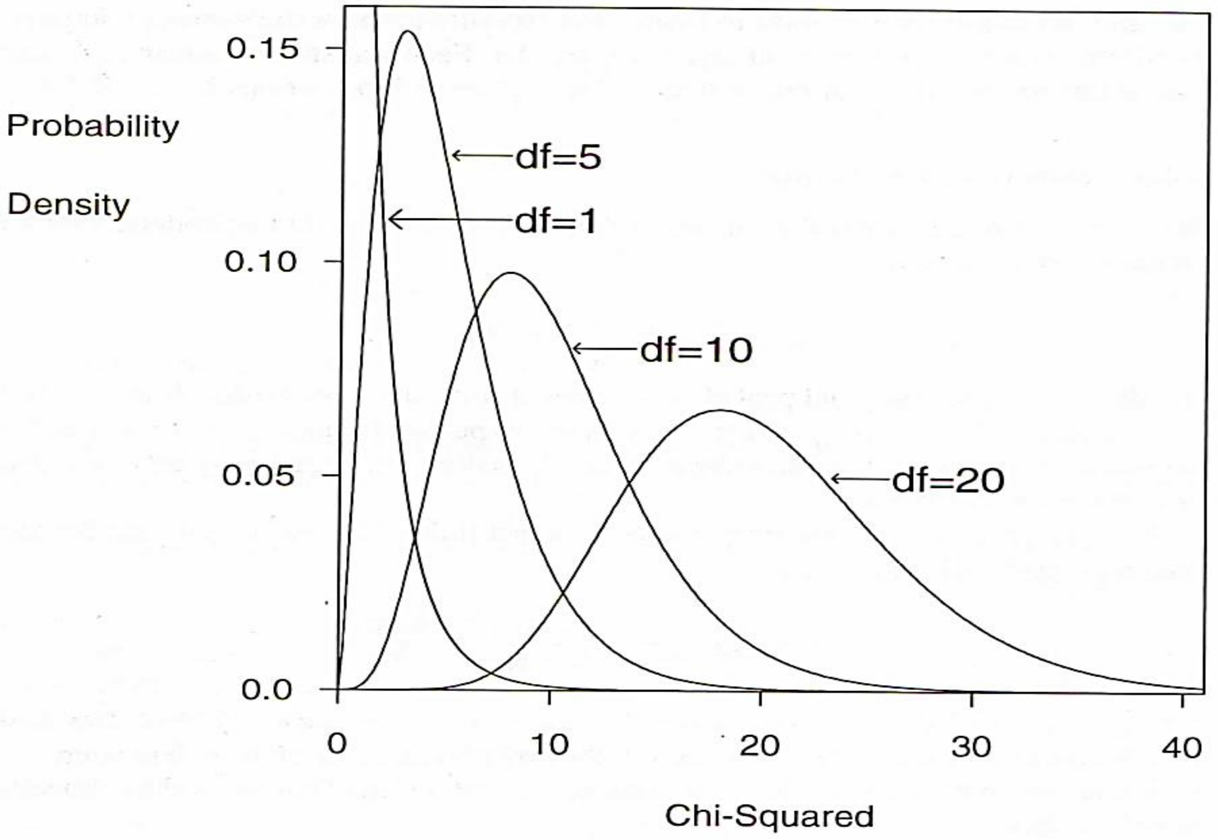


Figure 2.1 Examples of chi-squared distributions.

1. 피어슨 통계량과 카이제곱 분포

카이제곱 분포

- 카이제곱 분포는 자유도(degrees of freedom : df)에 의해서 분포 개형이 결정됨

$$\mu = df \ , \ \sigma = \sqrt{2df}$$

- 자유도가 커짐에 따라 분포 모양은 종모양(Bell-shaped)에 가깝게 됨

2. 가능도비 검정 (Likelihood Ratio Test)

가능도비 검정의 일반 형태

$$\Lambda = \frac{\text{모수가 } H_0\text{를 만족할 때의 최대가능도값}}{\text{모수에 대한 제한조건이 없는 상황에서 구한 최대가능도값}}$$

- 가능도비 Λ는 1을 초과할 수 없음
- 모수가 H_0를 만족하지 않을 때 가능도비 Λ는 1보다 훨씬 작아질 것임
  ➔ H_0 에 대한 강한 반증
- 가능도비 검정통계량은 $-2log(\Lambda)$ 와 동치
- Λ가 "작은" 값 ⇔ $-2log(\Lambda)$ 는 "큰" 값

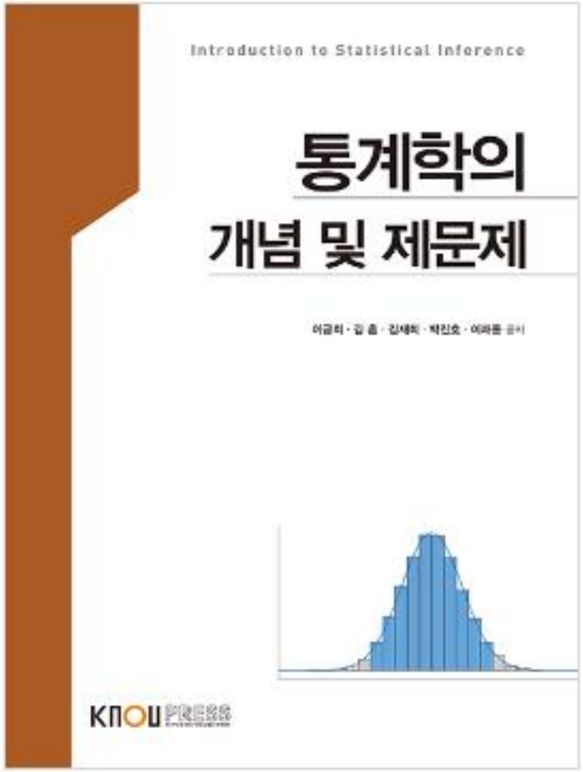
2. 가능도비 검정 (Likelihood Ratio Test)

이차원 분할표에 대한 가능도비 검정

- 가능도비 카이제곱 통계량 $G^2 = 2\sum n_{ij}\log(\frac{n_{ij}}{\widehat{\mu_{ij}}})$
- G^2은 근사적으로 자유도 $df = (I-1)(J-1)$인 카이제곱 분포를 따름
- $n_{ij} = \widehat{\mu_{ij}}$ 일 때, $G^2 = X^2 = 0$으로 최소값을 가짐
- G^2 이 클수록 H0에 대한 강한 반증이 됨
- X^2 와 G^2는 서로 별개의 통계량이지만 공통된 성질이 많고, 독립성이 검정에 대해서 대개 같은 결론을 줌

2. 가능도비 검정 (Likelihood Ratio Test)

가능도비 검정과 G^2에 대한 유도과정



통계학의 개념 및 제 문제
(이긍희 등, 2019) P279 ~ 281 참고

2. 가능도비 검정 (Likelihood Ratio Test)

독립성 검정을 위한 SAS 절차문

```
DATA party;
 INPUT gender $ id $ count @@ ;
 CARDS ;
 f   demo 495 f   repub 272 f   indep 590
 m demo 165 m repub 265 m indep 498
 ;
 PROC FREQ data=party order = data ;
    WEIGHT count ;
    TABLES gender*id/chisq expected nocol norown opercent ;
 RUN ;
```

2. 가능도비 검정 (Likelihood Ratio Test)

분석 결과

테이블 : gender * id				
gender	id			
	demo	repub	indep	총합
f	495 456.95	272 297.43	590 602.62	1357
m	330 368.05	265 239.57	498 485.38	1093
총합	825	537	1088	2450

통계량	자유도	값	유의확률
카이제곱	2	12.569	0.0019
우도비 카이제곱	2	12.601	0.0018

3. 잔차

분할표에서 잔차 계산의 의미

- 각 칸별로 관측값과 기대값을 비교하면 검증결과를 더 잘 이해할 수 있음
- $n_{ij} - \widehat{\mu_{ij}}$만을 고려하는 것은 충분한 정보를 주지 못함 (기대도수가 크게 되는 칸에서는 $n_{ij} - \widehat{\mu_{ij}}$ 도 커지는 경향)

3. 잔차

수정잔차 (Adjusted Residual)

$$r_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - p_{i+})(1 - p_{+j})}}$$

- 귀무가설 하에서 r_{ij} 는 표준정규분포 N(0, 1)을 따름
- $|r_{ij}| > 2 \ \text{or} \ 3$ ➔ H_0가 적합하지 않다는 것을 나타냄

3. 잔차

예제

성별	지지정당		
	민주당	공화당	무소속
여성	3.27	-2.50	-1.03
남성	-3.27	2.50	1.03

독립성 가설(H0)이 참일 때 기대되는 도수에 비해 여성은 민주당을, 남성은 공화당을 지지하는 경우가 월등히 많았음

3. 잔차

- SAS : PROC GENMOD에서 잔차값 제공

$$l_{ij} = \frac{n_{ij} - \widehat{\mu_{ij}}}{\sqrt{\widehat{\mu_{ij}}}}$$

- R : chisq.test 함수를 이용하면
  카이제곱 검정, 수정잔차 등을 구할 수 있음

- 교재 p63 R 프로그램 참고

4. 카이제곱의 분할

카이제곱 통계량의 특성

- 서로 독립인 카이제곱 통계량들은 다음 관계 만족

$$\chi_a^2 + \chi_b^2 = \chi_{a+b}^2$$

- $df > 1$ 인 카이제곱 통계량은 더 작은 자유도를 갖는 여러 카이제곱 통계량으로 분할 가능

- 카이제곱 통계량의 분할을 통해서 특정한 범주들 간의 차이 혹은 범주들 간의 그룹간의 차이를 밝혀낼 수 있음

4. 카이제곱의 분할

$2 \times J$ 분할표에서 독립성 검정을 위한 G^2통계량의 분할

- 검정통계량의 자유도는 $df = (J-1)$ 이므로 $J-1$개로 분할 가능
- 원래의 분할표의 첫 번째 두 열을 비교하는 G^2값,
  첫 번째 두 열을 합한 후 세 번째 열과 비교하는 G^2 값,
  같은 방법으로 $J-1$번째 열까지 합한 후 마지막 J 번째 열과 비교하는 G^2 값으로 분할
- 각 G^2는 자유도가 1이고,
  합하면 $2 \times J$ 분할표에서 G^2값과 일치함

4. 카이제곱의 분할

$2 \times J$ 분할표에서 G^2 통계량의 분할

성별	지지정당		
	민주당	공화당	무소속
여성	495	272	590
남성	330	265	498

$$G^2 = 12.60, \quad df = 2$$

4. 카이제곱의 분할

1단계

성별	지지정당	
	민주당	공화당
여성	495	272
남성	330	265

$$G^2 = 11.536, \quad df = 1$$

4. 카이제곱의 분할

2단계

성별	지지정당	
	민주당+공화당	무소속
여성	767	590
남성	595	498

$$G^2 = 1.065, \quad df = 1$$

4. 카이제곱의 분할

- G^2 의 부분들의 합은 원래 2x3 분할표에서 독립성 검정을 위한 G^2값과 일치함

- 민주당과 공화당은 지지하는데 성별의 차이가 있지만, (민주당+공화당)과 무소속은 성별의 차이가 없음

참고

- G^2 는 정확하게 분할되지만 X^2 통계량은 정확하게 분할되지 않음
- 하지만 각 분할된 표에서 X^2 통계량을 이용하여 추론하는 것은 타당함

5. 카이제곱 검정에 관한 추가사항

- **단순히 검정결과에 의존하지 말고 연관성의 본질을 연구해야 함**
  - 카이제곱 분해, 잔차 계산, 연관성의 정도를 나타내는 오즈비 추정

- **X^2과 G^2 카이제곱 검정을 적용할 수 있는 자료 유형에 대한 고려**
  - 칸의 수 $I \times J$ 에 비해 표본크기 n이 상대적으로 클 때
    X^2과 G^2의 분포가 카이제곱 분포에 잘 근사됨
  - 칸의 수 $I \times J$ 에 비해 n이 작다면
    X^2 통계량을 이용하거나
    소표본 정확검정(Exact Test)을 이용할 것

5. 카이제곱 검정에 관한 추가사항

순서형인 두 변수의 독립성 검정

- X^2 와 G^2 의 계산 $\left(\widehat{\mu_{ij}} = n_{i+}n_{+j}/n\right)$은 열과 행들의 주변합에 의존하지만 열과 행의 나열된 순서와는 무관함
  ➔ *두 변수의 범주를 명목형으로 다루고 있음*

- 순서형인 두 변수에 대해서 독립성 검정을 위해 X^2 나 G^2 을 이용하는 것은 정보 손실이 있게 됨
  ➔ *두 변수가 모두 순서형인 경우는 다른 방법도 가능*

02

제 2장. 이차원 분할표(2)

순서자료에 대한 독립성 검정

순서형 자료의 독립성 검정

검정통계량 X^2와 G^2을 이용한 카이제곱 독립성 검정은
두 변수를 명목형 변수로 간주하고 있음

- 행 또는 열 변수가 순서형 변수일 때는 순서적 개념을 활용한 검정 통계량을 이용하는 것이 바람직함

1. 독립성에 대한 선형추세 대립가설

행변수 X와 열변수 Y가 순서형일 추세 연관성 분석

- X 수준이 증가할 때
  Y 반응수준이 높아지거나 낮아지는 경향이 있는 경우

- 일반적인 분석방법
  : 범주 수준에 점수(Score)를 부여하여
  선형추세(Linear Trend)나 상관관계를 측정함

1. 독립성에 대한 선형추세 대립가설

두 변수 X와 Y간의 양 또는 음의 선형 추세를 찾기 위한 검정 통계량

- 행 점수 : $u_1 \le u_2 \le u_3 \le \cdots \le u_I$
- 열 점수 : $v_1 \le v_2 \le v_3 \le \cdots \le v_J$

행(열) 점수가 범주의 수준과 같은
순서를 갖게 되는 경우를 단조점수(Monotone Score)라고 함

- 선택된 범주점수에 대한 X와 Y간의 피어슨 적률 상관계수

$$r = \frac{\sum_{ij} u_i v_j n_{ij} - (\sum_i u_i n_{i+})(\sum_j v_j n_{+j})/n}{\sqrt{[\sum_i u_i^2 n_{i+} - \frac{(\sum_i u_i n_{i+})^2}{n}][\sum_j v_j^2 n_{+j} - \frac{(\sum_j v_j n_{+j})^2}{n}]}}$$

1. 독립성에 대한 선형추세 대립가설

적률 상관계수 r 의 성질

① $-1 \leq r \leq 1$

② H0 : 상관계수 = 0 VS H1 : 상관계수 ≠ 0
귀무가설 하에서 $M^2 = (n-1)r^2$ 은
근사적으로 자유도 1인 카이제곱 분포를 따름

③ M^2 통계량은 r 또는 n 이 커짐에 따라 증가하게 되고
그 값이 클수록 독립성에 위배됨을 나타냄

④ $M = \sqrt{n-1} \times r$ 은 귀무가설 하에서
근사적으로 표준정규분포를 따름

2. 음주와 선천성 기형에 대한 연구

임산부의 알코올 섭취량과 신생아의 기형 여부

알콜량	기형여부		총계	기형비율 (%)	수정잔차
	없음	있음			
0	17,066	48	17,114	0.28	-0.18
<1	14,464	38	14,502	0.26	-0.71
1–2	788	5	793	0.63	1.84
3–5	126	1	127	0.79	1.06
≥6	37	1	38	2.63	2.71

2. 음주와 선천성 기형에 대한 연구

- 알콜량 : 순서형 변수
- 기형 여부 : 명목형이지만 “없음”을 “낮음”으로,
  “있음”을 “높음”으로 간주하여 순서형 변수로 줌
- 범주 점수
  : 알콜량(0, 0.5, 1.5, 4, 7), 기형여부 (0, 1)

2. 음주와 선천성 기형에 대한 연구

독립성 검정 결과

① $G^2 = 6.2\,(P값 = 0.185),\ X^2 = 12.1\,(P값 = 0.017)$
"n이 크더라도 X^2 및 G^2의 카이제곱 근사가 좋지 않은 경우"

② $M^2 = 6.570\,(P값 = 0.010)$
➔ 두 변수의 상관관계가 0이 아니라는 강한 증거를 제시함

③ 양이나 음의 연관성이 있을 때 M^2 을 사용한
순서형 검정법은 X^2 나 G^2 보다 높은 검정력을 가짐

3. 범주 점수의 선택

- **대개 점수 부여 방법의 차이가 분석 결과에 미치는 영향은 크지 않음**
  - 자료가 매우 불균형한 경우는 결과가 달라질 수 있음

3. 범주 점수의 선택

범주점수의 부여 방법

① 주관적인 범주점수 부여 방법

- 일반적으로 사용하는 점수 부여 방법
- 선택 가능한 여러 가지 범주 점수들 중 2~3가지를 선택한 후 이들에 대해 분석결과가 유사한지를 검토하는 것이 바람직함
- 정치철학(진보, 온건, 보수)의 경우처럼 범주로부터 직접적으로 점수 선택이 어려운 경우는 등 간격 점수를 부여하는 것이 합리적일 수 있음

3. 범주 점수의 선택

범주점수의 부여 방법

② 중간수위(Midrank) 부여 방법

- 각 관측개체에 순위를 매긴 후에 그 순위를 범주 점수로 사용함
- 한 범주에 속한 모든 개체에 대해서는
  해당 범주에 속한 개체들의 순위의 평균값을 부여함
- 알코올 소비량의 수준 0인 경우는 17,144명에 속해 있음
  ➔ 평균 순위는 $(1+17,144)/2=8,557.5$가 됨
- 알코올 소비량과 기형 Data는
  중간순위를 부여하는 것이 적절하지 않은 예임

중간순위 : (8,557.5, 24,365.5, 32,013.0, 32,473.0, 32,555.5)
마지막 세 범주에서 대단히 유사한 범주점수를 나타냄

연구자의 판단에 따라 범주간의 거리를
반영하는 점수를 선택하는 것이 바람직함

3. 범주 점수의 선택

다른 분석 방법

순서형 r x c 분할표에 대한 다른 검정법

- 순서형변수 연관성 측도를 활용하는 방법이 있음
- 감마(γ)와 켄달의 타우비(Kendall's τ_b) :
  켄달의 타우로 불리는 순서형 변수의 연관성 측도를 일반적인 분할표에 응용한 것

4) 코크란-아미티지 추세검정 (Cochran-Armitage trend test)

- 행변수 X가 설명변수, 열변수 Y가 반응변수인 경우
  ➔ 특별히 $I \times 2$ 분할표에 대해서 관심

- X의 수준에 따라 Y의 특정 범주가
  차지하는 비율이 어떻게 변화하는지를 알아보는데 관심이 있음

- 순서형 변수 X에 임의의 단조점수를 부여하고
  Y에 대해서도 임의의 점수를 부여한 후 M^2 계산
  ➔ M^2을 통해서 X수준 변화에 따라
  특정 범주 비율의 선형추세 여부를 알 수 있음
  ➔ M^2값이 클수록 선형추세의 기울기가
  0이 아님을 나타냄

5. SAS 실습

```
Data infants ;
INPUT malform alcohol count @@ ;
CARDS ;
  1 0   17066 1 0.5  14464 1 1.5  788 1 4.0  126 1 7.0 37
  2 0        48 2 0.5       38 2 1.5     5 2 4.0   1 2 7.0   1
;
RUN ;
PROC FREQ ;
  TABLES malform *alcohol / chisq cmh1 trend ;
  WEIGHT count ;
RUN ;
```

〈참고〉 “중간순위 부여”

```
TABLES malform*alcohol / chisq cmh1 scores=ridit ;
```

5. SAS 실습

수행 결과

malform * alcohol 테이블에 대한 통계량

통계량	자유도	값	확률
카이제곱	4	12.0821	0.0168
우도비 카이제곱	4	6.202	0.1846
Mantel-Haenszel 카이제곱	1	6.5699	0.0104
파이 계수		0.0193	
우발성 계수		0.0193	
크래머의 V		0.0193	

Cochran-Armitage 추세 검정	
통계량 (Z)	-2.5632
단측 Pr < Z	0.0052
양측 Pr > |Z|	0.0104

Cochran-Mantel-Haenszel 통계량 (테이블 스코어에 기반한)				
통계량	대립가설	자유도	값	확률
1	영(0)이 아닌 상관계수	1	6.5699	0.0104

03

제 2장. 이차원 분할표(2)

소표본에 대한 정확 추론

소표본에 대한 정확 추론

- X^2, G^2 와 같은 카이제곱 통계량은 표본크기 n이 증가함에 따라 근사적으로 카이제곱 분포를 따른다는 사실을 이용하여 가설검정 진행
- 표본크기 n이 작은 경우의 독립성 검정방법은?

1. 피셔의 정확 추론

2 X 2 분할표의 경우(Fisher 1934)

		Y		
		1	2	
X	1	n_{11}	n_{12}	n_{1+}
	2	n_{21}	n_{22}	n_{2+}
		n_{+1}	n_{+2}	n

- 행과 열의 주변합이 고정된 경우에 n_{11} 값은 나머지 세 개의 칸 도수를 결정함

1. 피셔의 정확 추론

- 오즈비 $\theta = 1$ (독립성 만족)일 때

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

" 초기하 분포 "

1. 피셔의 정확 추론

초기하 분포를 이용한 독립성 검증

- 독립성 검증 P -값
  : 현재 관측된 결과 또는 그 결과보다 대립가설을
  더 지지하는 결과들에 대한 초기하 분포의 확률값

- $\hat{\theta} = (n_{11}n_{22})/(n_{12}n_{21})$
  " n_{11} 값이 커질수록 더 큰 오즈비 $\hat{\theta}$ 을 갖게 됨"
  ➔ P-값
  : 현재의 n_{11}의 관측값보다
  더 큰 값을 갖게 될 초기하 분포의
  오른쪽 꼬리 분포의 확률이 됨

2. 피셔의 차 맛보기 실험

실제 먼저 부은 것	추측 결과		
	우유	홍차	
우유	?		4
홍차			4
	4	4	8

n_{11} = 0, 1, 2, 3, 4

- n_{11} = 4 일 확률

$$P(4) = \frac{\binom{4}{4}\binom{4}{4-4}}{\binom{8}{4}} = \frac{1}{70} = 0.014$$

2. 피셔의 차 맛보기 실험

- n_{11} = 3 일 확률

$$P(3) = \frac{\binom{4}{3}\binom{4}{4-3}}{\binom{8}{4}} = \frac{16}{70} = 0.229$$

n_{11}	$P(n_{11})$
0	0.014
1	0.229
2	0.514
3	0.229
4	0.014
계	1

2. 피셔의 차 맛보기 실험

- 2X2 분할표에서
  H_0 : 서로 독립 ⇔ H_0 : $\theta = 1$ (θ: 오즈비)

① 가설검정 H_0 : $\theta = 1$, H_1 : $\theta > 1$

P-값 = $P(\hat{\theta} \geq \widehat{\theta_{obs}})$

= 0.229 + 0.014 = 0.243

H_0 를 기각할 만한 충분한 증거는 아님

② 가설검정 H_0 : $\theta = 1$, H_1 : $\theta \neq 1$

P-값 = 양쪽 꼬리분포의 확률

= P(0) + P(1) + P(3) + P(4)

= 0.486

2. 피셔의 차 맛보기 실험

SAS 프로그램

실제	추측 결과	
	우유	홍차
우유	3	1
홍차	1	3

```
DATA tea ;
INPUT first $ second $ count @@ ;
CARDS ;
milk milk 3 milk tea 1
tea milk 1 tea tea 3
;
PROC FREQ ;
 TABLE first*second / exact;
 WEIGHT count ;
RUN ;
```

2. 피셔의 차 맛보기 실험

분석 결과

통계량	자유도	값	확률
카이제곱	1	2.000	0.157
우도비 카이제곱	1	2.093	0.148

Fisher의 정확 검정	
(1,1) 셀 빈도(F)	3
하단측 p값 Pr <= F	0.9857
상단측 p값 Pr >= F	0.2429
테이블 확률 (P)	0.2286
양측 p값 Pr <= P	0.4857

표본 크기 = 8

03 강

다음시간안내

삼차원 분할표

수고하셨습니다.