13강

통계적 비교 I I

통계·데이터과학과 장영재 교수

한국방송통신대학교

목차

01

분할표

범주형 데이터의 검정

- 범주형 데이터 분석은 조사변수의 속성, 즉 범주에 따라 분류하고 해당 범주에 속하는 도수를 조사하여 이를 분석하는 방법을 의미

- 각 변수의범주에 따라 도수를 기입하여 데이터의 분포를 나타낸 분할표를 토대로 분석

➡ 변수들 조합의 빈도수를 가지고 이 변수들이 독립적인지, 또는 데이터들이 이론적 분포와 일치하는지 등을 분석

분할표의 예시

일원분할표(one-way contingency table)

과목명	등록생 수
전공 이론	70명
전공 응용	100명
교양	80명
합계	250명

이원분할표(two-way contingency table)

학과	등록	미등록	계
통계학과	20	30	50
데이터과학과	13	12	25
합계	33	42	75

02

독립성 검정

독립성 검정이란

- 두 범주형 변수가 독립인지를 검정하는 방법
- 두 범주형 변수 A 와 B 의 범주가 각각 r, c 개인 r×c 분할표에서 각 범주 조합에 속할 확률이 p_{ij} 일 때, 아래와 같은 확률분포를 고려

구분	변수 B				행의 합
	B_1	B_2	$\cdots$	B_c	
변수 A A_1	p_{11}	p_{12}	$\cdots$	p_{1c}	$p_{1\cdot}$
A_2	p_{21}	p_{22}	$\cdots$	p_{2c}	$p_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	p_{r1}	p_{r2}	$\cdots$	p_{rc}	$p_{r\cdot}$
열의 합	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot c}$	1

* A_i에 속하는 사건 A_i와 B_j에 속하는 사건 B_j 가 독립이라면 다음이 성립

$$P(A_i \cap B_j) = P(A_i) \cdot P(B_j)$$

독립성 검정의 가설과 검정 방법

독립성 검정의 가설

H_0 : 변수 A와 B는 독립이다. 즉, 모든 i, j에 대하여

$$p_{ij} = p_{i\cdot} \cdot p_{\cdot j} \quad (i = 1, \cdots, r, \quad j = 1, \cdots, c)$$

H_1 : 변수 A와 B는 독립이 아니다(서로 관련이 있다).

분할표 및 검정통계량

관찰도수	변수 B				행의 합
	B_1	B_2	$\cdots$	B_c	
변수 A A_1	O_{11}	O_{12}	$\cdots$	O_{1c}	$T_{1\cdot}$
A_2	O_{21}	O_{22}	$\cdots$	O_{2c}	$T_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	$T_{r\cdot}$
열의 합	$T_{\cdot 1}$	$T_{\cdot 2}$	$\cdots$	$T_{\cdot c}$	n

$$E_{ij} = n\left(\frac{T_{i\cdot}}{n}\right)\left(\frac{T_{\cdot j}}{n}\right) = T_{i\cdot}\left(\frac{T_{\cdot j}}{n}\right)$$

$$\chi^2_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

독립성 검정의 가설과 검정 방법

- 검정 방법 : 아래를 만족하면 유의수준 α에서 H_0 기각

$$\chi^2_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} > \chi^2_{(r-1)(c-1),\,\alpha}$$

- 독립성 검정에서 χ^2 분포를 이용하려면 모든 기대도수가 적어도 5 이상이어야 하며, 5 보다 작을 경우, 인접 범주와 합하여 분석하는 것을 권고

03

적합도 검정

적합도 검정이란

- 각 범주형 변수의 범주에 따라 관찰된 도수를 토대로 모집단이 특정 분포를 따르는가를 검정하는 방법

- 변수 X에 관한 (이산형) 확률분포

X	$P(X = x_i)$
x_1	p_1
x_2	p_2
$\vdots$	$\vdots$
x_k	p_k
	1.0

적합도 검정의 가설과 검정 방법

적합도 검정의 가설 및 검정통계량

이산형 변수 X의 가능한 값이 n개 있고 $P(X = x_i)$가 $p_1, p_2, \cdots p_k$

모집단에서 표본추출된 n개 데이터에 대한 관찰도수가 $O_1, O_2, \cdots O_k$

모집단 확률분포가 $p_{10}, p_{20}, \cdots p_{k0}$라고 할 때의 귀무가설 및 대립가설

$$H_0 : (p_1, p_2, \cdots, p_k) = (p_{10}, p_{20}, \cdots, p_{k0})$$

$$H_1 : \text{적어도 하나의 } p_i \text{는 } p_{i0} \text{와 다르다}$$

$$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$ 단, $(E_1, E_2, \cdots, E_k) = (np_{10}, np_{20}, \cdots, np_{k0})$

적합도 검정의 가설과 검정 방법

- 검정 방법 : 아래를 만족하면 유의수준 α에서 H_0 기각(k는 범주의 개수)

$$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} > \chi^2_{k-1,\,\alpha}$$

- 단, H_0에서 p개의 모수 추정이 필요할 경우에는 아래 기준을 적용

$$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} > \chi^2_{k-p-1,\,\alpha}$$

- 적합도 검정에서 χ^2분포를 이용하려면 모든 기대도수가 적어도 5 이상이 어야 하며, 5 보다 작을 경우, 인접 범주와 합하여 분석하는 것을 권고

04

R을 이용한 실습

독립성 검정

- xtabs 함수는 각 범주값이 저장되어 있는 객체, 즉 범주형 변수를 가지고 분할표를 생성

```
dept <- c(rep("Stat",50),rep("DS",25))
regi <- c(rep("Y",20),rep("N",30),rep("Y",13),rep("N",12))
deptregi <- data.frame(dept,regi)
rtable <-xtabs(~dept+regi, data=deptregi)
rtable
 regi
dept    N  Y
  DS   12 13
  Stat 30 20
```

독립성 검정

- chisq.test 함수로 카이제곱 검정을 실시
- 'correct=F'는 연속성 수정(continuity correction)을 하지 않음을 의미(연속성 수정은 통상 2X2 분할표의 경우에 실시)

```
ctest <-chisq.test(rtable, correct=F)
ctest

        Pearson's Chi-squared test


data:  rtable
X-squared = 0.97403, df = 1, p-value = 0.3237
```

적합도 검정

- 은행 전국 60개 지점의 당일 부도수표를 관측 (0, 1, 2, 3개 이상)
- m은 분할표를 이용하여 구한 표본평균
- **dpois 는 포아송 분포를 의미**

```
catnum <- c(0:3)
obs <-c(33, 15, 9, 3)
m <- sum(catnum*obs)/sum(obs)
pprob <- round(dpois(catnum, m),3)
pprob


[1] 0.497 0.348 0.122 0.028
```

적합도 검정

- 분포함수를 이루기 위해 합이 1이 되도록 조정
- 전국 60개 지점의 부도수표 기댓값 계산
- 기대도수를 산출하고 5미만의 기대도수를 갖게되는 범주는 인접한 범주로 병합

```
pprob[4] <- 1-sum(pprob[1:3])
pprob
[1] 0.497 0.348 0.122 0.033
pprob*60
[1] 29.82 20.88  7.32  1.98
```

적합도 검정

- 자유도 : 병합한 범주의 개수(3)-추정모수의 개수(1)-1=1
- H0 : 포아송분포를 따른다 → 기각할 수 없음

```
# 기대도수 5 미만 범주를 병합하여 재 분석
obs1 <- c(33, 15, 12)
pprob1 <- pprob[1:3]
pprob1[3] <- 1-sum(pprob[1:2])
pprob1
[1] 0.497 0.348 0.155
ctest1 <- chisq.test(obs1, p=pprob1)
ctest1$statistic > qchisq(0.95,1)
X-squared
   FALSE
```

정리하기

- 범주형 데이터(categorical data)의 분석은
  조사변수를 범주에 따라 분류하고 해당 범주에 속하는 도수를
  조사하여 분석하는 방법을 의미한다.
- χ^2 검정통계량은 관찰수와 기대도수의 차를 제곱해서
  기대도수로 나눈 것의 총합으로 계산된다.
- 표본으로부터 χ^2 검정통계량을 구했을 때, 주어진 유의수준 α와
  자유도 ν 에 대한 $\chi^2_{\nu,\alpha}$값을 비교하여 χ^2 값이 더 크면 귀무가설을
  기각한다. 즉, '두 변수가 서로 독립임'을 기각(독립성 검정)하거나
  '모집단의 분포가 이론분포와 일치함'을 기각(적합도 검정)한다.

정리하기

- 독립성 검정은 모집단이 두 변수에 의해 범주화되었을 때, 이 변수들이 독립인지 검정하는 통계적 절차이다.
- 적합도 검정은 관찰된 표본으로부터 그 모집단의 분포가 이론분포를 따르는가를 검정하는 통계적 절차이다.

14강

회귀모형 I

다음 시간 안내