R을 이용한 기초통계학

7강: 분할표 분석

수원대학교
데이터과학부
김 진 흠

서울대학교
보건환경연구소
이 보 라

The 19th ASIAN INSTITUTE IN STATISTICAL GENETICS AND GENOMICS

INTRODUCTORY STATISTICS USING R

두 변수의 연관분석

- 분할표 분석: 두 범주형 변수 간의 연관성을 **분할표**로 분석
- 상관분석: 두 연속형 변수 간의 연관성을 **선형관계**의 강도로 분석
- 회귀분석: 두 연속형 변수 간의 연관성을 **함수적 관계**로 분석

범주형 변수

- 범주형 변수(categorical variable) 또는 질적 변수(qualitative variable)
- 측정된 척도가 여러 범주로 이루진 변수

범주형 변수

- 명목형 변수(nominal variable)
  - 순위의 개념이 **없는** 변수
  - 예: 성별(남,여), 결혼 상태(미혼, 기혼), 흡연 상태(과거 흡연, 현재흡연, 비흡연)
- 순서형 변수(ordinal variable)
  - 순위의 개념이 **있는** 변수
  - 예: 나이(미성년, 성년, 장년), 혈압(<70, 70-90, 91-110, 111-130, >130), 환자의 병의 호전 상태(나빠짐, 변화 없음, 좋아짐)

분할표

- 칸(cell): 범주형 변수들의 범주들의 조합의 결합
- 이차원 분할표(two-way contingency table): **두 개**의 범주형 변수에 대해 분류한 분할표
- 다차원 분할표(multi-way contingency table): **세 개 이상**의 범주형 변수들에 대하여 분류한 분할표

이차원 분할표

- 범주형 변수들의 **관찰도수**를 정리한 표
- X, Y: 범주형 변수
- $I \times J$ 분할표: X는 I개의 수준, Y는 J개의 수준

	Y				
X	1	2	…	J	Total
1	n_{11}	n_{12}	…	n_{1J}	n_{1+}
2	n_{21}	n_{22}	…	n_{2J}	n_{2+}
⋮	⋮	⋮	n_{ij}	⋮	⋮
I	n_{I1}	n_{I2}	…	n_{IJ}	n_{I+}
Total	n_{+1}	n_{+2}	…	n_{+J}	n

범주형 자료분석: 분석방법

- 카이제곱검정: Pearson 검정, **가능도비 검정**
- Fisher 정확(exact) 검정
- 만텔-핸젤(Mantel-Haenszel) 검정
- 로그-선형모형
- 로지스틱 회귀모형

예제

- 아스피린 복용과 심근경색(myocardial infarction, MI)의 관련성 자료
- Q: 정기적인 **아스피린**의 복용이 **MI를** 감소시키는가?
- 대상: 하버드 의과대학에 있는 의사보건연구그룹
- 기간: 5년
- 처리(treatment): 아스피린과 위약(placebo)
- 방법: 눈가림법(blind)

예제

- 아스피린복용과 심근경색증의 분할표

	심근경색증		
처리	예	아니요	합
위약	189	10,845	11,034
아스피린	104	10,933	11,037

예제 ➜ R code로 이동

```
> mi <- matrix(c(189,104,10845,10933),ncol = 2)
> mi
      [,1]  [,2]
[1,]   189 10845
[2,]   104 10933
> dimnames(mi) <- list(treat = c("placebo","aspirin"),
+                      "mypcarial infraction" = c("yes","no"))
> mi
         mypcarial infraction
treat     yes    no
  placebo 189 10845
  aspirin 104 10933
```

이차원 분할표

- X와 Y가 각각 두 개의 범주를 가진 분할표

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

이차원 분할표: 결합확률

- 결합확률(joint probability)
  - $\pi_{ij} = P(X = i, Y = j)$
- $\{\pi_{ij}\}$는 X와 Y의 결합분포(joint distribution). $\sum_{i=1}^{2} \sum_{j=1}^{2} \pi_{ij} = 1$을 만족

이차원 분할표: 주변확률

- 주변확률(marginal probability)
  - 행 변수에 대한 주변확률: $\{\pi_{i+}\}, \pi_{i+} = \sum_{j=1}^{2} \pi_{ij}$
  - 열 변수에 대한 주변확률: $\{\pi_{+j}\}, \pi_{+j} = \sum_{i=1}^{2} \pi_{ij}$
- 주변분포(marginal distribution): 결합분포의 행과 열의 합의 분포

분할표에서의 확률구조

- 결합확률, 주변확률

	Y		
X	1	2	합
1	π_{11}	π_{12}	$\boldsymbol{\pi_{1+}}$
2	π_{21}	π_{22}	$\boldsymbol{\pi_{2+}}$
합	$\boldsymbol{\pi_{+1}}$	$\boldsymbol{\pi_{+2}}$	1

이차원 분할표: 조건부 확률

- 조건부 확률(conditional probability): Y는 **반응**변수이고 X는 **설명**변수인 경우 X가 주어졌을 때 Y의조건부 확률은

$$P(Y = j|X = i) = \frac{P(X=i,Y=j)}{P(X=i)} = \frac{\pi_{ij}}{\pi_{i+}} = \pi_{j|i}$$

- 조건부 분포(conditional distribution): 조건부 확률의 분포

분할표에서의 확률구조

- 조건부 확률

	심근경색증		
처리	예	아니요	합
위약	$\pi_1(\pi_{1\|1})$	$1-\pi_1(\pi_{2\|1})$	1
아스피린	$\pi_2(\pi_{1\|2})$	$1-\pi_2(\pi_{2\|2})$	1

분할표에서의 확률구조

➜ R code로 이동

```
> prop.table(mi,1)
         mypcarial infraction
treat            yes        no
  placebo 0.01712887 0.9828711
  aspirin 0.00942285 0.9905771
> prop.table(mi,2)
         mypcarial infraction
treat           yes       no
  placebo 0.6450512 0.4979796
  aspirin 0.3549488 0.5020204
```

실습: 분할표

- Data: Seat-belt usage in California (82명 조사)

	Child	
Parent	buckled	unbuckled
buckled	56	8
unbuckled	2	16

실습: 분할표➜ R code로 이동

```
> ex1 <- matrix(c(56,2,8,16),nrow = 2)
> dimnames(ex1) <- list(parent = c("buckled","unbuckled"),
+                       child=c("buckled","unbuckled"))
> ex1
           child
parent      buckled unbuckled
  buckled        56         8
  unbuckled       2        16
```

실습: 분할표에서의 확률구조

```
> prop.table(ex1)
           child
parent         buckled  unbuckled
  buckled   0.68292683 0.09756098
  unbuckled 0.02439024 0.19512195
> prop.table(ex1,1)
           child
parent        buckled unbuckled
  buckled   0.8750000 0.1250000
  unbuckled 0.1111111 0.8888889
> prop.table(ex1,2)
           child
parent         buckled unbuckled
  buckled   0.96551724 0.3333333
  unbuckled 0.03448276 0.6666667
```

독립성과 동질성 검정

- X와 Y가 "통계적으로 독립"
- Y는 반응변수, X도 반응변수
  - $\pi_{ij} = \pi_{i+}\pi_{+j}, i = 1,2; j = 1,2$
- Y는 반응변수, X는 독립변수
  - Y의 조건부 확률이 X의 각각의 수준에서 동일
  - $\pi_{j|1} = \pi_{j|2}, j = 1,2$

검정법

- 동질성 검정: 두 그룹 간의 비율을 비교하기 위한 검정
- 독립성 검정
- 두 검정통계량이 서로 일치

Pearson의 카이제곱 통계량

- 정 의: $X^2 = \sum \frac{(\text{관찰도수} - \text{기대도수})^2}{\text{기대도수}}$
- $X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$: $I \times J$ 분할표에서 독립성 검정을 위한 통계량
- H_0하에서 (i, j) 칸의 기대도수: $\hat{\mu}_{ij} = \frac{n_{i+} n_{+j}}{n}$

Pearson의 카이제곱 통계량

- $|n_{ij} - \hat{\mu}_{ij}| \uparrow \Rightarrow H_0$ 기각
- $P - 값 = P(X^2 \geq x_0^2 | H_0)$
- X^2 통계량은 표본이 커지면 (대개 $\hat{\mu}_{ij} \geq 5$) 근사적으로 자유도가 $(I-1)(J-1)$인 카이제곱분포를 따름. 즉 $X^2 \sim \chi^2((I-1)(J-1))$

실습: 독립성과 동질성 검정

➜ R code로 이동

```
> chisq.test(ex1)

        Pearson's Chi-squared test with Yates' continuity correction

data:  ex1
X-squared = 35.995, df = 1, p-value = 1.978e-09

> res <- chisq.test(ex1)
> res$p.value
[1] 1.977918e-09
```

Fisher의 정확 검정

- X^2, G^2는 대표본 카이제곱 통계량!
- 표본크기가 작은 경우?
- 정확(exact)분포 이용
- 'exact'는 근사적인 분포를 사용하지 않고 통계량의 정확한 분포를 사용한다는 의미

Fisher의 정확 검정

- H_0: 두 반응변수 X, Y가 서로 독립이다
- 행 주변합과 열 주변합이 **고정**된 경우: **초기하분포**(hypergeometric distribution)
- 확률분포는 H_0 하에서 첫째 칸 (1,1)의 도수가 n_{11}일 확률

Fisher의 정확 검정

- 주변합 $n_{1+}, n_{2+}, n_{+1}, n_{+2}$이 고정되어 있다고 가정

X	Y		합
	1	2	
1	n_{11}	$n_{12}(=n_{1+}-n_{11})$	n_{1+}
2	$n_{21}(=n_{+1}-n_{11})$	$n_{22}(=n_{2+}-n_{+1}+n_{11})$	n_{2+}
합	n_{+1}	n_{+2}	n

- n_{11}의 값만 알면 나머지 n_{12}, n_{21}, n_{22}의 값을 알 수 있으므로 $\boldsymbol{n_{11}}$의 분포만 구하면 됨

Fisher의 정확 검정

- n_{11}의 분포는 **초기하분포**
- $P(n_{11}) = \dfrac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{21}}}{\binom{n}{n_{+1}}}$

Fisher의 정확 검정

- P−값: 관측값 n_{11}보다 실현 **가능성이 적거나 같은** 도수들에 대응되는 확률의 합 즉,
- P−값은 $P(y) \leq P(n_{11})$를 만족하는 모든 첫째 칸 도수 y에 대한 확률의 합

Fisher의 차 맛보기 실험

- Fisher(1935). “The Design of Experiments”
- 런던 근처의 로뎀스테드의 실험연구소
- 8 컵의 차를 맛보는 실험을 시행
  - 처음 4 컵 : 우유 + 차
  - 나중 4 컵 : 차 + 우유
- 각 유형마다 **4개**의 컵이 있음을 **통지**한 후에 랜덤한 순서로 맛을 본 후에 **우유**를 먼저 넣은 컵을 선택

Fisher의 차 맛보기 실험

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

Fisher의 차 맛보기 실험

- H_0: 실제와 추측 간에 상관이 없다
- 열 주변합과 행 주변합이 각각 4로 고정
- n_{11}의 H_0 하에서 초기하분포를 따름
- P –값: $P(3) = 0.229$ 보다 작거나 같은 모든 확률들의 합
- 즉 $P(0) + P(1) + P(3) + P(4) = 0.486$

Fisher의 차 맛보기 실험

- Fisher 자료와 같은 주변합을 가진 초기하분포

n_{11}	확률
0	.014
1	.229
2	.514
3	.229
4	.014
Total	1.00

실습: Fisher의 차 맛보기 실험

➜ R code로 이동

```
> ex2 <- matrix(c(3,1,1,3),nrow = 2)
> dimnames(ex2) <- list(Guess = c("Milk","Tea"),
+                       Truth = c("Milk","Tea"))
> ex2
      Truth
Guess  Milk Tea
  Milk    3   1
  Tea     1   3
```

```
> fisher.test(ex2)

        Fisher's Exact Test for Count Data

data:  ex2
p-value = 0.4857
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.2117329 621.9337505
sample estimates:
odds ratio
  6.408309

> fisher.test(ex2,alt = "greater")

        Fisher's Exact Test for Count Data

data:  ex2
p-value = 0.2429
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.3135693       Inf
sample estimates:
odds ratio
  6.408309
```

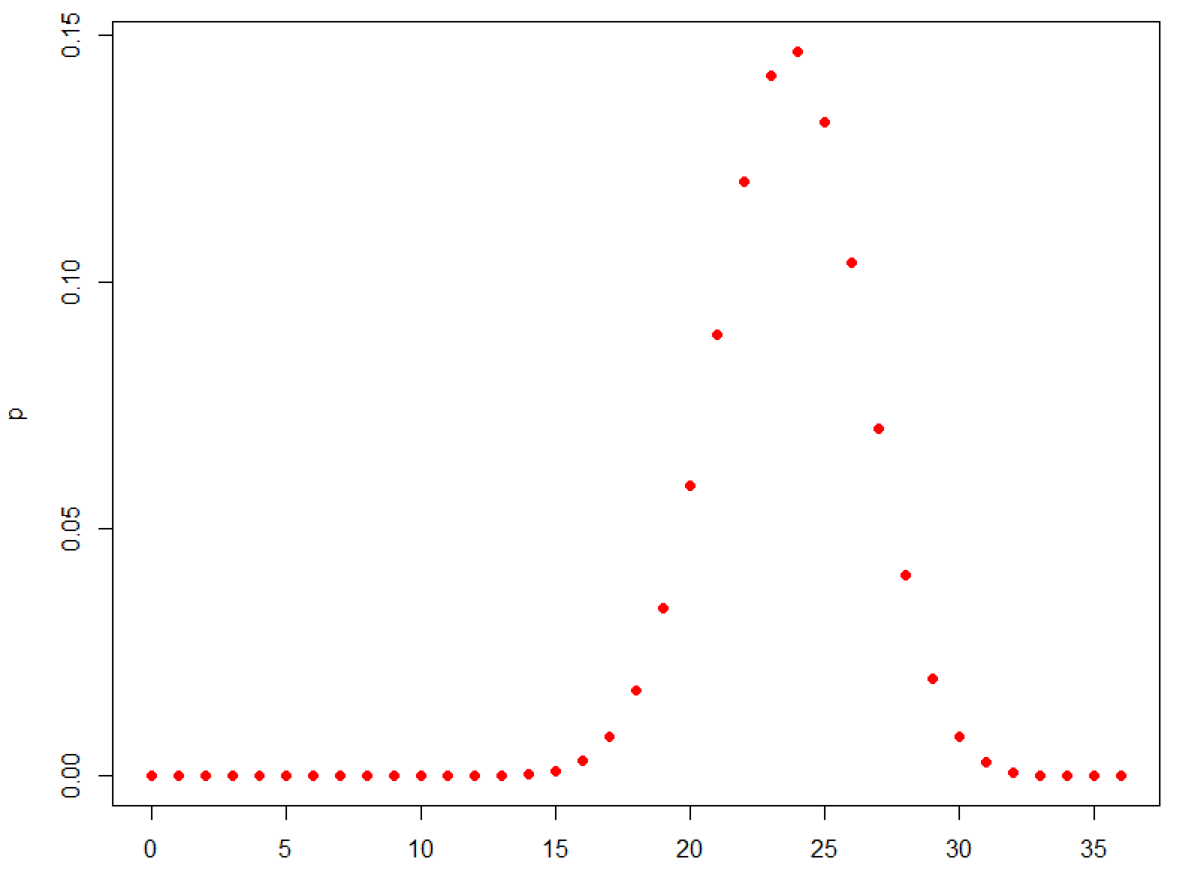
Fisher의 정확 검정

Age	Disease		Total	Proportion
	Yes	No		
Old	30	184	214	14.01
Young	6	106	112	5.3
Total	36	290	326	

- 30보다 크거나 같은 경우: 30, 31,32,⋯, 35,36 ⇒ 0.0116
- 30보다 작은 경우: 17,⋯, 1, 0 ⇒ 0.0126

초기하분포 ➜ R code로 이동

```
> x <- 0:36
> p <- choose(214,x) * choose(112,36 - x) / choose(326,36)
> plot(x,p,col = "red", pch = 16)
```


0.00
0.05
0.10
0.15
p
0
5
10
15
20
25
30
35
x

```
> observed.p <- p[x == 30]
> observed.p
[1] 0.008079815
> exact.P <- sum(p[p <= observed.p])
> exact.P
[1] 0.02423025
> d <- matrix(c(30,6,184,106),ncol = 2)
> fisher.test(d)

        Fisher's Exact Test for Count Data

data:  d
p-value = 0.02423
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.127473 8.719466
sample estimates:
odds ratio
   2.87246
```

상관분석과 회귀분석

- 상관분석(correlation analysis)
  - 두 변수 간의 선형성을 표본 상관계수를 이용해 분석
- 회귀분석(regression analysis)
  - 두 개 이상의 변수 간의 함수 관계를 선형 모형을 사용해 분석

표본상관계수

- Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- $r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$
- $r > 0 \Rightarrow$ 양의 상관관계 $\Leftrightarrow x\uparrow \& \, y\uparrow$
- $r < 0 \Rightarrow$ 음의 상관관계 $\Leftrightarrow x\uparrow \& \, y\downarrow$
- 일반적으로,
  - $|r| \approx 1 \Rightarrow$ 강한 상관관계
  - $|r| \approx 0 \Rightarrow$ 약한 상관관계

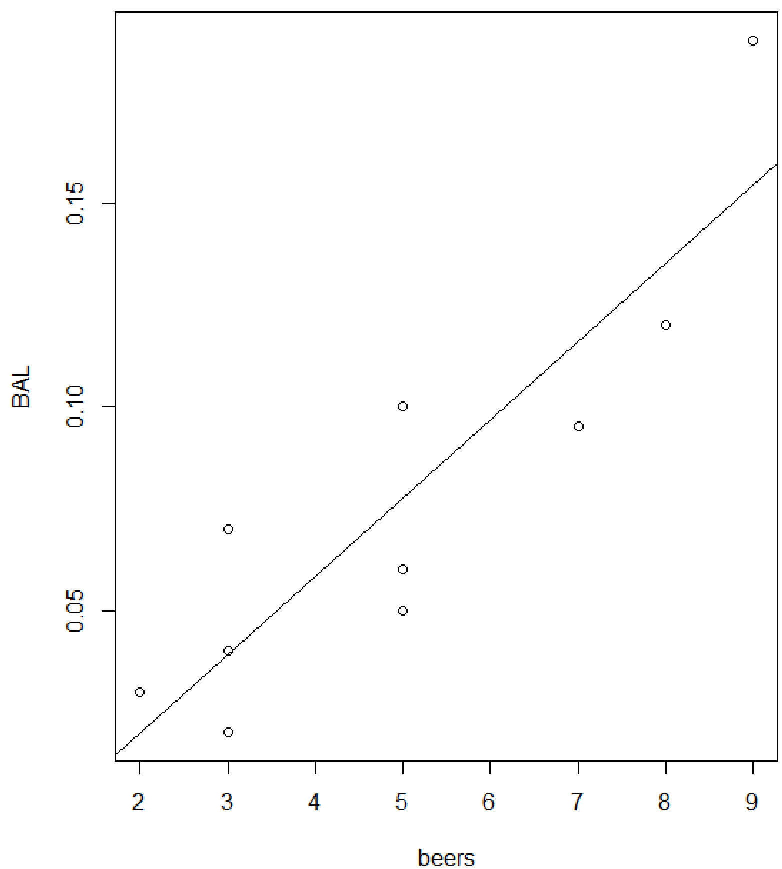
상관분석

- r 값은 두 변수 간의 **직선 관계**를 나타내는 측도
- 주어진 x 값으로부터 y 값을 **예측**하기 위해서는 x와 y 간의 보다 정확한 함수 관계를 찾는 것이 필요함
- **여러 개의 x 변수**들과 y 간의 상관관계를 알기 위해 서는 r 값만 사용하는 것은 부적절함

실습: 상관분석

➜ R code로 이동

```
> beers <- c(5,2,9,8,3,7,3,5,3,5)
> BAL <- c(0.10,0.03,0.19,0.12,0.04,0.095,0.07,0.06,0.02,0.05)
> cor(beers,BAL)
[1] 0.8882323
> ex3 <- data.frame(cbind(beers,BAL))
> head(ex3)
  beers   BAL
1     5 0.100
2     2 0.030
3     9 0.190
4     8 0.120
5     3 0.040
6     7 0.095
> plot(BAL ~ beers,data=ex3)
> res <- lm(BAL ~ beers)
> abline(res)
```



0.05
0.10
0.15
BAL
2
3
4
5
6
7
8
9
beers

요약

- 독립성과 동질성 검정법
  - 근사검정: 카이제곱 검정법
  - 정확검정: Fisher 검정법
- 상관분석