교차분석 개념

변수 간 분석

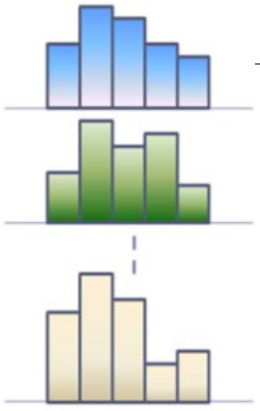
X Y	범주형	측정형
범주형	교차분석*)	분산분석*)
측정형	로지스틱 회귀분석	상관분석*)

교차표

- 두 범주형 요약 빈도표를 교차
- 교차표 작성 시 원인(설명)에 해당되는 변수를 행으로 넣는다
- 결합확률밀도함수, 주변확률밀도함수 개념
- 이진형(성공, 실패)
- n_{ij} : i행 j열의 셀 빈도 / π_{ij} : i행 j열의 비율

X Y	범주1	범주2	…	범주c	총합
범주1	n11	n12		n1c	n1+
범주2	n21	n22		n2c	n2+
…	…	…	…	…	…
범주r	n_{r1}	n_{r2}		n_{rc}	
총합	n_{r1}	n_{+2}	…	n_{+c}	n_{++}

동질성 검정 Homogeneity



각 행의 분포는 동일한가?

귀무가설 : $\pi_{ij} = \pi_{kj}$ for all j – 모든 행 분포는 동일하다.

대립가설 : 적어도 하나 이상의 행 분포는 동일하지 않다.

적합성 검정 Goodness of fits

- 귀무가설 : 데이터는 임의 분포를 따른다. (정규성 검정의 경우 정규분포)
- 대립가설 : 귀무가설에 설정된 분포를 따르지 않는다.
- 교차표의 첫행은 관측빈도, 두번째 행은 귀무가설에 설정된 분포에 의해 계산된 이론(기대) 빈도, 이 교차표를 이용하여 분포의 적합성 검정을 할 수 있음.

독립성 검정 Independence

두 범주형 변수는 독립인가?

- 귀무가설 : $\pi_{ij} = \pi_{i+}\pi_{+j}$ (결합확률밀도함수는 주변확률밀도함수의 곱이다)
- 대립가설 : 두 변수는 연관관계가 있다.

검정통계량

동질성, 독립성 검정의 검정통계량은 동일하다.

관측빈도 : O_{ij}, 기대빈도 : $E_{ij} = n_{++} * \frac{n_{ij}}{n_{i+}} * \frac{n_{ij}}{n_{+j}}$

검정통계량 $TS = \sum\sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(df = (r-1)(c-1))$

*) 2X2 교차표에서 검정통계량은 자유도 1인 χ^2 분포를 따른다. 이는 두 모비율 차이 검정에서 사용되는 통계량이 표준정규분포(N(0, 1))를 따름 〈=〉 표준정규분포의 제곱은 카이제곱 분포 – 그러므로 2X2 교차표 분석과 두 모비율 차이 z-검정과 동일 (그래서 R software는 두 모비율 검정 시 검정통계량으로 카이제곱 통계량을 제시함) 당연히 유의확률은 동일함

카이제곱 활용

- 귀무가설이 기각되면 행퍼센트를 활용하여 행 범주별 열 비율의 차이에 대하여 설명한다.
- 어떤 비율에 무게를 두어 해석하는지는 연구 목적에 따라 다름

예제 shop.csv

고객의 성향(1=보수, 2=전통, 3=현대)과 직업형태(1=무직, 2=파트타임, 3=풀타임)와 관계가 있는지 검정하시오. 그리고 연수입($1,000)도 조사하였다.

```
shop=read.csv("shop.csv")
names(shop)
shop.table=table(shop$성격,shop$직업)
margin.table(shop.table,1) # marginal of row
margin.table(shop.table,2) # marginal of col
prop.table(shop.table) # percentage
prop.table(shop.table,1) # row percentage

library(gmodels)
CrossTable(mt$, mt$, expected=T,
        chisq=T, prop.t=F, prop.c =F, prop.chisq=F)
```

```
> table(shop$성격,shop$직업)
    1   2   3
 1 157  44 217
 2 219  53 264
 3 256 102 524
> margin.table(shop.table,1) # marginal of row
 1   2   3
418 536 882
> margin.table(shop.table,2) # marginal of col
  1   2   3
632 199 1005
```

```
             | shop$직업
   shop$성격 |         1 |         2 |         3 | Row Total |
-------------|-----------|-----------|-----------|-----------|
           1 |       157 |        44 |       217 |       418 |
             |   143.887 |    45.306 |   228.807 |           |
             |     0.376 |     0.105 |     0.519 |     0.228 |
-------------|-----------|-----------|-----------|-----------|
           2 |       219 |        53 |       264 |       536 |
             |   184.505 |    58.096 |   293.399 |           |
             |     0.409 |     0.099 |     0.493 |     0.292 |
-------------|-----------|-----------|-----------|-----------|
           3 |       256 |       102 |       524 |       882 |
             |   303.608 |    95.598 |   482.794 |           |
             |     0.290 |     0.116 |     0.594 |     0.480 |
-------------|-----------|-----------|-----------|-----------|
Column Total |       632 |       199 |      1005 |      1836 |
-------------|-----------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  23.09461     d.f. =  4     p =  0.0001212324
```

카이제곱 검정결과 유의확률이 <0.001이므로 귀무가설 기각

성격과 직업간에는 관계가 있음

성격별로 풀타임이 산호되고 있으나 전통적인 성격 소유자는 풀타임과 무직이 비슷한 비율

현대적 성격 소유자의 정규직 비율이 가장 높고 전통적인 성격이 무직이 가장 높음

풀타임 다른 성격에 비해 현대적이 사람이 많이 근무하고 있음 (열의 해석 – 기대빈도와 상대빈도의 차이) (상대빈도-기대빈도) 양의 차이가 크면 그 범주 선호

(Data 빈도가 주어진 경우)

```
mt=matrix(c(157,44,217,219,53,264,256,102,524),ncol=3,byrow=TRUE)
colnames(mt)=c("NO","PT","FT")
rownames(mt)=c("C","T","M")
as.table(mt) #freq. table
prop.table(mt,1)
(1/margin.table(mt))*as.array(margin.table(mt,1)) %*%
  t(as.array(margin.table(mt,2))) # row percent
summary(mt) #chisq test
```

근사 통계량 문제

Cochran 법칙

- 기대빈도가 5 이하인 셀의 개수가 전체 셀 개수의 20%를 넘지 않으면 근사 분포 χ^2 사용 가능

문제 발생 시 해결

1) 만약 20%를 넘으면 행이나 열의 유사 범주를 합쳐 기대빈도의 셀을 합친다. (예) 변수가 리커트 척도인 경우 “만족”, “매우만족”에서 기대빈도 5이하 문제가 발생하면 두 범주를 합쳐 “만족이상”으로 하여 셀의 개수를 줄여 빈도 크기를 늘린다.
2) Fisher Exact 검정

Fisher Exact test

영국인들은 밀크 티 마시는 것을 좋아한다. 그래서 귀족들은 잔에 밀크를 먼저 넣었는지, 티를 먼저 넣었는지 알 수 있다고 주장하였다. 초기하분포(Hyper Geometric) 활용

유의확률 (셀 1,1이 a일 경우) $p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}$

	범주1	범주2
범주1	a	b
범주2	c	d

a의 셀 값이 가장 적게 배치되도록 교차표를 조정한후 위의 유의확률을 a, a-1,…,0까지 구하여 모두를 더하면 최종 유의확률이 된다. (단측가설의 유의확률 계산방법)

예제 (wikipedia)

	남자	여자
다이어트	1	9
없음	11	3

```
> fisher.test(rbind(c(1,9),c(11,3)),alternative="less")

        Fisher's Exact Test for Count Data

data:  rbind(c(1, 9), c(11, 3))
p-value = 0.00138
```

귀무가설이 기각되어 성별과 다이어트 유무에는 관계가 있음

남자(1/12)에 비해 여자(9/12)가 더 많이 다이어트 경험이 있음

Case Study 💾 bank.csv

성별에 따라 운영하는 기업형태 (1=개인, 2=공동, 3=기업) 차이는 있는가?

Case Study2 💾 sport.csv

야구 경기에 따라 경기의 박진감(승부가 뒤집히는 경우)은 있나?

경기 초, 경기 후반에 따른 차이는 있는가?

- 1열 – 1=경기 초 리드하고 최종에서 진 경우 2=경기 초에 리드하고 최종에서 이긴 경우
- 2열 – 1=농구(1쿼터), 2=야구(1~3회), 3=풋볼(1쿼터), 4=아이스하키(1쿼터)
- 3열 – 1=경기 후반 리드하고 최종에서 진 경우 2=경기 후반에 리드하고 최종에서 이긴 경우
- 4열 – 1=농구(4쿼터), 2=야구(7~9회), 3=풋볼(1쿼터), 4=아이스하키(3쿼터)

McNemar 검정 (짝진 표본) 동일 개체

2 X2 교차표에서 행렬이 동일 하나의 범주형 변수여서 실험 전후의 변화를 비교하거나 두 범주형 변수의 주변 확률분포가 동일한지 검정

(Wikipedia) 메르스 사태가 학생들의 학교 출석률의 차이를 보였나?

전 후	실패 (출석)	성공(결석)
실패 (출석)	a 101	b 121
성공 (결석)	c 59	d 33

귀무가설 : $H_0 : p_b = p_c$ (전 후 성공비율이 동일하다)

대립가설 : $H_0 : p_b \neq p_c$ (전 후 성공비율 다르다)

검정통계량 : $TS = \frac{(b-c)^2}{(b+c)^2} = \frac{(121-59)^2}{(121+59)^2} = 21.35 \sim \chi^2(1)$

귀무가설 기각, 메르스가 학교 출석율에 유의한 영향을 주었고 후의 결석율이 전보다 높다.

전의 결석율=(59+33)/(314), 후의 결석율=(121+33)/314

```
ds=matrix(c(101, 121, 59, 33),
       nrow = 2, byrow=T,
       dimnames = list("Pre" = c("Absent", "Present"),
                       "Post" = c("Absent", "Present")))
prop.table(ds,1)
mcnemar.test(ds)
```

```
> prop.table(ds,1)
         Post
Pre          Absent   Present
  Absent  0.4549550 0.5450450
  Present 0.6413043 0.3586957

            McNemar's Chi-squared test with continuity
correction

McNemar's chi-squared = 20.6722, df = 1, p-value =
5.45e-06
```