교차분석과
chi-square 분석

교차분석과 chi-square 분석

1) 교차표 작성/분석
   - data.frame() 이용 교차표 작성
   - package 이용 교차표 작성
   - 교차표 분석(학력수준과 진학 여부 교차분석)
2) Chi-square 가설검정
   - 교차분석/ Chi-square 보고서 작성법
   ① 적합성 검정
   ② 독립성 검정
   ③ 동질성 검정

교차 분석(Cross Table Analyze)

- 범주형 자료(명목척도 또는 서열척도)를 대상으로 두 개 이상의 변수들에 대한 관련성 체크.
- 결합분포를 나타내는 교차 분할표를 작성.
- 변수 상호간의 관련성 여부를 분석하는 방법.
- 빈도분석의 특성별 차이를 분석하기 위해 수행하는 분석 방법.
- 빈도분석결과에 대한 보충자료를 제시하는 데 효과적.
- 빈도분석과 함께 고급 통계 분석의 기초 정보를 제공.

교차 분석 시 고려사항

- 교차 분석에 사용되는 변수는 값이 10 미만인 범주형 변수여야 함.
- 비율척도인 경우는 코딩변경(리코딩)을 통해서 범주형 자료로 변환하여 적용 가능.
  - ex) 나이: 10~19세는 1, 20~29세는 2, 30~39세는 3 …

교차표 작성 / 분석

- data.frame() 이용 교차표 작성

```
setwd("c:/workspaces/Rwork/data")
data <- read.csv("cleanDescriptive.csv", header=TRUE)
data # 확인
head(data) # 변수 확인

x <- data$level2 # 학력수준 리코딩 변수
y <- data$pass2 # 대학진학 리코딩 변수

# 학력수준(독립변수) -> 진학여부(종속변수)
result <- data.frame(Level=x, Pass=y ) # 데이터 프레임 생성 - 데이터 묶음

dim(result) # 차원보기 ->  248   2
table(result) # 교차표 보기
#            Pass
# Level     불합격 합격
# 고졸         40   49
# 대졸         27   55
# 대학원졸     23   31
```

교차표 작성 / 분석

- package 이용 교차표 작성

```
# 교차표 작성을 위한 패키지 설치
install.packages("gmodels")
library(gmodels) # CrossTable() 함수 사용

# diamonds 데이터 사용을 위한 ggplot2 패키지 설치
install.packages("ggplot2")
library(ggplot2)

# diamond의 cut과 color에 대한 교차표 생성
CrossTable(x=diamonds$color, y=diamonds$cut, chisq = TRUE)
```

교차표 작성 / 분석

- package 이용 교차표 작성

Total Observations in Table: 53940

제목 없음 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

diamonds$color	diamonds$cut Fair	Good	Very Good	Premium	Ideal	Row Total
D	163	662	1513	1603	2834	6775
	7.607	3.403	0.014	9.634	5.972	
	0.024	0.098	0.223	0.237	0.418	0.126
	0.101	0.135	0.125	0.116	0.132	
	0.003	0.012	0.028	0.030	0.053	
E	224	933	2400	2337	3903	9797
	16.009	1.973	19.258	11.245	0.032	
	0.023	0.095	0.245	0.239	0.398	0.182
	0.139	0.190	0.199	0.169	0.181	
	0.004	0.017	0.044	0.043	0.072	
F	312	909	2164	2331	3826	9542
	2.596	1.949	0.333	4.837	0.049	
	0.033	0.095	0.227	0.244	0.401	0.177
	0.194	0.185	0.179	0.169	0.178	
	0.006	0.017	0.040	0.043	0.071	
G	314	871	2299	2924	4884	11292
	1.575	23.708	20.968	0.473	30.745	
	0.028	0.077	0.204	0.259	0.433	0.209
	0.195	0.178	0.190	0.212	0.227	
	0.006	0.016	0.043	0.054	0.091	
H	303	702	1824	2360	3115	8304
	12.268	3.758	0.697	26.432	12.390	
	0.036	0.085	0.220	0.284	0.375	0.154
	0.188	0.143	0.151	0.171	0.145	
	0.006	0.013	0.034	0.044	0.058	
I	175	522	1204	1428	2093	5422
	1.071	1.688	0.090	1.257	2.479	
	0.032	0.096	0.222	0.263	0.386	0.101
	0.109	0.106	0.100	0.104	0.097	
	0.003	0.010	0.022	0.026	0.039	
J	119	307	678	808	896	2808
	14.772	10.427	3.823	11.300	45.486	
	0.042	0.109	0.241	0.288	0.319	0.052
	0.074	0.063	0.056	0.059	0.042	
	0.002	0.006	0.013	0.015	0.017	
Column Total	1610	4906	12082	13791	21551	53940
	0.030	0.091	0.224	0.256	0.400	

교차표 작성 / 분석

- 학력수준과 대학진학여부 교차분석(Package 이용)

```
# 학력수준(독립변수) : y -> 진학여부(종속변수) : x
# 학력수준이 대학 진학에 영향을 미친다.
x <- data$level2 # 행 - 리코딩 변수 이용
y <- data$pass2 # 열 - 리코딩 변수 이용

CrossTable(x,y) # x:학력수준, y:대학진학
```

교차표 작성 / 분석

● 부모의 학력수준과 자녀의 대학진학 여부

Total Observations in Table: 225

x	y 실패	합격	Row Total
고졸	40	49	89
	0.544	0.363	
	0.449	0.551	0.396
	0.444	0.363	
	0.178	0.218	
대졸	27	55	82
	1.026	0.684	
	0.329	0.671	0.364
	0.300	0.407	
	0.120	0.244	
대학원졸	23	31	54
	0.091	0.060	
	0.426	0.574	0.240
	0.256	0.230	
	0.102	0.138	
Column Total	90	135	225
	0.400	0.600	

- 기대치 비율 예 (1행2열)
- 기대치 : 89(행합)*135(열합)/225(전체합) =53.4
- 기대치 비율 : (49-53.4)^2/53.4=0.363

관측치
기대치비율(χ^2)=(관측치-기대치)^2/기대치
행비율
열비율
셀비율

관측치
(관측치-기대치)^2/기대치
행비율
열비율
셀비율

관측치
(관측치-기대치)^2/기대치
행비율
열비율
셀비율

전체 관측치
전체 열비율

교차표 작성 / 분석

❖논문에서 교차분석에 대한 해설 예

<교차분석 해설>----------

부모의 학력수준에 따른 자녀의 대학진학여부를 설문 조사한 결과 학력수준에 상관없이 대학진학 합격률이 평균 60%로 학력수준별로 유사한 결과가 나타났다. 전체 응답자 225명을 대상으로 고졸 39.6%(89명) 중 55.1%가 진학에 성공하였고, 대졸 36.4%(82명) 중 68.4%가 성공했으며, 대학원졸은 24%(54명) 중 57.4%가 대학진학에 성공하였다. 특히 대졸 부모의 대학진학 합격률이 평균보다 조금 높고, 고졸 부모의 대학진학 합격률이 평균보다 조금 낮은 것으로 분석된다.

---

chi-square 검정

- Chi-square 검정
  - 교차 분석으로 얻어진 교차 분할표를 대상으로 유의 확률을 적용하여 변수 간의 독립성 및 관련성 여부 등을 검정하는 분석 방법.
  - 범주(Category)별로 관측빈도와 기대빈도가 차이가 있는지 검정.
  - 카이 제곱 분포에 기초한 통계적 방법(카이 제곱 분포표 이용).
  - χ^2 = Σ (관측값 - 기댓값)^2 / 기댓값.
  - 분석을 위해서 교차 분할표 작성.
  - 교차분석은 검정 통계량으로 카이 제곱 사용(=카이 제곱 검정).
  - 검증 유형 분류 : 일원 카이 제곱 검정, 이원 카이 제곱 검정

chi-square 검정

1. 일원카이제곱 : 교차분할표 이용 안함(한 개 변인)
   - 적합성 검정 : 실제 표본이 내가 생각하는 분포와 같은가? 다른가?
     예) 관찰도수가 기대도수와 일치하는지를 검정

2. 이원카이제곱 : 교차분할표 이용
   1) 독립성 검정 : 두 변인는 서로 관련성이 있는가 없는가?
      - 한 모집단으로부터 하나의 표본이 추출된 경우
      - 예) 흡연량과 음주량 사이에 관련성이 있는가?
      - 귀무가설 : 흡연과 음주량은 관련성이 없다.(독립적이다.)

   2) 동일성 검정 : 두 집단의 분포가 동일한가? 다른 분포인가?
      - 두 개 이상의 범주형 자료가 동일한 분포를 갖는 모집단에서 추출된 것인지 검정하는 방법
      - 두 개 이상의 모집단에서 각 표본이 추출된 경우
      - 귀무가설 : 집단 간의 비율이 동일하다.

chi-square 검정

- Chi-square 검정 절차

1. 가설을 설정한다.
2. 유의수준을 결정한다.
3. 기각값(카이제곱 분포표 참조)을 결정한다.
   - 자유도(df)와 유의수준으로 기각값 결정
4. 관찰도수에 대한 기대도수를 구한다.
5. 검정통계량 x^2의 값을 구한다.
6. 귀무가설의 채택 또는 기각 여부를 판정한다.
7. 카이제곱 검정 결과를 설명한다.

chi-square 검정

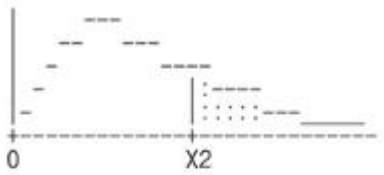
- 카이제곱 분포표

자유도

유의수준

자유도 = n-1
(n은 표본수)

CHI-SQUARE TABLE: VALUES OF CHI-SQUARE (ALPHA) OF THE CHI-SQUARE DISTRIBUTION



DF	X2(.995)	X2(.99)	X2(.975)	X2(.95)	X2(.05)	X2(.025)	X2(.01)	X2(.005)
1	0.000	0.000	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928

chi-square 검정

1. 일원카이제곱 검정

(1) 적합성 검정 - chisq.test() 이용

- 귀무가설 : 기대치와 관찰치는 차이가 없다.

  예) 도박사의 주사위는 게임에 적합하다.

- 대립가설 : 기대치와 관찰치는 차이가 있다.

  예) 도박사의 주사위는 게임에 적합하지 않다.

```
# 주사위의 관찰치가 기대치와 차이가 있는가? 또는 없는가?
# 60회 주사위를 던져서 나온 관측도수/기대도수
# 관측도수 : 4, 6, 17,16 ,8,9
# 기대도수 : 10,10,10,10,10,10
chisq.test(c(4,6,17,16,8,9)) # p-value = 0.01439
# 해설 : 도박사의 주사위는 게임에 적합하지 않다.
```

chi-square 검정

● p값 해석 방법

<해설> p값이 0.05미만이기 때문에 유의미한 수준에서 귀무 가설을 기각할 수 있다.
따라서 '도박사의 주사위는 게임에 적합하지 않다.'라는 대립가설을 채택한다.
(귀무 가설 기각, 대립가설 채택)

● 유의수준과 유의확률

유의수준(Confidence level) : 0.05(100개 중 5개(100*0.05) 허용 기준치(허용 오차)
유의확률 : p-value 귀무 가설이 나올 수 있는 확률
p-value < 0.05 경우 : 유의확률은 유의수준 보다 적다.(귀무 가설 기각)

● 검정통계량 해석 방법

검정통계량 : X-squared = 14.2, df = 5
자유도(df) : 관측치가 n 인 경우 df = n - 1
자유도(degree of freedom)란 검정을 위해서 n개의 표본(관측치)를 선정할 경우
n번째 표본은 나머지 표본이 정해지면 자동으로 결정되는 변인의 수를 의미
자유도(df) 5인 경우, X-squared 기각값(역) : x^2 >= 11.071(chi-square 분포표 참고)
χ^2 값이 11.071 이상이면 귀무 가설을 기각할 수 있다는 의미

chi-square 검정

(2) 선호도 분석
귀무가설 : 기대치와 관찰치는 차이가 없다.
예) 맥주의 선호도에 차이가 없다.
대립가설 : 기대치와 관찰치는 차이가 있다.
예) 맥주의 선호도에 차이가 있다.

```
data <- textConnection(
"맥주종류  관측도수
1   12
2   30
3   15
4   7
5   16")
x <- read.table(data, header=T)

chisq.test(x$관측도수) # X-squared = 18.375, p-value = 0.001042
# 해설 : 맥주의 선호도에 차이가 있다.
```

chi-square 검정

- 선호도 분석 결과
  - 검정통계량 :

    X-squared = 18.375, df = 4
  - p-value 해석 :

    p값이 0.05미만이기 때문에 유의미한 수준에서 귀무가설을 기각할 수 있다. 따라서 '맥주의 선호도에 차이가 있다.'라는 대립가설을 채택할 수 있다. (귀무가설 기각, 대립가설 채택)

chi-square 검정

2. 이원카이제곱 검정

1) 독립성 검정(관련성 검정) - 교차테이블 이용

- 동일 집단의 두 변인을 대상으로 관련성이 있는가? 없는가?를 검정하는 방법.

예) 귀무가설 : 부모의 학력수준과 자녀의 대학진학 여부와 관련성이 없다.

- 두 변인은 독립적이다.

예) 대립가설 : 부모의 학력수준과 자녀의 대학진학 여부와 관련성이 있다.

- 두 변인은 독립적이지 않다.

```
CrossTable(x, y, chisq = TRUE) # p = 0.2507057
```

chi-square 검정

- 독립성 검정(관련성 검정) 결과

x	y 실패	합격	Row Total
고졸	40 0.544 0.449 0.444 0.178	49 0.363 0.551 0.363 0.218	89 0.396
대졸	27 1.026 0.329 0.300 0.120	55 0.684 0.671 0.407 0.244	82 0.364
대학원졸	23 0.091 0.426 0.256 0.102	31 0.060 0.574 0.230 0.138	54 0.240
Column Total	90 0.400	135 0.600	225

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 2.766951 d.f. = 2 p = 0.2507057

<검정 결과 해설>

✓Chi^2 = Σ [(관측값 - 기댓값)2 / 기댓값]
✓d.f. = (행수-1)*(열-1) = (3-1)*(2-1) = 2
-> 두 값만 구하면 나머지는 저절로 구해진다.
✓ p = 유의수준 : 0.05이하이면 귀무가설 기각

자유도에 따른 Chi^2 분포도
-> 자유도가 클수록 정규분포에 가까워진다.
유의수준 0.05에서,
-> 자유도 : 2인 경우, 기각역 : x2 >= 5.99,
-> 자유도 : 6인 경우, 기각역 : x2 >= 12.59
자유도가 2인 경우 χ2 값이 5.99이상이면 귀무가설 기각(카이제곱 분포표 참조)
해설 : Chi^2값이 5.99 이하이고, 유의수준이 0.05 이상으로 분석되어 귀무가설을 기각할 수 없다. 따라서 부모의 학력수준과 자녀의 대학 진학 변인 간의 관련성은 없는 것으로 분석된다.

chi-square 검정

❖논문에서 교차분석표와 Chi-square 검정에 대한 해설 예

<교차분석표와 카이제곱 검정결과 해설>----------

'부모의 생활수준과 자녀의 대학진학 여부와 관련성이 있다.'를 분석하기 위해서 자녀를 둔 A회사 225명의 부모를 표본으로 추출한 후 설문 조사하여 교차분석과 카이제곱 검정을 실시하였다.

분석 결과를 살펴보면 부모의 생활수준과 자녀의 대학진학 여부의 관련성은 유의미한 수준에서 차이가 없는 것으로 나타났다.($X^2=2.767$, $p>0.05$)

따라서 귀무가설을 기각할 수 없다. 다음 <표>에서 부모의 생활 수준과 자녀의 대학 진학 여부에 대한 교차표와 카이제곱 검정결과를 제시하고 있다.

---

chi-square 검정

<논문에서 카이제곱 검정 결과 제시방법>

카이제곱 검정결과를 논문에서 제시할 경우 교차표와 카이제곱 검정통계량 함께 제시

학력수준		실패	진학	X-squared	유의확률(p)
고졸	관찰빈도	40	49	2.766951	0.2507057
	기대빈도	36	54		
대졸	관찰빈도	27	55		
	기대빈도	33	49		
대학원졸	관찰빈도	23	31		
	기대빈도	21	32		

chi-square 검정

<실습> 교육수준과 흡연율 간의 관련성 분석

1. 파일 가져오기

```
setwd("c:/workspaces/Rwork/data")
smoke <- read.csv("smoke.csv", header=TRUE)
# 변수 보기
head(smoke) # education, smoking 변수
names(smoke)
[1] "education" "smoking"
```

- 변수 모델링

  객체를 대상으로 분석할 속성(변수)을 선택하여 속성 간의 관계 설정 과정

  예) smoke 객체에서 education, smoking 속성을 분석대상으로 하여 교육수준이 흡연율과 관련성이 있는가를 education -> smoking 형태로 기술한다.
  education은 영향을 미치는 변수로 독립변수라 하며, 영향을 받는 smoking은 종속변수라고 한다.

chi-square 검정

2. 코딩 변경 - 변수 리코딩 <- 가독성 제공
education(독립변수) : 1:대졸, 2:고졸, 3:중졸
smoke(종속변수): 1:과다흡연, 2:보통흡연, 3:비흡연

```
table(smoke$education, smoke$smoking)
smoke$education2[smoke$education==1] <- "대졸"
smoke$education2[smoke$education==2] <- "고졸"
smoke$education2[smoke$education==3] <- "중졸"

smoke$smoking2[smoke$smoking==1] <- "과대흡연"
smoke$smoking2[smoke$smoking==2] <- "보통흡연"
smoke$smoking2[smoke$smoking==3] <- "비흡연"
smoke # 가독성을 위한 변수값 변경 결과
```

chi-square 검정

3. 교차표 작성

table(smoke$education2, smoke$smoking2)

	과대흡연	보통흡연	비흡연
고졸	22	21	9
대졸	51	92	68
중졸	43	28	21

chi-square 검정

4. 독립성 검정

```
library(gmodels) # CrossTable() 함수 사용
CrossTable(smoke$education2, smoke$smoking2, chisq = TRUE)
Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  18.91092      d.f. =  4      p =  0.0008182573
```

chi-square 검정

2) 동질성 검정 - 교차테이블 이용

- 두 집단의 분포가 동일한가? 분포가 동일하지 않는가?를 검정하는 방법.

- 즉, 동일한 분포를 가지는 모집단에서 추출된 것인지를 검정하는 방법.

예) 귀무가설 : 집단 간의 비율이 동일하다.

예) 교육방법에 따른 만족도에 차이가 없다.

예) 대립가설 : 집단 간의 비율이 동일하지 않다.

예) 교육방법에 따른 만족도에 차이가 있다.

chi-square 검정

1. 파일 가져오기

```
setwd("c:/workspaces/Rwork/data")
data <- read.csv("homogenity.csv", header=TRUE)
head(data) # 변수 보기
data <- subset(data, !is.na(survey), c(method, survey))
```

chi-square 검정

2. 변수리코딩 - 코딩 변경

```
# method: 1:방법1, 2:방법2, 3:방법3
# survey: 1:매우 만족, 2:만족, 3:보통, 4: 불만족, 5: 매우 불만족

# 교육방법2 필드 추가
data$method2[data$method==1] <- "방법1"
data$method2[data$method==2] <- "방법2"
data$method2[data$method==3] <- "방법3"

# 만족도2 필드 추가
data$survey2[data$survey==1] <- "매우 만족"
data$survey2[data$survey==2] <- "만족"
data$survey2[data$survey==3] <- "보통"
data$survey2[data$survey==4] <- "불만족"
data$survey2[data$survey==5] <- "매우 불만족"
```

chi-square 검정

3. 교차분할표 작성

table(data$method2, data$survey2) # 교차표 생성 -> table(행,열)

	만족	매우만족	매우불만족	보통	불만족	
방법1	8	5	6	15	16	-> 50
방법2	14	8	6	11	11	-> 50
방법3	7	8	9	11	15	-> 50

주의 : 반드시 각 집단별 길이(50)가 같아야 한다.

chi-square 검정

4. 동질성 검정 - 모수 특성치에 대한 추론검정

```
chisq.test(data$method2, data$survey2)
        Pearson's Chi-squared test

data:  data$method2 and data$survey2
X-squared = 6.5447, df = 8, p-value = 0.5865
```

<해설>

유의수준 0.05에서 x2값이 6.545, 자유도 8, 그리고 유의확률 0.586을 보이고 있다. 즉 6.545 이상의 카이제곱값이 얻어질 확률이 0.586라는 것을 보여주고 있다.

이 값은 유의수준 0.05보다 크기 때문에 귀무가설을 기각할 수 없다. 따라서 '교육방법에 따른 만족도에 차이가 없다.'라고 할 수 있다.