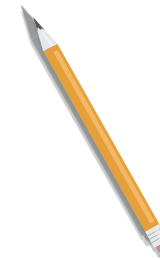


4강

범주형 자료 분석

정보통계학과
이태림 교수



목차

1. 범주형 자료와 분할표

범주형 자료
분할표 정의

2. 범주형 자료의 검정

동질성 검정

독립성 검정

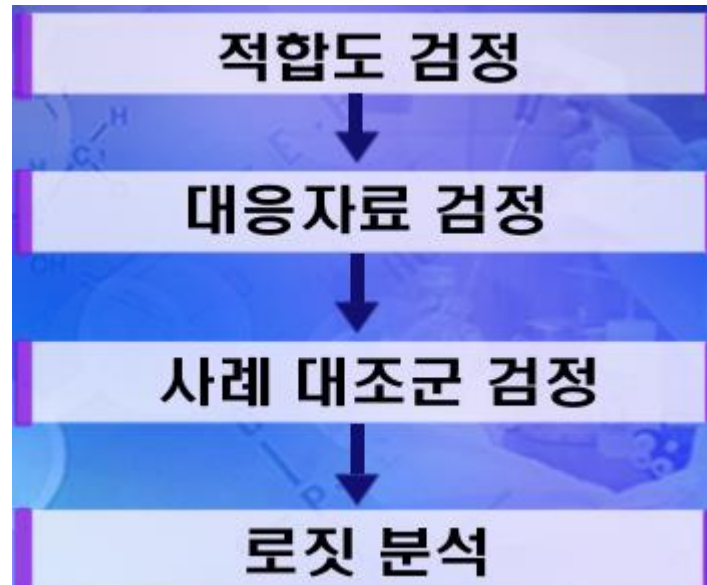
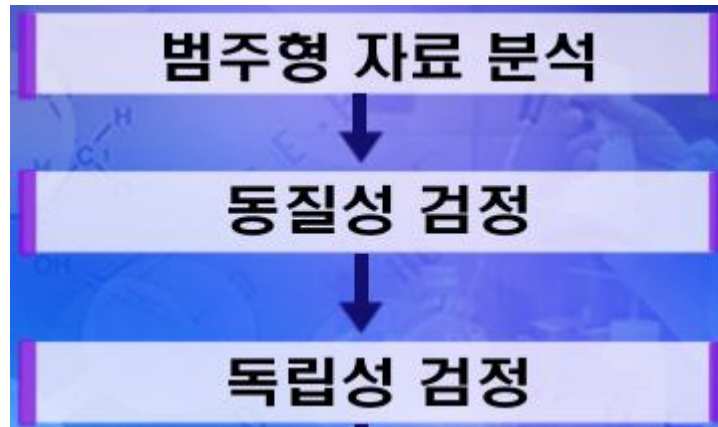
적합도검정

대응자료 및
사례대조군 검정

3. 로짓분석

4. 통계패키지 수행

R 자료분석



I 범주형 자료(Categorical Data Analysis)

4강 범주형 자료의 분석

- 사회조사, 임상시험에서 얻는 특성에 의해
구분된 자료

예) 성별(남:1, 여:2),
종교(기독교:1, 불교:2, 카톨릭교:3), 인종

- ➡ 반응인자에 대한 설명인자의 영향을 평가
- ➡ 덧셈, 뺄셈 등의 연산을 정의할 수 없다.

유형

이분형 척도
명목형 척도
순서형 척도

I 분할표(contingency table)

4강 범주형 자료의 분석

관찰치를 몇 개의 범주로 분할하여 그 해당도수로 자료를 정리해 놓은 표

유형

- 일원분할표(one-way contingency table)
- 이원분할표(two-way contingency table)
- 삼원분할표(three-way contingency table)

I 분석: 독립성 검증

4강 범주형 자료의 분석

예제



10마리의 실험용 쥐를 다섯 마리씩 임의로 두 군으로 나누어 한 군에는 특정약물을 투여하고 (처리군), 다른 군에는 아무 처리도 하지 않는다. (대조군) 13주 후 생존여부를 관찰하여 특정 약물의 효과를 검정한다.

분할표

	생존	사망	계
대조군	4	1	5
처리군	2	3	5
계	6	4	10



예 4.1

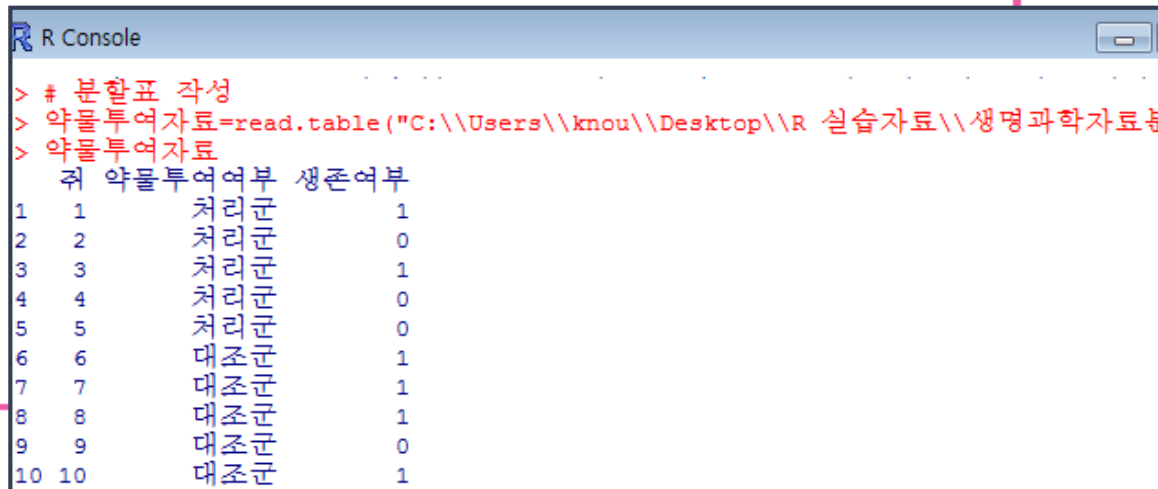
P 127

10마리의 실험용 쥐를 5마리씩 나누어서 두 군을 구성하고 한 군에는 특정 약물을 투여하고(처리군), 다른 군에는 아무 처리도 하지 않았다(대조군). 그리고 13주 후 생존 여부(생존=1, 사망=0)를 관찰하여 <표 4.1>과 같은 원자료를 얻었다. 이 자료에 대한 분할표를 R 프로그램을 이용하여 구해 보자.

표 4.1 실험용 쥐의 13주 후 생존 여부

쥐	약물투여 여부	생존 여부	쥐	약물투여 여부	생존 여부
1	처리군	1	6	대조군	1
2	처리군	0	7	대조군	1
3	처리군	1	8	대조군	1
4	처리군	0	9	대조군	0
5	처리군	0	10	대조군	1

```
> #===== 4.1 범주형 자료와 분할표
> # <표 4.1> 실험용 쥐 자료 읽기
> 약물투여자료=read.table("e:\\WORK\\약물투여자료.txt",header=T)
> 약물투여자료
  약물투여여부 생존여부
1      처리군      1
2      처리군      0
3      처리군      1
4      처리군      0
5      처리군      0
6      대조군      1
7      대조군      1
8      대조군      1
9      대조군      0
10     대조군      1
> attach(약물투여자료)
```



R Console

```
> # 분할표 작성
> 약물투여자료=read.table("C:\\Users\\knou\\Desktop\\R 실험자료\\생명과학자료용
> 약물투여자료
  쥐 약물투여여부 생존여부
1  1      처리군      1
2  2      처리군      0
3  3      처리군      1
4  4      처리군      0
5  5      처리군      0
6  6      대조군      1
7  7      대조군      1
8  8      대조군      1
9  9      대조군      0
10 10     대조군      1
```



```
> attach(약물투여)
> 약물투여=table(약물투여)
> colnames(약물투여)=c("사망", "생존")
> rownames(약물투여)=c("대조군", "처리군")
> 약물투여
> 주변표1=margin.table(약물투여, 1)
> 주변표1
> 주변표2=margin.table(약물투여, 2)
> 주변표2
```

R Console

```
> attach(약물투여자료)
> 약물투여=table(약물투여여부, 생존여부)
> colnames(약물투여)=c("사망", "생존")
> rownames(약물투여)=c("대조군", "처리군")
> 약물투여
```

	생존여부	
약물투여여부	사망	생존
대조군	1	4
처리군	3	2

```
> 주변표1=margin.table(약물투여, 1)
> 주변표1
```

약물투여여부	
대조군	처리군
5	5

```
> 주변표2=margin.table(약물투여, 2)
> 주변표2
```

생존여부	
사망	생존
4	6

>분할표=addmargins(약물투여)

>분할표

백분율표 작성

>대조군백분율=약물투여

>처리군백분율=약물투여

>행_백분율=cbind

>행_백분율

>사망백분율=약물투여

>생존백분율=약물투여

>열_백분율=cbind

>열_백분율

R Console

```
> 분할표=addmargins(약물투여)
```

```
> 분할표
```

	생존여부		
약물투여여부	사망	생존	Sum
대조군	1	4	5
처리군	3	2	5
Sum	4	6	10

```
> # 백분율표 작성
```

```
> 대조군백분율=약물투여[1,]/주변표1[1]
```

```
> 처리군백분율=약물투여[2,]/주변표1[2]
```

```
> 행_백분율=cbind(대조군백분율,처리군백분율)
```

```
> 행_백분율
```

	대조군백분율	처리군백분율
사망	0.2	0.6
생존	0.8	0.4

```
> 사망백분율=약물투여[,1]/주변표2[1]
```

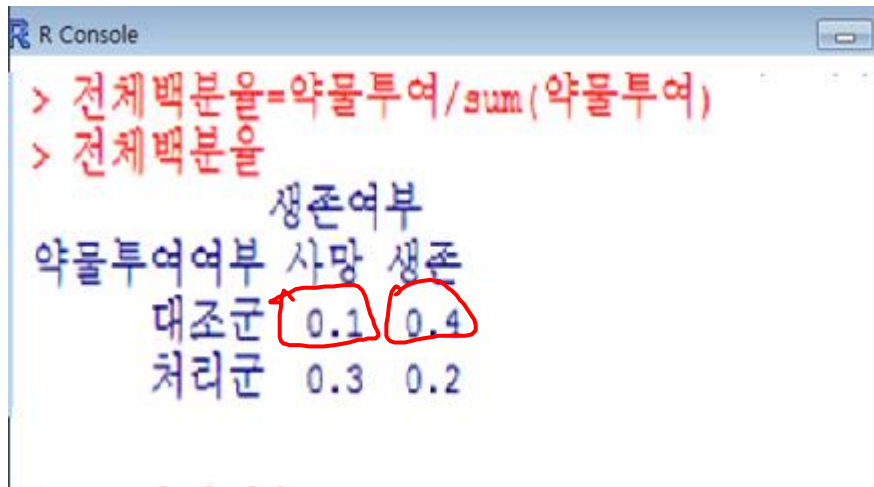
```
> 생존백분율=약물투여[,2]/주변표2[2]
```

```
> 열_백분율=cbind(사망백분율,생존백분율)
```

```
> 열_백분율
```

	사망백분율	생존백분율
대조군	0.25	0.6666667
처리군	0.75	0.3333333

> 전체백분율=약물투여/sum(약물투여)
> 전체백분율



R Console

```
> 전체백분율=약물투여/sum(약물투여)  
> 전체백분율
```

	생존여부	
약물투여여부	사망	생존
대조군	0.1	0.4
처리군	0.3	0.2

예제

유전자 종류	A	B	C	계
관측도수	18	55	27	100
이론도수	$25 \left[\frac{1}{4} \right]$	$50 \left[\frac{1}{2} \right]$	$25 \left[\frac{1}{4} \right]$	100

검정 : 생물학적 이론에 의한 귀무가설

$$H_0 : P_A = \frac{1}{4} \quad P_B = \frac{1}{2} \quad P_C = \frac{1}{4}$$

가설

$$H_0 : P_{ij} = P_i \times P_j \text{ [독립성 검정]}$$

$$H_1 : H_0 \text{ 가 아니다}$$

검정통계량

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2$$

여기서 $E_i = n_i P_i$ [일원분할표]

$$= \frac{O_i \times O_j}{N}$$

기각역

$$\chi_0^2 \geq \chi_{\alpha}^2 \text{ 이면 } H_0 \text{ 기각}$$

한 변수의 각 수준에 따른 다른 변수의 분포가
같은가의 검정

예제 : 이중눈가림 동질성 검정

비타민 C가 감기 치료에 효과가 있는지 알아보기 위하여
스키를 타는 사람 279명을 대상으로 대조군 140명에게는
당사자들이 모르게 플라시보(placebo)를 주고, 대조군과
독립적으로 선택된 처리군 139명에게는 당사자들이 모르게
매일 비타민 1g을 투여하여 검진 의사로 피진찰자가
어떤 약을 복용했는지 모르는 상태에서 진찰하여
결과를 얻는다.

비타민 C 복용과 감기에 관한 이중 눈가림 연구

	감기 걸림	감기 안걸림	계
대조군(placebo)	31(24.1)	109(115.9)	140
비타민C 복용군	17(23.9)	122(115.1)	139
계	48	231	279

가설

$$H_0 : \pi_C = \pi_p$$

$$H_1 : \pi_C \neq \pi_p$$

기대치 산출

$$E(X_1) = 140 \times \hat{\pi}_i = 140 \times \frac{48}{279} = 24.1$$

$$E(X_2) = 139 \times \hat{\pi}_i = 139 \times \frac{48}{279} = 23.9$$

비타민 C 복용과 감기에 관한 이중 눈가림 연구

	감기 걸림	감기 안걸림	계
대조군(placebo)	31(24.1)	109(115.9)	140
비타민C 복용군	17(23.9)	122(115.1)	139
계	48	231	279

가설

$$H_0 : \pi_C = \pi_p$$

$$H_1 : \pi_C \neq \pi_p$$

검정통계량

$$\chi^2_0 = \sum_i \sum_j \frac{(O_i - E_i)^2}{E_i} = \frac{(31 - 24.1)^2}{24.1} + \frac{(109 - 115.9)^2}{115.9} + \dots = 4.81$$

비타민 C 복용과 감기에 관한 이중 눈가림 연구

	감기 걸림	감기 안걸림	계
대조군(placebo)	31(24.1)	109(115.9)	140
비타민C 복용군	17(23.9)	122(115.1)	139
계	48	231	279

가설

$$H_0 : \pi_C = \pi_p$$

$$H_1 : \pi_C \neq \pi_p$$

검정통계량

$$x_0^2 = \sum_i \sum_j \frac{(O_i - E_i)^2}{E_i} = \frac{(31 - 24.1)^2}{24.1} + \frac{(109 - 115.9)^2}{115.9} + \dots = 4.81$$

검정결과

$x_0 = 4.81 > x_{(1)} = 3.841$ H_0 를 기각하여 비타민 C의 복용에 따라 감기에 걸리는 비율이 다르다고 결론 짓는다.

동질성 검정

```
# 카이검정
chisq.test(비타민)
# 관찰도수
chisq.test(비타민)
# 기대도수
chisq.test(비타민)
# 피어슨잔차
chisq.test(비타민)
```

```
R Console
> # 카이검정
> chisq.test(비타민효과)

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

data:  비타민효과
X-squared = 4.1407, df = 1, p-value = 0.04186

> # 관찰도수
> chisq.test(비타민효과)$observed
      감기여부
비타민복용여부 감기 결핍 감기 안 결핍
대조군          31      109
처리군          17      122

> # 기대도수
> chisq.test(비타민효과)$expected
      감기여부
비타민복용여부 감기 결핍 감기 안 결핍
대조군  24.08602    115.914
처리군  23.91398    115.086

> # 피어슨잔차
> chisq.test(비타민효과)$residuals
      감기여부
비타민복용여부 감기 결핍 감기 안 결핍
대조군   1.408787   -0.6421849
처리군  -1.413846    0.6444908
```

동질성 검정

비타민 C복용과 감기에 관한 이중 눈가림 연구

비타민효과=matrix(c(31,109,17,122),nrow=2,byrow=T)

R Console

```
>
> 비타민효과=matrix(c(31,109,17,122),nrow=2,byrow=T)
> dimnames(비타민효과)=list(비타민복용여부=c("대조군","처리군"),감기여부=
> 비타민효과
```

		감기여부	
비타민복용여부	감기	결핍	감기 안 결핍
대조군	31	109	
처리군	17	122	

```
> 분할표_비타민효과=addmargins(비타민효과)
> 분할표_비타민효과
```

		감기여부		
비타민복용여부	감기	결핍	감기 안 결핍	Sum
대조군	31	109		140
처리군	17	122		139
Sum	48	231		279



예 4.3

굿맨과 크루스컬은 6,800명을 대상으로 눈색과 머리색을 조사하여 얻은 자료를 가지고 <표 4.10>과 같은 관찰표를 얻었다. 눈색과 머리색에 따라 분류하여 3×4 분할표를 구성할 때 눈색이 머리색에 영향을 주는가? 즉, 서로 독립적인가?

	금발	검정	빨강	갈색	계
초록	1768	807	189	47	2811
검정	976	1387	746	53	3132
파랑	115	438	288	16	857
계	2829	2632	1223	116	6800

예제 - Goodman과 Kruskal의 독립성 검정

가설

H_0 : 눈색과 머리카락색은 독립이다

$$P_{ij} = P_i \times P_j$$

H_1 : 눈색과 머리카락색은 서로 관련이 있다

검정통계량

$$x_0^2 = \sum \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 1073.5076$$

검정결과

$x_0^2 = 1073.5 > x_\alpha = 14.45$ 이므로 H_0 를 기각

➡ 눈색과 머리카락색이 유의하게 서로 영향을 준다

```
> # 모자이크그림(mosaic plot)  
> mosaicplot(t(눈색_머리색),shade=T,main="")
```

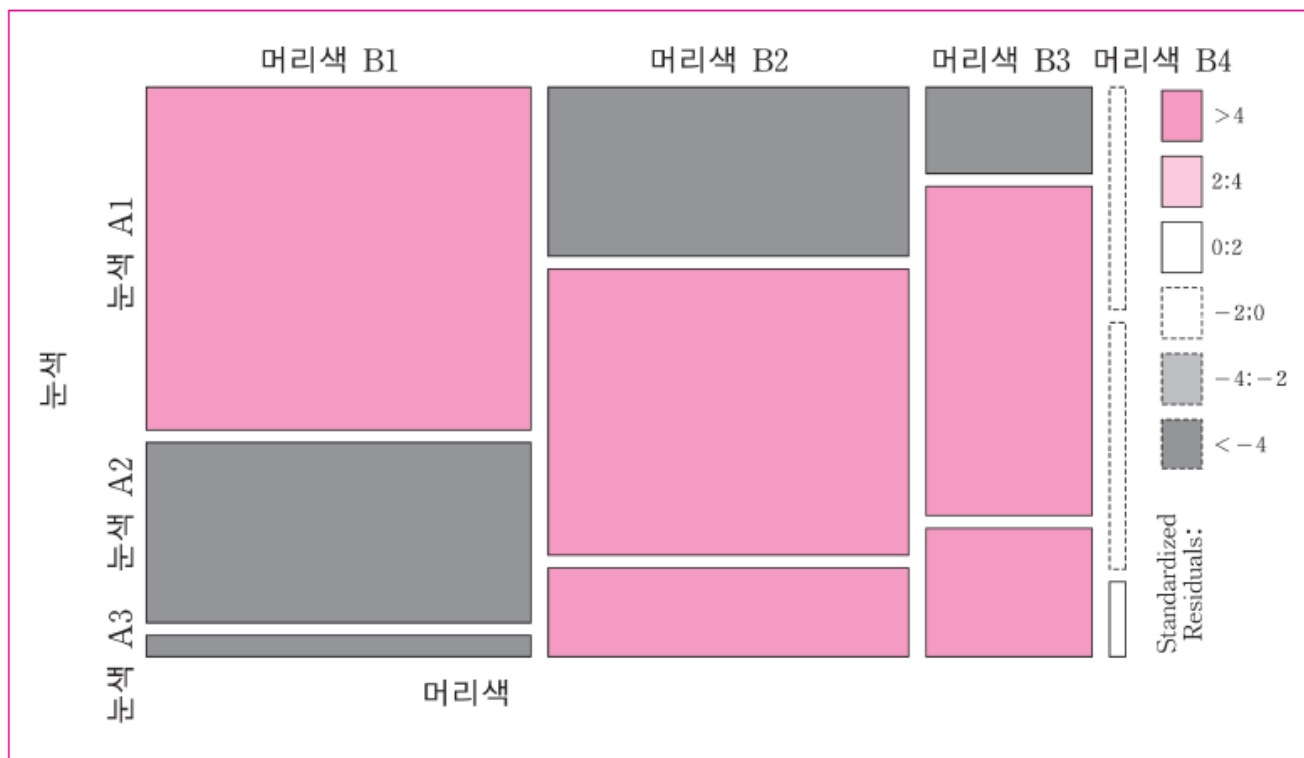


그림 4.4 눈색과 머리색 분할표에 대한 모자이크그림

독립성 검정

Goodman과 Kruskal의 독립성 검정

```
눈색_머리색=matrix(c(1768,807,189,47,946,1387,746,53,115,438,288,16),nrow=3,ncol=4)
```

```
dimnames(눈색_머리색)=list(눈색=c("눈색 A1","눈색 A2","눈색 A3"),
                             머리색=c("머리색 B1","머리색 B2","머리색 B3","머리색 B4"))
```

눈색	머리색 B1	머리색 B2	머리색 B3	머리색 B4
눈색 A1	1768	807	189	47
눈색 A2	946	1387	746	53
눈색 A3	115	438	288	16

```
> 분할표_눈색_머리색=addmargins(눈색_머리색)
```

눈색	머리색 B1	머리색 B2	머리색 B3	머리색 B4	Sum
눈색 A1	1768	807	189	47	2811
눈색 A2	946	1387	746	53	3132
눈색 A3	115	438	288	16	857
Sum	2829	2632	1223	116	6800

```
> # 카이제곱검정  
> chisq.test(눈색_머리색)
```

Pearson's Chi-sq

```
data: 눈색_머리색  
X-squared=1073.5, df=6,
```

```
> # 관찰도수
```

```
> chisq.test(눈색_머리색)$observed  
머리색
```

눈색	머리색 B1	머리색 B2	머리색 B3	머리색 B4
눈색 A1	1768	807	189	47
눈색 A2	946	1387	746	53
눈색 A3	115	438	288	16

```
> # 기대도수
```

```
> chisq.test(눈색_머리색)$expected  
머리색
```

눈색	머리색 B1	머리색 B2	머리색 B3	머리색 B4
눈색 A1	1169.4587	1088.0224	505.5666	47.95235
눈색 A2	1303.0041	1212.2682	563.2994	53.42824
눈색 A3	356.5372	331.7094	154.1340	14.61941

```
> # 피어슨잔차
```

```
> chisq.test(눈색_머리색)$residuals  
머리색
```

눈색	머리색 B1	머리색 B2	머리색 B3	머리색 B4
눈색 A1	17.502565	-8.519654	-14.079133	-0.13752858
눈색 A2	-9.890092	5.018483	7.697865	-0.05858643
눈색 A3	-12.791799	5.836008	10.782543	0.36107650

$$\lambda = \frac{P[\text{①의 예측 오류}] - P[\text{②의 예측 오류}]}{P[\text{①의 예측 오류}]} = 0.1924$$

→ 눈의 색을 알고 머리
예측할 때보다 예측오
줄일 수 있다.

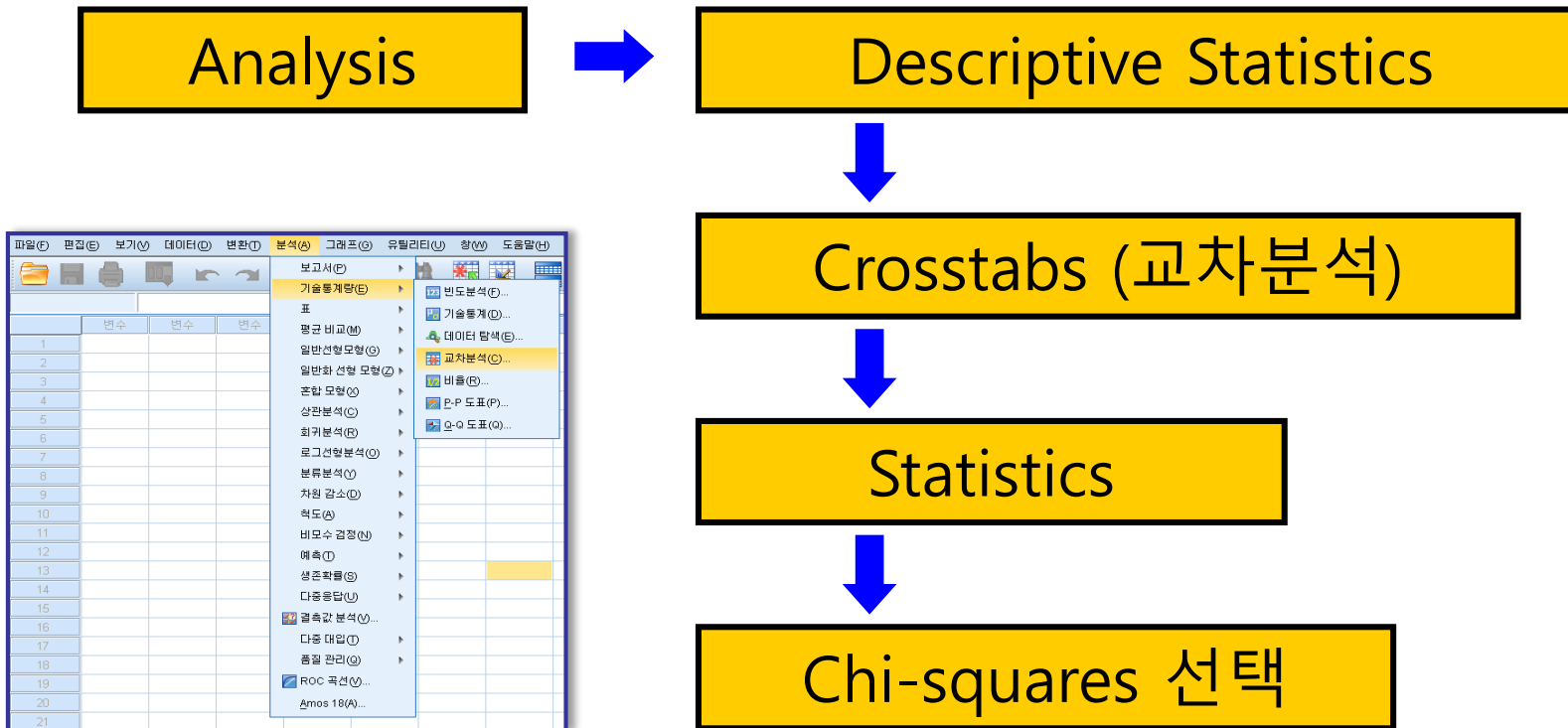
```
> install.packages("DescTools")
> library(DescTools)
> Lambda(눈색_머리색)
[1] 0.2076188
> # lambda의 95% 신뢰구간
> Lambda(눈색_머리색, conf.level=0.95)
      lambda   lwr.ci   upr.ci
0.2076188 0.1871747 0.2280629
> # Lambda(R|C)
> Lambda(눈색_머리색, direction="row")
[1] 0.2241003
> # Lambda(C|R)
> Lambda(눈색_머리색, direction="column")
[1] 0.1923949
```



■ 지역별 정당 선호도 조사 분할표

정당 지역	A 정당	B 정당	C 정당	합
경기	158	53	62	273
서울	172	128	83	383
영남	95	162	27	284
호남	40	21	99	160
합	465	364	271	1100

- 이차원분할표에서 독립성검정을 실시하는 수행 절차



I 분할표 자료 입력

4강 범주형 자료의 분석

표5-7.sav [데이터집합3] - SPSS Statistics Data Editor

파일(F) 편집(E) 보기(V) 데이터(D) 변환(T) 분석(A) 그래프(G) 유틸리티(U) 추가 기능(O) 창(W) 도움말(H)

15: 지지자수 표사: 3 / 3 변수

	지역	정당	지지자수	변수	변수	변수	변수
1	1	1	158				
2	1	2	53				
3	1	3	62				
4	2	1	172				
5	2	2	128				
6	2	3	83				
7	3	1	95				
8	3	2	162				
9	3	3	27				
10	4	1	40				
11	4	2	21				
12	4	3	99				

데이터 보기 변수 보기

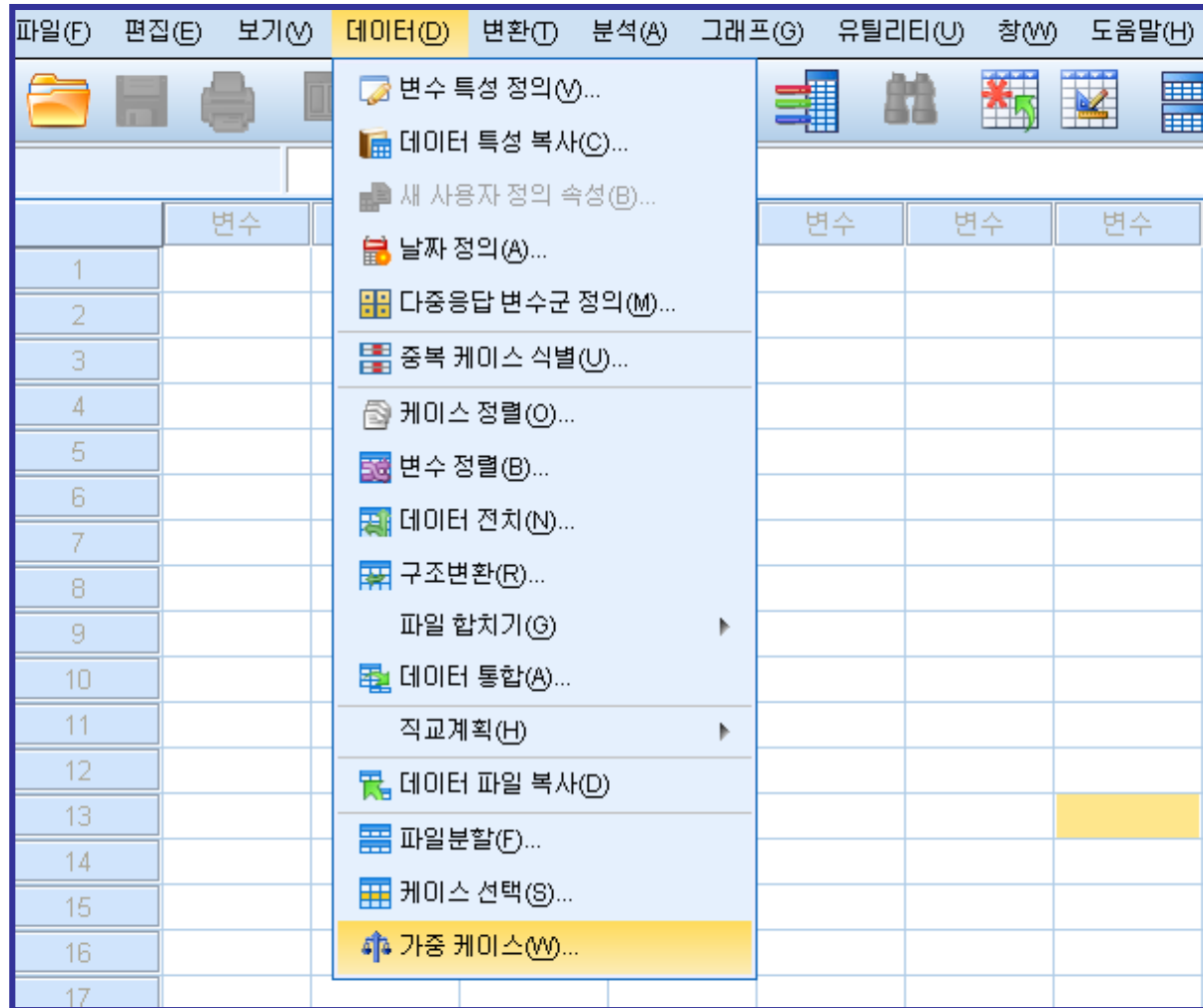
SPSS Statistics Processor is ready

가중 설정

정당 지역	A 정당	B 정당	C 정당	합
경기	158	53	62	273
서울	172	128	83	383
영남	95	162	27	284
호남	40	21	99	160
합	465	364	271	1100

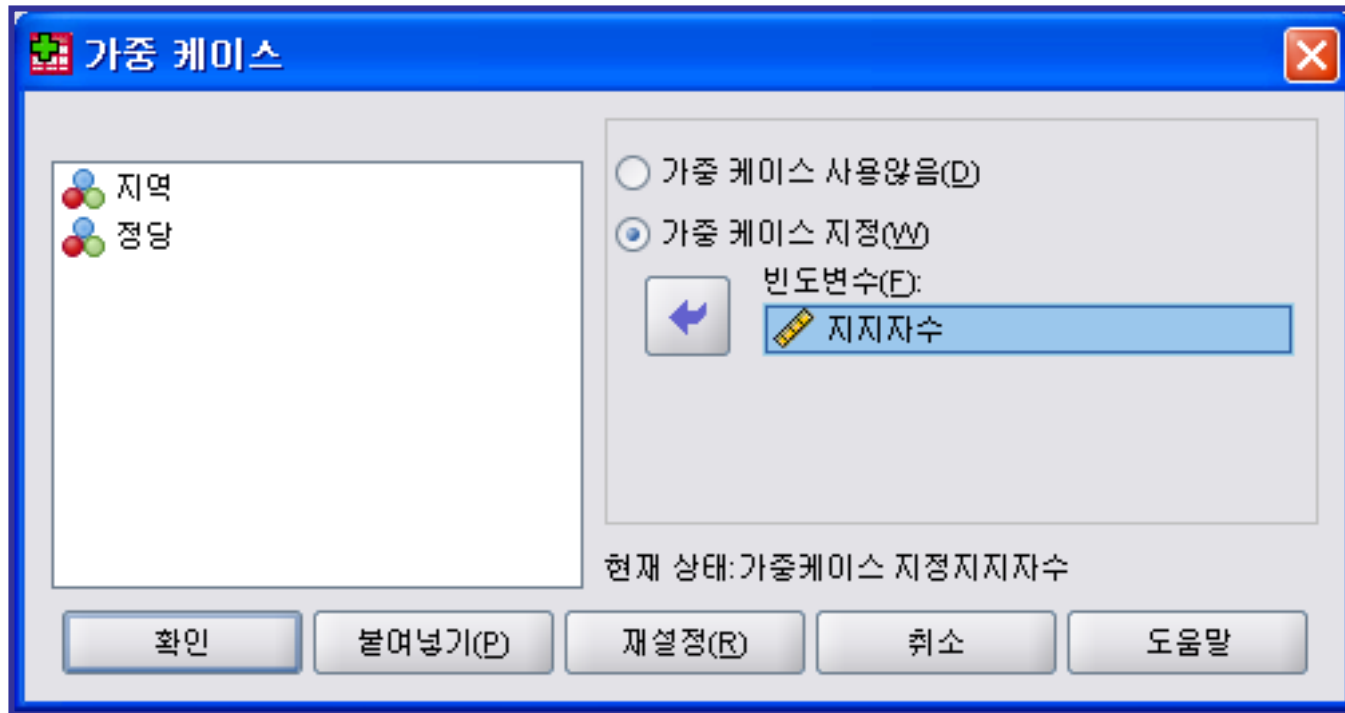
I 데이터-가중케이스를 선택

4강 범주형 자료의 분석



I 변수 “지지자수”를 가중변수로 선택

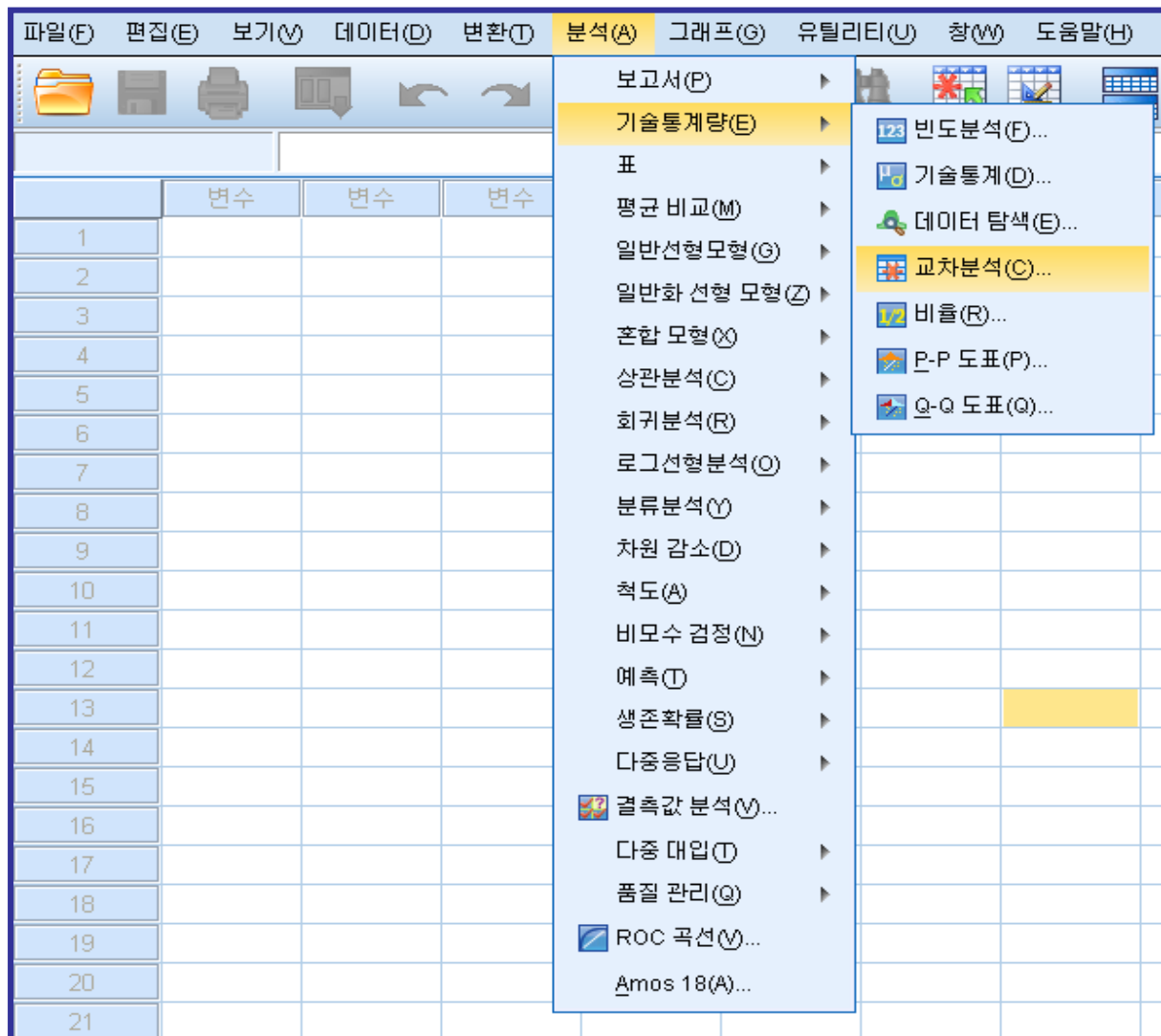
4강 범주형 자료의 분석





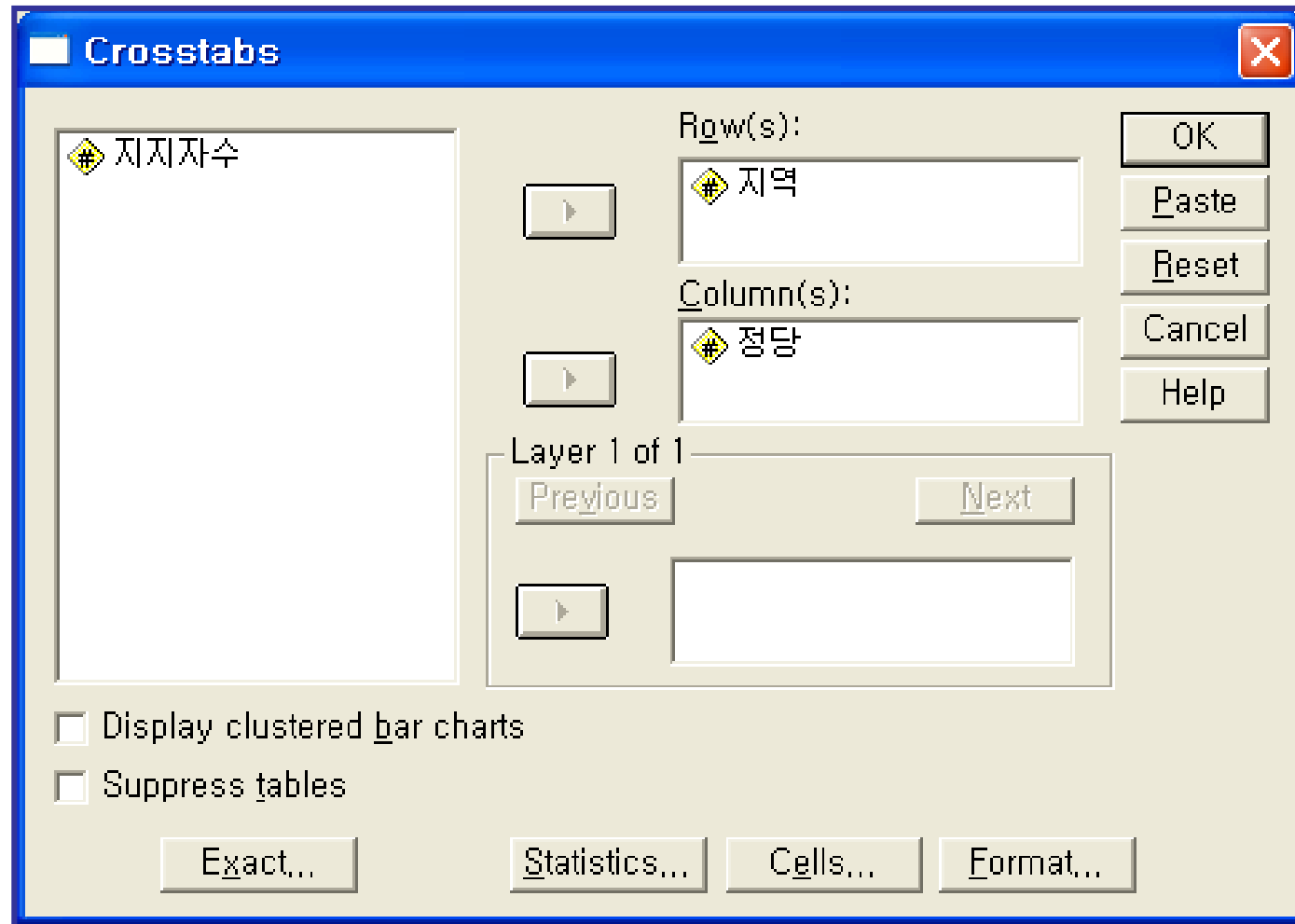
교차분석 메뉴 선택(분석-기술통계량-교차분석)

4강 범주형 자료의 분석



I 교차분석 대화상자

4강 범주형 자료의 분석



Crosstabs: Statistics

☒ Chi-square

☐ Correlations

Nominal

☐ Contingency coefficient

☐ Phi and Cramer's V

☐ Lambda

☐ Uncertainty coefficient

Ordinal

☐ Gamma

☐ Somers' d

☐ Kendall's tau-b

☐ Kendall's tau-c

Nominal by Interval

☐ Eta

☐ Kappa

☐ Risk

☐ McNemar

☐ Cochran's and Mantel-Haenszel statistics

Test common odds ratio equals: 1

Continue

Cancel

Help

I 독립성 검정 결과

4강 범주형 자료의 분석

지역 * 정당 교차표

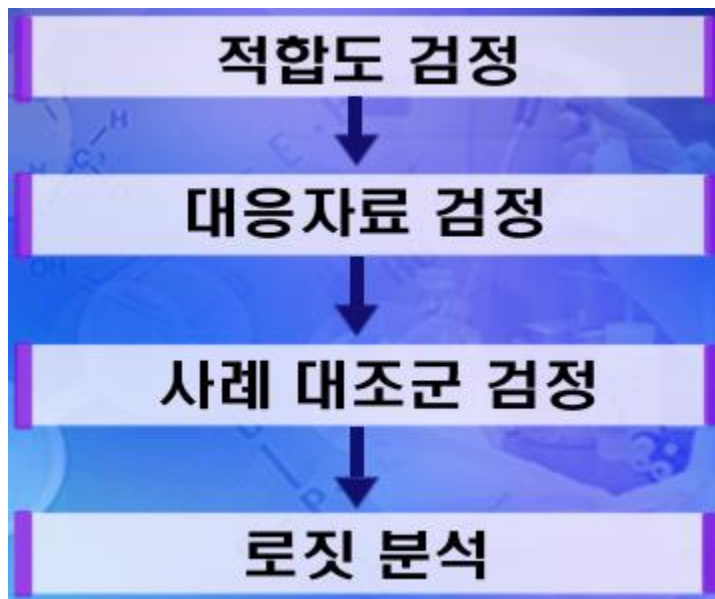
빈도

		정당			전 체
		A정당	B정당	C정당	
지역	경기	158	53	62	273
	서울	172	128	83	383
	영남	95	162	27	284
	호남	40	21	99	160
전 체		465	365	271	1100

카이제곱 검정

	값	자유도	점근 유의확률 (양측검정)
Pearson 카이제곱	235.051^a	6	.000
우도비	216.721	6	.000
선형 대 선형결합	59.606	1	.000
유효 케이스 수	1100		

a. 0셀(.0%)은(는) 5보다 작은 기대 빈도를 가지는 셀입니다.
최소 기대빈도는 39.42입니다.



I 적합도 검정(Coodness of fit)

4강 범주형 자료의 분석

- 관찰된 도수가 정규분포 또는 이항분포 등의 이론분포와 일치하는가의 검정

예

- 재배한 강낭콩이 멘델의 유전법칙을 따르는가?
- 2차 세계대전 당시 한 지역에 투하된 폭탄의 분포가 포아송 분포를 따르는가?
- 여성 운전자들의 승용차 구매유형이 우리나라 전체 구매 선호도와 일치하는가?

I 적합도 검정(Goodness of fit)

4강 범주형 자료의 분석

변수 X 에 대한 확률분포

X	1	2	3	...	k	계
$P(X = x)$	p_1	p_2	p_3	...	p_k	1.0

모집단의 확률분포

$$(p_{10}, p_{20}, \dots, p_{k0})$$

변수 X 에 대한 확률분포

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

H_1 : 적어도 하나의 p_i 는 가정된 p_{i0} 와 다르다.

검정통계량

$$\chi_0^2 = \sum \frac{(O_i - E_i)^2}{E_i} > \chi_\alpha^2 (k - 1) \text{ 이면 } H_0 \text{ 기각}$$

2차 세계대전 당시 런던 남부에 투하된 폭탄의 분포



예 4.4

2차대전 당시 런던 남부지역에 떨어진 V_1 폭탄의 지역적 분포를 조사하였더니 <표 4.13>과 같았다. 전체 지역은 $1/4\text{km}^2$ 를 단위지역으로 하였을 때 576개의 단위지역으로 구성되고, 이 지역에 떨어진 폭탄의 총 개수는 537개였다. <표 4.13>에서 보듯이 576개 단위지역 중 폭탄이 하나도 안 떨어진 곳은 229개이고 폭탄이 1개만 떨어진 단위지역은 211개이다. “폭탄 수의 분포가 포아송 분포를 이루는가?” 혹은 “평균이 1인 포아송 분포를 이루는가?” 하는 질문에 대한 해답을 알아보자.

$$X \sim B\left(537, \frac{1}{576}\right)$$

$$X \sim P\left(\frac{537}{576}\right)$$

표 4.13 런던 남부지역에 투하된 V_1 폭탄의 분포

폭탄의 수	단위지역의 수	폭탄의 수	단위지역의 수
0	229	3	35
1	211	4	7
2	93	≥ 5	1
		총합	576

I 적합도 검정(Goodness of fit)

4강 범주형 자료의 분석

예 2차 세계대전 당시 런던 남부에 투하된 폭탄의 분포

포아송 분포하의 기대도수

표 4.14 X 의 관찰도수와 $P(1)$ 아래에서의 기대도수

i	관찰도수(O_i)	$P_r(X = i)$	기대도수(E_i)
0	229	0.3679	211.90
1	211	0.3679	211.90
2	93	0.1839	105.95
3	35	0.0613	35.32
4	7	0.0153	8.83
≥ 5	1	0.0037	2.10
	576	1.0000	576

I 적합도 검정(Goodness of fit)

4강 범주형 자료의 분석

```
> #===== 4.2.3 적합도검정
> # [예 4.5]
> 폭탄수=c(0, 1, 2, 3, 4, 5)
> 관찰도수=c(229, 211, 93, 35, 7, 1)
> names(관찰도수) <- 폭탄수
> 확률=c(dpois(0:4, 1), 1-sum(dpois(0:4, 1)))
> chisq.test(관찰도수, p=확률)
```

Chi-squared test for given probabilities

data: 관찰도수

X-squared=3.9309, df=5, p-value=0.5594

Warning message:

In chisq.test(관찰도수, p=확률) :

카이제곱 approximation은 정확하지 않을수도 있습니다

검정통계량

➡ 폭탄 투하 분포는 포아송 분포를 따른다고
결론짓는다.

I 대응자료 및 사례-대조군 검정 (McNemar test)

4강 범주형 자료의 분석



예 4.5



선거유세의 효과를 파악하는 방법으로 유권자 100명을 확률표본으로 추출하여 유세 이전의 지지 (1=여당, 0=야당)와 유세 이후의 지지 (1=여당, 0=야당)를 물어볼 수 있다. 수집되는 원자료는 <표 4.17>의 형태를 취하게 될 것이다. 이 자료를 바탕으로 유세 이전과 이후 유권자의 지지변화가 있었는지 알아보자.

원자료

유권자	이전	이후
1	1	0
2	0	0
3	1	1
⋮	⋮	⋮
100	0	1

분할표

		유세 이후		
		여당	야당	계
유세 이전	여당	63	4	67
	야당	21	12	33
	계	84	16	100



대응자료 및 사례-대조군 검정 (McNemar test)

4강 범주형 자료의 분석

원자료

유권자	이전	이후
1	1	0
2	0	0
3	1	1
⋮	⋮	⋮
100	0	1

분할표

		유세 이후		
		여당	야당	계
유세 이전	여당	63	4	67
	야당	21	12	33
	계	84	16	100

가설

H_0 : 유세 이전과 이후에 유권자의 지지지는 변함없다

H_1 : H_0 가 아니다.

검정통계량

$$\chi_0^2 = \frac{(b - c)^2}{b + c} \sim \chi^2(1) \text{ 이라는 이론에 의하여}$$

검정결과

$$\chi_0^2 = 11.56 > \chi^2(1) = 3.84 \quad \chi^2 = 11.56 \quad p\text{-값} : 0.0007$$

이므로 H_0 를 기각

➔ 유세를 함으로써 여당 지지율이 달라졌다고 결론짓는다.

독립성 검정

대응자료 χ^2 검정(McNemar test)

```
선거유세효과<-  
matrix(c(63,4,21,12),2,2,byrow=T,dimnames=list(유세이  
전=c("유세전 여당","유세전 야당"),유세이후=c("유세후 여당","  
유세후 야당")))
```

선거유세효과

McNemar

mcnemar

```
R Console  
  
> #===== 4.2.4 대응자료 및 사례-대조군 검정  
> # [예 4.5]  
> # McNemar 검정  
> 선거유세효과<-matrix(c(63,4,21,12),2,2,byrow=T,dimnames=list(유세이전=  
c("유세전 여당","유세전 야당"),유세이후=c("유세후 여당","유세후 야당")))  
> 선거유세효과  
      유세이후  
유세이전 유세후 여당 유세후 야당  
유세전 여당      63      4  
유세전 야당      21     12
```

독립성 검정

```
# McNemar 검정
mcnemar.test(선거유세효과)

# 정확 McNemar
library(exact2x2)
exact2x2(선거유세효과, paired=T)
```

```
R Console

> # McNemar 검정 (연속성 수정)
> mcnemar.test(선거유세효과)

McNemar's Chi-squared test with continuity correction

data: 선거유세효과
McNemar's chi-squared = 10.24, df = 1, p-value = 0.001374

> # 정확 McNemar 검정
> library(exact2x2)
필요한 패키지를 로딩중입니다: exactci
> exact2x2(선거유세효과, paired=T)

Exact McNemar test (with central confidence intervals)

data: 선거유세효과
b = 4, c = 21, p-value = 0.0009105
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.04753664 0.56452522
sample estimates:
odds ratio
 0.1904762
```



세 가지 McNemar 검정을 통하여 유세 이전과 이후에 유권자의
지지에 변화가 일어났음을 알 수 있다.

I 사례 대조군 연구 검정(case-control test)

4강 범주형 자료의 분석

예



예 4.6

1972년에 편도적출과 호지킨병의 관계를 규명하는 연구보고가 있었다. 미국 국립암 연구소에서 치료를 받은 174명의 호지킨병 환자들을 사례군으로 하고 환자들의 형제자매 472명을 대조군으로 구성하였다. 그리고 이 자료를 기초로 한 추후의 연구 보고에서는 사례와 대조를 1 : 1 대응으로 하기 위하여 사례 1명에 대응되는 대조는 나이 차이가 5년 이내이고 같은 성별을 가진 형제자매 중 사례와 나이가 가장 가까운 사람으로 정하였다. 이렇게 하여 85쌍의 사례-대조군 자료가 <표 4.19>와 같이 관찰되었다. 85쌍의 사례-대조군 자료를 기초로 하여 호지킨병과 편도적출이 독립적인가를 검정하는 문제를 생각해 보자.



대조군

분할표

	편도 적출	편도 비적출	계
편도 적출	26	15	41
편도 비적출	7	37	44
계	33	52	85

사례군

호지킨병 환자

가설

$$H_0: P(T | \text{사례}) = P(T | \text{대조})$$

호지킨 림프종과 편도적출은 서로 독립이다

사례대조군 연구검정

잘못 적용된 통계적 절차: 카이제곱검정

호지킨병_잘못된자료

```
=matrix(c(41,44,33,52),nrow=2,byrow=T)
```

dimr

대조군

호지킨

분할표

잘못된

분할표

```
R Console
> 호지킨병_잘못된자료=matrix(c(41,44,33,52),nrow=2,byrow=T)
> dimnames(호지킨병_잘못된자료)=list(그룹=c("사례군","대조군"),편도적출여부=c("유","무"))
> 호지킨병_잘못된자료
```

그룹	편도적출여부	편도적출 유	편도적출 무
사례군		41	44
대조군		33	52

```
> 분할표_호지킨병_잘못된자료=addmargins(호지킨병_잘못된자료)
> 분할표_호지킨병_잘못된자료
```

그룹	편도적출여부	편도적출 유	편도적출 무	Sum
사례군		41	44	85
대조군		33	52	85
Sum		74	96	170

사례대조군 연구검정

카이검정

chisq

카이

chisq

R Console

```
> # 카이검정  
> chisq.test(호지킨병_잘못된자료, correct=F)
```

Pearson's Chi-squared test

```
data: 호지킨병_잘못된자료  
X-squared = 1.5315, df = 1, p-value = 0.2159
```

```
> # 카이검정 (연속성 수정)  
> chisq.test(호지킨병_잘못된자료)
```

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

```
data: 호지킨병_잘못된자료  
X-squared = 1.1726, df = 1, p-value = 0.2789
```

사례대조군 연구검정

올바른 통계적 절차: 카이제곱검정

호지킨병=ma

dimnames(
적출 무"),대조

호지킨병

분할표_호지킨

분할표_호지킨

R Console

```
> # 올바른 통계적 절차: 카이제곱검정  
> 호지킨병=matrix(c(26,15,7,37),nrow=2,byrow=T)  
> dimnames(호지킨병)=list(사례군=c("편도적출 유","편도적출 무"),호지킨병=c("유","무"))  
> 호지킨병
```

사례군	대조군	편도적출 유	편도적출 무
편도적출 유		26	15
편도적출 무		7	37

```
> 분할표_호지킨병=addmargins(호지킨병)  
> 분할표_호지킨병
```

사례군	대조군	편도적출 유	편도적출 무	Sum
편도적출 유		26	15	41
편도적출 무		7	37	44
Sum		33	52	85

사례대조군 연구검정

```
# McNemar 검정
mcnemar.test(호지킨병,correct=F)
# McNemar 검정(연속성 수정)
mcnemar.test(호지킨병,correct=T)
```

```
R Console
>
> # McNemar 검정
> mcnemar.test(호지킨병,correct=F)

McNemar's Chi-squared test

data: 호지킨병
McNemar's chi-squared = 2.9091, df = 1, p-value = 0.08808

>
> # McNemar 검정(연속성 수정)
> mcnemar.test(호지킨병,correct=T)

McNemar's Chi-squared test with continuity correction

data: 호지킨병
McNemar's chi-squared = 2.2273, df = 1, p-value = 0.1356
```

사례대조군 연구검정

정확 McNemar 검정

library(exact2x2)

exact2x2(호지킨병,paired=T)

```
> # 정확 McNemar 검정  
> library(exact2x2)  
> exact2x2(호지킨병,paired=T)
```

```
Exact McNemar test (with central confidence intervals)
```

```
data: 호지킨병
```

```
b = 15, c = 7, p-value = 0.1338
```

```
alternative hypothesis: true odds ratio is not equal to 1
```

```
95 percent confidence interval:
```

```
0.8224084 6.2125863
```

```
sample estimates:
```

```
odds ratio
```

```
2.142857
```

I 사례 대조군 연구 검정(case-control test)

4강 범주형 자료의 분석

대조군

예 분할표

사례군

	편도 적출	편도 비적출	계
편도 적출	26	15	41
편도 비적출	7	37	44
계	33	52	85

가설

$$H_0: P(T | \text{사례}) = P(T | \text{대조})$$

호지킨 림프종과 편도적출은 서로 독립이다

H_1 : 편도적출 여부에 따라 호지킨 림프종 발생이 달라진다

검정통계량

$$\chi_0^2 = \frac{(b - c)^2}{b + c} = 2.91 < \chi_\alpha^2(1) = 3.84$$

검정결과

H_0 기각할 수 없어 편도적출 여부가 림프종 발생에 영향을 준다는 충분한 증거가 없다.

로짓 분석(Logit Analysis)

4강 범주형 자료의 분석

- 반응변수가 범주형이고 이에 대한 설명변수가 범주형과 이산형이 혼합된 경우 관련성 여부를 규명하기 위한 범주형 자료의 회귀분석 방법



	1	0
1	a	b
0	c	d

오즈(odds)

$$odds = \frac{p}{1-p} \quad (0 < p < 1) = \frac{ad}{bc}$$

로짓(logit)

$$logit = \ln\left(\frac{p}{1-p}\right) \quad (0 < p < 1)$$

$\rightarrow \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$ 식으로 자료를 적합

I 로짓 분석(Logit Analysis)

4강 범주형 자료의 분석

예

바르샤바의 3918명의 소녀들을 대상으로 조사한 초경자료를 가지고 범주형 자료의 회귀분석인 로짓 분석을 실시, 3918명의 소녀에게 현재 연령과 월경을 하고 있는지의 여부를 묻은 자료를 원자료로 한다.

평균연령	초경 경험자	군의 총수	경험률	Logit(p)
9.21	0	376	0.000	-
10.21	0	200	0.000	-
10.58	0	93	0.000	-
10.83	2	120	0.017	-4.076
⋮	⋮	⋮	⋮	⋮
17.58	1049	1049	1.000	-

로짓모형

$$\text{logit}(p) = -21.2258 + 1.6319 \times X$$

→ X가 1단위 증가할 때 로그오즈는 1.6319 변한다.

→ 나이가 한살 증가할 때 초경 경험자의 오즈는 5.114[=exp(1.6319)] 배 많다고 할 수 있다.

예

바르샤바의 3918명의 소녀들을 대상으로 조사한
초경자료를 가지고 범주형 자료의 회귀분석인
로짓 분석을 실시, 3918명의 소녀에게 현재
연령과 월경을 하고 있는지의 여부를 묻은
자료를 원자료로 한다.

로짓모형

$$\text{logit}(p) = -21.2258 + 1.6319 \times X$$

➔ X가 1단위 증가할 때 로그오즈는 1.6319 변한다.

➔ 나이가 한살 증가할 때 초경 경험자의 오즈는
5.114[=exp(1.6319)] 배 많다고 할 수 있다.

유효 중앙값 (ED₅₀)

50%가 초경을 경험한 소녀들의 나이

적용 예

100명의 고열 환자에서 해열제를 투여했을때
환자 50%의 열이 떨어지는 효과를 보여주는
약의 용량 ED₅₀(Effective Dose of 50%)가 된다.

예 4.8

모건(B. J. T. Morgan)은 폴
초경자료(g:\\WORK\\바르샤바
석(로짓 분석)을 소개하였다. <
있는지의 여부를 묻은 관찰조사
떠한 관계가 있는지 알아보자.

표 4.23 바르샤바 소녀의 초경연령

평균 연령	초경경험자	군의 총수	초경경험률	Logit(p)
9.21	0	376	0.000	—
10.21	0	200	0.000	—
10.58	0	93	0.000	—
10.83	2	120	0.017	-4.076
11.08	2	90	0.022	-3.784
11.33	5	88	0.057	-2.809
11.58	10	105	0.095	-2.251
11.83	17	111	0.153	-1.710
12.08	16	100	0.160	-1.658
12.33	29	93	0.312	-0.792
12.58	39	100	0.390	-0.447
12.83	51	108	0.472	-0.111
13.08	47	99	0.475	-0.101
13.33	67	106	0.632	-0.541
13.58	81	105	0.771	1.216
13.83	88	117	0.752	1.110
14.08	79	98	0.806	1.425
14.33	90	97	0.928	2.554
14.58	113	120	0.942	2.781
14.83	95	102	0.931	2.608
15.08	117	122	0.959	3.153
15.33	107	111	0.964	3.283
15.58	92	94	0.979	3.829
15.83	112	114	0.982	4.025
17.58	1,049	1,049	1.000	—

로짓분석

- > 확률=초경경험자/군총수
- > # 나이별 초경비율 그리기
- > plot(평균연령, 확률)

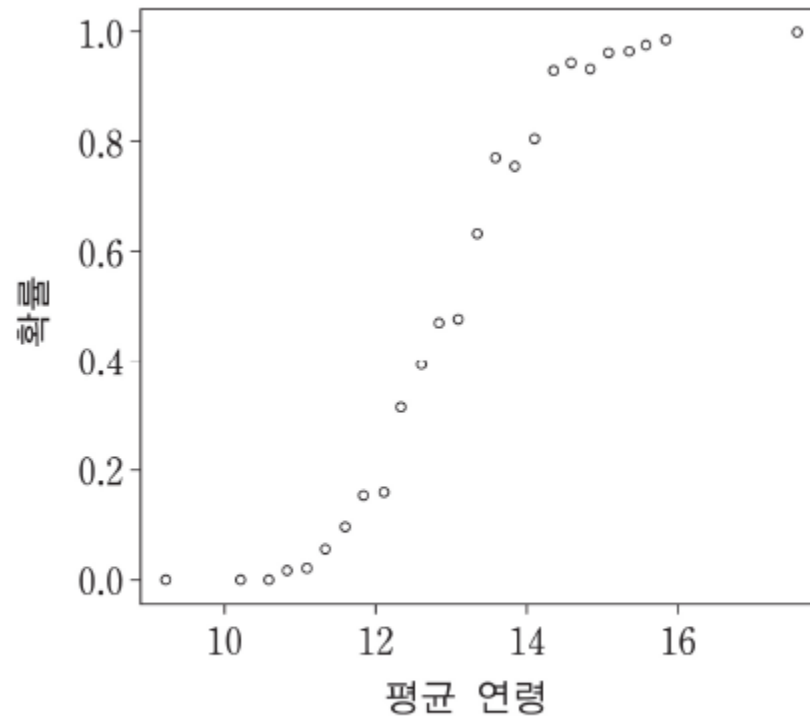


그림 4.6 나이별 초경경험률

로짓분석

```
> # 로짓분석
> 로짓 <- glm(확률 ~ 평균연령, data=초경연령자료, weight=군총수,
  family="binomial")
> summary(로짓)
```

```
Call:
glm(formula=확률 ~ 평균연령, family="binomial", data=초경연령자료,
  weights=군총수)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.0363	-0.9953	-0.4900	0.7780	1.3675

Coefficients:

$$\text{logit}(p) = -21.2264 + 1.6320x$$

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-21.22639	0.77068	-27.54	<2e-16 ***
평균연령	1.63197	0.05895	27.68	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family t

Null deviance: 3693.884 on 24 degrees
Residual deviance: 26.703 on 23 degrees
AIC: 114.76

```
> ## 오즈비
> exp(coef(로짓)) [-1]
평균연령
5.113931e+00
```

오즈비

```
## 오즈비  
exp(coef(로짓))
```

```
> ## 오즈비  
> exp(coef(로짓)) [-1]  
평균연령  
5.113931e+00
```


예

비타민 E의 용량에 따른 임신한 쥐의 수



용량 (mg)	쥐의 수	임신한 쥐의 수
3.75	5	0
5.0	10	2
6.25	10	4
7.5	10	8
10.0	11	10
15.0	11	11

로짓 모형식

$$\ln\left(\frac{p_x}{1-p_x}\right) = -11.88 + 14.707 \log_{10}x$$

➡ 비타민 E 용량이 10배 증가함에 따라 로그 오즈가 14.707배 증가함을 알 수 있다.

예

비타민 E의 용량에 따른 임신한 쥐의 수

용량 (mg)	쥐의 수	임신한 쥐의 수
3.75	5	0
5.0	10	2
6.25	10	4
7.5	10	8
10.0	11	10
15.0	11	11

로짓 모형식

$$\ln\left(\frac{p_x}{1-p_x}\right) = -11.88 + 14.707 \log_{10}x$$

$$\rightarrow \text{ED}_{50} : -11.88 + 14.707X = 0$$

$$X = 0.808, \quad 10^{0.808} = 6.427$$

ED₅₀ = 6.45mg 이다.

로짓분석



예 4.9

힐리(M. J. R Healy)는 비타민 E의 용량에 따른 임신한 쥐의 숫자를 <표 4.24>와 같이 발표하고 있다. 비타민 E와 임신에 대한 관계를 알아보자.

표 4.24 비타민 E 용량에 따른 임신한 쥐의 숫자

용량(mg)	쥐의 숫자	임신한 숫자
3.75	5	0
5.0	10	2
6.25	10	4
7.5	10	8
10.0	11	10
15.0	11	11

로짓분석

예) 비타민 E

용량=c(3.75, 5, 6.25, 7.5, 10, 15)

쥐숫자=c(5, 10, 10, 10, 11, 11)

임신숫자=c(0, 2, 4, 8, 10, 11)

확률2=임신숫자/쥐숫자

로짓2 <- glm

summary(로

R Console

```
> 용량=c(3.75, 5, 6.25, 7.5, 10, 15)
> 쥐숫자=c(5, 10, 10, 10, 11, 11)
> 임신숫자=c(0, 2, 4, 8, 10, 11)
> 확률2=임신숫자/쥐숫자
> 로짓2 <- glm(확률2~log10(용량), family="binomial")
```

경고메시지:

```
In eval(expr, envir, enclos) :
  binomial glm에서 number of success가 정수가 아닙니다
> summary(로짓2)
```

Call:

```
glm(formula = 확률2 ~ log10(용량), family = "binomial")
```

Deviance Residuals:

1	2	3	4	5	6
-0.23099	0.11888	-0.09852	0.15618	-0.16518	0.08464

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-12.42	10.39	-1.195	0.232
log10(용량)	15.35	12.76	1.203	0.229

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 4.29706 on 5 degrees of freedom
Residual deviance: 0.13603 on 4 degrees of freedom
AIC: 6.3152

Number of Fisher Scoring iterations: 6

로짓분석

```
> 용량=c(3.75,5,6.25,7.5,10,15)  
> 쥐숫자=c(5,10,10,10,11,11)  
> 임신숫자=c(0,2,4,8,10,11)
```

```
> 확률2=임신숫자/쥐숫자  
> # 용량의 상용로그별 임신비율 그리기  
> plot(log10(용량),확률2,ylab="확률")
```

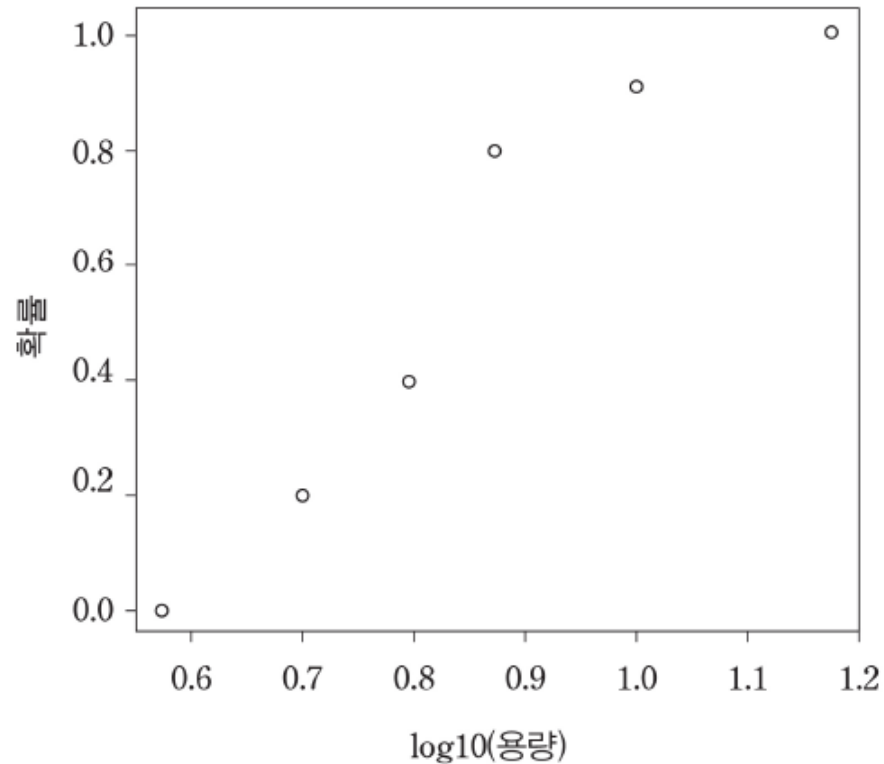


그림 4.7 용량의 상용로그별 임신비율

로짓분석

```
> 로짓2 <- glm(확률2 ~ log10(용량), weight=쥐숫자, family="binomial")  
> summary(로짓2)
```

Call:

```
glm(formula=확률2 ~ log10(용량), family="binomial", weights=쥐숫자)
```

Deviance Residuals:

1	2	3	4	5	6
-0.5613	0.2695	-0.3534	0.5266	-0.4628	0.3128

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-11.884	3.301	-3.601	0.000317 ***
log10(용량)	14.707	4.022	3.656	0.000256 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family is not estimated)

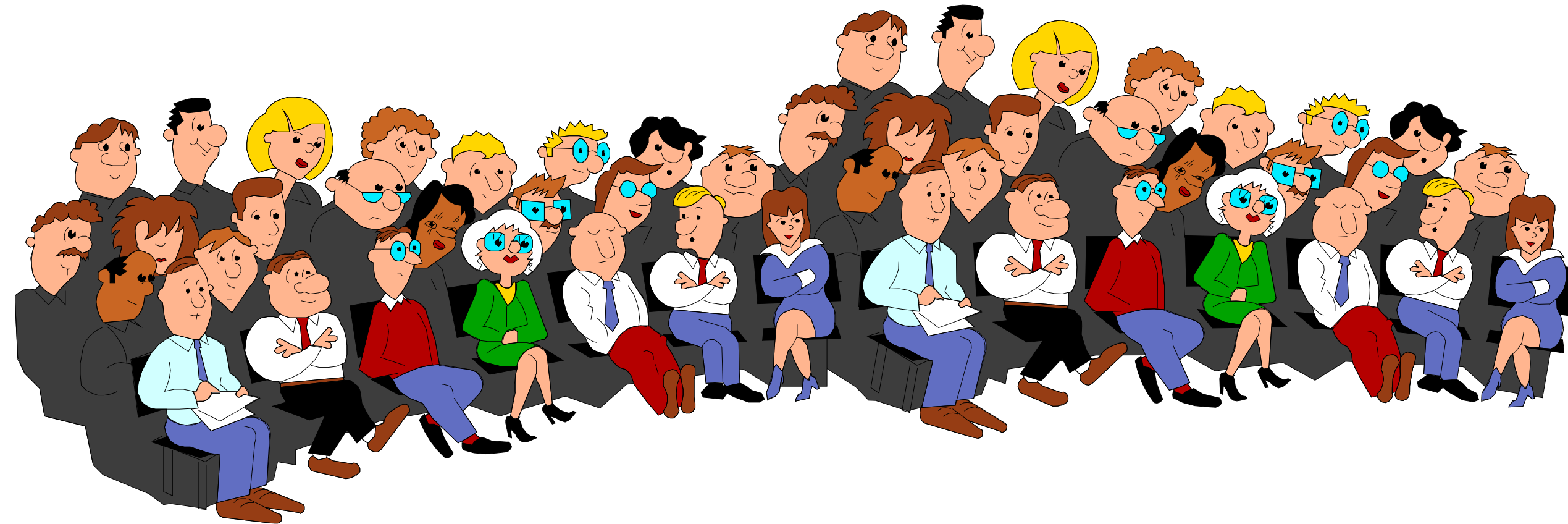
Null deviance: 35.849 on 5 degrees of freedom
Residual deviance: 1.102 on 4 degrees of freedom
AIC: 14.564

Number of Fisher Scoring iterations: 5

```
> ## 오즈비  
> exp(coef(로짓2)) [-1]  
log10(용량)  
2.439749e+06
```

$$\ln\left(\frac{p_x}{1-p_x}\right) = -11.884 + 14.707\log_{10}x$$

Q & A



다음시간에는



▶ 4강 범주형자료 분석

5강 공분산 분석

6강 반복측정 자료분석