교차표 Cross Table

두 범주형 변수의 정보를 요약한 표, 행 범주와 열 범주, 셀은 관측치

■교차표

- π_{ij}: (X, Y) 결합밀도함수
- π_{i+}: (X) 주변밀도함수

X \ Y	1	2	...	C	Total
1	π_{11}	π_{12}	...	π_{1c}	π_{1+}
2	π_{21}	π_{22}	...	π_{2c}	Π_{2+}
...	...	...	...	...	...
R	π_{r1}	π_{r2}	...	π_{rc}	Π_{r+}
Total	π_{+1}	π_{+2}	...	π_{+c}	π_{++}

■Homogeneity (동질성)

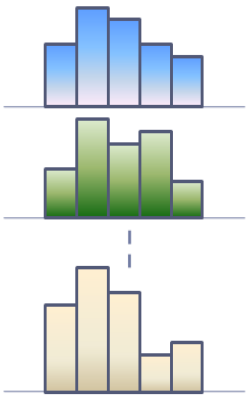
- 각 행에 대해 열의 분포가 동일한가?

$$H_0 : \pi_{ij} = \pi_{kj} \ for \ j = 1,2,\ldots,c \text{ and } k \neq i$$

■Independence (독립성)

- (X, Y)는 서로 독립인가? $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$

■검정통계량 $\dfrac{\sum_{i=1}^{k}(O_i - E_i)^2}{E_i} \sim \chi^2(k-1)$



예제 데이터

예제

학위에 따른 진출 분야의 차이? 해석은 행 퍼센트를 이용한다.

Major	Oil	Chemical	Electrical	Computer
Business	30	15	15	40
Engineering	30	30	20	20

```
ds=c(30,30,15,30,15,20,40,20)
rn=c("B", "E")
cn=c("O", "C", "E", "Com")
CT=matrix(ds, nrow=2, ncol=4, dimnames=list(rn,cn))
```


Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

```
> CT
   O  C  E Com
B 30 15 15  40
E 30 30 20  20
```

```
> prop.table(CT, 1)    # 2=열%, 1=행%, none=셀%
     O    C    E Com
B 0.3 0.15 0.15 0.4
E 0.3 0.30 0.20 0.2
```

```
prop.table(CT, 1)    # 2=열%, 1=행%, none=셀%
chisq.test(CT)
```

```
data:  CT
X-squared = 12.381, df = 3, p-value = 0.006186
```

o 유의확률이 0.62%이므로 귀무가설 (전공과 진출분야 연관관계 없음) 기각됨

o 비즈니스 전공은 컴퓨터 분야, 공학 전공은 오일, 화학 분야 진출이 높다 (행 퍼센트)

교차분석

Data on the marital status of men and women ages 20 to 29 were obtained as part of a national survey. The results from a sample of 350 men and 400 women follow. These data are representative of results published in the *U.S. Current Population Report* (*The Statistical Abstract of the United States*, 1999).

	Marital Status		
Gender	**Never Married**	**Married**	**Divorced**
Men	234	106	10
Women	216	168	16

a. Use $\alpha = .01$ and test for independence between marital status and gender. What is your conclusion?

b. Summarize the percent in each marital status category for men and for women.

CASE 15.1 Predicting the Outcomes of Basketball, Baseball, Football, and Hockey Games from Intermediate Results*

© Dave Sanford/ Allsport/Getty Images

Some basketball fans generally believe that it doesn't pay to watch an entire game because the outcome is determined in the last few minutes (some say the last 2 minutes) of the game. Is this really true, and, if so, is basketball different in this respect from other professional sports played in North America? For example, is it true that the team that leads a baseball game after seven innings almost always wins the game? To address these questions, three researchers tracked basketball, baseball, football, and hockey games. The results (whether the early-game leader and whether the late-game leader won) of games during the 1990 season (for baseball and football) and during the 1990–1991 season (for basketball and hockey) were recorded. Early-game leaders are defined as the teams that are ahead after one quarter of basketball and football, one period of hockey, or three innings of baseball. Late-game leaders are defined as the teams that are ahead after three quarters of basketball and football, two periods of hockey, or seven innings of baseball.

The data were recorded in the following way:

Column 1: Results of games where
2 = early-game leader wins and
1 = early-game leader loses

Column 2: Early-game leader game where 1 = basketball, 2 = baseball, 3 = football, and 4 = hockey games

Column 3: Results of games where
2 = late-game leader wins and
1 = late-game leader loses

Column 4: Late-game leader game where 1 = basketball,
2 = baseball, 3 = football,
and 4 = hockey games

Can we infer from these data that all four professional sports experience the same proportion of early-game leaders winning the game?

Can we infer from these data that all four professional sports experience the same proportion of late-game leaders winning the game?

```
table.sp=table(ds.sp$Game..Late,ds.sp$Late) #교차표작성
```

교차표 그래프 표현 (Simpson's Paradox)

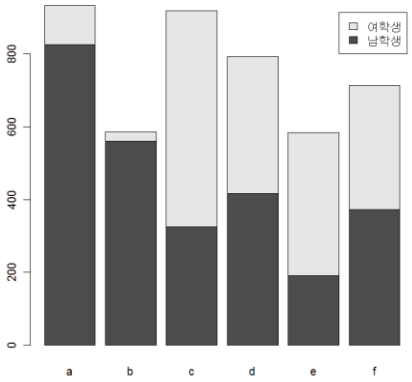
(예제 데이터 중심) Google : Berkeley's Gender discrimination

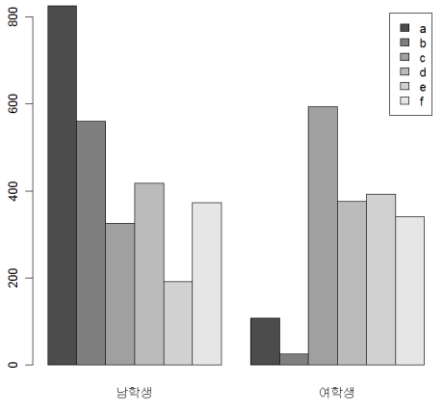
여학생들이 주장했다. 남학생 지원자 8,442 명 중 44% (2,691 명)가 합격, 4,321 명 중 35% (1,835)가 합격 => 이는 성 차별이라고 주장

(학과별로 작성된 표)

```
> data=c(825,108,560,25,325,593,417,375,191,393,373,341)
> cn=c("a","b","c","d","e","f")
> rn=c("남학생","여학생")
> table=matrix(data,nrow=2,ncol=6,dimnames=list(rn,cn))
> table
         a   b   c   d   e   f
남학생 825 560 325 417 191 373
여학생 108  25 593 375 393 341
> chisq.test(table)

        Pearson's Chi-squared test

data:  table
X-squared = 1068.372, df = 5, p-value < 2.2e-16

> prop.table(table,1)
                a          b         c         d          e         f
남학생 0.30657748 0.20810108 0.1207729 0.1549610 0.07097733 0.1386102
여학생 0.05885559 0.01362398 0.3231608 0.2043597 0.21416894 0.1858311
> prop.table(table,2)
               a          b         c         d         e        f
남학생 0.8842444 0.95726496 0.3540305 0.5265152 0.3270548 0.522409
여학생 0.1157556 0.04273504 0.6459695 0.4734848 0.6729452 0.477591
```

```
barplot(table, legend=rownames(table))
barplot(t(table),beside=T,legend=colnames(table))
```





Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	272	6%	341	**7%**

빈도 분할

교차표 셀 빈도

$f_{ij} = \mu + R_i + C_j + e_{ij}, \quad e_{ij} \sim N(0, \sigma^2)$ 가법모형

추정치

- o $\hat{\mu} = \bar{y}_{..}, \quad \hat{R}_i = \bar{y}_{i.} - \bar{y}_{..}, \quad \hat{C}_j = \bar{y}_{.j} - \bar{y}_{..}$
- o 위의 추정치는 이상값에 양향을 받으므로 적절한 추정치가 아님

Median Polish

가법모형 Robust 추정방법

단계 1 : 행 효과 (R_i)

- o $\tilde{y}_{i.}$: i 행의 관측치 중앙값을 행 효과의 추정치로 사용
- o $y_{ij}^{(1)} = y_{ij} - \tilde{y}_{i.}$
- o $y_{ij}^{(1)}$ 에는 $\mu + C_j$ 만 남아 있음

단계 2 : 열 효과 (C_j)

- o $\tilde{y}_{.j}^{(1)}$: j 열의 관측치 중앙값을 열 효과의 추정치로 사용
- o $y_{ij}^{(2)} = y_{ij}^{(1)} - \tilde{y}_{.j}^{(1)}$

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

모형 적합성 검토

o 가법모형 vs. 승법모형 $f_{ij} = \mu * R_i * C_j + e'_{ij}$

o 두 모형의 잔차 관계 : $e_{ij} = \dfrac{\text{행효과}(R_i) * \text{열효과}(C_j)}{\text{공통효과}(\mu)} + e'_{ij}$

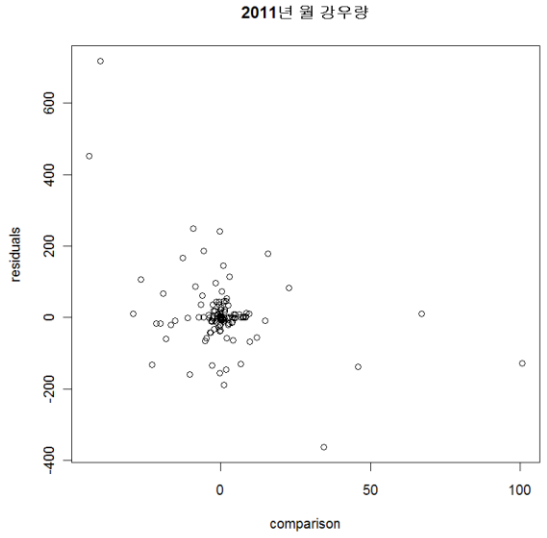
o 비교 comparison= $\dfrac{\text{행효과}(R_i) * \text{열효과}(C_j)}{\text{공통효과}(\mu)}$

o 승법모형과 비교의 산점도가 패턴이 있다면 승법모형은 적합하지 않다. 왜냐하면, 만약 승법모형이 적합하다면 오차항만 남아 있어야 하는데…

o 산점도의 기울기가 양의 기울기 (+1)를 가지면 승법모형이 적합 => 승법모형이 적합한 경우에는 데이터를 로그 변환하여 승법모형을 적합시키면 된다.

2011 대도시 강우량 데이터

```
rainfall=read.csv("rainfall.csv")
data=as.matrix(data.frame(rainfall[,2:11]))

cn=c("서울","인천","강릉","대전","전주","광주","제주","부산","울산","대구
rn=c("1월","2월","3월","4월","5월","6월","7월","8월","9월","10월","11월
table.rainfall=matrix(data,nrow=12,ncol=10,dimnames=list(rn,cn))

ct.fit<-medpolish(table.rainfall)
attach(ct.fit)
comparison=matrix(row,ncol=1)%*%matrix(col,nrow=1)/overall
plot(residuals~comparison, main="2011년 월 강우량")
```



(일정한 패턴을 보이지 않음)

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

```
Overall: 80.01875

Row Effects:
      1월       2월       3월       4월       5월       6월       7월
-73.59375 -21.06875 -56.86875  21.06875  40.81875 227.88125 342.61875
      8월       9월      10월      11월      12월
156.29375 -31.19375 -37.81875  28.40625 -65.06875

Column Effects:
     서울      인천      강릉      대전      전주      광주      제주
 -9.37500 -10.27500  23.53750  -2.93750  -0.10000   0.40000   8.03750
     부산      울산      대구
  3.44375  -5.34375  -0.96875
```

o 행 효과 (월별) 7 월에 비가 많이 오고 1 월 강우량이 가장 낮음

o 열 효과 (도시) 강릉에 비가 많이 오고 서울이 가장 낮음

o (잔차) 대전 지역 7,8 월에 비가 많이 옴, 제주는 7 월 비가 적게 옴

```
Residuals:
        서울     인천      강릉      대전      전주      광주      제주
1월   11.850   10.850  -16.9625    0.5125   -1.8250    0.5750    1.1375
2월  -20.475  -16.175   61.5125  -11.2125   -0.4500    0.4500  -31.8875
3월    0.825    1.625  -20.1875   -1.2125    0.4500   -0.4500    1.2125
4월   18.387   36.788    8.3750  -27.1500  -20.1875    1.5125  -57.8250
5월  -58.063  -66.362  -56.3750   44.1000    2.0625   21.6625  -62.7750
6월  105.975    9.975   11.5625   86.6375 -156.0000 -188.3000   82.6625
7월  717.737  451.838 -127.9750  167.6000   44.4625 -145.5375 -362.3750
8월  -60.137  -17.637 -138.6500  186.9250  241.5875  145.7875  178.7500
9월  -13.850  -12.150  248.5375   45.8125   22.4750  -35.7250   -9.0625
10월  -0.825   -1.625   -0.9375   -2.2625   -7.5000  -22.1000   -1.1375
11월 -42.850  -42.450    0.9375   -2.2875   10.5750   27.9750  114.8375
12월   1.525    2.425   66.5125   -0.5125   -4.3500   -2.0500   36.9125
```

(1990~2011 년) 소비지출

가계수지항목별	2004			2005	
	전가구	근로자가구	근로자외가구	전가구	근로자가
⊟ 소비지출 (원)	1,963,316	2,005,758	1,906,569	2,035,256	2,0
⊞ 식료품 (원)	532,452	546,711	513,309	539,260	5
⊞ 주거 (원)	64,627	63,675	65,902	69,498	
⊞ 광열·수도 (원)	97,477	94,888	100,951	102,561	1
⊞ 가구집기가사용	77,928	81,004	73,845	84,950	
⊞ 의류및신발 (원	101,523	104,133	98,047	106,698	1
⊞ 보건의료 (원)	93,705	91,552	96,609	102,446	1
⊞ 교육 (원)	219,833	220,422	218,983	229,747	2
⊞ 교양오락 (원)	93,918	98,053	88,411	98,275	1
⊞ 교통·통신 (원)	339,047	356,895	315,078	356,166	3
⊞ 기타소비지출	342,806	348,427	335,437	345,656	3

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

대응분석 correspondence analysis

개념

분할표로 나타내어지는 자료의 행과 열범주를 저차원 공간상(2 차원)의 점들로 동시에 나타내어, 그들의 관계를 탐구하려는 탐색적 자료 분석 기법

- 단순 대응 분석 : 행과 열범주를 나타내는 변수가 둘뿐인 이원분할표의 대응분석
- 다중 대응 분석 : 범주를 나타내는 변수가 두 개인 이변량 대응분석을 포함하여 변수가 셋 이상인 다원분할표의 대응분석

워크시트 상의 이원표 검정과의 비교

- 일반적으로 2 차원 분할표의 경우는 독립성, 동질설 검정의 경우 카이제곱검정을 실시하고 행(열) 퍼센트에 의해 그 의미를 충분히 파악할 수 있음
- 그러나 r×c 가 복잡해지면 카이제곱검정의 신뢰도가 떨어지게 되므로 대응일치 분석이 효과적 일 수 있음
- 카이제곱검정의 경우 결과가 P-value=0.000 으로서 매우 종속적인 관계가 있음을 보여주지만 의미해석이 simple 하지 못할 수 있음
- 이에 대응분석은 대칭도를 통해 맵핑 결과를 명확하고 단순하게 알 수 있음

기원

대응분석이 다변량 자료분석에서 널리 알려진 때는 1980 년대이다.
대응분석의 수리적인 기원은 1930 년대 Hirshfeld 의『상관관계와 분할표의 연관성』이라는 논문이고, 1940-1950 년대에 몇몇 학자들에 의해서 더욱 발전되었다.
대응분석의 기하적인 면은 1960 년대 프랑스에서 Jean-Paul Benzecri 에 의해서 발전되었다.
일본에서는 1950 년대 Chikio Hayashi 에 의해서 수량화 제 3 방법으로 개발되어 발전되었고, 프랑스에서는 1960 년대 Jean-Paul Benzecri 가 이끄는 자료분석 모임이 다양한 분야로부터 수집된 자료를 분석하는데 대응분석 기법을 응용하고 발전시켰다.

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

기호

- n_{ij}: (i, j) 셀 관측빈도
- n_{i+}: (i)번째 행의 관측빈도 합
- n_{+j}: (j)번째 열의 관측빈도 합

방법

- (i, j) 셀의 빈도 nij (≥ 0)의 i 번째 행 (ni1, …, nic)은 총빈도가 ni+ = ni1+ … + nic 이고 C 개 범주를 갖는 다항 분포
- Multinomial 분포의 대응 확률은 상대빈도 fij = nij / n: 이것을 행 프로파일 (row profile)이라 정의
- 각 행의 상대적 빈도 fi+ = fi1 + … + fic 를 선형계수로 (주성분 분석과 유사) 하여 좌표 계산
- ri =(fi1 / fi+ , fi2 / fi+ , …, fic / fi+)는 C 차원 가중 Euclid 공간의 좌표
- 가중(weighted) Euclid 공간이란 두 개의 좌표 r_a, r_b 사이의 거리가 다음과 같이 정의

$$d(r_a, r_b) = \sqrt{\sum_j \left(\frac{f_{aj}}{f_{a+}} - \frac{f_{bj}}{f_{b+}}\right)^2 / f_{+j}}$$

$f_{+j} = f_{1j} + \ldots + f_{rj}$ ==>

- 같은 방식으로 열 프로파일의 좌표 및 개체 거리 계산
- 행, 열 프로파일을 각각 2 차원 공간에 표현하거나 동시에 표현

예제

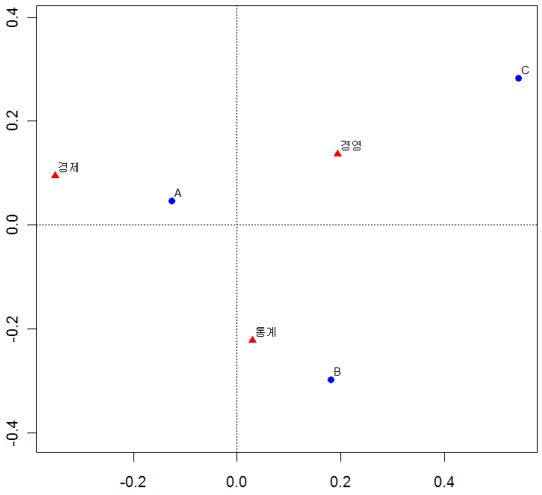
학과 영어 등급	경영	경제	통계
A	78	65	68
B	22	8	30
C	20	2	7

```
> chisq.test(table) #카이제곱검정

        Pearson's Chi-squared test

data:  table
X-squared = 21.946, df = 4, p-value = 0.0002054
```


Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

```
> prop.table(table, 1) # 행퍼센트
        경영       경제       통계
A 0.3696682 0.30805687 0.3222749
B 0.3666667 0.13333333 0.5000000
C 0.6896552 0.06896552 0.2413793
> prop.table(table, 2) # 열퍼센트
        경영       경제       통계
A 0.6500000 0.86666667 0.64761905
B 0.1833333 0.10666667 0.28571429
C 0.1666667 0.02666667 0.06666667
```

```
fit.ds=ca(table)
print(fit.ds) # basic results
summary(fit.ds) # extended results
plot(fit.ds) # symmetric map
```



15 대 대통령선거 득표율

🖫대통령선거.csv 데이터는 제 15 대 대통령선거에서 입후보자들의 지역별 득표이다.

(1) 카이제곱검정을 실시하여 후보자간 지역별 득표의 차이가 있는지 검정하고 행퍼센트, 열퍼센트 이용하여 해석하시오.

(2) median polish 분석을 실시하고 해석하시오.

(3) 대응분석을 실시하여 대응 산점도를 그리고 해석하시오.