## 2.1. Two-way Contingency Table (이차원 분할표) 맛보기

### 2.1.1. 예제

하나의 범주형 자료에 정리 방법으로 사용되는 것이 빈도표(혹은 다양한 차트)를 작성하는 것이다. 예를 들어 정보통계학과 학생 120 에 대한 출신지 조사 결과 다음을 얻었다.

| 출신지 | 대전 | 충남 | 기타 지역 |
|---|---|---|---|
| 빈도(비율) | 40(33.3%) | 30(25%) | 50(41.7%) |

**동일** 학생 120 명들에 대해 OO 후보 지지여부를 물어 아래 결과를 얻었다.

| OO 후보 | 지지 | 반대 |
|---|---|---|
| 빈도(비율) | 80(66.7%) | 40(33.3%) |

두 범주형 변수간의 연관성(association)을 알아보기 위해 한 범주형 변수에 대한 빈도표는 열로, 다른 범주형 변수에 대한 빈도표는 행으로 하여 교차표(cross-tabulation)를 작성하게 되는데 이를 이차원 분할표(two-way contingency table)이라 한다. 일반적으로 영향을 미친다고 생각되는 변수(~따라서)를 행으로, 영향을 받는다고 생각되는 것을 변수(~차이가 있다)를 열로 하여 교차표를 작성하면 된다.

위의 예제에서 출신지별 OO 후보 지지 여부 차이가 있는지 알아보기 위하여 분할표를 작성하여 보자. 위의 두 표만으로는 분할표를 작성할 수 없다. 조사할 때 학생들의 (출신지, OO 후보 지지여부)를 조사하여 분할표를 작성해야 한다.

| OO 후보<br>출신지 | 지지 | 반대 | 합계 |
|---|---|---|---|
| 대전 | 30(75%) | 10(25%) | 40 |
| 충남 | 10(33.3%) | 20(66.7%) | 30 |
| 기타 지역 | 40(80%) | 10(20%) | 50 |
| 합계 | 80 | 40 | 120 |

괄호 안에 표시된 비율은 행 비율(row percentage)로 출신지별 후보 지지여부의 차이를 알 수 있다. 대전 출신자와 기타 지역 출신자의 OO 후보 지지도가 높고 충남 지역 출신자들은 반대 비율이 높음을 알 수 있다.

실제 출신지별 후보 지지 여부의 차이는 유의한가? (통계적 가설 검정➔분할표 검정)

### 2.1.2. 분할표 확률 구조

두 개의 범주형 변수를 각각 X 와 Y 로 표시하고 각각 I, J 수준을 갖고 있다고 하자. X 를 행으로 Y 를 열로 하여 분할표를 만들면 IxJ 개의 결합 조건이 존재한다. 이를 IxJ 분할표(contingency table) 혹은 교차표(cross-tabulation table)라 한다.

| X \ Y | 1 | 2 | … | C | **Total** |
|---|---|---|---|---|---|
| 1 | $\pi_{11}$ ($\pi_{1\|1}$) | $\pi_{12}$ ($\pi_{2\|1}$) | … | $\pi_{1c}$ ($\pi_{c\|1}$) | $\pi_{1+}$ |
| 2 | $\pi_{21}$ | $\pi_{22}$ | … | $\pi_{2c}$ ($\pi_{c\|2}$) | $\pi_{2+}$ |
| … | … | … | … | … | … |
| R | $\pi_{r1}$ | $\pi_{r2}$ | | $\pi_{rc}$ | $\pi_{r+}$ |
| **Total** | $\pi_{+1}$ | $\pi_{+2}$ | | $\pi_{+c}$ | $\pi_{++}$ |

$\pi_{ij}$ 는 (X, Y)가 (i 행, j 열)에 속할 확률로 $\pi_{ij} = \Pr(X = i, Y = j)$ 이다.

$\pi_{ij}$ : Joint distribution of (X, Y) (결합 밀도 함수)

$\pi_{i+} = \sum_j \pi_{ij}$ : Marginal distribution of (X) (주변 밀도 함수)

$\pi_{+j} = \sum_i \pi_{ij}$ : Marginal distribution of (Y) (주변 밀도 함수)

**예제를 살펴보면… (3x2 분할표)**

2.1.1.예제는 출신지에 따른 후보 지지여부의 차이가 있는지 알아보기 위하여 모집단으로부터 표본 120 명을 추출하여 조사한 것이다. 그러므로 분할표의 셀의 빈도는 $\pi_{ij}$ 의 추정치로 사용될 수 있다. ➔ $p_{ij} = \frac{f_{ij}}{n} = \hat{\pi}_{ij}$

표본에서의 결합밀도 함수는

$\hat{\pi}_{11} = 30/120 = 0.25$ (대전, 지지), $\hat{\pi}_{12} = 10/120 = 0.08$ (대전, 반대)

$\hat{\pi}_{21} = 10/120 = 0.08$, $\hat{\pi}_{22} = 20/120 = 0.17$

$\hat{\pi}_{11} = 40/120 = 0.33$, $\hat{\pi}_{32} = 10/120 = 0.08$ (기타, 반대)  모두의 합은 당연히 ?

지지여부 변수 Y 에 대한 주변 밀도함수는

$\hat{\pi}_{+1} = 80/120 = 0.67$ (지지), $\hat{\pi}_{+2} = 40/120 = 0.33$ (반대)

**HOMEWORK#1-1** 출신지 변수 X 에 대한 주변 밀도함수?

X=i 가 주어졌을 때 Y 의 조건부 확률 분포 함수(conditional)

$\hat{\pi}_{1|2} = 10/30 = 0.33$ (충남 출신자 중 지지하는 사람 비율)

$\hat{\pi}_{2|2} = 20/30 = 0.67$ (충남 출신자 중 반대하는 사람 비율)

### 2.1.3. Independence (독립성)

**Definition (정의)**

두 변수의 joint probability 가 각 변수의 marginal probability 의 곱과 같다면 두 변수는 통계적(서로)으로 독립(statistically independent)한다.

$$\pi_{ij} = \pi_{i+}\pi_{+j}$$

(cf) $p(AB) = P(A)P(B)$ 이면 A, B 는 서로 독립이다.

**Definition (정의)**

X 의 값이 주어졌을 경우 변수 Y 의 조건부(conditional) 확률은 다음과 같이 정의한다.

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}}$$

(cf) B 가 주어진 경우 A 조건부 확률은 $p(A \mid B) = \frac{P(AB)}{P(B)}$ 이다.

**Theorem (정리)**

두 변수 서로 독립이면 조건부 확률에 대해 다음이 성립한다.

$$\pi_{j|i} = \pi_{+j}$$

**HOMEWORK#1-2** 위 정리를 증명하시오.

### 2.1.4. 반응 변수에 대한 확률 모형

범주형 자료 분석의 경우 반응 변수에 대한 확률 모형은 반응 변수 수준이 2 개인 경우 이항 분포나 포아송 분포를 가정하거나 3 개 이상인 경우는 다항 분포를 가정하게 된다. 다행히도 범주형 자료 분석의 추론의 경우 어떤 확률 모형을 가정하든 동일한 결과를 가져오므로 어떤 확률 모형을 가정할 것인지에 대한 걱정은 할 필요가 없다.

## 2.2. 2x2 분할표 (Comparing proportions)

다음은 2x2 분할표이다.

| X \ Y | 1 | 2 |
|---|---|---|
| 1 | $\pi_{11}$ | $\pi_{12}$ |
| 2 | $\pi_{21}$ | $\pi_{22}$ |

### 2.2.1. Difference of proportions (비율 차이 검정)

2x2 분할표의 경우 두 변수간 연관성(association) 분석하는 경우 $\chi^2$-검정 대신 두 집단간 비율 차이 검정으로 대신할 수 있다. 변수의 수준이 2 개인 경우인 경우 binary(이진, dichotomous) 변수라 하고 3 개 이상인 경우를 poly-chotomous 라 한다. 이진 변수의 경우 일반적으로 성공, 실패로만 나눌 수 있으므로 (Bernoulli 시행) 성공 확률이 $p$ 이면 실패 확률은 $(1-p)$ 이다. 그러므로 2x2 분할표를 다음과 같이 쓸 수 있을 것이다.

2x2 분할표의 경우 행의 변수 X=1 일 경우 성공률이 $\pi_1$ 이면 실패율은 $(1-\pi_1)$ 이고 X=2 일 경우 성공률이 $\pi_2$ 이면 실패율은 $(1-\pi_{21})$ 이다.

| X \ Y | 성공 | 실패 |
|---|---|---|
| 1 | $\pi_1 (= \pi_{1\|1} = \pi_{11})$ | $(1-\pi_1)$ |
| 2 | $\pi_2 (= \pi_{1\|2} = \pi_{21})$ | $(1-\pi_2)$ |

**Theorem2-1** 2x2 분할표에서는 두 반응 변수가 서로 독립이다 ←→ $\pi_1 = \pi_2$

[증명] [Tip]X 와 Y 가 독립이다.$(P(XY) = P(X)P(Y))$ ←→ $P(Y=1 \mid X=1) = P(Y=1 \mid X=2)$

이런 식으로 증명하시오.

Example Harvard 대학에서 심장 마비 증세에 아스피린(aspirin)이 효과가 있는지 알아보기 위하여 한 그룹에는 위약을 다른 그룹에는 아스피린을 투여하여 5년이 지난 후 심장 마비 발생 여부를 조사하여 다음 표를 얻었다. (Alan Agresti textbook page 17)

| 약 \ 치명적 심장마비 | 발생 | 건강 |
|---|---|---|
| Placebo (위약) | 189 | 10,845 |
| 아스피린 | 104 | 10,933 |

약의 종류와 심장 마비 발생 여부와 연관성이 있는가? (즉 약의 종류에 따른 심장 마비 여부 차이는 있는가?) 분석은 위약 복용자의 심장 마비 발생율과 아스피린 복용자의 심장 마비 발생율의 차이가 있는지 검정하는 것과 동일하다.

위약 복용자 심장 마비 발생률 $\pi_1$ 의 추정치는 $\hat{\pi}_1 = p_1 = \frac{189}{11{,}034} = 0.0171$ 이고 아스피린 복용자 심장 마비 발생율 $\pi_{12}$ 의 추정치는 $\hat{\pi}_2 = p_2 = \frac{104}{11{,}037} = 0.0094$ 이다.

**비율 차이 검정 순서**

- 귀무가설: $H_0 : \pi_1 = \pi_2$ (위약 복용자 심장마비 발생율은 아스피린 복용자의 그것과 동일하다)
- 대립가설: $H_a : \pi_1 \neq \pi_2$ (같지 않다)
- 통계량: $(p_1 - p_2) = (\frac{\sum x_{1i}}{n_1} - \frac{\sum x_{2i}}{n_2})$

**Recall** $\sum x_{1i} \sim Binomial(n_1, \pi_1)$ : Bernoulli 시행

**Theorem2-2** 두 집단이 서로 독립이고 표본의 크기$(n_1, n_2)$가 크다면

$$(p_1 - p_2) \sim Normal(\pi_1 - \pi_1, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}) \text{ 이다.}$$

[증명]

Example 심장 마비 증상에 대한 아스피린 효과 실험. (계속)

○ 검정통계량: $T = \dfrac{(p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} = \dfrac{(0.0171 - 0.0094)}{\sqrt{\dfrac{0.0171(1-0.0171)}{11{,}034} + \dfrac{0.0094(1-0.0094)}{11{,}037}}} = 5.13$

○ 결론: 유의수준을 5%라 (←→신뢰수준 95%) 하면 기각 값은 (critical value) 1.96 이다. 검정 통계량의 값이 기각 값보다 크므로 귀무가설을 기각한다. 심장 마비 발생율은 차이는 있다. 그러므로 아스피린 복용자의 심장 마비 발생율이 위약 복용자보다 낮으므로 아스피린은 심장 마비 억제 효과가 있음을 알 수 있다.

○ 신뢰 구간: $(p_1 - p_2) \pm z_{\alpha/2}\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$

$$0.0077 \pm 1.96(0.0015) \Rightarrow (0.005, 0.011)$$

### 2.2.2. Relative risk 와 Odds Ratio

**Relative risk**

두 비율의 값의 차이는 두 비율이 0.5 근처일 경우보다는 0 이나 1 일 경우 더 중요한 의미를 갖는다. 예를 들어 두 집단의 비율의 차이가 0.0077 인 경우 (0.0171, 0.0094)가 (0.5, 0.5077) 보다는 상대적 중요성을 갖는다. 이를 개념화 한 값이 relative risk (상대 위험도)이다.

$$\text{상대 위험도} = \frac{\pi_1}{\pi_2}$$

Example 심장 마비 증상에 대한 아스피린 효과 실험. (계속)

$$\text{상대 위험도 추정치} = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{0.0171}{0.0094} = 1.82$$

위약 복용 그룹의 심장 마비 발생율이 82%나 높다. (0.5, 0.5077) 경우에는 상대 위험도가 1.015 로 같은 0.0077 차이이지만 상대 위험도는 1.5% 밖에 되지 않는다.

## Odds

Odds 는 성공 확률을 실패 확률로 나눈 값으로 다음과 같으며 축구나 농구 등 둘이 하는 경기에서 배팅(betting)을 하는 경우 이익 배당의 근거가 된다. $\pi$ 성공 확률이라고 하면 Odds 는 다음과 같이 정의되며 반응 변수의 (실패율 대비)성공률이라고 해석될 수 있다.

$$\text{Odds} = \frac{\pi}{1-\pi}$$

Example 한국과 폴란드의 경기에서 한국이 이길 확률을 0.1라고 하면 한국의 Odds는

$$\frac{\pi}{1-\pi} = \frac{0.1}{0.9} = \frac{1}{9} \text{ 이다.}$$

Odds 의 의미는 한 번 이기기 위해서는 9 번 지는 경기를 한다는 것이다. 즉 질 가능성이 9 배가 된다는 것이다. 폴란드의 Odds 는 9 이므로 이길 가능성이 9 배가 된다는 것이다. 이 Odds 는 축구나 농구와 같이 두 팀이 하는 경기의 betting 의 배당금 배분의 근거가 된다. 한국에 거는 사람은 1$을 걸면 9$ 배당금을 받지만 폴란드에 거는 사람은 9$을 걸어야 1$을 배당 받게 된다.

**Odds ratio**

$$\text{Odds ratio } \theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

Odds ration 의 값은 음의 실수 값을 가지며 두 반응 변수가 서로 독립이면 ($\pi_1 = \pi_2$) Odds ratio 는 1 이다. Odds ratio 가 1 보다 크면 1 행 변수의 (반응변수 1: 예 위약) 성공률이 2 행 변수의 (반응변수 2: 예 아스피린) 성공률이 보다 높다는 것이고 1 보다 작으면 반응변수 2 의 성공률이 더 높음을 의미한다. Odds ratio 가 1 에서 멀어질수록 두 반응 변수는 독립성에서 멀어진다.

**Property2-1** 2x2 분할표에서 행과 열을 바꾸어도 Odds ratio는 변하지 않는다.

[증명]

위 property 의 의미는 반응 변수(행)와 설명 변수(열)가 바뀌어도 Odds ratio 가 동일하므로 변수를 구별할 필요는 없다. (relative risk 는 행과 열이 바뀌면 달라진다)

**Odds ratio 추정치**

Odds ratio 추정치 $\hat{\theta} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$ where $n_{ij}$ 는 i 행 j 열 빈도

Example 심장 마비 증상에 대한 아스피린 효과 실험. (계속)

위약 복용자 Odds ratio $\frac{189}{10845}=0.0174$ ➔ 심장 마비 발생(성공)률은 0.0174 로 1.74 명의 심장 마비가 일어나는 동안 100 명은 무사하다.

아스피린 복용자 Odds ratio $\frac{104}{10933}=0.0095$ ➔ 심장 마비 발생(성공)률은 0.0095 로 0.95 명의 심장 마비가 일어나는 동안 100 명은 무사하다.

Odds ratio 추정치 $\hat{\theta}=\frac{n_{11}n_{22}}{n_{12}n_{21}}=\frac{189\times10933}{10845\times104}=1.832$

위약 복용자의 심장 마비 발생률은 아스피린 복용자의 심장 마비 발생율보다 83% 높다. (0.5, 0.5077) 경우에는 상대 위험도가 1.015 로 같은 0.0077 차이이지만 상대 위험도는 1.5% 밖에 되지 않는다.

**Odds ratio 추론**

두 반응 변수가 서로 독립인지 (연관성 검정) 어떻게 검정할 수 있을까?

$\theta$ 의 값은 0 과 $\infty$ 을 가지고 독립인 경우는 1 이다. 그러므로 좌우 비대칭 형태의 분포를 가지므로 $\theta$ 대신 $\ln(\theta)$ 생각해보자. 두 변수가 독립이면 $\ln(1)=0$ 이고 (한 개념에서) 좌우 대칭의 형태를 갖는다. (예: $\ln 4=1.39$, $\ln(1/4)=-1.39$)

표본의 크기가 커지면 $\ln(\hat{\theta})\underrightarrow{app}Normal(\ln(\theta),\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}})$

그러므로 $\ln(\theta)$ 의 95% 신뢰구간은 $(\ln(\hat{\theta})\pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}})$

Odds ration $\theta$ 의 95% 신뢰구간은 $e^{\ln(\hat{\theta})\pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}}$ 이다.

Example 심장 마비 증상에 대한 아스피린 효과 실험. (계속)

$\ln(\hat{\theta})=\ln(1.832)=0.605$ 이므로

$\ln(\theta)$ 의 95% 신뢰구간은 $(0.605\pm1.96\sqrt{\frac{1}{189}+\frac{1}{10933}+\frac{1}{104}+\frac{1}{10845}})$ ➔(0.365, 0.846)

$\theta$ 의 95% 신뢰구간은 $(e^{0.365},e^{0.846})$ ➔(1.44, 2.33)

신뢰구간이 1 을 포함하고 않고 1 이상이므로 위약의 심장 마비 발생율이 더 높다고 결론 지을 수 있다.

**Relative risk 와 Odds ratio**

$$\text{Odds ratio} = \text{Relative risk} \times \left(\frac{1-\pi_{1|2}}{1-\pi_{1|1}}\right)$$

두 반응 변수의 성공 확률이 0 에 가까우면(아스피린 예제의 경우) Odds ratio 는 Relative risk 와 유사한 값을 갖는다. (아스피린 예: $1.83 \approx 1.82$)

두 반응 변수의 성공의 상대적 비교를(예: 위약과 아스피린간 심장 마비 발생율의 상대적 비교) 할 때는 Relative risk 값을 가지고 해야 한다. 즉 위약이 아스피린에 비해 1.82 배 심장 마비 걸릴 가능성이 높다고 말할 수 있다. Odds ratio 는 (실패율 대비) 성공률의 비율이므로 앞에 해석과는 거리가 멀다.

위의 관계식을 이용하면 Relative risk 를 구할 수 없는 경우 Odds ratio(반응변수와 설명변수를 바꾸어도 계산이 가능)를 이용하여 계산 가능하다.

Example 흡연이 폐암에 영향을 미치는지 알아보기 위하여 폐암 증상으로 병원을 찾아온 환자 200명과 이 환자들에 일반 환자 2명을 짝지어 일반 환자 400명의 흡연 여부를 조사하였다. (이런 연구를 retrospective study 혹은 case-control study라 한다.)

| 폐암 증상 / 흡연 | 폐암 환자 | 일반 환자 (대조군) |
|---|---|---|
| 예 | 140 | 100 |
| 아니오 | 60 | 300 |

흡연에 따른 폐암 발생 비율에 대해 알아보도록 하자. 흡연자의 폐암 발생 비율($\pi_{1|1}$)과 비흡연자의 폐암 발생 비율($\pi_{1|2}$)을 이용한다? 그러나 폐암 환자의 비율이 1/3 로 고정되어 있으므로 $\pi_{1|1}$, $\pi_{1|2}$ 을 구하는 것은 의미가 없다.

대신 폐암 환자 중 흡연 비율($140/200 = 0.7$)과 일반 환자 중 흡연 비율($100/400 = 0.25$)을 구할 수 있다. 즉 설명 변수가 주어졌을 경우 반응 변수의 조건부 확률을 구할 수 있다. Odds ratio 의 경우는 반응 변수와 설명 변수의 구별이 없으므로 Odds ratio 의 추정치는 $\hat{\theta} = \frac{140 \times 300}{60 \times 100} = 7$ 이다. 흡연 경험자의 Odds ratio 추정 값은 폐암 발생 환자의 경우 [140/200]/[60/200]=2.33 이고 폐암 미발생자의 경우 Odds ratio 는 0.33 이다.

의학 연구 결과 폐암 발생율은 매우 낮으므로 Odds ratio 를 relative risk 로 해석하여 흡연 경험자가 비흡연자에 비해 약 7 배 폐암 발생율이 높다고 말할 수 있다.

HOMEWORK#2-1 Theorem2-1, Theorem2-2, Property2-1을 증명하시오. =

HOMEWORK#2-2 다음은 조사 결과에 대해 답하시오. [Clogg and Shockey 1988]

| 총기등록법 \ 사형 제도 | 찬성 | 반대 |
|---|---|---|
| 찬성 | 784 | 236 |
| 반대 | 311 | 66 |

1) 반응변수와 설명변수가 무엇인지 밝히시오.
2) 두 변수의 연관성을 비율 차이 검정에 의해 분석하시오. [유의수준=5%]
3) Relative risk 추정치를 구하고 해석하시오.
4) Odds ratio 추정치를 구하고 해석하시오.
5) Odds ratio 를 이용하여 두 변수가 서로 독립인지 검정하시오. [유의수준=5%]

HOMEWORK#2-1 Theorem2-1, Theorem2-2, Property2-1을 증명하시오.

HOMEWORK#2-2 다음은 조사 결과에 대해 답하시오. [Clogg and Shockey 1988]

| 총기등록법 \ 사형 제도 | 찬성 | 반대 |
|---|---|---|
| 찬성 | 784 | 236 |
| 반대 | 311 | 66 |

6) 반응변수와 설명변수가 무엇인지 밝히시오.
7) 두 변수의 연관성을 비율 차이 검정에 의해 분석하시오. [유의수준=5%]
8) Relative risk 추정치를 구하고 해석하시오.
9) Odds ratio 추정치를 구하고 해석하시오.
10) Odds ratio 를 이용하여 두 변수가 서로 독립인지 검정하시오. [유의수준=5%]

페이지 14 의 예제 연구를 retrospective study(look into the past)라 (이를 case-control 연구라고도 한다.) 하는데 이 경우 반응 변수의 주변 분포(marginal dist.)가 주어져 있다.

이미 실험 결과가 나타난(폐암) 조사자를 대상으로 폐암 발생 여부를 물으므로 반응 변수의 주변 확률이 주어지고 일반적인 방법으로 Odds ratio 를 구하지 못해 Odds 와 Relative Risk 의 관계식에 의해 $\hat{\theta}$(Odds ratio 의 추정치) 구했다.
위의 연구 방법을 관측 연구라 하는데 이는 실제 영향을 제대로 파악하지 못하는 문제가 있다. 즉 흡연이 폐암 발생에 직접적인 영향을 미쳤는지는 알 수 없다. 관측 연구의 또 다른 방법은 독립 변수와 반응 변수의 그룹을 동시에 조사하는 Cross-sectional 연구가 있다. 이것이 우리가 일반적으로 보는 2x2 분할표이다.
폐암이 흡연 발생 여부에 영향을 미치는 것을 보기 위해서는 실험을 해야 한다. 건강 상태가 양호한 사람들을 두 집단(실험군, 대조군) 나눈 후 실험 집단은 흡연하게 하고, 대조군에는 흡연하지 못하게 하여 일정 시간(5 년)이 지난 후 두 집단의 흡연 비율의 차이를 검정하면 된다. 이런 실험 방법을 실험 연구라 한다. Cohort 연구는 피시험자가 어느 그룹에 속할지 정한 후 실험을 한다는 것이 실험 연구와 다르다. 이 두 연구를 prospective study 라 한다.
Homework #2-2 의 예제는 prospective study 예제이다. 그러므로 일반적인 방법에 의해 Odds ratio 의 추정치를 구하면 된다. Odds ratio 와 relative risk 의 관계식을 이용하여야 하는 예제는 Homework#3 이다.

## 2.4. IxJ contingency table 분석하기

지금까지는 IxJ 분할표의 특별한 케이스인 2x2 분할표에 대한 검정을 살펴보았는데, 이를 정리하면 다음과 같다.

1) 분할표를 작성할 때는 항상 행은 설명변수, 열은 종속변수로 한다.

2) 2x2 분할표에서 설명변수와 종속변수의 독립성 검정은 각 수준(이를 집단)의 성공률의 차이 검정과 같다. ($\pi_{y=1|x=1} = \pi_{y=1|x=2}$ ←→X 와 Y 는 독립)

3) 2x2 분할표에서 설명변수와 종속변수의 독립성 검정은 각 설명변수의 Odds ratio $\theta$ 가 1 인 것을 검정하는 것과 같다. Odds ratio 의 신뢰구간을 구하여 $\theta = 1$ 을 포함하고 있으면 서로 독립이고, $\theta > 1$ 이면 분자 집단 성공 확률이 높다고 하고, $\theta < 1$ 이면 분모 집단의 성공 확률이 높다. ($\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = 1$ ←→X 와 Y 는 독립)

4) Odds ratio 와 relative risk 의 관계 Odds ratio=Relative risk $\times (\frac{1-\pi_{1|2}}{1-\pi_{1|1}})$ 와 odds ratio 는 설명변수와 반응변수의 위치에 상관없이 계산될 수 있으므로 retrospective study(이 경우

종속변수 각 수준의 응답자가 정해져 있어 일반 분할표 검정이 불가능)의 분할표 분석이 가능하다. (페이지 14 참고)

5) 설명변수와 종속변수가 모두 순서형인 경우 “설명변수 X 가 증가함에 따라 Y 가 증가한다고 할 수 있나?” 두 변수간의 상관 관계 분석이 가능하다. concordant 한 짝이 많으면 X 가 증가하면 Y 가 증가한다고 볼 수 있을 것이다. 반대로 discordant 한 짝이 많으면 X 증가에 따라 Y 는 감소한다고 결론 지을 수 있다. 물론 tied 가 많은 경우는 X 와 Y 의 관계(association)가 없다고 할 수 있다. (페이지 17 참고)

이 절에서는 IxJ 분할표 검정을 의한 방법들을 살펴보기로 하자. 이 검정들을 좁은 의미의 IxJ 분할표 범주형 자료 분석이라 한다. 물론 이 절에서 언급되는 방법들은 2x2 분할표에도 적용할 수 있다. 다음은 IxJ 분할표이다. 귀무가설(모집단) 하에서 설정된 결합확률 $\pi_{ij}$ 로부터 계산된 기대빈도를 $E_{ij}$ 라 하자.

| Y / X | 1 | 2 | … | C | **Total** |
|---|---|---|---|---|---|
| 1 | $\pi_{11}(E_{11})$ | $\pi_{12}(E_{12})$ | … | $\pi_{1c}(E_{1c})$ | $\pi_{1+}$ |
| 2 | $\pi_{21}(E_{21})$ | $\pi_{22}(E_{22})$ | … | $\pi_{2c}(E_{2c})$ | $\pi_{2+}$ |
| … | … | … | … | … | … |
| R | $\pi_{r1}$ | $\pi_{r2}$ | | $\pi_{rc}$ | $\pi_{r+}$ |
| **Total** | $\pi_{+1}$ | $\pi_{+2}$ | | $\pi_{+c}$ | $\pi_{++}$ |

### 2.4.1. Goodness of fit

1900 Karl Pearson 에 제안한 방법으로 multinomial(다항) distribution 의 확률이 귀무가설에서 설정한 값과 동일한지를 검정한다.

1) 귀무가설(일정한 분포함수를 갖는다) 하에서 셀의 기대 확률을 $\pi_i$ 이라 하면 이 셀의 기대 빈도(expected frequency)는 $E_i = n\pi_i$ 이자.

2) 실제 자료에 의해 관측된 빈도 관측 빈도 (observed frequency) $O_i$ 라 하자.

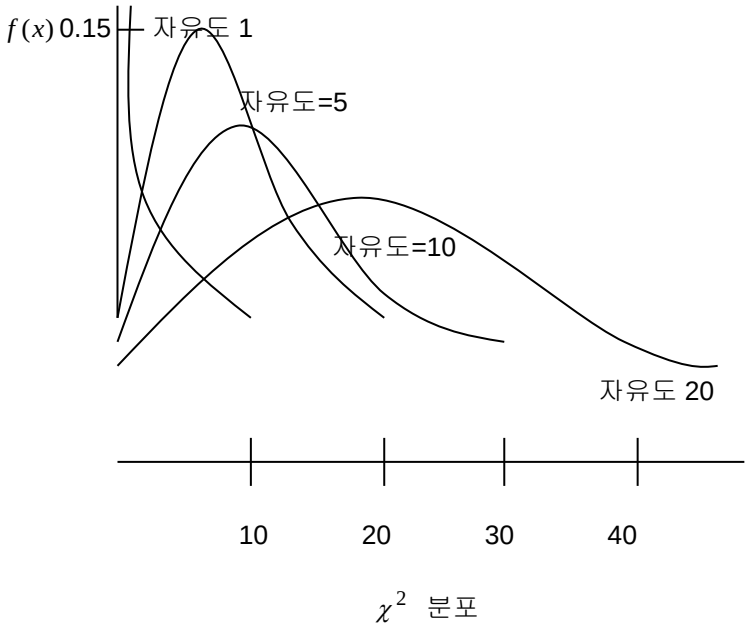
3) ( $E_{ii}$ - $O_i$ )의 값들이 크면 귀무가설을 기각하게 될 것이고 그렇지 않으면 귀무가설을 채택한다.

**Pearson Chi-square Statistic ( $\chi^2$ -검정 통계량)**

다음을 (피어슨) $\chi^2$ -통계량이라 하면 표본의 크기가 크면 근사적으로 $\chi^2$ (자유도=(r-1)(c-1)) 분포를 따른다.

$$T = \sum_i \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(df = (c-1))$$

$O_i$ 는 관측 빈도, $E_i = n\pi_i$ , $n$ =총 응답자 수, $\pi_i$ =귀무가설 하에서 i 번째 셀의 기대 확률, 그러므로 $E_i$ 는 기대 빈도이다.




$\chi^2$ 분포

**예제 1: Mendel 의 유전 법칙**

이 방법에 대한 초기 예제는 Mendel 의 유전 법칙에 관한 것이다. 2 세대 잡종은 노란 75%, 초록 25%라는 이론의 사실 여부를 알아보기 위하여 $n = 8023$ 을 조사하였더니 6022 가 노란색, 2001 이 초록색이었다.

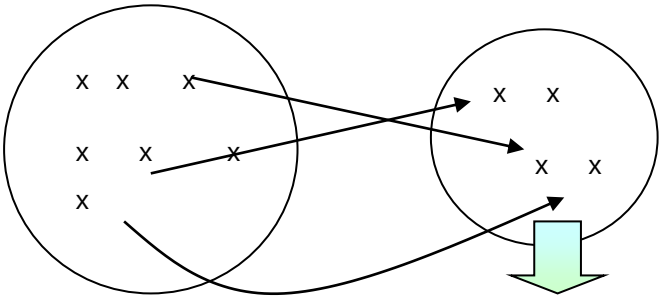
| | 노란색 | 초록색 |
|---|---|---|
| 관측 빈도 | 6022 | 2001 |
| 기대 확률 | 0.75 | 0.25 |

| | | |
|---|---|---|
| | 6017.3 | 2005.8 |

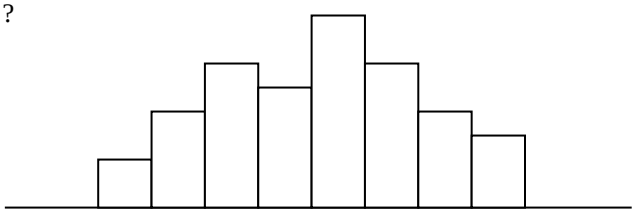
검정통계량 $T = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{(6022-6017.3)^2}{6017.3} + \frac{(2001-2005.8)^2}{2005.8} = 0.015$

기각치 $\chi^2(df = 1, \alpha = 0.05) = 3.84$ 보다 작으므로 귀무가설 채택.

**예제 2: 어떤 자료가 임의의 분포를 따르는가?**

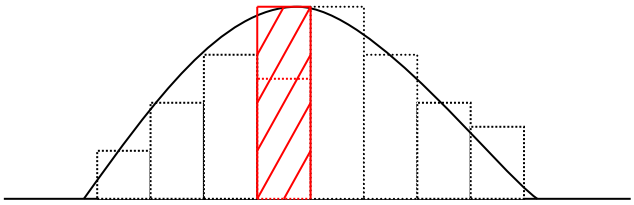


population: $f(x)$?



| 구간 | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ | $C_8$ |
|---|---|---|---|---|---|---|---|---|
| 관측 도수 | $O_1$ | $O_2$ | $O_4$ | $O_4$ | $O_5$ | $O_6$ | $O_7$ | $O_8$ |

모집단의 분포가 $f(x)$? (예: 정규분포) 따를까?



| 구간 | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ | $C_8$ |
|---|---|---|---|---|---|---|---|---|
| 기대 도수 | $E_1$ | $E_2$ | $E_4$ | $E_4$ | $E_5$ | $E_6$ | $E_7$ | $E_8$ |

표본 분포가 설정한 모집단 분포와 동일하다면…

관측 도수와 (observed frequency) 기대 도수는 (expected frequency) 비슷한 값일 것이다. 즉 $O_1 \approx E_1$, $O_2 \approx E_2, \ldots, O_k \approx E_k$ (위 예에서는 $k=8$)

검정통계량 (test statistics) ? $T = \dfrac{\sum_{i=1}^{k}(O_i - E_i)^2}{E_i} \sim \chi^2(df = k-c-1)$ c=모수 추정 개수

이를 $\chi^2$- 적합성 검정 방법이라 한다.

**예제 3: 주사위 예제**

주사위로 게임을 하려고 주사위를 하나 샀다. 이 주사위 각 면이 나올 확률이 동일한지 (fair) 알아보기 위하여 실험을 하기로 하였다. 주시위를 1,000 번 던져 다음 결과가 나왔다.

| 눈금 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 빈도 | 150 | 160 | 165 | 155 | 170 | 200 |

- 귀무가설: 각 눈금이 나올 확률은 모두 1/6 로 같다.

  X=주사위 눈금 ➔ $f(x) = 1/6 \quad \text{for } x = 1,2,...,6$

- 대립가설: 각 눈금이 나올 확률이 모두 1/6 은 아니다.

  주사위 눈금 X 는 귀무가설의 확률 분포 $f(x)$ 를 따르지 않는다.

- 검정통계량

| 눈금 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 관측 빈도 $(O_i)$ | 150 | 160 | 165 | 155 | 170 | 200 |

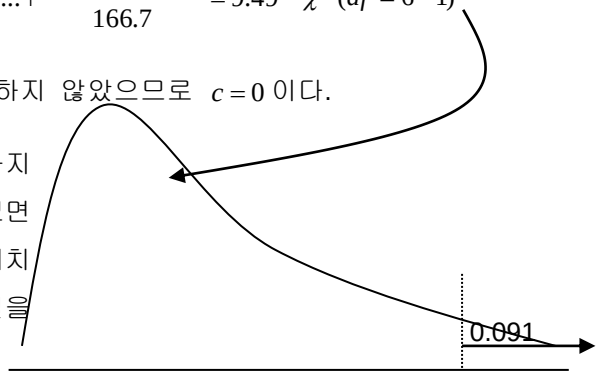
| 기대 빈도 $(E_i)$ | 166.7 | 166.7 | 166.7 | 166.7 | 166.7 | 166.7 |
|---|---|---|---|---|---|---|

기대 빈도는 귀무가설이 맞다는 가정 하에서 계산한다.

- 검정통계량: $T = \frac{(150-166.7)^2}{166.7} + \frac{(160-166.7)^2}{166.7} + ... + \frac{(200-166.7)^2}{166.7} = 9.49 \sim \chi^2(df = 6-1)$

  기대 빈도 계산을 위하여 어떤 모수도 추정하지 않았으므로 $c = 0$ 이다.

- 결론: p-값이 0.091 이므로 귀무가설을 기각하지 못한다. 주사위는 fair 하다. 다른 측면에서 보면 검정 통계량(9.49)이 유의수준 5%의 임계치 (critical value) 11.07 보다 작으므로 귀무가설을 기각하지 못한다.




**예제 4: Binomial Distribution**

베르누이 시행을 n 번 독립적으로 시행했을 경우 X 를 성공한 회수라 하면

$$f(x) = \binom{n}{c} p^x (1-p)^{n-x} \quad \text{for } x = 0,1,...,n \;\rightarrow\; \text{Binomial (n, p)}$$

평균 $E(X) = np$, 분산 $V(X) = np(1-p)$

남녀 출산 비율이 0.5 인지 알아보기 위하여 아이들이 3 명이 1,000 가구를 대상으로 남자 아이의 수를 조사하여 다음 표를 얻었다.

| 남자 아이 수 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 빈도 | 100 | 350 | 400 | 150 |

- 귀무가설: 남자 아이 수는 이항 분포(n=3, p=0.5)를 따른다.

$$\text{X=남자 아이 수} \;\rightarrow\; f(x) = \binom{3}{x} (0.5)^x (1-0.5)^{3-x} \quad \text{for } x = 0,1,2,3$$

- 대립가설: X 는 이항 분포를 따르지 않는다.
- 검정통계량

| 남자 아이 수 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 관측 빈도 $(O_i)$ | 100 | 350 | 400 | 150 |
| 기대 확률 | 0.125 | 0.375 | 0.375 | 0.125 |
| 기대 빈도 $(E_i)$ | 125 | 375 | 375 | 375 |

기대 빈도는 귀무가설이 맞다는 가정 하에서 계산한다.

$$f(x=0)=\binom{3}{0}(0.5)^0(1-0.5)^3=0.125\ ,\quad f(x=1)=\binom{3}{1}(0.5)^1(1-0.5)^3=0.375$$

- 검정통계량: $T=\frac{(100-125)^2}{125}+\ldots+\frac{(150-125)^2}{125}=13.3\sim\chi^2(df=4-1)$

  기대 빈도 계산을 위하여 어떤 모수도 추정하지 않았으므로 $c=0$ 이다.

- 결론: p-값이 0.004 이므로 귀무가설을 기각한다. (혹은 검정 통계량 값 13.3 이 임계치 7.82 보다 크므로) 그러므로 남자 아이의 수는 성공 확률이 0.5 인 이항 분포를 따르지 않는다.

**예제 5. Poisson Distribution**

Poisson 분포는 $n\to\infty, p\to 0$ 인 이항 분포로부터 유도된다. [$\lambda=np$ 수리 통계 참고]

X 를 관심이 있는 사건이 발생할 회수라 하면

$$f(x\mid\lambda)=\frac{e^{-\lambda}\lambda^x}{x!}\ \text{ for } x=0,1,2,\ldots,\ \text{평균}\ E(X)=\lambda\ \text{분산}\ V(X)=\lambda$$

하나 은행에서 지난 한달 동안 조사하였더니 한 시간에 평균 6 명의 고객이 방문하고 그 방문 회수는 포아송 분포를 따르고 있음을 알았다. 그럼 10 분 동안 고객이 한 명도 찾아오지 않을 확률은?

$\lambda = np = 6 \times 1/6 = 1$ ➔ $P(X = 0 \mid X \sim Poisson(\lambda = 1)) = \dfrac{e^{-1}(1)^0}{0!} = 0.36788$

다음은 한남대학교 정문을 통과하는 차량의 수가 Poisson 분포를 따르는지 알아보기 위하여 1 분마다 차량 통과 회수를 300 회 조사하였다. 아래 자료를 이용하여 Poisson 분포를 따르는지 검정하시오. (유의수준=0.05)

| 통과 차량 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| 관측 빈도 | 20 | 54 | 74 | 67 | 45 | 25 | 11 | 4 |

- 귀무가설: 위의 자료는 Poisson 분포를 따른다.
- 대립가설: Poisson 분포를 따르지 않는다.

각 셀의 기대 빈도를 구하기 위해서는 Poisson 분포의 모수를 (λ) 알아야 한다.

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0,1,2,\mathrm{K}$$

표본 자료로부터 모수 p의 추정치를 $(\hat{\lambda})$ 구하면 $\hat{\lambda} = (0 \times 20 + 1 \times 54 + \mathrm{K} + 7 \times 4)/300 = 2.67$

그러므로 기대 확률과 기대 빈도는 다음 Poisson 확률 분포에 의해 계산하면 된다.

$$p(x) = \frac{e^{-2.67}2.67^x}{x!}$$

| 통과 차량 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| 관측 빈도 | 20 | 54 | 74 | 67 | 45 | 25 | 11 | 4 |
| 기대 확률 | 0.069 | 0.185 | 0.247 | 0.22 | 0.147 | 0.078 | 0.035 | 0.013 |
| 기대 빈도 | 20.7 | 55.5 | 74.1 | 66 | 44.1 | 23.4 | 10.5 | 3.9 |

- 검정 통계량: $T = \dfrac{(20-20.7)^2}{20.7} + \dfrac{(54-55.5)^2}{55.5} + \mathrm{K} + \dfrac{(4-3.9)^2}{3.9} = 0.234$

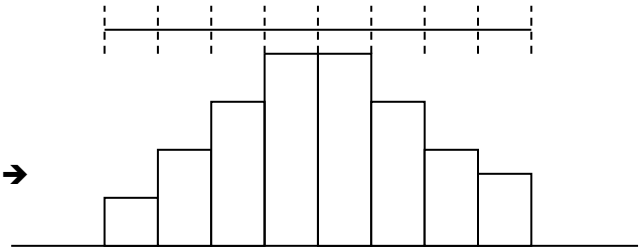
결론: 표로부터 임계치는 $\chi^2$(자유도=8-1-1=6, $\alpha = 0.05$)=12.59 이므로 귀무가설이 채택되고 이 자료는 Poisson 분포를 따른다고 할 수 있다. (자유도 계산 시 1 을 더 빼 주는 이유는 포아송 분포의 모수 $\lambda$를 알지 못하므로 자료를 이용하여 추정하였기 때문이다.

**예제 6. 정규분포**

이산형 확률 모형에 대한 적합성 검정의 경우는 구간(셀)을 분석자가 나눌 필요는 없다. 이항분포나 포아송 분포의 예를 보면 변수가 가질 수 있는 값이 이산이므로 각 값을 셀로 설정하면 된다. 그러나 연속형의 경우는 한 값에 대한 확률은 존재하지 않는다. 그러므로 자료에 의해 구간을 나누고 그 구간을 셀로 하여 적합성 검정을 실시하면 된다.

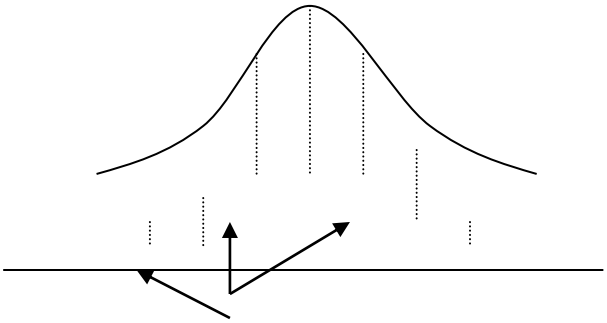
자료의 범위

표본으로부터 히스토그램 ➔



귀무가설에서

설정한 분포 ➔



기대 확률

기대 확률을 이용하여 기대 빈도( $E_i$ )를 구하고 히스토그램의 관측 빈도( $O_i$ )를 이용하여 적합성 검정을 실시하면 된다.

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$ ➔ 평균= $\mu$ , 표준 편차= $\sigma$

다음 키 (inch) 자료가 정규분포를 따름을 보이시오. (n=19)

65 69 56 62 63 57 59 62 62 59

51 64 56 66 72 64 67 57 66

- 귀무가설: 자료는 정규분포를 따른다. / 대립가설: 정규분포를 따르지 않는다.
- 정규분포의 모수는 평균과 표준편차다. ➔ 추정치: $\bar{x}=62, s=5.19$ 2 개 추정
- 범위=72-51=21 ➔ 구간의 폭 $21/6 \approx 4$
- 자료로부터 빈도표 (히스토그램) 만들기

| 셀(구간) | 51-54 | 55-58 | 59-62 | 63-66 | 67-70 | 71-74 |
|---|---|---|---|---|---|---|
| 관측 빈도 | 1 | 4 | 5 | 6 | 2 | 1 |
| 연속 구간 | 50.5-54.5 | 54.5-58.5 | 58.5-62.5 | 62.5-66.5 | 66.5-70.5 | 70.5-74.5 |

연속 구간? 이산 구간을 연속일 때 $\Pr(x=54)=0$ 그러나 $x=54$가 존재

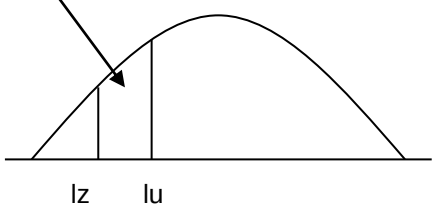
- 기대확률 구하기 ➔ SAS 정규 분포 함수 이용하는 프로그램

```
data prob;
    input l u @@;
    lz=(l-62)/5.19; lu=(u-62)/5.19;
    p=probnorm(lu)-probnorm(lz);
    f=p*19;
    cards;
    50.5 54.5
    54.5 58.5
    58.5 62.5
    62.5 66.5
    66.5 70.5
    70.5 74.5
run;

proc print data=prob;
    var p f;
run;
```

표준화 ➔ 표준 정규분포 이용

표준 정규분포 누적 밀도 함수




| Obs | p | f |
|---|---|---|
| 1 | 0.06086 | 1.15642 |
| 2 | 0.17582 | 3.34058 |
| 3 | 0.28834 | 5.47841 |
| 4 | 0.26867 | 5.10471 |
| 5 | 0.14222 | 2.70221 |
| 6 | 0.04273 | 0.81179 |

| 셀(구간) | 51-54 | 55-58 | 59-62 | 63-66 | 67-70 | 71-74 |
|---|---|---|---|---|---|---|
| 관측 빈도 | 1 | 4 | 5 | 6 | 2 | 1 |
| 기대 확률 | 0.06 | 0.18 | 0.29 | 0.27 | 0.14 | 0.04 |
| 기대 빈도 | 1.16 | 3.34 | 5.48 | 5.1 | 2.7 | 0.8 |

- 검정 통계량: $T = \frac{(1-1.16)^2}{1.16} + \frac{(4-3.34)^2}{3.34} + ... + \frac{(1-0.8)^2}{0.8} = 0.58$
- 임계치: $\chi^2(df = 6-1-2, \alpha = 0.05) = 7.81$ ($\chi^2$-분포표)
- 검정 통계량이 임계치보다 작으므로 귀무가설은 채택되고 자료는 정규분포를 따르고 있다고 할 수 있다.

**HOMEWORK #4-1**

여러분이 가진 동전의 앞면과 뒷면이 나올 확률이 동일한지 알아보는 실험을 하고 그 동전이 fair 한지 검정하시오.

**HOMEWORK #4-2**

다음은 120 개 지역에 폭격 회수를 조사한 자료이다. Poisson 분포를 따르고 있음을 보이시오.

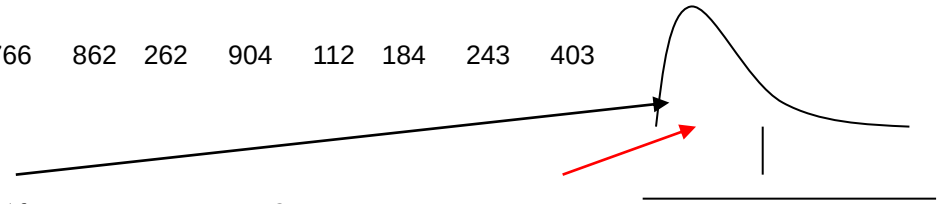
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 폭격 | 24 | 16 | 16 | 18 | 15 | 9 | 6 | 5 | 3 | 4 | 3 | 0 | 1 |

**HOMEWORK #4-3**

다음은 전구 수명 자료이다. 지수분포를 따름을 보이시오.

23 1326 290 47 32 446 11 486 773 18

338 348 766 862 262 904 112 184 243 403



$f(x|\beta)=\frac{1}{\beta}e^{-x/\beta}$ (평균, 분산)= $(\beta,\beta^2)$ SAS 함수: $PROBGAM(x/\beta,1)$ $x$

### 2.4.2. Independence Test

귀무가설: 두 변수는 서로 독립이다. ( $\pi_{ij}=\pi_{i+}\pi_{+j}$ from P(AB)=P(A)P(B)) 그러므로 귀무가설 하에서 각 셀의 기대빈도는 $E_{ij}=n\pi_{i+}\pi_{+j}$ (n 는 총응답자 수) 이에 대한 검정으로

**1900 Karl Pearson 의 $\chi^2$-검정 이용한다.** 표본의 크기가 크다면 다음이 성립한다.

$$T=\sum_{i,j}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}=\sum_{i,j}\frac{(n_{ij}-E_{ij})^2}{E_{ij}}\sim\chi^2(df=(r-1)(c-1))$$

**Likelihood Ratio Chi-Square**

$$\Lambda=\frac{\prod_{i,j}(n_{i+}n_{+j})^{n_{ij}}}{n^n\prod_{i,j}n_{ij}^{n_{ij}}}$$ from $\prod_{i,j}\pi_{ij}^{n_{ij}}$ ~ Multinomial

$$G^2=-2\log(\Lambda)=2\sum_{i,j}(O_{ij}\log(\frac{O_{ij}}{E_{ij}}))\sim\chi^2(df=(I-1)(J-1))$$

- Karl Pearson Chi-square 통계량과 LR Chi-square 통계량은 asymptotically equivalent 하다.
- Karl Pearson Chi-square 통계량과 LR Chi-square 통계량은 분할표의 범주들의 순서에 invariant 하다. 범주의 순서를 바꾸어도 계산된 검정 통계량 값은 변하지 않는다.

**예제**

년 소득에 따른 직업 만족도의 차이는 있는가를 알아보기 위한 예제 [계속]

| | 직업 만족도 | | | |
|---|---|---|---|---|
| | Very Dissatisfied | Little Dissatisfied | Moderately Satisfied | Very Satisfied |
| <6,000 | 20 | 24 | 80 | 82 |
| 6,000~15,000 | 22 | 38 | 104 | 125 |
| 15,000~25,000 | 13 | 28 | 81 | 113 |
| >25,000 | 7 | 18 | 54 | 92 |

```
data one;
    do income=1 to 4;
    do job=1 to 4;
    input nij @@;output;
    end;
    end;
    cards;
20  24  80  82
22  38  104 125
13  28  81  113
7   18  54  92
;
run;
proc freq data=one;
    weight nij;
    table income*job/chisq expect;
run;
```

실제 분석에서는 다음과 같이 프로그램을 하는 것이 좋다. nocol 옵션은 열 퍼센트를 출력하지 말라는 것이고 nopercent 는 백분율을 출력하지 말라는 것이다.

```
proc freq data=one;
    weight nij;
    table income*job/chisq nocol nopercent;
run;
```

$$\chi^2 = \frac{(20-14.175)^2}{14.175} + \frac{(24-24.693)^2}{24.693} + \ldots + \frac{(92-78.193)^2}{78.193} = 11.98$$

| income | job | | | | |
|---|---|---|---|---|---|
| 기대빈도 / 백분율 / 행 백분율 | 1 | 2 | 3 | 4 | 총합 |
| 1 | 20 | 24 | 80 | 82 | 206 |
| | 14.175 | 24.693 | 72.935 | 94.198 | |
| | 2.22 | 2.66 | 8.88 | 9.10 | 22.86 |
| | 9.71 | 11.65 | 38.83 | 39.81 | |
| | 32.26 | 22.22 | 25.08 | 19.90 | |
| 2 | 22 | 38 | 104 | 125 | 289 |
| | 19.887 | 34.642 | 102.32 | 132.15 | |
| | 2.44 | 4.22 | 11.54 | 13.87 | 32.08 |
| | 7.61 | 13.15 | 35.99 | 43.25 | |
| | 35.48 | 35.19 | 32.60 | 30.34 | |
| 3 | 13 | 28 | 81 | 113 | 235 |
| | 16.171 | 28.169 | 83.202 | 107.46 | |
| | 1.44 | 3.11 | 8.99 | 12.54 | 26.08 |
| | 5.53 | 11.91 | 34.47 | 48.09 | |
| | 20.97 | 25.93 | 25.39 | 27.43 | |
| 4 | 7 | 18 | 54 | 92 | 171 |
| | 11.767 | 20.497 | 60.543 | 78.193 | |
| | 0.78 | 2.00 | 5.99 | 10.21 | 18.98 |
| | 4.09 | 10.53 | 31.58 | 53.80 | |
| | 11.29 | 16.67 | 16.93 | 22.33 | |
| 총합 | 62 | 108 | 319 | 412 | 901 |
| | 6.88 | 11.99 | 35.41 | 45.73 | 100.00 |

income * job 테이블에 대한 통계량

| 통계량 | 자유도 | 값 | 확률값 | |
|---|---|---|---|---|
| 카이제곱 | 9 | 11.9886 | 0.2140 | $\chi^2$ |
| 우도비 카이제곱 | 9 | 12.0369 | 0.2112 | |
| Mantel-Haenszel 카이제곱 | 1 | 9.5455 | 0.0020 | $G^2$ |
| 파이 계수 | | 0.1154 | | |
| 분할 계수 | | 0.1146 | | |
| 크래머의 V | | 0.0666 | | |

표본 크기 = 901

**표본의 크다? (대표본: large sample)**

일반적으로 각 셀의 기대치가 5 이상이면 대표본으로 간주하여 $\chi^2$-분포를 따른다고 한다. R. A. Fisher 에 의하면 5 미만이 셀의 개수가 전체 셀의 개수의 20%을 넘지 않으면 근사 통계량으로 $\chi^2$-분포를 사용해도 된다고 했다. 그래서 SAS 출력 결과에 항상 5 미만인 셀의 비율이 나타난다.

```
data one;
    do income=1 to 4;
    do job=1 to 4;
    input nij @@;output;
    end;
    end;
    cards;
20  24  3   2
22  38  1   5
13  28  1   3
7   18  4   92
;
run;
proc freq data=one;
    weight nij;
    table income*job/chisq nocol nopercent;
run;
```

```
경고: 셀들의 25%가 5보다 작은 기대도수를 가지고 있습니다.
      카이제곱 검정은 올바르지 않을 수 있습니다.

표본 크기 = 281
```

만약 대표본 조건을 만족하지 못하면...

1) 총 응답자의 수를 늘려라. (n 을 크게 하면)

2) 열이나 행의 인접 범주 2~3 를 합쳐 열이나 행의 범주 수를 줄여라. 예를 들어 페이지 17 예제를 보면 (Moderate Satisfied + Very Satisfied)=Satisfied 로 합치거나 (15,000~25,000)+25,000 이상=15,000 이상으로 합쳐 셀의 수를 줄일 수 있다. 반드시 인접 범주를 합치면 합치는 범주가 새로운 개념을 나타낼 수 있어야 한다. 직업이 범주인 경우 (공무원, 전문직, 무직, 자영업, 회사원)인 경우는 두 개의 범주를 합쳐 하나로 만들기에는 다소 어려운 점이 있다.

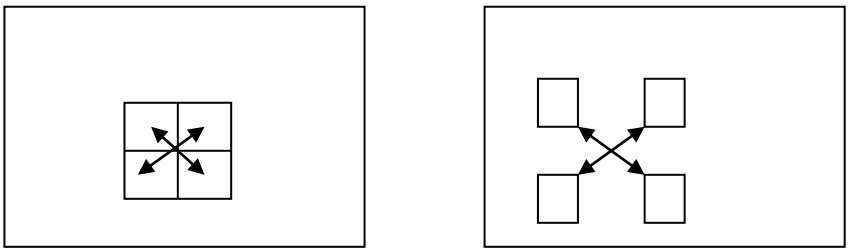
3) Fisher 의 Exact 를 검정 방법을 사용한다. (추후 논의)

**HOMEWORK #4-4**

Homework3-1 에 대해 독립성 검정(남편의 만족도와 아내의 만족도)인 $\chi^2$-검정 하시오. 일단 셀의 기대 빈도 크기에 대한 경고(warning)를 무시하고 분석한 결과를 해석하시오. 그리고 셀의 기대 빈도 5 이하인 셀이 많이 나오면 범주를 합쳐 분석하고 결과를 해석하시오.

## 2.3. Summary Measures of Association

### 2.3.1. Odds Ratio for IxJ 분할표



일반적으로 IxJ 분할표의 Odds Ratio 들을 정보의 희생 (loss of information) 없이는 하나의 값으로 표현할 수 없다. 만약 왼쪽 그림과 같이 인접한 열과 행들의 Odds ratio 들을 구하는 경우는 그 값의 크기가 유사하면 weighted average 에 의해 대표 값을 구할 수 있다. 이 부분에 대해서는 나중에 다루기로 한다.

### 2.3.2. Measure of Ordinal Association

종속 변수와 설명 변수 모두가 범주형이되 순서형인 경우 “설명변수 X 가 증가함에 따라 Y 가 증가한다고 할 수 있나?” 모두 측정형 변수인 경우는 상관 분석이나 회귀분석을 이용하여 이 질문에 대답할 수 있다. 그러나 순서형 범주형은 엄밀히 말하면 metric 이 아니므로 동일한 분석은 불가능하다. 대신 순서형이 “monotonicity”인(X, Y 가 크기 순으로 정렬할 수 있음) 성질을 이용하여 유사한 개념 “높은 반응 변수를 갖는 개체(subject)는 무엇인가?” 분석한다. 두 변수(X, Y)의 각 수준에서 두 개체의 짝을 순서화 하는 경우 만약 X 의 값(수준)이 클 때 Y 의 값이 큰 개체들의 짝을 concordant, X 의 값이 클 때 Y 값이 작은 개체들의 짝을 discordant, X 와 Y 에서 동일한 분류를 갖는 개체의 짝을 tied 라고 한다. concordant 한 짝이 많으면 X 가 증가하면 Y 가 증가한다고 볼 수 있을 것이다. 반대로 discordant 한 짝이 많으면 X 증가에 따라 Y 는 감소한다고 결론 지을 수 있다. 물론 tied 가 많은 경우는 X 와 Y 의 관계(association)가 없다고 할 수 있다.

**예제**

년 소득에 따른 직업 만족도의 차이는 있는가를 알아보기 위한 조사 결과 다음을 얻었다고 하자. [General Social Survey, Norusis, 1988]

| | 직업 만족도 | | | |
|---|---|---|---|---|
| | Very Dissatisfied | Little Dissatisfied | Moderately Satisfied | Very Satisfied |
| <6,000 | 20 | 24 | 80 | 82 |
| 6,000~15,000 | 22 | 38 | 104 | 125 |
| 15,000~25,000 | 13 | 28 | 81 | 113 |
| >25,000 | 7 | 18 | 54 | 92 |

소득이 범주형 변수로 분류되어 있고 직업 만족도는 Likert 척도로 조사되어 있으므로 둘 다 순서형 범주형 자료이다. 만약 소득을 분류하지 않고 금액으로 조사된 자료를 이용하여 두 변수의 관계를 분석하려면 Logistic 방법이다. [나중에 다루기로 한다]

(<6,VD) 셀의 개체와 (6-15,LD)의 개체 짝은 concordant 하다. 즉 (<6,VD)에서 20 개 개체, (6-15,LD)에서 38 개 개체의 쌍 760 개는 모두 concordant 하다. 그러므로 (<6,VD) 셀의 개체는 파랑 셀 부분의 개체들과 짝을 이루면 그 짝은 모두 concordant 이다.

같은 방법으로 Concordant 짝의 수를 계산하면

$$C = 20(38+104+125+28+81+113+18+54+92)+24(104+125+81+113+54+92)$$
$$+80(125+113+92)+22(28+81+113+18+54+92)+38(81+113+54+92)$$
$$+104(113+92)+13(18+54+92)+28(54+92)+81*92 = 109{,}520$$

같은 방법으로 Discordant 짝의 수를 계산하면

$$D = 24(22+13+7)+80(22+38+13+28+7+18)+...+113(7+18+54) = 84{,}915$$

Intuitively, C>D 이면 X 값의 증가함에 따라 Y 값이 증가한다고 할 수 있다. 즉 X 는 Y 에 영향에 양의 영향을 미친다. (소득 수준이 높아질수록 직업의 만족도는 높다.)

어떻게 검정할 것인가? 다음을 고려해보자.

$$\gamma = \frac{\Pi_c - \Pi d}{\Pi_c + \Pi d}$$ where $\Pi_c$=모집단 concordant, $\Pi_d$=모집단 discordant

$-1 \le \gamma \le 1$의 값을 갖고 Gamma 의 추정치는 $\hat{\gamma} = \frac{C-D}{C+D}$ 이다.

만약 $\gamma = 1$이면 완전한 선형 관계이고 X 가 증가하면 Y 도 증가한다.

만약 $\gamma = -1$이면 완전한 선형 관계이고 X 가 증가하면 Y 는 감소한다.

만약 $\gamma = 0$ 이면 X 와 Y 는 관계가 없다.

위의 예제에서 $\hat{\gamma} = 0.127$ 약한 양의 상관 관계가 존재한다. (자세한 내용은 추후 토론)

2x2 분할표의 경우 $\gamma$는 다음과 같이 Yule's Q 로(이는 벨기에 통계학자 Quetelet 을 기리기 위하여) 단순화 된다.

$$Q = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{11}\pi_{22} + \pi_{12}\pi_{21}}$$

HOMEWORK#3 흡연에 따른 폐암 발생 여부를 조사하기 위해 폐암 환자와 그렇지 않은 환자의 흡연 정도를 조사한 자료이다. [Doll & Hill 1988]

| 일인 흡연량 \ 폐암 | 폐암 | 일반 환자 |
|---|---|---|
| None | 7 | 61 |
| <5 | 55 | 429 |
| 5-14 | 489 | 570 |
| 15-24 | 475 | 431 |
| 25-49 | 293 | 154 |
| 50+ | 38 | 12 |

1) 반응변수와 설명변수가 무엇인지 밝히시오.
2) 일일 담배량 5 미만과 5 이상으로 재그룹 하여 2x2 분할표를 만든 후
   A. 두 변수의 연관성을 비율 차이 검정에 의해 분석하시오.
   B. Relative risk 추정치를 구하고 해석하시오.
   C. Odds ratio 추정치를 구하고 해석하시오.
   D. Odds ratio 를 이용하여 두 변수가 서로 독립인지 검정하시오. [유의수준=5%]
3) Ordinal Association Measure 인 $\gamma$ 를 구하고 해석하시오.

### 2.4.2. Independence Test (계속)

$$T = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i,j} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2 (df = (r-1)(c-1))$$

**2x2 분할표**

Example(계속) 심장 마비 증세에 아스피린의 효과 (페이지 10) 두 변수의 독립성(심장마비 증상에 아스피린 약의 효과가 있는가?)을 알아보는 방법은 3가지이다.

| 치명적 심장마비 / 약 | 발생 | 건강 |
|---|---|---|
| Placebo (위약) | 189 | 10,845 |
| 아스피린 | 104 | 10,933 |

1) 집단의 성공률 차이 검정(두 집단 모비율 차이 검정: 검정통계량 T=5.13, 페이지 11) 모비율 차이에 대한 95% 신뢰구간은 (0.005, 0.011) 의미는?

2) Odds ratio $\hat{\theta}=1.83$ (심장 마비 발생 확률이 83% 높다. 페이지 13) Odds ratio 에 대한 95% 신뢰구간은 (1.44, 2.33)이다. 의미는?

3) $\chi^2$ 검정: 귀무가설: 약 변수와 심장 마비 변수는 서로 독립이다.

DRUG ATTACK

| 행 백분율 / 빈도 | 마비 | 정상 | 총합 |
|---|---|---|---|
| 아스피린 | 104 | 10933 | 11037 |
| | 0.94 | 99.06 | |
| 위약 | 189 | 10845 | 11034 |
| | 1.71 | 98.29 | |
| 총합 | 293 | 21778 | 22071 |

DRUG * ATTACK 테이블에 대한 통계량

| 통계량 | 자유도 | 값 | 확률값 |
|---|---|---|---|
| 카이제곱 | 1 | 25.0139 | <.0001 |
| 우도비 카이제곱 | 1 | 25.3720 | <.0001 |

p-값이 0.001 보다 더 작으므로 귀무가설을 기각하고 두 변수는 관계(association)가 있다고 말한다. 귀무가설이 기각되면 행 퍼센트를 관찰하여 크기나 크기 순서를 따져 해석하면 된다. 2x2 분할표의 경우는 하나의 행 퍼센트만 비교하면 된다. 아스피린 복용자의 마비 비율이 0.94, 위약 복용자는 1.71 이므로 아스피린 복용자는 위약 복용자보다 54.9% (위약 복용자는 81.9% 높다) 밖에 되지 않는다. (relative risk 와 동일하다) Recall: 성공률이 내우 낮을 때는 relative risk 와 odds ratio 는 거의 같다.

```
DATA ONE;
    DO DRUG="위약      ", "아스피린";
        DO ATTACK="마비", "정상";
            INPUT NIJ @@;OUTPUT;
        END;
    END;
    CARDS;
    189 10845
    104 10933
;
RUN;
PROC FREQ DATA=ONE;
    WEIGHT NIJ;
    TABLE DRUG*ATTACK/CHISQ NOCOL NOPERCENT;
RUN;
```

**Residual**

앞 절에서는 2x2 분할표에서는 $\chi^2$ 검정 결과를 relative risk 개념에서 해석하는 방법을 살펴보았으나 이런 해석은 IxJ 분할표에서는 불가능하다. 분할표의 검정 통계량의 값에 가장 영향을 많이 미치는 셀이 어느 것인가 알아보는 통계량으로 잔차 개념을 사용한다. 다음을 셀의 수정 잔차라 (adjusted residual) 하고 귀무가설(두 변수가 독립) 하에서는 표준정규분포를 따른다고 한다. 그러므로 수정 잔차의 절대값이 2~3 을 벗어나면 그 셀이 귀무가설을 기각하는데 많은 영향을 미쳤다고 할 수 있다.

$$\frac{O_{ij}-E_{ij}}{\sqrt{E_{ij}(1-p_{i+})(1-p_{+j})}} \sim Normal(0,1) \quad \text{under } H_0$$

일반 표준화 잔차(standardized residual)는 분산이 1 보다 작은 문제가 있어 수정 잔차를 사용한다. $e_{ij}=(O_{ij}-E_{ij})/\sqrt{E_{ij}} \sim Normal(0,<1)$

Example 다음은 성별에 따른 정당 지지 성향의 차이가 있는지 알아보기 위하여 총 980 명을 대상으로 실시한 설문 조사 결과를 정리한 것이다.

| | 공화당 | 무소속 | 민주당 | 합계 |
|---|---|---|---|---|
| 남성 | 191 | 47 | 165 | 403 |
| 여성 | 225 | 73 | 279 | 577 |
| 합계 | 416 | 120 | 444 | 980 |

우선 앞 절에서 살펴 보았던 독립성 검정을 위한 $\chi^2$-검정을 실시해 보자.

GENDER PARTY

| 빈도 / 행 백분율 | 공화당 | 무소속 | 민주당 | 총합 |
|---|---|---|---|---|
| 남성 | 191<br>47.39 | 47<br>11.66 | 165<br>40.94 | 403 |
| 여성 | 225<br>38.99 | 73<br>12.65 | 279<br>48.35 | 577 |
| 총합 | 416 | 120 | 444 | 980 |

GENDER * PARTY 테이블에 대한 통계량

| 통계량 | 자유도 | 값 | 확률값 |
|---|---|---|---|
| 카이제곱 | 2 | 7.0095 | 0.0301 |
| 우도비 카이제곱 | 2 | 7.0026 | 0.0302 |

p-값이 0.03 으로 0.05 보다 적으므로 귀무가설(성별 변수와 정당 변수는 서로 독립)이 기각되고 성별에 따라 지지 정당이 달라진다고 결론 내릴 수 있다. 그러면 어떻게 다른가? 이 답변을 위하여 행 퍼센트를 보면 된다. 남성의 경우 공화당 지지율이 가장 높고 여성의 경우 민주당 지지율이 높다고 말하면 된다.

만약 위의 결과를 행 퍼센트 개념이 아니라 수정 잔차 개념에서 해석하여 보자.

(1 행, 1 열)의 수정 잔차 값: $\dfrac{191-171}{\sqrt{171(1-416/980)(1-403/980)}} = 2.62$

(1 행, 2 열)의 수정 잔차 값: $\dfrac{47-49.3}{\sqrt{49.3(1-120/980)(1-403/980)}} = -0.46$

같은 방법으로 구하면...

| | 공화당 | 무소속 | 민주당 |
|---|---|---|---|
| 남성 | 191 | 47 | 165 |
| | (2.62) | (-0.46) | (-2.29) |
| 여성 | 225 | 73 | 279 |
| | (-2.62) | (-0.46) | (2.29) |

공화당과 민주당에서 성별 차이가 나타나고 있다( 수정 잔차가 2 이상의 값) 여성은 민주당 지지자가 많고 남성은 공화당 지지자가 많은 반면 여성은 공화당 지지자가 남성은 민주당 지지자가 적다.

수정 잔차와 행 퍼센트의 해석 방법에는 큰 차이가 없고 수정 잔차는 수작업으로 계산해야 한다는 불편함으로 인하여 주로 행 퍼센트에 의해 결과를 해석하게 된다. 일반적으로 이 방법을 따르는데 때로는 간과되는 부분(민주당 지지의 차이 부분)이 생긴다.

**HOMEWORK #5-1**

다음은 학교 창립 이념에 따른 정신 분열의 기원에 의견의 차이가 있는지 알아보기 위하여 조사한 결과 이다. [Gallagher et al. 1987]

| 창립이념 \ 정신분열기원 | 유전적인 | 환경적인 | 유전+환경 |
|---|---|---|---|
| 다방면 | 90 | 12 | 78 |
| 의학적 | 13 | 1 | 6 |
| 정신분석학적 | 19 | 13 | 50 |

1) 독립성 검정을 위한 $\chi^2$-검정을 실시하시오,

2) 행 퍼센트를 이용하여 결과를 해석하시오.

3) 수정 잔차를 이용하여 결과를 해석하시오.

**Partitioning Chi-square**

IxJ 분할표를 $\chi^2$ 분포의 성질에 따라 분할할 수 있다. 자유도가 각각 $a$ , $b$ 이고 서로 독립인 Chi-square 변수의 합은 자유도가 $(a+b)$인 Chi-square 를 따른다.

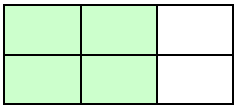
왜 하는가? 분할표를 몇 개의 sub 분할표로 나누어 분석함으로써 관심이 있는 범주들간의 차이 혹은 범주들간 그룹 간의 차이를 볼 수 있다.

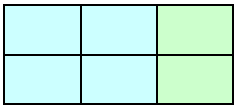
IxJ 분할표를 다음 논리에 의해 분할하면 각 분할표의 $G^2$ (L-R Chi-square) 통계량의 합은 IxJ 분할표의 $G^2$ 통계량의 합과 같다. (*: Pearson 의 Chi-square 에서는 성립하지 않으나 Pearson 의 chi-square 통계량과 L-R chi-square 통계량은 근사적으로 equivalent 하므로 상관없다. 페이지 27)

- sub 분할표의 자유도의 합은 IxJ 분할표의 자유도와 같다.
- IxJ 분할표 각 셀은 sub 분할표에 단 한 번만 나타난다.
- IxJ 분할표 주변 합은 sub 분할표에 단 한 번만 나타난다.

뭔가 무척이나 복잡하다. 간단한 sub 분할표 만드는 방법을 살펴 보면 다음과 같다.

2x3 분할표





Example 성별에 따른 지지 여당 차이 [페이지 32]

GENDER PARTY

| 빈도 / 행 백분율 | 무소속 | 민주당 | 총합 |
|---|---|---|---|
| 남성 | 47<br>22.17 | 165<br>77.83 | 212 |
| 여성 | 73<br>20.74 | 279<br>79.26 | 352 |
| 총합 | 120 | 444 | 564 |

성별에 따른 무소속 vs 민주당 지지의 차이는 없다.

GENDER * PARTY 테이블에 대한 통계량

| 통계량 | 자유도 | 값 | 확률값 |
|---|---|---|---|
| 카이제곱 | 1 | 0.1618 | 0.6875 |
| 우도비 카이제곱 | 1 | 0.1612 | 0.6881 |

GENDER * PARTY 교차표

GENDER PARTY

| 빈도 / 행 백분율 | 공화당 | 민주당+? ぜ宥? | 총합 |
|---|---|---|---|
| 남성 | 191<br>47.39 | 212<br>52.61 | 403 |
| 여성 | 225<br>38.99 | 352<br>61.01 | 577 |
| 총합 | 416 | 564 | 980 |

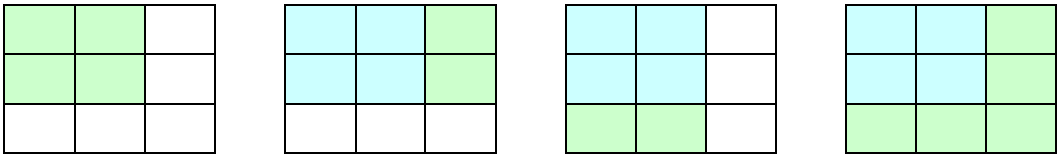
성별에 따른 공화당 vs 비공화당의 차이는 존재하고 남자가 공화당 지지 비율이 높다.

=7.0026 (페이지 33)

GENDER * PARTY 테이블에 대한 통계량

| 통계량 | 자유도 | 값 | 확률값 |
|---|---|---|---|
| 카이제곱 | 1 | 6.8528 | 0.0089 |
| 우도비 카이제곱 | 1 | 6.8414 | 0.0089 |

3x3 분할표



**HOMEWORK #5-2**

Homework #5-1 의 3x3 분할표를 위와 같이 분할하고 각 sub 분할표에 대해 $\chi^2$ 검정을 실시하고 해석하시오. [Gallagher et al. 1987]

### 2.4.3. More on Linear Association

변수 X, Y 가 모두 순서형 변수이면 그 변수들간 선형 관계를 (linear association) 이용하여 두 변수의 독립성(independence ←→관계) 분석할 수 있다.

행 변수 (독립변수, X)의 범주를 크기 순으로 정렬 하고 각 범주를 $u_1 \le u_2 \le u_3 \ldots$로 점수화 하고 열 변수 (반응변수, Y) 범주를 크기 순으로 정렬한 후 각 범주를 $v_1 \le v_2 \le v_3 \ldots$ 를 점수화 하자. 두 범주의 개념 차이가 크기가 크면 점수의 차이를 크도록 점수화 한다. 이 점수를 이용하여 두 변수간의 가중 상관계수를 구하면 다음과 같다. 이를 Pearson cross moment correlation coefficient 라 한다.

$$r = \frac{\sum_{i,j} u_i v_j n_{ij} - (\sum_i u_i n_{i+})(\sum_j v_j n_{+j})/n}{\sqrt{[\sum_i u_i^2 n_{i+} - \frac{(\sum_i u_i n_{i+})^2}{n}][\sum_j v_j^2 n_{+j} - \frac{(\sum_j v_j n_{+j})^2}{n}]}}$$

여기서 $n_{ij}$ 는 관측 빈도 $O_{ij}$ , $n$ 은 표본의 총 크기를 의미한다. 물론 우리는 손으로 이것을 계산할 필요는 없다. 상관 계수의 값은 –1 과 1 사이의 값을 갖고 0 이면 두 변수는 상관 관계가 없다(독립성)고 결론 내린다.

Pearson 상관 계수 이외에도 두 순서형 변수의 상관 관계에 대한 계산 값은 여러 개 존재하는데 이는 대부분 Pearson 상관 계수로부터 유도되었다. Phi Coefficient(파이 계수), Cramer's V, 분할 계수가 그 예이다. 상관 계수의 유의성은 검정은 Mantel-Haenszel Chi-Square (M-H) 검정 통계량이라 불리는 $M^2$에 의해 실시한다.

$$M^2 = (n-1)r^2 \sim \chi^2(df = 1) \quad \text{when n is large.}$$

Example 년 소득에 따른 직업 만족도의 차이는 있는가 [예제 계속]

| | 직업 만족도 | | | |
|---|---|---|---|---|
| | Very Dissatisfied | Little Dissatisfied | Moderately Satisfied | Very Satisfied |
| <6,000 | 20 | 24 | 80 | 82 |
| 6,000~15,000 | 22 | 38 | 104 | 125 |
| 15,000~25,000 | 13 | 28 | 81 | 113 |
| >25,000 | 7 | 18 | 54 | 92 |

```
DATA ONE;
       DO INCOME=1 TO 4;
       DO JOB=1 TO 4;
              INPUT NIJ@@;OUTPUT;
       END;
       END;
```

```
CARDS;
20      24      80      82
22      38      104     125
13      28      81      113
7       18      54      92
;
RUN;
PROC FREQ DATA=ONE;
       WEIGHT NIJ;
       TABLE INCOME*JOB /CHISQ MEASURES CL NOPERCENT NOCOL;
RUN;
```

INCOME JOB

| 빈도 / 행 백분율 | 1 | 2 | 3 | 4 | 총합 |
|---|---|---|---|---|---|
| 1 | 20<br>9.71 | 24<br>11.65 | 80<br>38.83 | 82<br>39.81 | 206 |
| 2 | 22<br>7.61 | 38<br>13.15 | 104<br>35.99 | 125<br>43.25 | 289 |
| 3 | 13<br>5.53 | 28<br>11.91 | 81<br>34.47 | 113<br>48.09 | 235 |
| 4 | 7<br>4.09 | 18<br>10.53 | 54<br>31.58 | 92<br>53.80 | 171 |
| 총합 | 62 | 108 | 319 | 412 | 901 |

INCOME * JOB 테이블에 대한 통계량

| 통계량 | 자유도 | 값 | 확률값 |
|---|---|---|---|
| 카이제곱 | 9 | 11.9886 | 0.2140 |
| 우도비 카이제곱 | 9 | 12.0369 | 0.2112 |
| Mantel-Haenszel 카이제곱 | 1 | 9.5455 | 0.0020 |
| 파이 계수 | | 0.1154 | |
| 분할 계수 | | 0.1146 | |
| 크래머의 V | | 0.0666 | |

- 두 변수의 상관 관계 검정은 Measures 옵션에 의해 출력된 아래 결과 중 Pearson 상관계수, Spearman 상관 계수를 이용하면 된다. 양의 상관 관계가 존재하므로 (신뢰한계(구간)가 0 을 포함하고 있지 않고 값이 양의 값이다) 소득이 높아질수록 직업 만족도가 높아짐을 알 수 있다.
- 두 변수의 독립성 검정을 위한 $\chi^2$ 검정 결과는 유의하지 않았으므로 Chi-square 검정 결과는 소득 수준은 직업 만족도에 영향을 미치지 않는다.
- 독립성을 검정을 위한 $\chi^2$ 검정 결과의 해석은 행 퍼센트로 하고 상관 관계를 위한 검정 결과는 상관 계수 해석 방법과 동일하게 한다. 독립성 검정을 분할표 검정 모두에 적용되지만, 상관 관계 분석은 두 변수 모두 순서형일 때 적용 가능하다.

| 통계량 | 값 | 점근표준오차 | 95% 신뢰한계 | |
|---|---|---|---|---|
| 감마 | 0.5313 | 0.0935 | 0.3480 | 0.7146 |
| Kendall의 타우-b | 0.3373 | 0.0642 | 0.2114 | 0.4631 |
| Stuart 타우-c | 0.4111 | 0.0798 | 0.2547 | 0.5675 |
| Somers D C\|R | 0.2569 | 0.0499 | 0.1592 | 0.3547 |
| Somers D R\|C | 0.4427 | 0.0837 | 0.2786 | 0.6068 |
| Pearson 상관계수 | 0.3776 | 0.0714 | 0.2378 | 0.5175 |
| Spearman 상관계수 | 0.3771 | 0.0718 | 0.2363 | 0.5178 |

빨간 박스 안의 통계량들은 페이지 17 의 $\gamma$ (Gamma)과 유사한 것으로 두 순서형 변수간의 선형 관계를 검정하는 통계량이다. (Concordant, Discordant) 페이지 17 의 Gamma 값(0.127)과 위의 감마가 일치하지 않는 것은 조금 다른 계산 공식을 사용하였기 때문이다. 수작업 할 필요는 없으니 SAS 출력 결과를 이용하여 해석하면 된다.

In SAS, $C = \sum_i \sum_j n_{ij} \left( \sum_{k>il>j} \sum n_{kl} + \sum_{k<il<j} \sum n_{kl} \right)$, $D = \sum_i \sum_j n_{ij} \left( \sum_{k<il>j} \sum n_{kl} + \sum_{k>il<j} \sum n_{kl} \right)$

**점수를 바꾸면**

순서형 변수에 점수를 부여하는 것은 다소 임의적이나 등간 (equal-distance) 점수이면 (이를 monotonic score 라 함) 검정 통계량의 값은 변하지 않으므로 별 문제는 없다. 물론 각 범주의 점수를 등간으로 하지 않으면 변하지만...

INCOME JOB

| 빈도\| 행 백분율\| | 10\| | 20\| | 30\| | 40\| | 총합 |
|---|---|---|---|---|---|---|
| 10 | 20 | 24 | 80 | 82 | 206 |
| | 9.71 | 11.65 | 38.83 | 39.81 | |

INCOME * JOB 테이블에 대한 통계량

| 통계량 | 자유도 | 값 | 확률값 |
|---|---|---|---|
| 카이제곱 | 9 | 11.9886 | 0.2140 |
| 우도비 카이제곱 | 9 | 12.0369 | 0.2112 |
| Mantel-Haenszel 카이제곱 | 1 | 9.5455 | 0.0020 |
| 파이 계수 | | 0.1154 | |
| 분할 계수 | | 0.1146 | |
| 크래머의 V | | 0.0666 | |

**Cochran Armitage Trend**

반응 변수가 이진 (binary : 예 성공/확률) 변수이고 설명 변수가 순서형인 경우 설명 변수 범주의 크기에 따라 반응변수의 비율이 어떻게 변하는지 알아보는 것이다. 검정 통계량은 다음과 같다.

$$T = \frac{\sum_{i=1}^{R} n_{i1}\left(R_i - \overline{R}\right)}{\sqrt{p_{\cdot 1}\left(1-p_{\cdot 1}\right)s^2}} \quad \text{where} \quad s^2 = \sum_{i=1}^{R} n_{i\cdot}\left(R_i - \overline{R}\right)^2$$

Row scores :
$$R1_i = \sum_{k<i} n_{k\cdot} + \left(n_{i\cdot}+1\right)/2 \quad i = 1, 2, \ldots, R$$

Column scores :
$$C1_j = \sum_{l<j} n_{\cdot l} + \left(n_{\cdot j}+1\right)/2 \quad j = 1, 2, \ldots, C$$

```
DATA PAIN;
   INPUT DOSE ADVERSE $ COUNT @@;
   CARDS;
0 NO 26 0 YES  6
1 NO 26 1 YES  7
2 NO 23 2 YES  9
3 NO 18 3 YES 14
4 NO  9 4 YES 23
;

PROC FREQ DATA=PAIN;
    WEIGHT COUNT;
    TABLES DOSE*ADVERSE /TREND CHISQ NOPERCENT NOCOL;
RUN;
```

| 빈도 \| 행 백분율 | No | Yes | 총합 |
|---|---|---|---|
| 0 | 26<br>81.25 | 6<br>18.75 | 32 |
| 1 | 26<br>78.79 | 7<br>21.21 | 33 |
| 2 | 23<br>71.88 | 9<br>28.13 | 32 |
| 3 | 18<br>56.25 | 14<br>43.75 | 32 |
| 4 | 9<br>28.13 | 23<br>71.88 | 32 |
| 총합 | 102 | 59 | 161 |

약의 복용량(0, 1, 2, 3, 4)에 따른 부작용(No:없음/Yes: 있음)의 차이를 살펴본 것이다.

Dose * Adverse 테이블에 대한 통계량

| 통계량 | 자유도 | 값 | 확률값 |
|---|---|---|---|
| 카이제곱 | 4 | 26.6025 | <.0001 |
| 우도비 카이제곱 | 4 | 26.6689 | <.0001 |
| Mantel-Haenszel 카이제곱 | 1 | 22.8188 | <.0001 |
| 파이 계수 | | 0.4065 | |
| 분할 계수 | | 0.3766 | |
| 크래머의 V | | 0.4065 | |

Cochran-Armitage 추세반영 검정

| | |
|---|---|
| 통계량 (Z) | -4.7918 |
| 단측 Pr < Z | <.0001 |
| 양측 Pr > \|Z\| | <.0001 |

비율의 추세(trend) 직선 dl 유의하고 음의 통계량 값을 가지므로 부작용 비율(No/Yes)은 낮아지는 경향이 있다.

**HOMEWORK #5-3**

Homework #-3-1 자료를 이용하여 다음에 답하시오. 셀의 기대 빈도에 대한 경고는 무시하고 다음을 실시하시오.

- 남편 만족도와 아내 만족도는 서로 독립인가? 독립성 $\chi^2$-검정: 이미 숙제 했음
- 설명 변수(남편 만족도), 반응변수(아내 만족도)에 대한 상관 관계 분석을 실시하고 해석하시오.
- 아내 만족도를 두 범주(Never vs 나머지 3 범주)로 분할표를 다시 작성하고 Cochran-Armitage trend 분석을 실시하시오.

**HOMEWORK #5-3**

Homework #-3-1 자료를 이용하여 다음에 답하시오. 셀의 기대 빈도에 대한 경고는 무시하고 다음을 실시하시오.

- 남편 만족도와 아내 만족도는 서로 독립인가? 독립성 $\chi^2$-검정: 이미 숙제 했음
- 설명 변수(남편 만족도), 반응변수(아내 만족도)에 대한 상관 관계 분석을 실시하고 해석하시오.
- 아내 만족도를 두 범주(Never vs 나머지 3 범주)로 분할표를 다시 작성하고 Cochran-Armitage trend 분석을 실시하시오.

### 2.4.4. Exact Test

R. A. Fisher "전통적인 분석 도구가 실제적인 연구에 항상 적합한 것은 아니다. 대포로 참새를 잡으려고 할 뿐 아니라 명중시키지 못하기도 한다. 대표본 자료 분석 방법은 간단한 실험 자료에 적합하지 않다."

지금까지 살펴본 $\chi^2$-검정 통계량은 근사 통계량이었다. 셀의 기대 빈도가 5 미만인 셀이 (Thin cell) 없다면 (혹은 R. A. Fisher 는 5 미만 셀의 수가 전체 셀의 20% 넘지 않으면)

검정통계량 $\sum_i \sum_j \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ 은 ($O_{ij}$ 는 $n_{ij}$ 와 같다) $\chi^2(df=(r-1)(c-1))$ 에 근사한다.

그러나 만약 표본의 크기가 적거나 (소표본) thin cell 의 조건이 만족되지 않으면? (해결 방법은 페이지 30 참고: 1)셀 합치기 2)표본 늘리기) 우리는 더 이상 $\chi^2$-분포를 사용하지 못한다. 그래서 R. A. Fisher 는 Exact (더 이상 근사 분포가 아니다) Test 를 제안하였다.

**Fisher's exact test**

2x2 분할표를 먼저 생각해 보자.

| X \ Y | 1 | 2 | Total |
|---|---|---|---|
| 1 | $n_{11}$ | $n_{12}$ | $n_{1+}$ |
| 2 | $n_{21}$ | $n_{22}$ | $n_{2+}$ |
| Total | $n_{+1}$ | $n_{+2}$ | $n$ |

두 변수가 독립이면 주변 빈도 (marginal frequency)의 조건부 확률로부터 구할 수 있다. 다음 초기하 분포를 (hyper-geometric distribution) 생각해보자. 두 변수가 독립일 때(odds ratio $\theta$=1) (1,1) 셀의 빈도가 $n_{11}$ 일 확률밀도함수는

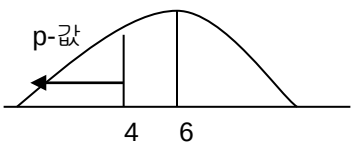
$$\frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}}$$

총 표본 $n$로부터 변수 Y 의 1 범주 주변 빈도 합 $n_{+1}$만큼 뽑을 때 변수 X 의 범주 1 에서 $n_{11}$명을 뽑고 변수 X 의 범주 2 에서 $(n_{+1}-n_{11})=n_{21}$명을 뽑을 확률이다. 주변 빈도 합이 주어지면(총 4 개) $n_{11}$의 값만 주어져도 다른 3 셀의 빈도를 계산할 수 있다.

독립성 검정을 위하여 p-값 개념을 이용하자. 표본으로부터 얻어진 결과가 나올 확률과 그 이상(혹은 이하, 귀무가설의 모수 위치에 따라 결정)의 결과들이 나올 확률을 p-값이라 한다. 다음 예를 들어보자.

Example A 선수는 승률이 0.6이라고 주장한다. 이 주장의 진실여부를 알아보기 위하여 10게임을 조사하였더니 A가 4번 이겼다. A의 주장은 사실인가?

- 귀무가설: A 의 승률은 0.6 이다. $p=0.6$
- 대립가설: $p>0.6$ (반대의 경우는 결코 발생할 수 없다. 왜냐하면 $\hat{p}=0.4$ 이므로)
- 검정통계량: 대표본일 경우 우리는 $\hat{p}\sim$ z-분포에 근사한다는 사실을 이용하여 가설검정할 수 있으나 n=10 개인 소표본인 경우에는 이 근사 통계량을 사용할 수 없다.
- 대신 p-값을 구해보자. $p-value=\Pr(X\le 4 \mid X\sim Binomial(10,0.6))$ : 귀무가설 하에서 측정된 4 번 이하 이길 확률을 p-값이라 한다.




- $p-value=\Pr(X\le 4 \mid X\sim B(10,0.6))=0.16624$ In SAS, p=PROBBNML(0.6,10,4)

- p-값이 유의수준 0.05 보다 크므로 귀무가설을 기각하지 못한다. A 의 승률이 0.6 이 아니라고 말할 근거가 없다.

이제 분할표 독립성 검정을 위한 검정통계량으로 돌아가자. Fisher's Tea Drinker 예제를 중심으로 검정 방법을 살펴보자.

Example 영국 여성은 차를 마실 때 그 차에 우유를 먼저 부었는지 차를 먼저 부었는지 알 수 있다고 주장하였다. 이에 R. A. Fisher는 이 주장에 대한 검정을 위하여 여성 8명을 대상으로 조사를 실시하여 다음 결과를 얻었다.

| 실제 \ 추측 | 우유 | 차 | Total |
|---|---|---|---|
| 우유 | 3 | 1 | 4 |
| 차 | 1 | 3 | 4 |
| Total | 4 | 4 | 8 |

- 귀무가설: 두 변수는 서로 독립이다. 실제와 추측간에는 상관이 없다. 즉 여성의 추측은 실제 차의 상태와 관계가 없다. ($\theta = 1$)
- 대립가설: 두 변수는 관계가 있다. 즉 실제 차의 상태와 예측 간에는 상관 관계가 존재한다. ($\theta > 1$)
- 검정통계량: 소표본이므로 분할표 독립성 검정을 위한 $\chi^2$ -검정 방법을 사용할 수 없다.
- p-값: $$p-value = \Pr(n_{11} \geq 3 \mid n_{11} \sim HG(8,4,4,n_{11})) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} + \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = 0.229 + 0.014 = 0.243$$

유의수준 0.05 보다 크므로 귀무가설을 기각하지 못한다. 즉 영국 여성은 우유나 차 중 어느 것을 먼저 따랐는지 맞출 수 있다는 주장을 지지할 수 없다.

```
DATA ONE;
        INPUT TEA $ GUESS $ NIJ @@;
        CARDS;
        M M 3 M T 1 T M 1 T T 3
;
RUN;
PROC FREQ DATA=ONE;
        WEIGHT NIJ;
        TABLE TEA*GUESS/EXACT NOCOL NOPERCENT;
RUN;
```

경고: 셀들의 100%가 5보다 작은 기대도수를 가지고 있습니다.
카이제곱 검정은 올바르지 않을 수 있습니다.

Fisher의 정확 검정

| | |
|---|---|
| (1,1) 셀 빈도(F) | 3 |
| 하단측 p값 Pr <= F | 0.9857 |
| 상단측 p값 Pr >= F | 0.2429 |
| 테이블 확률 (P) | 0.2286 |
| 양측 p값 Pr <= P | 0.4857 |

2 배

**Comments**

1) Randomization Test: 유의수준 0.05 에서 기각역은 어떻게 설정되어야 하는가? 이 질문은 (1,1) 셀의 빈도가 얼마 이상이 나오면 귀무가설을 기각해야 하는가와 같다. (1,1) 셀이 4 가 나올 확률은 0.014 이다. 그러므로 4 가 나오면 기각한다고 하면 아직 유의 수준 0.05 가 안된다. 만약 (1,1) 셀이 3 이상(3 과 4)이 나오면 귀무가설을 기각한다고 하면 0.243 으로 0.05 를 넘는다. 그러면 어쩌라. 이때 필요한 개념이 Randomization 이다. (1,1)셀이 4 가 나오면 귀무가설을 무조건 기각하고 3 이 나오면 0.157 의 확률로 귀무가설을 기각하면 된다. 0.157 확률을 가진다? 난수표를 이용하여 3 자리 임의의 수를 뽑는다. 이 값이 0.157 보다 작으면 귀무가설을 기각하면 된다.

$$\Pr(reject\ H_0) = E(\Pr(reject\ H_0)\,|\,n_{11}) = 0.014 + 0.157 \times 0.229 = 0.05$$

2) 만약 (1,1) 셀의 크기가 (2,1)의 셀의 크기보다 적으면 p-값을 계산할 때 (1,1) 셀의 빈도 이하인 경우 확률을 다 더하면 된다. 즉 대립 가설은 $\theta < 1$ 이다. 위의 SAS 출력에서 하단측 p-값에 해당. $p-value = \Pr(n_{11} \le 3\,|\,n_{11} \sim HG(8,4,4,n_{11})) = 0.9857$

3) Exact Test 를 IxJ 분할표에도 확대되었다. [자세한 내용은 생략] SAS 에서 Exact 옵션을 쓰면 IxJ 분할표에 대한 Exact Test 검정 결과를 출력한다. 출력 결과는 양측 검정 결과만 출력한다. [SAS 는 Mehta and Patel (1983) 의 network algorithm 사용] Mehta, C.R. and Patel, N.R. (1983), "A Network Algorithm for Performing Fisher's Exact Test in r×c Contingency Tables," Journal of the American Statistical Association, 78, 427-434.

**HOMEWORK #6-1**

다음은 치료 종류(일반 치료, 방사선 치료)와 후두암 치료의 관계를 알아보고자 조사한 자료이다. [Mandenhall et.al (1984)]

| 치료방법 \ 암 치료 | Yes | NO |
|---|---|---|
| 일반 | 21 | 2 |
| 방사선 | 15 | 3 |

- 수작업 계산하여 p-값을 구하시오.
- 유의수준 0.05 에서 기각역을 구하시오.
- SAS 를 이용하여 Fisher Exact test 를 실시하고 결과를 해석하시오.

**HOMEWORK #6-2**

Homework #-3-1 자료에서 SAS 를 이용하여 Exact Test 를 실시하고 결과를 해석하시오.