강원대학교
AI 소프트웨어학과

머신러닝1
– 기초통계 –
추론통계(분산의 분포– Chi-Square)

KNU 강원대학교

추론통계

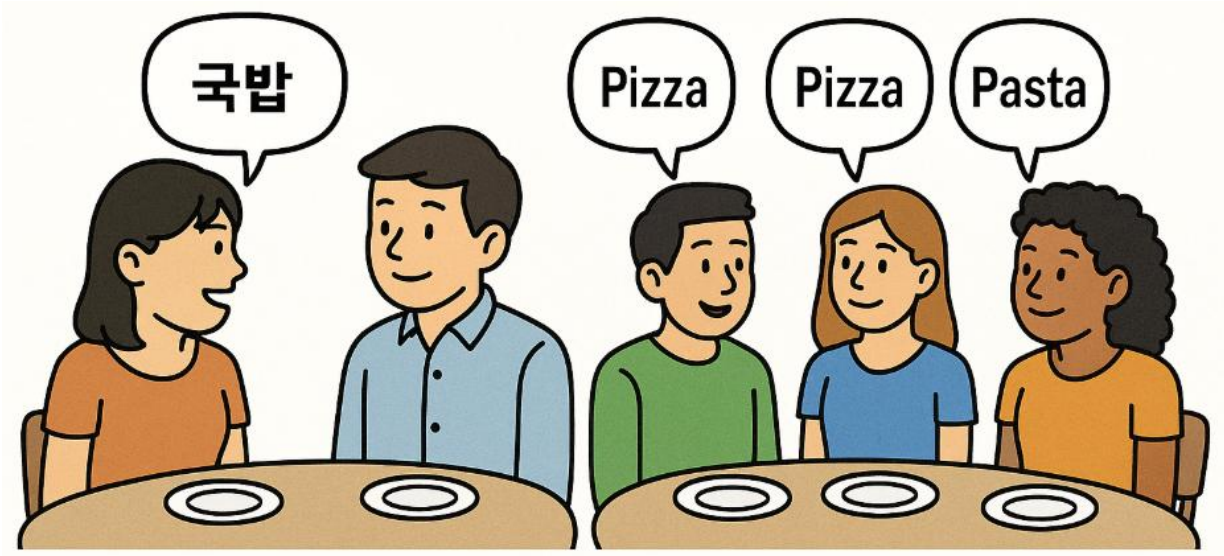
분산의 분포

- 분산 분포는 확률 변수의 분산(또는 표준편차의 제곱)에 대한 통계적 분포를 나타냄 ➔ 변동성
  - 임상 시험: 두 가지 치료법의 효과를 비교하는 경우 어떤 치료법이 더 나은 평균 결과를 나타내는지 뿐만 아니라 어떤 치료법이 더 일관된(더 낮은 분산) 결과를 나타내는지 알고 싶음
  - 교육: 서로 다른 두 가지 교육 간의 시험 점수를 비교할 때 일관된 차이를 이해하면 해당 교육이 학생 전체에 걸쳐 얼마나 잘 작동하는지 나타낼 수 있음
  - 제조: 품질 관리에서는 단순히 높은 평균 품질이 아닌 일관되게 높은 품질의 제품을 원하기 때문에 평균 품질 수준 뿐만 아니라 편차도 아는 것이 중요한 경우가 많음
  - 재무: 포트폴리오 관리에서 평균은 기대 수익을 제공할 수 있지만 분산 또는 표준 편차는 관련 위험에 대한 아이디어를 제공함

추론통계

가설 검정 및 추론통계 (연관관계 파악)

- 카이제곱분포(Chi-Square Distribution) : 카이제곱 검정은 두 범주형 변수 사이에 유의미한 연관성을 확인하는 데 사용
  - 분할표를 분석하고 관찰된 빈도 분포가 예상 빈도 분포와 다른지 여부를 테스트하는 데 자주 사용



추론통계

Chi-Square Test

- Chi-Square Test : 카이제곱 검정은 두 범주형 변수 사이에 유의미한 연관성이 있는지 확인하는 데 사용됨
- 분할표를 분석하고 관찰된 빈도 분포가 예상 빈도 분포와 다른지 여부를 테스트하는 데 자주 사용됨

  → 데이터 간의 빈도 차이를 보기 위해 만들어진 검정
- 이는 명목 또는 순서 측정값이 있는 변수 간의 비율의 차이를 통해 관계를 연구하는 데 특히 유용함

가설 설정 방법

- 귀무가설(H_0) : 두 범주형 변수 사이에 유의미한 관계나 연관이 없다.
- 대립가설(H_1) : 두 범주형 변수 사이에 중요한 관계 또는 연관성이 있다.

추론통계

Chi-Square Test

- 실제로 관측된 빈도와 기대되는 빈도 사이에 차이가 있는지 판단하는 방법
- 표준 정규분포를 따르는 $Z_i \sim N(0,1)$ 평균이 0, 분산이 1인 연속형 확률 변수
- Z통계량에 제곱을 하면 $Z^2 \sim \chi^2(1)$ 자유도 1의 카이제곱 분포를 따름 → 자유도로 모든것이 결정됨
- Z통계량의 값을 제곱한 확률 변수는 카이제곱 분포 → 표준화된(즉, 모수의 영향을 제거한) 분산 추정치의 분포로서, 자유도라는 단일 모수로 완전히 설명되는 분포
- 카이제곱 값이 클수록 → 분산이 큼 → 기대값에서 멀리 퍼짐 → 통계적으로 차이가 존재함

$$Z_k^2 = \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2 \rightarrow Z_1^2 + Z_2^2 + Z_3^2 + \cdots + Z_k^2$$

빈도의 차이(관측 값-기대빈도 값) → 표본이 클 때, 표준화를 통해 정규 분포로 근사 가능 →제곱을 통해 양수 값을 가지게 변환

오차 $= O_i - E_i$ → $\frac{O_i - E_i}{\sqrt{E_i}} \approx Z_i \sim N(0,1)$ → $\chi^2 = \sum\left(\frac{O_i - E_i}{\sqrt{E_i}}\right)^2$

추론통계

Chi-Square Test

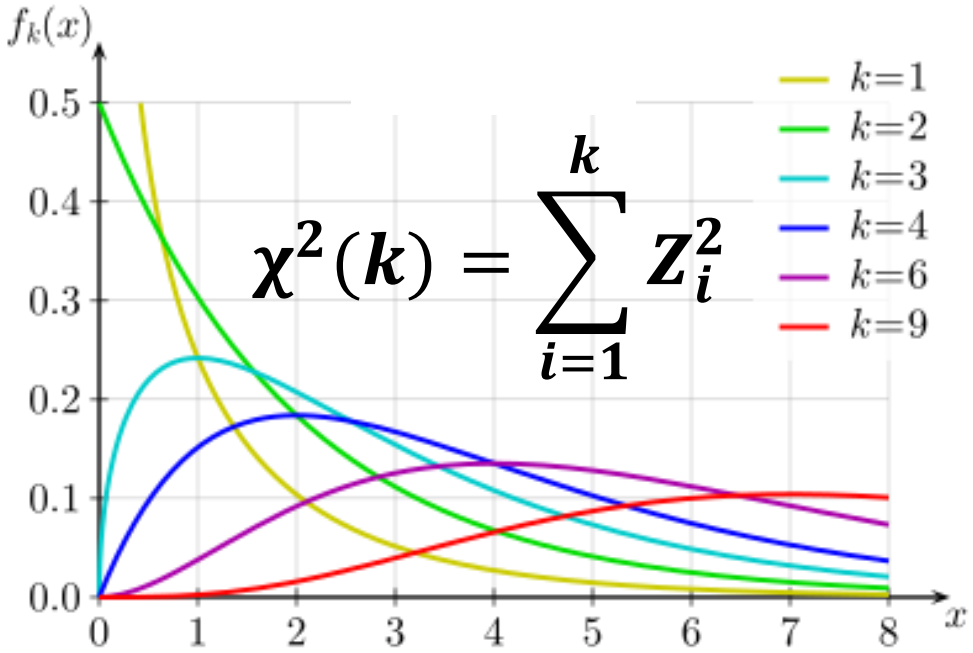
- 범주에서 기대빈도와 관측빈도의 차이가 얼마나 큰지를 보는 것으로 기대빈도 값이 클수록 차이에 덜 민감하게, 작을수록 더 민감하게 반영되도록 설계됨
- 희귀한 사건의 과잉반응 방지 또는 소수 범주 검출 강화

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$$\chi^2 = \sum \left(\frac{O_i - E_i}{\sqrt{E_i}}\right)^2 = \frac{(O_i - E_i)^2}{E_i}$$

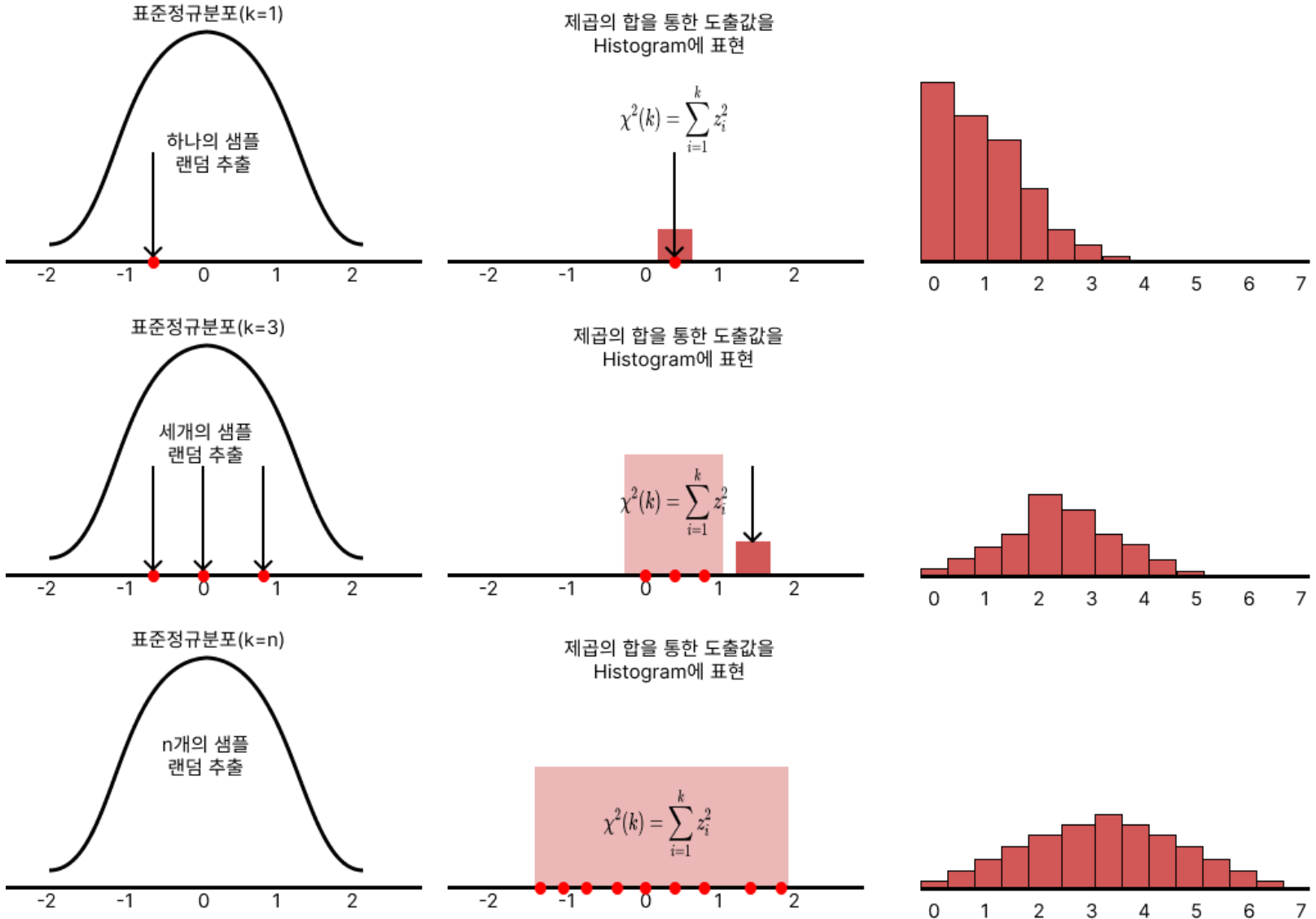
Chi-Square Test

- **Chi-Square Test**
  - 카이제곱 분포는 서로 독립인 정규분포를 따르며 확률변수들을 각각 제곱한 후 합하여 얻어지는 분포
  - 양의 정수 k에 대해 k개의 독립적인 표준정규분포를 따르는 확률변수 $Z_1, \dots Z_k$를 정의하면 자유도 k의 카이제곱 분포는 확률 변수의 분포 → 자유도k인 카이제곱 분포를 따름



추론통계

Chi-Square Test



- **카이제곱 검정은 음수가 나올 수 없으므로 단측(우측)검정이고,**
- **K값이 증가하면 정규분포와 유사한 형태를 가짐**

추론통계

Chi-Square Test

- 카이제곱 분포는 오차 혹은 편차를 분석할 때 사용함
- 카이제곱 분포를 이용해 오차나 편차 검증하면 → 우연히 발생하는 오차인지 숨겨진 의미가 있는 오차나 편차 인지 알 수 있음
- 두 가지 검정
  - 적합도 검정(goodness-of-fit-test) → 기대되는 빈도의 분포와 관찰한 빈도의 분포를 비교
  - 독립성 검정(chi-square independence test) → 범주형 변수가 여러 개인 경우에 사용하는 분석방법
- 두 경우 모두 다음의 아래의 통계량 공식을 사용함

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

추론통계

Chi-Square Test : 적합도 검정(goodness-of-fit-test)

- 적합도 검정(goodness-of-fit-test) → 기대되는 빈도의 분포와 관찰한 빈도의 분포를 비교
- 분석하려는 범주형 변수의 각 범주에 대해 기대되는 빈도를 계산함
- 실제 데이터에서 각 범주의 관찰된 빈도와 예상 빈도를 비교하고, 이를 통해 각 범주 간의 차이를 계산함

가설 설정 방법

- 귀무가설(H_0) : 주어진 데이터는 분포가 적합하다.→ 실제 관찰된 빈도와 기대되는 이론적인 빈도 간에는 유의미한 차이가 없다.
- 대립가설(H_1) : 주어진 데이터는 분포가 적합하다. → 실제 관찰된 빈도와 기대되는 이론적인 빈도 간에 유의미한 차이가 있다.

추론통계

Chi-Square Test : 적합도 검정(goodness-of-fit-test)

- 적합도 검정(goodness-of-fit-test) → 독립변수가 하나이고, 기대되는 빈도의 분포와 관찰한 빈도의 분포를 비교
  - 바구니 안에 100개의 아이스크림이 있고, 이 200개의 아이스크림은 4가지 맛을 가지고 있다.
  - 해당 바구니 안에 아이스크림이 골고루 섞여 있다고 할 수 있을까?

	메로나	쌍쌍바	붕어싸만코	바밤바
관찰값	45	47	59	49
기대빈도 값	50	50	50	50

추론통계

Chi-Square Test : 적합도 검정(goodness-of-fit-test)

- 적합도 검정(goodness-of-fit-test) → 독립변수가 하나이고, 기대되는 빈도의 분포와 관찰한 빈도의 분포를 비교
  - 바구니 안에 100개의 아이스크림이 있고, 이 200개의 아이스크림은 4가지 맛을 가지고 있다.
  - 해당 바구니 안에 아이스크림이 골고루 섞여 있다고 할 수 있을까?

	메로나	쌍쌍바	붕어싸만코	바밤바
관찰값	45	47	59	49
기대빈도 값	50	50	50	50

$$\chi^2 = \frac{(45-50)^2}{50} + \frac{(47-50)^2}{50} + \frac{(59-50)^2}{50} + \frac{(49-50)^2}{50} = 2.32$$

- 총 4개의 카테고리 → 자유도(k-1)→ 3

$$\chi^2(3)_{0.95} = 7.815$$

- 카이제곱 통계량 > 기준값 : 귀무가설 기각(적합(일정)하지 않다.)
- 카이제곱 통계량 < 기준값 : 귀무가설 채택(적합(일정)하다.)

Chi-Square Test : 적합도 검정(goodness-of-fit-test)

- H_0: 하나의 주머니에 세 가지맛의 사탕의 분포는 차이가 없다(일정하다)
- H_1: 하나의 주머니에 세 가지맛의 사탕의 분포는 차이가 있다(일정하지 않다)

```
candy_data <- data.frame(
       Color = c("Red", "Blue", "Green"),
       Observed = c(30, 50, 20))

total_candies <- sum(candy_data$Observed)

expected <- rep(total_candies/3, 3)

test_result <- chisq.test(candy_data$Observed, p = expected / sum(expected))

print(test_result)
```

추론통계

```
std_residuals <- (candy_data$Observed - expected) / sqrt(expected)

# 각 셀의 p-value 계산
chi_square_values <- std_residuals^2
p_values <- 1 - pchisq(chi_square_values, df = 2) #df = 자유도

candy_data$Residuals <- std_residuals
candy_data$Chi_square <- chi_square_values
candy_data$p_value <- p_values

print(candy_data)
```

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

- **잔차가 양수일 경우 : 관측값 > 기대값 (더 많이 나옴)**
- **잔차가 음수일 경우 : 관측값 < 기대값(덜 나옴)**

추론통계

Chi-Square Test : 적합도 검정(goodness-of-fit-test)

- H_0: 하나의 주머니에 세 가지맛의 사탕의 분포는 차이가 없다(일정하다)
- H_1: 하나의 주머니에 세 가지맛의 사탕의 분포는 차이가 있다(일정하지 않다)

```
observed <- table(데이터$변수)

expected <- rep(sum(observed) / length(observed), length(observed))

test_result <- chisq.test(x = observed, p = expected / sum(expected))

# 5. 결과 출력
print(test_result)
```

추론통계

Chi-Square Test : 독립성 검정(chi-square independence test)

- 교차 분석(cross tabulation analysis) → 범주형 변수가 여러 개인 경우에 사용하는 분석방법
- 여러 범주형 변수의 범주 간 차이가 기대빈도 값에서 유의하게 벗어나는지를 판단 → 변수 간의 연관관계 파악

Subject ID	Age Group	Sex
GW	young	F
JA	middle	F
TJ	young	M
JMA	young	M
JMO	middle	F
JQA	old	F
AJ	old	F
MVB	young	M
WHH	old	F
JT	young	F
JKP	middle	M

변수를 가지는 데이터 셋

Age Group / Sex	Female	Male	Total
young	2	3	5
middle	2	1	3
old	3	0	3
Total	7	4	11

교차 테이블(Cross-tabulation)

추론통계

Chi-Square Test : 독립성 검정(chi-square independence test)

- 두 개의 변수가 범주형을 가질 경우 → 두 개의 범주형 요인들을 카운트해 서로 연관성이 있는지 판단함(두 개 이상도 가능함)
- 변수가 가지는 고유의 값 → 범주형
- 범주형 변수의 연관관계를 파악함

$$\chi^2 = \sum \frac{(O-E)^2}{E} \qquad E_{ij} = \frac{(\text{행의 합}_i) \times (\text{열의 합}_j)}{\text{전체갯수}} \qquad \text{자유도}(k) = (row-1)(column-1)$$

O : 실제관측값
E : 예상개수(연관이 없다고 할 경우)

아이스크림 맛	선호 색깔
명목형	명목형
명목형	명목형

	초콜릿	바닐라	딸기	Total
빨간색	25	15	20	25+15+20
파란색	30	35	25	30+35+25
초록색	15	20	15	15+20+15
Total	25+30+15	15+35+20	20+25+15	200

Chi-Square Test : 독립성 검정(chi-square independence test)

	초콜릿	바닐라	딸기	Total
빨간색	25	15	20	25+15+20
파란색	30	35	25	30+35+25
초록색	15	20	15	15+20+15
Total	25+30+15	15+35+20	20+25+15	200

$R_{red} = 60, R_{Blue} = 90, R_{Green} = 50$

$C_{ch} = 70,\ C_{Va} = 70,\ C_{St} = 60$

$$E_{red,ch} = \frac{(60) \times (70)}{200} \qquad E_{red,ch} = \frac{(60) \times (70)}{200} = 21$$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(25-21)^2}{21} + .. + .. + .. +$$

- 카이제곱 통계량 > 기준값 : 귀무가설 기각(독립이 아니다)
- 카이제곱 통계량 < 기준값 : 귀무가설 채택(독립이다)

Chi-Square Test : 독립성 검정(chi-square independence test)

- H_0: 좋아하는 색깔과 선호하는 아이스크림 맛은 서로 관련성이 없다.
- H_1: 좋아하는 색깔과 선호하는 아이스크림 맛은 서로 관련성이 있다.

```
observed <- matrix(c(25, 15, 20, 30, 35, 25, 15, 20, 15), nrow = 3, byrow = TRUE)

rownames(observed) <- c("Red", "Blue", "Green")
colnames(observed) <- c("Chocolate", "Vanilla", "Strawberry")
chi_squared <- chisq.test(observed)
print(chi_squared)
```

Chi-Square Test : 독립성 검정(chi-square independence test)

- H_0: 성별에 따라 좋아하는 음식은 서로 관련성이 없다.
- H_1: 성별에 따라 좋아하는 음식은 서로 관련성이 있다.

```
data <- data.frame(Gender = c("Male", "Female", "Male", "Male", "Female",
"Female", "Male", "Male", "Female", "Female"),  Food = c(“국밥”, “마라탕”, “국밥”,
“피자”, “피자”, “국밥”, “국밥”, “마라탕”, “피자”, “피자”))

cross_tab <- table(data$Gender, data$Food)
cross_tab

chi_square_test_result <- chisq.test(cross_tab)
print(chi_square_test_result)
```

Chi-Square Test : 독립성 검정(chi-square independence test)

- H_0: 성별에 따라 좋아하는 음식은 서로 관련성이 없다.
- H_1: 성별에 따라 좋아하는 음식은 서로 관련성이 있다.

```
#gender_food.csv파일을 이용해 카이제곱 검정하기

cross_tab <- table(data$Gender, data$Food)
cross_tab

chi_square_test_result <- chisq.test(cross_tab)
print(chi_square_test_result)
```

Chi-Square Test : 사후 검정

- 성별(남/여)에 따라 어떤 음식(국밥, 마라탕, 피자)을 더 선호하는가를 판단하기 위해
- 카이제곱 분석의 사후검정 방법으로 그룹과 그룹의 차이를 비교함
- Bonferroni Correction Method 방법론을 사용해 그룹 간에 더 큰 영향력이 있는 집단을 판단함
- 비교 대상을 기준으로 항목을 조합에서 나오는 횟수만큼 나눠 우연히 발생하는 항목을 무시함

```
library(chisq.posthoc.test)
cross_tab <- table(data$Gender, data$Food)
cross_tab

results <- chisq.posthoc.test(cross_tab, method = "bonferroni")
results

#조합 공식
choose(n, k)
```

$$\text{조합} = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

n : 항목의 수
k : 비교 대상

- 잔차가 양수일 경우 : 관측값 > 기대값 (더 많이 나옴)
- 잔차가 음수일 경우 : 관측값 < 기대값(덜 나옴)

Chi-Square Test : 예제

- **부부의 집안일에 대해서 판단하기**

```
#카이제곱 분석
data=read.csv("경로~~/housetasks.csv", row.names = 1)
chi_square_test_result <- chisq.test(data)
print(chi_square_test_result)

#사후검증
results <- chisq.posthoc.test(data, method = "bonferroni")
results

#시각화
library(corrplot)
corrplot(chi_square_test_result$residuals, is.cor = FALSE)
```