

## Chapter 1

# 교차분석과 카이제곱검정

### 학습내용

교차분석은 두 개 이상의 범주형 변수를 대상으로 교차분할표를 작성하고, 이를 통해서 변수 상호 간의 관련성 여부를 분석한다. 교차분석은 특히 빈도분석 결과에 대한 보충자료를 제시하는데 효과적으로 이용할 수 있다. 또한 카이제곱검정은 교차분석으로 얻어진 교차분할표를 대상으로 유의확률을 적용하여 변수들 간의 독립성 및 관련성 여부 등을 검정하는 분석방법이다.

### 학습목표

- 교차분석을 위해서 두 개 이상의 데이터프레임을 생성할 수 있다.
- 두 개 이상의 범주형 변수를 대상으로 교차분할표를 작성할 수 있다.
- 연구 환경에서 연구가설과 귀무가설을 진술 할 수 있다.
- 연구 환경에서 일원카이제곱의 적합도 검정을 수행할 수 있다.
- 연구 환경에서 이원카이제곱의 독립성 검정을 수행할 수 있다.

## Chapter 1 구성

### 1.1 교차분석

### 1.2 카이제곱검정

### 1.3 교차분석과 검정보고서 작성

### 연습문제

## 1.1 교차분석

교차분석(Cross Table Analyze)은 범주형 자료(명목척도 또는 서열척도)를 대상으로 두 개 이상의 변수들에 대한 관련성을 알아보기 위해서 결합분포를 나타내는 교차분할표를 작성하고 이를 통해서 변수 상호 간의 관련성 여부를 분석하는 방법이다. 또한 교차분석은 빈도분석의 특성별 차이를 분석하기 위해 수행하는 분석방법으로 빈도분석 결과에 대한 보충자료를 제시하는데 효과적이다. 따라서 교차분석은 빈도분석과 함께 고급통계분석의 기초정보를 제공한다.

### 【교차분석 고려사항】

교차분석에 사용되는 변수는 값이 10 미만인 범주형 변수(명목척도, 서열척도)이어야 한다. 또한 비율척도인 경우는 코딩변경(리코딩)을 통해서 범주형 자료로 변화해야 한다. 예를 들면 연령인 경우 10~19세는 1, 20~29세는 2, 30~39세는 3 등으로 범주화하여 변경한다.

#### 1.1.1 데이터프레임 생성

교차분할표를 작성하기 위해서는 연구 환경에서 해당 변수를 확인(독립변수와 종속변수)하여 모델링한 후 범주형 데이터로 변환하는 변수 리코딩 과정을 거친다. 끝으로 대상 변수를 분할표로 작성하기 위해서 데이터프레임을 생성해야 한다.

### 【변수 모델링】

변수 모델링이란 특정 객체를 대상으로 분석할 속성(변수)을 선택하여 속성 간의 관계를 설정하는 일련의 과정을 의미한다. 여기서는 속성은 변수 또는 변인이라고도 한다. 변수 모델링에 관한 예를 들면 smoke 객체에서 education과 smoking 속성을 분석대상으로 하여 교육수준(education)이 흡연율(smoking)과 관련성이 있는가를 모델링할 경우 'education -> smoking' 형태로 기술한다. 이때 education은 영향을 미치는 변수로 독립변수에 해당되고, 영향을 받는 smoking은 종속변수에 해당된다.

<실습> 변수 리코딩과 데이터프레임 생성

단계 1 : 실습파일 가져오기

```
setwd("c:/Rwork/Part-III")
```

```
data <- read.csv("cleanDescriptive.csv", header=TRUE)
```

```
data # 확인
```

```
head(data) # 변수 확인
```

```
resident gender age level cost type survey pass cost2 resident2 gender2 age2 level2
pass2
1          1          1  50          1  5.1          1          1          2          2    특별시    남자    장년층    고졸
```

```

      실패
2      2      1 54      2 4.2      1      2      2      2      광역시      남자 장년층      대졸
      실패

```

단계 2 : 변수 리코딩

코딩 변경(리코딩)한 학력수준(level2)과 진학여부(pass2) 변수를 사용한다.

```
x <- data$level2 # 리코딩 변수 이용
```

```
y <- data$pass2 # 리코딩 변수 이용
```

단계 3 : 데이터프레임 생성

교차분할표 작성을 위한 데이터프레임 생성 방법은 다음 형식과 같다.

형식) data.frame(칼럼명=x, 칼럼명=y) (단, x, y는 명목척도 변수)

```
result <- data.frame(Level=x, Pass=y ) # 데이터프레임 생성
```

```
dim(result) # 차원보기
```

```
[1] 248  2
```

<해설> 부모의 학력수준이 자녀의 대학진학 여부와 관련이 있는지를 분석하기 위해서 학력수준(독립변수)과 진학여부(종속변수) 변수를 대상으로 데이터프레임을 생성하는 과정이다.

### 1.1.2 교차분석

교차분할표를 통해서 범주형 변수의 관계를 분석하는 방법으로 이전에 작성한 데이터프레임을 이용하여 교차분석을 수행한다.

<실습> 교차분할표 작성

단계 1 : 기본 함수 이용 교차분할표 생성

```
table(result) # 교차 빈도수
```

	Pass	
Level	실패	합격
고졸	40	49
대졸	27	55
대학원졸	23	31

<해설> result 데이터프레임 객체를 대상으로 table() 기본함수를 이용하여 두 개 이상의 변수의 결합분포를 나타내는 교차분할표를 작성한다.

단계 2 : 패키지 이용 교차분할표 생성

```
install.packages("gmodels") # gmodels 패키지 설치
```

```
library(gmodels) # CrossTable() 함수 사용
```

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

데이터 셋	diamonds 데이터 셋에 관한 설명
	<p>약 5만4천개의 다이아몬드에 관한 속성을 기록한 데이터 셋으로 53,940개의 관측치와 10개의 변수로 구성되어 있다. 주요 변수에 대한 설명은 다음과 같다.</p> <p>price : 다이아몬드 가격(\$326~\$18,823)</p> <p>carat : 다이아몬드 무게(0.2~5.01)</p> <p>cut : 컷의 품질(Fair, Good, Very Good, Premium Ideal)</p> <p>color : 색상(J:가장나쁨 ~ D:가장 좋음)</p> <p>clarity : 선명도(I1:가장나쁨, SI1, SI1, VS1, VS2, VVS1, VVS2, IF:가장 좋음)</p> <p>x: 길이 (0-10.74mm), y : 폭(0-58.9mm), z : 깊이 (0-31.8mm),</p> <p>depth : 깊이 비율 = <math>z / \text{mean}(x, y)</math></p>

```
# diamonds의 cut과 color에 대한 교차분할표 생성
```

```
CrossTable(x=diamonds$color, y=diamonds$cut)
```

<해설> 다이아몬드의 컷(cut) 품질과 색상(color)의 속성을 갖는 두 변수는 모두 명목척도로 구성되어 있다. 두 변수에 대한 교차분할표는 gmodels 패키지에서 제공되는 CrossTable() 함수를 이용하여 생성할 수 있다.

CrossTable() 함수를 실행하면 교차분할표를 이루고 있는 각 셀에 대한 데이터의 설명이 다음과 같이 제공된다. 여기서 각 셀에 나타난 데이터의 의미는 다음과 같다. 가장 첫 번째 줄은 관측치를 의미하고, 두 번째 줄은 카이제곱의 결과(기대치 비율), 세 번째 줄은 현재 행의 비율, 네 번째는 현재 열의 비율, 마지막 줄은 전체 비율에서 현재 셀의 값이 차지하는 비율을 의미한다.

```

Cell Contents
|-----|
|                                     N | 관측치(Row Total)
|      Chi-square contribution | 카이제곱( $X^2$ ) 적용
|          N / Row Total | 행비율(현재 행 차지 비율)
|          N / Col Total | 열비율(현재 열 차지 비율)
|          N / Table Total | 전체비율(전체에서 셀 비율)
|-----|

```

```
Total Observations in Table: 53940 # 교차분할표에서 전체 관측치
```

	diamonds\$cut					
diamonds\$color	Fair	Good	Very Good	Premium	Ideal	Row Total
D	163	662	1513	1603	2834	6775
	7.607	3.403	0.014	9.634	5.972	
	0.024	0.098	0.223	0.237	0.418	0.126
	0.101	0.135	0.125	0.116	0.132	
	0.003	0.012	0.028	0.030	0.053	

이하 생략 ...

<실습> 패키지이용 부모학력수준과 자녀 대학진학여부 교차분할표 작성

# 변수 모델 : 학력수준(독립변수) -> 진학여부(종속변수)

x <- data\$level2 # 학력수준 리코딩 변수

y <- data\$pass2 # 진학여부 리코딩 변수

CrossTable(x,y) # 교차분할표 작성(x:부모학력수준, y:자녀대학진학)

<해설> 부모의 학력수준과 자녀의 대학진학여부 교차분할표는 행(부모학력수준)과 열(자녀대학진학여부)에 의해서 다음 표 12-1과 같은 표 형태로 만들어진다.

표 12-1. 부모의 학력수준과 자녀의 대학진학여부 교차분할표

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

<해설> 교차분할표에서 기대비율은 카이제곱 식( $\chi^2 = \sum (\text{관측값} - \text{기댓값})^2 / \text{기댓값}$ )에 의해서 구해진 결과이다. 기대비율에 대한 기댓값은 「현재 행의 합 ×

현재 행 비율」 수식에 의해서 구한다. 예를 들면 교차분할표에서 부모의 학력수준이 고졸이면서 자녀가 대학에 실패할 경우의 기대치는  $35.244 = 89(\text{관측치 행의 합}) * 0.396(\text{행비율})$  계산되며, 소숫점 이하를 절사하면 기대치는 35가 된다.

연구자는 변수들에 대한 결합분포를 나타내는 교차분할표의 결과를 토대로 변수 간의 관련성을 다음과 같이 전반적으로 진술할 필요가 있다.

#### <논문/보고서에서 교차분할표 해석>

부모의 학력수준에 따른 자녀의 대학진학여부를 설문조사한 결과 학력수준에 상관없이 대학진학 합격률이 평균 60.0%로 학력수준별로 유사한 결과가 나타났다. 전체 응답자 225명을 대상으로 고졸 39.6%(89명) 중 55.1%가 진학에 성공하였고, 대졸 36.4%(82명) 중 68.4%가 성공했으며, 대학원졸은 24%(54명) 중 57.4%가 대학진학에 성공하였다. 특히 대졸 부모의 대학진학 합격율이 평균보다 조금 높고, 고졸 부모의 대학진학 합격율이 평균보다 조금 낮은 것으로 분석된다.

## 1.2 카이제곱검정

카이제곱검정(chi-square test)은 범주(Category)별로 관측빈도와 기대빈도의 차이를 통해서 확률 모형이 데이터를 얼마나 잘 설명하는지를 검정하는 통계적 방법이다.

일반적으로 교차분석으로 얻어진 분할표를 대상으로 유의확률을 적용하여 변수들 간의 독립성(관련성) 여부를 검정하는 분석방법으로 사용된다.(교차분석은 카이제곱검정 통계량을 사용하기 때문에 교차분석을 카이제곱 검정이라고 한다.) 카이제곱검정의 유형에는 적합도 검정, 독립성 검정, 동질성 검정으로 분류된다.

#### 【카이제곱검정 중요사항】

- 카이제곱검정을 위해서는 교차분석과 동일하게 범주형 변수를 대상으로 한다.
- 집단별로 비율이 같은지를 검정(비율에 대한 검정)하여 독립성 여부를 검정한다.
- 유의확률에 의해서 집단 간의 ‘차이가 있는가?’ 또는 ‘차이가 없는가?’로 가설을 검정한다.

<실습> CrossTable()함수 이용 카이제곱 검정

CrossTable()함수에 'chisq=TRUE' 속성을 적용하면 카이제곱 검정 결과를 볼 수 있다. 다음은 교차분할표 작성에서 이용된 diamonds의 cut과 color 변수에 대한 카이제곱 검정을 수행하는 R 코드이다.

```
CrossTable(x=diamonds$cut, y=diamonds$color, chisq = TRUE)
```

```
Pearson's Chi-squared test
```

```
-----  
--
```

$$\chi^2 = 310.3179 \quad \text{d.f.} = 24 \quad p = 1.394512e-51$$

<논문/보고서에서 카이제곱검정 해석>

cut과 color변수를 대상으로 카이제곱 검정을 수행한 결과 p(유의확률)값이 0.05보다 현저하게 적은 값으로 나타났다. 이는 유의확률(p)이 유의수준(0.05)보다 적다는 의미로 ‘두 변인은 서로 독립적이다.’라는 귀무가설을 기각할 수 있다. 따라서 ‘두 변인은 서로 독립적이지 않다.’라는 대립가설을 채택할 수 있다. 여기서 ‘독립적이다’라는 의미는 두 변인 사이에는 전혀 관련이 없다는 의미를 내포하기 때문에 ‘두 변인은 서로 관련성이 있다.’라는 의미로 해석할 수 있다.

### 1.2.1 카이제곱검정 절차와 기본가정

카이제곱검정은 유의수준과 유의확률 값에 의해서 가설을 검정해야 하기 때문에 가장 먼저 가설을 설정해야 한다. 카이제곱검정 절차는 다음과 같다.

단계 1. 가설을 설정한다.

귀무가설( $H_0$ ) : ~같다. ~다르지 않다. ~차이가 없다. ~효과가 없다.

대립가설( $H_1$ ) : ~같지 않다. ~다르다. ~차이가 있다. ~효과가 있다.

단계 2. 유의수준( $\alpha$ )을 결정한다.

일반사회과학분야 :  $\alpha=0.05$ , 의·생명과학분야 :  $\alpha=0.01$

단계 3. 자유도(df)와 유의수준( $\alpha$ )에 따른  $\chi^2$  분포표에 의한 기각값 결정한다.

단계 4. 관찰도수에 대한 기대도수를 구한다.

단계 5. 검정통계량  $\chi^2$ 의 값을 구한다. ( $\chi^2 = \sum (\text{관측값} - \text{기댓값})^2 / \text{기댓값}$ )

단계 6.  $\chi^2$  검정통계량과 기각값을 비교하여 귀무가설 채택 여부를 판정한다.

단계 7. 카이제곱 검정 결과를 진술한다.

#### 【카이제곱검정 기본가정】

- 변인의 척도 제한 : 종속변인은 범주형(명목척도, 서열척도) 변인을 사용한다.
- 기대빈도의 크기 : 사례수가 30보다 큰 경우 5미만의 기대빈도의 셀이 전체의 20%보다 많으면 사례수를 증가시킨 후 다시 검정을 수행한다.

### 1.2.2 카이제곱검정 유형

카이제곱검정 유형은 교차분할표 이용 여부에 따라서 크게 일원카이제곱검정과 이원카이제곱검정으로 분류된다.

#### (1) 일원카이제곱검정

교차분할표를 이용하지 않는 카이제곱검정으로 한 개의 변인(집단 또는 범주)을 대상으로 검정을 수행한다. 관찰도수가 기대도수와 일치하는지를 검정하는 적합도 검정(test for goodness of fit)이 여기에 속한다.

## (2) 일원카이제곱

교차분할표를 이용하는 카이제곱검정으로 한 개 이상의 변인(집단 또는 범주) 대상으로 검정을 수행한다. 분석 대상의 집단 수에 의해서 독립성 검정과 동질성 검정으로 나누어진다.

- **독립성 검정(test of independence)** : 한 집단 내에서 두 변인의 관계가 독립인지를 검정하는 방법이다. 독립성 검정을 위한 귀무가설의 예는 다음과 같다.

귀무가설( $H_0$ ) : ‘두 사건은 관련성이 없다.’

- **동질성 검정(test of homogeneity)** : 두 집단 이상에서 각 범주(집단) 간의 비율이 서로 동일한지를 검정하는 방법이다. 즉 두 개 이상의 범주형 자료가 동일한 분포를 갖는 모집단에서 추출된 것인지 검정하는 방법이다. 동질성 검정을 위한 귀무가설의 예는 다음과 같다.

귀무가설( $H_0$ ) : ‘모든 표본들의 비율은 동일하다.’

## 1.2.3 일원카이제곱검정

한 개의 변인을 대상으로 검정을 수행하기 때문에 교차분할표를 이용하지 않고 검정을 수행한다. 일원카이제곱검정에서는 적합도 검정과 선호도 분석에서 주로 이용된다.

## (1) 적합도 검정

적합도 검정은 `chisq.test()` 함수를 이용하여 관찰빈도와 기대빈도 일치여부를 검정한다. 다음은 적합도 검정을 위한 가설의 예시이다.

## &lt;적합도 검정 가설 예&gt;

귀무가설 : 기대치와 관찰치는 차이가 없다.

예) 주사위는 게임에 적합하다.

대립가설 : 기대치와 관찰치는 차이가 있다.

예) 주사위는 게임에 적합하지 않다.

## &lt;실습&gt; 주사위 적합도 검정

60회 주사위를 던져서 나온 관측도수와 기대도수가 다음과 같이 나온 경우 이 주사위는 게임에 적합한 주사위 인가를 일원카이제곱검정 방법으로 분석한다.



표 12-2. 주사위 눈금의 관측도수와 기대도수

주사위 눈금	1	2	3	4	5	6
관측도수	4	6	17	16	8	9
기대도수	10	10	10	10	10	10

```
chisq.test(c(4,6,17,16,8,9))
```

chisq.test()함수에 의해서 분석한 적합도 검정의 통계량은 다음과 같다.

Chi-squared test for given probabilities

```
data: c(4, 6, 17, 16, 8, 9)
```

```
X-squared = 14.2, df = 5, p-value = 0.01439
```

<해설> 카이제곱검정 결과를 해석하는 방법은 다음과 같이 유의확률(p-value)로 해석하는 방법과 검정통계량(X-squared, df)으로 해석하는 방법이 있다.

<유의확률 해석하는 방법>

유의확률(p-value : 0.01439)이 0.05미만이기 때문에 유의미한 수준( $\alpha=0.05$ )에서 귀무가설을 기각할 수 있다. 따라서 ‘주사위는 게임에 적합하다.’라는 귀무가설을 기각하고 대립가설(주사위는 게임에 적합하지 않다.)을 채택할 수 있다.

<검정통계량 해석하는 방법>

검정통계량 : X-squared = 14.2, df = 5

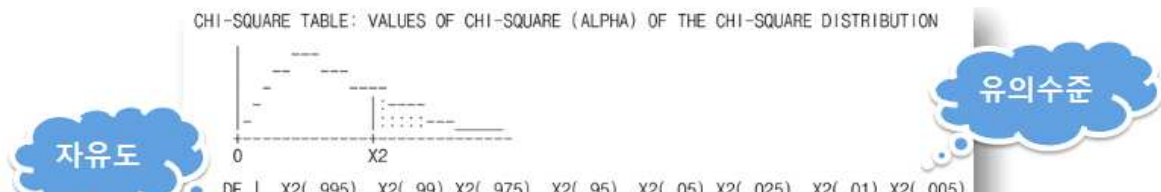
카이제곱(X-squared)은 관측값과 기댓값을 이용하여 다음과 같은 수식에 의해서 구해진다.

$$\chi^2 = \sum (\text{관측값} - \text{기댓값})^2 / \text{기댓값}$$

자유도(df : degree of freedom)란 검정을 위해서 n개의 표본(관측치)을 선정한 경우 n번째 표본은 나머지 표본이 정해지면 자동으로 결정되는 변인의 수를 의미하기 때문에 자유도는 N-1로 표현된다. 교차분할표에서 자유도(df) = (행수-1) \* (열수-1)로 구해진다.

카이제곱검정 절차의 ‘단계 3’에 의해서 자유도(df)와 유의수준( $\alpha$ )에 따른  $\chi^2$  분포 표에 의한 기각값을 결정할 수 있다. 검정통계량이 자유도(df)가 5이고, 유의수준이 0.05인 경우 chi-square 분포표(그림 12-1)에 의하면 임계값이 11.071에 해당된다. 그러므로 X-squared 기각값(역)은  $\chi^2 \geq 11.071$ 이 된다. 즉  $\chi^2$  값이 11.071 이상이면 귀무가설을 기각할 있다는 의미이다. 따라서 X-squared 검정통계량이 14.2 이기 때문에 기각역에 해당되어, 귀무가설을 기각하고, 대립가설을 채택할 수 있다.

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



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

그림 12-1. chi-square분포 표

## (2) 선호도 분석

선호도 분석은 적합도 검정과 마찬가지로 관측빈도와 기대빈도의 차이를 통해서 확률 모형이 주어진 자료를 얼마나 잘 설명하는지를 검정하는 통계적 방법이다. 단지 차이점은 분석에 필요한 연구 환경과 자료라고 볼 수 있다. 다음은 선호도 분석을 위한 가설의 예시이다.

## &lt; 선호도 분석 가설 예 &gt;

귀무가설 : 기대치와 관찰치는 차이가 없다.

예) 스포츠음료에 대한 선호도에 차이가 없다.

대립가설 : 기대치와 관찰치는 차이가 있다.

예) 스포츠음료에 대한 선호도에 차이가 있다.

<실습> 5개의 스포츠 음료에 대한 선호도에 차이가 있는지 검정

```
data <- textConnection(
  "스포츠음료종류 관측도수
  1 41
  2 30
  3 51
  4 71
  5 61
  ")
x <- read.table(data, header=T)
x
```

스포츠음료종류 관측도수		
1	1	41
2	2	30
3	3	51
4	4	71
5	5	61

`chisq.test(x$관측도수) # 선호도 분석의 검정 통계량 확인`

Chi-squared test for given probabilities

data: x\$관측도수  
X-squared = 20.4882, df = 4, p-value = 0.0003999

<유의확률 해석 방법>

유의확률(p-value : 0.0003999)이 0.05미만이기 때문에 유의미한 수준( $\alpha=0.05$ )에서 귀무가설을 기각할 수 있다. 따라서 ‘스포츠음료에 대한 선호도에 차이가 없다..’라는 귀무가설을 기각하고 대립가설(스포츠음료에 대한 선호도에 차이가 있다.)을 채택할 수 있다. 검정통계량으로 해석하는 방법은 적합도 검정과 동일하기 때문에 생략한다.

#### 1.2.4 이원카이제곱검정

한 개 이상의 변인(집단 또는 범주) 대상으로 교차분할표를 이용하는 카이제곱검정 방법으로 분석 대상의 집단 수에 의해서 독립성 검정과 동질성 검정으로 나누어진다.

(1) 독립성 검정(관련성 검정)

동일 집단의 두 변인을 대상으로 관련성이 있는가? 또는 없는가?를 검정하는 방법이다. 다음은 독립성 검정을 위한 가설의 예시이다.

<독립성 검정 가설 예>

귀무가설 : 경제력과 대학진학 합격률과 관련성이 없다.(=독립적이다.)

대립가설 : 경제력과 대학진학 합격률과 관련성이 있다.(=독립적이지 않다.)

<실습> 부모의 학력수준과 자녀의 대학진학 여부와 독립성이 있는지 검정

#### 【독립성 검정 가설】

- 연구가설( $H_1$ ) : 부모의 학력수준과 자녀의 대학진학 여부는 관련성이 있다.
- 귀무가설( $H_0$ ) : 부모의 학력수준과 자녀의 대학진학 여부는 관련성이 없다.

※ 논문이나 보고서에서는 귀무가설을 기각하고, 연구가설을 채택하는 것이 목적이다. 또한 대립가설 용어를 연구가설로 표현한다.

```
# 독립변수(x)와 종속변수(y) 생성
x <- data$level2 # 부모의 학력수준
y <- data$pass2 # 자녀의 대학진학여부
CrossTable(x, y, chisq = TRUE)
```

Pearson's Chi-squared test

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

#### <유의확률 해석 방법>

유의확률(p-value : 0.2527)이 0.05이상이기 때문에 유의미한 수준( $\alpha=0.05$ )에서 귀무가설을 기각할 수 없다. 따라서 ‘부모의 학력수준과 자녀의 대학진학 여부는 독립성이 없다.’라는 귀무가설을 기각할 수 없기 때문에 두 변인 간에 관련성이 없는 것으로 해석할 수 있다.

#### <검정통계량 해석 방법>

검정통계량 :  $\chi^2 = 2.766951$  d.f. = 2

검정통계량의 자유도(df)가 5이고, 유의수준이 0.05인 경우 chi-square 분포표에 의하면 임계값이 5.99에 해당된다. 그러므로 X-squared 기각값(역)은  $\chi^2 \geq 5.99$ 가 된다. 즉  $\chi^2$  값이 5.99 이상이면 귀무가설을 기각할 있다는 의미이다. 하지만 검정통계량의  $\chi^2=2.766951$ 이기 때문에 귀무가설을 기각할 수 없다.

※ 카이제곱 검정통계량의 자유도(df)가 클수록 정규분포에 가까워진다.

### 1.3 교차분석과 검정보고서 작성

카이제곱검정은 교차분석으로 얻어진 교차분할표를 대상으로 유의확률을 적용하여 변수들 간의 독립성(관련성) 여부를 검정하기 때문에 논문이나 보고서에서는 다음 표 12-3과 같이 교차분할표와 카이제곱 검정통계량을 함께 제시한다.

표 12-3. 교차분석과 카이제곱검정 결과

학력수준		실패	합계	X-squared	유의확률(p)
고졸	관찰빈도(%)	40(44.9%)	49(55.1%)	2.766951	0.2507057
	기대빈도	35	54		
대졸	관찰빈도(%)	27(32.9%)	55(67.1%)		
	기대빈도	30	52		
대학원졸	관찰빈도(%)	23(42.6%)	31(57.4%)		
	기대빈도	25	29		

[관찰빈도-기대빈도] 값이 작을수록 카이제곱 값도 작아져서 귀무가설이 채택될 가능성이 높다. 기대빈도는 표 12-1 부모의 학력수준과 자녀의 대학진학여부 교차분할표에서 설명한 내용을 참고하기 바란다.

연구자는 교차분석과 카이제곱검정 결과를 토대로 가설검정의 연구 환경과 검정결과를 다음과 같이 종합적으로 진술할 필요가 있다.

#### <논문/보고서에서 교차분석과 카이제곱검정 결과 해석>

'부모의 학력수준과 자녀의 대학진학 여부와 관련성이 있다.'를 분석하기 위해서 자녀를 둔 A회사 225명의 부모를 표본으로 추출한 후 설문조사하여 교차분석과 카이제곱 검정을 실시하였다. 분석결과를 살펴보면 부모의 학력수준과 자녀의 대학진학 여부의 관련성은 유의미한 수준에서 차이가 없는 것으로 나타났다. ( $\chi^2 = 2.766951$ ,  $p > 0.05$ ) 따라서 귀무가설을 기각할 수 없기 때문에 부모의 학력수준과 자녀의 대학진학 여부와는 관련성이 없는 것으로 분석되었다.

#### (2) 동질성 검정

두 집단의 분포가 동일한가? 분포가 동일하지 않는가?를 검정하는 방법이다. 즉 동일한 분포를 갖는 모집단에서 추출된 것인지를 검정하는 방법이다. 다음은 동질성 검정을 위한 가설의 예시이다.

#### <동질성 검정 가설 예>

귀무가설 : 직업유형에 따라 만족도에 차이가 없다.

대립가설 : 직업유형에 따라 만족도에 차이가 있다.

<실습> 교육센터에서 교육방법에 따라 교육생들의 만족도에 차이가 있는지 검정

#### 【동질성 검정 가설】

- 연구가설( $H_1$ ) : 교육방법에 따라 만족도에 차이가 있다.
- 귀무가설( $H_0$ ) : 교육방법에 따라 만족도에 차이가 없다.

단계 1 : 파일 가져오기

```
setwd("c:/Rwork/Part-III")
data <- read.csv("homogeneity.csv", header=TRUE)
head(data) # 변수 보기
# survey변수의 NA값을 제외하여 서브 셋 작성
data <- subset(data, !is.na(survey), c(method, survey))
```

단계 2 : 코딩 변경(변수리코딩)

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

```
# method2 필드 추가
```

```
data$method2[data$method==1] <- "방법1"
data$method2[data$method==2] <- "방법2"
data$method2[data$method==3] <- "방법3"
# survey2 필드 추가
data$survey2[data$survey==1] <- "매우만족"
data$survey2[data$survey==2] <- "만족"
data$survey2[data$survey==3] <- "보통"
data$survey2[data$survey==4] <- "불만족"
data$survey2[data$survey==5] <- "매우불만족"
```

단계 3 : 교차분할표 작성

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

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

※ 주의 : 반드시 각 집단별 길이(50)가 같아야 한다.

단계 4 : 동질성 검정 - 모수 특성 치에 대한 추론검정

```
chisq.test(data$method2, data$survey2)
```

Pearson's Chi-squared test

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

#### <동질성 검정 해석>

유의수준 0.05에서  $\chi^2$ 값이 6.545, 자유도(df) 8, 그리고 유의확률(p-value) 0.586을 보이고 있다. 즉 6.545 이상의 카이제곱 값이 얻어질 확률이 0.586라는 것을 보여주고 있다. 이 값은 유의수준 0.05보다 크기 때문에 귀무가설을 기각할 수 없다. 따라서 '교육방법에 따른 만족도에 차이가 없다.' 라고 말할 수 있다.

## 【1장 연습문제】

01. 교육수준(education)과 흡연율(smoking) 간의 관련성을 분석하기 위한 연구가설을 수립하고, 각 단계별로 가설을 검정하시오. [독립성 검정]

귀무가설( $H_0$ ) :

연구가설( $H_1$ ) :

단계 1 : 파일 가져오기

```
setwd("c:/Rwork/Part-III")
```

```
smoke <- read.csv("smoke.csv", header=TRUE)
```

```
head(smoke)
```

단계 2 : 코딩 변경

education 칼럼(독립변수) : 1:대졸, 2:고졸, 3:중졸

smoke 칼럼(종속변수): 1:과다흡연, 2:보통흡연, 3:비흡연

단계 3 : 교차분할표 작성

단계 4 : 독립성 검정

단계 5 : 검정결과 해석

02. 나이(age3)와 직위(position) 간의 관련성을 단계별로 분석하시오. [독립성 검정]

단계 1 : 파일 가져오기

```
setwd("c:/Rwork/Part-III")
```

```
data <- read.csv("cleanData.csv", header=TRUE)
```

```
head(data)
```

단계 2 : 코딩 변경(변수 리코딩)

```
x <- data$position # 행 - 직위 변수 이용
```

```
y <- data$age3 # 열 - 나이 리코딩 변수 이용
```

단계 3 : 산점도를 이용한 변수간의 관련성 보기 - plot(x,y) 함수 이용



단계 4 : 독립성 검정

단계 5 : 검정결과 해석

03. 직업유형에 따른 응답정도에 차이가 있는가를 단계별로 검정하시오.[동질성 검정]

단계 1 : 패키지 설치 및 로딩

```
install.packages("XLConnect")
```

```
library(XLConnect)
```

단계 2 : 파일 가져오기

```
setwd("c:/Rwork/Part-III")
```

```
response <- read.csv("response.csv", header=TRUE)
```

단계 3 : 코딩 변경 - 리코딩

job 칼럼 : 1:학생, 2:직장인, 3:주부

response 칼럼 : 1:무응답, 2:낮음, 3:높음

단계 4 : 교차분할표 작성

단계 5 : 동일성 검정

단계 6 : 검정결과 해석

## Chapter 2

### 집단 간 차이 분석

#### 학습내용

통계학이란 논리적 사고와 객관적인 사실을 바탕으로, 일반적이고 확률론적 결정론에 의해서 인과관계를 규명한다. 특히 연구목적에 의해 설정된 가설들에 대하여 분석결과가 어떤 결과를 뒷받침하고 있는지를 통계적 방법으로 검정할 수 있다. chapter11의 기술 통계분석 방법은 수집된 자료의 특성을 쉽게 파악하기 위해서 자료를 정리 및 요약하는 통계학 영역이라면, 집단 간 차이 분석은 모집단에서 추출한 표본의 정보를 이용하여 모집단의 다양한 특성을 과학적으로 추론하는 학문 영역이다. 이장에서는 추론통계학을 기반으로 집단 간 차이에 대한 분석방법을 알아본다.

#### 학습목표

- 추정과 가설 검정에 대한 개념을 설명할 수 있다.
- 단일집단 비율 검정과 단일집단 평균 검정의 차이점을 설명할 수 있다.
- 두 집단 비율검정과 두 집단 평균 검정에 대한 사례를 각각 설명할 수 있다.
- 분산분석의 검정 방법을 이용하여 세 집단 이상의 평균 차이 검정을 수행하고 결과를 해석할 수 있다.

#### Chapter 2 구성

##### 2.1 추정과 검정

##### 2.2 단일집단 검정

##### 2.3 두 집단 검정

##### 2.4 세 집단 검정

#### 연습문제

## 2.1 추정과 검정

통계 조사에서 조사 대상이 되는 전체 집단을 모집단이라 하고, 모집단에서 뽑은 일 부 자료를 표본이라고 한다. 또한 모집단과 표본에 포함되어 있는 자료의 갯수를 각각 모집단의 크기, 표본이 크기라고 한다. 그리고 모집단으로부터 추출된 표본 으로부터 모수와 관련된 통계량(statistic)들의 값을 계산하고 이것을 이용하여 모 집단의 특성(모수)을 알아내는 과정을 추론 통계분석이라고 한다.

어떤 모집단에서 조사하고자 하는 특성을 나타내는 확률변수를  $X$ 라고 할 때,  $X$ 의 평균, 분산, 표준편차를 각각 모평균(  $\mu$  ), 모분산( $\sigma^2$ ), 모표준편차( $\sigma$ )라 한다. 또한 어떤 모집단에서 크기가  $n$ 인 표본을 임의추출 하였을 때 이 표본에 대한 평균, 분산, 표준편차를 표본평균(  $\bar{x}$  ), 표본분산( $S^2$ ), 표본표준편차( $S$ )라고 한다.

### 2.1.1 점 추정과 구간 추정

추론 통계분석 과정은 모집단에서 추출한 표본으로부터 얻은 정보를 이용하여 모집 단의 특성을 나타내는 값을 확률적으로 추측하는 추정(estimation)과 유의수준과 표본의 검정통계량을 비교하여 통계적 가설의 진위를 입증하는 가설 검정(hypotheses testing)로 나눌 수 있다.

추정 방법에는 점 추정과 구간 추정으로 분류할 수 있는데, 점 추정은 하나의 값을 제시하여 모수의 참값을 추측하고, 구간 추정은 하한값과 상한값의 신뢰구간을 지 정하여 모수의 참값을 추정하는 방식이다.

표 13-1. 점 추정과 신뢰구간 추정

구분	점 추정	신뢰구간 추정
방법	<ul style="list-style-type: none"> <li>하나의 값을 제시하여 모수의 참값을 추측하는 방법</li> </ul>	<ul style="list-style-type: none"> <li>하한값과 상한값의 구간을 지정 모수의 참값을 추정하는 방법</li> </ul>
특징	<ul style="list-style-type: none"> <li>추정값과 모수의 참값 사이의 오차범위 제공 안함</li> </ul>	<ul style="list-style-type: none"> <li>추정값과 모수의 참값 사이의 오차범위 제공</li> </ul>

점 추정 방식을 적용하여 가설을 검정할 경우 제시된 하나의 값과 표본에 의한 검 정통계량을 직접 비교하여 일치하면 귀무가설이 기각되지만, 일치하지 않으면 귀무 가설이 채택된다. 따라서 점 추정 방식에 의한 가설검정은 귀무가설의 기각율이 낮 다고 볼 수 있다. 또한 검정통계량과 모수의 참값 사이의 오차범위를 확인할 수 없 다.

한편 구간추정 방식으로 가설을 검정할 경우 오차범위에 의해서 결정된 하한값과 상한값의 신뢰구간과 검정통계량을 비교하여 가설을 검정하게 된다. 일반적으로 추론 통계분석에서는 구간추정 방식을 더 많이 이용한다. 오차범위는 모표준편차( )가 알려지지 않은 경우 표본의 표준편차(S)를 이용하여 추정한다.

### 2.1.2 모 평균의 구간추정

「우리나라 전체 중학교 2학년 남학생의 평균 키를 알아보기 위해서 중학교 2학년 남학생 10,000명을 대상으로 키를 조사한 결과 표본평균( )은 165.1cm 이고 표본 표준편차(S)는 2cm였다.」 이러한 연구 환경에서 표본평균( $\bar{X}$ )을 이용하여 모평균( $\mu$ )에 대한 신뢰도 95% 신뢰구간을 추정하는 방법에 대해서 알아본다.

정규분포  $N(\mu, \sigma^2)$ 을 따르는 모집단에서 크기가  $n$ 인 표본  $X_1, X_2, \dots, X_n$ 을 임의추출 할 때 표본평균( )는 정규분포  $N(\mu, \frac{\sigma}{n})$ 을 따르므로  $\bar{X}$ 를 표준화한 확률변수  $Z$

는 표준정규분포  $N(0,1)$ 을 따른다.

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

이때 표준정규분포에서  $P(-1.96 \leq Z \leq 1.96) = 0.95$  이므로  $Z$  수식을 적용하면

$$P(-1.96 \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96) = P(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}) = 0.95 \text{ 이다.}$$

다.

따라서 모평균  $\mu$ 가  $\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}$ 에 포함될 확률은 95%이고, 이 범위를 모평균의 신뢰도 95%의 신뢰구간이라고 한다. 이러한 의미는 모집단으로부터 크기가  $n$ 인 표본을 임의 추출하는 일을 반복할 경우 이들 중 95%는 모평균  $\mu$ 를 포함한다는 의미이다. 여기서 모표준편차  $\sigma$ 의 값이 알려지지 않은 경우 표본의 크기가  $n$ 이 충분히 클 때( $n \geq 30$ )에는 표본표준편차  $S$ 를 사용한다.

풀이)  $n=10,000$ ,  $\bar{X}=165.1$ ,  $S=2\text{cm}$  이므로 평균 키  $\mu$ 의 신뢰도 95% 신뢰구간은

$$165.1 - 1.96 \frac{2}{\sqrt{10000}} \leq \mu \leq 165.1 + 1.96 \frac{2}{\sqrt{10000}}$$

따라서 모평균  $\mu$ 의 신뢰구간은  $165.0608 \leq \mu \leq 165.1392$ 가 된다. 신뢰도와 모평균  $\mu$ 의 신뢰구간을 정리하면 다음 표 13-2와 같다.

1) 중심극한정리는 표본평균들이 이루는 분포는 샘플 수가 충분히 큰(케이스의 수가 30개 이상) 경우에는 모집단의 분포에 관계없이 정규분포에 접근한다는 이론이다.

표 13-2. 신뢰도와 모평균 신뢰구간

신뢰도	모평균( )의 신뢰구간
95%	$P(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}})$
99%	$P(\bar{X} - 2.58 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 2.58 \frac{\sigma}{\sqrt{n}})$

<실습> 우리나라 중학교 2학년 남학생 평균 신장 표본 조사

우리나라 중학교 2학년 남학생 평균 신장 표본 조사를 위한 검정통계량은 다음과 같다.

- 전체 표본 크기(N) : 10,000명
- 표본평균(X) : 165.1cm
- 표본표준편차(S) : 2cm

# 신뢰수준 95%의 신뢰구간 구하기

N = 10000 # 표본 크기(N)

X = 165.1 # 표본평균(X)

S = 2 # 표본표준편차(S)

low <- X - 1.96 \* S/sqrt(N) # 신뢰구간 하한값

high <- X + 1.96 \* S/sqrt(N) # 신뢰구간 상한값

low; high

[1] 165.0608

[1] 165.1392

<해설> 모평균( $\mu$ )에 대한 신뢰수준 95%의 신뢰구간 추정 수식을 적용하여 하한값과 상한값의 범위(165.0608~165.1392)를 계산한 결과 표본평균의 신장(165.1cm)이 신뢰구간에 포함되는 것으로 나타난다. 즉 표본평균의 신장은 95% 신뢰수준에서 우리나라 전체 중학교 2학년 남학생 평균 신장의 신뢰구간에 포함된다고 할 수 있다.

신뢰수준 95%의 모평균 신뢰구간 :  $165.0608 \leq \mu \leq 165.1392$

신뢰수준 95%는 신뢰구간이 모수를 포함할 확률을 의미하고, 신뢰구간은 오차범위에 의해서 결정된 하한값 ~ 상한값을 의미한다.

<실습> 신뢰구간으로 표본오차 구하기

low - X # 0.0392 = 신뢰구간 하한값 - 표본평균

high - X # 0.0392 = 신뢰구간 상한값 - 표본평균

# 백분율 적용

(low - X) \* 100 # -3.92

(high - X) \* 100 # +3.92

<해석> 표본오차는 표본이 모집단의 특성과 정확히 일치하지 않아서 발생하는 확률의 차이를 의미한다. 신뢰구간의 하한값에서 평균 신장을 빼고, 상한값에서 평균 신장을 뺀 값을 백분율로 적용하면  $\pm 3.92$ 의 표본오차가 나온다. 표본오차를 적용하여 검정통계량을 다음과 같이 해석할 수 있다. 즉 우리나라 중학교 2학년 남학생 평균 신장이 95% 신뢰수준에서 표본오차  $\pm 3.92$  범위에서 165.1cm로 조사되었다면 실제 평균키는 165.0608cm ~ 165.1392cm 사이에 나타날 수 있다는 의미이다.

### 13.1.3 모 비율의 구간추정

제품의 불량률, 대선 후보 지지율 등과 같이 모집단에서 어떤 사건에 대한 비율을 모비율(p)이라고 한다. 이러한 모비율 추정은 모집단으로부터 임의추출한 표본에서 어떤 사건에 대한 비율을 표본비율( $\hat{p}$ )이라고 하는데 이러한 표본비율을 이용하여 모비율을 추정할 수 있다.

「A반도체 회사의 사원을 대상으로 임의 추출한 150명을 조사한 결과 90명이 여자 사원이다.」 이러한 연구환경에서 표본비율을 이용하여 모비율의 신뢰도 95%의 신뢰구간을 추정하는 방법에 대해서 알아본다.

모비율이 p인 모집단에서 크기가 n인 표본을 임의추출한 경우, n이 충분히 크면 확

률변수 
$$\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \quad (q=1-p)$$
는 표준정규분포  $N(0,1)$ 을 따른다. 또한 표본의 크

기 n이 충분히 클 때  $\hat{p}$ 의 분산  $\frac{pq}{n}$ 에서 p, q 대신에 표본비율  $\hat{p}$ ,  $\hat{q}(\hat{q}=1-\hat{p})$ 을 사

용한 
$$\frac{\hat{p} - \hat{p}}{\sqrt{\frac{\hat{p}\hat{q}}{n}}}$$
 표준정규분포  $N(0,1)$ 을 따른다.

이때 표준정규분포에서  $P(-1.96 \leq z \leq 1.96) = 0.95$  이므로 Z 수식을 적용하면

$$P(-1.96 \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \leq 1.96) = P(\hat{p} - 1.96 \frac{\hat{p}\hat{q}}{n} \leq p \leq \hat{p} + 1.96 \frac{\hat{p}\hat{q}}{n}) = 0.95 \text{ 이다.}$$

따라서 모비율 p가  $\hat{p} - 1.96 \sqrt{\frac{\hat{p}\hat{q}}{n}} \leq p \leq \hat{p} + 1.96 \sqrt{\frac{\hat{p}\hat{q}}{n}}$ 에 포함될 확률은 95%이

고, 이 범위를 모비율 p의 신뢰도 95%의 신뢰구간이라고 한다. 신뢰도와 모비율 p

의 신뢰구간을 정리하면 다음 표 13-3과 같다.

표 13-3. 신뢰도와 모비율 신뢰구간

신뢰도	모비율(p)의 신뢰구간
95%	$-1.96 \sqrt{\frac{\hat{p}\hat{q}}{n}} \leq p \leq \hat{p} + 1.96 \sqrt{\frac{\hat{p}\hat{q}}{n}}$
99%	$\hat{p} - 2.58 \sqrt{\frac{\hat{p}\hat{q}}{n}} \leq p \leq \hat{p} + 2.58 \sqrt{\frac{\hat{p}\hat{q}}{n}}$

풀이)  $n=150$ ,  $\hat{p} = 90/150 = 0.6$  이므로 전체 사원의 여자 사원 비율  $p$ 의 신뢰도 95% 신뢰구간은  $0.6 - 1.96 \sqrt{\frac{0.6 \times 0.4}{150}} \leq p \leq 0.6 + 1.96 \sqrt{\frac{0.6 \times 0.4}{150}}$  따라서  $0.596864 \leq p \leq 0.603136$ 이다.

## 2.2 단일집단 검정

한 개의 집단과 기존 집단과의 비율 차이 검정과 평균 차이 검정에 대해서 알아본다. 비율 차이 검정은 기술통계량으로 빈도수에 대한 비율에 의미가 있으며, 평균 차이 검정은 표본평균에 의미가 있다.

### 2.2.1 단일집단 비율검정

단일 집단의 비율이 어떤 특정한 값과 같은지를 검정하는 방법으로 검정 방법 중에서 가장 간단한 방법으로 분석절차는 다음과 같다.

분석할 데이터를 대상으로 결측치와 이상치를 제거하는 전처리 과정을 거친 후 기술통계량으로 빈도분석을 계산하고, 이를 `binom.test()` 함수의 인수로 사용하여 비율 차이 검정을 수행한다.

비율 차이 검정 통계량을 바탕으로 귀무가설의 기각여부를 결정한다.

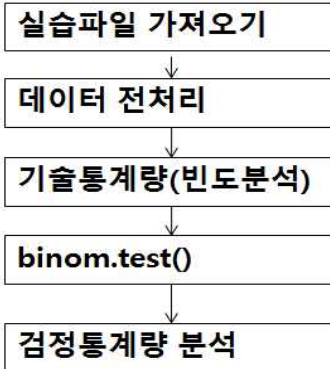


그림 13-1. 단일집단 비율검정 절차

#### <연구가설>

연구가설( $H_1$ ) : 기존 2016년도 고객 불만율과 2017년도 CS교육 후 불만율에 차이가 있다.  
 귀무가설( $H_0$ ) : 기존 2016년도 고객 불만율과 2017년도 CS교육 후 불만율에 차이가 없다.

#### <연구환경>

2016년도 114 전화번호 안내고객을 대상으로 불만을 갖는 고객은 20%였다. 이를 개선하기 위해서 2017년도 CS교육을 실시한 후 150명 고객을 대상으로 조사한 결과 14명이 불만을 갖고 있었다. 기존 20% 보다 불만율이 낮아졌다고 할 수 있는가?

#### (1) 단일표본 대상 기술통계량

분석 대상의 단일표본을 대상으로 빈도분석을 통해서 불만율에 대한 비율을 계산한다.

<실습> 단일표본 빈도수와 비율 계산

단계 1 : 실습데이터 가져오기

```
setwd("c:/Rwork/Part-III")
data <- read.csv("one_sample.csv", header=TRUE)
head(data)
x <- data$survey
```

단계 2 : 빈도수와 비율 계산

```
summary(x) # 결측치 없음
length(x) # 150개
table(x) #table(x, useNA="ifany") # 시리얼 데이터와 NA 갯수 출력 시
x
0 1
14 136 -> 0:불만족(14), 1: 만족(136)
```

단계 3 : 패키지 이용 빈도수와 비율 계산

```
install.packages("prettyR")
library(prettyR) # freq() 함수 사용
freq(x)
Frequencies for x
  1  0  NA
136 14  0 <- 빈도수
% 90.7 9.3 0 <- 비율 제공
```

<해설> table()함수나 패키지에서 제공되는 함수를 이용하여 단일집단을 대상으로 기술통계량을 구한다.

#### (2) 이항분포 비율검정



명목척도의 비율을 바탕으로 `binom.test()` 함수를 이용하여 2)이항분포의 양측 검정을 통해서 검정 통계량을 구한 후 이를 이용하여 가설을 검정한다. `binom.test()` 함수의 사용을 위한 형식은 다음과 같다.

`help(binom.test)` # 함수 형식 보기

```
형식) binom.test(x, n, p = 0.5,
                 alternative = c("two.sided", "less", "greater"),
                 conf.level = 0.95)
```

`binom.test()` 함수의 형식을 적용한 예로, 150명 고객을 대상으로 136명이 만족, 14명이 불만족으로 집계된 경우 전체 고객 중 80% 이상이 만족하고 있는지를 알아보기 위해서 첫 번째 인수는 성공 횟수 136, 두 번째 인수는 시행 횟수 150(136+14)이 되도록 14를 지정하고, 세 번째 인수 `p`는 136명의 만족 고객이 전체 80% 이상의 만족율을 나타내는지를 검정하기 위해서 0.8을 지정한다.

예) `binom.test(c(136, 14), p=0.8)`

<실습> 만족율 기준 비율검정

만족율 80% 이상을 기준으로 양측검정을 실시한다.

단계 1 : 양측검정 : 기존 80% 만족율 기준 검정 실시

`binom.test(c(136, 14), p=0.8)`

`binom.test(c(136, 14), p=0.8, alternative="two.sided", conf.level=0.95)`

- `alternative="two.sided"` 속성은 양측검정을 의미하고, `conf.level=0.95`는 95% 신뢰수준을 의미한다. 이 두 속성은 기본값으로 지정되어 있기 때문에 생략이 가능하다. 만약 `conf.level=0.99`로 지정하면 99%신뢰수준으로 검정통계량이 구해진다.

<검정결과 해설>

만족 고객 136명을 대상으로 95% 신뢰수준에서 양측 검정을 실시한 결과 검정 통계량 `p-value` 값은 0.0006735로 유의수준 0.05보다 작기 때문에 기존 만족률(80%)과 차이가 있다고 볼 수 있다. 즉 기존 2016년도 고객 불만율과 2017년도 CS교육 후 불만율에 차이 있다고 볼 수 있다. 하지만 양측가설 검정 결과에서는 기존 만족률보다 ‘크다’ 혹은 ‘작다’라는 방향성은 제시되지 않는다. 귀무가설을 부정하는 대립가설은 3가지가 있다. 즉 귀무가설이 ‘모평균 = 상수’ 일 때 ‘모평균 = 상수’가 아닐 때 양측가설 검정이고, 방향성이 있는 경우 단측가설 검정을 수행한다.

2) 정규분포와 마찬가지로 모집단이 가지는 이상적인 분포형태로 정규분포가 연속변량인 반면에 이항분포는 이산변량이며, 그래프는 좌우대칭인 종 모양의 곡선 형태를 갖는다.

단계 2 : 방향성을 갖는 단측가설 검정

`binom.test(c(136,14), p=0.8, alternative="greater", conf.level=0.95)`

`alternative="greater"` 속성은 방향성을 갖는 연구가설을 검정할 경우 이용된다. 즉 95% 신뢰수준에서 전체 150명 중에서 136명의 만족 고객이 전체 비율의 80% 보다 더 큰 비율인가를 검정하기 위한 속성이다.

<검정결과 해설>

‘기존 2016년도 고객 불만율에 비해서 2017년도 CS교육 후 불만율이 더 낮다.’라는 방향성 (`greater`)이 있는 연구가설을 검정한 결과 검정통계량  $p$ -value값은 0.0003179으로 유의수준 0.05보다 작기 때문에 기존 만족율 보다 80% 이상의 효과를 얻을 수 있다고 볼 수 있다. 결과적으로 기존 20% 보다 불만율이 낮아졌다고 할 수 있다. 따라서 귀무가설이 기각되고, 연구가설이 채택된다.(CS교육에 효과가 있다고 볼 수 있다.)

### 2.2.2 단일집단 평균검정(단일표본 T검정)

단일 집단의 평균이 어떤 특정한 집단의 평균과 차이가 있는지를 검정하는 방법으로 분석절차는 다음과 같다.

분석할 데이터를 대상으로 전처리 후 평균 차이 검정을 위해서 기술통계량으로 평균을 구한다. 평균차이 검정은 정규분포 여부를 판정한 후 결과에 따라서 T검정 또는 웰콕스(`wilcox`) 검정을 수행한다. 만약 정규분포인 경우에는 모수 검정인 T 검정을 수행하지만 정규분포가 아닌 경우에는 비모수 검정인 웰콕스 검정으로 평균차이 검정을 실시하여 귀무가설의 기각여부를 결정한다.

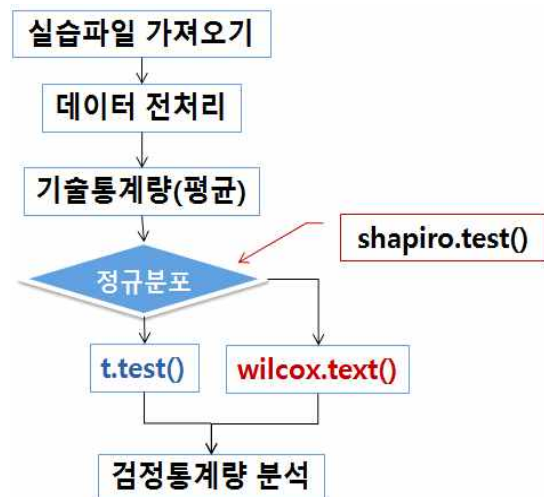


그림 13-2. 단일집단 평균검정 절차

<연구가설>

연구가설( $H_1$ ) : 국내에서 생산된 노트북과 A회사에서 생산된 노트북의 평균 사용 시간에 차이가 있다.  
 귀무가설( $H_0$ ) : 국내에서 생산된 노트북과 A회사에서 생산된 노트북의 평균 사용 시간에 차이가 없다.

<연구환경>

국내에서 생산된 노트북 평균 사용 시간이 5.2시간으로 파악된 상황에서 A회사에서 생산된 노트북 평균 사용시간과 차이가 있는지를 검정하기 위해서 A회사 노트북 150대를 랜덤으로 선정하여 검정을 실시한다.

### (1) 단일표본 평균 계산

데이터의 전처리 과정을 통해서 outlier를 제거한 후 변수에 대한 대푯값의 성격을 갖는 평균을 계산한다.

#### <실습> 단일표본 평균 계산하기

단계 1 : 실습파일 가져오기

```
setwd("c:/Rwork/Part-III")
data <- read.csv("one_sample.csv", header=TRUE)
str(data) # 150
head(data)
x <- data$time
head(x)
```

단계 2 : 데이터 분포/결측치 제거

```
summary(x) # NA-41개
mean(x) # error
```

단계 3 : 데이터 정제

```
mean(x, na.rm=T) # NA 제외 평균(방법1)
x1 <- na.omit(x) # NA 제외 평균(방법2)
mean(x1)
[1] 5.556881
```

<해설> 단일집단 평균차이 검정을 수행하기 전에 단일집단을 대상으로 평균에 관한 통계량을 계산한다.

### (2) 평균(mean) 검정통계량 특징

평균 검정통계량은 비율척도와 같은 수치기반 데이터에 의미가 있다. 특히 분포의 중심위치를 나타내는 대푯값의 성격을 가지며, 정규분포에서 도수분포곡선이 평균값을 중앙으로 하여 좌우대칭인 종 모양을 형성한다. 또한 집단 간의 평균에 차이가 있는지를 검정하는 용도로 사용된다.

## (3) 정규분포 검정

단일표본 평균 차이 검정을 하기 전에 데이터의 분포 형태가 정규분포 인지를 먼저 검정해야한다. 정규분포 검정은 stats패키지에서 제공하는 shapiro.test()함수를 이용할 수 있다. 검정 결과가 유의수준 0.05보다 큰 경우 정규분포로 본다.

## &lt;실습&gt; 정규분포 검정

귀무가설( $H_0$ ) : x의 데이터 분포는 정규분포이다.

```
shapiro.test(x1) # x1 데이터에 대한 정규분포 검정
Shapiro-Wilk normality test
data: x1
W = 0.9914, p-value = 0.7242
```

<해설> 검정통계량 p-value값은 0.7242로 유의수준 0.05보다 크기 때문에 x1 객체의 데이터 분포는 정규분포를 따른다고 할 수 있다. 따라서 모수 검정인 T-검정으로 평균차이 검정을 수행해야한다.

## (4) 정규분포 시각화

정규분포 검정 결과를 시각화하여 x1 변량의 정규분포 형태를 확인할 수 있다.

## &lt;실습&gt; 정규분포 시각화

```
hist(x1) # x1객체 데이터 분포보기
qqnorm(x1)
qqline(x1, lty=1, col='blue')
```

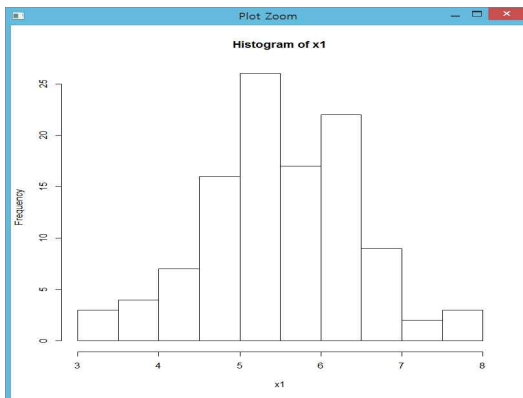


그림 13-3. x1변량의 히스토그램

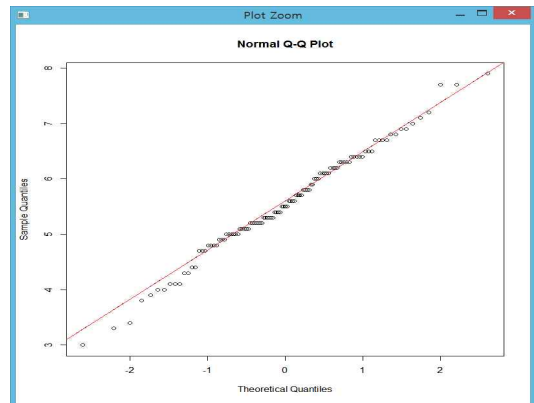


그림 13-4. qqnorm()과 qqline()함수 이용 정규분포 시각화

<해설> 히스토그램을 이용하여 x1객체의 데이터 분포 형태를 확인하면 오른쪽으로 약간 편향된 분포를 보이지만 대체적으로 평균을 중심으로 균등하게 종 모양형태로 나타내고 있다. 또한 stats패키지에서 정규성 검정을 위해서 제공되는 qqnorm()과 qqline()함수를 이용하여 정규분포를 시각화할 수 있다.

## (5) 평균차이 검정

모집단에서 추출한 표본 데이터의 분포 형태가 정규분포 형태를 갖는다면 T-검정을 수행한다. T-검정은 모집단의 평균값을 검정하는 방법으로 stats패키지에서 제공하는 `t.test()` 함수를 이용할 수 있다.

**help(t.test)** # 함수 형식 보기

```
형식) t.test(x, y = NULL,
            alternative = c("two.sided", "less", "greater"),
            mu=0, paired=FALSE, var.equal=FALSE, conf.level = 0.95, ...)
```

`alternative` 속성을 이용하여 양측검정과 단측검정을 할 수 있고, `conf.level` 속성으로 신뢰수준을 지정하여 평균차이 검정을 수행할 수 있다. 또한 <sup>3)</sup>`mu` 속성은 비교할 기존 모집단의 평균값을 지정한다.

## &lt;실습&gt; 단일표본 평균 차이 검정

단계 1 : 양측검정 : x1객체와 기존 모집단의 평균 5.2시간 비교

```
t.test(x1, mu=5.2) # x1 : 표본집단 평균, mu=5.2, 기존 모집단의 평균값
t.test(x1, mu=5.2, alter="two.side", conf.level=0.95)
```

One Sample t-test

```
data: x1
t = 3.9461, df = 108, p-value = 0.0001417
alternative hypothesis: true mean is not equal to 5.2
95 percent confidence interval:
 5.377613 5.736148
sample estimates:
mean of x
 5.556881
```

## &lt;검정결과 해설&gt;

기존 노트북 평균 사용시간 5.2시간과 x1 데이터의 평균을 기준으로 95% 신뢰수준에서 양측검정을 실시한 결과 검정통계량 p-value값은 0.0001417로 유의수준 0.05보다 작기 때문에 국내에서 생산된 노트북과 A회사에서 생산된 노트북의 평균 사용 시간에 차이가 있다고 볼 수 있다.

## &lt;검정통계량 해설&gt;

검정통계량은  $t = 3.9461$ ,  $df = 108$ ,  $p\text{-value} = 0.0001417$ 이며, 95% 신뢰수준

3)  $\mu$ 는 그리스 알파벳의 열두째 글자로  $\mu$ 를 의미한다.

에서 신뢰구간은 5.377613 ~ 5.736148(구간추정)이고,  $x_1$  변수의 평균은 5.556881(점추정)으로 나타났다.

단계 2 : 방향성을 갖는 단측가설 검정

```
t.test(x1, mu=5.2, alter="greater", conf.level=0.95)
```

```
data: x1
```

```
t = 3.9461, df = 108, p-value = 7.083e-05
```

```
alternative hypothesis: true mean is greater than 5.2
```

```
95 percent confidence interval:
```

```
5.406833      Inf
```

```
sample estimates:
```

```
mean of x
```

```
5.556881
```

<검정결과 해설>

기존 노트북 평균 사용시간 5.2시간과  $x_1$  데이터의 평균을 기준으로 95% 신뢰수준에서 ‘국내에서 생산된 노트북 평균 사용 시간 보다 A회사에서 생산된 노트북의 평균 사용 시간이 더 길다’라는 방향성(greater) 갖는 연구가설을 검정한 결과 p-value값은  $7.083 \times 10^{-5}$  (0.00007083)으로 유의수준 0.05보다 매우 작기 때문에 A회사에서 생산된 노트북의 평균 사용 시간이 국내에서 생산된 노트북 평균 사용시간 보다 더 길다고 할 수 있다.

단계 3 : 귀무가설 임계값 계산

stats 패키지에서 제공하는 qt()함수를 이용하면, 귀무가설의 임계값을 확인할 수 있다. 즉 pt()함수에서 p-value와 자유도(df)를 인수로 지정하여 함수를 실행하면 귀무가설을 기각할 수 있는 임계값을 얻을 수 있다.

```
형식) qt(p-value, df)
```

```
qt(7.083e-05, 108) #
```

```
[1] -3.946073
```

<해설>

검정통계량의 p-value와 자유도(df)를 이용하여 귀무가설의 임계값을 계산한 결과 -3.946073으로 나타난다.(임계값은 절대값) 따라서 t 검정통계량이 3.946 이상이면 귀무가설을 기각할 수 있다. 실제  $t=3.946073$ 이기 때문에 귀무가설을 기각할 수 있다.

<실습> T 검정 변수 보기

t.test()함수에 의해서 검정한 결과를 특정 변수에 저장한 후 검정통계량과 관련 정

보를 확인할 수 있다.

```
result <- t.test(x1, mu=5.2, alter="greater", conf.level=0.95)
names(result)
[1] "statistic" "parameter" "p.value"    "conf.int"  "estimate"
[6] "null.value" "alternative" "method"    "data.name"
attach(result)
statistic # t = 3.94606
parameter # df = 108
p.value # p-value = 7.083346e-05
conf.int # conf.level = 0.95
estimate # mean of x = 5.556881
null.value # mean = 5.2
alternative # greater
method # One Sample t-test
data.name : "x1"
detach(result)
```

<해설> result변수에 저장된 칼럼명을 확인하면 각 칼럼에 저장된 T-검정 관련 정보를 확인하면 다음과 같다.

#### (6) 단일집단 t-검정 결과 작성

논문이나 보고서에서 단일표본 평균검정 결과를 제시하기 위해서는 다음과 같은 형식으로 일목요연하게 기술하는 것이 좋다.

#### 【단일집단 t-검정 결과 정리 및 기술】

가설 설정	연구가설(H1) : 국내에서 생산된 노트북과 A회사에서 생산된 노트북의 평균 사용 시간에 차이가 있다.
	귀무가설(H0) : 국내에서 생산된 노트북과 A회사에서 생산된 노트북의 평균 사용 시간에 차이가 없다.
연구환경	국내에서 생산된 노트북 평균 사용 시간이 5.2시간으로 파악된 상황에서 A회사에서 생산된 노트북 평균 사용시간과 차이가 있는지를 검정하기 위해서 A 회사 노트북150대를 랜덤으로 선정하여 검정을 실시한다.
유의수준	$\alpha = 0.05$
분석방법	단일표본 T검정
검정통계량	$t = 3.9461, df = 108$
유의확률	$P = 0.00007083$
결과해석	유의수준 0.05에서 귀무가설이 기각되었다. 따라서 국내에서 생산된 노트북과 A 회사에서 생산된 노트북의 평균 사용 시간에 차이를 보인다고 할 수 있다. 즉 국내에서 생산된 노트북의 평균 사용 시간은 5.2이며, A회사에서 생산된 노트북의 평균 사용 시간은 5.556으로 국내 평균 사용 시간 보다 더 길다고 할 수 있다.

## 2.3 두 집단 검정

독립된 두 집단 간의 비율 차이 검정과 평균 차이 검정에 대해서 알아본다. 비율 차이 검정은 기술통계량으로 빈도수에 대한 비율에 의미가 있으며, 평균 차이는 표본평균에 의미가 있다.

### 2.3.1 두 집단 비율검정



두 집단을 대상으로 비율 차이 검정을 통해서 두 집단의 비율이 같은지 또는 다른지를 검정하는 방법으로 분석절차는 다음과 같다.

분석할 데이터를 대상으로 결측치와 이상치를 제거하는 전처리 과정을 거친 후 비교 대상의 두 집단을 분류하고, 이를 `prop.test()` 함수의 인수로 사용하여 비율 차이 검정을 수행한다.(단일표본 이항분포 비율검정은 `binom.test()` 함수를 이용하지만 독립표본 이항분포 비율검정은 `prop.test()` 함수를 이용한다.)  
비율 차이 검정 통계량을 바탕으로 귀무가설의 기각여부를 결정한다.

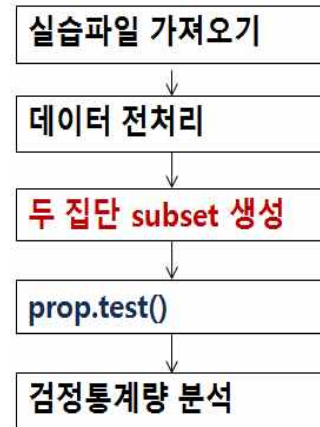


그림 13-5. 두 집단 비율 검정 절차

#### <연구가설>

연구가설( $H_1$ ) : 두 가지 교육방법에 따라 교육생의 만족도에 차이가 있다.  
귀무가설( $H_0$ ) : 두 가지 교육방법에 따라 교육생의 만족도에 차이가 없다.

#### <연구환경>

IT교육센터에서 PT를 이용한 프레젠테이션 교육방법과 실시간 코딩 교육방법을 적용하여 교육을 실시하였다. 2가지 교육방법 중 더 효과적인 교육방법을 조사하기 위해서 교육생 300명을 대상으로 설문을 실시하였다. 조사한 결과는 다음 표13-4와 같다.

표 13-4 교육방법과 만족도 교차분할표

교육방법 \ 만족도	만족	불만족	참가자
PT 교육	110	40	150
코딩교육	135	15	150
합 계	245	55	150

#### (1) 집단별 subset 작성과 교차분석

교육방법에 따라서 두 집단으로 subset을 작성한 후 전처리 과정을 통해서 데이터를 정제한다.

#### <실습> 두 집단 subset 작성과 교차분석 수행

단계 1 : 실습파일 가져오기

```
setwd("c:/Rwork/Part-III")
data <- read.csv("two_sample.csv", header=TRUE)
data
head(data) # 변수명 확인
```

단계 2 : 두 집단 subset 작성 및 데이터 전처리

```
x <- data$method # 교육방법(1, 2) -> NA 없음
y <- data$survey # 만족도(1: 만족, 0:불만족)
```

단계 3 : 집단별 빈도분석

```
table(x) # 교육방법1과 교육방법2 모두 150명 참여
  1   2
150 150
table(y) # 교육방법 만족(1)/불만족(0)
  0   1
55 245
```

단계 4 : 두 변수에 대한 교차분석

```
table(x, y, useNA="ifany") # useNA="ifany" : 결측치 까지 출력
      y
x     0   1
1    40 110
2    15 135
```

<해설> 집단 간의 비율 차이를 분석하기 전에 교육방법과 만족도 칼럼을 추출하고, 빈도분석과 교차분석을 통해서 집단 간의 차이를 검정통계량으로 미리 알아본다.

(2) 두 집단 비율 차이검정

명목척도의 비율을 바탕으로 prop.test()함수를 이용하여 두 집단 간 이항분포의 양측 검정을 통해서 검정 통계량을 구한 후 이를 이용하여 가설을 검정한다. prop.test() 함수의 사용을 위한 형식은 다음과 같다.

```
help(prop.test) # 함수 형식 보기
```

```
형식) prop.test(x, n, p = NULL,
               alternative = c("two.sided", "less", "greater"),
               conf.level = 0.95, correct = TRUE)
```

prop.test()함수의 형식을 적용한 예는 다음과 같다. 첫 번째 벡터는 PT 교육과 코딩 교육 방법에 대한 만족 수 이고, 두 번째 벡터는 두 교육방법에 대한 변량의 길이이다.

```
예) prop.test(c(110,135),c(150,150))
```

<실습> 두 집단 비율 차이 검증

PT 교육방법과 코딩 교육방법에 따른 만족도에 차이가 있는지를 검정한다.

단계 1 : 양측검정

```
prop.test(c(110,135),c(150,150)) # 교육방법에 따른 만족도 차이 검정
prop.test(c(110,135),c(150,150), alternative="two.sided", conf.level=0.95)
2-sample test for equality of proportions with continuity correction
```

```
data: c(110, 135) out of c(150, 150)
X-squared = 1.8237, df = 1, p-value = 0.0003422
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.25884941 -0.07448392
sample estimates:
 prop 1    prop 2
0.7333333 0.9000000
```

<검정 결과 해설>

PT 교육방법과 코딩 교육방법에 따른 만족도에 차이가 있는지를 검정하기 위해서 95% 신뢰수준에서 양측검정을 실시한 결과 검정통계량 p-value값은 0.0003422로 유의수준 0.05보다 작기 때문에 두 교육방법 간의 만족도에 차이가 있다고 볼 수 있다. 즉 「두 가지 교육방법에 따라 교육생의 만족율에 차이가 있다.」 라는 연구 가설이 채택된다.

<검정통계량 해설>

검정통계량은  $X\text{-squared} = 1.8237$ ,  $df = 1$ ,  $p\text{-value} = 0.0003422$ 이며, 95% 신뢰수준에서 신뢰구간은  $-0.25884941 - 0.07448392$ 이고, 첫 번째 교육방법의 비율은 0.7333333, 두 번째 교육방법의 비율은 0.9000000으로 나타났다.

<X-squared 검정통계량으로 가설검정>

신뢰수준 95%에서  $df$ (자유도)가 1이면 X-squared 기각값(3.841)보다 X-squared 검정통계량(1.8237)이 더 크기 때문에 귀무가설을 기각할 수 있다.

단계 2 : 방향성을 갖는 단측가설 검정

첫 번째 교육방법(PT 교육)이 두 번째 교육방법(코딩 교육) 보다 클 것으로 가정하고 방향성(greater)을 갖는 연구가설을 검정한다.

```
prop.test(c(110,135),c(150,150), alter="greater", conf.level=0.95)
```

#### <검정 결과 해설>

PT 교육방법이 코딩 교육방법 보다 만족도가 더 클 것으로 가정하고 95% 신뢰수준에서 방향성(greater)을 갖는 연구가설을 검정한 결과 p-value값은 0.9998로 유의수준 0.05보다 크기 때문에 첫 번째 교육방법인 PT 교육방법이 두 번째 교육방법인 코딩 교육 방법 보다 만족도가 더 크다고 볼 수 없다. 즉, 코딩 교육방법이 PT교육 방법 보다 교육생들에게 만족도가 더 높은 것으로 분석된다.

### 2.3.2 두 집단 평균검정(독립표본 T검정)

두 집단을 대상으로 평균 차이 검정을 통해서 두 집단의 평균이 같은지 또는 다른지를 검정하는 방법으로 분석절차는 다음과 같다.

분석할 데이터를 대상으로 결측치와 이상치를 제거하는 전처리 과정을 거친 후 비교 대상의 두 집단을 분류하고, 평균 차이 검정을 위해서 기술통계량으로 평균을 구한다. 독립표본 평균검정은 두 집단 간 동질성 검증(정규분포 검정) 여부를 판정한 후 결과에 따라서 T-검정 또는 웰콕스(wilcox) 검정을 수행한다. 두 검정 방법의 선택은 단일표본 평균검정과 동일하다. 두 집단 평균차이 검정 통계량을 바탕으로 귀무가설의 기각여부를 결정한다.

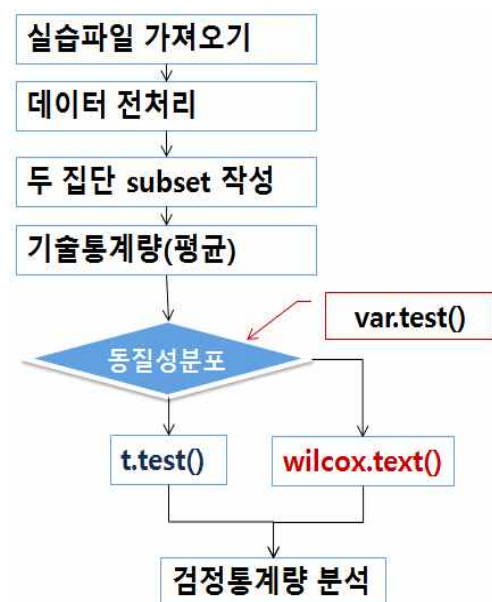


그림 13-6. 두 집단 평균검정 절차

#### <연구가설>

연구가설( $H_1$ ) : 교육방법에 따른 두 집단 간 실기시험의 평균에 차이가 있다.  
 귀무가설( $H_0$ ) : 교육방법에 따른 두 집단 간 실기시험의 평균에 차이가 없다.

## &lt;연구환경&gt;

IT교육센터에서 PT를 이용한 프레젠테이션 교육방법과 실시간 코딩 교육방법을 적용하여 1개월 동안 교육받은 교육생 각 150명을 대상으로 실기시험을 실시하였다. 두 집단간 실기 시험의 평균에 차이가 있는가 검정한다.

## (1) 독립표본 평균 계산

데이터의 전처리 과정을 통해서 outlier를 제거한 후 변수에 대한 대푯값의 성격을 갖는 평균을 계산한다.

## &lt;실습&gt; 독립표본 평균 계산 수행

단계 1 : 실습파일 가져오기

```
data <- read.csv("c:/Rwork/Part-III/two_sample.csv", header=TRUE)
data
head(data) #4개 변수 확인
summary(data) # score - NA's : 73개
```

단계 2 : 두 집단 subset 작성 및 데이터 전처리

```
result <- subset(data, !is.na(score), c(method, score))
# c(method, score) : data의 전체 변수 중 두 변수만 추출
# !is.na(score) : na가 아닌 것만 추출
```

단계 3 : 정제된 데이터를 대상으로 subset 생성

```
result # 방법1과 방법2 혼합됨
length(result$score) # 227
```

단계 4 : 데이터 분리

```
a <- subset(result, method==1) # 교육방법 별로 분리
b <- subset(result, method==2)
```

```
a1 <- a$score # 교육방법에서 점수 추출
b1 <- b$score
```

단계 5 : 기술통계량

```
length(a1); # 109
length(b1); # 118
mean(a1) # 5.556881
```

```
mean(b1) # 5.80339
```

<해설> 각 교육방법 별로 실기시험 점수를 추출하여 평균을 계산하면, 집단별 평균의 차이를 볼 수 있다.

## (2) 동질성 검정

모집단에서 추출된 표본을 대상으로 분산 동질성 검정을 통해서 등분산 가정과 등분산 가정되지 않음(이분산)에 따라서 검정방법이 달라진다. 등분산은 모집단에서 추출된 표본이 균등하게 추출된 경우이고, 이분산은 추출된 표본이 특정 계층으로 편중되어 추출되는 경우이다. 동질성 검정은 stats패키지에서 제공하는 var.test()함수를 이용할 수 있다. 검정 결과가 유의수준 0.05보다 큰 경우 두 집단 간 분포의 모양이 동질하다고 할 수 있다.

동질성 검정의 귀무가설 : 두 집단 간 분포의 모양이 동질적이다.

```
var.test(a1, b1) # a1과 b1 집단 간의 동질성 검정
```

```
F test to compare two variances
```

```
data: a1 and b1
```

```
F = 1.2158, num df = 108, denom df = 117, p-value = 0.3002
```

```
alternative hypothesis: true ratio of variances is not equal to 1
```

```
95 percent confidence interval:
```

```
0.8394729 1.7656728
```

```
sample estimates:
```

```
ratio of variances
```

```
1.215768
```

<해설> 검정통계량 p-value값은 0.3002로 유의수준 0.05보다 크기 때문에 두 집단 간의 분포형태가 동질하다고 볼 수 있다.

## (3) 두 집단 평균 차이검정

두 집단 간의 동질성 검정에서 분포형태가 동질하다고 분석되었기 때문에 t.test() 함수를 이용하여 두 집단 간 평균 차이 검정을 수행한다.

<실습> 두 집단 평균 차이검정 수행

단계 1 : 양측검정

```
t.test(a1, b1)
```

```
t.test(a1, b1, alter="two.sided", conf.int=TRUE, conf.level=0.95)
```

```
# p-value = 0.0411 : 두 집단 간 평균에 차이가 있다.
```

단계 2 : 방향성을 갖는 연구가설 검정

```
t.test(a1, b1, alter="greater", conf.int=TRUE, conf.level=0.95)
# p-value = 0.9794 : a1을 기준으로 비교 -> a1이 b1보다 크지 않다.
t.test(a1, b1, alter="less", conf.int=TRUE, conf.level=0.95)
# p-value = 0.02055 : a1을 기준으로 비교 -> a1이 b1보다 작다.
```

#### <검정 결과 해설>

프레젠테이션 교육방법과 실시간 코딩 교육방법 간의 실기점수의 평균에 차이가 있는지를 검정하기 위해서 95% 신뢰수준에서 양측검정을 실시한 결과 검정통계량  $p$ -value값은 0.0411로 유의수준 0.05보다 작기 때문에 두 집단 간의 평균에 차이가 있는 것으로 나타났다. 또한 방향성을 갖는 연구가설을 수행한 결과 a1 집단의 평균이 b1 집단의 평균보다 더 작은 것으로 나타났다. 따라서 「교육방법에 따른 두 집단 간 실기시험의 평균에 차이가 있다.」 라는 연구가설이 채택된다.

#### (4) 두 집단 평균 차이검정 결과 작성

논문이나 보고서에서 독립표본 평균검정 결과를 제시하기 위해서는 다음과 같은 형식으로 기술한다.

#### 【독립표본 t-검정 결과 정리 및 기술】

가설 설정	연구가설(H1) : 교육방법에 따른 두 집단 간 실기시험의 평균에 차이가 있다.
	귀무가설(H0) : 교육방법에 따른 두 집단 간 실기시험의 평균에 차이가 없다.
연구환경	IT교육센터에서 PT를 이용한 프레젠테이션 교육방법과 실시간 코딩 교육방법을 적용하여 1개월 동안 교육받은 교육생 각 150명을 대상으로 실기시험을 실시하였다. 두 집단간 실기시험의 평균에 차이가 있는가 검정한다.
유의수준	$\alpha = 0.05$
분석방법	독립표본 T검정
검정통계량	$t = -2.0547, df = 218.192$
유의확률	$P = 0.0411$
결과해석	유의수준 0.05에서 귀무가설이 기각되었다. 따라서 교육방법에 따른 두 집단 간 실기시험의 평균에 차이가 있다. 라고 말할 수 있다. 단측검정을 실시한 결과 첫 번째 교육방법이 두 번째 교육방법 보다 크지 않은 것으로 나타났다. 즉 실시간 코딩 교육방법이 교육효과가 더 높은 것으로 분석된다.

#### 2.3.3 대응 두 집단 평균검정(대응표본 T검정)

대응표본 평균검정(Paired Samples t-test)은 동일한 표본을 대상으로 측정된 두 변수의 평균 차이를 검정하는 분석방법이다. 일반적으로 사전검사와 사후검사의 평균 차이를 검증할 때 많이 이용한다.(예 : 교수법 프로그램을 적용하기 전 학생들의 학습력과 교수법 프로그램을 적용한 후 학생들의 학습력에 차이가 있는지를 검정한다)

다.)

분석할 데이터를 대상으로 전처리 과정을 거친 후 전과 후로 두 집단을 분류하고, 평균 차이 검정을 위해서 기술통계량으로 평균을 구한다.

대응표본 평균검정은 독립표본 평균검정 방법과 동일하다.

대응 표본 두 집단 평균차이 검정 통계량을 바탕으로 귀무가설의 기각여부를 결정한다.

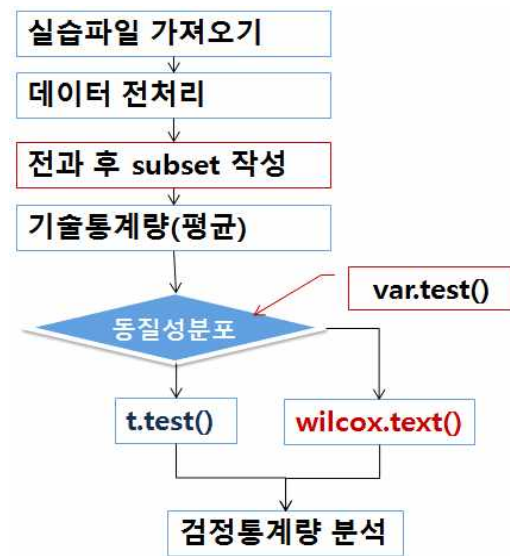


그림 13-7. 대응 두 집단 평균검정 절차

#### <연구가설>

연구가설( $H_1$ ) : 교수법 프로그램을 적용하기 전 학생들의 학습력과 교수법 프로그램을 적용한 후 학생들의 학습력에 차이가 있다.

귀무가설( $H_0$ ) : 교수법 프로그램을 적용하기 전 학생들의 학습력과 교수법 프로그램을 적용한 후 학생들의 학습력에 차이가 없다.

#### <연구환경>

A교육센터에서 교육생 100명을 대상으로 교수법 프로그램 적용 전에 실기시험을 실시한 후 1개월 동안 동일한 교육생에게 교수법 프로그램을 적용한 후 실기시험을 실시한 점수와 평균에 차이가 있는지 검정한다.

#### (1) 대응표본 평균 계산

대응되는 두 집단의 subset을 생성한 후 두 집단 간의 평균 차이 검정을 위해서 집단 간 평균을 계산한다.

#### <실습> 대응표본 평균 계산하기

단계 1 : 실습파일 가져오기

```
setwd("c:/Rwork/Part-III")
```

```
data <- read.csv("paired_sample.csv", header=TRUE)
```

단계 2 : 대응 두 집단 subset 생성



data 객체의 before와 after 칼럼을 대상으로 after 칼럼의 결측치를 제거하여 subset을 생성한다.

```
result <- subset(data, !is.na(after), c(before,after)) # subset 작성
```

```
result # 결측데이터 4개
```

```
# 동일한 사람에게 두 번 질문
```

```
x <- result$before # 교수법 적용 전 점수
```

```
y <- result$after # 교수법 적용 후 점수
```

```
x; y
```

단계 3 : 기술통계량 계산

대응표본인 경우에는 서로 짝을 이루고 있기 때문에 서로 표본수가 같아야한다.

```
length(x) # 96 -> 4개 결측치 제거
```

```
length(y) # 96
```

```
mean(x) # 5.16875
```

```
mean(y) # 6.220833 -> 1.052 정도 증가
```

## (2) 동질성 검정

독립표본의 동질성 검정과 동일하게 stats패키지에서 제공하는 var.test()함수를 이용한다. 또한 검정 결과가 유의수준 0.05보다 큰 경우 두 집단 간 분포의 모양이 동질하다고 할 수 있다.

동질성 검정의 귀무가설 : 두 집단 간 분포의 모양이 동질적이다.

```
var.test(x, y, paired=TRUE)
```

F test to compare two variances

data: x and y

F = 1.0718, num df = 95, denom df = 95, p-value = 0.7361

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.7151477 1.6062992

sample estimates:

ratio of variances

1.071793

<해설> 검정통계량 p-value값은 0.7361로 유의수준 0.05보다 크기 때문에 두 집단 간의 분포형태가 동질하다고 볼 수 있다.

## (3) 대응 두 집단 평균 차이검정

대응되는 두 집단 간의 동질성 검정에서 분포형태가 동질하다고 분석되었기 때문에 `t.test()` 함수를 이용하여 대응 두 집단 간 평균 차이 검정을 수행한다.

<실습> 대응 두 집단 평균 차이검정 수행

단계 1 : 양측검정

```
t.test(x, y, paired=TRUE) # p-value < 2.2e-16
```

단계 2 : 방향성을 갖는 연구가설 검정

```
t.test(x, y, paired=TRUE, alter="greater", conf.int=TRUE, conf.level=0.95)
```

#p-value = 1 -> x을 기준으로 비교 : x가 y보다 크지 않다.

```
t.test(x, y, paired=TRUE, alter="less", conf.int=TRUE, conf.level=0.95)
```

# p-value < 2.2e-16 -> x을 기준으로 비교 : x가 y보다 적다.

<검정 결과 해설>

교수법 프로그램을 적용하기 전 시험성적과 교수법 프로그램을 적용한 후 시험성적의 평균에 차이가 있는지를 검정하기 위해서 95% 신뢰수준에서 양측검정을 실시한 결과 검정통계량 p-value값은 2.2e-16로 유의수준 0.05보다 매우 작기 때문에 두 집단 간의 평균에 차이가 있는 것으로 나타났다. 또한 방향성을 갖는 연구가설을 검정한 결과 x 집단의 평균이 y 집단의 평균보다 더 작은 것으로 나타났다. 따라서 「교수법 프로그램을 적용하기 전 학생들의 학습력과 교수법 프로그램을 적용한 후 학생들의 학습력에 차이가 있다.」라는 연구가설이 채택된다.

<기술통계량 해설>

대푯값의 성격을 갖는 평균 통계량에서 교수법 프로그램 적용 전 평균(5.16875)과 교수법 프로그램 적용 후 평균(6.220833)을 비교한 결과 교수법을 적용한 후 시험성적이 1.052 점수가 향상된 것으로 나타났다.

(4) 대응표본 평균검정 결과 작성

논문이나 보고서에서 대응표본 평균검정 결과를 제시하기 위해서는 다음과 같은 형식으로 기술한다.

**【대응표본 t-검정 결과 정리 및 기술】**

가설 설정	연구가설(H1) : 교수법 프로그램을 적용하기 전 학생들의 학습력과 교수법 프로그램을 적용한 후 학생들의 학습력에 차이가 있다.
	귀무가설(H0) : 교수법 프로그램을 적용하기 전 학생들의 학습력과 교수법 프로그램을 적용한 후 학생들의 학습력에 차이가 없다.
연구환경	A교육센터에서 교육생 100명을 대상으로 교수법 프로그램 적용 전에 실기시험을 실시한 후 1개월 동안 동일한 교육생에게 교수법 프로그램을 적용한 후 실기시험을 실시한 점수와 평균에 차이가 있는가 검정한다.
유의수준	$\alpha = 0.05$
분석방법	대응표본 T검정
검정통계량	$t = -2.6424, df = 95$
유의확률	$P = < 2.2e-16$
결과해석	유의수준 0.05에서 귀무가설이 기각되었다. 따라서 교수법 프로그램 적용 전과 적용 후의 두 집단 간 학습력의 평균에 차이가 있다. 라고 말할 수 있다. 또한 단측검정을 실시한 결과 교수법 프로그램 적용 전 학습력이 교수법 프로그램 적용 후 학습력 보다 크지 않은 것으로 나타났다. 즉 교수법 프로그램이 학습력에 효과가 있는 것으로 분석된다.

## 2.4 세 집단 검정

독립된 세 집단 이상의 집단 간 비율 차이 검정과 평균 차이 검정에 대해서 알아본다. 비율 차이 검정은 기술통계량으로 빈도수에 대한 비율에 의미가 있으며, 세 집단의 평균 차이 검정은 분산분석이라고 한다.

### 2.4.1 세 집단 비율검정

세 집단을 대상으로 비율 차이 검정을 통해서 세 집단 간의 비율이 같은지 또는 다른지를 검정하는 방법으로 분석절차는 다음과 같다.

분석할 데이터를 대상으로 결측치와 이상치를 제거하는 전처리 과정을 거친 후 비교 대상의 세 집단을 분류하고, 이를 `prop.test()`함수의 인수로 사용하여 비율 차이 검정을 수행한다. (두 집단과 세 집단 이상의 비율검정은 `prop.test()` 함수를 이용한다.)

비율 차이 검정 통계량을 바탕으로 귀무가설의 기각여부를 결정한다.

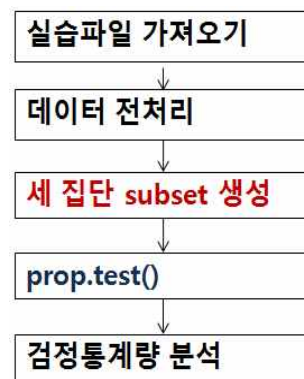


그림 13-8. 세 집단 비율 검정 절차

<연구가설>

연구가설( $H_1$ ) : 세 가지 교육방법에 따른 집단 간 만족율에 차이가 있다.  
 귀무가설( $H_0$ ) : 세 가지 교육방법에 따른 집단 간 만족율에 차이가 없다.

#### <연구환경>

IT교육센터에서 3가지 교육방법을 적용하여 교육을 실시하였다. 3가지 교육방법 중 더 효과적인 교육 방법을 조사하기 위해서 교육생 150명을 대상으로 설문을 실시하였다. 조사한 결과는 다음 표 13-5와 같다.

표 13-5. 교육방법과 만족도 교차분할표

교육방법 \ 만족도	만족	불만족	참가자
PT 교육	34	16	50
코딩교육	37	13	50
혼합교육	39	11	50
합 계	110	40	150

#### (1) 세 집단 subset 작성과 기술통계량 계산

비율검정을 위한 데이터 셋을 대상으로 전처리 과정을 통해서 데이터를 정제한 후 비율 검정에 필요한 기술통계량을 계산한다.

#### <실습> 세 집단 subset 작성과 기술통계량 계산하기

단계 1 : 파일 가져오기

```
setwd("c:/Rwork/Part-III")
data <- read.csv("three_sample.csv", header=TRUE)
head(data)
```

단계 2 : 세 집단 subset 작성(데이터 전처리)

```
method <- data$method # 교육방법
survey <- data$survey # 만족도
method; survey
```

단계 3 : 기술통계량(빈도수)

```
table(method, useNA="ifany") # 세 그룹 모두 관찰치 50개
method
  1  2  3
50 50 50
```

```
table(method, survey, useNA="ifany") # 교육방법과 만족도 교차분할표
```

```
      survey
method 0  1
1  16 34  <- 방법1(불만족16, 만족 34)
2  13 37  <- 방법2(불만족13, 만족 37)
3  11 39  <- 방법3(불만족11, 만족 39)
```

<해설> 비율검정을 위한 데이터 셋을 대상으로 세 집단으로 분류하여 비율 검정에 필요한 기술통계량을 계산한다.

## (2) 세 집단 비율 차이검정

명목척도의 비율을 바탕으로 prop.test() 함수를 이용하여 세 집단 간 이항분포의 양측 검정을 통해서 검정 통계량을 구한 후 이를 이용하여 가설을 검정한다. 세 집단 비율 차이 검정을 위한 prop.test() 함수의 형식을 적용한 예는 다음과 같다. 즉 첫 번째 벡터는 방법1, 방법2, 방법3에 대한 만족 수 이고, 두 번째 벡터는 세 교육방법에 대한 변량의 길이이다.

예) prop.test(c(34,37,39), c(50,50,50))

<실습> 세 집단 비율 차이 검증

세 교육방법의 만족도에 차이가 있는지를 검정한다.

### ■ 양측검정

```
prop.test(c(34,37,39), c(50,50,50))
```

```
prop.test(c(34,37,39), c(50,50,50), alternative="two.sided", conf.level=0.95)
```

3-sample test for equality of proportions without continuity correction

```
data: c(34, 37, 39) out of c(50, 50, 50)
```

```
X-squared = 1.2955, df = 2, p-value = 0.5232
```

```
alternative hypothesis: two.sided
```

```
sample estimates:
```

```
prop 1 prop 2 prop 3
```

```
0.68 0.74 0.78
```

## <검정 결과 해설>

세 교육방법에 따른 만족도에 차이가 있는지를 검정하기 위해서 95% 신뢰수준에서 양측검정을 실시한 결과 검정통계량 p-value 값은 0.5232로 유의수준 0.05보다 크기 때문에 세 교육방법 간의 만족도에 차이가 있다고 볼 수 없다. 즉 「세 가지 교육방법에 따른 집단 간 만족율에 차이가 없다.」라는 귀무가설을 기각할 수 없다.

<X-squared 검정통계량으로 가설검정>

신뢰수준 95%에서  $df$ (자유도)가 2이면 X-squared 기각값(5.991)보다 X-squared 검정통계량(1.2955)이 더 작기 때문에 귀무가설을 기각할 수 없다. 또한 각 교육방법에 따른 만족도의 비율을 68%(prop 1), 74%(prop 2), 78%(prop 3)로 서로 다른 비율의 차이를 나타내고 있다.

#### 2.4.2 분산분석(F 검정)

분산분석(ANOVA Analysis)은 T 검정과 동일하게 평균에 의한 차이 검정방법이다. 차이점은 T 검정이 두 집단 간의 평균차이를 검정했다면 분산분석은 **세 집단 이상**의 평균 차이를 검정한다. (예 : 의학연구 분야에서 개발된 3가지 치료제가 있다고 가정할 때, 이 3가지 치료제의 효과에 차이가 있는지를 검정한다.) 분산분석은 가설검정을 위해 F분포를 따르는 F 통계량을 검정통계량으로 사용하기 때문에 F 검정이라고 한다.

##### 【분산분석 중요사항】

- 1개의 범주형 독립변수와 종속변수간의 관계를 분석하는 일원분산분석과 2개 이상의 독립변수가 종속변수에 미치는 효과를 분석하는 이원분산분석으로 분류한다.
- 독립변수는 명목척도(성별), 종속변수는 등간척도나 비율척도로 구성되어야 한다.
- 마케팅전략의 효과, 소비자 집단의 반응 차이 등과 같이 기업의 의사결정에 도움을 주는 비계량적인 독립변수와 계량적인 종속변수 간의 관계를 파악할 때 이용한다.

세 집단 이상을 대상으로 집단 간의 평균 차이 검정을 수행하는 분산분석(F 검정)을 위한 분석절차는 다음과 같다.

분석할 데이터를 대상으로 전처리 과정을 거친 후 집단 간 차이 검정을 위해서 세 집단을 분류하고, 평균 차이 검정을 위한 기술통계량으로 평균을 구한다.

분산분석에서 집단 간의 동질성 여부를 검정하기 위해서는 `bartlett.test()` 함수를 이용한다. (두 집단은 `var.test()` 함수를 이용하고, 분산분석은 `bartlett.test()` 함수를 이용한다.) 집단 간의 분포가 동질한 경우 분산분석을 수행하는 `aov()` 함수를 이용하며, 그렇지 않은 경우에는 비모수 검정방법인 `kruskal.test()` 함수를 이용하여 분석을 수행한다. 마지막으로 `TukeyHSD()` 함수를 이용하여 사후검정을 수행한다.

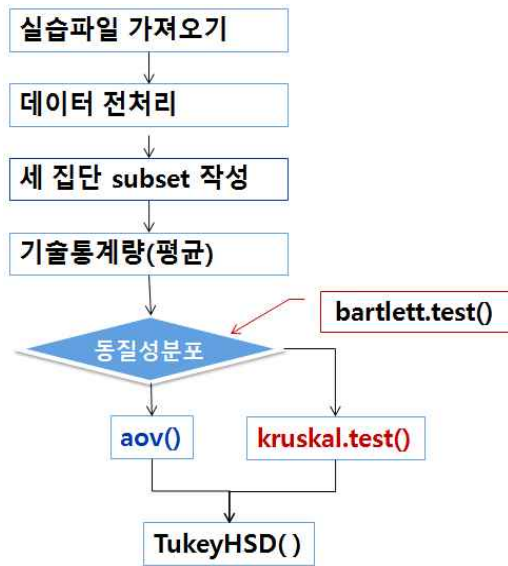


그림 13-9. 분산분석 검정 절차

#### <연구가설>

연구가설( $H_1$ ) : 교육방법에 따른 세 집단 간 실기시험의 평균에 차이가 있다.  
 귀무가설( $H_0$ ) : 교육방법에 따른 세 집단 간 실기시험의 평균에 차이가 없다.

#### <연구환경>

세 가지 교육방법을 적용하여 1개월 동안 교육받은 교육생 각 50명씩을 대상으로 실기시험을 실시하였다. 세 집단 간 실기시험의 평균에 차이가 있는지를 검정한다.

#### (1) 데이터 전처리

분석할 데이터를 대상으로 NA와 outline를 제거하여 데이터를 정제한다.

#### <실습> 데이터 전처리 수행

단계 1 : 실습파일 가져오기

```
setwd("C:/Rwork/Part-III")
```

```
data <- read.csv("three_sample.csv", header=TRUE)
```

```
head(data) # 변수명 확인 -> no method survey score
```

단계 2 : 데이터 전처리 : NA, outline 제거

```
data <- subset(data, !is.na(score), c(method, score))
```

```
head(data) # method, score
```

단계 3 : 차트이용 outlier 보기(데이터 분포 현황 분석)

```
plot(data$score) # 산점도 이용 outlier 확인(50이상 발견)
```

```
barplot(data$score) # 막대 차트 이용 outlier 확인
```

```
mean(data$score) # 평균 통계량 : 14.45
```

단계 4 : 데이터 정제(outlier 제거 : 평균(14) 이상 제거)

```
length(data$score) # outlier 제거 전 관측치 91개
```

```
data2 <- subset(data, score <= 14) # 14이상 제거
```

```
length(data2$score) #88 (3개 제거)
```

단계 5 : 정제된 데이터 확인

```
x <- data2$score
```

```
boxplot(x) # 박스 차트에서 정제 데이터 확인
```

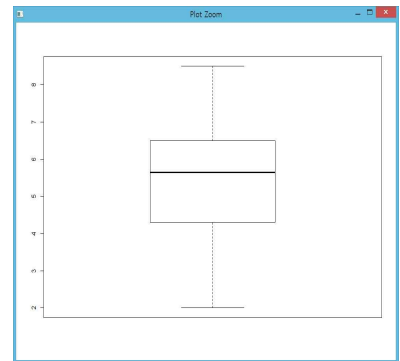


그림 13-10. 박스 차트에서 정제 데이터 확인

<해설> subset()함수를 이용하여 데이터를 전처리하는 과정에서 특정 데이터 셋의 칼럼만 추출이 가능하고, 또한 NA가 포함되지 않은 관측치만 선별하여 데이터를 정제할 수 있다. 정제된 데이터는 박스 차트(그림 13-10)를 이용하여 확인한다.

(2) 세 집단 subset 작성과 기술통계량

<실습> 세 집단 subset 작성과 기술통계량 구하기

단계 1 : 세 집단 subset 작성

```
# 코딩 변경 - 변수 리코딩(method: 1:방법1, 2:방법2, 3:방법3)
```

```
data2$method2[data2$method==1] <- "방법1"
```

```
data2$method2[data2$method==2] <- "방법2"
```

```
data2$method2[data2$method==3] <- "방법3"
```

단계 2 : 교육방법 별 빈도수

```
table(data2$method2) # 교육방법 별 빈도수
```

```
방법1 방법2 방법3
```

```
31 27 30
```

단계 3 : 교육방법을 x변수에 저장



```
x <- table(data2$method2)
```

단계 4 : 교육방법에 따른 시험성적 평균 구하기

```
y <- tapply(data2$score, data2$method2, mean)
```

```
y
```

```
방법1    방법2    방법3
4.187097 6.800000 5.610000
```

단계 5 : 교육방법과 시험성적으로 데이터프레임 생성

```
df <- data.frame(교육방법=x, 성적=y)
```

```
df # 교육방법에 따른 시험성적 평균 교차표
```

```
교육방법.Var1 교육방법.Freq    성적
방법1          방법1          31 4.187097
방법2          방법2          27 6.800000
방법3          방법3          30 5.610000
```

<해설> 교육방법에 따라서 세 집단으로 subset을 작성한 후 각 방법에 대한 빈도수를 구한다.

### (3) 세 집단 간 동질성 검정

분산분석의 동질성 검정은 stats패키지에서 제공하는 bartlett.test()함수를 이용한다. 또한 검정 결과가 유의수준 0.05보다 큰 경우 세 집단 간 분포의 모양이 동질하다고 할 수 있다.

동질성 검정의 귀무가설 : 세 집단 간 분포의 모양이 동질적이다.

세 집단 간 동질성 검정을 수행하기 위한 bartlett.test()함수의 형식은 다음과 같다.

```
형식) bartlett.test(종속변수 ~ 독립변수, data=dataset )
```

<실습> 세 집단 간 동질성 검정 수행

```
bartlett.test(score ~ method, data=data2)
```

```
Bartlett test of homogeneity of variances
```

```
data: score by method2
```

```
Bartlett's K-squared = 3.3157, df = 2, p-value = 0.1905
```

※ 틸드(~)를 이용하여 분석 식을 작성하면 집단별로 subset를 만들지 않고 사용할 수 있다.

<해설> 검정통계량  $p$ -value값은 0.1905로 유의수준 0.05보다 크기 때문에 세 집단 간의 분포형태가 동질하다고 볼 수 있다.

#### (4) 분산분석(세 집단 간 평균 차이검정)

세 집단 간의 동질성 검정에서 분포형태가 동질하다고 분석되었기 때문에 `aov()` 함수를 이용하여 세 집단 간 평균 차이 검정을 수행한다. 만약 동질하지 않은 경우에는 `kruskal.test()` 함수를 이용하여 비모수 검정을 수행한다.

#### <실습> 분산분석 수행

```
help(aov) # 형식) aov(종속변수 ~ 독립변수, data=data set)
result <- aov(score ~ method2, data=data2)
names(result)
```

# `aov()`의 결과값은 `summary()` 함수를 사용해야  $p$ -value값을 확인한다.

```
summary(result)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
method2	2	99.37	49.68	43.58	9.39e-14 ***
Residuals	85	96.90	1.14		

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

#### <검정 결과 해설>

교육방법에 따른 세 집단 간의 실기시험 평균에 차이가 있는지를 검정하기 위해서 95% 신뢰수준에서 양측검정을 실시한 결과 검정통계량  $p$ -value값은  $9.39e-14$ 로 유의수준 0.05보다 매우 작기 때문에 세 교육방법 간의 평균에 차이가 있다고 볼 수 있다. 즉 「교육방법에 따른 세 집단 간 실기시험의 평균에 차이가 있다.」라는 연구가설을 채택한다.

#### <F 검정통계량으로 가설검정>

분산분석에서 신뢰수준 95%에서는  $-1.96 \sim +1.96$ 의 범위가 귀무가설의 채택역이다. 따라서 F 검정통계량이 채택역에 해당하지 않으면 귀무가설을 기각할 수 있다. 현재 F 검정 통계량 43.58은  $\pm 1.96$  보다 크기 때문에 귀무가설을 기각하고, 연구가설이 채택된다. 분산분석에서 F 검정통계량과 유의수준  $\alpha$ 의 관계는 다음 표 13-6과 같다.

표 13-6. 분산분석에서 F 검정통계량과 유의수준  $\alpha$ 의 관계

F값(절대치)	유의수준 $\alpha$ (양측검정 시)
F값(절대치) $\geq 2.58$	$\alpha = 0.01$ (의·생명분야)
F값(절대치) $\geq 1.96$	$\alpha = 0.05$ (사회과학분야)
F값(절대치) $\geq 1.645$	$\alpha = 0.1$ (기타 일반분야)

## (5) 사후검정

분산분석의 결과를 대상으로 각 집단별로 평균의 차에 대한 비교를 통해서 사후검정을 수행할 수 있다.

<실습> 사후검정 수행

```
TukeyHSD(result) # 분산분석의 결과로 사후검정
$method2
```

```

              diff      lwr      upr    p adj
방법2-방법1  2.612903  1.9424342  3.2833723 0.0000000
방법3-방법1  1.422903  0.7705979  2.0752085 0.0000040
방법3-방법2 -1.190000 -1.8656509 -0.5143491 0.0001911
```

# 사후검정 시각화

```
plot(TukeyHSD(result)) # diff : 폭 크기
```

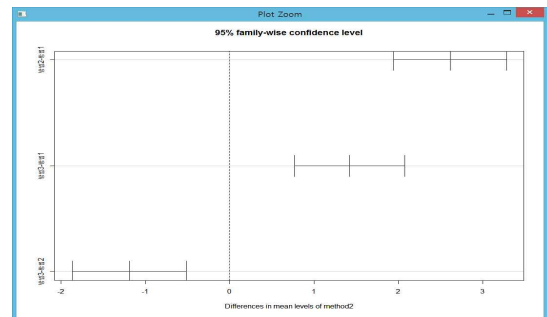


그림 13-11. 사후검정 결과 시각화

<검정 결과 해설>

분산분석의 사후 검정(Post Hoc Tests)은 분산분석에서 ‘3가지 교육방법에 따른 실기시험의 평균에 차이가 있다.’라는 결론을 내렸다면 구체적으로 어떤 교육방법 간에 차이가 있는지는 보여주는 부분이다. 여기서 방법2와 방법1의 집단 간 평균의 차(diff)가 가장 큰 것으로 나타났다.

## (6) 분산분석 검정 결과 작성

논문이나 보고서에서 분산분석 검정 결과를 제시하기 위해서는 다음과 같은 형식으로 기술한다.

## 【분산분석 검정 결과 정리 및 기술】

가설 설정	연구가설(H1) : 교육방법에 따른 세 집단 간 실기시험의 평균에 차이가 있다.
	귀무가설(H0) : 교육방법에 따른 세 집단 간 실기시험의 평균에 차이가 없다.
연구환경	세 가지 교육방법을 적용하여 1개월 동안 교육받은 교육생 각 50명씩을 대상으로 실기시험을 실시하였다. 세 집단간 실기시험의 평균에 차이가 있는가 검정한다.
유의수준	$\alpha = 0.05$
분석방법	ANOVA 검정
검정통계량	$F = 43.58, Df = 2, \text{Sum Sq} = 99.37, \text{Mean Sq} = 49.68$
유의확률	$P = 9.39\text{e-}14 ***$
결과해석	유의수준 0.05에서 귀무가설이 기각되었다. 따라서 교육방법에 따른 세 집단 간 실기시험의 평균에 차이가 있는 것으로 나타났다. 또한 사후검정을 위한 Tukey 분석을 실시한 결과 ‘방법2-방법1’의 평균 점수의 차이가 가장 높은 것으로 나타났다.

## 【제13장 연습문제】

01. 중소기업에서 생산한 HDTV 판매율을 높이기 위해서 프로모션을 진행한 결과 기존 구매비율 보다 15% 향상되었는지를 각 단계별로 분석을 수행하여 검정하시오.

연구가설(H<sub>1</sub>) : 기존 구매비율과 차이가 있다.

귀무가설(H<sub>0</sub>) : 기존 구매비율과 차이가 없다.

조건) 구매여부 변수 : buy (1: 구매하지 않음, 2: 구매)

(1) 데이터셋 가져오기

```
setwd("c:/Rwprk/Part-III")
```

```
hdtv <- read.csv("hdtv.csv", header=TRUE)
```

(2) 빈도수와 비율 계산

(3)가설검정

02. 우리나라 전체 중학교 2학년 여학생 평균 키가 148.5cm로 알려져 있는 상태에서 A중학교 2학년 전체 500명을 대상으로 10%인 50명을 표본으로 선정하여 표본평균신장을 계산하고 모집단의 평균과 차이가 있는지를 각 단계별로 분석을 수행하여 검정하시오.

(1) 데이터셋 가져오기

```
setwd("c:/Rwprk/Part-III")
```

```
stheight <- read.csv("student_height", header=TRUE)
```

```
height <- stheight$height
```

(2) 기술통계량 평균 계산

(3) 정규성 검정

(4) 가설검정

**03. 대학에 진학한 남학생과 여학생을 대상으로 진학한 대학에 대해서 만족도에 차이가 있는가를 검정하시오.**

힌트) 두 집단 비율 차이 검정

조건1) 파일명 : two\_sample.csv, 변수명

조건2) 변수 : gender(1,2), survey(0,1)

**04. 교육방법에 따라 시험성적에 차이가 있는지 검정하시오.**

힌트) 두 집단 평균 차이 검정

조건1) 파일 : twomethod.csv

조건2) 변수 : method : 교육방법, score : 시험성적

조건3) 모델 : 교육방법(명목) -> 시험성적(비율)

조건4) 전처리 : 결측치 제거