13. 집단 간 차이 검정

chap13_Ttest_Anova 수업내용

1) 단일 집단 검정

2) 두 집단 검정

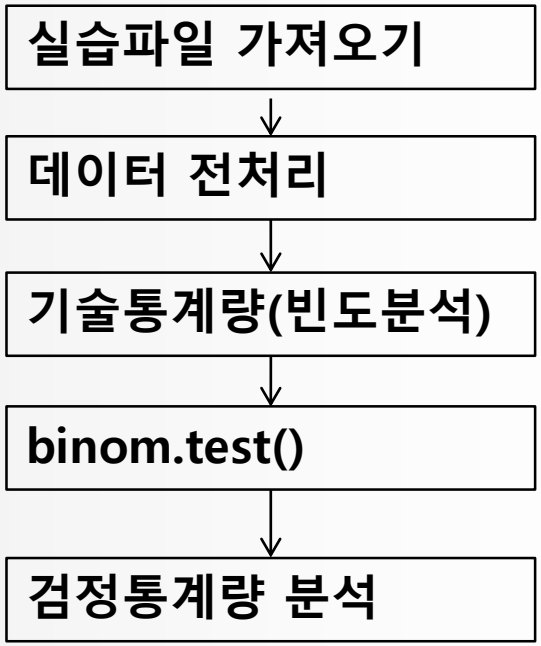
3) 두 집단 이상 검정(분산 분석)

단일집단 비율검정

```
##################################################
# 추론통계학 분석 - 1-1. 단일집단 비율검정
##################################################
 #  방법 : 1개 집단의 비율과 기존 집단과의 비율 차이 분석
 #  작업절차
#       1. 실습데이터 가져오기
#       2. 빈도수와 비율계산
#       3. binom.test() 이용
##################################################
```

단일집단 비율검정

- 분석절차



단일집단 비율검정

<연구가설>

- 연구가설(H_1) : 기존 2014년도 고객 불만율과 2015년도 CS교육 후 불만율에 차이가 있다.
- 귀무가설(H_0) : 기존 2014년도 고객 불만율과 2015년도 CS교육 후 불만율에 차이가 없다.

<연구환경>

2014년도 114 전화번호 안내고객을 대상으로 불만을 갖는 고객은 20%였다. 이를 개선하기 위해서 2015년도 CS교육을 실시한 후 150명 고객을 대상으로 조사한 결과 14명이 불만을 갖고 있었다. 기존 20% 보다 불만율이 낮아졌다고 할 수 있는가?

대상 파일 : c:/Rwork/Part-III/one_sample.csv

해당 변수 : survey(만족도)

변수 척도 : 명목척도(y/n)

가정 : 기존 불만율과 CS교육 후 불만율 분석

단일집단 비율검정

1. 실습데이터 가져오기

```
getwd()

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

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

head(data)

x <- data$survey # 만족도 변수
```

단일집단 비율검정

2. 빈도수와 비율 계산

```
summary(x) # 결측치 없음
length(x) # 150개
table(x)
#x
#  0   1
# 14 136 -> 0:불만족(14), 1: 만족(136)

#table(x, useNA="ifany") # 시리얼 데이터와 NA 개수 출력 시
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
```

단일집단 비율검정

3. 가설검정 : binom.test() 함수 : 명목척도(y/n) 대상

 # 이항분포 개념

1. 정규분포와 마찬가지로 모집단이 가지는 이상적인 분포형

2. 정규분포가 연속변량, 이항분포는 이산변량

3. 그래프는 좌우대칭인 종 모양 곡선

 # binom.test() 함수 이용 가설검정

help(binom.test) # 함수 형식

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

형식) binom.test(만족수, 불만족수, p = 확률)

단일집단 비율검정

1) **만족율 기준 검정**

 # 양측검정

binom.test(c(136,14), p=0.8) # 기존 80% 만족율 기준 검증 실시

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

alternative="two.sided" : 양측검정-> p-value = 0.0006735

해설 : 기존 만족율(80%)과 차이가 있다. -> 연구가설 채택

 # 단측검정

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

alternative="greater" : 단측검정-> 방향성 # p-value = 0.0003179

 # 해설 : CS교육을 통해서 기존 만족율(80%) 이상의 효과를 얻을 수 있다고

 # 볼 수 있다. 따라서 기존 20% 보다 불만율이 낮아졌다고 할 수 있다.

단일집단 비율검정

2) 불만족율 기준 검정

 # 양측검정

binom.test(c(14,136), p=0.2) # 기존 20% 불만족율 기준 검증 실시

binom.test(c(14,136), p=0.2, alternative="two.sided", conf.level=0.95)

alternative="two.sided" : 양측검정-> p-value = 0.0006735

해설 : 기존 불만족율(20%)과 차이가 있다. -> 연구가설 채택

 # 단측검정

binom.test(c(14,136), p=0.2, alternative="greater", conf.level=0.95)

alternative="greater" : 단측검정-> 방향성 # p-value = 0.9999

불만족율 20% 보다 크지 않다.

binom.test(c(14,136), p=0.2, alternative="less", conf.level=0.95)

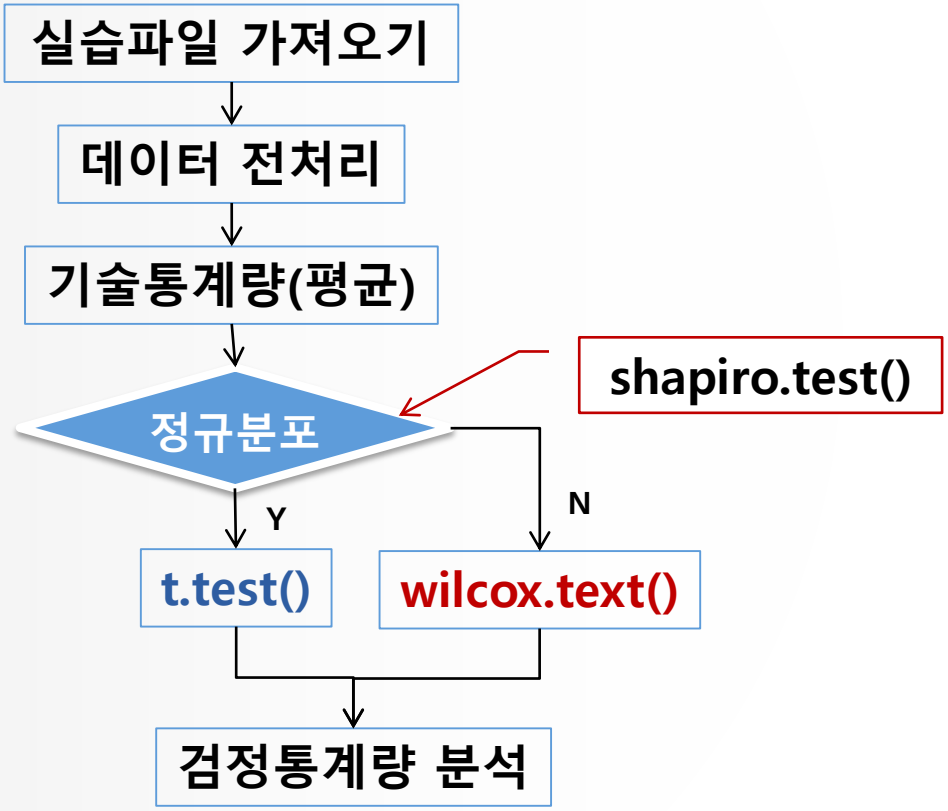
p-value = 0.0003179 -> 불만족율 20% 보다 적다.

단일집단 평균검정

```
##################################################
# 추론통계학 분석 - 1-2. 단일집단 평균 검정(단일표본 T검정)
##################################################
# 방법 : 1개 집단의 평균과 어떤 특정한 값과 차이가 있는지 검증
# 작업절차
#    1. 실습파일 가져오기
#    2. 데이터 분포 및 결측치 제거(데이터 정제)
#    3. 정규분포 검정 : 모집단의 특성 반영 유무
#    4. 가설검정(모수/비모수) -> t.test()/wilcox.text()
##################################################
```

단일집단 평균검정

- 분석절차



단일집단 평균검정

<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 노트북150대를 랜덤으로 선정하여 검정을 실시한다.

대상 파일 : c:/Rwork/Part-III/one_sample.csv

해당 변수 : time

변수 척도 : 비율척도(직접 입력한 수치 데이터)

가정 : 기존 노트북 평균 사용시간 vs A회사 노트북 평균 사용시간

검정 : 노트북 평균 사용시간 수집 -> 평균 -> 정규성 검정 -> T검정

단일집단 평균검정

1. 실습파일 가져오기

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

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

head(data)

x <- data$time # 노트북 사용 시간

head(x)
```

단일집단 평균검정

2. 데이터 분포 /결측치 제거

```
summary(x) # NA-41개
mean(x) # error
mean(x, na.rm=T) # NA 제외 평균(방법1)
# 데이터 정제 -> 5.556881
x1 <- na.omit(x) # NA 제외 평균(방법2)
X1


# 평균(mean) 특징
# 평균 모양 : 양측에 대한 균형
# 대상 : 수치 데이터 -> 비율(ratio)
# 적용 : 평균 차이 검정으로 의사결정
# 평균 검정 : 평균에 의미가 있는가 검정, 평균을 중심으로 종 모양 형성
# 왜도 : 한쪽으로 치우쳐진 정도
```

단일집단 평균검정

3. 정규분포 검정

```
# 정규분포(바른 분포) : 평균에 대한 검정
# 정규분포 검정 귀무가설 : 정규분포와 차이가 없다.
# shapiro학자가 만든 함수 이용 : shapiro.test()
shapiro.test(x1) #  x1 데이터에 대한 정규분포을 검정하는 함수
# W = 0.9914, p-value = 0.7242 <- 정규분포
# 검정결과 분석 : 0.05보다 작으면 정규분포가 아닌 것으로 판단
# 명목척도 -> 보기 항목으로 정규분포가 그려지기 때문에 의미 없음
# 비율척도, 수치 기반 척도(평균에 의미 있는 척도) -> 정규분포 검정 필요

# 정규분포(모수검정) -> t.test()
# 비정규분포(비모수검정) -> wilcox.test()

hist(x1) # 정규분포 형태
```

단일집단 평균검정

4. 가설검정 - 모수/비모수

 # t.test()

- 모집단의 평균값을 검정하는 함수

- 예) 기존평균사용시간 5.2시간 기준으로 검정(같다 vs 차이)

help(t.test)

t -> student에서 t

1) 양측검정

t.test(x1, mu=5.2) # mu(그리스 로마 - 평균) : 기존 5.2시간 기준 검정

x1 : 표본집단 평균, mu=5.2, 모집단의 평균값

 # 정제 데이터와 5.2시간 비교

t.test(x1, mu=5.2, alter="two.side", conf.level=0.95)

 # p-value = 0.0001417

해설 : 평균 사용시간 5.2시간과 차이가 있다.(귀무가설 기각)

단일집단 평균검정

- 점추정 vs 구간추정

```
#alternative hypothesis: true mean is not equal to 5.2
#95 percent confidence interval:
#  5.377613 5.736148  -> 구간추정(95% 신뢰구간 추정)
#sample estimates:
#   mean of x
# 5.556881     -> 점추정 : mean값과 직접비교하여 추정


# 점추정(point) vs 구간추정(interval estimation)
# 점추정 : 모수를 하나의 값으로 추정(평균이나 중위수 사용)
# 구간추정 : 모수가 포함될 것이라고 제시하는 구간추정(신뢰구간)
```

단일집단 평균검정

2) 단측검정

```
t.test(x1, mu=5.2, alter="greater", conf.level=0.95)
# p-value = 7.083e-05 = 0.00007083
# 해설 : A회사 노트북의 평균 사용시간은 5.2시간 보다 더 길다.


# 검정 결과를 변수에 저장하여 특정 변수 확인하기
result <- t.test(x1, mu=5.2, alter="greater", conf.level=0.95)
names(result)
str(result)
result$p.value #  7.083346e-05 -> 세밀한 정보 제공
```

단일집단 평균검정

【단일표본 t-검정 결과 정리 및 기술】

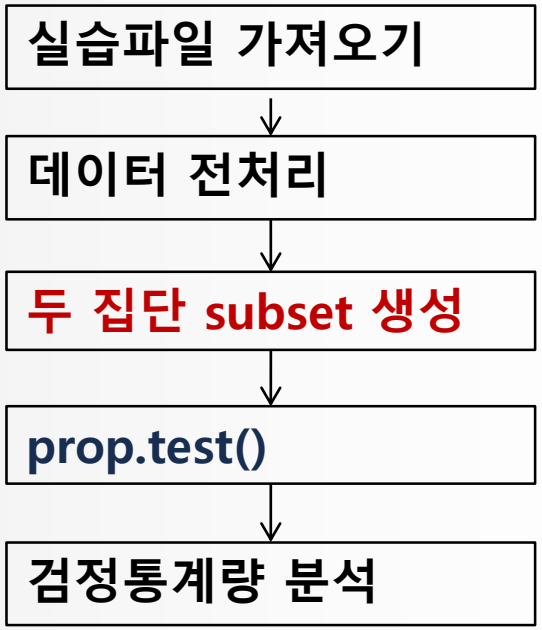
1) 가설 설정	연구가설(H1) : 국내에서 생산된 노트북과 A회사에서 생산된 노트북의 평균 사용 시간에 차이가 있다.
	귀무가설(H0) : 국내에서 생산된 노트북과 A회사에서 생산된 노트북의 평균 사용 시간에 차이가 없다.
2) 연구환경	국내에서 생산된 노트북 평균 사용 시간이 5.2시간으로 파악된 상황에서 A회사에서 생산된 노트북 평균 사용시간과 차이가 있는지를 검정하기 위해서 A회사 노트북150대를 랜덤으로 선정하여 검정을 실시한다.
3) 유의순준	**α = 0.05**
4) 분석방법	**단일표본 T검정**
5) 검정통계량	**t = 3.9461, df = 108**
6) 유의확률	***P* = 0.0001417**
7) 결과해석	**유의수준 0.05에서 귀무가설이 기각되었다. 따라서 국내에서 생산된 노트북과 A회사에서 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 5.56으로 국내 평균 사용 시간 보다 더 길다고 할 수 있다.**

두 집단 비율검정

```
##################################################
# 추론통계학 분석 - 2-1. 두 집단 비율 검정
##################################################
# 방법 : 두 집단 간 비율 차이에 관한 분석
# 작업절차
#     1. 실습파일 가져오기
#     2. 두 집단 subset 작성(데이터 정제,전처리)
#         -> 데이터 정체, 전처리
#         -> 기술통계량 - 빈도수
#         -> 두 변수(집단)에 대한 교차분석
#     3. 두 집단 비율차이 검정
#         -> prop.test()
##################################################
```

두 집단 비율검정

- 분석절차



두 집단 비율검정

<연구가설>

- 연구가설(H_1) : 두 가지 교육방법에 따라 교육생의 만족율에 차이가 있다.
- 귀무가설(H_0) : 두 가지 교육방법에 따라 교육생의 만족율에 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 표와 같다.

대상 파일 : c:/Rwork/Part-III/two_sample.csv
해당 변수 : method(명목척도), survey(명목척도)
변수 척도 : 명목척도 : 빈도수(기술통계량)

두 집단 비율검정

<설문조사 교차표>

교육방법W만족도	만족	불만족	참가자
PT교육	110	40	150
코딩교육	135	15	150
합계	245	55	300

두 집단 비율검정

1. 실습데이터 가져오기

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

두 집단 비율검정

2. 두 집단 subset 작성

```
data$method # 1, 2 -> 노이즈 없음
data$survey # 1(만족), 0(불만족)
# 데이터 정체/전처리
x<- data$method # 교육방법(1, 2) -> 노이즈 없음
y<- data$survey # 만족도(1: 만족, 0:불만족)
x;y
```

두 집단 비율검정

1) 데이터 확인

 # 교육방법 1과 2 모두 150명 참여
table(x) # 1 : 150, 2 : 150
 # 교육방법 만족/불만족
table(y) # 0 : 55, 1 : 245

2) data 전처리 & 기술통계량 -> 빈도수 -> 정규성 검정 필요 없음

두 변수에 대한 교차분석
table(x, y, useNA="ifany") # 결측치 까지 출력

```
################################
#  y
#x    0    1
#  1  40   110  -> 방법A - 110 만족
#  2  15   135  -> 방법B - 135 만족
################################
```

두 집단 비율검정

3. 두 집단 비율차이검증 - prop.test()

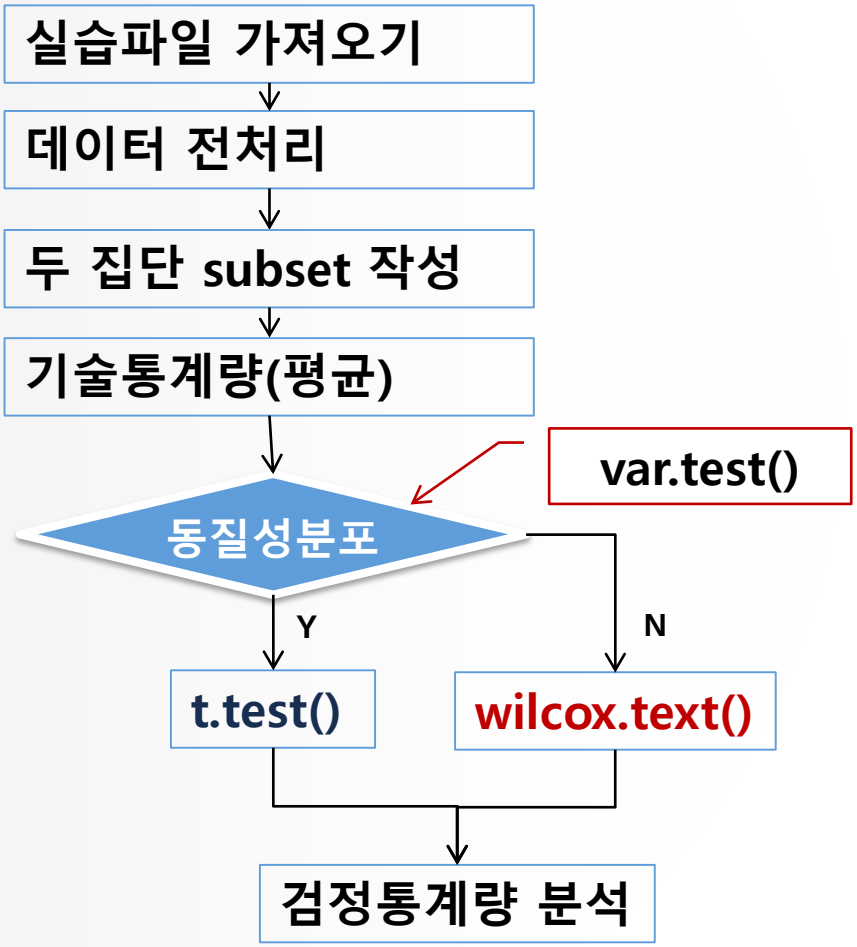
```
help(prop.test) # prop.test(x,n,p, alternative, conf.level, correct)

# 양측검정
prop.test(c(110,135),c(150,150)) # 방법A 만족도와 방법B 만족도 차이 검정
# p-value = 0.0003422
#sample estimates: 집단 간 비율
#  prop 1    prop 2
#0.7333333 0.9000000
prop.test(c(110,135),c(150,150), alternative="two.sided", conf.level=0.95)
# 해설) p-value = 0.0003422 - 두 집단간의 만족도에 차이가 있다.

# 단측검정
prop.test(c(110,135),c(150,150), alter="greater", conf.level=0.95)
# 해설) p-value=0.9998 : 방법A가 방법B에 비해 만족도가 낮은 것으로 파악
```

두 집단 평균검정

```
##################################################
# 추론통계학 분석 - 2-2. 두 집단 평균 검정(독립표본 T검정)
##################################################
# 방법 : 두 집단 간 평균 차이에 관한 분석
# 작업절차
#    1. 실습파일 가져오기
#    2. 두 집단 subset 작성(데이터 정제,전처리)
#    3. 두 집단 간 동질성 검증(정규분포 검정)
#         -> var.test()
#    4. 두 집단 평균 차이검정
#         -> t.test() or wilcox.test()
##################################################
```

두 집단 평균검정

- 분석절차



두 집단 평균검정

<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.

<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.

 # 대상 파일 : c:/Rwork/Part-III/two_sample.csv

 # 해당 변수 : method(명목척도), score(비율척도)

 # 대상 변수 : 교육방법, 시험성적

 # 모형(모델) : 교육방법(A/B) -> 시험성적(비율-성적)

두 집단 평균검정

1. 실습파일 가져오기

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

2. 두 집단 subset 작성(데이터 정제,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
# 위에서 정제된 데이터를 대상으로 subset 생성
result # 방법1과 방법2 혼합됨
length(result$score) # 227
```

두 집단 평균검정

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

2) 교육방법에서 점수 추출
a1 <- a$score
b1 <- b$score

# 기술통계량 -> 평균값 적용 -> 정규성 검정 필요
length(a1); # 109
length(b1); # 118
```

두 집단 평균검정

3. 분포모양 검정 : 두 집단의 분포모양 일치 여부 검정

귀무가설 : 두 집단 간 분포의 모양이 동질적이다.

두 집단간 동질성 비교(분포모양 분석)

var.test(a1, b1) # p-value = 0.3002 -> 차이가 없다.

동질성 분포 : t.test()

비동질성 분포 : wilcox.test()

4. 가설검정 – 두 집단 평균 차이검정

t.test(a1, b1)

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

p-value = 0.0411 - 두 집단간 평균에 차이가 있다.

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 : a1이 b1보다 작다.

두 집단 평균검정

【독립표본 t-검정 결과 정리 및 기술】

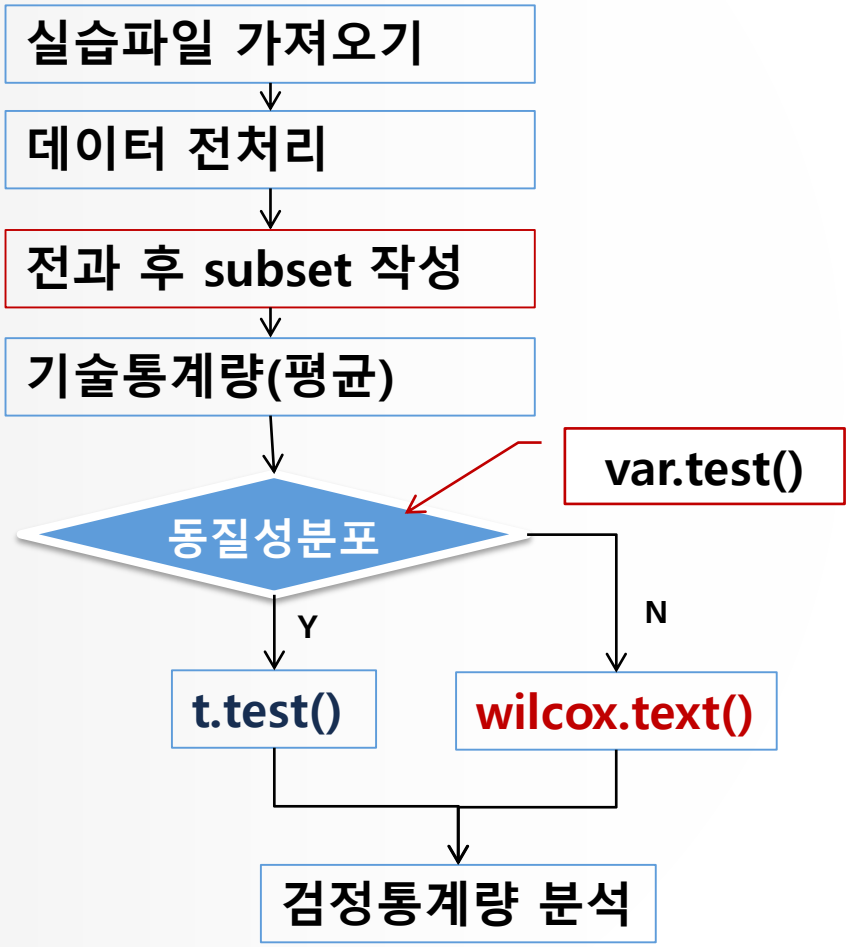
1) 가설 설정	연구가설(H1) : 교육방법에 따른 두 집단 간 실기시험의 평균에 차이가 있다.
	귀무가설(H0) : 교육방법에 따른 두 집단 간 실기시험의 평균에 차이가 있다.
2) 연구환경	IT교육센터에서 PT를 이용한 프레젠테이션 교육방법과 실시간 코딩 교육방법을 적용하여 1개월 동안 교육받은 교육생 각 150명을 대상으로 실기시험을 실시하였다. 두 집단간 실기시험의 평균에 차이가 있는가 검정한다.
3) 유의순준	**α = 0.05**
4) 분석방법	**독립표본 T검정**
5) 검정통계량	**t = -2.0547, df = 218.192**
6) 유의확률	***P* = 0.0411**
7) 결과해석	**유의수준 0.05에서 귀무가설이 기각되었다. 따라서 교육방법에 따른 두 집단 간 실기시험의 평균에 차이가 있다라고 말할 수 있다. 단측검정을 실시한 결과 교육방법1이 교육방법2보다 크지 않은 것으로 나타났디. 즉 실시간 코딩 교육방법이 교육효과가 더 높은 것으로 분석된다.**

대응 두 집단 평균검정

```
##################################################
# 추론통계학 분석 – 2-3. 대응 두 집단 평균 검정(대응표본 T검정)
##################################################
# 방법 : 대응되는 두 집단간 평균 차이에 관한 분석
# 작업절차
#    1. 실습파일 가져오기
#    2. 두 집단 subset 작성(데이터 정제, 전처리)
#    3. 두 집단 간 동질성 검증(정규분포 검정)
#       -> var.test(x,y paired=TRUE)
#    4. 두 집단 평균 차이검정
#       -> t.test(x,y, paired=TRUE)
#       -> wilcox.test(x,y, paired=TRUE)
##################################################
```

대응 두 집단 평균검정

- 분석절차



대응 두 집단 평균검정

1. 실습파일 가져오기

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

2. 두 집단 subset 작성

1) 데이터 정제

```
# subset(x, subset, select, ..) -> subset은 반드시 논리적이어야 함
result <- subset(data, !is.na(after), c(before,after))
# data 테이블을 대상으로 after 결측치 제거하여 subset 생성
result # 결측 데이터 4개
```

대응 두 집단 평균검정

2) 동일한 사람에게 두 번 질문

x <- result$before # 교수법 적용 전 점수

y <- result$after # 교수법 적용 후 점수

x;y # 대응포본인 경우 표본수가 같아야 한다. -> 짝을 이루어야 되기 때문에

length(x) # 96 -> 4개 결측치 제거

length(y) # 96

mean(x) # 5.16875

mean(y) # 6.220833 -> 1.052 정도 증가

3. 분포모양 검정 : 두 집단의 분포모양 일치 여부 검정

var.test(x, y, paired=TRUE) # p-value = 0.7361 -> 차이가 없다.

동질성 분포 : t.test()

비동질성 분포 : wilcox.test()

대응 두 집단 평균검정

4. 가설검정

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

단측검정 - 방향성 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.

<해설>

교수법 프로그램을 적용하기 전 시험성적과 교수법 프로그램을 적용한 후 시험성적을 비교한 결과 교수법을 적용한 후 시험성적이 약 1.052 점수가 향상된 것으로 나타났다.

3) 대응표본 t-검정

【대응표본 t-검정 결과 정리 및 기술】

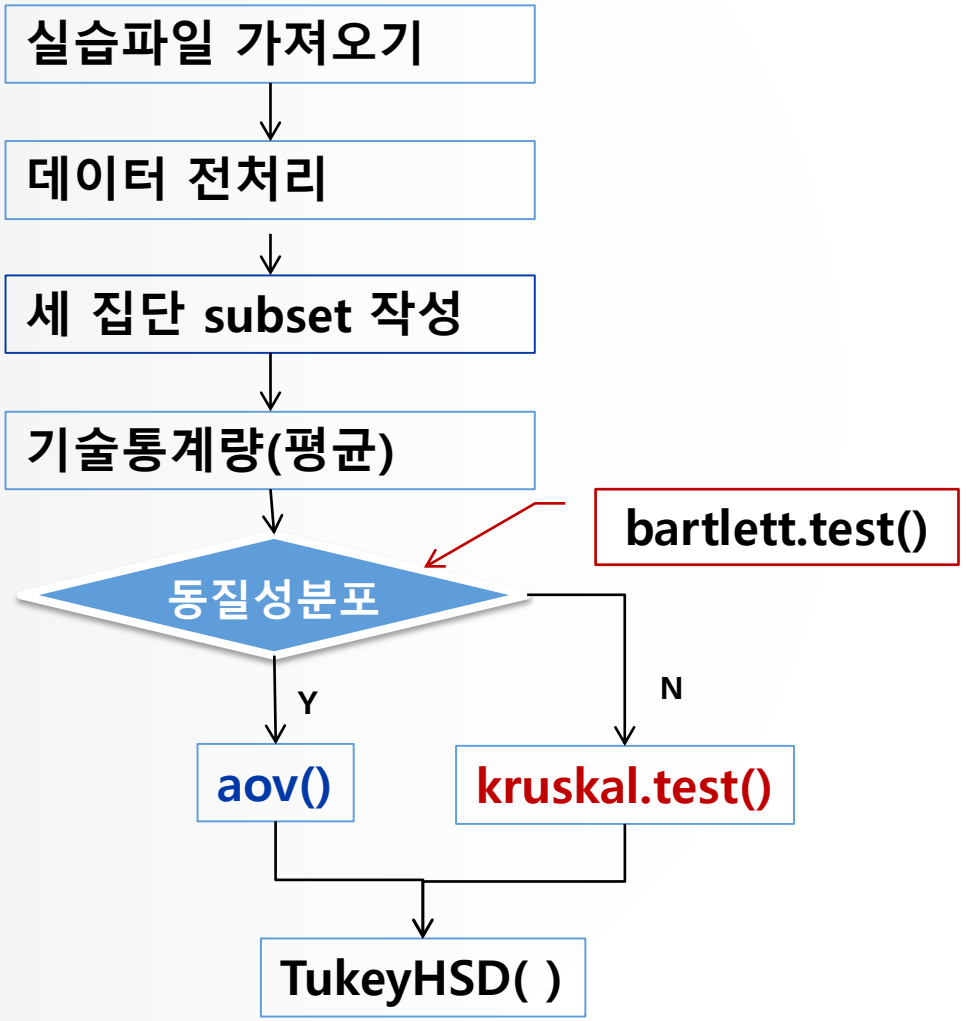
1) 가설 설정	연구가설(H1) : 교수법 프로그램을 적용하기 전 학생들의 학습력과 교수법 프로그램을 적용한 후 학생들의 학습력에 차이가 있다.
	귀무가설(H0) : 교수법 프로그램을 적용하기 전 학생들의 학습력과 교수법 프로그램을 적용한 후 학생들의 학습력에 차이가 없다.
2) 연구환경	A교육센터에서 교육생 100명을 대상으로 교수법 프로그램 적용 전에 실기시험을 실시한 후 1개월 동안 동일한 교육생에게 교수법 프로그램을 적용한 후 실기시험을 실시한 점수와 평균에 차이가 있는가 검정한다.
3) 유의순준	α = 0.05
4) 분석방법	대응표본 T검정
5) 검정통계량	t = -13.6424, df = 95
6) 유의확률	*P* = < 2.2e-16
7) 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 프로그램 이 학습력에 효과가 있는 것으로 분석된다.

두 집단 이상 평균 검정

```
##############################################
# 추론통계학 분석 - 3. 세 집단 평균 검정(분산 분석)
##############################################
#  방법 : 세  집단(이상)간 평균 차이에 관한 분석
#  작업절차
#   1. 파일 가져오기
#   2. 데이터 정제/전처리 - NA, outline 제거
#   3. 세집단 subset 작성
#      -> 코딩 변경
#      -> 기술통계량(빈도수)
#      -> 교차표 작성
#   4. 세집단 동질성 검정 : bartlett.test()
#   5. 분산검정 : aov() or kruskal.test()
#   6. 사후검정 : TukeyHSD()
##############################################
```

두 집단 이상 평균검정

- 분석절차



두 집단 이상 평균검정

<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 집단간 실기시험의 평균에 차이가 있는가 검정한다.

 # 대상 파일 : c:/Rwork/Part-III/two_sample.csv

 # 해당 변수 : method(명목척도), score(비율척도)

 # 대상 변수 : 교육방법, 시험성적

 # 모형(모델) : 교육방법(A/B) -> 시험성적(비율-성적)

두 집단 이상 평균 검정

1. 파일 가져오기

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

2. 데이터 정제/전처리 - NA, outline 제거

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

# 차트이용 - ontline 보기(데이터 분포 현황 분석)
plot(data$score) # 차트로 outline 확인 : 50이상과 음수값
barplot(data$score) # 바 차트
boxplot(data$score) # 박스 차트
mean(data$score) # 14.45
```

세 집단 평균 검정

```
# outline 제거 - 평균(14) 이상 제거
length(data$score)#91
data2 <- subset(data, score <= 14) # 14이상 제거
length(data2$score) #88(3개 제거)

####### 정제된 데이터 보기 ########
x <- data2$score
boxplot(x)
plot(x)
bp <- boxplot(data2$score) # 차트 결과 저장
```

세 집단 평균 검정

3. 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"
table(data2$method2) # 교육방법 별 빈도수
#방법1 방법2 방법3
#  31    27    30

x <- table(data2$method2)
#교육방법에 따른 시험성적 평균 구하기
y <- tapply(data2$score, data2$method2, mean)
#   방법1    방법2    방법3
# 4.187097 6.800000 5.610000
out <- data.frame(교육방법=x, 시험성적=y)
out # 교육방법에 따른 시험성적 평균 교차표
#       교육방법.Var1 교육방법.Freq 시험성적
#방법1         방법1           31   4.187097
#방법2         방법2           27   6.800000
#방법3         방법3           30   5.610000
```

세 집단 평균 검정

4. 동질성 검정 - 정규성 검정

bartlett.test(종속변수 ~ 독립변수) # 독립변수 - 세 집단
bartlett.test(score ~ method, data=data2)
#Bartlett's K-squared = 3.3157, df = 2, p-value = 0.1905

 # data2의 테이블을 대상으로
3집단 이상인 경우 : (종속변수 ~ 독립변수) 분석식으로 표현
~ : 틸드 -> 집단별로 subset를 만들지 않고 사용하도록 편의성 제공

귀무가설 : 세 집단 간 분포의 모양이 동질적이다.
해설 : 유의수준 크기 때문에 귀무가설을 기각할 수 없다.

동질한 경우 aov() 사용 : aov - Analysis of Variance(분산분석)
동질하지 않은 경우 - kruskal.test()

세 집단 평균 검정

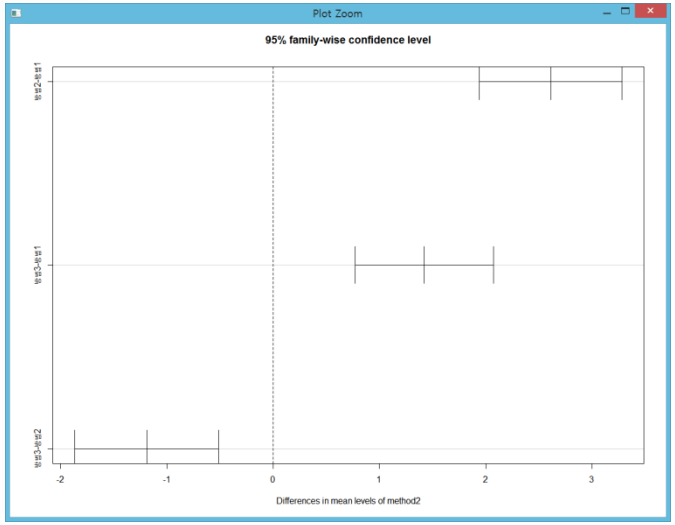
5. 분산검정

```
help(aov)
# 분산분석 결과를 result에 저장
# 귀무가설 : 세 집단의 평균에 차이가 없다.

data2$method2 <- factor(data2$method2)
# factor() : method가 집단 구성변수라는 것을 명시

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

# aov()의 결과값은 summary()함수를 사용해야 p-value 확인
summary(result) # Pr(>F) : 9.39e-14 -> 귀무가설 기각
# 해설 : 0.05보다 현저하게 작음
# 교육방법에 따라서 시험성적 평균에 차이가 있다.
```

세 집단 평균 검정

6. 사후검정

```
# 집단간 차이 상세보기 -> A!=B!=C, A==B!=C, A!=B==C
TukeyHSD(result) # 분석분석의 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
# 교육방법 간 비교  -> p값(tapply 차이 검정) -> 4.187097 6.800000 5.610000

# 해석) A B C 집단간 모두 차이가 있다.
plot(TukeyHSD(result))
 # 그래프 보기(lwr~upr변수 이용)
```



두 집단 이상 평균 검정

【분산분석 결과 정리 및 기술】

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