Ch13 집단간 차이분석

1. 추정과 검정

1.1 점 추정과 구간 추정

[표13.1] 점추정과 신뢰구간 추정

1.2 모평균의 구간 추정

표본 평균이 따르는 분포

표본의 크기 n이 충분히 클 때 (n>= 30)

[표 13.2] 신뢰도와 모평균 신뢰구간

실습 (우리나라 중학교 2학년 남학생의 평균 신장 표본조사)

```
N = 10000
X = 165.1
S = 2
low <- X -1.96 * S / sqrt(N)
high <- X + 1.96 # S / sqrt(N)
low; high
```

실습 (신뢰구간으로 표본오차 구하기)

```
high - X
(low - X) * 100
(high - X) * 100
```

1.3 모비율의 구간추정

[표 13.3] 신뢰도와 모비율 신뢰구간

2. 단일 집단 검정

한 개의 집단과 기존 집단과의 비율 차이 검정과 평균 차이 검정
비율차이 검정: 빈도수에 대한 비율에 의미
평균 차이 검정: 표본 평균에 의미

2.1 단일집단 비율 검정

연구가설

(1) 단일 표본 대상 기술통계량

실습 (단일 표본 빈도수와 비율계산)

1단계: 실습 데이터 가져오기

```
setwd("C:/Rwork/ ")
data <- read.csv("one_sample.csv", header = TRUE)
head(data)

x <- data$survey
```

2단계: 빈도수와 비율계산

```
summary(x)
length(x)
table(x)
```

3단계: 패키지를 이용하여 빈도수와 비율 계산

```
install.packages("prettyR")
library(prettyR)
freq(x)
```

prettyR 패키지

(2) 이항분포 비율 검정

binom.test()함수

정규분포 vs 이항분포

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

실습 (불만율 기준 비율검정)

1단계: 양측 검정

```
binom.test(14, 150, p = 0.2)
binom.test(14, 150, p = 0.2, alternative = "two.sided", conf.level = 0.95)
```

양측 검정 결과는 기존 불만율보다 '크다' 또는 '작다'는 방향성은 제시되지 않음. 따라서 방향성을 갖는 단측 가설 검정을 통해 기존 집단과 비교하여 신규 집단의 불만율이 개선되었는지를 확인해야 한다.

2단계: 방향성을 갖는 단측 가설 검정

1) 2020년 불만율 > 2019년 불만율
2) 2020년 불만율 < 2019년 불만율

```
binom.test(c(14, 150), p = 0.2,
           alternative = "greater", conf.level = 0.95)

binom.test(c(14, 150), p = 0.2,
           alternative = "less", conf.level = 0.95)
```

2.2 단일 집단 평균 검정(단일 표본 T-검정)

단일 집단의 평균이 어떤 특정한 집단의 평균과 차이가 있는지를 검정

평균차이검정은 정규분포 여부를 판정한 후 결과에 따라 T-검정 또는 윌콕스(Wilcox) 검정을 시행.
정규분포이면 모수 검정인 T-검정
정규분포가 아닌경우 비모수 검정인 윌콕스(Wilcox)검정으로 평균차이 검정을 시행

(1) 단일 표본평균 계산
outlier를 제거한 후 평균 계산

실습 (단일 표본 평균 계산하기)
1단계: 실습파일 가져오기

```
setwd("C:/Rwork/")
data <- read.csv("one_sample.csv", header = TRUE)
str(data)
head(data)

x <- data$time
head(x)
```

2단계: 데이터 분포 확인/결측치 제거

```
summary(x)
mean(x)
```

3단계: 데이터 정제

```
mean(x, na.rm = T)
x1 <- na.omit(x)
mean(x1)
```

na.omit()함수: na제외 함수

(2) 평균 검정 통계량의 특징

(3) 정규분포 검정

단일 표본평균 차이 검정을 수행하기 전에 데이터의 분포 형태가 정규분포 인지를 먼저 검정해야 한다. (정규성 검정)

정규분포 검정은 stats패키지에서 제공하는 shapiro.test()함수 사용
검정 결과가 유의수준 0.05보다 큰 경우 정규분포

실습 (정규분포 검정)

```
shapiro.test(x1)
```

(4) 정규분포 시각화

정규분포 검정 결과를 시각화하여 x1변량의 정규분포 형태 확인

실습 (정규분포 시각화)

```
par(mfrow = c(1, 2))
hist(x1)
```

Normal Q-Q(Quantile-Quantile) Plot
qqnorm()함수
qqline()함수

(5) 평균 차이 검정

모집단에서 추출한 표본 데이터의 분포 형태가 정규분포 형태를 가지면, T-검정을 수행
T-검정은 모집단의 평균값을 검정
stats패키지에서 제공하는 t.test()함수 이용

형식: t.test(x, y=NULL, alternative=c("two.sided", "less", "greater"), mu=0, paired=FALSE, var.equal=FALSE, conf.level=0.95,...)
where
alternavtive속성: 양측검정 또는 단측 검정 수행 가능
conf.level속성: 신뢰수준 지정

mu속성: 비교할 기존 모집단의 평균값 지정

실습 (단일 표본평균 차이 검정)

1단계: 양측 검정 – x1객체의 기존 모집단의 평균 5.2시간 비교

```
t.test(x1, mu = 5.2)
qqnorm(x1)
qqline(x1, lty = 1, col = "blue")
t.test(x1, mu = 5.2, alter = "two.side", conf.level = 0.95)
```

2단계: 방향성을 갖는 단측 가설 검정

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

3단계: 귀무가설의 임계값 계산

```
qt(7.08e-05, 108)
```

stats패키지에서 제공하는 qt()함수: 귀무가설의 임계값(귀무가설을 기각할 수 있는 임계값 계산)

형식: qt(p-value, df, lower.tail=T)
Where
Lower.tail=T (default): P(X <=x)
Lower.tail=F : P(X > x)

통계 분포 함수 내 접두어
d(ensity): 확률 밀도 함수 값 구하기 P{X=x}
p(probability): 누적분포 함수에 의한 누적확률 구하기 P(X < x)
q(uantile): 누적확률에 해당하는 분위 구하기
r(andom): 난수생성

(6) 단일 집단 T-검정 결과 작성

[표 13.4] 단일 집단 T-검정 결과 정리 및 기술

3. 두 집단 검정

독립된 두 집단 간의 비율 차이 검정과 평균 차이 검정
비율차이 검정: 기술통계량으로 빈도수에 대한 비율에 의미
평균 차이 검정: 표본평균에 의미

3.1 두 집단 비율 검정

두 집단의 비율이 같은지 또는 다른지를 검정
Prop.test()함수 이용 비율차이 검정

단일표본 이항분포 비율 검정: binom.test() 함수 이용
독립 표본 이항분포 비율 검정: prop.test() 함수 이용

귀무가설, 연구가설

[표 13.5] 교육 방법과 만족도 교차분할표

(1) 집단별 subset작성과 교차분석
Subset 구성, 전처리 과정을 통해 데이터 정제

실습 (두 집단의 subset 작성과 교차분석 수행)

1단계: 파일 가져오기

```
setwd("C:/Rwork/ ")
data <- read.csv("two_sample.csv", header = TRUE)
head(data)
```

2단계: 두 집단의 subset 작성 및 데이터 전처리

```
x <- data$method
y <- data$survey
```

3단계: 집단별 빈도분석

```
table(x)
table(y)
```

4단계: 두 변수에 대한 교차분석

```
table(x, y, useNA = "ifany")
```

useNA ="ifany" : 결측치까지 출력

(2) 두 집단 비율 차이 검정

명목척도의 비율을 바탕으로 prop.test()함수를 이용하여 두 집단 간 이항분포의 양측 검정을 통해서 검정 통ㄱ메량을 구한 후 이를 이용하여 가설 검정

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

실습 (두 집단 비율 차이 검정)
PT교육 방법과 코딩 교육 방법에 따른 만족도에 차이가 있는지 검정

1단계: 양측검정

```
prop.test(c(110, 135), c(150, 150),
          alternative = "two.sided", conf.level = 0.95)
```

2단계: 방향성을 갖는 단측 가설 검정

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

3.2 두 집단 평균 검정 (독립 표본 T-검정)

두 집단을 대상으로 평균 차이 검정을 통해서 두 집단의 평균이 같은지 또는 다른지를 검정

독립 표본평균 검정은 두 집단간 분산의 동질성 검증(정규성 검정)여부를 판정한 후 정규분포이면 T-검정, 정규분포가 아니면 윌콕스(Wilcox)검정을 수행

두 검정 방법의 선택은 단일 표본평균 검정과 동일

연구가설
H0: 평균에 차이가 없다.
H1: 평균에 차이가 있다.

(1) 독립 표본평균 계산

Outlier 제거 후 독립 표본평균 계산

실습 (독립 표본 평균 계산)

1단계: 파일 가져오기

```
data <- read.csv("C:/Rwork/Part-III/two_sample.csv", header = TRUE)
head(data)
summary(data)
```

2단계: 두 집단의 subset작성 및 데이터 전처리

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

!is.na() 속성: NA가 아닌 것만 추출

3단계: 데이터 분리

```
a <- subset(result, method == 1)
b <- subset(result, method == 2)
a1 <- a$score
b1 <- b$score
```

4단계: 기술통계량

```
length(a1)
length(b1)
mean(a1)
mean(b1)
```

(2) 동질성 검정

모집단에서 추출된 표본을 대상으로 분산의 동질성 검정

var.test() 함수: F검정 이용

분산의 동질성 검정
분산의 동질성 검정은 등분산 가정과 등분산 가정되지 않음(이분산)에 따라서 결과가 상이
등분산: 모집단에서 추출된 표본이 균등하게 추출된 경우
이분산: 추출된 표본이 특정 계층으로 편중되어 추출되는 경우

실습 (두 집단 간의 동질성 검정)

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

(3) 두 집단 평균 차이 검정

두 집단 간의 동질성 검정에서 분포의 형태가 동질하다고 분석 → t.test()함수 이용 두 집단 간 평균 차이 검정

실습 (두 집단 평균 차이 검정)

1단계: 양측검정

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

2 sample t-test

2단계: 방향성을 갖는 단측 가설 검정

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

alter = “greater” case
alter = “less” case

(4) 두 집단 평균 차이 검정 결과 작성

[표 13.6] 독립 표본 t-검정 결과 정리 및 기술

3.3 대응 두 집단 평균 검정 (대응 표본 T-검정)

대응 표본평균 검정(Paired samples t-test): 동일한 표본을 대상으로 측정된 두 변수의 평균 차이를 검정
사전검사와 사후검사의 평균 차이를 검증할 때 많이 사용
예) 교수법 프로그램을 적용하기 전 학생들의 학습력과 교수법 프로그램 적용 후 학생들의 학습력에 차이가 있는지 검정

대응 표본평균 검정은 독립 표본평균 검정 방법과 동일

연구가설
H0: 학생들의 학습력 차이가 없다.
H1: 학생들의 학습력 차이가 있다.

(1) 대응 표본평균 계산

대응하는 두 집단의 subset을 생성한 후 두 집단 간의 평균 차이 검정을 위해서 집단 간 평균 계산

실습 (대응 표본평균 계산)

1단계: 파일 가져오기

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

2단계: 대응 두집단 subset 생성

```
result <- subset(data, !is.na(after), c(before, after))
x <- result$before
y <- result$after
x; y
```

3단계: 기술통계량 계산

```
length(x)
length(y)
```

mean(x)
mean(y)

* 대응 표본은 짝을 이루기 때문에 서로 표본 수가 같아야 한다.

(2) 동질성 검정

독립 표본의 동질성 검정과 동일하게 var.test()함수 이용

# 실습: 대응표본의 동질성 검정
var.test(x, y, pared = TRUE)

검정 결과가 유의수준 0.05보다 큰 경우 두 집단 간 분포의 모양이 동질

(3) 대응 두 집단 평균 차이 검정

대응되는 두 집단 간의 동질성 검정에서 분포 형태가 동질하다고 분석되었기 때문에 t.test()함수 이용 대응 두 집단 간 평균 차이 검정

실습: 대응 두 집단 평균 차이 검정

1단계: 양측 검정

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

paired=TRUE 속성 추가

2단계: 방향성을 갖는 단측 가설 검정

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

t.test(x, y, paired = TRUE,

```
        alter = "less",
        conf.int = TRUE, conf.level = 0.95)
```

alter = “greater” case
alter = “less” case

(4) 대응 표본평균 검정 결과 작성

[표 13.7] 대응 표본 t-검정 결과 정리 및 기술

4. 세 집단 검정

독립된 세 집단 이상의 집단 간 비율 차이 검정과 평균 차이 검정

4.1 세 집단 비율 검정

세 집단을 대상으로 비율 차이 검정을 통해서 세 집단 간의 비율이 같은지 또는 다른지를 검정

데이터 대상 전처리 과정(결측치와 이상치 제거) 후 비교대상의 세 집단 분류 후 prop.test()함수로 비율 차이 검정
두 집단 및 세 집단 이상의 비율 검정은 prop.test()함수 이용

연구가설
H0: 세 가지 교육 방법에 따른 집단 간 만족율에 차이가 없다.
H1: 세 가지 교육 방법에 따른 집단 간 만족율에 차이가 있다.

(1) 세 집단 subset 작성과 기술통계량 계산

실습 (세 집단 subset작성과 r기술통계량 계산)

1단계: 파일 가져오기

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

2단계: 세집단 subset 작성(데이터 전처리)

```
method <- data$method
survey <- data$survey
method; survey
```

3단계: 기술통계량(빈도수)

```
table(method, useNA = "ifany")
table(method, survey, useNA = "ifany")
```

(2) 세 집단 비율 차이 검정
세 집단 간 이항분포의 양측 검정을 통해서 가설 검정

예) prop.test(c(34, 37, 39), c(50, 50, 50))
세 교육 방법에 대한 변량의 길이(성공횟수), 방법에 대한 만족 수(시행횟수)

실습 (세 집단 비율 차이 검정)

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

4.2 분산분석(F-검정)

분산분석(ANOVA Analysis):
T-검정과 동일하게 평균에 의한 차이 검정 방법
두 집단 이상의 평균 차이를 검정

만일 ANOVA 가 아닌 여러 번 t 검정을 하면 안되나?
→ 1종 오류가능성이 증대됨

예) 세 집단을 비교하기 위해서는 세 번의 독립표본 t 검정을 수행하여야 함.
각 t 검정에서 유의수준을 0.05로 설정하였다면, 세 번 모두 귀무가설이 맞는데 귀무가설을 기각하지 않은 옳은 결정을 할 확률은
0.95 × 0.95 × 0.95 = 0.86, 1종 오류는 0.14

m개의 집단을 두 개씩 t 검정을 수행한다면 1종오류는?
→$1-(1-\alpha)^{m(m-1)/2}$

분산분석에서 집단 간의 동질성 여부를 검정하기 위해서 barlett.test()함수 사용
cf) 두 집단 간 동질성 검정 시 var.test()함수 사용, 분석분석은 barlett.test()함수 사용

집단 간의 분포가 동질한 경우 분산분석을 위해 aov()함수 사용
집단 간의 분포가 동질하지 않는 경우 kruskal.test()함수 사용

사후비교(post hoc comparisons)
• 귀무가설이 기각된 경우 어떤 집단간 차이가 있는지 2개씩 짝지어 차이를 분석함
• m개의 집단이면 $_mC_2$ 번 비교

• 집단 i와 집단 j를 비교하는 경우
$H_0 : \mu_i = \mu_j$ $H_1 : \mu_i \neq \mu_j$

사후비교 방법
• Least Squate Difference method: 최소유의차, 검정력이 낮음
• Bonferroni's method LSD: 방법을 보완
• Scheffe's method: 사회과학에서 많이 사용
• Tukey's method : 검정력이 높음

TukeyHSD()함수를 이용하여 사후검정 수행

연구가설

H0: 교육 방법에 따른 세 집단 간 실기시험의 평균에 차이가 없다.
H1: 교육 방법에 따른 세 집단 간 실기시험의 평균에 차이가 있다.

(1) 데이터 전처리
데이터 정제 (NA, 이상치 제거)

실습 (데이터 전처리 수행)

1단계: 파일 가져오기

```
data <- read.csv("C:/Rwork/three_sample.csv")
head(data)
```

2단계: 데이터 전처리 (NA, 이상치 제거)

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

3단계: 차트이용 outlier보기(데이터 분포 현황 분석)

```
par(mfrow = c(1, 2))
plot(data$score)
barplot(data$score)
mean(data$score)
```

4단계: 데이터 정제(이상치 제거, 평균(14)이상 제거)

```
length(data$score)
data2 <- subset(data, score <= 14)
length(data2$score)
```

5단계: 정제된 데이터 확인

```
x <- data2$score
par(mfrow = c(1, 1))
boxplot(x)
```

(2) 세 집단 subset작성과 기술 통계량

실습 (세 집단 subset작성과 기술통계량 구하기)
1단계: 세집단 subset 작성

```
data2$method2[data2$method == 1] <- "방법1"
data2$method2[data2$method == 2] <- "방법2"
data2$method2[data2$method == 3] <- "방법3"
```

2단계: 교육 방법별 빈도수

```
table(data2$method2)
```

3단계: 교육 방법을 x변수에 저장

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

4단계: 교육 방법에 따른 시험성적 평균 구하기

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

lappy()함수

5단계: 교육방법과 시험성적으로 데이터프레임 생성

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

(3) 세 집단 간 동질성 검정

barlett.test()함수 사용
검정 결과가 유의수준 0.05보다 큰 경우 세 집단 간 분포의 모양이 동질하다고 할 수 있다.

형식: barlett.test(종속변수 ~ 독립변수, data=dataset)

실습 (세 집단간 동질성 검정 수행)

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

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

(4) 분산분석 (세 집단 간 평균 차이 검정)

세 집단 간의 동질성 검정에서 분포 형태가 동질하다고 분석되었기 때문에 aov()함수를 이용하여 세 집단 간 평균 차이 검정
동질하지 않다면 kruskal.test()함수 이용하여 비모수 검정을 수행

실습 (분산분석 수행)

```
help(aov)
result <- aov(score ~ method2, data = data2)
names(result)

summary(result)
```

p-value 이용 검정

(5) 사후검정

집단별로 평균의 차에 대한 비교를 통해 사후검정을 수행

실습 (사후검정 수행)

1단계: 분산분석 결과에 대한 사후검정

```
TukeyHSD(result)
```

2단계: 사후검정 시각화

```
plot(TukeyHSD(result))
```

(6) 분산분석 검정 결과 작성

[표 13.10] 분산분석 검정 결과 정리 및 기술

연습문제 풀기