Data Analytics - Midterm

광고홍보학과 14학번 송현지

2019년 4월 23일

## 1번 ~ 4번

**1번**

중앙집중치는 평균, 중앙값, 최빈값을 구한다.

* 평균 : 32.42
* 중앙값 : 32.50
* 최빈값은 32라고 나오지만, 모든 수의 빈도가 1이므로 이 경우 최빈값은 없다.

region <- c(32, 37, 38, 30, 42, 35, 26, 28, 36, 33, 27, 25)  
df <- data.frame(region)  
df

## region  
## 1 32  
## 2 37  
## 3 38  
## 4 30  
## 5 42  
## 6 35  
## 7 26  
## 8 28  
## 9 36  
## 10 33  
## 11 27  
## 12 25

# 평균, 중앙값  
summary(df)

## region   
## Min. :25.00   
## 1st Qu.:27.75   
## Median :32.50   
## Mean :32.42   
## 3rd Qu.:36.25   
## Max. :42.00

# 최빈값 함수  
getMode <- function(v) {  
 uniqv <- unique(v)  
 uniqv[which.max(tabulate(match(v, uniqv)))]  
}  
  
# 최빈값  
getMode(df$region)

## [1] 32

**2번**

* 변량 : 28.62879
* 표준편차 : 5.350588

# 변량   
var(df$region)

## [1] 28.62879

# 표준편차   
sd(df$region)

## [1] 5.350588

**3번**

* 답 : 용산구, 서대문구
* 풀이 -

은 12.75이다.

sort(df$region)

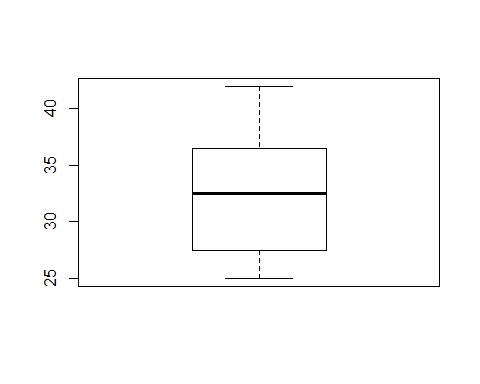
## [1] 25 26 27 28 30 32 33 35 36 37 38 42

# IQR  
iqr <- IQR(df$region)  
  
1.5\*iqr

## [1] 12.75

이에 대해 boxplot을 그려보면, 아래와 같이 나온다. 극단치는 표시되지 않으나 아슬아슬하게 극단치의 경계에 있어 outlier로 고려될 수 있는 것은 **‘용산구’, ‘서대문구’** 등이 있다.

boxplot(df$region)$stats



## [,1]  
## [1,] 25.0  
## [2,] 27.5  
## [3,] 32.5  
## [4,] 36.5  
## [5,] 42.0

**4번**

outlier에 가장 영향을 많이 받는 것은 ’3번 mean(평균)’이다.

## 5번 ~ 8번

**5번**

* 답 : 0.02275013

표준 정규분포표에 따라 이므로 는 이다. 아래처럼 pnorm으로 구하면 더 정확한 값을 구할 수 있다.

pnorm은 X가 q이하일 확률을 계산한다. 4400시간 이상일 확률이므로 1에서 pnorm의 값을 뺀다.

x <- 1 - pnorm(q=4400, mean=4000,sd=200)

**6번**

* 답 : 34.02 ~ 35.98

이다. 표준오차 se는 이므로 0.5이다. 상한선에 대해 qnorm(alpha/2, lower.tail = F)는 표준정규분포에서 상위 α/2 = 0.025에 해당하는 값은 1.959964이다. 모표준편차를 5로 가정하였으므로 오차 한계는 이다.표본평균을 기준으로 error 값을 더하고, 빼서 신뢰구간을 구한다. 이에 따라 구한 모평균에 대한 95% 신뢰구간은 (34.02, 35.98)이다.

sigma <- 5  
n <- 100  
alpha <- 0.05 # 신뢰수준   
  
se <- sigma/sqrt(n) # 표준오차   
  
error <- qnorm(alpha/2, lower.tail = F)\*se  
error

## [1] 0.979982

# 신뢰 하한선  
lower <- round(35-error, digits=3)  
lower

## [1] 34.02

# 신뢰 상한선   
upper <- round(35+error, digits=3)  
upper

## [1] 35.98

#신뢰 구간   
c(lower, upper)

## [1] 34.02 35.98

**7번**

이 미지수이므로 t-test를 사용한다. 이다.

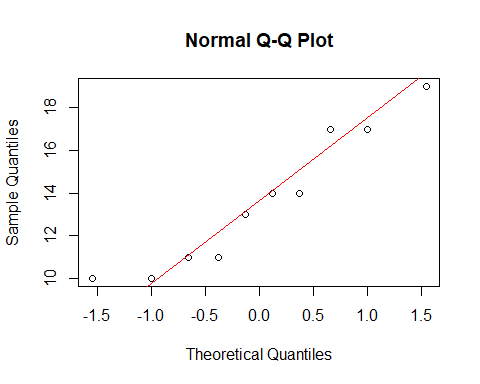
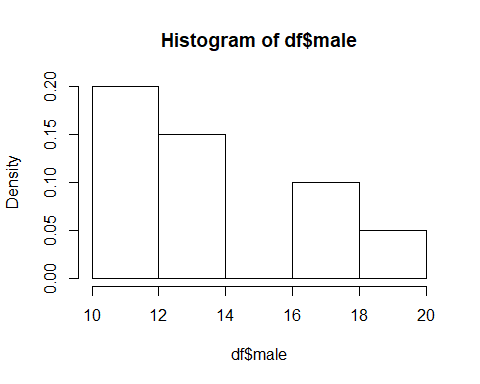
* 귀무가설은 으로 설정하고, 이에 대한 대립가설은 으로 설정한다. 이때 α에 대한 기각역은 이다.
* 주어진 자료로부터 이므로 t 관측값은
* 이다.
* 은 t분포표에 따라 2.131이다. t 관측값은 이보다 더 작으므로 신뢰수준 95%에서 귀무가설을 기각할 수 없다. 즉, 모평균은 35가 아니라고 할 수 없다.

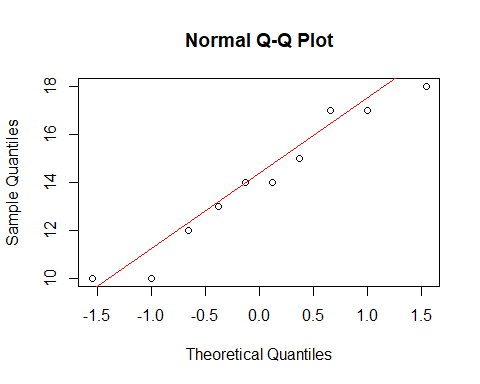
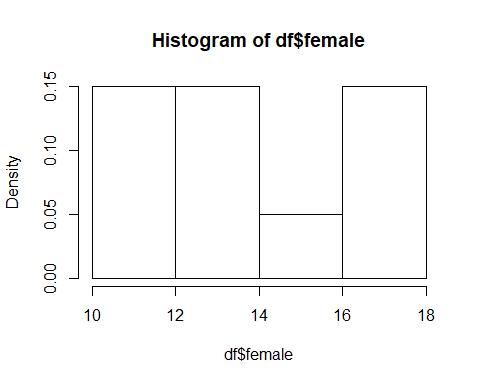
## 8번

먼저 남성과 여성의 히스토그램을 보면 남성은 편향, 여성은 일부 균등해보인다. 하지만 표본의 크기가 적어 판단하기 어렵다.

male <- c(11, 17, 14, 10, 13, 11, 17, 10, 19, 14)  
female <- c(10, 13, 12, 15, 18, 14, 14, 17, 17, 10)  
  
df <- data.frame(male, female)  
df

## male female  
## 1 11 10  
## 2 17 13  
## 3 14 12  
## 4 10 15  
## 5 13 18  
## 6 11 14  
## 7 17 14  
## 8 10 17  
## 9 19 17  
## 10 14 10

**male** 

**female** 

two-sample t-test를 하기 위해서는 3가지 가정이 필요하다. 독립성, 정규성, 동질성(등분산성)이다. 이 문제의 경우, 성별은 서로 독립적이므로 독립성을 충족하고, 동질성은 문제에서 가정하고 있으므로, 정규성 가정을 판단한다. 표본의 크기는 각각 10명 씩인 소표본에 해당하므로 샤피로 테스트를 사용한다.

아래 두 p-value 모두 유의수준 0.05보다 크므로 정규성을 만족한다고 가정한다.

# 정규성 검증  
shapiro.test(df$male)

##   
## Shapiro-Wilk normality test  
##   
## data: df$male  
## W = 0.90822, p-value = 0.269

shapiro.test(df$female)

##   
## Shapiro-Wilk normality test  
##   
## data: df$female  
## W = 0.93613, p-value = 0.5108

이제 t.test()를 실행한다. p-value는 0.7706으로 신뢰수준 95%에서 귀무가설을 기각한다. 즉, 남녀 성별에 따른 승진 소요 기간에는 차이가 없다고 할 수 없다.

t.test(df$male, df$female, var.equal = T)

##   
## Two Sample t-test  
##   
## data: df$male and df$female  
## t = -0.29596, df = 18, p-value = 0.7706  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -3.239485 2.439485  
## sample estimates:  
## mean of x mean of y   
## 13.6 14.0

effect size도 보자. cohen’s d는 0.1323565이므로 p-value는 높고, d 값은 훨씬 작아 유의미하다.

cohens\_d <- function(x, y) {  
 lx <- length(x)- 1  
 ly <- length(y)- 1  
 md <- abs(mean(x) - mean(y)) ## mean difference (numerator)  
 csd <- lx \* var(x) + ly \* var(y)  
 csd <- csd/(lx + ly)  
 csd <- sqrt(csd) ## common sd computation  
   
 cd <- md/csd ## cohen's d  
 return(cd)  
 }  
  
result <- cohens\_d(df$male, df$female)  
result

## [1] 0.1323565

## 9번

3가지 가정 검정

**정규성**

각각 소표본이므로 샤피로 테스트를 사용한다.

control <- c(0,1,0,3,1,2,4,2)  
day\_1 <- c(1,3,2,2,4,6,3,4)  
day\_2 <- c(5,4,7,8,6,3,2,5)  
day\_3 <- c(7,1,6,9,10,12,8,7)  
  
df <- data.frame(control,day\_1, day\_2, day\_3)  
df

## control day\_1 day\_2 day\_3  
## 1 0 1 5 7  
## 2 1 3 4 1  
## 3 0 2 7 6  
## 4 3 2 8 9  
## 5 1 4 6 10  
## 6 2 6 3 12  
## 7 4 3 2 8  
## 8 2 4 5 7

shapiro.test(df$control)

##   
## Shapiro-Wilk normality test  
##   
## data: df$control  
## W = 0.93444, p-value = 0.5573

shapiro.test(df$day\_1)

##   
## Shapiro-Wilk normality test  
##   
## data: df$day\_1  
## W = 0.95201, p-value = 0.7314

shapiro.test(df$day\_2)

##   
## Shapiro-Wilk normality test  
##   
## data: df$day\_2  
## W = 0.9828, p-value = 0.9754

shapiro.test(df$day\_3)

##   
## Shapiro-Wilk normality test  
##   
## data: df$day\_3  
## W = 0.93714, p-value = 0.5832

모두 유의수준 0.05보다 높으므로 정규성을 만족한다.

**동질성**

동질성 검증을 실행한다. Pr이 0.3531로 유의수준 0.05보다 크므로 F-value가 높게 책정되었다. 따라서 세 집단 간 분산에는 큰 차이가 없으므로 등분산을 가정한다.

library(car)

## Loading required package: carData

y <- c(df$day\_1, df$day\_2, df$day\_3)   
group <- as.factor(c(rep(1, length(df$day\_1)),   
 rep(2, length(df$day\_2)),  
 rep(3, length(df$day\_3))))   
  
leveneTest(y, group)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 2 1.0943 0.3531  
## 21

## 10번