Result 6.2. (Independent t-tests)

(1) 표본평균 $\bar{X}_1 \sim N_p(\mu_1, \frac{\Sigma}{n_1})$, $\bar{X}_2 \sim N_p(\mu_2, \frac{\Sigma}{n_2})$이다.

$X_{1i} \perp X_{2j}$이므로 $\bar{X}_1 \perp \bar{X}_2$, $\bar{X}_1 - \bar{X}_2 \sim N_p(\mu_1 - \mu_2, \Sigma(\frac{1}{n_1} + \frac{1}{n_2}))$.

한편, $(n_1-1)S_1 \sim Wishart_p(n_1-1, \Sigma)$, $(n_2-1)S_2 \sim Wishart_p(n_2-1, \Sigma)$.

(2) $(n_1-1)S_1 + (n_2-1)S_2 \sim Wishart_p(n_1+n_2-2, \Sigma)$이다.

$\bar{X}_1 - \bar{X}_2 \sim N_p(\mu_1 - \mu_2, \Sigma(\frac{1}{n_1} + \frac{1}{n_2}))$이고, $(\bar{X}_1 - \bar{X}_2) \perp \{(n_1-1)S_1 + (n_2-1)S_2\}$.

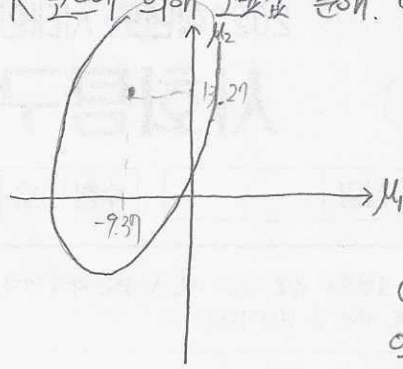
따라서 이를 통해 계산한 Hotelling's T^2는

$$T^2 = \{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)\}^T \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} S_p^{-1} \{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)\},$$

$S_p = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2}$에서 $T^2 \sim \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p, n_1+n_2-p-1}$이다.

Result 6.3.

6.1 R 코드에 의해 고유값 분해. 타원은 다음과 같다.



데이터 타원이 (0,0)을 포함하지 않으므로 귀무가설을 기각하며, 이는 e.g. 6.1의 결과와 일관된다.

6.2. R 코드 참고

6.3. R 코드 참고

6.4. R 코드 참고

6.5. R 코드 참고

6.6. $X_2 := \begin{pmatrix} 3 & 3 \\ 1 & 6 \\ 2 & 3 \end{pmatrix}$ $\tilde{X}_2 := \begin{pmatrix} 1 & -1 \\ -1 & 2 \\ 0 & -1 \end{pmatrix}$ $\bar{X}_2 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$

(a) $X_3 := \begin{pmatrix} 2 & 3 \\ 5 & 1 \\ 3 & 1 \\ 2 & 3 \end{pmatrix}$ $\tilde{X}_3 := \begin{pmatrix} -1 & 1 \\ 2 & -1 \\ 0 & -1 \\ -1 & 1 \end{pmatrix}$ $\bar{X}_3 = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$

$S_2 = \frac{1}{2} \tilde{X}_2^T \tilde{X}_2 = \begin{pmatrix} 1 & -3/2 \\ -3/2 & 3 \end{pmatrix}$ $S_3 = \begin{pmatrix} 2 & -4/3 \\ -4/3 & 4/3 \end{pmatrix}$

$S_{pooled} = \frac{1}{5}(2S_2 + 3S_3) = \begin{pmatrix} 8/5 & -7/5 \\ -7/5 & 2 \end{pmatrix}$

(b) $T^2 = (2-3 \quad 4-2)\left(\frac{7}{12} S_{pooled}\right)^{-1} \begin{pmatrix} 2-3 \\ 4-2 \end{pmatrix} = \frac{120}{31} < 45.$

R 코드에서의 45보다 작으므로 H_0 기각.

(c) $\mu_{21} - \mu_{31} \in (2-3) \pm 6.480741 \approx (-7.48, 5.48)$

$\mu_{22} - \mu_{32} \in (4-2) \pm 7.245688 \approx (-5.76, 9.24)$

6.7. $H_0: \mu_1 = \mu_2$의 가설검정 시행시 검정통계량 $16 > 6 =$ 임계값이므로 귀무가설 기각.

선형결합: $S_p^{-1}(\bar{X}_1 - \bar{X}_2) = \begin{pmatrix} 0.0017 \\ 0.0026 \end{pmatrix}$

6.13.

(a) (1st variable)

(b)
$$\begin{pmatrix} 6 & 4 & 8 & 2 \\ 3 & -3 & 4 & -4 \\ -3 & -4 & 3 & -4 \end{pmatrix} = \text{mean} + \text{F1 eff} + \text{F2 eff} + \text{err}$$

$SS_{tot} =$

$$= \underset{\text{mean}}{\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}} + \underset{\text{F1}}{\begin{pmatrix} 4 & 4 & 4 & 4 \\ -1 & -1 & -1 & -1 \\ -3 & -3 & -3 & -3 \end{pmatrix}} + \underset{\text{F2}}{\begin{pmatrix} 1 & -2 & 4 & -3 \\ 1 & -2 & 4 & -3 \\ 1 & -2 & 4 & -3 \end{pmatrix}}$$

$$+ \begin{pmatrix} 0 & 1 & -1 & 0 \\ 2 & -1 & 0 & -1 \\ -2 & 0 & 1 & 1 \end{pmatrix}$$

$SS_{tot} = 220$ $\quad SS_{F2} = 90$

$SS_{mean} = 12$ $\quad SS_{Res} = 14$

$SS_{F1} = 104$

(2nd variable)

$$\begin{pmatrix} 8 & 6 & 12 & 6 \\ 8 & 2 & 3 & 3 \\ 2 & -5 & -3 & -6 \end{pmatrix} = \underset{}{\overset{\text{mean}}{\begin{pmatrix} 3 & 3 & 3 & 3 \\ 3 & 3 & 3 & 3 \\ 3 & 3 & 3 & 3 \end{pmatrix}}} + \overset{\text{F1}}{\begin{pmatrix} 5 & 5 & 5 & 5 \\ 1 & 1 & 1 & 1 \\ -6 & -6 & -6 & -6 \end{pmatrix}}$$

$$+ \underset{\text{F2}}{\begin{pmatrix} 3 & -2 & 1 & -2 \\ 3 & -2 & 1 & -2 \\ 3 & -2 & 1 & -2 \end{pmatrix}} + \underset{\text{err}}{\begin{pmatrix} -3 & 0 & 3 & 0 \\ 1 & 0 & -2 & 1 \\ 2 & 0 & -1 & -1 \end{pmatrix}}$$

$SS_{tot} = 440$ $\quad SS_{F2} = 54$

$SS_{mean} = 108$ $\quad SS_{Res} = 30$

$SS_{F1} = 248$

6.14.

(a) (1st variable)

$$\begin{pmatrix} 14 & 6 & 8 & 16 \\ 1 & 5 & & \\ & & & \end{pmatrix}$$

multivariate_hw7

Na SeungChan

2024-12-10

Q 6.1

```
x1_c <- c(6, 6, 18, 8, 11, 34, 28, 71, 43, 33, 20)
x1_s <- c(25, 28, 36, 35, 15, 44, 42, 54, 34, 29, 39)
x2_c <- c(27, 23, 64, 44, 30, 75, 26, 124, 54, 30, 14)
x2_s <- c(15, 13, 22, 29, 31, 64, 30, 64, 56, 20, 21)

df <- tibble(x1_c, x1_s, x2_c, x2_s)
```

```
diff_df <- df %>%
  transmute(diff_x1 = x1_c - x1_s, diff_x2 = x2_c - x2_s)
```

```
mean_df <- colMeans(diff_df)
mean_df
```

```
##   diff_x1   diff_x2
## -9.363636 13.272727
```

```
var_df <- cov(diff_df)
var_df
```

```
##           diff_x1   diff_x2
## diff_x1 199.25455  88.30909
## diff_x2  88.30909 418.61818
```

교재의 S_d와 데이터 매트릭스를 직접 입력해 계산한 S_d의 값이 약간 다르다. 문제풀이에서는 데이터 매트릭스를 직접 입력해 계산한 값을 활용해 계산한다.

```
sqrt(eigen(var_df)$values)
```

```
## [1] 21.20732 12.96620
```

```r
eigen(var_df)$vectors
```

```
##           [,1]       [,2]
## [1,] 0.3324812 -0.9431099
## [2,] 0.9431099  0.3324812
```

Q 6.2

```r
crit_v1 <- sqrt(var_df[1]/11)*qt(1/80, df = 10, lower.tail = FALSE)
BCI_V1 <- c(mean_df[1] - crit_v1, mean_df[1] + crit_v1)
crit_v2 <- sqrt(var_df[4]/11)*qt(1/80, df = 10, lower.tail = FALSE)
BCI_V2 <- c(mean_df[2] - crit_v2, mean_df[2] + crit_v2)
```

```r
BCI_V1
```

```
##    diff_x1    diff_x1
## -20.573107   1.845835
```

```r
BCI_V2
```

```
##   diff_x2   diff_x2
## -2.974903 29.520358
```

example 6.1의 동시 신뢰구간은 diff_x1의 경우 (-22.46, 3.74), diff_x2의 경우 (-5.71, 32.25)이다. 여기서 구한 본페르니 신뢰구간이 동시 신뢰구간에 비해 V1과 V2 모두 upper bound가 작고 lower bound가 커서 신뢰구간이 더 좁은 것을 볼 수 있다.

Q 6.3

```r
diff_q3 <- diff_df %>% slice(-8)
xbar <- colMeans(diff_q3)
Smtx <- cov(diff_q3)
```

위와 같이 이상치를 제거한 데이터프레임을 구하였고, 평균벡터와 분산행렬을 구하였다.

동시 신뢰구간

```r
crit <- sqrt(qf(0.05, 2, 8, lower.tail = FALSE) * (2*9/8))
SCI_V1 <- c(xbar[1] - crit*sqrt(Smtx[1]/10), xbar[1] + crit*sqrt(Smtx[1]/10))
SCI_V2 <- c(xbar[2] - crit*sqrt(Smtx[4]/10), xbar[2] + crit*sqrt(Smtx[4]/10))
```

```r
eigen(Smtx)
```

```
## eigen() decomposition
## $values
## [1] 228.2318 106.4793
##
## $vectors
##            [,1]       [,2]
## [1,] -0.4960999 -0.8682654
## [2,]  0.8682654 -0.4960999
```

본페르니 신뢰구간

```r
crit_v1 <- sqrt(Smtx[1]/10)*qt(1/80, df = 9, lower.tail = FALSE)
BCI_V1 <- c(xbar[1] - crit_v1, xbar[1] + crit_v1)
crit_v2 <- sqrt(Smtx[4]/10)*qt(1/80, df = 9, lower.tail = FALSE)
BCI_V2 <- c(xbar[2] - crit_v2, xbar[2] + crit_v2)
```

```r
SCI_V1
```

```
##     diff_x1     diff_x1
## -23.7000163  -0.2999837
```

```r
BCI_V1
```

```
##    diff_x1    diff_x1
## -21.917997  -2.082003
```

```r
SCI_V2
```

```
##   diff_x2   diff_x2
## -5.503711 22.703711
```

```r
BCI_V2
```

```
##   diff_x2   diff_x2
## -3.355587 20.555587
```

동시 신뢰구간 SCI_V1, SCI_V2가 (0,0) point를 포함하지 않으므로 이상치를 기각한 경우에도 example 6.1과 같이 귀무가설을 기각하게 된다. 이상치 유무와 관계없이 귀무가설을 기각하지만, 이 경우에는 동시 신뢰구간이 0을 포함하는 상황 자체가 존재하지 않아 example 6.1에 비해 유의수준이 높을 것이다.

Q 6.4

```r
tdf <- log(df) %>% transmute(dx1 = x1_c - x1_s, dx2 = x2_c - x2_s)
tdf
```

```
## # A tibble: 11 x 2
##       dx1     dx2
##     <dbl>   <dbl>
##  1 -1.43   0.588
##  2 -1.54   0.571
##  3 -0.693  1.07
##  4 -1.48   0.417
##  5 -0.310 -0.0328
##  6 -0.258  0.159
##  7 -0.405 -0.143
##  8  0.274  0.661
##  9  0.235 -0.0364
## 10  0.129  0.405
## 11 -0.668 -0.405
```

로그 변환된 데이터프레임을 계산하였다.

(a)

```r
xbar <- colMeans(tdf)
Smtx <- cov(tdf)
```

Hotelling's T^2 test

```r
test_statistics <- 11 * t(xbar) %*% solve(Smtx) %*% (xbar)
critical_value <- qf(0.05, 2, 9, lower.tail = FALSE) * (2*10/9)
test_statistics
```

```
##          [,1]
## [1,] 10.21541
```

```r
critical_value
```

```
## [1] 9.458877
```

```r
test_statistics > critical_value
```

```
##      [,1]
## [1,] TRUE
```

검정 결과 차이가 있다는 귀무가설을 기각하게 된다.

(b)

```
crit_v1 <- sqrt(Smtx[1]/11)*qt(1/80, df = 10, lower.tail = FALSE)
BCI_V1 <- c(xbar[1] - crit_v1, xbar[1] + crit_v1)
crit_v2 <- sqrt(Smtx[4]/11)*qt(1/80, df = 10, lower.tail = FALSE)
BCI_V2 <- c(xbar[2] - crit_v2, xbar[2] + crit_v2)
```

```
BCI_V1
```

```
##        dx1        dx1
## -1.09448795 -0.02190232
```

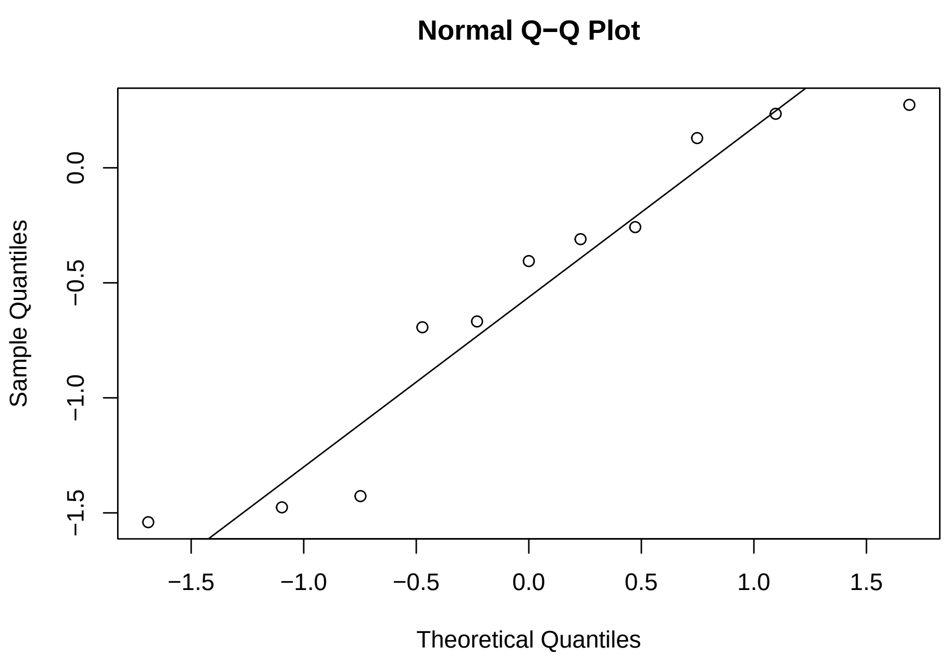
```
BCI_V2
```

```
##         dx2         dx2
## -0.04498455  0.63604112
```

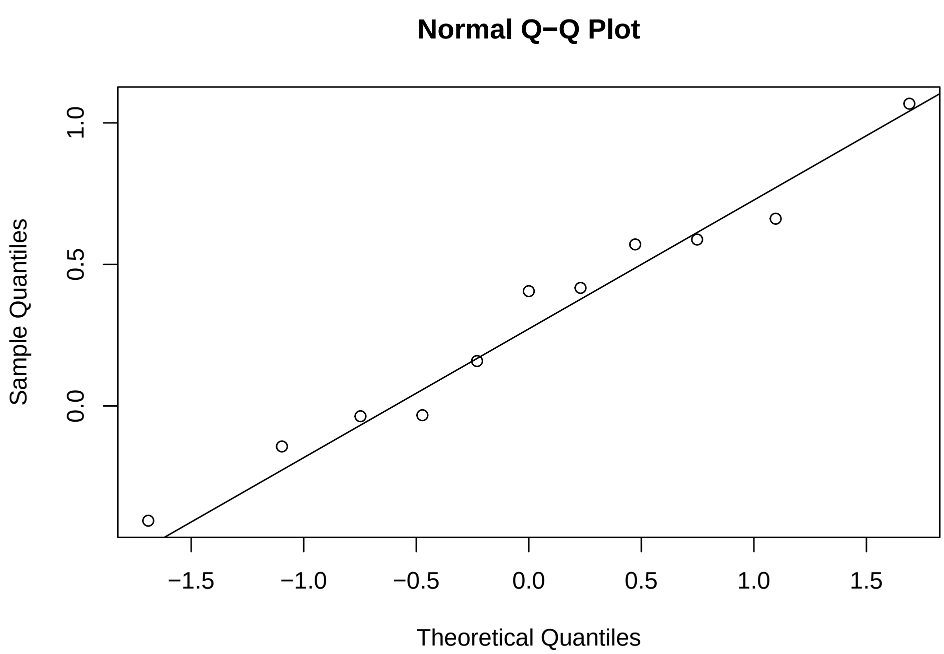
본페르니 신뢰구간은 위와 같다.

(c)

```
qqnorm(tdf$dx1); qqline(tdf$dx1)
```

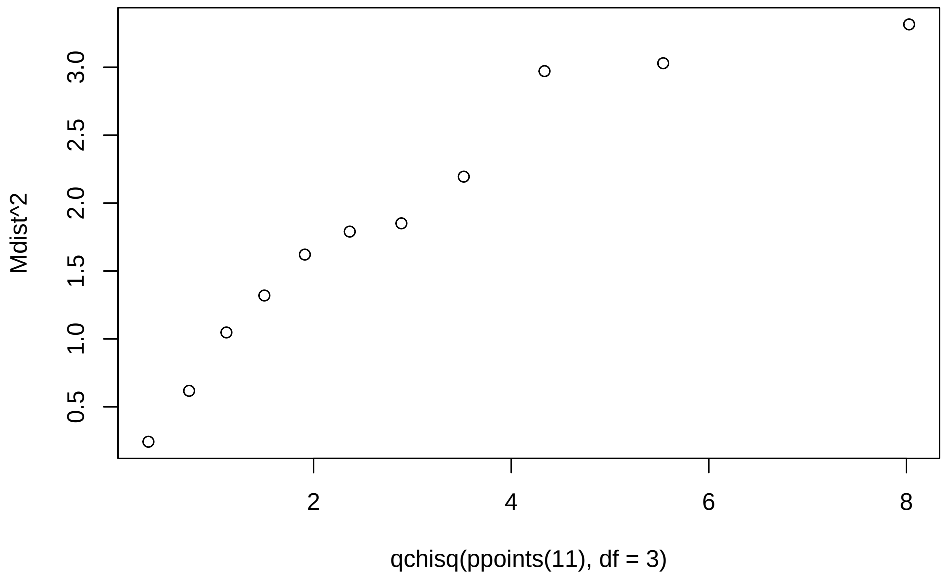


```r
qqnorm(tdf$dx2); qqline(tdf$dx2)
```



marginally normal이라고 볼 수도 있을 것 같다.

```r
Xc <- t(t(tdf) - xbar)
Mdist <- sqrt( diag( Xc %*% solve(Smtx) %*% t(Xc) ) )
qqplot( qchisq(ppoints(11), df = 3), Mdist^2)
```



n = 11은… 정규성 여부를 판단하기 어렵지만, 직선을 따른다고 말하기 어려워 보인다. 즉 joint normal이라고 말하기 어렵다.

Q 6.5

```
xbar <- c(46.1, 57.3, 50.4)
Smtx <- matrix(c(101.3, 63.0, 71.0, 63.0, 80.2, 55.6, 71.0, 55.6, 97.4), nrow=3)
n <- 40
Cmtx <- matrix(c(1, 0, -1, 1, 0, -1), nrow = 2)
```

(a)

```
test_statistics <- n * t(Cmtx %*% xbar) %*% solve(Cmtx %*% Smtx %*% t(Cmtx)) %*% (Cmtx %*% xbar)
critical_value <- qf(0.05, 2, 38, lower.tail = FALSE) * (2*39/38)
test_statistics > critical_value
```

```
##      [,1]
## [1,] TRUE
```

가설 검정 결과 귀무가설을 기각한다. 즉 유의수준 0.05에서 평균이 다르다고 볼 만한 충분한 증거가 있다.

(b)

```
#(mu1 vs. mu3)
xbar[1] - xbar[2]
```

```
## [1] -11.2
```

```
sqrt(critical_value * Smtx[1] / 40)
```

```
## [1] 4.107007
```

Q 6.6

```
qf(0.01, 2, 4, lower.tail = FALSE)*(2*5/4)
```

```
## [1] 45
```

```
sqrt(45*7/12)*sqrt(8/5)
```

```
## [1] 6.480741
```

```
sqrt(45*7/12)*sqrt(2)
```

```
## [1] 7.245688
```

Q 6.7

```
n1 <- 45
n2 <- 55
xb1 <- c(204.4, 556.6)
xb2 <- c(130.0, 355.0)
Sm1 <- matrix(c(13825.3, 23823.4, 23823.4, 73107.4), nrow = 2)
Sm2 <- matrix(c(8632.0, 19616.7, 19616.7, 55964.5), nrow = 2)
Sp <- ((n1-1)*Sm1 + (n2-1)*Sm2)/(n1 + n2 - 2)
```

```
critical_value <- qf(0.05, 2, 97, lower.tail = FALSE) * (98*2/97)
test_statistics <- t(xb1 - xb2) %*% solve((1/n1 + 1/n2)*Sp) %*% (xb1 - xb2)
```

```
critical_value
```

```
## [1] 6.244089
```

```r
test_statistics
```

```
##          [,1]
## [1,] 16.06622
```

```r
solve(Sp) %*% (xb1 - xb2)
```

```
##            [,1]
## [1,] 0.00170252
## [2,] 0.00259163
```

Q 6.13