Project 2

Code ▾

Jonathan Feng

PART 1

Use Hotelling's T^2 test to test for a difference in the mean score vector of the boys and the mean vector of the girls.

Hide

```
testdata <- read.table("C:\\Users\\Taterthot\\Desktop\\da 410\\project2\\testscoredata.txt", header = TRUE)[,-1]
testdata
```

math <dbl>	reading <dbl>	sex <chr>
83.16	79.67	boy
102.51	101.13	boy
81.63	80.53	boy
88.25	84.58	boy
81.47	76.52	boy
87.19	84.70	boy
88.66	85.86	boy
79.35	81.03	boy
83.35	80.44	boy
86.58	84.67	boy

1-10 of 62 rows Previous **1** 2 3 4 5 6 7 Next

Hide

```
testmanova <- manova(cbind(testdata$math,testdata$reading)~testdata$sex)
summary(testmanova, test= 'Hotelling-Lawley')
```

```
             Df Hotelling-Lawley approx F num Df den Df    Pr(>F)    
testdata$sex  1          0.30593   9.0249      2     59 0.0003805 ***
Residuals    60                                                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

h0: There is no difference between the mean vectors for math and reading across both genders.

ha: There is a difference between the mean vectors for math and reading across both genders.

Due to the p value of our hotelling lawley manova having a p-value less than 0.05 level of significance, we can conclude that there is a significant difference between the mean vector of the boys and girls scores for the tests they took.

PART 2

Suppose we have gathered the following data on female athletes in three sports. The measurements we have made are the athletes' heights and vertical jumps, both in inches. The data are listed as (height, jump) as follows.

Hide

```
sp<-data.frame(
  sport=c('B','B','B','B','B','T','T','T','T','S','S','S','S','S','S'),
  height=c(66,65,68,64,67,63,61,62,60,62,65,63,62,63.5,66),
  jump=c(27,29,26,29,29,23,26,23,26,23,21,21,23,22,21.5)
)
sp
```

sport <chr>	height <dbl>	jump <dbl>
B	66.0	27.0
B	65.0	29.0
B	68.0	26.0
B	64.0	29.0
B	67.0	29.0
T	63.0	23.0
T	61.0	26.0
T	62.0	23.0
T	60.0	26.0
S	62.0	23.0

1-10 of 15 rows Previous **1** 2 Next

Use R to conduct the MANOVA F-test using Wilks' Lambda to test for a difference in (height, jump) mean vectors across the three sports.

Hide

```
summary(manova(cbind(height, jump) ~ sport, data = sp), test = "Wilks")
```

```
          Df    Wilks approx F num Df den Df    Pr(>F)    
sport      2 0.035879   23.536      4     22 1.117e-07 ***
Residuals 12                                              
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

h0: There is no difference between the mean vectors for jump and height across all three sports

ha: There is a difference between the mean vectors for jump and height across all three sports

Due to the p-value being less than 0.05 level of significance we can conclude and reject the null hypothesis and accept that there is a difference in height and jump mean vectors across all three sports for female athletes.

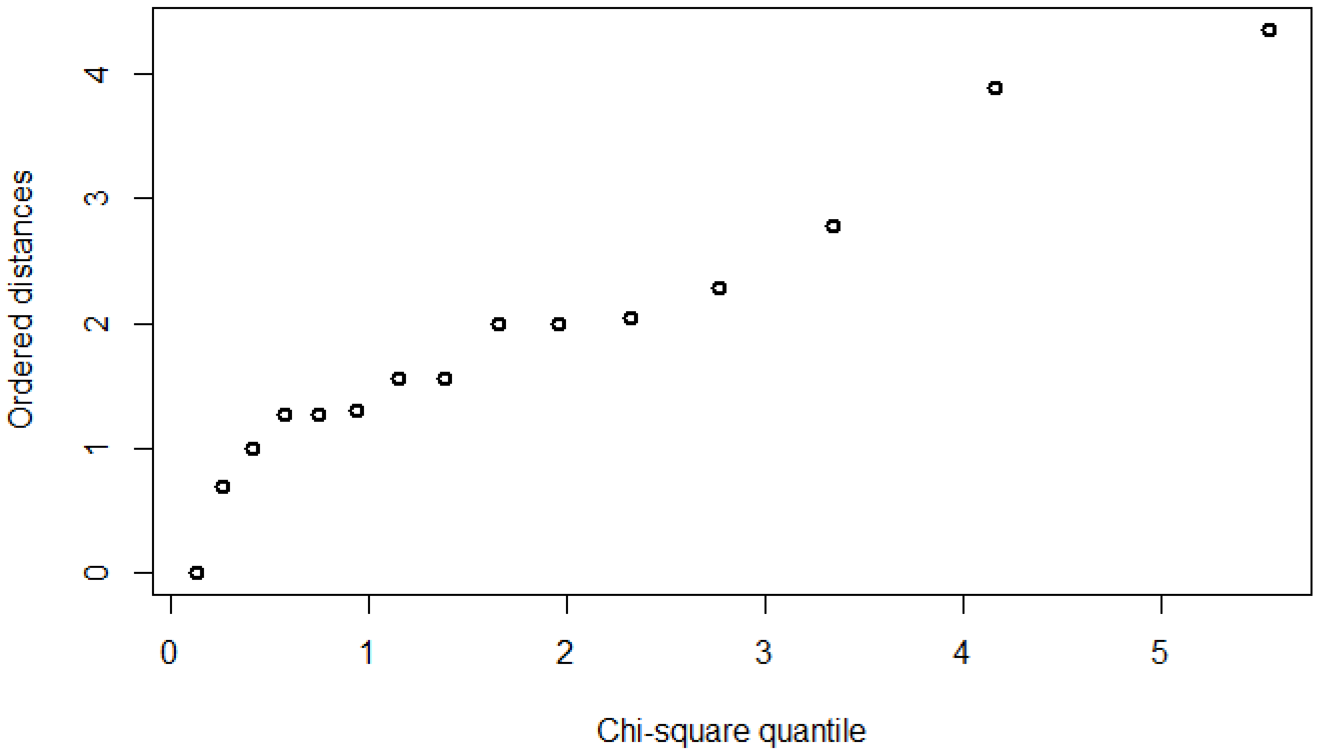
State the assumptions of your test and check to see whether assumptions are met. Do you believe your inference is valid? Why or why not?

Hide

```
chisplot <- function(x) {
  if (!is.matrix(x))
    stop("x is not a matrix")
  ### determine dimensions
  n <- nrow(x)
  p <- ncol(x)
  xbar <- apply(x, 2, mean)
  S <- var(x)
  S <- solve(S)
  index <- (1:n) / (n + 1)
  xcent <- t(t(x) - xbar)
  di <- apply(xcent, 1, function(x, S)
    x %*% S %*% x, S)
  quant <- qchisq(index, p)
  plot(
    quant,
    sort(di),
    ylab = "Ordered distances",
    xlab = "Chi-square quantile",
    lwd = 2,
    pch = 1
  )
}
```

Hide

```
chisplot(residuals(manova(cbind(height, jump) ~ sport, data = sp), test = "Wilks"))
```



Hide

```
by(sp[,-1], sp$sport, var)
```

```
sp$sport: B
       height jump
height    2.5 -1.5
jump     -1.5  2.0
----------------------------------------------------------------
sp$sport: S
          height       jump
height  2.641667 -1.0416667
jump   -1.041667  0.8416667
----------------------------------------------------------------
sp$sport: T
          height jump
height  1.666667   -2
jump   -2.000000    3
```

We use a chi-squared to test for normality and we need to test and it seems fairly non-normal so the assumptions are not met as it is not linear. There are more than 2 independent variables, they are categorical, and are independent observations. There is an adequate sample size and there are no outliers, there should not be any multicollinearity as well. When it comes to the variance when we see the matrices using by() it does not seem like they are similar, so thuse we can assume that it is non-normal, thus assumptions are not met.

Use R to examine the sample mean vectors for each group.

Hide

```
aggregate(x = sp$height, by = list(sp$sport), FUN = mean)
```

```
Group.1                x
<chr>              <dbl>
B               66.00000
S               63.58333
T               61.50000
3 rows
```

Hide

```
aggregate(x = sp$jump, by = list(sp$sport), FUN = mean)
```

```
Group.1                x
<chr>              <dbl>
B               28.00000
S               21.91667
T               24.50000
3 rows
```

It seems like the means for the female basketball player is taller and can jump higher than the other groups, while track atheletes have the lowest average height and a decent jump and softball players having a decent height and the lowest jump.