DA410_Project2_MattGraham

This is our first project, analyzing air pollution, mortality rates, and relevant parameters.

```
library(nnspat)  # used for dist2full()
library("dplyr")  # used to select numeric datatypes
library("ggplot2")
library(reshape)  # used for melting matricies
```

Part 1

Scores by sex

```
testdata <- read.table("C:/mattgraham93.github.io/school/22_3_DA410/data/testscoredata.txt",
header=TRUE)
testdata.noIDs <- testdata[,-1]  #to remove the ID numbers
as.data.frame(testdata.noIDs)
```

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

Hotelling's test

```
summary(manova(cbind(math, reading) ~sex, data=testdata.noIDs), test="Hotelling")
```

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

Hotelling's Analysis

At alpha = 0.05 and p-value < 0.05, we can conclude there is sufficient evidence to state there are differences between mean math and reading scores between the recorded sexes.

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:

```
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)

sports <- as.data.frame(cbind(sport, jump, height))
sports$jump <- as.numeric(sports$jump)
sports$height <- as.numeric(sports$height)

sports
```

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

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

Wilks' Lambda test

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

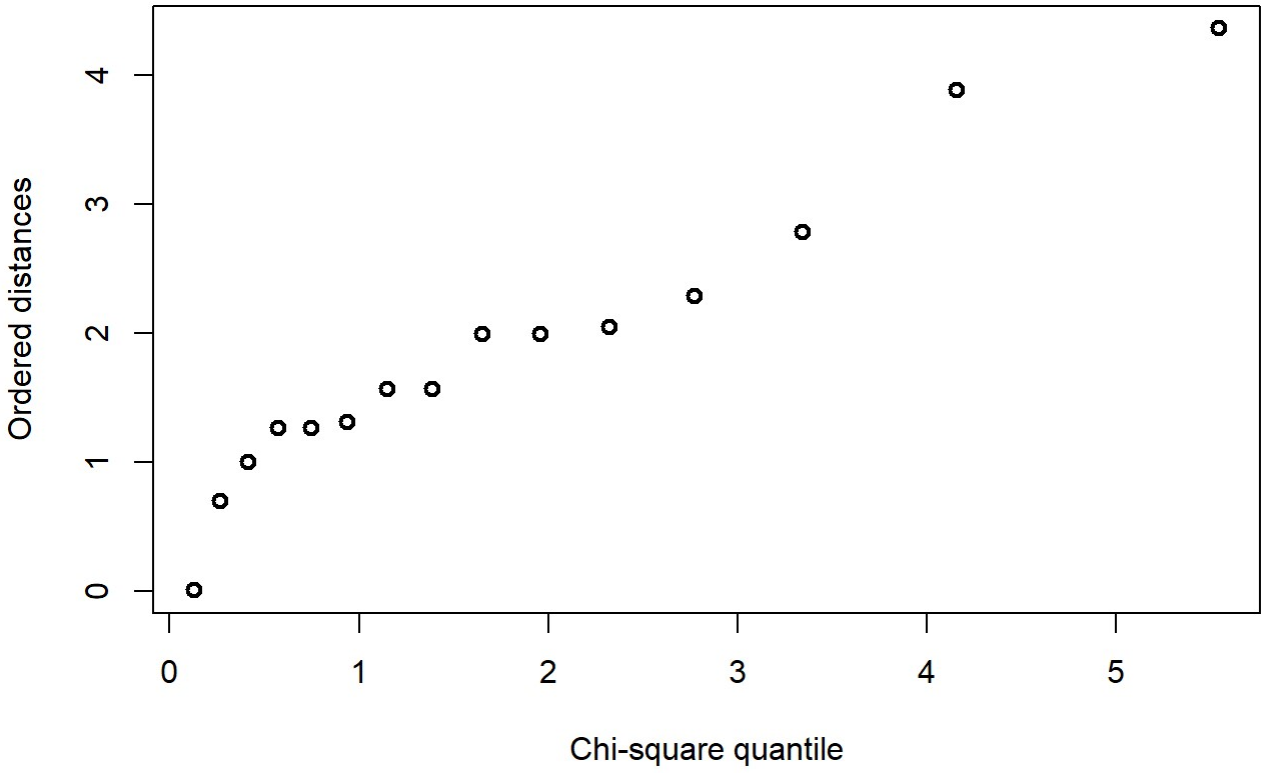
```
##           Df    Wilks approx F num Df den Df    Pr(>F)
## sport      2 0.035879   23.536      4     22 1.117e-07 ***
## Residuals 12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Assumption check

```
sport.manova <- manova(cbind(height, jump) ~ sport)

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)
}

chisplot(residuals(sport.manova))
```



Normality analysis

There is a bit of a sine pattern on our plot. This does make me second-guess the assumption of our tests. However, As this is a mostly-linear line, we can assume we have met all our assumptions and can continue stating there are differences between recorded sexes.

Mean analysis

```
my.n <- nrow(sports)  # number of individuals

# Sample mean vectors for the SAT scores data:
as.data.frame(
  sports %>%
    group_by(sport) %>%
    summarise_at(vars("jump", "height"), mean)
)
```

sport <chr>	jump <dbl>	height <dbl>
B	28.00000	66.00000
S	21.91667	63.58333
T	24.50000	61.50000

3 rows

Above, we can see the differences in our height vs. jump averages. We can note our softball players are between track and basketball players in terms of height. They are, however, lowest when jumping. Meanwhile, the inverse of that statement is true for our shorest track folks. Our basketball players are both the tallest and jump the highest.