da410\_project2

Jason Grahn

1/23/2019

# Part 1:

## The problem statement

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. Make sure you include clear command lines and relevant output/results with hypotheses, test result(s) and conclusion(s)/interpretation(s).

## The data

Download testscoredata.txt and read it in R

scores <-   
 readr::read\_table2(file = here::here("/project2/testscoredata.txt"),  
 col\_names = TRUE) %>%   
 mutate(subject = factor(subject),  
 sex = factor(sex))  
  
head(scores,5)

## # A tibble: 5 x 4  
## subject math reading sex   
## <fct> <dbl> <dbl> <fct>  
## 1 1 83.2 79.7 boy   
## 2 2 103. 101. boy   
## 3 3 81.6 80.5 boy   
## 4 4 88.2 84.6 boy   
## 5 5 81.5 76.5 boy

## The hypothesis

: The mean scores between sexes is equal ( ).

: The mean scores between sexes is not equal ( ).

#sample covariance matrix  
scores.S <- scores %>%   
 select(math, reading) %>%   
 var()  
  
round(scores.S,3)

## math reading  
## math 26.909 25.672  
## reading 25.672 27.335

#build the manova binding  
scores.manova <- stats::manova(cbind(math, reading) ~ sex, scores)  
scores.manova

## Call:  
## stats::manova(cbind(math, reading) ~ sex, scores)  
##   
## Terms:  
## sex Residuals  
## resp 1 19.976 1621.466  
## resp 2 3.7684 1663.6793  
## Deg. of Freedom 1 60  
##   
## Residual standard errors: 5.198502 5.265737  
## Estimated effects may be unbalanced

## The test

#use the HL method to test results.  
tidy.score.hotel <- broom::tidy(scores.manova,   
 test = "Hotelling-Lawley",  
 intercept = FALSE)

## The result:

tidy.score.hotel

## # A tibble: 2 x 7  
## term df hl statistic num.df den.df p.value  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 sex 1 0.306 9.02 2 59 0.000381  
## 2 Residuals 60 NA NA NA NA NA

## The conclusion

With a of 3.805224610^{-4}, we can reject and conclude that the mean vectors between the sexes is not equal.

# Part 2:

## The problem statement

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:

Basketball Players: (66, 27), (65, 29), (68, 26), (64, 29), (67, 29)  
Track Athletes: (63, 23), (61, 26), (62, 23), (60, 26)  
Softball Players: (62, 23), (65, 21), (63, 21), (62, 23), (63.5, 22), (66, 21.5)

## A

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. Make sure you include clear command lines and relevant output/results with hypotheses, test result(s) and conclusion(s)/interpretation(s)

### The data

Download testscoredata.txt and read it in R

#imput data  
sport <- as.factor(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)  
  
#make a table  
sports <- tibble(sport, height, jump) %>%   
 #and give the sports friendly names  
 mutate(sport = case\_when(sport == "B" ~ "Basketball",  
 sport == "T" ~ "Track",  
 sport == "S" ~ "Softball"))  
head(sports, 5)

## # A tibble: 5 x 3  
## sport height jump  
## <chr> <dbl> <dbl>  
## 1 Basketball 66 27  
## 2 Basketball 65 29  
## 3 Basketball 68 26  
## 4 Basketball 64 29  
## 5 Basketball 67 29

#making a tidy table for later use  
tidy.sports <- sports %>% gather(key = var\_name, value = measurement, -sport)

## The hypothesis

: The mean difference between sports is equal ( ).

: The mean difference between sports is not equal ( ).

sports.manova <- stats::manova(cbind(height, jump) ~ sport, sports)  
sports.manova

## Call:  
## stats::manova(cbind(height, jump) ~ sport, sports)  
##   
## Terms:  
## sport Residuals  
## resp 1 45.62500 28.20833  
## resp 2 101.02500 21.20833  
## Deg. of Freedom 2 12  
##   
## Residual standard errors: 1.533197 1.329421  
## Estimated effects may be unbalanced

### The test

tidy.sports.wilks <- broom::tidy(sports.manova,   
 test = "Wilks",  
 intercept = FALSE)

### The result:

tidy.sports.wilks

## # A tibble: 2 x 7  
## term df wilks statistic num.df den.df p.value  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 sport 2 0.0359 23.5 4 22 0.000000112  
## 2 Residuals 12 NA NA NA NA NA

Wilks’ score 0.036.

### The conclusion

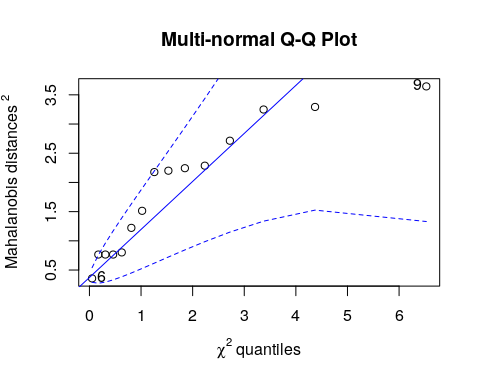
With a of 1.116880510^{-7}, we can safely reject and conclude that the mean vectors aross the sports are not equal.

## B

### The assumptions

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?

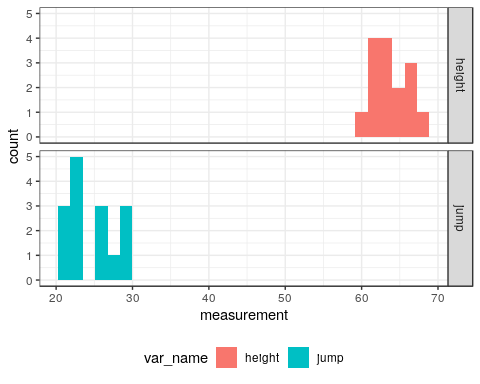
#one way to check for nomalicy is the qqplot. We'll use RVAideMemoire::mqqnorm for this  
RVAideMemoire::mqqnorm(sports %>% select(height, jump))



## [1] 9 6

While the displayed confidence interval is quite spread, we see most values do reside near the line and can conclude normalicy.

tidy.sports %>%   
 ggplot(aes(x = measurement)) +  
 geom\_histogram(aes(fill = var\_name)) +  
 facet\_grid(var\_name ~ .) +  
 theme\_bw() +  
 theme(legend.position="bottom")



Histograms for either variable indicate we may not have normal distributions. Ideally we would have more samples in order to conduct a thorough analysis.

We assume independence as we do not have evidence of the sampling methodology.

#checking for covariance  
sports.S <- sports %>%   
 select(height, jump) %>%   
 var()  
  
#Covariance across all data  
round(sports.S,3)

## height jump  
## height 5.274 1.756  
## jump 1.756 8.731

#covariance within each   
by(sports %>% select(height, jump), sport, var)

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

The data shows little covariance across the entire data as well as within each sport group.

#checking for correlation   
sports.R <- sports %>%   
 select(height, jump) %>%   
 cor()  
  
#correlation across all data  
round(sports.R,3)

## height jump  
## height 1.000 0.259  
## jump 0.259 1.000

by(sports %>% select(height, jump), sport, cor)

## sport: B  
## height jump  
## height 1.0000000 -0.6708204  
## jump -0.6708204 1.0000000  
## --------------------------------------------------------   
## sport: S  
## height jump  
## height 1.0000000 -0.6985857  
## jump -0.6985857 1.0000000  
## --------------------------------------------------------   
## sport: T  
## height jump  
## height 1.0000000 -0.8944272  
## jump -0.8944272 1.0000000

The data shows little correlation across the entire data, though we do see strong relationships within each sport group; specifically within “T” (Track).

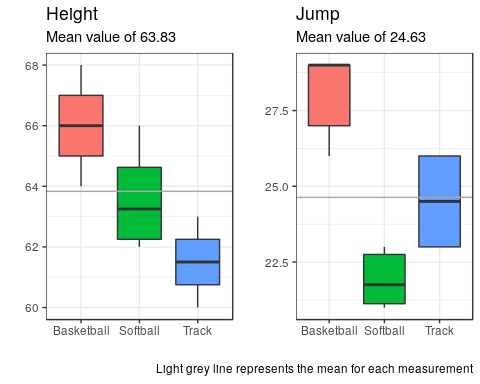
## C

Use R to examine the sample mean vectors for each group. Make sure you include clear command lines and relevant output/results. Also comment on the differences among the groups in terms of the specific variables.

# Sample mean vectors for the sports data  
xbar.sat <- apply(sports %>% select(height, jump), 2, mean)  
xbar.sat

## height jump   
## 63.83333 24.63333

# do some boxplots showing mean distributions  
height\_box <- sports %>%   
 ggplot(aes(x = sport, y = height, fill = sport)) +  
 geom\_boxplot() +  
 geom\_hline(yintercept = xbar.sat[[1]], color = "dark grey") +  
 theme\_bw() +  
 theme(legend.position="none") +  
 labs(title = "Height",  
 subtitle = "Mean value of 63.83",  
 y = "",  
 x = "",  
 caption = " ")  
  
jump\_box <- sports %>%   
 ggplot(aes(x = sport, y = jump, fill = sport)) +  
 geom\_boxplot() +  
 geom\_hline(yintercept = xbar.sat[[2]], color = "dark grey") +  
 theme\_bw() +  
 theme(legend.position="none") +  
 labs(title = "Jump",  
 subtitle = "Mean value of 24.63",  
 y = "",  
 x = "",  
 caption = "Light grey line represents the mean for each measurement")  
  
cowplot::plot\_grid(height\_box,jump\_box)



Plotting the sample mean for each measurement on top of boxplots, we clearly see the distributions for each sport in each variable do not overlap. This reinforces the hypothesis test showing rejection of .

# a quick table of means.  
sports %>%   
 group\_by(sport) %>%   
 summarize(mean\_height = round(mean(height),3),  
 mean\_jump = round(mean(jump),3))

## # A tibble: 3 x 3  
## sport mean\_height mean\_jump  
## <chr> <dbl> <dbl>  
## 1 Basketball 66 28   
## 2 Softball 63.6 21.9  
## 3 Track 61.5 24.5