Project 2 Report

Lohith Kovuri

April 14, 2024

1 Analysis A: Bone Data

The goal of the study is to find if there is a significant difference in the mean mineral content in various bones between the dominant and non-dominant sides. Since the sample size is small (n = 25), we can use Hotelling's two sample T^2 test to complete the study goal. The normality of the data from both the groups is proven by the tables 1 and 2 to do the hypothesis test.

	Test	H	p value	MVN
1	Royston	7.226713	0.06358168	YES
2	Henze-Zirkler	0.6090604	0.4166623	YES

Table 1: DominantBones multivariate Normality test Test H

		p value	MVN	
1	Royston	2.005838	0.5276596	YES
2	Henze-Zirkler	0.72039	0.1787508	YES

Table 2: Non Dominant Bones multivariate Normality test

The result of the Hotelling's two sample T^2 test is given below which concludes that, There is no significant evidence that the bone mineral strength for dominant and non dominant sides are different.

T^2 = 0.29523, df1 = 3, df2 = 46, p-value = 0.8286

From the 95 % confidence intervals below we can state that there is no difference between the mean mineral strengths of the dominant and non dominant bones as all the variables contain 0 in the range.

	Bonferroni Correction		
	Radius	humerus	Ulna
UpperLimit	0.103	0.250	0.084
LowerLimit	-0.052	-0.134	-0.063

Table 3: confidence Interval based on Bonferroni's cor- rection

Both the multivariate T test and the confidence inter- vals agree with one another, as the assumptions made for the test are valid and the data is fairly normal for the test to hold true.

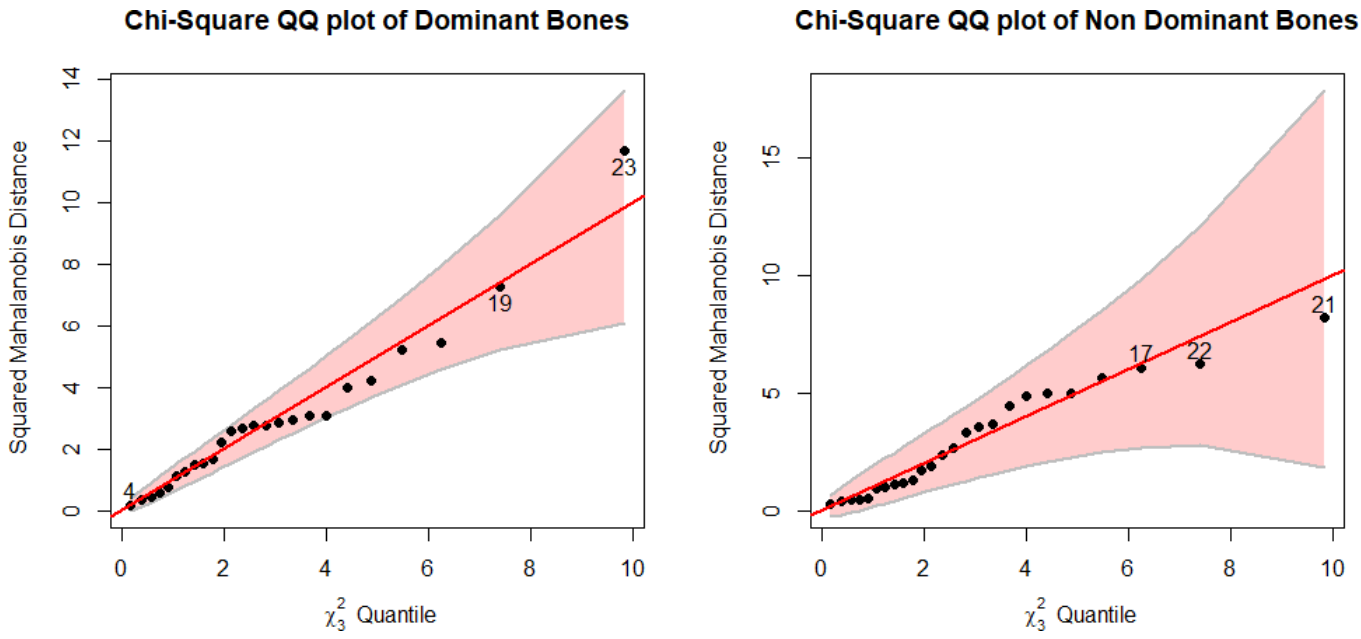

Figure 1: Chisquare qq plots

2 Analysis B1: Triathlon Data

Figure 5 and 6 are the bivariate and multivariate plots for the triathlon data. The goal of the study is to find if there is a significant difference in the average perfor- mance between the two age categories across three differ- ent sports. Since the sample size is less (n = 20), we can use Hotelling's two sample T^2 test to complete the study goal. The normality of the data from both the groups is proven by the tables 4 and 5 to do the hypothesis test. The age category 1 as shown in Fig 2 has an outlier but it is of less significance in this study.

	Test	H	p value	MVN
1	Royston	4.094535	0.2535422	YES
2	Henze-Zirkler	0.5689392	0.4647631	YES

Table 4: Age Category 1 multivariate Normality test The result

of the Hotelling's two sample T^2 test is given below which concludes that, There is significant evidence that the 2 Age categories have different average performance.

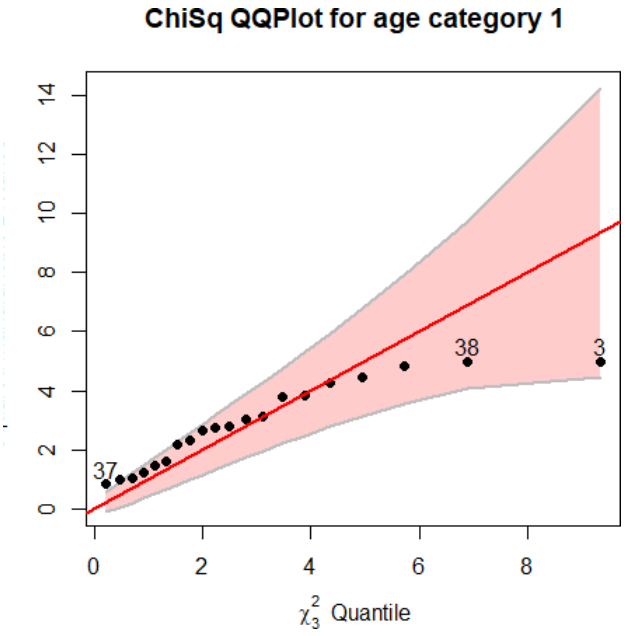

Figure 2: Chisquare qq plot for Age Category 1

T^2 = 17.81, df1 = 3, df2 = 36, p-value = 2.961e-07

From the 95 % confidence intervals in table 6 we can state that there is no difference between the mean min- eral strengths of the dominant and non dominant bones as all the variables contain 0 in their upper and lower limits.

	Test	H	p value	MVN
1	Royston	7.202779	0.06624402	YES
2	Henze-Zirkler	0.8202952	0.05494819	YES

Table 5: Age Category 2 multivariate Normality test

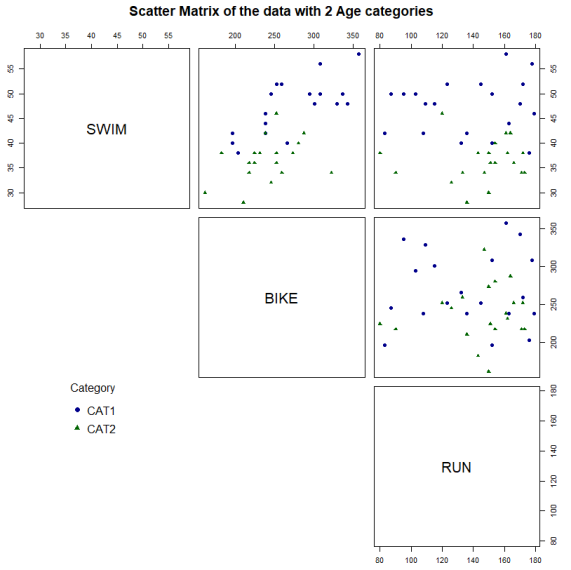

Figure 3: Bivariate Scatter plot

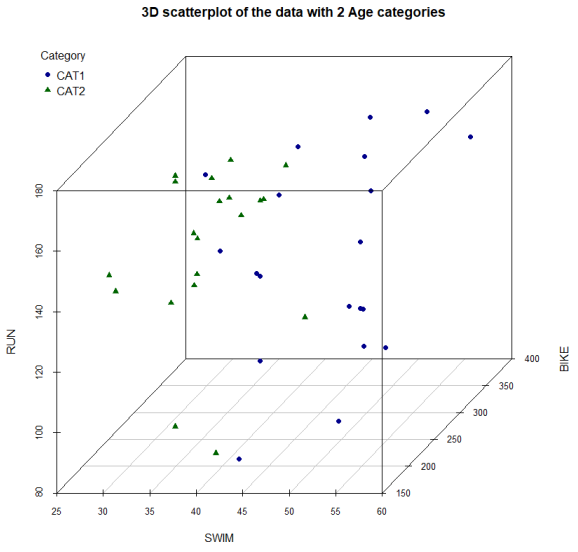

Figure 4: 3D Scatter plot for 2 Age Categories

	Swim	Bike	Run
UpperLimit	14.892	66.25	14.772
LowerLimit	7.108	-2.550	-31.172

Table 6: Bonferroni Correction for Categories 1 and 2

Both the multivariate T test and the confidence inter- vals agree with each other, as the assumptions made for the test are valid and the data is fairly normal for the test to hold true.

3 Analysis B2: Triathlon Data

Figure 5 and 6 are the bivariate and multivariate plots for the triathlon data. The goal of the study is to find

if there is a significant difference in the average performance between the two age categories across 3 different sports. Since the sample size is large (n = 60), we can use Wilk's Lambda test to complete the study goal. The normality of the data from both the groups is proven by the tables 4, 5 and 7 to do the hypothesis test. We assume that the covariances are equal and the groups are independent for the Wilk's Lambda test.

	Test	H	p value	MVN
1	Royston	0.9461821	0.8217762	YES
2	Henze-Zirkler	0.5870892	0.4136919	YES

Table 7: Age Category 3 multivariate Normality test

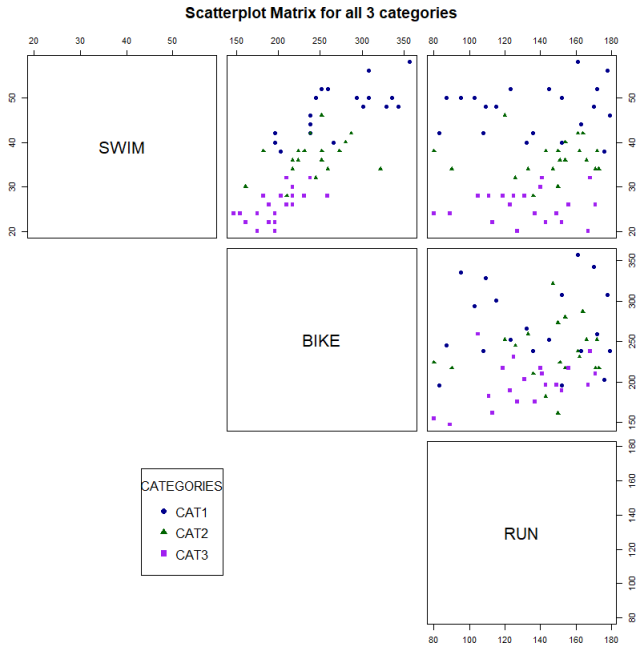

Figure 5: Bivariate Scatter plot

From the wilks lambda approximation in table 8 we can state that there is significant difference in average performance between the different age categories.

	Swim	Bike	Run
UpperLimit	25.81	107.33	29.91
LowerLimit	17.58	36.16	-20.71

Table 9: Bonferroni Correction for Categories 1 and 3

	Swim	Bike	Run
UpperLimit	14.81	75.48	38.11
LowerLimit	6.58	4.31	-12.51

Table 10: Bonferroni Correction for Categories 2 and 3

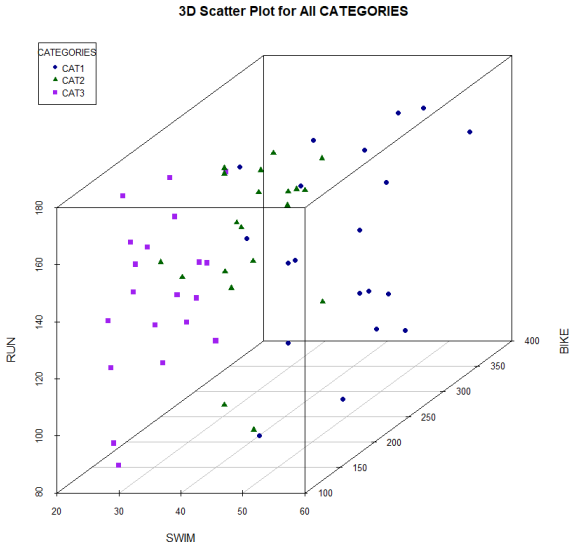

Figure 6: 3D Scatter plot for 3 Age Categories Upon

constructing 95 % Confidence Intervals in the tables 6 , 9 and 10 we can say that there is difference in average performance in both swimming and Biking since there is no 0 in the range of any pairs of groups. However there is no difference in performance for Running accross the 3 pairs of age groups since 0 is in the range of all the three ranges.

The assumptions we made are valid for the data hence the manova test and the confidence intervals both state that atleast one group has significant difference.

Pair	Wilks approx	F	num df	den df	Pr(>F)
cat1-cat2	0.46	21.28	3	55	8.407e-09
cat1-cat3	0.18	78.71	3	55	<2.2e-16
cat2-cat3	0.48	19.08	3	55	3.935e-08

Table 8: Pairwise Wilk's Lambda test

4 R-codes

```
      # Load required libraries
library(readxl)
library(MVN)
library(heplots)
source("Dependencies.R")

 # Read data from Excel files
BonesData = read_excel("./Data/BonesData.xlsx")
TriathlonData = read_excel("./Data/TriathlonData.
    red↪ xlsx")

# Extract 3 variables from Bones Data
x1 = data.frame(BonesData$DomRadius, BonesData$
    red↪ DomHumerus, BonesData$DomUlna)
x2 = data.frame(BonesData$NonDomRadius, BonesData$
    red↪ NonDomHumerus, BonesData$NonDomUlna)

# Load DescTools library
library(DescTools)

# Perform Hotelling's T-squared test and Chi-square
    red↪ test
HotellingsT2Test(x1, x2)
HotellingsT2Test(x1, x2, test = 'chi')

# Set plot layout to 1 row and 2 columns
par(mfrow = c(1, 2))

# Create Chi-square QQ plot for dominant bones and
    red↪ test multivariate normality
cqplot(x1, id.n = 3, main = "Chi-Square␣QQ␣plot␣of␣
    red↪ Dominant␣Bones")
mvn(x1, mvnTest = "royston")
mvn(x1, mvnTest = "hz")

# Create Chi-square QQ plot for non-dominant bones
    red↪ and test multivariate normality
cqplot(x2, id.n = 3, main = "Chi-Square␣QQ␣plot␣of␣
    red↪ Non␣Dominant␣Bones")
mvn(x2, mvnTest = "royston")
mvn(x2, mvnTest = "hz")

# Compute confidence intervals for two-sample
    red↪ Hotelling's T-squared test
MVN2Sample.HT.CIs.f(x1, x2, conf.level = 0.95,
    red↪ alpha = .05, mu0 = rep(0, ncol(x1)),
                ContrastMAT = NULL, SigDig = 3,
                    red↪ var.eq = TRUE)

# Set plot layout to 1 row and 1 column
par(mfrow = c(1, 1))

# Subset TriathlonData excluding CAT3
df = subset(TriathlonData, TriathlonData$CATEGORY !
    red↪ = "CAT3")

# Define colors and shapes for each category
combinations_colors <- c("CAT1" = "darkblue", "CAT2
    red↪ " = "darkgreen")
combinations_shapes <- c("CAT1" = 19, "CAT2" = 17)

# Create scatterplot matrix with colors and shapes
    red↪ for combinations
pairs(~ SWIM + BIKE + RUN, data = df,
      col = combinations_colors[df$CATEGORY],
      pch = combinations_shapes[df$CATEGORY],
      main = "Scatter␣Matrix␣of␣the␣data␣with␣2␣Age␣
          red↪ categories",
      lower.panel = NULL)

# Add legend for combinations
legend("bottomleft", legend = unique(df$CATEGORY),
    red↪ title = "Category",
      bty = "n", col = combinations_colors, pch =
          red↪ combinations_shapes, inset = c
          red↪ (0.12, 0.22), cex = 1)

# Load scatterplot3d library
library(scatterplot3d)

# Create a new data frame with selected variables
new_data = data.frame(
  Combination = df$CATEGORY,
  SWIM = df$SWIM,
  BIKE = df$BIKE,
  RUN = df$RUN
)

# Create a 3D scatterplot
scatterplot3d(new_data$SWIM, new_data$BIKE, new_
    red↪ data$RUN,
              color = combinations_colors[new_data$
                  red↪ Combination],
              pch = combinations_shapes[new_data$
                  red↪ Combination],
              xlab = "SWIM",
              ylab = "BIKE",
              zlab = "RUN",
              main = "3D␣scatterplot␣of␣the␣data␣with
                  red↪ ␣2␣Age␣categories")

# Add legend for combinations
legend("topleft",
       legend = unique(new_data$Combination),
       col = combinations_colors,
       pch = combinations_shapes,
       title = "Category",
       bty = "n",
       inset = c(0.001, 0.01),
       cex = 1)

# Subset data for CAT1, CAT2, and CAT3
x1 = subset(data.frame(TriathlonData$SWIM,
    red↪ TriathlonData$BIKE, TriathlonData$RUN),
    red↪ TriathlonData$CATEGORY == "CAT1")
x2 = subset(data.frame(TriathlonData$SWIM,
    red↪ TriathlonData$BIKE, TriathlonData$RUN),
    red↪ TriathlonData$CATEGORY == "CAT2")
x3 = subset(data.frame(TriathlonData$SWIM,
    red↪ TriathlonData$BIKE, TriathlonData$RUN),
    red↪ TriathlonData$CATEGORY == "CAT3")

# Load DescTools library
```

```
library(DescTools)

# Perform Hotelling's T-squared test and Chi-square test for CAT1 and CAT2
HotellingsT2Test(x1, x2)
HotellingsT2Test(x1, x2, test = 'chi')

# Set plot layout to 1 row and 1 column
par(mfrow = c(1, 1))

# Create Chi-square QQ plot for age category 1 and test multivariate normality
cqplot(x1, id.n = 3, main = "ChiSq QQPlot for age category 1")
mvn(x1, mvnTest = "royston")
mvn(x1, mvnTest = "hz")

# Create Chi-square QQ plot for age category 2 and test multivariate normality
cqplot(x2, id.n = 3, main = "ChiSq QQPlot for age category 2")
mvn(x2, mvnTest = "royston")
mvn(x2, mvnTest = "hz")

# Create Chi-square QQ plot for age category 3 and test multivariate normality
cqplot(x3, id.n = 3, main = "ChiSq QQPlot for age category 3")
mvn(x3, mvnTest = "royston")
mvn(x3, mvnTest = "hz")

# Compute confidence intervals for two-sample Hotelling's T-squared test for CAT1 and CAT2
MVN2Sample.HT.CIs.f(x1, x2, conf.level = 0.95,
    alpha = .05, mu0 = rep(0, ncol(x1)),
                    ContrastMAT = NULL, SigDig = 3,
                        var.eq = TRUE)

# Set plot layout to 1 row and 1 column
par(mfrow = c(1, 1))

# Assign TriathlonData to df
df = TriathlonData

# Define colors and shapes for each category
combinations_colors <- c("CAT1" = "darkblue", "CAT2
    " = "darkgreen", "CAT3" = "purple")
combinations_shapes <- c("CAT1" = 19, "CAT2" = 17,
    "CAT3" = 15)

# Create scatterplot matrix with colors and shapes for combinations
pairs(~ SWIM + BIKE + RUN, data = df,
      col = combinations_colors[df$CATEGORY],
      pch = combinations_shapes[df$CATEGORY],
      main = "Scatterplot Matrix for all 3
          categories",
      lower.panel = NULL)

# Add legend for combinations
legend("bottomleft", legend = unique(df$CATEGORY),
    title = "CATEGORIES",
       col = combinations_colors, pch = combinations
           _shapes, inset = c(0.22, 0.12), cex
           = 1)

# Load scatterplot3d library
library(scatterplot3d)

# Create a new data frame with selected variables
new_data = data.frame(
  Combination = df$CATEGORY,
  SWIM = df$SWIM,
  BIKE = df$BIKE,
  RUN = df$RUN
)

# Create a 3D scatterplot
scatterplot3d(new_data$SWIM, new_data$BIKE, new_
    data$RUN,
              color = combinations_colors[new_data$
                  Combination],
              pch = combinations_shapes[new_data$
                  Combination],
              xlab = "SWIM",
              ylab = "BIKE",
              zlab = "RUN",
              main = "3D Scatter Plot for All
                  CATEGORIES")

# Add legend for combinations
legend("topleft",
       legend = unique(df$CATEGORY),
       col = combinations_colors,
       pch = combinations_shapes,
       title = "CATEGORIES",
       inset = c(0.001, 0.01),
       cex = 0.8)

# Convert first column of new_data to factor
new_data[, 1] = as.factor(new_data[, 1])

# Load car library for MANOVA
library(car)

# Compute means and variance-covariance matrices per group
VMeans = statList(new_data[, -1], new_data[, 1],
    FUN = colMeans)
VMat = statList(new_data[, -1], new_data[, 1], FUN
    = var)

# Sample sizes
Ns = table(new_data[, 1])

# Number of variables and groups
p = 3
g = 3

# Total sample size
n = nrow(new_data)

# Manually compute the W matrix
W = (Ns[1] - 1) * VMat[[1]] + (Ns[2] - 1) * VMat
    [[2]] + (Ns[3] - 1) * VMat[[3]]
```

```
# Load heplots and MVN libraries
library(heplots)
library(MVN)

# Split data by group
new_data_gr = split(new_data, new_data[, 1])

# Set plot layout to 2 rows and 2 columns
par(mfrow = c(2, 2))

# Check multivariate normality for each group
for (i in 1:g) {
  X = new_data_gr[[i]][, -1]
  cqplot(X, id.n = 3)
  print(mvn(X, mvnTest = "royston"))
  print(mvn(X, mvnTest = "hz"))
}

# Extract response variables as a matrix
Y = as.matrix(new_data[, -1])

# Extract grouping variable as a factor
Gr = as.factor(new_data[, 1])

# Fit a linear model
LM.res = lm(Y ~ Gr)

# Perform MANOVA and summarize results using Wilks'
    red↪ lambda
SUM = summary(Manova(LM.res), "Wilks")
SUM

# Extract the within-group sum of squares and cross
    red↪ -products matrix
W = SUM$SSPE

# Load biotools library for pairwise comparisons
library(biotools)

# Perform pairwise comparisons using Wilks' lambda
    red↪ and Bonferroni adjustment
mvpaircomp(LM.res, factor1 = "Gr", test = "Wilks",
    red↪ adjust = "bonferroni")

# Perform univariate ANOVA for each variable
summary.aov(LM.res)

# Number of comparisons
k = p * g * (g - 1) / 2

# Compute critical value for Bonferroni adjustment
ta = qt(.05 / (2 * k), df = n - g, lower.tail =
    red↪ FALSE)

# Simultaneous confidence intervals
i = 2 # Group 2
j = 3 # Group 3
v = 1 # Variable
LL = VMeans[[i]] - VMeans[[j]] - ta * sqrt((1 / Ns[
    red↪ i] + 1 / Ns[j]) * diag(W) / (n - g))
UL = VMeans[[i]] - VMeans[[j]] + ta * sqrt((1 / Ns[
    red↪ i] + 1 / Ns[j]) * diag(W) / (n - g))
cbind(LL, UL)

i = 1 # Group 1
j = 3 # Group 3
v = 1 # Variable
LL = VMeans[[i]] - VMeans[[j]] - ta * sqrt((1 / Ns[
    red↪ i] + 1 / Ns[j]) * diag(W) / (n - g))
UL = VMeans[[i]] - VMeans[[j]] + ta * sqrt((1 / Ns[
    red↪ i] + 1 / Ns[j]) * diag(W) / (n - g))
cbind(LL, UL)
```