DiFrank, Annie

Course #537

HW# 3

February 13, 2025

HW3

2025-02-11

## Problem 1

dat<-read.table(file="data/middleschool.txt", sep=",", header=TRUE)  
colnames(dat)

## [1] "ID" "Math" "ELA" "Science"   
## [5] "SocialStudies" "school"

head(dat)

## ID Math ELA Science SocialStudies school  
## 1 1 78.80 81.26 77.81 91.92 A  
## 2 2 96.64 90.50 87.93 100.00 A  
## 3 3 79.43 79.09 84.19 87.05 A  
## 4 4 85.19 80.48 85.48 89.87 A  
## 5 5 73.96 81.50 76.13 91.08 A  
## 6 6 79.42 77.55 78.05 88.53 A

1. **Test and design**

test: Hotelling’s T2 Test

* design: two sample

1. **Performing the test**

Assumptions: both populations are multivariate normal, both have the same variance-covariance matrix, and the two samples are mutually independant.

library(ICSNP)

## Loading required package: mvtnorm

## Warning: package 'mvtnorm' was built under R version 4.3.3

## Loading required package: ICS

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

#Filter data for School A and School B  
school\_A <- dat[dat$school == "A", ]  
school\_B <- dat[dat$school == "B", ]  
school\_A <- school\_A %>% select("Math","ELA","Science","SocialStudies")  
school\_B <- school\_B %>% select("Math","ELA","Science","SocialStudies")  
#we have large N for both samples (100 and 150), so can assume normality   
  
HotellingsT2(as.matrix(school\_A),as.matrix(school\_B))

##   
## Hotelling's two sample T2-test  
##   
## data: as.matrix(school\_A) and as.matrix(school\_B)  
## T.2 = 0.52139, df1 = 4, df2 = 245, p-value = 0.7201  
## alternative hypothesis: true location difference is not equal to c(0,0,0,0)

1. **Interpretation**

The p-value tells us how likely it is to see a difference as larger or larger than the observed one if the two schools actually performed the same. The alternative test checks whether the overall performance difference across subjects is significantly different from zero. Since the p-value (0.72) is much greater than the standard 0.05, we fail to reject the null hypothesis. There is not statistically significant evidence that one school performed significantly better than the other overall (across the four classes). Based on the data, the two school appear to have similar academic performance. This result agrees with my findings in HW2, where the bon ferroni confidence interval for each subject between schools overlapped, which was not suggestive of any significant difference between the schools in each subject.

## Problem 2

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

#head(Pottery)

dat2<- as.matrix(Pottery[,-1])  
site <- Pottery[,1]  
#table(site)

1. **Mathematical hypothesis to test and perform an appropriate hypothesis test:**

H0:μ1=μ=μ3=μ4

HA: At least one site’s mean composition is different

Where μi represents the mean vector of chemical compositions for the ith site.

Test: MANOVA, as we are dealing with multiple dependent variables across different groups.

site <- as.factor(site)  
  
out<- manova(dat2~site)  
  
#view results  
summary(out, test = "Wilks")

## Df Wilks approx F num Df den Df Pr(>F)   
## site 3 0.012301 13.088 15 50.091 1.84e-12 \*\*\*  
## Residuals 22   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Because the p value is less than 0.05, we reject Ho meaning at least one site has significantly different mean chemical composition.

1. **Investigating which components are different using pair-wise comparisons**

library(emmeans)

## Warning: package 'emmeans' was built under R version 4.3.3

## Welcome to emmeans.  
## Caution: You lose important information if you filter this package's results.  
## See '? untidy'

#mean of each variable   
Almeans<- emmeans(out, "site", weights=c(1,0,0,0,0))  
Almeans

## site emmean SE df lower.CL upper.CL  
## AshleyRails 17.3 0.663 22 15.95 18.7  
## Caldicot 11.7 1.048 22 9.53 13.9  
## IsleThorns 18.2 0.663 22 16.81 19.6  
## Llanedyrn 12.6 0.396 22 11.74 13.4  
##   
## Results are averaged over the levels of: rep.meas   
## Confidence level used: 0.95

Femeans<- emmeans(out, "site", weights=c(0,1,0,0,0))  
Mgmeans<- emmeans(out, "site", weights=c(0,0,1,0,0))  
Cameans<- emmeans(out, "site", weights=c(0,0,0,1,0))  
Nameans<- emmeans(out, "site", weights=c(0,0,0,0,1))  
  
#adjust alpha accordingly due to multiple comparisons   
p<- 5 #number of variables   
g<- 4 #number of groups   
alpha.old <- 0.05  
#number of comparison   
nc <- p\*g\*(g-1)/2  
#new significance level   
alpha.new <- alpha.old/nc

Obtain the contrasts:

Alcont <- contrast(Almeans, "pairwise")  
Fecont <- contrast(Femeans, "pairwise")  
Mgcont <- contrast(Mgmeans, "pairwise")  
Cacont <- contrast(Cameans, "pairwise")  
Nacont <- contrast(Nameans, "pairwise")

Pairwise differences for Al

al\_pw <- confint(Alcont, level = 1-alpha.new, adj="none")  
al\_pw

## contrast estimate SE df lower.CL upper.CL  
## AshleyRails - Caldicot 5.620 1.240 22 1.18 10.06  
## AshleyRails - IsleThorns -0.860 0.937 22 -4.22 2.50  
## AshleyRails - Llanedyrn 4.756 0.772 22 1.99 7.52  
## Caldicot - IsleThorns -6.480 1.240 22 -10.92 -2.04  
## Caldicot - Llanedyrn -0.864 1.120 22 -4.87 3.15  
## IsleThorns - Llanedyrn 5.616 0.772 22 2.85 8.38  
##   
## Results are averaged over the levels of: rep.meas   
## Confidence level used: 0.998333333333333

There are significant differences in Aluminum content between AshleyRails (AR) and Caldicot, AR and Llanedyrn, and IsleThorns (IT) and Llanedyrn.

Pairwise differences for Fe

fe\_pw <- confint(Fecont, level = 1-alpha.new, adj="none")  
fe\_pw

## contrast estimate SE df lower.CL upper.CL  
## AshleyRails - Caldicot -3.903 0.590 22 -6.02 -1.789  
## AshleyRails - IsleThorns -0.200 0.446 22 -1.80 1.398  
## AshleyRails - Llanedyrn -4.860 0.368 22 -6.18 -3.544  
## Caldicot - IsleThorns 3.703 0.590 22 1.59 5.817  
## Caldicot - Llanedyrn -0.957 0.533 22 -2.87 0.953  
## IsleThorns - Llanedyrn -4.660 0.368 22 -5.98 -3.344  
##   
## Results are averaged over the levels of: rep.meas   
## Confidence level used: 0.998333333333333

There are significant difference in Iron content between AR and Caldicot, AR and LLanedyrn, Caldicot and IT, and IT and Llanedyrn.

Pairwise differences for Mg

mg\_pw <- confint(Mgcont, level = 1-alpha.new, adj="none")  
mg\_pw

## contrast estimate SE df lower.CL upper.CL  
## AshleyRails - Caldicot -3.249 0.701 22 -5.758 -0.74  
## AshleyRails - IsleThorns -0.068 0.530 22 -1.965 1.83  
## AshleyRails - Llanedyrn -4.220 0.436 22 -5.783 -2.66  
## Caldicot - IsleThorns 3.181 0.701 22 0.672 5.69  
## Caldicot - Llanedyrn -0.971 0.633 22 -3.238 1.30  
## IsleThorns - Llanedyrn -4.152 0.436 22 -5.715 -2.59  
##   
## Results are averaged over the levels of: rep.meas   
## Confidence level used: 0.998333333333333

There are significant differences in magnesium content between AR and IT, AR and Llanedyrn, Caldicot and IT, and IT and Llanedyrn.

Pairwise differences for Ca

ca\_pw <- confint(Cacont, level = 1-alpha.new, adj="none")  
ca\_pw

## contrast estimate SE df lower.CL upper.CL  
## AshleyRails - Caldicot -0.2430 0.0405 22 -0.3879 -0.0981  
## AshleyRails - IsleThorns 0.0260 0.0306 22 -0.0836 0.1356  
## AshleyRails - Llanedyrn -0.1501 0.0252 22 -0.2404 -0.0599  
## Caldicot - IsleThorns 0.2690 0.0405 22 0.1241 0.4139  
## Caldicot - Llanedyrn 0.0929 0.0366 22 -0.0381 0.2238  
## IsleThorns - Llanedyrn -0.1761 0.0252 22 -0.2664 -0.0859  
##   
## Results are averaged over the levels of: rep.meas   
## Confidence level used: 0.998333333333333

There are significant differences in Cadmium content between AR and Caldicot, AR and Llanedyrn, Caldicot and IT, and IT and Llanedyrn.

Pairwise differences for Na

na\_pw <- confint(Nacont, level = 1-alpha.new, adj="none")  
na\_pw

## contrast estimate SE df lower.CL upper.CL  
## AshleyRails - Caldicot -0.002 0.0796 22 -0.287 0.2831  
## AshleyRails - IsleThorns -0.006 0.0602 22 -0.222 0.2096  
## AshleyRails - Llanedyrn -0.203 0.0496 22 -0.380 -0.0252  
## Caldicot - IsleThorns -0.004 0.0796 22 -0.289 0.2811  
## Caldicot - Llanedyrn -0.201 0.0719 22 -0.458 0.0569  
## IsleThorns - Llanedyrn -0.197 0.0496 22 -0.374 -0.0192  
##   
## Results are averaged over the levels of: rep.meas   
## Confidence level used: 0.998333333333333

There are significant differences in sodium content between AR and Llanedyrn and IT and Llanedyrn.

1. **Assumptions**

The assumptions are that when comparing group means the covariance matrices of any two different populations are equal, each population is multivariate normal, and the samples are mutually independent.

## Problem 3

dat3<- read.table("data/T6-17.dat", header=F)  
colnames(dat3) <- c("Location","Variety","Yield","SdMatKer","Size")  
#dat3  
  
dat3$Location = as.factor(dat3$Location)  
dat3$Variety = as.factor(dat3$Variety)

1. **Performing MANOVA**  
   Testing for a location-variety interaction effect, location effect, and variety effect

corn <- as.matrix(dat3[,3:5])  
 location <- dat3$Location  
 variety <- dat3$Variety  
   
 #First, will test for the interaction effect.   
 loc\_var\_int <- lm(corn ~ location \* variety)  
 fitcorn <- manova(loc\_var\_int)  
 summary(fitcorn, test="Wilks")

## Df Wilks approx F num Df den Df Pr(>F)   
## location 1 0.106516 11.1843 3 4 0.020502 \*   
## variety 2 0.012444 10.6191 6 8 0.001928 \*\*  
## location:variety 2 0.074300 3.5582 6 8 0.050794 .   
## Residuals 6   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Here we see that (just barely) there is no evidence of an interaction effect, but main effects of both factors are significant.

Looking at the additive model (main effects of location and variety):

corn\_add <- lm(corn ~ location + variety)  
fitaddcorn <- manova(corn\_add)  
summary(fitaddcorn,test = "Wilks")

## Df Wilks approx F num Df den Df Pr(>F)   
## location 1 0.233390 6.5693 3 6 0.025249 \*   
## variety 2 0.084307 4.8881 6 12 0.009488 \*\*  
## Residuals 8   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Both factor effects have significant (<0.05) p values, indicating that the response mean differs among the various groups.

1. **Investigating whether the interaction shows up for some variables but not others by running three univariate ANOVA models**

summary.aov(loc\_var\_int)

## Response Yield :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## location 1 0.701 0.701 0.0404 0.84743   
## variety 2 196.115 98.057 5.6460 0.04177 \*  
## location:variety 2 205.102 102.551 5.9048 0.03824 \*  
## Residuals 6 104.205 17.367   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Response SdMatKer :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## location 1 162.07 162.07 2.7617 0.14761   
## variety 2 1089.01 544.51 9.2786 0.01459 \*  
## location:variety 2 780.70 390.35 6.6517 0.03003 \*  
## Residuals 6 352.10 58.68   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Response Size :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## location 1 72.521 72.521 4.5882 0.07594 .  
## variety 2 284.102 142.051 8.9872 0.01567 \*  
## location:variety 2 85.952 42.976 2.7190 0.14435   
## Residuals 6 94.835 15.806   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The interaction between location and variety is significant for the Yield and SdMatKer variables, but not for the Size variable.

1. **Interpretation**

The effect of the variety variable (The level of the variety factor) on the outcome (peanut measurements) depends on the value of the location variable (location of peanut), and vice versa. With this significant interaction, you can’t interpret the effects of each variable seperately because they influence eachother. For example, If location is = 1, whether variety is = 6 or 8 will lead to a different effect on the response variable(s).

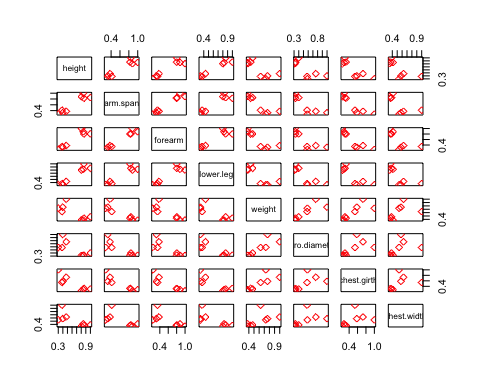
## Problem 4

library(datasets)  
Harman23.cor$cov

## height arm.span forearm lower.leg weight bitro.diameter  
## height 1.000 0.846 0.805 0.859 0.473 0.398  
## arm.span 0.846 1.000 0.881 0.826 0.376 0.326  
## forearm 0.805 0.881 1.000 0.801 0.380 0.319  
## lower.leg 0.859 0.826 0.801 1.000 0.436 0.329  
## weight 0.473 0.376 0.380 0.436 1.000 0.762  
## bitro.diameter 0.398 0.326 0.319 0.329 0.762 1.000  
## chest.girth 0.301 0.277 0.237 0.327 0.730 0.583  
## chest.width 0.382 0.415 0.345 0.365 0.629 0.577  
## chest.girth chest.width  
## height 0.301 0.382  
## arm.span 0.277 0.415  
## forearm 0.237 0.345  
## lower.leg 0.327 0.365  
## weight 0.730 0.629  
## bitro.diameter 0.583 0.577  
## chest.girth 1.000 0.539  
## chest.width 0.539 1.000

1. **Plot to visualize the correlation matrix**

pairs(Harman23.cor$cov, pch=5, col="red")



corr\_data <- Harman23.cor$cov

There are general positive correlations between height/length-related measurements (height, arm span, forearm, lower leg). There seems to be slightly negative correlations between length related measurements and weight/bitro/chest related measurements. There are mildly positive correlations between the weight variable and chest and bistro measurements, but they don’t seem to be as strong as the positive relationships between the height/length measurements.

1. **Performing PCA**

#Checking if we need to scale the variables due to differences in scales/variances  
round(apply(corr\_data,2,sd),3)

## height arm.span forearm lower.leg weight   
## 0.271 0.296 0.304 0.279 0.222   
## bitro.diameter chest.girth chest.width   
## 0.246 0.267 0.217

All the standard deviations are generally similar. They are all human body measurements, so the scales aren’t super different.

data\_pca <- prcomp(corr\_data)  
summary(data\_pca)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 0.6907 0.18301 0.16453 0.1073 0.07623 0.05325 0.04022  
## Proportion of Variance 0.8528 0.05987 0.04839 0.0206 0.01039 0.00507 0.00289  
## Cumulative Proportion 0.8528 0.91267 0.96106 0.9817 0.99204 0.99711 1.00000  
## PC8  
## Standard deviation 3.47e-17  
## Proportion of Variance 0.00e+00  
## Cumulative Proportion 1.00e+00

0.8528 of the variance is captured by the first PC and 0.05987 of the variance is captured by the second PC. Together, these capture 0.91267 of the variance.

1. **Retaining the first two PCs and interpreting their loadings in the context of the problem**

round(data\_pca$rotation[,1:2],3)

## PC1 PC2  
## height -0.377 0.163  
## arm.span -0.423 -0.084  
## forearm -0.432 0.017  
## lower.leg -0.391 0.177  
## weight 0.283 0.305  
## bitro.diameter 0.301 0.191  
## chest.girth 0.346 0.397  
## chest.width 0.222 -0.805

The first PC can be constructed as

*Y1 = -0.377(height) -0.423(arm.span) - 0.432(forearm) -0.391(lower.leg) + 0.283(weight) + 0.301(bitro.diameter) + 0.346(chest.girth) + 0.222(chest.width)*

There are strong negative loadings for height-related variables and positive loadings for weight-related variabnles. It looks like this can be interpreted as a contrast between the height and weight traits. Individuals with higher values for height-related traits will have lower PC1 scores, and vice versa.

This represents a body size distinction between taller and leaner individuals (negative PC1 scores) versus broader and heavier individuals (positive PC1 scores).

The second PC can be constructed as

*Y2= 0.163(height) -0.084(arm.span) + 0.017(forearm) +0.177(lower.leg) + 0.305(weight) + 0.195(bitro.diameter) + 0.397(chest.girth) - 0.805(chest.width)*

The PC has more moderate positive loadings for weight-related variables, but chest width has a very strong negative loading. We also see small positive loadings for some height-related variables.

This suggests that PC2 differentiates individuals based on body proportions rather than overall size. Higher PC2 indicates larger chest girth and weight, lower indicates wider chest width but lower values for weight-related traits.

This represents a body shape distinction, contrasting people with wide chests but lower weight (negative PC2 scores) to those with larger chest girth and overall weight (positive PC2 scores).