Mahalanobis Distance

Vaishnavi

2025-01-13

# Loading the necessary libraries  
library(haven) # For importing Stata files

## Warning: package 'haven' was built under R version 4.2.3

library(MASS) # For Mahalanobis distance computation  
  
  
data <- read\_dta("D:/Desktop/Yale/Stats for business and society/random\_groups\_data\_with\_social\_group.dta") # Import the Stata file  
View(data)

### Part 1: Creating the dataset

# Data setup with covariates for two treatment groups (T1, T2) and control (C)  
T1 <- data.frame(group = 'T1', cov1 = rnorm(30), cov2 = rnorm(30))  
T2 <- data.frame(group = 'T2', cov1 = rnorm(33), cov2 = rnorm(33))  
C <- data.frame(group = 'C', cov1 = rnorm(33), cov2 = rnorm(33))  
  
# Combine the groups into one dataset  
data\_combined <- rbind(T1, T2, C)  
  
# Calculate the covariance matrix  
# Select the covariates (exclude the 'group' column)  
covariates <- data\_combined[, c('cov1', 'cov2')]  
cov\_matrix <- cov(covariates)

### Part 2: Compute the Mahalanobis distance

# Mean vector of the control group  
control\_mean <- colMeans(C[, c('cov1', 'cov2')])  
  
# Mahalanobis distance for each observation in the combined dataset to the control group mean  
mahalanobis\_distance <- mahalanobis(covariates, center = control\_mean, cov = cov\_matrix)  
  
# Add the Mahalanobis distance to the dataset for reference  
data\_combined$mahalanobis\_distance <- mahalanobis\_distance

### Part 3: Inspect the results

# View the dataset with Mahalanobis distance  
head(data\_combined)

## group cov1 cov2 mahalanobis\_distance  
## 1 T1 1.7821821 1.3707613 4.6461790  
## 2 T1 -0.8072203 -0.4905972 1.3135177  
## 3 T1 0.3982021 0.4833838 0.2050832  
## 4 T1 -0.2681862 0.2258440 0.1546486  
## 5 T1 1.0514648 1.1883424 2.0121655  
## 6 T1 0.4739625 -1.7936978 3.2021231

# Converting group/treatment variable to factor  
data\_combined$group <- factor(data\_combined$group)  
  
# Checking the structure of the data  
str(data\_combined)

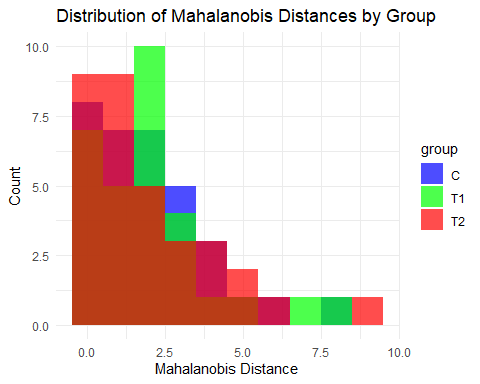
## 'data.frame': 96 obs. of 4 variables:  
## $ group : Factor w/ 3 levels "C","T1","T2": 2 2 2 2 2 2 2 2 2 2 ...  
## $ cov1 : num 1.782 -0.807 0.398 -0.268 1.051 ...  
## $ cov2 : num 1.371 -0.491 0.483 0.226 1.188 ...  
## $ mahalanobis\_distance: num 4.646 1.314 0.205 0.155 2.012 ...

### Part 4: Visualisations

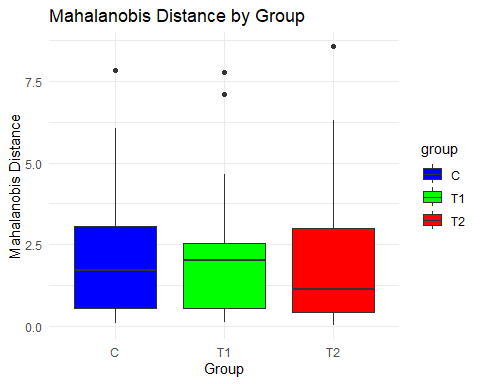
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.3

# 4.1 Histogram by group  
ggplot(data\_combined, aes(x = mahalanobis\_distance, fill = group)) + geom\_histogram(binwidth = 1, alpha = 0.7, position = 'identity') + labs(title = "Distribution of Mahalanobis Distances by Group",  
x = "Mahalanobis Distance", y = "Count") + theme\_minimal() + scale\_fill\_manual(values = c("blue", "green", "red"))

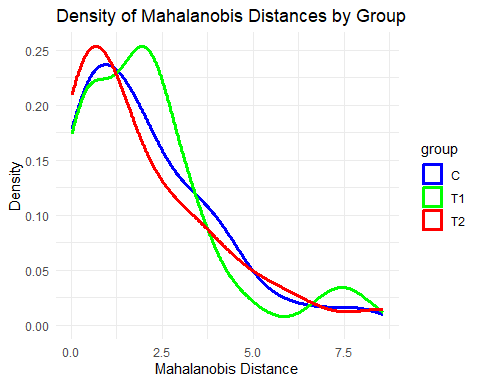


# 4.2 Boxplot by group  
ggplot(data\_combined, aes(x = group, y = mahalanobis\_distance, fill = group)) + geom\_boxplot() +  
labs(title = "Mahalanobis Distance by Group", x = "Group", y = "Mahalanobis Distance") +  
theme\_minimal() + scale\_fill\_manual(values = c("blue", "green", "red"))



# 4.3 Density plot of Mahalanobis distances by group  
ggplot(data\_combined, aes(x = mahalanobis\_distance, color = group)) + geom\_density(size = 1.2) + labs(title = "Density of Mahalanobis Distances by Group", x = "Mahalanobis Distance", y = "Density") +theme\_minimal() + scale\_color\_manual(values = c("blue", "green", "red"))

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



### Interpretation of the visualizations

### 1. The histogram shows the frequency distribution of Mahalanobis distances for each group.

### 2. The boxplot displays the spread of Mahalanobis distances for each group. A smaller distance indicates greater similarity to the control group.

### 3. The density plot provides a smooth estimate of the distribution, showing where the majority of Mahalanobis distances fall for each group.