Confirmation analysis

Maksym Polovnikov

Introduction

The aim of this assignment is to get familiar with multivariate confirmation analysis. As such, the objective of confirmation analysis is to test hypotheses or theories model. These models could be based on a previous empirical research. For this purpose, you will use multivariate analysis of variance (MANOVA) to test your hypothesis about given data.

The deadline for this assignment is 4.11.2019.

Input data

In this tutorial, we will work with a dataset that aims to the quality of potatoes growing in Oregon (for more details see [1]). Each potato is determined by its size, area, holding temperature, holding period, and cooking method. Overall quality consists of three aspects: texture score, flavor score, and moistness score. In this context, the properties of potatoes could be seen as independent variables; on the other hand, the quality of potatoes are dependent variables. For loading of the dataset, you can use a prepared function `Load`:

```
knitr::opts_chunk$set(echo = TRUE)

#load the dataset
Load <- function(path = "potato.dat"){
  potatoData <- read.table(path, header = FALSE)
  colnames(potatoData) <- c("Area", "Size" ,"Temp", "Period", "Method", "Texture", "Flavor", "Moistness"
  potatoData$Area <- as.factor(potatoData$Area)
  potatoData$Size <- as.factor(potatoData$Size)
  potatoData$Temp <- as.factor(potatoData$Temp)
  potatoData$Period <- as.factor(potatoData$Period)
  potatoData$Method <- as.factor(potatoData$Method)
  return(potatoData)
}

source("boxTest.R")
library(mvnormtest)
library(MVN)
```

```
## Warning: package 'MVN' was built under R version 3.5.2
```

```
## sROC 0.1-2 loaded
```

```
#load the dataset
potatoData <- Load()
```

MANOVA

There are four main theoretical issues to be considered before running MANOVA. Not surprisingly, MANOVA has similar assumptions to ANOVA but extended to the multivariable case:

- **Independence**: Observations should be statistically independent.
- **Random sampling**: Data are randomly sampled from the population of interest.

- **Multivariate normality**: Dependent variables are multivariate normally distributed within each group of the independent variables, which are categorical.
- **Homogeneity of covariance matrices**: The population covariance matrices of each group are equal.

Take the initial assumptions granted. The assumption of multivariate normality can be tested using R with a test known as the Shapiro test implemented in mshapiro.test in `mvnormtest` package. The assumption of equality of covariance matrices is often tested using Box's test that is implemented in `BoxMTest` function in `boxTest.R` file or as a `BoxM` function in biotools package.

Step by Step

You should go through the following steps:

1. Formalize your MANOVA hypothesis.
2. Visualize your data using plot, boxplot, etc. If data have more than two dimensions project them to lower dimmensionionality. (Optional: use a dimensionality reduction method instead.)
3. Check the assumptions for MANOVA model.
4. Use MANOVA to test your hypothesis.
5. Discuss the obtained results (the meaning of the individual statistics, comparison with the visual analysis ad 2, practical implications), compare with the results of simpler MANOVA alternatives (e.g. repeated ANOVA).

Submission Form

Submit your solution to the upload system. Submit the directory you have downloaded with only this file modified. Write all your code and answers directly into this file and leave the others unmodified.

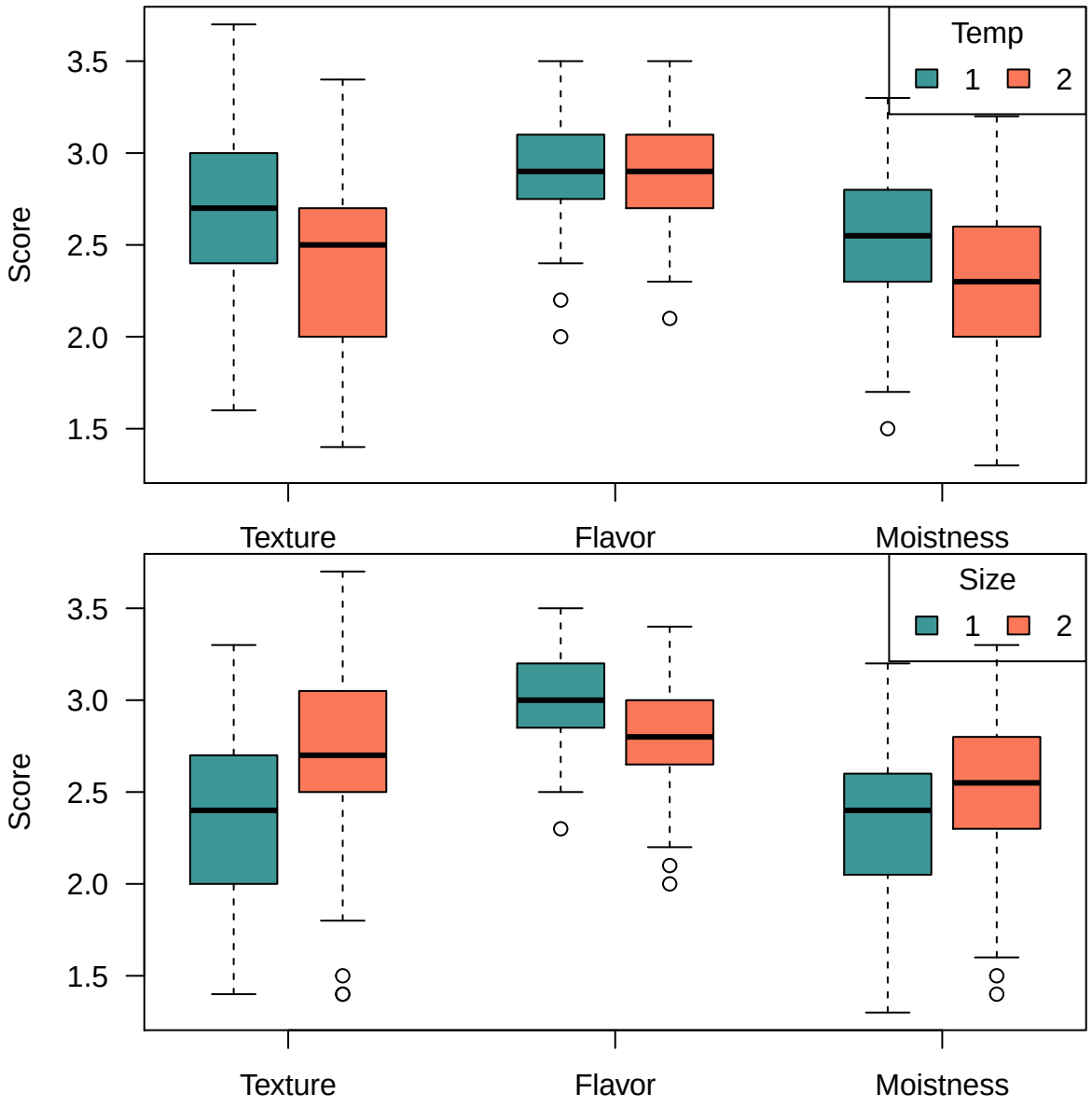
Hypothesis formulation

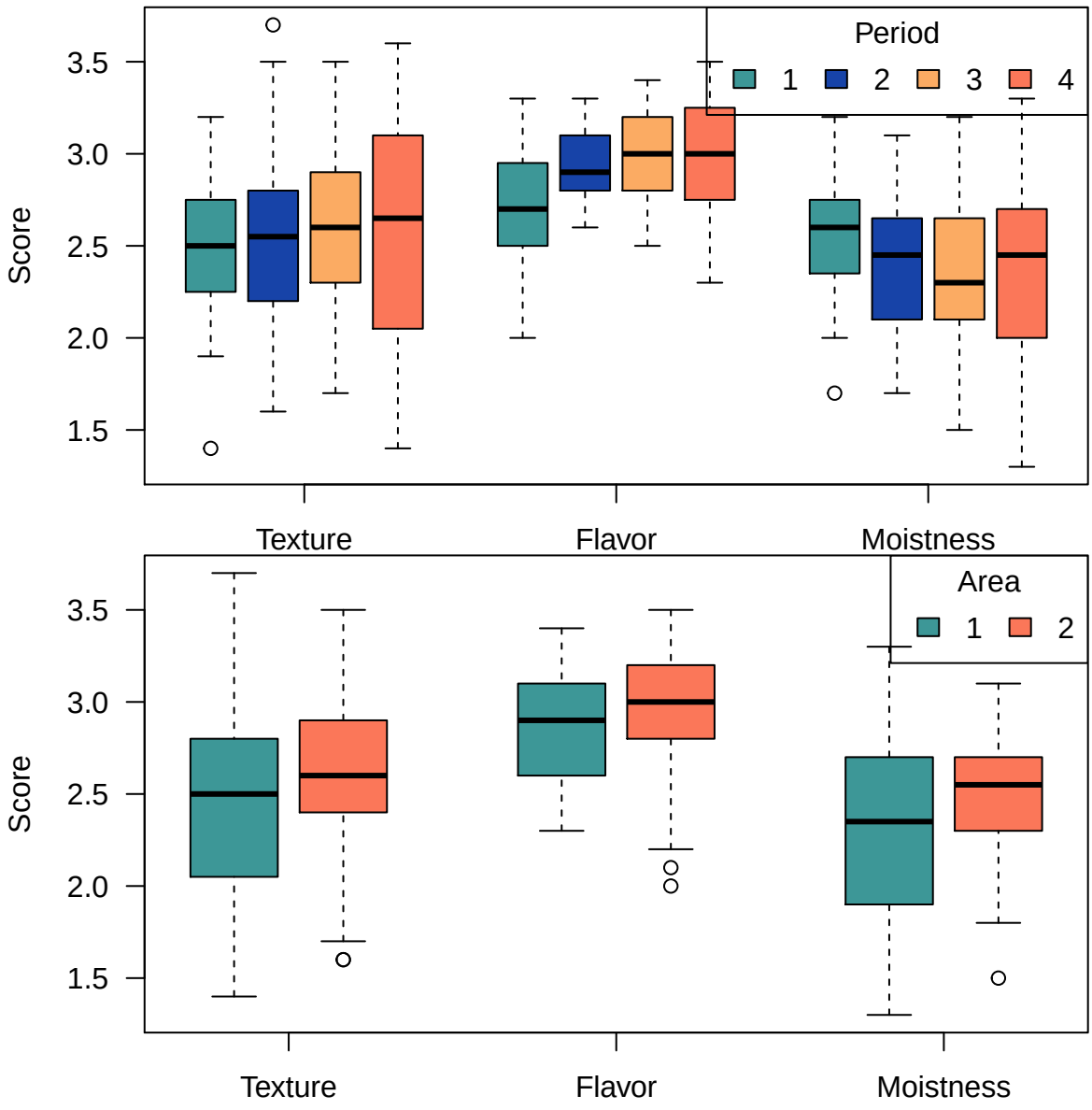
Let $\mu_{\mathbf{i}} \in \mathbb{R}^3$ denote the vector of population means of each score for the group i. Then hypothesis can be stated as follows:

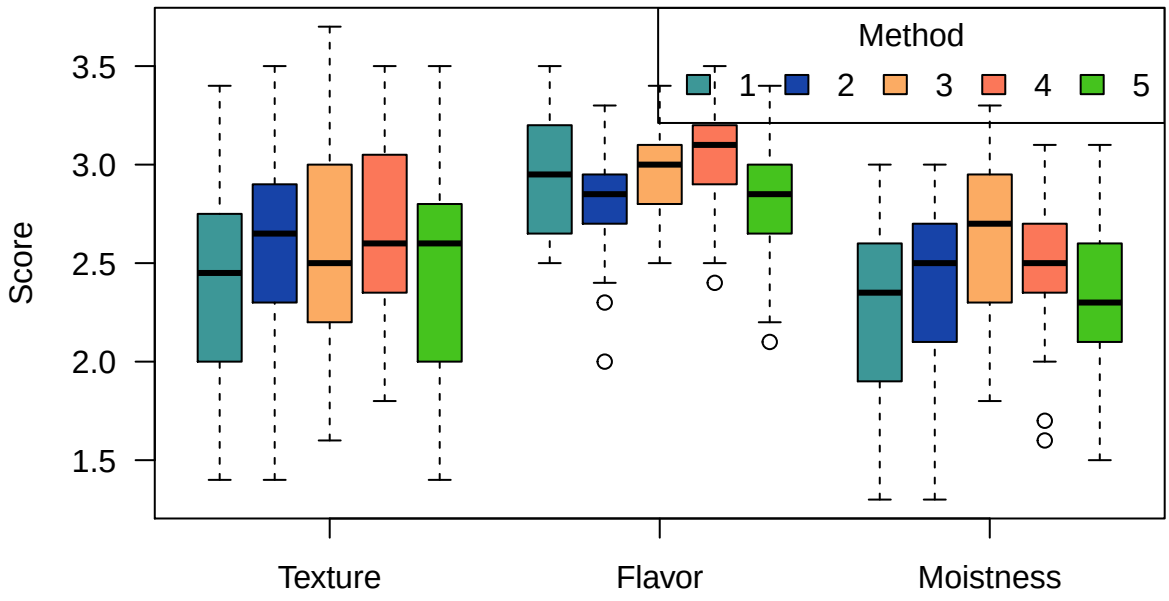
$H_0 : \mu_{\mathbf{1}} = \mu_{\mathbf{2}} = \cdots = \mu_{\mathbf{g}}$, $H_a : \mu_{\mathbf{ik}} \neq \mu_{\mathbf{jk}}$ for at least one $i \neq j$ and at least one k

This way, we are testing whether all potato groups do not vary in quality. In order to do this, we should specify the grouping of data samples. Possible options are: * Perform one-way MANOVA for each independent categorical variable or selected categorical variables * Perform multivariate MANOVA for each independent variables and their interactions To take a more simple way, for the first approuch we will select only one variable and it's "period" due to seasonality.

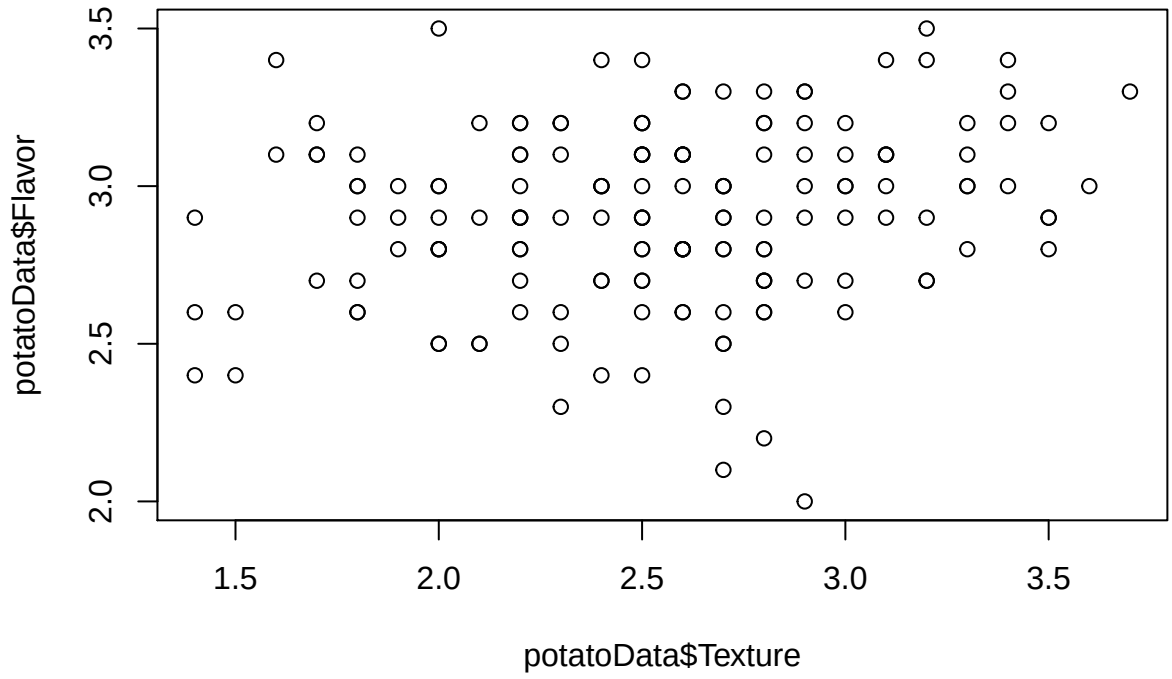
Independet variables visualisation



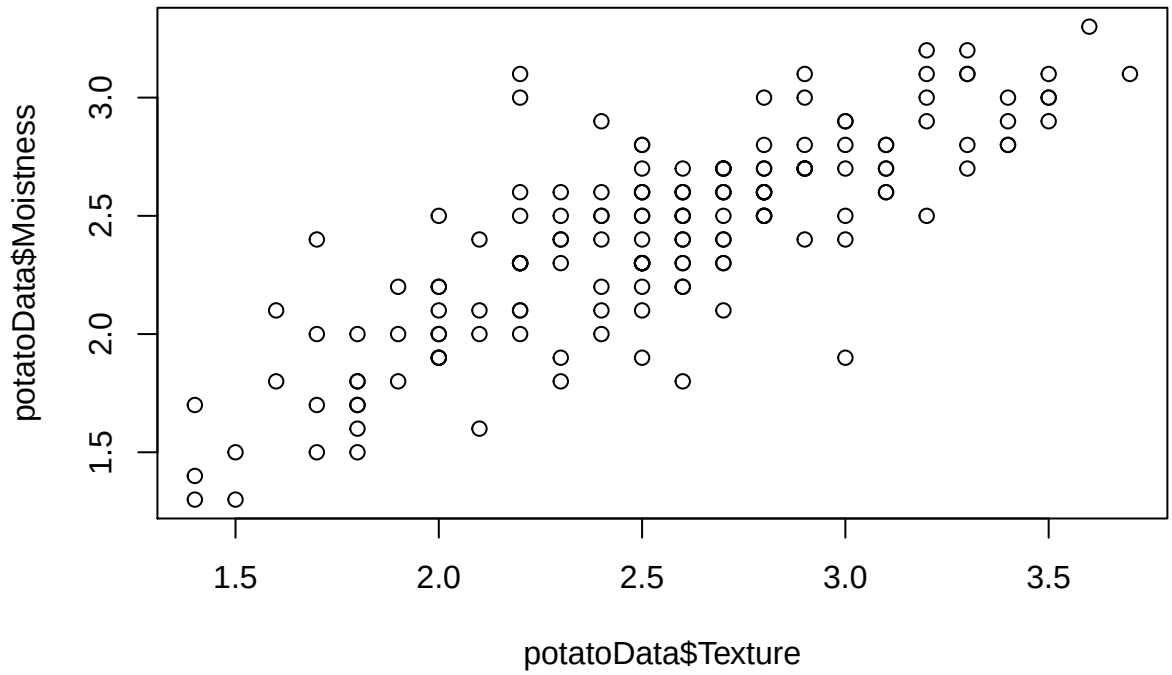

Score
3.5
3.0
2.5
2.0
1.5
Period
1
2
3
4
Texture
Flavor
Moistness
Score
3.5
3.0
2.5
2.0
1.5
Area
1
2
Texture
Flavor
Moistness



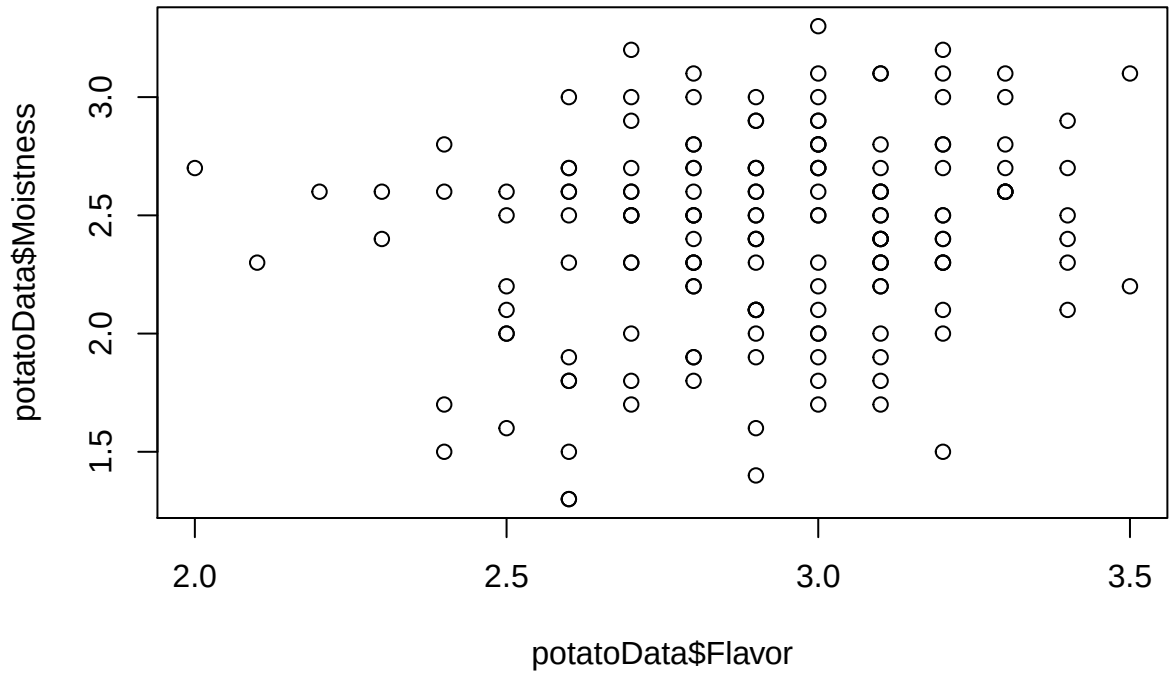

```
plot(potatoData$Texture, potatoData$Flavor)
```




```
plot(potatoData$Texture, potatoData$Moistness)
```




```
plot(potatoData$Flavor, potatoData$Moistness)
```




```
cor(potatoData[, c("Texture", "Flavor", "Moistness")])
```

```
##             Texture    Flavor Moistness
## Texture   1.0000000 0.1952333 0.8114908
## Flavor    0.1952333 1.0000000 0.1976560
## Moistness 0.8114908 0.1976560 1.0000000
```

Assumption verification

Multivariate normality test

```
selected.cat <- potatoData[c("Period", "Texture", "Flavor", "Moistness")]
U1 <- as.matrix(selected.cat[selected.cat$Period == 1, c("Texture", "Flavor", "Moistness")])
U2 <- as.matrix(selected.cat[selected.cat$Period == 2, c("Texture", "Flavor", "Moistness")])
U3 <- as.matrix(selected.cat[selected.cat$Period == 3, c("Texture", "Flavor", "Moistness")])
U4 <- as.matrix(selected.cat[selected.cat$Period == 4, c("Texture", "Flavor", "Moistness")])
mshapiro.test(t(U1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.9656, p-value = 0.2589
```

```
mshapiro.test(t(U2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.92189, p-value = 0.008834
```

```
mshapiro.test(t(U3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.97554, p-value = 0.5282
```

```
mshapiro.test(t(U4))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.97538, p-value = 0.5229
```

We can see a strong evidence against null hypothesis of multivariate normality in the second group. In other cases it is ok, which can be due outliers. We can remove tha, with Mehalanobis' distance:

```
dependent <- c("Texture", "Flavor", "Moistness")
x <- selected.cat[selected.cat$Period == 2, dependent]
D2 <- mahalanobis(x, colMeans(x), cov(x))
sort(D2, decreasing = TRUE)
```

```
##        148         68         48        108         66         26
## 11.3906623  9.5090492  7.8913194  6.4329094  6.0475351  5.1170776
##         88        126        127          6         50        130
##  4.9976864  4.4164019  4.0340225  3.8030422  3.5969999  3.5802902
##          9        129         28         29        150         89
##  3.4904100  3.2939837  2.6432978  2.4933582  2.2428986  2.1926776
##         47         49         67         46        106         10
##  2.1792274  2.1792274  2.1749529  2.1467769  2.0630838  2.0383663
```

```
##        146         70          8         69         30        147
##  1.9920605  1.6534340  1.6033937  1.5956768  1.5912468  1.5884358
##        107         86          7        109        128        149
##  1.0934094  0.8497191  0.7796454  0.7432331  0.6834748  0.6705454
##         27        110         87         90
##  0.6628984  0.6628984  0.6557041  0.2189677
```

```
selected.cat <- selected.cat[-c(148, 68, 48, 108), ]
U2 <- as.matrix(selected.cat[selected.cat$Period == 2, c("Texture", "Flavor", "Moistness")])
mshapiro.test(t(U2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.96619, p-value = 0.3307
```

Homogeniety of covariance matrices

```
X <- as.matrix(selected.cat[c("Texture", "Flavor", "Moistness")])
l <- selected.cat$Period
BoxMTest(X, l)
```

```
## ------------------------------------------------
##  MBox Chi-sqr. df P
## ------------------------------------------------
##    32.1048    30.9581          18       0.0291
## ------------------------------------------------
## Covariance matrices are significantly different.
```

```
## $MBox
##        1
## 32.10484
##
## $ChiSq
##        1
## 30.95807
##
## $df
## [1] 18
##
## $pValue
##          1
## 0.02911022
```

MANOVA application

```
res.man <- manova(cbind(Texture, Flavor, Moistness) ~ Period, data = selected.cat)
summary(res.man)
```

```
##            Df Pillai approx F num Df den Df    Pr(>F)
## Period      3 0.4008   7.8129      9    456 9.682e-11 ***
```

```
## Residuals 152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case we can see some difference in scores and that's why MANOVA tells us that there is a strong evidence against the null hypothesis.

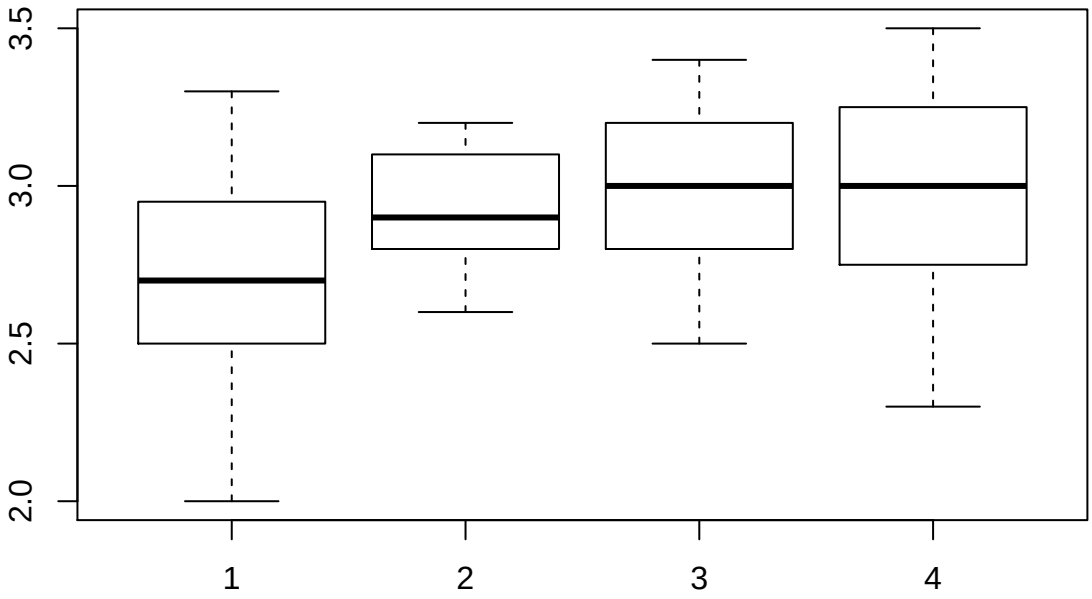
Results

As a result we can reject null hypothesis with high significance level, which means that period variable affects quality scores.

```
summary.aov(res.man)
```

```
##  Response Texture :
##              Df Sum Sq  Mean Sq F value Pr(>F)
## Period        3  0.159 0.052834  0.2092 0.8899
## Residuals   152 38.393 0.252585
##
##  Response Flavor :
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## Period        3  1.9952 0.66508  9.6398 7.282e-06 ***
## Residuals   152 10.4870 0.06899
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Moistness :
##              Df  Sum Sq Mean Sq F value Pr(>F)
## Period        3  1.0984 0.36613  2.1135 0.1008
## Residuals   152 26.3316 0.17323
```

```
plot(selected.cat$Period, selected.cat$Flavor)
```



ANOVA performed on each dependent variable confirms the result and additionally suggests that period variable affects only flavor score.

References

[1] Mackey, Stockman: Cooking Quality of Oregon-Grown Russet Potatoes, American Potato Journal, pp. 395–407, 1958.