A Nonparametric Alternative to MANOVA

Assoc. Prof. Dr Hasan BULUT
<hasan.bulut@omu.edu.tr >

10 November 2025

Introduction

- Multivariate normality tests
- Homogeneity of covariance matrices
- One-way MANOVA
- Permutation tests
- PERMANOVA
- Distance measures
- PISA Example

Multivariate Normality Test

Important of Multivariate Normality

Many multivariate statistical analysis methods, such as MANOVA and Hotelling T^2 Test, require multivariate normality (MVN) assumption. If the data are multivariate normal (exactly or approximately), such multivariate methods provide more reliable results.

The performances of these methods dramatically decrease if the data do not come from multivariate normal. So, researchers should check whether data are multivariate normal or not before continuing with such parametric multivariate analyses.

Hypothesis

H_0 : The multivariate data comes from multivariate normal distribution.

H_1 : The multivariate data does not come from multivariate normal distribution.

Mardia's Multivariate Normality Test

This test bases on multivariate skewness and kurtosis of squared Mahalanobis distances. MARDIA (1970) proposed test statistics based multivariate **skewness** α_3 as below:

$$\frac{n}{6}\alpha_3 \sim \chi^2_{p*(p+1)*(p+2)/6}$$

where

$$\alpha_3 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}^2$$

For small samples, the power and the type I error could be violated. Therefore, Mardia (1974) introduced a correction term into the skewness test statistic, usually when **n < 20**, in order to control type I error. We can write the corrected skewness statistic for small samples as below:

$$\frac{nk}{6}\alpha_3 \sim \chi^2_{p*(p+1)*(p+2)/6}$$

where

$$k = (p+1)(n+1)(n+3)/(n(n+1)(p+1) - 6.$$

Mardia's test statistic based on multivariate **kurtosis** is below:

$$\alpha_4 = \frac{1}{n}\sum_{j=1}^{n} m_{ii}^2 \sim N(p(p+2), 8p(p+2)/n)$$

We can perform Mardia's MVN test in a real data in R programming language by using the function `mvn` in the package **`MVN`**(Korkmaz, Goksuluk, and Zararsiz 2014).

Iris data

```
data(iris)
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

MVN Test for Setosa Subset

```
library(MVN)
# setosa subset of the Iris data
setosa <- iris[1:50, 1:4]
str(setosa)
```

```
## 'data.frame':    50 obs. of  4 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
```

```
mvn(data = setosa,mvn_test = "mardia")$multivariate_normality
```

```
##             Test Statistic p.value     Method      MVN
## 1 Mardia Skewness    25.664   0.177 asymptotic ✓ Normal
## 2 Mardia Kurtosis     1.295   0.195 asymptotic ✓ Normal
```

MVN Test for Iris Species (Setosa, Versicolor, Virginica)

```
mvn(data = iris,subset = "Species",
    mvn_test = "mardia")$multivariate_normality
```

```
##        Group            Test Statistic p.value      MVN
## 1     setosa Mardia Skewness    25.664   0.177 ✓ Normal
## 2     setosa Mardia Kurtosis     1.295   0.195 ✓ Normal
## 3 versicolor Mardia Skewness    25.185   0.194 ✓ Normal
## 4 versicolor Mardia Kurtosis    -0.572   0.567 ✓ Normal
## 5  virginica Mardia Skewness    26.271   0.157 ✓ Normal
## 6  virginica Mardia Kurtosis     0.153   0.879 ✓ Normal
```

Multivariate Shapiro-Wilk Test

We use Shapiro-Wilk test to determine whether or not a variable comes from univariate normal distribution. Shapiro–Wilk's statistic for testing the hypothesis of univariate normality based on a random sample of size n, $x_1, x_2, \ldots, x_n$, is defined as:

$$W_X = \frac{\hat{\sigma}_X^2}{s_X^2}$$

where s_X^2 is a sample variance, and $\hat{\sigma}_X^2 = [\sum_{i=1}^{n} a_i x_{(i)}]^2$, $x_{(i)}$ is the i_{th} order statistic a_i is a special parameter.

(Villasenor Alva and Estrada 2009) proposed a multivariate extension of this test. They proposed to calculate W_{X_i} statistic for each variable separetely. Then the multivariate Shapiro-Wilk test statistic is average of W test seperately the univariate normality for all p variables, and then we calculate the average of W_{X_i} values.

In R, we can perform the multivariate Shapiro-Wilk test by using the function `mvShapiro.Test()` in the package **`mvShapiroTest`**.

MVN Test for Iris Data

```
library(mvShapiroTest)
mvShapiro.Test(as.matrix(iris[,1:4]))
```

```
##
##  Generalized Shapiro-Wilk test for Multivariate Normality by
##  Villasenor-Alva and Gonzalez-Estrada
##
## data:  as.matrix(iris[, 1:4])
## MVW = 0.97327, p-value = 1.655e-06
```

Homogeneity of Covariance Matrices

Introduction

Like univariate parametric tests, parametric multivariate tests assume the data sets come from multivariate normal distribution with the same (common) covariance matrices. The second assumption is called as homogeneity scatter assumption. If this assumption is not met, it is called the multivariate Behrens-Fisher problem.

Actually, Under multivariate normality, we can use "Welch tests" when there is a multivariate Behrens-Fisher problem in our data. So, the assumption of homogeneity of scatters is important to decide which hypothesis tests will we use. For this reason, we should investigate homogeneity of covariance matrices before any multivariate test.

Hypothesis

$$H_0 = \Sigma_1 = \Sigma_2 = \ldots = \Sigma_g$$

$$H_1 = \text{At least one of } \Sigma_j\text{'s is different from others, (j=1,2,...,g)}$$

Here, g is a group number.

Box-M Test

Box proposed a test statistics to test the null hypothesis as below:

$$U = -2(1 - c_1)ln(M) \sim \chi^2_{\frac{p(p+1)(g-1)}{2}}$$

We reject the null hypothesis, when $U > \chi^2_{\frac{p(p+1)(g-1)}{2};\alpha}$

$$M = \frac{|\boldsymbol{S}_1|^{\frac{n_1-1}{2}}|\boldsymbol{S}_2|^{\frac{n_2-1}{2}} \dots |\boldsymbol{S}_g|^{\frac{n_g-1}{2}}}{|\boldsymbol{S}_{pl}|^{\sum_{i=1}^{g}\frac{n_i-1}{2}}}$$

$$\boldsymbol{S}_{pl} = \frac{\sum_{i=1}^{g}(n_i-1)\boldsymbol{S}_i}{\sum_{i=1}^{g}(n_i-1)} = \frac{\boldsymbol{E}}{df_E}$$

$$c_1 = \begin{cases} \left[\frac{2p^2+3p-1}{6(p+1)(g-1)}\right]\left[\sum_{i=1}^{g}\frac{1}{n_i-1} - \frac{1}{\sum_{i=1}^{g}(n_i-1)}\right] & , \quad n_i's \ are \ different \\ \left[\frac{(g+1)(2p^2+3p-1)}{6g(p+1)(n-1)}\right] & , \quad n_1 = n_2 = \cdots = n_g = n \end{cases}$$

Box's M Test in R

We can perform Box's M Test in R by using the function `BoxM()` in the package **`MVTests`** (Bulut 2019b)

```
library(MVTests)
```

```
##
## Attaching package: 'MVTests'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     iris
```

```
## The following object is masked from 'package:datasets':
##
##     iris
```

```
result<-BoxM(data = iris[,-5],group = iris$Species)
summary(result)
```

```
##        Box's M Test
##
## Chi-Squared Value = 140.943 , df = 20  and p-value: <2e-16
```

->

or directly,

```
BoxM(data = iris[,-5],group = iris$Species)
```

```
## $Chisq
## [1] 140.943
##
## $df
## [1] 20
##
## $p.value
## [1] 3.352034e-20
##
## $Test
## [1] "BoxM"
##
## attr(,"class")
## [1] "MVTests" "list"
```

One Way MANOVA

Introduction

Multivariate Analysis of Variance (MANOVA) is an extension of Analysis of Variance (ANOVA) that allows for the comparison of mean vectors across two or more independent groups. While ANOVA tests for differences in means of a single dependent variable across groups, MANOVA does this for multiple dependent variables simultaneously.

The primary purpose of MANOVA is to determine if there are any statistically significant differences in the mean vectors of multiple dependent variables among different groups. MANOVA tests the null hypothesis that the mean vectors of the dependent variables are the same across all groups.

Hypothesis

Null hypothesis (H_0): The mean vectors are equal across the groups.

Alternative hypothesis (H_A): At least one group's mean vector is different.

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_g$$

$$H_1 : \text{At least one of the } \mu_j\text{'s is different from others. (j=1,2,...,g)}$$

Assumptions

- **Multivariate Normality**: The dependent variables should be normally distributed within each group.
- **Homogeneity of Covariance Matrices**: The covariance matrices of the dependent variables should be equal across groups.
- Independence: Observations should be independent of each other.
- Linearity: Relationships among dependent variables should be linear.
- No Multicollinearity: Dependent variables should not be highly correlated.

Variability Matrices

MANOVA tests base on Within-Subjects Matrix (**W**) which represents the **variability within each group,** and Between-Subjects Matrix (**B**) which Represents the **variability between the group means themselves.** Like ANOVA, we divide the total matrix (T) into W and B matrices in MANOVA. We can calculate the total matrix as below:

$$\mathbf{T} = \sum_{k=1}^{g} \sum_{i=1}^{n_k} (\mathbf{x}_{ik} - \bar{\mathbf{X}})(\mathbf{x}_{ik} - \bar{\mathbf{X}})^T$$

We can calculate W and B matrices as below:

$$\mathbf{W} = \sum_{k=1}^{g} (n_k - 1)\mathbf{S}_k$$

$$\mathbf{B} = \mathbf{T} - \mathbf{W}$$

Test Statistics

MANOVA uses several test statistics to determine the significance of the differences in mean vectors among groups. The most common test statistics are:

- Wilks' Lambda
- Pillai's Trace
- Hotelling-Lawley Trace
- Roy's Largest Root (Rencher 2002).

Wilks' Lambda Statistic

$$\Lambda = \frac{\det(\mathbf{W})}{\det(\mathbf{T})}, 0 \leq \Lambda \leq 1$$

where $\mathbf{W}$ is the within-groups sum of squares and cross-products matrix, and $\mathbf{T}$ is the total sum of squares and cross-products matrix.

Λ statistic approximates χ^2 distribution as below:

$$L = -\left[(n-1) - \frac{p+g}{2}\right] ln(\Lambda) \sim \chi^2_{p(g-1)}$$

Pillai's Trace Test Statistic

$$T = \sum_{j=1}^{p} \frac{\lambda_i}{1 + \lambda_i}$$

where λ_i is i_{th} eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$. This statistic has an approximation to the F distribution.

Hotelling-Lawley Trace Test

$$T_0 = \mathrm{tr}(\mathbf{W}^{-1}\mathbf{B})$$

This statistic has an approximation to the χ^2 distribution as below:

$$nT_0^2 \sim \chi^2_{p(g-1)}.$$

This statistic has also another approximation to F distribution.

Roy's Largest Root Test

$$\theta_{\max} = \frac{\lambda_{\max}}{1 + \lambda_{\max}}$$

where $\lambda_{\max}$ is the largest eigenvalue. This statistic has an approximation to the F distribution.

MANOVA in R

The function `manova()` in **`stats`** package can be used to perform MANOVA in R (Team 2023; Bulut 2019a).

```
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```r
library(stats)
result.manova<-manova(cbind(Sepal.Length,
                            Petal.Length,
                            Sepal.Width,
                            Petal.Width) ~ Species,
                      data = iris)
```

Default result based on Pillai test

```
summary(result.manova)  # default test="Pillai"
```

```
##            Df Pillai approx F num Df den Df    Pr(>F)
## Species     2 1.1919   53.466      8    290 < 2.2e-16 ***
## Residuals 147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Result based on Pillai test

```
summary(result.manova,test = "Pillai")
```

```
##            Df Pillai approx F num Df den Df    Pr(>F)
## Species     2 1.1919   53.466      8    290 < 2.2e-16 ***
## Residuals 147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Result based on Roy test

```
summary(result.manova,test = "Roy")
```

```
##            Df    Roy approx F num Df den Df    Pr(>F)
## Species     2 32.192     1167      4    145 < 2.2e-16 ***
## Residuals 147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Result based on Hotelling-Lawley test

```
summary(result.manova,test = "Hotelling-Lawley")
```

```
##            Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
## Species     2           32.477   580.53      8    286 < 2.2e-16 ***
## Residuals 147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Result based on Wilks test

```
summary(result.manova,test = "Wilks")
```

```
##            Df    Wilks approx F num Df den Df    Pr(>F)
## Species     2 0.023439   199.15      8    288 < 2.2e-16 ***
## Residuals 147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Permutation Tests

Introduction

Permutation tests are non-parametric statistical tests and we use them to determine the significance of observed data. The basic logic of permutation tests is to compare observed data with the distribution of values obtained by randomly shuffling the data.

Scenario

Let's assume we want to investigate whether there is a difference between the PISA mathematics scores of Turkish and Slovak students. For this investigation, the null hypothesis would state that there is no difference.

We can define the null hypothesis as below:

$$H_0 : \text{There is no difference between the PISA mathematics scores of Turkish and Slovak students.}$$

or

$$H_0 : \mu_1 = \mu_2$$

Algorithm of Permutation Test

- **Step 0:**
- Collect the sample data from both Turkish and Slovak students.
- Compute the t-statistic for the observed data like in independent t test. Let this be denoted as t^*.

Permutation Procedure

- **Step 1: Combine** the data from both groups into a single dataset.
- **Step 2:** Randomly **shuffle** the combined dataset.
- **Step 3: Split** the shuffled data back into two groups, maintaining the original sample sizes for each group.
- **Step 4:** Calculate the t-statistic for the shuffled data. Denote this recalculated statistic as $t_i^*, i = 1, 2, \ldots, M$.

Repeat the permutation procedure **(Steps 1-4) M** times (e.g., 1000, 3000, 10,000 iterations) to generate a distribution of t_i^* statistics.

Calculate p-value

In permutation tests, the p value is proportion of permuted t_i^* statistics that are more extreme than the observed t^* statistic.

$$\text{p-value} = \frac{\#(t_i^* > t^*)}{M}$$

- If the p-value < 0.05, reject the null hypothesis (H_0).
- Conclude that there is a statistically significant difference between the PISA mathematics scores of Turkish and Slovak students.

Advantages of Permutation Tests

1. **Non-Parametric:** They do not rely on assumptions about the distribution of the data.

2. **Exact P-Values:** They provide exact p-values given a sufficient number of permutations (*Applied Multivariate Statistical Analysis* 2007).

Disadvantages of Permutation Tests

1. **Computationally Intensive:** They can require long computation times, especially with large datasets or a high number of permutations.

2. **Sample Size Sensitivity:** The accuracy of the test can be sensitive to the sample size; small sample sizes might not provide reliable results.

PERMUTATIONAL MANOVA: PERMANOVA

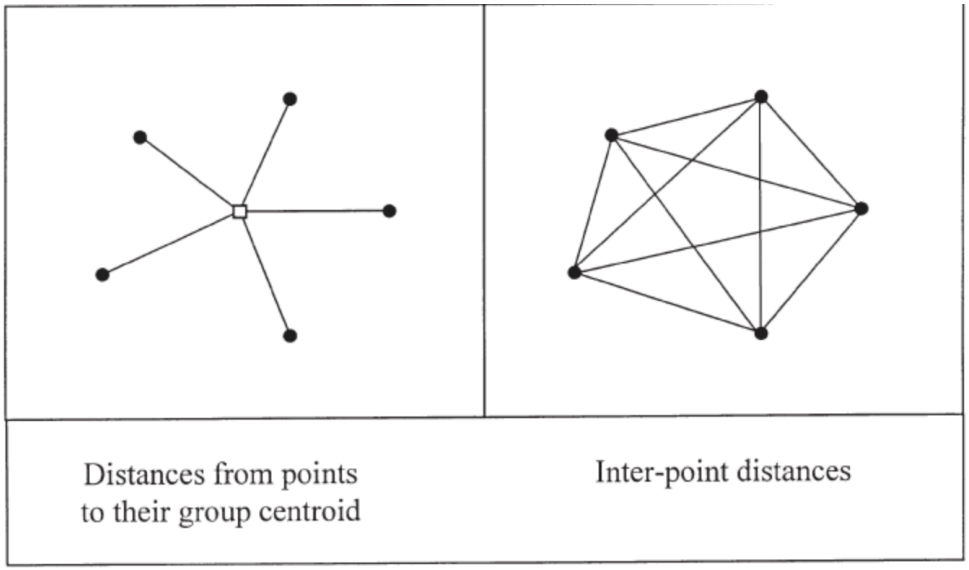
Introduction

PERmutational Multivariate ANalysis of VAriance (PERMANOVA) is a permutation-based technique. For this reason, there are no distributional assumptions about multivariate normality or homogeneity of variances in PERMANOVA (Bakker 2024).

Theory

The key idea behind PERMANOVA is that **the variation within a group can be calculated directly from a distance matrix.**

Specifically, the sum of squared differences between points and their centroid is equal to the sum of the squared interpoint distances divided by the number of points.



PERMANOVA, similar to MANOVA, partitions the total variation into two components: variation **among groups** and **within groups**. The test statistic used in this method, called **pseudo-F**, is calculated in the same way as the conventional ANOVA F-statistic. It represents the ratio of between-group variation to within-group variation, and higher values indicate stronger group differences.

The significance of this statistic is evaluated using a permutation procedure: the group labels of the observations are randomly reassigned many times, and the pseudo-F value is recalculated for each permutation to create a reference distribution.

Basic Procedure

The basic procedure for PERMANOVA is as follows.

1. Convert data matrix to a distance matrix, using an appropriate distance measure.
2. Square the distance matrix.
3. Calculate three partitions of the variation: SST, SSW, and SSB.

- Total variation (total sum of squares; *SST*) – the sum of squared distances divided by the number of points.
- Variation within groups (sum of squares within groups; *SSW*) – calculated identical to *SST* but separately for each group. Add the value from each group together to yield the overall *SSW*.
- Variation between groups (sum of squares between groups; *SSB*) – calculated by simple subtraction (i.e., *SSB* = *SST* – *SSW*).

4. Calculate the test statistic. The test statistic is termed a 'pseudo-*F* statistic' to distinguish it from the traditional parametric univariate *F*-statistic, but is calculated using the same formula:

$$PseudoF = \frac{SSB/(g-1)}{SSW/(n-g)}$$

where *SSB* and *SSW* are as calculated above, g is the number of groups, and n is the total number of sample units.

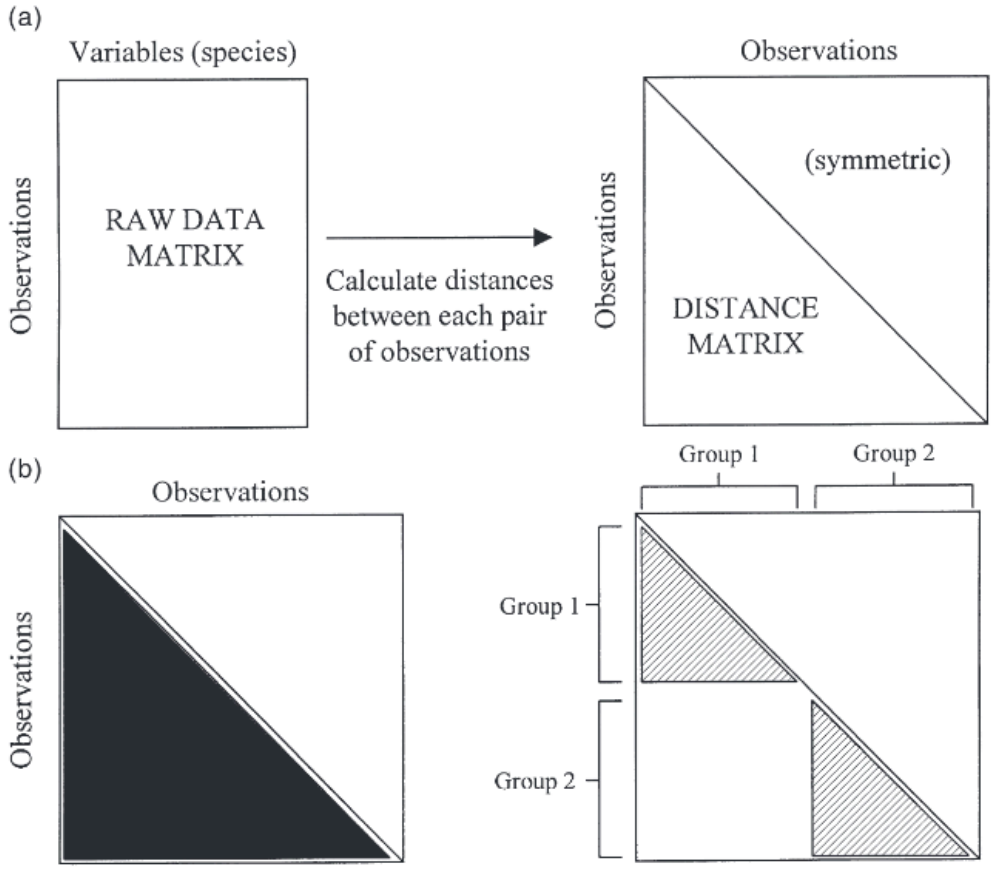


Fig. 3. Schematic diagram for the calculation of (a) a distance matrix from a raw data matrix and (b) a non-parametric MANOVA statistic for a one-way design (two groups) directly from the distance matrix. SS_T, sum of squared distances in the half matrix (■) divided by N (total number of observations); SS_W, sum of squared distances within groups (▨) divided by n (number of observations per group). $SS_A = SS_T - SS_W$ and $F = [SS_A/(a-1)]/[SS_W/(N-a)]$, where a = the number of groups.

5. Assess statistical significance via a permutation test:

    - *Permute group identities.*
    - *Recalculate pseudo-F statistic*
    - *Repeat the specified number of times.The permutations produce a distribution of pseudo-F values against which the actual pseudo-F statistic value is compared.*

6. Calculate the p-value as the proportion of permutations that yielded a pseudo-F value equal or greater than the actual data did.

$$p = \frac{\#(PseudoF_i \geq PseudoF)}{M}, i = 1, 2, \ldots, M$$

where M is the permutation number.

Basic Example by Hand

Let's define an imaginary dataset in R.

```
x1<-c(1,3,5,9,10,11)
x2<-c(4,2,3,12,8,11)
group<-c("A","A","A","B","B","B")
data<-data.frame(x1,x2,group)
rownames(data)<-c("a1","a2","a3","b1","b2","b3")
```

```
data
```

```
##    x1 x2 group
## a1  1  4     A
## a2  3  2     A
## a3  5  3     A
## b1  9 12     B
## b2 10  8     B
## b3 11 11     B
```

```r
# Calculate distance matrix
D<-dist(data[,1:2],diag = TRUE)  # eucledian distances
D
```

```
##           a1        a2        a3        b1        b2        b3
## a1  0.000000
## a2  2.828427  0.000000
## a3  4.123106  2.236068  0.000000
## b1 11.313708 11.661904  9.848858  0.000000
## b2  9.848858  9.219544  7.071068  4.123106  0.000000
## b3 12.206556 12.041595 10.000000  2.236068  3.162278  0.000000
```

```r
# Calculate squared distance matrix
D^2
```

```
##     a1  a2  a3  b1  b2  b3
## a1   0
## a2   8   0
## a3  17   5   0
## b1 128 136  97   0
## b2  97  85  50  17   0
## b3 149 145 100   5  10   0
```

$$SST = \frac{\sum d_{ij}^2}{n} = \frac{8 + 17 + 128 + \ldots + 17 + 5 + 10}{6} = 174.8333$$

```r
n<-nrow(data)
SST<-sum(D^2)/n    # (8+17+5+...+10)/6
SST
```

```
## [1] 174.8333
```

```
##        a1   a2   a3   b1   b2   b3
## a1      0
## a2      8    0
## a3     17    5    0
## b1    128  136   97    0
## b2     97   85   50   17    0
## b3    149  145  100    5   10    0
```

$$SSW = \sum_{k=1}^{g} \frac{\Sigma\, d_{ij}^2}{n_k} = \frac{8+17+5}{3} + \frac{17+5+10}{3} = \frac{30}{3} + \frac{32}{3} = 20.67$$

```
SSW=(8+17+5)/3+(17+5+10)/3
SSW
```

```
## [1] 20.66667
```

$$SSB = SST - SSW$$

```
SSB=SST-SSW
SSB
```

```
## [1] 154.1667
```

Often, we can summarize this information in an ANOVA table:

Source	**df**	**SS**	**MS**	**PseudoF**
Between	g-1=1	154.1667	154.1667	29.8387
Within	n-g=4	20.6667	5.1667	
Total	n-1=5	174.8333		

This table contains all of the data that we usually see in an ANOVA output except the *P*-value.

Statistical significance will be determined by **permuting** the group identities, recalculating the pseudo-*F* statistic for each permutation, and comparing the observed value against the distribution of values obtained via permutation. For this purpose, we can use R programming language.

PERMANOVA in R

We can use two popular functions to perform PERMANOVA in R:

- `adonis2()` in the package **vegan** (Oksanen et al. 2022).
- `PERMANOVA()` in the package **PERMANOVA** (Vicente-Gonzalez and Vicente-Villardon 2021).

adonis2() function

The usage of function `adonis2()` is:

```
adonis2(
formula,
data,
permutations = 999,
method = "bray",...,)
```

- `formula` – model formula such as `Y ~ A + B * C`.
  - *If Y is a dissimilarity matrix (i.e., output from `dist()` or `vegdist()`), the pseudo-F statistic is calculated directly.*
  - *If Y is a data frame or matrix containing data, `vegdist()` is applied to calculate the distance matrix first and then the pseudo-F statistic is calculated.*
- `data` – the data frame in which the explanatory variables are located. Sample units are assumed to be in the same order in `data` as in the rows of Y.
- `permutations` – number of permutations to conduct to assess the significance of the pseudo-*F* statistic. Default is 999 permutations.
- `method` – method of calculating the distance matrix.

Basic Example with the function adonis2()

```
# when we use data directly
library(vegan)
```

```
## Loading required package: permute
```

```
adonis2(data[,1:2]~group,
       data=data,
       method = "euclidean")
```

```
## 'nperm' >= set of all permutations: complete enumeration.
```

```
## Set of permutations < 'minperm'. Generating entire set.
```

```
## Permutation test for adonis under reduced model
## Permutation: free
## Number of permutations: 719
##
## adonis2(formula = data[, 1:2] ~ group, data = data, method = "euclidean")
##          Df SumOfSqs      R2      F Pr(>F)
## Model     1  154.167 0.88179 29.839    0.1
## Residual  4   20.667 0.11821
## Total     5  174.833 1.00000
```

```r
# when we use distance matrix and factor
D<-dist(data[,1:2])
grp<-as.factor(data$group)
adonis2(D~grp,data=data,method = "euclidean")
```

```
## 'nperm' >= set of all permutations: complete enumeration.
```

```
## Set of permutations < 'minperm'. Generating entire set.
```

```
## Permutation test for adonis under reduced model
## Permutation: free
## Number of permutations: 719
##
## adonis2(formula = D ~ grp, data = data, method = "euclidean")
##          Df SumOfSqs      R2      F Pr(>F)
## Model     1  154.167 0.88179 29.839    0.1
## Residual  4   20.667 0.11821
## Total     5  174.833 1.00000
```

`adonis2()` function returns an ANOVA table with p-value calculated from permutations. But it returns partial R^2 values in the ANOVA table instead of means of SS. These R^2 values are calculated as below:

$$R2_{grp} = \frac{SSB}{SST} = \frac{154.167}{174.833} = 0.88179$$

$$R2_{Residual} = \frac{SSW}{SST} = \frac{20.667}{174.833} = 1 - R2_{grp} = 0.11821$$

Distance Measures

Especially for nonparametric or permutational tests, we use distance measures instead of mean or centroids. for this reason, we will discuss about common distance measures in this section. In this section, we introduce distance measures of:

- Euclidean Distance
- Manhattan Distance
- Minkowski Distance
- Mahalanobis Distance
- Bray-Curtis Distance

Euclidean Distance

The Euclidean distance between two points $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ in n-dimensional space is defined as:

$$d_E(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Manhattan Distance

The Manhattan distance (or L1 distance) is given by:

$$d_M(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Minkowski Distance

The Minkowski distance is a generalization of both the Euclidean and Manhattan distances and is defined as:

$$d_{Min}(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

For $p = 2$, it is the Euclidean distance. For $p = 1$, it is the Manhattan distance.

Mahalanobis Distance

The Mahalanobis distance takes into account the correlations of the data set and is defined as:

$$d_{Mah}(x, y) = \sqrt{(\mathbf{x} - \mathbf{y})^T \mathbf{S}^{-1} (\mathbf{x} - \mathbf{y})}$$

where $\mathbf{S}$ is the covariance matrix of the data.

Bray-Curtis Distance

The Bray-Curtis distance is particularly useful when dealing with **count** data or **proportional** data. The Bray-Curtis distance (or dissimilarity) is a measure often used in ecology and other fields to quantify the compositional dissimilarity between two different sites or samples. It ranges from 0 to 1, where 0 indicates identical compositions and 1 indicates complete dissimilarity.

$$d_{BC}(x, y) = 1 - \frac{2 * C_{xy}}{S_x + S_y}$$

where:

- C_{xy}: The sum of the lesser values for the species found in each site.
- S_x: The total number of specimens counted at site x
- S_y: The total number of specimens counted at site y

Distance Measures in R

We can use `vegdist()` function in the **vegan** package to calculate Euclidean, Manhattan, Minkowski, Mahalanobis, and Bray-Curtis distances.

PISA 2022 Example

Import data

```
url<-"https://raw.githubusercontent.com/hsnbulut/PISA_Example/main/dataPISA.csv"
dataPISA<-read.csv(url,header = TRUE,sep = ";")
head(dataPISA,3)
```

```
##   CNT gender mathematics reading science
## 1 TUR   Male      372.86  397.79  413.01
## 2 TUR Female      446.69  448.86  463.50
## 3 TUR Female      429.03  481.61  464.37
```

Github Page: https://github.com/hsnbulut

Factorization of Country and Gender

```
dataPISA$CNT<-as.factor(dataPISA$CNT)
dataPISA$gender<-as.factor(dataPISA$gender)
head(dataPISA,3)
```

```
##   CNT gender mathematics reading science
## 1 TUR   Male      372.86  397.79  413.01
## 2 TUR Female      446.69  448.86  463.50
## 3 TUR Female      429.03  481.61  464.37
```

Structure of data

```
str(dataPISA)
```

```
## 'data.frame':    1500 obs. of  5 variables:
##  $ CNT        : Factor w/ 5 levels "CHE","CZE","FRA",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ gender     : Factor w/ 2 levels "Female","Male": 2 1 1 1 1 2 2 2 2 2 ...
##  $ mathematics: num  373 447 429 405 388 ...
##  $ reading    : num  398 449 482 417 373 ...
##  $ science    : num  413 464 464 349 399 ...
```

Compare PISA Scores According to Country

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

$$H_1 : \text{At least one of the } \mu_j\text{'s is different from others. (j=1,2,...,5)}$$

Assumptions controls:

- Multivariate Normality
- Homogeneity

Multivariate Normality

Multivariate Normality for CZEch Students

```
library(mvShapiroTest)
mvShapiro.Test(as.matrix(subset(dataPISA,
                                CNT == "CZE")[,3:5]))
```

```
##
##  Generalized Shapiro-Wilk test for Multivariate Normality by
##  Villasenor-Alva and Gonzalez-Estrada
##
## data:  as.matrix(subset(dataPISA, CNT == "CZE")[, 3:5])
## MVW = 0.99471, p-value = 0.4048
```

Multivariate Normality for ITAlian Students

```
mvShapiro.Test(as.matrix(subset(dataPISA,
                                CNT == "ITA")[,3:5]))
```

```
##
##  Generalized Shapiro-Wilk test for Multivariate Normality by
##  Villasenor-Alva and Gonzalez-Estrada
##
## data:  as.matrix(subset(dataPISA, CNT == "ITA")[, 3:5])
## MVW = 0.99493, p-value = 0.4706
```

Multivariate Normality for French (FRA) Students

```
mvShapiro.Test(as.matrix(subset(dataPISA,
                            CNT == "FRA")[,3:5]))
```

```
##
##  Generalized Shapiro-Wilk test for Multivariate Normality by
##  Villasenor-Alva and Gonzalez-Estrada
##
## data:  as.matrix(subset(dataPISA, CNT == "FRA")[, 3:5])
## MVW = 0.99501, p-value = 0.4954
```

Multivariate Normality for Swiss (CHE) Students

```
mvShapiro.Test(as.matrix(subset(dataPISA,
                            CNT == "CHE")[,3:5]))
```

```
##
##  Generalized Shapiro-Wilk test for Multivariate Normality by
##  Villasenor-Alva and Gonzalez-Estrada
##
## data:  as.matrix(subset(dataPISA, CNT == "CHE")[, 3:5])
## MVW = 0.99177, p-value = 0.0235
```

Multivariate Normality for Turkish (TUR) Students

```
mvShapiro.Test(as.matrix(subset(dataPISA,
                                CNT == "TUR")[,3:5]))
```

```
##
##  Generalized Shapiro-Wilk test for Multivariate Normality by
##  Villasenor-Alva and Gonzalez-Estrada
##
## data:  as.matrix(subset(dataPISA, CNT == "TUR")[, 3:5])
## MVW = 0.98794, p-value = 0.0002384
```

Mardia's Tests for MVN

```
MVN::mvn(dataPISA[,-2],subset = "CNT",
        mvn_test = "mardia")$multivariate_normality
```

```
##    Group            Test Statistic p.value          MVN
## 1    CHE Mardia Skewness    17.372   0.067     ✓ Normal
## 2    CHE Mardia Kurtosis    -1.277   0.202     ✓ Normal
## 3    CZE Mardia Skewness    10.578   0.391     ✓ Normal
## 4    CZE Mardia Kurtosis    -0.066   0.948     ✓ Normal
## 5    FRA Mardia Skewness     7.617   0.666     ✓ Normal
## 6    FRA Mardia Kurtosis    -2.379   0.017 ✗ Not normal
## 7    ITA Mardia Skewness    15.161   0.126     ✓ Normal
## 8    ITA Mardia Kurtosis    -0.587   0.557     ✓ Normal
## 9    TUR Mardia Skewness    26.116   0.004 ✗ Not normal
## 10   TUR Mardia Kurtosis     0.193   0.847     ✓ Normal
```

PERMANOVA

```
str(dataPISA)
```

```
## 'data.frame':    1500 obs. of  5 variables:
##  $ CNT        : Factor w/ 5 levels "CHE","CZE","FRA",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ gender     : Factor w/ 2 levels "Female","Male": 2 1 1 1 1 2 2 2 2 2 ...
##  $ mathematics: num  373 447 429 405 388 ...
##  $ reading    : num  398 449 482 417 373 ...
##  $ science    : num  413 464 464 349 399 ...
```

```
library(vegan)
permanova.test2<-adonis2(dataPISA[,3:5]~CNT,
                      data=dataPISA,
                      method = "euclidean")
```

```
permanova.test2
```

```
## Permutation test for adonis under reduced model
## Permutation: free
## Number of permutations: 999
##
## adonis2(formula = dataPISA[, 3:5] ~ CNT, data = dataPISA, method = "euclidean")
##            Df SumOfSqs      R2      F Pr(>F)
## Model       4  1354406 0.03518 13.628  0.001 ***
## Residual 1495 37143855 0.96482
## Total    1499 38498261 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

References

Applied Multivariate Statistical Analysis. 2007. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72244-1.

Bakker, Jonathan D. 2024. *Applied Multivariate Statistics in r*. University of Washington.

Bulut, Hasan. 2019a. "An r Package for Multivariate Hypothesis Tests: MVTests" 14. https://doi.org/10.12739/NWSA.2019.14.4.2A0175.

———. 2019b. "AN r PACKAGE FOR MULTIVARIATE HYPOTHESIS TESTS: MVTESTS." *NWSA Academic Journals* 14 (4): 132–38. https://doi.org/10.12739/nwsa.2019.14.4.2a0175.

Korkmaz, Selcuk, Dincer Goksuluk, and Gokmen Zararsiz. 2014. "MVN: An R Package for Assessing Multivariate Normality." *The R Journal* 6 (2): 151. https://doi.org/10.32614/rj-2014-031.

MARDIA, K.V. 1970. "Measures of Multivariate Skewness and Kurtosis with Applications." *Biometrika* 57 (3): 519–30. https://doi.org/10.1093/biomet/57.3.519.

Oksanen, Jari, Gavin L. Simpson, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, et al. 2022. "Vegan: Community Ecology Package." https://CRAN.R-project.org/package=vegan.

Rencher, Alvin C. 2002. "Methods of Multivariate Analysis." *Wiley Series in Probability and Statistics*, February. https://doi.org/10.1002/0471271357.

Team, R Core. 2023. "R: A Language and Environment for Statistical Computing." https://www.R-project.org/.

Vicente-Gonzalez, Laura, and Jose Luis Vicente-Villardon. 2021. "PERMANOVA: Multivariate Analysis of Variance Based on Distances and Permutations." https://CRAN.R-project.org/package=PERMANOVA.

Villasenor Alva, José A., and Elizabeth González Estrada. 2009. "A Generalization of ShapiroWilk's Test for Multivariate Normality." *Communications in Statistics - Theory and Methods* 38 (11): 1870–83. https://doi.org/10.1080/03610920802474465.