

LAB 02 STATISTICAL MODELS

The MATLAB code necessary for extracting all the conclusions presented below can be found within a code script included in the same zip file as this document. The graphs provided are screen captures of the figures that result from executing our code, which you can verify by running it yourself.

PROBLEM 1

- File SM_GaitAnalysis_4.xlsx contains data from patients with different levels of knee problems. We wish to investigate whether, on average, time and velocity measurements obtained by analyzing the way they walk (gait) differ depending on a certain splitting of these patients in four groups (A, B, C and D):

a) State the null and alternative hypotheses for this problem.

Having $g=4$ groups in this problem. And them being of the form:

Population 1: $\mathbf{X}_{11}, \mathbf{X}_{12}, \dots, \mathbf{X}_{1n_1}$

Population 2: $\mathbf{X}_{21}, \mathbf{X}_{22}, \dots, \mathbf{X}_{2n_2}$

\vdots

Population g : $\mathbf{X}_{g1}, \mathbf{X}_{g2}, \dots, \mathbf{X}_{gn_g}$

We can decompose each sample the following way:

$$\mathbf{X}_{\ell j} = \boldsymbol{\mu} + \boldsymbol{\tau}_{\ell} + \mathbf{e}_{\ell j}, \quad j = 1, 2, \dots, n_{\ell} \quad \text{and} \quad \ell = 1, 2, \dots, g$$

where we are saying that each sample from each population group g (in this case we have $l=4$) and dimension n_l (in this problem we have $j=3$) is equal to: an overall mean μ + a TI that represents the l th treatment effect + an e_{lj} that are independent $N(0, E)$ variables (the error for each sample).

After having presented our data, now we can proceed to state our null hypothesis and alternative hypothesis. That will be:

$$H_0 : \vec{\tau}_{\ell} = \vec{0} \quad \forall \ell \in [1, \dots, g]$$
$$H_A : \exists \ell' \text{ so that } \vec{\tau}_{\ell'} \neq \vec{0}$$

The null hypothesis represents that all groups in our data have no treatment effect. That is, that all treatment effects are equal between populations and 0. While the alternative hypothesis is that there is at least one group that differs from 0 and the other groups treatment effects.

b) Indicate what method would you use to answer the research questions of this Problem.

The method we would use is a MANOVA test. We want to test within 4 groups, 3 different features and their overall at the end. The features we will be testing are: the average swing time, the average time of both feet on the ground and the average velocity. These features were obtained by analyzing the walk of each sample. We assume all samples are identically distributed and independent in each population and samples from different groups are also independent between them.

c) If your method requires uni- or multivariate normality in any way, assess whether the data fulfills such assumption(s), at significance level $\alpha = 0.05$. Make sure to state your conclusion.

To perform a MANOVA test it is important that each population from the data follows multivariate normality. Since we have a small n , we need to assess whether the data appear to violate the assumption that it comes from a normal distribution or not. To do so, we will perform the following tests.

1) QQ-plots

a) GROUP A

The QQ plot for the first feature (time of swing) in Group A is the following:

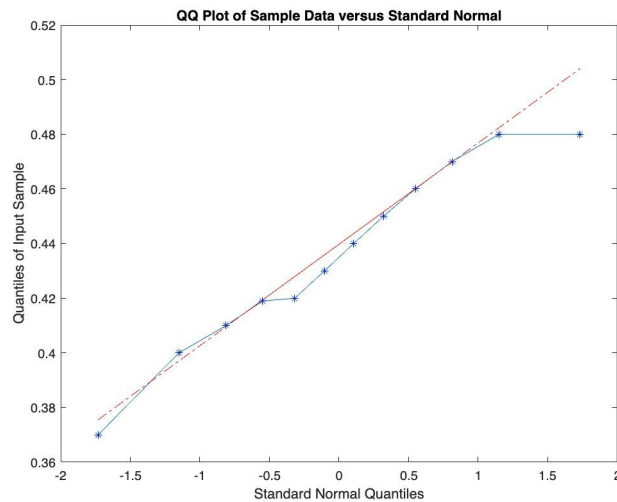


Figure 1.1: QQ-plot for Group A feature 1 (tSWING)

By looking at Figure 1.1, we can see that the relationship between the two variables is pretty straight and follows a slope very near 1. To check if we are right when saying that we don't have enough evidence to reject that it follows a normal distribution we need to look at the correlation coefficient. By applying the function **corr(X, Y)** being X the input samples of the time of swing and Y the standard normal distribution quantiles, we get that $R = 0.98333$. By looking at the table (at the end of the document) Figure 1 and knowing that $n=12$ and significance level is 0.05 we do the following comparison:

$$0.9833 > 0.9389$$

We went with $n=15$ for comparison because it's a tougher test than $n=10$. The gap between the critical value for $n=15$ and our calculated R is quite large (if we used $n=12$, we would still reach the same conclusion). Therefore, we don't have enough evidence to reject that the distribution of the time of swing data on Group A follows a normal distribution.

The QQ plot for the second feature (average time of both feet on the ground) in Group A is the following:

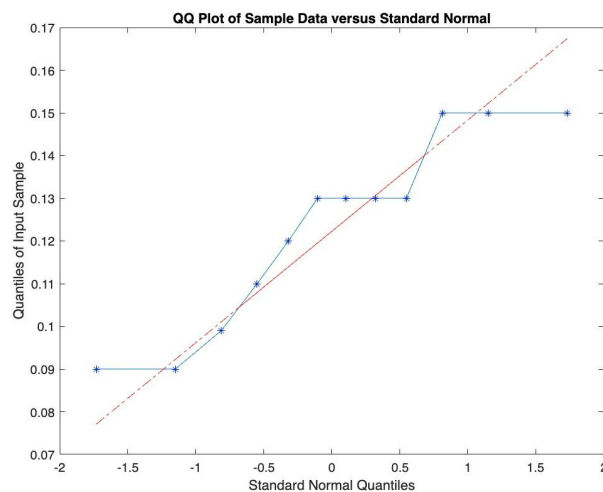


Figure 1.2: QQ-plot for Group A feature 2 (tDBLSTANCE)

In this figure we can't see as clearly as before if it follows a normal distribution or not. Again, we will check Figure 1 and compare the correlation coefficient between the two distributions that construct the previous plot. Computing the $n=12.5$ critical value as the average of $n=15$ and $n=10$ critical values at significance level 0.05, we get 0.92935. Also, by computing the correlation coefficient of this graph using the same function as before we get $R = 0.9542$. Therefore,

$$0.9542 > 0.92935$$

We could have also applied the same method as before of comparing it to $n=15$, but we wanted to be more accurate since the value had less difference. From the comparison above, we can, again, say that we don't have enough evidence to reject that it follows a normal distribution.

The QQ plot for the third feature (average velocity) in Group A is the following:

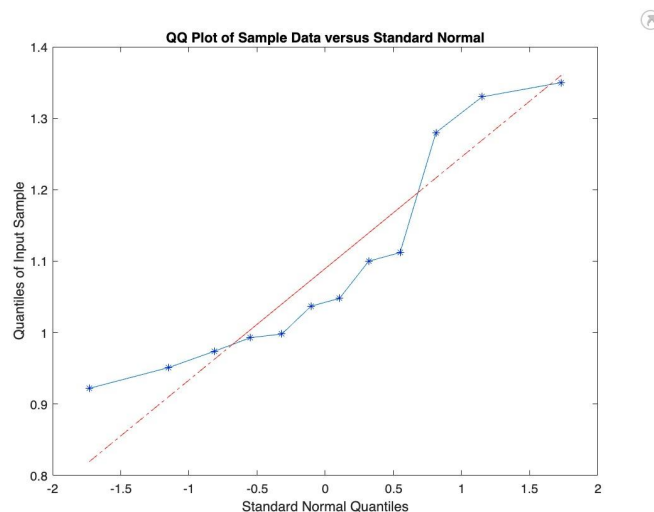


Figure 1.3: QQ-plot for Group A feature 3 (Velocity)

Doing the same procedure as before and using the function in MATLAB we compute the correlation coefficient between the two distributions obtaining $R=0.9371$. Since this value is also lower than the one used before, we will also use the critical value of $n=12.5$ to do the comparison, getting the following:

$$0.9371 > 0.92935$$

Since we know that the critical value when $n=12.5$ is higher than when $n=12$ (because as n grows more clearly needs to be seen that is a normal distribution) we can reach the same conclusion as with the other two features, that we don't have enough evidence to reject that it follows a normal distribution.

All features in group A don't reject the normal distribution assumption at significance level 0.05. For the next groups, we will keep the explanation more simple, since we think that the Group A explanation has left the steps we follow clearer enough.

b) GROUP B

The QQ plot for the first feature (time of swing) in Group B is the following:

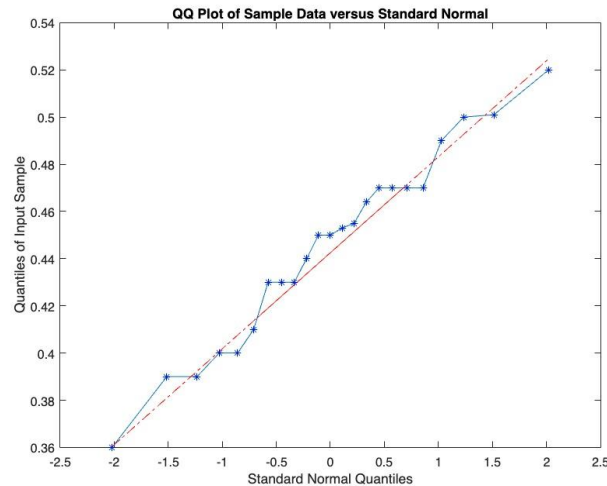


Figure 1.4: QQ-plot for Group B feature 1 (tSWING)

We have computed $R = 0.9896$ and we will be using $n=25$ critical value with significance level 0.05. Therefore, comparing them, we get:

$$0.9896 > 0.9591$$

From this comparison we are able to say that we don't have enough evidence to reject that it doesn't follow a normal distribution.

The QQ plot for the second feature (average time of both feet on the ground) in Group B is the following:

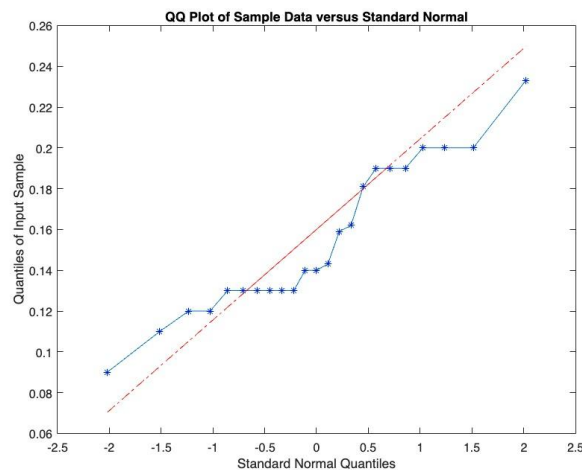


Figure 1.5: QQ-plot for Group B feature 2 (tDBLSTANCE)

By computing the correlation coefficient we get $R=0.9646$. Using the same critical value as before for the same reasons we have stated in all the exercise we know that:

$$0.9646 > 0.9591$$

From this comparison we are able to say, again, that we don't have enough evidence to reject that it doesn't follow a normal distribution.

The QQ plot for the third feature (average velocity) in Group B is the following:

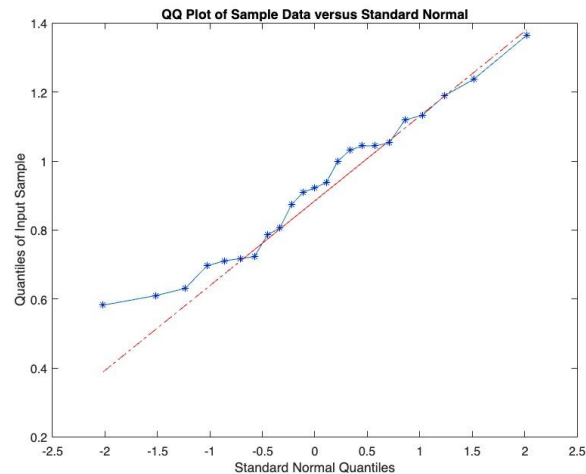


Figure 1.6: QQ-plot for Group B feature 3 (Velocity)

By computing the correlation coefficient we get $R=0.9872$. Using the same critical value as before for the same reasons we have stated in all the exercises before we know that:

$$0.9872 > 0.9591$$

From this comparison we are able to say, again, that we don't have enough evidence to reject that it doesn't follow a normal distribution.

All features in group B don't have enough evidence to reject the normal distribution assumption at significance level 0.05.

c) GROUP C

The QQ plot for the first feature (time of swing) in Group C is the following:

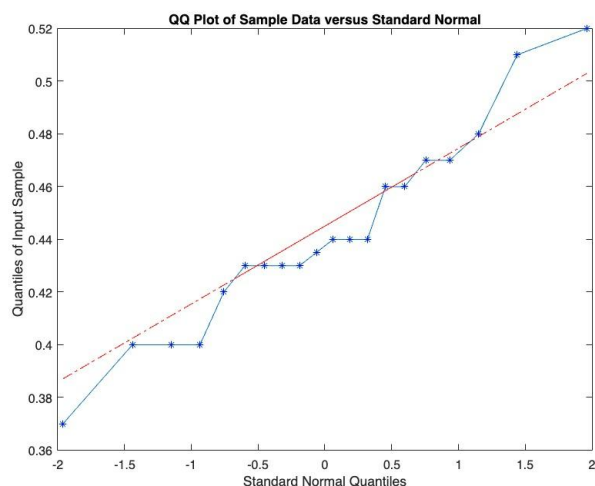


Figure 1.7: QQ-plot for Group C feature 1 (tSWING)

The QQ plot for the second feature (average time of both feet on the ground) in Group C is the following:

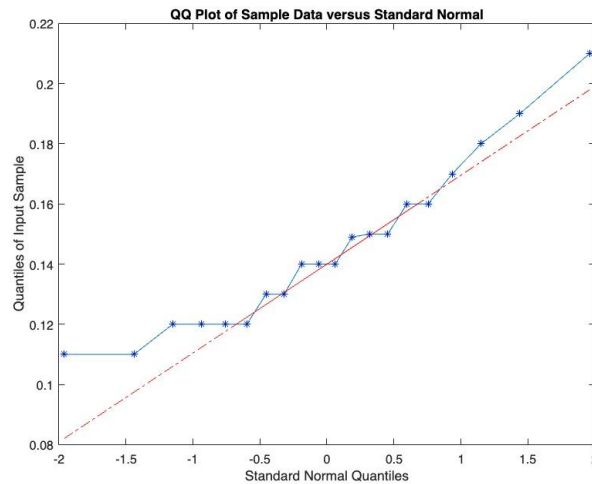


Figure 1.8: QQ-plot for Group C feature 2 (tDBLSTANCE)

The QQ plot for the third feature (average velocity) in Group C is the following:

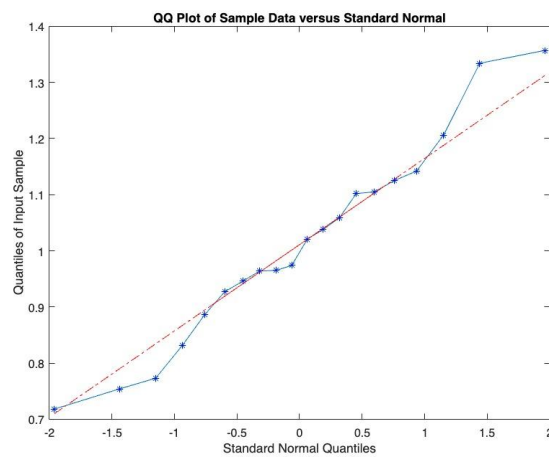


Figure 1.9: QQ-plot for Group C feature 3 (Velocity)

We have computed the values R for each feature of the Group C distribution:

For the first feature $R = 0.9805$. For the second feature $R = 0.9687$ and for the third feature $R = 0.9892$. Comparing them to the critical value of $n=20$ and significance level 0.05 (since the number of samples in Group C is 20) we get the following comparisons:

$$\begin{aligned} 0.9805 &> 0.9508 \\ 0.9687 &> 0.9508 \\ 0.9892 &> 0.9508 \end{aligned}$$

All features in group C don't have enough evidence to reject the normal distribution assumption at significance level 0.05.

d) GROUP D

The QQ plot for the first feature (time of swing) in Group D is the following:

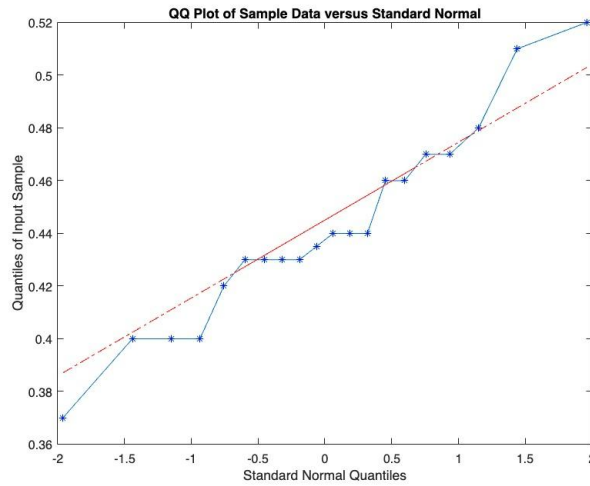


Figure 1.10: QQ-plot for Group D feature 1 (tSWING)

The QQ plot for the second feature (average time of both feet on the ground) in Group D is the following:

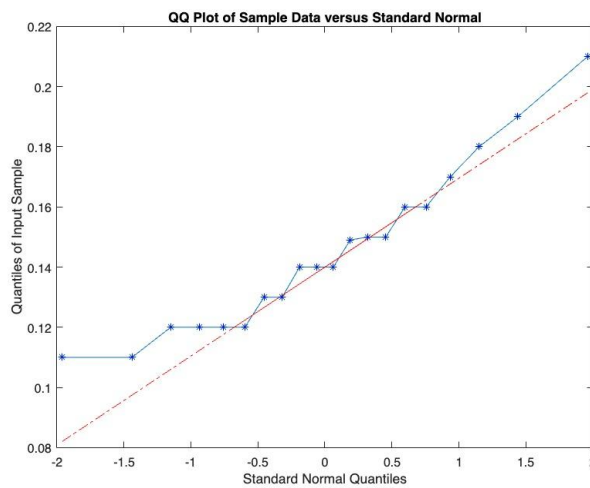


Figure 1.11: QQ-plot for Group D feature 2 (tDBLSTANCE)

The QQ plot for the third feature (average velocity) in Group D is the following:

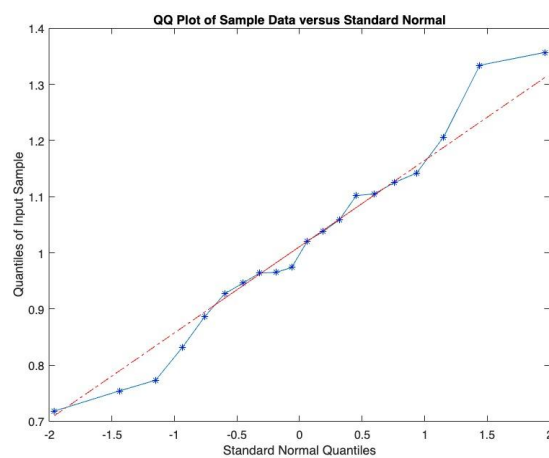


Figure 1.12: QQ-plot for Group D feature 3 (Velocity)

We have computed the values R for each feature of the Group D distribution:

For the first feature $R = 0.9743$. For the second feature $R = 0.9100$ and for the third feature $R = 0.9757$. Comparing them to the critical value of $n=26$ and significance level 0.05 (since the number of samples in Group D is 26) we get the following comparisons:

$$\begin{aligned} 0.9743 &> 0.96032 \\ 0.9100 &< 0.96032 \\ 0.9757 &> 0.96032 \end{aligned}$$

Computation of $n=26$ critical value at 0.05 significance level.

$$\begin{aligned} \alpha=0.05, n=25 &\rightarrow 0.9591 \\ \alpha=0.05, n=30 &\rightarrow 0.9652 \\ \alpha=0.05, n=26 &\rightarrow 0.9591 + (0.9652-0.9591)/5 = 0.96032 \end{aligned}$$

In Group D we have not enough evidence to reject that it can be a normal distribution in features 1 and 3 (that is, time of swing (tSWING) and Velocity). However, in feature 2 (tDBLSTANCE) we find a correlation coefficient lower than the critical value. Therefore, we have enough evidence to reject that it is a normal distribution at significance level 0.05.

Since we have found a feature in a Group that rejects a normal distribution, we will not proceed to further tests to check normality because we have rejected it already.

d) If your method requires any assumptions regarding variances or covariances, assess whether the data fulfills such assumption(s). Make sure to state your conclusion.

To perform a MANOVA test, we need also to assume that the covariance matrices between groups are the same. Therefore, we will need to compute each covariance matrix for each group and compare them.

%with matlab function
gA = cov(groupA)

gA = 3x3
0.0012 -0.0001 -0.0002
-0.0001 0.0005 -0.0027
-0.0002 -0.0027 0.0223

gC = cov(groupC)

gC = 3x3
0.0014 0.0003 -0.0031
0.0003 0.0007 -0.0040
-0.0031 -0.0040 0.0308

gB = cov(groupB)

gB = 3x3
0.0016 0.0007 -0.0041
0.0007 0.0013 -0.0071
-0.0041 -0.0071 0.0465

gD = cov(groupD)

gD = 3x3
0.0012 0.0010 -0.0023
0.0010 0.0028 -0.0069
-0.0023 -0.0069 0.0228

As we can see above, the covariance matrices are very similar between them. Therefore, there is no sufficient evidence to reject that the covariances are different. By using the following property to know if they are similar and comparing all sample variances of each of the populations:

$$1/4 \leq S_{1ii}^2/S_{2ii}^2 \leq 4$$

We get that we can assume that the covariance matrices between populations are the same.

e) Based on the results from (c) and (d), discuss whether the method selected in (a) can be applied or not. In case it cannot be applied because your data does not fulfill the required assumptions, try to fix it by choosing one of

the following options (you are free to choose any of the above solutions, but the choice must be justified and you must justify also that, after applying the selected solution, your data now fulfills the required assumptions):

By the conclusions extracted from the previous studies, we can't perform the method selected in (a) since all the assumptions needed to be able to do it are not satisfied. As we have stated, there is one QQ-plot that we could reject normal distribution.

- Apply a transformation (e.g. see Section 4.8 from Johnson-Wichern's book).
- Discard the patient group(s) that do not fulfill the assumptions.
- Identify and eliminate outliers.

After trying the first and third method and checking again the QQ-plot, we have reached the conclusion that **the best method for this dataset is to discard the patient group that does not fulfill the assumptions. That is, discard group D.**

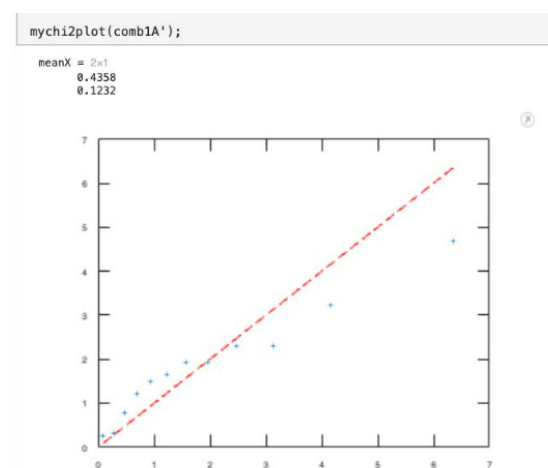
As you can see in the code we provided in the MATLAB file, the other methods that are not the one we have chosen still didn't pass the new test and in both cases we ended up having enough evidence to reject that it was a normal distribution. Although having a significantly higher value of correlation coefficient, it wasn't enough to get to the desired conclusion to be able to perform further tests.

Once the patient group that didn't let the test be performed is discarded, we can follow with the further test to assess normality in all individual populations. From the tests made in c) we know that all populations have univariate normal distribution. Now we need to check bivariate normal distribution and, finally, multivariate normality for the whole dataset.

To do so, we need to perform Chi-squared plots. In particular we need to perform 4 chi-square plots for each population.

1) GROUP A CHI-SQUARE PLOTS

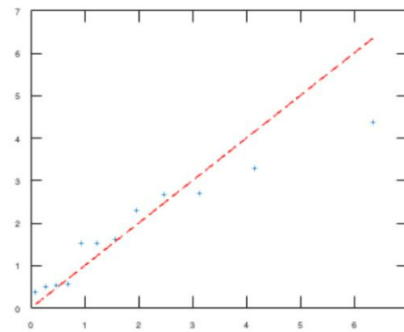
For features 1 and 2:



For features 2 and 3:

```
mychi2plot(comb3A');
```

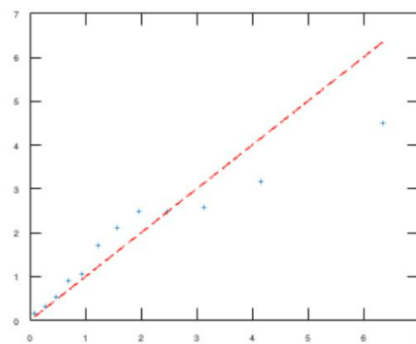
```
meanX = 2x1  
0.1232  
1.0912
```



For features 3 and 1:

```
mychi2plot(comb2A');
```

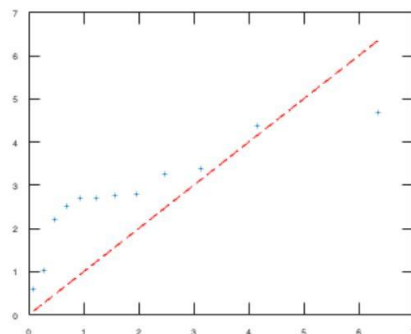
```
meanX = 2x1  
0.4358  
1.0912
```



The three features in Group A:

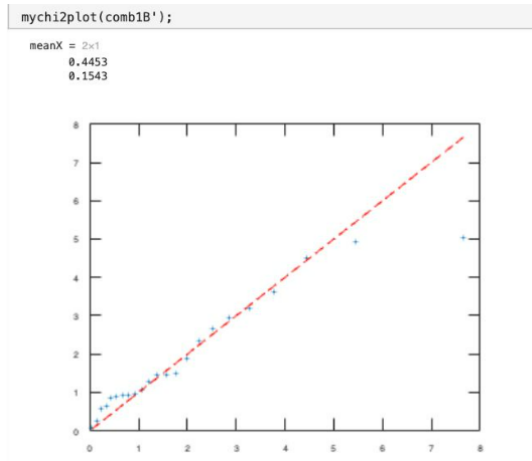
```
mychi2plot(groupA');
```

```
meanX = 3x1  
0.4358  
0.1232  
1.0912
```

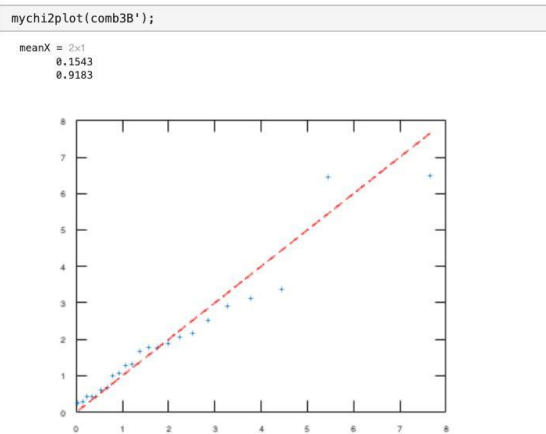


2) GROUP B CHI-SQUARE PLOTS

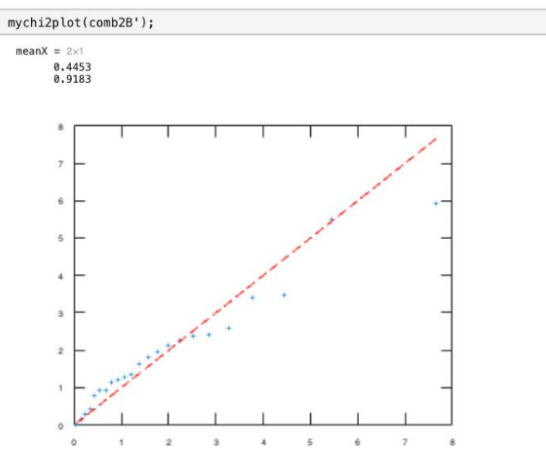
For features 1 and 2:



For features 2 and 3:



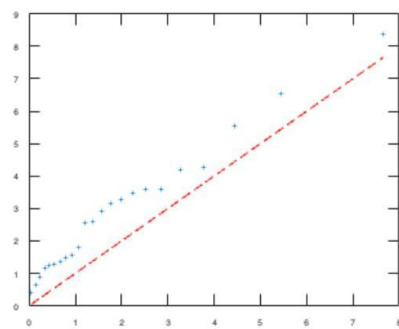
For features 3 and 1:



The three features in Group B:

```
mychi2plot(groupB');
```

```
meanX = 3x1  
0.4453  
0.1543  
0.9183
```

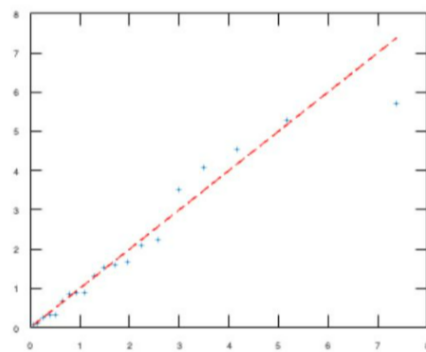


3) GROUP C CHI-SQUARE PLOTS

For features 1 and 2:

```
mychi2plot(comb1C');
```

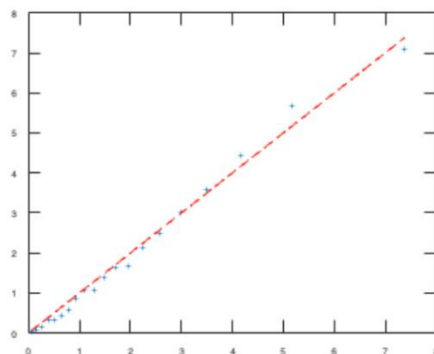
```
meanX = 2x1  
0.4418  
0.1450
```



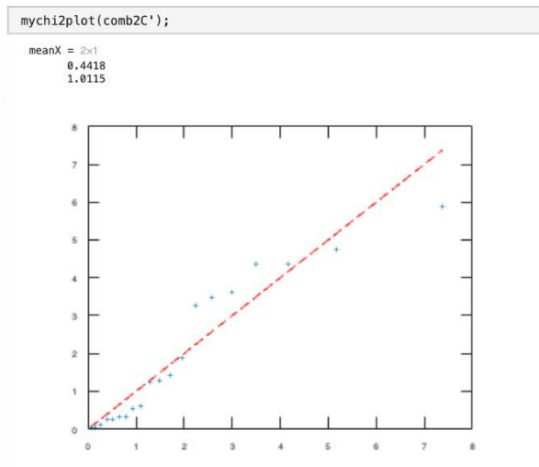
For features 2 and 3:

```
mychi2plot(comb3C');
```

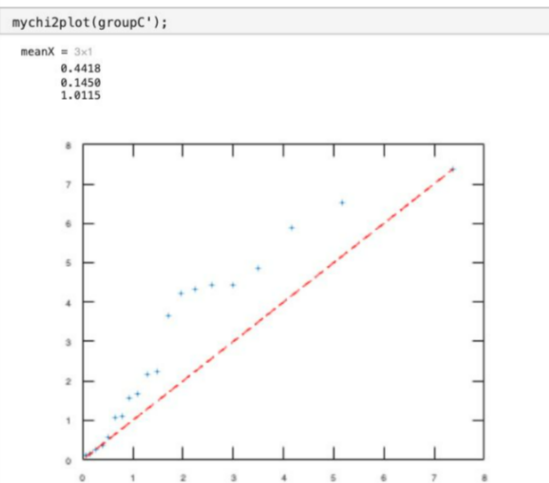
```
meanX = 2x1  
0.1450  
1.0115
```



For features 3 and 1:



The three features in Group C:



As we can see in the plots above, we don't have enough evidence to reject that they form a normal distribution. Once all Chi-squared plots are computed and certain, we can follow to the next steps on testing.

f) Apply the method selected in (a) to answer the research question, at significance level $\alpha = 0.05$. If you performed any fixing in (e), then use the fixed data. Indicate your conclusion.

```
% f) Apply the method selected in (a) to answer the research question, at significance level  $\alpha = 0.05$ . if you performed any fixing in (e),
n = length(datawithoutD(:,1));
g = length(datawithoutD(i,:));
N = length(groupA) + length(groupB) + length(groupC)

N = 55
```

```
p=3;
means = zeros(3, 3);

for i=1:3
    means(i, 1) = mean(groupA(:, i));
end

for i=1:3
    means(i,2) = mean(groupB(:,i));
end

for i=1:3
    means(i,3) = mean(groupC(:,i));
end

for i=1:3
    mymeanov(i) = mean(datawithoutD(:,i));
end

diffTotal = group_means - overall_means;
mySSB = length(groupA) * (diffTotal * diffTotal') + length(groupB)*(diffTotal * diffTotal') + length(groupC)*(diffTotal * diffTotal');
mySSW = zeros(3,3);

for i = 1:g
    for j = 1:12
        mygData = datawithoutD(1:j, :);
        mygData = (mygData - means(:,i))*(mygData - means(:,i))';
        mySSW = mySSW + mygData;
    end
end
```

```
end
```

```
mylambda = det(mySSW)/det(mySSB + mySSW)
```

```
mylambda = 0.8203
```

```
DFSSb = g - 1;
DFSSw = N - g;
s = min(p,DFSSw);

DF1 = p*DFSSb;
DF2 = s*(DFSSw-((p-DFSSb+1)/2))-((p*DFSSb-2)/2);
```

```
% using Wilks' Lambda
```

```
F = ((1-sqrt(mylambda))/DF_1)/(sqrt(mylambda)/DF_2)
```

```
F = 1.7352
```

```
% Calculate the critical F-value
```

```
Fcritic = finv(0.95, DF_1, DF_2)
```

```
Fcritic = 2.1906
```

By looking at the photos above, we can conclude, by seeing the observed statistic and the critical value of the F distribution that:

$$1.7352 < 2.1906$$

Therefore, we don't have enough evidence to reject the null hypothesis.

PROBLEM 2

- We wish to assess the following research question: does the (average) difficulty of the labs in Statistical Models vary along the course (at significance level $\alpha = 0.01$)?. To test that, we have collected the marks obtained by students in the 5 labs of the subject during the last year as our sample data. File SM22_LabGrades.xlsx contains this data. A few remarks:

- Each row corresponds to a student that took the subject in course 22-23.
- Each column corresponds to one of the 5 labs of the subject (note that even though there were 6 labs, only 5 had deliverables and are the ones provided in the table).
- The marks of the labs are considered as a measurement of achievement of a given student. Therefore, you are expected to consider the lab marks as repeated measurements of achievement for each student. More difficult labs are expected to imply lower marks. Once the above is clear, you are asked to answer the following:

a) State the null and alternative hypotheses for the research question indicated at the very beginning of the problem.

In order to test whether the (average) difficulty of the labs in Statistical Models varies along the course, we can state the null and alternative hypotheses as follows (*where $\mu_1, \mu_2, \mu_3, \mu_4$, and μ_5 represent the population means of the lab marks*) :

Null Hypothesis (H_0): There is no significant variation in the average difficulty of the labs in Statistical Models along the course. In mathematical terms: **$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$**

Alternative Hypothesis (H_a): There is a significant variation in the average difficulty of the labs in Statistical Models along the course. In mathematical terms: **H_a : At least one of the population means ($\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$) is different from the others.**

b) Propose a model to test the hypotheses stated in (a). State what assumptions are required in order to apply your model correctly.

In this problem, since we have the same subjects for the 5 population means we will be comparing, we will be using the Repeated Measures method. The three assumptions for a single-group univariate repeated-measures analysis are:

1. Independence of the observation
2. Multivariate normality
3. Sphericity in the data

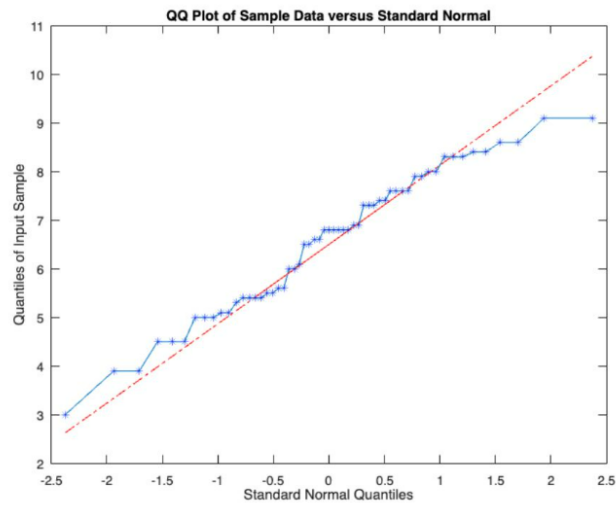
If all these assumptions can be made in a certain level of confidence, the Repeated Measures analysis can be performed over a specific dataset.

c) Check whether your data fulfills the assumptions that you stated in (b) (use significance level $\alpha = 0.01$ in your assessments). If your data does not hold the required assumptions: 1. Check if there is a different method that could be used to solve this problem for which all the required assumptions hold. In such a case, repeat (b) and (c) for that method. 2. If the above also fails, try fixing the data with one of the suggestions from Problem 1(e). You can choose any of the options but you must retain at least 3 of the considered groups.

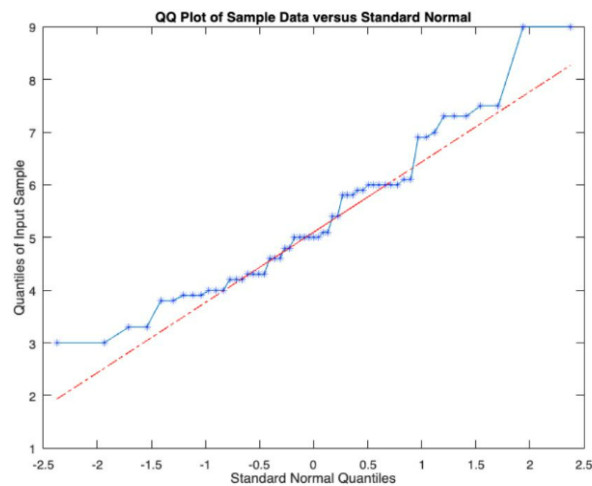
We know that the first assumption is achieved because we are working with the following dataset: Represents different students (each row corresponds to a different student) and there's no inherent reason to expect that the lab marks of one student are correlated with the lab marks of another student.

To check Multivariate normality, we can check that each population (each Lab) of the database provided follows a univariate normal distribution by checking the QQ-plots and its correlation value.

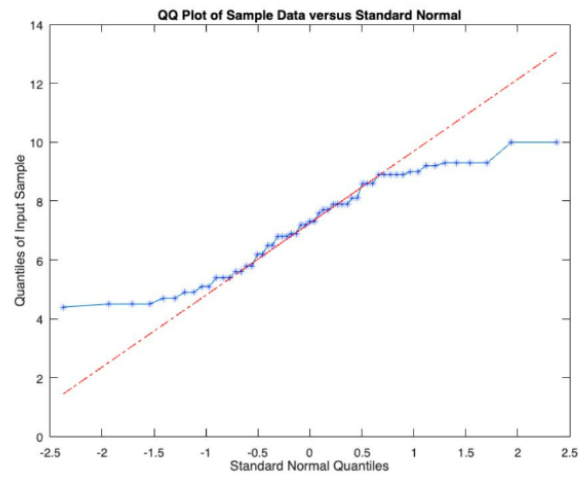
QQ-plot Lab1



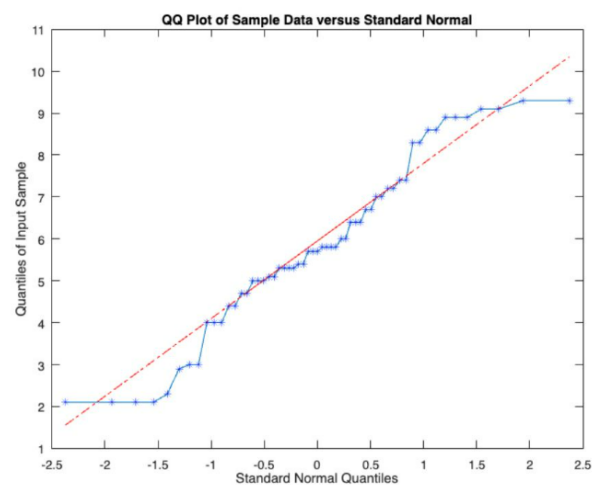
QQ-plot Lab2



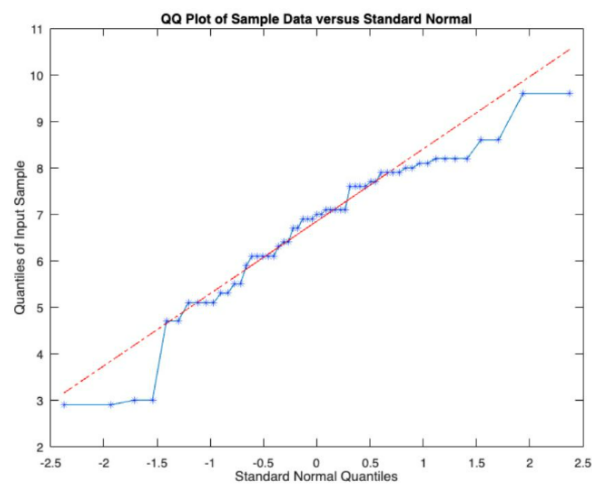
QQ-plot Lab3



QQ-plot Lab4



QQ-plot Lab5



Computing the correlation coefficients we have:

R for Lab1 = 0.9766
R for Lab2 = 0.9883
R for Lab3 = 0.9730
R for Lab4 = 0.9822
R for Lab5 = 0.9713

Knowing $n=57$ and looking at Figure 1, we can do the following comparisons and reach some conclusions:

$0.9766 > 0.9705$
 $0.9883 > 0.9705$
 $0.9730 > 0.9705$
 $0.9822 > 0.9705$
 $0.9713 > 0.9705$

Computing critical value for $n=57$ at significance level 0.01

$\alpha=0.01$ and $n=55 \rightarrow 0.9695$

$\alpha=0.01$ and $n=60 \rightarrow 0.9720$

$\alpha=0.01$ and $n=81 \rightarrow ((0.9720-0.9695)/5)*(2) + 0.9695 = 0.9705$

Therefore, **the conclusion we reach is that we don't have enough evidence to reject that we find ourselves in front of a normally distributed dataset.** The second assumption is well made.

The last assumption we need to check is whether there is sphericity in the data. To do so:

```
function covtesting(variance1, variance2)
    div = variance1/variance2;
    if div < 1/4 || div > 4
        disp("diff variances");
    else
        disp("same variances");
    end
end
```

```
%check sphericity
variancelab1 = var(data(:,1));
variancelab2 = var(data(:,2));
variancelab4 = var(data(:,3));
variancelab5 = var(data(:,4));
variancelab6 = var(data(:,5));

covtesting(variancelab1, variancelab2);
covtesting(variancelab1, variancelab4);
covtesting(variancelab1, variancelab5);
covtesting(variancelab1, variancelab6);
covtesting(variancelab2, variancelab4);
covtesting(variancelab2, variancelab5);
covtesting(variancelab2, variancelab6);
covtesting(variancelab4, variancelab5);
covtesting(variancelab4, variancelab6);
covtesting(variancelab5, variancelab6);
```

```
variancelab1 = 0.0001;
correlation_coef
```

```
same variances
same variances
same variances
same variances
same variances
same variances
same variances
same variances
same variances
same variances
```

Therefore, **we can conclude there is sphericity in our dataset.**
All assumptions have been checked.

d) Test the hypotheses from (a) and indicate what can you conclude about the research question, i.e. are there labs significantly more difficult than others? If so, which ones?

```
%D test hypothesis
meanlab1 = mean(data(:,1))
meanlab2 = mean(data(:,2))
meanlab4 = mean(data(:,3))
meanlab5 = mean(data(:,4))
meanlab6 = mean(data(:,5))
means = [meanlab1, meanlab2, meanlab4, meanlab5, meanlab6];

overallmean = mean(alldata)

rowmeans = [];
for i=1:n
    rowmeans(i) = mean(data(i,:));
end

SSw = mySSw(data, means, n, g)

SSsubj = mySSsubj(data, rowmeans, overallmean, n, g)

SSres = SSw - SSsubj

SSb = mySSb(data, means, overallmean, n, g)

mystat = (SSb/(g-1))/(SSres/((n-1)*(g-1)))

fcrit = finv(0.99, g-1, n-g+1)
```

```

g = 5
meanlab1 = 5.2842
meanlab2 = 6.5526
meanlab4 = 7.1860
meanlab5 = 5.8316
meanlab6 = 6.6860

overallmean = 6.3081

SSw = 745.0586

SSsubj = 260.8074

SSres = 484.2512

SSb = 128.1728

mystat = 14.8222

fcrit = 3.6954

```

The code we have provided above computes the statistics we need from the samples and then compares it to the critical value of the distribution it follows (an F distribution with $g-1$ and $n-g+1$ degrees of freedom and significance level 0.01). Computing both values, we get:

$$\text{mystat} = 14.822222 > 3.69 = \text{Fcrit}$$

Looking at the numbers, we can see that the observed statistic is higher than the critical value. Therefore, we have enough evidence to reject the null hypothesis.

We need to do post hoc to know which pairs are the ones that are different.

```
%postHoc
MSw= SSw/(g*(n-1));
postHoc(meanlab1, meanlab2, length(data(:,1)), length(data(:,2)), MSw, 0.01, g, n);
postHoc(meanlab1, meanlab4, length(data(:,1)), length(data(:,3)), MSw, 0.01, g, n);
postHoc(meanlab1, meanlab5, length(data(:,1)), length(data(:,4)), MSw, 0.01, g, n);
postHoc(meanlab1, meanlab6, length(data(:,1)), length(data(:,5)), MSw, 0.01, g, n);
postHoc(meanlab2, meanlab4, length(data(:,2)), length(data(:,3)), MSw, 0.01, g, n);
postHoc(meanlab2, meanlab5, length(data(:,2)), length(data(:,4)), MSw, 0.01, g, n);
postHoc(meanlab2, meanlab6, length(data(:,2)), length(data(:,5)), MSw, 0.01, g, n);
postHoc(meanlab4, meanlab5, length(data(:,3)), length(data(:,4)), MSw, 0.01, g, n);
postHoc(meanlab4, meanlab6, length(data(:,3)), length(data(:,5)), MSw, 0.01, g, n);
postHoc(meanlab5, meanlab6, length(data(:,4)), length(data(:,5)), MSw, 0.01, g, n);
```

```

dont have enough evidence to reject
dont have enough evidence to reject
dont have enough evidence to reject
dont have enough evidence to reject
dont have enough evidence to reject
dont have enough evidence to reject
dont have enough evidence to reject
dont have enough evidence to reject
dont have enough evidence to reject
dont have enough evidence to reject
dont have enough evidence to reject

```

e) Compute confidence intervals for the difference of the average lab marks (for all pairs of labs). Indicate whether these intervals confirm your findings in (d).

The confidence intervals we have found are the following:

```
%E Compute confidence intervals for the difference of the average lab
[from, to] = myCI(meanlab1, meanlab2, n, g, SSres)
[from, to] = myCI(meanlab1, meanlab4, n, g, SSres)
[from, to] = myCI(meanlab1, meanlab5, n, g, SSres)
[from, to] = myCI(meanlab1, meanlab6, n, g, SSres)

[from, to] = myCI(meanlab2, meanlab4, n, g, SSres)
[from, to] = myCI(meanlab2, meanlab5, n, g, SSres)
[from, to] = myCI(meanlab2, meanlab6, n, g, SSres)

[from, to] = myCI(meanlab4, meanlab5, n, g, SSres)
[from, to] = myCI(meanlab4, meanlab6, n, g, SSres)

[from, to] = myCI(meanlab5, meanlab6, n, g, SSres)

function mvaaanlot(x)
```

```
from = -2.3569
to = 3.0854
from = -2.9902
to = 2.4520
from = -1.6358
to = 3.8064
from = -2.4902
to = 2.9520
from = -1.7218
to = 3.7205
from = -0.3674
to = 5.0748
from = -1.2218
to = 4.2205
from = 0.2659
to = 5.7082
from = -0.5884
to = 4.8538
from = -1.9428
to = 3.4994
```

When 0 is included in the CI means we are likely seeing that there is not a significant difference between our sample and the population parameter. Our confidence interval coincides with the post hoc test we did before.

PROBLEM 3

- Data in file SM23_Lab_2_Fish.mat corresponds to an experiment carried out to compare 3 different methods to cook fish. For that purpose, 36 fish samples were randomly split in 3 groups of 12 samples. Samples in each group were cooked according to one of the methods, which we will call A, B, and C. To compare the 3 methods, judges were asked to taste the cooked fish samples and rate them in terms of 4 variables: Aroma, Flavor, Texture and Moisture.

a) State the null and alternative hypotheses required for the comparison indicated above.

Having $g=3$ groups in this problem. And them being of the form:

Population 1: $\mathbf{X}_{11}, \mathbf{X}_{12}, \dots, \mathbf{X}_{1n_1}$

Population 2: $\mathbf{X}_{21}, \mathbf{X}_{22}, \dots, \mathbf{X}_{2n_2}$

\vdots

Population g : $\mathbf{X}_{g1}, \mathbf{X}_{g2}, \dots, \mathbf{X}_{gn_g}$

Being $n=4$ (since we value 4 features, that is 4 dimensions).

We can decompose each sample the following way:

$$\mathbf{X}_{\ell j} = \boldsymbol{\mu} + \boldsymbol{\tau}_{\ell} + \mathbf{e}_{\ell j}, \quad j = 1, 2, \dots, n_{\ell} \quad \text{and} \quad \ell = 1, 2, \dots, g$$

where we are saying that each sample from each population group g (in this case we have $l=3$) and dimension n_l (in this problem we have $j=4$) is equal to: an overall mean μ + a TI that represents the l th treatment effect + an $e_{\ell j}$ that are independent $N(0, E)$ variables (the error for each sample).

After having presented our data, now we can proceed to state our null hypothesis and alternative hypothesis. That will be:

$$H_0: \vec{\tau}_{\ell} = \vec{0} \quad \forall \ell \in [1, \dots, g]$$
$$H_A: \exists \ell' \text{ so that } \vec{\tau}_{\ell'} \neq \vec{0}$$

The null hypothesis represents that all groups in our data have no treatment effect. That is, that all treatment effects are equal between populations and 0. While the alternative hypothesis is that there is at least one group that differs from 0 and the other groups treatment effects. We are trying to check if the cooking methods differ between each other or not.

b) Indicate what method would you use to test the hypotheses stated in (a). Indicate what assumptions would be needed in order for your method to be valid. For this problem, you do not need to test whether the assumptions hold; we will assume that they do.

We would use MANOVA since we are dealing with 3 different populations with 4 variables. The assumptions for this problem are:

1. Each population (in this case, each group of fish A,B and C) come from a multivariate normal distribution
2. All samples are i.i.d for each group
3. All samples within the same group are independent from each other
4. The variance of all groups is the same

If all of them are satisfied and not rejected, we can perform a test with MANOVA.

c) Test the hypotheses stated in (a) at significance level $\alpha = 0.05$. Provide all the relevant numerical details and clearly state your conclusion (both in terms of the null and in plain language).

By observing the matrices and putting the data onto the matrices manually. We have reached the following test for the hypothesis stated in (a):

```
my_lambda = det(my_SSW)/det(my_SSB + my_SSW)
DFSSb = g - 1;
DFSSw = N - g;
s = min(p,DFSSb);

DF1 = p*DFSSb;
DF2 = s*(DFSSw-((p-DFSSb+1)/2))-((p*DFSSb-2)/2);

% using Wilks' Lambda

F = ((1-sqrt(my_lambda))/(DF1))/(sqrt(my_lambda)/DF2)

% critical F-value
fcrit = finv(0.95, DF1, DF2)
```

my_lambda = 0.4727

F = 128.7533

fcrit = 2.0970

Comparing the values, we get:

$$2.0970 < 128.7533$$

By seeing the values, we have enough evidence to accept the null hypothesis with a lot of confidence. Since there is a lot of difference between the statistic and the critical value.

d) If you find a significant difference in (c), then test what are the methods that differ and check also if it is possible to assess in which dimension. Otherwise (if your results in (c) do not support a significant difference), compute 95% confidence regions for the average rating differences between the 3 cooking methods.

We would need to compare method A-B A-C B-C (multivariate analysis) with F distribution and then do the univariate with t student.

| Table 4.2 Critical Points for the Q-Q Plot Correlation Coefficient Test for Normality | | | |
|---|------------------------------|-------|-------|
| Sample size <i>n</i> | Significance levels α | | |
| | .01 | .05 | .10 |
| 5 | .8299 | .8788 | .9032 |
| 10 | .8801 | .9198 | .9351 |
| 15 | .9126 | .9389 | .9503 |
| 20 | .9269 | .9508 | .9604 |
| 25 | .9410 | .9591 | .9665 |
| 30 | .9479 | .9652 | .9715 |
| 35 | .9538 | .9682 | .9740 |
| 40 | .9599 | .9726 | .9771 |
| 45 | .9632 | .9749 | .9792 |
| 50 | .9671 | .9768 | .9809 |
| 55 | .9695 | .9787 | .9822 |
| 60 | .9720 | .9801 | .9836 |
| 75 | .9771 | .9838 | .9866 |
| 100 | .9822 | .9873 | .9895 |
| 150 | .9879 | .9913 | .9928 |
| 200 | .9905 | .9931 | .9942 |
| 300 | .9935 | .9953 | .9960 |

Figure 1: Critical Point for the Q-Q Plot Correlation Coefficient Test for Normality