**MACHINE LEARNING FROM DATA**

**Fall 2018**

**Report: Lab Session 0 – Exploratory data analysis**

**Names:**

**Group:**

1. **Instructions**

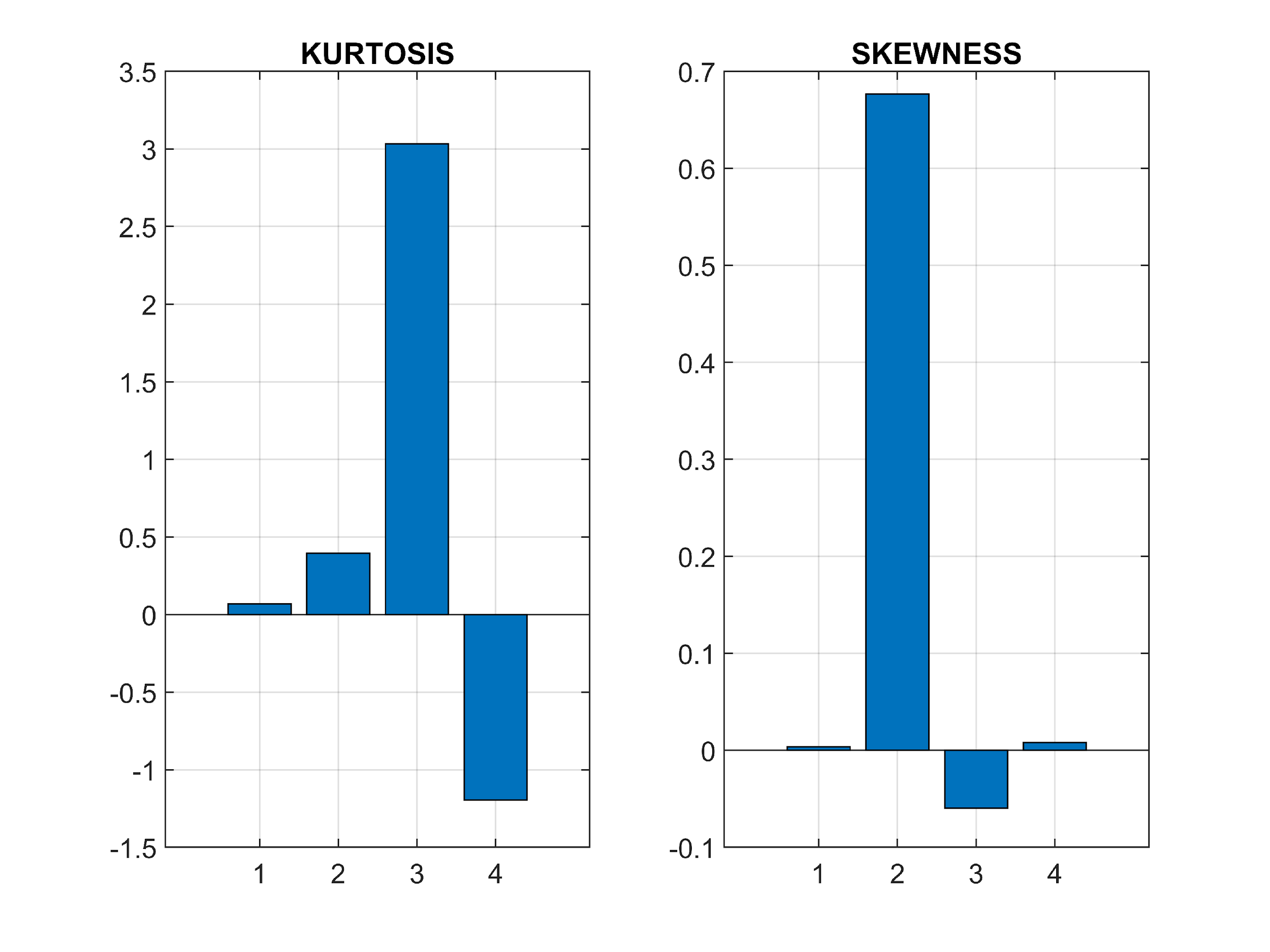
* Answer the questions
* Save the report, convert to pdf and upload the pdf file

1. **Questions**

Q1. Briefly describe the conclusions of your analysis (you can insert plots)

As expected, observing the four different kind of plots, we can assume a Gaussian distribution only for the randomly generated Gaussian distribution.

Command bar(kurtosis(XX')-3) is used => the plot show the difference from 3 of kurtosis => first one is gaussian, second Rayleigh, third Laplacian, fourth uniform



In particular, from the plot above, it is possible to identify with sufficient confidence from which of the four distributions the samples are actually taken. The second one to begin with shows a particular skewness, which leads to believe to have been originated from a Rayleigh distribution as the other three are symmetric around their mean; the third one instead shows a kurtosis quite close to three, which suggest a gaussian behaviour, while the one of the fourth one is quite negative, a trait that between the four is only shown by the uniform; a last we can confidently say that the first one is actually the Laplacian as it is the only one left.

Q2. For each class and each feature, analyse histograms, cdfs and normal plots. Can we assume a Gaussian distribution for any of the features?

Q3. Analyse kurtosis and skewness values for each feature and class.

1. Sepal length:

Kurtosis is inferior to 3 for the three classes but at different values (about 2.65, 2.40 and 2.91 respectively) the last one being suitable to result from a gaussian.

Skeweness of sepal length distribution is slightly positive and similar for all the classes (about 0.12, 0.10 and 0.11 recpectively).

1. Sepal width:

No general consideration can be done about the kurtosis of this feature over the three classes as values vary widely (about 3.69, 2.55 and 3.51 respectively), we can see that the ones of class 1 and 3 are similar and denote more extreme deviations in sepal length for these classes.

For class one skewness related to sepal width (about 0.10) is very similar to the one related to sepal length while we see that class 2 and 3 skewness have both a moderate absolute value but different sign (respectively -0.35 and 0.35).

1. Petal length:

For petal length kurtosis is pretty high for the first class and close to the gaussian reference for classes 2 and 3 (about 3.81, 2.93 and 2.74 respectively).

Skewness is almost zero (about 0.07) for class one, while the same scheme as before repeat for skewness of class 2 and 3 petal length: quite high but opposite values (about -0.59 and 0.53).

1. Petal width:

Kurtosis values (about 4.30, 2.51 and 2.33 respectively) show us that petal length distribution is very outlier-prone for class 1 while classes 2 and 3 show frequent modestly sized deviations.

For petal width skewness is very low for classes 1 and 2 (about -0.03 and -0.13) but very large and positive for class 1 (about 1.16).

Q4. Analyze boxplots by feature. Are there ‘significant’ differences between the classes?

The classes do present significant differences. With the exception of feature 2 where the median values and quartiles of the three classes appear to be similar, it is possible to observe a clear trend in the other three boxplots. In fact class 1 appears to have a median value much lower than that of classes 2 and 3 up to the point that for features 3 ad 4 even the 95% confidence interval results disjoint from that of classes 2 and 3, showing great difference between the classes.

Q5. Analyze the scatter plot. Are features related in any way? What can you say about the separability of the classes?

There appears to be a strong relation between some of the features (e.g. the scatter plot of features 3 and 4 presents a clear tendency along the main diagonal, impling a relation of direct proportionality between the two). As for the classes’ separability we can say that class 1 seems to be easily distinguishable from the other two as in almost every scatter plot it is clearly divided from the others. On the other hand, even if we can still see some differences between the behaviour of class 2 and 3 it appears way harder to separate the two as in almost every plot the points of the two classes appear mixed together. Features 3 and 4 should anyway give the best results in this attempt as a simple linear boundary can be determined in this plane to separate correctly many of the points in the correct classes.???

Q6. Edit the script ML\_Lab0\_irisdataset.m. Choose one feature (among the four available) and compute the feature mean and confidence intervals at confidence levels 95%, 99% and 99.9% for the three classes.

Hint: use Matlab functions tinv and var

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature #3 | Mean | CI at 95% | CI at 99% | CI at 99,9% |
| Class 1 | 1.4640 | [1.4147, 1.5133] | [1.3982, 1.5298] | [1.3781, 1.5499] |
| Class 2 | 4.2600 | [4.1265, 4.3935] | [4.0819, 4.4381] | [4.0274, 4.4926] |
| Class 3 | 5.5520 | [5.3952, 5.7088] | [5.3428, 5.7612] | [5.2788, 5.8252] |

Q7. Copy the code used to answer Q6.

i\_feat = 3

alphas=[0.05, 0.01, 0.001]

for j = 1:length(alphas)

alfa=alphas(j)

for i\_class = 1:N\_class

i\_class

index=find(Labels==i\_class);

df=length(index)-1

M\_mean=mean(X(index,i\_feat))

S\_deviation=sqrt(var(X(index,i\_feat)))

P=1-(alfa/2)

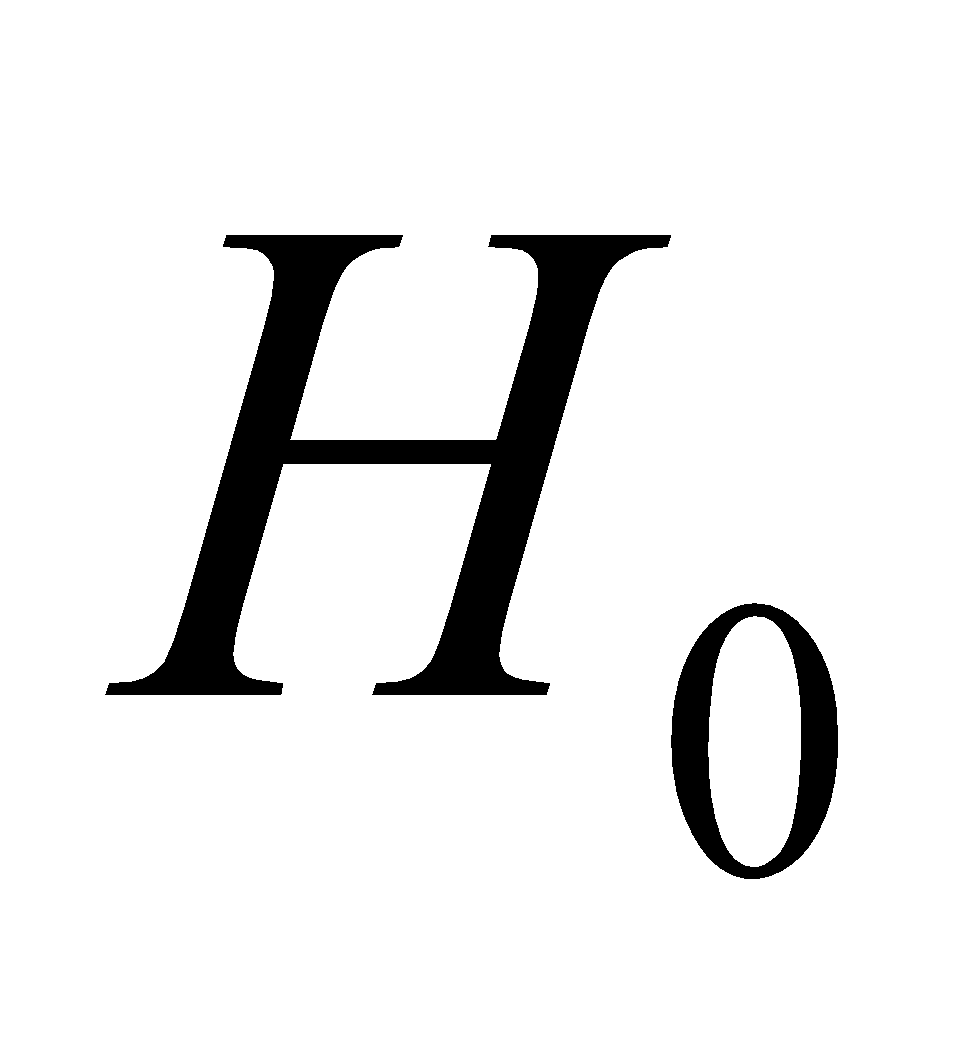
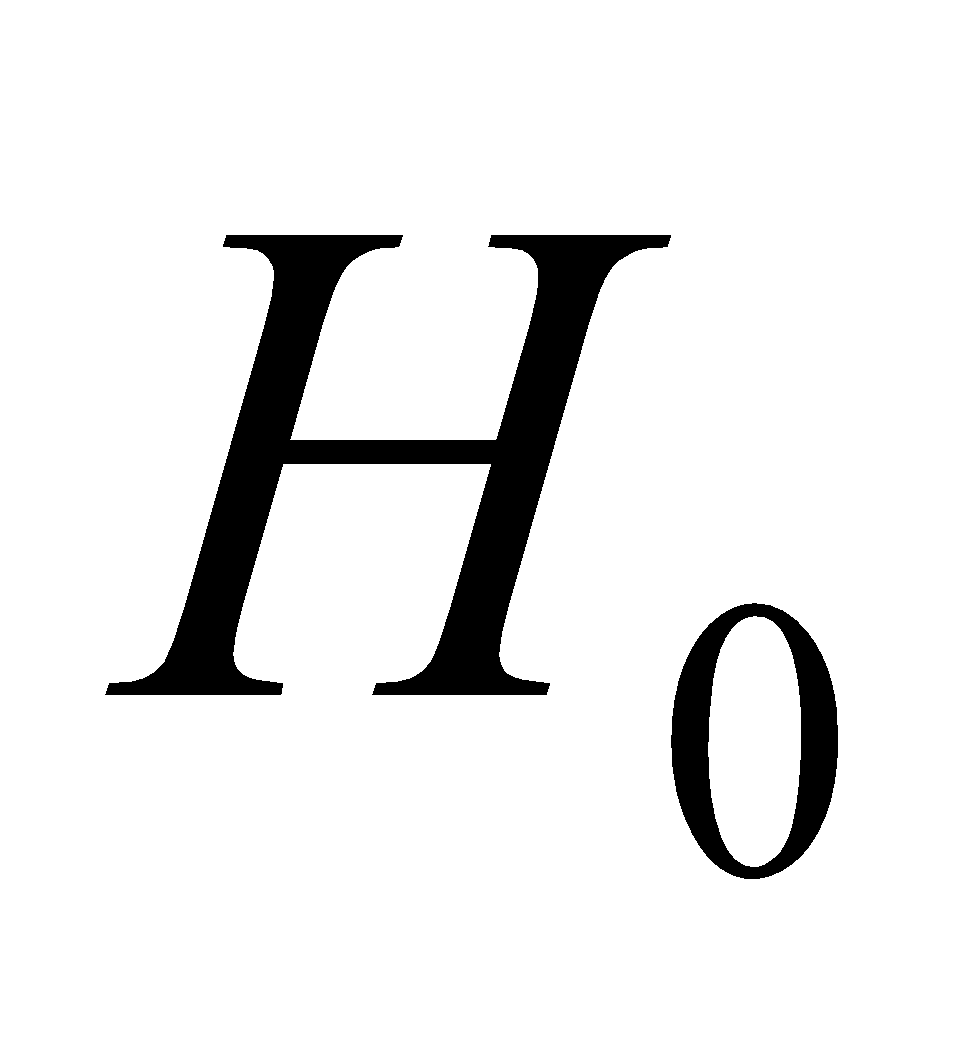
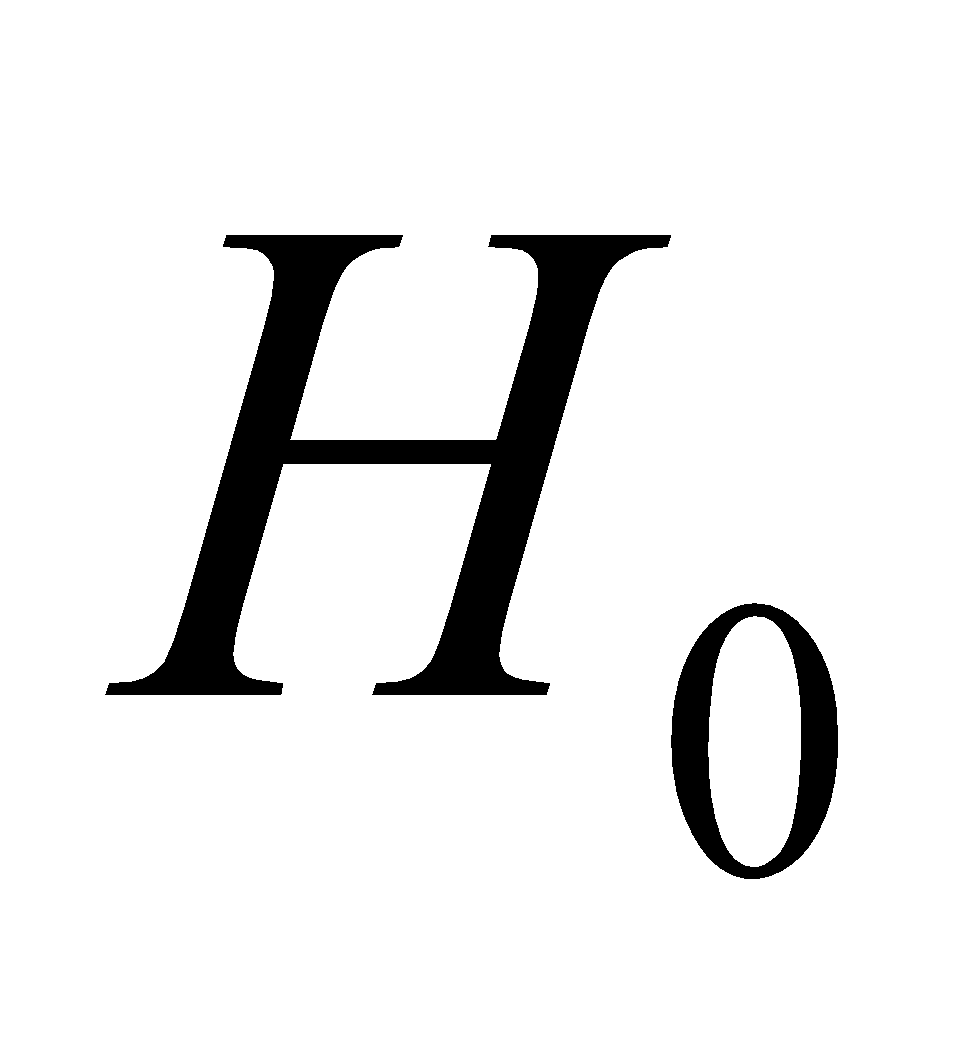
t\_alfa\_2=tinv(P,df)

Confidence\_I=[M\_mean-(t\_alfa\_2\*S\_deviation/sqrt(df+1));M\_mean+(t\_alfa\_2\*S\_deviation/sqrt(df+1))]

end

end

Q8. Choose one feature K (among the four available). Edit the script ML\_Lab0\_irisdataset.m to conduct the following hypothesis tests, using a chi-squared test

* Null hypothesis : Feature K from class 1 comes from a Gaussian distribution at the significance level 0.001
* Null hypothesis : Feature K from class 2 comes from a Gaussian distribution at the significance level 0.001
* Null hypothesis : Feature K from class 3 comes from a Gaussian distribution at the significance level 0.001

Complete the following table with the decisions (acceptance/rejection) for the null hypothesis H0 (feature Gaussianity), p-value and degrees of freedom for *α* = 0,001.

Explain the meaning of the p-value and interpret the results accordingly.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature # 3 | Acceptance / rejection of *H*0 | *p*-value | Degrees of freedom |
| class 1 | H0 accepted | 0.2645 | 3 |
| class 2 | H0 accepted | 0.4050 | 3 |
| class 3 | H0 accepted | 0.0381 | 4 |

Q9. Copy the code used to answer Q8.

i\_feat = 3;

for i\_class=1:N\_class

i\_class

index=find(Labels==i\_class);

V=X(index,i\_feat);

[H,P,STATS] = chi2gof(V,'ALPHA',0.001,'nbins' ,10)

end