Homework 1

1 Question no. 1

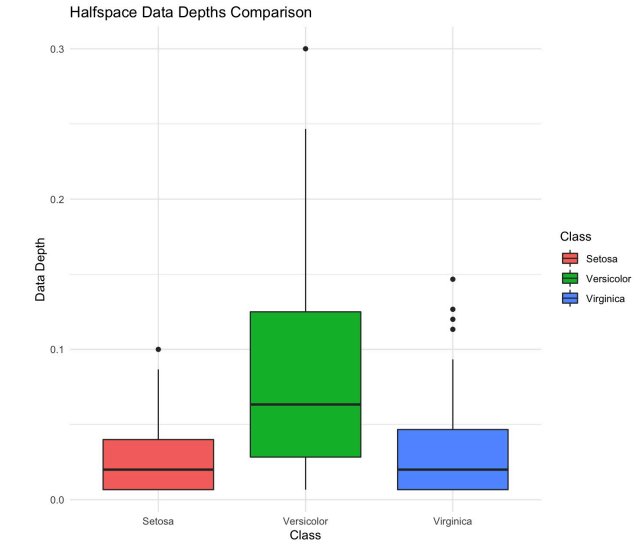
Download Iris data and check whether the observations associated with Iris setosa, Iris virginica and Iris versicolor obtained from the same distribution or not.

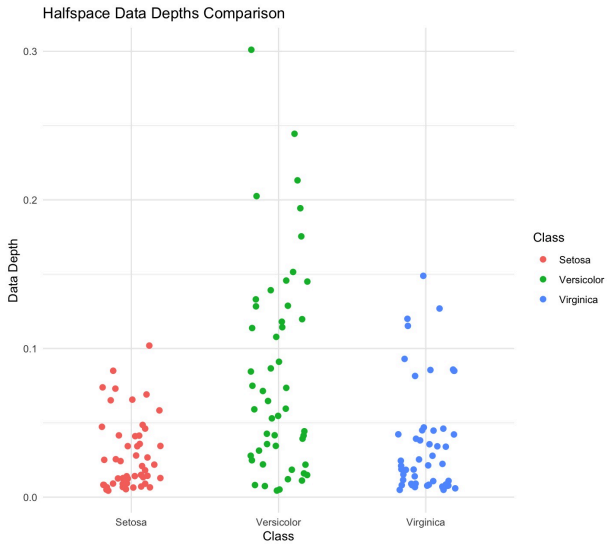
1.1 About the dataset

The iris dataset has 150 rows and 5 columns. With three species of flowers and each having 4 variables.

1.2 Data depth

In this study, we focus on the Iris dataset which includes measurements of sepal length, sepal width, petal length and petal width for three distinct species of iris flowers: setosa, versicolor, and virginica. Our primary goal is to examine the distributional patterns of these variables across the three species.
To achieve this, we employ the concept of data depth—a statistical measure that evaluates the centrality of individual data points within a dataset. By calculating and visualizing the data depth for each variable across iris species, we seek to discern similarities or differences in their distributions.
Here, we use Tukey depth where it is a measure of centrality based on the idea of how deep a point is within a dataset. It is defined as the minimum number of points, over all possible subsets required to "surround" a given point. The function is determining how central each point in setosa dataset is within the entire Iris dataset based on the Tukey depth measure. The result is a numerical vector of depths, one for each row in setosa. Higher depths generally indicate that a point is more central within the dataset.Higher data depth values indicate greater centrality within the distribution, while lower values may highlight potential variations or outliers.Based on the numerical values of Tukey depth we made the below plots.





But unfortunately we cannot infer from the plot completely that whether the variable is from different distribution,so moved to non parametric test.The box and jitter plot of half-space depth reveals distinct central tendencies for the setosa, virginica, and versicolor species. While these observations suggest potential differences in their underlying distributions, further statistical analysis is warranted to establish the certainty of such distinctions.

1.3 The hypothesis

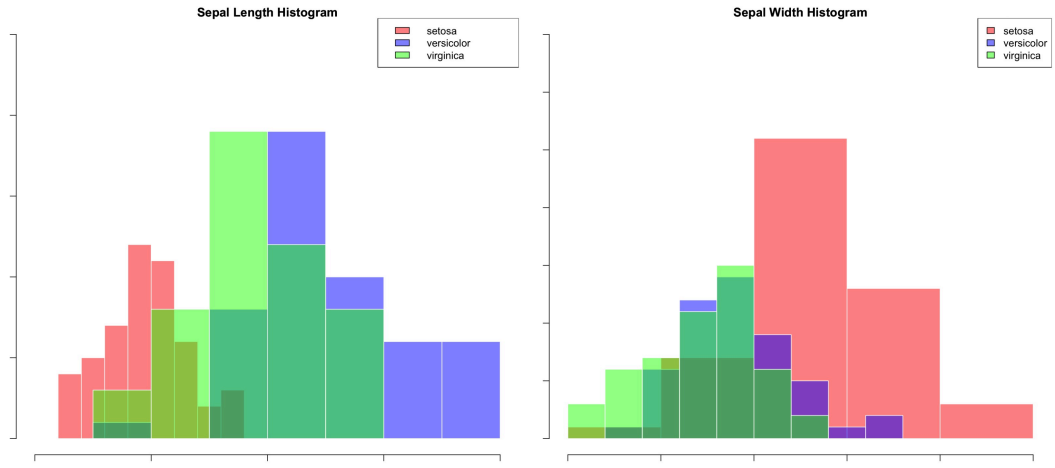
Null hypothesis : The data is coming from same distribution.
Alternate hypothesis : The data is not coming from the same distribution.
The test statistic for the Kruskal-Wallis test is

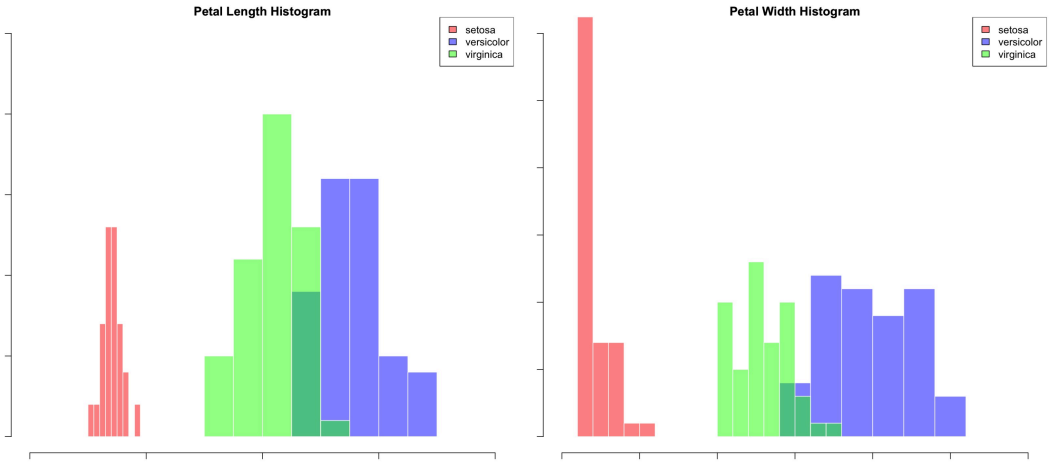
$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} (T_i^2 - N_i)$$

Where:

- n is the total number of observations across all groups.
- k is the number of groups.
- N_i is the number of observations in group i.
- T_i is the sum of ranks in group i.

After conducting the experiment, from the p values we conclude that the variables are from different distribution.
Also created a composite histogram by plotting the distributions of the four features in the Iris dataset for each of the three species on the same chart.





1.4 Conclusion

Firstly to check whether the three variables are from same distribution we checked data depth method by Tukey depth and plotted the graph but cannot make a completely affirm decision . So used Non parametric method of Kruskal-Wallis test and from the p values

	versicolor_vs_virginica	virginica_vs_setosa	setosa_vs_versicolor
Sepal.Length	5.764909e-07	6.212319e-17	8.129789e-14
Sepal.Width	4.522701e-03	6.954776e-09	2.088715e-13
Petal.Length	8.871573e-17	5.496344e-18	5.482553e-18
Petal.Width	9.419319e-17	2.358117e-18	2.214977e-18

From here we concluded that the variables are from different distribution and after plotting their histogram we confirm that they are from different distribution.

2 Question no. 2

Download a multivariate (i.e, dimension is strictly greater than one) data and compute/draw multivariate quantile contours when

$$||u|| = \frac{i}{10}, \tag{1}$$

where i = 1, . . . , 9. Using those contours, describe various features of the data set.

2.1 About the dataset

We took a "glass" dataset from chemometrics library that is already inbuilt in R. It had 180 rows and 13 columns or variables. Here we considered 2 variables namely MgO and Cl concentration,and the data dimension reduced to (180,2).

2.2 Theory

Mahalanobis depth is based on an outlyingness measure, viz. the Mahalanobis distance (Mahalanobis 1936) between z and a center of X, µ(X) say:

$$d^2_{\text{Mah}}(z; \mu(X), \Sigma(X)) = (z - \mu(X))^T \Sigma(X)^{-1} (z - \mu(X)).$$

The depth of a point z w.r.t. X is then defined as

$$D_{\text{Mah}}(z|X) = \frac{1}{1 + d^2{}_{\text{Mah}(z;\mu(X),\Sigma(X))}}$$

To determine quantiles (isolines) of a multivariate normal distribution.

The contour line is an ellipsoid. The reason is when we look at the argument of the exponential in the pdf of the multivariate normal distribution: the isolines would be lines with the same argument. Then we get

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c$$

where Σ is the covariance matrix. That is exactly the equation of an ellipse; in the simplest case, μ = (0,0) and Σ is diagonal, so we get

$$\left(\frac{x}{\sigma_x}\right)^2 + \left(\frac{y}{\sigma_y}\right)^2 = c$$

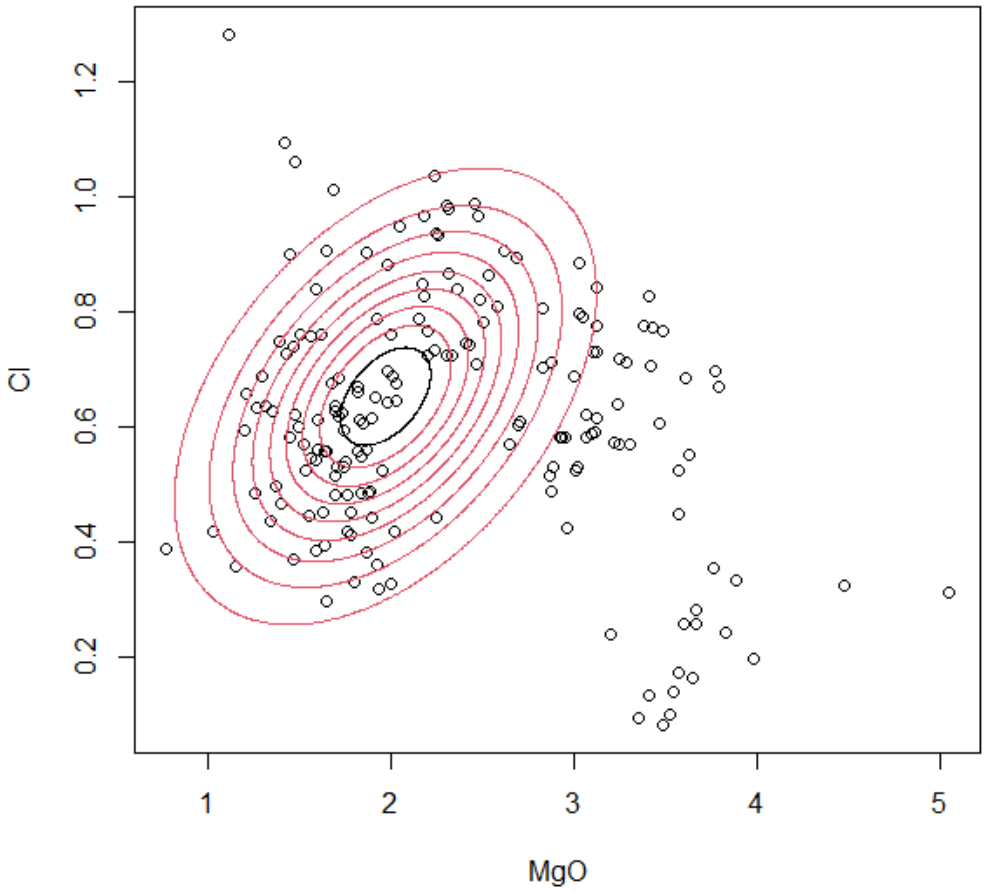
If Σ is not diagonal, diagonalizing we get the same result. Now, we would have to integrate the pdf of the multivariate inside (or outside) the ellipse and request that this is equal to the quantile we want. I would change variables in the pdf to $z^2 = \left(\frac{x}{\sigma_x}\right)^2 + \left(\frac{y}{\sigma_y}\right)^2$ integrate in the angle and then for z from 0 to $\sqrt{c}$
$1 - \alpha = \frac{1}{2\pi} \int_0^{\sqrt{c}} dz\, z e^{-\frac{z^2}{2}} \int_0^{2\pi} d\theta = \int_0^{\sqrt{c}} z e^{-\frac{z^2}{2}}\, dz$

Then you substitute $s = -\frac{z^2}{2}$.
$\int_0^{\sqrt{c}} ze^{-\frac{z^2}{2}}\, dz = \int_0^{-\frac{c}{2}} e^s\, ds = (1 - e^{-\frac{c}{2}})$

So in principle, you have to look for the ellipse centered in μ , with axis over the eigenvectors of Σ and effective radius -2 $\ln \alpha$

$(x - \mu)^T \Sigma^{-1} (x - \mu) = -2 \ln \alpha$

2.3 Interpretation



From the graph we see the 9 quantile contours, the smallest ellipse is 0.1 quantile and the subsequent biggers ellipse are of increasing quantile 0.2,0.3,.. till 0.9 .

2.4 Findings and Conclusion

Here for each value of α in the equation, we get a different quantile from the above equation $-2\log(\alpha)$. Here for $\alpha = 0.90$, we get the outer contour plot of the 90th percentile of the data, and continuingly for $\alpha = 0.1$, we get the 10th percentile of the data.

3 Refrences

Oleksii Pokotylo, Pavlo Mozharovskyi, Rainer Dyckerhoff,Depth and Depth-Based Classification with R Package ddalpha