Data Science 3 Project: Homework 1

Group 7 : Aindrila, Suchandra, Chirag, Paritosh, Vanshika

24 Jan, 2024

1 Question 1:

The IRIS data is downloaded. The IRIS dataset contains 50 observations on 4 variables for 3 different species. We have to check whether the distribution of data for these 3 species is identical or not.

Testing whether two datasets come from the same distribution is a common statistical problem. Several statistical tests are available for this purpose, and the choice of test depends on the nature of your data and assumptions about the distributions. There are various ways through which the distributions can be checked whether they are identical or different. Here are some commonly used tests:

- **Half Space Depth:** Half-space depth is a concept in mathematical and computational geometry employed to gauge the depth of a point within a set of points in a multidimensional space, particularly in the realms of data analysis and geometric algorithms.

  In a d-dimensional space with a given set of points, the half-space depth of a point serves as a metric for how deeply the point resides within the convex hull of the point set. Essentially, it denotes the fraction of half-spaces containing the point when considering all possible half-spaces defined by pairs of points from the set.

  This concept finds applications in robust statistics, machine learning, and outlier detection. It is utilized to pinpoint central or typical points within a dataset and to assess the depth of a point in the data distribution. Notably, the half-space depth provides a geometric measure of centrality and exhibits lower sensitivity to outliers compared to other statistical measures like the mean or median.

  At first, we consider iris setosa and iris versicolor. We combine both of these 4-dimensional datasets. Then, we compute the half-space depth value of the combined sample with respect to both setosa and versicolor successively. If both these datasets follow the same distribution, depths should approximately be the same and they should cluster around the $y = x$ line. Finally, this procedure is repeated for iris setosa, Iris virginica and iris versicolor, iris virginica.

  From 1, we can see that the distributions of Iris setosa and Iris Versicolor are different as the depth points do not cluster around the $y = x$ line. Similar results can be drawn from Figure 2 and Figure 3.

  Hence, we conclude that the observations associated with Iris setosa, Iris virginica and Iris versicolor obtained do not come from the same distribution.

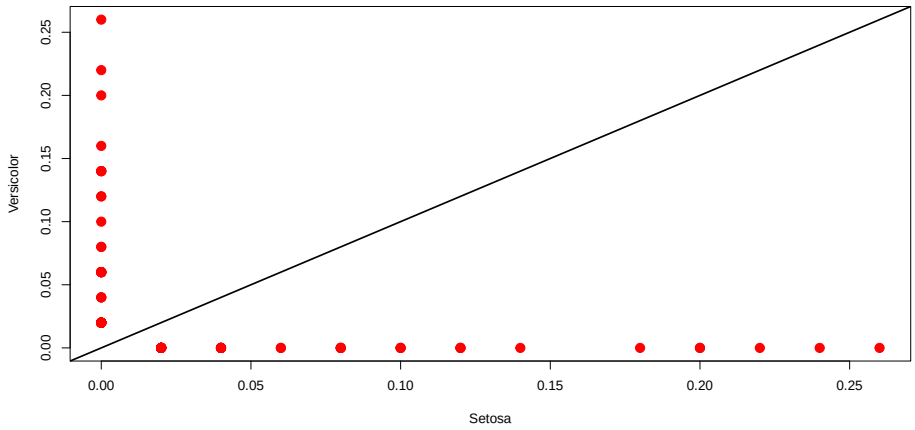

Figure 1: Half Space Depth of Iris Setosa and Iris Versicolor

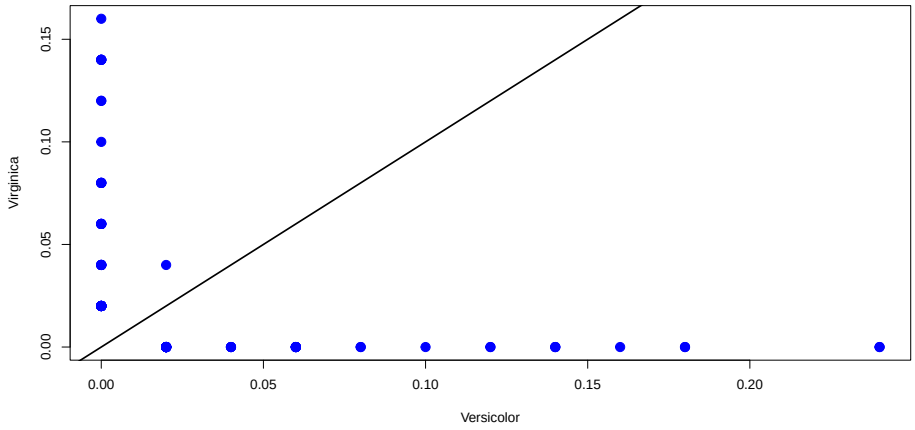

Figure 2: Half Space Depth of Iris Versicolor and Iris Virginica

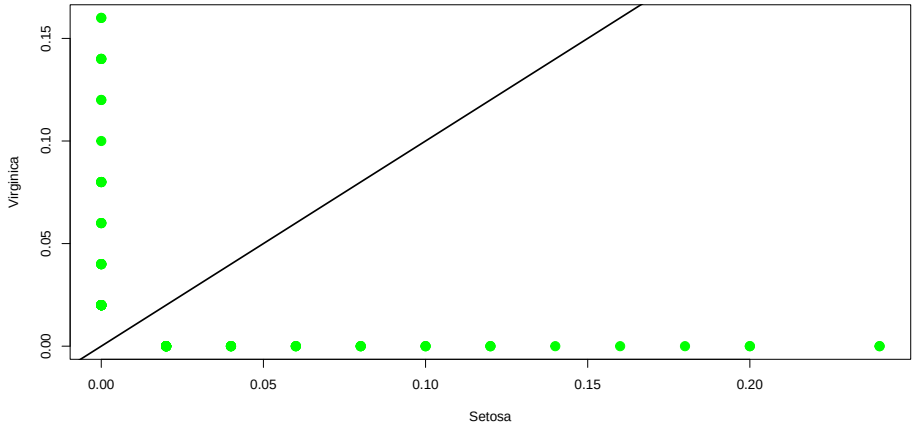

Figure 3: Half Space Depth of Iris Setosa and Iris Virginica

2 Question 2:

Our chosen bivariate data `https://github.com/AindrilaGarai/Data-Science-Lab-3/blob/main/ASSIGNMENT%201/gpa_study_hours_data.csv` contains 2 columns, namely GPA and Study hours for 193 observations. Let X be the independent variable i.e., Study hours, and Y be the dependent variable indicated by GPA.

Quantiles serve as statistical benchmarks that delineate specific points in a dataset, dividing it into intervals with predetermined proportions of data falling below each point. Commonly used quantiles include quartiles (25th, 50th, 75th percentiles) and deciles (10th, 20th, ..., 90th percentiles). Quantile contours, evident in scatterplots, are lines connecting data points sharing the same quantile value, offering a visual representation of the distribution of bivariate data across various quantile levels. For instance, the 0.5 quantile contour corresponds to the line linking points at the median.

In the context of the provided data, quantile contours are drawn for $i/10$ th quantiles, where i ranges from 1 to 9 in Figure 4. The contour level i corresponds to $i/10$th quantile for $i = 1, \ldots, 9$. For instance, if the 0.9th quantile contour encloses a small region, it signifies the concentration of the top 10 of data points. Overall, quantile contours serve as a comprehensive visual tool, depicting numerous statistical properties about the distribution of data, encompassing insights into patterns, clusters, and outliers at specific quantile levels.

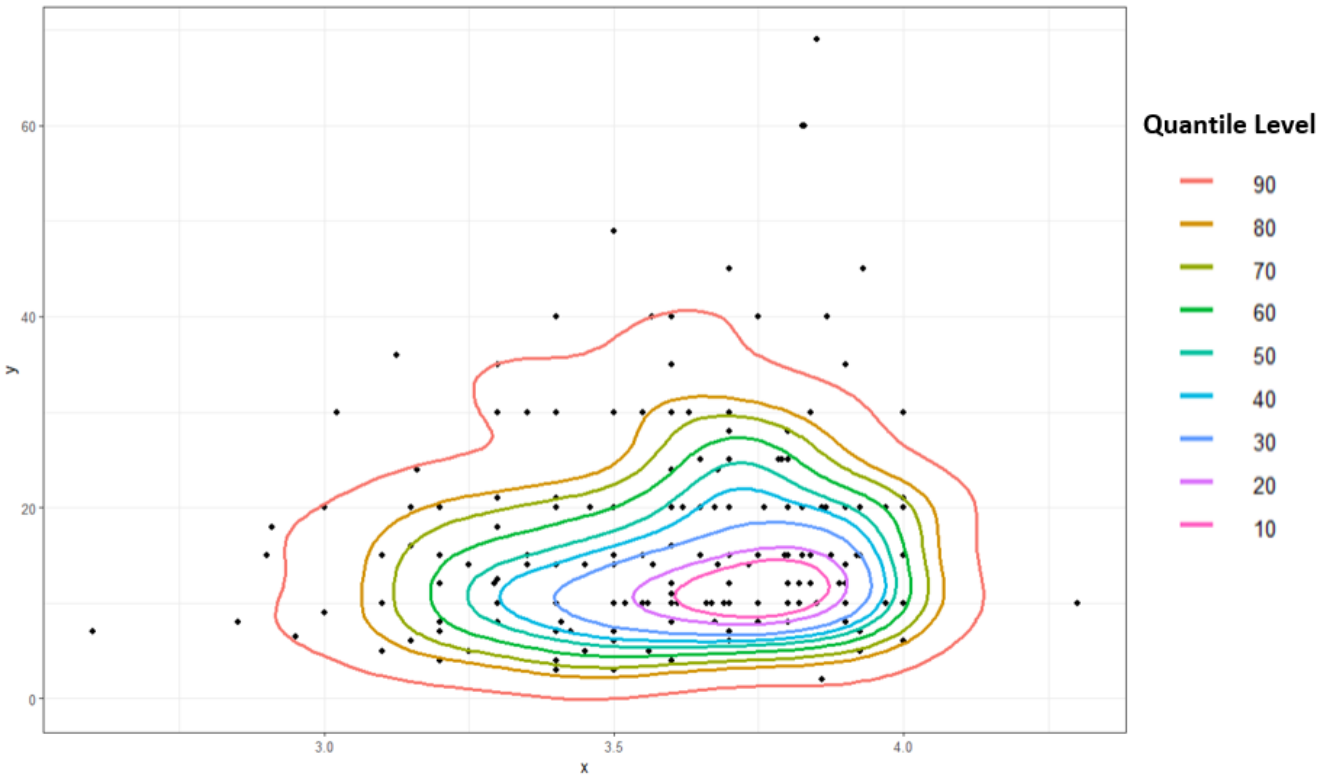

Figure 4: Quantile Contour Plot

Analyzing quantile contours allows for the observation of how points with similar quantile values are dispersed across the scatterplot, providing insights into the data's shape and spread at different quantile levels. These contours are invaluable for identifying patterns, clusters, or outliers at specific quantile thresholds. Particularly useful in multivariate scenarios, quantile contours offer a visual summary of how data points are distributed across diverse quantile levels, aiding in the identification of patterns and variations within the dataset.

- **Inferring Dependence structure:**

  Quantile contour plots can provide insights into the dependence structure between variables by visualizing the joint distribution of the variables.

  If the contours are elliptical and symmetric, it suggests a linear relationship between variables. The orientation and elongation of the ellipse indicate the strength and direction of the correlation. If the contours are not elliptical, it suggests non-linear or non-parametric dependence between variables. Areas, where contours overlap, indicate regions of higher density, suggesting stronger dependence between variables. Here, our plotted contours are almost elliptical, but neither symmetric nor overlapped, rather elongated to the left, indicating a nonlinear relationship between the two variables.

  While quantile contour plots visually represent the dependence structure, additional statistical analyses may be needed for a more comprehensive understanding. We have interpreted the dependence structure considering statistical measures, such as correlation coefficients, to quantify the strength of relationships. Numerically, using the "cor" function in R, we find that the correlation between X and Y is 0.1330, indicating a negligible correlation between them, which depicts there is no noticeable linear relation between the variables. This result supports our inference from the quantile contour plot.

- **Identifying Outliers of the data:**

  Outliers can be identified by looking for regions in the plot where the contours are spread out or have irregular shapes. Points that fall outside the main cluster of contours may be considered outliers.

  From the contour plot, we can observe 5 bivariate observations that do not belong to the cluster of observations and lie far from them. These points may be considered as outliers in the given data.

  This inference is supported by the values that we obtain numerically through the box plot. Figure 5 demonstrates the boxplots for the 2 variables X and Y. The boxplot for X has only one outlier whereas the boxplot for Y has 6 outliers. Hence we can conclude that observations having X values less than 2.7 or Y values greater than 38 can be considered as outliers.

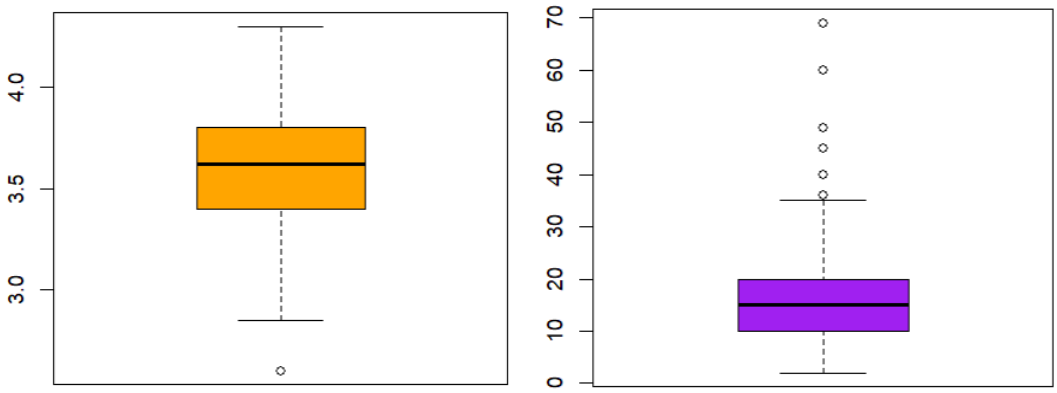

Figure 5: Boxplot to detect Outliers

- **Identifying Median of the data:**

  The median represents the middle value when the data is ordered, but in bivariate data, it is about finding a point where roughly 50% of the data lies on either side. Quantile contours

in bivariate data provide information about the distribution of the data at different quantile levels. The median represents the 50th percentile, which can be inferred from these contours.

The 0.5 quantile contour corresponds to the line that connects points in the scatterplot with the same value as the median. This contour will enclose the region containing approximately 50% of the data points. The point where the 0.5 quantile contour intersects is an estimate of the median in bivariate space.

Using package "rsdepth" in R, the median for this data comes out to be (3.609975,15.002513). From Figure 5 and from the quantile contour plots, we can also visualize that the median lies very close to the obtained value.

- **Identifying Mode of the data:**

  The mode of bivariate data refers to the combination of values that occurs most frequently in the dataset. To identify the mode in bivariate data, we need to examine the frequency of each pair and determine which combination occurs with the highest frequency. If there are multiple combinations with the same highest frequency, the dataset may be considered multimodal. The mode of a distribution can be inferred from the peak of the contour plot. The highest point in the plot represents the mode of the distribution. Finding the mode of bivariate (two-variable) data numerically involves identifying the combination of values for the two variables where the joint probability density is maximized. One common approach is to estimate the mode from the bivariate kernel density estimate.

  In the context of a quantile contour plot, the "peak" refers to the highest point or region on the plot, where the contours are closest together. The peak corresponds to the mode or the region of highest density in the joint distribution of the variables.

  A quantile contour plot represents regions of the bivariate distribution at different quantiles. Each contour line encloses a specific percentage of the data. The closer the contours are to each other, the higher the data density in that region.

  Identifying the peak on the contour plot is important for understanding where the most probable values are concentrated. The highest point on the plot generally corresponds to the mode of the distribution. In other words, it represents the combination of values for the variables where the joint probability is maximized.

  From the contour plot, we can observe that the mode lies somewhere around (3.75,9.8), which can be validated by the mode = (3.759896,10.723958) computed numerically using R. This mode is calculated using the "MASS" package, where after estimating the kernel density function, we find the grid point with the highest density in the kernel density estimate, which becomes the estimated mode for the given data.

- **Identifying Skewness of the data:**

  Skewness is a measure of the asymmetry of a probability distribution. In the context of bivariate data, skewness can be assessed for each variable separately. Bivariate skewness, however, refers to the asymmetry in the joint distribution of two variables.

  To calculate the bivariate skewness, you might want to use methods like the skewness coefficient or other measures that capture the departure from symmetry in two dimensions. One common approach is to use the third standardized moment, which is a measure of skewness. If the right tail of the distribution is longer or fatter than the left tail, the distribution is positively skewed. In a quantile contour plot, this might be reflected in a longer or stretched-out tail on the right side of the plot. If the left tail of the distribution is longer or fatter than the right tail, the distribution is negatively skewed. In a quantile contour plot, this might be reflected in a longer or stretched-out tail on the left side of the plot.

From the quantile contour plots, we can see a stretched-out tail on the left side of the plot, indicating negatively skewed data.

From the "fastmatrix" package in R, we have calculated Mardia's multivariate skewness for the dataset.

$$\text{Skewness} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{z}_i^T\mathbf{S}^{-1}\mathbf{z}_i)^{\frac{3}{2}}$$

where $\mathbf{z}_i$ is the vector of standardized deviations of the i-th observation from the mean, and $\mathbf{S}$ is the covariance matrix.

Mardia's skewness suggests positive skewness if the value is greater than 1. The skewness value obtained for this data is 3.397901.

- **Identifying Kurtosis of the data:**

  If the distribution has heavy tails and a sharp peak, it is leptokurtic. For a sharper and more peaked distribution in the quantile contour plot, with closely spaced contours near the peak, it would be leptokurtic. In a quantile contour plot, this might be reflected in closely spaced contours near the peak. If the distribution has light tails and a flatter peak with more spread-out contours near the peak, it is platykurtic. In a quantile contour plot, this might be reflected in more spread-out contours near the peak.

  While these visual cues can give you an idea, they are not a substitute for actual numerical estimates. Mardia's kurtosis coefficient measures the "tailedness" or the shape of the distribution's tails in a multivariate setting. For a dataset with p variables, the kurtosis is given by the formula:

$$\text{Kurtosis} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{z}_i^T\mathbf{S}^{-1}\mathbf{z}_i)^2 - p(p+2)$$

  The term $p(p+2)$ is subtracted to adjust for the kurtosis of a multivariate normal distribution.

  Here: n is the number of observations, $\mathbf{z}_i$ is the standardized vector of deviations for the i-th observation, and $\mathbf{S}$ is the covariance matrix.

  The kurtosis value obtained for this data is 11.04057.

  It's important to note that Mardia's skewness and kurtosis coefficients are often used in conjunction with statistical tests to assess the multivariate normality assumption. If the coefficients are significantly different from those of a multivariate normal distribution, it suggests a departure from normality. Since here the skewness value is 3.397901 and the kurtosis value is 11.04057, we may infer that the distribution of the data is not multivariate normal.

- **Identifying Inter Quantile Range (IQR) of the data:**

  The interquartile range (IQR) is a measure of statistical dispersion, specifically a measure of the spread or variability of a dataset. It is defined as the difference between the third quartile (Q3) and the first quartile (Q1) in a dataset.

$$\text{IQR} = Q_3 - Q_1$$

  The interquartile range is particularly useful because it is less sensitive to outliers than the range. It provides a measure of the spread of the middle 50% of the data, making it a robust

statistic for describing the variability in a dataset. We can numerically calculate multivariate quantiles using "optim" function in R.

In boxplots, the interquartile range is often represented by the length of the box. Outliers may be identified based on a multiplier (e.g., 1.5 times the IQR) from the quartiles. Data points beyond this range are considered potential outliers.