Classifying Iris Flower Species: A Study of Statistical Methods and Feature Effectiveness

Richard Meyer and Anett Martinez-Vazquez

Abstract—In this report, the Iris dataset, containing measurements of petal and sepal features for the three species—Versicolor, Setosa, and Virginica—was analyzed using statistical methods, including Exploratory Data Analysis (EDA), confidence interval estimation, and hypothesis testing. It was revealed through EDA that petal features, particularly petal length and width, provide the clearest separation among the species. In contrast, sepal features were found to exhibit significant overlap. By calculating the confidence intervals, it was confirmed that no overlap exists in the population means for petal features, validating their effectiveness. The hypothesis tests further supported the significant differences in petal length between species. Based on these observations, it was concluded that petal features are the most effective for classifying the three species, highlighting their reliability in distinguishing Iris species while underscoring the limitations of sepal-based classification.

Index Terms—Exploratory Data Analysis, Confidence interval, Hypothesis Test, Classification

I. INTRODUCTION

The Iris Flower Dataset, a widely utilized multivariate dataset, consists of 150 samples from three distinct Iris species: Setosa, Virginica, and Versicolor. Statistical methods are applied in this project to classify these species based on four measurements that facilitate accurate flower identification. Patterns and relationships that effectively distinguish between the species are examined through classification methods. Various visualization techniques, feature analyses, and statistical methods are employed throughout the report to provide insights into the dataset's characteristics and its potential for species classification based on these measurements.

The overall goal of this report is to analyze the Iris dataset using various statistical methods, including exploratory data analysis (EDA), confidence interval estimation, and hypothesis testing. These methods are applied to classify the Iris flower species based on their features. Scatter plots and histograms are observed to visually reveal feature relationships and distributions. Feature analysis and confidence interval estimation are performed to validate the observed behaviors identified through EDA. Hypothesis tests are conducted to compare feature differences among the three species. Based on these observations, effective classification guidelines for distinguishing the three species are proposed.

II. METHODS

Before beginning, an approach to the dataset must be established in order to gain an understanding of how the statistical methods may be applied to organize these features. Three classifier models were trained on the dataset: Exploratory Data Analysis (EDA), confidence interval estimation, and the use of hypothesis testing. Each of these methods will be examined to understand how they function, their utility, and their contribution to the context of this report. In this section, the general methods used to analyze the Iris dataset are outlined. The data was accessed through the 'data(iris)' function in RStudio, and all analyses were performed using R with custom scripts.

A. Exploratory Data Analysis Method

EDA is a fundamental method used to analyze the key characteristics of a dataset. This method involved the use of graphical and statistical techniques to uncover patterns, relationships, and irregularities within the dataset, providing valuable insights before applying more formal statistical modeling or hypothesis testing. Common EDA techniques included in this report are scatterplots, histograms, boxplots, and correlation analysis. These plots will help identify trends, distributions, and potential outliers within the dataset, as will be observed later in this report. Specifically in this report, the EDA method will be used to visualize feature pairs such as sepal length and width, or petal length and width, and to identify clear patterns and clustering behaviors among the three Iris species. While EDA is essential for understanding and summarizing the dataset, it generally focuses on pairwise relationships between features. Therefore, relying solely on EDA may result in missing more complex multivariate interactions. More specifically, EDA alone may not reveal how combinations of all features together could offer better separation. Additionally, EDA does not provide formal hypothesis testing or validation of findings. In the context of this report, it will be shown that EDA is useful for revealing visual patterns to the analyst but fails to provide formal statistical proof to confirm the observed relationships between the features. Thus, the dataset will need to be investigated using other methods.

B. Confidence Interval Estimation

To prevent these potential errors, the dataset was also investigated using the confidence interval estimation method. Confidence interval (CI) estimation is a statistical approach used to determine the range within which a population parameter, such as the mean, is likely to fall based on a sample from that population. A confidence interval consists of an upper and lower bound around the sample estimate, providing an interval estimate rather than a single-point estimate. The width of the interval reflects the precision of the estimate; a narrower interval indicates greater precision, while a wider interval suggests less precision. For instance, if the population mean of a single feature for a species in the dataset is estimated, the confidence interval provides a range within which the true mean of the entire population is likely to lie, based on the sample data. In the context of this project, confidence interval estimation was used to assess the reliability of the sample means for the given features of each Iris species. This was accomplished using the following equation, which applies the statistical formula for the confidence interval:

$$CI = \bar{x} \pm Z \cdot \frac{s}{\sqrt{n}} \quad (1)$$

Where $\bar{x}$ is the sample mean is the Z-score for the desired confidence level, s is the sample standard deviation, and n is the sample size. By calculating the confidence intervals for these features, the level of uncertainty surrounding the estimated means for each species was determined. This approach allowed for an evaluation of whether the differences between species are statistically significant. If the confidence intervals for two species do not overlap, this would suggest a significant difference between their means, thereby supporting the visual observations made during exploratory data analysis.

C. Hypothesis Test Method

The use of hypothesis testing as a statistical method for analyzing the dataset was mentioned earlier. This method involves formulating two opposing hypotheses: the null hypothesis H_O and the alternative hypothesis H_a. The null hypothesis, generally and for the purpose of this report, represents the default assumption that no effect or difference exists, while the alternative hypothesis represents the claim being tested, such as the presence of a difference or effect. A test statistic is calculated from the sample data, and a p-value is obtained to determine whether the null hypothesis should be rejected or not. The significance level α, is used as a threshold for interpreting the p-value. If the p-value is less than α, the null hypothesis is rejected in favor of the alternative hypothesis, indicating that the observed data provide sufficient evidence to support the claim being tested. In this project, hypothesis testing is applied to evaluate differences in feature values among the three Iris species. The equation used to conduct

a comparing of population means, also known as the t-test is calculated as:

$$t = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} \left(\bar{x}_1 - \bar{x}_2 \right) \quad (2)$$

Where $\bar{x}_1$ and $\bar{x}_2$ represent the sample means of the two groups of features being compared. The sample variances of the two groups are represented by s_1^2 and s_1^2, while n_1 and n_2 are the sample sizes of the two groups. Based on the test statistic outputs for the t-tests, the decision rule can be used to determine whether these differences can be rejected. The decision rule states that if the absolute value of ttt is greater than the critical value, which is based on the degrees of freedom and α, the null hypothesis H_O is rejected. As previously mentioned, in order to assess whether the null hypothesis can be rejected, the degrees of freedom (df) must be determined. To do so, equation 3 is used:

$$df = \sum_{i=1}^{k} n_i - k \quad (3)$$

The population means of features are compared between pairs of Iris species using the hypothesis test method. Later in the report, it will be shown how performing such tests allows for the evaluation of whether differences in sample means are statistically significant or could have occurred due to random sampling variability. By applying hypothesis tests to feature values across species, a layer of statistical rigor is added to the analysis, strengthening the argument that certain features are more effective discriminators than others.

III. RESULTS AND DISCUSSION

A. Exploratory Data Analysis

As seen in Figure 1, the scatter plots for all feature pairs demonstrate a clear separation of the three species across all feature combinations. It is evident from this plot that petal length versus petal width provides the clearest separation among the three species. In contrast, sepal length and sepal width show the greatest overlap, suggesting that these two features may not be effective for classifying the species. Notably, Versicolor and Virginica overlap in most feature combinations and are most distinguishable in petal-related features than in sepal-related features. While Setosa is clearly separated from Versicolor and Virginica across most feature pairs due to its smaller values. To further analyze the features, it may be helpful for the dataset to be separated into subsets for each species. The box plots, figure 2, confirm that Setosa is clearly separated from Versicolor and Virginica across all features. Similar to the scatterplot shown in Figure 1, petal length and petal width display distinguishing characteristics among the species. This observation further supports the conclusion
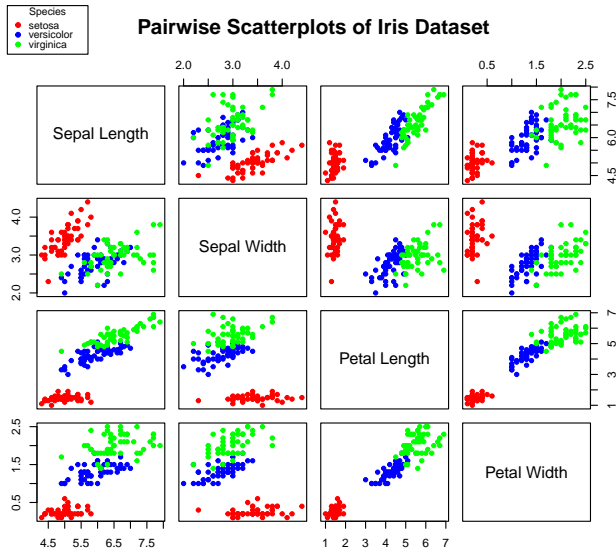
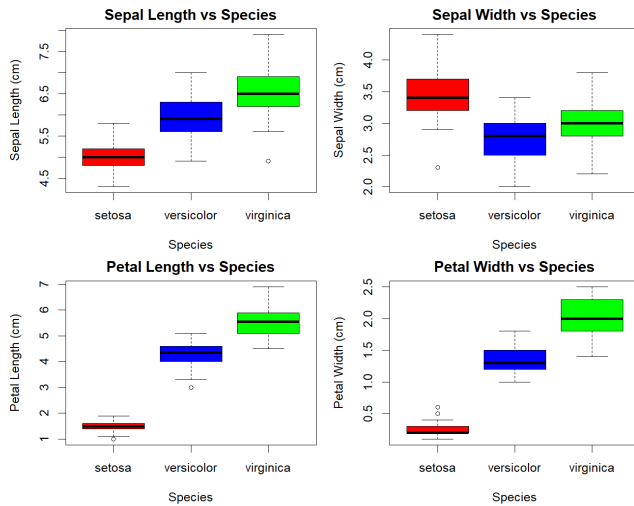Fig. 1. Pairwise scatterplot of Iris dataset features

Fig. 2. Box Plots of Sepal and Petal Features by Species

that sepal length and sepal width exhibit significant overlap between Versicolor and Virginica, making them less effective for species classification. Overall the exploratory data analysis reveals distinct patterns across the four features of the Iris dataset. Setosa is easily distinguishable from Versicolor and Virginica due to its smaller sepal length, smallest petal length, and smaller petal width, as well as its notably larger sepal width. In contrast, Versicolor and Virginica exhibit significant overlap in sepal length, sepal width, and petal width, making classification between them more challenging. However, petal length proves to be a highly effective feature for distinguishing all three species, as Virginica shows consistently larger values compared to Versicolor. These observations highlight the relative importance of petal-related features for species

classification.

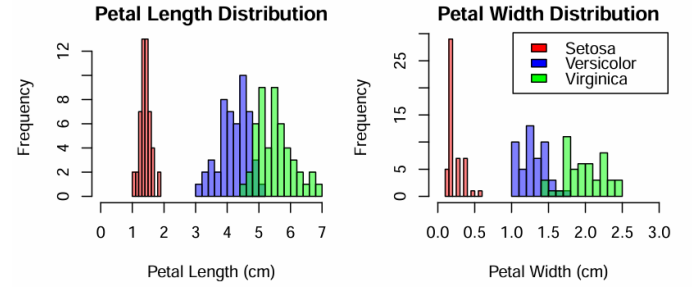### B. Confidence Interval Estimation Analysis

Fig. 3. Histograms of Petal Length and Width Distribution

The histogram presented in Figure 3 displays notable observations including Setosa having smaller petal lengths and widths, with tight clustering around 1.5 cm for petal length and 0.2 cm for petal width. Versicolor displays a broader range of values for both petal length and petal width, while Virginica exhibits the largest petal length distribution among the three species. The histogram further reveals that the petal length samples for each species are roughly normally distributed, confirming the appropriateness of using a t-test to compare the species' petal length means.

TABLE I
CONFIDENCE INTERVAL ESTIMATES FOR PETAL LENGTH BY SPECIES

Species	Mean	Lower Bound	Upper Bound
Setosa	1.46	1.41	1.51
Versicolor	4.26	4.13	4.39
Virginica	5.5	5.4	5.71

The confidence intervals for petal length indicate that a clear separation exists between species. It should be noted that none of the intervals overlap, which confirms that the population means for petal length differ significantly across all species. The petal lengths of Setosa are tightly clustered around the mean, reflecting low variability. In contrast, Versicolor petal lengths exhibit greater variability, as evidenced by a wider confidence interval around the mean. This formally suggests that Versicolor petal lengths are significantly larger than those of Setosa. Based on the confidence intervals, Virginica petal lengths demonstrate the greatest variability among the three species.

To compute the confidence interval estimate for the population mean of a species subset feature, a function was created to perform this calculation. The desired confidence level was set to a default of 0.95. The sample mean for the given data points was calculated, followed by the computation of the standard deviation to determine the standard error of the mean. The qt() function was used to obtain the t-distribution critical value corresponding to the desired confidence level. The confidence interval, representing the range around the sample mean where

the true population mean is likely to lie, was then calculated. This interval indicates where the true population mean of petal length for each species is expected to lie with 95% certainty. Since the confidence intervals for Setosa, Versicolor, and Virginica did not overlap, it can be concluded that the population means are significantly different across the three species.

C. Hypothesis test Analysis

Previously, it was determined that petal length is a strong feature for classifying Iris flower species, but this conclusion requires validation. Hypothesis testing is utilized to determine if a significant difference exists between the means of two samples. Since small sample sizes are involved, and the petal length in each species sample was found to be normally distributed with unknown population parameters, a t-test is appropriate for comparing means. To compute the hypothesis test, the pooled standard deviation was first calculated to combine the variability of the two samples. Next, the t-statistic was calculated to measure the difference in means relative to variability. The degrees of freedom were then determined, as they are required for calculating p-values and confidence intervals using the t-distribution. The p-value was computed to assess the strength of evidence against the null hypothesis, representing the probability of observing a result as extreme as the one obtained under the null hypothesis. Finally, the confidence interval was calculated to estimate the true difference in population means.

The hypothesis tests provide strong evidence of significant differences in petal length between the three Iris species. For the comparison between Virginica and Versicolor, a one-tailed t-test was conducted with the null hypothesis stating that the mean petal length of Virginica is less than or equal to that of Versicolor. The resulting p-value of $1.59e^{-22}$ was found to be far below the significance level of 0.05, leading to the rejection of the null hypothesis. As a result, it was concluded that Virginica has a significantly greater mean petal length than Versicolor. The same test was conducted to compare the petal length of Versicolor and Setosa. The null hypothesis proposed that the mean petal length of Versicolor is less than or equal to that of Setosa. With a p-value of $2.7e^{-62}$, which is significantly less than 0.05, the null hypothesis was rejected, confirming that Versicolor has a significantly greater mean petal length than Setosa. These results support the conclusion that petal length is an effective feature for distinguishing between the three species.

D. Proposed Classification Standards

The results obtained can be distinguished using certain rules, with petal features proving to be the most effective for classifying the three Iris flower species. When petal length is used, a length under 2 cm should be classified as Setosa. A petal length between 2 cm and 5 cm, inclusive, indicates Versicolor. Any petal length exceeding 5 cm is classified as Virginica. For classification based on petal width, observations under 0.8 cm should be classified as Setosa. Petal widths between 0.8 cm and 1.75 cm, inclusive, correspond to Versicolor, while those greater than 1.75 cm indicate Virginica.

The use of sepal features for classification is not recommended, as additional features may be required for further separation. However, if sepal features must be used, the following criteria can be applied: a sepal length of less than 5.2 cm should be classified as Setosa. A sepal length between 5.2 cm and 6.3 cm, inclusive, should be identified as Versicolor, while any sepal length exceeding 6.3 cm is classified as Virginica. For classification based on sepal width, any width greater than 3.5 cm should be classified as Setosa. If the sepal width falls between 2.1 cm and 3.5 cm, inclusive, additional feature information would be needed to distinguish between Versicolor and Virginica. Refer to table 2 for a concise organization of the proposed classification guidelines.

TABLE II
CLASSIFICATION GUIDELINES FOR IRIS SPECIES

Feature (cm)	Setosa	Versicolor	Virginica
Petal Length	<2	2 − 5	>5
Petal Width	<0.8	0.8 − 1.75	>1.75
Sepal Length	<5.2	5.2 − 6.3	>6.3
Sepal Width	>3.5	requires more info	requires more info

CONCLUSION

The analysis of the Iris dataset revealed that petal-related features, specifically petal length and petal width, are the most effective for distinguishing between the three Iris flower species: Setosa, Versicolor, and Virginica. Setosa was found to be easily classified due to its distinct separation from Versicolor and Virginica based on petal length and petal width. In comparison, Versicolor and Virginica exhibited some overlap in sepal features, requiring classification based on petal features to ensure accurate identification. Confidence interval analysis confirmed that the population means for petal length and width differ significantly across the species, with no overlap in intervals, thereby validating the effectiveness of these features for classification. Hypothesis testing further supported these findings by demonstrating statistically significant differences in petal length between all species pairs, with p-values less than 0.05 in all tested cases. Specifically, Virginica's petal length was shown to be significantly larger than Versicolor's, and Versicolor's petal length was significantly larger than Setosa's. Based on these observations, classification rules were established to distinguish between the three species. While sepal features provide some discriminatory power, significant overlap between Versicolor and Virginica renders them less effective for classification. Consequently, the classification rules prioritize petal features. These findings highlight the importance of combining statistical methods with

visual analyses to develop robust classification guidelines. Overall, the results emphasize the critical role of petal features in reliably identifying Iris species and the limitations of sepal-based classification.

IV. CONTRIBUTION

This report and its results were achieved through the collaborative efforts of team members Richard Meyer and Anett Martinez-Vazquez, as outlined below.

TABLE III
COLLABORATION DETAILS

Collaborator	Contribution
Anett Martinez-Vazquez	Provided clear interpretations of the results and ensured the accuracy of key observations.
Richard Meyer	Responsible for data collection and the development of essential functions for analysis.

```r
1   data(iris)
2
3   # feature analysis: scatterplot matrix
4
5   species_colors <- c("setosa" = "red", "versicolor" =
        "blue", "virginica" = "green")
6
7   pairs(
8     iris[, 1:4],
9     col = species_colors[iris$Species],
10    pch = 19,
11    main = "Pairwise_Scatterplots_of_Iris_Dataset",
12    labels = c("Sepal_Length", "Sepal_Width", "Petal_
        Length", "Petal_Width")
13  )
14
15  # feature analysis: feature vs species boxplots
16
17  par(mfrow = c(2,2))
18
19  boxplot(Sepal.Length ~ Species, data = iris,
20          main = "Sepal_Length_vs_Species",
21          col = c("red","blue","green"),
22          ylab = "Sepal_Length_(cm)")
23  boxplot(Sepal.Width ~ Species, data = iris,
24          main = "Sepal_Width_vs_Species",
25          col = c("red","blue","green"),
26          ylab = "Sepal_Width_(cm)")
27  boxplot(Petal.Length ~ Species, data = iris,
28          main = "Petal_Length_vs_Species",
29          col = c("red","blue","green"),
30          ylab = "Petal_Length_(cm)")
31  boxplot(Petal.Width ~ Species, data = iris,
32          main = "Petal_Width_vs_Species",
33          col = c("red","blue","green"),
34          ylab = "Petal_Width_(cm)")
35
36
37  par(mfrow= c(2,2))
38
39
40  # petal length histogram for each species
41  hist(iris$Petal.Length[iris$Species == "setosa"],
42       breaks = 10, col = rgb(1, 0, 0, 0.5),
43       xlim = c(0, 7), main = "Petal_Length_
          Distribution",
44       xlab = "Petal_Length_(cm)", freq = TRUE)
45  hist(iris$Petal.Length[iris$Species == "versicolor"
        ],
46       breaks = 10, col = rgb(0, 0, 1, 0.5),
47       add = TRUE)
48  hist(iris$Petal.Length[iris$Species == "virginica"],
49       breaks = 10, col = rgb(0, 1, 0, 0.5),
50       add = TRUE)
51  legend("topright", legend = c("Setosa", "Versicolor"
        , "Virginica"),
52         fill = c("red", "blue", "green"))
53
54  # petal width histogram for each species
55  hist(iris$Petal.Width[iris$Species == "setosa"],
56       breaks = 10, col = rgb(1, 0, 0, 0.5),
57       xlim = c(0, 3), main = "Petal_Width_
          Distribution",
58       xlab = "Petal_Width_(cm)", freq = TRUE)
59  hist(iris$Petal.Width[iris$Species == "versicolor"],
60       breaks = 10, col = rgb(0, 0, 1, 0.5),
61       add = TRUE)
62  hist(iris$Petal.Width[iris$Species == "virginica"],
63       breaks = 10, col = rgb(0, 1, 0, 0.5),
64       add = TRUE)
65  legend("topright", legend = c("Setosa", "Versicolor"
        , "Virginica"),
66         fill = c("red", "blue", "green"))
67
68  # function for finding confidence interval estimate
        of the mean
69
70  confidence_interval <- function(data, confidence_
        level = 0.95) {
71    n <- length(data)
72    sample_mean <- mean(data)
73    standard_error <- sd(data) / sqrt(n)
74
75    alpha <- 1 - confidence_level
76    critical_value <- qt(1 - alpha / 2, df = n - 1)
77
78    margin_of_error <- critical_value * standard_error
79
80    lower_bound <- sample_mean - margin_of_error
81    upper_bound <- sample_mean + margin_of_error
82
83    return (list(
84      mean = sample_mean,
85      lower_bound = lower_bound,
86      upper_bound = upper_bound,
87      confidence_level = confidence_level
88    ))
89  }
90
91  # extract petal length data for each species
92
93  setosa_petal_length <- iris$Petal.Length[iris$
        Species == "setosa"]
94  versicolor_petal_length <- iris$Petal.Length[iris$
        Species == "versicolor"]
95  virginica_petal_length <- iris$Petal.Length[iris$
        Species == "virginica"]
96
97  # calculate confidence interval estimate for
        population mean of each species petal length at
        confidence level 0.95
98
99  ci_setosa <- confidence_interval(setosa_petal_length
        )
100 ci_versicolor <- confidence_interval(versicolor_
        petal_length)
101 ci_virginica <- confidence_interval(virginica_petal_
        length)
102
103 # hypothesis test for two population means (t-test)
104
105 hypothesis_test_means <- function(sample1, sample2,
        alternative = "two-sided", confidence_level) {
106   # relevant sample sizes, means, and variances
107
108   n1 <- length(sample1)
109   n2 <- length(sample2)
110
111   mean1 <- mean(sample1)
112   mean2 <- mean(sample2)
113
114   var1 <- var(sample1)
115   var2 <- var(sample2)
116
117   # pooled standard deviation
118
119   pooled_sd <- sqrt(((n1 -1) * var1 + (n2-1) * var2)
          / (n1 + n2 - 2))
120
121   # standard error of mean difference
122
123   standard_error <- pooled_sd * sqrt(1/ n1 + 1 / n2)
124
125   # t statistic
126
127   t_statistic <- (mean1 - mean2) / standard_error
```

```r
128
129    # degrees of freedom
130
131    df <- n1 + n2 - 2
132
133    # p value calculation
134
135    if (alternative == "greater") {
136      p_value <- pt(t_statistic, df, lower.tail =
                FALSE)
137    } else if (alternative == "less") {
138      p_value <- pt(t_statistic, df, lower.tail = TRUE
                )
139    } else {
140      p_value <- 2 * pt(-abs(t_statistic), df)
141    }
142
143    # confidence interval for mean difference
144
145    alpha <- 1 - confidence_level
146    critical_value <- qt(1 - alpha / 2, df)
147    margin_of_error <- critical_value * standard_error
148
149    ci_lower <- (mean1 - mean2) - margin_of_error
150    ci_upper <- (mean1 - mean2) + margin_of_error
151
152    return(list(
153      t_statistic = t_statistic,
154      df = df,
155      p_value = p_value,
156      confidence_interval = c(ci_lower, ci_upper),
157      mean_difference = mean1 - mean2
158    ))
159  }
160
161  # perform hypothesis test for population means of (
          virg vs vers) and (vers vs seto)
162
163  ht_virginica_versicolor <- hypothesis_test_means(
          virginica_petal_length, versicolor_petal_length,
           alternative = "greater", confidence_level =
          .95)
164  ht_versicolor_setosa <- hypothesis_test_means(
          versicolor_petal_length, setosa_petal_length,
          alternative = "greater", confidence_level = .95)
```