Statistics

Assignment-1

Submitted By:

HUSSAIN MURTAZA ALI

DS COHORT 4 VIOLET GROUP

Introduction:

In this assignment, I've performed a comprehensive statistical analysis of the Iris Species dataset. This dataset consists of measurements of four features of three different species of Iris flowers i.e. (Iris Setosa, Iris Versicolor, and Iris Virginica). These features are as follows,

1. Sepal Length (cm)
2. Sepal Width (cm)
3. Petal Length (cm)
4. Petal Width (cm)

Task 1: Descriptive Statistics

Sepal Length (cm)		***Sepal Width (cm)***		***Petal Length (cm)***		***Petal Width (cm)***	
Mean	5.84	**Mean**	3.05	**Mean**	3.76	**Mean**	1.20
Median	5.80	**Median**	3.00	**Median**	4.35	**Median**	1.30
Mode	5.00	**Mode**	3.00	**Mode**	1.50	**Mode**	0.20
Standard Deviation	0.83	**Standard Deviation**	0.43	**Standard Deviation**	1.76	**Standard Deviation**	0.76
Variance	0.6811222	**Variance**	0.1867507	**Variance**	3.0924249	**Variance**	0.5785316

Mean:

These are the average measurements of these four attributes mentioned below. We can see that the Mean of Sepal Length is greater than that of other attribute's means.

- **Sepal Length (cm):** 5.84
- **Sepal Width (cm):** 3.05
- **Petal Length (cm):** 3.76
- **Petal Width (cm):** 1.20

Median:

These are the middle values of each attribute. We use median when there are outliers in the dataset, because if there's any outlier, it'll impact the mean, but in this case, there are not many outliers, as we can see that the Medians of all the attributes are near the Means, except for Petal Length (cm), as there are some outliers between **2.2-3.4** cm.

- **Sepal Length (cm):** 5.80
- **Sepal Width (cm):** 3.00
- **Petal Length (cm):** 4.35
- **Petal Width (cm):** 1.30

Mode:

These are the most frequently appearing values in a dataset.

- **Sepal Length (cm):** 5.00
- **Sepal Width (cm):** 3.00
- **Petal Length (cm):** 1.50
- **Petal Width (cm):** 0.20

Standard Deviation:

This shows us how far are the datapoints from their mean. It shows us the spread of datapoints.

- **Sepal Length:** 0.82 cm
  The relatively small Standard Deviation shows that sepal length values tend to gather closely around the mean Sepal Length of this dataset (i.e. 5.84 cm). This implies that sepal length data points are not highly spread out and are relatively consistent.
- **Sepal Width:** 0.43 cm
  Sepal Width also has a relatively small Standard Deviation that tells us that the datapoints are very close to the Mean Sepal Width (i.e., 3.05 cm). They are not very spread-out suggesting consistency in datapoints.

- **Petal Length:** 1.76 cm

  Petal length has a relatively higher Standard Deviation. This implies that the datapoints are relatively spread out from Mean Petal Length (i.e., 3.76 cm). This shows that there's a lot of variation in these datapoints.

- **Petal Width:** 0.76 cm

  Petal Width has a relatively low Standard Deviation compared to Petal, instilling the fact that the the datapoints are closer to Mean Petal Width (i.e., 1.20 cm)

Variance:

- **Sepal Length:** 0.68
- **Sepal Width:** 0.18
- **Petal Length:** 3.09
- **Petal Width:** 0.57

Task 2: Correlation Analysis

	Sepal Length (cm)	*Sepal Width (cm)*	*Petal Length (cm)*	*Petal Width (cm)*
Sepal Length (cm)	1			
Sepal Width (cm)	-0.10936925	1		
Petal Length (cm)	0.871754157	-0.420516096	1	
Petal Width (cm)	0.817953633	-0.35654409	0.962757097	1

Sepal Length – Sepal Width Correlation:

The Correlation coefficient between Sepal Length and Sepal Width is **-0.109.** This is a weak negative correlation between the two attributes, which means that if the Sepal length would increase, then the Sepal width would decrease.

Sepal Length – Petal Length Correlation:

The Correlation coefficient between Sepal Length and Petal length is **0.871.** This shows strong positive correlation between the two attributes. It means that if the Sepal Length is increased, the Petal length will also increase. This indicates a strong linear positive correlation.

Sepal Length – Petal Width Correlation:

The Correlation coefficient between Sepal Length and Petal width is **0.817.** This shows strong positive correlation between the two attributes. It means that if the Sepal Length is increased, the Petal Width will also increase. This indicates a strong linear positive correlation.

Sepal Width – Petal Length Correlation:

The Correlation coefficient between Sepal Width and Petal Length is **-0.42.** This shows moderate negative correlation between the two attributes. It means that if the Sepal Width is increased, the Petal length will also decrease moderately. This indicates a moderate negative correlation.

Sepal Width – Petal Width Correlation:

The Correlation coefficient between Sepal Width and Petal Length is **-0.356.** This shows moderate negative correlation between the two attributes. It means that if the Sepal Width is increased, the Petal Width will also decrease moderately. This indicates a moderate negative correlation.

Petal Length – Petal Width Correlation:

The Correlation coefficient between Petal Length and Petal Width is **0.962.** This shows a very high positive correlation between the two attributes. It means that if the Petal Length is increased, the Petal Width will also increase significantly. This indicates a very high positive linear correlation.

Task 3: Hypothesis Testing

t-Test: Paired Two Sample for Means		
	Sepal Length (cm) Setosa	***Sepal Length (cm) versicolor***
Mean	5.006	5.936
Variance	0.12424898	0.266432653
Observations	50	50
Pearson Correlation	-0.080849727	
Hypothesized Mean Difference	0	
df	49	
t Stat	-10.14589948	
P(T<=t) one-tail	6.20957E-14	
t Critical one-tail	1.676550893	
P(T<=t) two-tail	1.24191E-13	
t Critical two-tail	2.009575237	

Null Hypothesis (Ho): Mean Sepal Length of Iris-Setosa is equal to Mean Sepal Length of Iris Versicolor.

$$\mu(Iris\ setosa) = \mu(Iris\ versicolor)$$

Alternate Hypothesis (Ha): Mean Sepal Length of Iris-Setosa is not equal to Mean Sepal Length of Iris Versicolor.

$$\mu(Iris\ setosa) \neq \mu(Iris\ setosa)$$

From this table, we can see that the **P(T<=t) two-tail** is very small, **1.24E-13** which is **smaller** than 0.05, hence, we will **reject** our Null Hypothesis (Ho). This shows that there's a huge difference between the mean Sepal Length of Iris Setosa and Iris Versicolor.

Task 4: Regression Analysis

Regression Statistics	
Multiple R	0.10936925
R Square	0.011961633
Adjusted R Square	0.005285698
Standard Error	0.825874775
Observations	150

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	6.481223211	0.481295118	13.46621431	1.72623E-27
Sepal Width (cm)	-0.208870294	0.156040557	-1.338564142	0.182765215

Significance F
0.1827652

Multiple R: This number **0.1094** tells you that there is a very weak relationship between Sepal Width and Sepal Length. They don't change much together.

R Square: The number **0.01196** means that only about 1.20% of the Sepal Length can be explained by Sepal Width. In other words, sepal width doesn't have much influence on sepal length.

Coefficient: Coefficient for Sepal width is **-0.208,** which means, that if the Sepal Length increases by one unit, there would be a decrease of **0.208 cm** in Sepal Width.

P-Value: The P-Value for Sepal Width is **0.1828**, which is relatively high. This suggests that the relationship between sepal width and sepal length isn't strong enough to be considered significant.

Regression Equation:

Sepal Length = 6.4812 - 0.2089 × Sepal Width

Now, for example, if we have a Sepal width of **4.5 cm,** we can predict the estimated Sepal Length by using this equation would be **5.541 cm**

S.L = 6.4812 – 0.2089 * 4.5

S.L = 6.4812 – 0.94005

S.L = 5.541 cm

Bonus Task: Scatter Plot Chart for Sepal Length and Sepal Width