# 

**Correlation Analysis**

Name \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_ Period: \_\_\_\_\_\_

**For each activity below:**

✔️ means to complete this task

✏️ means to write an answer here

# **What is Correlation Analysis?**

Correlation analysis is a method in statistics that helps us understand how two or more variables are related to each other linearly. It tells us if the variables tend to go up and down together, go in opposite directions, or don't have any clear connection. Correlation can help determine if there is a pattern relating to changes in the variables.

For example, we can use correlation analysis to see if there's a connection between a person's height and weight. If we find a strong correlation, it means that as a person's height increases, their weight also tends to increase.

While correlation measures the strength and direction of a statistical relationship between variables, it does not provide insight into the underlying mechanisms that might be driving the relationship, and other outside factors could be responsible for both variables changing. Therefore, **a correlation between two variables does not necessarily indicate a causal relationship**.

# 

# **The Iris Dataset**

For this activity, we will use the Iris dataset that is available in Posit Cloud. It contains data on various characteristics of iris flowers collected and observed in 1935. The iris dataset consists of 150 observations, representing three different species of iris flowers: *setosa*, *versicolor*, and *virginica*. For each observation, there are 5 variables with the first 4 being numerical and the last one being categorical:

1. Sepal Length (in centimeters)

2. Sepal Width (in centimeters)

3. Petal Length (in centimeters)

4. Petal Width (in centimeters)

5. Species

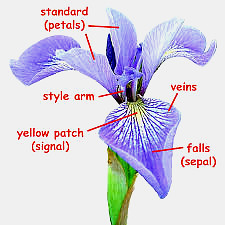


Image courtesy Flickr, Used under CC-BY-NC-2.0 license

# 

# **What are you trying to assess from the Iris dataset?**

Correlation analysis on the Iris dataset will allow us to uncover the relationships between the various measurements (features) of iris flowers. By calculating the correlation coefficients between sepal length, sepal width, petal length, and petal width, we can determine the strength and direction of the associations between these characteristics. A positive correlation coefficient close to 1 would indicate that as one measurement increases, the other tends to increase as well. Conversely, a negative correlation coefficient close to -1 suggests that as one measurement increases, the other tends to decrease. A correlation coefficient close to 0 would indicate a weak or negligible linear relationship between the variables, suggesting that variation in one variable is unrelated to variation in the other.

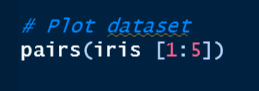
# **Activity #1: Visualizing Correlation in the Iris Dataset**

1. ✔️ Log into your Posit Account and create a new Posit Cloud Project. Name it “Iris\_Activity\_Code” on the top left. Click “Ctrl + S” to save the file into the “Iris Posit Cloud” folder you just created.
2. ✔️ Access the Iris Dataset from the Posit Cloud.

* Since this is a built-in dataset In Posit Cloud, the way to load it will be very different
* In your code, type **iris** (make sure all letters of “iris” are lowercase)

# 

1. ✔️ Take a look at the dataset. You can also look at the dataset by typing and running the **view(iris)** command. If you want to see the names of the variables, type the **names(iris)** command
2. ✏️ Does anything catch your eye when viewing the dataset? What variables appear to be correlated? Why?
3. ✔️ From the last lesson, you learned how to create scatter plots to explore relationships between two variables and detect outliers. But you could only display one scatter plot at a time. What if we want to see all possible scatter plots of the dataset to save time? To do this, type the code below



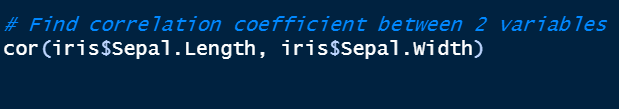
* The pairs() function generates a matrix of scatter plots, showing the relationships between all pairs of variables in the dataset.

1. ✏️ View the plots. Do any of the plots show a pattern? Do any of the plots appear to show a linear relationship?

# **Activity #2: Performing Correlation Analysis with the Iris Dataset**

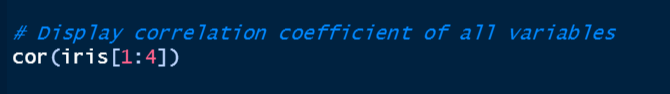
You created scatter plots and saw that some points may follow a pattern. However, you are not sure how correlated the variables of the Iris dataset are. To validate your observations, you can execute specific commands to obtain both the correlation coefficient and the outcomes of the correlation test.

1. ✔️ To find the linear correlation coefficient between 2 numerical variables on Posit Cloud, type the code in this format



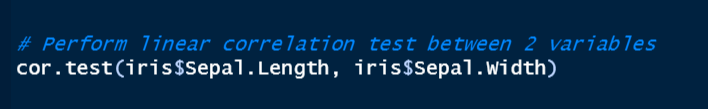
* The cor() function is used to find the correlation coefficient
* Remember what the “$” sign does in the function from your previous activity. If you cannot remember, ask your teacher for help

1. ✏️ Run the code above. What was the correlation coefficient? Was it closer to 0 or 1?
2. ✏️ Try this with a different combination of variables. What was the correlation coefficient? Was it closer to 0 or 1?
3. ✔️ To display all the linear correlation coefficients between 2 numerical variables on Posit Cloud, type the code in this format



* [1:4] is used to select columns of the dataset to use. Here we selected columns #1 - 4 because they are the only columns that display numerical variables.

1. ✏️ Paste the table of all the correlation coefficients below.
2. ✏️ Which two variables have the highest correlation coefficient? What does this mean? Note: ignore correlation coefficients between the variable and itself, That will always be 1.
3. ✏️ Are there 2 variables that have a correlation coefficient close to 0? What does this mean?
4. ✏️ Which 2 variables have the most negative correlation coefficient? What does this mean?
5. ✔️ To perform a **correlation test** that gives us more information on the correlation between 2 numerical variables on Posit Cloud, type the code in this format



* Correlation Coefficient - This number tells us how strong the linear direction between the two variables is. The value can be between -1 and 1. If it's close to 1, it means they have a strong positive relationship (both increase together), and if it's close to -1, it means they have a strong negative relationship (one increases while the other decreases). If the value is around 0, it means there's no significant relationship between them.
* P-value - This number tells us if the relationship we found is reliable or just occurring by chance. A low p-value (usually below 0.05) means the relationship is statistically significant and not likely to be due to random chance. A high p-value suggests that the relationship may not be reproducible (if you chose a different sample of iris flowers, for instance) or trustworthy.
* T-value - This value is used to calculate the P-value above. For this class, it is not important that you understand the details of how this is done. Be careful not to confuse the T-value with the P-value or the Correlation Coefficient.

1. ✏️ What was the correlation coefficient and p-value between Sepal Length and Sepal Width? Is the correlation significant?
2. ✏️ Try this with a different combination of variables. What variables did you perform your correlation test on? What were the correlation coefficient(s) and p-value(s)? Is the correlation significant?

# **Activity #3: Performing Correlation Analysis by species.**

Now that you have analyzed the entire dataset, let's say you want to look for correlation within each species. To start, let’s display the data points on a scatter plot but color them so that we identify which points are which species.

1. ✔️ Use this code to create scatter plots with the data points colored by species:



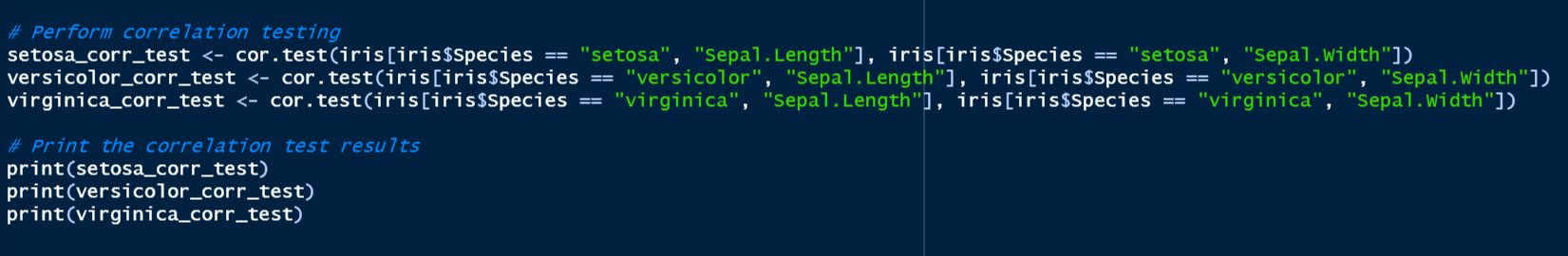
* create an object "species\_colors" that maps each species (setosa, versicolor, and virginica) to a specific color (red, blue, and green, respectively).

1. ✏️ View the plots. Do any of them follow a pattern? Does the relationship appear to be stronger for any species?
2. ✔️ Start correlation analysis by finding the correlation coefficient between sepal length and sepal width for each species. To do this, type in the code below:



* Create variables to hold the correlation coefficients for each species and run the code to get the correlation coefficients.

1. ✏️ What were the correlation coefficients of the sepal length and sepal width for each species?
2. ✏️ Which species has the highest correlation coefficient? Which species has the lowest correlation coefficient?
3. ✔️ Using sepal length and sepal width for each species, perform a correlation test using this code:



1. ✏️ What were the correlation coefficients and p-values of the different species' sepal length and sepal width? Can you tell if the relationship you found is reliable or just occurring by chance?
2. ✏️ How do your results for individual species compare with the initial results looking at all the species together as a group? Can you suggest a reason for the pattern that you see?
3. ✔️ Redo steps #2 and #3 of Activity #3 but use different variables instead of sepal length and sepal width.
4. ✏️ Which species of iris flowers shows the strongest correlation relationship between the two variables you chose? What does this strong positive correlation mean?
5. ✏️ Which species of iris flowers show the most negative correlation relationship between the two variables you chose? What does this strong negative correlation mean for the iris flowers?
6. ✏️ Are there any species of iris flowers that show a weak or no significant relationship between the two variables you chose? Can you explain why some measurements might not be strongly related to each other?
7. ✏️ Are there any correlations that surprise you or go against your initial expectations? Can you think of possible reasons why these unexpected correlations might exist?
8. ✏️ Do the scatter plots and correlations provide any insights into how the different measurements of iris flowers are related to each other and contribute to their overall characteristics?