Bivariate Analysis

Bivariate Analysis

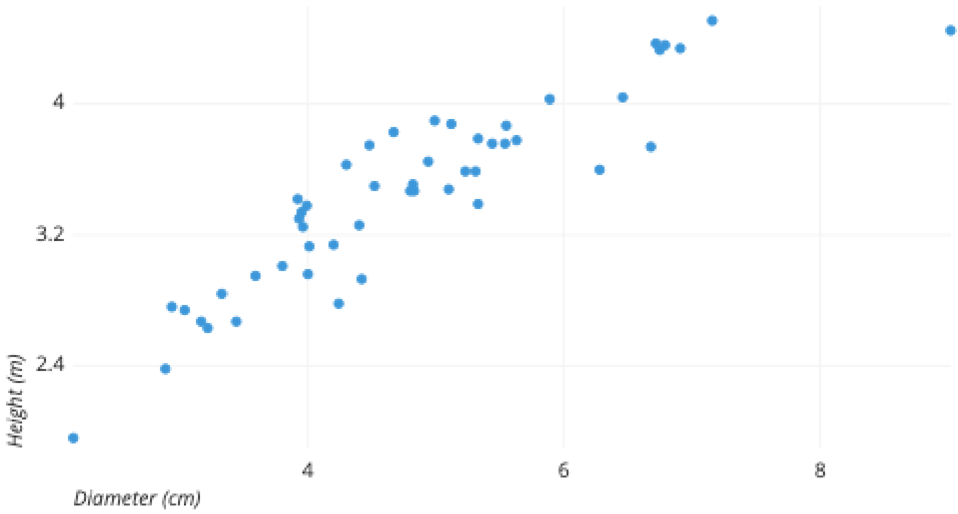
This type of analysis involves identifying relationship between two variables, one dependent variable and one independent variable. The main aspect is to find how a change in independent variable influence the dependent variable. We can use a scatter plot to find the relationship between two variable. Regression analysis is one of the commonly used ML technic to find a function that can give the relationship between these two variables.

Types of bivariate analysis

Many kinds of bivariate analysis can be used to determine how two variables are related. Here are some of the most common types.

1. **Scatterplots**

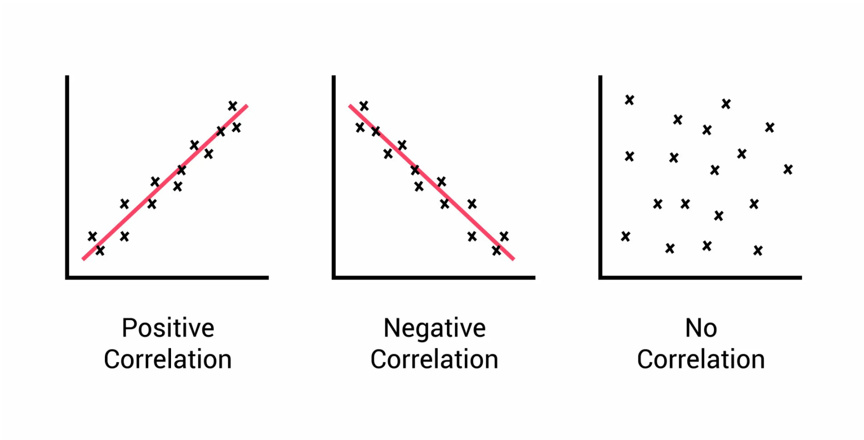
A scatterplot is a graph that shows how two variables are related to each other. It shows the values of one variable on the x-axis and the values of the other variable on the y-axis.



The pattern shows what kind of relationship there is between the two variables and how strong it is.

2. **Correlation**

Correlation is a statistical measure that shows how strong and in what direction two variables are linked. Usually the correlation is calculated using the Pearson's correlation coefficient.



Positive Correlation

Negative Correlation

No Correlation

A positive correlation means that when one variable goes up, so does the other. A negative correlation shows that when one variable goes up, the other one goes down.

Pearson's Correlation - Pearson's Correlation measures the strength of the relationship between two features. The scores are then ranked to identify the significant feature. The score lies between -1 and 1. Pearson's Coefficient is calculated using the below formula:

$$r = \frac{\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum \left(x_i - \bar{x}\right)^2 \sum \left(y_i - \bar{y}\right)^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

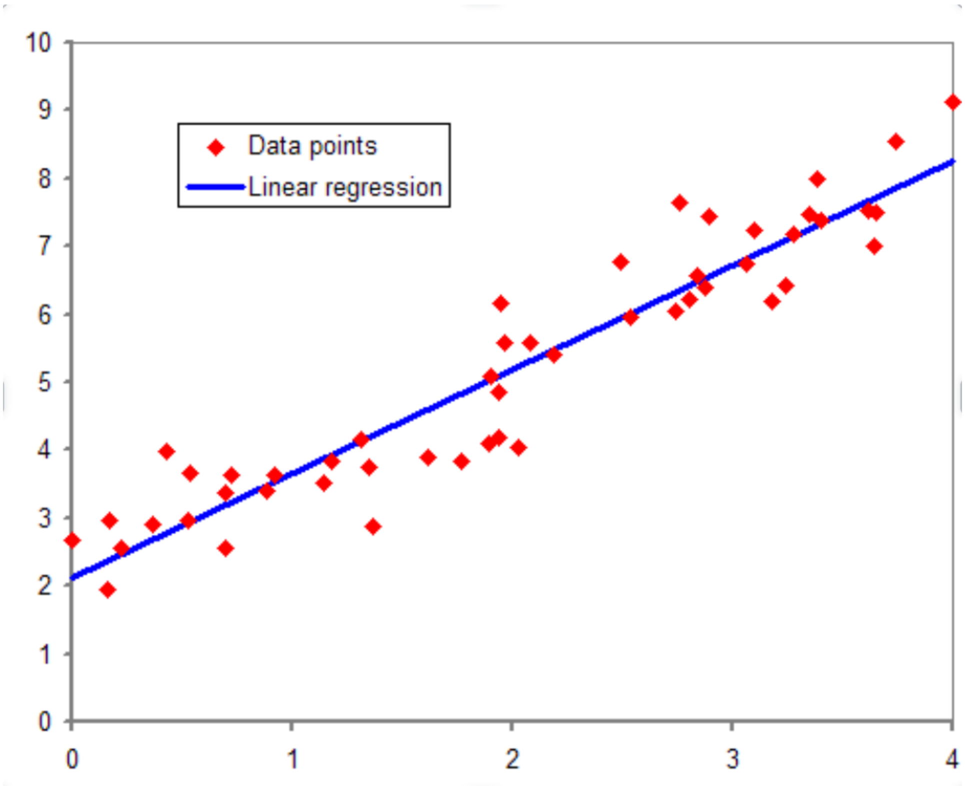
$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

3. **Regression**

This kind of analysis gives you access to all terms for various instruments that can be used to identify potential relationships between your data points. This function would help to predict one variable on the availability of the other.



The equation for that curve or line can also be provided to you using regression analysis. Additionally, it may show you the correlation coefficient. The most commonly used types of regression analysis are the linear regression and logistic regression.

4. **Chi-square test**

The **chi-square test** is a statistical method for identifying disparities in one or more categories between what was expected and what was observed. The test's primary premise is to assess the actual data values to see what would be expected if the null hypothesis was valid.The features that are to be compared for correlation are drawn as a pivot table and the values of the pivot table denotes the corresponding frequencies. Then the sum of each of the class of feature are evaluated. Now the expected value is the values based on the totals. Each cell is filled with the expected value. Now the following formula is used to to Chi-square value,

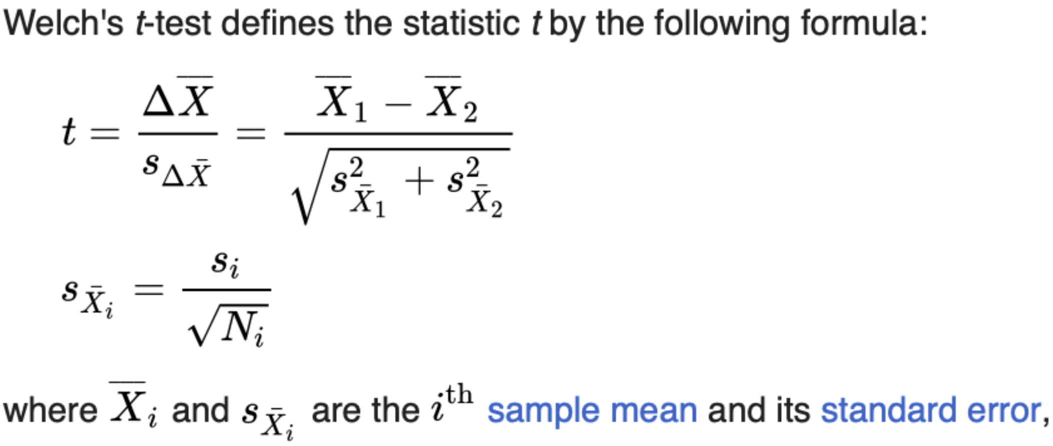
$$X^2 = \sum \frac{(\text{Observed value - Expected value})^2}{\text{Expected value}}$$

Higher the Chi-square value, higher is the deviation, which means that these features are independent of each other

Researchers use this statistical test to compare categorical variables within the same sample group. It also helps to validate or offer context for frequency counts.

5. **T-test**

A t-test is a statistical test that compares the means of two groups to see if they have a big difference. This analysis is appropriate when comparing the averages of two categories of a categorical variable.This test is used to identify the deviation between two features which has unequal sample size and unequal variances. The following is the formula,



Welch's *t*-test defines the statistic *t* by the following formula:

$$t = \frac{\Delta\overline{X}}{s_{\Delta\bar{X}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_{\bar{X}_1}^2 + s_{\bar{X}_2}^2}}$$

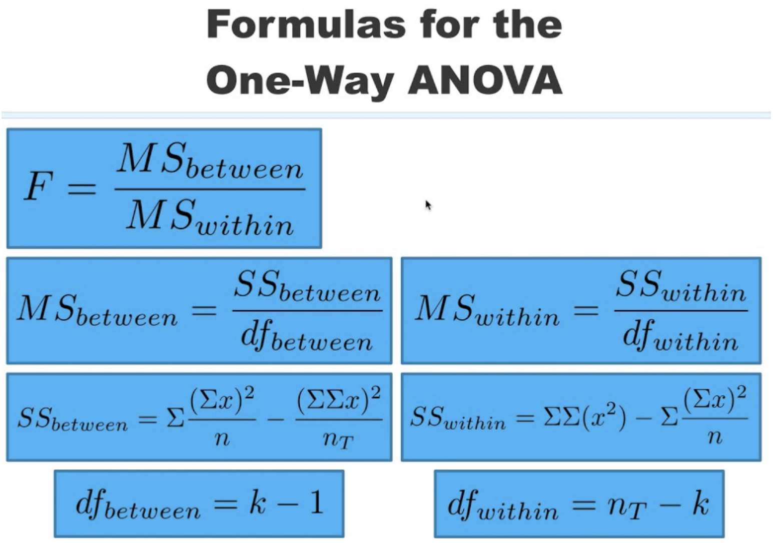
$$s_{\bar{X}_i} = \frac{s_i}{\sqrt{N_i}}$$

where $\overline{X}_i$ and $s_{\bar{X}_i}$ are the i^{th} sample mean and its standard error,

Here si is the standard deviation and Ni is the sample size of that feature

6. **ANOVA (Analysis of Variance)**

The ANOVA test determines whether the averages of more than two groups differ from one another statistically. This comparison of averages of a numerical variable for more than two categories of a categorical variable is appropriate.



Formulas for the One-Way ANOVA

$$F = \frac{MS_{between}}{MS_{within}}$$

$$MS_{between} = \frac{SS_{between}}{df_{between}}$$

$$MS_{within} = \frac{SS_{within}}{df_{within}}$$

$$SS_{between} = \Sigma\frac{(\Sigma x)^2}{n} - \frac{(\Sigma\Sigma x)^2}{n_T}$$

$$SS_{within} = \Sigma\Sigma(x^2) - \Sigma\frac{(\Sigma x)^2}{n}$$

$$df_{between} = k - 1$$

$$df_{within} = n_T - k$$

ANOVA test is of three types:

One-way - Variance between the groups of one independent variable.

Two-way - Variance between the groups of two independent variable

N-way - Variance between the groups of N independent variable