

**2020-1-HR01-KA226-HE-094713**

**O3 - Distance learning curricula in Machine Learning**

**Bridging Statistics & Machine Learning**

October 2022

**2020-1-HR01-KA226-HE-094713**

**CODEIN**

Cloud cOmputing for Digital Education INnovation

|  |  |
| --- | --- |
| Work Package: | Intellectual Outputs |
| Product/Deliverable: | O3 - Distance learning curricula in Machine Learning |

|  |  |  |  |
| --- | --- | --- | --- |
| Version: | 1 | Date: | October, 2022 |
| Type: | e-book | | |
| Distribution: | Project Partners | | |
| Responsible Partner: | Polytechnic of Šibenik | | |
| Author: | All Partners | | |
| Contributors: | All Partners | | |
| Approved by: | Quality Assurance Team | Date: | 30/11/2022 |

**Identification Sheet**

|  |  |
| --- | --- |
| **Project Code** | **2020-1-HR01-KA226-HE-094713** |
| **Project Acronym** | **CODEIN** |
| **Project Full Title** | **Cloud cOmputing for Digital Education INnovation** |

|  |  |
| --- | --- |
| **Keywords** | Machine Learning, Data Analysis, Descriptive Statistics, Hypothesis Testing, Regression Analysis, Classification |
| **Abstract** | The purpose of this document is to combine traditional statistical analysis methods with modern machine learning techniques in a distance learning setting. The content is designed to make statistical concepts more understandable and to show their application in machine learning, with the goal of providing learners with the skills necessary for data analysis and predictive modeling. |

|  |  |
| --- | --- |
| **Disclaimer** | This project has been funded with support from the European Commission. This publication reflects the author's views only, and the Commission cannot be held responsible for any use that may be made of the information contained therein. |

**Contents**

[Introduction 6](#_Toc158052670)

[The Simplicity of Data - Descriptive Statistics Unveiled 6](#_Toc158052671)

[Measures of Central Tendency: The "Heart" of Your Data 7](#_Toc158052672)

[Mean 7](#_Toc158052673)

[Median 7](#_Toc158052674)

[Mode 8](#_Toc158052675)

[Measures of Dispersion 9](#_Toc158052676)

[Range 9](#_Toc158052677)

[Variance 9](#_Toc158052678)

[Standard deviation 10](#_Toc158052679)

[Standard Error 11](#_Toc158052680)

[Histograms and Box Plots 11](#_Toc158052681)

[Hypothesis Testing 14](#_Toc158052682)

[Z – test 16](#_Toc158052683)

[T-test 17](#_Toc158052684)

[Chi-Square test 18](#_Toc158052685)

[Regression analysis 21](#_Toc158052686)

[Linear Regression 22](#_Toc158052687)

[Multiple Linear Regression 23](#_Toc158052688)

[Performing Multiple Linear Regression in Excel 23](#_Toc158052689)

[Classification and its applications 26](#_Toc158052690)

[Logistic Regression 27](#_Toc158052691)

[References 30](#_Toc158052692)

## Introduction

Ever feel lost in the sea of data tools out there? Every website you visit offers a magic wand to sort your numbers out. However, hear this: you do not always need to dive into heavy-duty coding to get the gist of your data. Often, a few neat tricks on good old Excel can take you a long way.

Choosing the best tool can be a head-scratcher with all the chatter. We are here to ease that headache. Our aim? To hand you the tools that help you make sense of information, plain and simple. We are all about helping folks make wiser decisions by getting good with data.

Think of statistics as the first rung on the ladder. It's the nuts and bolts of making sense of all those numbers. Just like you cannot build a house without a solid foundation, you cannot get good with data without a grip on the basics.

By marrying the old-school smarts of statistics with our shiny, new tech, we create a killer combo for slicing and dicing data. It's like getting the hang of a game's basics to crush it later. Stick to the simple stuff first, and you will set yourself up for the big league of machine learning.

## The Simplicity of Data - Descriptive Statistics Unveiled

Descriptive statistics make information more comprehensible using simplified data summaries, which bring out the main points of the data immediately. These statistics aim to give a quick snapshot of the main features of the data so that anyone can quickly grasp the most relevant information.

Central tendency measures and dispersion measurements are some statistical tools used in descriptive statistical analysis. Mean, median and mode are examples of central tendency measures. They help us locate an average or a typical case in discussions.

Dispersion tools like range, variance, and standard deviation tell us how widely scattered data values are from each other. The range shows us the difference between the smallest and largest values, while variance and standard deviation show us how far away data deviates from its average value.

Combining these tools helps to effectively communicate the narrative the data conveys, enabling decisions and conclusions to be drawn from a straightforward, informed perspective. Descriptive statistics are essential for anyone looking to understand or convey information through data, whether in academics, business, policy-making, or everyday life.

## Measures of Central Tendency: The "Heart" of Your Data

In the story of your data set, the measures of central tendency are the protagonists, leading you to the central theme around which all other data points revolve. This section unfolds the tale of three such measures: the mean, median, and mode. Understanding these measures matters because they give you a quick and clear summary of where most of your data points cluster. They cut through the noise and offer a glimpse into the big picture without getting bogged down by every detail. By mastering the mean, median, and mode, you can swiftly identify the central point around which your data pivots, providing a starting point for further analysis or a superficial understanding for rapid communication. This knowledge is indispensable for anyone who needs to make sense of numbers, from researchers and analysts to students and everyday data enthusiasts.

### Mean

The **mean** is one of the most fundamental concepts in the world of statistics, often called the **arithmetic average**. It represents the "balance point" of a dataset, providing a single value that summarizes the entire set. To calculate the mean, you take the sum of all the values in your data and divide by the count of the data points.

For example, let's consider a simple dataset: 2, 4, 4, 6, and 8. These numbers are the individual data points. The mean is (2 + 4 + 4 + 6 + 8) divided by 5, which equals 4,8. This tells you that if the data were perfectly balanced, each value would be 4,8. Therefore, the mean is particularly useful for understanding the central tendency of data when the dataset is symmetrical and free of outliers, as outliers can skew the mean significantly.

### Median

The **median** is a form of measure of central tendency that is used in statistics and probability theory. It represents the value separating the higher half from the lower half of a data sample, a population, or a probability distribution. Unlike the mean, which provides an average of all values, the median offers a middle point in an ordered set of numbers, representing the typical value in cases where the data set includes outliers that would skew the mean.

When the number of data points is odd, the median is the number that falls in the middle of the set when ordered from the lowest to the highest. For example, in an ordered set like {1, 3, 5}, the median is 3, since it is the second number in the sequence with one number before and after it.

On the other hand, when the dataset contains an even number of points, there is no single middle value. Instead, the median is calculated by taking the average of the two middle numbers. For example, in an ordered sequence of four numbers like {1, 3, 5, 7}, the middle numbers are 3 and 5. Their average, (3+5)/2, equals 4, which is the median of this set.

The median is a useful statistical tool because it is not affected by extreme values and provides a more accurate reflection of the dataset's central location, particularly in skewed distributions. It is widely used in fields like real estate to report house prices, in economics to determine the middle income, and in other areas of research where a central tendency measure is required that is resistant to skewness or outliers.

### Mode

The **mode** is a statistical term that refers to the most frequently occurring number found in a set of numbers. It represents the values with the highest frequency in a data set. This measure of central tendency is handy for categorical data, where numerical operations such as finding an average may not be meaningful.

The mode can reveal the most common category or value in a given dataset. For example, in a data set of the numbers {2, 4, 4, 6, 8}, 4 is the mode because it appears more often than any other number. A dataset may have one mode (unimodal), two modes (bimodal), or more than two modes (multimodal), and it is also possible for a dataset to have no mode at all if no number repeats.

The mode concept is easy to understand and can be quickly identified by observation. It's widely used in everyday contexts, such as determining the most common shoe size sold in a store, the most popular car color, or the most frequent birth month in a group of people. Because extremely high or low values do not influence it, it can better measure what is typical in cases where outliers may distort the average or median.

## Measures of Dispersion

Measures of dispersion are statistical tools that describe the spread of data points around a central value, such as the mean or median. These measures are essential in understanding the dataset's distribution and identifying how varied or dispersed the data is.

### Range

The **range** is indeed one of the most straightforward measures of variability to understand. It quickly estimates the spread of numbers in a data set. The range is calculated as the difference between the highest and lowest values in the data set. In your given example with data points 10, 15, 20, 25, and 30, the range is computed by subtracting the smallest number (10) from the most significant number (30), resulting in a range of 20.

This measure gives you a single number that represents the span of your data. It tells you how much variety there is in the values of your data set. However, it does not explain how the numbers in the data set are distributed between the lowest and highest values. It is most effective in conveying the extent of variability when the data set has no outliers since outliers can dramatically affect the range, making it less representative of the typical spread of values.

### Variance

**Variance** is a statistical measure that tells us how much the individual data points in a set differ from the average value of the set. It is a numerical value representing the degree of spread in a data set.

If you have gathered data from every group member you're interested in, then you should calculate the **population variance (σ²)** using the formula (1). This gives you a precise measure of variability for the entire group since nothing and no one is left out of your dataset.

**σ²**=Σ(*xi*−*μ*) ²/*N* (1)

The N is the number of data points in the entire population, and μ is the mean (average) of the entire population.

If you are analyzing a set of data that represents just a portion of a larger group, you should calculate the **sample variance (s²)** using formula (2). This helps to understand how your sample data is spread around the mean, considering that you're working with only a part of the whole.

**s²**=Σ(*xi*−*x*ˉ) ²/(*n*−1) (2)

The n is the number of data points in the sample, and x̄ (x-bar) is the mean (average) of the sample. The (n-1) in the denominator is known as Bessel's correction, and it corrects the bias in estimating the population variance from a sample.

To understand the sample and population variance calculations, let us use an example with the following data points: {10, 15, 20, 25, 30}.

First, we find the mean of these data points, which is the sum of the numbers divided by the count of the numbers. Adding them up gives us 100, and since there are 5 data points, we divide 100 by 5 to get a mean of 20.

Next, we calculate the squared difference for each data point from the mean, which is the data point minus the mean squared. Doing this for our data set, we get the following squared differences: 100, 25, 0, 25, and 100. Adding these squared differences up, we get a total of 250.

For population variance, we divide the sum of these squared differences by the total number of data points (N), which in our case is 5. This gives us a population variance (σ²) of 50.

For sample variance, we divide the sum of squared differences by the total number of data points minus one (n-1), which accounts for the degrees of freedom in a sample. For our 5 data points, this gives us a denominator of 4, resulting in a sample variance (s²) of 62.5.

In summary, with our data set, the population variance is 50, and the sample variance is higher at 62.5, reflecting the adjustment for using a sample to estimate the variance for a larger population.

### Standard deviation

Standard deviation is a statistical measure for determining how a group of values varies or spreads. It is widely used to describe how the numbers deviate from an average (mean). It just says if data points are closer to the mean or scattered over a wider range. There are, however, two slightly different formulas for calculating standard deviation depending on whether one deals with a whole population or only sample data. The population variance is divided by its square root to get σ, which is the population standard deviation. The sample variance is divided by its square root to obtain s –, which is the standard deviation in samples. This step allows us to understand variance and express it in terms of original units and variables to help simplify the analysis of initial data sets.

### Standard Error

The **standard Error (SE)** is a statistical term representing the precision with which a sample's mean estimates the population mean. It essentially indicates the reliability of the mean as an estimate of the expected value. The standard Error tells us the expected fluctuation in the sample means if we were to take multiple samples from the same population. A smaller SE suggests our sample mean is a good estimate of the population mean, while a larger SE indicates more variability and less certainty about our estimate.

The standard Error (3) formula is the sample's standard deviation divided by the sample size's square root (n).

Standard error (SE) = (3)

This calculation hinges on the principle that the larger the sample size, the smaller the standard Error, and consequently, the closer the sample mean is to the population mean.

Using the standard Error, statisticians can gauge the extent to which the average of any given sample will likely differ from the actual population average. This is crucial in hypothesis testing and constructing confidence intervals, providing a mathematical basis for making inferences about the population from the sample.

## Histograms and Box Plots

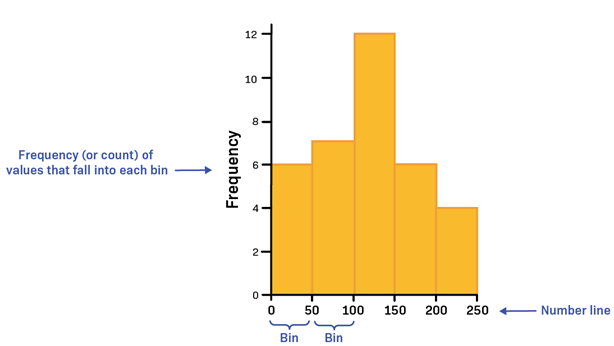
Visual tools like histograms and box plots are invaluable for understanding the distribution and spread of data. They turn complex data sets into visual stories that are easy to understand and provide clear insights into the critical features of the data.

**Histograms** are bar graphs that display the frequency of data within specific ranges or bins. They help reveal patterns that might not be obvious from looking at a list of numbers, such as the central tendency, variability, and any outliers or unusual observations.

Bins, or intervals in a histogram, act like designated meeting points where numbers gather. The height of each bin represents its popularity, showing how many numbers fall into each category.

Imagine a histogram as a snapshot of a party for data points. A tall bin is like a popular group at the party—lots of numbers huddled together, chatting away. It is crowded because many data points share the same range of values. A short bin, on the other hand, is like a quiet corner—fewer data points more breathing room.

Each bin groups numbers within a specific range and the height of the bars in a histogram visually conveys this concentration of numbers. So when you look at a histogram, you see where numbers in the data set have decided to 'hang out' about each other.

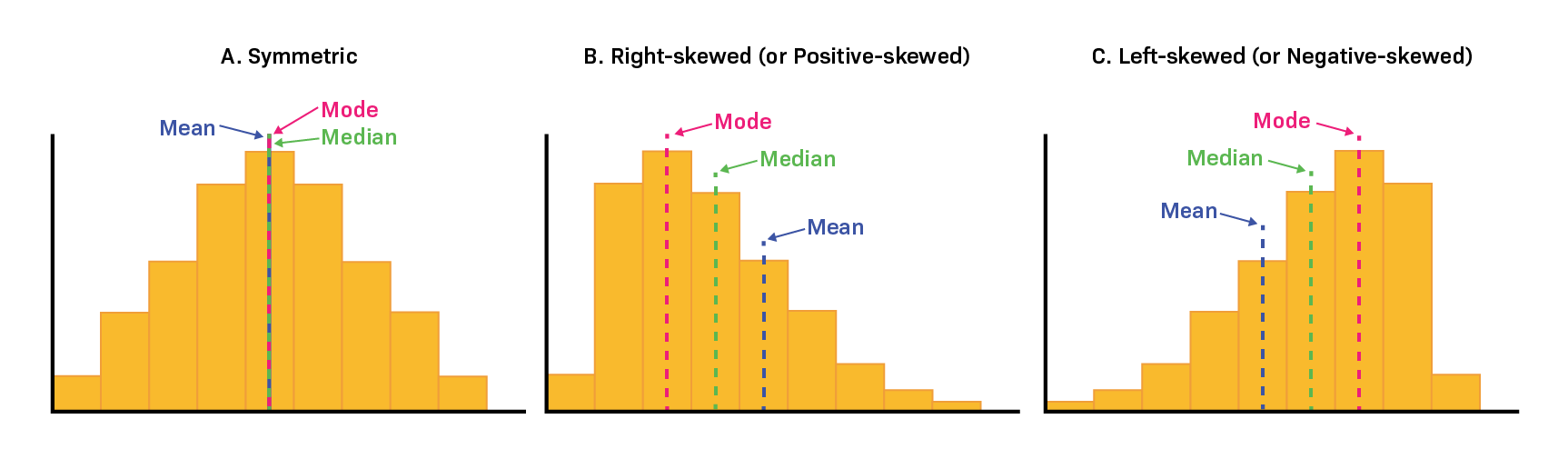


*Figure 1 Histogram plot*

Symmetry and skewness in a histogram reflect the balance and distribution of data values. A symmetric histogram is one where the left and right sides are mirror images of each other, much like a perfectly balanced seesaw. This indicates that the data is evenly distributed around the central value, with no particular direction being favored for spreading data points.

On the other hand, a skewed histogram tilts to one side, either left or right, indicating that the data is not evenly spread. If a histogram leans to the right, with a long tail extending in that direction, it is a right or positive skew. This suggests that there are several values significantly higher than the mode. Conversely, if the histogram stretches to the left, it exhibits a left or negative skew, indicating lower-than-typical values.

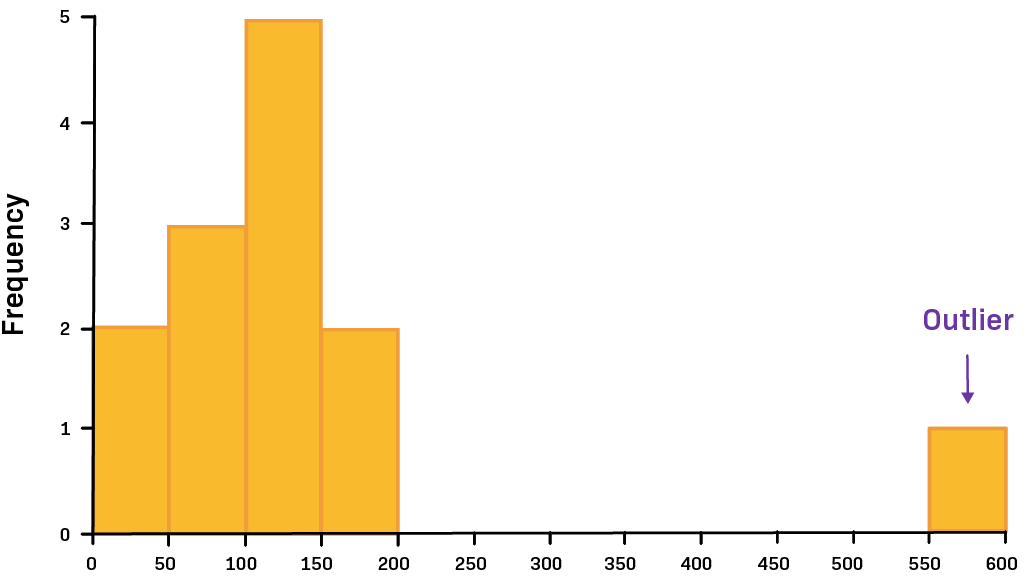
In essence, skewness tells us about the direction and relative strength of a dataset's deviation from the norm, providing insights into the nature of the underlying distribution.



*Figure 2 Symmetry and skewness in a histogram*

Outliers in a dataset are like the mavericks or the loners at the edges of a distribution. In a histogram, they appear as isolated bars that stand apart from the cluster of other bars, representing the bulk of the data. These bars are like distant outposts on a graph, indicating that the data points they represent significantly differ from most observations, which are grouped together in the central, more crowded bins.

These isolated data points can be exceptionally high or low in value compared to the rest of the data and can substantially impact the overall analysis. Outliers can affect the mean and standard deviation of the data, leading to potential misinterpretations if not adequately accounted for or understood. They are essential to identify because they may indicate variability in the data collection process, experimental errors, or novel phenomena worth investigating.

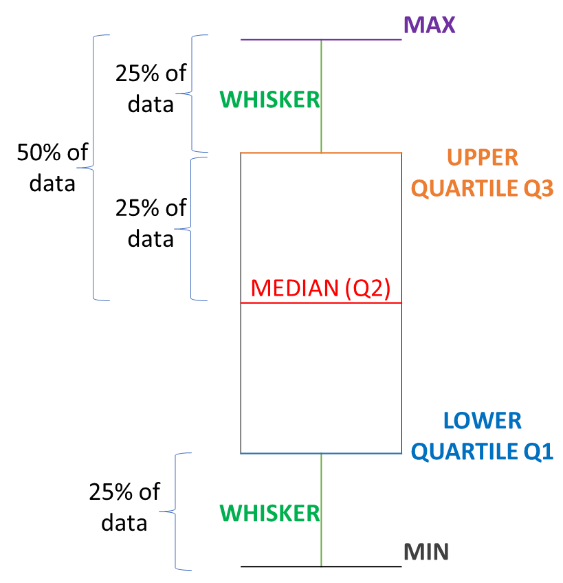


*Figure 3 Isolated data points (Outliers)*

**Box plots**, also known as box-and-whisker plots, summarize data from a five-number summary: minimum, first quartile (Q1), median, third quartile (Q3), and maximum. They provide a quick snapshot of the distribution's shape, central value, and variability.

It is an excellent visual tool for identifying outliers and understanding the spread and symmetry of the data. Here's how to read one (Figure 4):

* The section from the minimum value to Q1 (**first quartile**) represents the lowest 25% of the data. This part of the plot gives you a quick view of the bottom quarter of your data points.
* From Q1 to the median (**second quartile**), you find the next chunk of 25% of data values. This range, including Q1 and the median, encapsulates the lower half of the data set.
* The median is the middle point of the data and essentially cuts the dataset in half, with 50% of the data falling below it.
* Moving from the median to Q3, we encounter the **third quartile**, which covers the next quarter of the data, taking us to 75% of the way through the data set. This quartile demonstrates the middle to upper range of the data values.
* The stretch from Q3 to the maximum value, or the **fourth quartile**, contains the highest 25% of data points in the set, completing the data's picture.
* By dividing the data into these quartiles, a box plot can show the range and distribution of the data at a glance, from the lowest to the highest values, including any potential asymmetry or outliers that might affect the analysis.



*Figure 4 Box plot*

## Hypothesis Testing

Hypothesis testing is a fundamental method in statistics that is used to make inferences about populations based on sample data. It involves making an initial assumption or hypothesis about a population parameter and then using statistical analysis to determine whether there is enough evidence in the sample data to reject this initial hypothesis.

Here are real-life scenarios where hypothesis testing might be applied:

* *In education, a teacher hypothesizes that 60% of the students at their college come from lower-middle-class families. By testing this hypothesis, the teacher can validate or challenge assumptions about the socioeconomic status of the student body, which may influence educational strategies and support services.*
* *In healthcare, a doctor might hypothesize that a '3D approach' consisting of Diet, Dose, and Discipline effectively manages diabetes in 90% of cases. Hypothesis testing can be used to evaluate the effectiveness of this treatment regimen, potentially influencing medical practice and patient care guidelines.*

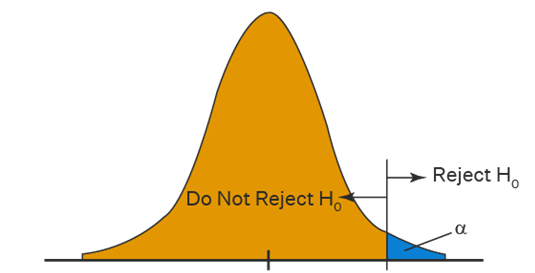
Hypothesis testing typically involves contrasting two opposing theories: the null hypothesis and the alternative hypothesis.

The null hypothesis (H₀) posits that there is no significant effect or difference. For example, it might suggest that a new drug has no effect on a disease or that there is no difference between the means of two different populations. It represents a statement of "no change" or "no difference" that the test seeks to challenge.

The alternative hypothesis (H₁) proposes that there is an effect or a difference. It contradicts H₀ and is what researchers hope to support, indicating, for instance, that a new teaching method is better than the traditional one or that a policy change has had a positive impact.

The data collection and analysis phase is where the evidence to test these hypotheses is gathered and scrutinized. Researchers collect data in a way that is consistent with the assumptions of their test and then apply statistical analyses to determine if the observed data falls within a range that would be considered normal under the null hypothesis.

Decision-making in hypothesis testing (Figure 5) is fundamentally about assessing whether the data collected provides sufficient evidence to reject H₀ in favor of H₁. This decision is often based on calculating a p-value, which reflects the probability of obtaining the observed results or more extreme ones if H₀ were true. If the p-value is less than the pre-determined significance level α (alpha), typically set at 0.05, the null hypothesis is rejected, suggesting that the alternative hypothesis may be true.



*Figure 5 Hypothesis testing*

Through the following six steps, researchers can make informed decisions about the validity of their hypotheses backed by statistical evidence.

1. **Formulate Hypotheses**: Begin by stating the null hypothesis (H₀), a statement of no effect or no difference, and the alternative hypothesis (H₁), which suggests a potential effect or difference.
2. **Choose Significance Level (Alpha)**: Select an alpha level, such as 0.05, 0.01, or 0.10, determining the probability threshold for rejecting the null hypothesis.
3. **Collect and Analyze Data**: Gather the necessary data through experiments, surveys, or other means and perform a preliminary analysis to understand its characteristics and patterns.
4. **Calculate Test Statistic and P-Value**: Use the data to calculate a test statistic, which, in turn, helps you calculate the p-value—the probability of observing the test results under the assumption that the null hypothesis is true.
5. **Make a Decision**: Compare the p-value with the chosen alpha level to decide whether to reject or not reject the null hypothesis.
6. **Draw a Conclusion**: Based on the decision, conclude whether the evidence supports the alternative hypothesis or if the null hypothesis stands.

Choosing the proper statistical test can feel overwhelming, but it is all about matching the test to your research question. A Z-test might be your go-to if you're looking at the difference between two group means under normal conditions. A T-test is more appropriate for comparing means without that normality or with smaller samples. Moreover, when you are curious about the relationship between categorical variables, the Chi-square test is the one you would consider. Each test serves its purpose based on the data and the hypothesis you are exploring.

### Z – test

The Z-test is a statistical test that answers the question: *"How unusual is our sample?"* By calculating a Z-score, we determine the number of standard deviations our sample mean (**X̄)** is from the population mean (**μ**). This score incorporates the variability of the data (σ) and the number of observations (**n**). We convert this Z-score into a probability (**p-value**) using statistical tables or software tools to determine whether the result is statistically significant.

For instance, if we want to test whether the average birth weight of babies in Zagreb deviates from the Croatian average, we set this as our null hypothesis. We then collect data on Zagreb birth weights and apply the formula:

Z = / (σ/) (4)

This formula plugs in our sample mean, the known population mean, the standard deviation, and the size of our sample to compute the Z-score. A high absolute value of the Z-score (usually greater than 1.96 or less than -1.96) may indicate that the difference is significant, leading us to reconsider the null hypothesis.

The Z-test formula is not just for determining how extreme a sample mean is but also a powerful tool for planning studies. When researchers design an experiment or survey and want to ensure their findings are precise, they use the Z-test formula to calculate the necessary sample size. This calculation helps to achieve the desired confidence level and margin of error before collecting any data.

For estimating a population proportion, the sample size formula is:

n = Z² p (1-p)/ E² (5)

Here, '**n**' is the sample size you aim for to be confident about your results. '**Z**' corresponds to your chosen confidence level (like 1.96 for 95% confidence). The '**p**' is the proportion of the population you expect to exhibit the feature or outcome you study. If you are unsure what '**p**' to use, 0.5 is a conservative estimate that maximizes the required sample size. The '**E**' is your acceptable margin of Error (such as 0.05 for a 5% margin). Using this formula ensures that your study has enough power to detect an actual effect or difference when it exists.

### T-test

The T-test is ideal for comparing the means of two small sample groups to see if they have a statistically significant difference. It is beneficial when dealing with populations where the variance is unknown, the sample size is small, and the data follows a normal distribution.

The formula for a T-test when you're comparing two independent samples (like '*different fertilizers' effects on tomato yield'*) is given by:

​ t = (6)

This equation measures the difference between the two sample means (**X̄1** and **X̄2**) relative to the variability observed in the samples, standardized by the sample sizes (n1 and n2).

Here, and represent the variances of each sample, a measure of dispersion in the data.

Researchers convert the calculated t-value to a p-value using statistical software or tables to conclude the T-test. This p-value helps determine whether the observed differences in sample means could be due to chance. For instance, Farmer Joe can use the T-test to validate his organic fertilizer claim scientifically. If the p-value is low enough (typically less than 0.05), it suggests that the difference in tomato yields is statistically significant, supporting Joe's hypothesis that organic fertilizer leads to better yields than traditional fertilizer.

### Chi-Square test

The Chi-Square test is a standard statistical method used to assess whether there is a significant association between two categorical variables. It is beneficial in research when the objective is to compare the observed frequency of events against the expected frequency under the assumption that no relationship exists between the variables.

Here is the Chi-Square test formula:

*χ*2=∑ (7)

The **O** represents the **Observed Frequency**, which is the count of events or characteristics seen in the data. The **E** stands for the **Expected Frequency**, the count of events you would expect to find if the null hypothesis is true, meaning there is no association between the variables.

An example hypothesis that can be tested using the Chi-Square test is: "*There is an association between smoking status and the presence of disease*." This hypothesis posits that smoking could have a relationship with the incidence of a particular disease. Researchers collect data on the smoking status (like in Table 1) and the presence or absence of the disease. The Chi-Square test can then analyze the observed counts against the counts that would be expected if smoking did not affect the disease's presence to determine if there is a statistically significant association.

Expected frequencies

Observed frequencies

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Has Disease** | **Doesn't Have Disease** | **Row Totals** |
| **Smoker** | 70 | 30 | **100** |
| **Non-Smoker** | 40 | 60 | **100** |
| **Column Totals** | 110 | 90 | **200** |
| **Expected total - Smoker** | **55** | **45** |  |
| **Expected total - Non-Smoker** | **55** | **45** |  |

*Table 1 Chi-Square contingency table*

The expected frequency for each cell in a contingency table can be calculated using the formula:

Expected=(Row Total×Column Total) / Grand Total (8)

The row totals are the total number of smokers and non-smokers (100 each), the column totals are the total number of individuals with and without the disease (110 and 90, respectively), and the grand total is the sum of all individuals surveyed (200).

Chi-Square for smokers with disease: *χ*2=≈ 4,09

Chi-Square for smokers without disease: *χ*2=≈ 5,00

Chi-Square for non-smokers with disease: *χ*2=≈ 4,09

Chi-Square for non-smokers without disease: *χ*2=≈ 5,00

Adding these values together gives the total Chi-Square statistic:

*χ*2 =4.09+5.00+4.09+5.00=18.18

This total Chi-Square statistic can then be used to determine if the observed frequencies differ significantly from the expected frequencies, indicating an association between smoking status and the presence of the disease.

To conclude from the Chi-Square statistic, you would compare the calculated Chi-Square value to a critical value from the Chi-Square distribution table at a chosen significance level (usually 0.05 for a 95% confidence level) and with the appropriate degrees of freedom.

For the Chi-Square test, the degrees of freedom are calculated as:

Degrees of freedom=(number of rows−1)×(number of columns−1) (9)

Given that you have two rows ("*smoker* "and "*non-smoker* ") and two columns ("*has the disease* "and "*does not have the disease* "), the degrees of freedom (df) for your test would be: df=(2−1)×(2−1)=1

Using a standard Chi-Square distribution table, you can find the critical value for 1 degree of freedom at the 0.05 significance level, typically 3.841.

Therefore, since your Chi-Square statistic (18.18) is substantially more significant than the critical value (3,841), you can conclude that there is a statistically significant association between smoking status and the presence of the disease in your data. This means that the difference between the observed and expected frequencies is not due to chance alone, and there is likely a real relationship between smoking and disease presence in the population from which your sample was drawn.

## Regression analysis

Observed frequencies

Regression analysis is a powerful statistical tool used to model and analyze the relationships between a dependent variable and one or more independent variables. A regression analysis's key components and concepts are **dependent** and **independent variables** (s).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age** | **Has Job** | **Owns House** | **Credit Rating** | **Approved?** |
| Young | No | No | Fair | No |
| Young | No | No | Good | No |
| Young | Yes | No | Good | Yes |
| Middle | No | No | Good | No |
| Middle | Yes | Yes | Excellent | Yes |
| Old | Yes | No | Good | Yes |
| Old | No | No | Fair | No |

**Independent Variables (X)**

**Dependent (Y) Variables (X)**

**Labels**

*Table 2 Regression analysis example - Loan approval criteria analysis*

The dependent variable (Y) is the variable you try to predict or explain. It is dependent on the independent variables. For example, the dependent variable is whether a loan is approved, which can be quantified as a binary outcome (0 for not approved, 1 for approved).

Independent variables (s) (X) are the variables that influence the dependent variable. They are called independent because the dependent variable does not influence them. For example, from Table 2, independent variables could include "*age* "(numerical), "*employment status* "(categorical: "*has a job* "or "*has no job* "), and homeownership status (categorical: owns a house or does not own a house).

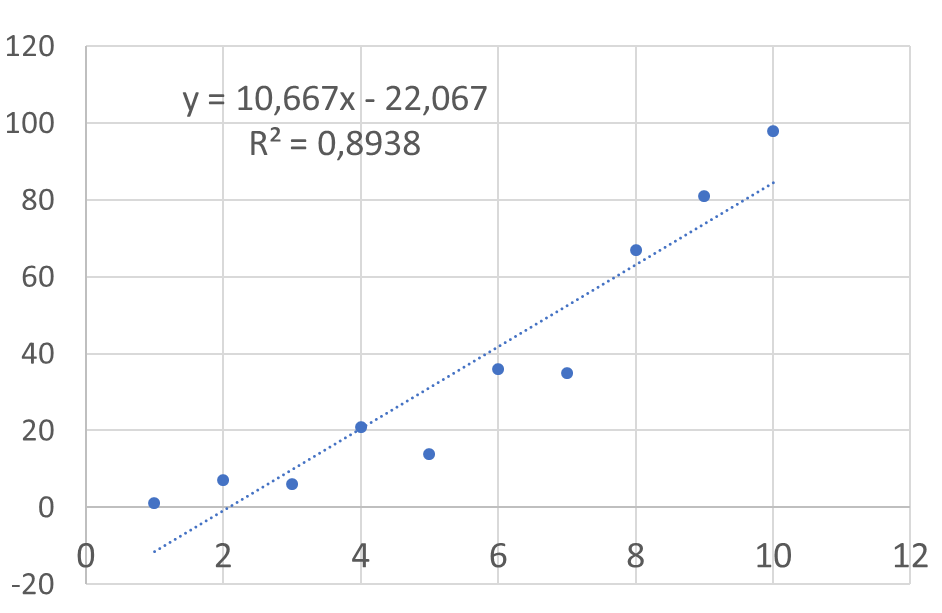
Two primary types of regression analysis are used to predict outcomes: simple linear Regression, which uses one independent variable to forecast the dependent variable's changes, and multiple linear Regression, which involves multiple independent variables for more complex insights.

### Linear Regression

Linear Regression is a valuable technique for analyzing and predicting the relationship between two variables and can be applied in a wide range of fields, from finance and economics to social sciences and engineering.

The goal of linear Regression is to find the best linear relationship between X and Y, which can be used to predict the value of Y for a given value of X. We can use linear Regression to find the equation of this regression line, which is of the form Y = b0 + b1\*X, where b0 is the intercept and b1 is the slope of the line. This equation is used to predict the value of Y for any given value of X.

Let us look at the scatter plot in Figure 6 to understand how linear regression works. This scatter plot has data points representing the relationship between two variables, X and Y. The blue line in the plot represents the best linear fit of the data, also known as the regression line.



*Figure 6 Linear Regression example*

R-squared (R²) is a statistical measure that represents how well the linear regression model fits the data.

In simpler terms, R² tells us how much of the dependent variable (Y) variation can be explained by the independent variable (X) in a linear regression model. A high R² value (close to 1) means that the model explains a large proportion of the variation in the dependent variable, while a low R² value (close to 0) means that the model does not explain much of the variation in the dependent variable.

For example, if the R² value is 0.89, 89% of the variation in the dependent variable can be explained by the independent variable in the linear regression model.

R² is a helpful tool for evaluating the performance of a linear regression model and comparing the performance of different models. However, it is essential to note that R² is just one measure of model performance and should be used with other metrics and visualizations to evaluate the model entirely.

### Multiple Linear Regression

Multiple Linear Regression (MLR) is a cornerstone statistical method for interpreting the dynamic between one dependent and several independent variables. This analytical approach is foundational in predictive analytics and modeling complex relationships. The essence of MLR is encapsulated in the formula:

Y=β0​+β1​X1​+β2​X2​+...+βn​Xn​+ε (10)

Expected frequencies

Where:

* Y represents the outcome or the dependent variable whose variance we seek to explain.
* X1, X2, ..., Xn signify the independent variables hypothesized to influence Y.
* β0 is the y-intercept, marking the expected value of Y when all independent variables are zero.
* β1, β2, ..., βn are each independent variable's regression coefficients, quantifying their impact on Y.
* ε denotes the error term, accounting for the variation in Y not explained by the independent variables.

By applying MLR, researchers and data analysts can isolate and quantify the contribution of each variable to the prediction, providing a multifaceted understanding of the factors at play.

### Performing Multiple Linear Regression in Excel

Calculating the linear regression coefficients typically involves statistical software that can handle matrix algebra, as the calculation by hand is complex and time-consuming, significantly as the number of independent variables increases.

As an illustration, we can check the following data from Table 4. The air we breathe in cities is influenced by various factors—each with its own story. The constant flow of vehicles, the looming presence of industrial sectors, and the refreshing presence of urban green spaces all contribute to the air's composition. By incorporating variables such as vehicle emissions, industrial pollution, and tree coverage into our analysis, we can understand their individual and collective impact on urban air quality, as measured by the Air Quality Index (AQI).

**Independent Variables (X)**

**Dependent (Y) Variables (X)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **City** | **Vehicle Emissions (kTons/year)** | **Industrial Emissions (kTons/year)** | **Tree Coverage (%)** | **Air Quality Index (AQI)** |
| **A** | 150 | 45 | 40 | 80 |
| **B** | 200 | 70 | 25 | 110 |
| **C** | 180 | 50 | 35 | 90 |
| **D** | 220 | 65 | 20 | 115 |
| **E** | 140 | 35 | 45 | 75 |
| **F** | 195 | 60 | 28 | 100 |
| **G** | 160 | 40 | 38 | 85 |
| **H** | 175 | 55 | 32 | 95 |
| **I** | 210 | 75 | 22 | 120 |
| **J** | 130 | 30 | 50 | 70 |
| **K** | 190 | 48 | 33 | 92 |
| **L** | 205 | 63 | 27 | 105 |
| **M** | 170 | 52 | 36 | 88 |
| **N** | 215 | 68 | 24 | 112 |
| **O** | 155 | 47 | 41 | 82 |

*Table 4 Multiple Linear Regression in Excel example - Air Quality Index (AQI) analysis*

Here are the steps to perform Multiple Linear Regression in Excel for Table 4:

1. Organize your data: Arrange your independent variables in adjacent and dependent variables in a separate column.
2. Go to the Data tab and click on Data Analysis. If you do not see this option, you must install the free Analysis ToolPak first.
3. Select Regression and click OK.
4. For the Input Y Range, fill in the array of values for the **dependent** variable (**Air Quality Index**). For Input X Range, fill in the array of values for the **independent** variables (**Vehicle Emissions, Industrial Emissions, Tree Coverage**). Check the box next to Labels so Excel knows we included the variable names in the input ranges. For Output Range, select a cell where you would like the output of the Regression to appear. Then click OK.

The output will automatically appear as follows:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Regression Statistics* |  |  |  |  |  |  |  |  |
| Multiple R | 0,989127352 |  |  |  |  |  |  |  |
| R Square | 0,978372918 |  |  |  |  |  |  |  |
| Adjusted R Square | 0,972474623 |  |  |  |  |  |  |  |
| Standard Error | 2,533092713 |  |  |  |  |  |  |  |
| Observations | 15 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 3 | 3193,017854 | 1064,3393 | 165,8738 | 1,94556E-09 |  |  |  |
| Residual | 11 | 70,58214563 | 6,4165587 |  |  |  |  |  |
| Total | 14 | 3263,6 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95,0%* | *Upper 95,0%* |
| Intercept | 93,99888397 | 42,92424536 | 2,1898785 | 0,050979 | -0,476743071 | 188,474511 | -0,476743071 | 188,474511 |
| Vehicle Emissions (kTons/year) | 0,055995172 | 0,124334102 | 0,4503605 | 0,661197 | -0,217662341 | 0,329652684 | -0,217662341 | 0,329652684 |
| Industrial Emissions (kTons/year) | 0,408047448 | 0,174830689 | 2,3339578 | 0,039592 | 0,023247697 | 0,792847199 | 0,023247697 | 0,792847199 |
| Tree Coverage (%) | -0,94667812 | 0,516494918 | -1,83289 | 0,094 | -2,08347577 | 0,190119529 | -2,08347577 | 0,190119529 |

Here is how to interpret the most relevant numbers in the output.

**Regression Statistics:**

* Multiple R (0.9891): The correlation coefficient indicates a strong positive correlation between the predictors and the AQI.
* R Square (0.9783): This value explains that 97.83% of the variance in AQI can be predicted from the three variables in the model.
* Adjusted R Square (0.9725): Adjusted for the number of predictors in the model, this high value still indicates that the model explains a large proportion of the variance.
* Standard Error (2.5331): This shows the average distance that the observed values deviate from the regression line. A smaller number suggests a better fit.
* Observations (15): The dataset's total number of observations (cities).

**ANOVA (Analysis of Variance):**

* The ANOVA table breaks down the variability in AQI into components due to the Regression (explained by the model) and the residuals (unexplained by the model).
* Significance F (1.94556E-09): This is a very small p-value, which means the regression model is statistically significant.

**Coefficients:**

* **Intercept** (93.9989): When vehicle emissions, industrial emissions, and tree coverage are at zero, the starting value for AQI is predicted to be around 94, which is the base level of AQI without these urban factors.
* **Vehicle Emissions** (0.0560): For every kiloton/year increase in vehicle emissions, AQI is predicted to increase by 0.0560, but this is not statistically significant (p-value: 0.6612).
* **Industrial Emissions** (0.4080): For every kiloton/year increase in industrial emissions, AQI is predicted to increase by 0.4080, which is statistically significant (p-value: 0.0396).
* **Tree Coverage** (-0.9467): For every percentage increase in tree coverage, AQI is predicted to decrease by 0.9467. This relationship is near the typical significance level (p-value: 0.0940), suggesting that more tree coverage is associated with better air quality, although this finding is at the marginal edge of being considered statistically significant.

## Classification and its applications

Classification algorithms have emerged as a cornerstone in predictive analytics, with a wide range of applications spanning various domains. At its core, the classification process involves assigning labels to new observations based on knowledge gleaned from a dataset containing pre-labeled examples. This method has revolutionized how we approach problem-solving in areas such as image recognition, where algorithms can identify and categorize elements within a photo, and medical diagnosis, where predictive models classify patient data to assist in the early detection of diseases.

The success of a classification algorithm is deeply rooted in the quality of its training data. The adage "*good data in, accurate results out*" is especially pertinent here; classification accuracy is contingent on the representativeness and granularity of the training set. This highlights the importance of meticulously curated datasets that reflect the complexity and diversity of real-world scenarios.

Selecting the appropriate algorithm for a classification task is a decision of paramount importance. Factors include the dataset's size and dimensionality, the desired balance between bias and variance, and computational efficiency. It is often advisable to begin with simple, interpretable, and widely utilized algorithms such as Logistic Regression, which models the probability of a binary outcome; k-nearest Neighbors (k-NN), which classifies an observation based on the majority label of its closest neighbors; Naive Bayes, which applies Bayes' theorem with the assumption of independence between predictors; and Decision Trees, which segment the data into branches to form a tree of decisions.

These algorithms serve as a foundation for building more complex models. However, their strength lies in their simplicity and robustness, making them a suitable starting point for many classification tasks. By effectively classifying data, we can unlock actionable insights and make informed decisions, whether determining the product category based on customer reviews or diagnosing a patient based on symptomatic data.

### Logistic Regression

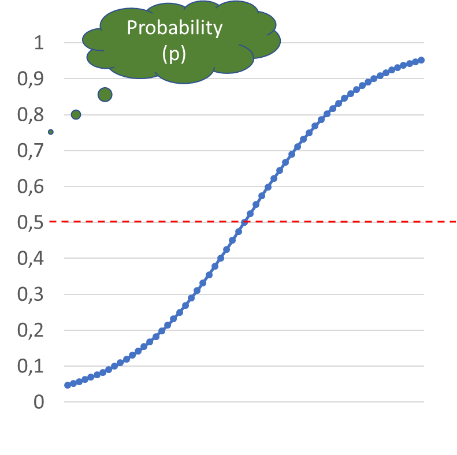
Logistic Regression is a predictive analysis technique used for classification problems. It predicts the probability of the occurrence of an event by fitting data to a logistic function. The general form of the logistic regression model is expressed mathematically as:

(11)

Where:

* p(y=1∣x) is the probability that the dependent variable equals a case, given the values of independent variables.
* ​b0 is the intercept term, which is the prediction value when all the independent variables xi are held at zero. The intercept is estimated during model training through maximum likelihood estimation.
* b1, b2, … , bn are the coefficients of the independent variables. These coefficients represent the change in the log odds of the dependent event occurring for a one-unit change in the respective independent variable.
* ​the independent variables influencing the prediction are x1, x2, … , xn.

The beauty of Logistic Regression lies in its ability to provide probabilities and classify new data using a logistic curve (Figure 7), which is bound between 0 and 1. This allows for a transparent, interpretable model where the coefficients indicate the direction and strength of the relationship between predictors and the outcome.



*Figure 7 Logistic curve*

As an illustration, we can check the following data from Table 5.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Average Daily Emissions (tons)** | **Number of Factories** | **Environmental Regulations (1=Yes, 0=No)** | **Average Temperature**  **(C)** | **Annual Rainfall (L/m2)** | **Population Density (people / m2)** | **Urbanized Area (%)** | **High Pollution Level (1=Yes, 0=No)** |
| 85 | 10 | 0 | 25,56 | 20 | 1295 | 50 | 1 |
| 30 | 2 | 1 | 21,11 | 50 | 388 | 20 | 0 |
| 90 | 12 | 0 | 26,67 | 25 | 1424 | 60 | 1 |
| 25 | 1 | 1 | 18,33 | 75 | 336 | 10 | 0 |
| 88 | 15 | 0 | 27,78 | 30 | 1554 | 70 | 1 |
| 20 | 0 | 1 | 20 | 80 | 259 | 5 | 0 |
| 95 | 18 | 0 | 29,44 | 15 | 1683 | 80 | 1 |
| 28 | 3 | 1 | 22,22 | 65 | 362 | 15 | 0 |

*Table 5 Environmental Data for Logistic Regression Analysis on Pollution Levels*

Through the lens of a logistic regression model, we can systematically approach the critical question: What elevates the probability of a region experiencing high pollution levels? Our illustrative dataset encapsulates a range of predictive variables, from the tangible—like the volume of daily emissions and the proliferation of factories—to the more abstract, such as environmental regulations and the degree of urbanization. The beauty of Logistic Regression lies in its ability **to handle binary outcomes**—such as whether a **location is classified as having high pollution levels (1) or not (0)**—with a finesse that linear models cannot afford. It elegantly captures the non-linear probabilities and provides odds ratios that offer real-world interpretability.

For instance, in our dataset, the variable "*Environmental Regulations*" emerges not merely as a policy indicator but as a beacon revealing the efficacy of regulatory frameworks in mitigating pollution. Similarly, "*Population Density*" and "*Urbanized Area*" percentages become proxies for human activity and its environmental footprint.

Maximum likelihood estimation (MLE) is employed to solve the Logistic Regression equation (11), which involves finding the parameter values that best fit the data. This technique determines the parameters by calculating the likelihood of the observed data under various hypothetical parameters and selecting the set that maximizes this likelihood. Programming languages like Python are ideal for implementing Logistic Regression and performing MLE, thanks to libraries such as *statsmodels* and *scikit-learn*. These libraries provide comprehensive functions that handle complex calculations, enabling users to concentrate on the analysis and interpretation of the model's outcomes.

In conclusion, these coefficients have been calculated for a specific example using Python, with certain limitations in mind:

* b0 - Intercept (const): 8.6274
* b1 - Average Daily Emissions (tons): -0.3523
* b2 - Number of Factories: -0.2066
* b3 - Environmental Regulations (1=Yes, 0=No): -26.8086
* b4 - Average Temperature (C): -0.3180
* b5 - Annual Rainfall (L/m2): -0.2306
* b6 - Population Density (people/m2): 0.0806
* b7 - Urbanized Area (%): -0.7232

However, due to the small size of the dataset and the potential for overfitting, these coefficients should be interpreted with caution. Further validation with more data is advisable to ensure the model's reliability and applicability to broader contexts.

Based on the coefficients derived from previous analysis, we can determine the probability of a high pollution level occurring under specific conditions. The logistic regression equation is given by:

(12)

The p(y=1∣x) is the probability of a high pollution level (y=1) given the independent variables (x). The coefficients represent the intercept and slopes for the respective variables in the dataset. Once the probability is calculated, if p is greater than 0.5, the model predicts a high likelihood of a pollution level being classified as *'high'* under the given circumstances. Conversely, if p is less than 0.5, the pollution level is predicted to be *'low*.'

## References

1. dsd