Supervised Learning
Course Summary - Master's Degree

Carlos Alberto Botina Carpio

Universidad Internacional de la Rioja

carlos.botina621@comunidadunir.net

November 17, 2025

Abstract

This document contains a summary of the Supervised Learning course syllabus for the Master's Degree. It includes a summary of the main topics covered during the sessions, as well as additional explanations and extensions of the concepts and techniques referenced in class. The purpose of this document is to serve as study material and reference for the course contents. Notice that this document is fully customized to the author's needs, placing greater emphasis on topics that the author struggles with or has not yet mastered, while covering more briefly those topics that are already well understood by the author.

Contents

Lecture 002

1 Descriptive Data Analysis

1.1 Types of Variables

One of the first stages of descriptive analysis is to recognize the different types of data, as these observations are a representation of the world around us through what are called variables. A variable is a characteristic or attribute of an object that can be observed and measured, assuming different values. Variables can be represented in different forms: numerical (discrete and continuous) or categorical (nominal and ordinal). Figure 1 summarizes the classification of variable types.
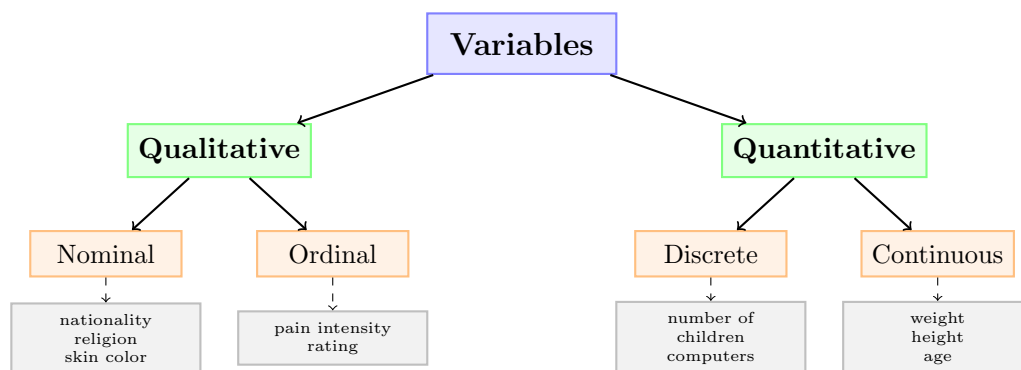
Figure 1: Classification of variable types

1.1.1 Qualitative Variable

A qualitative variable expresses a characteristic in a non-numerical form, representing a quality. Examples: color, gender, blood type, marital status. It can be of two types:

- **Nominal:** Values are codes or names that cannot be ordered. Examples: nationality, religion, skin color.

- **Ordinal:** Values imply an ordering from greater to lesser or from better to worse. Examples: pain intensity (absent, mild, moderate, severe, very severe), rating (excellent, good, fair, poor).

1.1.2 Quantitative Variable

A quantitative variable is expressed by means of a number. It can be of two types:

- **Discrete:** Only admits a finite set of numerical values, usually integers. Examples: number of children per household, number of computers in a classroom, number of electrons in an atom.

- **Continuous:** Can take infinitely many possible values within a range or interval on the real number line. These variables represent measurements whose values cannot be enumerated. Examples: weight (kg), height, age, the time it takes an athlete to run 100 meters.

1.2 Characterization of Variable Distributions

1.2.1 Histogram

A histogram is a graphical representation of the distribution of numerical data. It displays the frequency of data points within specified intervals (bins). The x-axis represents the intervals, and the y-axis represents the frequency or relative frequency. The bars are adjacent to each other, unlike bar charts for categorical data.

Histograms are used for continuous or discrete quantitative variables. The bin width affects the appearance and interpretation. The area of each bar is proportional to the frequency, with no gaps between bars (unless a bin has zero frequency). Histograms reveal the shape, center, and spread of the data distribution.

To read a histogram, examine the overall shape (symmetric, skewed, or multimodal), identify the center, assess the spread, and look for outliers or unusual patterns.

Figure 2 shows an example of a histogram. Its shape depends on the number of intervals, and the symmetry can vary; the more similar the values on both sides of the center, the more symmetric the distribution (Rodríguez Ojeda, 2007):

- If the height or y-axis of the bars are similar to each other, we would have a uniform distribution.

- If the heights are greater in the central zone, a "bell" shape is formed, which can be symmetric or asymmetric, toward the positive side (to the right) or negative (to the left).

- If there are bars far from the group, they are considered atypical data, which are probably due to measurement errors and can be discarded, as they do not belong to the group that is desired to be characterized (Rodríguez Ojeda, 2007).
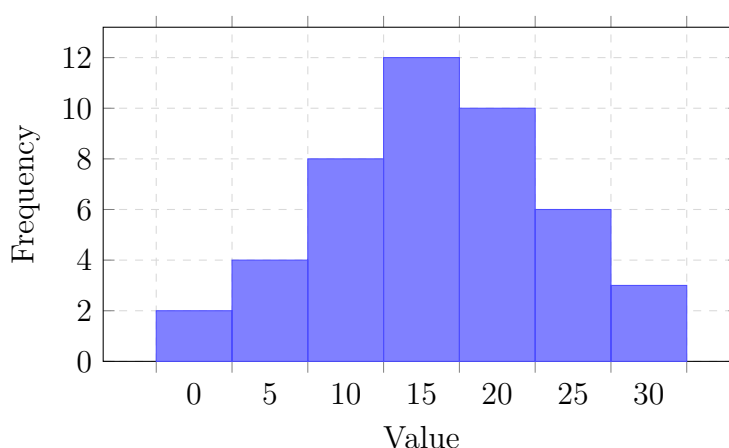

Figure 2: Example of a histogram showing a bell-shaped distribution

2 Exploratory Data Analysis

2.1 Measures of Dispersion

Data tends to cluster toward the center of the frequency distribution, but extreme values can be far from this central tendency. Measures of dispersion quantify this distance from the average.

2.1.1 Range

The difference between the upper and lower limits of a dataset. Limited utility as it is easily affected by extreme values. The interquartile range (IQR) is commonly used to suppress the influence of extremes (Witte y Witte, 2017).

2.1.2 Quartiles and Interquartile Range

Quartiles divide the distribution into four equal parts (25% each). Q1 (25th percentile), Q2 (median, 50th percentile), and Q3 (75th percentile). The IQR equals Q3 - Q1, representing the range for the middle 50% of data, effectively removing the upper and lower 25%. The IQR is resistant to extreme values and tends to be less than half the range (Witte y Witte, 2017).
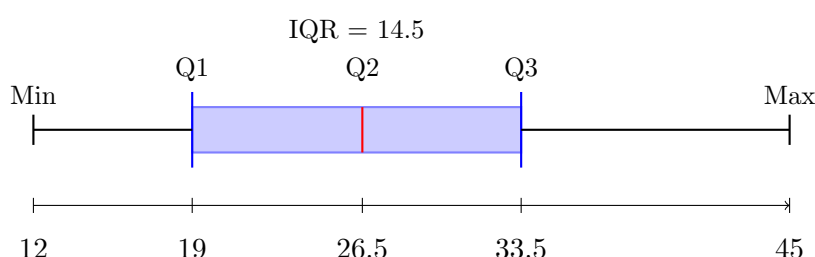

Figure 3: Box plot showing quartiles and interquartile range. The box represents the IQR (middle 50% of data), Q1 and Q3 are the box edges, and Q2 (median) is the red line.

2.1.3 Variance

The average of the squared deviations from the arithmetic mean. Deviations are squared because their sum equals zero (negative cancel positive). Notation: s^2 (sample) and σ^2 (population). Useful for comparing variability between distributions, but has squared units making interpretation complex (Spiegel y Stephen, 2009).

The formulas for variance are:

$$\text{Population variance: } \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{1}$$

$$\text{Sample variance: } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2}$$

where N is the population size, n is the sample size, μ is the population mean, and $\bar{x}$ is the sample mean.

Example: Calculating Variance Consider the following dataset representing test scores: 75, 82, 68, 90, 85.
Note: This example calculates sample variance (s^2), treating the data as a sample from a larger population. If this were the entire population, we would use population variance (σ^2) and divide by N instead of $(n-1)$.

Step 1: Calculate the mean

$$\bar{x} = \frac{75 + 82 + 68 + 90 + 85}{5} = \frac{400}{5} = 80 \tag{3}$$

Step 2: Calculate deviations from the mean

$$x_1 - \bar{x} = 75 - 80 = -5 \tag{4}$$
$$x_2 - \bar{x} = 82 - 80 = 2 \tag{5}$$
$$x_3 - \bar{x} = 68 - 80 = -12 \tag{6}$$
$$x_4 - \bar{x} = 90 - 80 = 10 \tag{7}$$
$$x_5 - \bar{x} = 85 - 80 = 5 \tag{8}$$

Step 3: Square each deviation

$$(x_1 - \bar{x})^2 = (-5)^2 = 25 \tag{9}$$
$$(x_2 - \bar{x})^2 = (2)^2 = 4 \tag{10}$$
$$(x_3 - \bar{x})^2 = (-12)^2 = 144 \tag{11}$$
$$(x_4 - \bar{x})^2 = (10)^2 = 100 \tag{12}$$
$$(x_5 - \bar{x})^2 = (5)^2 = 25 \tag{13}$$

Step 4: Sum the squared deviations

$$\sum_{i=1}^{5}(x_i - \bar{x})^2 = 25 + 4 + 144 + 100 + 25 = 298 \tag{14}$$

Step 5: Calculate sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{5-1} \times 298 = \frac{298}{4} = 74.5 \tag{15}$$

Therefore, the sample variance is $s^2 = 74.5$ (units squared).

2.1.4 Standard Deviation

To resolve the problem of squared units of measurement, the square root of the variance is calculated, always taking the positive value. This produces a new measure, known as standard deviation, which describes variability in the original units of measurement. Like the previous measures, standard deviation must be distinguished: symbolized by s for the sample and σ for the population.

The formulas for standard deviation are:

$$\text{Population standard deviation: } \sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{16}$$

$$\text{Sample standard deviation: } s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{17}$$

Example: Calculating Standard Deviation Using the previous variance example (test scores: 75, 82, 68, 90, 85), where we calculated $s^2 = 74.5$:

$$s = \sqrt{s^2} = \sqrt{74.5} \approx 8.63 \tag{18}$$

Therefore, the sample standard deviation is $s \approx 8.63$ test score points.

Interpretation The standard deviation measures the typical distance of data points from the mean. In this example:

- The mean test score is 80 points

- The standard deviation is approximately 8.63 points

- This means that, on average, test scores deviate from the mean by about 8.63 points

- Most scores (approximately 68% in a normal distribution) fall within one standard deviation of the mean, i.e., between $80 - 8.63 = 71.37$ and $80 + 8.63 = 88.63$ points

> **Curious Fact**
>
> A larger standard deviation indicates greater variability (scores are more spread out), while a smaller standard deviation indicates less variability (scores are clustered closer to the mean).

2.2 Outlier Detection

An outlier is a numerical observation that is distant from the rest of the data. Outliers in a dataset can cause problems in statistical analyses by generating misleading conclusions. Usually, an outlier can indicate a measurement error or a long-tailed distribution in the population, although they can occur by chance.

2.2.1 Box Plot

A box plot is a graphical representation of numerical data through its quartiles. The boxes in this standardized method contain the following elements: the median (Q2), the interquartile range, and the quartiles (Q1 and Q3). Box plots also have lines extending from the boxes (whiskers) that indicate variability outside the upper and lower quartiles, hence the term box-and-whisker plot. Depending on the type of representation, the whiskers can extend to the minimum value at one end and the maximum value at the other, representing the full range.

For outlier identification, John Tukey proposed whiskers with a length of 1.5 times the IQR (Tukey, 1977). That is, from the hinges (Q1 and Q3), 1.5 times the IQR is added or subtracted respectively. Then, values beyond these whiskers are considered outliers and are represented as individual points.

To construct box plots, the median, first and third quartiles, and the interquartile range of each attribute are used. Then, for the whiskers, 1.5 times the IQR is added above and subtracted below the box. Finally, the last value within those limits is taken, and that is where the whiskers are graphed.

Example: Box Plot with Outliers   Consider the dataset: 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 60, 70.

Calculating the quartiles:

- Q1 = 16, Q2 (median) = 23, Q3 = 30

- IQR = Q3 - Q1 = 30 - 16 = 14

- Lower whisker limit: Q1 - 1.5 × IQR = 16 - 21 = -5 (minimum within limit: 10)

- Upper whisker limit: Q3 + 1.5 × IQR = 30 + 21 = 51

- Outliers: 60 and 70 (both exceed the upper whisker limit)

Note: The whiskers extend to 10 and 32, not to the calculated limits (-5 and 51). This is because whiskers are drawn to the last data value within the limits, not to the limits themselves. Since -5 is below any actual data point, the lower whisker goes to the minimum value (10). Similarly, the upper whisker stops at 32, which is the last value within the upper limit (51). Values beyond these limits (60 and 70) are shown as outlier points.

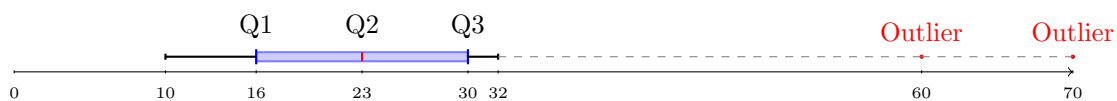Figure 4 shows the box plot for this dataset, with outliers represented as individual points.

Figure 4: Box plot showing quartiles, IQR, whiskers (extending 1.5 × IQR from Q1 and Q3), and outliers (60 and 70) represented as individual red points.

2.3   Scatter Plots

The most useful graphical representations for describing the behavior of a set of variables are scatter plots, also known as dot plots or scatter diagrams. These graphs show the relationship between numerical variables, and they allow discovering and confirming anticipated relationships between two associated sets of data. Numerical variables are represented by coordinates in n-dimensional Euclidean space.

2.3.1   Two-dimensional Scatter Plots

For the case of two variables, the datasets must have the same length: one for the X-axis (horizontal) and one for the Y-axis (vertical). In Python, a scatter plot of two numerical variables X and Y can be obtained using the `plot(X, Y)` command from the matplotlib library.

The representation of each variable on each axis allows identifying the relationship of increase or decrease between the dependent and independent variables. Scatter plots are suitable when you have paired numerical data and want to see if one variable affects another. However, it is necessary to understand that correlation is not causation, and another unnoticed variable may influence the results. In the case of a database, numerical variables correspond to dataset attributes and are used to define aspects of an instance, such as size, shape, and color.

Correlation vs. Causation Example: Consider a scatter plot showing ice cream sales (X-axis) and drowning deaths (Y-axis) over time. The plot would likely show a positive correlation: as ice cream sales increase, drowning deaths also increase. However, this does not mean that ice cream sales *cause* drowning deaths.

Why correlation is not causation: Both variables are influenced by a third, hidden variable: the season (summer). In summer:

- More people buy ice cream (hot weather)

- More people go swimming (hot weather)

- More swimming leads to more drowning incidents

The correlation exists because both variables respond to the same underlying factor (summer weather), not because one directly causes the other. This illustrates that:

- **Correlation** means two variables tend to change together

- **Causation** means one variable directly causes changes in another

- A correlation can exist without causation when a third variable (confounding variable) influences both

To establish causation, controlled experiments or additional evidence are needed, not just observational correlation.

Example: Scatter Plot Figure 5 shows an example of a two-dimensional scatter plot displaying the relationship between two variables. Each point represents a pair of values (x, y), allowing visualization of patterns, trends, and potential correlations in the data.
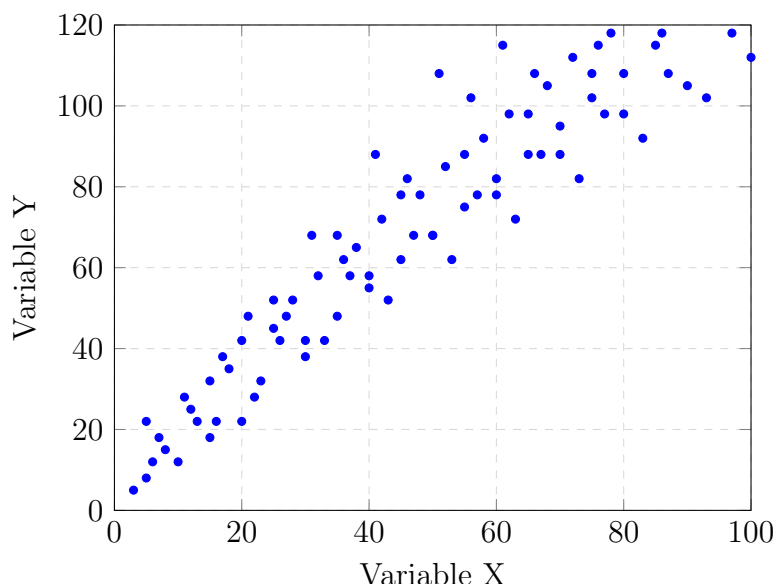

Figure 5: Example of a two-dimensional scatter plot showing a positive correlation between Variable X and Variable Y.

2.3.2 Scatter Plot Matrix

The scatter plot matrix (SPLOM—scatterplot matrix) presents multiple adjacent scatter plots for all variable comparisons in a single display. To some extent, it addresses one of the geometric problems for datasets with more than three attributes by presenting two or more pairs of variables in the same graph, which is very useful for identifying dependencies between data.

One of the disadvantages is that SPLOM requires a large amount of screen space, and the formation of multivariate associations remains a challenge. Additionally, it is necessary to incorporate additional statistical measures to organize SPLOM and guide the observer through an exploratory analysis of high-dimensional datasets.

2.4 Correlation between Variables

2.4.1 Linear Correlation

Correlation is known as a statistical measure that quantifies the degree of joint variation that exists between two variables and, specifically, evaluates the increasing or decreasing trend of the data. The types of correlation can be observed in Table 1.

Type of Correlation	Description
Positive correlation	The values of the variables increase together, given that an increase in Y (dependent variable) depends on an increase in X (independent variable).
Negative correlation	One value decreases as the other increases; therefore, an increase in the values of variable X will cause a decrease in the values of variable Y.
Null or no correlation	There is no relationship between the variables, therefore, the variables are independent. The graph does not follow any trend, causing the points to be totally dispersed.
Linear	There is a relationship between the variables, which is linear.
Non-linear	Although there is a relationship between the variables, it is not linear. It can be exponential, U-shaped, etc.

Table 1: Types of correlation

Given the subjectivity in interpreting scatter plots, it is necessary to use a coefficient that allows measuring the degree of dependence between variables. The linear correlation coefficient represents the behavior of a dependent variable Y with respect to an independent variable X.

The strength of the correlation is determined by the proximity of the points to each other in the graph. If the points are closer together, the strength will be greater (strong correlation). If they are not so concentrated, the strength will be weak. If a pattern of very dispersed points is observed, the correlation strength will be null (no correlation).

2.4.2 Pearson's Correlation Coefficient

Pearson's correlation coefficient is a statistic whose values range between -1 and 1 ($-1 \leq r \leq 1$). A correlation close to 1 indicates a positive association (both variables increase together). A correlation close to -1 indicates a negative association (one increases as the other decreases). A value near 0 indicates no linear relationship, although a non-linear relationship may exist. This coefficient can only be used for quantitative and continuous variables.

The correlation coefficient is calculated as:

$$r = \frac{\text{Cov}(X,Y)}{S_x \cdot S_y} \tag{19}$$

where:

- r is Pearson's correlation coefficient

- $\text{Cov}(X,Y)$ is the covariance between X and Y

- S_x is the standard deviation of variable X

- S_y is the standard deviation of variable Y

This coefficient allows determining the existence and intensity of linear associations between variables. The sign is interpreted the same as covariance: positive (direct correlation), negative (inverse correlation), or null (no correlation). It measures how close the points are to a straight line, reflecting the data trend.

Curious Fact: Covariance

Covariance measures how two variables vary together. It is calculated as the average of the products of deviations from the mean: $\text{Cov}(X,Y) = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})$. A positive covariance means that when X is above its mean, Y tends to be above its mean too (they move together). A negative covariance means when X is above its mean, Y tends to be below its mean (they move in opposite directions). However, covariance is difficult to interpret because its magnitude depends on the units of measurement. This is why Pearson's correlation coefficient normalizes covariance by dividing by the product of standard deviations, making it unitless and easier to interpret.

Example: Calculating Pearson's Correlation Coefficient   Consider the following paired dataset:

X	Y
2	4
4	6
6	8
8	10
10	12

Step 1: Calculate the means

$$\bar{x} = \frac{2 + 4 + 6 + 8 + 10}{5} = \frac{30}{5} = 6 \tag{20}$$

$$\bar{y} = \frac{4 + 6 + 8 + 10 + 12}{5} = \frac{40}{5} = 8 \tag{21}$$

Step 2: Calculate deviations and their products

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (2-6)(4-8) + (4-6)(6-8) + (6-6)(8-8) + (8-6)(10-8)... \tag{22}$$

$$= (-4)(-4) + (-2)(-2) + (0)(0) + (2)(2) + (4)(4) \tag{23}$$

$$= 16 + 4 + 0 + 4 + 16 = 40 \tag{24}$$

Step 3: Calculate standard deviations

$$S_x = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2} = \sqrt{\frac{(-4)^2 + (-2)^2 + 0^2 + 2^2 + 4^2}{4}} = \sqrt{\frac{40}{4}} = \sqrt{10} \approx 3.16 \tag{25}$$

$$S_y = \sqrt{\frac{1}{n-1} \sum (y_i - \bar{y})^2} = \sqrt{\frac{(-4)^2 + (-2)^2 + 0^2 + 2^2 + 4^2}{4}} = \sqrt{\frac{40}{4}} = \sqrt{10} \approx 3.16 \tag{26}$$

Step 4: Calculate covariance

$$\text{Cov}(X,Y) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{40}{4} = 10 \tag{27}$$

Step 5: Calculate Pearson's correlation coefficient

$$r = \frac{\text{Cov}(X,Y)}{S_x \cdot S_y} = \frac{10}{3.16 \times 3.16} = \frac{10}{10} = 1.0 \tag{28}$$

Therefore, $r = 1.0$, indicating a perfect positive linear correlation. This makes sense as the data points lie exactly on a straight line with a positive slope.

2.5 Covariance Matrix

Statistics allow understanding relationships between multiple variables simultaneously. If covariance is positive, the linear association is positive. If covariance is negative, the linear association is negative (small X values correspond to large Y values, and vice versa).

The covariance matrix is a square matrix containing variances on the main diagonal and covariances in the off-diagonal elements. The matrix is symmetric: the covariance between X and Y equals the covariance between Y and X, so each pair appears twice.