Mathematics for Machine Learning

E & ICT Academy IIT Kanpur

Submitted By
Diwakar Jaiswal

Machine Learning — Probability & Statistics

Machine Learning is an interdisciplinary field that uses statistics, probability, algorithms to learn from data and provide insights which can be used to build intelligent applications. In this article, we will discuss some of the key concepts widely used in machine learning.

Probability and statistics are related areas of mathematics which concern themselves with analyzing the relative frequency of events.

Probability deals with predicting the likelihood of future events, while statistics involves the analysis of the frequency of past events.

Probability

Most people have an intuitive understanding of degrees of probability, which is why we use words like “probably” and “unlikely” in our daily conversation, but we will talk about how to make quantitative claims about those degrees [1].

In probability theory, an event is a set of outcomes of an experiment to which a probability is assigned. If E represents an event, then P(E) represents the probability that E will occur. A situation where E might happen (success) or might not happen (failure) is called a trial.

This event can be anything like tossing a coin, rolling a die or pulling a colored ball out of a bag. In these examples the outcome of the event is random, so the variable that represents the outcome of these events is called a random variable.

Joint Probability

Probability of events A and B denoted by$P(A \text{ and } B)$ or $P(A \cap B)$is the probability that events A and B both occur. $P(A \cap B) = P(A). P(B)$. This only applies if Aand Bare independent, which means that if Aoccurred, that doesn’t change the probability of B, and vice versa.

Conditional Probability

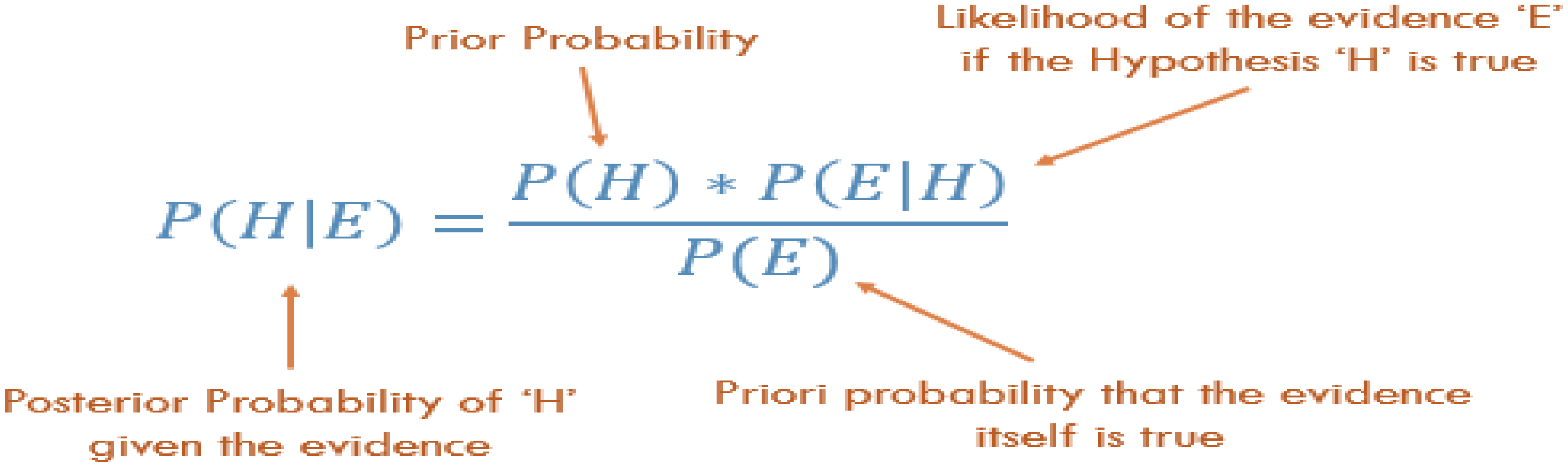
Let us consider A and B are not independent, because if A occurred, the probability of B is higher. When A and B are not independent, it is often useful to compute the conditional probability, $P(A|B)$, which is the probability of A given that B occurred: $P(A|B) = P(A \cap B)/ P(B)$.

The probability of an event A conditioned on an event B is denoted and defined $P(A|B) = P(A \cap B)/P(B)$

Similarly, $P(B|A) = P(A \cap B)/ P(A)$. We can write the joint probability of as A and B as $P(A \cap B)= p(A).P(B|A)$, which means : “The chance of both things happening is the chance that the first one happens, and then the second one given the first happened.”

Bayes’ Theorem

Bayes’s theorem is a relationship between the conditional probabilities of two events. For example, if we want to find the probability of selling ice cream on a hot and sunny day, Bayes’ theorem gives us the tools to use prior knowledge about the likelihood of selling ice cream on any other type of day (rainy, windy, snowy etc.).



$$P(H|E) = \frac{P(H) * P(E|H)}{P(E)}$$

Prior Probability

Likelihood of the evidence ‘E’ if the Hypothesis ‘H’ is true

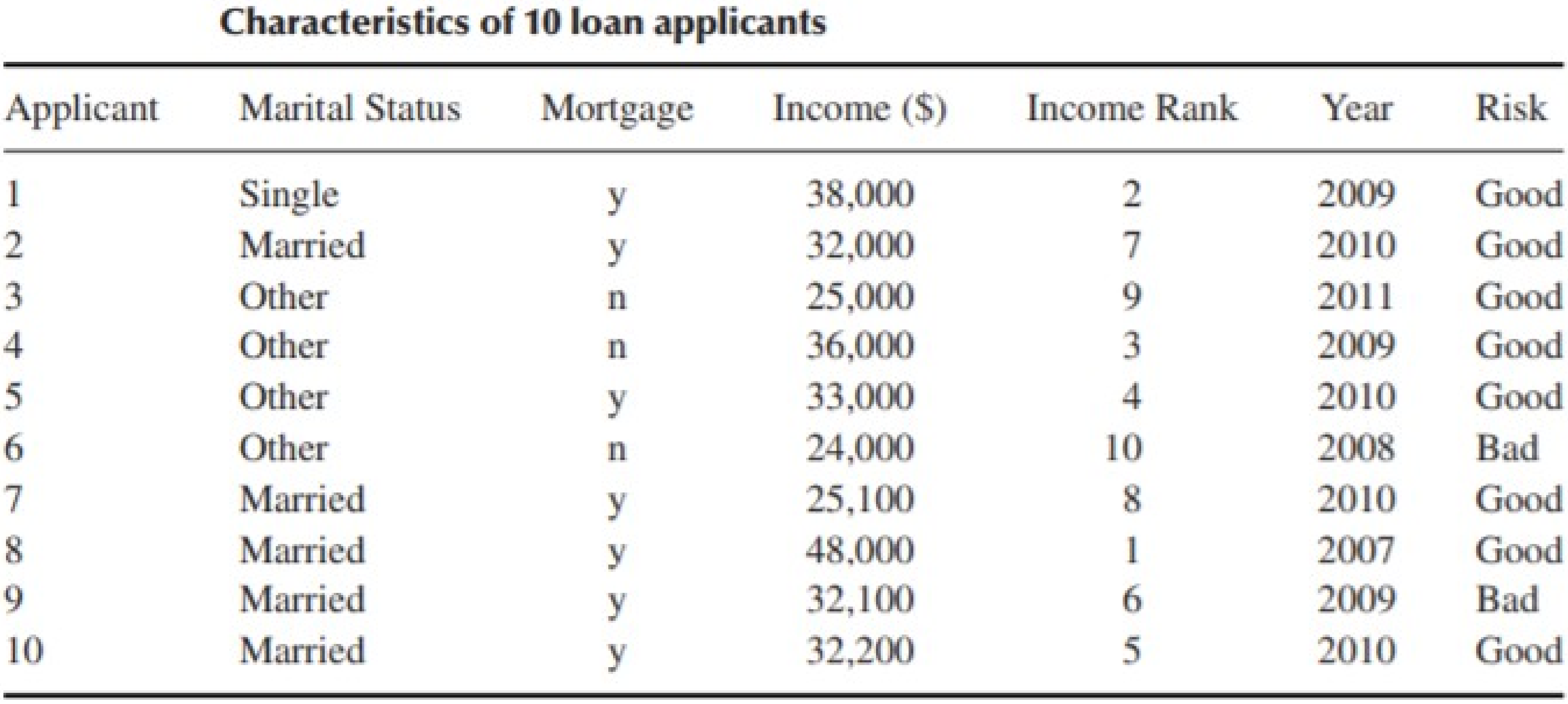
Posterior Probability of ‘H’ given the evidence

Priori probability that the evidence itself is true

where Hand E are events, P(H|E) is the conditional probability that event H occurs given that event E has already occurred.

Descriptive Statistics

Descriptive statistics refers to methods for summarizing and organizing the information in a data set. We will use below table to describe some of the statistical concepts [4].



Characteristics of 10 loan applicants

Applicant	Marital Status	Mortgage	Income ($)	Income Rank	Year	Risk
1	Single	y	38,000	2	2009	Good
2	Married	y	32,000	7	2010	Good
3	Other	n	25,000	9	2011	Good
4	Other	n	36,000	3	2009	Good
5	Other	y	33,000	4	2010	Good
6	Other	n	24,000	10	2008	Bad
7	Married	y	25,100	8	2010	Good
8	Married	y	48,000	1	2007	Good
9	Married	y	32,100	6	2009	Bad
10	Married	y	32,200	5	2010	Good

Elements: The entities for which information is collected are called the elements. In the above table, the elements are the 10 applicants. Elements are also called cases or subjects.

Variables: The characteristic of an element is called a variable. It can take different values for different elements.e.g., marital status, mortgage, income, rank, year, and risk. Variables are also called attributes.

Variables can be either qualitative or quantitative.

Qualitative: A qualitative variable enables the elements to be classified or categorized according to some characteristic. The qualitative variables are marital status, mortgage, rank, and risk. Qualitative variables are also called categorical variables.

Quantitative: A quantitative variable takes numeric values and allows arithmetic to be meaningfully performed on it. The quantitative variables are income and year. Quantitative variables are also called numerical variables.

Discrete Variable: A numerical variable that can take either a finite or a countable number of values is a discrete variable, for which each value can be graphed as a separate point, with space between each point. ‘year’ is an example of a discrete variable.

Continuous Variable: A numerical variable that can take infinitely many values is a continuous variable, whose possible values form an interval on the number line, with no space between the points. ‘income’ is an example of a continuous variable.

Population: A population is the set of all elements of interest for a particular problem. A parameter is a characteristic of a population.

Sample: A sample consists of a subset of the population. A characteristic of a sample is called a statistic.

Random sample: When we take a sample for which each element has an equal chance of being selected.

Measures of Center: Mean, Median, Mode, Mid-range

Indicate where on the number line the central part of the data is located.

Mean: The mean is the arithmetic average of a data set. To calculate the mean, add up the values and divide by the number of values.The sample mean is the arithmetic average of a sample, and is denoted $\bar{x}$ (“x-bar”). The population mean is the arithmetic average of a population, and is denoted μ (“myu”, the Greek letter for m).

Median: The median is the middle data value, when there is an odd number of data values and the data have been sorted into ascending order. If there is an even number, the median is the mean of the two middle data values. When the income data are sorted into ascending order, the two middle values are $32,100 and $32,200, the mean of which is the median income, $32,150.

Mode: The mode is the data value that occurs with the greatest frequency. Both quantitative and categorical variables can have modes, but only quantitative variables can have means or medians. Each income value occurs only once, so there is no mode. The mode for year is 2010, with a frequency of 4.

Mid-range: The mid-range is the average of the maximum and minimum values in a data set. The mid-range income is:

mid-range(income) = (max(income) + min(income))/2 = (48000 + 24000)/2 = $36000

Measures of Variability: Range, Variance, Standard Deviation

Quantify the amount of variation, spread or dispersion present in the data.

Range: The range of a variable equals the difference between the maximum and minimum values. The range of income is:

range(income) = max (income) − min (income) = 48,000 − 24,000 =$24000

Range only reflects the difference between largest and smallest observation, but it fails to reflect how data is centralized.

Variance: Population variance is defined as the average of the squared differences from the Mean, denoted as σ^2 (“sigma-squared”):

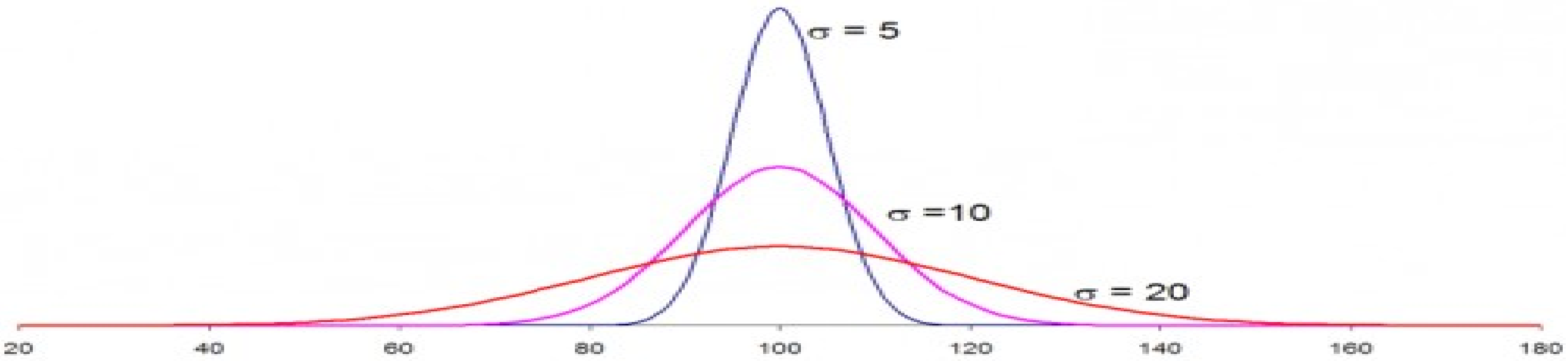
$$\sigma^2 = \frac{\sum (x-\mu)^2}{N}$$

Larger Variance means the data are more spread out.

Standard Deviation

The standard deviation or sd of a bunch of numbers tells you how much the individual numbers tend to differ from the mean.

The population standard deviation is the square root of the population variance: sd= $\sqrt{\sigma^2}$



Three different data distributions with same mean (100) and different standard deviation (5,10,20)

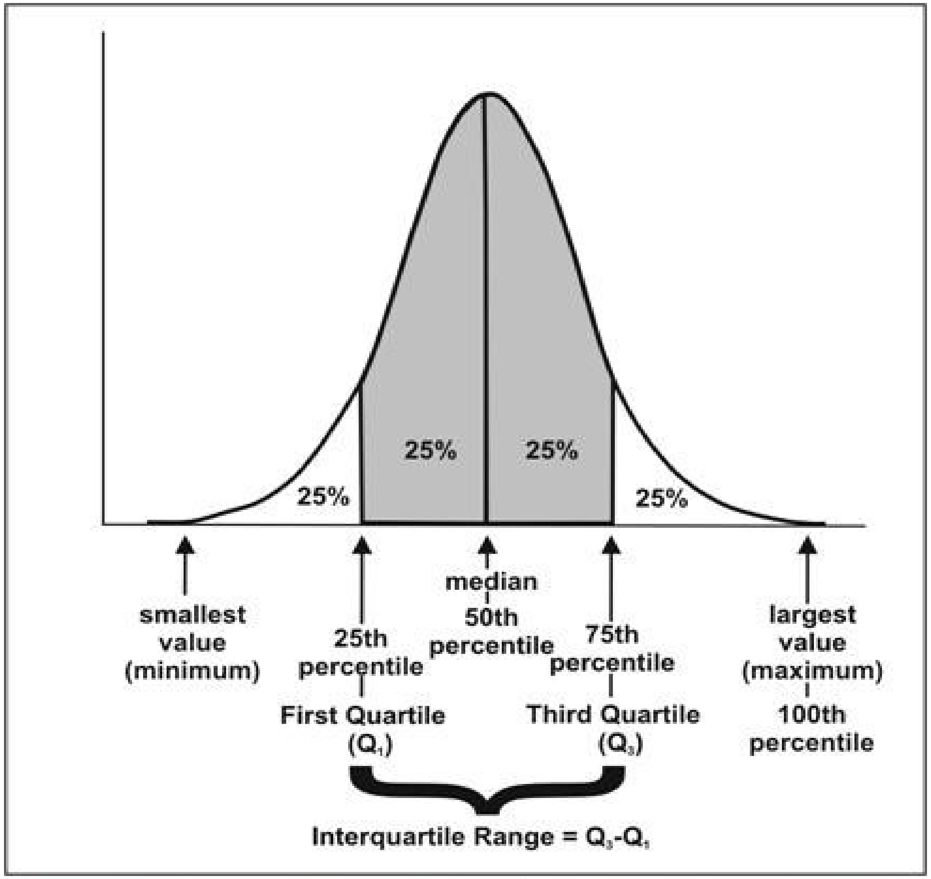
The smaller the standard deviation, narrower the peak, the data points are closer to the mean. The further the data points are from the mean, the greater the standard deviation.

Measures of Position: Percentile, Z-score, Quartiles

Indicate the relative position of a particular data value in the data distribution.

Percentile: The pth percentile of a data set is the data value such that p percent of the values in the data set are at or below this value. The 50th percentile is the median. For example, the median income is $32,150, and 50% of the data values lie at or below this value.

Percentile rank: The percentile rank of a data value equals the percentage of values in the data set that are at or below that value. For example, the percentile rank. of Applicant 1’s income of $38,000 is 90%, since that is the percentage of incomes equal to or less than $38,000.



Interquartile Range (IQR): The first quartile (Q1) is the 25th percentile of a data set; the second quartile (Q2) is the 50th percentile (median); and the third quartile (Q3) is the 75th percentile.

The IQR measures the difference between 75th and 25th observation using the formula: IQR = Q3 − Q1.

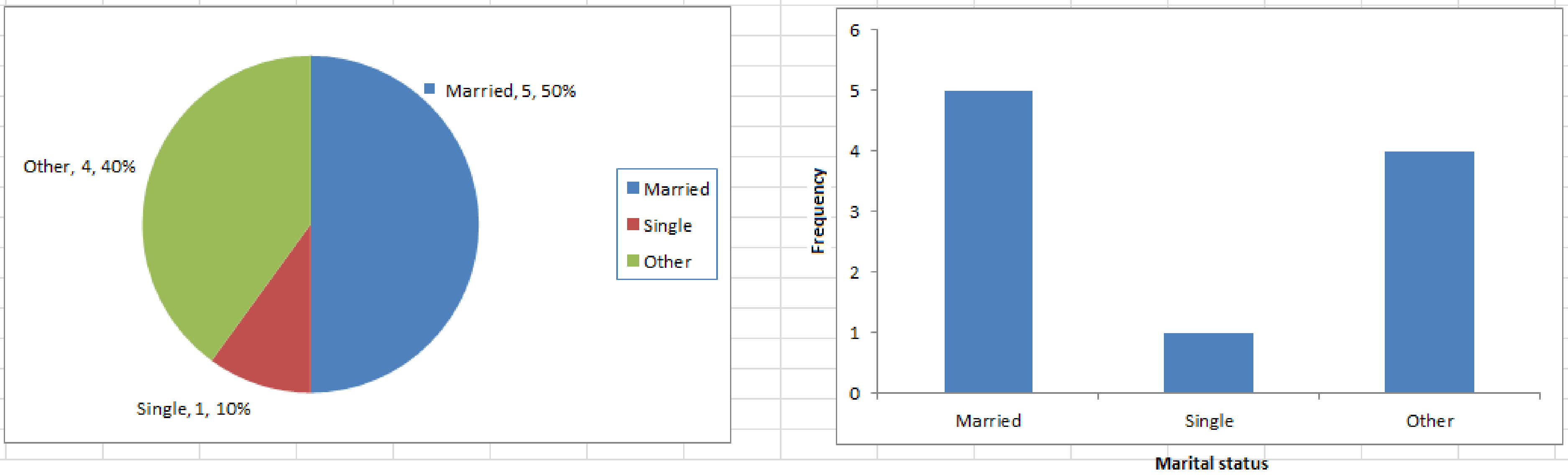
A data value x is an outlier if either $x \leq Q1 - 1.5(IQR)$, or $x \geq Q3 + 1.5(IQR)$.

Z-score: The Z-score for a particular data value represents how many standard deviations the data value lies above or below the mean.

$$\text{Z-score} = \frac{x - \bar{x}}{s}$$

Uni-variate Descriptive Statistics

Different ways you can describe patterns found in uni-variate data include central tendency : mean, mode and median and dispersion: range, variance, maximum, minimum, quartiles , and standard deviation.



Pie chart [left] & Bar chart [right] of Marital status from loan applicants table.

The various plots used to visualize uni-variate data typically are Bar Charts, Histograms, Pie Charts. etc.

Bi-variate Descriptive Statistics: Bi-variate analysis involves the analysis of two variables for the purpose of determining the empirical relationship between them. The various plots used to visualize bi-variate data typically are scatter-plot, box-plot.

Scatter Plots: The simplest way to visualize the relationship between two quantitative variables , x and y. For two continuous variables, a scatter-plot is a common graph. Each (x, y) point is graphed on a Cartesian plane, with the x axis on the horizontal and the y axis on the vertical. Scatter plots are sometimes called correlation plots because they show how two variables are correlated.

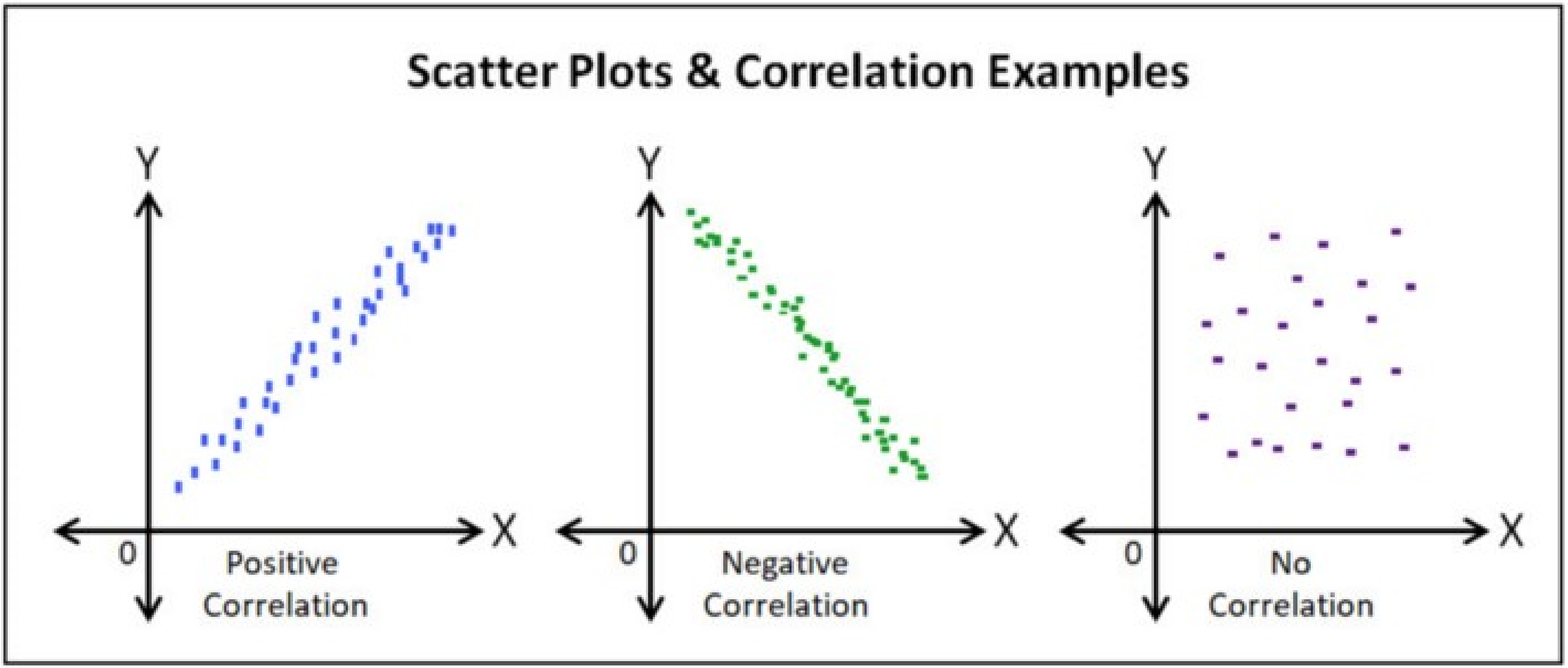
Correlation: A correlation is a statistic intended to quantify the strength of the relationship between two variables. The correlation coefficient r quantifies the strength and direction of the linear relationship between two quantitative variables. The correlation coefficient is defined as:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{(n-1)\,s_x s_y}$$

where sx and sy represent the standard deviation of the x-variable and the y-variable, respectively. −1 ≤ r ≤ 1.

If r is positive and significant, we say that x and y are positively correlated. An increase in x is associated with an increase in y.

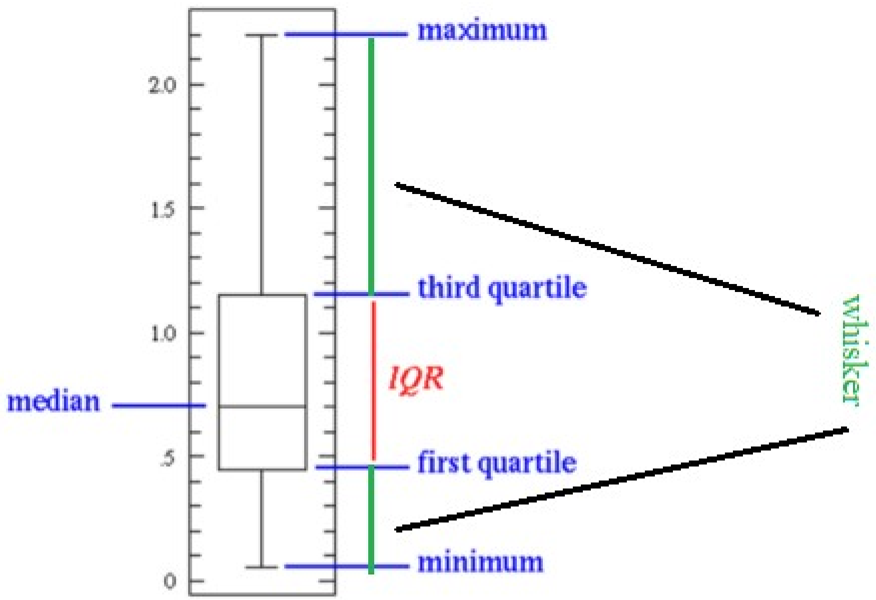
If r is negative and significant, we say that x and y are negatively correlated. An increase in x is associated with a decrease in y.



Positive correlation (r > 0), Negative correlation (r < 0), No correlation (r = 0)

Box Plots

A box plot is also called a box and whisker plot and it's used to picture the distribution of values. When one variable is categorical and the other continuous, a box-plot is commonly used. When you use a box plot you divide the data values into four parts called quartiles. You start by finding the median or middle value. The median splits the data values into halves. Finding the median of each half splits the data values into four parts, the quartiles.



Box plots are especially useful for indicating whether a distribution is skewed and whether there are potential unusual observations (outliers) in the data set.

The five-number summary of a data set consists of the minimum, Q1, the median, Q3, and the maximum.

Thank You !