BASIC STATISTICS

1.) Basic Concepts:

Statistics: is a science that analyzes information variables (for instance, population age, height of a basketball team, the temperatures of summer months, etc.) and attempts to extract conclusions based on the behavior of these variables. Statistics is one of the sciences that allow us to know, or at least to understand, the reality in which we live. Through statistics, we can obtain very valuable information that will help us to make decisions regarding any aspect of our life. The purpose of statistics is to analyze past information to help us make decisions for the future.

Random variable: A set of the different numerical values that adopt a quantitative character. It is the piece of data susceptible to acquire different values in different and specific circumstances. Statistics is the quantitative study of variables, therefore, these values may be considered as the raw material for statistic studies. Any variable that has a specific probability law associated; each of the values that may take has a corresponding specific probability.

The variables may be qualitative or quantitative.

Qualitative variables (or categorical): those variables that are not in numerical form, but appear as categories or attributes (gender, profession, eye color). Quantitative variables: those variables that may be expressed numerically (temperature, salary, numbers of goals in a soccer match, etc.). Quantitative variables may be defined, according to the type of values that represent as:

- **Discrete**: Those values that represent isolated values (natural numbers) and that cannot take any intermediate value between two established consecutive values. For instance; number of goals, number of children, number of bought records, number of heartbeats...
Continuous: Those values that represent infinite values (real numbers) in a given interval, so that they can represent any intermediate value, in theory at least, in their range of variation. For instance; size, weight, blood pressure, temperature...

Frequency: Number of times a datum is repeated. There are two types of frequencies:

- **Absolute frequency**: the absolute frequency of a statistical variable is the number of times that value of the variable appears in the sample.
- **Relative frequency**: Absolute frequency is a measure influenced by the size of the simple. Increasing the sample size also increases the absolute frequency. This correlation makes it a measure not useful to compare. That is why it is necessary to introduce the concept of relative frequency, or the quotient obtained dividing the absolute frequency over the sample size.

The following concepts have to be considered when studying the behavior of a variable:
(Components of a Statistical Study)

Population: is the set of all the elements that possess certain properties and are the elements desired to study a particular phenomenon (homes, number of screws manufactured yearly in a plant, flipping a coin, etc.). Statistic population or universe is the reference set used to make the observations.

Individual: a statistical unit, or individual, is each of the elements that make up the statistic population. The individual is an observable entity that does not have to be a person. It can be an object, a living being and even an abstract concept.
**Sample:** is the population subset under study used to extract conclusions regarding the characteristics of the population. The sample must be representative, in the sense that the conclusions obtained from it must be applicable to the entire population. Samples can be probabilistic or non probabilistic. A **probabilistic sample** is chosen by means of mathematical rules, and therefore the probability of selecting each of the units is known in advance. A **non probabilistic sample** is not ruled by mathematical probability rules, and therefore, while it is possible to calculate the size of the sample error when working with probabilistic samples, it is not possible to do so with the non probabilistic samples.

The more basic probabilistic sample is the **simple random simple**, in which all the components or units of the population have the same opportunities to be selected.

**Census:** We say we are conducting a census when we are observing all the elements that make up the statistic population.

**Parameter:** is a characteristic of a population, summarized for its study. It is considered a true value of the characteristic under study.
**Probability:** Is the set of possibilities that an event occurs or not at a given time.

These events may be measurable in a scale from 0 to 1 (the scale can also be expressed in percentages ranging from 0% and 100%), where the event that cannot occur has an assigned probability of 0 and an event that can occur with certainty has an assigned probability of 1, and the remaining events will have assigned probabilities between “cero and one”, that will be greater the greater the probability of occurrence is.

Example: When flipping a coin we wish to know what is the probability of it falling as heads or as tails, that is, there is a 0,5 (50%) of it being heads or of 0,5 (50%) being tails.

The experiment must be random, that is, several results may occur within a possible set of solutions, and this must be true even when doing the experiment under the same conditions. Therefore, we do not know a priori which event will occur.

Example: Christmas lottery.

There are experiments that are not random and therefore the laws of probability cannot be applied to them.

**Probability distribution model:** specification of the values of the random variable with their respective probabilities.

2.) **Measures of random variables**

Oftentimes it is more expeditious, easy and precise, to study a variable using numerical values than the visual description of a variable by means of tables and graphics, since numerical values give us an idea of the location or of the center of the data (position measures), and using quantities that inform us about the concentration of the observations around said center (dispersion or variability measures)

**a) Measures of central tendency:**

Carry information about the middle values of the data series. A central tendency measure is a value representative of a set of data and that tends to be positioned, according to its magnitude, in the center of the data set a

**Mean:** is the weighted average value of the data set of values that the statistical value represents. The mean is the sum of all the variables divided over the total number of available data. The mean is calculated utilizing the following formula;
If the \( x_i \) value of the \( X \) variable is repeated a \( n_i \) number of times, this is expressed in the arithmetic mean formula as:

\[
\bar{X} = \frac{x_1 + x_2 + x_3 + \ldots + x_{n-1} + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}
\]

Where \( x_i \) are the variables, \( n_i \) the times the variable \( x_i \) appears, and \( N \) the sum of all the \( n_i \). That is:

\[ N = \sum n_i \]

The arithmetic mean is also called the distribution’s center of gravity.

**Median:** Is one of the most representative calculations of the sample. The median is the value of the intermediate element once all the elements have been ordered. The median is calculated ordering the data in increasing order and taking the value positioned in the middle, that is, the value that has 50% of observations on the left and 50% on the right.

Its location is established dividing the number of values by 2: \( \frac{n}{2} \)

When there is an odd number of values for the variable, the median will be, precisely, the central value, the value whose cumulative absolute frequency coincides with the expression \( \frac{n}{2} \). Therefore the median coincides with one value of the variable.

The problem arises when there is an even number of values for the variable. If the result of \( \frac{n}{2} \) is a value lower than the cumulative absolute frequency, the value of the median will be the variable which absolute frequency fulfils the following condition: \( N_{i-1} < \frac{n}{2} \leq N_i \Rightarrow Me = x_i \).
However, if the value is: \( \frac{N}{2} = n \), to obtain the median we will have to use the following formula: 

\[
Me = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2} + 1}}{2}
\]

Legend: median

**Mode**: Is the most frequent value of the statistical variable and the highest value of the histogram.

Example: The mode for the set 2,2,5,7,9,9,9,10,10,11,12 and 18 is \( = 9 \). Example: The set 3,5,8,10,12,15 and 16 does not a mode. Example: The set 2,3,4,4,4,5,5,7,7,7 and 9 has two modes, 4 and 7. It is a bimodal set.

A distribution with one sole mode is called *unimodal*.
b) Other measures:

These measures carry information regarding the manner of distribution of the remaining values of the data series.

Los Quantiles (quartiles, deciles, percentiles) are localization measures. They carry information regarding the value of the variable that will occupy the position (expressed as a percentage) we are calculating within the whole set of variables.

We can say that the quantiles are positioning measures that divide the distribution in a given number of parts so that each of them contains the same value of the variable.

The more important are:

QUARTILES, divide the distribution in 4 equal parts (three divisions). Q₁, Q₂, Q₃, corresponding to 25%, 50%, 75%.

DECILEES, divide the distribution in 10 equal parts (9 divisions). D₁, ..., D₉, corresponding to 10%, ..., 90%.

PERCENTILEES, when they divided the distribution in 100 parts (99 divisions). P₁, ..., P₉₉, that correspond to 1%, ..., 99%.

There is a value in which the quartiles, the deciles and the percentiles coincide when they are equal to the Median, such as:

\[
\frac{2}{4} = \frac{5}{10} = \frac{50}{100}
\]

Quartiles: The quartiles are the three values that divided the ordered data set in four, percentually equal, parts.

There are three quartiles, usually represented by Q₁, Q₂, Q₃:

The first quartile, Q₁, has the lowest value that is greater than a one fourth of the data; that is, the variable’s value that is greater than 25% of the observations and is smaller than the 75% of the observations.

The second quartile, Q₂, (that coincides, it is identical or similar to the median, Q₂ = Md), is the lowest value that is greater than half of the data, that is 50% of the observations have a greater value than the median and 50% have a lower value.

The third quartile, Q₃, has the lowest value that is greater than three fourths of the data, that is, the value of the variable that has a value greater 75% of the observations and of a lower value than 25% of the observations.
**Deciles:** The deciles are nine numbers that divide the succession of ordered data in ten, percentually equal, parts. They are also a particular case of percentiles, since a decile can be defined as “a percentile in which the value that indicates its proportion is a multiple of ten. Percentile 10 is the first decile; percentile 20 is the second decile, etc”.

The first decile $D_1$: indicates there is only a 10% probability for the variable’s value to be below said figure.

The fifth decile $D_5$, also called “Base Case”, also indicates there is 50% probability for the value to be above as for the value to be below this figure. It represents the Median of the distribution.

**Percentiles or centiles:** The percentiles are, perhaps, the most utilized measures for location or classification purposes (in the case of people when the characteristics are weight, height, etc.).

The percentiles are numbers that divided the succession of ordered data in one hundred, percentually equal, parts. These are the 99 values that divided in one hundred equal parts the set of ordered data. The Percentile is, simple, the value of the trajectory of a variable, that encompasses a specific proportion of the population.

The percentiles ($P_1$, $P_2$, ..., $P_{99}$), read as first percentile,..., percentile 99, show the variable that leaves behind a cumulative frequency equal to the percentile’s value:

The first percentile is greater than one percent of the values and lower than the remaining ninety-nine.

The percentile 60 is the value of the variable that is greater than 60% of the observations and lower than 40% of the observations.

The 99 percentile is greater than 99% of the data set and is lower than the remaining 1%.

c) **Dispersion measures:**

Those measures that allow us to relate the distance of the variable’s values to a given central value, or that allow us to identify the concentration of data in certain sector of the trajectory of the variable. They study the distribution of the values of the series, analyzing if said values are more or less concentrated or more or less disperse.

**Range:** Measures the amplitude of the sample’s values. It is calculated as the difference between the highest and the lowest value.

$$R_e = x_{\text{max}} - x_{\text{min}}$$
**Variance:** Measures the distance between the values in the series and the mean. It is the sum of the square of the differences between each value and the mean, multiplied by the number of times each value has repeated. The result obtained is then divided over the sample size.

\[ S_x^2 = \sigma_x^2 = \frac{\sum (x_i - \bar{x}) \cdot n_i}{N} \]

The variance will always be greater than zero. The closest it is to zero the more concentrated are the values of the data series around the mean. The greater the variance, the more dispersed the values.

**Standard deviation:** is the square root of the variance. It expresses the dispersion of the distribution and it is expressed in the same units of measurement as the variable. The standard deviation is the most utilized measure of dispersion in statistics.

\[ \sigma = \sqrt{\text{var}(X)} = +\sqrt{\text{var}(X)} \]

3.) Distributions of probability

As mentioned before, a **random variable** is the variable that can represent different values, or set of values, with different probabilities. Random variables have 2 important characteristics: its values and the probabilities associated to these values.

A table, graphic, or mathematical expression that shows the probabilities each random variable has of adopting different values is called a **probability distribution of the random variable**.

The statistical inference (that is, the process done by the “Riskmeter”) relates to the conclusions that may be extracted from a population of observations based on an observation simple; in this case we wish to know something about a distribution based on a random sample of said distribution.

In this manner we see we are working with **random samples of a population** that is larger than the obtained simple; said isolated random simple is nothing more than one of the many different samples that could have been obtained through the selection process. That is why using the **distributions of probability** has such relevance.
**Discrete distributions:**

Are those distributions in which the variable can adopt a specific number of values. The most noteworthy distributions amongst the existing ones are:

**Bernouilli;** the model followed by an experiment that is done only once and can have two solutions: true or false:

When the solution is **true** (success) the variable equals 1

When the solution is **false** (failure) the variable equals 0

Because there are only two possible solutions they are complementary events:

- The probability of success is called "p"
- The probability of failure is called "q"

When: \[ p + q = 1 \]

The Bernouilli distribution is then applied to experiments that are done one time only and have two possible results, failure or success, and hence the variable can only have two values: 1 or 0.

Example: flipping a coin.

**Binomial;** the binomial distribution is based on the Bernouilli distribution. It is applied when the Bernouilli experiment is done an "n" number of times, each of the assays being independent from the previous one. The variable then can adopt values between:

- 0: if all the experiments have been failures
- n: if all the experiments have been successes

Example: flipping a coin repeatedly.

**Poisson;** the Poisson distribution is based on the binomial distribution.

The Poisson distribution is applied in the cases when using a binomial distribution the experiment is done a high "n" number of times and the probability of success "p" per assay is low. The following condition must be met:

\[ p < 0,10 \]

\[ p \times n < 10 \]

Example: number of errata per page in a book
**Continuous distributions:**

Are those that present an infinite number of possible solutions.

Types of distributions:

**Uniform:** a distribution that may adopt any value within an interval (all the values have the same probability).

Characteristics:

- All the possible values the variable may adopt, located between the maximum and minimum quantities, present the same possibilities of being reached.
- The entrepreneur identifies a value range for the variables.
- Exogenous variables.
- Function parameters the entrepreneur can identify and quantify.

**Normal:** It is used to measure and represent many variables such as weight, height, exam scores..., in which the distribution is symmetrical from a central value, around which it takes values with a great probability of existing with hardly any extreme values.

It is the most used distribution model. The importance of the normal distribution is mainly due to the many variables associated to natural phenomena that follow the normal distribution model (sizes, weights, breadth, consumption of a given product, exam scores, degree of adaptation to a environment, etc.).

This distribution is also characterized by the arrangement of the values in a bell shape called **Gaussian distribution**, around a central value that coincides with the middle value of the distribution.

Of the total values, 50% are located to the right of this central value and the other 50% are to the left.
This distribution is defined by two parameters:

\[ X: N(\mu, \sigma^2) \]

represents the mean value of the distribution and it is precisely the value at the center of the curve (in the Gaussian distribution).

\[ \sigma^2 \]: is the variance. It indicates if the values are more or less near the central value: a low variance value indicates the values are close to the mean; if it is high it indicates the values are very dispersed.

When the distribution mean equals 0 and the variance equals 1 is called a "standard normal distribution". The advantage of using this distribution is that there are tables with the cumulative probability for each point of the curve of this distribution.

Characteristics:

- Pre-established minimum
- Pre-established maximum
- All values between the minimum and maximum values of the distribution are equally probable.

**Triangular**; the triangular distribution is useful for an initial approximation in situations in which we do not have reliable data. It allows us to estimate the duration of the activities of a Project using three estimation degrees: optimistic, very pessimistic and pessimistic.
Characteristics:

- A distribution function commonly applied to sales variables and market costs.
- Endogenous variables, the entrepreneur can use them to negotiate.
- The entrepreneur can identify and quantify the function parameters.

Practical example:

<table>
<thead>
<tr>
<th>Value of the variable ($X_i$)</th>
<th>Absolute frequency</th>
<th>Relative frequency</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.20</td>
<td>5</td>
<td>5/20 = 25%</td>
<td>5</td>
</tr>
<tr>
<td>1.7</td>
<td>4</td>
<td>4/20 = 20%</td>
<td>9</td>
</tr>
<tr>
<td>2.35</td>
<td>3</td>
<td>3/20 = 15%</td>
<td>12</td>
</tr>
<tr>
<td>2.01</td>
<td>7</td>
<td>7/20 = 35%</td>
<td>19</td>
</tr>
<tr>
<td>0.94</td>
<td>1</td>
<td>1/20 = 5%</td>
<td>20</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>20</strong></td>
<td><strong>100%</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Measures of central tendency:**

Mean $\longrightarrow \quad X = \frac{(1.20*5)+(1.7*4)+(2.35*3)+(2.01*7)+(0.94*1)}{20} = 1.743$

Median $\longrightarrow \quad$ Ordering the set: 0.94 – 1.20 – 1.7 – 2.01 – 2.35

$\quad$ Med. = 1.7

Mode $\longrightarrow \quad$ Mode = 2.01 (the most often repeated value considering it has the highest frequency)
Other measures

(P75 = Third quartile)

\[
P_{75} = \frac{3 \times n}{4} = \frac{3 \times 20}{4} = 15
\]

Observing the table of cumulative frequency we realize that for X_i = 2.01, 75% observations are below and 25% are above.

Dispersion measures:

Variance

\[
\sigma^2 = \frac{[(1.2-1.743)^2 \times 5] + [(1.7-1.743)^2 \times 4] + \ldots + [(0.94-1.743)^2 \times 1]}{20} = 0.586
\]

Standard deviation

\[
\sigma = 0.7655
\]

Range

\[
R = 2.35 - 0.94 = 1.41
\]