

Teaching statistics and technology through English

Lourdes Emperatriz Paredes Castelo
Johanna Enith Aguilar Reyes
Geoconda Marisela Velasco Castelo
María Yadira Cárdenas Moyano

Teaching statistics and technology through English

Lourdes Emperatriz Paredes Castelo
Johanna Enith Aguilar Reyes
Geoconda Marisela Velasco Castelo
María Yadira Cárdenas Moyano

Este libro ha sido debidamente examinado y valorado en la modalidad doble par ciego con fin de garantizar la calidad científica del mismo.

© Publicaciones Editorial Grupo Compás
Guayaquil - Ecuador
compasacademico@icloud.com
<https://repositorio.grupocompas.com>



Paredes, L., Aguilar, J., Velasco, G., Cárdenas, M. (2024) Teaching statistics and technology through English. Editorial Grupo Compás

© Lourdes Emperatriz Paredes Castelo
Johanna Enith Aguilar Reyes
Geoconda Marisela Velasco Castelo
María Yadira Cárdenas Moyano
Escuela Superior Politécnica de Chimborazo (ESPOCH)

ISBN: 978-9942-33-819-8

El copyright estimula la creatividad, defiende la diversidad en el ámbito de las ideas y el conocimiento, promueve la libre expresión y favorece una cultura viva. Quedan rigurosamente prohibidas, bajo las sanciones en las leyes, la producción o almacenamiento total o parcial de la presente publicación, incluyendo el diseño de la portada, así como la transmisión de la misma por cualquiera de sus medios, tanto si es electrónico, como químico, mecánico, óptico, de grabación o bien de fotocopia, sin la autorización de los titulares del copyright.

INTRODUCTION TO STATISTICS

The origin of statistics is as old as civilization, it reached an important development with the emergence of the States, in this event it became a decisive instrument. The conception of statistics has evolved throughout history, where initially it was limited to the collection and ordering of data on aspects of interest, in this context statistics became closely linked to the theory of probabilities, becoming a branch of applied mathematics, using mathematical principles and models applicable in all areas.

The field of statistics dates back to 1654, with roots in gambling, it has developed in the field of study of methods and tests to quantitatively define the variability in data, the probability of results, the error and uncertainty associated with the results, in this context statistical methods are used extensively in the scientific process from the design of research questions, data analysis and the final interpretation of results.

In general, the scientific researcher designs the research studies established in the nature of the arguments to be investigated, polishes the research plan according to the concepts of statisticians in order to increase the possibility that the findings will be useful.

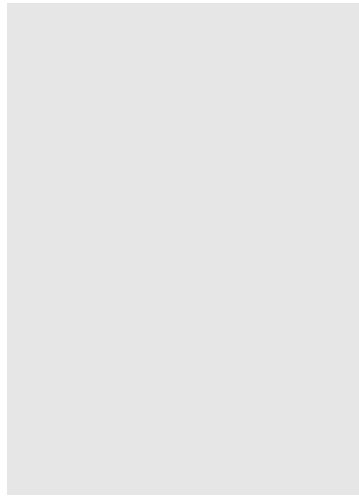
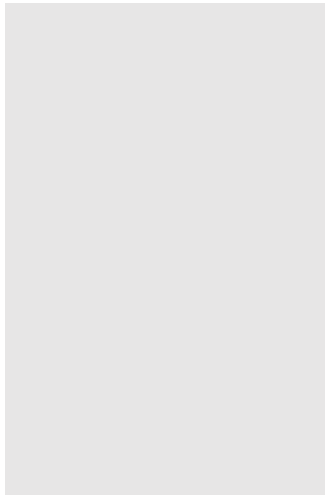
Current statistics is the consequence of the alliance of two disciplines that have unfolded, the first is probabilistic calculus, whose origin dates back to the 19th century as the theory of games of chance, the second is statistics or state science, which describes data; the integrations of these two lines of thought give rise to the science that studies how to generate conclusions from empirical research through the use of mathematical models.

CREATE YOUR KNOWLEDGE

1. READ AND COMPLETE THE CHART

STATISTICS	HAS DEVELOPED	

STATISTICS IS THE ALLIANCE OF



2. ANSWER THESE QUESTIONS

a. What is probabilistic calculus?

b. What is probabilistic statistics or state science?

3. WRITE DOWN A GLOSSARY

1	_____
2	_____
3	_____
4	_____
5	_____
6	_____
7	_____
8	_____
9	_____
10	_____
11	_____
12	_____
13	_____
14	_____
15	_____

DEFINITION OF STATISTICS

Statistics is considered as a scientific method, applicable to a great variety of areas of knowledge, whose usefulness is transcendental, because it goes beyond the simple description, discovering laws and tendencies, one of the main examples that can be cited is the case of the statistician Ernest Engel (1821-1896).

Statistics is a mathematical discipline whose purpose is focused on the **interpretation** of **numerical data** that are detached from **empirical events**, it is responsible for the study of **random events or experiments**, it collects and organizes a large number of data with the purpose of obtaining some **consequence** (Martínez, 2020, Matus,2010).

Statistics studies scientific methods for collecting, organizing, summarizing and analyzing information, as well as for obtaining valid conclusions and making reasonable decisions based on such analysis (Murray & Spiegel, 2005), and represents an area of science that deals with the design of experiments, data analysis and inferences about the population from the information generated in a sample.

Authors such as Hernández and Oteyza (2015), Guerra (2003) and García et al., (2002), mention statistics as a discipline in charge of collecting, organizing, describing and interpreting data, in this context it is concluded that statistics is a fundamental tool of numerical analysis that allows generating knowledge about an event.

With the aforementioned, statistics provides basic tools for research, this discipline contributes to the discovery of relationships between facts and the basis of these discoveries, it is a way of approaching the knowledge of reality.

2. READ THE SECOND PARAGRAPH EXPLAIN THE MEANING OF THE WORDS IN BLOND. GIVE EXAMPLES

WORD	MEANING	EXAMPLES

<i>Statistics studies scientific method</i>	TO	

3. WRITE DOWN A GLOSSARY

- 1 _____
- 2 _____
- 3 _____
- 4 _____
- 5 _____
- 6 _____
- 7 _____
- 8 _____
- 9 _____
- 10 _____
- 11 _____

12

13

14

15

IMPORTANCE OF STATISTICS IN SCIENTIFIC RESEARCH

Statistics is the usual link that is present in almost all scientific research, in which the treatment, interpretation and prediction of data is involved; the statistical study tends to become especially important when carrying out research in scientific fields such as medicine, where a bad interpretation generates adverse consequences for the population.

In the researcher's work, statistics is a fundamental support tool in the investigation of phenomena, it is logic with a strong ingredient of arithmetic processes, which creates material on which inference is based and uncertainty is measured, statistics contributes to the research process in the design stage, data collection plan, analysis of results up to the evaluation of the uncertainty associated with the inference drawn from them.

Statistics, being a tool for data analysis and interpretation, has acquired relevance in all areas of work, being a clear factor in predictions and decision making based on data, the most relevant task of statistics is to provide quantitative alternatives, which translate into objective conclusions, in this sense, statistical techniques allow the scientific researcher to quantify the probability of occurrence of a phenomenon.

TASK 1. READ AND COMPLETE THE CHART



CLASSIFICATION OF STATISTICS

A particularity of practical statistical work is the analysis or processing of numerical data; any conclusion generated by a statistical procedure necessarily involves the analysis of quantitative quantities or characteristics, which is why it is important to use statistics as an auxiliary tool that contributes to the research process.

At present, a wide range of statistical techniques have been developed for data analysis, generally divided into two groups:

- **Descriptive statistics:** it is in charge of gathering, presenting and organizing data, it allows the scientist to adhere the most significant properties of a data set, applying measures such as average, mean, standard deviation; these measures provide a general sense of the study group. The essential objective of this type of statistics is the characterization of numerical data set, which highlights the properties of the data set (Cardenas, 2014).
- **Statistical Inference:** is concerned with analyzing data from a sample, in order to draw conclusions from a study population, it is used to model patterns in data, make judgments about data, identify relationships between variables in the data set and make inferences about broader populations based on a sample of data, Figure 1- 1 shows the process of statistical inference.

1. COMPLETE THE CHART WITH SIMILARITIES AND DIFFERENCES



2. WRITE DOWN THE EXAMPLES IN THE CORRECT COLUMN

Year-over-year pricing changes, month-over-month sales growth, the number of users, or the total revenue per subscriber, if you say A and B are independent but and you use statistical methods to reject the null hypothesis that says A is dependent on B, clinical trials use inferential statistics to determine whether a new drug is effective in treating a particular medical condition.

Descriptive statistics	Statistical Inference

3. WRITE DOWN A GLOSSARY

1	_____
2	_____
3	_____
4	_____
5	_____
6	_____
7	_____
8	_____
9	_____
10	_____
11	_____
12	_____
13	_____
14	_____
15	_____

DATA

A data is an element of a whole set that has a value or peculiarity with which it is distinguished from the others (Hernández & Oteyza, 2015), the number of data can be counted, however; *the values that data acquire are not always expressed through a numerical measure, but by an attribute*, example: colors (blue, red, yellow, etc), level of education (elementary, basic, secondary, higher, postgraduate), in this context the data are classified into: quantitative and qualitative.

- **Quantitative Data:** are those that can be counted or measured through a numerical expression, e.g.: ages of students (can be expressed in years 19 years, 25 years, etc.), the price of a good or services (expressed in monetary value 19 USD, 35.5 USD). They are subdivided into continuous data that can take any value of the real numbers (height of a person 1.70 meters) and discrete when they only take integer values (5 tables, 3 chairs).

- **Qualitative data:** data whose values cannot be quantified, but express a level of quality or indicate an attribute by which each element can be identified, i.e., data that can be identified by the following criteria identify them by gender (male, female), marital status (single, married, widowed, divorced, common-law)

1. EXPLAIN THIS SENTENCE IN YOUR OWN WORDS A data is an element of a whole set that has a value or peculiarity with which it is distinguished from the others

DATA: _____

ELEMENT: _____

VALUES: _____

MY DEFINITION:

2. COMPLETE THE ALPHABET USING WORDS RELATED WITH DATA

A _____

C _____

E _____

G _____

I _____

K _____

M _____

O _____

Q _____

S _____

U _____

W _____

B _____

D _____

F _____

H _____

J _____

L _____

N _____

P _____

R _____

T _____

V _____

Y _____

3. IDENTIFY THE FOLLOWING MEASURES AS EITHER QUANTITATIVE OR QUALITATIVE:

- a. The blood pressure in (mmhg).
- b. The number of times a child brush his/her teeth.
- c. Whether or not someone fail in an exam.
- d. Weight of babies at birth.
- e. The time to run a certain distance.
 - The 30 high-temperature readings of the last 30 days.
 - The scores of 40 students on an English test.
 - The blood types of 120 teachers in a middle school.
 - The last four digits of social security numbers of all students in a class.
 - The numbers on the jerseys of 53 football players on a team.
 - The genders of the first 40 newborns in a hospital one year.
 - The natural hair color of 20 randomly selected fashion models.
 - The ages of 20 randomly selected fashion models.
 - The fuel economy in miles per gallon of 20 new cars purchased last month.
 - The political affiliation of 500 randomly selected voters.

4. WRITE DOWN A GLOSSARY

1	
2	_____
3	_____
4	_____
5	_____
6	_____
7	_____
8	_____
9	_____
10	_____

11 _____

12 _____

13 _____

14 _____





15 _____

DATA COLLECTION TECHNIQUES

Data collection refers to a systematic approach to gathering and measuring information from different sources, allowing the researcher to answer relevant questions, evaluate results and better anticipate trends.

There are different data collection techniques, the choice of technique depends on the type of variable, the desired accuracy, the point of collection and the skills of the interviewer.

1. READ AND COMPLETE

TECHNIQUE	MEANING	CHARACTERISTICS
 Observations		
 Surveys		
 Interviews and Focus Groups		
 Forms		

 <p>Online Tracking</p>		
--	--	--

1. **Observation:** allows to know the behavior of the object of study directly, the most appropriate way to apply this technique is to record the observations in field notes or on a platform. This technique is characterized by not being instructive and requires evaluating the behavior of the object of study continuously without intervening.

2. **Surveys:** this technique consists of obtaining data directly from the subjects of study, in order to achieve the desired results it is essential to have clear research objectives, it is also important to develop the questionnaires to apply the surveys carefully, defining what type of questionnaire is the most efficient for data collection, in this sense the most popular questionnaires are:
 - a. *Open-ended questionnaires: they are applied to get to know in depth the person's perspective on a specific topic, analyze his/her opinions and obtain detailed information.*
 - b. *Closed Questionnaires: they are applied to obtain a sufficient amount of information, the answers of the respondents are limited, they may contain multiple choice or dichotomous questions (yes/no).*

3. **Focus group:** this is a qualitative technique that consists of holding a meeting where people give their opinion and seek to solve a specific problem. One of the qualities of this technique is the possibility of obtaining several points of view on the same subject in order to arrive at an optimal solution.
4. **Interview:** this method consists of gathering information by clarifying a question, through interpersonal communication, the sender has verbal answers from the receiver on a specific topic or problem.

2. USE THE INFORMATION TO WRITE ABOUT SURVEY

TECHNIQUE	KEY FACTS	EXAMPLES
-----------	-----------	----------

<p>SURVEY (mail, self-completion, online, email)</p>	<ol style="list-style-type: none"> 1. Responses can be analyzed with quantitative methods (numerical values to Likert-type scale) 2. Generally easier to analyze. 3. Pre-test/Post- test can be compared and analyzed. 	<p>A satisfaction survey or opinion survey: A needs analysis survey; Market research.</p>
---	---	---

3. WRITE MORE EXAMPLES

<i>Open-ended questionnaires</i>	<i>Closed Questionnaires</i>

4. WRITE DOWN A GLOSSARY

1 _____

2 _____

3 _____

4 _____

5 _____
6 _____
7 _____
8 _____
9 _____
10 _____
11 _____
12 _____
13 _____
14 _____
15 _____

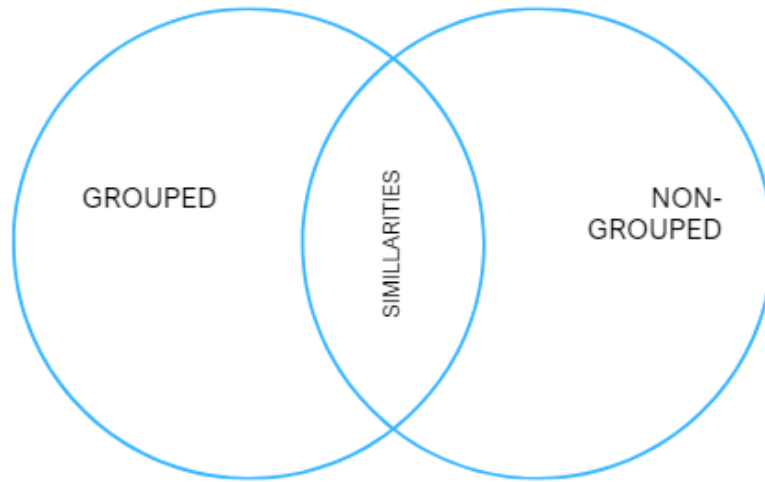
DATA TYPES

In statistics, data can be differentiated into grouped and non-grouped, in statistics there are no established standards to define the appropriate use of grouped or non-grouped data, however; the suggestion is that when the total data (N) is equal to or greater than 20, a grouped data distribution should be used.

- **Grouped data:** data whose main characteristic is the frequency with which they are presented, i.e., data that are counted and classified, either by their quantitative or qualitative qualities, which is why they can be grouped.
- **Non-grouped data:** data that lack frequency.

1. COMPLETE THE DIFFERENCES BETWEEN GROUPED AND NON-GROUPED DATA

DIFFERENCES

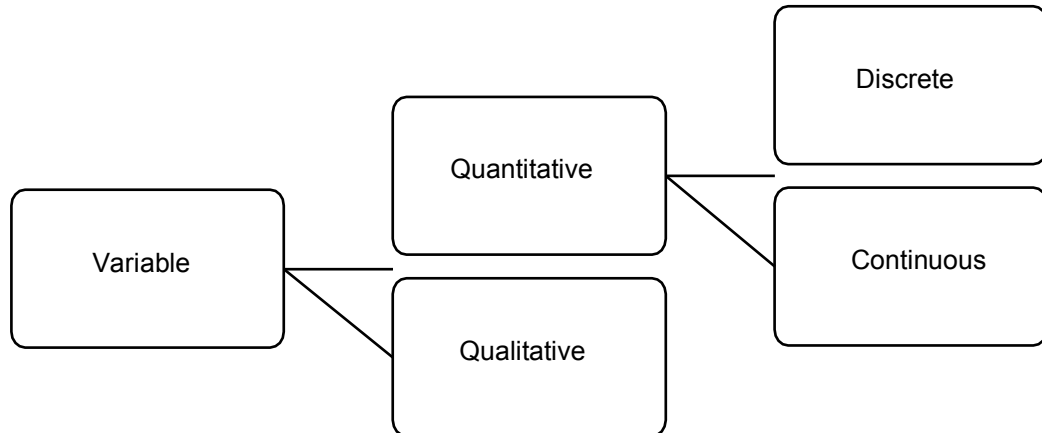


2. WRITE DOWN A GLOSSARY

- 1 _____
- 2 _____
- 3 _____
- 4 _____
- 5 _____
- 6 _____
- 7 _____
- 8 _____
- 9 _____
- 10 _____
- 11 _____
- 12 _____
- 13 _____
- 14 _____
- 15 _____

VARIABLES

A statistical variable represents a characteristic of a sample or population of data that can adopt different values, variables are classified as quantitative and qualitative.



Quantitative Variables

- Continuous variable: they are those that arise through the measurement process, they accept decimal numbers.
- Discrete variable: they arise through the counting process; they accept only integer values.

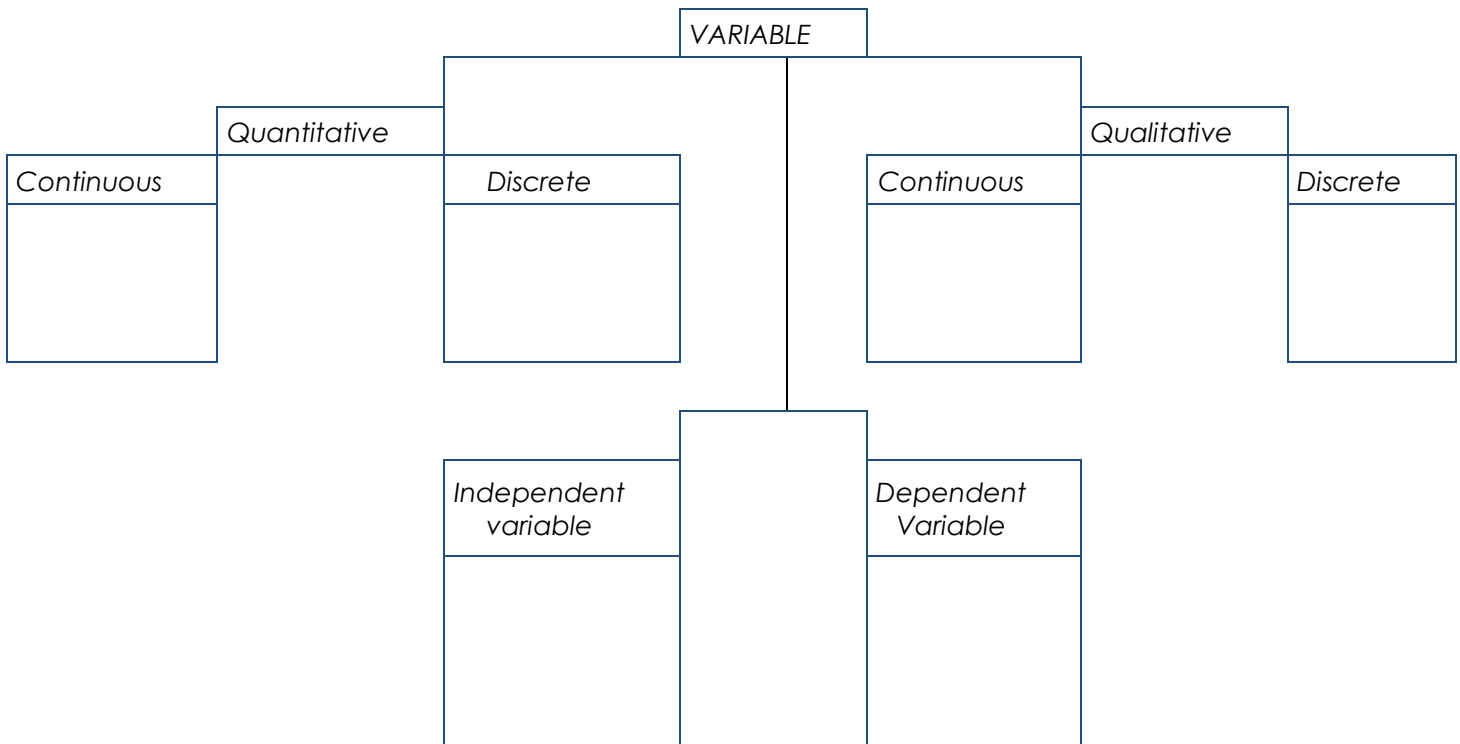
Qualitative Variables

- Continuous variable: they are those that arise through the measurement process, they accept decimal numbers.
- Discrete variable: they arise through the counting process, they accept only integer values.

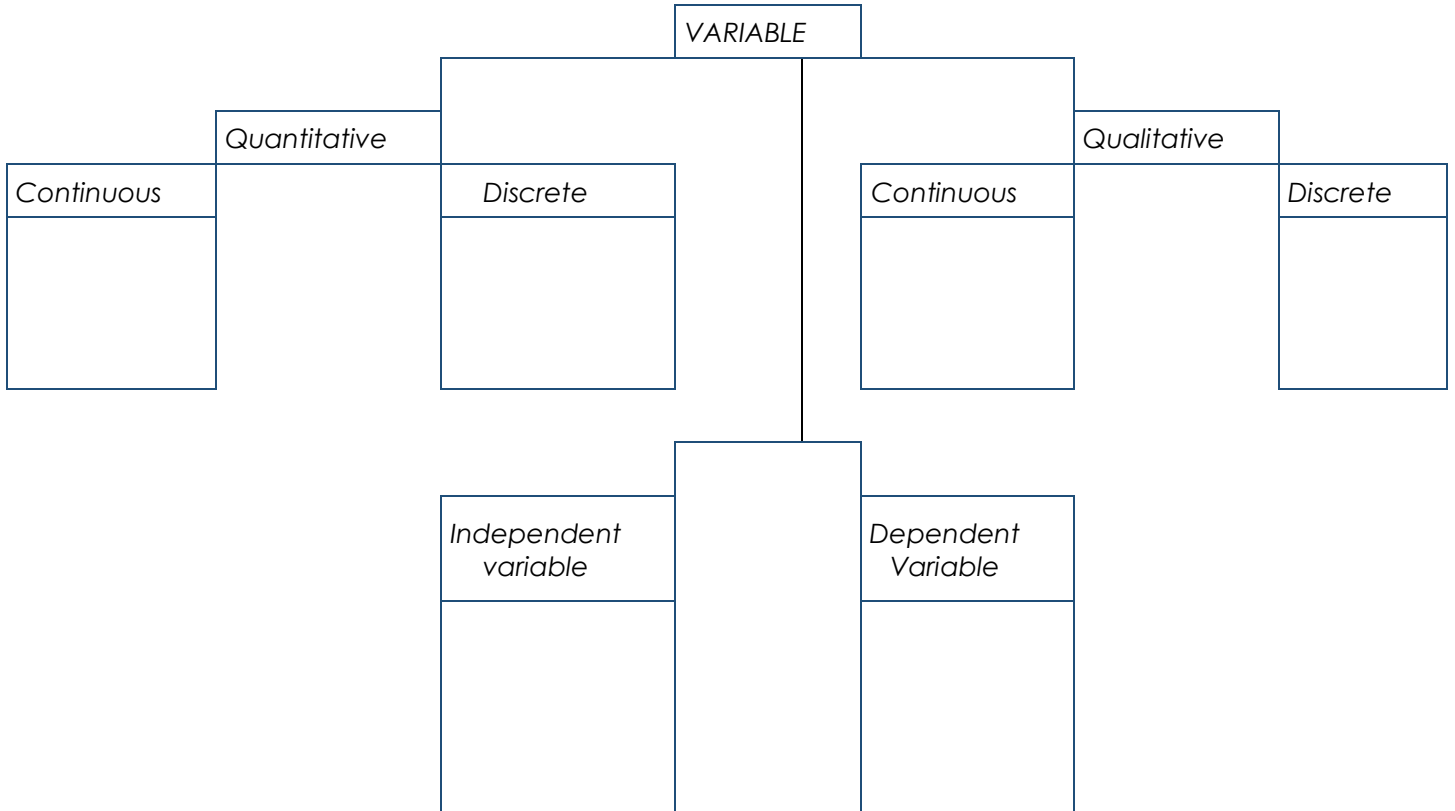
From the perspective of statistical tests, variables are classified into:

- **Independent variable:** it is explanatory of the dependent variable, for example, the scores of a memory test can be explanatory of the normal or impaired cognitive state.
- **Dependent Variable:** is the variable explained by the independent variable, for example, the cognitive state of a person can be explained by a set of independent variables such as age, level of education, social economic level, among others.
- **Intervening Variables:** known as control variables, e.g. the presence of depression may influence the test result, deviating the actual results.

1. READ AND COMPLETE THE CHART



2. WRITE EXAMPLES FOR EACH ONE



3. WRITE DOWN A GLOSSARY

- 1 _____
- 2 _____
- 3 _____
- 4 _____
- 5 _____
- 6 _____
- 7 _____

8	
9	<hr/>
10	<hr/>
11	<hr/>
12	<hr/>
13	<hr/>
14	<hr/>
15	<hr/>

MEASURING SCALES

Data collection requires one of the scales of measurement be it nominal, ordinal, interval, ratio, the scale of measurement determines the amount of information contained in the data and indicates the most appropriate way of summarizing and analyzing the data statistically (Anderson et al., 2012).

In this sense, it can be said that the scales or levels of measurement refer to the relationship between the values assigned to the attributes of a variable.

1. READ THE DEFINITION AND WRITE EXAMPLES

- Nominal scale: when the data of a variable is composed of labels or names used to identify an attribute of the element.

Example 1.1

What is the degree of discomfort with the noise generated by vehicular movement?

- 1. Mild*
- 2. Moderate*
- 3. Severe*

Example 1 _____
What is your gender?

Example 2 _____
¿Do you live in Riobamba?

Example 3 _____

Example 4 _____

Example 5 _____

This subtype of nominal scale is known as a dichotomous nominal scale.

- **Ordinal scale:** when the data exhibit the properties of nominal data and their rank order

Example 1

In rating the service excellent, good or bad, giving each of them a number where 3 is excellent, 2 good, 1 is bad.

What respondents do is choose among satisfaction options, but of course the answer to the question "how much exactly?" remains unanswered. Understanding the various scales of measurement help researchers obtain data that can be applied to advantage in the future. Therefore, an ordinal scale is used as a parameter to understand whether variables are larger or smaller. The central tendency of the ordinal scale is median.

Example 2 _____

Example 3 _____

Example 4 _____

The Likert scale: is an example of why the interval difference between ordinal variables cannot be concluded. In this scale in fact, the response options are usually polar, such as, for example, something like "totally satisfied" or "totally dissatisfied".

Example 1

How satisfied are you with our products?

Fully satisfied

1. Satisfied
2. Neutral
3. Dissatisfied
4. Fully satisfied

<i>EXAMPLE:</i>
1. <i>Satisfied</i>
2. <i>Neutral</i>
3. <i>Dissatisfied</i>
4. <i>Fully satisfied</i>

<i>EXAMPLE:</i>
1. <i>Satisfied</i>
2. <i>Neutral</i>
3. <i>Dissatisfied</i>
4. <i>Fully satisfied</i>

<i>EXAMPLE:</i>
1. <i>Satisfied</i>
2. <i>Neutral</i>
3. <i>Dissatisfied</i>

4. *Fully satisfied*

Interval scale: they present all the properties of ordinal data and the interval between values are expressed in terms of a fixed unit of measurement, interval data are always numerical. The interval scale is the type of question most frequently used in a study or research. To obtain any type of response, it is essential that the question requested requires respondents to answer on a numerical scale where the difference between the two numbers is the same.

Example 1

How was your experience with the food at the ESPOCH restaurant?

Much	Little	Neutral
Spicy		
Boring Pleasant		

Example 2 _____

Example 3 _____

Example 4 _____

Ratio scale: has all the properties of interval data and the ratio of two values is significant. For the measurement of variables such as distance, height, weight and time, the ratio scale is used.

2. WRITE DOWN A GLOSSARY

1	
2	_____
3	_____
4	_____
5	_____
6	_____
7	_____
8	_____
9	_____
10	_____
11	_____
12	_____
13	_____
14	_____
15	_____

POPULATION AND SAMPLE

Population

The population is defined as the total of elements or associated data in a study, set of elements or sets that present a common characteristic (Lind et al, 2012). In statistics, the population is the total of individuals or set of individuals that present a characteristic

trait that is intended to be studied, in this context there are two types of populations:

- **Finite statistical population:** number of values that have an end, e.g. the number of trees in the city of Riobamba, the number of students at the Escuela Superior Politécnica de Chimborazo.
- **Infinite statistical population:** this is a population that has no end, e.g. the number of stars.

The size of the population represents the total number of entities, phenomena, events that conform it, symbolized by the letter N.

1. *WRITE EXAMPLES*

Finite statistical population		Infinite statistical population:

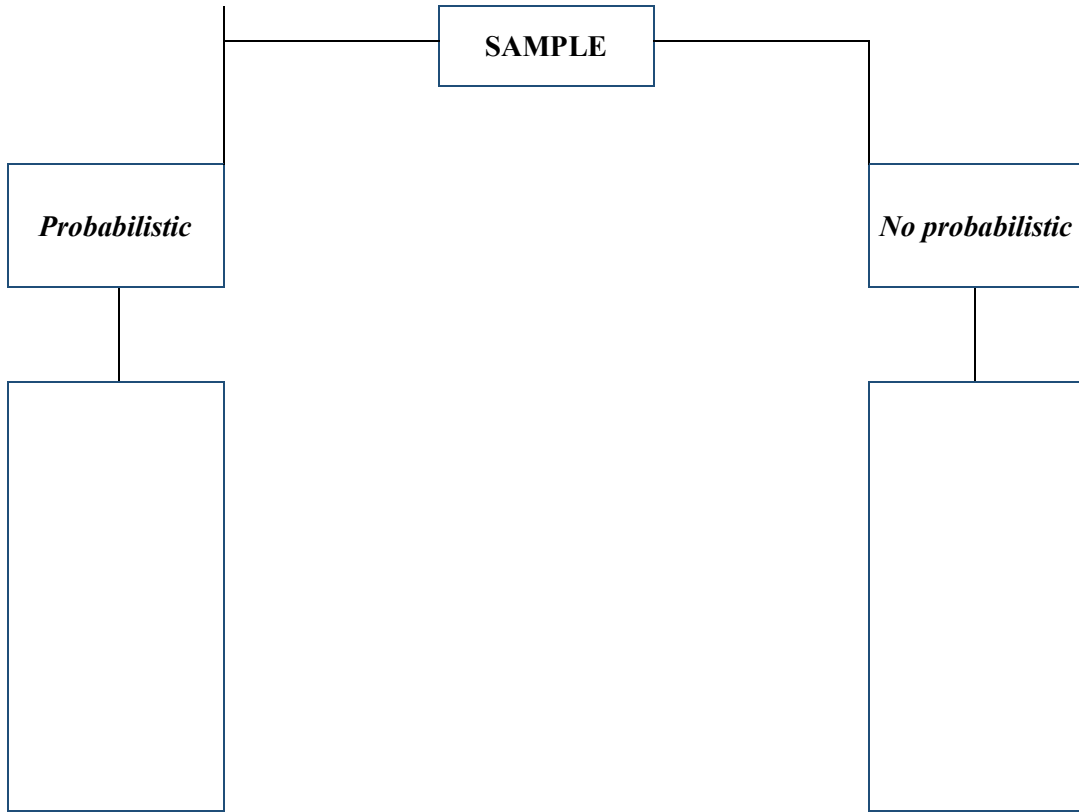
Each of the entities, phenomena or events that make up the population is called an element.

Sample

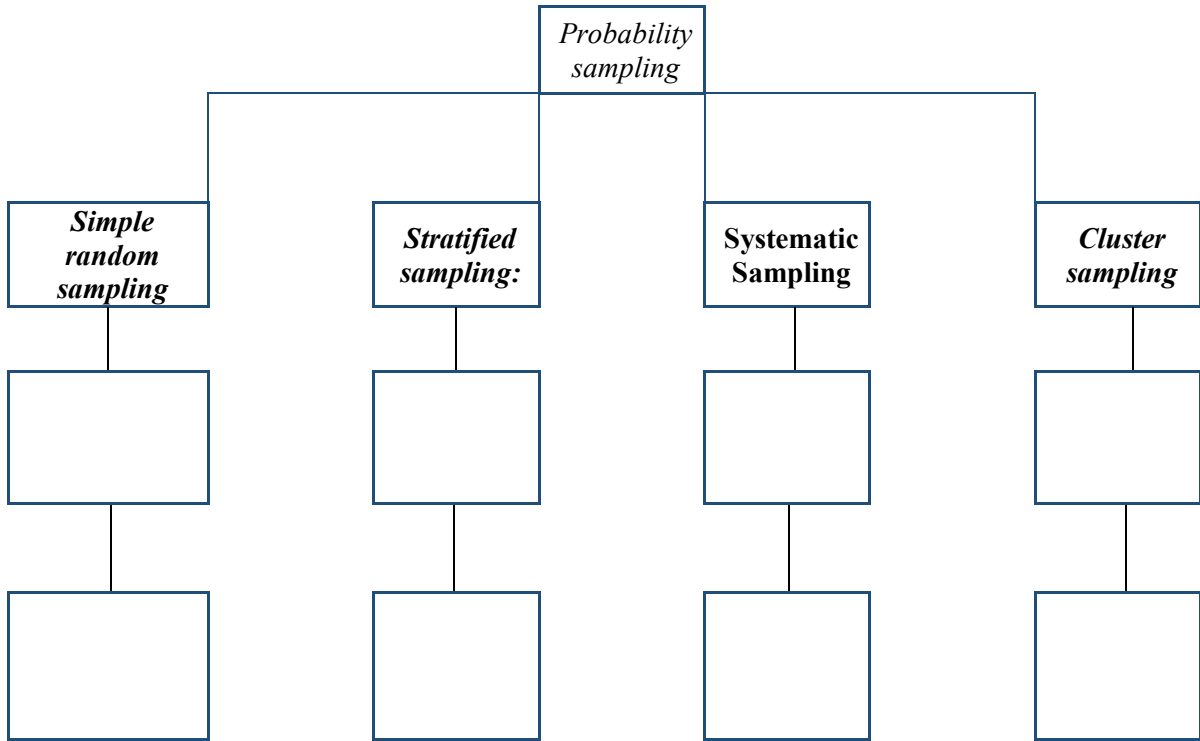
The sample is a portion or part of the population of interest, sampling serves to learn something about a population is often used in administration, agriculture, politics, and government actions, is a subset of the population.

In statistics, the term sample is used to refer to any subset of a population. There are two types of sampling: random or probabilistic and non-probabilistic, each of which includes different types of sampling that distinguish according to the characteristic factors of the population.

2. COMPLETE CHART



Probability sampling is subdivided into four main types:



- a. **Simple random sampling:** is a probability sampling procedure in which each element of the target population has the same probability of being selected; this type of sampling is not used in consumer research.

Example 1

COVID - 19 vaccine coverage in ABC school among 1500 students, sample 60 children, list all children, number them from 1 to 1500 and randomly select 60.

Example 2

Example 3

Example 4

Example 5

- a. Stratified sampling:** a sampling procedure in which the target population is separated into exclusive, homogeneous segments (strata) and then a simple random sample is selected from each segment or stratum. The samples selected from the different strata are combined into a single sample, this procedure is called random quota sampling.

Example 1

To obtain a sample of 100 individuals from a population of 1000, the population is divided into the following strata:

- b.** Draw 1: 300 individuos
- c.** Draw 2: 500 individuos
- d.** Draw 3: 200 individuos

Example 2

Example 3

Example 4

By means of the stratified sampling provided, the sample obtained from each stratum will be representative of each one of them and will give the following results:

Excerpt	Individuals	Percentage	Sample
Excerpt 1:	300	30%	30
Excerpt 2:	500	50%	50
Estrato 3:	200	20%	20

However, this sample cannot be considered completely probabilistic for all strata, since the individuals in the group with the smallest number of people are more likely to be selected for the sample than the other strata.

a. Systematic Sampling: consists of a random selection of the first element for the sample, then subsequent elements are selected using fixed or systematic intervals until the desired sample size is achieved. From a technical point of view, this sampling does not create a truly random sample, only the selection of the first element of systematic sampling is a probability selection, some elements will have a zero probability of selection.

Example 1

Draw a sample of 10 people from a total population of 100 and the first individual selected for the sample is number 3. From this, using an interval of 4 decided by the researcher, the next individuals will be selected until the sample is complete, so that they will be numbers 7, 11, 15, etc.

Example 2

Example 3

Example 4

Example 5

a. Cluster sampling: called cluster sampling, where the elements of the population are randomly selected by groupings (clusters), the elements of the sample are selected from the population individually, one at a time.

The sampling units or groups can be spaced, as occurs in physical or geographical units, e.g. states, provinces, cantons, districts, parishes; based on an organization, higher education institutions, school level; the heterogeneity of the group is

fundamental for a good design of cluster sampling, the elements within each group must be as heterogeneous as the target population itself, the dimensions of this type of sampling are based on the number of stages of the sample design and on the proportional representation of the groups in the sample.

Example 1

Example 2

Example 3

Example 4

3. WRITE DOWN A GLOSSARY

1	_____
2	_____
3	_____
4	_____
5	_____
6	_____
7	_____
8	_____

9	
10	_____
11	_____
12	_____
13	_____
14	_____
15	_____

SAMPLE SIZE

The sample is a selection of respondents chosen to represent the total population, its size translates into a significant representation of the population, which complies with peculiarities related to the research. To determine the size of the sample within an investigation, it is important to take into account the objectives and circumstances in which the research work is carried out.

a. Sample for a finite population

$$n = \frac{N * Z^2 * p * q}{e^2 * ((N(-1)) + Z^2 * p * q)}$$

- n = sample size.
- N = population or universe.
- Z = confidence level.
- p = probability in favor.
- q = probability against.
- e = sampling error.

The confidence level Z is a necessary constant value, the most common confidence levels are:

Level of Confidence	Value of Z	Sampling error
90%	1.645	10%
95%	1.960	5%
99%	2.576	1%

Example of finite sample calculation

A higher education institution has 6530 students, the researcher for a study of satisfaction with the quality of education, takes a sample of the universe to investigate, assigning a confidence level of 95%, the probability p is unknown.

In this case the value of Z is 1.96, e = 5%; N 6530, then the formula 1.1 is applied, not knowing the probability of occurrence p, it is assumed that both p and q represent 50%.

$$n = \frac{6530 * 1.96^2 * 0.5 * 0.5}{0.05^2 * (6530 - 1) + 1.96^2 * 0.5 * 0.5}$$

$$n = 363$$

EXAMPLE

Thirty people from a population of 300 were asked how much they had in savings. The sample mean (\bar{x}) was \$1,500, with a sample standard deviation of \$89.55. Construct a 95% confidence interval estimate for the population mean.

Solution

EXAMPLE

Solution

EXAMPLE

Solution

INFINITIVE SAMPLE

When the population is unknown (number), the researcher should apply the following formula:

$$n = \frac{Z^2 * p * q}{e^2}$$

Where:

n = sample size sought.

e = maximum allowable estimation error

p = probability in favor.

q = probability against.

Example 1

How many people would we have to study to know the prevalence of diabetes?

Example 2

For a market research work in Peru (infinite population 24'000,000 inhabitants), among other things, we want to know how many people will travel abroad to work, with the decision to settle permanently in the country of destination. What should be the sample size for a survey confidence level of 95.5% and a possible margin of error of 4%?

Example 3

When the value of P and Q are unknown or when the survey covers different aspects in which these values may be unequal, it is convenient to take the most appropriate case, that is, the one that needs the maximum sample size, which is the case for $P = Q = 50$, then, $P = 50$ and $Q = 50$.

1. WRITE DOWN A GLOSSARY

1	_____
2	_____
3	_____
4	_____
5	_____
6	_____
7	_____
8	_____
9	_____
10	_____
11	_____
12	_____
13	_____
14	_____
15	_____

SCIENTIFIC RESEARCH

<hr/> <i>RESEARCH</i> <hr/>	<i>MEANS</i>	<hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/>	
<hr/> <i>IT</i> <hr/>	<i>IS CONSIDERED</i>	<hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/>	<i>TO</i> <hr/> <hr/> <hr/>
<hr/> <i>IT</i> <hr/>	<i>IS AN ACTIVITY THAT</i>	<hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/> <hr/>	

Research means a series of methods aimed at the individual acquisition of knowledge not already possessed, it implies the search for new relationships between already known elements, novel aspects, that is to say; research constitutes the generation of knowledge.

Research is considered as the basis that allows building a guide, which provides science not only descriptive but causal and conceptual schemes of the environment (Perez 2008; Sautu et al., 2014), in other words it is the guide in relation to events and their reciprocal causal connections, the nature of science and consequently research have been explained by the branch of philosophy called Philosophy of science, which constitutes a discipline of human reasoning to understand what is the foundation of science.

From the academic perspective, research is an activity that is executed in a systematic, controlled and critical manner. The purpose of scientific research according to the authors

Cruz et al., (2014); Mora & Sepúlveda (1999) is to discover, describe, interpret facts or phenomena, as well as to establish relationships between facts or phenomena, generate, disseminate knowledge, produce theories and solve practical problems.

Scientific research is a dynamic procedure, characterized by its rigor and conduct in the acquisition of new knowledge (Monroy & Nava, 2018), its main function focuses on describing, understanding, controlling, predicting phenomena, behaviors, facts, authors such as Rodolfo Mondolfo (1961) state that research arises when there is awareness of a problem, that the human being tends to seek the solution to it, the inquiry conducted to reach that solution represents the research itself.

Scientific research is research conducted for the purpose of contributing to science by systematically collecting, interpreting and evaluating data in a planned manner (Capalar & Dönmez, 2016).

The starting point of the research is the existence of a problem, which must be defined, examined, evaluated and critically analyzed, for the emergence of its solution.

1. READ AND COMPLETE THE CHART



2. WRITE DOWN A GLOSSARY

1	_____
2	_____
3	_____
4	_____
5	_____
6	_____
7	_____
8	_____
9	_____
10	_____
11	_____
12	_____
13	_____
14	_____
15	_____

CLASSIFICATION OF SCIENTIFIC RESEARCH

Classification of scientific research

By purpose

For its depth

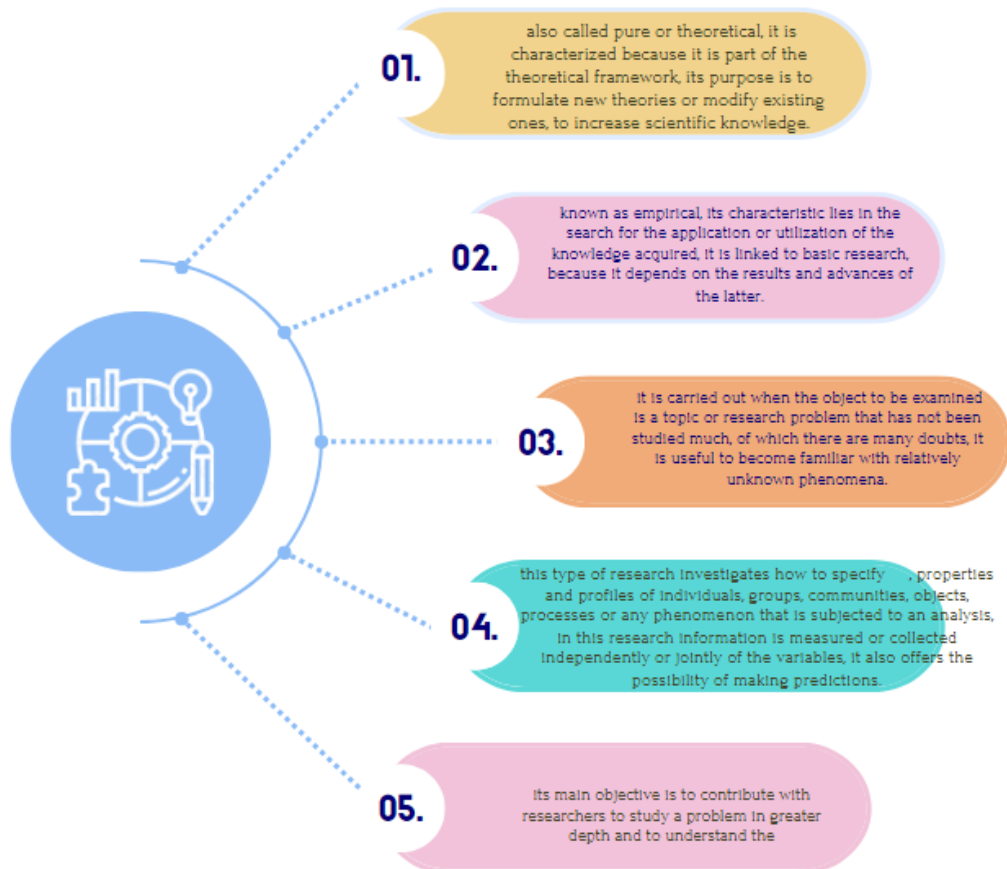
For its development
framework

1. MATCH THE NAME AND THE DESCRIPTION

Basic research Applied Research Exploratory research

Descriptive research Explanatory research

RESEARCH



Laboratory research: it is an experimental research, which resorts to reasoning, through experiments it seeks to give an answer to a hypothesis.

Field Research: is the collection of data from primary sources for a specific purpose, this type of research uses instruments such as files or statistical representations that allow collecting and analyzing the data to be studied. It is executed in the place of the facts, it involves taking information from direct source, without manipulating or controlling the variables, this type of research allows observing a phenomenon in real conditions (Monroy

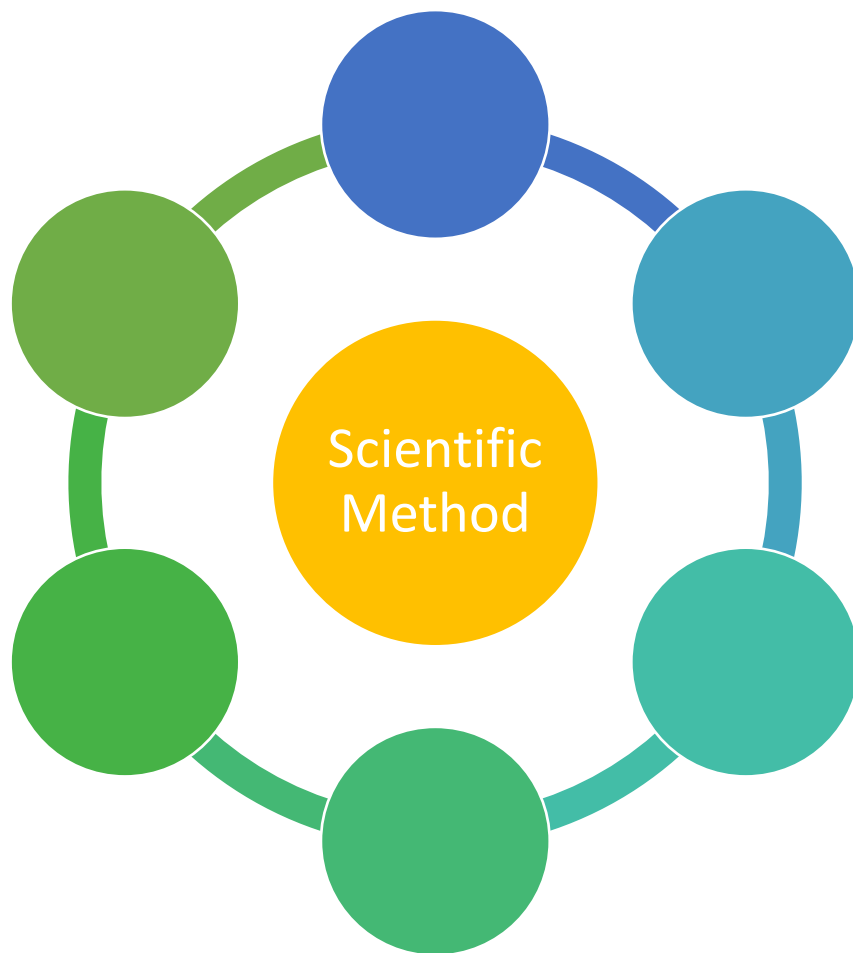
& Nava, 2018).

2. WRITE DOWN A GLOSSARY

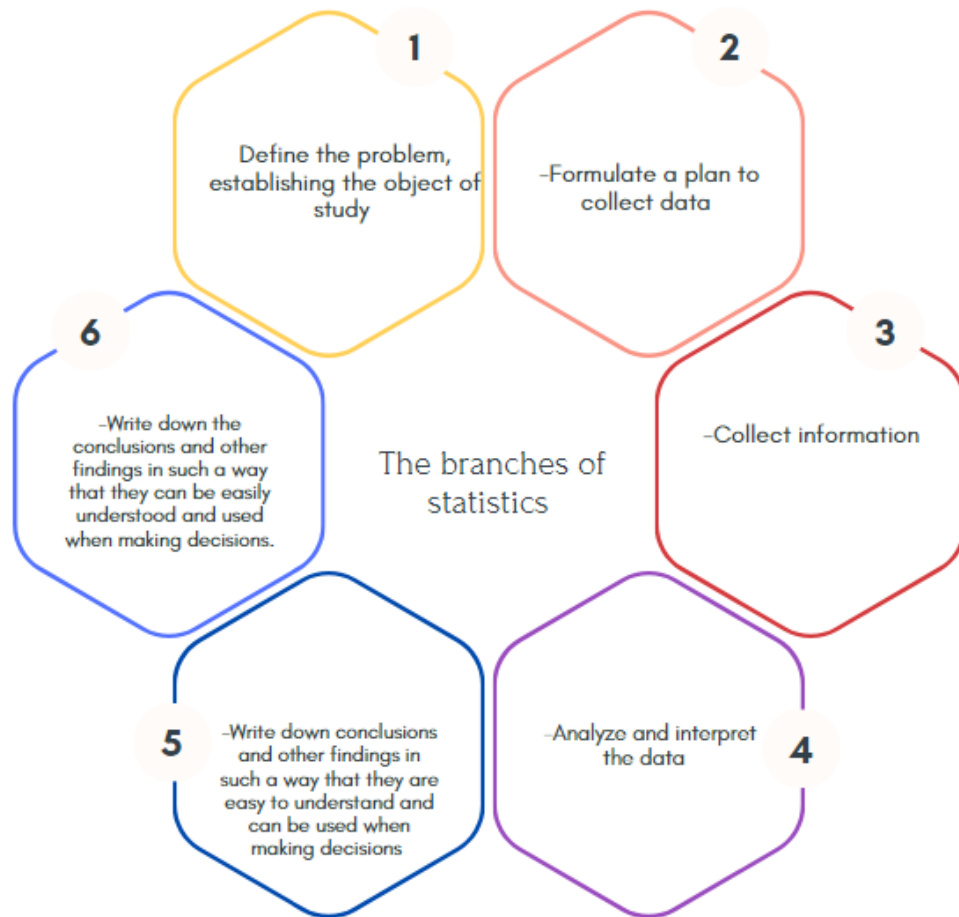
1	
2	_____
3	_____
4	_____
5	_____
6	_____
7	_____
8	_____
9	_____
10	_____
11	_____
12	_____
13	_____
14	_____
15	_____

SCIENTIFIC METHOD

The scientific method is a systematic process that is subject to notions and rules to achieve a predetermined end, it seeks to establish the procedures to be followed, in general terms it can be said that the scientific method is applied to a complete cycle of research, in search of solutions to each problem of knowledge, authors such as Monroy (2008) state that the scientific method is a procedure that requires systematization of thought to develop a reflective research.



The branches of statistics use the scientific method, which is adapted to it in five steps:



1. Is it considered purely mathematical?

WHY

3. WRITE DOWN A GLOSSARY

1	
2	<hr/>
3	<hr/>
4	<hr/>
5	<hr/>
6	<hr/>
7	<hr/>
8	<hr/>
9	<hr/>
10	<hr/>
11	<hr/>
12	<hr/>
13	<hr/>
14	<hr/>
15	<hr/>

RESEARCH APPROACHES

Since research is a set of systematic, critical and empirical processes applicable to a problem, there are two traditional research approaches: quantitative approach and qualitative approach, both approaches are used for the generation of knowledge, however, in the last decade researchers have tended to the combined use of both approaches generating the mixed approach, arguing that testing a theory with both approaches generates more reliable results.

Description of the approaches

<i>RESEARCH APPROACHES</i>	<i>Quantitative Approach</i>	
	<i>Qualitative approach</i>	
	<i>Mixed</i>	

QUANTITATIVE APPROACH

The quantitative approach is sequential evidential, quantitative research considers that knowledge should be objective, it is generated from a deductive process, by means of numerical measurement and inferential statistical analysis, it tests hypotheses previously formulated, this approach is associated with norms and practices of positivism (Hernández et al., 2017).

In this context, the quantitative research approach uses data collection based on numerical calculations and statistical research to establish patterns of behavior and test theories.

a. Characteristics of the quantitative approach

The main characteristics of the quantitative approach are:

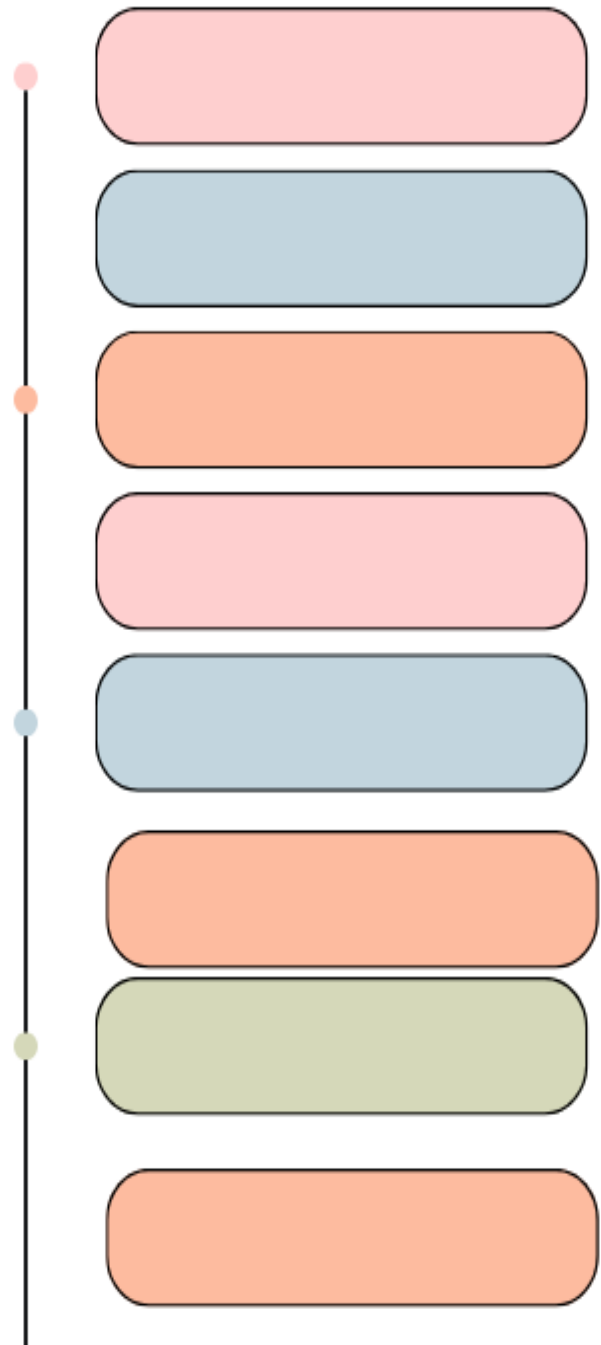
- a) It shows the need to evaluate and calculate magnitudes of the phenomena to be investigated.
- b) The statement of the problem to be investigated is based on specific questions.
- c) With the problem statement, the investigator (researcher) takes into consideration the research background, builds the literature review, is based on existing theories

as a guide, from which hypotheses are derived and tested.

- d) Hypotheses are generated prior to data collection and analysis.
- e) The data collected are based on the measurement of variables contained in the hypothesis; standardized processes accepted in the scientific community are used for data collection.
- f) Data analysis is performed using statistical methods.
- g) Reliance on experimentation or causality testing.
- h) Initial predictions allow the interpretation of quantitative analyses, as an explanation of the results of existing knowledge.
- i) It must be objective, avoiding biases that influence the result.
- j) This type of research follows a structured pattern.
- k) It tries to generalize the results in a sample to the universe or population.
- l) It aims to confirm and predict the phenomena analyzed, identifying causal relationships between elements.
- m) It identifies universal and causal laws.
- n) Explains how reality is conceived with a research approach.

1. WRITE A SUMMARY

Characteristics of the quantitative approach

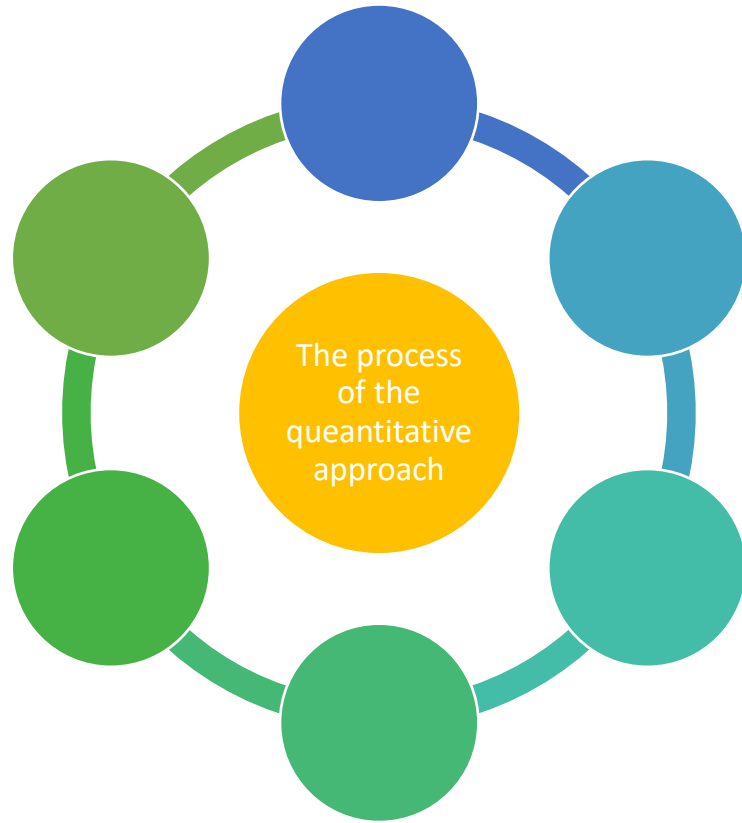


Quantitative approach process

The process of the quantitative approach consists of 10 phases listed below:

- Idea
- Problem statement.
- Literature review and development of the theoretical framework
- Visualization of the scope of the study
- Elaboration of hypotheses and definition of variables
- Development of the research design
- Definition and selection of the sample
- Data collection
- Data analysis
- Elaboration of the report of results

2. SUMMARIZE THE PROCESS OF THE QUANTITATIVE APPROACH



3. WRITE DOWN A GLOSSARY

1	
2	<hr/>
3	<hr/>
4	<hr/>
5	<hr/>
6	<hr/>
7	<hr/>
8	<hr/>
9	<hr/>

10	
11	
12	
13	
14	
15	

QUALITATIVE APPROACH

The qualitative approach is based on data collection without numerical measurement to discover or refine research questions in the process of interpretation, it aims at describing the qualities of a phenomenon, through this approach theories and hypotheses are generated.

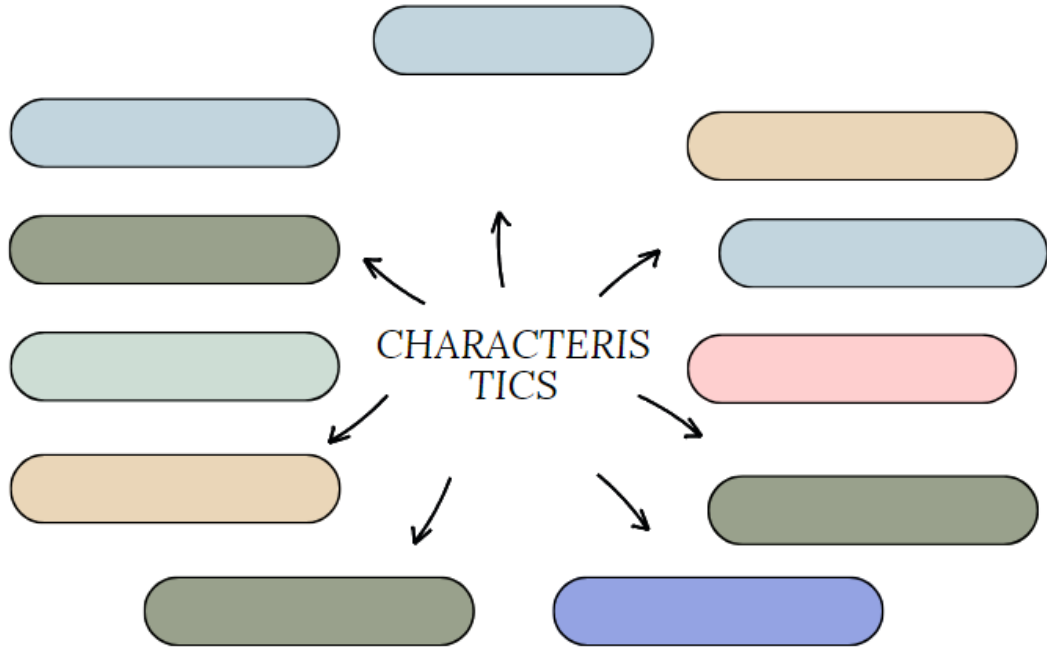
This approach is also called naturalistic, phenomenological, interpretative or

ethnographic research, which includes a variety of non-quantitative techniques, among its main characteristics:

- The researcher poses a problem, but does not follow a clearly defined process
- The research is initiated by examining facts, it is based on logic and the inductive process - Qualitative studies do not test hypotheses
- It is based on non-standardized, non-predetermined collection methods, the collection consists of obtaining the points of view of the participants.
- The main collection techniques in this approach are unstructured observation, open interviews, document review, evaluation of individual experiences, group reflection
- The inquiry is flexible, its purpose is to reconstruct reality
- It values the natural progress of processes, there is no manipulation of reality
- It centrally interprets the meaning of actions.

1. COMPLETE THE CHART

Qualitative approach



2. WRITE DOWN A GLOSSARY

- 1 _____
- 2 _____
- 3 _____
- 4 _____
- 5 _____
- 6 _____
- 7 _____
- 8 _____
- 9 _____
- 10 _____

- 11 _____
- 12 _____
- 13 _____
- 14 _____
- 15 _____

DIFFERENCES BETWEEN QUANTITATIVE AND QUALITATIVE RESEARCH

The use of both approaches in a research study makes it possible to correct the biases of each of the methods applied; Table 1-2 identifies the main differences between these approaches.

Quantitative Approach	Qualitative Approach
Based on probabilistic induction Pervasive and controlled measurement Objective Inference beyond the data - Confirmatory, inferential, deductive Results-oriented - Robust and repeatable data - Generalizable - Particularistic - Static reality	- Focused on phenomenology and understanding - Uncontrolled naturalistic observation - Subjective - Inference from your data - Exploratory, inductive and descriptive - Process oriented - Deep and rich data - Non-generalizable - Holistic - Static reality

1. WRITE DOWN MORE DIFFERENCES

Quantitative Approach	Qualitative Approach

In the quantitative approach, the approaches to be investigated from the beginning of the study are delimited and specific, the hypotheses are established prior to data collection and analysis, while the qualitative approach focuses on reconstructing reality as observed by previously defined social actors. Quantitative research is objective and seeks to generalize the results found in a sample to a population, while qualitative research is subjective and does not seek to generalize the results to a population.

2. COMPLETE THE RELATION

The hypotheses are established

Qualitative approach focuses on

Quantitative
research is

Qualitative research
is

ETHICAL GUIDELINES FOR THE PRACTICE OF STATISTICS IN RESEARCH.

The ethical component is fundamental in all the researcher's actions, ethical problems arise in statistics due to the importance of statistics in the collection, analysis, presentation and interpretation of data; in a statistical study, unethical behavior causes an inappropriate sampling, as a consequence, an erroneous development of statistical results is generated.

EXPAIN

**ETHICAL
PROBLEMS**

**UNETHICAL
BEHAVIOR**

As statistical research work progresses, it is advisable to be fair, meticulous, objective and neutral in data collection, analysis and reporting. According to authors such as Seltzer (2005), the ethical challenges in statistics can be expressed as: using adequate methodology, protecting confidentiality.

1. WRITE DOWN A GLOSSARY

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

BASIC GUIDE TO ETHICS APPLIED TO RESEARCH STATISTICS

Ethics is applied in statistics because it is important to have moral values in order not to deceive people with false data that in some cases are generated in the statistical context, for this reason every statistical process must present:

Respect for people,

- Aim to do good,
- Promote justice,
- Promote full development,
- Be transparent,

Statistical

process

must

present

- Avoid selection and exclusion of data at the convenience of the researcher,

- Avoid performing only those statistical analyses that seem to favor a particular hypothesis, examine alternative applications to what has been observed,

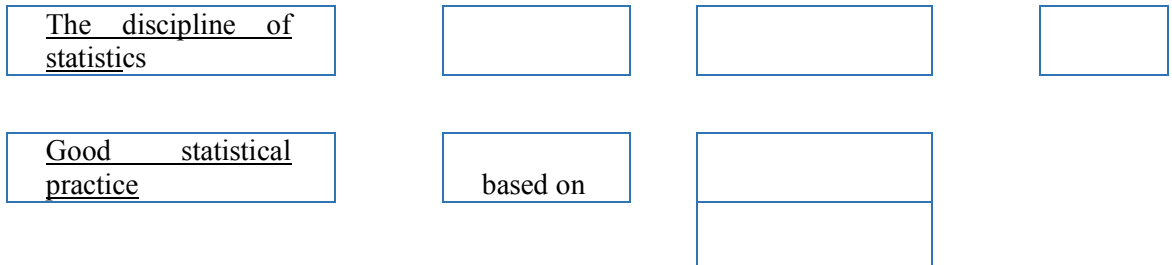
- The narrative should be in accordance with the expected results, that is, the explanations and interpretations of the results should respond to the evidence,

- The narrative should be in accordance with the expected results, that is, the explanations and interpretations of the results should respond to the evidence.

The discipline of statistics links the ability to observe with the ability to gather evidence and make decisions, providing a foundation for building a more informed society.

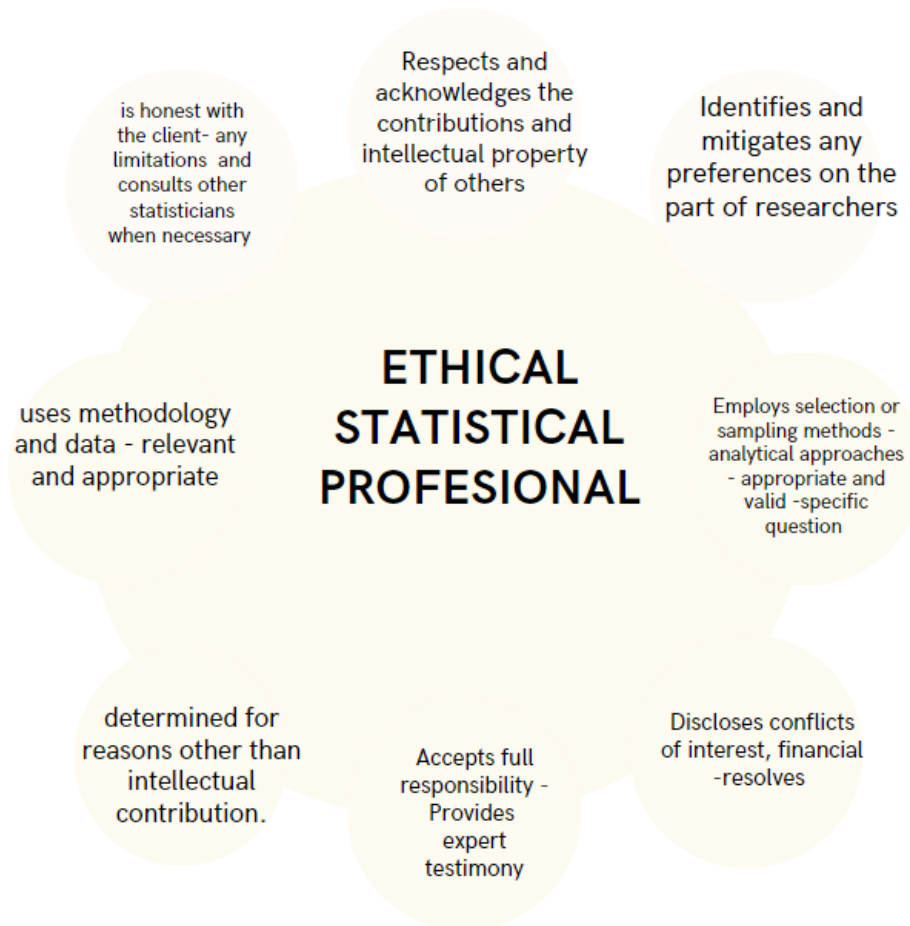
Because society depends on informed judgments supported by statistical methods, all statistical professionals, regardless of training, occupation, or position, have an obligation to work in a professional, competent, and ethical manner and to discourage any professional and scientific misconduct.

Good statistical practice is fundamentally based on transparent assumptions, reproducible results and valid interpretations. In some situations, the principles of the guidelines may conflict, requiring individuals to prioritize the principles according to the context.



ETHICAL STATISTICAL PROFESIONAL

The ethical statistician uses methodology and data that are relevant and appropriate, without



favoritism or bias, and in a manner intended to produce valid, interpretable and reproducible results.

1. LOOK AT THE CHAT AND WRITE ABOUT ETHICAL STATISTICAL PROFESSIONAL

2. WRITE DOWN A GLOSSARY

1	<hr/>
2	<hr/>
3	<hr/>
4	<hr/>
5	<hr/>
6	<hr/>
7	<hr/>
8	<hr/>
9	<hr/>
10	<hr/>
11	<hr/>
12	<hr/>
13	<hr/>
14	<hr/>
15	<hr/>

DATA AND METHOD INTEGRITY

The ethical statistician is frank about any known or suspected restriction, impairment or bias in the data that may affect the integrity or confidentiality of the statistical analysis. The objective and valid interpretation of the results demands that the in-depth analysis records and acknowledges the level of security and probity of the data.

From an ethical point of view the statistical researcher must:

Recognize the statistical and substantive assumptions made in the execution and interpretation of any analysis. In reporting the validity of the data used, recognizes data editing procedures, including imputation mechanisms and missing data.

Reports the limitations of statistical inference and possible sources of error, in publications, reports or testimony, identifies who is responsible for the statistical work if it would not otherwise be evident.

Reports the sources and assessed adequacy of the data; explains all the data considered in a study and explains the samples actually used.

Reports clearly and completely the steps taken to preserve the integrity of the data and valid results; when appropriate, addresses possible confounding variables not included in the study.

In publications and reports, conveys findings in a way that is honest and meaningful to the user/reader, this includes tables, models and graphs, identifies the ultimate financial sponsor of the study, the stated purpose and intended use of the study results.

When reporting analyses of volunteer or other data that may not be representative of a

defined population, includes appropriate disclaimers and, if used, appropriate weighting To aid peer review and replication, shares data used in analyses whenever possible/permissible

1. READ AND COMPLETE

<i>Recognize statistical assumptions</i>	TO	
<i>Reports the limitations of statistical inference</i>	TO	
<i>Reports the sources and the evaluated adequacy of the data</i>	TO	
<i>It reports clearly and completely the steps taken.</i>	TO	
<i>In publications and reports, conveys findings in an honest and meaningful manner</i>	TO	
<i>When reporting analyses of volunteer data or other data that may not be representative of a defined population, include appropriate disclaimers and, if used, appropriate weighting</i>	TO	

*Quickly correct
any errors
discovered*

TO

RESPONSIBILITIES TO SCIENCE, PUBLIC, FUNDER, CUSTOMER

The ethical statistician supports valid inferences, transparency and good science in general, taking into account the interests of the public, funder, client or customer (as well as professional colleagues, patients, the public, and the scientific community).

To the extent possible, presents a client or employer with choices among valid alternative statistical approaches that may vary in scope, cost, or accuracy, strives to explain any expected adverse consequences of not complying with an agreed-upon sampling or analysis plan.

Applies sampling and statistical analysis procedures in a scientific manner, without predetermining the outcome, strives to make new statistical knowledge widely available to provide benefits to society at large and beyond his/her own scope. Understands and complies with the confidentiality requirements of data collection, disclosure and dissemination and any restrictions on its use established by the data provider (to the extent required by law), and protects the use and disclosure of data accordingly. Protects the privileged information of the employer, customer or funder.

1. READ AND WRITE A SUMMARY

RESPONSABILITIES

SCIENCE

PUBLIC

FUNDER

CUSTOMER

2. WRITE DOWN A GLOSSARY

1	
2	<hr/>
3	<hr/>
4	<hr/>
5	<hr/>
6	<hr/>
7	<hr/>
8	<hr/>
9	<hr/>
10	<hr/>
11	<hr/>
12	<hr/>
13	<hr/>
14	<hr/>
15	<hr/>

RESPONSIBILITIES WITH RESEARCH SUBJECTS

The ethical statistician:

- Protects and respects the rights and interests of human and animal subjects at all stages of their participation in a project;
- Safeguards the contents of administrative records and subjects of physically or psychologically invasive research; - Protects and respects the rights and interests of human and animal subjects at all stages of their participation in a project; - Safeguards the contents of administrative records and subjects of physically or psychologically invasive research.

1. WRITE DOWN MORE RESPONSIBILITIES

- 1) _____
- 2) _____
- 3) _____
- 4) _____
- 5) _____

The statistical researcher from an ethical perspective:

- Keeps informed
- Adheres to applicable rules, approvals, and guidelines for the protection and welfare of human and animal subjects.
- Strives to avoid the use of excessive or inappropriate numbers of research subjects and excessive risk to research subjects.
- Makes informed recommendations about study size.
- Protects the privacy and confidentiality of research subjects.
- Ensures that data are obtained from subjects directly, from others, or from

existing records.

- Anticipates and seeks approval for secondary and indirect uses of data, including linkage to other data sets, by obtaining approvals from research subjects, and obtains appropriate approvals to allow peer review and independent replication of analyses.
- Knows the legal limitations on privacy and confidentiality assurances and does not overpromise or assume legal privacy and confidentiality protections where they may not apply. Considers whether appropriate approvals were obtained from research subjects before participating in a study involving human subjects or organizations, before analyzing data from such a study, and when reviewing manuscripts for publication or internal use.

READ AND WRITE A SUMMARY

*The statistical
researcher from an
ethical perspective:*

WRITE DOWN A GLOSSARY

1	_____
2	_____
3	_____

4 _____
5 _____
6 _____
7 _____
8 _____
9 _____
10 _____
11 _____
12 _____
13 _____
14 _____
15 _____

Responsibilities with colleagues in the research team

Scientific and statistical practice often takes place in teams made up of professionals with different professional standards, the statistician must know how to work ethically in this environment. In this context the statistician recognizes that other professions have standards and obligations, that research practices and standards may differ between disciplines and that the statistician has no obligations to the standards of other professions that conflict with these guidelines, ensures that all discussions and reports of statistical design and analysis are consistent, avoids compromising scientific validity for convenience, and strives to begin transparency in the design, execution and reporting or presentation of all analyses.

WRITE A SUMMARY

*Responsibilities with
colleagues in the
research team*

--

Responsibilities with respect to allegations of misconduct

The ethical statistician understands the difference between questionable scientific practices and practices that constitute misconduct, avoids both, but knows how each should be handled.

Avoid tolerating or appearing to condone incompetent or unethical practices in statistical analysis, recognizes that differences of opinion and honest mistakes do not constitute misconduct; they merit discussion, but not accusation, knows the definitions and related operations of misconduct. If involved in a misconduct investigation, follows prescribed procedures.

Maintain confidentiality during an investigation, but disclose the results of the investigation honestly to the appropriate parties and stakeholders once they are available. Following a misconduct investigation, supports appropriate efforts by all involved, including those reporting the potential scientific error or misconduct, to resume their careers as normally as possible. Avoids and acts to discourage retaliation or harm to the employability of those who responsibly bring possible misconduct to the attention of others.

Responsibilities

WRITE DOWN A GLOSSARY

- 1 _____
- 2 _____
- 3 _____
- 4 _____
- 5 _____
- 6 _____
- 7 _____
- 8 _____
- 9 _____
- 10 _____
- 11 _____
- 12 _____
- 13 _____
- 14 _____
- 15 _____

CHAPTER 2

Considerations in Applied Descriptive Statistics

Considerations in applied descriptive statistics

Descriptive Statistics

*Descriptive
Statistics*

*is
focus
on*

*Descriptive
statistics*

*is
used
for*

Once collected the data of the study variables, and following the research objectives the next step is the statistical analysis in order to describe the sample, variables studied, that is; we proceed to the study of the relationship between dependent and independent variables (Mias, 2018).

Descriptive statistics is focused on describing, summarizing, visualizing the distribution of data, as well as the organization, dispersion in relation to measures of central dispersion; it estimates descriptive statistics such as mean, median, mode, range, standard deviation (standard), variance, case count, percentages and percentiles. Then the data can be represented by frequency histograms, bar charts, pie charts or pyramid charts among others.

Descriptive statistics is used for an exploratory analysis of the behavior of variables, data, the main statistics are summarized as follows:

- **Measures of central tendency:** these are measures of centralization, it is a number located towards the center of the distribution of values of a series of data, among these measures are: mode (for all scales), median (ordinal and interval), mean (ordinal and interval).

- **Measures of dispersion:** called measures of variability, they represent the degree to which a distribution is compressed or stretched. The main measures of dispersion are the range (includes maximum and minimum), standard deviation and variance.

- **Ratios, proportions and rates:** in statistics the ratio is used as indexes, the rate is a measure of comparison of data between different times and populations, proportions represent relative frequencies, which estimate the probabilities of occurrence of an event.

- **Relative risk:** is a measure of effect that establishes in relative terms the relationship between the probability of occurrence of an event in the exposed group and the

probability of occurrence of the same event in the non-exposed group.

WRITE DOWN A GLOSSARY

1	_____
2	_____
3	_____
4	_____
5	_____
6	_____
7	_____
8	_____
9	_____
10	_____
11	_____
12	_____
13	_____
14	_____
15	_____

Description and measurement of univariate data

Data are the representation of attributes or variables that represent facts; when analyzed and

processed, they are transformed into information (Martínez, 2020). In this context it is important to know some elements such as the properties of the sigma operator, interpretation of measures of central tendency and dispersion for simple and compound series, graphical representations.

Notation and use of Sigma

In the statistical context it is generally necessary to calculate the sum of a set of number, the Greek letter sigma, represented by the symbol Σ , which is used in statistics to denote a set of elements to be summed.

Example 1

A studied variable can take different values, and it can be represented as x_i and each of its values as X_1, X_2, X_3 , if these letters are assigned a number, for example: 15, 10, 20; they could be a large amount, when adding up these values that the variable takes, it would be represented as follows:

$$X_1 + X_2 + X_3 \rightarrow 15 + 10 + 20$$

Using the Sigma: 3

$$\sum_{i=1}^4 x_i = x_1 + x_2 + x_3 = 15 + 10 + 20 = 45$$

$$\sum_{i=1}^4 x_i = 45$$

The symbol reads: the sum of the values of X when it goes from one to 4.

Ejemplo 2.2

Assuming that n numbers multiplied each by 3, we seek to add to obtain the result of n products, it would be expressed as follows:

$$3X_1 + 3X_2 + 3X_3 + \dots + 3X_N$$

Applying the common factor, the expression is as follows:

$$3(x_1+x_2+x_3+\dots+x_n)$$

Using sigma is reduced to: n

$$\sum_{i=1}^n 3x_i$$

The constant 3 can be left out of the sigma and does not alter the product:

$$3 \sum_{i=1}^n x_i$$

Ejemplo 2.3

When each of the values that the variable takes, a constant c must be added to it.

$$X_1 + CX_2 + CX_3 + C \dots X_n + C$$

For the sum of these elements, the following should be done:

$$\sum_{i=1}^n (x_i + c) = \sum_{i=1}^n X_i + \sum_{i=1}^n C = \sum_{i=1}^n X_i + n + C$$

The expression is reordered:

In this case the set X_i can be abbreviated and the C is expressed as the product of n

$$3(x_0 + c) = 3x_0 + 3c = 3x_0 + n * c$$

ADD MORE EXAMPLES

Time Series

A statistical series is a set of numbers or terms that measure the variations of a phenomenon; there are simple, frequency, class and frequency series.

Simple series

A simple series represents a set of numbers of the variations of a particular phenomenon that make up a group. Example: the heights of 18 people are shown below:

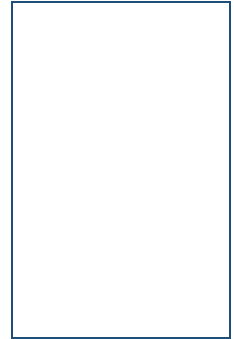
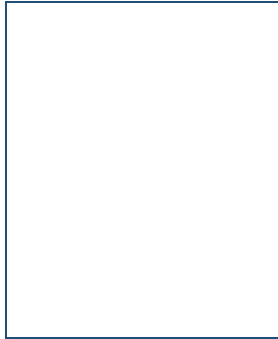
1.60	1.61	1.55
1.70	1.55	1.61
1.55	1.60	1.61
1.55	1.55	1.60
1.70	1.70	1.70
1.45	1.70	1.70

The set of data is the result of the measurements of the heights of 15 people, these data are called simple series, in this example the amount of data is very small, but when the number of data is very large it can be deduced with the simple series of characteristics of the phenomenon.

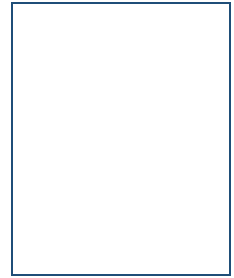
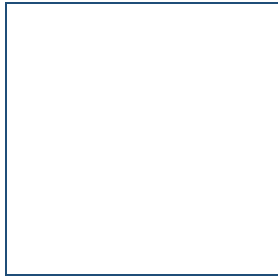
In this case, it can be concluded that 3 people are 1.60 m tall, 5 people are 1.55 m tall, 3 people are 1.61 m tall, 6 people are 1.70 m tall, and only 1 person is 1.45 m tall.

EXAMPLES

Time series



Simple series



WRITE DOWN A GLOSSARY

- 1 _____
- 2 _____
- 3 _____
- 4 _____
- 5 _____
- 6 _____
- 7 _____
- 8 _____
- 9 _____
- 10 _____
- 11 _____
- 12 _____
- 13 _____

14

15

Frequency series

A frequency series refers to the number of times a value or term is repeated in a simple series, i.e., how frequent is the data within the series.

The sum of the frequencies is equal to the number of terms of the simple series, the frequencies are only the grouping of the equal terms of the series, which does not modify the original or total quantity.

Class and frequency series

When having a very large number of observations of a fact or phenomenon, for example 1000 students of the ESPCOH, in this case we would have the simple series (1000 data), but to work the information more quickly and easily we proceed to form a frequency table, however, the data would be many, in this case the information is more compact forming groups or class intervals. **Class:** se define como ciertos grupos o intervalos en los que se concentran las frecuencias observadas de la serie.

The main steps in the elaboration of a class and frequency distribution for sample data are as follows:

Establish the class intervals 2 into which the data are grouped

- Sort the data into classes by counting marks

- Count the number of frequencies in each class and write it down

- Present the results in a table.

To establish the class intervals of the series it is essential to perform the following steps:

Determine the amplitude of the data variance (variable path), which is one plus the difference between the largest and smallest of the scores.

$$AV = 1 + (\textit{puntuación mayor} - \textit{puntuación menor}) \quad (2.1)$$

Deciding the number of classes or intervals, there are several criteria for the number of classes, an empirical approximation is obtained from \sqrt{n} , where n = number of data; example $n = 1000$; $\sqrt{1000} = 31.62 \approx 32$, it is important to round the result to a whole number; another criterion is the individual decision of each researcher depending on his experience.

- Calculate the size of each class: it consists of dividing the amplitude of the variation of the data by the number of classes, it should be divided by the number of classes rounded. $Tc = Av/Nc$

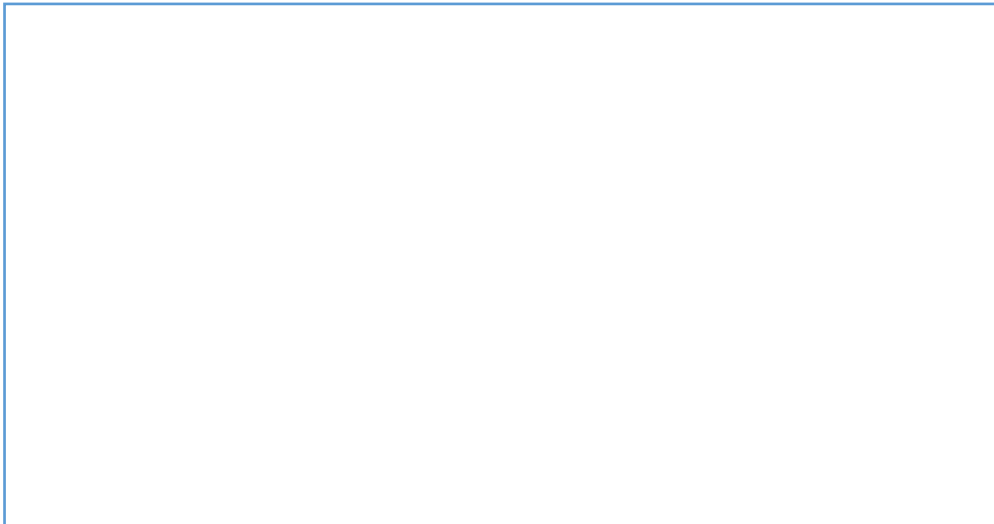
- To tabulate data or determine class boundaries, start with the first interval, ensuring that the minimum observation is included. Classes should be close to each other so that there are no considerable gaps. The numbers that limit a class are called class boundaries.

EXAMPLE

The following data are presented for the ages of 24 people.

15	28	33
20	27	32
18	16	15
17	40	17
16	45	18
32	40	20
42	30	21
33	33	25

It is requested to build a series of classes and frequencies.





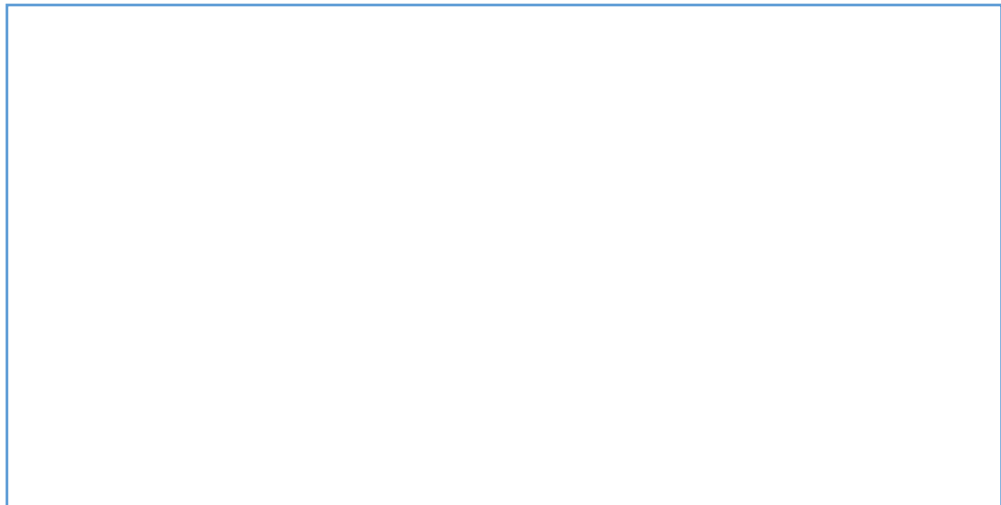
- **Determine the amplitude of data variation**

- Decidir el número de clases o intervalos

$$Nc = \sqrt{n}$$

$$Nc = \sqrt{24}$$

Since it is a non-integer result, it must be rounded up, never down.



WRITE DOWN A GLOSSARY

- 1 _____
- 2 _____
- 3 _____
- 4 _____
- 5 _____
- 6 _____
- 7 _____
- 8 _____
- 9 _____
- 10 _____
- 11 _____
- 12 _____
- 13 _____
- 14 _____
- 15 _____

Measures of central tendency

MEASURE

MODE

The measure of central tendency called central tendency parameter, is a number that is located towards the center of the distribution of the values of a series of observations in which it is located in the data set, in this context the statistics of central tendency are: mode, median and mean; they are used to describe characteristics of centrality of variables (Puente, 2018).

Mode

The mode is the value or category of the variable that is repeated more times or has a higher frequency, this mode is considered as absolute; there may be other modes that are called relative and its characteristic is the value of the variable that has a higher frequency than the previous and subsequent values.

This statistic can be used with the variables of nominal measurement, ordinal, interval and ratio, the calculation is made from the frequency tables, the result can vary according to the grouping of the intervals, in the categorical variables the value of the mode is calculated by observation of the frequency table.

It is important to mention that not all sets of scores have a mode, sometimes there is no mode or there may be more than one, below are some examples of modes.

EXAMPLES

We have the following scores 2,6,6,7,7,8,8,8,9,10; the mode in this case is 8, because it is the most repeated score in the series.

EXAMPLE

In the case where all scores or items in a group have the same frequency, i.e.;

.....
.....
.....

EXAMPLE

In the case where two consecutive scores have equal frequency, which is greater than any other score, the mode is considered as the average of the two scores, then the following set of scores 0, 1, 1, 1, 1, 2, 2, 2, 2, 3, 4 is presented; in this case

.....
.....
.....

EXAMPLE

In the case of a group of data there are two non-consecutive data and they have the same frequency, and it is greater than any frequency of the rest of the data, the existence of two modes is evident. 9,9,10,11,12,12,13;

.....
.....
.....

WRITE DOWN A GLOSSARY

1	
2	_____
3	_____
4	_____
5	_____
6	_____
7	_____
8	_____
9	_____
10	_____
11	_____
12	_____
13	_____
14	_____
15	_____

Mode for a series of classes and frequencies

In this type of series, the same criterion for the mode of the previous cases is applied, i.e., the value that is repeated the most times is sought, except that in this scenario the values are grouped in classes or intervals.

Process of calculating the mode

The frequency table is ordered from smallest to largest. 2. The mode is in that category that has the largest number of cases. 3.

$$M = L + D1 * a / (D1 + D2) \text{ Where:}$$


L_{i-1} = Lower bound of the modal interval a = amplitude of the intervals $D1$ = Absolute frequency difference between modal interval and the previous one $D2$ = Absolute frequency difference between modal interval and the next one.

EXAMPLE

The following is a frequency table of the weight of 95 ESPOCH students.

Intervals	Frecency
45 – 55 kg	23
55 – 65 kg	21
65 – 75 kg	32
75 – 85 kg	19
Total	95

1. The frequency table is ordered from smallest to largest.

Intervals	Frequency
65 – 75 kg	
45 – 55 kg	
55 – 65 kg	
75 – 85 kg	
Total	95

- 2.

1. The mode is in the category with the highest number of cases. The interval with the highest number

of cases is 65 - 75 kg.

The value is calculated.

EXAMPLE

EXAMPLE

Median

The median in a set of data is the value that is in the middle of the other values, that is, when ordering the numbers from smallest to largest, this value is in the middle, some of the characteristics of this measure of central tendency are:

The operations for its calculation are simple to perform.

The median does not depend on the values of the variables, only on their order.

Generally its values are integers.

It can be calculated even though the previous and following numbers have no limits.

Median for a simple series

To find the median it is first necessary to order the values from smallest to largest, then you must separate half of the values to obtain the median, the procedure to identify the median is summarized as follows:

- | |
|---|
| 1. Order or sort the values from largest to smallest or smallest to largest. |
| 2. Count the values to determine if there is an even or odd number of data. |
| 3. In the case of odd cases the median is the central value, but if the number is an even value, the median is the average of the central values. |

Formula determines the position of the median, but not its value; this is found by finding the position of the median.

EXAMPLE

- Finding the median of group 2, 3, 4, 5, 6.
First we define the position of the median with the formula $n+)$; thus for 2 five values, the position of the median is 3, in this sc

EXAMPLE

The following data are presented 9, 10, 11, 12, 13,14.

When applying the formula $n+)$; is 3.5, in this position there are two intermediate values 2, in this case it is 11 and 12; adding and calculating the average of these values the median is 11.5.

EXAMPLE

EXAMPLE

Median for a series of classes and frequency

The median is the value of the variable that leaves below 50% of the cases, the median can be used with variables that at least have an ordinal level of measurement, but its use is more appropriate with intervals, in nominal variables it cannot be used because the cases cannot be ordered.

The median formula is a derivation of the percentile formula and is developed with a rule of three.

Calculation process

The frequency table is ordered from smallest to largest.

The accumulated frequencies are calculated.

The number of cases is divided by two and the interval that contains them is called the critical interval (CI).

The exact value of the median is then calculated.

$$M_e = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} * a$$

Where:

Me = median

Li-1 = lower limit of CI

N/2 = half of the cases

Ni-1 = total cases under the CI

a = width of CI

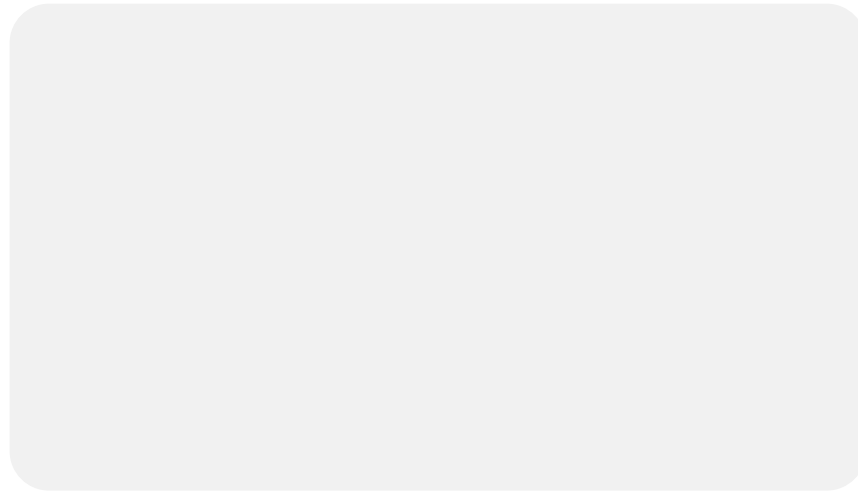
ni = Frequency of the critical interval

EXAMPLE
EXPLAIN

Intervals	Frequency	Cumulative frequency
45 – 55 kg	23	23
55 – 65 kg	21	44
65 – 75 kg	32	76
75 – 85 kg	19	95
Total	95	

EXPLANATION

FORMULA



EXAMPLE

Intervals Stature cm	Frequencies	Frequencies accumulate
118 – 126	3	3
127 – 135	5	8
136 – 144	9	17
145 – 153	12	29
154 - 162	5	34
163 - 171	4	38
172 - 180	2	40
Total	40	

EXPLANATION

A vertical line on the left side of the form is connected to the 'EXPLANATION' header. To the right of this line are ten horizontal blue lines, providing a space for writing an explanation.

FORMULA



Mean

The mean, also known as average, is the value obtained by dividing the sum of all the numbers that make up the cluster, some of the characteristics of this measure of central tendency are:

- Consider all elements
- The numerator of the formula is the number of values

The arithmetic mean is the value obtained by adding all the data and dividing the result by the total number of data. We denote the mean with the symbol

Arithmetic mean for a simple series

La media se calcula sumando los valores de la serie, de los cuales se obtiene el promedio, dividiendo entre el número de datos que se toma en cuenta en la suma, la media de las n medidas se determina de la siguiente manera:

$$x = \frac{21 + 22 + \dots + 2n}{n}$$

EXAMPLE

The following data are presented: 8, 5, 6, 7, 10, 11, 15

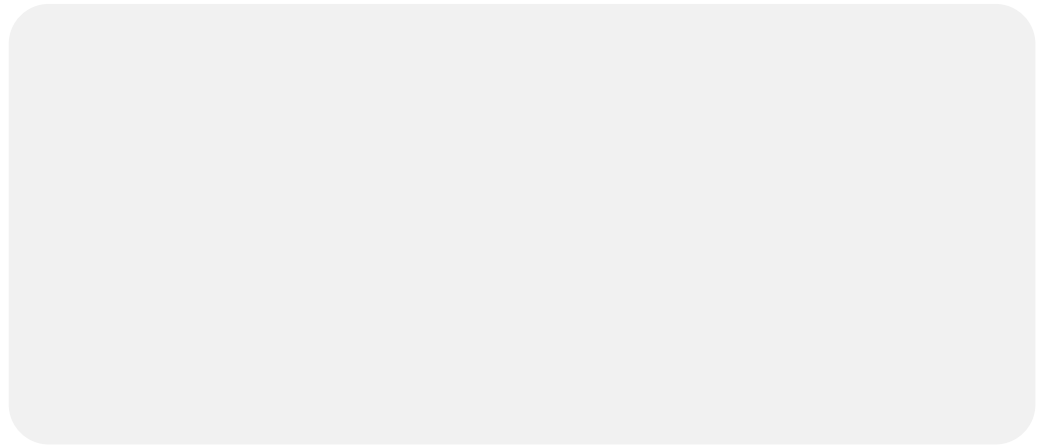
$$x = \frac{8 + 5 + 5 + 7 + 10 + 11 + 15}{7}$$

$$x = \frac{62}{7}$$

$$x = 8.86$$

EXAMPLE

The heights of 10 students of a Chemistry course at ESPOCH were collected as follows: 1.70; 1.60; 1.55; 1.60; 1.75; 1.63; 1.69; 1.59; 1.61; 1.68.



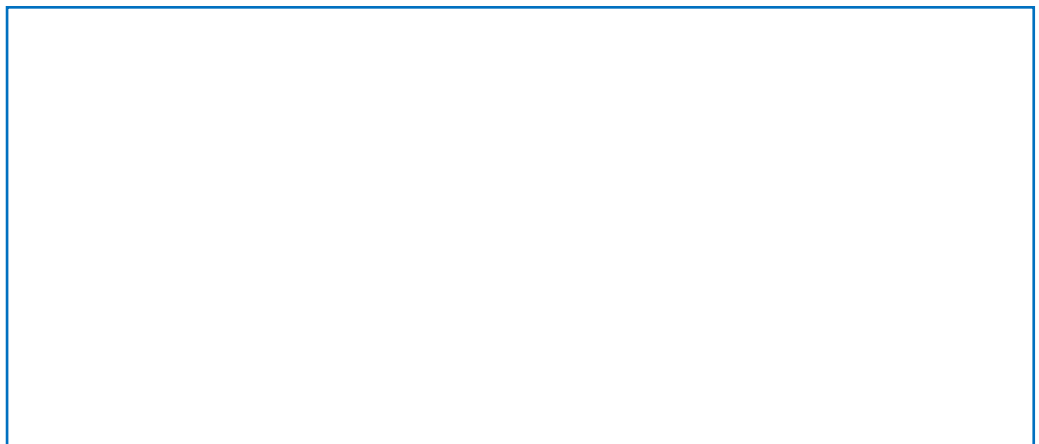
SOLUTION:

EXAMPLE



SOLUTION:

EXAMPLE



SOLUTION:

EXAMPLE

SOLUTION:

Arithmetic mean for a series of frequencies

The mean for a series of data organized by frequencies is calculated as follows:

$$\bar{X} = \frac{\sum_{i=1}^n f_i Y_i}{n}$$

Where

\bar{X} = Media

Y1 = class mark

n = Total cases

EXAMPLE

Intervals	Frequency f_i	Class Mark (yi)	$f_i * y_i$
45 – 55 kg	23	50	1150
55 – 65 kg	21	60	1260
65 – 75 kg	32	70	2240
75 – 85 kg	19	80	1520
Total	95		6170

$$x = \frac{6170}{95} = 64.95\text{kg}$$

EXAMPLE

According to the following data determine the mean

Intervals	Frequency
50-55	6
55-60	13
60-65	9
65-70	8
70-75	4
n=	40

With the data obtained, we proceed to calculate the class mark and the frequency for the variable as shown below:

Intervalos	f_i	Clases (yi)
50-55	6	
55-60	13	
60-65	9	
65-70	8	
70-75	4	
Total	40	

$$\bar{X} = \frac{2467}{40} = 61,68 \quad \underline{\hspace{2cm}}$$

EXAMPLE

According to the following data determine the mean

Intervals	Frequency
50-55	
55-60	
60-65	
65-70	
70-75	
n=	40

With the data obtained, we proceed to calculate the class mark and the frequency for the variable as

WRITE DOWN A GLOSSARY

- 1 _____
- 2 _____
- 3 _____
- 4 _____
- 5 _____
- 6 _____
- 7 _____
- 8 _____
- 9 _____
- 10 _____
- 11 _____
- 12 _____
- 13 _____
- 14 _____
- 15 _____

Comparison of mode, median and mean

The calculation of the mean, median and mode is a mechanical procedure, and a comparison between these three measures is presented below, where their advantages and limitations are determined.

Measure	Concept	vantage	Disadvantage
Media	It is the data that is defined as the average of all the scores.	<ul style="list-style-type: none">- Reflects the value - Attractive mathematical properties- Attractive mathematical properties- Attractive mathematical properties	<ul style="list-style-type: none">- May be overly influenced by extremes
Median	It is the value that divides the series into two equal parts.	<ul style="list-style-type: none">- Less sensitive to extremes than average	<ul style="list-style-type: none">- Difficult to determine if there is a large amount of data
Mode	It is the value with the highest frequency	<ul style="list-style-type: none">- Typical value, more values gathered at this point than at any other point.	<ul style="list-style-type: none">- It does not lend itself to mathematical analysis.

Other position measurements

In statistics, position measures are statistical indicators that allow data to be summarized in a single one, or to divide their distribution into intervals of equal size, among the most usual measures are:

Quartile:

is a point on a numerical scale that groups a series of data by dividing them into four equal parts, there are three quartiles (Q1, Q2, Q3).

Quartile:

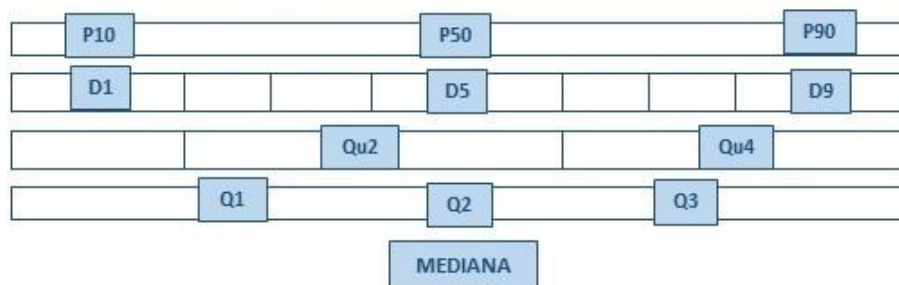
is a point on a numerical scale that groups a series of observations, dividing the distribution into five parts; therefore, there are four quintiles, it is of less use than the quartile.

Decile:

is a point on the numerical scale that divides the data into ten equal parts, there are nine deciles and they are represented as follows D1, D2, D3, D4, D5, D6, D7, D8, D9; D5 corresponds to the median.

Percentile

it is a statistic of central tendency, it divides the set of data and its distribution in one hundred equal parts, there are 99 percentiles, it has in turn an equivalence with the deciles and quartiles.



Graphical representation and equivalences of quartiles, deciles, percentiles.

Cuartil

$$Q = \frac{a(n+1)}{4} =$$

Decil

$$Q = \frac{a(n+1)}{10} =$$

Percentile

$$Q = \frac{a(n+1)}{100} =$$

Where:

Q1 = Cuartil. D1 = Decil. P1 = Decil.

α = The Q, D, P that can be calculated

.n = number of data.

EXAMPLE

Quartiles for a simple data series:

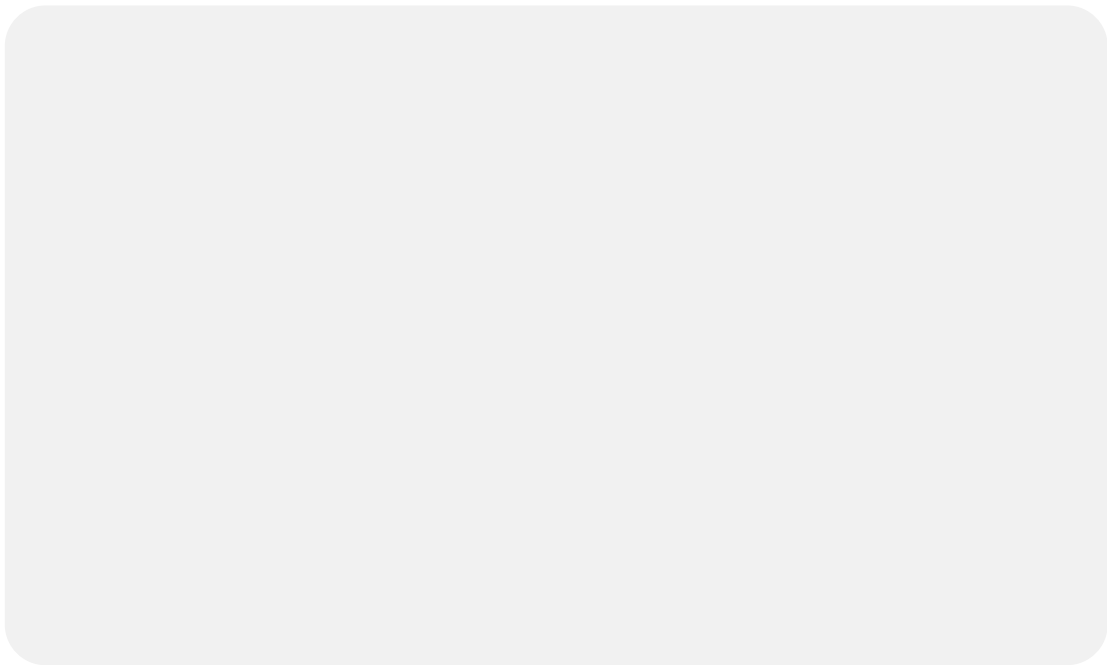
The following data are presented 1,4,5,7,10,10,13,14,16,20,23,25

$$Q = \frac{a(n+1)}{4}$$
$$Q_1 = \frac{1(11+1)}{4} = \frac{12}{4} = 3$$
$$Q_2 = \frac{2(11+1)}{4} = \frac{12}{2} = 6$$
$$Q_3 = \frac{3(11+1)}{4} = \frac{12 * 3}{4} = 9$$

To determine the value of Q1, its position is 3, in this case quartile 1 corresponds to 5; Q2 = 13 and Q3 = 20.

It is important to mention that the quartile, decile or percentile value indicates the place within the data series where each one of them is located; it is important to order the data in ascending or descending order.

With the following data: 1,4,5,7,7,10,13,14,16,20,23,25, calculate the decile 1, 5,9



To determine the value of D1, its position is 1.2, it is a number with decimals, then decile 1 corresponds to the value between position 1 and 2 $(1+4)/2 = 2.5$

D2 corresponds to position 6, i.e. 13; D9 corresponds to the position between 10 and 11 because it is a decimal number $(23+25)/2=24$.

With the following data: 1,4,5,7,7,10,13,14,16,20,23,25, calculate the 10th, 50th, 90th percentile.



D1 (decil 1) =

D5 (decil 5) =

D8 (decil 8) =

WRITE DOWN A GLOSSARY

1	_____
2	_____
3	_____
4	_____
5	_____
6	_____
7	_____
8	_____
9	_____
10	_____
11	_____
12	_____
13	_____
14	_____
15	_____

MEASURES OF DISPERSION

From the statistical point of view the dispersion measures represent the degree to which a distribution is spread or compressed (Gamero, 2017); this type of measures aims to express in a single value the dispersion that a set of data has, among the most used measures are: range of variation, variance, standard deviation, coefficient of variation.

***DISPERSION
MEASURES
REPRESENT***

Range or path

The range (R) or path of a variable X in statistics represents the difference between the maximum value and the minimum value of a set of data, mathematically it is represented by:

$$R = \max(X) - \min(X) \text{ (2.9)}$$

The range is expressed in the same unit of measurement as the variable of analysis, considering as an example the following data: 19, 25, 36; the difference between the largest and smallest number is $36 - 19 = 17$.

In general, the range is expressed by establishing the difference between the largest number and the smallest number of a set of data, one of the advantages of using the range as a measure of dispersion is the fact that it is easy to calculate, even when dealing with a very large set of data, its main limitation is that it considers extreme data, which can generate a false image of the set, this limitation generates the use of other paths such as those shown in Table 2

RANGE REPRESENTS

ROUTE	DESCRIPTION
Intercentile defendant	Difference between the last and first percentile, which excludes for the calculation of the 2% path of the observations, which admits extreme values. $R_{>} = C_{<<} - C_{>}$ (2.10)
Interdecyl defendant	Difference between the last and the first decile, which is why it is eliminated for the calculation of 20% of the observations. $R_D = D_{<} - D_{>}$ (2.11)
	Interquartile range Difference between the last and the first quartile, reason for which it is eliminated for the calculation of 50% of the observations. $R_{@} = Q_3 - Q_1$ (2.12)

In this context, the intequantylic runs eliminate outlier observations to study the dispersion of the distribution of the central zone, containing 98%, 80% and 50% of the observed data, thus creating an advantage over the run (R) of dispensing with extreme values; however, there is still the general disadvantage of not considering all the available information.

.....
<i>DIFERENCES</i>

WRITE DOWN A GLOSSARY

- 1 _____
- 2 _____
- 3 _____
- 4 _____
- 5 _____
- 6 _____
- 7 _____
- 8 _____
- 9 _____
- 10 _____
- 11 _____
- 12 _____
- 13 _____
- 14 _____
- 15 _____

Variance

VARIANCE

IS

-

-

TYPES

DIFFERENCES

Variance is a measure of dispersion, it represents the variability of a set of data in relation to their mean; it is a measure of variation equal to the square of the standard deviation, in the statistical context there are two types of variance: population variance and sample variance.

The variance of a sample its calculation is similar to that of the standard deviation with small differences: 1. The deviations are squared before being summed. 2. The average is obtained using n-1 instead of n, because it provides a better calculation of the variance.

Variance of a population

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - u)^2}{n - 1}$$

(2.13)

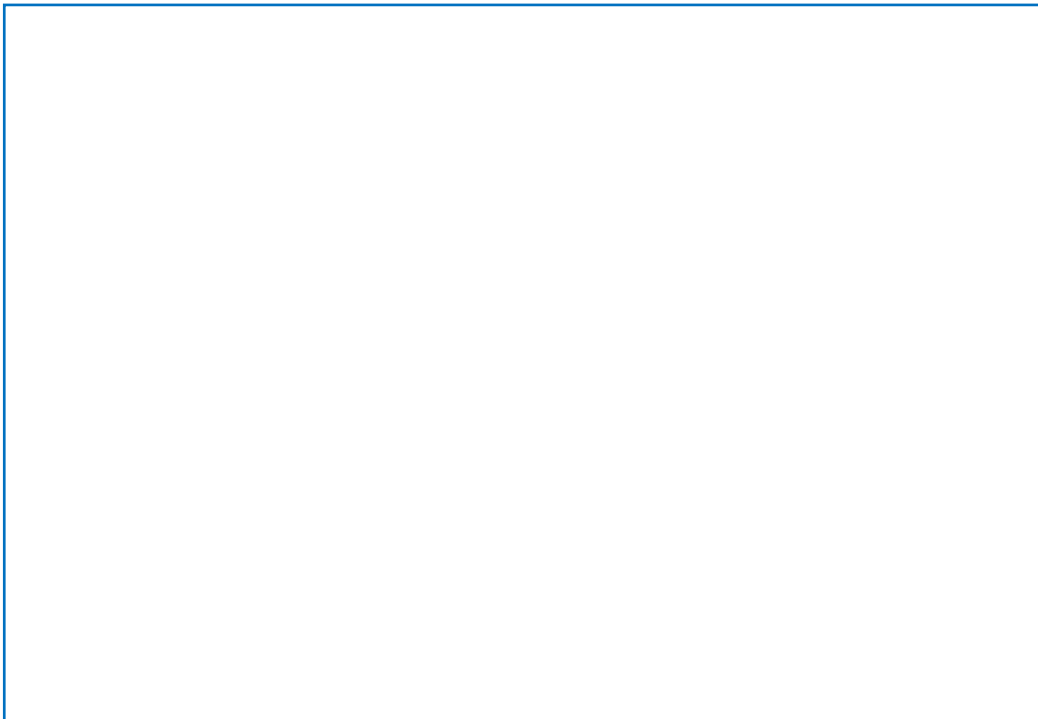
Dónde

X_i = values of X.

μ = value of the population mean.

n = number of elements.

EXAMPLES



Variance of a sample for simple series

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - u)^2}{n - 1}$$

Where

X_i = Values of X. X

u = Value of the mean.

n = number of elements.

Variance of a sample for frequency series

$$S^2 = \frac{\sum_{i=1}^N (x_i - X)^2}{n - 1}$$

where

X_i = Values of X. X

X = Value of the mean.

n = Number of items.

f_i = Frequencies.

Variance of a sample for frequency series and intervals

$$S^2 = \frac{\sum_{i=1}^N f_i (y_i - X)^2}{n - 1}$$

where

Y_i = Values of the class mark or midpoint of the class mark.

X = Value of the mean.

n = Number of elements.

f_i = Frequencies.

EXAMPLE

Variance for a simple series:

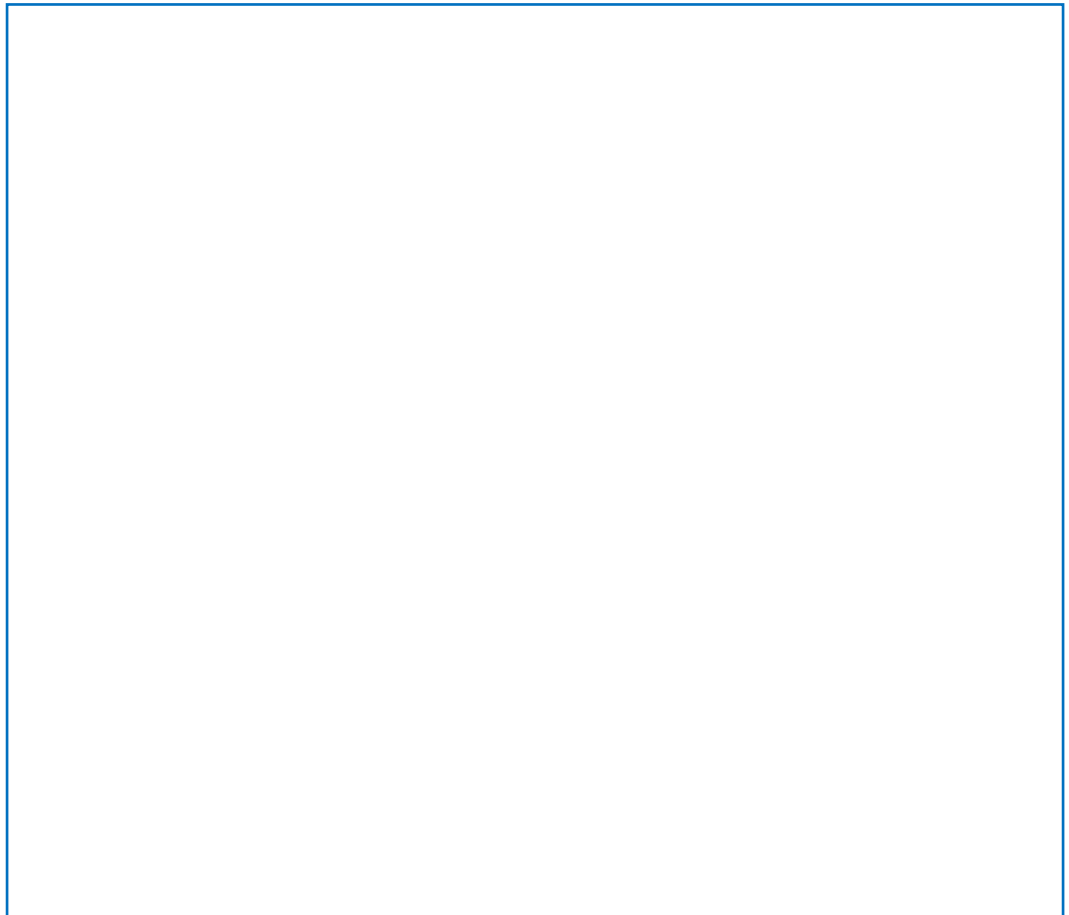
The ages of 5 students (14, 16, 18, 20, 22) are presented, calculate the variance for the sample.

- Calculate the mean, in this case there are 5 data, therefore $n = 5$.
- Subtract the mean from each value in the data set, giving the respective deviations.
- Square each deviation
- Divide the result of the summation by $n-1$, if all observations equal a population.

EXAMPLE

Variance for a frequency series:

Ages	Frequencies
14	5
15	2
16	6
17	7
19	2
20	1
Σ	23



- **Subtract the mean from each value in the data set and this gives the deviations from the mean.**

- **Square each deviation.**

- **Multiply the frequency corresponding to each of the squared deviations and add them together.**

EXAMPLE

Subtract the mean from each class mark, which gives the deviations from the mean.

Intervals	Frequency f_i	Class Mark (y_i)	$f_i * y_i$	$y_i - \bar{X}$
-----------	-----------------	-------------------------	-------------	-----------------

Total				
--------------	--	--	--	--

- Square each of the deviations with r.

Intervals	Frequency f_i	Class Marks (y_i)	$f_i * y_i$	$y_i - \bar{X}$	$(y_i - \bar{X})^2$
-----------	-----------------	--------------------------	-------------	-----------------	---------------------

Total					
--------------	--	--	--	--	--

Multiply each squared deviation by the frequency.

Intervals	frequency f_i	Class Marks (y_i)	$f_i * y_i$	$y_i - X$	$(y_i - X)$	$f_i(y_i - X)^2$
<hr/>						
<hr/>						
Total						
<hr/>						

- Apply the variance formula.

WRITE DOWN A GLOSSARY

- 1 _____
- 2 _____
- 3 _____
- 4 _____
- 5 _____
- 6 _____
- 7 _____
- 8 _____
- 9 _____
- 10 _____
- 11 _____
- 12 _____
- 13 _____
- 14 _____
- 15 _____

Standard deviation

*The
standard
deviation*

represents

*A low
standard
deviation*

represents

The standard deviation, also called standard deviation, represents the square root of the variance, this measure is used to quantify the variation or dispersion of a set of numerical data. A low standard deviation represents that most of the data are clustered close to their mean, if the standard deviation is high it shows that the data are spread over a wider range.

$$S = \sqrt{S^2}$$

Where

S = standard deviation.

S²= Variance.

EXAMPLE

From example, calculate standard deviation

$$S = \sqrt{3820.62}$$

$$S = 61.81$$

EXAMPLE

From example,

EXAMPLE

From example,

Coefficient of Variation

The coefficient of variation is the standardization of the standard deviation by eliminating the unit of measurement of the variable, this coefficient is also called Pearson's coefficient of variation. Its calculation is based on dividing the standard deviation by the mean of a set of data.

The coefficient of variation can be expressed as CV or R

$$CV = \frac{s}{\bar{x}}$$

EXAMPLE

Using the data of example 2.21, calculate the coefficient of variation

$$X = \frac{6170}{95} = 64.95kg$$

$$S = 61.81$$

$$CV = \frac{61.81}{64.95} = 0.9516 * 100 = 95.16\%$$

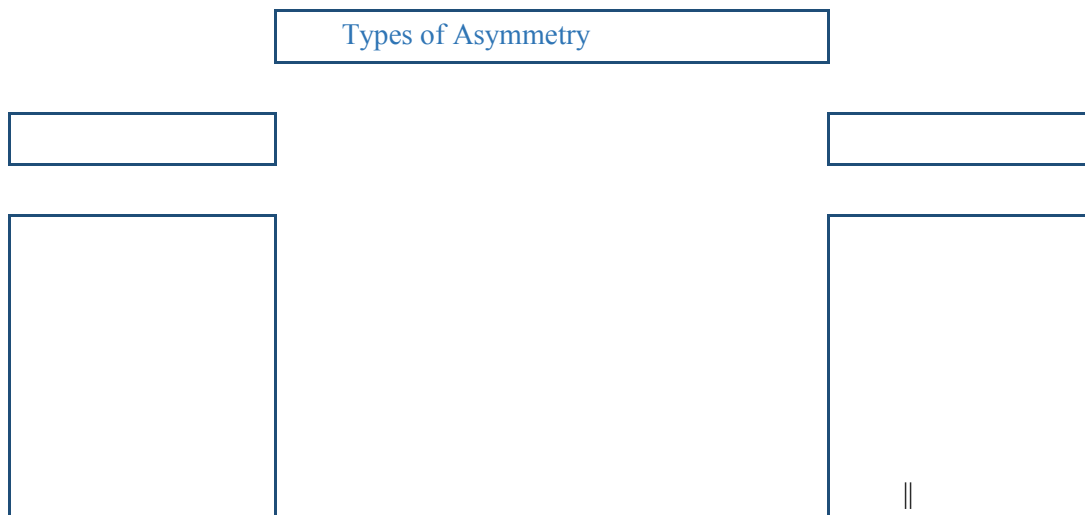
The coefficient of variation establishes the relationship between the standard deviation and the mean, the closer it is to zero, the lower the dispersion of the variable, it has no unit of measurement because the units of measurement of the numerator and denominator being the same are eliminated.

EXAMPLE

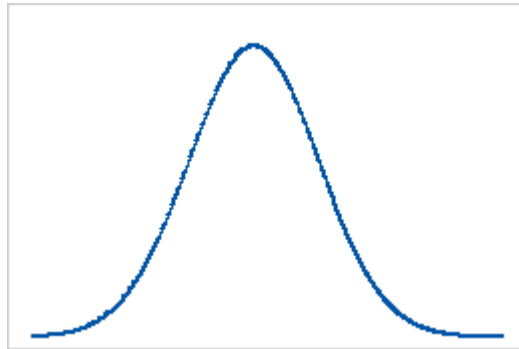
EXAMPLE AND EXPLANATION

Measures of skewness and kurtosis

Measures of skewness and kurtosis



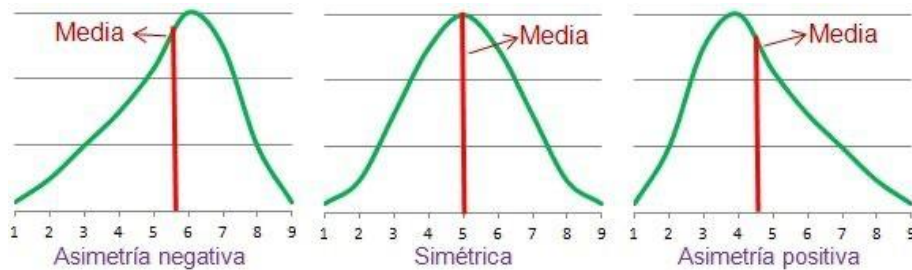
Skewness and kurtosis inform about the shape of the distribution of a variable, these measures allow to know the homogeneity of the data without the need of a graph (Gamero, 2017).



Graphical representation of skewness and kurtosis.

Measures of asymmetry

Skewness is an indicator that makes it possible to establish the degree of symmetry of a probability distribution of a random variable, without the need for graphic representation; perfect skewness exists when the values of the mean, median and mode coincide. Skewness is positive if the mean is greater than the median and negative if the median is greater than the mean, it is a relative value and therefore can be used for comparison purposes.



Types of Asymmetry

$$CA = \frac{\sum_{i=1}^N (X_i - \bar{X})^3}{N * S^3}$$

If the asymmetry coefficient is = 0 then there is a symmetric distribution, if it is greater than zero it is a positive asymmetric distribution and if it is less than 0 it is a negative asymmetric distribution.

Nowadays with statistical programs this measure can be calculated automatically in softwares such as STATA, SPSS.

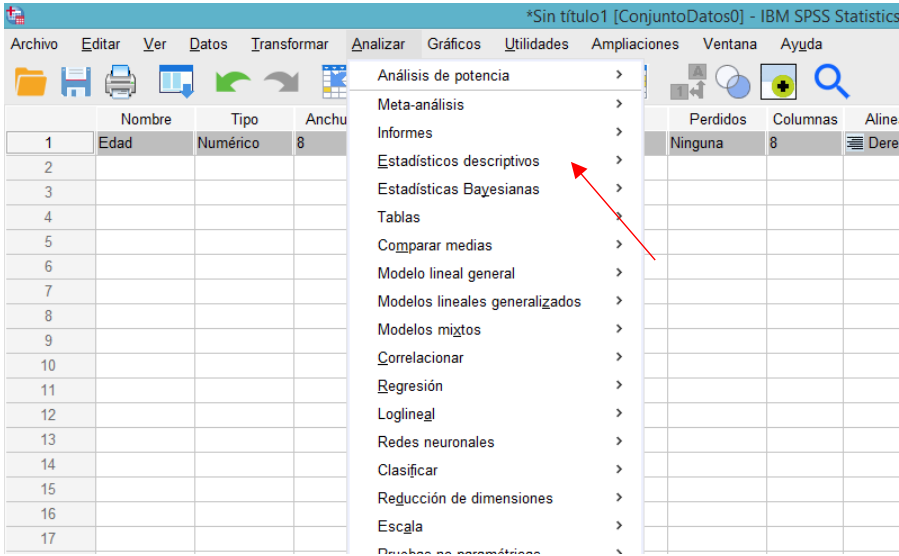
EXAMPLE

The following data are presented, calculate the descriptive statistics using a statistical program

Create the variable age.

When creating the variable identify what type of variable it is, in this case it is numeric; once the variable is created enter the variable data as shown below:

Once the data has been entered, click on analyze, a series of tabs will be displayed as shown below:



Select descriptive statistics, including measure of skewness.

Statistics		
Age		
N	Valid	33
	Lost	0
Media		34,8788
Median		37,0000
Mode		39,00
Desv. Deviation		7,12723
Variance		50,797
Asymmetry		-,982
Standard error of skewness		,409
Range		26,00

In this case the skewness is

GRAPHIC:

Measure of kurtosis

Kurtosis

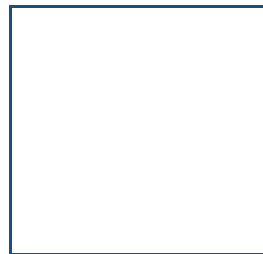
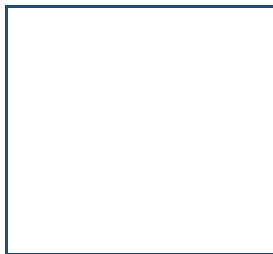
is



Leptokurtic

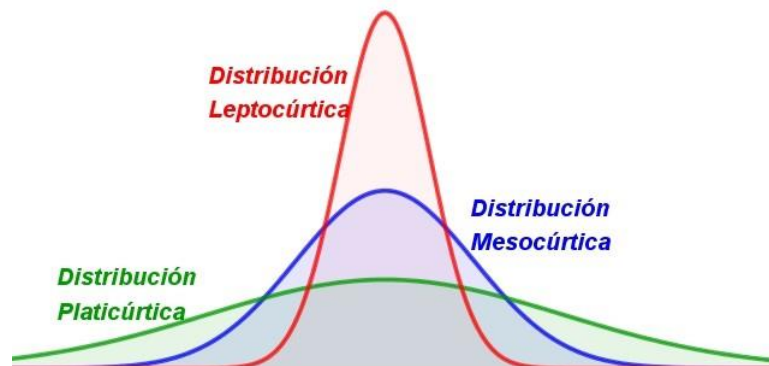
Mesocurtic

Platicurtica



The kurtosis of a statistical variable is a characteristic of the shape of the frequency or probability distribution (Chuhutin & Jespersen, 2017), it is a statistical measure that defines the degree of concentration that the values of a variable present around the central zone of the frequency distribution, it is also called a measure of pointing, Figure 2-3 shows the types of kurtosis that can occur.

Types of Kurtosis.



- **Leptokurtic:** when there is a large number of values concentrated around its mean ($\alpha > 3$).
- **Mesocurtic:** when there is a normal concentration of values around their mean ($\alpha = 3$).
- **Platicurtica:** originates with the existence of a low concentration of values around its mean ($\alpha < 3$).

Depending on the type of data the calculation formula varies

$$\alpha_4 = \frac{1}{N} * \frac{\sum_{i=1}^n (X_i - \bar{X})^4}{S^4}$$

WHERE

S4 = standard deviation raised to the fourth power.

X= mean

N = number of elements in the data set.

EXAMPLE

Danilo works in a maintenance company ABC S.A, his job is to provide gardening care services, below is the number of appointments he made in the last three working days of 8 working hours.

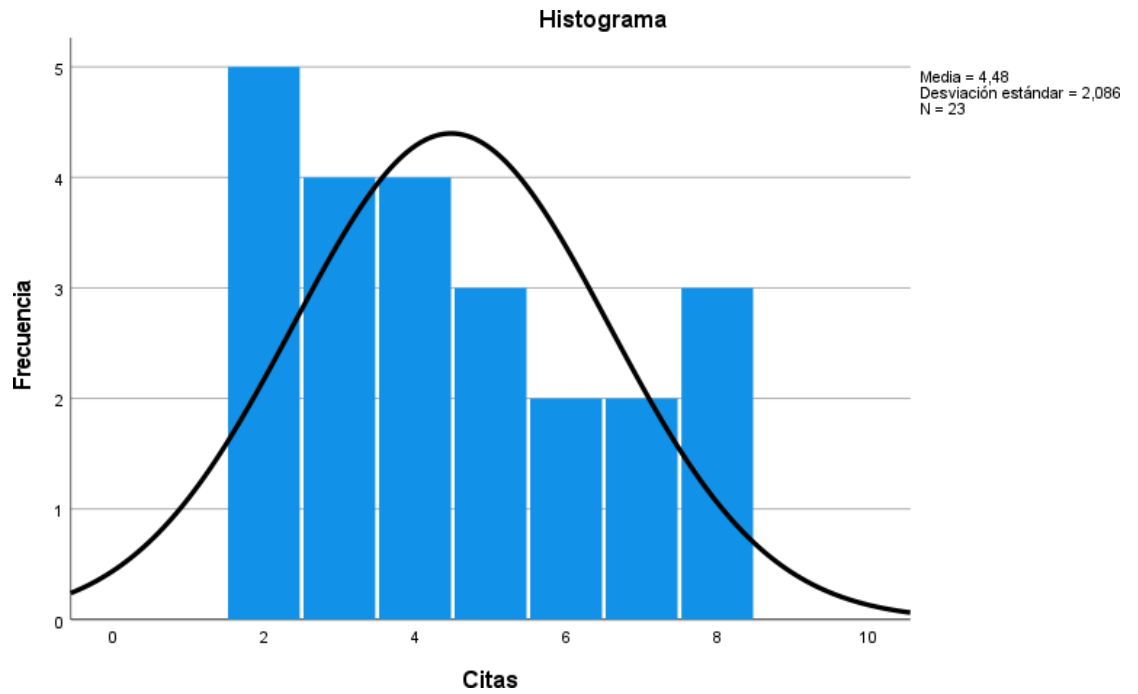
9	5	2	6	2	8	3	7
3	6	3	4	7	5	2	4
5	4	8	3	8	4	2	2

Frequency table.

Quotes					
		Frequency	Percentage	Percentage valid	Percentage accumulated
Valid	2	5	21,7	21,7	21,7
	3	4	17,4	17,4	39,1
	4	4	17,4	17,4	56,5
	5	3	13,0	1,0	69,6
	6	2	8,7	8,7	78,3
	7	2	8,7	8,7	87,0
	8	3	13,0	13,0	100,0
	Total		23	100,0	100,0

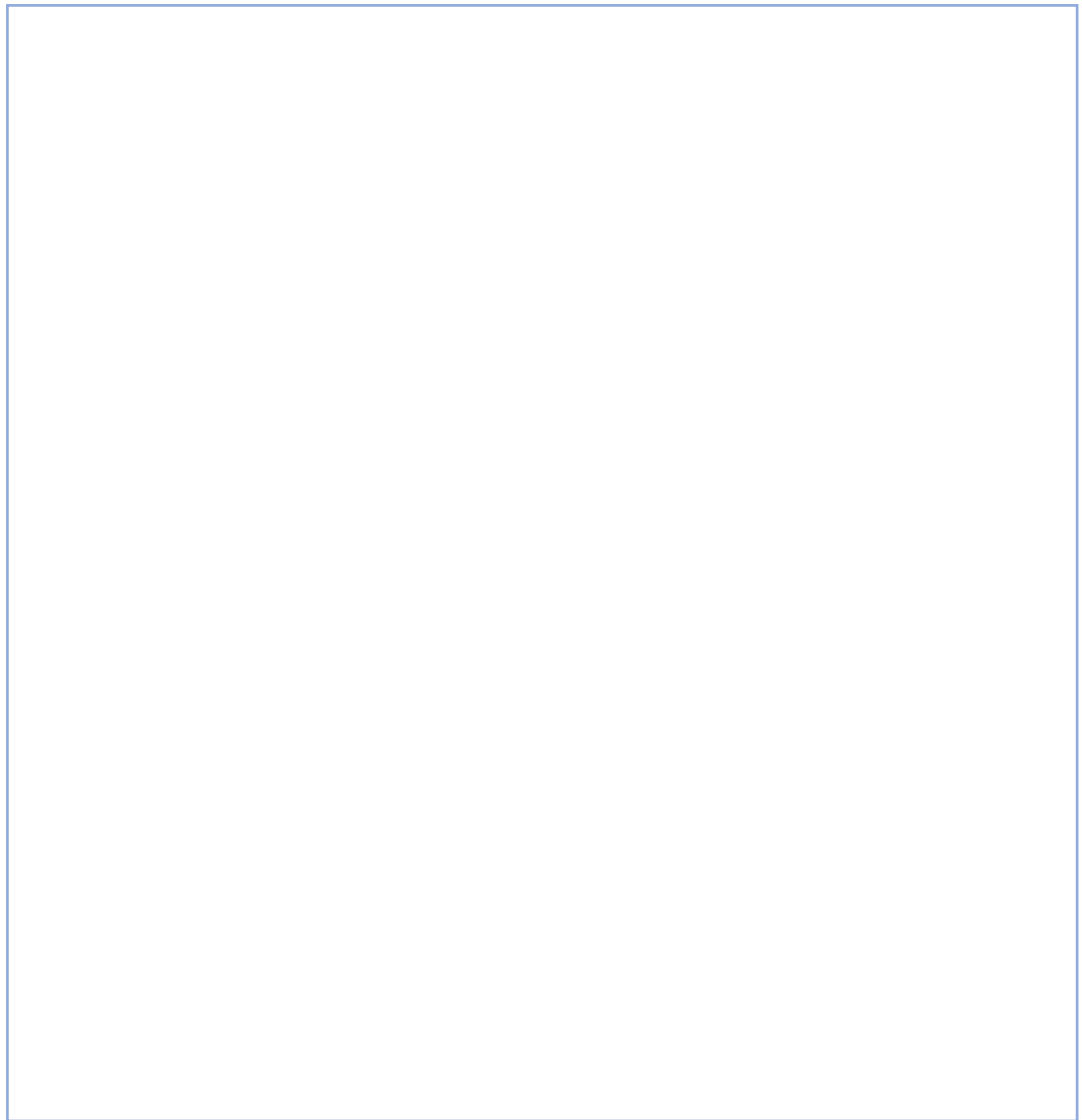
Cumulative percentage

Statistics		
Cites		
N	Valid	23
	Lost	0
Media		4,48
Standard error of the mean		,435
Median		4,00
Moda		2
Desv. Deviation		2,086
Variance		4,352
Curtosis		-1,056
Standard error of kurtosis		,935
Minimum		2
Maximum		8
Addition		103



Distribution chart.

MORE EXAMPLES



WRITE DOWN A GLOSSARY

- 1 _____
- 2 _____
- 3 _____
- 4 _____
- 5 _____
- 6 _____
- 7 _____
- 8 _____
- 9 _____
- 10 _____
- 11 _____
- 12 _____
- 13 _____
- 14 _____
- 15 _____

Bivariate data

Bivariate data ARE _____
EXAMPLES _____
two variables are measured _____
comes from _____
two variables are measured resulting data _____

Two qualitative variables

In statistics, bivariate data are data for each of the two study variables, where each value of the variables is combined with each other, e.g. age and gender, weight and height, it is usual for researchers to be interested in more than a single variable that they can measure in their research. When two variables are measured in a single experimental unit, the resulting data are called bivariate data.

When bivariate data come from two quantitative variables, it is important to analyze the relationship between them, the relationship can be linear, quadratic, exponential, logarithmic, trigonometric, among others, from a statistical perspective the relationship that interests is the linear one, so a linear regression and correlation analysis is carried out.

Bivariate data are expressed through pairs of values; ordered pairs type (a,b); when these values are obtained it refers to a two-dimensional distribution of the information analyzed, each of the two variables can be quantitative or qualitative.

Two qualitative variables

When there are bivariate data of two qualitative variables, it is common to arrange them in a cross-

tabulation or contingency table, which is a particular means to simultaneously represent two characteristics observed in the same population.

A sample of 30 teachers belonging to private and public IES, have 3 types of dedication: Full-time, part-time and part-time as shown in the following table.

Bivariate Data Qualitative Variables.

Type of Entity Time of dedication		
1	Public	Prof_TC
2	Public	Prof_MT
3	Public	Prof_TP
4	Public	Prof_TP
5	Private	Prof_TP
6	Private	Prof_TP
7	Private	Prof_TP
8	Private	Prof_TP
9	Public	Prof_TP
10	Private	Prof_TC
11	Public	Prof_MT
12	Public	Prof_TC
13	Private	Prof_MT
14	Private	Prof_TC
15	Private	Prof_TC
16	Public	Prof_TC
17	Public	Prof_TC
18	Private	Prof_MT
19	Private	Prof_TC
20	Private	Prof_MT

These 30 data can be summarized in a 2*3 crosstabulation table, where the two rows represent the type of entity, and the three columns represent the time of dedication of the teachers, the contingency table shows the frequency for each cross-category.

Cross Table Qualitative Variables.

Entity Type * Teacher Type					
Count		Tipo de docente			
		Prof_MT	Prof_TC	Prof_TP	Total
Type		3	0	0	3
Entity	Private	0	4	4	12

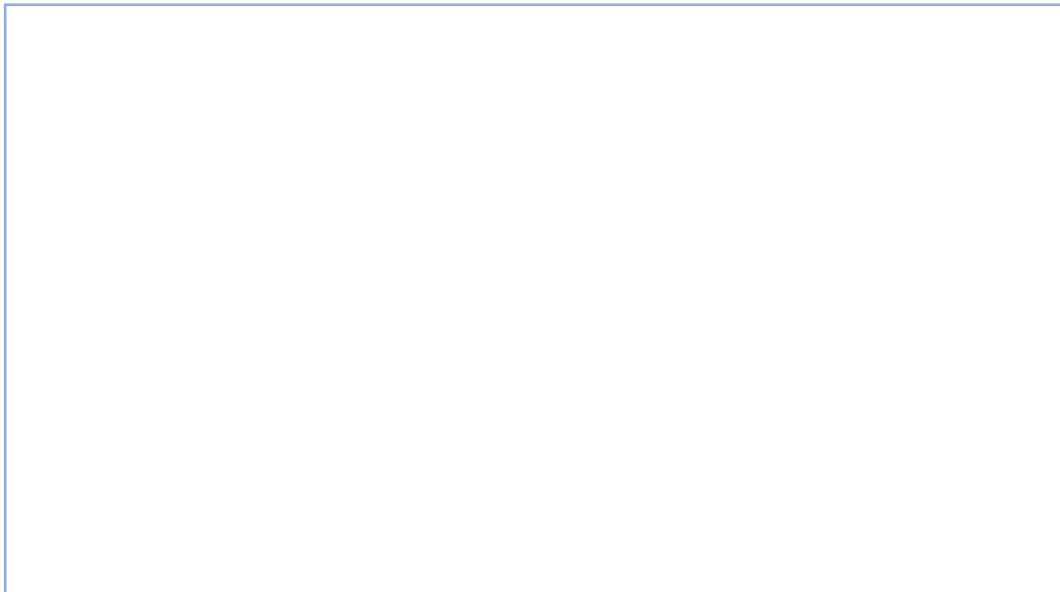
	Public	0	1	4	3	8
Total		3	5	8	7	23

A table with the percentages of the cross categories can also be generated

Cross Table Qualitative Variables percentages.

Entity Type * Teacher Type											
Count											
		Prof_MT				Prof_TC		Prof_TP		Total	
		N		N	%	N	%	N	%	N	%
Type Entity		3		0		0		0		3	
	Private	0		4		4		4		12	
	Public	0		1		4		3		8	
Total		3		5		8		7		23	

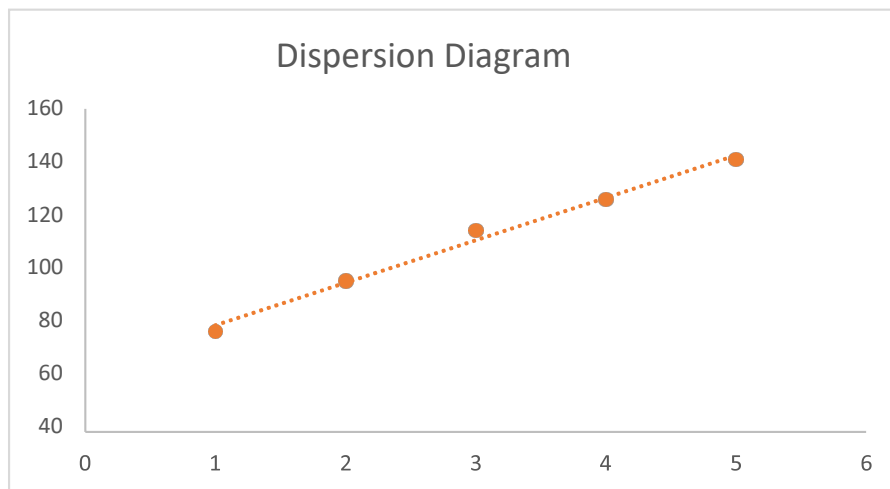
Bar chart



Two quantitative variables

When there are bivariate data of two quantitative variables it is customary to express the data mathematically as ordered pairs (x, y) , where x is the input variable (independent variable) and Y is the output variable (dependent variable), the sample data are represented in a scatter plot.

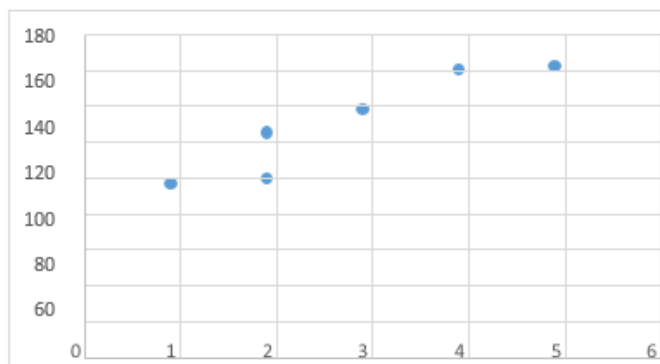
A scatter plot is a graph of all ordered pairs of bivariate data on a coordinate axis system, the X variable is located on the horizontal axis and the output variable on the vertical axis.



EXAMPLE

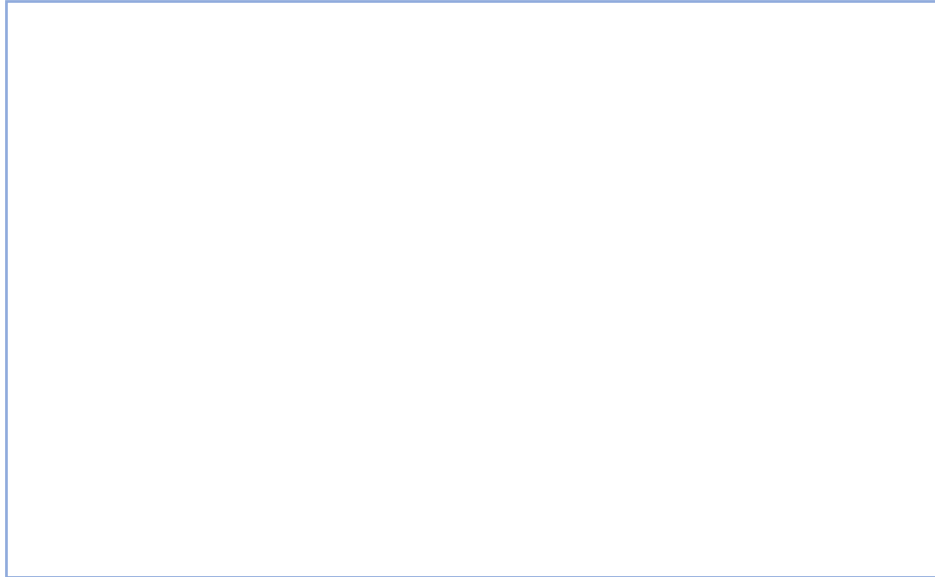
The number of members of a family (X) is measured, as well as the amount they spend in a week on food.

X	2	2	3	4	1	5
Y	95.75	119.10	113.83	159.20	86.85	160.82



Solution: Mark the X and Y axes, plot the points using the (x,y) coordinates for each of the six ordered pairs, the scatter plot is as shown in the figure.

EXERCISE



Linear regression

Linear
regression

IS

It

can help

Linear
regression
techniques

allows
describes

Linear regression is a statistical modeling technique used to describe a continuous response variable as a function of one or more predictor variables. It can help to understand and predict the behavior of complex systems or to analyze experimental, financial and biological data.

Linear regression techniques allow the creation of a linear model. This model describes the relationship between a dependent variable y (also known as the response) as a function of one or several independent variables X_i (called predictors) (Puente et al., 2017). The general equation corresponding to a linear regression model is:

$$Y = \beta_0 + \sum \beta_i X_i + \varepsilon \quad (2.21)$$

Where β_0 represents the linear parameter estimates to be calculated and ε represents the error terms.

There are several types of linear regression such as simple linear regression, multiple linear regression, and multivariate linear regression.

EXERCISES

EXERCISE 1

The following simple series are presented and the median is requested to be calculated.

- a. 2, 3, 3,4
- b. 1, 18, 19, 20
- c. 5.2, 6.5, 8.1, 9.1, 10.1,15.5
- d. 1, 2, 3, 3, 3,4, 7
- e. 9, 40, 80, 81, 100

ANSWER:

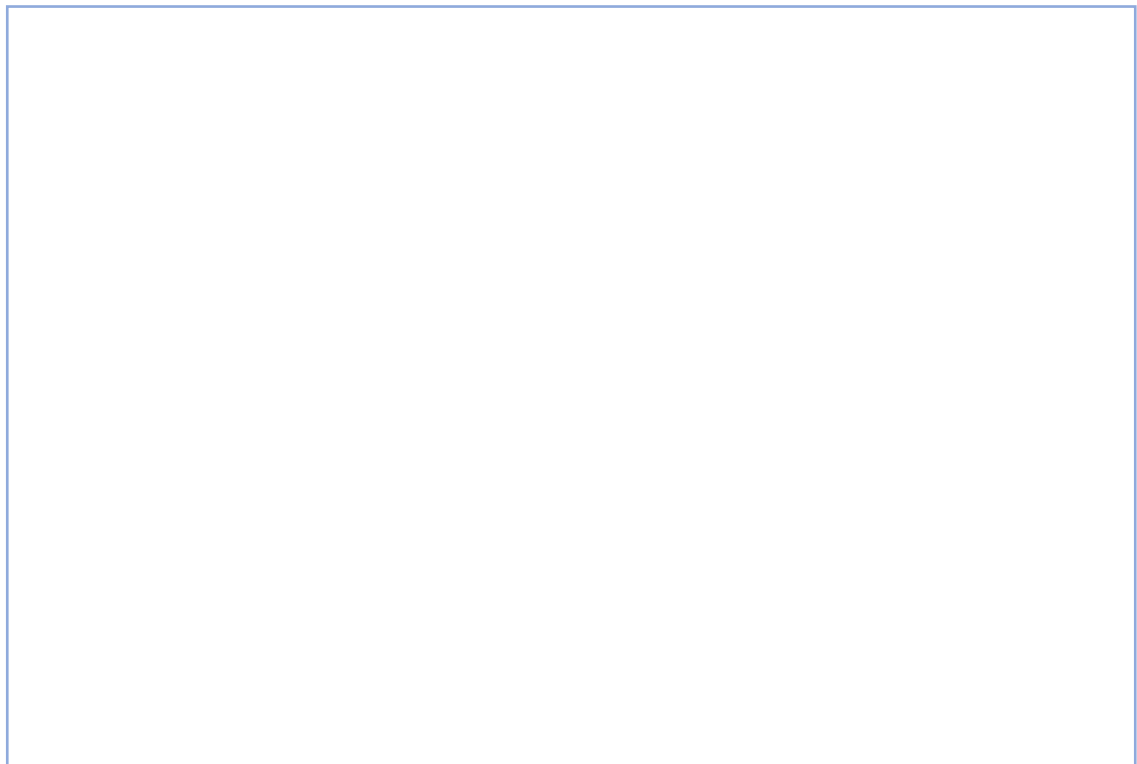
EXERCISE 2

The mathematics final exam scores of 40 ESPOCH remedial students are presented in the following table.

8	5	6	10
7	10	6	9
6	10	3	7
4	10	3	8
3	10	5	8
10	9	5	6
9	8	9	5
7	7	10	5
7	6	10	4
6	5	2	2

From the simple series, construct the series or frequency table

- Construct the table of classes and frequencies following the steps of determining amplitude of variation of, number of classes, class size and tabulation of data
- Mention which are the boundaries of the second class.

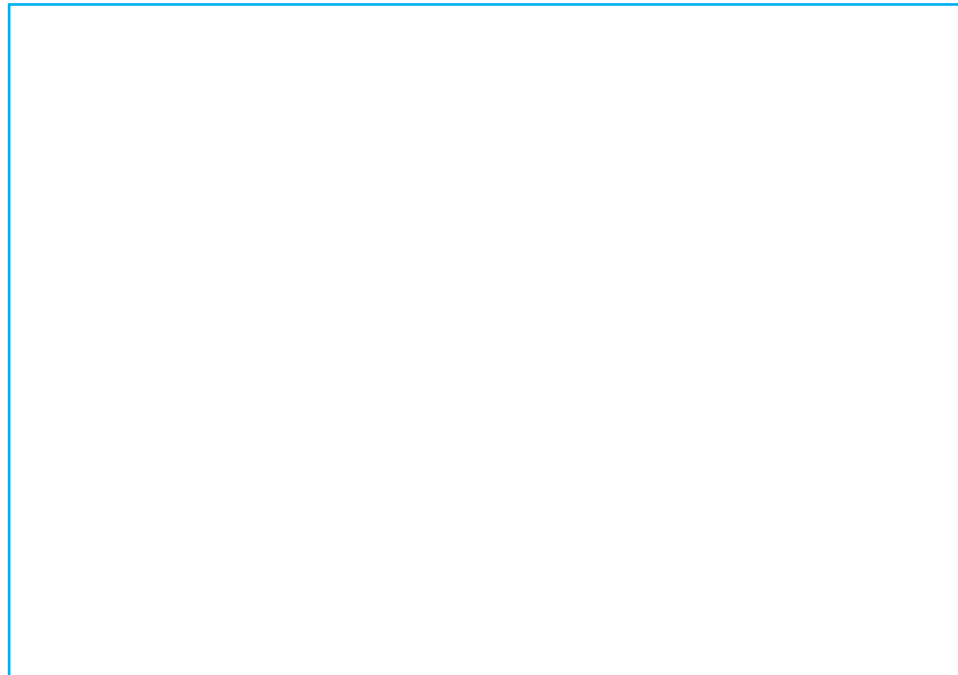


EXERCISE 3

The following table presents data on the height of 36 students of the Electronic Engineering course.

150	160	161
151	160	161
153	159	162
156	167	160
159	167	160
160	169	160
160	145	165
165	150	170
165	151	165
170	152	166
171	150	166
172	153	163

- From the simple series, construct the series or frequency table
- Construct the table of classes and frequencies following the steps of determining amplitude of variation of, number of classes, class size and tabulation of data
- Mention which are the boundaries of the second class.



EXERCISE 4

- Find the median, mean and mode of the following values: 1.2, 1.5, 1.6, 2.1, 2.4, 2.7, 2.8, 3.0, 3.0, 3.0, 3.0, 3.1, 3.1, 3.1, 3.1, 3.4.

- Find the mean, median and mode of the following data: 3, 5, 2, 6, 5, 5, 9, 5, 2, 8

EXERCISE 5

Encontrar la moda, mediana y media de las puntuaciones que se presentan en la siguiente distribución de frecuencia agrupadas.

Interval	Frequencies
40-48,1	3
48,1-56,1	8
56,1-64,1	11
64,1-72,1	32
72,1-80,1	21
80,1-88,1	18
88,1-96,1	14
96,1-104,0	1

EXERCISE 6

The blood ph values of 32 individuals are:

7.33	7.33	7.37	7.32
7.31	7.32	7.35	7.33
7.26	7.35	7.34	7.32
7.33	7.39	7.32	7.4
7.37	7.33	7.29	7.33
7.27	7.38	7.35	7.32
7.30	7.33	7.38	7.34
7.33	7.31	7.32	7.33

Calculate the measures of central tendency, dispersion, make the frequency table.

EXERCISE 7

With the following data calculate the measures of central tendency, dispersion, deciles, and percentiles.

Intervals	Frequencyf_i
1500-2321	9
2321-3142	1
3142-3963	10
3963-4787	0
4787-5606	4
Total	24

EXERCISE 8

Five children aged 2, 3, 5, 7, and 8 years weigh, respectively, 14, 20, 32, 42, and 44 kilograms.

Find the equation of the regression line of age on weight. What would be the approximate weight of a six-year-old child?

EXERCISE 9

The blood ph values of 32 individuals are:

7.33	7.33	7.37	7.32
7.31	7.32	7.35	7.33
7.26	7.35	7.34	7.32
7.33	7.39	7.32	7.4
7.37	7.33	7.29	7.33
7.27	7.38	7.35	7.32
7.30	7.33	7.38	7.34
7.33	7.31	7.32	7.33

Calculate the measures of central tendency, dispersion, make the frequency table.

EXERCISE 10

With the following data calculate the measures of central tendency, dispersion, deciles, percentiles.

Intervals	Frequencyfi
1500-2321	9
2321-3142	1
3142-3963	10
3963-4787	0
4787-5606	4
Total	24

EXERCISE 11

Five children aged 2, 3, 5, 7, and 8 years weigh, respectively, 14, 20, 32, 42, and 44 kilograms.

Find the equation of the regression line of age on weight. What would be the approximate weight of a six-year-old child?

WRITE DOWN A GLOSSARY

- 1 _____
- 2 _____
- 3 _____
- 4 _____
- 5 _____
- 6 _____
- 7 _____
- 8 _____
- 9 _____
- 10 _____
- 11 _____
- 12 _____
- 13 _____
- 14 _____
- 15 _____

Statistical software used for data analysis

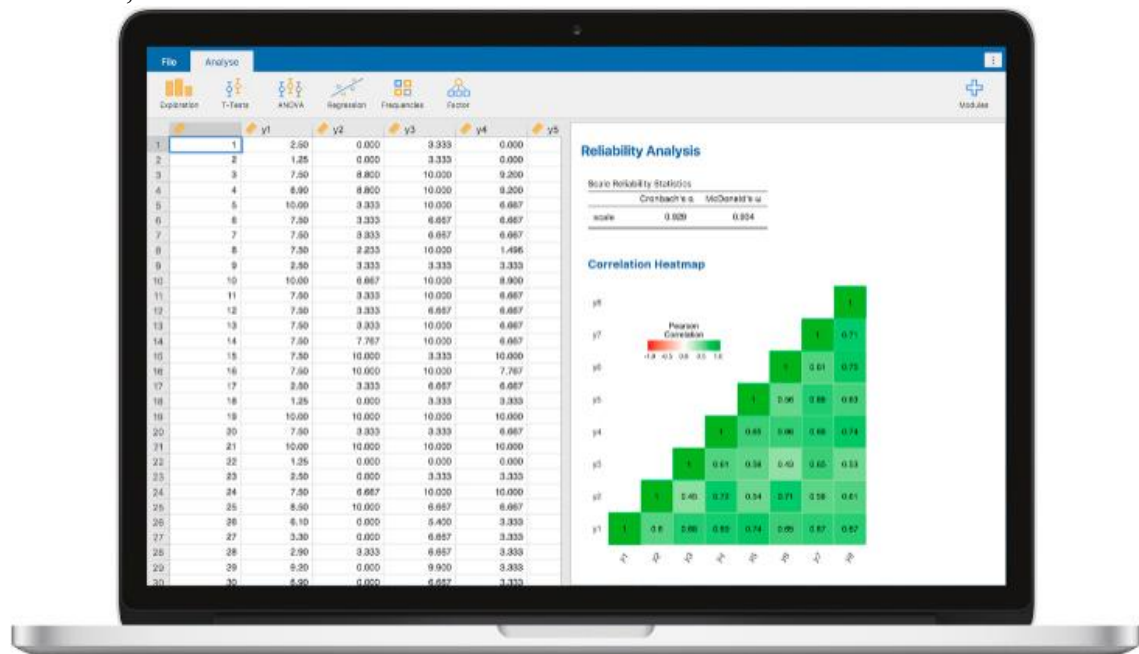
Software JAMOVI

English translation-jamovi is a free and open source software for data analysis and statistical testing. The lead developers of jamovi are Jonathon Love, Damian Dropmann and Ravi Selker, who are developers of the JASP project. jamovi is a fork of JASP.

jamovi is a free and open source distribution software. It aims to simplify some aspects of the use of R. In addition, it can be described as an advanced spreadsheet that allows complex statistical calculations to be performed in a simple and efficient way, using R as the underlying infrastructure and taking advantage of all its benefits.

It includes tools to perform most of the most classic statistical tests in social sciences, as well as the most advanced tests that can be implemented using R, and as an interface a spreadsheet with advanced capabilities to manipulate data, including filtering, selection, combination, etc. operations. It also includes the option of installing additional modules to perform more specific tasks, such as Bayesian models or visual analysis, among others.

One of the interesting features of Jamovi is that it has been adopted as a tool for teaching statistics, and has generated a community around Jamovi that has produced different didactic resources, in the form of examples, manuals, online courses, etc.



What is SPSS software?

SPSS is a software used for data analysis, either for social sciences or health sciences. It is characterized, like the rest of the programs, by having statistical analysis, such as descriptive statistics and differential statistics. Even so, the main difference between the SPSS program and the others is that it covers a larger amount of data, that is to say, it has the capacity to manage large volumes.

- Descriptive statistics is the set of data during an investigation, represented in graphs and/or tables. Basically, they allow us to describe what we have, for example: we have a sample of a patient with X characteristics.

- Differential statistics goes one step further, as different analyses are made to test certain hypotheses, crossing variables, for example: when we want to see if a drug has an effect or not on the improvement of a patient.

Stata: How is Stata software different?

- The Stata statistical package was created in 1985 by StataCorp. Like SPSS, it is a statistical software for data analysis and, although it is also used by research institutions, it has small differences with respect to other programs: - It is distributed in more than 200 countries, so you have easy access; - You can use it on any system, whether Windows, Mac or Linux; - Everything you need is in one package, not in modules; - Flexibility in licensing the program; - When it comes to data management, the tools are infinite; - It is available in more than 200 countries, so you have easy access; - Flexibility when it comes to licensing the program; - When it comes to data management, the tools are infinite.

A: How is R software different?

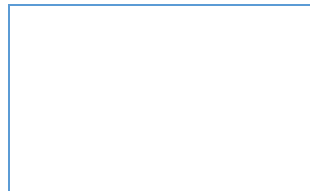
It is a programming environment that has a great advantage: it is totally free. In 1993, Robert Gentleman, together with Ross Ihaka, developed this tool in order to contribute new functions accessible to everyone. In fact, the name "R" comes from the old programming language "S".

This software is recommended for all data scientists (Data Scientist) who want to go deeper into the tool. To help you make up your mind, here are the benefits of using this statistical program:

- You will create high quality exportable graphs - You will consume few computer resources.
- You will work with other languages: it allows you to read data from other programs such as SPSS.
- You will create interactive web applications (it offers you that possibility) - You will solve complex statistical problems

STATISTICAL ANALYSIS PACKAGES

***STATISTICAL
ANALYSIS
PACKAGES
IS***



THEY ARE



What is statistical software?

Statistical software is a computer program whose main purpose is to automate and perform complex data analysis. This type of tools usually provide useful mechanisms for the organization, interpretation and presentation of data selected for statistical analysis in an easy and simple way. This type of programs are generally used by specialists in the area of statistics. However, since they are

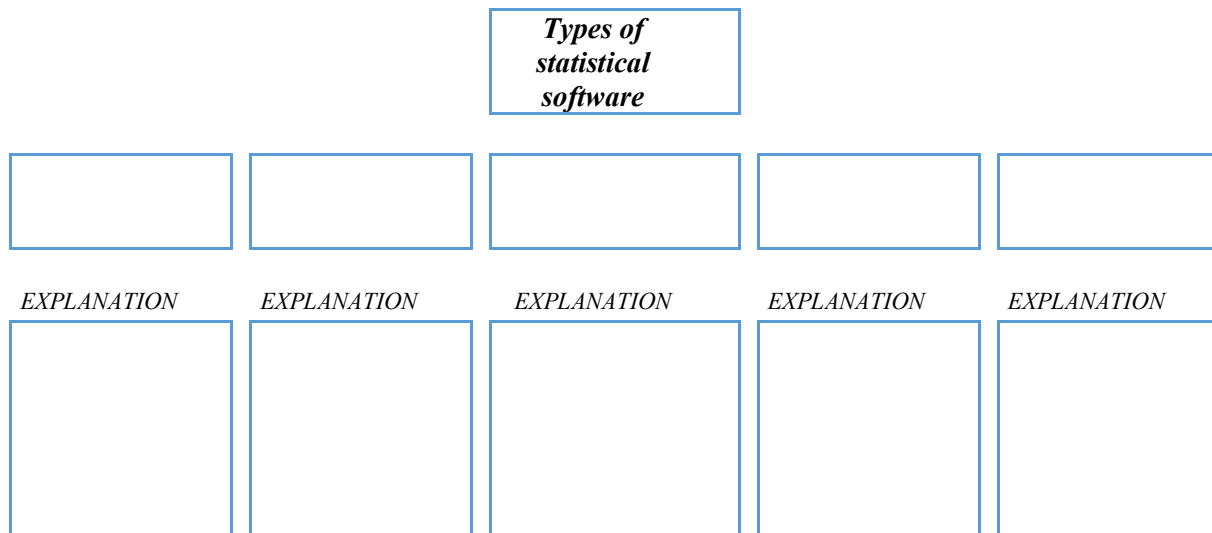
easy to access and easy to use, they can be used by anyone who requires these tools to enhance their research.

What is statistical software for?

As we have already mentioned, this type of statistical software is used to automate and facilitate data processing. Whatever the data analysis to be carried out, these tools make it easy to identify correlations, regressions and variations between the variants studied.

From the simplification of these data, decision making in a given investigation is made possible. In addition, this type of software allows easy identification of duplicate data or unusual information that may appear during data curation. At the same time, they provide detailed reports and easy-to-interpret graphs.

TYPES OF STATISTICAL SOFTWARE



There are countless such tools on the market. Below, we will detail the most common and easy to use statistical software:

SPSS

This statistical software was acquired by IBM in 2009. SPSS was originally developed to support the analysis and management of data related to social science research. Currently, it has become widespread in the academic world, covering various fields of study.

This program is very easy to use, although it is always advisable to have a basic knowledge of data analysis. It is a good option if you need to carry out descriptive statistics or factor analysis, among other methodologies.

Excel

This tool is undoubtedly the best known of all. Its use is very standardized in all fields of study and does not require expensive courses to take full advantage of this tool.

With this tool, you will be able to obtain and filter the most useful data for your research. At the same time, you will be able to make graphs that condense this information, so your audience will be able to read the data quickly and easily.

SAS

It is among the best statistical software available today. It is the favorite of the strongest industries in the market, such as Bank of America, Orlando and Honda, among others.

The innovation of this program is that it incorporates artificial intelligence to automate data analysis. Therefore, this tool allows advanced analysis, multivariate, predictive and data management. Like Python, the SAS user must have basic knowledge of statistics and data analysis, but it will be worth the knowledge and investment.

STATA

This tool, created in 1985, is a statistical software package that provides reliable results for your research projects. Besides, it is an easy-to-use program for all kinds of people, regardless of the level of knowledge they have about data analysis.

So, if you are looking for a user-friendly and easy-to-use tool, you should definitely count on STATA. Despite this, the information it provides is 100% reliable.

JAMOVI

<https://estudiosestadisticos.ucm.es/aplicaciones-informaticas-a-la-estadistica>

I. SPSS PACKAGE

SPSS PACKAGE ENVIRONMENT

- The graphical user environment.
- Getting help.
- The graphical type and interactive editors.

FILE MANIPULATION AND DATA TRANSFORMATION

- Data entry.
- Sorting, selection and weighting of cases.
- File segmentation. o - Summary files and combination of files.
- Variable manipulation. File segmentation.
- Summary files and file combination.
- Manipulation of variables.
- Variable manipulation. Functions.
- Emodification of the values of a variable.

REPORT GENERATION

Summaries with deblocked statistics.

Statistical tables.

Multiple responses.

THE SPSS COMMAND LANGUAGE

Control structures.

SYSTEM SAS

SAS SYSTEM ENVIRONMENT

- The graphical user environment.
- Management of SAS data libraries.
- The SAS SQL query window.

WORKING WITH DATA FILES

- Acquiring data from external data files.
- Specifying external file structure and reading control.
- Importing external files with the graphical external file environment.
- Acquiring data from SAS dataset.

PROGRAMMING IN THE DATA STEP

- The processing of the DATA step. Flowchart.
- Generic structure of the DATA step.
- Specific treatment of variables according to their attributes. Functions.
- Manipulation of observations.

COMBINATION OF SAS DATA SETS

- SAS dataset concatenation.
- SAS dataset interleaving.
- SAS dataset merging.
- SAS dataset updating.

SAS SESSION ANALYSIS

- Analysis of the SAS session with the SAS log.
- Routing of SAS output and SAS log.

- Diagnosis and debugging of errors using the log window.

CORE PROCEDURES

- Introduction to the PROC step. Common sentences.
- Creation of standard and customized reports.
- Creation of summary tables.
- Deblockediva statistics procedures.

GRAPHICS WITH THE SAS SYSTEM

- Elementary graphics.
- Advanced graphics.

SAS MACRO LANGUAGE

- Macro variables. Scope of Macro variables.
- Particular aspects of Macro processing.
- Macro expressions and Macro dimensioning.
- Storing and reusing Macros
- Error messages and debugging.

BIBLIOGRAPHIC REFERENCES

American Statistical Association (2016). Ethical Guidelines for Statistical Practice. Obtenido de: <https://higherlogicdownload.s3.amazonaws.com>

Anderson, D., Sweeney, D. & Williams, T. (2012). Estadística para los negocios (11ava ed). CENGAGE Learning.

Capalar, C., & Dönmez, A. (2016). What is scientific reseach and how can it be done?. Turkish Journal of anaesthesiology and reanimation, 44(4), 212.

Cardenas, R. (2014). Estadística en la Educación. UNID Editorial Digital Chuhutin, A., Hansen, B., & Jespersen, S. N. (2017). Precision and accuracy of diffusion kurtosis estimation and the influence of b-value selection. NMR in Biomedicine, 30(11), e3777.

Cruz, C., Oliverez, S & González, M.(2014) Metodología de la Investigación. Grupo editorial Patria.

Díaz, M. (2019). Estadística Inferencial Aplicada. Universidad del Norte Gamero, C. (2017). Estadística I. Elementos de Estadística Descriptiva y teoría de Probabilidad. Universidad de Málaga.

García, J., Ramos, C & Ruiz, G. Estadística Administrativa. Universidad de Cadiz.

Guerra, C. (2003). Estadística. Editorial Felix Varela

Hernández, C. & Oteyza, E. (2015). Probabilidad y Estadística. PersonEdición.

Hernández, M., Hernández, S & Papiá, M. (2019) Estadística inferencial 2: Aplicaciones para ingeniería. Editorial Patria

Linas, H. (2017). Estadística Inferencial. Universidad del Norte

Lind, D., Marchal, W., & Wathen, S. (2012). Estadística aplicada a los negocios y la economía (15 va ed). Mc Graw Hill

Martínez, E. (2020). Estadística. Ediciones UAPA. Matus, R. (2010) Estadística. Instituto Politécnico

Mias, C. (2018). Metodología de la investigación, estadística aplicada: Guía práctica para investigación. Editorial Brujas.

Mondolfo, R. (1961). Problemas y métodos de investigación en la historia de la Filosofía. Edit Universitaria Buenos Aires.

Monroy, M. & Nava, N. (2018). Metodología de la Investigación. Lapislazuli Ediciones.

Monroy, S. (2008). Estadística Descriptiva. Instituto Politécnico nacional Mora, M. & Sepulveda, P. (1999). Metodología de la Investigación Murray & Spiegel, Estadística (4ed). Mc Graw Hil

Pérez, H. (2008). Estadística para las ciencias sociales, del comportamiento y la salud, 3ed. Cengage Learning

Puente, C. (2018). Estadística descriptiva e inferencial. Ediciones IDT.

Puente, M., Viñan, J & Aguilar, J. (2017). Planeación Financiera y presupuestaria. Escuela Superior Politécnica de Chimborazo.

Sautu, R., Freidin, B., Najmias, C., Otamendi, M., Paredes, D., Ballesteros, M., ... & Trepiana, M. (2014) Metodología de la Investigación. Universidad de Buenos Aires

-Camacho Rosales J. "Estadística con SPSS (versión 11) para Windows ". Ra-ma. 2002.

Ferran Aranaz M. "Curso de SPSS para Windows". Mcgraw-Hill. 2002.

FIGUEIRA LÓPEZ, E. 2001. "Análisis de datos con SPSSWIN". Alianza. 2001.

GEORGE, D. & MALLERY, P. "SPSS for Windows Step by Step". Allyn and Bacon. 2000.

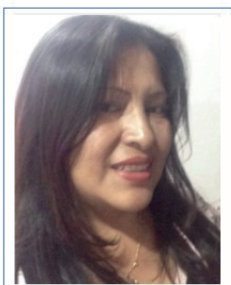
GREEN,S.B. & SALKIND, N.J."Using SPSS for Windows: Analyzing and Understanding Data".Prentice Hall.2003

KIRKPATRICK, L.A. & FEENEY, B.C. "A Simple Guide to SPSS for Windows, 5e". Duxbury Press. 2003.

NORUSIS, M.J. "SPSS 11.0 Guide to Data Analysis". Prentice Hall. 2002.

SHANNON, D.M. & DAVENPORT, M.A. "Using SPSS to Solve Statistical Problems: A Self-Instruction Guide". Prentice Hall. 2001.

SPSS Inc. "SPSS 11.0 Syntax Refence Guide". Prentice Hall. 2001.



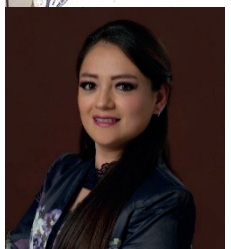
Lourdes Emperatriz Paredes Castelo

Ingeniera en Sistemas Informáticos (ESPOCH). Magister en Informática Educativa (ESPOCH), Docente – Investigador en la Escuela Superior Politécnica de Chimborazo (ESPOCH) hasta la presente fecha. Actualmente investigadora del grupo de Ciencia de datos CITED – ESPOCH.



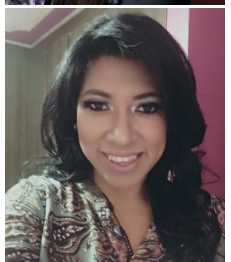
Johanna Enith Aguilar Reyes

Ingeniera en Estadística Informática (ESPOCH). Magister en Gestión y liderazgo educacional (UTPL). Magister en Estadística aplicada (UPEC), Docente – Investigador en la Escuela Superior Politécnica de Chimborazo (ESPOCH) hasta la presente. Actualmente investigadora del grupo de investigación CITED – ESPOCH.



Geoconda Marisela Velasco Castelo

Ingeniera Mecánica (ESPOCH). Magister en Física Aplicada Mención productivos (ESPOCH), Docente – Investigador en la Escuela Superior Politécnica de Chimborazo (ESPOCH) hasta la presente. Actualmente investigadora del grupo de investigación CITED – ESPOCH.



María Yadira Cárdenas Moyano

Magister en Lengua Extranjera de la Universidad Técnica de Ambato, Investigador en la Escuela Superior Politécnica de Chimborazo (ESPOCH) hasta la presente.

ISBN: 978-9942-33-819-8



   @grupocompas.ec
compasacademico@icloud.com