Quantitative Analysis: the guide for beginners

Book · June 2019

CITATION
1

READS
24,901

1 author:

Julián Cárdenas
University of Valencia
42 PUBLICATIONS **563** CITATIONS

SEE PROFILE

QUANTITATIVE ANALYSIS

The guide for beginners

JULIÁN CÁRDENAS

QUANTITATIVE ANALYSIS

THE GUIDE FOR BEGINNERS

By

Julián Cárdenas

julian.cardenas@onlinebschool.com

www.networksprovidehappiness.com

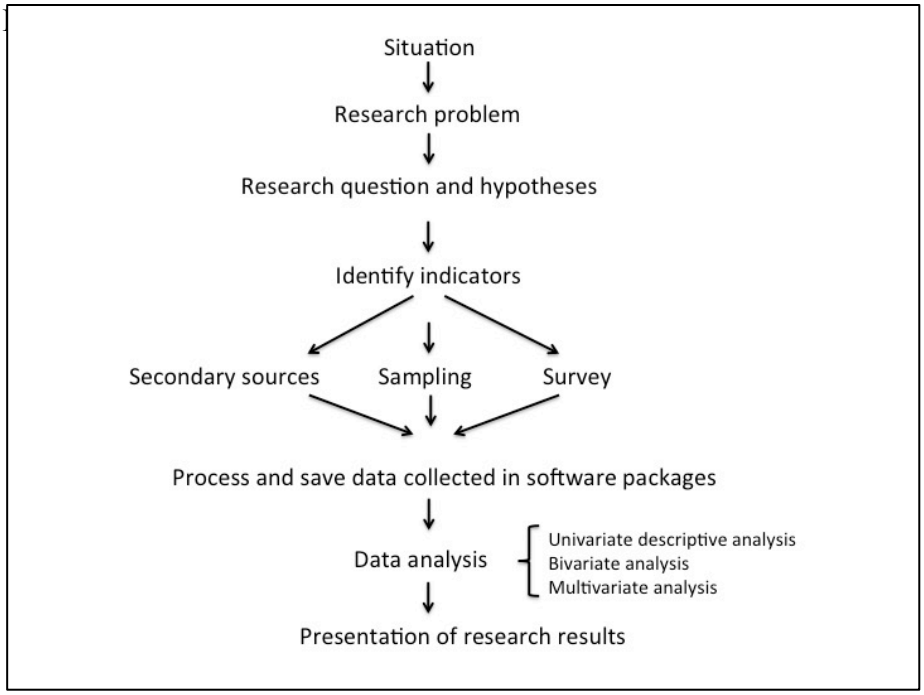
To the beginners

Index

Introduction. The research process

The best way to increase our knowledge and get control of our lives is to do research. Results based on empirical data allows us to make more reasonable decisions, find out what happens beyond our eyes, and predict what will occur in the future. Why has life expectancy increased significantly in recent decades? Why is it cheaper to travel today than 30 years ago? Why has the number of homicides declined in most countries? Research had something to do with it. For all this and many more, doing research is worth it.

Research is a process designed to resolve questions based on the collection and analysis of data. Doing research requires time, human and material resources, but above all to know the process of conducting a research. The following table displays the major phases of research.



Source: own elaboration

The origin of any research is the problem. The problem (also called issue) is a concise relevant and contextualized aspect of a situation that interest to investigate. The research question is what we want to find out about the problem and it is the most relevant phase of a research process. Answering the research question is the goal to achieve. The research question is built on literature review, critical observation of reality, and discussion with experts and actors from the field. Subsequently, the phenomena of interest pointed in the research question are observed, measured, and quantified by indicators. Once we know what to measure, data is collected. Data can be obtained from primary sources (the researcher generates the data) or secondary sources (the data was generated by other researchers outside the project). One way to generate the data is through a survey. This

research technique is a set of questions and response categories applied to a sample of actors. Data collected either by survey or through secondary sources are processed in variable and data sheets. To process, save and later analyze the data, computer software packages are employed, such as Excel, SPSS, PSPP or Stata. Data analysis is executed according to the type of variables and the number of variables to analyze. The statistical techniques to analyze one variable are tables of frequencies, mean and standard deviation; and to analyze the relationship between two variables are cross-tabulation tables, correlations and one-way ANOVA, among others. Data analysis aims to answer the research question(s) posed and to identify other trends. After analyzing data, results are be presented in a research report or article. Thus, the research process is completed.

The objective of this handbook is that readers become capable to conduct research following a quantitative methodology. This is a manual to understand and practice all the phases of the research process, and to show how to analyze data through basic statistical techniques. The present book is composed of 11 chapters or sessions. Each chapter represents a phase of the research, and it displays the main ideas, suggested bibliography and practical exercises. This book is intended for anyone who has a need or desire to conduct research applying quantitative methods and analyzing quantitative data. Before to start, what is quantitative data?

1. Types of research methods and main research techniques

Research is a process of asking questions and answering them by collecting and analyzing data. Data can be converted into numbers, words or images. When data are numbers (or the information collected is transformed into numerical scales), we are conducting research with quantitative data. When the data collected are words or images (not transformed into numerical scales), we conduct research with qualitative data. This is the main difference between quantitative and qualitative research. In quantitative research, data to answer the research question are numbers. In qualitative research, the data collected are words or images that are not synthesized in numbers.

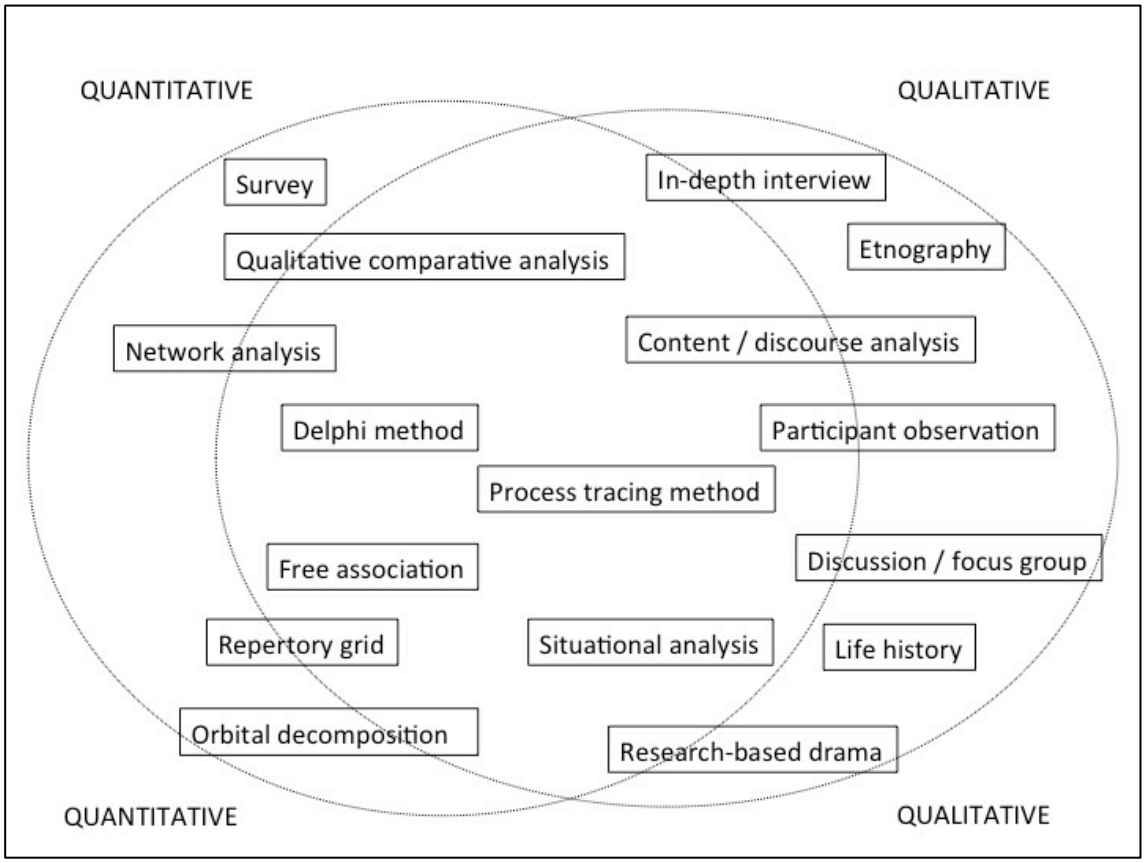
The method is a stage of the research process focused on how the research question will be answered. Behind the methods, there are a series of paradigms and theories that indicate the strengths of each one. The decision to opt for one method or another depends on the objectives of the research, the available data, and obviously the possible resources to carry out one type of research or another. These research methods are also known as research approaches.

Quantitative method	Qualitative method
Based on the positivism logical that seeks to find laws that explain the reality	Based on phenomenology that aims to understand in depth the point of view of others
Directed to measurable and quantifiable data	Directed to the experiences of the participants
Usually employed for explanation purposes	Usually employed for comprehension purposes
Search relationships between phenomena	Search the depth understanding of phenomena
Focused on the outcomes	Focused on the process
If the study is based on representative samples, results are generalizable to the population. Allow making inferences	Research results are not generalizable to the population, although they are transferable
Work with many cases	Work with few cases
Statistical analysis	Content analysis
Identification of trends, comparison of groups, relationships between variables	Identification of categories and description of themes
Numerical data	Data in words or images

Despite these distinctions features between quantitative and qualitative method, both approaches are usually combined and integrated. Many studies are a combination of quantitative (numbers) and qualitative (words or images) data. The combination of quantitative and qualitative research methods is called mixed methods. For example, first, numerical data are collected and analyzed to appreciate to what extent a phenomenon arises and to select a segment of the population. Subsequently, in-depth interviews are applied to people from the chosen population segment and the responses of the interviewees are compared with the numerical data of the general population. Therefore, quantitative and qualitative methods are not opponents. Teaching separately quantitative and qualitative methods is due more to operational processes that facilitate understanding and organization. In practice, quantitative and qualitative methods are combined very frequently.

The research techniques are the tools to collect data. Figure 2 represents the main social research techniques and their position in the spectrum of quantitative and qualitative methodology. There is a high overlap between quantitative and qualitative methods because research techniques use very often elements of both approaches.

Figure 2. Social research techniques and types of research methods



Source: own elaboration

2. How to define and delimit a research problem

The first step in research is to define and delimit a research problem, and the second one is to formulate a research question based on the research problem. "A research problem is a definite or clear expression statement about an area of concern, a condition to be improved upon, a difficulty to be eliminated, or a troubling question that exists in scholarly literature, in theory, or within existing practice that points to a need for meaningful understanding and deliberate investigation" (Bryman, 2007). The research problem is what we already know, and the research question is what we want to know or find out about the problem. The research problem and the research question interact throughout this initial phase until we can define and delimit the problem, and then until we manage to formulate the research question (Figure 3).

Figure 3. Problem and research question

How to move from general situations or ideas to a defined and delimited research problem

Hundreds of situations happen around us: people living in the streets, absenteeism in school, drug trafficking, child prostitution, happiness, presidential elections, and attendance at football stadiums, among others. Situations are very broad ideas that must be defined and delimited to find a specific problem to investigate. We have to transform general situations or topics of interests into a defined and delimited research problem. To do that, we should follow a protocol of two major steps.

1. Description of the situation – List of problems

The first step is to make a general description of the situation we observe in order to list more specific problems. These problems are identified through an analysis of actors involved, points of view, experiences, context, prejudices, causal factors, possible consequences, and suggestive theories. Useful in this description are images, videos, direct observation of the situation, conversations with people involved, with experts, and literature review on the situation. For example, if there are many people living in the street we have to know what policies exist in that place to solve the problem, what has been studied on it, what other problems are associated with that situation such as alcoholism or isolation, what other actors are involved around this situation, support institutions,

what homeless think about it, what the experts argue, why people help them or not. Once we have more knowledge of the situation and a list of possible problems to investigate, we must define and delimit them.

2. Definition and delimitation of the research problem

In this second step, we have to choose one of the problems identified in the situation and define seven aspects of this problem. These aspects are queries that answering them help us define and delimit the problem.

a. Conceptual delimitation: define the problem in a concise way. We have to point out:
   - How this problem is usually named in the literature
   - How this problem is defined in other studies

   In this way we define **what to research**.

   For example, if we are interest in people living in the streets, we have to be aware that this problem is named homelessness, and we have to define this concept on basis of previous studies.

b. Relevance of the problem: why this problem is important to investigate. We have to indicate:
   - Risks of the problem
   - Associated economic, human and political costs
   - If the problem has increased in recent years or decreased
   - If the problem has spread in different places
   - If it is a recognized or unknown problem.

   In this way we aim to define **why to conduct research on this problem**.

c. Applications: what are possible practical, theoretical and academic uses of the research. We have to answer this queries:
   - What are the impacts of increasing knowledge of this problem?
   - To what fields or areas would increasing knowledge of this problem contribute?

   In this way it is possible to define **the what for**.

d. Physical-geographical delimitation. It is necessary to expose:
   - Where this problem occurs or occurred.
   - Where will the problem be studied? The problem can happen in many places, but we have to clarify why we are interested in that place.

   In this way it is delimited **where the problem happens and is studied**.

e. Temporal delimitation: it is possible to show when this problem occurs, and also at what moment, stage or years it is interesting to investigate it. The problem can happen over time or be eternal, but we must point out why we are interested in

that time or specific years. In this way, we delimit **when the problem happens and is studied**.

f. Background: this aspect refers to previous studies on the problem. We have to expose:
   - What is known about this problem and what is not known?
   - How it has been investigated: data, methods?
   - What debates are on this problem in the literature?

   We, thus, identify what is not yet known about this problem and define **what can be the contribution of our research and the lack that will cover**.

g. Interrogative formulation: we must question the problem about the causes, consequences, possible solutions, and comparisons with other places or other moments of time. Thus, we define **what we want to find out**. This aspect is a way to formulate possible research questions, which it is the matter of the next chapter.

It is recommended to write all these aspects of the problem, and then evaluate if the problem is well defined and delimited, and if it is worth doing research on it. Only through writing it is possible to improve and correct the definition and delimitation of the problem. A paragraph of maximum 10 lines can be enough to appreciate if the problem is well defined and delimited.

How to know if the problem is well defined and delimited

For a research problem to be well defined and delimited, it must meet the following conditions:

- ✓ Concise: the problem must be very clearly posed and stated, and easily understandable.
- ✓ Relevant: although importance of the problem can be very relative and subjective, the problem must be presented so that it is worth investing resources to investigate it. For example, it may seem more important to study poverty than leisure spaces in parks. However, the relevance of a research is determined by the detail of costs, risks, advantages, extension and possible contributions. It is the section where our research idea is "marketed".
- ✓ Contextualized: the problem must be delimited in geographical space, time and background. The exposure of these aspects is what gives a specific vision and the integral time of what is analyzed.
- ✓ Specific: it is recommended that the research problem be something concrete since research must be achievable, and trying to cover too much is one of the main mistakes when defining and delimiting the research problem.

Another way to assess whether the problem is well defined and delimited is to formulate research questions with the problem. The next chapter aims to show how to formulate research questions.

Other resources:

To understand the situations and to be able to detect research problems, some other tools such as problem tree, a map of actors and the SWOT analysis (strengths, weaknesses, opportunities, threats) are also useful.

Exercise

From this photograph that represents a situation, define and delimit a research problem. Apply the steps suggested in the chapter.

Figure 3. Situation

References

Bryman, Alan (2007) "The Research Question in Social Research: What is its Role?" *International Journal of Social Research Methodology* 10: 5-20.

3. How to formulate research questions

The research question (RQ) is the goal that we aim to answer, and the guide during the entire research process. We cannot start to write or prepare a research proposal or research project if we have not formulated a research question. We cannot choose techniques, theories or data if we do not have a research question. It is better to waste days, weeks, months or years looking for a research question than to begin a study without having a research question. If the research question changes once the investigation initiated, we must restart the research process and review the problem, state-of-art and methods previously selected. Finding a "good" research question is a challenge for all those who initiate a thesis or research proposal.

Following a 3-step protocol and applying six strategies is the most efficient way to formulate research questions and avoid being lost and aimless during the writing of the research proposal and the fieldwork. But before writing the research question, it is necessary to take into account the conditions that must be met.

Conditions of a research question:
The research question must be:

- Concise: simple and clear language. Anyone, even without expertise in the field, must understand the research question. Short and direct phrases are appropriate instead of complex and pretentious sentences.
- Achievable: the research question must have a possible answer and data collection to answer it must be feasible.
- Relevant: answering the question should contribute to solve problems, increase the lack of knowledge, generate new debates, and produce other impacts at a theoretical, empirical, political, economic and social level.

If your research question does not meet any of these conditions, it is not a worthy question that deserves to conduct a professional research.

3 steps to formulate a research question
There is a protocol to devise research questions. The 3 steps are:

1) **Define a research problem**: this phase is detailed in the previous chapter. For example, happiness is not a problem, it is too general, and it must be defined more specifically. Following this example, we note that there are places in Latin America with many social conflicts and high inequality, but its population feels very satisfied and manifests to be happy. Data from the World Values Survey or Happy Planet Index indicate that Panama, Colombia, Venezuela, Ecuador, Costa Rica and Honduras are countries with a very high level of happiness compared to the regions of Europe and North America.

2) **Delimit the research problem**: the delimitation of the problem addresses to detail the action(s) that occur (e.g., an increase of the problem) and the actors involved (e.g., people, organizations, countries). It is also recommended to specify the place and time of the problem. A phrase should summarize the problem to study. For example, the level of happiness is high in the countries of Latin America between the years 2000 and 2017.

3) **Apply six strategies to interrogate the defined and delimited problem**. These strategies are addressed to formulate research questions:

    a. **Description**: if very little is known about the problem or there is not enough information, an exploratory or descriptive investigation should be carried out. In this case, the questions to the problem can be written starting with "to what extent" or "is there". The research will aim to find out if the problem exists or not, or to what extent occurs the research problem. For example:

    To what extent is there a high level of happiness across all the countries of Latin America between 2000 and 2017?

    The purely descriptive research questions are not usually suitable for a doctoral dissertation or a competitive research proposal since they only detail the problem but do not explain it, nor relate it to another phenomenon, nor compare it. The following strategies do allow formulating questions for causal, relational and comparative research.

    b. **Causes**: interrogate about one or several causes or explanations of the problem. If there are several possible causal conditions of the problem, the question can be written with "why" or "what are the factors" or "under what conditions" or "which are reasons that explain or influence". If we want to find out if a specific cause explains or influences the problem, then we should place the possible cause at the beginning of the research question and then the research problem. Some examples of research questions about the causes of the problem:

    Why is the level of happiness high in the countries of Latin America between 2000 and 2017?

    What factors explain the high level of happiness in the countries of Latin America between 2000 and 2017?

    Does the level of religiosity influence (or explain) the level of happiness in the countries of Latin America between 2000 and 2017?

If we aim to evaluate a policy or the introduction of a change, or it is already known that something is the cause of the problem but we want to know the process that connects the cause to the problem, then the question should be written using the "how". For example:

How has the policy of raising consumer taxes affected happiness levels in the countries of Latin America between 2000 and 2017?

How does collective social capital influence the level of happiness of the population in the countries of Latin America between 2000 and 2017?

c. **Consequences**: interrogate about the consequences of the problem in some area. It is advisable to place the problem at the beginning of the question and then add the phenomenon where it impacts, effects or have consequences. For example:

What effects does the level of happiness have on the development of the IQ of young people between 18 and 30 years of age in Latin American countries?

How does the high level of happiness affect public spending on healthcare in the countries of Latin America?

To what extent the level of happiness impact on labor productivity in the countries of Latin America?

d. **Solution**: think of a solution to the problem, or an intervention that affects the problem. Ask what would happen whether that solution or intervention is carried out, and how or to what extent would affect the problem. For example:

Will the change in working hours for schoolchildren reduce happiness levels in Latin American countries?

Happiness levels would be reduced in the countries of Latin America whether birth control policies were applied?

To what extent the implementation of policies for reducing income inequality affect happiness levels in Latin American countries?

These types of research questions are more common in studies that apply experimental methods, make projections, interventions, social actions, or are design-based.

e. **Comparison-place**: interrogate whether the problem happens elsewhere, or whether there are similarities and differences between places. These research questions are formulated for comparative studies.

What are the differences and similarities in the level of happiness among the working class in the countries of Latin America?

Are there different levels of happiness in the Latin American countries among the urban and rural population?

To these comparative questions can be added questions about the causes or consequences. For example:

Why some Latin American countries have higher levels of happiness than others?

f. **Comparison-time**: interrogate whether the problem arose before, or whether a problem happens nowadays. These questions are appropriate for longitudinal studies.

Has the high level of happiness in the countries of Latin America been stable in the last hundred years?

How have happiness levels changed (or evolved) in Latin American countries after the 2008 financial crisis?

It is recommended to formulate all possible research questions by applying the six strategies, always taking into account the three conditions that research questions must meet: concise, achievable, relevant. Once the various research questions have been written, submit them to validation through discussion with experts and review if they have already been widely studied. If a research questions have been excessively examined, and the answers are already known and do not imply any innovation, it is better to discard them and continue looking for others. Of all the research questions formulated, the researcher could select one or several, as long as they are interconnected. That is, research could be addressed to investigate the causes, consequences of a problem, and also make comparisons between places and in time.

Once the research question that we want to answer has been identified, we must add the analysis units, which are the actors that will be analyzed (for example, young people between 18-30 years old, or countries), and the geographical and temporal space that will be analyzed (for example, in the countries of Latin America in 2017, in the countries of Latin America in the last hundred years, in the countries of Latin America after the financial crisis of 2008). Therefore, a final research question would be:

How does the high level of happiness of Latin American countries affect labor productivity between 2000 and 2017?

Why is the level of happiness of young people between 18 and 30 years olds residing in Colombia higher than that of those who reside in Peru between 2000 and 2017?

The main mistakes in the formulation of research questions come from skipping the steps. We must follow each of the steps in order. The research question is the first and most important achievement to be successful in writing of a research project, disseration or thesis. Without a research question, there is no research project or study ready to start. Writing the research question in infinitive, we can get the research objective, for example:

Objective:
Find out the causes that explain why the level of happiness of young people between 18 and 30 years olds residing in Colombia is higher than that of those who reside in Peru between 2000 and 2017

Hypothesis

The tentative answers to the research question are the hypotheses. They are tentative because they are not based on data but on conjectures made from the literature review and observation of reality. The hypotheses are what we want to test or refute with data collection and data analysis.

For example, if the research question is:

Why is the level of happiness of young people between 18 and 30 years of age residing in Colombia higher than that of those who reside in Peru between 2000 and 2017?

Possible hypotheses based on previous knowledge of the subject, after reviewing literature, and after some observations are:

- Hypothesis 1: *The higher level of literacy among young Colombians between 18 and 30 years of age explains the higher level of happiness concerning young Peruvians*

- Hypothesis 2: *The participation of young people in non-governmental organizations explains that in Colombia people are happier than in Peru*

The research question and the hypotheses will help us identify the indicators that we have to measure, which is the aim of the next chapter.

Exercise:

Apply the 3-step protocol to formulate research questions from the research problem defined and delimited in the previous chapter. Write as many research questions as you can. Then, assess the research questions based on compliance with the three indispensable conditions. Of all the research questions, present one in class to discuss with your colleagues.

4. How to identify indicators from the research question and hypothesis. The process of operationalization

The great challenge in a research study is to identify how to measure the concepts of our research questions and hypotheses. How is school performance measured? How is the economic development of a country measured? How is sustainable development in education measured at a country level? How is income inequality among the inhabitants of a country measured? Abstract concepts such as school performance, economic development, sustainable development or income inequality cannot be observed directly, and therefore must be measured by indicators. An indicator (also known as measure) is a measurable variable used as a representation of an associated concept. School performance at a student level is measured by indicators such as grade point average in one year, ratio of class attendance, and number of awards obtained. Economic development is measured by indicators such as Gross Domestic Product (GDP) and GDP per capita. Sustainable development in education at a country level can be measured by indicators such as literacy rate in the country, number of university students per inhabitant, and number of teachers per enrolled students. Income inequality of a country is measured by indicators such as the GINI index and the 90/10 ratio. In more quantitative research, indicators are usually aimed at obtaining numerical data. Qualitative research also uses indicators, although they are more aimed at obtaining non-numerical data.

How we measure a concept of interest is key because it is how we observe the phenomenon, what we look at, what we will look for in the field and how we value it. Therefore, the process of identifying indicators is one of the essential parts of the research process. The transition from abstract concepts to measurable indicators is known as operationalization process.

The following protocol helps to move from the research question to the indicators.

Protocol to identify indicators:

1. Formulate hypothesis

A hypothesis is a tentative answer to the research question, and is based on literature review and direct observation of reality. Writing the hypothesis facilitates the identification of the main concept(s) that we have to measure. A hypothesis is usually composed of 1 or 2 concepts. When the hypothesis is descriptive (or comparative) there is only one concept in the sentence. When the hypothesis is relational (causal or associative), there are two concepts in the sentence. For example:

- In the following hypothesis, there is only one concept "quality of life": *Barcelona has a higher quality of life than Madrid*. In this example, Barcelona and Madrid are the cases of analysis, but not concepts to be measured.
- In the following hypothesis, there are two concepts "quality of life" and "drug consumption": *The increase in the quality of life is associated to an increase in*

drug use in the young population. In this example, young people are the cases of analysis, but not the concept to be measured.

2. Define the concepts

The concepts must be defined using previous studies and theories. Therefore, a literature review is necessary at this stage. The definition must always be explicit. It is not convenient that we define vaguely concepts. The researcher is responsible for deciding which definitions from previous studies and theories take, and whether complement or modify the definitions. A good strategy is to present several definitions and use all of them to identify several indicators.

For example, several authors have defined "quality of life" as follows:

- *Quality of life is defined as personal well-being derived from satisfaction or dissatisfaction with areas that are important to him or her* (Ferrans, 1990).
- *Quality of life is the multidimensional evaluation, according to intrapersonal and socio-normative criteria, of the personal and environmental system of an individual* (Lawton, 2001).
- *Quality of life is equivalent to the sum of the scores of life conditions objectively measurable in a person, such as physical health, living conditions, social relations, functional activities or occupation* (Urzúa M and Caqueo-Urízar, 2012).

Which of these definitions is used in the research is a decision of the researcher, and is conditioned by the objective of the research and by what can be measured. That is, there must be an agreement on how the concept is defined and how it will be measured.

3. Break the concept into dimensions

There are very general and abstract concepts, such as "quality of life", that need to be broken or divided into dimensions, which are the various parts that make it up. For instance, the concept "quality of life" can be broken or divided in the following dimensions according to studies in the European Union by Eurostat (goo.gl/fNsZJP)

Dimensions of quality of life:

- *Material living conditions (income, consumption and material conditions)*
- *Productive or main activity*
- *Health*
- *Education*
- *Leisure and social interactions*
- *Economic and physical safety*
- *Governance and basic rights*
- *Natural and living environment*
- *Overall experience of life*

The number of dimensions identified in each concept depends on the researcher's interests and definitions used. The more dimensions are identified, the more complete the research is, as it covers more areas, but the more indicators will be necessary.

The process of breaking the concept into dimensions is done taking into account:

- Previous studies and theories
- The research question and hypothesis
- The identified dimensions must be independent of each other

4. Identify indicators for each dimension of the concept

Each dimension must be measured by at least one indicator. It is recommended to use several indicators simultaneously for each dimension to be more accurate. Indicators are measures that serve as clues to observe reality; the more clues, the more precise the measurement is. For example, how do we measure the dimension "leisure and social interactions"? Possible indicators would be:

- *Frequency with which people engage in leisure activities (compared to spending time at home for example), such as going to the cinema, attending live performances, visiting cultural sites and attending live sports events*
- *Activities with people, that is, getting together with relatives and friends*
- *Activities for people, that is, one's involvement in voluntary and charitable activities beyond one's work*
- *Supportive relationships, shown by one's ability to get help and personal support in case of need*

The same process must be done for each of the dimensions of the concept "quality of life". The set of indicators identified are the ways in which we measure and observe "quality of life". The indicators will be used to compare the level of quality of life and to test or refute the hypotheses.

It is important when identifying indicators:

- The indicators are features of the reality and must be concise, specific, timely, referred to the concept, comparable, available, observable, measurable and quantifiable.
- The population of study: measuring the quality of life in cities is not the same as in companies. The indicators may be similar but measures have to be adapted. For example, "quality of life in leisure" can be measured in cities by indicators such as "number of cinemas and theaters per number of inhabitants in the city". In contrast, the "quality of life in leisure" at an organizational level can be measured in companies by the indicator "availability of a space for recreational activities such as games, reading or talks within the company".
- Indicators will be transformed into survey questions, queries for an in-depth interview or guide of observations for carrying out ethnographies. For example, the indicator "availability of a space for recreational activities such as games, reading or talks within the company" would be transformed into a survey question:

In your current workplace, is there any space reserved for recreational activities such as games, reading or talks? 1) Yes, 2) No, 3) DK / NA.

When writing a final report or article, a table showing this process of operationalization should be included. For example:

Table 2. Indicators to measure quality of life

TABLE 1. Domains and core indicators of quality of life.

Domains	*Indicators*	*Item examples*
Emotional well-being (EW)	Mental stability; satisfaction, self-concept; lack of stress/negative feelings	He/she shows symptoms of depression.
Interpersonal relations (IR)	Social relationships; family relationships; to have stable and clearly identifies friends; to have positive and gratifying social contacts	He/she complains about his/her relationships with friends.
Material well-being (MW)	Housing conditions; workplace conditions; Services conditions; employment; incomes/salary; possessions	His/her incomes are not enough to afford whims.
Personal development (PD)	Education; learning opportunities; work abilities; functional abilities (personal competency; adaptive behavior, etc.); activities of daily living.	He/she is involved in the development of his/her individual planning.
Physical wellbeing (PW)	Health care; sleep; health consequences (sorrow, medication, etc.); health; mobility; technical assistance	He/she has sleep problems.
Self-determination (SD)	Autonomy; goals and personal preferences; decisions; choices	Other people decide how to spend his/her money.
Social inclusion (SI)	Participation; integration; supports	His/her family supports him/her.
Rights (RI)	Knowledge of rights; defense of rights; exercise of rights; privacy; respect	He/she suffers exploitation, violence or abuse.

Source: Verdugo et al. (2010) "Development of an objective instrument to assess quality of life in social services: Reliability and validity in Spain". *International Journal of Clinical and Health Psychology* 10(1): 105-123.

The success in the search for indicators lies in reading, reading and reading literature related to the concept. It is really helpful reviewing how other studies have measured the concept of interest. The high number of studies is our best tool for identifying indicators.

The process of operationalization is possibly one of the most complicated phases of a research process and, at the same time, the most relevant since everything depends on how our phenomena of interest and concepts are measured. Everything can be measured, and by measuring phenomena, we can analyze them, criticize them and intervene in them.

Once the indicators have been identified, the next step will be either to design an instrument of data collection (for example a survey) or to use secondary sources (for example, online databases) to collect data that refer to the indicators.

Exercise

Identify indicators based on the research question you made in the previous chapter. Create a table like Table 2, which presents the dimensions and indicators of one of the main concepts of your research question.

5. How and where to collect data from secondary sources

In the research process, two types of data sources can be distinguished: primary and secondary.

- Primary sources or primary data are those collected or produced by the researcher or group conducting the research. Whether we conduct an interview or a survey and we collect the information, we are using primary sources.
- Secondary sources or secondary data are those collected or produced by people or institutions that are not directly involved in the present research. When we use statistical data from the World Bank, when we use documents from archives or when we review bibliographic information, we are using secondary sources.

This chapter explains how and where to collect data from secondary sources.

What information or data should we collect?

The data that we search and collect have to refer to the indicators. We have to collect information or data on the indicators that we have identified in the previous phase. If we collect data for the simple fact of collecting, it can lead us to an eternal and aimless work. For that reason, it is convenient to have a guide. This guide is the set of indicators identified in the previous phase. If during data collection we find other useful data, obviously we can collect them, but it is always convenient to search for data about the indicators that measure the concepts of our research questions.

Which data from secondary sources is more reliable?

In the era of fake news and post-truth, it is not easy to rely on data from secondary sources. However, three significant criteria are used to identify which data from secondary source can be reliable:

a) The number of times a secondary source has been used by studies published in scientific journals. The more a secondary source is used in scientific papers, the more reliable is the source.
b) Data published in high-impact scientific journals. Several experts review if data used in an article is reliable. The better the scientific journal, the stricter review. Thus, data published in the top scientific journals have more reliability.
c) Status and reputation of the institutions that collected the data. The better the reputation of an institution, the more reliable the data from this source. In the following section, some high-status secondary sources are presented, which are reliable to develop your research.

Where to collect data from secondary sources?

There are hundreds or thousands of institutions that regularly collect data on various topics. The Internet is an immensely rich, heterogeneous and cost-effective information media. Therefore, it is important to know which are the main secondary sources in our

subject or field of study. The following list presents several secondary sources to obtain empirical data.

International institutions - These organizations make available data where the units of analysis are usually countries or regions. They are very useful to carry out research that aims to compare countries or study the evolution over time of a phenomenon. Some of the main international institutions with accessible databases are:

- World Bank - possibly the largest source of data by countries on political, economic, social, health and environmental indicators. Free access and availability of historical data. https://data.worldbank.org/
- CIA Factbook - data compiled by the Central Intelligence Agency of the United States. It offers guides for each country where both quantitative and qualitative data on various topics are presented. Free access. https://www.cia.gov/library/publications/the-world-factbook/
- ECLAC - United Nations Regional Economic Commission for Latin America. It contains data on various topics, especially economic and commercial, and data based on censuses and surveys conducted at the national and local levels. They also make reports where the compiled data is synthesized and analyzed. Free access. http://www.cepal.org/en/datos-y-estadisticas
- UNESCO Institute for Statistics (UIS) – a source of data that allows comparison between countries and longitudinal studies on education, science and technology, culture and communication. Free access. http://www.uis.unesco.org/Pages/default.aspx
- Web of Science of Thomson Reuters – a source of data on publications of articles and scientific books. Some university libraries have access to this bibliometric source since it requires a subscription. http://isiknowledge.com/wos A free access source of data on scientific journals is Scimago: http://www.scimagojr.com/

International surveys - Some research groups collect data through an identical survey in several countries. These quantitative studies allow comparative research between countries and longitudinal analyses, as well as analyze the opinions of the population of a specific country.

- World Values Survey - contains opinions on values, beliefs and social behaviors. Online analysis and maps can be executed on the website and availability to download the data in Excel, SPSS, R, Stata, SAS and ASCII. http://www.worldvaluessurvey.org/WVSContents.jsp
- European Social Survey – a study of opinions on various social issues developed in 36 European countries. http://www.europeansocialsurvey.org/data/
- Latinobarómetro - a public opinion study carried out in 18 Latin American countries. Availability of online analysis and complete data to download. http://www.latinobarometro.org/latOnline.jsp

- Latin American Public Opinion Project (LAPOP) - surveys conducted in various Latin American countries on the opinion, attitudes and behavior of the population in various political, social and economic topics.
  http://www.vanderbilt.edu/lapop/

National surveys - Institutions, usually public, that collect information about the population of a country. They are useful when research is focused on a specific country or city.

- Statistical Institute - usually include information obtained by the census and surveys of various subjects. For Germany:
  https://www.destatis.de/EN/Homepage.html
- Market commission - includes data on large companies. For Germany:
  https://www.bafin.de/

There are thousands secondary reliable sources. One option to identify secondary sources of data on a specific topic is to ask experts in the field and to review the data sources of previous studies, usually mentioned in the methods section.

Several secondary sources are usually employed in the same research study. For instance, to measure inequality, the GINI index is used, obtained from the World Bank. This data indicates that income inequality in 2014 is 44.1 in Peru and 53.5 in Colombia. Therefore there is higher income inequality in Colombia than in Peru. To collect information about opinions regarding inequality, the Latinobarómetro can be used, specifically, the question referred to "What is the main problem in the country?" In 2015, only 0.2% of the respondents in Peru and 2% in Colombia indicated that inequality is the country's main problem.

All secondary sources used in the research must be cited in the bibliography.

What are the advantages and disadvantages of using secondary sources instead of primary sources?

The main advantages are:

- Access to more data in less time and spending fewer resources
- Allows comparisons in time (longitudinal) and between several cases (for instance, countries)

The main disadvantages are:

- The researcher loses control over the data collection
- The data may not make explicit reference to the indicator identified and can be inappropriate to answer the research question
- Some secondary sources of information are unreliable

Of all these advantages and disadvantages, the one that influences the most is the availability of economic resources to carry out the research. Whether financial resources are limited, secondary data are employed, especially to develop macro-level analysis.

In the next chapter we will focus on how to design by ourselves an instrument for gathering information, specifically a survey, and thus, collecting primary data.

Exercise

From the indicators identified in the previous chapter, collect data about them based on information from secondary sources. They can be official databases, online data, books, files or any type of secondary source.

6. How to design a survey

We employ a survey when we want to collect empirical data that are not available. A survey is a research technique to collect quantitative data through a questionnaire and a sample of actors. A questionnaire refers to the set of questions and response categories.

You can use an existing questionnaire (with minor adaptations) or design a new one. The most advisable is to use existing questionnaires and adapt them minimally to the place, time and population to which it is addressed since the cost in time and resources is higher when designing it from scratch.

If we choose to design a questionnaire, these are the main steps to follow.

From the indicators to the questionnaire:

1. The indicators
The first step in the creation of a questionnaire is to have a set of indicators. If we want to measure "quality of life" and some of its indicators are "access to public gas", "family income" and "access to daycare centers"; the survey questions will be aimed at obtaining information about these indicators. Without indicators, we cannot begin to design a questionnaire.

2. Writing questions
For each indicator, at least one question is necessary. For example, to measure "access to public gas", the survey question would be:

Does the dwelling in which you live have access to public gas?

You can also include the indicator within a question with several options.

Of the following services to which do you have access in your home? Select as many options you want.
1. Electric light
2. Potable water
3. Gas
4. Solar energy
5. Garbage collection

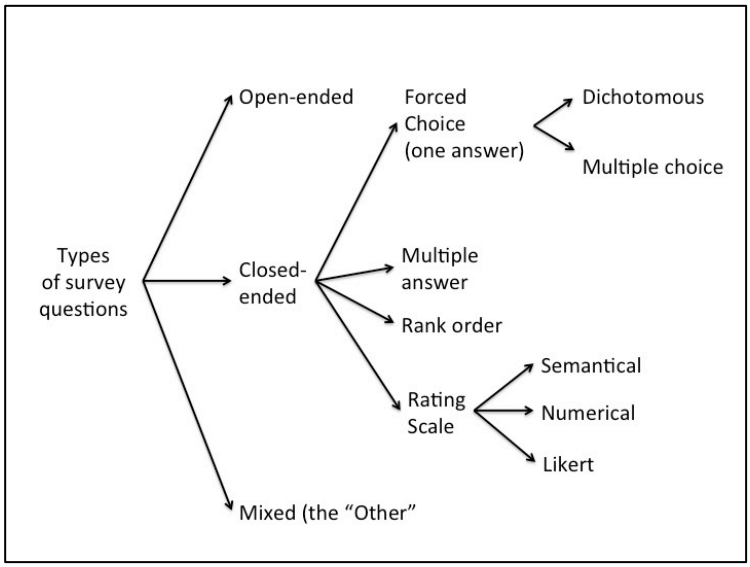
The best learning strategy to design survey questions is to review other questionnaires that have been applied and have high legitimacy, and thus formulate questions in a similar way. It is recommended in the field of Sociology, for instance, to review the World Values Survey, European Social Survey and Latinobarómetro.

The most common mistakes when writing survey questions and that we should avoid are:

- Questions that combine 2 or more topics in the same statement. An example of a badly worded question would be: *How satisfied are you with the government's educational and economic policy?*
- Questions that require high memory effort of the respondent. For example: *How many times have you given tips in the last two years?*
- Questions that assess the knowledge of respondents abruptly and lead the respondent to feel uncomfortable. For example: *Do you know what the functions of the Prime Minister are?*
- Too general questions. For example: *Do you like the President?*
- Questions written in negative. For example: *Have you not visited your doctor in the last month?*
- Questions that affect the sensitivity of interviewees. For example: *Are you racist?*
- Questions that contain some of the options included in the answers. For example: *What means of transportation such as bus or taxi do you use to go to work?*
- Abstract or polysemic words that can be interpreted in different ways. For example: *Do you think the country has developed in the last year?*
- Ambiguous words and phrases. For example: *How much have you progressed in mathematics?*
- Implicit judgments. For example: *As you already know...*
- Value judgments: For example: *Although religion has become a controversial issue in our country, how often do you attend religious services?*

There are several types of survey questions (Figure 4).

Figure 4. Types of survey questions



Source: own elaboration

Likert scale survey questions are those in which respondents have to select a grade on a scale that ranges from one extreme to another, such as from "strongly agree" to "strongly disagree". This type of questions was originally invented by psychologist Remis Likert,

who in 1932 wanted to measure attitudes avoiding the dichotomous questions of Yes or No. An example of Likert scale question:

To what extent do you agree or disagree with the policy of extradition of drug traffickers to the United States?
1- Totally agree
2- Agree
3- Neither agree nor disagree
4- Disagree
5- Completely disagree
0- No response / Do not know (NR / DK)

The response categories (or choices) may include a scale from 1 to 5, from 1 to 10, or from 1 to 4 as decided by the researcher. The longer the scale, the more nuanced responses. If the total of categories in the scale is even number, such as 1 to 4, the intermediate category is eliminated (*neither agree nor disagree*) and it obliges the respondent to make a decision for or against something.

In summary, to ask questions of opinion, conformity or agreement, the use of Likert scale questions is highly recommended.

3. Writing response categories

The response categories or choices accompany each closed question and must fulfill two characteristics: categories must be exhaustive and mutually exclusive.

Exhaustive: the full range of possible answers must be listed for the respondent. The response categories of the following question are not exhaustive.

Regardless of where you access the Internet, what do you use it for?
1. To use email / messenger
2. To search for information
3. To entertain
4. To work
5. To do paperwork
0. NR / DK

In this question, some items or categories, such as “to make purchases” or “to study” are missing. Since there are multiple and very diverse uses of the Internet, it is convenient to place the “Other” category. It is recommended to place the category “Other” in a survey question when response options are almost infinite, thus we comply with the criteria of exhaustiveness. The correct way to write the response categories would be:

Regardless of where you access the Internet, what do you use it for?
1. To use e-mail / messenger
2. To search for information
3. To entertain
4. To work
5. To study
6. To do paperwork
7. To make purchases
8. Other
0. NR / DK

Mutually exclusive: response categories cannot overlap since the respondent must select one, and only one of the choices. The response categories of the following question do not meet the criterion of mutual exclusivity.

How often do you talk about politics with your friends?
1. Very frequently
2. Frequently
3. Pretty much
4. Almost never
5. Never
6. I do not talk about politics
0. NR / DK

The categories *2-Frequently* and *3-Pretty much* are very similar and overlap, so it is convenient to discard one of them. Also, if you never talk about politics with your friends, you could select the categories *5-Never* and *6-I do not talk about politics*. Therefore, it is also convenient to eliminate one of them. The question and its categories are correctly written as follows:

How often do you talk about politics with your friends?
1. Very frequently
2. Frequently
3. Almost never
4. Never
5. NR / DK

4. Sections of a questionnaire

It is recommended to sort the questions by thematic blocks, and establish sections following this order:

Presentation - every questionnaire must contain a welcome message where the survey is presented: research objective, subsequent use of the data and offer an email or telephone number in case the respondent wants to contact the person responsible for the study. Also, if you do not ask for the name of the respondent, which is advisable, in this welcome message it must be made explicit that anonymity of respondent is guaranteed. This presentation should be very short and motivate respondents to answer the questions.

Initial questions – place important questions at the beginning to ensure that the respondent answer at the time that has the most attention.

Critical questions - those more problematic or invasive.

Sociodemographic questions - are those questions about the social and demographic profile of respondent: gender, year of birth,[1] place of residence, educational level, income level and occupation, among others. The researcher decides which may be relevant for their study, although it is recommended to include all of them, as they are explanatory of the responses or opinions collected.

It is recommended that the questionnaire is not excessively long in order to avoid the respondent's lack of attention or survey abandonment, especially if the questionnaire is distributed online.

5. Coding

All questions and response categories must be associated with a code. Numbered letters should be used for survey questions and numbers for response categories.

6. Validation of the survey

Any new survey that we design must be validated. The validation process of a survey consists of two phases: external and internal.

External validation is the review by experts on the subject of the survey and on research methods to assess if questions are related to indicators, if questions are correctly formulated, and if the questionnaire shows any errors or possible improvements.

Internal validation is a pilot test of the survey with a small group of potential respondents. The questionnaire is applied to a small part of the sample to know if they understand all

[1] To ask about age of the respondent it is better to question about year of birth rather than directly how old the person is, it increases the response rate since many people do not like to confess their age and less to a stranger.

the questions, if they are represented in response categories, if the time reference is clear, and if they detect any error or possible improvement.

The study population and the sample to which we will apply the survey is an essential aspect of writing the questions. What procedure and sample size should be used, it is the purpose of next chapter.

Exercise 1

From the indicators identified in the previous phase, formulate survey questions to collect information about each of indicators. If your indicators refer to measures at country level (e.g., GDP per capita), modify the indicator to measure the same concept but at an individual or organizational level (e.g., income per year). Consider all of the aspects outlined in this lesson before writing your survey questions.

Exercise 2

Review the World Values Survey and select those questions that would help you in your questionnaire, for example, sociodemographic questions. You can copy or adapt them if you think convenient.

Exercise 3

From the questions written in exercise 1 and those adapted or copied in exercise 2, build the questionnaire in Google Drive, in the Forms option. There are several tutorials on Youtube on how to build a questionnaire in Google Drive. Once finished, distribute it among your instructor to obtain external validation, and to at least 5 of your contacts to perform an internal validation pilot test.

7. How to select and calculate samples from a population

What is a sample and what is a population or universe

The universe or study population is the set of cases or actors (people, organizations, countries) that share some characteristic(s) and that are those units of analysis with which the research question will be answered (or the hypotheses will be tested). A sample is a limited number of actors (or cases) taken from that population. We study samples due to the impossibility of studying the entire population or universe.

The main objective of sampling is the possibility of generalizing, i.e., drawing general conclusions based on the study of a few cases. How a representative sample is obtained, how the sample cases are selected, and what size a sample should have is important since bad sampling leads to bad conclusions, and therefore to worthless research.

There are three issues that we must ask ourselves when sampling from a population: 1) representativeness of the sample, 2) sampling types, 3) sample size.

1. Representativeness of the sample. Do we want a representative sample?

If we want to generalize the results of our study, that is, to point out that what was studied with a sample can be attributed to the whole population, we need representative samples. We get a representative sample when the main characteristics of the population are present in proportion in the sample. It is the researcher, based on the research question and hypothesis, the one who decides which variables or characteristics are taken into account to carry out the sampling. For example, if we are conducting an election poll in Spain and we consider that voting behavior depends extremely on gender then, we should select a sample in which men and women are represented in proportion.

The proportion can be extracted equally from the number of categories of the variable, or in a proportion equivalent to what extent different categories are in the population. For example, if we consider that the variable gender is important in our research and its two categories are male and female, the sample should contain both men and women. If there are 800 men and 200 women in the universe or population, following an **equitable criterion**, the sample should consist of 50% men and 50% women. On the other hand, following an **equivalent criterion**, the sample should contain 80% of men and 20% of women. The final decision –whether the sample consists of 50% men and 50% women (equitable criterion), or 80% men and 20% women (equivalent criterion), will be made by the researcher based on the research objective and the hypotheses to be tested. Experimental studies use more the equitable criterion. Election polls employ more the equivalent criterion.

2. Sampling types. What kind of procedure will we follow to select the cases (or actors) of a sample?

There are two major sampling strategies or methods: non-random sampling and random sampling. They are also commonly referred to as non-probabilistic and probabilistic sampling.

a. Non-random sampling: the researcher deliberately chooses cases and not chance. There are several subtypes:
   i. For convenience or accident: the researcher chooses those cases (or actors) who are around and available. For example surveying pedestrians on the street.
   ii. By quotas: the researcher chooses the cases taking into account that they are from the various categories of a variable. For example: survey a group of men and a group of women proportionally, that is, with quotas of both categories of the gender variable.
   iii. Intentional or judgment: the researcher selects the cases following the own criteria or that of experts. For example, survey key actors following the judgment of an expert on the subject.
   iv. Snowball: the researcher begins with a small group of actors and expands the sample by asking those initial participants to identify others that should participate in the study. Useful for investigating clandestine populations or those difficult to access by the researcher.
   v. Volunteering: the subjects or actors are those who come to be studied after a call of the researcher. Used in experimental investigations where candidates are voluntarily presented for research.

b. Random sampling: actors are chosen randomly and entirely by chance. All actors in the population should have the same chances of being elected. There are several subtypes or procedures to select a random sample:
   i. Simple random: a “raffle” is made between all the actors of the population. The list of the entire population is needed. All the actors are listed and chosen randomly.
   ii. Systematic randomization: having the enumerated list of the population, "jumps" are given in the list following a fixed interval to select the actors that will compose the sample. An example step-by-step: a) from a population of 10,000 actors we have to select a sample of 200; b) we divide the size of the population by the size of the sample and thus obtain the interval, e.g., 10,000 / 200 = 50, i.e., we will give jumps in 50 on the list of the population; c) we choose an actor from the list at random and from it we count the number of the interval (50) to choose the next one, and from that, counting 50 more we get the next one, thus until reaching the size of the sample. Useful in telephone surveys using the directory.
   iii. Random stratified: the list of the population is also needed and a “raffle” of the actors is carried out, but selection arises from categories of a relevant variable in

the research. For example: randomly select men and women in a proportion equivalent to the population.

iv. Cluster random sampling: in this case, a "raffle" is made of the places, organizations or clusters to which actors belong. Then, actors that belong to these organizations are chosen. For example, we have to survey the students of a country but there is not a centralized registry of all the students of the country. Because a list of universities can be obtained, we use these clusters (organizations) to execute the "raffle". Universities are selected at random and students of the chosen universities are interviewed. If later subdivisions of these clusters or organizations are drawn, it is a multi-stage cluster sampling (several stages). For example, following the previous example, as the number of students at universities is high, in a second stage, faculties of the selected universities are drawn. From the chosen faculties, classrooms are selected at random in this third stage and there the students are surveyed.

Different types of sampling are usually combined. The most used in macro sociological research is the stratified multistage cluster random sampling, although all other methods are equally valid and legitimate. It is key to justify why one procedure or another is selected. It is recommended to review other empirical studies to check how they justify the use of a sampling strategy.

3. Sample size. How large should the sample be?

If we want the sample to be representative, we must follow a calculation process of the size taking into account 4 factors:

a) Size of the population: the larger a population, the larger the sample size. Although, the sample size does not change much for populations larger than 20,000.
b) Heterogeneity (or response distribution): refers to the level of dispersion of the population in some relevant variable in the research. The more heterogeneous a population is, the larger the sample size should be. The more homogeneous a population, the smaller the sample size. The maximum heterogeneity is 50% (also expressed as 0.5). If it is unknown the level of heterogeneity or there is no a variable to measure the degree of diversity, it is typically used 50%.
c) Sampling error (or standard error or margin of error): refers to how much the results of the sample vary with respect to the universe or population. For example, if the average number of times a sample goes to the cinema per month is 2.5, and the average number of times the total population goes to the cinema per month is 1, then the sampling error is the difference between 2.5 and 1. The researcher chooses the sampling error before carrying out the study, and standard sampling errors are used. In sociological research, it is commonly used a sampling error of 3% or 5%. The smaller the sampling error, the larger the sample size, since to be more precise (less error) we have to study more actors of the population (larger sample size).

d) Level of confidence: is the probability that the result obtained is within the confidence interval. Confidence levels of 95% are often used in sociological research. It is a bit complex to understand but with an example it is clearer. We study a sample using a confidence level of 95% and we obtain that the average consumption of cigarettes per day is 5.5 and the standard deviation is 1.5. The 95% confidence interval is expressed as the mean +/- 1.96 standard deviations, that is, the interval will go between approximately 8.5 and 2.5. That is, there is a 95% probability that the average number of cigarettes consumed per day in the entire population is between 8.5 and 2.5. If we use higher confidence levels like 99% to get more accuracy (higher exactitude), the sample size will be larger.

The formula for calculating the sample size is:

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

n: sample size
N: size of the population
Z: confidence level
e: sampling error
p: heterogeneity

If the size of the population is very large, more than 100,000 people, the formula can be simplified as follows:

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

Nowadays there are online calculators to find out the size of the suitable sample taking into account these factors. It is easy to use them and allows you to play with different levels to know the most suitable sample size. It even allows us to calculate the sampling error if we already have chosen or defined the sample size.
http://www.raosoft.com/samplesize.html

Exercises
From your research question and study population, answer the following questions:

- Would you choose to analyze a representative or non-representative sample? What advantages would it have in one case and another?
- Which sampling type do you think is the most appropriate? Why? Consider the possible access to the target population, available resources, complete knowledge of the population and other factors to justify your response
- Calculate the sample size on the basis of the conditions of your study.

8. Data analysis: types of statistical analysis

Statistics is a branch of mathematics employed in research to analyze quantitative data. But before to start with statistics, a short glossary

Glossary

- Cases: they are the actors or subjects of analysis. Cases can be individuals, objects, organizations, and countries. Examples of cases: people who have answered a questionnaire, countries that had a civil war.

- Variables: they are the characteristics, qualities or attributes of cases. A variable is a characteristic or feature of the actors (cases) that is liable to vary or be capable of varying in value. Variables derive from each survey question or indicator. Examples of variables: gender, educational level, height, weight, economic benefits at the end of the year, geographic extent, GDP, GINI index, number of protests.

- Response categories or values: they are response options, choices or values of a variable. Example: male, female (for the variable gender); no formal education, primary, secondary and university studies (for the variable educational level); 180 cm, 181 cm (for the variable height); 83 kg, 123 kg (for the variable weight); 1,000 euros (for the variable economic benefits at the end of the year); 600 km (for the variable geographical extension); 0.23 and 0.53 (for the variable GINI index variable); 0, 1, 2, 3 ... (for the variable number of protests).

Now, let's start with statistical data analysis:

The analysis of the data depends on two major factors:
A. Types of variables
B. Number of variables simultaneously analyzed

A. Types of variables

There are 3 types of variables according to response categories or measurements:

- Nominal: are those variables whose response categories do not have a pre-established order or internal hierarchy. Examples:
  - Variable *gender*: 1- Male, 2-Female
  - Variable *marital status*: 1-Single, 2-Married, 3-Live as a couple, 4-Divorced, 5-Widowed, 6-Other
  - Variable *belonging to a sport club*: 1-Yes, 2-No

- Ordinals: variables whose response categories have a pre-established order or internal hierarchy. Examples:
  - Variable *educational level*: 1-No formal education, 2-Primary, 3-Secondary, 4-University
  - Variable *degree of agreement with tax reform*: 1-Strongly agree, 2-Agree, 3-Neither agree nor disagree, 4-Disagree, 5-Strongly disagree
  - Variable *age group*: 1-Under 18 years, 2-Between 18 and 35 years old, 3-Between 36 and 50 years old, 4-Between 51 and 65 years old, 5-Over 65 years old
- Scale: variables whose response categories have a pre-established order or internal hierarchy, and the gap between one category and another is the same.
  - Variable *year of birth* (open question):
  - Variable *number of times you have attended cinema during last month* (open question):

B. Number of variables simultaneously analyzed

Depending on the number of variables simultaneously analyzed, there are 3 types of analysis:

- Univariate descriptive analysis: a single variable is analyzed and the purpose is to describe central values and distribution of responses.
- Bivariate analysis: two variables are simultaneously analyzed and the purpose is to test relational hypotheses (causal or associative), i.e., find relationships between two variables.
- Multivariate analysis: more than two variables are analyzed and the purpose is to test relational hypotheses (causal or associative), i.e., find relationships between variables.

Statistical techniques

The main statistical analysis techniques for each type of analysis are:

Type of statistical analysis	Statistical techniques
Univariate analysis	▪ Frequency tables ▪ Mean, standard deviation
Bivariate analysis	▪ Contingency table (cross-tabulation) ▪ Bivariate correlations ▪ One-way ANOVA
Multivariate analysis	▪ Multiple linear regression ▪ Binary and multiple logistic regression ▪ Logit and probit models ▪ Discriminant analysis ▪ Factor analysis ▪ Cluster analysis

	▪ Multidimensional scaling ▪ Two-way ANOVA ▪ MANOVA ▪ Structural equations

Source: own elaboration

The following chapters will explain how to use statistical techniques to analyze quantitative data. Statistics allows us to describe results and make inferences. Descriptive statistics synthesize and visualize results. Inferential statistics draw conclusions and identify relationships from data collected.

9. Univariate descriptive analysis

The statistics used to analyze a single variable in order to describe the data collected depends on the type of variable: nominal, ordinal or scalar.

Variable	Frequency table	Central tendency	Dispersion	Graph
Nominal	Yes	*Mode**	-----	Pie or bar charts
Ordinal	Yes	*Median**	*Range**	Bar chart
Scale **	No	Mean	Standard deviation	Histogram

* They have a limited use, so we do not present them in this handbook.
** Ordinals of more than 5 categories can be treated as scale variables.

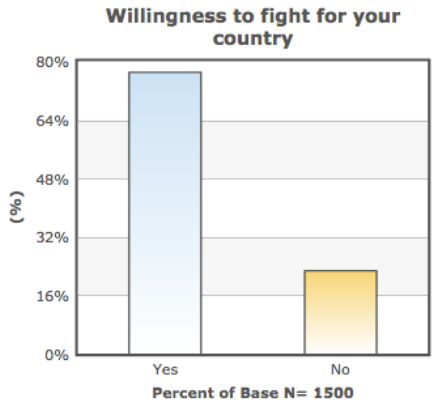
Frequency table
The frequency table is an analysis where the responses of the collected data are expressed in the number of times they occur (absolute frequency) and the percentage represented by these responses (relative frequency). The relative frequency (or percentage of responses) is calculated by dividing the number of responses in a category by the total number of cases.

Example of univariate descriptive analysis using frequency table and bar chart

Variable: "Of course, we all hope that there will not be another war, but if it were to come to that, would you be willing to fight for your country?"
Cases: people living in Kazakhstan (2011)

	Absolute frequency (number of cases)	Relative frequency (percentage)
Yes	1,158	77.2 %
No	342	22.8 %
Total	1,500	100.0 %

Source: World Values Survey



Frequency tables and charts should be accompanied by a text to report the main findings, that is, the main values or response categories are highlighted and an interpretation is made. Example:

> *The majority of respondents in Kazakhstan (77.2%) are willing to fight for the country. This result suggests a high national sense of belonging, which can be explained by the compulsory enlistment of men in military services (conscription).*

Mean (average)
The mean (also called popularly average or in statistical language, arithmetic mean) is the value that represents and synthesizes a set of data. It is a measure of the central tendency of a variable. It is calculated from the sum of the values of the answers divided by the total number of cases.

Standard deviation
The standard deviation is the value that measures the dispersion of values or responses of a variable. It is the average of the distances of the values of each case with respect to the average. That is, an average is calculated of how far the responses of each case are away from the average. The higher the standard deviation, the more dispersed the answers obtained, that is, the more heterogeneous is the opinion or behavior of the cases analyzed.

The standard deviation squared is the variance. Both the standard deviation and the variance are widely used in multivariate data analysis.

Histogram
The histogram is the graph used to visualize the results of a scale variable, and is characterized by including the normality curve. The normal curve, also known as the Gaussian bell, is a representation of the distribution of the data.

Example of univariate descriptive analysis using the mean and standard deviation:

Variable: "On a scale from 1 to 10, where 1 means your country is not democratic at all and the 10 means your country is totally democratic. Where would you position your country?"

This question was asked in three different countries. Because it is an ordinal variable of more than 5 categories, it is treated as a scale variable and the mean and standard deviation are calculated.

Results:

Table. How democratic your country is

Country	Mean (average)	Standard deviation (SD)
Peru	5.26	2.05
Chile	5.78	1.98
Mexico	5.00	2.46

Source: own elaboration from Latinobarómetro (2015)

The interpretation of these results should report the main findings and also include an interpretation:

> *According to the opinions of the respondents in their respective countries, the most democratic country is Chile (mean = 5.78), followed by Peru (mean = 5.26), and the least democratic of the three analyzed is Mexico (mean = 5.00). Although in Mexico is where opinions diverge most (standard deviation = 2.46). The numerous corruption scandals in Mexico might explain the skepticism of Mexicans with current democratic institutions.*

To set that a mean and a standard deviation are high, medium or low, we should compare the results between the categories (groups) of another variable, like we just did. In the previous example, we have compared the results of the question about "how democratic your country is" between three groups (countries). To appreciate if these differences are statistically significant, we have to carry out more complex techniques such as one-way ANOVA, which we will see in the following chapters.

In conclusion, the type of analysis to describe a single variable depends on the type of variable.

Exercise
Check the World Values Survey website and click on "Online Analysis". Choose a topic and analyze all the variables of this topic. Present the results of the univariate descriptive analysis for each variable. Consider the type of variable to select the type of analysis (frequency table, mean and standard deviation) and the type of graphs (pie chart, bar chart or histogram).

10. How to analyze contingency tables (crosstabs)

Contingency tables (also known as cross-tabulation, crosstabs, pivot table and two-way tables) are possibly the most used statistical technique in data analysis. Contingency tables show frequency distribution of two variables simultaneously and allow analyzing association between two variables using the relative frequencies (percentages).

Conditions for using contingency tables:

- This is a bivariate technique, i.e., **only two variables** are analyzed simultaneously.[2]
- Cross-tabulation is employed **with nominal and ordinal variables**. Nominal variables are those without any established internal order (e.g., gender, marital status, party affiliation), and ordinal variables are those with an established internal order (e.g., educational level, interest in politics). Scale variables, those with numeric responses, (e.g., age) do not usually fit in contingency tables. If we want to use age or another scale variables in a contingency table we must recode it by ranges (e.g. 18-35 years, 36-64 years, more than 64 years). When recoding a scale variable by ranges, it becomes ordinal, and therefore an analysis of contingency tables can be applied. Example:

Table 9. Contingency table: interest in politics by age groups

		Age groups			Total
		< 35	35-65	> 65	
Interest in politics	Very interested	53 10.0%	225 21.9%	148 30.5%	426 20.8%
	Rather interested	218 41.1%	445 43.3%	188 38.7%	851 41.6%
	Not very interest	160 30.2%	278 27.0%	130 26.7%	568 27.8%
	Not interested at all	99 18.7%	80 7.8%	20 4.1%	199 9.7%
	Total	530 100.0%	1028 100.0%	486 100.0%	2044 100.0%

Source: own elaboration based on World Values Survey

Which variable in rows and which on in columns?

Because one variable is studied in terms of another, the researcher must distinguish between dependent (or explained) variable, and independent (or explanatory) variable. If a hypothesis is causal or explanatory, independent (or explanatory) variable is placed in columns, and dependent variable (or explained) in rows. If a hypothesis is associative (and not causal or explanatory), the researcher must decide which variable is placed in rows and which in columns.[3] Contingency tables are composed of response categories of

[2] A 3-way contingency tables (with 3 variables) can also be built and analyzed, but they are barely used.
[3] Although contingency tables allow verification of associative and causal hypotheses, showing causality requires more advanced multivariate techniques such as regressions.

the variable in rows, response categories of the variable in columns, and cells between each pair of response categories. Cells contain two values: the number of cases that match in each pair of categories, and the percentage that these cases represent (see Table 9).

How to read a contingency table to test if two variables are associated

Two variables are associated whether frequency distribution of the variable placed in rows is different according to the response categories of the other variable placed in columns. That is, if the results of one variable are different in the response categories of the other variable, the two variables are associated. On the other hand, if the results of one variable are similar in the response categories of the other variable, the variables are not associated.

Column percentages are necessary to analyze contingency tables and test association between variables. Percentages based on column total are obtained by dividing the cases of a cell into the cases of the column total. For example, in the previous contingency table where the data of "interest in politics" and "age groups" were crossed, there are 53 people under 35 years old who are very interested in politics. To calculate what percentage they represent, 53 is divided into 530, which is the total number of respondents under 35 years of age. In this example, 10,0% (= (53/530)*100) represents the total of number of respondents under 35 years old who are very interested in politics. In the cell below, 41.1% is the total of those under 35 years who are rather interested in politics, and it has been obtained by dividing 218 (people under 35 years old interested in politics) into 530 (total people under 35 years). Software packages execute all these calculations, although it is always the researcher who must indicate what percentage is calculated, where variables are placed (either rows or columns), and how results are interpreted.

Recommendation:

A) Contingency tables should be read row by row, and from right to left. This is the best way to find out if the responses of a variable in rows are repeated equally or differently in the response categories of a variable in columns.

- If percentages are very different in the same row, association between variables is strong.
- If values are slightly different in the same row, association between variables is weak.
- And if values are very similar or equal in the same row, there is no association between variables, i.e., one phenomenon does not affect the other one.

B) To report the findings from contingency table, first describe similarities and differences of percentages row by row, and then indicate whether there is association between the two variables.

Example in 3 steps:

1. Hypothesis and selection of the two variables

For example, we want to analyze if "belief in God" explains "interest in politics". Our hypothesis is that people who believe in God have more interest in politics. We deal with two variables "interest in politics" and "belief in God", which are ordinal and nominal variables respectively, so a contingency table is appropriate to test association.

- Variable "interest in politics" has four response categories: 1-very interested, 2-rather interested, 3-not very interested, 4-not interested at all
- Variable "belief in God" has two categories: 1-yes I believe, 2-I do not believe.

To test this hypothesis, we use data from the World Values Survey conducted in Germany in 2013.

Before running an analysis of cross-tabulation, frequency tables of each variable are separately presented. This presentation is not necessary but it is displayed here to distinguish between univariate analysis of frequency tables and bivariate analysis of contingency tables.

Table 10. Frequency table: interest in politics

	Absolut frequency (cases)	Percentage
Very interested	426	20,8%
Rather interested	852	41,6%
Not very interested	568	27,8%
Not interested at all	199	9,7%
Total	2045	100,0%

Table 11. Frequency table: belief in God

	Absolut frequency (cases)	Percentage
Yes, I believe	1286	65,1%
No, I do not believe	690	34,9%
Total	1976	100,0%

2. Building a contingency table

According to our initial hypothesis "belief in God" explains "interest in politics", therefore, "belief in God" is the explanatory (or independent variable) and "interest in politics" is here the explained (or dependent variable). Thus, variable "belief in God" goes in columns and variable "interest in politics" in rows.

Column percentages must be calculated to analyze a contingency table. Although computer software packages run these operations, it is detailed here how column percentages are calculated.

- The value of each cell display cases that belong to each pair of categories. For example, in the top-left cell, 283 are the respondents who "believe in God" and are simultaneously are "very interested in politics".
- To calculate column percentages, cases in each cell are divided into the total number of cases in the column. For example, the number of respondents who "believe in God" and are "very interested in politics" is divided into the total number of respondents who "believe in God", and multiply by 100 to express it in percentage, i.e., (283/1286) * 100 = 22.0%. Other example, number of respondents who "do not believe in God" and are "very interested in politics" is divided into the total number of respondents who "do not believe in God", and multiply by 100, i.e., (122/691) * 100 = 17.7%.

After calculating column percentages for each cell, results are displayed such as in Table 12. Without column percentage, contingency tables cannot be interpreted because total number of cases is not equal in each column.

Table 12. Contingency table: interest in politics by belief in God

		Belief in God		Total
		Yes, I believe	No, I do not believe	
Interest in politics	Very interested	283 22,0%	122 17,7%	405 20,5%
	Rather interested	528 41,1%	306 44,3%	834 42,2%
	Not very interested	356 27,7%	190 27,5%	546 27,6%
	Not interested at all	119 9,3%	73 10,6%	192 9,7%
	Total	1286 100,0%	691 100,0%	1997 100,0%

Source: own elaboration from World Values Survey (2013)

4. Reading, description and interpretation of results in a contingency table

Remember, a contingency table is read row by row, and from right to left. Following previous example, this is a template to describe and interpret results:

> 20.5% of respondents in Germany are very interested in politics. This percentage is higher in people who do believe in God than those who do not believe (22% of those who do believe in God are very interested in politics, compared to 17.7 of those who do not believe in God). 42.2% of total

> respondents are rather interested in politics, and this percentage is slightly higher for those who do not believe in God (41.1% vs. 44.3%). 27.6 of respondents said they are not very interested in politics, this percentage is almost the same for believers and non-believers in God (27.7% and 27.5% respectively). Finally, 9.7% of the respondents are not interested in politics, and this percentage is slightly higher for non-believers in God (9.3% of people who do believe in God are not interested in politics, compared to 10.6% of those who do not believe in God). Responses about "interest in politics" vary among the categories of "belief in God"; therefore, these two variables are associated. It is a weak relationship because differences between believers and non-believers in God are only higher in the category of very interested. In the other categories of "interest in politics", the differences between those who do believe in God and do not believe in God are low or nonexistent. These results help us to understand that interest in politics does not depend so much on questions of religious faith, and that religious debate does not have much influence on political mobilization. Future analyzes should delve into other aspects to understand interest in politics, perhaps educational level or income level.

There is no rule of thumb to indicate whether differences between percentages are high or low since depend on the sample size and the number of categories. Therefore, there are statistics that test if two variables are associated. One of the most used is Pearson's chi-square, which contrast whether differences observed for each pair of response categories (cells) differ from those expected whether variables were independent. When significance of Chi-square is less than 0.05, the two variables are associated.

5. Conclusions:

a. Contingency table (aka crosstab) is a bivariate technique that analyzes two variables simultaneously to test if variables are associated
b. Contingency tables are designed for nominal and ordinal variables, but not scalar variables (unless are recoded into ranges).
c. We have to distinguish between the variable we want to explain (dependent variable) that goes in rows, and the explanatory variable (or independent) that is placed in columns.
d. Column percentage must be calculated in order to read a contingency table.
e. Contingency table is read row by row and from right to left.
f. The goal is to find out if percentages of the variable to be explained (the one that goes in rows) differ a lot, little or nothing between categories of the explanatory variable (the one that goes in columns). If there are high percentage differences, the two variables are associated, i.e., one variable explains the other. If there are no percentage differences, there is no relationship between the variables, i.e. variables are independent. And if percentage differences are small or occur only in some categories, the relationship between the variables is weak.

11. Bivariate correlations

The bivariate correlation is a statistical technique designed to find out:

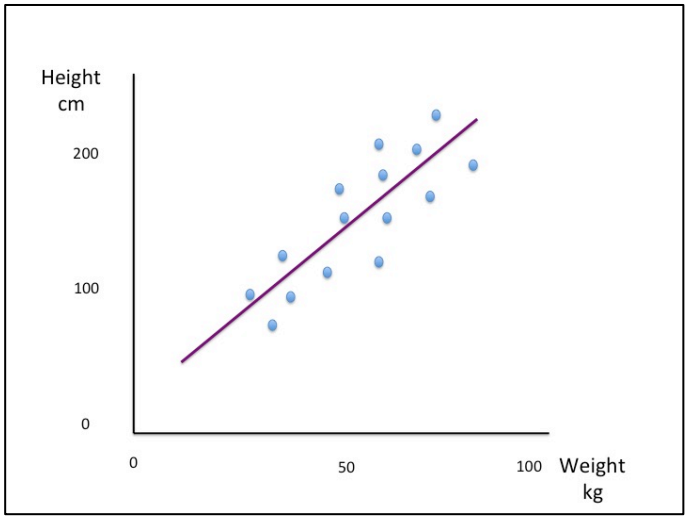
a) Whether two variables are related to each other
b) Whether the relationship is strong, moderate or weak, and
c) What direction does the relationship have?

The coincidences often hide associations between phenomena. Correlation is the most used technique to measure linear association in all sciences.

The analysis of correlations indicates association or relationship between two variables, and does not imply causality.

The correlation is based on the linear association, that is, when the values of one variable increase, the values of the other variable increase or decrease proportionally. For example, height and body weight have a positive linear relationship: when height increases, body weight tends to increases. If we make a plot of points with both variables, the point cloud will resemble a diagonal, which indicates that there is a positive correlation between these two variables.

Figure 6. Correlation between height and body weight



Source: own elaboration

There are 2 major types of correlations: Pearson correlation and Spearman correlation. Both are based on the same information, although they use different formulas. The Pearson correlation is more appropriate when the variables follow the normal curve. The Spearman correlation is more convenient to use when the variables do not follow the normal curve. In general, there are usually not many differences between the results, although they can vary especially when working with small size samples.

Correlation is used in statistical data analysis when working with ordinal or scale variables. The ordinal and scale variables are those whose categories have an internal order. If we include a nominal variable we must recode it to a dummy variable. A dummy variable is one that have only two categories or values, 1 and 0. The category or value 1 indicates presence of the phenomenon, and the category or value 0 indicates absence of the phenomenon.

How to analyze the bivariate correlation in 2 steps

The main advantage of using correlation analysis is that all information on the existence of relationship, strength and direction is synthesized in a correlation coefficient (r) and a level of significance (sig.).

1. The level of significance indicates whether or not there is a relationship between two variables. The most used level of significance is 0.05, which refers to the 95% level of confidence, which is the probability that the result is not due to chance. When the significance obtained from correlating two variables is less than 0.05, there is a significant correlation between these two variables. If there is a significant correlation, i.e., significance is lower than 0.05, we move to step 2.

2. The correlation coefficient (r) indicates how strong or weak a correlation is. This correlation coefficient can range between -1 and +1. The further away from 0, the stronger the relationship between the two variables. The closer to 0 indicates that the relationship between the variables is weak. If the correlation coefficient is very close to 0, both variables are not correlated. On the other hand, the sign (positive or negative) of the correlation coefficient indicates the direction of the relationship.

Several examples to understand it better:

Example 1:
The sample (N) consists of 2249 respondents in Colombia (World Values Survey 2005)

We analyze the relationship between "Ideology" and "Importance of God in life":

- Ideology is a scale variable that ranges from 1 to 10, where 1 means far-left and 10 means far-right.
- Importance of God in life is a scale variable where 1 means not at all important and 10 very important.

Table. Correlation between ideology and importance of God in life

		Ideology
Importance of God in life	Pearson correlation coefficient (r)	0.124
	Significance	0.000
	N	2249

There is a significant correlation between Ideology and Importance of God in life because the significance is 0.000 and therefore less than 0.05, so this result is not due to chance. The Pearson correlation (r = 0.124) indicates that it is a weak relation because it is close to 0. The positive sign of this correlation indicates that the more right the people in Colombia are, the more importance they give to God in the lifetime. And the same if we read it in an inverse way, the less importance of God in life, the more left are the Colombian people.

Warning: it is necessary to take into account how the categories of variables are ordered because the interpretation of the direction of the relation is based on the order of the categories. That is, the positive sign of a correlation indicates that the results resemble a diagonal upwards, but it is the researcher who must check how the results are interpreted. For example, the importance of God in life was asked on a scale from 1 to 10 where 1 means not at all important and 10 means very important, and ideology on a scale from 1 to 10 where 1 means far-left and 10 means far-right. Because the value of the Pearson correlation (r = 0.124) is positive, it indicates that the more importance of God in life, the more right are people, and, the lesser importance of God in life, the more tendency to be leftist. But if the variable importance of God in life had been asked on a scale from 1 to 10, but where 1 means very important and 10 means not at all important (unlike how it was originally done), the value of the correlation of Pearson between importance of God in life and ideology would have been (r = -0,124), that is, same value but with a negative sign. Negative sign (-) indicates that by increasing the values of the variable importance of God in life, the values of the variable ideology descend. The result is the same, but we must be very attentive to the order of the categories of the variables to correctly interpret the direction of a significant correlation.

Example 2:

Sample (N): 3017 people in Colombia (World Values Survey 2005)

We analyze the correlation between “Age" and “Interest in politics”

- Age is a scale variable
- Interest in politics is an ordinal variable where the categories are: 1-very interested, 2-somewhat interested, 3-not very interested, 4-not at all interested

		Interest in politics
Age	Pearson correlation coefficient (r)	0.013
	Significance	0.467
	N	3017

There is no significant correlation between age and interest in politics because the significance is higher than 0.05 (Sig = 0.467). When age increases, interest in politics

does not increase or decrease. Therefore, we should look for other variables if we want to understand what the interest in politics is related to, since age does not correlate with interest in politics.

The use of correlation is useful for characterizing and profiling. For example, it would make it possible to identify what are the characteristics of the religious people. In addition, correlations allow analyzing relationships between phenomena (or variables). For example, is there a relationship between investment in education and crime reduction? To what extent does foreign tourism reduce poverty? To answer these research questions, correlations can be essential.

QUANTITATIVE ANALYSIS

THE GUIDE FOR BEGINNERS

By

Julián Cárdenas

julian.cardenas@onlinebschool.com

www.networksprovidehappiness.com