Chapter 1: Why Statistics is hard[1]

I have often conducted the following test in class. On the first day of the course, I ask students to imagine an alternative reality where Statistics is not a compulsory subject in Psychology. Then I ask them to imagine they have a button in front of them that, if pressed, will make this alternative reality come true. You may have guessed that I later ask how many of them actually pressed the button and we have a nice discussion about why they did that, with the rather conspicuous intention of trying to convince them that this would not be such a good idea.

Not that I am very successful in changing their minds. Not many students are enthusiastic about the contents of a Statistics course in Psychology before it begins. Indeed, only a few of them seem to change their mind once the course has ended. In fact, I sometimes get requests for information about candidates for jobs related to data analysis and Psychology but very rarely can I recommend any recent graduate that may be suitable. In my experience, interest in Statistics among Psychology students is not great.

However, I don't exactly blame them . The opinion that Statistics is a difficult and dry subject, or worse, deceitful and confusing, is common among laymen. In the statement, probably falsely attributed to Mark Twain, "There are lies, damned lies, and statistics", statistics is regarded as the grounds used by politicians, charlatans and vendors, etc., to justify any idea, whether true or false, using numbers that nobody really understands or bothers to comprehend. Jokes about Statistics can fill many books (I love the ones by Forges, one of my favorite authors of all time), but I am sure you can find many yourself on the Internet.

[1] . When searching through books on introductory statistics, I found that the title of the first chapter of "Discovering Statistics Using IBM SPSS" by Andy Field is "Why is my evil lecturer forcing me to learn Statistics?" I guess all teachers of Statistics in Psychology share the impression that their students believe their classes are a fate worse than death.

I sometimes meet colleagues who teach Statistics in Engineering, Medicine, Chemistry, and Journalism, etc., and who also find it difficult to spark much of an interest in Statistics among their students. Heck, even Math students are usually not very enthusiastic about Statistics.

Considering that Statistics is a curricular component of many degrees, this is indeed an extremely curious phenomenon. It may be easier for us to list the exceptions (think Law, Philosophy or Linguistics) than to list degrees that require a course on Statistics. However, in reality, the fact that they don't study it doesn't mean that they shouldn't. For example, statistical legal arguments are often used in courts, while automatic language translators are based on sophisticated algorithms that use, guess what?, statistical analyses of the use of words in sentences. Philosophers also study the problem of how to acquire knowledge as one of their core issues and, since statistics is the main tool of many disciplines for that purpose, it would certainly not harm them to learn a little bit about it. As we will see later in this text, the Philosophy of Science is an important subject for understanding numerous concepts in Statistics.

So, if Statistics is such an important topic, why does everybody dislike it so much? While I can't claim to have the perfect answer, here are my five cents on three reasons for disliking Statistics.

Three reasons for disliking Statistics

Here are three reasons that I think partially explain why Statistics is distrusted, disliked or misunderstood by almost everybody who gets exposed to them.

Statistics are used for deception

First, people don't trust Statistics for a good reason: they are constantly misused, and often with malign intention. This feeling is captured in the statement "The only statistics you can trust are those you have falsified yourself". I think everybody has once had the feeling they have been deceived by tricky Statistics that try to make us see reality according to the dark interests of evil entities – perhaps

governments, politicians, businesses, or vendors of fake products.

True as it may be, I think the above argument should be taken with a pinch of salt. Actually, with a lot of salt. Of course, I agree with the opinion that Statistics are used for this purpose on many occasions but we should not extend that argument to every situation in which statistics are used. Incredible advances in industry, medicine, telecommunications, physics, and all areas in general, have Statistics at their core. Every time you take a pill, buy clothes, travel on public transport, click on a website banner, or do countless other everyday activities, some statistic has been calculated so that the pill will have its intended effect, the clothes in your size will be available in sufficient numbers, the buses will run with the expected level of occupancy, and so on. Don't throw the baby out with the bathwater: although Statistics can, are and will be used for deception, we do need them. You may be so used to them that you think they are unnecessary but without them our daily activities would be unpredictable in ways that would make them unbearable.

In fact, knowing Statistics is how to avoid getting manipulated. Understanding how data are created is also the way to understand how they are fabricated. If so many aspects of our lives depend on statistics, we will be better off knowing more about them, not less.

Statistics are as hard as Math

Second, most people dislike Statistics because they think the subject is like Mathematics in some way. Psychology students are just one group but the list is not restricted to them. Here I'm not talking about the layman who is exposed to the media but not to Statistics. I am referring to undergraduate students who perform successfully in other subjects but when they have to take a course on Statistics they simply hate it straight away. Why this feeling is so general has baffled me for many years and I still don't have a proper answer for it. However, I do have a few theories, which I will discuss below:

One possible explanation is that those who are unhappy with Mathematics are also unhappy with Statistics. Many people don't like Mathematics at all and it could be claimed that, since many students who choose Psychology as their main degree fall

into this category, this explanation appears to be correct. However, although I used to believe this explanation, I have changed my opinion over the years. This is because current courses on Statistics in Psychology have reduced their mathematical component so much that they can hardly be considered "Mathematics": there are no calculations (we use software instead), we don't show many formulas, and we use examples based on real data. Very little math there, then.

A second explanation I have recently come to appreciate is that, for some people, the problem is not that Statistics is like Mathematics but that Statistics is not like Mathematics enough. Or at least, it is not like the Mathematics that is taught at certain levels. It seems to me that many people see Mathematics as a set of procedures that can be rigidly applied to a set of problems with certain features to provide a number as the solution. Actually, the number is the solution and finding and applying the formula is the algorithm.

But when we use Statistics, the number is not the solution. Performing the calculations correctly is necessary but not sufficient. To reach a conclusion, as a final step you need to interpret the result.

Why is it so difficult to interpret statistical results? The answer to this question takes us to our third explanation of why Statistics is perceived as difficult for many people.

Interpreting the results

The third explanation of why Statistics is perceived as difficult is related to the second. Earlier we saw that results obtained using Statistics need further steps so that they can be used, which means that the numbers must refer to the context in which they were calculated. Without a context to guide our interpretation of the results of statistical analyses, drawing conclusions is not so straightforward a process as that used in exercises intended for learning Mathematics. Moreover, in Psychology, conclusions from statistical analyses can be more ambiguous than those of other disciplines, so results often leave students and researchers feeling disconcerted.

Let me illustrate this with an example: imagine you test a sample of students for their numerical ability and the average result is 50. What does this mean? Well, basically nothing, without a

context. So, to interpret this figure we need to know more about the test itself: what is the normal average of those who take this test? Is it 200 or 30? If it is 200, our students are probably not very gifted. But if it is 30, they are. But what is the age of the students? Are they relatively young? If so, should they be compared to students of the same age or not? Also, what were the test conditions? Were the students motivated or not? Without knowing all these details, statistical results are very often meaningless and therefore of little value.

In my teaching, I try to make up for this lack of context by elaborating descriptions of the examples. So, I expand a little on the story behind the data. I also try to find real data, and never invent them for the sake of illustrating a statistical analysis. After all, as a book on data examples states[1], students exposed to fabricated data may think: "If the technique is as important as the teacher claims, how is that he/she has been unable to find a real example?" But time is limited and in my lectures I often feel I don't dwell enough on examples. However, if you are interested in knowing further details, you are welcome to chip in and ask me for them.

In summary, statistics are not useful on their own. In the next section I will introduce other elements that are needed to make useful interpretations of their outcomes.

What it is necessary to interpret statistical results

The knowledge needed to understand statistical results in Psychology – and in other subjects – is based on four elements:

- Research design: This is the plan or strategy behind a study. It ensures it is conducted coherently and that it addresses the research problem correctly. The component of the Psychology curriculum that covers these issues is generally called Design Research Methods or Research Methods.
- Measurement: A study needs to measure the objects or subjects that compose it. In Psychology, measurement is

[1] Hand et al. (1994). *A handbook of small data sets*. Chapman and Hall.

associated with questionnaires and psychological tests. However, other types of measurements are also used, including survey questions, counting events in video-recordings, electro-physiological sensors, and narratives, etc. In Psychology, measurement is called Psychometrics and students quite often find this subject interesting because it can be used in practice. For example, psychological diagnostics are based on tests based on the methods used in Psychometrics.

- Statistics: The data collected must later be analyzed. This course is mainly concerned with this aspect. However, since we will always use examples, we will sometimes mention elements of research design or measurement.

- Subject knowledge: The above three points are related to what is called research methodology. However, methodology is not sufficient on its own: we need questions about the world. Sound research questions derive from practical experience and previous research, etc. However, experts with years of experience often remark at this point that they could benefit from methodological knowledge.

As you can see, like in many other fields, research in Psychology involves four types of knowledge. Until you have a reasonable understanding of these four types, you may not understand the importance of statistics and research in sciences such as Psychology.

Chapter 2: Why do applied psychologists need to know Statistics?

Many first-year students of Psychology probably think that the professional career of the applied psychologist follows these steps:

- They learn what is needed to work as a psychologist at the university. The knowledge they acquired on their degree is rather general but their Master's provides the skills required for a specific field of work.
- They start working in a position where clinical experience helps to refine their knowledge.
- During their professional life in a specific field, new treatments or methods arise, and they may have to decide whether they are worth using. They can learn these methods by attending short courses in which experts explain how they can be applied in practice.

In short, most of their knowledge derives from other people who perform the hard work of separating the wheat from the chaff.

Notice that this plan avoids the need for Statistics – the psychologist does not need to know much about that subject because somebody else is doing the hard work for them. Of course, they will also learn for themselves from clinical experience – nobody can do that for them – but they do not need Statistics for that since this process will basically be applied on a case-by-case basis.

No doubt some of you will follow this plan, so you may claim there is no reason for you to attend this course. This view is not very realistic, however. In fact, you may increase your chances of having a successful career if you study Statistics not as a subject

you must pass but as an aspect that could increase your chances of earning a good salary. Here is a list of reasons why I believe Statistics can be of great interest to you:

- You never know.
- There are fads.
- Everybody is a psychologist.
- Be a psychologist, but of a different type.

I will discuss these reasons separately below.

You never know

Earlier I described the plan some (perhaps many) of you may have in mind as a future professional career. Plans are nice to make, but good planning must always consider that things rarely work out as intended. In summary, you don't know if you will end up working as a psychologist or even if you will want to work as a psychologist once you finish your degree (nobody knows)[1][2].

A recent report on statistics on employment[3] and the affinity of jobs to the degrees studied sets Psychology in an intermediate position, i.e., roughly 52% of Psychology graduates are employed in positions related to that subject, and roughly 60% have a job. Note that these values are not the worst: in other cases, only about 10% of people are working in areas related to their degree and some degrees have less than 20% of their graduates in any form of employment. Of course, the opposite also occurs in some cases: for example, 93% of those who studied Medicine are working, and all of those are working in areas related to their degree.

[1]. Actually, whenever you analyze cases individually and then draw conclusions from them, you are using some sort of cognitive-based Statistics. Unfortunately, as Kahneman (2011) showed, this kind of process tends to be strongly biased.

[2]. Sometimes I hear students complaining that we do not train them for real jobs. The reason for this is that we do not know where you will end up working, so general knowledge is the best bet.

[3]. You can find a nice plot with this information here: http://www.elmundo.es/especiales/educacion/empleo-universidad.html

Well, never mind. These percentages are not new. For many years, psychologists and other graduates have been working in areas that are not "theirs". In Psychology, I have long heard complaints about this state of affairs as a sort of societal blindness. It is often said, for example, that there are many places where psychologists are needed but that government, companies, and schools, etc., are unwilling to devote the resources to hire them. Sooner or later, they hope something will happen to open the eyes of the world to the importance of Psychology, and that things will then change radically for our profession. Well, let's hope so. Meanwhile, you'd be better off taking things as they are rather than as you wish they were.

Psychology is a degree subject that can help students develop general skills that can be applied in many places. Many jobs exist in which the human factor is the key to success. Whatever the goal of the organization you end up working for, humans will be a part of it, either as workers, customers, patients, or users, etc. Of course, you may have to learn a few new things that you did not study on your degree but, as the other graduates will have to learn things about human nature that you will already have studied, you are not necessarily in a worse position than them.

So, what about Statistics? Well, people working in recruitment often recommend that we should learn general skills that will make us good candidates for different jobs. These general skills help us to find a job when we are starting out on our professional career but they also help us to start a new career if needed. It is said nowadays that there is more uncertainty than ever in employment and there are no guarantees that we will stay in the same place forever.

Statistics is probably one of the general skills you will find applicable to more situations – notice how it is studied in almost every degree you can imagine. Knowing these general skills may help you find good jobs in many places. And these are often better jobs than those you are offered as a standard psychologist.

There are fads

One of the things I find most curious about Psychology is the constant flow of fads we are exposed to. I don't want to mention any but you can do your own research, find out what is the current one, and decide whether you like it. If you don't, don't worry, another one will come along soon to replace it.

Of course, many of these fads are not new: the wrapping may be slightly different but the same concepts have been around for years. Some of them may be zombie ideas: ideas that were killed off by past research or practice but still manage to walk around and eat your brain when you're not looking.

Should you follow fads? Well, there are good reasons for doing so: you sometimes get more rewards by following the crowds, even when they are walking in the wrong direction, than by calling them fools.

Sometimes you must really question yourself: if so many people are so willing to accept this new trend, treatment, approach, or school of therapy, etc., might I be missing something important? Is this fad a good one?

If you are in this situation, I recommend you look at the evidence. Find out if scientific papers have been published on it in respectable journals, look at the sample sizes, try to find independent studies about them. Do your homework. Don't trust pundits. Don't read books by people who just want to sell their ideas without producing any proof. Beware of courses in which people introduce fantastic new methods or therapies. All of them are selling their products and, of course, are so enthusiastic about them. If you were a doctor, would you believe everything you were told by pharmaceutical companies[1]?

Statistics and other methodological subjects may save you a great deal of time, money and effort throughout your professional career. If you develop a certain skepticism and have the proper tools to exercise it, you may find that you avoid wasting a lot of energy in the future.

[1] Actually, pharmaceutical companies probably have to abide by tougher tests than those we have in Psychology because the research they conduct is carefully regulated and any new medicine needs to pass controls that are not so common in our field.

Everybody is a psychologist

Naive concepts of psychology are very common. Everybody has their own. Some are simple, while others are sophisticated, harmful, willful, positive, or negative. Don't think you don't have your own: you have. We all have.

Very often you will find yourself discussing these concepts with non-psychologists. Do you think you will end up as the winner in those discussions? Spoiler alert: it's not that easy.

Don't even try to tell them: "I am a psychologist with a diploma, and you are not". That isn't going to work. People really enjoy challenging psychologists.

In such cases, try the Science argument: "I know this works this way because there was this experiment with this number of subjects, etc., etc., and the results were these".

Sometimes you will win the day in that way[1].

Be a psychologist, but of a different type

Psychologists are traditionally identified as therapists working individually on patients' problems. However popular this view may be, psychologists also work in many other areas in which skills for data analysis are the keys to success. Moreover, new technologies have brought opportunities for psychologists that should not be dismissed unthinkingly. Below are some examples of what I mean.

Recycling

Psychologists could save the world. Some of the greatest challenges facing modern societies are not individual, and

[1] You would be worse off if you were a football coach. Everyone thinks they know more than football coaches.

psychologists could collaborate to convince people to help solve such problems.

A few years ago, on a visit to a university in the UK, I had a brief chat with a Ph.D. student. Her thesis was on the psychological aspects of recycling. I found this topic mind-blowing. It was so full of possibilities!

Unfortunately, I have not been in contact with her since but I have seen several papers on that topic. Understanding what the psychological or practical barriers of proper recycling are could make such a difference! Simply implementing some of those ideas in campaigns and measuring the impact in terms of kilos of recycling would be a great job for psychologists working for the government. Then they could talk with industry and retail companies to figure out how to get the best out of people in this aspect.

Statistics would probably be needed, but wouldn't the rewards be worth the effort?

In the last three years I have been working on a subject that has several similarities with recycling, i.e., cycling, which is another hugely important topic nowadays. Mobility is responsible for a large proportion of emissions that lead to global warming as well as health problems in urban areas. The wider use of bicycles may help to make our lives better in numerous ways. However, there are many rough edges to smooth out before significant numbers of people will choose bicycles as their chosen mode of transport. Some of those are practical but there are a few psychological ones too. I have been analyzing surveys on these issues and will use some data from them in my classes.

Human Computer Interaction and Human Factors

I wrote my Ph.D. thesis on a topic you may think is unusual for a psychologist: the title was "Formal methods for describing human-computer interfaces". At the time, this was quite an unknown topic in Spain. However, I spent some time working with psychologists in this area in the UK and Australia, where it was also a new field of application. This was before the Internet and smartphones were so prevalent, and many questions were raised about how to design these new tools. "User-Centered Design",

"mental models of tasks", "intelligent interfaces", and "design for all" were the buzzwords of the moment.

In the last two decades, this area of work has come of age and there are many Master's, conferences, books and, more importantly, jobs, in it. Psychologists help to design and evaluate new technologies, providing expertise about how people think or feel and incorporating this knowledge into real products. Big companies now have teams of HCI researchers who can potentially impact the experience of users of their products. Their work is based on understanding user behavior and real use in areas such as ubiquitous computing, social and collaborative computing, interactive visualization, and visual analytics.

Over the years I have moved onto other areas but occasionally I have met students who have developed an interest in those topics by themselves and have contacted me about them. I have sometimes been in touch with them afterwards and they seem to be doing quite well by and large: there don't seem to be many Spanish psychologists who are interested in this topic and I guess the lack of competition helps.

What is the typical work of a psychologist as user-interface expert? Since the Internet is now the main channel of communication, I would say that their main activity is evaluating how easy-to-use web sites are. Conducting interviews and sometimes simply observing what people do and taking notes can take you far. However, you can also collect clicks and count time spent on pages or unfinished transactions, etc. Of course, Statistics is an important component of these last tasks since the critical elements measured can be used to design or redesign web sites.

Information technologies and Psychology

Given how important information technologies are today, it would be strange if psychologists did not have opportunities to apply their knowledge to them. For example, I know some former students who have worked at companies feeding contents onto websites or managing discussion lists. I wouldn't be surprised if more psychologists get jobs like those in the future.

Below I have selected an example I have used in past Statistics classes to illustrate the kind of expertise a statistician-psychologist can bring to companies working in these fields.

Finding couples online is now a multimillion dollar business. Web sites offer the chance to find people from the information provided by users. This information may be extensive or minimal (just a photograph). If things work out, people go on dates and rate them – so we have information about how things went.

Obviously, Psychology and Statistics play a key role in this process. You need Psychology to define what information to request and how to request it. People can lie, so you need to think about mechanisms for correcting this (or maybe not). In short, you need a theory on romantic relationships and sex appeal, etc. With this theory, you can develop the questions and create the method for doing the matching. Apparently, one of the most common complaints users of these sites have is the time wasted on dates that they knew were not going anywhere after a conversation of just five minutes. Good sites can help to minimize such wasted effort.

Of course, the data for testing and improving your theories are available. People report on the success of their dates (or you can simply check whether they are looking for a new one after a short time). The possibilities are endless[1] and, I reckon, full of fun. Yes, once again, statistical analysis is involved, but so what?

[1] . If you are interested in this topic, after a cursory Google search I found this site: https://www.luvze.com/about-us/. I cannot offer any assessment of the quality of the site because I have not explored it thoroughly, but I do like their motto: "The important things in life deserve data".

2 Causality

Somewhere between 1900 and 1912 in this country, according to one sober medical scientist, a random patient, with a random disease, consulting a doctor chosen at random had, for the first time in the history of mankind, a better than fifty-fifty chance of profiting from the encounter.

Attributed to Lawrence J. Henderson

After reading the above statement attributed to Lawrence J. Henderson, did you think that, before the 20th century, the safest course of action if you felt sick was not to go to the doctor's? If so, you may be jumping to conclusions. Without denying that many treatments applied until a century ago were often worse than the illnesses themselves (Bryson 2019), the above statement omits a hugely important detail, i.e. the odds of getting better *without visiting the doctor*, which could have been much worse than those when visiting the doctor.

The above statement can be made more general: to know whether the effect of a treatment is beneficial, a comparison must always be made between receiving or not receiving that treatment. Although this may not seem obvious to you now, I hope it will once you have finished reading this chapter.[1]

Treatment is an action applied to an individual unit which, in psychology or medicine, is usually a person but in other disciplines could be a plant, an animal, a substance or an object. Examples of treatments in Psychology are clinical therapies, educational methods, and personal choices such as studying at university, following a vegetarian diet, or making teenagers have dinner every night with all the family. In all these cases, our interest lies in determining whether the treatment is effective or not, i.e. whether a certain therapy improves the patient's well-being, an educational method improves academic performance, a personal choice such as continuing your studies helps you get a better-paid job, becoming vegetarian means you will live longer, or teenagers having dinner with all the family will have fewer problems in adolescence (Meier and Musick 2014).

Unfortunately, before 1912 the odds of a statistician knowing how to analyze data to make a correct comparison between two conditions were even lower than the odds of benefiting from going to the doctor's. Moreover, this assumes that the data on the effectiveness of the treatments have been collected systematically, or even just collected. In reality, the progress in statistical theory that makes it possible to compare the effect of treatments (medical or otherwise) as well as the technical and logistical infrastructure required to collect these data have a history of less than a hundred years. And we still have a long way to go, judging by the fact that the Nobel Prize in Economic Sciences awarded in 2021 went to scientists working on answering causal questions using observational data[2].

I wouldn't be shocked if the above surprises you. After all, clarifying whether an action produces an effect is no more than establishing whether a causal relationship exists between two events – and children already learn to do that through play when they are little. However, despite how apparently simple it is to know whether one thing causes another, philosophers have been discussing causality for

centuries without settling the issue. And later, when statisticians joined the conversation, their input was not as decisive as one might expect. As Holland (1986) says: *"The reaction of many statisticians when confronted with the possibility that their profession could contribute to the debate on causality is immediately to deny that such a possibility exists"*. By way of example, he cites the following excerpt from an article on causality written by Barnard (1982): *"That correlation is not causation is perhaps the first thing that should be said."*, which, though true, is insufficient without saying what causality is.

So what is the problem? Why these doubts? Shouldn't scientists ever claim to have identified the cause of this or that phenomenon? And if so, are there no exceptions? Below I will try to answer these questions.

2.1 Correlation is not causation

"Correlation is not causation" refers to the fact that when we observe that two events usually occur together, or that one follows on quickly from the other, we should not jump to the conclusion that one causes the other. Or in statistical terms, if two variables are related to each other, this does not mean that one is necessarily the cause of the other. That correlation is not causation is a cliché found in almost any introduction to a book on statistics, usually accompanied by examples such as:

- There is a relationship between the consumption of ice cream and the number of people who drown in swimming pools. However, that does not mean that eating ice cream is a cause of drowning but that in summer both events tend to increase whereas in other seasons they tend to decrease.

- People who listen to disco music tend to have more sexual relations. However, that does not mean that this type of music has such an irresistible influence that it increases such behavior exponentially but that the people who prefer disco music are younger than those who prefer ballroom dancing music, and age is, of course, related to sexual desire.[3]

The explanation is that both these examples contain an intermediate variable (the confounding variable) that affects the two observed variables and accounts for the association or relationship between them. However, if we ignore the confounding variable and focus only on what is observed, we could invent some causal explanation[4] and even propose measures to act on the cause in an attempt to influence the (supposed) consequences. For example, we could propose banning children from eating ice cream so that they won't drown in swimming pools, and we could use music as sex therapy for couples with sexual problems. Does this sound ridiculous to you? Here is a real example: since many studies show a negative relationship between having dinner with the family and drug abuse, in the United States there are public campaigns to encourage families to have dinner with their teenage children to prevent addiction (Meier and Musick 2014). Can you see the intermediate variable here?

Since the problem resides in the intermediate variable, the way to ascertain whether a relationship is causal is to study the association between the two variables in a pure way without interference from other variables. This is achieved by using laboratories that are isolated as much as possible from external factors that could confound the observed relationship. *As much as possible* is the key phrase here. Although external variables can sometimes be removed, in Psychology this kind of control is more difficult than in disciplines such as Physics or Chemistry, where scientific principles have been demonstrated in this way for centuries. To advance knowledge in

sciences that do not lend themselves so easily to research in the laboratory, statistics itself has had to progress in methods that can test causality outside the laboratory, an approach that began to bear fruit only a few decades ago. For this reason, it was common among statisticians to avoid the issue of causality beyond providing examples in which correlation was not causation.

Similarly, although introductory books and courses on statistics would provide examples of why correlation (or association in general) was not causation early on, it was common for them not to discuss the issue afterwards (Cummiskey et al. 2020). In my opinion, this was disappointing for the students because whenever practical examples were discussed in class, the conclusions were always of the kind: "Since we can see that there is smoke (correlation or association), there must be a fire (causality), but we can't affirm as much because correlation is not causation and so all we can say is that perhaps this causes that but…". This is all highly evasive and uncertain, as you can see.

Fortunately, the topic of causality is becoming an established component of introductory courses on statistics. However, there is still a long way to go before we can find the best approach to presenting it. In my opinion, the best way is to start with how we typically draw causal inferences in everyday life and, despite the usefulness of this type of reasoning, examine its limitations.

2.2 Motivating example

Students with a long road of studying ahead of them may believe at some point that taking food supplements could help them to manage the road more easily. Indeed, if you search the internet, you will find

a great many foods, including avocados, fish and berries, that are promoted as being beneficial for cognitive skills [5].

Let's imagine that a student decides to test whether eating a lot of avocados is the secret weapon he needs in his first year at university, starts eating them every day and, at the end of the year, passes all his exams with honors.

Do you think this test proves that the student's magical diet improves his academic performance? I imagine you will say 'no', but why not?

2.3 Comparisons

Remember I said earlier that in order to conclude that a treatment has an effect, an element of comparison (a control) is needed, which in this case would be that the student has spent the first year without taking avocados[6] and then taken the exams. In other words, to be able to conclude that avocados are the real cause of the improvement in the student's grades, we would need him to do the same things twice except when it comes to the avocados[7]. Since this is impossible, it leads to what Holland (1986) calls *the fundamental problem of causal inference*:

It is impossible to observe the value of Yt(u)Y_t(u) y Yc(u)Y_c(u) on the same unit and, therefore, it is impossible to observe the effect of tt on uu.

In the above, Yt(u)Y_t(u) is the value of the variable of interest after a treatment is applied on a unit, Yc(u)Y_c(u) is the value of the variable of interest after a control treatment is applied on the same unit; and tt is the difference between the two previous variables (i.e. the treatment effect). We cannot observe the difference *on the same unit*

because it is impossible for it to undergo both treatments under identical conditions.

If we apply this to our case, *it is impossible to observe both the effect of eating and not eating avocados on the same student*, so we cannot know if the student would have obtained the same grades if he hadn't followed the diet.

"Hang on a minute," you may be saying, "exactly the same situation is impossible but something similar might be." Indeed, one way the student could conduct his test is by *creating two situations that are largely similar to each other but that differ in the avocado diet.* If this could be managed, this comparison could be made.

I see two ways in which our student could create these two situations: First, he could use himself and look at the results he was getting in the past or he could use the results he gets at a future moment in similar situations. Second, he could look for someone similar to himself but who does not follow the same diet. Both strategies resemble what people do in everyday life to evaluate the effect of a certain treatment. It is interesting, therefore, to analyze them to understand how we can make causal inferences without using laboratories – and to see how sometimes those inferences are incorrect.

2.3.1 The student compares with himself

Let's begin with the case where our student assesses the effect of his dietary change by comparing his freshman-year grades with his senior high-school grades. Suppose that in his high-school year he didn't perform well, noticed his memory sometimes faltered, often felt anxious and overwhelmed, and obtained results that were not very good (but good enough to get to University). In this case, he may

well conclude, “Thanks to avocados, in my first year at university I found everything much easier and more fun.”

Another possible situation is that of a student who compares his first year at university – his avocado year – with his second year in which he turned his diet towards more sustainable foods for the planet and found that, by doing so, his grades dropped dramatically. In this case, the student would surely conclude that his brain was made to not function properly without a daily supply of unsaturated fats and so continued to eat avocados every day until the end of his days.

I think you might easily be able to deduce that, although the student was strongly convinced of the effectiveness of that creamy green manna, his experiments do not provide solid evidence of the effect of avocados on memory. For example, alternative explanations for his good results in the first year of university include: a) he has chosen a major that is easier than his high school subjects, b) it is easier *for him* because he is now focused on things that interest him, c) his personal situation has changed and he is now more focused on the task, and d) he is now more mature in general, so he worries more about his diet and studies more. In short, many things could have happened in addition to the avocado that might explain his results.

In summary, these kinds of individual experiments can be good as a source of inspiration: if this student tells you that avocados have been good for him, you may reasonably be encouraged to try them too, but don’t have high expectations. However, if the student writes a book, teaches a course, sets up a foundation, or founds a sect dedicated to disseminating the benefits of avocados on educational performance by providing only the experiment we have described, I would advise you not to take it too seriously.[8].

It is true that, in our daily lives, we are all exposed to anecdotal information from other people based on “experiments” of this kind.

And it is true that some of these contributions can be useful to us at certain times. But we must always be wary of something that seems obvious in principle but is sometimes not so obvious in practice – just because someone is convinced that something has worked out for them, it does not mean that this is truly what happened or, more importantly, that it will also work out for you.

Why, then, do we take such anecdotes so seriously? What is it about them that makes them so convincing?

Experiments that are conducted in this way are convincing because we unconsciously believe in the *stability* of the investigated unit from one condition to another. For example, in the above cases, we might accept that the student is the same in both moments and that the difference we observe between them is a product of the diet only. Or, even if we don't really believe that the student is literally the same in both moments, we dismiss the idea that the differences between the two are important enough to affect the results. Of course, this is not the case here: the student in his first year at university may be so different in numerous important aspects from the person he was a year earlier or a year later that drawing causal inferences can be very misleading.

Note also that, though not valid in this case, there are other contexts in which stability can be defended. Take for example a chemical study that uses water to determine how temperature change affects its volume. If we compare the volume of a bucket of water at two different temperatures, we are confident that the water is essentially the same substance in both cases and we can draw our conclusions in the assurance that our results will not be skewed. With people, on the other hand, stability is often difficult to uphold since people tend to be significantly different from each other and from one time to another.

Another strategy we could employ is to use someone who is similar to our student as the unit of comparison to determine whether avocados have the same effect on him. Next we will see how far this strategy can take us.

2.3.2 The student compares himself with another student

Another strategy the student could use to decide whether eating avocados every day affects his memory and therefore his grades is to compare himself with another student who does not follow the same diet. In theory, you might think this test would be easy to manage since the supply of students is large and it should be easy to enroll one in such a test. However, remember that we need someone who is as similar as possible to the individual unit being tested (our student) and who is willing to do pretty much the same things our student would do over the same period. Let's see some options.

A convenient situation would be one in which our student had a twin who studied the same subject, liked to eat the same things, chose the same courses, lived in the same place, and did the same activities, etc. What chance does the student have of achieving something like that? Very little, really. However, were this to happen, we would be facing one of the best possible situations for obtaining causal conclusions because both the external and internal factors would be similar and would therefore not confound our causal inferences. The only thing that would vary between the two students would be the treatment tested, i.e., whether they eat avocados or don't. Obviously, "similar" here is relative since there would always be little, unavoidable differences between what happens to one and what happens to the other – beyond the fact that genetic equality between twins is not absolute – and we would have to think carefully whether some of those differences were significant enough to render the results

unconvincing. Nevertheless, research with twins is a favorite strategy in Psychology research thanks to the advantages outlined above and, despite its limitations, is certainly an opportunity worth exploring if available.

However, since most of us don't have a twin around to test whether a treatment works, we have no choice but to turn to someone else. Therefore, the student who is interested in evaluating how avocados affect memory should search for someone with a similar age, similar gender, similar eating habits and similar course, and who ticks as many of the other boxes he can think of, until he is fully satisfied – which I think is very difficult to achieve.

In summary, an experiment that uses only two people seems unlikely to ever look convincing. Perhaps in our everyday lives we can take for granted that "if my roommate has done well, it will work for me", but in reality we all know that no two people are the same. So if a person who has eaten avocados every day does well in his exams and another who has not eaten them does not, this evidence is not sufficiently convincing no matter how similar they are. Again, if you treat it with caution, seeing how a treatment works with others is fine as long as you don't expect too much.

And again, in other disciplines, comparing two units of research may also be reasonable: an experiment in chemistry may be convincing with only two similar water samples provided we are convinced of their stability. With people, on the other hand, and with living beings in general, this is surely not the case.

So, what can we do in Psychology? As we will see below, one way to obtain more convincing evidence of the effectiveness of a treatment is by comparing the average effects on various subjects.

2.3.3 Comparing with the average effect

If you have followed the above explanations, I think you will accept that drawing causal conclusions about the effect of food on academic performance from just two students, each of which is assigned a different treatment, is difficult but may be acceptable in other contexts. For example, in a chemistry experiment carried out in a laboratory in which everything is controlled and pure substances are used, using two units may be enough to determine the effects of the treatment on the control. In this case, we would be applying what Holland (1986) calls *the scientific solution* to the problem of causal inference. In his own words:

Science has progressed a lot using this approach. The scientific solution is something very common also in our daily life. We all use it to make causal inferences that appear in our lives (p. 947).

Indeed, the scientific solution is the intuitive solution for human beings as it is easier to understand than the one required when the assumption of stability cannot be held, i.e., when the units have *variability*, we need another solution that is much less intuitive for humans. This is the *statistical solution*, which involves calculating the *average causal effect* obtained from the difference between the mean values in the variable of interest – exam grades in our case –computed on a sufficient number of people who follow the treatment versus others who do not.

The above solution may not seem like much to you[9], just as Holland thinks it may not seem much to many people. However, in his opinion, its interest resides in the fact that *the statistical solution replaces the impossible-to-observe causal effect of tt on a specific unit with the possible-to-estimate average causal effect of tt over a population of units*, and therefore provides a solution to the fundamental problem of causal inference.

The key to the statistical solution is that it evens out differences between individual units: if we use only one student per condition, there will always be some characteristic that makes them distinct from each other and could provide an alternative explanation for the results found. For example, a student in the treatment group could do more sport than another and we could not rule out that this was the real cause of his grades. However, if we select appropriately (and 'appropriately' is the key word) a sufficient number of students in both conditions, we will clearly balance the students who practice sports in both groups and the average causal effects between them will be comparable with respect to that variable. The crucial point is to assign people in such a way that balance is achieved, and the way to do this is through what we call randomized experiments.

2.3.4 Randomized experiments

In statistics, randomization means placing research units randomly across treatment groups. In the example of the student and the avocados, suppose we get a large enough number[10] of students to volunteer to participate in our experiment. If we want to do an experiment with randomization, we will make a list of the volunteers and assign them to one condition or another via a totally random procedure (for example, throwing a dice).

The idea of experiments with randomization is often associated with Fisher (R. A. Fisher 1935), who must have developed it during his 14 years as a senior scientist at the Rothamsted Experiment Station. On this farm, he had access to multiple data that had been stored for decades to help in the discovery of crops that were more productive and more resistant to diseases, and that needed fewer resources to grow. Since genetic variability causes each plant to have individual properties, rather than performing an experiment with individual

units, each experiment used a sufficiently large number of randomly selected units assigned to each condition. In this way, the influence of the individual variability of each plant and of other factors that could confound the effect of those being systematically studied (type of fertilizer, temperature, water, etc.), was reduced.

The design of experiments following Fisher's approach is at the core of much of the technical and scientific progress made in the last century. It's not that experiments weren't being conducted before Fisher: as I said earlier, their basic ideas are relatively intuitive and had been in use since ancient times. However, it was not until roughly 100 years ago that the fundamentals of designing and analyzing data from units that hold inherent variability (such as those found in agriculture, biology, medicine, sociology and psychology) were established.[11]

Fisher was convinced that an experiment that randomly assigned units to conditions was the only way to make causal inferences, and that studies that did not do this were not sufficiently conclusive. Therefore, when, towards the end of his life, he was involved in research to evaluate the effects of smoking on cancer, he rejected the "simple conclusion that the products of combustion reaching the surface of the bronchi induce, albeit after a long interval, the development of cancer" (Ronald A. Fisher 1958). In short, Fisher repeated the well-known statistician's adage that correlation is not causation and that just because smokers tended to develop lung cancer more often than non-smokers, there was no basis for claiming that smoking causes cancer. In order to affirm that claim, it was necessary to conduct an experiment in which a group of individuals were randomly assigned to be lifetime smokers and another group of individuals were assigned not to be, and then to follow both groups until the end of their lives to observe the consequences. Without doing this, Fisher claimed, it would be impossible to exclude a third

factor that might explain the correlation found in many studies. Fisher even hypothesized that a genetic trait may cause certain people to smoke and be more likely to develop lung cancer, and thus explain the apparent association between the two[12].

Fortunately, the evidence accumulated by epidemiologists, as well as a deeper understanding of the rules that govern causal inferences, managed to convince society of the causal relationship in spite of Fisher[13]. However, it was not until the 1970s that the way in which causal inferences can be drawn from non-experimental data was systematized more rigorously, as we will see below.

2.3.4.1 Experiments with irregular randomization

Despite the undeniable superiority of experiments with random assignment of units to conditions for making causal inferences, many researchers have no choice but to base their conclusions on data that have been collected differently. In the example of tobacco and cancer, for instance, the proposed experiment could clearly not be carried out for ethical reasons since it would mean imposing smoking on people who in principle would never have smoked of their own free will just for the sake of science. Studies that do not use random assignment to conditions are called *observational studies* and statisticians, in agreement with Fisher, deemed that their results could be interpreted only in terms of associations or correlations rather than in terms of causes and consequences.

However, in a series of articles published in the early 1970s, (Imbens and Rubin 2015; Rubin 2005, 2007), Rubin put forward the view that randomized experiments and observational studies were not in entirely different categories but differed only in what he called the *mechanism of assigning units to units terms*. This mechanism could be

ignorable in the case of experiments, which means that the way in which subjects are assigned to conditions enables causal inferences to be drawn. Without randomized assignment, however, the way in which research units end up falling into the experimental conditions is not ignorable since it can make units in one condition different from those in the other by reasons other than the factor to be studied. If this happened, we would find ourselves, following Rubin's terminology, with an *irregular assignation mechanism* that needed to be studied in each case to observe its effect on the causal conclusions. Examples of irregular assignment mechanisms are enabling the subjects to choose which group they are in, having different ages or genders in the groups, or allowing subjects to discontinue treatment before they have completed it based on a characteristic such as low motivation.

For example, suppose that those who volunteer to participate in the avocados and memory study can choose whether to be in the treatment group or the control group. If we allow this, we are opening the door for students with a certain profile to choose to follow the diet, and so those who choose not to follow it would have a different profile. Factors that occur to me that could tip the balance are the subjects' economic level (avocados are expensive and one per day may be over their budget), academic performance (subjects with lower grades may feel more attracted to miracle treatments), personality, intelligence, place of residence, preference for green foods, etc. Obviously, some of these factors are more credible than others. On the other hand, despite our suspicions, some of them may not arise. So it is the job of the statistician in charge of the analysis to verify them and introduce any necessary corrections.

One of Rubin's contributions for solving this problem involves identifying each person's *propensity* to receive the treatment based on their personal characteristics regardless of whether they have

received it or not. The comparison would then be made by matching people with similar propensities to receive the treatment, though they were finally allocated to different conditions. Remember how at the beginning of this chapter I mentioned that an intuitive way to see the causal effect of a treatment is to observe its effect on two units that are as similar as possible (twins, in the ideal situation). Rubin's procedure is a general solution for matching subjects who are as similar as possible to each other.

2.4 Consequences for the psychologist

Although the above is mainly a discussion of methodology, I believe it still has valuable lessons for applied psychologists.

Below I discuss two scenarios: a) the psychologist who works in programs that address people in groups, and b) the psychologist who works with individual clients (patients).

2.4.1 Evaluating groups

The first situation describes the psychologist who offers "treatments" to groups of people generally through talks, workshops, or retreats. Examples include talks on stress management to executives, gender violence to adolescents, and rehabilitation or prevention to addicts. During these group sessions, the psychologist presents information, leads a practical activity and, at the end of the activity, ascertains the participants' opinions and level of awareness or knowledge. This final part of the "treatment" *may be mandated by their employer or sponsor*

but if it is not, it is still a good idea to introduce it to check that everything is working properly.

Here are several suggestions in light of the above:

- If you wish to see the effect of your intervention on the participants, you must collect the responses of those who attended the course *and of those who didn't* so that you can compare. Collecting information only from those who have participated in the course would not provide interesting information. The ideal situation occurs when there are more people interested in participating than seats available and a draw must be made to choose those who finally participate. A comparison between participants and those on the waiting list is especially interesting because, in theory, the only substantial difference between the two groups is the treatment.

- If the participants are selected according to a particular characteristic (e.g. age, gender, motivation or salary) you would need to also collect this information from the group of non-participants and base your comparison mainly on those whose characteristics are similar to those who have participated on the course.

- If participation is voluntary, it is interesting to collect information from those who are not interested in participating. Again, a comparison of the results of your program should be made between those with similar characteristics and those who wished to participate but in the end decided not to.

- It is not good a good idea to make evaluations voluntary at the end of the course because those who provide feedback

may be outnumbered by those who benefitted least or most from the course, which may lead to opinions that are too pessimistic or too optimistic. If necessary, the profiles of those who do not answer and those who do could be compared in order to observe any differences.

2.4.2 Evaluating individuals

The psychologist is often visualized as a person sitting in a chair listening to patients and giving advice to them about their problems. Many professionals believe this type of activity cannot be evaluated objectively at all since it is so specific to each patient that no general conclusions can be drawn. It is like the student who tries to assess for himself whether avocados benefit memory by eating them for a year and then seeing if his grades improve or worsen: with so many uncontrolled factors and without valid comparisons to check his results against, his conclusions would fail to convince a moderately demanding critic.

Don't think that this problem is exclusive to psychology: read a little about the history of medicine and you will find that before the 20th century many doctors were developing therapies on their own and, to test whether they worked, they tested them on their patients to see what happened. Bryson 2019 provides some curious examples of this form of demonstrating the effects of therapies. This author also asserts that the less scrupulous doctors were very pleased to attribute their successes to themselves and their failures to their patients who "had not put enough effort into their recovery".

Modern doctors rarely act in this way. Nowadays, if a treatment has not been evaluated in what is known as a *randomized clinical trial*, the risks involved in applying it are so great that they rarely take the

chance. Moreover, although untested methods are sometimes essayed in special cases, to do so it is often necessary to obtain approval from special committees that oversee such cases. This does not mean that when using a well-known, already-tested method, doctors don't adapt it to their or their patient's characteristics if they consider it necessary. Moreover, their own experience, material resources or situation may encourage them to select method X over method Y *from among those that have shown acceptable efficacy in clinical trials*.

There is no room here to describe what a randomized clinical trial consists of. Moreover, the details of such trials are closely linked to the medical field and, more specifically, to drugs. However, studies of this type have also evaluated psychological therapies, and my advice to you is that your clinical practice, if this is the professional path you eventually take, should be guided as much as possible by them. It is true that, as the amount of resources used in these studies is far from those used in pharmacological studies, these studies may be limited in comparison. Nevertheless, their conclusions are more credible than those of a psychologist that attends to patients who arrive at their office one by one. Perhaps Psychology is not the same as other disciplines and the same criteria to which those disciplines are accustomed cannot yet be applied in our case. However, that does not mean that we cannot copy what has worked well in other disciplines to bootstrap ours.

---

1. Not receiving the treatment is usually called "control", though, in truth, belonging to the control group sometimes

means receiving some kind of treatment that is less effective than the one given to members of the treatment group or is even ineffective.↩

2. The topic seems complicated, but even children can understand it.↩

3. For more examples, visit spurious correlations.↩

4. The tendency to invent stories to causally connect two events is called narrative fallacy.↩

5. I'm sure you don't need me to tell you that much of this information is misleading and that if you want to learn more about it you should only find articles like those in scientific databases such as Solomon et al. (2002).↩

6. Well, just one now and then would be OK.↩

7. Like in *Groundhog Day*.↩

8. Don't think this is an exaggeration: there are many examples of diets, learning methods, and self-help books on the internet and in bookstores that are supported only by the fact that the treatment worked for the author, guru or pundit who then decided to set up in business as a consequence.↩

9. I have the impression that some people misunderstand scientific claims about health, education or psychology because they have not internalized the statistical solution: for example, some people refuse to take a drug because it is not guaranteed to have no negative side effect, despite the fact that the average effect in the population is known to be positive and there is no reason for it to be different in their case. On the other hand, the *scientific* solution, even when applied incorrectly, is more convincing and it is not

uncommon to hear "*my friend took it and it felt good so I'm going to take it too*." ↩

10. Large enough? Can't you be more specific? I'm afraid not. The rules for choosing sample size are rather long and this is not a good time to explain them.↩

11. Wikipedia provides this example from the Old Testament: King Nebuchadnezzar proposed that some Israelites should eat "a daily quantity of food and wine from the king's table." Although Daniel preferred a vegetarian diet, the official was concerned that the king "would see you as worse than other young men your age. Then the king would take my head off because of you." So Daniel proposed the following controlled experiment: "Test your servants for ten days. Give us nothing but vegetables to eat and water to drink. Then compare our appearance with that of the young men who eat the royal food and treat your servants according to what you see." (Daniel 1, 12–13)↩

12. As you may have already suspected, Fisher was a smoker. ↩

13. Fisher would be very unhappy to learn that, even though everyone recognizes his genius, he is always presented in anecdotes as an example of the idea that "even geniuses can be wrong".↩

4 The scientific method

At the heart of the novelist's craft lies an optimism which thinks that knowledge we gather from our everyday experience, if given proper form, can become valuable knowledge about reality.

Orhan Pamuk, The Naive and the Sentimental Novelist

Psychology is an empirical science like biology, economics, sociology and chemistry, etc., in contrast with the formal sciences (mainly Mathematics). In empirical sciences, theories must be contrasted with reality to see if they are correct. In formal science, this is not necessary: we don't need to check mathematical statements such as "two plus two equals four" against reality to be sure that they are true.

The scientific method is how empirical sciences contrast theories with reality to determine to what extent they are good descriptions of it.

In other words, in the empirical sciences there is a theory (in our imagination) that will fit better or worse with reality. The aim of the scientific method is to examine how good this fit is while bearing in mind that it is always possible to imagine a different theory that may fit worse, just as well, or better. Ultimately, in the empirical sciences theory and reality are two different entities. In the formal sciences, however, there is no such difference.

I am not an expert in Philosophy, so the above explanation, though it works for me, may not be as complete as it could be. After all, the scientific method has a long history. However, although I could go on to describe it without referring to the historical story behind it, I find that knowing something about it has helped me to understand it better. And, since we teachers tend to think that what has gone well

for us will go well for our students, I will provide some brief notes about it.

You may remember the name David Hume from previous years. Perhaps you know that some of his contributions concerned the problem of using induction as the method for obtaining knowledge about the world (another method is deduction, which is more relevant to formal sciences). Induction is what we do when we observe cases in reality and extract a generalization such as: since all the swans I have seen in my life are white, all swans are white. Induction, Hume said, is the way we learn things about the world and, although it generally works in everyday life, it has a fundamental flaw when it comes to producing general knowledge: there is never any absolute certainty that knowledge obtained from it is absolutely true. Following the previous example, *there is always a chance of encountering a black swan even though only white swans have been seen for many centuries*.[1]

The problem with induction as a way of obtaining knowledge is that the result it produces is fragile: at any moment, what we take for granted can disappear because reality provides us with a piece of new evidence that contradicts what had previously been accepted. Look at

it from your own point of view as a university student: what would be the point of studying for a degree if what you learn on it is so unstable that it can change radically shortly after finishing it?[2]

Fortunately, a couple of centuries later, another philosopher developed an idea that, in theory, would make induction-based knowledge more resilient and strengthen the foundations of empirical sciences. That philosopher was Karl Popper.

Popper[3] agreed with Hume that we cannot definitively prove something is true by using empirical observations or experiments. However, what we can prove instead is that something is true up to now – as long as something does not happen that contradicts it. However, the opposite, proving that something is false, is definitely possible. Therefore, what he proposed is that, instead of focusing on proving that something is true, we can strengthen our knowledge by rejecting things that could be in contradiction with what we know so far. Thus, the scientific method of empirical science begins with what is considered true at a time, i.e. current theory, and tests it to see whether we can find facts that contradict it. If the facts do not contradict it, then we can continue to believe in it. But if they do, it is time to develop new theories that take those contradictory facts into account.

Shadish, Cook and Campbell (Shadish et al. 2002) explain Popper's idea as follows:

The ruling out of alternative hypotheses is closely related to a falsificationist logic popularized by Popper (1959). Popper noted how hard it is to be sure that a general conclusion (e.g., all swans are white) is correct based on a limited set of observations (e.g., all the swans we've seen were white). After all, future observations may change (e.g., some day I may see a black swan). So confirmation is logically difficult. By contrast, observing a disconfirming instance (e.g., a black swan) is

sufficient, in Popper's view, to falsify the general conclusion that all swans are white. Accordingly, Popper urged scientists to try deliberately to falsify the conclusions they wish to draw rather than only to seek information corroborating them. Conclusions that withstand falsification are retained in scientific books or journals and treated as plausible until better evidence comes along.

The scientific method, therefore, is based on falsificationist logic. We will see how this method works in the next section.

4.1 The steps of the Scientific Method

To apply the "falsificationist" logic of the scientific method we need:

- A theory or general statement we wish to test.
- A hypothesis derived from this theory.
- A method for testing this hypothesis (measurement methods, study design, measurement tools, etc.).
- A hypothesis test, normally based on statistical analysis (the main focus of this course).
- A discussion of the credibility of the theory (after seeing the results).

We will discuss these steps one by one and then see how this structure is used in empirical scientific documents.

4.1.1 Theory

For the purposes of this course, a theory is a general statement or a statement that summarizes or generalizes a series of observations of reality. Be careful not to confuse theories with facts: a theory is not a fact but an explanation or summary of the facts. If a theory is good, we can use it not only to explain what we observe but also to predict what will happen in similar situations in the future. In reality, theories are not true or false in the strict sense but simply better or worse summaries of the facts used to predict what will happen under certain conditions.

A fairly common but quite dangerous mistake is to act as if the theory is above reality and, if it does not match reality, to deny that reality. As a psychologist, you will find that this problem occurs very often in certain people. By way of example, this article illustrates a theory held by some people which states that our minds have the power to cure our ailments. The article describes the agony suffered by the author's father, who had a wound that grew gangrenous, refused to receive medical treatment because he was so convinced of his theory, and died in excruciating pain as a result. If the father had known, as I hope you do, that theories can and must be reviewed if the facts contradict them, he should not have died in that way.

Where do facts in science come from? There are various sources: new fields of knowledge sometimes begin with casual observations of reality observed in the laboratory, as occurred with Pasteur, who observed that chickens became immunized after attenuated exposure to bacteria. Or in real life when, for example, the inhabitants of a region notice that certain schools produce the most successful students in a particular area. In other cases, facts are derived from earlier research conducted in a formal way, such as when psychologists endeavoring to identify predictors of mental illness explore a range of socioeconomic, family and personal, etc., backgrounds. Whether facts come from one source or another,

theories try to give them order in such a way that what appears to be a series of disconnected facts turns out to be a coherent set.

Another source of fact is people's experience. We all have our own perceptions about how things seem to work and we develop theories that stem from these perceptions. The advantage scientists have is that they collect facts systematically whereas when we act as individuals, we have access only to certain facts, which are often biased, and it is difficult for us to draw generalizable knowledge. An intermediate situation, however, is professional practice. When psychologists (or doctors) treat patients, they do so systematically to a certain extent, and since the situation is the same, they usually behave in a certain manner, using the same protocols, etc. However, since many factors in professional practice are not systematic, theories drawn from such conditions can be very limited. In early Psychology, several theories were developed from clinical practice that, although they still have their followers, are now seen as problematic. However, don't think that psychologists are the only professionals who have encountered this problem: the history of medicine has numerous examples of treatments that were based on this type of evidence but which today we find creepy [@ bryson2019body].

There are grand theories but there are also small ones. Psychologists are often associated with grand theories such as psychoanalysis, behaviorism, cognitivism, and neuroscience, etc. In practice, however, small theories are also very useful, e.g. limiting the number of hours of television viewing improves academic performance in children; reading at night helps you sleep well; exercising helps with mental tasks, etc. You may think that a good, comprehensive, theory is all that should be needed, since small theories can be drawn from grand ones, but in fact the grand theories are often not specific

enough, while stretching them leads nowhere. Therefore, both grand and small theories are needed.

Although some theories can provide details about internal mechanisms that explain how a cause is connected to an effect, this is not always the case. For example, a clear theory sometimes exists about why a certain drug should cure a certain disease and this theory can be used to set up an experiment for testing whether the drug actually works in practice. However, the theory may be based simply on the observation that people who live in a certain place where the medicine's active component is naturally present never get the disease[4]. In the latter case, perhaps there is no detailed explanation as to why the component works. However, it is still an interesting theory to test and, if the test is successful, the missing details can be sought in the future. This procedure is very common in both literature and psychological practice because, very often in Psychology, although the theory is quite weak we can still prove a relationship between two events.

At this point, I think it is useful to present an example of a theory we can use in our Statistics class. Let's look at the following statement:

- A family's socioeconomic level predicts the academic achievement of the subjects.

Although, as you can see, this is not a very complicated theory to understand, I think it requires a little reflection to grasp its meaning. For example, it does not assert that your family's socioeconomic level predicts your intelligence but your *academic achievement* – which we know is partly a consequence of intelligence but also of other causes. Similarly, the theory does not explain in detail why one thing leads to another, though it would be nice, from the point of view of Psychology, to know the details that explain this relationship. A natural second step, therefore, once we are convinced that the

general theory holds, is to try to understand the basis for such a relationship. For example, we could see whether the children of families with a high socio-economic level:

- feel greater psychological pressure to achieve good grades,
- have better resources at their disposal (such as their own room),
- do not need to work,
- take advantage of vacation periods to do academic activities, or
- combine several of the above situations as well as others.

Note that each situation above leads to different solutions. If the situation is one of psychological pressure, everything would fall on the families and the solution would be to teach them how to motivate their children to study. If the problem is financial, on the other hand, scholarship programs that allow those with fewer resources to continue studying would be best and it would unnecessary to convince families to put more pressure on their children. It is common to observe interventions based on psychological theories, or on details of those theories, that have not really been proven and that, despite all the money invested in them, have no results.

Since this is a subject on methodology, here is not the place to delve into detailed theories. We will therefore work with a more general approach. Remember that it is assumed that, if we have a theory at a particular time, it is because there is a sufficiently broad body of prior knowledge to support it and we believe the theory will be upheld empirically, i.e. in real examples. The aim of a study or scientific investigation is to verify with a real example that a general theory is confirmed. We will see this process in the next section.

4.1.2 Hypothesis

A hypothesis is a prediction drawn from a theory about the world. Hypotheses must be specific enough to be compared to reality. That is, they must be concrete examples of the theories, with names, places, and dates. Think of the theory as the spirit and the hypothesis as the flesh. Hypotheses refer to reality while theories only exist in the world of ideas.

Let's go back to the example of socioeconomic status and academic achievement. If we wish to establish an investigation in relation to this topic, we must do so with a concrete example and a concrete method, while always taking into account the resources we have available.

In our case, the available resource we are going to use is the General Social Survey (GSS), a 1972 study based on a series of surveys with a representative sample of Americans on all kinds of social, demographic, and personal issues. The data from the Survey are publicly available[5]. Also, fortunately, we have a smaller but fairly complete version from 1993 in SPSS format.

To formulate our hypothesis, we must first specifically define how we intend to evaluate the concepts contained in the theory in our example. Obviously, the simplest way would be to consider how to do it and then conduct the study by asking participants exactly what you want to know – assuming that they will know the answer, of course. In this case, however, since we start from already-collected data, we will need to adapt our hypotheses to them. Although this procedure does not match how investigations are supposed to be carried out, the truth is that collecting data correctly is expensive in time and money, and so it is interesting to try to adapt our hypotheses and test

them using available databases. In our case, the GSS provides a representative sample of the entire United States and uses interviewers who have been trained specifically for the purpose. Although the questions they ask are not exactly what we would like to ask, the opportunity the GSS offers us is worth taking advantage of.

The GSS data file provides several variables related to socioeconomic status and academic performance. Two of these variables seem particularly applicable to this problem. These are:

- Income: this variable summarizes family income in dollars in 1991 in four groups: 1 = $24,999 or less; 2 = $25,000 to $39,999; 3 = $40,000 to $59,999; and 4 = $60,000 or more.

- The highest year of school completed: this variable has values ranging from 9 to 20, with typical values of 12 (subjects who did not attend high school) and 16 (subjects who have obtained a university degree).

Since our theory suggests that these variables are related, it appears that we could use them to test our hypothesis. However, if we think a little harder, it seems logical to restrict our study to subjects who are old enough to have had the opportunity to reach their highest school level possible (say, those aged 24 and above) but not old enough to be independent of family income for some time (I have chosen 28 as this upper limit). The first restriction should be obvious: if someone is too young, they may not yet have reached their highest educational level simply because they are not old enough. We also need the second restriction because we want to use the socioeconomic level of the subjects' families, not their own, and so we exclude subjects who will probably be more or less financially independent.

Our hypothesis for the study is therefore: “Those surveyed between 24 and 28 years of age with a higher family income will have completed more school years than those with a lower family income”.

As you can see, the scope of this hypothesis is more restricted than that of our theory. While our theory could be applied to many different settings with different variables and different groups of people, in a specific study we almost always limit ourselves to specific circumstances, specific variables, and specific people. A study therefore does not serve to definitively accept or reject a theory but to provide examples in which the predictions of that theory are right and in this way increase support (or otherwise) for that theory.

4.1.2.1 The statistical hypothesis

The above hypothesis still lacks an element to make it manageable in practice. As you can imagine, even if our theory were correct in general, it would not mean that all subjects with a higher socio-economic level spend more years in the school system than those with a lower level. What it would mean is that, in *statistical terms*, they would spend more years. This statement needs a little explanation.

Relationship may at first seem a simple word to define but, in fact, it is not. Here is a possible interpretation of the meaning of the *relationship* between the variables considered in our case:

- All those in high family-income categories have studied for more years than all those in low family-income categories.

This is a very strong statement. It means that the relationship between these two variables is deterministic, i.e. if we know the values in one variable, we can automatically predict, or almost

predict, those in the other. Therefore, subjects in the $60,000+ group should have, say, a minimum of 16 years of schooling, while those with fewer years of schooling would never reach that income level. This is not realistic. Most of the time, relationships in the real world are stochastic[6], which means that they are not deterministic or perfect.

A more reasonable definition of our relationship would be:

- The mean number of school years of subjects whose families are in high-income categories will be higher than that of subjects whose families are in low-income categories.

Notice that we have introduced a statistical term into our definition: the mean. This is one reason why you need to know statistics to establish hypotheses. In this definition, not all subjects in a high-income family group will necessarily have more years in school than those in the low-income family group: although there are exceptions that do not follow the rule, we would still say that the relationship exists stochastically.

Another possibility is to use another statistic, such as the median:

- The median number of school years of subjects whose families are in high-income categories will be higher than that of subjects whose families are in low-income categories.

We could also think of other definitions based, for example, on quartiles, ranges, or something else. However, these other possibilities are rarely used in introductory courses on statistics, which base explanations of techniques mainly on means and correlations and the definition of relationship generally, but not exclusively, on those two statistics.

I hope the above explanation clarifies the difference between theory and hypothesis. As you can see, hypotheses link the theories to concrete examples. Theories are more abstract and can lead to different hypotheses that may apply to different cases or examples.

4.1.3 Hypothesis tests

Having seen earlier how a study hypothesis is defined in statistical terms, you probably now expect me to move on to describe how to test it. Before that, however, we ne.ed to take another step. If you go back and read the introduction to this chapter, you will recall that philosophers have argued that scientists shouldn't focus on testing hypotheses they believe are true but to test those that are supposedly false. This procedure can be rather confusing at first so it is worth spending time trying to understand it well. Let's see how it works.

To follow Popper's falsificationist approach, we need two hypotheses. The first one derives from the theory we consider to be true and is called **the study hypothesis**. The second one derives from denying the first hypothesis, so we believe it will not actually be confirmed. We will see examples of these two hypotheses later. Note that the second hypothesis could derive from a different theory from the one we believe in but if it doesn't, it doesn't matter too much[7]. The hypothesis that is in opposition to the study hypothesis is called the **null hypothesis**.

4.1.3.1 The study hypothesis

The study hypothesis[8] is the one based on the theory we believe is correct. We can set our study hypothesis in this case as follows:

- The average number of school years completed is higher for subjects in higher family-income groups than it is for subjects in lower family-income groups.

Now let's look at the null hypothesis.

4.1.3.2 The null hypothesis

The theory we believe is that there is a relationship between socioeconomic status and academic achievement. What could be another possible theory? Let's keep it simple: another theory could be that such a relationship does not exist.

We already have a hypothesis that derives from the theory we believe in. The null hypothesis should be drawn from the theory we do not believe in. This hypothesis could be:

- The average number of school years completed is the same for all individuals in the various income categories.

This is our null hypothesis and the aim of our study is to show that it **does not** hold true in our example[9].

Since both hypotheses cover the full range of possibilities, rejecting the null hypothesis means that the study hypothesis is the good one, which reinforces our confidence in the theory that supports our study hypothesis. Provided that the contrary is not observed in new studies, this enables us to assert that this is a valid theory.

One aspect that is often difficult to appreciate is that, since we are testing the null hypothesis, our outcome must be phrased in

reference to it. The correct way to report a hypothesis test is therefore to say whether the null hypothesis has been rejected. A result in line with what was expected in our study would therefore be:

- The null hypothesis stating equality in the average number of school years for subjects in different income categories is rejected.[10]

After we have set the hypotheses, our next step is to conduct the study to see which of the two hypotheses best fits the facts. To do so, we need to design a study, collect the data, and analyze them. These steps will be described below.

4.1.4 Study design

Now that we have the two hypotheses, it is time to test which one is more compatible with the data. To do so, we need to conduct a study or do research that provides such data.

How to design a study properly is a very broad subject that can easily fill the whole year of a course. Indeed, in Psychology it is common to offer a course on Research Methods that covers all the aspects needed to carry out a suitable study, i.e. sample size, participant selection, and experimental control, etc. In statistics courses such as this one, we usually use data that have already been collected and don't need to explain design-related issues. However, if you are interested in understanding how one was carried out, do not hesitate to ask.

Another key aspect of any study is defining how to measure the variables you wish to use in it. In Psychology, the course on Psychometrics will show you to how to do this correctly. Again, on this Statistics course, although we won't spend too much time on these issues, you are welcome to ask if you wish.

Another group of skills that are also useful in research is related to equipment. Some studies may use special instruments such as video cameras, electronic sensors, and eye-movement tracker devices, etc. Measurements taken with such fancy equipment may look more objective than those taken with simple equipment, but this is not necessarily true. Nevertheless, you may have the opportunity to learn how to use those devices in specific courses.

In summary, I'd like to mention that a Statistics course does not normally dwell on these topics for long. Since in this course we focus on statistical tests, many examples and exercises will generally be introduced without providing many details about how the study was conducted or whether it would have been possible to do it better.

4.1.4.1 Budget

Once you've designed a study, it is time to carry it out. Of course, if you don't have the resources to do that, you will need to pay for them. For this reason, acquiring funding is an extremely important aspect of conducting good research.

Books on experimental design don't often mention the practical aspects involved in implementing a study. However, budget matters a great deal. Most of a researcher's time is spent filling out grant applications and then managing research projects if they are successful.

The costs involved mean that the smart thing to do to take advantage of the resources that are more easily at your disposal. Very often, therefore, you don't design your study in the way that you would do ideally but in a way that makes things easy for you. So, you use the people you know, the organizations you are in contact with, the place where you work, etc. Much good research is conducted by being "opportunistic", i.e., by adapting your research goals to the means available around you. In this course, we will often use this approach as we will use datasets that have been collected for other purposes and think of theories and hypotheses that can be tested with them.

4.1.4.2 The GSS93

The example I am using for this chapter, the GSS93 dataset, is an example of opportunistic research. Of course, there are many things I won't be able to do as I would like with this dataset, so my research won't be as dazzling as I had hoped, but you have to learn to live with what you have. In any case, as you will learn over time, each research method usually has some limitation and, in most cases, only the convergence of results between studies conducted in different ways provides enough evidence to consider a problem satisfactorily resolved.

4.1.5 The results

After an investigation is complete, it is time to analyze the data it has produced. This step usually involves some form of statistical analysis, which is the main focus of this course. Below is a list of steps usually followed to carry out this statistical analysis:

- Select a suitable test for the data.
- Check the assumptions.
- Establish statistical hypotheses.
- Check the probability of the null hypothesis and estimate the effect.
- Interpret the results (including post hoc tests).

4.1.5.1 Selecting the statistical test

As we will see during this course, most of the difficulty boils down to selecting a suitable test for the type of data your study has produced and then applying that test. The proper test for a particular type of data depends primarily on the level of measurement of the variables you have used in your study and the role they play. The measurement levels we will take into account are:

- Categorical
- Ordinal
- Numeric (including interval and ratio scales)

The roles these variables can play are:

- Independent
- Dependent

In this course, **we focus mainly on the statistical analysis of two variables**, one of which is independent and the other is dependent. Both dependent and independent variables can be categorical,

ordinal, or numeric. By listing all the combinations of possible levels of measurement for these two variables (CC, CO, CN, OC, OO, ON, NC, NO, NN), we have at least nine possibilities, which would lead to nine different tests (plus some special cases). Since all these tests take some time to explain, an introductory Statistics course that discussed a test for each combination would take far too long. Moreover, some of these tests are considered advanced material and discussions of them are considered unsuitable before other tests have been fully understood. Introductory Statistics courses therefore usually focus on just a small number of tests that are fairly common and flexible and are therefore still valid in situations that are not exactly what they were designed for. The chapter on "Selecting the statistical test" in these course materials provides a fairly comprehensive list of these tests. However, to preview this topic, we will see a short example of how to select a suitable test for our study on income and academic achievement. Remember that the two variables in our study were family income and the number of school years completed. Let's see what types of variables these are:

- Income has four ordered categories: families in category 1 ($24,999 or less) earn less than those in category 2 ($25,000 to $39,999); and those in category 2 earn less than those in category 3, and so on. This variable is measured at the ordinal level and adopts the role of independent variable.

- The highest school year is a numerical variable. This is the dependent variable.

We need a suitable statistical technique to observe the relationship between these two variables. However, and here is the first difficulty, in a Statistics course such as this one, no specific statistical test is usually studied for the situation in which we have an ordinal independent variable and a numerical dependent variable. In this course, therefore, we will study a suitable technique for the situation

in which both variables are numerical, or in which both are ordinal, but we won't consider the combination of an ordinal and a numeric variable.

Does this mean that sometimes you won't be able to do any statistical analysis because you haven't been taught the proper technique for the variables you have? The answer is no because, as we will see below, there is a degree of flexibility in how we use the techniques you will learn on this course, so it will always be possible to perform an analysis that, though not ideal, is good enough in practice.

If the statistical technique for the type of variables we have in our data is not known, what we can do is treat one of the variables as if it were measured at a lower level than it was actually measured. So, ordinal categorical variables can be treated as categorical variables, or numeric variables can be treated as ordinal variables. In our case, for example, the income variable can be treated as a categorical variable so that we can apply the technique known as analysis of variance. This is one of the statistical tests normally included in an introductory course on Statistics. Another acceptable solution is to treat the years in school as an ordinal variable and apply an ordinal correlation between the two variables.

4.1.5.2 Checking the assumptions

The statistical tests we will see on this course usually assume that the data meet some statistical assumptions or conditions. Common assumptions in many cases are therefore that: the population from which the data are extracted follows the normal distribution; there is equality of variances in the groups analyzed; or the number of cases in the sample is large enough, etc. As we will see, these assumptions must be checked before the statistical tests are performed.

Books on statistics used to assert that if the assumptions were not met, the analysis had to be stopped without any conclusion being reached. This is a very serious situation, for example, for a PhD student who has been collecting data for a year and wants to finish their PhD thesis as soon as possible!

My opinion, however, is that such a radical approach is not necessary and that in general we can take alternative routes rather than discard data when assumptions are not met. Some of these routes will be discussed in the chapter on the steps involved in a statistical test. However, since statisticians know of even more alternative paths, you can always seek expert help in the future if you have any queries about this type of problem.

Below will see how to establish statistical hypotheses.

4.1.5.3 Establishing statistical hypotheses

Earlier we had a section on establishing hypotheses; now we have a section on statistical hypotheses. You may be wondering what the difference is between these two types of hypotheses – and for good reason because they may seem very similar and you won't actually find this distinction in statistical manuals. However, my experience in supervising the theses of seniors has led me to believe that it may be useful to make this distinction.

The process involved in initiating the research that will end with an undergraduate thesis generally begins with reading and reflection on the problem of interest to the student. Once this is done, it is time for a meeting to see how the hypotheses the student has prepared can be combined with a statistical test[11]. This process is difficult to summarize in a textbook or teach on a short course. In reality, even

seasoned researchers sometimes turn to statistical experts for guidance, in areas outside their comfort zone, on how to select a correct statistical test.

In fact, the statistics toolbox has many options and you will always be able to find the best test for a specific objective. However, this "best" test may be difficult to understand and interpret, or software may be unavailable for your application. For this reason, it is often best to configure the hypotheses in such a way that the student can use statistical tests that are taught in introductory courses on Statistics. Meetings with students serve that purpose.

In our example on income and academic achievement, we mentioned that Analysis of Variance and ordinal correlations may be suitable tests. For Analysis of Variance, the null hypothesis would be:

H0:μ1=μ2=μ3=μ4H_{0}: \mu_{1} = \mu_{2}= \mu_{3}=\mu_{4}

The null hypothesis is traditionally represented by H0H_ {0}, which stands for "all means are equal". Note that $ mu $ signifies the mean number of school years of the *population*[12] of people in each family income group, which, in this case, are 1 = $24,999 or less; 2 = $25,000 to $39,999; 3 = $40,000 to $59,999; and 4 = $60,000 or more.

The study hypothesis is traditionally represented by H1H_ {1}, which stands for "at least one mean is different from the others"[13].

If you remember our hypotheses, you will see that the hypotheses of the Analysis of Variance do not match exactly. Our null and study hypotheses were therefore:

H0:μ1≥μ2≥μ3≥μ4H_{0}: \mu_{1} \geq \mu_{2} \geq \mu_{3} \geq \mu_{4}

H1:μ1<μ2<μ3<μ4H_{1}: \mu_{1} < \mu_{2} < \mu_{3} < \mu_{4}

The null hypothesis states that "the mean of years studied is greater or equal for each group than for the groups with a lower income level than that group", and the study hypothesis states that "the mean number of years studied at each income level is higher than for the groups with an income level lower than that level". Therefore, the average number of years of study of the low-income group will be lower than the average number of years of study of the intermediate-income group, which, in turn, will be lower than that of the high-income group. As you can see, while our original hypotheses establish a decreasing order in the number of years studied, when we use the standard set of hypotheses of the analysis of variance we can only test whether the means are different.[14] As the statistical hypothesis does not match the study hypothesis perfectly, we must be careful how we interpret the results.

Does this mean that there are no suitable statistical methods for testing our hypotheses? Not at all. As I mentioned earlier, there are many more statistical methods than it would be reasonable to include in an introductory course on Statistics, so we study only a selection of them. However, those we include are not only applicable to many situations but also to other, quite different, situations and produce similar results to those of the most advanced techniques, which we won't cover on this course. Of course, I will avoid setting exams or exercises that cannot be solved using the techniques you learn on this course. When writing your report, or writing reports for other subjects, however, you may come across situations that can be difficult to judge. If this happens, feel free to ask how to proceed.

I mentioned earlier that another possible strategy for testing the relationship between these two variables is to use an ordinal correlation. The disadvantage of this option, however, is that it considers the dependent variable as ordinal rather than continuous. If we proceed in this way, our statistical hypotheses will therefore be:

H0:ρ≤0 H_{0}: \rho \leq 0

H1:ρ>0 H_{1}: \rho >0

The Greek letter ρ\rho stands for the Spearman correlation. The H0H_{0} above means “the correlation is zero or below zero” whereas the H1H_{1} means “the correlation is above zero”. You can see that this formulation looks much simpler than that for Analysis of Variance. In some cases, therefore, I would recommend that you use this approach depending on your interest and sophistication in matters of Statistics. However, if I were performing this analysis myself, I would choose Anova for this situation since it enables the problem to be investigated more thoroughly.

In summary, every statistical test is useful for testing certain types of hypotheses. Deciding which is the right statistical test for your problem involves analyzing whether the statistical hypotheses associated with that test are suitable. Although sometimes there is a perfect match, in other situations there is more than one option and you have to work out which one is the easiest or simplest. In this course, exercises and exams are set so that choosing the right statistical test should be relatively straightforward. However, if you conduct a research project, no matter how small it may seem, this choice can be more problematic and you may need assistance from somebody with experience.

Now that we have set our hypotheses, it is time to see how close they are to reality, i.e., to the results of our study or research. As we will see, these results are key to deciding which hypothesis (null or study) is the most plausible.

4.1.5.4 Performing the analyses

After we have conducted a study and gathered our data, it is time to look at our results. Data can be explored or summarized in many different ways. However, if we are testing hypotheses, introductory Statistics courses usually narrow down the options to just a few (as we will see later). In our example, we have already taken this decision in our statistical hypotheses, since we have chosen the averages of the groups as the relevant value to look at.

The table below shows the results of the average number of years of education across the various levels of family income:

Income level	**School years**
Low	13.29
Intermediate	13.84
High	14.37
Very high	14.64
Total	13.76

From the above table we see that the average number of years in the education system increases by about half a point for each income category. This means that the average effect of being one step higher in income is roughly half a year of education. We can be more specific and calculate the effect size of the difference between, for example, being in the low-income category and being in the intermediate-income category by subtracting one from the other: 13.84-13.29 = 0.55 years difference. Notice, however, that, as this value is calculated

in a sample rather than the population, there is an uncertainty associated with it, so we can calculate intervals for where the value of the population could be. Since learning to calculate a confidence interval for the value of a population, given the result obtained in a sample, *is a central aspect of an Inference Statistics course, we will see how to do that later*.

Meanwhile, I can inform you that the 95% confidence interval for this difference ranges from -1.92 to 0.80. Since this interval includes the value zero, we conclude that we cannot rule out that the effect is null. That is, although in our sample the difference in years of study between these two income levels is half a year, this difference is not significant enough to assert that the difference does exist in the population as a whole. Sample-based estimates are subject to error – or difference – with respect to the population values. This error is called the sampling error and calculating it is also a component of this course.

The sampling error allows us to calculate the level of significance, which is one of the concepts most used to interpret the results of the statistical tests we will perform on this course.

4.1.5.5 The p value (or significance value)

Introductory math courses often include topics on probability. This concept is applicable to games of chance such as flipping a coin, throwing dice, or playing roulette. However, in statistics, we use probability to analyze real-life problems and treat them as if they follow the rules of games of chance.

Just to remind you, a probability is a number between 0 and 1 that expresses our confidence that something is true or will happen,

where 1 means absolute confidence and 0 means total lack of confidence. However, true or false are not the only options. Intermediate values are also possible. For example, 0.5 means that we are undecided[15].

In statistics, the p-value, or significance value, is a concept expressed in terms of probabilities. It is defined as follows:

The p-value is the probability of obtaining the results that have occurred in a sample if the null hypothesis (for the entire population) is true.

In our example, the significance value is the probability that the values in 4.1 occur if the null hypothesis (that all the averages of numbers of years studied for each level of family income are the same) is true (*for the population*).

Let's see how this works: if the means obtained in the study are fairly similar to each other, the p-value will be close to 1, which indicates that the null hypothesis is correct (i.e. there are no differences between the average number of school years completed across levels of family income).

On the other hand, if the average number of years of study were very different between income groups, the probability would be close to zero and *we would reject the null hypothesis*. In this case, we would conclude that, given the result in the sample, there must be differences *in the population* between the average number of school years completed across levels of family income.

Remember that the null hypothesis is the one we believe is not true, so we usually want the significance value to be as close to zero as possible.

Each statistical test uses a specific method to calculate the significance value. Sometimes this calculation is easy and can be done by hand but, most of the time, we need some form of assistance. In the past, this assistance was provided by tables of significance values contained in books on statistics. Nowadays, however, we use computers.

Continuing with our example, and omitting all details of the calculations involved, I used Analysis of Variance to test the null hypothesis that all the averages of the numbers of years of study for subjects in the various economic levels are equal, and obtained the result: pp = 0.094 (pp is the symbol commonly used for p-value or significance). This significance seems quite low, so it appears that the logical thing is to reject the null hypothesis. However, before we can reach this decision, we need a new concept. This is the level of significance we wish to adopt when taking our decision.

4.1.5.6 The level of significance

I previously stated that the null hypothesis is rejected when the significance value is close to zero. But how close is close? 0.01? 0.001? or 0.0000001? In other words, we need a specific cut-off point to decide when a significance value is low enough to reject the null hypothesis. The level of significance is the name given to that cut-off point.

In our example, you may have concluded that if the significance value is 0.094, it is time to claim victory and reject the null hypothesis. Since you may think this value is quite low, you say “since it is quite impossible to get this result if the null hypothesis is true, we can reject it.”

However, taking the decision to reject the null hypothesis when the significance value is 0.094 means that, although we would have very strong reasons for rejecting the null hypothesis, our confidence is not absolute. So, even if this result points in the direction we want (that of rejecting the null hypothesis), we are not completely sure of our decision. To be absolutely sure that we are correct in rejecting the null hypothesis, we should have obtained a significance level of 0.

This is because we want to be absolutely sure of every decision we make, don't we? Isn't science about absolute truths? And isn't it therefore unacceptable for science to give us answers we cannot totally trust?

Well, actually, no. Science is not in the business of providing absolute truths.

In fact, if we wished to use significance levels that are equal to zero, we would never reject the null hypothesis because (and I will save you the mathematical proof), this is not possible. It is true that we can obtain very low significance values that are close to zero. But exactly zero? No way! We therefore need to set a limit to significance because it is impossible to reject a null hypothesis with absolute certainty. We call this limit the level of significance.

The most common level of significance used is 0.05. The rule is: if the p-value is less than 0.05, we reject the null hypothesis. A second option is 0.01, which is preferable with large sample sizes or for special reasons[16]. On very rare occasions, we see other levels of significance but the above two are sufficient for a course such as ours.

In our example, since we obtained a probability of 0.094, and since 0.094 is greater than 0.05, we say that the differences between the means shown in the one for means are not large enough for us to reject the null hypothesis,. In our case, *we cannot therefore reject the*

null hypothesis regarding equality in the number of years of study completed by subjects in the different levels of family income.

However, if we don't reject the null hypothesis, does that mean we have to accept a null hypothesis that supports a theory we don't believe in? Well, this is a fairly common confusion. Without explaining in too much detail, non-rejection of the null hypothesis does not provide enough information to determine whether the null hypothesis is true, and so it should not be interpreted in that way. For this reason, *the result of a hypothesis test is to reject or fail to reject the null hypothesis, never to accept the null hypothesis*. In our case, we see that the result of our study does not allow us to reject the null hypothesis and what we must do now is find a justification for what has happened: Was the study wrong? Was our sample not correct? Is this a special case where the general theory is not applicable? The last step in the scientific method is called the Discussion. This is where we ponder the conclusions we can gather from our study. We will see this step in the next section.

4.1.6 The discussion

As this has been quite a long chapter, let me remind you about the steps of the scientific method. We begin with a theory, plan our specific hypotheses and methodology, and then conduct the study. This produces data that enable us to determine which hypothesis is the most compatible with them. As you will recall, we exemplified this process with data about numbers of school years and income levels before our results led us to *not* reject our null hypothesis.

Obviously, this is a simplified version of the steps taken in a real study since there is usually more than one analysis and the results often

produce conflicting messages that sometimes back up our theories but sometimes do not. Also, the results apply to the specific study, with a specific sample, and in specific conditions (measurement instruments, etc.). Drawing conclusions from all this information is not simple, which is why we need the final part of the scientific method, i.e. the Discussion.

The Discussion is where you compare your starting point with your results and reach one or more conclusions. Bear in mind that what is expected here is not a simple "yes" or "no" about your results. Here, you are expected to speculate on whether the evidence supports your theory, consider any limitations of your method, discuss other ways in which the study could have been carried out, consider whether it would be worth spending more money on it, and comment on future research on the topic, etc.

A fairly common mistake is to confuse the previous section (Results) with the Discussion. Although it is true that in the Results section you provide some sort of conclusion (whether the hypotheses are rejected), this should be fairly concise and technical. It is in the Discussion section that you elaborate on your results.

Let's continue with our example in which we fail to reject the null hypothesis. Our conclusion should be to reject our theory, right?

Not so fast!

The idea that a single result can lead us to disprove a theory is too dramatic; just as a single result cannot lead us certify that a theory is true. In the words of Shadish, Cook and Campbell (p. 28) (Shadish et al. 2002).

As theories are often not fully specified down to the last detail, there can always be some conditions, subjects, methods, measurements, samples, etc. that can provide seemingly dissonant results that with some effort

the researcher can accommodate with the theory. … Up to a point: if negative results are what you get most of the time … it's time to start over and think of a new theory.

Of course, it may be that our results are in line with what we expected. In this case, a discussion could, for example, point to future studies that would increase the scope of the theory, clarify some of its aspects, or attempt to respond to possible criticisms.

In fact, since studies are based on a series of assumptions, methods, and measurements, etc., there is always the possibility that weaknesses will be found in it. Criticism is part of the scientific process and blindly accepting a result as final is not recommended. Therefore, empirical evidence and the theories themselves are usually constructed in a process that uses the results of several studies to adjust the theory until a substantial body of theory and results is accumulated.

As an example of criticism that could be applied in our case, remember that we used the number of years of study as a measure of academic performance. However, this may not be a good measure of performance because some students may study until the last possible year with mediocre grades, while others may study for fewer years but get good enough grades to obtain a well-paid job. Studying for many years may not actually be a sign of good academic performance, so our study could be criticized from this perspective. To solve this problem and avoid this criticism, other researchers who would like to do better than us may collect data on student grades (though those researchers may be subjected to other criticisms for different reasons).

Another problem with using the number of years spent in the education system as an indicator of academic performance is that we had to select respondents who were older than a certain age, though

not by much because income for older subjects could be caused by their own academic level rather than the other way around (i.e. the academic level causing the income). Because of this selection, the sample of subjects we used for our study was fairly small and the effect we wished to find may have been too weak to appear. In conclusion, since our research design was not sound enough, the study does not provide sufficient reasons to abandon our theory – at least for now[17].

4.2 Introduction, method, results and discussion (IMRAD)

The scheme I have used to present the scientific method is close to what is called IMRD[18]. This is the basic structure used in scientific-empirical documents for sciences such as biology, psychology and chemistry[19]. The sections of a document used to follow this scheme are:

- Introduction: In this section the state of the art is established, a review of the literature is conducted, gaps in prior knowledge or possible new applications are identified, the method for conducting the analysis is justified, and the study objectives and hypotheses are established. In summary, this section comprises the theoretical component of a study. The hypothesis to be tested is also often mentioned at the end of this section – but not in great detail.

- Method: In this section the authors describe what they did in the study and what steps they took. This section can also be called Materials and Methods, Procedure, Experiments, or

Methodology. In Psychology, within the Methods section it is common to see the following sub-sections: Participants or Subjects, Description of the Sample, Measurements, Design, and Materials.

- Results: This section shows the statistical analyses of the data collected in the study. This is where statistical hypotheses are established and put to the test.

- Discussion: This final part of an article compares the initial theory or state of the art introduced at the beginning with the results obtained. If the results are what you expected, it's time to think of ways to expand or improve your current research. Hints can also be given on applications of the results. There is also usually a comment on limitations, where the authors point out any problems with their own study and suggest ways to tackle them.

In many scientific and academic papers, as well as in student essays and Bachelor's or Master's theses, you will find the four sections above. The Introduction is based on the subject of the document: in Psychology this could be social psychology, basic psychology, or psychobiology, etc. The Methods section is related to subjects taught in the Methodology Department, i.e. research design and psychometrics. The Results section is the statistics component. Finally, the Discussion is where you can show your ingenuity and intelligence, explain your results, and describe how the world can be a better place by using them.

1. When, in the 17th century and after many years of believing that swans were always white, black swans were found by Europeans in Australia, the term black swan began to be used to refer to the occurrence of unexpected events.↩

2. The fragility of established knowledge does not only have negative consequences: in a famous best-seller, Taleb (2007) explains that Hume inspired him to successfully invest in the stock market by taking advantage of black swans in the economy.↩

3. Popper lived in the 20th century and Hume died in the eighteenth century. Therefore, it took two centuries to arrive at a more or less accepted philosophical solution to the problem of induction.↩

4. From personal experience, I know that psoriasis on the skin has been treated with tar for many years, but I believe the mechanism by which it works is still not very well understood.↩

5. http://gss.norc.org/↩

6. Probabilistic or statistical.↩

7. Although sometimes we have two different theories that generate conflicting hypotheses and our result will enable us to decide which theory is better, our study often focuses only on testing whether a specific theory works or does not work.↩

8. I use the term “study hypothesis” because to me it seems closer to the real meaning. However, in most places it is called an “alternative hypothesis”.↩

9. As we will see, there are certain exceptions to this rule. In rare cases, we may think that the null hypothesis is correct and the study/alternative hypothesis is wrong. For example, if we were considering how the sign of the zodiac affects income, most of us would believe that this effect does not exist, so people with different signs would earn roughly the same as the average salary. In this case, the null hypothesis would be correct.↩

10. Note that the phrase says *reject equality*, which means that the average number of school years *is different* for those with different incomes.↩

11. Provided that the degree assignment is a research study (it could be a different type of work).↩

12. It is population because the hypotheses do not refer to our sample of particular cases but to the population the sample is supposed to represent, i.e. people between 24 and 28 years old in the United States in 1993.↩

13. Since expressing this in formulas can be quite difficult, this hypothesis is usually expressed in words↩

14. We could perform what is called a planned comparisons test. This type of analysis of variance may not be covered in introductory courses on statistics.↩

15. Probability understood as a number has a very short history. It was not until the 18th century that expressing chance as a fraction of a total showed up for the first time in the writings of mathematicians. Thinking in such terms sems not to come naturally, which may explain why people often struggle with it is application to real-life problems.↩

16. However, so as not to complicate my students' lives more than they already are, the practical exercises on this course will always use 0.05 as the level of significance.↩

17. You may be surprised that in our example, which, after all, was made up, I didn't find what I was supposed to find. However, if I had used an example in which the null hypothesis was rejected and my theory was "supported" by the data, I would not have been able to discuss the case where this did not occur. Very often, textbooks just show the ideal situation in which everything happens as it should, leaving students with the impression that if things don't work out that way, it must be because they did something wrong.↩

18. Or, sometimes, IMRAD.↩

19. In some non-empirical disciplines such as Mathematics and Philosophy, however, this scheme is not so dominant.↩

5 Selecting the statistical test

Articulated by J.W. Tukey (Tukey 1977), the distinction between the confirmatory and the exploratory approaches to data analysis is extremely important in statistics. Exploratory analysis is analogous to the work of a detective: it provides a general idea of the mystery, sets out the clues that enable the mystery to be solved, and suggests possible ways of organizing the enigma. Confirmatory analysis, on the other hand, is the part of the trial that tries to find evidence that is as conclusive as possible. While exploratory analysis is conducted in an open way, with graphs, quick summaries of data, and a broad perspective, confirmatory analysis focuses much more on specific aspects and looks for evidence that corroborates hypotheses derived from theories. The first part of introductory statistics courses usually focuses on the exploratory part, while the second usually focuses on the confirmatory part, i.e., making judgments as to whether a theory fits or does not fit satisfactorily with the data.

Confirmatory data analysis mainly involves:

- identifying a specific hypothesis to be tested,
- finding the correct statistical test for the hypothesis, and
- applying it and interpreting the result.

Confirmatory data analysis is more rigid than exploratory analysis and comprises a series of steps that must be followed in a certain way. In this chapter we will focus on finding the correct test for the hypothesis. In the next chapter we will describe the steps that are common to the various tests.

To determine which statistical technique is to be used, you must know certain selection rules. Since these rules are based mainly on the type of variables to be analyzed, I will first describe these variables and then list the tests from which we will choose the most suitable one for each situation.

5.1 Types of variables

The simplest situation in confirmatory analysis is to test the effect of one variable on another. We therefore have two variables to take into account. More advanced statistical tests may include three or even more variables – though I warn you that going beyond four or five variables becomes rather difficult to manage. In this course we will primarily consider tests with only two variables.

Two characteristics of the variables need to be taken into account when selecting the correct test[1]:

- the level of measurement of each variable, and
- the role each variable plays in the analysis.

We will discuss these two characteristics separately.

5.1.1 The level of measurement

When measuring a characteristic of something, we can achieve different levels of quality of measure depending on how we perform the measurement.

Note that the characteristic measured and the measure of this characteristic are two different things. Note also that the same

characteristic can be measured with different methods and that these methods can result in different levels of measurement. For example, the temperature of three objects can be measured by touching them with our hand or by using a thermometer. In the first case, the measurement will report the order (i.e., which object is the warmest, which one is the second warmest, and which one is the third), whereas the thermometer will provide a value on a numerical scale.

We will distinguish between three types of variables and three levels of measurement:

- Categorical variables.
- Ordinal variables, which include:
  - ordered categorical variables, and
  - range variables.
- Numerical variables.

The best way to explain these three levels is by using an example. Let us now consider the variable *academic achievement at university* and the various ways in which we could measure it.

5.1.1.1 Categorical variables

- Type of degree: {Engineer, Licentiate}.

As you may know, some university faculties produce Engineers while others produce Licentiates. Engineers are associated with technical degrees, whereas Licentiates are associated with studies in Psychology, Medicine, Law, and Biology, etc. Having one title or the other may not be so consequential though some people might claim

that it is better to get one title or the other. In any case, the variable *type of degree* with categories {Engineer, Licentiate} is an example of a binary or dichotomous variable, i.e. a categorical variable with only two possible values. Other categorical variables may have more than two categories: for example, the title with all the possible degree subjects as categories: {Psychology, Medicine, Economics, Chemistry, etc.}. Note that measuring with categorical variables is equivalent to classifying objects or subjects within categories. In theory, categorical variables are the easiest to collect since they involve simply observing whether someone or something apparently has a certain characteristic. Some studies are therefore based on categorical data and a strategically placed observer who records what they see. However, analyzing categorical data, especially when we have more than two categorical variables and want to see their inter-relationship, can be quite complicated. It is therefore advisable to propose studies based, as far as possible, on variables measured on other scales, especially numerical ones.

Note that taking into account further characteristics of the measured objects and then the level of measurement may be a different approach from what we originally considered. For example, we could order the degrees according to the grades required to enter, which would make the level of measurement ordinal. Another, more imaginative, method could be to assign numbers to the degrees so that they can be treated as numerical variables.

In theory, categorical variables are the easiest to measure. Very often, an observer can judge quickly whether an object falls into a particular category. For this reason, many studies in psychology are based on categorical data that simply use an observer who records what they observe. However, *analyzing* categorical data, especially when we have more than two categorical variables, can be quite complicated.

It is therefore advisable to propose studies based, as far as possible, on variables measured on other scales, especially numerical ones.

An important practical difference in statistical analysis is whether the variable has two categories or more than two. At first glance, it may appear that there is no fundamental difference between these two situations. In fact, many statistical tests have two versions: one that I call the "short" version and one that I call the "long" version. The short version capitalizes on the fact that variables with two categories are easier to analyze and interpret than variables with more than two categories. This difference will also be used when selecting the right statistical test.

5.1.1.2 Ordinal variable

Ordinal variables can show up in two different ways: as categories that are ordered or as categories that are ranked.

- Achievement as {Excellent, Good, Pass, Fail}.

The values of this variable are categories but there is also a rank, or order, in them. If the measured variable is academic performance in a subject, each category indicates a better performance than those below it. One way to think about this type of variable is to consider that a numerical variable underlies the ordered categories (in this case, academic performance or grades) that is divided into segments. For example, it is clear that someone with Excellent has a better grade than someone with Good, or Pass (and, of course, Fail).

One characteristic of these categories is that we are unsure whether the difference between two consecutive categories is equal to the difference between two other categories. For example, in Spain, academic grades usually go from 0 to 10: Fail goes from 0 to 5; Pass

goes from 5 to 7; Good goes from 7 to 9; and Excellent goes from 9 to 10. In this case, thanks to the extra knowledge of what each category means in terms of numbers, we can make a reasonable estimate of the numerical value obtained by the subjects. If we have a list of grades that uses the categories but not the numbers, we could use the intermediate values on the numerical scale as follows: {Excellent = 9.5, Good = 8, Pass = 6, Fail = 2.5}. From these intermediate values we can obtain an estimate of the difference between the categories (Pass-Fail = 6-2.5 = 3.5; Good-Pass = 8-6 = 2; Excellent-Good = 9.5-8 = 1.5), which makes it easy to see than the distance between each of those categories. Note that *we see the difference because we have extra knowledge of the numerical values behind those categories*. If this were not the case, it would have been impossible to estimate the distance between them. For example, if a film critic rates films using the Fail, Pass, Good, Excellent scale, it would be impossible for us to tell whether the difference between Good and Excellent is the same, twice as much, or half as much as that between Pass and Excellent.

Note that there is a great deal of similarity between categorical variables and ordered categorical variables. In fact, determining what type of variable you have can sometimes be quite challenging. Also, although we know there is some order in the categories, we may ignore it and use the variables as if they were just categorical because then we can use a statistical test or plot that we like more than the one we would use for ordinal variables. Downgrading the level of measurement of a variable is technically acceptable and may be convenient in some cases. The opposite, i.e. using the ordinal variable as if it were numerical, is also done in practice but there is much concern about whether it is appropriate to do so.

An important case of ordered categorical variables are the so-called Likert scales. These derive from questions on questionnaires or surveys that ask for agreement/disagreement – on a scale of 1 to 5 or

1 to 7. Sometimes, however, other values are used for questions such as: “Do you feel happy today?”, or “Do you agree with the government’s actions in relation to the environment?”, etc. Typically, the lowest value would mean “No, not at all” or “Never” and the highest value would mean “Yes, absolutely” or “Always”. Of course, the middle value would mean being in the middle. Note that, in Likert scales, there is usually an odd number of values so that respondents have the choice of giving a value that is exactly in the middle.

In Psychology it is common for Likert-type questions to be part of a questionnaire that has several questions referring to the same concept. This is how questionnaires that measure aspects of people’s personality and attitudes, etc., are constructed. Responses to individual questions on a questionnaire are combined for each person (using a simple or a weighted sum) to provide an overall score for each person that measures their tendency to anxiety, their level of extraversion/introversion, or their attitudes towards certain groups of people.

- Grades as a position with respect to the other students {First, Second, Third, etc.}.

Let’s say you are the best student in the class: you are number one. The next student would be number two, and the next would be number three, and so on. There may be two students with exactly the same score. This situation, which is called a tie, can be resolved by giving both students an intermediate position – for example 5.5 for those tied in fifth place.

Another version would be to count the percentage of students below a certain student and report it. This is a percentile: the best student would get percentile 100 and the worst would get percentile 0.

The results of psychological questionnaires are often given in the form of percentiles since this scale is easy to understand and we don't need to know all the details of the questionnaire to interpret the result. For example, one questionnaire on Extroversion may have 50 questions that are answered on a Likert scale from 1 to 5 while another questionnaire has 70 questions answered on a scale from 1 to 7. If one person has a score of 150 on the first questionnaire and another has a score of 200 on the second, it is difficult to know whether they have similar levels of Extroversion. On the other hand, if both have a similar percentile score (say 50) in the respective questionnaires, we know that they are similar in this aspect.

In both cases we are speaking of a rank variable. Although rank variables convey some information about distance, this may not be as straightforward as we might hope. For example, student number one is 10 and student number two is 6: clearly, student number one is well above student number two but percentiles do not convey this information.

Rank and percentile variables have different properties. If we focus on the percentage of cases above or below, or at a certain position, and we do not take into account the criteria we used to make that order, we can see that the distance between two ranges has a meaning that is constant. For example, the distance between the first and the second is a position, and the distance between the second and the third is also a position. In terms of percentiles, a position means the same thing: for example, if there are a thousand people in our sample, each percentile point means ten people, so the person with the 100th percentile is above 10 more people than the one with the 99th percentile, and this person is above 10 more people the person with the 98th percentile.[2]

However, if the ranks derive from ordering the subjects based on a numerical (implicit) variable, then a rank conveys ordinal

information about that variable. Let's consider the result of a race in which the first runner arrives half an hour earlier than the second, and the third arrives just one minute after the second. From the point of view of time spent, the difference between the first and the second is much greater than the difference between the second and the third, which also suggests a great difference in terms of their running ability. In this case, the ranges simply provide information of order *relative to the original variable, i.e. race time.*

Very often, you will find that the variables of ranges or percentiles are treated as numerical variables, which corresponds to a measurement level that is higher than ordinal. Those who rigidly follow the classification in measurement levels proposed by Stevens often frown when they see this. However, other authors, myself included, believe that Steven's classification does not perfectly fit every situation.

5.1.1.3 Numerical variables

- Grades as numerical values [0–10}.

Students' grades in Spain usually range between 0 and 10. Therefore, just by calculating the difference score between the grades of two students, we know the distance between them. Note that there are two types of numerical values: integers, if they are rounded to the nearest integer; or continuous, if several decimal values are (or can be) reported. This may seem a minor difference but statistical plots for integers are sometimes much less useful than they are for continuous values. In general we prefer to measure variables so that they are numerical. Often, however, this is not possible.[3]

In any case, although it is important to think carefully about the levels of measurement of our variables, a typical course on Statistics usually avoids, as far as possible, going into this type of detail. It will therefore not be necessary for us to thoroughly investigate the true meaning of the measures but simply to observe whether we are dealing with categories, whether they are ordered, whether the data are ranges or positions and, finally, whether they are numbers. Although this is certainly not the be all and end all of the subject in terms of types or levels of measurement of variables, it will suffice for now.

5.1.2 Variable types in practice

As we will see later, the most useful types of variables in this course are categorical and numerical, since many of the statistical tests we will see use only variables of these types. It is not that we will not use techniques specifically tuned for ordered categories or rank variables, but these are of less use in introductory courses on statistics.

Note also that we will sometimes apply techniques designed, in theory, for categorical or numerical variables to variables with ordered categories. Since some flexibility exists regarding how the level of measurement of variables can be used when selecting which statistical test to apply, we will sometimes do this for the sake of simplicity or convenience.

5.1.3 The role of the variables

A recurring theme in knowledge is that of causality. Often we want to obtain knowledge based on causes and consequences, but this is not easy to achieve. It is true that, since childhood, humans are able to interpret that, for two events that occur in succession or interrelatedly, one event must be caused by the other. However, if you have ever played with small children, you will have realized how easy it is to trick them and make them believe they are affecting something as a result of their actions when in reality it is you who is secretly pulling the strings. Progressively, as children grow older, they develop a more acute sense of causality and it becomes more difficult to deceive them. Don't think that the problem ever disappears, however: correctly identifying cause and consequence is a problem that accompanies all of us all our lives.

Moreover, as psychologists, on many occasions you may find yourself struggling with the attribution of false causes by other individuals. People easily see causes where there are none and react in ways that do not benefit them, or worse, may even harm them. Making them see that can keep you busy for much of your professional life.

But how do we determine cause and consequence? In many courses on Statistics, the phrase "correlation is not causality" is invoked but no clear path to determining causality is given. Although it is true that if two things occur next to each other, we must not automatically deduce that one causes the other, we are often left in the dark about how to correctly make this deduction. Unfortunately, I don't intend to provide a full answer to this issue here either. A first step in determining causality, therefore, is to follow the steps I presented earlier on the subject of the Scientific Method. However, although that may point you in the right direction, it is still not enough.

So, although in this course you will apply statistical analyses that often provide information that could be interpreted, to a certain extent, as supporting the existence of cause-and-effect relationships,

since claiming that such a relationship actually exists is like treading on thin ice, we don't generally use these words but others. Instead of cause and effect, we will use the terms **independent variable** and **dependent variable**, while below is a (probably non-exhaustive) list of other possible terms we could use:

- "Cause": Independent variable, Treatment, Experimental variable, Predictor, Factor (if the variable is categorical), Covariate (if it is numeric), Grouping (same as Factor), Controlled, Explanatory, Manipulated, Regressor, Input.

- "Consequence": Dependent variable, Outcome, Predicted, Explained, Response, Effect, Output. Note that, in some tests, the variables do not have different roles. In such cases, the above names do not apply and we generally say, "all variables are independent". By way of example, correlation and regression share numerous aspects. When analyzing correlations, however, we think only in terms of relationships between the variables, whereas when analyzing regressions we have to indicate which variable is dependent and which one is independent.

I must confess that I often actively contribute to the confusion that is probably caused by having so many different names for what conceptually is very similar. When I am disciplined enough to maintain consistency, however, my two favorite names are Independent/Dependent (because they are very general and can be used in many contexts) and Explanatory/Explained (because I think they convey the spirit of the concept).

Finally, I would like to reiterate that, when we are performing statistical analysis, the names for cause and effect must be used with restraint because, I repeat, statistical analyses *per se* are just one part

of a whole process that enables us to ascertain whether a relationship is causal.

5.1.4 Choosing the level of measurement

In a Statistics course like this one, we usually proceed with tests with the variables as they are given to us. We do this because we assume that the decisions on the level of measurement were taken when the research study was designed and there is little we can do about it. However, in a real study, you will sometimes have the chance to decide how to measure the same characteristic in different ways and, therefore at different levels. This is important because measuring your variables in a certain way can make them much more convenient to analyze than when they are measured in other ways. Of course, this assumes that you have the option of choosing the measurement level of your variables, which is not always true. However, if you can, knowing certain rules can be very helpful.

With dependent variables, the general rule is that the higher the level you achieve, the better your measurement is. So, if you are able to measure your variable so that the result produces a numerical variable, by all means go ahead. Since other levels of measurement make the analysis more complicated, try to avoid them if possible.

With independent variables, the rule is less clear. Numerical independent variables are technically superior because more advanced analyses are possible, but categorical independent variables are easier to analyze. Introductory courses on statistics usually begin with tests for independent variables that are categorical. Then, depending on the time available, they go on to discuss numerical independent variables. Ordinal variables are the most challenging to use. Although some tests for analyzing them very

much resemble those for categorical and numerical variables, they have several limitations that advise against using them if possible. In this course, we will see techniques that use ordinal dependent variables. The disadvantage of these techniques is that they must be interpreted using ranks, understanding of which is often not very intuitive. It is less common to use independent ordinal variables as ordinal In practice, they are often treated simply as categorical or numerical variables because statistical techniques for this situation are not usually discussed in introductory courses on Statistics.

A more controversial solution is when an ordinal dependent variable such as a Likert scale is used as if it were numerical. This approach is typical in surveys where, in theory, large samples make it acceptable since it is assumed that equivalent results would be obtained using more sophisticated approaches. In this course, for the sake of simplicity, we will sometimes proceed in this way.

5.1.5 Identifying the level of measurement

I am aware that students sometimes have problems identifying a variable's level of measurement. I guess the difficulties stem mainly from something I mentioned earlier: since a variable can be measured in different ways, in order to decide what the variable's level of measurement is, you have to judge how the measured values relate to the characteristic you wish to measure, which is sometimes hard to understand without thorough knowledge of that characteristic. However, some simple guidelines may help out here. The first thing to consider is whether the result of the measurement is a list of categories. If it is, you should next consider whether there is an order in the categories. Questions in surveys that are answered using a Likert scale such as {1=Never, 2=Sometimes, 3=Always}, where the question could be "I wear a seat belt when traveling by coach", are

examples of ordered categories. Finally, if you have excluded the two previous levels of measurement in your variable, you probably have a numerical variable. You then still need to check whether the variable is based on ranks such as first, second, and third, etc., which would make the variable ordinal.

5.2 List of statistical tests

Table 3 shows a list of the statistical tests we will discuss on this course. The columns are for level of measurement, statistical test, aim, an example, and other tests related to the tests in the list. Note that, in some cases, some tests use only one variable rather than two, so dependent and independent roles do not apply. Finally, in some cases, the number of possible categories sometimes matters, as in the t-test for comparison of groups, where the independent variable is categorical with two values and the dependent variable is numeric. Analysis of variance, on the other hand, is the same as the t-test but with an independent categorical variable with any number of groups and a dependent variable that is numerical. In this case, the t-test is considered a "short" version of analysis of variance, which is considered the "long" version.

One sample tests					
Test	**Aim**	**Independent variable**	**Dependent variable**	**Example**	**Related tests**

One sample t-test	To compare a theoretical or populational value with the result obtained in a sample	Numerical		Amount of salt in a sample of bread loaves with a reference value
One sample proportion	To compare whether the proportion or percentage of cases in one category of the independent variable is different from one value	Categorical		Proportion of women in a job different from 0.50

Tests for categorical independent and numerical dependent variables

Test	**Aim**	**Independent variable**	**Dependent variable**	**Example**	**Related tests**
Two groups t-test	To compare the means in the dependent variable of the cases in each group	Categorical (only two groups/categories)	Numerical	Effect of medicine in a group of people versus a	This is a short version of Analysis of Variance (see below)

				placebo group	
Analysis of Variance (ANOVA)	To test the differences among several means	Categorical	Numerical	Mean taps per minute for three doses of caffeine	This is a long version of the two groups t-test (see above)
T-test for comparing repeated or dependent measures	To test whether the difference between two repeated or dependent measurements is zero	*Two Repetitions*	Numerical	Body weight of a sample of people before and after a treatment	This is a short version of Analysis of Variance for repeated measures (see below)
Repeated measures Analysis of Variance	To test there are differences among several repeated or dependent measures	*Two or more repetitions*	Numerical	Differences in depression feelings in four times for a group of people	This is a long version of the t-test for repeated/dependent measures

Tests for categorical independent and categorical dependent variables

Test	**Aim**	**Independent variable**	**Dependent variable**	**Example**	**Related tests**

Two groups proportion test	To compare whether the proportion of cases in one category of the independent variable is different from the proportion of cases in another category of the independent variable for a category of the dependent variable	Categorical (only two categories)	Categorical (we calculate proportions or percentages)	Differences in the proportion of people who died in the sinking of the Titanic between those traveling in first class versus those traveling in third class
Chi-square test	To test the association between two categorical variables	Two categorical variables	Association between class and survival in the Titanic	

Tests for categorical independent and ordinal dependent variables

Test	**Aim**	**Independent variable**	**Dependent variable**	**Example**	**Related tests**

Mann-Whitney U test	To test rank/ordinal scores in two groups	Categorical (only two groups/categories)	Ordinal	Testing whether life is more exciting for men or for women	This test is like a two-group t-test when the dependent variable is ordinal. It is a short version of the Kruskal-Wallis' test
Kruskal-Wallis test	To test rank/ordinal scores in two or more groups	Categorical	Ordinal	Testing whether life is exciting depending on marital status	This test is like an Analysis of Variance when the dependent variable is ordinal.
Wilcoxon test	To test rank/ordinal scores in the same subjects in two repeated/dependent measures	Two Repetitions	Ordinal	Preferences between two types of music for each subject	This test is an alternative to the t-test for repeated/dependent measures. It is a short version of Friedman's test.
Friedman's test	To test rank/ordinal scores in the same	Two or more repetitions	Ordinal	Preferences between three types	This test is an alternative to the Anova test with

	subjects in several repeated/dependent measures			of music for each subject	dependent measures. It is a long version of the Wilcoxon test.

Tests for association between ordinal variables

Test	Aim	Independent variable	Dependent variable	Example	Related tests
Spearman correlation	To test the relationship between two ordinal measures	Two ordinal variables		There is no dependent variable but your theory may set one variable as a cause and the other as a consequence	This is the Pearson correlation after the numerical variables are transformed into ranks

Tests for numerical independent and numerical dependent variables

Test	Aim	Independent variable	Dependent variable	Example	Related tests
Pearson's correlation	To test the relationship between two	Two numerical variables	If your theory sets one	Correlation between body	

	numerical variables		variable as dependent, you should use simple regression rather than this test	weight and height in a sample of people	
Simple regression	To predict the values in the dependent variable from the values in the independent variable	Numerical	Numerical		Pearson's correlation and simple regression share many features
Multiple regression	To predict the values in the dependent variables from those in several independent variables	One or more numerical variables	Numerical		This is an extended version of simple regression

Multivariate tests

Test	**Aim**	**Independent variable**	**Dependent variable**	**Example**	**Related tests**

Cluster analysis (K-means)	To identify groups of subjects with similar characteristics	Several variables (numerical but ordinal are also used in practice)	
Principal Components Analysis	To find groups of highly correlated variables	Several variables (numerical but ordinal are also used in practice)	This test starts with Pearson's correlations

---

1. Many statistical packages, including SPSS, use some type of variable classification that restricts the analyses that are permitted. However, since such classifications are not universally accepted and are sometimes plain wrong, we will not pay much attention to them.↩

2. Tukey (1977) advocated a new type of data called proportion/percentage that would serve for variables between 0 and 1 or between 0 and 100. This idea has not yet been applied in introductory courses on Statistics but is quite common in more advanced courses.↩

3. In general, we treat a variable with many different numerical values as if it were numeric. However, if we think carefully about some of these variables, we may hesitate before doing so. An example I remember from my student days is the score

of 10 given by teachers who believed that this grade was for only very special students. For these teachers, the grade scale followed a more or less homogeneous upward progression until it reached the threshold from 9.9 to 10. In their opinion, the difference between these two grades was worth much more than the difference between two other grades separated by one-tenth of a point, so, for them, one of the criteria usually regarded as an indicator of numerical variables – equality in the differences between consecutive values on the scale – did not apply.↩

4. A more thorough treatment of this problem is normally the subject of courses on Experimental Methods, which are usually supplementary to courses on Statistics.↩

6 Steps in statistical tests

In the previous chapter we saw how to select the test to use in each case. At the end of that chapter is a table with the tests we will see on this course.

As you can see, the list is rather long and you may feel overwhelmed by all the contents you need to learn. Fortunately, as we will see in this chapter, many of those tests are applied by following similar steps, so the effort needed to learn new tests is not excessive.

In this chapter, I will first provide a list of the general steps you need to follow to apply a statistical test and provide an introduction for each one. Note that there is relevant material on this in the "Results" section of chapter 4, which I will sometimes refer you to rather than repeating the same things here. This will help to keep this chapter as short as possible.

Assuming that you have selected the right test for your problem, the steps you should take are as follows:

- Draw a plot/graphical representation of your data.
- Check your assumptions.
- Set the statistical hypothesis.
- Calculate the P-value and the confidence interval of the effect.
- Interpret your results.
- Perform *post-hoc* tests.

Below I will explain each of these steps.

6.1 Plots/graphical representations of data

An often-repeated phrase is that no statistical analysis should be performed without visually representing the data. With very small data sets, it is possible to keep each value in mind individually. However, when the data set grows in size, the only realistic way to observe them is with a graph. The problem is, finding the right graph for each situation, applying it, and interpreting the result is a fairly complex problem in itself (@ young2011visual). Sometimes the source of this difficulty stems from the software (e.g. SPSS or Excel) used for the analysis since it may not be able to draw a suitable graph for your data. In other cases, the problem stems from the level of measurement of the variables or their characteristics. On this course, I will recommend a basic chart to apply to every test we see.

What may we find in a plot?

- Data errors, typos or, simply, very strange values: finding someone in your sample who is over 100 years old, sleeps 24 hours a day, or earns 100,0000€ a month is so strange that we would be urged to check whether there isn't some sort of error behind these values. Be careful, though, because sometimes these strange values may not be errors but valid values that should be dealt with appropriately. These values are called outliers. What we do with them is hugely important but difficult to summarize in an introductory course on Statistics.[1]

- The effects to be analyzed: plots that show means, prediction lines or other elements give a hint of the result you can

evaluate visually. This may make it much easier to interpret the results but also help to see relationships that cannot be easily summarized with a number.

- Individual values: looking at specific values is sometimes as interesting as knowing a summary or trend. Plots make it possible to single out interesting values and, if necessary, label them individually.

For a simple example, look at Figure 6.1, which shows the relationship between a movie's budget and the average IMDb (Internet Movie Database) rating for a group of movies. Note that the horizontal axis has values ranging from zero or practically zero to 200 (millions dollars). Each point on that graph is a movie but, since those with a low budget are so numerous and overlap, we can only see the black shaded area. However, we can also see that some of the highest-rated films in this group (with an average rating close to 10) are also in that group of low-budget films. On the other hand, we can also see that films with the largest budgets are not guaranteed popular success (though the trend, as illustrated by the *average* line, does go very slightly upwards).

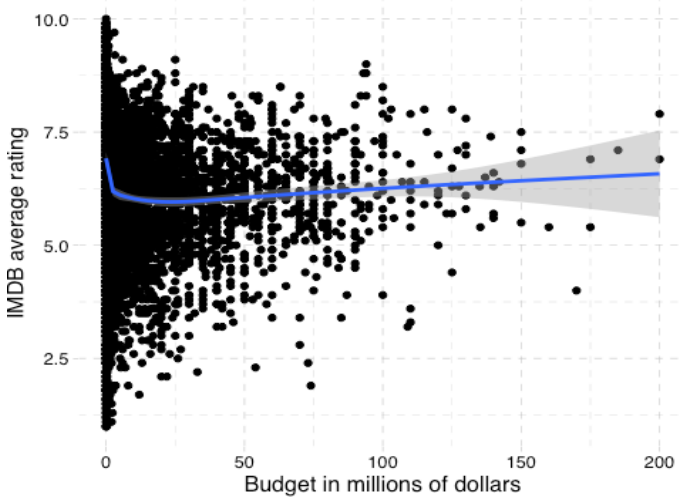

Figure 6.1: Ratings of IMDB movies and their budget.

6.2 Checking the test assumptions

The statistical tests we will see on this course are based on principles or assumptions that guarantee that their results are correct. Most sites where these statistical tests are explained recommend that those assumptions should be verified in some way, otherwise you run the risk that your results will not be valid. For the type of tests often seen on an introductory course such as this one, the two most common assumptions are:

- If the dependent variable is continuous, the sample of data analyzed must come from a population whose distribution follows the normal distribution.
- If there are groups, the sub-populations from which the samples of each group are drawn must have the same variance/standard deviation.

Don't worry if you don't understand what these assumptions mean at the moment. We'll see what they mean in more detail later but, until then, I would like to comment on one aspect that causes a lot of confusion: what happens if the assumptions are not met?

Places where these tests are explained often convey the impression that if the assumptions are satisfied, the data analysis cannot continue and the data must be discarded. However, in my opinion and that of others (e.g. Moore (2010)), this reaction is too extreme. Before reaching this extreme, you should take into account the following:

- The assumptions don't have to be absolutely true. In practice it is sufficient to approximately verify the assumptions. For example, in the two assumptions above, if a) the dependent variable is approximately normal, and b) the groups have

approximately the same variance, we can continue with our analysis[2].

- Several methods, including statistical tests, graphs, prior knowledge, and mathematical logic, can be used to check whether the assumptions in a given situation are sufficiently verified. Any of these methods can provide an answer to the problem but these answers may not be the same: sometimes some methods can contradict others.

- Certain relatively recent advances in statistics solve the problem of non-fulfillment of assumptions fairly conveniently. For example, the assumption of homogeneity of variances is less important than it was a number of years ago because new, adjusted formulas that take into account this problem have been developed. However, many old handbooks, and people who learned statistics from them, have not yet incorporated these new methods and claim to still be indispensable. A similar situation occurs with the assumption about the normal distribution of the population: with large samples[3] and continuous variables, the results are robust even if the assumption of normality is not satisfied.

- When the assumptions are not met, using an alternative, "assumption-free", statistical test is sometimes suggested as a remedy. However, although there may be strong reasons for using these alternative tests, they often have disadvantages that must also be taken into account. Since the standard tests are known to be good enough even if the assumptions are not met, it seems inadvisable to stop using them unless the reasons for doing so are very strong.

- In view of the relatively advanced statistical maturity required to judge whether a violation of the assumptions is a

major problem in a given situation, it would be absurd to introduce many exercises where the assumptions are not met and then have to reject their results on that basis. On this course, therefore, although we will discuss this situation in theory and several examples will be presented, you will need to conduct this assessment yourselves only if judging whether the assumptions have been met is straightforward.

In summary, although checking the assumptions of statistical tests is important and we will see how to do in each of the tests we will study on this course, there are ways to bypass their non-fulfilment and still maintain confidence in your results.

6.3 Setting the statistical hypothesis

In Chapter 3 we introduced the so-called falsificationist approach to testing hypotheses. There we saw that to proceed with a test we usually need to set a null hypothesis and a study hypothesis. The aim of the test is to reject the null hypothesis. If we are successful, this will increase the confidence we have in the theory on which our study hypothesis is based. The section Setting the statistical hypotheses explains this step well enough so, in order not to repeat those explanations, I will just discuss a few additional features that were not covered there. If you feel you need to, please re-read that section before continuing here.

The additional features of statistical hypothesis not discussed earlier are:

- Sometimes we do not want to reject the null hypothesis.
- You do not accept the null hypothesis but you fail to reject it.

- The null hypothesis is sometimes too easy to reject.

Although the logic and procedure of a hypothesis test is based on rejecting the null hypothesis, sometimes we have a good reason to act differently, i.e. when the theory we believe in points to no effect or no difference. For example, suppose we are testing how the signs of the zodiac influence whether a marriage will be successful: if you are like me, I imagine you believe your sign of the zodiac is not very important in romantic relationships. So, if we were to test whether the percentage of divorced people is different for people with different signs, we would expect the differences to be zero. In this case, we would use the same hypothesis-testing process as described above, but our expected result would be not to reject the null hypothesis.

If the result of a hypothesis test is non-rejection, we say, "I have failed to reject the null hypothesis." I'm afraid this is a rather difficult sentence to understand, but if you check the statistics books you will see that everyone agrees that this is the correct way to express it. But why is that? Why don't we accept the null hypothesis if we fail to reject it?

Well, it may seem illogical but accepting the null hypothesis is not a valid option.

It is not easy to give an easily comprehensible explanation since this procedure clashes with how we usually do it. However, bear with me while I try to explain: suppose we evaluate the effect of a random reward (a lottery for a holiday in Spain, for example) on the districts of a town where you want to improve the quantity and quality of garbage recycling. We could select several districts, measure their recycling levels, and announce the prize. Let's say we focus on the paper/cardboard category. Now, if you think about the result of this imaginary experiment, you will understand that it would be very odd, even if the treatment had no effect, to find the exact difference

between the amount of paper recycled before and after announcing the prize was zero for each neighborhood since, for whatever reason, there will always be a slight variation in the amount of paper recycled. However, let's assume that the difference is generally small *taking into account the number of neighborhoods used in a study* and the statistical test tells us that it is not significant: in this case, we would conclude that we cannot reject the null hypothesis.

However, now suppose that someone is strongly convinced that the rewards are effective despite the results of this study and manages to obtain funds to repeat the study nationally using many more districts than before. And suppose that the difference found is again small but, *as the number of neighborhoods used in this study is much higher than before*, the test now tells us that it is significant and we can reject the null hypothesis.

Contradictory? Actually, no – provided that in the first study you did not accept the null hypothesis. If you said you were unable to reject it, you left open the possibility that in a study with a larger sample, or in one with better measurements, the hypothesis could be rejected and so there is no contradiction.

What conclusion, therefore, can we draw from a study? If our result could be a different one, what confidence can we have in our hypothesis tests? I think I already answered this question when I quoted the book by Shadish et al. (2002), but I will repeat the relevant section again below:

Conclusions that resist falsification are preserved in scientific books or journals and treated as plausible until better evidence is presented.

That is, rejecting the null hypothesis means resisting falsification and therefore maintaining the scientific theory we wished to test *as long as no better evidence* is presented (since no statement based on

empirical data can be treated as though it were eternal). On the other hand, not rejecting the null hypothesis has a more ambiguous interpretation. But again, a sentence from the book I mentioned earlier is appropriate here:

… if negative results are what you get most of the time … it's time to start over and think of a new theory.

That is, when we do not reject the null hypothesis, what we must do is *not draw big conclusions for the time being*. However, if this happens on more occasions, it is time to make up your mind and think about a new theory to substitute for the one you have.

The process involved in establishing a statistical hypothesis sometimes derives from a null hypothesis that is so easy to reject that we must view the result with much suspicion. The idea is to establish null hypotheses that are sufficiently challenging. Otherwise, we could end up saying: "Yes, we reject the null hypothesis, but so what?".

Here is an example of what I mean: suppose a company with franchises all over the world is offering courses aimed at improving children's mathematical ability based on a revolutionary method invented by a foreign expert. Suppose also that, to prove the effectiveness of its method, the company provides a study that demonstrates that those who take their courses are better on average than those who do not (and that, moreover, the difference is statistically significant). Based on that study, should you regret not taking those courses when you were a child? The answer is "not really", because, when all is said and done, being somewhat better than average is not such a big improvement and if you were obliged to stop doing other important things to reach that level, it might even have been counterproductive.

If the company really wanted to impress, the null hypothesis should be based not on the average score of those who have not taken their courses but on at least one or two standard deviations above the average. Now that would be a level worth making the effort to reach (if, of course, you are interested in mathematics).

In Statistics classes we often don't have much time to think in great detail about what a suitable null hypothesis would be for our problem. However, it is important for you to pay attention to this aspect when you read results of studies that are important for your work.

6.4 Calculating the p-value and the confidence interval of the effect

We have already described the concept of the p-value and confidence level in general terms in the chapter on the Scientific Method. Remember that the p-value indicates the probability of obtaining the results we obtain in a study if the null hypothesis were true.

Remember, too, that if the probability is low, we reject the null hypothesis. But if it is high, we cannot reject the null hypothesis.

In reality, this procedure is not without controversy and there is a growing trend to provide what are called measures of effect size. However, as Shadish et al. (2002) wrote:

…few parties in the debate believe that the null hypothesis significance test should be eliminated entirely.

So what do they recommend?

- … we recommend that results are reported first as estimates of effect size accompanied by 95% confidence intervals, followed by the exact probability level of Type I error from a significance test of the null hypothesis.*

This paragraph needs a little explanation: there are two interesting parts to it:

- … exact probability level of Type I error from a significance test of the null hypothesis.*

In this sentence, Shadish et al. (2002) are referring to the bad habit of expressing the p-value simply as ">.05" (i.e. do not reject the null hypothesis), or "<0.05" (i.e. reject the null hypothesis). You will find this way of expressing the result of statistical analysis in many articles but, as these authors indicate, if the significance of the statistical test is, for example, p-value=0.00001 rather than p-value = 0.04, we can see that the results are more robust in the first case than in the second, and that the null hypothesis is more clearly rejected. Obtaining the exact p-value is not a problem nowadays, since the software usually generates it automatically. It is therefore a good idea to use it when you communicate the result of a statistical test (for example, in the report).

- … results are first reported as effect size estimates accompanied by 95% confidence intervals *

As you can see, Shadish et al. (2002) recommends that the effect size and confidence intervals should be reported in addition to the p-value. I'll postpone explaining effect size for now but I'll briefly explain the idea of confidence intervals.

In a later chapter, we will see how to calculate confidence intervals. Until then you will easily understand how they are used if you see the results of political polls on television or in other media. These polls

generally report an estimate of the percentage of votes a political party may get in a future election. Let's say that the percentage of votes for a certain political party is 51%. If this result is broadcast before an election, one might automatically think that that political party is due to win the elections – if they were held the next day. However, depending on the sample size for this estimate, plus other factors, the 95% confidence interval for this value could range from 31% to 71%. This very wide interval is probably due to a very small sample being used. If we know this, the impression that the political party is going to win the election is no longer so outstanding. Consider instead the following confidence interval {50.5%, 51.5%} at the same 95% confidence level: these values indicate that there is a 95% probability that the result of the elections will be between 50.5% and 51.5%, and the impression that that political party will be the winner is much stronger. If the confidence interval is narrower for a similar study, the usual reason is that the sample size is much larger, which enables the percentages to be estimated more accurately. As you can see, providing a confidence interval for a parameter is more informative than simply providing the parameter, so it is recommended that, if possible, we do.

6.5 Interpreting the results

After performing a statistical test, the next step is to interpret the result. Depending on the test used, this may or may not be relatively easy. For example, if the test is a difference between two means, and this difference is significant according to the p-value, it is relatively straightforward to interpret its meaning. Let's say that method A for teaching reading produces children who read at X speed, while method B produces children who read at speed X + 10. The difference between method A and B would be 10, so we would conclude that

method B is the best, since those who use that method read faster than those who use method A (assuming that the difference is significant).

Easy, isn't it?

Yes. In this example, interpreting the result is easy. In other cases, however, the test results are not just a difference. For example, interpreting the results of the Chi-square test involves looking at what are called standardized residual values, which are obtained by calculating the difference from the expected values. As you can imagine, a little explanation is required to understand what this means.

Fortunately, the tests taught in an introductory statistics course such as this one are relatively easy to interpret – though not always, so sometimes you will need to learn a few tricks to be able to do that.

6.6 Post hoc tests

The list of statistical tests informs us that some of the tests are "short" or "long" versions of others. In the short versions, we work with two groups/treatments/repetitions, while in the long versions we work with more than two. This is the case, for example, of the t-test and the analysis of variance, the former being a short version and the latter a long version of the same test. When we work with the short version, the result directly indicates the result we want – if there are differences between the two groups. The example in which we compared two teaching methods reflects this situation.

On the other hand, when we use the long version of the test, as we have more than two values to summarize, we can calculate more than one difference. For example, we can have five methods for teaching

reading. In this case, the statistical test provides a global assessment of the differences between the methods. However, if the results are significant, we do not know between which pairs of groups these differences are significant. It may be that all the methods are different from each other, or that four of them are roughly the same and only one stands out from the rest. We cannot tell from the results of analysis of variance where the significant results come from.

Post hoc tests are the answer to this problem: if the overall statistical test for a "long" test is significant, we can run a new round of statistical tests to see between which pairs of groups, treatments or conditions there are differences.

---

1. If you have an outlier, an easy way to see its effect is to repeat the analysis with and without it and observe whether the results are very different. If they are not, you can choose either to include or exclude the outlier since it has no effect. If they are, you can report both results.↩

2. I understand that judging when the assumptions are *approximately satisfied* is not easy for a beginner in statistics, but don't worry about that just yet.↩

3. When a sample is large will be explained in due course.↩

4. At least, this is my view. I don't usually set exercises in which you have to decide whether assumptions are satisfied; if I do, I try to use examples in which the answer is as obvious as possible.↩

References

Barnard, G. A. 1982. “Causation.” In *Encyclopedia of Statistical Sciences*, edited by C. Read, S. Kotz & N. Johnson, 387–89. John Wiley & Sons.

Bryson, Bill. 2019. *The Body: A Guide for Occupants*. Random House.

Cummiskey, Kevin, Bryan Adams, James Pleuss, Dusty Turner, Nicholas Clark, and Krista Watts. 2020. “Causal Inference in Introductory Statistics Courses.” *Journal of Statistics Education* 28 (1): 2–8.

Fisher, R. A. 1935. *The Design of Experiments*. Edinburgh, UK; London, UK: Oliver; Boyd.

Fisher, Ronald A. 1958. “Cancer and Smoking.” *Nature* 182 (4635): 596–96.

Holland, Paul W. 1986. “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81 (396): 945–60.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus; Giroux. https://www.amazon.de/Thinking-Fast-Slow-Daniel-Kahneman/dp/0374275637/ref=wl_it_dp_o_pdT1_nS_nC?ie=UTF8&colid=151193SNGKJT9&coliid=I3OCESLZCVDFL7.

Lewis, M. 2017. *The Undoing Project: A Friendship That Changed the World*. Penguin Books Limited. https://books.google.es/books?id=-ltNvgAACAAJ.

Lewis, M. M. 2003. *Moneyball: The Art of Winning an Unfair Game*. Norton Paperback. W.W. Norton. https://books.google.es/books?id=RWOX_2eYPcAC.

Meier, Ann, and Kelly Musick. 2014. "Variation in Associations Between Family Dinners and Adolescent Well-Being." *Journal of Marriage and Family* 76 (1): 13–23.

Moore, David S. 2010. *The Basic Practice of Statistics*. Palgrave Macmillan.

Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association* 100 (469): 322–31.

———. 2007. "The Design Versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials." *Statistics in Medicine* 26 (1): 20–36.

Shadish, William R, Thomas D Cook, Donald Thomas Campbell, et al. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference/William R. Shadish, Thomas D. Cook, Donald T. Campbell.* Boston: Houghton Mifflin.

Solomon, Paul R, Felicity Adams, Amanda Silver, Jill Zimmer, and Richard DeVeaux. 2002. "Ginkgo for Memory Enhancement: A Randomized Controlled Trial." *Jama* 288 (7): 835–40.

Taleb, Nassim Nicholas. 2007. *The Black Swan: The Impact of the Highly Improbable*. Vol. 2. Random house.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Addison Wesley.

Weisberg, Sanford. 2005. *Applied Linear Regression*. Vol. 528. John Wiley & Sons.

---

1. ↩