THE PHILOSOPHY OF ECONOMETRICS

AN INTRODUCTION

Willem van der Deijl

OPEN PRESS TiU

ISBN 978-94-037-1258-1
DOI 10.26116/OKRV-2B42

Design, Typesetting, and Cover by Thomas F. K. Jorna
Coordination by Violet S. Zagt
Publishers of Trial and Error
https://trialanderror.org

Published by Open Press Tilburg University,
Tilburg, Nederland
https://tiu.trialanderror.org
info@openpresstiu.org

The Philosophy of Econometrics

An Introduction

Willem van der Deijl

Open Press Tilburg University

Contents

Preface

In the period of 2011-2013 I was a graduate student at the Erasmus Institute for Philosophy and Economics, and became increasingly interested in the philosophy and methodology of econometrics. I was able to do this as a result of my teachers and supervisors at the time: Julian Reiss, Marcel Boumans, Jack Vromen, Conrad Heilmann, at the philosophy faculty, and my econometrics teacher and supervisor Nalan Basturk at the economics faculty. Especially Julian Reiss' course on the philosophy and methodology of statistics was instrumental in developing my thought on the topic. Much of the content in this book came to me through these teachers.

After this period, my research direction took another course: the measurement and ethics of value. However, when I got the opportunity to teach a part of a course on the philosophy of science and statistics at Tilburg University, I gladly accepted it. I got to revisit my interest. Because I could find few introductory texts, I decided to write my own, encouraged by the course coordinator Michael Vlerick. These texts developed into this book. I owe much gratitude to the students who took this course in the previous years. They helped me point out places in which the explanation was not as clear as it could be, and corrected many of my type I and type II error confusions. I apologize in advance to future students for mixing these up.

I also want to thank Matteo Colombo, for reading the book, and the helpful comments that resulted from this. Finally, I would like to thank editor Daan Rutten and copy-editor Violet Zagt.

Introduction

1 • THE RHEINHART-ROGOFF CONTROVERSY

During the Great Recession that lasted from roughly 2009 until 2014, many governments believed that it was necessary to fight the long-run impact of the recession through austerity measures: cutting government costs. The argument for this was, that high government debts would have negative consequences for the prosperity of these countries in the future. The economic reasoning behind this argument was that if a country's debt would become too high relative to its GDP, investors would no longer believe that country would be able to pay off its loans. Consequently, the interest rates on these loans would increase, only making it more difficult to maintain the high level of debt.

In 2009, two Harvest economists, Reinhart and Rogoff, published a research paper on this issue, having used historical data to estimate at which point debt will have significant negative effects on economic growth (Reinhart and Rogoff 2010). Their estimate was around 90% of GDP, a number many European countries had exceeded or were on the verge of exceeding. This paper became an influential argument for austerity measures in these countries, which in turn led to many job losses and a significant drop in household purchasing power.

Four years later, Thomas Herndon, a graduate student, discovered that numerous coding errors were made in the usage of the dataset. In truth, the estimate of the effect of debt on long-run economic growth should be much lower than Reinhart and Rogoff had thought (Herndon, Ash, and Pollin 2014). This paper, which many governments had used to justify cutting governments costs, was simply incorrect. Of course, we cannot know what would have happened otherwise, but this error may have very well caused many people to lose their jobs, many (small) businesses to go bankrupt, and many people on governments programs to lose their benefits.

This example is one of human error, not necessarily one of an error in statistical reasoning, but it does illustrate that real-world economics almost always has high stakes. And real-world economics is almost always based on statistical reasoning. This shows that good economic estimates are important, and that errors may result in the very real loss of jobs, financial security, and even life (Arcà, Principe, and Van Doorslaer 2020). It also shows why we need good econometrics.

2 • POPULAR, BUT NOT UNCONTROVERSIAL

Econometrics describes the field within economics that concerns itself with statistical reasoning in the economic context. Within economics, econometrics has become extremely influential. Econometrics has always held an important promise: making economic insight more practically relevant. It is great to understand in the abstract how national debt and GDP interact, but abstract models cannot tell exactly how much national debt will lead to a decrease in how much GDP growth. Economists who do not want to engage in statistical analysis can sit in their armchairs and debate the badness of government debt, but unless we have a precise estimate of how much debt is how bad, these armchair debates will not be very useful for policymakers. Economic numbers change the world, but it is important to ensure that they do so for the better.

Statistical reasoning has not always been so prominent in economics. Around the 1950s, papers using statistics in economics were still quite rare. Before that, they were virtually non-existent. In fact, the first usage of econometrics is often attributed to a specific publication: Jan Tinbergen's *Business Cycles in the United States of America, 1919-1932* (1939a; 1939b), a report published in 1939 for the League of Nations, the forerunner of the UN. Tinbergen wrote his report with a specific goal in mind: understanding the real-world workings of the economies of powerful nations, to limit the risk of war. This contribution was among the reasons he won the first Nobel Memorial Prize in Economics in 1969.

Interestingly, however, Tinbergen's introduction of statistical techniques to economics was not exactly uncontroversial. No other than John Maynard Keynes wrote a highly critical review of Tinbergen's report in *The Economic Journal* (Keynes 1939). Statistical techniques, he wrote, may be interesting, but will not really tell us anything of value.[1] Was there any point to Keynes' skepticism? In order to do econometrics well, we need not only to understand how to *perform* econometrics, but to understand the type of reasoning it involves. Understanding Keynes' skepticism, but also other arguments about the nature of statistical reasoning, will be essential in doing so.

This is an introductory text to the philosophy of statistics, and to statistical reasoning in economics. We will start at the beginning: why are the typical methods the way they are? The reasoning behind the most common type of statistical reasoning, significance testing, is quite peculiar. What is its rationale? We will investigate this in Chapter 1. Then, we will look at economic practice: do economists indeed use the methods as intended, and if not, is that a bad thing? To do this, we will look at two prominent critics of econometric practice in economics, Deirdre McCloskey and Stephen Ziliak, in Chapter 2. In Chapter 3, we will look at some philosophical problems for significance testing, and in Chapter 4, we introduce an alternative way of thinking about statistical inference: Bayesianism. In Chapter 5, we will take a closer look at the debate between Keynes and Tinbergen, and see what we can learn about this with respect to econometric modeling. Chapter 6 considers the problem of multiple testing. In Chapter 7, we look at

[1] More precisely, what he wrote was: "Taking everything into account, the successful application of this method [statistical analysis to economic problems] to so enormously complex a problem as the Business Cycle does strike me as a singularly unpromising project in the present state of our knowledge." (Keynes 1939, 567)

the well-known saying that “correlation does not imply causality”: is this indeed true? What then is the value of correlations?

This text aims to train the reader in critical statistical reasoning. It is not an introduction to doing econometrics. We have to start somewhere, and consequently, a basic familiarity with statistical methods and econometric tools is assumed. That being said, chapter 1 does review the reasoning behind significance testing, because these concepts are so crucial to the philosophy of econometrics, and in most statistics courses, it is easy to lose track the precise meaning of all the basic concepts. However, we will not pay attention to how we can calculate p-values from t-tests, and knowing how to do this will not be necessary to understand the material in this book. Instead, we take a closer look at the methods and the philosophy behind our everyday econometric practices. By the end of this book, you will have a better understanding of why statistical methods are the way they are, but also what their limitations are, and how we should think critically about the lessons that they can teach us.

I The Problem of Induction, Popper, & Significance Testing

1 • ECONOMETRICS: RELIABLE AND OBJECTIVE?

Econometrics is a scientific practice: its aim is to learn about the world based on data. What makes it a *scientific* practice? Different from mere opinions, science aims, and more or less succeeds, to be **reliable** and **objective**. Can econometrics be reliable and objective? It appears so. After all, the data do not lie. Nevertheless, there are some preliminary concerns.

Let's start with **objectivity**: econometricians like to say that we should let the data "speak for itself". Something can be called objective, when it does not depend on the particular characteristics of the individuals involved, in this case, in the econometric analysis. However, econometricians have to make many choices, and these can be made differently by different econometricians. In particular:

- What **data** do we use? Does they fit our question well? Can we trust the dataset?
- Economic questions usually are quite complex, and we use economic **models** to simplify and make sense of these questions. People using **econometric models** in particular make one important choice when they decide which variables to include in the model. When we are interested in the relationship between unemployment and economic growth, we still need to ask which

other variables should be included. How objective are these choices?
- There are many different statistical **methods**, each with advantages and disadvantages. When econometricians make conclusions based on, for instance, a regression analysis, they typically ask: "How likely is it, that we observe a specific sample mean, if our null hypothesis about our econometric model is true?" To arrive at a hypothesis test, we need to make important **statistical assumptions**. How objective are the choices that econometricians have to make here?

An economist and a philosopher jointly conducted an experiment in 1995 to answer exactly this question: if we send the same data and the same question to the same world-renowned econometricians, will they get to the same answers? The answer was "no": six different teams came up with six different answers (Magnus and Morgan 1999). This casts some doubt on the idea that choices made by econometricians make are truly objective.

What about **reliability**? Reliability describes the degree to which the results of an analysis are the same if the analysis is repeated under the same circumstances. One reason for concern for reliability is the replication crisis that started in psychology: researchers attempted to replicate empirical studies that psychologists conducted, and found that a large percentage could not be replicated. While the **replication crisis** has hit psychology harder than economics, in a large study of replication attempts in 2016 found that only two-thirds of results obtained through experimental economics can be replicated (Camerer et al. 2016). This is much less than we would expect if the statistical assumptions were correct. **Observational studies**, i.e. studies based on data that did not result from the researcher's manipulation, cannot be replicated by definition, Still, there have been disconcerting signs about the number of studies that falsely reject hypotheses (Ioannidis and Doucouliagos 2013). This appears to be much higher than the 5% or 1% of type 1 errors that we would expect if our statistical models were correct. We thus need to think critically about the methods that we use, to make sure that we do not draw the wrong conclusions from data.

2 • LIES, DAMNED LIES, AND STATISTICS

Statistical reasoning is essentially probabilistic, and as we have seen, the choices that statisticians have to make are not always made in the same way by all econometricians. The fact that many of these choices are expressed in mathematical expressions which are difficult to comprehend to the untrained, likely also does not help. This has led many to hold a fairly skeptical attitude toward statistics. Perhaps you have heard of the expression: "there are lies, there are damned lies, and then there are statistics", which is said to have originated from novelist Mark Twain. Another one, more specific to econometric modeling, comes from Nobel Laureate Ronald Coase: "if you torture the data long enough, it will confess to anything" (discussed in Tullock 2001).

It is perhaps easy to cast such skeptic comments aside by saying that those who make them do not fully understand the practice. At the same time, becoming more experienced as an econometrician, you will find that choices between seemingly reasonable modeling options can sometimes make significant differences. The story of the experiment about objectivity in econometrics described above illustrates that econometricians do in fact make different choices. Rather than casting such comments aside, we should try to understand the reasonable grounds for skepticism well, exactly so we can become better at avoiding pitfalls. Econometrics is hard, and not just because of all the statistical mathematics it involves. However, we have already seen that it is also tremendously important. So, how can we make sure we do econometrics with as little error as possible? To answer that question, we need to look at some basics.

3 • THE BASICS: INDUCTIVE INFERENCE AND THE PROBLEM OF INDUCTION

Traditionally, philosophers and scientists have divided reasoning about the world into two broad categories: **deductive reasoning** and **inductive reasoning**. Deductive reasoning involves drawing conclusions from premises using the methods of logic: if done correctly, this means that if the premises used are true, then so is the conclusion. If all trees are green, and oak is a type of tree, all oaks are green. Deductive reasoning can be difficult: essentially the whole field of mathematics consists of deductive reasoning. Yet, deductive reasoning has its limits. Are all trees in fact green? This is something that deductive reasoning cannot tell us. Even if a particular deductive argument is valid, its conclusion is as good as its premises. Deductive reasoning cannot tell us anything about what the world is like. How do we know whether all trees are green, or not? We need to actually investigate some trees! We call this **inductive reasoning**: reasoning that aims to draw general conclusions based on observations. An example of inductive reasoning is this:

> Tree 1 is green;
>
> Tree 2 is green;
>
> Tree 3 is green;
>
> ....
>
> C: All trees are green.

There is an obvious problem with this type of reasoning: not all trees are green. However, to someone collecting data in a Canadian forest in summer, this very fact may be entirely lost. This investigator may only see green trees, and therefore miss out on the fact that not all trees are green. This is an example of something known as **the problem of induction**.

The problem of induction is sometimes called the **black swan problem**. This comes from a particular historical anecdote. In 1697, Dutch explorer Willem de Vlamingh was exploring the coast of Western Australia on a rescue mission for a lost ship. The ship was never found, but he did see something no European had ever seen before: black swans. Up until this time, Europeans had thought that they had good reasons to believe in a widespread idea: all swans are white. After all, a swan of another color had never been observed. But Willem de Vlamingh's observation proved this idea wrong. This example raises an important question: how can we ever be sure that the generalizations we make from our data – such as "all swans are white" – are correct? In brief, the problem of induction is essentially that there appears to be no satisfactory answer to this question. We can never find sufficient observations to confirm the claim that all swans are white, all trees are green, and that no unicorns exist.

The problem can be put in the following terms. **The problem of induction:** if we want to generalize based on observations, we must do so in the following way:

1. X1 is W (swan 1 is white);
2. X2 is W (swan 2 is white);
3. X3 is W (swan 3 is white);

C: All X's are W (all swans are white).

However, **C does not follow from 1-3.** No matter how many white swans you observe, it does not logically follow from these observations that all swans are white.

The problem of induction has been an important influence on statisticians, philosophers and scientists. How can we avoid stepping into the trap that the Europeans made: looking at the world, and assuming that what you see is all there is? As silly as the problem of the black swan looks to us now, looking back at it, we often make generalizations based on past observations that appear to be quite reliable:

- The temperature has been increasing in the past, we can expect it to continue in the future.
- The sun comes up every day, so the sun will come up tomorrow.
- When economic growth has been high in the past, so is employment, and vice versa. Thus, when economic growth will go down, so will employment (Okun's law).

Are these also vulnerable to the same mistake that the Europeans made about swans? We appear to have good reason to be confident about these particular judgments, despite the problem of induction. The reason for this is that there is a partial solution to the problem of induction, one that has had a tremendous influence on the field of statistics.

4 • POPPER AND THE PROBLEM OF INDUCTION

The partial solution to the problem of induction is accredited to Karl Popper. A first step to Popper's solution involves the observation that while generalizations cannot be verified, they *can* be **falsified**. Falsification, here, means finding evidence that shows that a particular theory or hypothesis is false. In other words, while observations cannot definitively proof that the statement "all swans are white" is true, they can be used to show that the statement is false. The problem of induction applies to positive general claims (e.g. "all swans are white", or "unemployment always goes down when economic growth goes up"), but not to the denial of these claims (e.g. "not all swans are white", or "unemployment does not always go down when unemployment goes up").

A second step is this: if falsification of a general claim fails repeatedly, at some point, we can get confident that the claim is true. Popper called this process **corroboration**. While we cannot verify or confirm generalizations this way, we could get a little more confident in their veracity. Accordingly, Popper argued that scientists should not be looking for the confirmation or verification of their theories, but instead attempt to falsify them. The effects of climate change on the environment, the repetitious movements of the sun, and Okun's law have all been heavily scrutinized at various points in the history of science, and they have withstood this scrutiny: they are heavily **corroborated**. This should give us some confidence in their truth.

5 • HYPOTHESIS TESTING

The claims that economists investigate are typically not generalizations.

Economists do not claim that "when GDP goes up, unemployment always goes down". Rather, they are interested in the claim that *often*, or *typically,* when GDP goes up, unemployment goes down. A counterexample (e.g. one country where GDP went up, even though unemployment went down for a while) does not refute Okun's law. This makes the application of Popper's solution to the problem of induction a bit challenging. After all, it appears that claims like "typically, when GDP goes up, unemployment goes down" cannot really be falsified like "all swans are white" can. Thus, as economic laws are all of this form, they appear all unfalsifiable. Popper thought that what makes a field of investigation scientific is exactly the fact that they base themselves on falsifiable claims. So, if economics bases itself on unfalsiable claims, we can question whether we should call it a scientific practice.

We can find a solution to this problem in significance testing, an idea developed by the statistician Ronald Fisher. In his book *The Design of Experiments* (1935), Fisher illustrated significance testing with the example of a lady who said she could taste in the tea she was drinking, whether the milk or the tea was poured first. Fisher proposed the following experiment: the lady would be given 8 cups, in 4 of which the tea had been poured first, and in 4 the milk had been poured first. Fisher proposed a basic hypothesis, one that stated that the lady had no special ability: the **null hypothesis,** H_0. If the null hypothesis were correct, the likelihood of her identifying any one cup correctly would be ½, and we would expect her to correctly identify 2 of the cups in which the milk had been poured first. If the lady would correctly identify all 4 cups in which the milk had been poured first, however, the odds of this happening if the null hypothesis were true would be very small. In fact, it would be about 1.4%; see Table 1.

This, as you will recognize, is of course the standard logic of significance testing in statistics in general, and econometrics in particular. Almost exactly like Popper's Falsificationist logic, Fisher wrote about the null hypothesis that it "is never proved or established, but is possibly disproved" (1935, 16).

The standard approach to reasoning in econometrics is called **clas-**

TABLE I.I Fisher's Lady Tasting Tea experiment

Correctly selected cups of tea where the milk was poured first	0	1	2	3	Correctly selected cups of tea where the milk was poured first
Probability of event	1.4%	22.9%	51.4%	22.9%	Probability of event
Probability of identifying at least so many correctly	100%	98.6%	75.7%	24.3%	Probability of identifying at least so many correctly

sical statistics, and it is heavily based on these central insights. After Ronald Fisher's explication of what he called **statistical significance**, significance testing as a practice was further developed by Jerzy Neyman and Egon Pearson. Neyman and Pearson developed the significance test, which included two new concepts: the **alternative hypothesis**, and the **cutoff value** or **rejection rate**, typically called the α-value. Neyman and Pearson had different ideas from Fisher about how statistical testing should be implemented in science, but a conglomeration of their approaches has led to classical statistics: the significance testing framework that you find in statistics textbooks, and throughout almost any scientific publication that uses statistics.

The basic concepts

Perhaps you are familiar with hypothesis testing, but because this idea is so central to the contemporary statistical methodology, we need to look at the concepts in more detail. What are the key elements of significance testing?

> **The null hypothesis.** On a given sample space, a null hypothesis describes a specific distribution of a variable of interest with a specific mean. We can call this the target variable. In econometrics, such target variables generally are effect sizes or correlations. In these cases, the null hypothesis typically states that there is no effect or no cor-

relation. In other words, the mean of the target variable is 0.

The alternative hypothesis. the alternative hypothesis describes an alternative distribution that may be true if the null hypothesis is not true. Alternative hypotheses come in three forms:

1. **Specific.** The original developers of the classical framework intended the alternative hypothesis to be specific. For example, such a specific alternative hypothesis could be that the correlation between two variables is .5, rather than 0 as stated by the null hypothesis.
2. **One-sided.** Such general alternative hypotheses may be formulated one-sidedly: hypothesizing that the target variable's mean is higher than the null hypothesis suggests. For example, the target variable has a mean higher than 0.
3. **Two-sided.** Most often in econometric practice, the alternative hypothesis is formulated even more generally, namely two-sidedly: the null hypothesis is not true, the true mean is either higher or lower than it proposes.

P-value. A p-value quantifies the probability that the difference between a sample and the null hypothesis is as large as observed, or that it is even larger than observed in the direction of the alternative hypothesis, under the assumption that the null hypothesis is true.

Cutoff value or **rejection rate.** The rejection rate $\alpha$is a probability value that delineates the p-value cut-off line between rejecting or not rejecting a null hypothesis. It is typically set at .05. This means that whenever a p-value is lower than .05, we reject the null hypothesis.

These four concepts together capture the process of significance testing: a researcher formulates a null hypothesis and an alternative hypothesis about a target variable in a given population, then gathers a sample that is assumed to be a random selection of data from a larger population. The researcher observes the target variable in this sample, and under some statistical assumptions, calculates the p-value. This is key to the statistical reasoning of hypothesis testing: if the p-value is lower than the rejection rate, the null hypothesis is rejected, and we have found evidence for the alternative hypothesis. If not, we simply **fail to reject**.

An important implication is that, if the statistical model of the null has been correctly estimated, the rejection rate, α, is exactly equal to the probability of falsely rejecting a correct hypothesis. This is called a type 1 error. However, the probability of making a type 2 error – the chance of failing to reject an incorrect hypothesis, varies greatly with sample size, and can only be calculated if we know or assume the probability distribution of the alternative hypothesis or hypotheses.

6 • DIFFERENCES BETWEEN HYPOTHESIS TESTING AND FALSIFICATION

Significance testing and falsification come from similar concerns, and similar ideas about how we can circumvent the problem of induction. There are also important differences. The most important difference is that while Popper's criterion of falsification only applies when scientific evidence is *incompatible* with a certain theory, statistical data is **never strictly incompatible** with a statistical hypothesis. Because the inferences that we draw on the basis of the data are statistical, they are at most, highly unlikely to occur if a hypothesis is true. This is important because it acknowledges that even if we reject a hypothesis, we cannot do so with absolute certainty. To compare this, consider the type of logical argument that Popper proposes:

1. Theory A (all swans are white) implies that event X cannot occur (we observe a black swan);

2. X occurs (we observe a black swan);

C: Therefore, Theory A is false

The logic of hypothesis testing is similar, but importantly different:

1. Theory B (99.9% of the swans in the UK are white) implies that X (we observe a black swan in the UK) is *unlikely* to occur
2. X occurs (we observe a random sample of 30 swans in the UK from which 15% are black)

C: Therefore, we have evidence that B is false

What is important here is that observing a black swan in the UK is *incompatible* with the theory that all swans are white, but not incompatible with the statistical hypothesis that 99.9% of the swans in the UK are white.

7 • THE LOGIC OF HYPOTHESIS TESTING

We have already seen that the proper use of statistics is a process that is quite complex and relies on many different questions: what data do we use, what assumptions do we make about the data, which hypothesis do we test, and which tests do we use? We can summarize the logic of hypothesis testing as follows:

- if we correctly model the statistical process that generated the data, and
- we perform the significance testing procedure correctly, and
- we reject our null hypothesis, then
- we can interpret the test as evidence against the hypothesis.

Think again about our swan problem. Does the statistical method, with all the choices it adds to the simple inductive logic that Popper was concerned with, not offer a solution? Would 17th-century European biologists have drawn better inferences about the color of swans if they had been familiar with hypothesis testing?

The answer is: yes, and no. On one hand, the standard statistical method makes assumptions about the sample and population: the sample is, at least under the most common statistical assumptions, randomly drawn from a larger population. Applying this to the swan case, we see that, if we model the sample of swans that European biologists had seen as randomly drawn from the overall population of swans, we would still have corroborated the theory that 99.9% of the swans (or all swans) are white. However, the samples that Europeans had seen were not drawn from the general population of swans, but only from the Western, northern hemisphere population. Thus, if we make certain assumptions about the world, we can derive conclusions from our observations, but these conclusions always depend on the correctness of the assumptions that we have made.

8 • HOW DOES IT HELP?

How does all this background help us to evaluate the reliability and objectivity of econometrics? Note that the most common inductive inference that econometricians make is a bit odd. Before studying statistics, most people would expect statistics to be a practice in which we let the data speak for itself. But what the discussion above has shown, is that this is not what significance testers do. Significance testers do not ask: does the data show that X? For instance, do the data show that GDP and unemployment correlate negatively? Instead, they ask: if we assume X, is it likely to find something along the lines of what we observe? This not only requires us to make important assumptions that will affect our conclusions, but it also poses our research questions in a bit of an odd way. As we have seen, there is an important rationale for this way of thinking, but as we shall see in the next chapters, this odd type of inductive reasoning is not without its problems.

LEARNING GOALS FOR THIS CHAPTER

After studying this chapter, you should be able to explain:

1. indications why econometrics may not be reliable, and indications it may not be objective;
2. what the problem of induction is;
3. the logic of hypothesis testing, and how this logic can be seen as a partial solution to the problem of induction;
4. the similarities and differences between hypothesis testing reasoning and Popper's Falsificationist reasoning;
5. The following concepts: reliability, objectivity, inductive inference, the problem of induction, falsification, corroboration, replication crisis, p-value, null hypothesis, alternative hypothesis, cutoff value.

2 Is Economics a Cult of Statistical Significance?

Remember the last time that you ran an econometric model, or used data to test a hypothesis. What was the first value that you looked at after doing the test? For many users and students of econometrics, and of statistics more generally, the answer to this question will be the p-value. The p-value has become so central to statistics, that some have started to be concerned about it: are we not looking at the p-value too much? We will look at the criticism of its usage, and see if it is valid.

As we saw in the previous chapter, the reasoning involved in hypothesis testing may be a bit odd, but it also has an important rationale. It can be used successfully, as the simple example of Fisher's lady tasting tea has illustrated. There is also sufficient ground to criticize this way of reasoning, as we discuss in the next chapter. We may thus take issue with the logic of significance testing. However, another problem can be that the way the reasoning is used in practice does not quite correspond to theory. People may misuse the method, misinterpret the meaning of p-values, and apply hypotheses in ways that do not correspond to their original aims. This is the topic of this chapter.

You may think that such interpretation mistakes are made by *some* users of statistics, but not by econometrics students, who have turned statistics into their specialty. However, in this chapter we will discuss the views of two economists who believe that almost all economists who use statistics, including some of the best econometricians, frequently make errors in the interpretation of important statistics. In particular, they argue that one specific mistake is very common, with disastrous consequences: mistaking statistical significance for **economic significance** (McCloskey and Ziliak 1996; Ziliak and McCloskey 2004; 2008). Before we discuss their views, we first take a look at some common misinterpretations of p-values.

I • MEANING AND MISINTERPRETATIONS OF THE P-VALUE

P-values are everywhere in statistics, so the question of what they are should be a simple one. Nevertheless, even many experienced users misinterpret the p-value. What follows are three common misinterpretations of the meaning of p-values (see, for instance, Dickson and Baird 2011).

"The p-value is the probability that the hypothesis is true."

This is an understandable mistake, even for experienced users of significance testing. It is especially damaging in cases in which H_0 is highly likely to be true. A psychological study, for example, found that a sample of subjects was able to correctly guess in 53% of the cases whether a randomly placed image was going to appear behind a right or a left screen, and found that this was significantly different from 50% at the .05 level (Bem 2011). They took this as evidence that humans had the ability of foresight.

Does this mean that the probability that humans do not have this ability (the null) has a likelihood of less than 5%? In other words, is there a 95% or higher chance that humans have the ability to see into the future?

It does not. Critics of the study said that this was likely a coincidental result. If these critics are right, the probability that people cannot see into the future is very high (instead of less than 5%), even after this experiment. But that does not change the p-value, however. Even if the result is coincidental, the p-value remains correct. It still reflects the probability something like this, or even more different from H_0, would be observed in a random trial, if H_0 were true. As we shall see later, calculating the probability that a hypothesis is true or false requires more information than just a p-value. So, the p-value is something really quite different from the probability that H_0 is true.

"The p-value is the probability that the results of the trial are due to chance."

The phrasing of being "due to chance" is too imprecise. The fact that something is due to chance means that it is coincidental, that it results from a probabilistic process. However, the calculation of p-values *assumes* that the data result from probabilistic events. Except in cases in which the data was deliberately rigged, p-values always result from chance.

Consider the following example (Dickson and Baird 2011): someone is flipping a coin under the assumption that the coin is fair, and finds that it results in 6 heads and 4 tails. The p-value of this trial would be 75%. However, this is not the probability that the result is due to chance. If we are dealing with a normal coin, we *know for a fact* that the result is due to chance. The probability that the result is due to chance, is therefore not 75%, but 100%. However, imagine now that we know that there is a magician who *is manipulating the coins*, such that she controls the exact outcome of the coin flips. The p-value would be *the same*. However, now we know that the probability that the result is due to chance is 0%. A p-value of 75% can thus be the result of a test that is completely cooked, that has left nothing to chance, or it can be a result of a completely random process. The p-value by itself does not tell us enough about the probability that the trial was the result of chance, and neither does it provide enough information to make an inference about this probability.

"The p-value signifies reliability: $1 - p$ is the reliability of the result."

If the hypothesis is correct, the p-value (and not $1 - p$) signifies the probability of finding the same result, or a result that is more deviant from H_0. This *does* tell us *something* about how likely we are to find something similar again, were we to repeat the experiment, *if* H_0 is true. However, because we do not know whether H_0 is correct, neither p, nor $1 - p$ are clear signifiers of reliability.

Reliability of statistically significant trials also depends on sample size. Consider the example of Paul the Octopus, who, during the world cup of 2010 was remarkably successful at predicting match results. In 8 consecutive tries, he got 8 correct. This signifies a highly statistically significant result. Under the null hypothesis that the octopus has a 50% chance of guessing correctly, making 8 consecutive guesses corresponds to a p-value of .0039. But, if we find that in a small sample (n=1), an octopus can predict World Cup results (8 tries), we would not, and should not, see this as highly reliable. We should certainly not see it as 99.41% reliable $(1 - p)$.

2 • WHAT IS ECONOMIC SIGNIFICANCE?

The usage of statistical significance in economic science is widespread. When people run statistical tests, or regression models, the first things their eyes move to are the p-values: are they statistically significant? According to some, this has led to an overemphasis on this statistic. In particular, the two economists Deirdre McCloskey and Stephen Ziliak have criticized the usage of statistical significance in economics for decades. The title of their book on this topic, "**The Cult of Statistical Significance**", is a reference to the importance that economists and other users of statistics attach to the p-value. According to McCloskey and Ziliak, most economists take the p-value to be much more important than it actually is.

The word **significance** means something like **importance**. If something is significant, **it matters**. But, it can matter in different ways. In the context of econometric evidence, there are at least two different ways in which it may matter.

Statistical significance indicates whether a statistical datum meets a pre-set threshold that allows us to reject our null hypothesis. It tells us whether the data we found is removed from the hypothesized value to an extent that would be statistically (highly) unlikely if our hypothesis would be true.

Economic significance indicates how important a finding is from the perspective of the person asking the question. So, something can be of economic significance because it is important to economic science, economic policy, or to our own lives.

Something can be an important finding for economic theory and policy for various reasons, for instance:

- because it has a **large effect** on a variable of interest;
- because it makes **a big difference** in people's lives (small effects can have big implications in people's lives);
- because it **changes the way we think** about important economic issues.

Economic and statistical significance are two different things, but the two often get confused. Note that statistical significance does not tell us anything about the size of the effect. A statistically significant finding may indicate a tiny, and for all intents and purposes unimportant, deviation from our hypothesized value. The most common way to mistake statistical significance for economic significance, is to interpret a low p-value as evidence that the result **really matters** for economics or policy. If you make this mistake, you mistake statistical significance for economic significance. To see what this mistake typically looks like, let's look at an example.

McCloskey and Ziliak discuss an example from a paper by Nobel Prize laureate Gary Becker, Michael Grossman, and Kevin Murphy. The paper aims to explain why people buy cigarettes. In doing so, they build an economic model based on data from different states in the United States. At some point in their analysis, they write that "the highly significant effects of the smuggling variables (...) indicate the importance of interstate smuggling of cigarettes" (Becker, Grossman, and Murphy 1990). Why is that a mistake? Statistical and economic significance are simply different things. So, to say that statistical significance indicates that a particular variable is important for a model is incorrect. The fact that an estimate is significant tells us nothing about whether it should be in the model. For something to be included in an economic model, it would have to do more than be statistically significant. It would have to **change the qualitative results** of the model in a theoretically plausible way, correlate **strongly** with variables of interest, or **greatly improve** the model fit. Statistically significant variables may not do either of these things. Moreover, some non-significant variables may actually have really good grounds to be in an econometric model. If a team of highly trained economists makes these mistakes, it seems everyone is liable to do so.

It goes too far to say that there is no relationship between economic and statistical significance at all. If something is not statistically significant, the found deviation may have well been the result of chance, even if the null is true. Statistical significance therefore does tell us something of importance. Some economists have therefore argued that statistical significance is **necessary** for economic significance. That means that if something is not statistically significant, it cannot be an important finding for economic science. We will see later that this is arguable. According to McCloskey and Ziliak, statistical significance is neither sufficient nor necessary for economic significance.

The big mistake that many economists, including Becker, Grossman, and Murphy make, is that they take statistical significance as **sufficient** for economic significance. That means that if something is statistically significant, it is also economically important. Let's examine this mistake in more detail, before we address the claim that statistical significance is also not necessary for economic significance.

3 • STATISTICAL SIGNIFICANCE IS NOT SUFFICIENT FOR ECONOMIC SIGNIFICANCE

There are two reasons why statistical significance is not sufficient for economic significance: one obvious, and another less obvious. First, some null hypotheses that we may formulate about economic phenomena are outright uninteresting. There is, for example, a statistically significant relationship between the sea levels in Venice, as the city is slowly sinking, and the price of bread, which is slowly increasing, but that is not an economically significant finding (Sober 2001). So some statistically significant findings are simply not important. I did warn you this was going to be obvious.

But there is also a second set of examples that show that not all statistically significant findings matter economically. Consider the following question: how can you make sure, as a statistician, that almost any result will be significant? The answer to that question is very simple: increase your sample size.

Strictly speaking, increasing your sample size will not guarantee that your findings will be significant, but it will increase the likelihood that any small difference from the null will be detected. Consider someone who performs a test to see whether a coin is rigged, and flips the coin 1,000,000,000 times. They find that in about 500,050,000 cases, it lands head. An average of 50.005%. With such large numbers, this difference will be significant (P<.00078). But, does this also mean the coin is rigged? We have really good evidence that the coin is not *exactly* fair, but is it also rigged? A deviation of 0.005% will hardly affect our conclusion on the coin is fair or not. In fact, it seems that it shows that the coin is almost perfectly fair. A statistically significant finding may thus result from a very small, irrelevant difference, or effect size, especially when the sample size is large.

McCloskey and Ziliak (1996) provide an economic example: purchasing power parity (PPP). The PPP theory states that, controlled for exchange rates, goods should have the same price in every country. Imagine we conduct a statistical test of this hypothesis and find a value that is significantly different from 1, where 1 indicates equal prices. If we look at the statistical significance only, we would then have to say that this is evidence against the theory: there are statistically significant differences in prices (controlled for exchange rates) between different countries. But, if the sample is sufficiently large, we may find a value of ".999" that is statistically significantly different from 1. However, if that would happen, we would not say that the theory should be rejected, but rather, we have found evidence in favor of the theory. After all, .999 is extremely close to 1, even if it is statistically significantly different from it. For all intents and purposes, .999 shows that purchasing power parity is correct, even though it is not supported by statistics (Ziliak and McCloskey 2008, 94–97).

So, statistical significance is not sufficient for economic significance. Many statistically significant findings are simply not important, noteworthy, or interesting, from the perspective of economic science.

4 • STATISTICAL SIGNIFICANCE IS NOT NECESSARY FOR ECONOMIC SIGNIFICANCE

Even if statistically significance is not sufficient for economic significance, is it necessary? In other words, are only statistically significant results economically important? McCloskey and Ziliak believe not: if a sample size is small, and a certain target variable (e.g. an effect size) fluctuates, results are less likely to be significant, but may nevertheless matter a lot for policy. It is good to look at a few examples here.

Weight loss pills

The first example comes from a lecture by Stephen Ziliak.[2] He lets us imagine that someone asks us for advice on how to lose weight. You know that there are two pills available. The first pill causes a mean weight loss of 5 lbs, with a standard deviation of 1 lbs, while the second pill causes a mean weight loss of 20 lbs, with a standard deviation of 14 lbs. The second is not statistically significant, while the first is. However, for someone in need of losing a lot of weight, the second pill may be more effective.

A very real version of this example popped up during the early days of the Covid-19 pandemic. A small trial showed that the drug Remdesivir had positive but statistically insignificant benefits for hospitalized Covid-19 patients (Wang et al. 2020). If you had been in the patients' life-threatening situation, and you would be given a choice of whether to take the drug or not on the basis of this information, would you

[2] https://www.youtube.com/watch?v=_gK5r7LFZZs&t=13m30s

take it? If your answer is “yes, if there is a chance that it helps”, you have already implicitly admitted that sometimes non-statistically significant results are economically significant. Later trials found positive, statistically significant effects of the drug (Ali et al. 2022).

We can also look at an example from economic research discussed by McCloskey and Ziliak (2004) that has been discussed widely in economics:

A wage subsidy for unemployed workers

Another example that McCloskey and Ziliak discuss comes from research about a subsidy that was implemented in the state of Illinois in the 1980’s. It gave workers who were on benefits a cash sum if they found work. In another experiment, it gave employees a cash sum when it hired a worker who was currently on beneifts. When successful these policies are particularly cost-effective, because it limits the amount of money spend on benefits, and only costs the amount of the cash sum, which was $500. In a statistical analyses, the authors of this research paper assessed the effects of the policy. It concluded that the second policy had a cost-effectiveness ratio of $4.29 per dollar spent on average. However, this was not statistically significant. Only in one sub-group: those of white women, the effect was larger, 7.07 per dollar spent, and statistically significant. The authors write:

“The fifth panel . . . shows that the overall benefit-cost ratio for the Employer Experiment is 4.29, but it is not statistically different from zero. The benefit-cost ratio for white women, . . . however, is 7.07, and is statistically different from zero. . . . The Employer Experiment affected only white women” (Woodbury and Spiegelman 1987, 527)

Woodbury and Spiegelman see the non-significant finding in the group as a whole as indicating that there is no finding. But, in this case, the p-value was .12. As Ziliak and McCloskey write:

“That is to say, the 4.29 benefit-cost ratio was in the pilot study statistically significant at about the .12 level. In other words, the estimate was not all that noisy. A pretty strong signal for a very strong employment program.” (Ziliak and McCloskey 2008, 99)

What does this discussion tell us? It shows that limiting scientific, or economic, findings only to those that are statistically significant, may blind us against really important empirical findings. **Statistical** insignificance often means that we need more research. But, it does not mean that these findings are not **economically** insignificant. If, in the case study of the employment subsidies, we would neglect the result based on its statistical insignificance, we may not go through with a policy that actually works well for workers and the government. After all, if you would live in a state or country where a government would needed to decide on a policy, and it would find that this policy would have an expected return of \$4.29 per \$1, with a p-value of .12, you would probably want your government to go through with this.

McCloskey and Ziliak conclude that the relationship between statistical significance and economic significance is quite weak. If our statistical results show a deviation from hypothesized values, statistical significance may give us some assurance that this is unlikely to happen if H_0 were true, but even some non-statistically significant findings may tell us something important, and many statistically significant findings may tell us nothing of interest.

5 • ECONOMIC SIGNIFICANCE, OOMPH, AND LOSS FUNCTIONS

Now we know that economic significance and statistical significance are not the same thing. But what *is* economic significance? What is important from the perspective of policy and economic theory, is highly contextual. That which is economically significant in one context is not in another. Two factors that determine economic significance, are **oomph** and **the probability of the effect**.

Oomph describes the size of the effect. Think about the examples that we have seen: if someone is able to undergo an effective treatment for a deadly disease, this is of tremendous importance. If a minimum wage law is implemented, but as a result of this many lose their job, this matters greatly, specifically to those who do lose their jobs. But, a coin deviating a tiny bit from being exactly fair is rarely important. The common denominator here is the size of the effect. McCloskey and Ziliak call this **oomph** – how important is the observed effect in case we are right, and in case we are wrong?

The oomph of a result can be analyzed for four different scenarios that you are well familiar with:

- correctly concluding that there is an effect;
- correctly concluding that there is no effect;
- incorrectly concluding that there is an effect: type 1 errors (falsely rejecting the null of no effect);
- incorrectly concluding that there is no effect: **type 2 errors** (falsely failing to reject the null of no effect).

There may be important consequences to all these outcomes. Think about the experimental medication example: if we tell someone a treatment may work, and it does, this is great, but if it ends up not working, we may have given someone false hope in the final days of her life. But if we falsely conclude a treatment has no effect, while in truth there is a real effect, we may have prevented someone from being cured. If we correctly conclude that a treatment does not work, there are no consequences.

The same type of consequences play a role in economic research. Think about the Reinhart-Rogoff controversy we discussed in the introduction of this book: the two economists who wrote an article in the American Economic Review who concluded that economic growth is slowed down after a certain debt-to-GDP ratio based on an erroneous analysis. This paper was widely cited by policymakers, and the article had an immense impact on the austerity measures in Europe, and on the treatment of Greece and Spain by the Eurozone group. Falsely concluding that there was a big effect had as a consequent that many of the harsh austerity measures that were implemented were in vain or even destructive to fragile and heavily damaged economies in the middle of an economic depression. If these findings had been correct, they would still have had a large effect, but the effect would ultimately have been positive for the economies of Greece and Spain.

This brings us to the second factor in determining economic importance: **the probability of the effect**, which describes how likely is it that the positive or negative consequences of an action will occur. What are the uncertainties that come with the effect? Here, p-values do play an important role. P-values provide important information about the probability of an effect. It is not just the size of an effect that matters, but also how certain we are of it. For example, if we expect that, on average, an employment subsidy of $500 will have an average return of $4.29 per dollar, it makes a difference if we are sure it will be between $4.19 and $4.39, or whether it may be between -$5.29 and $13.29, as this would mean that the policy may also cost more money than it saves. In other words, uncertainty matters.

These factors together make up a **loss function** that describes the consequences (i.e. the costs) of making errors of magnitude in our inductive inferences. Such a loss function, McCloskey and Ziliak argue, should play a crucial role in the statistical inferences we make: if we observe a non-significant effect on a treatment of a deadly disease in a small sample, we should conclude that there may be an effect, but if we observe a very small effect of little consequence in a large sample, we may sometimes conclude that there is no effect or not one that matters.

Examples of two loss functions can be found in figure 2.1. Both the red line and the blue line signify two possible loss functions of being wrong. The outcome is best when it is exactly right. In that case, the loss is 0. However, if the deviation increases in either direction, the loss increases. The red loss function signals that increased error results in exponentially more losses. In other words, being a little off is not so bad, but being off by a lot will make an exponentially large difference. In some cases, loss functions are asymmetrical: underestimation may be better than overestimation, or vice versa, underestimation may be worse. For instance, it may be better to overestimate a loss in GDP in times of a recession than to underestimate it, because an overestimation will result in government action, while an underestimation may not. It may be better to do too much than too little. While the two example loss functions in Figure 2.1 are symmetrical, signifying that over- and under-estimations are equally bad.

Importantly, loss functions represent **value-judgements**: they are

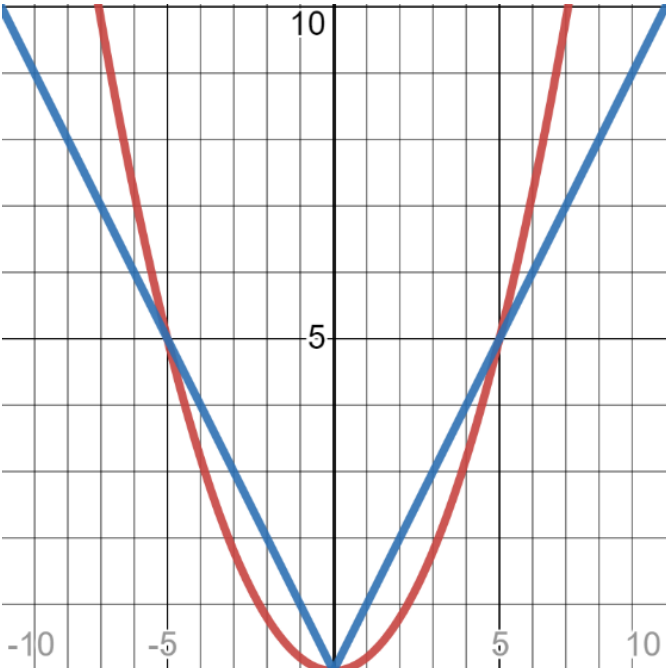


FIGURE 2.1 Two loss functions

not provided by the statistical theory itself. They are nevertheless important when we interpret the results. This insight is not new. McCloskey and Ziliak took it from two statisticians that we saw in the previous chapter: Jerzy Neyman and Egon Pearson. McCloskey and Ziliak cite them as follows:

> “Is it more serious to convict an innocent man or to acquit a guilty? That will depend on the consequences of the error; is the punishment death or fine; what is the danger to the community of released criminals; what are the current ethical views on punishment? From the point of view of mathematical theory all that we can do is to show how the risk of errors may be controlled and minimized. The use of these statistical tools in any given case, in determining just how the balance should be struck,

> must be left to the investigator" (Neyman and Pearson 1933, italics by; McCloskey and Ziliak 1996).

One of their students, Abraham Wald (who you might now from the Wald test), writes:

> "The statistician who wants to test certain hypotheses must first determine the relative importance of all possible errors, which will depend on the special purposes of his investigation" (Wald 1939, italics by; McCloskey and Ziliak 1996).

Both quotes get at something important: we cannot decide whether to accept or reject a conclusion based on the p-value alone. This is especially important for determining the required cut-off point, α, for our p-value. We need to know what the consequences are of being right or wrong, before we can say what the required cut-off value for p should be.

To summarize: according to McCloskey and Ziliak, econometrics should not just be used to identify statistically significant findings, but rather, it should assess the economic importance of the statistical differences from hypotheses. This has an important implication: If McCloskey and Ziliak are right, econometrics is not merely **technical** value-free practice: whether to reject or not does not only depends on the appropriate use of statistical methods, but also on the **values** of the loss function. Such values cannot be derived from data or from mathematical theory. Take, for instance, this question about loss: is it worse to conclude that there is no negative effect of minimum wages on employment, even if there may still be one, or is it worse to conclude the opposite, that there is a negative effect, even though there may not be one? This question is not just technical: it is at least partly determined by how bad we think it is to be unemployed and how bad it is to work for a low wage (as the quote by Wald illustrates well).

6 • ARE THINGS REALLY SO BAD?

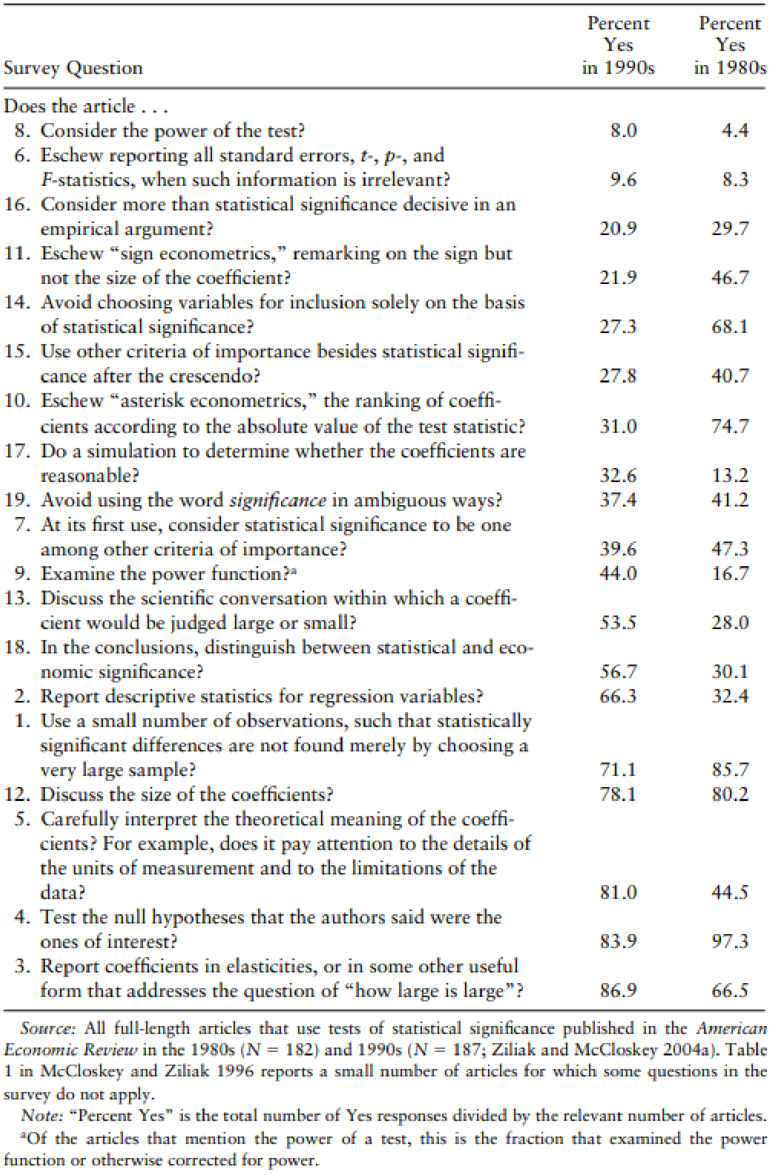
McCloskey and Ziliak have conducted some empirical research to analyze how often the interpretation problems that they are concerned occur. To do so, they analyzed papers published in the American Economic Review (AER). They analyzed papers from the 1980s and the 1990s (their book on the matter was published in 2008). What they find is summarized well in Table 2.1, which is taken from their book (Ziliak and McCloskey 2008, 81).

Things do not look good. In the 1990's, 62.8% (100-37.2) use the term significance in ways that is ambiguous: it is unclear whether it refers to economic or statistical significance. Furthermore, 21.9% (100-78.1) do not discuss the size of the coefficients that they have found, and 19% do not interpret the meaning of the coefficient. And remember, the American Economic Review is probably the most renowned journal of economics in the world.

It is clear why some of these findings are problematic to McCloskey and Ziliak, given the explanation that we have seen above. However, this may not be true for all of the problems that they observe. It is good to note here, that some of their criteria can also be seen as **controversial**. Not everything that they consider problematic is necessarily so. There is room for debate (see for example Hoover and Siegler 2008 for a critical analysis of some of these criteria). Let's look at some of them in detail.

> Question 11: Does the paper avoid "sign econometrics", remarking on the sign, but not on the size of the coefficient?

McCloskey and Ziliak find that only 19% of the articles in the 1990's in the AER *avoid* doing "sign economics". However, they argue, a sign only matters if the magnitude of a found effect is large enough. If not, a sign is not particularly important, and says nothing about economic significance.



Survey Question	Percent Yes in 1990s	Percent Yes in 1980s
Does the article . . .		
8. Consider the power of the test?	8.0	4.4
6. Eschew reporting all standard errors, *t*-, *p*-, and *F*-statistics, when such information is irrelevant?	9.6	8.3
16. Consider more than statistical significance decisive in an empirical argument?	20.9	29.7
11. Eschew "sign econometrics," remarking on the sign but not the size of the coefficient?	21.9	46.7
14. Avoid choosing variables for inclusion solely on the basis of statistical significance?	27.3	68.1
15. Use other criteria of importance besides statistical significance after the crescendo?	27.8	40.7
10. Eschew "asterisk econometrics," the ranking of coefficients according to the absolute value of the test statistic?	31.0	74.7
17. Do a simulation to determine whether the coefficients are reasonable?	32.6	13.2
19. Avoid using the word *significance* in ambiguous ways?	37.4	41.2
7. At its first use, consider statistical significance to be one among other criteria of importance?	39.6	47.3
9. Examine the power function?[a]	44.0	16.7
13. Discuss the scientific conversation within which a coefficient would be judged large or small?	53.5	28.0
18. In the conclusions, distinguish between statistical and economic significance?	56.7	30.1
2. Report descriptive statistics for regression variables?	66.3	32.4
1. Use a small number of observations, such that statistically significant differences are not found merely by choosing a very large sample?	71.1	85.7
12. Discuss the size of the coefficients?	78.1	80.2
5. Carefully interpret the theoretical meaning of the coefficients? For example, does it pay attention to the details of the units of measurement and to the limitations of the data?	81.0	44.5
4. Test the null hypotheses that the authors said were the ones of interest?	83.9	97.3
3. Report coefficients in elasticities, or in some other useful form that addresses the question of "how large is large"?	86.9	66.5

Source: All full-length articles that use tests of statistical significance published in the *American Economic Review* in the 1980s (N = 182) and 1990s (N = 187; Ziliak and McCloskey 2004a). Table 1 in McCloskey and Ziliak 1996 reports a small number of articles for which some questions in the survey do not apply.

Note: "Percent Yes" is the total number of Yes responses divided by the relevant number of articles.

[a] Of the articles that mention the power of a test, this is the fraction that examined the power function or otherwise corrected for power.

Table 2.1: Errors that McCloskey and Zilliak report in papers that they analyzed from the American Economic Review (from Ziliak and McCloskey 2008, tbl. 7.1)

> Question 8: Does the paper consider the power of the test?

The power of a test is $1-\beta$, where β is the probability of making a type 2 error (incorrectly not rejecting a true null hypothesis). While the results improved from the 1980s until the 1990s, only 8% of the papers in the AER consider the power of a test. Why is this an important problem according to McCloskey and Ziliak? Without power, we cannot calculate the uncertainty in our loss function. If we do not reject a hypothesis, how likely is it that we failed to correct a false hypothesis? There is a problem here that they acknowledge: we can only calculate the power of a test if we have a specific alternative hypothesis, something that is quite rare in most economics articles (generally, the alternative hypothesis is a non-specific one-sided, or two-sided alternative). However, because sample size is correlated to power, we can make inferences about power, and when we consider whether we should reject a hypothesis or not, it would be good to discuss this explicitly.

> Question 14: Does the paper avoid choosing variables for inclusion solely on the basis of statistical significance?

While this is a very common practice in economics (only 1 in 4 papers avoided doing this in the 1990's) there is no clear economic reason for this. In fact, even variables with coefficients insignificantly different from zero, that are genuinely unimportant in explaining a variable of interest, may have important interactions with variables of interest. Recall that this is exactly the error that Becker et al. made, in the article about cigarette addiction we discussed above.

> Question 10: Does the paper eschew "asterisk economics," the ranking of coefficients according to the absolute value of the test statistic?

Only 38.2% of the AER papers avoided asterisk economics. Asterisk economics occurs, for instance, when economists analyze p-values smaller than 0.01 as more important than p-values smaller than 0.05.

Why is it bad? Think about the reasoning behind significance testing that we have discussed in Chapter 1: we set a null hypothesis and rejection rate beforehand, and afterwards analyze if the p-value passes the critical value or not. Should it matter to the inference whether the p-value is smaller than the critical value, or much smaller? And does it signify anything that the p-value of one coefficient is larger than that of another? According to the logic of significance testing, these things should not matter. If it passes our critical value, that is all that matters, then we can reject the null hypothesis.

McCloskey and Ziliak take this practice to be a sign of making p-values too important: we have focused on that value so much that we have come to think of a lower p-value as better than a higher one. In fact, the p-value is only there to suggest whether there is a statistical deviation from the null or not. And if there is, we should focus on other things, such as the effect size. By ranking the p-values, we ascribe more value to the statistic than the inductive reasoning behind the statistic warrants.

> Question 1: Does the paper a small number of observations, such that statistically significant differences are not found merely by choosing a very large sample?

This is perhaps the most interesting, odd, and controversial criterion that McCloskey and Ziliak discuss. A common idea in statistics is that *more information*, more data, *is always better*. However, McCloskey and Ziliak seem to disagree: they think that economists are in error when they use a large number of observations instead of a small number. The reason for this is that the more data is used, the more likely it is we will find statistically significant results. Using a large data set thus seems to clash with Popper's idea of science: if economists want to reject a null hypothesis, and merely finding enough data makes it very likely to do so, even if the hypothesis is (roughly) right, perhaps economists are making things **too easy for themselves**. Think about the examples used above. If a coin is almost perfectly fair, but just a tiny bit biased, a large enough sample will automatically pick this out.

So, given that there will almost always be at least a small deviation from the null, a large sample almost always effectively guarantees statistically significant results.

Still, should we really prefer to have less data than more? Surely, that cannot be right? What would McCloskey and Ziliak say here? Of course, I do not know what they would say, but here is something that we can learn from this discussion: more data is of course better than less data, but if we focus on statistical significance as a criterion of scientific relevance, we should be very careful with drawing the conclusion that a significant finding signifies something of importance. In other words, if we are focusing too much on statistical significance as a criterion of economic significance, then large sample sizes may prompt us to draw highly misleading inferences. So, large sample sizes are good, because they help us estimate values of interest more precisely. However, if we do have a large sample size, our statistical significance test becomes less informative. If our sample size is large, it is all the more important to consider the size of our coefficients, and a significant p-value becomes less remarkable.

7 • CONCLUSION: IS ECONOMICS A CULT OF STATISTICAL SIGNIFICANCE?

So, what have we learned? First, we should be very careful with using statistical significance as an automatic criterion for relevance. Some significant results about economically relevant hypotheses may not be important themselves, because they are too small. Some non-significant results may be highly relevant from an economic perspective, even though their insignificance gives us reason to interpret them as highly uncertain outcomes.

A second thing we can learn from McCloskey and Ziliak, is that individual economists, using econometric techniques, often fail to appreciate the difference between statistical significance and economic significance, and consequently tend to overuse statistical significance as a criterion.

A third important idea that we have seen, is that the decision of which economic conclusions to accept based on the data is not merely a technical matter. It should involve another question: how important is the difference – is it a large difference, and would this magnitude make an important difference to our theories and policies? This may seem like a radical idea. After all, it is not only an acceptance of the claim that economics is not objective, but an endorsement thereof. Moreover, it incorporates an important subjective element in the study of economic data: **values**.

The idea that econometrics cannot escape subjectivity will be a theme in the next two chapters. In this chapter, we have discussed what may go wrong if economists misinterpret, or misuse the class book statistical method. In the next chapter, we will look at problems that do not have to do with the interpretation of the method, but instead with the method itself. Is significance testing really the best way to learn from the data? In Chapter 4, we will discuss an unapologetically **subjective** alternative: **Bayesianism**.

LEARNING GOALS FOR THIS CHAPTER

After studying this chapter, you should be able to explain:

1. the meaning of a p-value, and why the three examples of misinterpretations are indeed misinterpretations;
2. the difference between statistical and economic significance, and the arguments about why statistical significance is neither sufficient nor necessary for economic significance;
3. what a loss function is and why it is important for statistical inference;
4. why the problems that McCloskey and Ziliak discuss (asterisk econometrics, sign econometrics, not considering power, including variables based only on statistical significance, and using p-values when there is a large sample size) are problematic.

3 Problems for Significance Testing & Severe Testing

In Chapter 1, we discussed the relationship between significance testing, the problem of induction, and Popper's falsification argument. We saw that the logic of significance testing follows a similar logic as Popper's falsification method, but with important differences. Overall, we concluded that the logic of significance testing is at least supposed to work as follows.

1. If we correctly model the statistical process that generated the data (i.e. our technical and philosophical assumptions are correct), and
2. we perform the significance testing procedure correctly, and
3. we reject our null hypothesis, then
4. we can interpret the test as evidence against the hypothesis.

Importantly, there is a fifth step.

5. If we do not reject our null hypothesis, we cannot accept the null hypothesis.

In Chapter 2, we looked at the problem that economists and econometricians do not always correctly use this logic: they do not always follow the "official" reasoning, and derive conclusions that cannot be derived on the basis of this logic. In this chapter, we will look at this logic itself: if step 1, 2, and 3, are conducted correctly, *and* economists interpret the results correctly, is this a good way to get from the data to the truth?

We will identify **three serious philosophical problems** for significance testing. We will then look more closely at the logic behind significance testing, and see if there is something we can do to salvage it.

I • HOW ABOUT EVIDENCE IN FAVOR?

The first problem concerns step 5, mentioned above. It is also a problem for Popper's falsificationism. The key of the problem is this: according to both significance testing and falsificationism, evidence that fits with a theory or hypothesis never counts as evidence *in favor of a hypothesis*. This has an important rationale: avoiding the problem of induction. But, as we will see, it has some strange implications, which are difficult to defend.

To see this problem, ask yourself what happens when a p-value is larger than α? What if we find a p-value of 0.06 with an α of 0.05? We fail to reject. But what *do* we believe now? *Officially*, the logic of significance testing is conditional: **if** we reject our null, then we have found evidence. But what if we do not reject? Officially, the answer is that nothing happens. A p-value higher than 0.05 is not evidence *for* anything. It does not provide us with any reason to adjust our confidence in the null hypothesis, H_0. But this seems problematic for two reasons.

Common sense

First of all, it is intuitive that if we put a null hypothesis through a statistical test, and it is not rejected, we can be a little more confident in it. Consider the following example.

A friend tells you that she has seen a marvelous magician, who is not just doing tricks, but can actually read minds. Your interest is aroused, but you remain skeptical. As it turns out, a scientist has recently conducted an experiment with the magician, in which the scientists asked participants to think of one of 6 specific colors, after which the magician had to guess the color. The null hypothesis is that the magician has no special abilities, and that he will guess the colors correctly in one out of six of the cases. As it turns out, after a full day of testing with 50 participants, who all thought of a color 10 times, the scientist could not reject the null hypothesis.

What should this do our confidence in the belief that the magician cannot read minds (i.e. our belief in the null hypothesis)? Again, officially, the logic of significance testing tells us that we should just **fail to reject**, and that we cannot conclude anything from a p-value higher than α(higher than .05). However, it seems that we *should* now have a *higher confidence* in the hypothesis that the magician cannot read minds. Anytime I read about a magician claiming to read minds that fails to perform better than chance, my confidence that such a magician cannot read minds, and that mind reading does not exist, *increases*. However, according to the official logic of significance testing, this is not a proper conclusion.

The first way to phrase this problem is thus: common sense tells us that non-significant evidence is sometimes evidence *in favor* of a null hypothesis, but according to significance testing, we are never allowed to count non-significant evidence as evidence in favor of the null hypothesis. Is our commonsense judgment wrong, or is the principle of significance testing wrong?

The Principle of Total Evidence

There is a further reason to be skeptical about the idea that significance testing provides evidence only if it leads to a rejection: the **Principle of Total Evidence**. This principle states that if we judge what we should believe, we should use *all* the available evidence. This principle seems to be very reasonable, if not obvious. Surely, if we are rational, we should not ignore evidence when we want to make a judgement about what to believe. Scientists, policymakers, court judges and juries, and anyone who is evaluating evidence, should consider *all* evidence, not only a part of it.

The logic of significance testing appears to conflict with the principle of total evidence in cases that we fail to reject. Think about the magician case: when we are evaluating what to believe about the magician, we should incorporate the evidence that failed to reject the hypothesis that his correct guesses are due to chance. But, according to the logic of significance testing, when we fail to reject, the test is not evidence *for* anything. The logic of significance testing tells us to disregard evidence if it is non-significant, even if the quality of the evidence is good.

An easy fix?

Now, you may think that this is a problem that is easy to fix. We simply adjust the logic of significance testing: if $p > \alpha$, we consider it evidence in favor of the null hypothesis. This may be a good suggestion, but it is important to see that this is more complicated than it sounds. This solution faces two problems of its own.

First, it changes the logic of significance testing in an important way: it loses its similarity to Popper's falsification logic. This shared logic was there for an important reason: the problem of induction shows that evidence can never verify or confirm a hypothesis.

Second, where do we draw the line of when a p-value is evidence *in favor* of a null hypothesis? A p-value of 0.06 is hardly evidence for the null hypothesis. In fact, we may perhaps still interpret it as evidence *against* it. It is, after all, significant at a 0.1 level. How about a p-value of 0.06, or of 0.09, or of 0.11? Perhaps they are neither evidence for, nor evidence against a null hypothesis? But at what point does it become evidence? 0.2, 0.3, 0.4? This is unclear. In all of these cases, there is still a deviation from the null that we would only expect to occur sometimes. That, by itself, is not so clearly interpretable as evidence *in favor* of a hypothesis.

Recall that in Chapter 2, we discussed the problem that, sometimes, non-significant tests still count as evidence *against* the null hypothesis. Recall the Remdesivir example, the medication that showed some, non-significant positive results in treating the Covid-19 disease. In this case, non-significant evidence still seems to count (a little bit) against the null hypothesis that Remdesivir does not work at all.

The problem is thus this: highly significant results are clearly evidence against the null hypothesis, but slightly insignificant results may still count against them. Sometimes, non-significant results also count in favor of the null hypothesis, but there is no simple cut-off value at which the non-significant evidence becomes evidence in favor of the hypothesis.

In summary, the problem is not easily fixed. We will discuss a more serious attempt at a solution below.

2 • THE LOGIC FAILS

Unlikely events happen. There is a widely publicized story of a newlywed British couple that found a photo in one of their old photo albums in which they are building sand castles meters apart, years before they would meet (Ogrodnik 2014). The odds that you would marry an unidentified person that happened to end up in a photo in one of your old photo albums is not very high. Still, stories like this are not all that rare.

This results a serious problem for the logic of significance testing. Recall from Chapter 1 that the logic of significance testing was similar to, but also different from the logic of Popper's falsificationism. According to Popper, we should look for evidence that is *inconsistent* with a theory, in order to disprove it. If we have done so, we can know that the theory is incorrect. This logic is infallible: if we find evidence that is inconsistent with a theory, the theory *must* be wrong.

The logic of significance testing, on the other hand, cannot give us *inconsistencies,* because statistics are always probabilistic. It can only tell us that if a null hypothesis would be true, the observed evidence is very *improbable* to occur. The logic behind significance testing is thus: if, given a theory A, it would be very improbable to observe what we observe, we have good evidence against A. In a slogan, we can put this as follows: "evidence improbable according to a theory, then, evidence against this theory". This logic may appear to be solid, but it fails sometimes.

For example, we may have a theory that Jack is a football player (see Dickson and Baird 2011). Given that Jack is a football player, it is quite unlikely that he is a goalkeeper (about 1/11, p<.1). But, if we learn that Jack is a goalkeeper, this is *not* evidence that he is not a football player. In fact, the opposite is true.

This may seem like an unfair example, because a goalkeeper just is a soccer player. But, consider another example of why the logic of "evidence improbable according to a theory, then, evidence against this theory" misapplies. Bart won a lottery and believes that this was because "the universe wanted him to win". His logic is as follows: "I had a really bad day and asked the universe to help me. I bought a ticket never expecting to win the jackpot. After all, the probability that I would win the jackpot on coincidence is so low ($p < 0.0000001$), we would never expect it to happen. However, I won, and I really needed it. I can only explain it by thinking that the universe wanted me to win." There is something deeply unscientific about Bart's reasoning, but, there is nothing wrong with Bart's usage of the logic of significance testing. We can construct the hypotheses as follows, with $H0$ the null hypothesis and H_1 the alternative hypothesis.

H_0: luck determines whether Bart will win or not, and

H_1: the universe determines whether Bart will win.

The evidence (Bart winning) is highly improbable given H_0, while it is consistent with $H1$. So, is Bart's win evidence that the universe makes "decisions" about who will win the lottery? It obviously is not, but it does appear to follow the logic of significance testing.

We can think of another example: life on earth. According to our best theories of astronomy and biology, it is very unlikely that life develops on any planet. However, life has developed on earth. Following the logic behind significance testing, we have strong evidence that our best theories of astronomy and biology are therefore wrong.

Does that mean that the logic of significance testing is wrong? Its defenders can appeal to an important condition that should be satisfied before significance testing can take place: we should not formulate hypotheses *after* observing the data. Bart only started to consider the lottery as evidence for the hypothesis after he won. But, if he had conducted a real experiment, in which he first formulated the hypothesis and then saw the lottery result as a datum to test this hypothesis, the result would, in all likelihood, have failed this test.

Despite this rebuttal, the flaw in the logic behind significance testing is of serious concern. Unlikely events, such as an individual winning the lottery or the occurrence of life on earth, happen quite often. This does not disprove the theories that imply that these events are unlikely to occur. Yet, it is foundational to significance testing that, if a theory states that some event is improbable, the occurrence of this event is evidence against the theory.

3 • PRIOR PROBABILITY MATTERS

A final problem is that sometimes a p-value is very small, but we should still believe in the null hypothesis. In order to see this, let's go back to the magician example, but let's this time assume that the magician did pass the test. He guessed a statistically significant number of colors correctly. The most plausible explanation is still that he got very lucky. The test would be significant, but we should still not believe him. The reason for this is that it is just very unlikely that some magician is truly clairvoyant.

The final problem for significance testing is due to this aspect: according to the logic of significance testing, we only take into account the p-value of a test, but it does not matter whether the hypothesis under scrutiny is plausible or not. We can put this as follows: what we believe after observing the data should not only be determined by p-values (i.e. the data in relation to the hypothesis), but also by the **prior probabilities** of these hypotheses: how likely was the null hypothesis under consideration to begin with? If a hypothesis is very likely to be true, we should need much stronger evidence to reject it than if we are rejecting a incredibly unlikely hypothesis (e.g. that the universe determines who wins the lottery, or that people can read minds).

In Chapter 2, we discussed the psychological research investigating the hypothesis that people cannot see into the future (H_0) against the hypothesis that they can, under certain circumstances. In particular, this research gave individuals the task to choose whether they thought a certain image on a computer screen would pop-up left or right. In this case, whether or not we end up rejecting the hypothesis that our guesses will be random (i.e. not based on an ability to foresee the future), requires some pretty strong evidence. Not any significance test (with $\alpha$of .1, .05, or, .01) should be able to convince us here. Significance testing only gives us a binary choice: reject or not. However, if a hypothesis is very plausible (e.g. clairvoyance does not exist), we should need a little bit more than 1 significant test involving a sample of some small group of individuals to change our mind. One small p-value ($p < 0.01$) should not automatically lead to the rejection of the hypothesis. But significance testing says that it should.

The probability of H_0 or H_a prior to the evidence should thus make an important difference in the choice to reject H_0.

Base-rate Fallacy

We can show that prior probability matters by looking at numerical examples. One numerical example that illustrates this, is the **base rate fallacy**:

> Eve is in the hospital, being tested for a disease. The test that is used, has a high reliability, namely of 98%. That means that, if a person does not have the disease, it will come out as positive (i.e. as an indication that the person does have the disease) in only 2% of the cases. If a test is negative, the person can be sure that they do not have the disease: the false negative rate is 0%. This is similar to using a p-value .02, and having a power of 100%, under the null hypothesis that the person does not have the disease. In Eve's case, the test comes out as positive. This sounds like really bad news for Eve: the disease is quite awful, and it may take her a year to recover. However,

TABLE 3.1 The base-rate fallacy

	Has the disease	Does not have the disease	Total
Test positive	10	2,000	2,010
Test negative	0	97,990	97,990
Total	10	99,990	100,000

> now consider the fact that the disease is very rare: only 1 in 10,000 people suffer from it. If, in her town of 100,000 people, everyone were tested, we would expect to find the following results.
>
> That means that out of all the people who test positive for the disease (2,010), only .5% (10 individuals) have the actual disease. So, even though the test has high reliability and power, the chance of actually having the disease when the test comes out as positive is very small. How can this happen? The reason is that the base rate, i.e. the prior probability of having the disease, was extremely low.

In practice, we do not always know how many people have a disease, or more generally, how likely the null hypothesis is before we look at the data. However, the same logic that applies to the numerical case of the base rate fallacy also applies to cases in which we do not know for certain how likely the null hypothesis is before seeing the evidence. Even if we did not know that only 1 in 10,000 people has the disease, but rather that the disease was simply "extremely rare", we should have still been very careful with drawing conclusions on the positive test. The example shows that the prior probability matters in our assessment of statistical evidence. A rejection ($p < 0.05$) should not always mean that a null hypothesis is false. Whether it does depends on how likely it is that hypothesis was true from the start.

Lucia de Berk

The importance of considering the prior probability when evaluating statistical evidence is apparent when we consider a notorious court case in The Netherlands: the case of Lucia de Berk. De Berk was a nurse in a children's hospital, and was accused of playing a role in many of the deaths that occurred there. One of the main arguments for the accusation was statistical: Lucia was present at many of the incidents in the hospital. In fact, the co-occurrence of her presence and the incidents was statistically significant under the null hypothesis of it being a coincidence. Is this good evidence that Lucia de Berk is indeed a murderer?

To know this, it is not enough that there is a statistically significant co-occurrence between Lucia de Berk's shifts and hospital incidents. We also need to know what the probability is that a nurse working in a children's hospital is a murderer. That probability is very low. Thus, if we were testing whether her higher-than-average presence at deaths and near-deaths was due to chance, with the alternative being that she was a murderer, we should need strong evidence to convince us.

Lucia de Berk was in fact convicted of murdering patients in the hospital. After the conviction, mathematician Richard Gill played a crucial role as an expert witness and activist on behalf of Lucia de Berk. He argued that the proper p-value for the null hypothesis that Lucia's presence during the incidents at the hospital was solely due to chance if she were innocent, was 1/49 ($p = 0.02$), and thus statistically significant. He argued, however, that to determine to what extent the number of co-occurrences was evidence for the hypothesis that Lucia de Berk is a murderer, we needed to consider prior probability: how likely was it that she was a murderer before consideration of the evidence? If Dutch nurses in children's hospitals have a very low probability of being murderers, it may still be highly unlikely that she was a murderer, even with a p-value of .02 ($p < 0.05$) for the null hypothesis that her co-occurrence with the incidents is due to chance. This argument played a crucial role in Lucia de Berk's exoneration in 2010.

4 • SUB-CONCLUSION

All of the problems discussed in this chapter have something in common. Ultimately, we are interested in the question what the data should tell us to believe: which hypotheses are likely true, and which ones are not? Ultimately, we are interested in this probability:

$P(H|E)$: the probability that the hypothesis is true, given the evidence (or the data).

However, significance testing only provides us with another probability: the probability that we observed a discrepancy between the data and the hypothesis or a discrepancy that is even larger. Roughly speaking, this is an estimate of P(E|H).[3]

$P(E|H)$: the probability that we see the observed data if the hypothesis is true

The implicit underlying idea here appears to be that if $P(E|H)$ is small, then, it seems, $P(H|E)$ must also be small. However, the three arguments above show that a small P(E|H) does not imply the $P(H|E)$ is small. The two are still linked. Indeed, a low $P(E|H)$ in most circumstances indicates that we should decrease our confidence in H. For instance, if I think you are not a football enthusiast ($P(H)$ is low), but I then observe that you have three football shirts in your house, E, which is quite unlikely if you are not a football fan, I should decrease my confidence in my belief that you are not a football enthusiast. However, as discussion in this chapter shows, this does not apply as universally as we might think.

5 • RESEARCH DESIGN AND SEVERE TESTING

In the next chapter, we will see that some philosophers, statisticians, and econometricians think that the Principle of Total Evidence, the flawed logic behind significance testing, and the importance of prior probability provide a strong reason to move away from classical statistics. Instead of using significance testing as our main tool in statistical

[3] Strictly speaking, it tells us: P(d≥D|H), where d is a difference from the hypothesized value, and D is the observed difference.

reasoning, we should do something else. This conclusion, however, may be too hasty: significance testing can be salvaged. We can address all three problems through **research design**. **Deborah Mayo**, a philosopher of statistics, has defended some of the logic of significance testing by developing these ideas. According to her, the idea of significance testing does work, but it only works if we take into account the underlying logic of falsification: a test is only a test if it is hard to pass. The true motivation behind significance testing is not the statistical test itself, but a more fundamental idea. What truly matters is that a scientific idea passes a difficult test. She calls this **severe testing:**

Severe testing: if a hypothesis has a really good chance of not being rejected if it were true, then, a rejection is good evidence against the hypothesis.

Think about the test the magician underwent. If he truly did not have magical abilities, it would have been be very, very difficult for him to pass the test: the test was severe. And, accordingly, the magician failed the test. This fundamental idea can help us explain why significance testing sometimes does not work. The logic of significance testing looks very similar to the logic of severity. After all, low p-values indicate that the data would have been very difficult to have been brought about if the hypothesis is true. However, low p-values, as we have seen, can sometimes be a result of procedures that are *not very severe*. The concept of severe testing can also be reversed, when tests are not severe, they are insevere:

Insevere testing: If data **x** agree with a hypothesis H, but the method was practically incapable of finding flaws with H even if they exist, then **x** is poor evidence for H.

The case of the lottery winner was different from the magician example. One reason why Bart's winning of the lottery failed to be convincing evidence of the hypothesis that the universe made Bart win, is that the test was not very severe. This is not a good research design. The test was only formulated **after the evidence** was brought about, and it seems, the formulation of the hypothesis was partly based on the evidence ("oh, I won after I prayed to the universe, the universe must have made me win"). The test is thus very insevere.

When is a test a severe test? Severity is not just the relationship between some data and a hypothesis, as the p-value is. It is the relationship between 1) a research setup, 2) the evidence that rolls out of the research setup (i.e. the data), and 3) a hypothesis. In order to make sure a test is severe, we need to formulate the hypothesis first, and gather the data later. If this happens appropriately, we should at least observe our low p-value as a good reason to reduce our confidence in our null hypothesis, even if our confidence remains high. In case of the base rate fallacy, after observing a positive test statistic, we should still have a very high confidence that we are not ill. But, this probability is at least a lot reduced compared to the situation before the test.

In econometric practice, severity makes statistical inference more reliable. We implement severity in the following ways.

- Make sure that we formulate a hypothesis before we look at the data. If we let the data affect our hypotheses, and then we test those hypotheses on the same data, the test is not severe (Chapter 5).
- Do all the right diagnostic tests and robustness tests.
- Avoid doing multiple tests, only to then select the ones that are significant (multiple testing). We know that the more tests we run, the more likely it is that we will find low p-values. But as we shall see later in the book, avoiding multiple testing may be more difficult than it seems. We will look at this idea more closely in Chapter 6.

6 • CONCLUSION: WHERE THIS LEAVES US

The logic of significance testing faces some problems: 1) it can only tell us what not to believe, even if the data give us reason to believe certain things, 2) the logic fails in certain cases, and 3) the prior probability of a hypothesis matters in its evaluation, while hypothesis testing does not take it into account.

The logic of severity is a better alternative. However, does severity solve all the problems of significance testing? We can avoid the second problem of significance testing (**problem 2, the logic fails**), by simply making sure that our statistical tests are truly severe tests. The logic of significance testing may fail, but the logic of severity does not. Mayo even asserts that if a hypothesis passes a severe statistical test, this also counts as evidence **in favor** of the hypothesis. Therefore, it also deals with **problem 1 (How about evidence in favor?)**. So, the severe test that the magician went through, not only fails to be evidence for the mind-reading abilities of the magician, but also provides **positive** evidence for the hypothesis that he cannot read minds. But it does not tell us how much evidence it provides, and how strongly we now should believe in this hypothesis. So, our main problem (**problem 3, prior probability matters**) remains:

> **Main remaining problem**: we ultimately want to know *what we should believe* based on the data, and even severe tests only tell us that something *counts against* or *counts in favor* of a hypothesis, but not what we can conclude about the probability of the hypothesis itself: **P(H|E)**.

We are left with an important question: p-values may tell us that data provides a good reason to reduce our beliefs in a hypothesis, but what *should* we believe? In the next chapter, we will look at a radically different way of looking at statistical evidence that does take this into account.

LEARNING GOALS FOR THIS CHAPTER:

After having studies this chapter, you should be able to explain

1. The three different philosophical problems for the logic of significance testing. You should be able to explain what they are, why they pose a challenge to the logic of significance testing, and you should be able to provide some examples that illustrate the problem.
2. You should be able to explain why severe testing is a possible way to salvage the basic logic of significance testing

4 Bayesianism

So far, we have looked at classical statistical reasoning: the significance testing tools that are so widespread among contemporary social and natural sciences. In this chapter, we will look at an alternative, radically different theory of evidence, called **Bayesianism**. According to the Bayesian inferential logic, we should not ask the following question.

1. How likely is it that we find data that is so different from our expected value under the null hypothesis X?

Rather, Bayesianism proposes that we ask this question.

2. How likely is hypothesis X, in light of the available evidence?

As we shall see, answering Question 2 is more in line with what scientists, economists, and other users of statistics generally want to know. However, answering question 2 not only deviates substantially from significance testing, but also comes with a variety of challenges on its own.

1 • BAYES' THEOREM

How to calculate the probability that a hypothesis is true? The key mathematical formula underlying Bayesianism is Bayes' theorem. It looks as follows.

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

Bayes' Theorem tells us how we can calculate the probability a hypothesis is true, given the evidence.

In the previous chapter, we have actually already seen this formula in action, though we did not state it explicitly. It was used to illustrate an example in which a low $P(E|H)$ was compatible with a very high $P(H|E)$: the base rate fallacy. In this example, we knew exactly how to calculate $P(H|E)$.

$$\begin{aligned} P(H|E) &= 0.9999 (\text{probability of not having the disease}) \\ &* \frac{0.02 \ (\text{probability of a false positive result})}{\frac{2010}{100000} \ (\text{probability of positive result})} \\ &\approx 0.995 \end{aligned}$$

This tells us that, if the chance of having a disease is 0.00001 (1 - 0.9999), and we get a positive test with a 0.02 false positive rate, the chance of having the disease is 0.005 (1 - 0.995). This example illustrates how three pieces of information lead us to draw a conclusion from the observed evidence, which is exactly what we are interested in. The three pieces of information are 1) the probability that we observe the evidence under the null P(E|H), 2) the general probability that we observe a piece of evidence, and 3) the probability of a hypothesis being true in the first place, in this case, us not having the disease.

Bayes' theorem is a mathematical theorem, first derived by a Presbyterian minister, Thomas Bayes. The formula itself is merely a mathematical statement about the calculation of conditional probabilities. It was not originally designed as a theorem about evidence and hypotheses. The formula, however, was later adopted by a group of statisticians who call themselves Bayesian statisticians, who think this formula can be used to think about statistical inference more generally.

According to Bayesian statisticians, we should interpret Bayes' formula as a tool to describe how a rational person would alter their beliefs after observing some evidence. In practice, this can look as follows. Say you currently believe that 50% of your classmates failed an exam, and you have no reason to believe that you did better or worse than the rest. However, you do know you got Question 1 right. You also know that only 40% got the first question right. Of those who passed, 70% got the first question right. You want to know $P(H|E)$, the probability that you passed, given the evidence, E, that you got Question 1 right. Before we consider the evidence, the probability that you passed, $P(H)$, is 50%. The probability that you got Question 1 right if you also passed, $P(E|H)$, is 70%. Finally, $P(E)$ represents the probability that any student in the class gets Question 1 right, which is 40%. This gives us:

$$P\left(H|E\right) = 0.5\,\frac{0.7}{0.4} = 0.88$$

So after learning that you got Question 1 right, you should have a lot more confidence that you passed the test. The probability went from 50% to 88%.

The Bayesian school of statistics believes that we can use exactly this logic in scientific inference. In brief, the key distinguishing feature from classical statistics is as follows: Bayesians believe that, rather than following the logic of significance testing, statisticians should aim to estimate the probability that a hypothesis is true, given the evidence $P(H|E)$, using Bayes' theorem.

Bayesianism is therefore different from classical statistics. It tells us not to stop at the conclusion that the observed data or something even more unexpected are either unlikely or not unlikely to occur under the null hypothesis, but to go further and to estimate the probability that a hypothesis is true.

2 • EXAMPLE: A WHODUNNIT

Bayesian statistics in econometric practice can get quite technical. To illustrate Bayes' theorem, we will look at a detective example. Imagine that you are a detective in the early stages of a murder investigation. There are three suspects: i, ii, and iii. Before more evidence comes in, you think there is a 20% chance that each of these individuals has committed the murder, and a 40% chance that none did it. New evidence comes in: police officers have found shoes of type X on the murder scene. Suspect ii wore these shoes on the night in question. We know that 80% of murderers leave footprints at the scene; the other 20% make sure to cover up the evidence. We also know that 5% of the population currently wears shoes of type X. What should we believe about the guilt of Suspect ii?

The hypothesis (H) here is that Suspect ii is the murderer. The evidence (E) is that shoes of type X were found at the murder scene. We know that the probability the suspect is the murderer before we observe the evidence is 20%. We also know that the chance of finding these footprints if Suspect ii is the murderer, is 80%.

$$\begin{aligned} P(H) &: \ 0.20 \\ P(E|H) &: \ 0.80 \end{aligned}$$

The probability of observing shoes X ($P(E)$) depends on whether or not Suspect ii is the murderer. We think there is a 20% chance that suspect ii did it, and if so, there is an 80% chance that the suspect left their footprints. If not, we know that 5% of the population also wears these shoes, and 80% of those will also leave footprints. So if Suspect ii did not do it, the probability of which is currently 100%-20% = 80%, there is an 80% of 5% probability that they left footprint X there. So:

$$P(E) \ = \ 0.20 * 0.80 \ + \ 0.80 * 0.05 * 0.8 = 0.192$$

We can now fill in Bayes' formula:

$$P(H|E) = \ P\left(H\right) \frac{P(E|H)}{P\left(E\right)} = 0.2 * \frac{0.8}{0.192} = 0.83$$

Whereas we thought that there was a 20% chance that suspect ii did it before we observed the footprints, and if our assumptions about probabilities are correct, we should now think that the chance that suspect ii did it is 83%. We'd better keep our eye out for that one!

3 • SUBJECTIVE AND FREQUENTIST PROBABILITIES

How can we apply Bayesian logic to scientific practice? A first important challenge to applying the logic in the base rate case to scientific inference is this: what is $P(H)$? This factor represents the probability that a hypothesis is true, before we have seen the evidence. But what does that even mean? A hypothesis is either true or false. The probability to roll a "1" with a fair die is 1/6, which means that if you would throw it a thousand times, you would expect 1/6 of the results to be "1". But if we extend this logic to hypotheses, it becomes absurd: we cannot "roll" a hypothesis a thousand times, and see the results...

In the base rate example, we were trying estimating what we should believe in light of the evidence, given that we know how likely it is that our hypothesis is true in general. In real life, however, econometricians do not know what the general probability of a hypothesis is before they see the data. This is exactly what they are trying to find out. Think about the detective example, we assumed that there is a 20% chance that suspect ii has committed the murder, before finding the evidence. What was that based on, and how could the detective know that?

Beliefs that scientists have about the probability is that a certain hypothesis is true, before they see the data, are **subjective** beliefs: they do not indicate an objective probability of something that is already known, but they indicate what a person, such as the scientist, *believes*. When a Bayesian claims there is a 74% probability that a hypothesis is true before seeing any evidence, this means: I believe that there is a 74% chance that the hypothesis is true. The detective, for instance, believed that there was a 20% chance that Suspect ii was guilty. Perhaps she thought so on the basis of the way Suspect ii looked, behaved, or smelled. But there was no objective way of knowing this. The same is true for scientific hypotheses, for example: what is the chance that an increase in minimum wage increases wages, before we have seen the evidence?

Probabilities, in the classical framework, were not subjective, but had a clear objective meaning: if the hypothesis is true, we would expect a result like this, or even more different from the null hypothesis, to occur with a **frequency of p times**, just like rolling a die a thousand times. Because of this interpretation, classical statistics is often called **Frequentist statistics**.

Take our magician examples again. What is the probability that some magician can predict the color a participant is thinking of before you have seen the evidence? You probably think that this is quite low. But, this is a **subjective belief**, it represents what *you* think is likely. If you would say that that probability is 1%, this simply means that *you think* it is highly unlikely the magician can do this. If someone else says there is a 3% chance, there is not a sense in which either one of you is necessarily correct, while the other is not.

Subjective probabilities seem fuzzy. Is that fair to say? A first criticism of the Bayesian school is that it bases itself on subjective probabilities, which are arguably not clearly defined. What does it really mean to say you believe something with 1%, 30% or 80%? Science, classical statisticians say, should be based on clearly defined concepts. However, Bayesians typically respond that the fact that a belief is subjective, does not mean that it is necessarily vague. They may give the following answer. A subjective probability of event A can be defined as the implied probability of the willingness to bet on event A by a risk-neutral person. For example, if you believe that there is a 30% chance that Ajax will beat Chelsea, then, you would be willing to bet up to €.30, if your return would be €1, but not more than that. So, when a risk-neutral person bets €.30 on Ajax beating Chelsea, when the return is €1, we know that their subjective belief that Ajax will win is 30%. There is nothing vague about that.

Bayesian statisticians also have a strong argument for the claim that people will generally hold probabilities that are **rational**. This argument is called the **Dutch book argument**. The argument goes like this. Suppose that you have an irrational set of subjective beliefs. So you would, for example, believe that the chance that Ajax will win is 30%, but the chance that Chelsea will win, or that the game will be a tie, is 75%. The odds do not add up. According to the argument, you would then be willing to bet €0.75 on Chelsea winning or tying, and €0.30 on Ajax winning if the payoff is €1,-. But if that is so, anyone can make

TABLE 4.1 A Dutch book bet

	Ajax wins	A tie, or Ajax loses
Bet 1: costs €0.30	net income €1 (prize) - €0.30 = €0.70	Net loss: €0.30
Bet 2: costs €0.75	net loss €0.75	net income €1 (prize) - €0.75 = €0.25
Total	€0.70 - €0.75 = -€0.05	€0.25 - €0.30 = -€0.05

a set of bets with you from which you will surely lose money (see Table 4.1). If you were to make this set of bets and Ajax would win, you would win €0.70 (€1 - €0.30) from the first bet, but you would lose €0.75 from the second bet: a total loss of €0.05. If the match ended in a tie or Chelsea won, you would win €0.25 from the second bet, but you would lose €0.30 on the first. Again, you are losing €0.05. This type of bet is called a **Dutch Book**: a bet from which you lose no matter what happens. This is of course a silly example: few people would actually fall for this trap even if their beliefs were inconsistent. However, the general idea behind this argument is that, if you do not have consistent beliefs, you will act accordingly and make irrational mistakes. Not updating according to Bayes' theorem, is one way in which you subjective probabilities may be **irrational**. Overall, the argument shows that although subjective probabilities are subjective, they do have to abide by rational principles, because otherwise individuals will face avoidable costs. The fact that they are subjective, does not mean that a rational person can have any subjective probabilities they like.

4 • INTERPRETATIONS OF STATISTICS

The key takeway from this discussion is that Bayesians think that probabilities are 1) **subjective beliefs on which we act**, that 2) **are located in our mind**. Frequentists, on the other hand, think that probabilities are 1) **objective frequencies**, that are 2) **located in the world itself** (i.e. outside of the mind). This difference between Frequentists and Bayesians is a disagreement about **the interpretations of probabilities**. Consider this example. The weather reporter says there is "an 80% chance that it will rain tomorrow". Frequentists think that means:

> **Frequentists (probability of rain tomorrow):** if there are 100 days that have weather conditions similar to what we have now, in 80 of those, we will see rain in the day after.

Bayesians, on the other hand, think it means this:

> **Bayesians (probability of rain tomorrow):** I have a strong belief in the fact that it will rain tomorrow. If I would want to make a risk-neutral bet about it, I would bet €8 if I could win €10 if it rains tomorrow.

These interpretations may look compatible. In many cases they are. It may both be true that I would bet €8 on there being rain tomorrow, if there would be a €10 payout, and it may also be true that if we have weather conditions like we have today, in 80% of the cases there will be rain tomorrow. Both can be true at the same time. However, there are certain claims about probability that only make sense when using one of the approaches. One example of this is the truth of hypotheses. For Frequentists, **hypotheses are either true or false**. For **Bayesians, hypotheses have probabilities**. When you say that ascribe a high probability to the hypothesis that you have passed the test, it means that you believe you have probably passed it. According to Frequentists, you have either passed it or not, but it is meaningless to ascribe a probability to any hypothesis.

Bayesianism is a general approach to inductive inference with broad applicability. Bayesians, for example, can analyze the swan problem that we discussed in Chapter 1: building on inductive logic, it seemed plausible that all swans are white, but as the European explorers discovered more of the world, they learned that this was false. According to Bayesianism, should people who have only seen white swans believe all swans are white?

Bayesianism would say that if our starting probability that not all swans are white ($P(H)$ is smaller than 1), and if the probability that any observed swan is white ($P(E)$ is not 100%), P(H|E) may be very large, but will never be 100%. We should therefore also think that there is some small probability that not all swans are white.

A question for you: what happens with $P(H|E)$ in Bayes' formula when we observe a black swan after having seen only white swans?[4]

To understand the approach better, let's first take a closer look at the different parts of Bayes' formula.

$P(H)$: *priors*

The probability of a hypothesis being true before observation of the evidence is called the **prior**. We have encountered prior probabilities before, when we were analyzing why we still should have high confidence in some hypotheses while there is evidence against them with low p-values. Because probabilities in science are not as clear as they are in the base rate examples, priors represent subjective probabilities.

$P(E|H)$: *the likelihood of the evidence*

This probability is similar but not identical to a p-value. Remember, the p-value is the probability that the difference between the expected

[4] The answer is that the $P(E_{\text{a black swan}} | H_{\text{all swans are white}}) = 0$. Hence, Bayes' formula would have as an output that $P(H|E) = 0$.

result of a test and the measured result is as great as observed or greater, if the null hypothesis is true. The likelihood of the evidence is something slightly different: it denotes the probability that we observe what we observe, given that the hypothesis is true. In certain cases, like the base rate example case, we can specify this likelihood well: the possibility of observing a positive finding, even if a person does not have the disease. However, in other contexts, the p-value can serve as an approximation of the likelihood of the evidence.

$P(E)$*: the probability of the evidence occurring, unconditionally*

This term represents the probability that we would observe the evidence in any case, regardless of whether the hypothesis is true or not. This boils down to the probability that we observe the evidence, conditional on all hypotheses and the likelihood that they are true:

$$\begin{aligned} P(E) &= P(E|H_1)\,P(H_1) \\ &\quad + P(E|H_2)\,P(H_2) + P(E|H_3)\,P(H_3) + \ \ldots \\ &= \sum_i P(E|\,H_i)\,P(H_i) \end{aligned}$$

It is important to see that not only the prior of the hypothesis of interest plays a role here, but *also* the priors of all the other possible hypotheses.

Consider the example of the exam we discussed earlier. We knew that 50% of the people passed the exam, and of those people, 70% got Question 1 right. We also knew that in total 40% got the first question right. So, we can calculate how many of the students who failed, got the question right:

$$P(E) = P(E|H_1)\,P(H_1) + (E|H_2)\,P(H_2)$$

where H_1 is the hypothesis that you passed, and H_2 the hypothesis that you did not pass. E is the event of getting Question 1 correct.

$$\begin{aligned} 0.40 &= 0.5 + X * 0.5 \\ X &= 0.1 \end{aligned}$$

So, 10% of the students who did not pass the test, got Question 1 right. Often, we will only know the probabilities on the right-hand side, so we can calculate P(E) from those.

$P(H|E)$: *the posterior*

This is the probability that scientists are ultimately interested in: the probability that our hypothesis is true given the evidence. It is the output of our formula. How should we interpret the result? It essentially states that if we have a prior probability of $P(E|H)$, and our other probabilities are also correctly specified, we should, *rationally*, believe that $P(H|E)$ is the probability that the hypothesis is true after the evidence. This posterior is itself the **new prior** for future research (see Figure 4.1).

6 • BAYESIAN UPDATING

Bayesianism, unlike significance testing, tells us that *any new piece of evidence* should have an impact on our belief in a hypothesis. Whenever a relevant piece of data or an observation comes in, we should calculate a new posterior from a prior. In Chapter 3, we discussed the principle of total evidence: the idea that all evidence should be used in the evaluation of a hypothesis or theory. Bayesianism respects this principle. The process of using evidence to calculate a posterior from a prior is called **Bayesian updating**. Essentially, Bayesian updating means that the **posterior becomes the new prior** in future research. So, if new evidence comes in again, we use the previous posterior as our new prior (see Figure 4.1).

The fact that Bayesians update **on all evidence** that comes in, is an important difference with classical statistics. According to classical

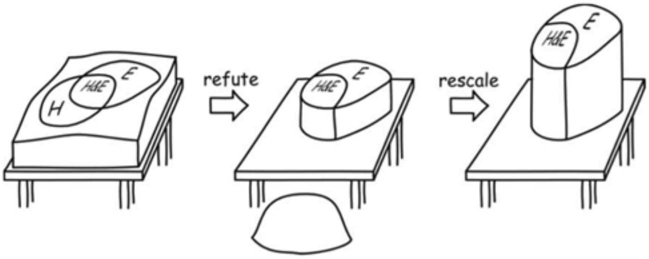


FIGURE 4.1 Bayesian updating. From Norton (2011)

statistics, we need to design a test beforehand. Only when all the data is in, can we evaluate the hypotheses. And, only when the data allow us to reject the null hypothesis, should we change our beliefs. Only then does it count as real evidence. Bayesianism tells us to update all the time. For example, every time we observe a white swan, this should affect our belief about the number of white swans, even if it does not come from a study of at least 30 individually randomly sampled swans.

7 • AN ECONOMETRIC EXAMPLE

Let's look at an econometric example: does a minimum wage increase unemployment? We may have different prior beliefs about this. Someone who likes neoclassical economics will say "yes, definitely", while more Keynesian-minded economists may say "no, not necessarily". We define two hypotheses:

> $H1$: minimum wage does affect unemployment;
>
> H_2: minimum wage will lower unemployment in an economically significant way.

We can then assign prior probabilities to these hypotheses. This will be different for different economists, and perhaps for different countries and contexts. But let's put them at $H_1 = 0.4$, and $H_2 = 0.6$. Now the evidence from Card and Krueger (1994) comes in: minimum wage affects unemployment in a statistically insignificant way, and in a way that is the opposite from what we would expect if H_2 would be true. We can say, that $PEH_1 = 0.2$: if a minimum wage does not have an economically significant influence on unemployment, we would expect to find something like this data, or even further removed from the null hypothesis in 20% of the cases. $P(E|H_2)$ on the other hand, is perhaps 0.01: if minimum wage would decrease unemployment in an economically significant way, then it is highly unlikely to observe the evidence that Card and Krueger find.[5] Applying the formula, we can now calculate how likely H_2 and H_1 are after the evidence:

$$P\left(H|E\right) = P\left(H\right)\frac{P\left(E|H\right)}{P\left(E\right)}$$
$$P\left(H_1|E\right) = 0.4 * \frac{0.2}{0.2 * 0.4 + 0.01 * 0.6} = \frac{0.08}{0.086} = 0.93$$
$$P\left(H_2|E\right) = 0.6 * \frac{0.01}{0.2 * 0.4 + 0.01 * 0.6} = \frac{0.006}{0.086} = 0.07$$

Someone who was first quite confident thatH_2 was correct, rather than H_1, with respective probabilities of 0.6 to 0.4, should now change their beliefs in the other direction. They should now think that H_1 is more likely (with 0.093 probability) and should believe that H_2 is now much less likely (0.07 probability).

The conditional probabilities in this example are still very simple. Below, I will look at an example that uses more sophisticated estimation techniques.

8 • IS BAYESIANISM UNSCIENTIFICALLY SUBJECTIVE?

Much like classical statistics, Bayesianism comes with philosophical problems. It is good to point out that while the mathematical formula

[5] It is important to note here that the $PEH_1 = 0.2$ can, under certain assumptions, be approximated by a p-value (which was, as you might remember, indeed .2). For the $P(E|H_2)$=0.01 calculation, we would need much more information, that the Card and Krueger paper does not provide, so in this case, this value is fictional.

itself is uncontroversial – everyone believes that Bayes' theorem correctly describes a mathematical reality – it *is* controversial whether, and to what extent, it should guide our scientific (and econometric) choices.

Above, we already discussed one problem for Bayesian philosophy: subjective probabilities are not clearly defined. Bayesians were able to deflect that criticism, although it is up to the reader to judge whether they did so satisfactorily.

The most important criticism of Bayesianism is related: when we are doing a Bayesian analysis, we need to specify a prior. Prior probabilities have an important influence on the outcome, so subjective beliefs of the researchers will affect the outcome. If we were to start with a high prior that a hypothesis is true, it would be more likely that the posterior would also be high than in other cases, and vice versa, low priors lead to low posteriors. For many proponents of classical statistics, the reliance on priors makes Bayesian analysis **unscientific**. Science requires objectivity, so this feature of Bayesian analysis is problematic.

Some Bayesians agree that science should be objective, and that the role of a scientist's own beliefs about the hypothesis should be minimized. One reassuring result is the **principle of stable estimation** (Savage 1963): no matter what your priors are, as the amount of available evidence approaches infinity, the effect of the priors on the posterior approaches zero. This is a reassuring result, but gathering infinite evidence takes infinite time. In cases of smaller sample sizes, the priors will have an important effect on the data.

A second way to respond to the critique is to point out that Bayesians can use **diffuse priors**, also called **flat**, or **uninformative priors**. Using diffuse priors entails dividing the prior probabilities equally over all available hypotheses. Say you are investigating a crime, and there are 10 suspects, and you are confident that one of the 10 has done it. In this case, using diffuse priors means that you assign the prior probability equally over all hypotheses, so 10% to each of the suspects.

This sounds like an objective solution, but it is arbitrary in its own way: what if one of the suspects, Mr. X, has a criminal record already, while another suspect is a high schooler with no record of serious wrongdoing whatsoever? Assigning the same prior probability to both may sound objective, but this objectivity may be illusionary. Sometimes we can have good reasons to assign different prior probabilities to hypotheses; neglecting these can hardly be called objective.

A final response to the problem of subjectivity of priors is that a Bayesian analysis should always include a **sensitivity analysis** to different priors. For example, we do a Bayesian analysis and find that the probability of Mr. X being a murder is 80%, if we use a prior of 60%. Next, we redo the analysis with different priors. We find that, if our prior probability is lower than 40%, the probability of Mr. X being a murderer drops below 50%. We see now that the result was mainly driven by our specification of the priors: our priors had a large impact on the outcome. Conversely, if priors have a minor impact on the posterior, sensitivity analysis provides us with a strong argument that our analysis is correct.

Despite these three arguments, a classical statistician may still find the usage of priors objectionable. The usage of priors is one of the main points of contention between the two schools of statistics.

9 • SMALL SAMPLES

As we saw above, Bayesian statisticians believe in continuous updating. That means that when evidence comes in, even if it is just a small sample or a single data point, we should update our beliefs, if only a little bit. This becomes of crucial importance in cases where the data samples are small, but the stakes are high.

Think, for example, about testing new medicines. This was very relevant in an example from 2008 in The Netherlands: a study about the treatment of pancreatitis with probiotics, substances that can be found in yoghurts, yoghurt drinks, and beverages like kombucha. There were 298 individuals with pancreatitis in the study (Besselink et al. 2008). Of the roughly 150 individuals that were treated with probiotics, 24 individuals died. In the control group, only 9 individuals died.

Halfway through the study, the researchers had already observed a difference, but that was non-significant ($p = 0.1 > 0.05$). In Chapter

3, we saw that classical statistics violates the principle of total evidence. If a difference is non-significant, we should disregard it. This is one of these cases where, according to classical statisticians, we cannot yet interpret the evidence because it is not significant, even though we can already observe a difference that may matter a lot. In hindsight, however, it is clear that if non-significant results had been taken more seriously, deaths may have been prevented.

In these cases, Bayesians argue, the prior probability that the treatment is harmless should play a role in determining whether to continue. What's more, the researchers should continuously update, and if the posterior that the treatment is beneficial (rather than harmful) drops below a certain level, the experiment should be terminated, no matter whether it is significant in the classical sense.

This example illustrates the importance of the Bayesian commitment to the principle of total evidence; classical statisticians believe that we can only start interpreting the evidence once the sampling is complete, but according to Bayesians, all evidence counts all the time.

There is an important reason classical statisticians think we should finish sampling. **Stopping rules**, which dictate the duration of the data-gathering process,have an important influence on the interpretation of the data. Consider the following strategies.

Strategy 1: collect data until you find a statistically significant finding.

Strategy 2: collect 30 data points and observe whether there is a difference.

In the first case, the chance of observing a significant finding is much larger than the second. If, in both cases, we end up with 30 data points and significant results, we should still have higher confidence in the second strategy: after all, the first strategy all but guarantees a significant result. Phrased in the terminology discussed in the previous chapter, Strategy 1 is less *severe* than Strategy 2.

For classical statistics, specifying the rules before data collection is an important part of the research design. For Bayesians, stopping rules may also matter. It may assign a different likelihood to evidence found with the first strategy. But it maintains that all evidence counts as evidence at all times, even if the sampling has not finished. In macroeconomics, the samples are often small too: think of yearly data points since 1960, and Bayesian and classical methods may therefore diverge in their outcomes. The next subchapter concerns a case study on this topic.

10 • ECONOMETRICS EXAMPLE: LONG-RUN GROWTH MODELS

Much like classical statistics, Bayesian statistics encompass a variety of statistical methods. Sometimes, such as in the example of the base rate fallacy, these methods are simple. Whenever econometric models are involved, however, Bayesian estimation gets increasingly complex. The following paragraphs briefly discuss one example of Bayesian econometric modeling, to give an impression of the process. However, if you are interested in the details, introductory textbooks on Bayesian econometrics are a much better place to look (e.g. Greenberg 2012).

Whereas classical statistics typically requires a certain minimum sample size for reliable results, Bayesian statistics does not. As such, Bayesian methods often find application in cases where only small sample sizes are available. One widely studied problem in economic modeling is the determination of the main causes of the long-term economic growth of a country. There are no more than 100 countries that collect data on economic growth, and have kept these data from the last 50 years. But, because long-run growth takes time, there are only few data points on which a model can be based. Many variables potentially affect economic growth, and a large sum of models can be created from these data. What is more, whether or not a variable works within a model, is dependent on the other variables. With different controls, we will find different parameter estimates, which will sometimes be statistically significant, and sometimes not (we will get back to the questions arising in modeling in Chapter 5).

Economist Xavier Sala-i-Martin tackled the problem of economic growth modeling in a 1997 paper called “I just ran two million regressions” (Sala-i-Martin 1997). He randomly selected a large set of economic models and computed a weighted average of their test statis-

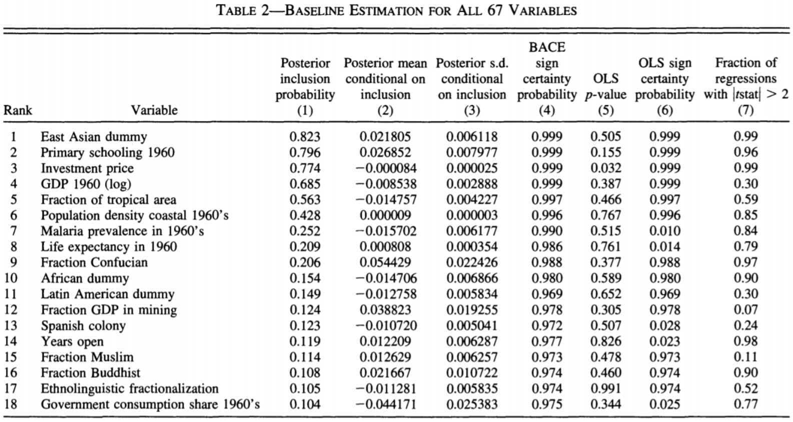


TABLE 2—BASELINE ESTIMATION FOR ALL 67 VARIABLES

Rank	Variable	Posterior inclusion probability (1)	Posterior mean conditional on inclusion (2)	Posterior s.d. conditional on inclusion (3)	BACE sign certainty probability (4)	OLS p-value (5)	OLS sign certainty probability (6)	Fraction of regressions with $\lvert t\text{stat}\rvert > 2$ (7)
1	East Asian dummy	0.823	0.021805	0.006118	0.999	0.505	0.999	0.99
2	Primary schooling 1960	0.796	0.026852	0.007977	0.999	0.155	0.999	0.96
3	Investment price	0.774	−0.000084	0.000025	0.999	0.032	0.999	0.99
4	GDP 1960 (log)	0.685	−0.008538	0.002888	0.999	0.387	0.999	0.30
5	Fraction of tropical area	0.563	−0.014757	0.004227	0.997	0.466	0.997	0.59
6	Population density coastal 1960's	0.428	0.000009	0.000003	0.996	0.767	0.996	0.85
7	Malaria prevalence in 1960's	0.252	−0.015702	0.006177	0.990	0.515	0.010	0.84
8	Life expectancy in 1960	0.209	0.000808	0.000354	0.986	0.761	0.014	0.79
9	Fraction Confucian	0.206	0.054429	0.022426	0.988	0.377	0.988	0.97
10	African dummy	0.154	−0.014706	0.006866	0.980	0.589	0.980	0.90
11	Latin American dummy	0.149	−0.012758	0.005834	0.969	0.652	0.969	0.30
12	Fraction GDP in mining	0.124	0.038823	0.019255	0.978	0.305	0.978	0.07
13	Spanish colony	0.123	−0.010720	0.005041	0.972	0.507	0.028	0.24
14	Years open	0.119	0.012209	0.006287	0.977	0.826	0.023	0.98
15	Fraction Muslim	0.114	0.012629	0.006257	0.973	0.478	0.973	0.11
16	Fraction Buddhist	0.108	0.021667	0.010722	0.974	0.460	0.974	0.90
17	Ethnolinguistic fractionalization	0.105	−0.011281	0.005835	0.974	0.991	0.974	0.52
18	Government consumption share 1960's	0.104	−0.044171	0.025383	0.975	0.344	0.025	0.77

TABLE 4.2 A table representing posterior estimates from a bayesian analysis. From Sala-i-martin (2004)

tics (i.e. coefficients and variance). In a follow-up article, Sala-i-Martin used Bayesian analysis to estimate the probability that certain variables belong in the true long-term growth model (Sala-I-Martin, Doppelhofer, and Miller 2004). Again, this was based on many model estimations, all of which were assigned a prior probability of being correct. For this, they used a diffuse prior in case models were equally large, and the larger the model, the smaller the prior. They found the 18 variables that most probably explain economic growth, as summarized in the following table.

As you can see in table 4.2, for each estimated variable, there is a posterior inclusion probability: the probability that this variable is included in the true model. As it turns out, 5 variables have a chance of belonging in the true model of over 50%.

II • BAYESIANISM VS. CLASSICAL STATISTICS

Bayesianism and the classical approach are generally seen as rivals. They differ in outlook on several matters, as we discussed in previous subchapters. The two crucial points of difference between Bayesian and classical statistics can be summarized as follows.

Small samples ($p > \alpha$): what to conclude?

Bayesian: We should learn from the data, even if they are not statistically significant. All evidence is evidence. This is particularly relevant in cases with small sample sizes, such as pharmaceutical trials. By updating continuously, we are doing the rational thing: taking into account all evidence. If we were to disregard the nonsignificant results, we are not abiding by the **principle of total evidence.**

Classical statistician: Good design is really important for good evidence. We should clearly specify the sample size before we do our tests and experiments. If we were to interpret our data halfway through, our analysis would be less rigorous and less severe. By specifying the sample size beforehand, we are making it difficult for the null hypothesis to be rejected. If we would not do so, a researcher can simply continue testing until they find what they like, making the result less rigorous. Specifying the tests beforehand, including the sample size, helps to avoid confirmation bias.

Priors: yes or no?

> **Bayesian:** Without priors, we are effectively blind. A piece of evidence may seem highly unlikely given a certain hypothesis, but if we do not know what the prior probability of the null hypothesis and its alternatives are, this by itself does not tell us much. We should be careful to interpret a statistically significant p-value as really strong evidence against a hypothesis if there are also strong reasons to believe the hypothesis is true. Unlikely events sometimes simply occur, so not all low p-values signify false null hypotheses. We should therefore consider the prior probability of a hypothesis before we judge whether it is likely true or false.
>
> **Classical statistician:** Priors are not scientific; they reflect a person's subjective view of the world. They may be

> affected by intuitions, prejudices, and biases. Therefore, they do not belong in science.
>
> **Bayesian:** Priors may be undesirable, but they are simply there and should affect the outcome. Even you, classical statistician, cannot say that you believe the psychologist who found statistically significant evidence for the hypothesis that humans can see into the future ($p < 0.01$). The only reason that you do not believe that, is because of your reliable, but subjective, prior beliefs. Moreover, the effect of priors fades away the more data becomes available, and we can use ignorance priors if need be. Not all priors necessarily reflect our personal beliefs.

12 • CONCLUSION

Ultimately, both classical and Bayesian statistics aim to draw reliable conclusions from data.

On one hand, classical statistics gives us only inverse probabilities: the probability that we find what we have seen (or something even more deviant from what we would expect), given that the null hypothesis is true. Often, what matters is the probability a hypothesis is true given the evidence.

On the other hand, classical statistics does give us a certain type of objectivity: the subjective beliefs of the researcher do not play a role in the procedure itself, even though they may affect the conclusions. Maybe it does not give us the probability a hypothesis is true given the evidence, but only because it recognizes that doing so would require subjective input, which should not be part of scientific analysis.

What is important to note is thatthe two approaches need not necessarily be in conflict. You do not have to decide whether you are a Bayesian or a classical statistician. You may also be more **pragmatic**. Bayesianism may be particularly useful when the data samples are small, or when we have good justifications for our priors. Classical statistics, on the other hand, may be more appropriate in fields with significant disagreements, where using priors may be controversial.

In econometrics, Bayesianism is used with increasing frequency, and I am sure that when you continue studying econometrics, you will encounter Bayesian analyses more and more.

LEARNING GOALS FOR THIS CHAPTER

After having studies this chapter, you should be able to

1. Explain Bayes' theorem, its central concepts (priors, posteriors, updating.) and how it can be applied to interpreting scientific evidence.
2. Conduct a basic Bayesian analysis, if the probabilities P(E|H), and P(E|-H) are given.
3. Explain arguments in favor and against Bayesianism.

5 Keynes, Tinbergen, & the Problems of Econometric Modeling

1 • THE STRUCTURE OF AN ECONOMETRIC MODEL

This is what the most basic econometric model looks like:

$$Y = \alpha + \beta_{\mathrm{i}} X_{\mathrm{i}} + e$$

There is a dependent variable Y, an intercept parameter α, and a set of independent variables X_{i}, which are assumed to linearly co-vary with Y in proportion to estimated parameters β_{i}. The error term e, captures all variance in Y that cannot be ascribed to variance in the independent variables, and is assumed to be randomly distributed.

So far in this book, we have looked at statistical reasoning in the abstract: Significance testing based on statistical evidence. In econometrics, however, statistical reasoning generally takes place within the context of economic models. Statistical modeling plays a crucial role in inductive inference in econometrics.

What is the role of modeling in our statistical inference? Once we know which dependent variable we want to study, testing models generally occurs in three steps. First, we assess which independent variables we use; these comprise the elements of X_{i}. Second, we estimate the model: we estimate α, β_{i}, and e, through, for example, ordinary least squares. Third, we assess the model fit, or the parameter of interest that we have estimated. The model fit is generally assessed through the method of least squares and an F-test. The method of least squares expresses in $0 \leq R2 \leq 1$ how much of the variance in Y is explained by the model. The F-test results in a p-value for the null hypothesis that the model does not explain Y at all. The slope parameter β_i represents effect sizes, whose statistical significance are tested through t-tests.

2 • EXPERIMENTAL AND NONEXPERIMENTAL DATA

When it comes to modeling, economics differs from many other social sciences in one important aspect: the type of data that it has at its disposal. Most other sciences build heavily on **experimental data**. By creating an experiment, knowledge is generated in a replicable way, and if the experimental design is done correctly, the experiment provides good evidence of a **treatment effect** in the groups under investigation. Experiments work well under one of the following conditions.

(a) Two treatment groups are **exactly** the same, except for the treatment

(b) Two treatment groups are **randomized**, such that we would expect any differences between the groups to be randomly distributed across the two treatment groups, except for the treatment.

In case of either (a) or (b), the treatment effect can be interpreted as a **causal effect**. After all, if the outcome variable is different in the treated group and the untreated (i.e. control) group, nothing but the treatment can explain this difference. This method is remarkable, and while it has its own problems (see for example the replication problem in psychology that we discussed in Chapter 1), experiments provide a solid basis for scientific inference.

While the very basics of the statistical tools used in econometrics and other sciences are the same, in macroeconomics, we generally **cannot do experiments**. After all, no two countries are the same, and we could not possibly randomize the countries of the world and treat half of them with some economic treatment that we are investigating. For example, we cannot randomly assign the economies of the world to two treatment groups, and impose austerity measures (i.e. government spending cuts, and increased taxes) on one group and not the other, to see how this plays out in the decades thereafter. The fact that this is impossible is a central challenge that has shaped econometrics.

There is an important caveat to the claim that macroeconomists cannot do experiments. There are rare occasions in which, without the intervention of researcher, settings arise that closely resemble conditions(a) and(b) above. We call these **natural experiments**. The example of the minimum wage increase in New Jersey discussed in Chapter 2 is one example of such a natural experiment. In this example, there were two very similar groups: fast food restaurant workers in west Pennsylvania and in East New Jersey. One of these groups underwent a treatment: an increase in the minimum wage. The subsequent difference or absence of difference between the groups, namely the difference in unemployment, must be the causal effect of this treatment. Another example of a natural experiment is found in an article by James Feyrer (2009). He investigated the impact of geographical distance on trade caused by the 7-year closing of the Suez Canal following hostilities between the oil-exporting nations in the middle east and the Western World.

These natural experiments, however, are imperfect. Strictly speaking, they neither satisfy the conditions (A) nor (B) that we discussed above: they neither represent perfectly equal treatment groups, nor truly randomized treatment groups. Therefore, we call them "**quasi-experiments**".

In the context of macroeconomics, econometrics has to make do with what is there: observational data. **Observational data** are defined by the fact that they are generated in a setting that is **not manipulated by the researchers** for the purpose of the research: they merely appear to the researcher, who can then observe them. Researchers may be involved in shaping economic policy, but they do so because they have economic expertise on what is best for the country, not, we may hope, because they are curious to learn what will happen if a certain policy is implemented.

Significance testing was developed as a method for doing experiments. As we shall discuss in this chapter, it is not exactly obvious that the statistical inference that applies to experimental data, also applies to observational data. In this Chapter, we will look at observational data generally. In Chapter 7, we will take a closer look at quasi-experiments.

An excellent strategy for learning about the methodological challenges that face a certain method, is to consider the time at which they were introduced to a science. The introduction of econometrics to economics was roughly in the **1930's**. Econometrics has since become an accepted part of economics. Consequently, discussions about its soundness and importance are rare. However, by the time of its introduction, economists heavily debated the issue of whether statistical data analysis should play a role in economics. We will look at arguably the most influential methodological debate between two economists about econometrics: the Keynes-Tinbergen debate.

3 • KEYNES AND TINBERGEN

John Maynard Keynes is perhaps the most well-known and renowned economist that has ever lived. He is well known for his macroeconomics book "*The General Theory of Employment, Interest, and Money*" (1936). He is less known for his earlier work on probability theory, and he is only a little known for his skeptical attitude towards econometrics and over-usage of mathematical tools in economics.

Keynes held a particular view on how the economic world worked. He believed that economics, like physics, was ultimately a matter of **laws**. Physics describes the world in **natural laws**, and economics in **economic laws**. However, there is an important difference between these types of laws. The physical world is, compared to the economic world, quite simple: there are a small number of variables that determine the behavior of a physical body, and the laws that describe them are stable – the same in all different contexts. Because of this, we can describe the physical world in mathematical form, but the same cannot be said for economics. Humans are not as simple as sub-atomic particles.

According to Keynes, economics may be guided by laws. For example, the relations described by the IS-LM model describe some economic laws.[6] But, ultimately, mathematical expressions of such laws are at best **mere approximations** of reality. The economic world is so complex, that economic laws, unlike physical laws, will remain uncertain and unstable. Even if there is some mathematical relationship between two economic variables in a particular time and place, changes in some third, fourth or fifth variables will cause render this relationship invalid. Economic laws are simply too complex to express in a simple linear model. We can summarize this point as follows.

Instability of economic laws (Keynes): because the economic realm is so complex, the laws that guide economic behavior will be unstable, as they depend on factors that change in different contexts, and uncertain, as the complexity makes it too difficult to estimate these unstable relationships.

This point has a significant implication for econometric modeling. Econometricians often use the term **"the true model"**. The true model is a faithful, correct description of reality. If Keynes was right, this has important implications for the concept of the true model in econometrics. The true model is either so complex that no econometric model will be able to estimate it, or the true model is simpler, but changes from time to time, and country to country. In both cases, the estimation of a true model is an extremely ambitious project, that may be practically impossible. At best, we may find **approximations** of the true model.

Keynes died before the first Nobel Prize in economics was awarded. If not, it seems likely he would have been among the first on the list of potential laureates. Instead, the first Nobel Prize in economics was awarded to **Jan Tinbergen**, together with another econometrician, Ragnar Frisch, ”for having developed and applied dynamic models for the analysis of economic processes”. In essence, the prize was awarded for their work on the usage of statistical techniques in economics.

Jan Tinbergen first studied physics, but moved to economics, because he was interested in making the world a better place: a place with

[6] It is important to note, however, that Keynes himself did not use the IS-LM graphs that have come to be associated with his work.

less poverty, and more peace. In the 1930s, the League of Nations, the predecessor of the United Nations, was convinced that economic stability was key to avoiding war in the future, and asked Tinbergen to conduct an analysis of business cycles (or "boom-bust cycles") using the best available data of the United States. Tinbergen used regression methods on time-series data. This method was relatively new at the time, because before that, data had been scarce, and computational power expensive: *computors,* at this time, were *people*, who would be doing math for the professors they worked for! Tinbergen's study was one of the first uses of statistical techniques in economics. In other words, it was one of the first econometric studies.

Keynes wrote a review of Tinbergen's study in The Economic Journal (Keynes 1939), that has become quite well known. The review was very critical. Keynes raised serious doubts about whether Tinbergen's statistical model could teach us anything. There is an important reason to go over Keynes' criticisms, namely that these criticisms potentially apply to *all* econometric models. So when you read the criticism below, think about which ones you think are relevant to econometric modeling today. I strongly recommend reading his review to anyone who works with econometrics. It is short and insightful. Keynes described numerous methodological problems with regressions methodology. We will go over them in turn.

4 • KEYNES' REVIEW

The aim of econometric models

Keynes assumed econometric models in general, and Tinbergen's model in particular, to have the following aims:

- estimating the relative effect of different independent variables on investment (his dependent variable);
- estimating their causal impact on investment (the dependent variable).

You may, at this point, already take issue with Keynes. We learn in our first statistics courses that correlation should not be confused with causality, and econometric models typically only tell us something about correlations. While that is true, Keynes does have an important point: if these factors do not represent causal effects, they are not really describing economic reality. Ultimately, we are interested in the question what happens to Y if X_i changes. Everything in the language of regression indicates that this is the aim. For example, the term "dependent variable" strongly indicates that it is dependent on the independent variable. Dependence implies causality. "Effect size" is a term that strictly speaking only refers to a correlation, but strongly indicates an actual *effect* of independent variables on dependent variables. We can call this way of describing economic models **causal language**, everything in this way of speaking reveals that we are ultimately interested in causal relationships.

What can we conclude from the statistical fit of a model?

According to Keynes, Tinbergen acknowledges that "no statistical model can prove a theory to be correct". This is in line with the falsificationist logic that we have seen before: even the best model fit is not proof that an econometric model is the true model. Also in line with the falsificationist logic, Tinbergen argues that a poor fit can show a model to be false or incomplete. Keynes disagrees with Tinbergen on this point. Keynes suggests that, because a model can only be tested **under a large set of assumptions**, even true models may fail to provide a good fit. If a model does not fit the data, the model may be incorrect, or some of the assumptions made in the process may be incorrect. So, even a bad fit does not prove that the model is incorrect.

Complete list of variables

Keynes notes that **the absence of a relevant variable** may lead to significant misspecification of the model, and of the individual coefficients of the independent variables that are in the model. Keynes concludes from this, that the econometric analysis is only valid if we have a *com-*

plete list of the variables that affect the dependent variable. But, if that is true, Keynes suggests that we already need to know everything about the world, namely all the independent variables that make up the true model, before we estimate our model. Often, we are not precisely sure what the complete list of variables is.

Measurability

A further concern is that not only do we need to have a complete list of relevant variables, but all these variables need to be measurable and available. Keynes argues that this may have important implications: some important variables, such as sentiments and expectations (e.g. Keynes' famous "animal spirits") may be very difficult to quantify, if not impossible. If these are included in the true model, and they are not available, what is the econometric model worth?

Multicollinearity

A significant concern of Keynes is multicollinearity, or has he calls it "the different factors are substantially independent of one another". If two independent variables "explain" the same variance, it is not clear to which of the two this variance should be ascribed.

Reverse causality

Similarly to multicollinearity, reverse causality may also create the problem that the coefficients are over- or underestimated.

Linearity assumption

A common complaint with simple econometric models is that they typically assume linearity. Why would we expect the economic world to be constructed out of linear relationships? In some instances, Keynes argues, it is highly unlikely that the variables interact linearly. However, the correct form does not simply present itself and is therefore often unknown. This may lead to misspecification of the model and, consequently, incorrect estimates.

Establishing the right time lag

Some economic variables do not impact growth (or investment, in Tinbergen's case) in the same period of time. In those cases, we need to establish a time lag: how long does it take to for the independent variable to have an impact on the dependent variable? Tinbergen tried different time lags until the fit was good, and then established that he should use that time lag. Keynes finds this a dubious method, because it might make it too easy to find a good fit (this is called overfitting, a term discussed shortly).

Sensitivity to specific data points

A further concern is that in the type of time-series analysis that is common in econometrics, the starting and end points of the data may have a large influence on the result. Keynes criticizes Tinbergen for using as a starting point for his analysis the year 1919, a boom period, and as an endpoint the year 1929. a recession period. This may have a significant impact on the coefficients that are found.

Invariance

Keynes finally points out that we can only observe the effect of independent variables on the dependent variable if the variables show sufficient variation. This point is quite important from a theoretical point of view: there may always be background factors, for example, trust in banks or the government, that will have a significant influence on the variables of macroeconomic models. If these background factors remain constant for significant periods, we cannot know how they impact the true model.

Conclusion: stability of found coefficients

Keynes' general conclusion aligns closely with his view on the economic world: Keynes was skeptical that Tinbergen's estimated model would hold true if more data became available. In other words, it may have been a description of the data from 1919 to 1929, but neither was it the true model of the United States economy, nor would it help us to predict the behavior of the economy in the future.

Keynes ends with a note about Tinbergen:

> "I have a feeling that prof. Tinbergen may agree with much of my comment, but that his reaction will be to engage another ten computors and drown his sorrows in arithmetic."

Tinbergen wrote a reply in which he explained his choices (Tinbergen 1940). He explained, for example, that linearity is not such a strange assumption if we consider the fact that any mathematical form can be approximated by a linear relation. Moreover, responding to the multicollinearity problem, Tinbergen explained that while the independent variables were not necessarily uncorrelated, he did assume that there was no shared covariance with the dependent variable.

Keynes wrote a reply to Tinbergen's reply (Keynes 1940), continuing the debate, and ending with the following challenge to Tinbergen:

> "Professor Tinbergen appeals to me several times to cook (or, should it be, eat?) more pudding myself before declaring it indigestible. I would ask in return for an experiment on his part. It will be remembered that the seventy translators of the Septuagint were shut up in seventy separate rooms with the Hebrew text and brought out with them, when they emerged, seventy identical translations. Would the same miracle be vouchsafed if seventy multiple correlators were shut up with the same statistical?"

Even putting aside his more particular point of contention, Keynes posed an interesting general challenge. According to the myth of the Septuagint, seventy translators translated the Hebrew bible to Greek in rooms that were completely independent of each other. They came out with seventy identical copies. Just to be clear: this is a myth, and there is no evidence this actually happened. Keynes wondered if the same is true for econometrics. If the same econometricians get the same econometric task, would they construct the same model, and find the same coefficients?

What is the main upshot from this interaction? Even though Keynes may be seen as a little uncharitable, he was right to point out that an econometrician makes many choices that may have a significant impact on the outcome. Remember, falsificationism, as a general theory of science, tells us that we should formulate theories and let the data tell us if they are right. Keynes shows us that there is an important step between formulating theories and testing them: **specifying an econometric model** that allows us to test our theory. In the specification of such a model, many choices have to be made that affect the outcome. Keynes rightly asks us to pause and assess the validity of this procedure.

5 • PHILOSOPHY OF SCIENCE: HOW SERIOUSLY SHOULD WE TAKE KEYNES' CONCERNS?

Keynes' Septuagint

In Chapter 1, we discussed that from 1995 to 1997, an econometrician and a philosopher jointly conducted an experiment on the basis of Keynes' open question (Magnus and Morgan 1999). Keynes was right to suggest that econometrics does not provide a singular method to analyze the data.

How about the more specific problems he raises with Tinbergen's econometric method? To see this, we need to introduce two central concepts from the philosophy of science: **under-determination** and **double counting**.

Under-determination

Remember from the discussion above, that the success of a model is typically assessed by its fit: how well does it capture the data. However, ultimately, we are interested in the underlying quantitative mechanisms or laws that explain the behavior of our dependent variable. Does a good fit (a high R^2, a low p-value for an F-test) tell us that the model indeed captures these economic relationships?

A significant problem with the idea that fit indicates the correctness of a model, is that it may be possible to explain the variance in one dependent variable with multiple models or theories. In fact, it turns out that statistical phenomenon can be described by infinite possible models. Consider the Lucia de Berk example from Chapter 4: her frequent presence at incidents in the hospitals in which she worked can be explained by numerous hypotheses. One of them is that Lucia de B. is a murderer, but many different hypotheses can explain this, for example (Philipse 2015, 30):

> H_1: the concurrence was a mere coincidence;
>
> H_2: at the times of concurrence, Lucia always shared her shifts with someone else who caused the incidents;
>
> H_3: Lucia was often on a night shift, and the risk of incidents is higher during the night;
>
> H_4: Lucia is a relatively incompetent nurse, so the risk of incidents during her shift on the ward is high;
>
> H_5: Lucia prefers to care for patients with complex disorders, and these patients have a greater risk of dying;
>
> H_6...

You get the point: there is no limit to the amount of theories that are compatible with the data. To some extent, we can solve this problem through critical thinking: perhaps not all alternative hypotheses are plausible. However, there is a deeper point: all data points can be described by an arbitrarily large number of theories. This is even true for our best theories. To give one extreme example, consider the following two theories about physics:

> Theory 1: our currently best theory of physics is correct;
>
> Theory 2: our currently best theory of physics is correct until 2030, after which the laws of physics will change.

Of course, Theory 2 is silly. However, it is important to note that both theory 1 and Theory 2 are consistent with the data we have about physical phenomena. There is an arbitrary element to theory (or hypothesis) development. Only our imagination limits the theories that we can come up with. We can, for example, also formulate a different theory:

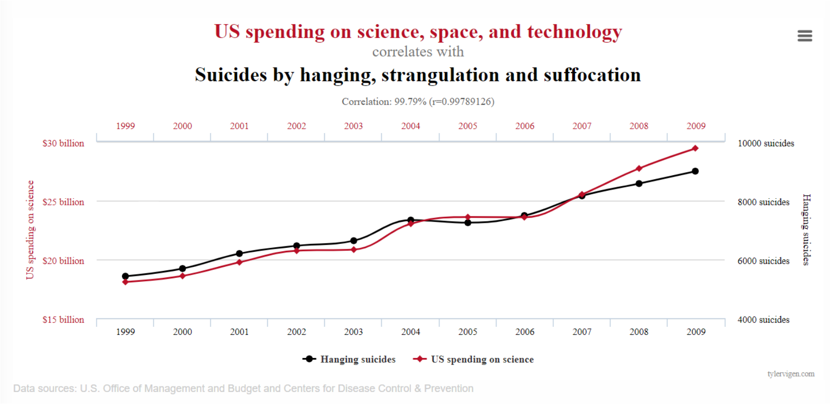
> Theory 3: our currently best theory of physics is correct until 2031, after which the laws of physics will change.

In this way, the list of theories that fit with the best available evidence is arbitrarily large.

In econometrics, we will often say that "the evidence points in the direction of...", or "our data indicates that...", but it is good to note that strictly speaking, data cannot really point in any direction. We can easily illustrate this with silly examples. You can, for example, find a spurious correlation generator online: https://www.tylervigen.com/spurious-correlations. This generator will give us correlations like those shown in the following graphs (figure 5.1 and 5.2).

These examples are clearly not serious. No one would make the mistake of actually interpreting these results as anything different than accidental correlations. But how do we make sure that we do not make similar mistakes in econometrics? Especially in time series data, it is not particularly difficult to find correlations. This is one reason why Tinbergen and Keynes are right to point out that data alone cannot prove any theory correct.

A variant of this problem applies to the process of specifying the



Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention

FIGURE 5.1 A spurious correlation (from https://www.tylervigen.com/spurious-correlations)

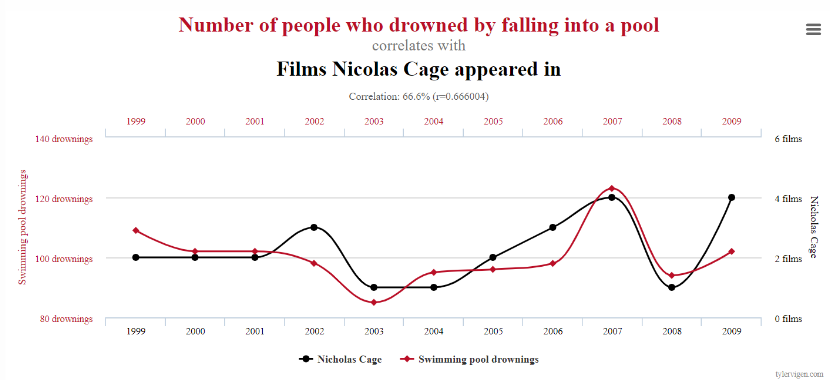


FIGURE 5.2 A spurious correlation (from https://www.tylervigen.com/spurious-correlations)

model shape. Keynes criticized Tinbergen for imposing linearity onto his model. But if Tinbergen had used more functional forms, this would also be reason for concern. If we start to include more functional forms, we are bound to arrive at models with a better fit. A more complex model that is fitted to a dataset will necessarily produce a better fit than a simpler model. If we make our model complex enough (i.e. with enough non-linear functional forms), we can guarantee a

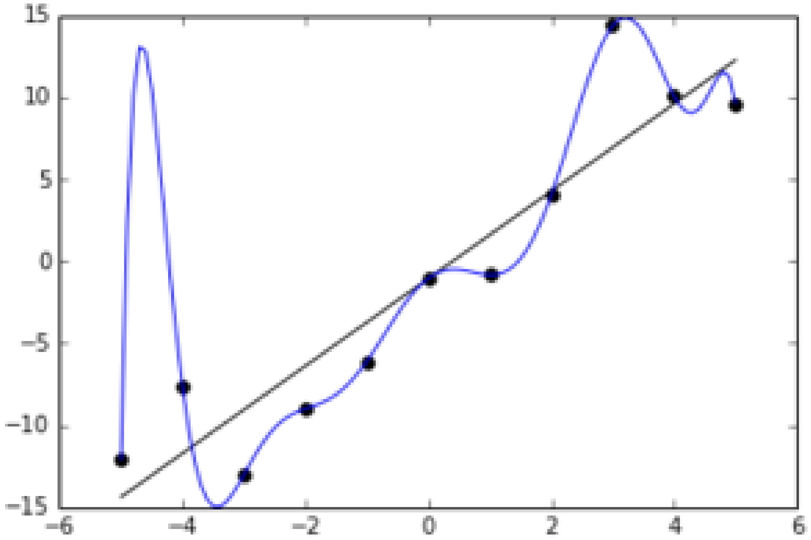


FIGURE 5.3 An overfitted model (from https://us.energypolicy.solutions/docs/comparing-results.html)

good model fit. This phenomenon is called **overfitting:** a good model fit that is driven by complexity. This can be illustrated by the following graph (figure 5.3):

The linear model has a worse fit than the blue model, as the blue model fits perfectly. Still, it is obvious that we should not expect the blue model to capture the true model.

We can come up with an arbitrary number of very complex models with different function forms and large sets of variables, and we are guaranteed that it will fit the data well. But such models will break down if we extrapolate them. These problems illustrate that that data alone will never be sufficient to tell us which theory we should believe. Philosophers of science call this the problem of under-determination.

> **Under-determination:** the available evidence will never be sufficient to determine which theory we should believe, because an arbitrarily large number of theories will be compatible with the data.

> **Overfitting:** adding complexity to the model, in particular adding more variables and functional forms, achieve an arbitrarily good fit that is unlikely to remain if new data comes in.

These two ideas tell us something important: model fit, by itself, does not guarantee that the econometric models are correct. The problem of under-determination thus shows us that we need to be careful: models should be simple and plausible. Even then, we cannot always trust our model fit. There may be many models with a good fit that are not the true model.

One reply to these concerns is that we need simple models. This is something that is already widely acknowledged in econometrics. For instance, while R^2 is taken to be a good estimate of model fit, when models get complex, econometric textbooks typically prescribe adjusted R^2 measures. **Adjusted** R^2 takes into account, or punishes, model size. So, in case a model fit, R^2, is really good, but only the result of the complexity of the model that we have created, our adjusted R^2 will not be very high.

Remember the concept of **Severity** from Chapter 2. We can use this concept in the context of modeling to assess how good econometrics is at modeling the world. In this context, severity means:

> **Severity (modeling):** if a theoretical model is incorrect, it is unlikely that an econometric model based on this theoretical model will fit the data well. If it is correct, it is likely to fit the data well.

Whether fit is a severe test for the model's accuracy, boils down to the issue of how easy it is to achieve a good fit, if the model is actually incorrect. One important aspect of this issue is that there are a number of features that Keynes mentions that actually increase this probability. For example, if we select time-lags based on which ones fit best with the data, and our beginning and end points are also selected based on what would work best with the data, we are making it more and more likely that the data will fit the model, even if it is not correctly specified. Such choices are stacking the deck in our favor: it will be more and more likely that our theory will fit the data, regardless of whether it is the true model. This brings us to the problem of double counting.

Double counting

There is another general concern with econometric modeling that Keynes does not explicitly address, but that play an important role in econometric methodology: in order to assess the model fit, we first need to estimate the parameters of the model α, β_i, and e. We want to know the model fit of this model: $\mathrm{Y} = \alpha + \beta_\mathrm{i}\mathbf{X}_\mathrm{i} + \mathrm{e}$. We first need to estimate parameter α, β_i, and e. These parameter values are estimated to maximize model fit. This process is called "model calibration". We can then assess how closely the model fits the data.

As you can see, this process uses the data in two ways: 1) to estimate the free parameters, **so that they model fit is maximized**, and 2) to **assess the model fit.** It is no wonder then, if the model fit is high: after all, the model was constructed such that it would best fit with the data.

This is a bit like shooting on a wall, and then drawing a bull's eye around the bullet holes afterwards. Doing so is not exactly good evidence that the person is a good shooter. Econometric modelling practice is not quite like that: in most contexts, a statistical model cannot achieve a perfect fit through calibration. However, because the data plays a role both in the construction and the testing of the model, we can no longer see it as a neutral arbiter. A better analogy, then, would be that of a figure skating match, in which one of the figure skaters is also part of the jury.

This phenomenon of using the same data to construct a model and using it to test the fit of the model, is called **double counting**. According to some philosophers, double counting is improper and should be avoided. Philosopher John Worrall, who is strongly inspired by Karl Popper, argues for the following requirement, for all scientists:

Use novelty requirement (Worrall): for data X to support hypothesis H, or for X to be a good test of H, H should not only agree with or “fit” the evidence X, but X itself must not have been used in H’s construction. (as summarized by Mayo 2010, 156)

The rationale behind this requirement should be clear. If we formulate hypotheses using the data, the question whether the model will “fit” with the data is not really a good test to assess whether it is true: it is too easy. Double counting makes a test of model fit less severe. The Use Novelty requirement has an intuitive appeal, but also far-reaching consequences. After all, almost all econometric modeling in practice violates it!

The Use Novelty requirement is controversial. There are many examples from scientific practice of discoveries being inspired by the same data that provided evidence for the hypothesis. For instance, although the movements of the planets in our solar system repeat themselves, we used these observations both when formulating theories of physics and as evidential support for them. Nevertheless, double counting will generally have a negative effect on the severity of model fit as a test for the model. It is better practice to use new evidence for the testing of a hypothesis, that was not yet used for the construction thereof.[7]

In macroeconomics, however, the data are scarce. And for severe tests we also need sufficient data – a small sample will not be likely to provide a severe test. New data only comes in once economies develop, go through new boom and bust cycles, and develop new technologies. However, certain economic events (e.g. oil crises) are quite rare, and it may take a lot of time before enough new data becomes available that can provide a severe test of the formulated hypotheses. For long-run

[7] If you have paid attention to Chapters 3 and 4, you will not be surprised that whether or not you agree with the use novelty requirement will dependent on whether you are a Bayesian or a Classical Statistician. Bayesians believe that all data should play a role in determining which theory is most plausible, while classical statisticians think that data should only count if meets certain quality standards, such as coming from a large enough, pre-determined sample, and only if it counts significantly against a null hypothesis. The Use Novelty Requirement therefore fits better with Classical Statistics, and not as well with Bayesian Statistics.

growth models, it is inevitably true that it takes a lot of time before new data points become available.

6 • SOLUTIONS? THEORY, PREDICTION, AND ROBUSTNESS CHECKING.

How can macroeconomics make sure that the models that they construct on the basis of the data are not "overfitted" and actually are indications of true underlying economic phenomena? In other words, how can we make sure that they do not only fit the existing data, but also capture the true model?

A first thing that economists can do, is make sure that the models that they construct are based on **solid economic theory**. If we have really good theoretical reasons to believe that the variable set X_i should be included in the model, we have more confidence that model fit is a good indication of the correctness of the model. However, theories are often not sufficiently developed to tell us the exact set of X_i, or the functional form of its coefficient.

A second thing is that true models should be able to **predict**. Predictive success is a clear indication that a model is not merely a description of a phenomenon in a particular time and place, but actually captures the real thing. A test of predictive success is, in fact, a way to incorporate the Use Novelty Requirement.

How good are economic models actually at prediction? In the wake of the 2008 financial and economic crisis, some commentators were quick to the charge that economists had not predicted it. In fact, Nobel Prize winner Paul Krugman wrote a critical piece in the New York Times, called "How did economists get it so wrong?". Krugman's explanation was very critical of the economic science:

> "As I see it, the economics profession went astray because economists, as a group, mistook beauty, clad in impressive-looking mathematics, for truth."

Krugman may have been a little unfair to the economists. Prediction is hard for any science. Climate Scientists also leave large margins of error in their projections, and weather forecasts still get it wrong sometimes, even for a couple of days into the future. Economic forecasts are being used with quite some success to predict economic growth for governments.

Whether a model is able to predict is generally a good test of that model, because it can generally be seen as a very severe test: a prediction test is very unlikely to be passed if a model is false. However, it still requires the availability of new data, so that a model can predict it. In some fields of research, such as long-run economic growth modeling, this may take decades.

A much more common countermeasure for overfitting is what economists call **robustness checking**. A **robustness check** is an adjustment of one of the modeling assumptions. The check is successful if the outcome of the model does not change in important ways. So, for example, if we are concerned about the correctness of a linearity assumption, we can assess whether our result changes if we change the functional form and use some quadratic relationships or interaction effects. If the result does not change, we can say that our model is robust to changes in functional form.

We can also control for the beginning and end points in the data. If we suspect, like Keynes did in the case of Tinbergen's model, that the result is driven by the specific beginning and endpoints, we simply use different beginning and endpoints. If the results do not change, they are robust to changes in the beginning and endpoints.

As a final example, Keynes was concerned with the specific variables in Tinbergen's model. This is actually a problem that often plays a role in critical analyses of empirical economics. If we used a different set of independent variables, the resulting coefficients could change. For example, if we are interested in the relationship between educational achievements and income, the coefficient that we use, changes if we include the variables "age" and "worker experience" in the model. In this case, our main interest is in one specific coefficient: the effect of educational achievement on income. This coefficient may or may not change if we add and deduct certain other independent variables in the model. If we find the same result in all these cases, we can say that our result is **robust to changes in the set of independent variables** (or **robust to changes in the covariates**).

Do robustness checks solve the problem of overfitting? Robustness checks are generally seen as good practice, and many journals will ask economists and econometricians to do them before publishing their articles. But robustness checks do not provide guarantees. Robustness checks are not based on new data, and can therefore not fully address concerns about double counting. Nevertheless, a robustness check is a way to make the result more severe. After all, robustness checks make it more difficult for incorrect models to pass. If a false model only fits because researchers have made an incorrect assumption, a robustness check may help discover this.

Importantly, robustness checks can only be conducted with respect to things that an economic modeler can actually control: the number of included variables, the data points included in the analysis, etc. The modeler can never check if the results are robust to changes that they do have under their control: such as the inclusion of **immeasurable variables**, or multicollinearity of variables included in the model.

7 • CONCLUSION AND BRIEF SUMMARY

The difficulty of econometric modeling should not be underestimated. As Keynes points out, the economic realm is complex. A model necessarily simplifies this complexity: it will always misrepresent the world to some extent, as econometric models will always make the economic world simpler than it actually is. Nevertheless, this does not mean that econometric models cannot correctly identify the relations between important economic variables, and their magnitude (the oomph, as McCloskey and Ziliak say, see chapter 2). But whether they do this depends on the quality of the model.

Models are typically assessed by their model fit, as we have seen, all econometric models involve double counting. Econometric models are based on observational data rather than experimental data, so econometricians need to use the data twice: once for estimating the variables, and once for testing it. Double counting stacks the deck in favor of the model and will make it likely that we will find a model with a good model fit, even if the model is not correct. Because of the problem of under-determination, an arbitrarily large number of models may fit the data, most of which will not be correct descriptions of the world. Some philosophers, such as John Worrall, think that we always need new data to test models built on old data. One way to do this is through predictive tests. However, if no future data is available, robustness tests can help assess the sensitivity of the model to some of the assumptions that a modeler needs to make.

LEARNING GOALS FOR THIS CHAPTER

After studying this chapter, you should be able to:

1. Describe the challenges that Keynes identified for Tinbergen's model, and, more generally, apply those challenges to econometric modellig more generally.
2. Explain the concepts of under-determination, use novelty, overfitting, and double counting, and explain their relationship
3. Explain different things a modeler may undertake to limit the challenges posed by Keynes, and the problem of under-determination.

6 Multiple testing, the file drawer problems, & meta-analysis

1 • MULTIPLE TESTING: THE BASICS

Many of the problematic examples that we have seen when we were looking at the classical approach to statistics in Chapter 3 had one thing in common: a significant test may sometimes look like strong evidence against a null hypothesis, but when it turns out that this actually comes from a large sample of tests, the evidence is no longer convincing. Mathematician Richard Gill calls this The Out of How Many Principle. We have seen this in the following examples.

- A test that people are able to see into the future led to a result that was significantly different from the null hypothesis that people cannot see into the future (Bem 2011; see Section 2.1). This looks at good evidence that people can see into the future. However, once we learned that the researcher conducted 10 different experiments, it was less surprising that 1 of them turned out to be significant. After all, we statistically *expect* a type 1 error once every twenty tests.
- Lucia de Berk was significantly more frequently present at incidents in the hospital than we would expect under the null hypothesis that she is innocent. However, when we realize she is one out of 70,000 hospital nurses working in the Netherlands at the time, we realize that

> it is not surprising that some of these 70,000 nurses are present at incidents more frequently than average. After all, we expect unlikely events to happen, when an event is frequently repeated.

With any significance test that a researcher employs, there is an expected number of false rejections, i.e. observing a p-value lower than the significance level, even though the null hypothesis is true. For example, if we do 1000 hypothesis tests of null hypotheses that are all certainly true, we would still expect a number of rejections.

> *100 psychic tests.* Imagine that we are testing which out the one hundred students in a class have psychic abilities. We will make each student predict a set of coin tosses in another room that you are not able to see. Our hypothesis is that no student is able to predict this, and all successes are a result of chance. We reject at an α of 0.05. What will we find? The expectation is that in these cases, we will find about 5 rejections of our null hypothesis: a result that should count as evidence against the null hypothesis according to the classical approach to statistics. This is because the α of 0.05 means that in 5% of the cases will you find a value that is this far removed from the hypothesized mean. If all hypotheses are true (i.e. no one is psychic), this corresponds to the expectation of false rejections.

In this example, it is not only likely that we will find rejections even if our hypothesis is true, but it is even *expected*. This is a problem: **the problem of multiple testing**. If we are doing a lot of tests, we will expect to find some rejections, even if the hypotheses we are testing are true. However, how do we know whether a rejection is simply a result of the number of tests we run, or whether our rejection signifies genuine evidence that our hypothesis is false.

Another way to put this is to formulate the problem in terms of

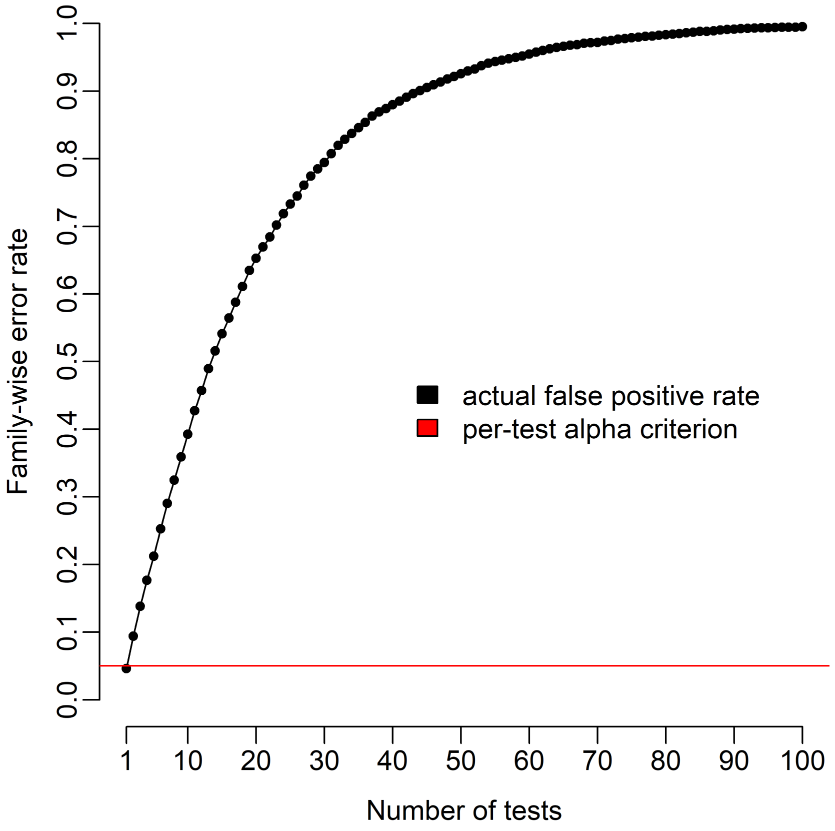


FIGURE 6.1 false positive rate and number of tests

the probability of making a false rejection (type 1 error). This type 1 error increases once we are conducting more and more hypotheses (see Figure 6.1). After 100 tests, the probability of making at least 1 type 1 error is almost 1.

You may think that this is another argument for Bayesianism and against classical statistics, but as we shall see later, it is actually a problem for both theories of statistical inference.

2 • CORRECTIONS AND FAMILIES OF TESTS

How to address the problem of multiple testing? There is a seemingly straightforward solution to the problem of multiple testing that is common among statisticians: we can correct our α to reflect the fact that this is not a singular test, but a test that is conducted in a **family of tests.** If we do so, we can keep the probability of incorrectly rejection a null hypothesis (type 1 errors) as low as 5%. These corrections thus make a distinction between the probability of making a type 1 error for hypothesis considered by itself, the standard α – which we can call α_{id} – that we are familiar with; and the probability of making a type 1 error considering that the test came from as set of tests, the **familywise error rate,** α_{fw}.

How does this work? In order to do so, we need to know (1) how many hypotheses belong to the **family of tests,** m; and (2) what is the desired false rejection (type 1 error) ratio for the family of tests, α_{fw}.

In the case we discussed above, the desired α was 5%, and the family of tests consisted out of 100 tests. We can now calculate the cutoff point of an individual p-value that would result in an overall probability of type 1 errors of 5% in all these tests, taken as a whole. This is called the Šidák correction:

$$\alpha_{id} = 1 - (1 - \alpha_{fw})^{1/m}$$

In this particular case:

$$1 - (1 - 0.05)^{1/100} = 0.0005128$$

So, if we would use an individual p-value cutoff point for all these 100 tests of 0.0005128, there would only be a 5% chance of finding a significant finding if all the null hypotheses were correct.

A simpler, and more common, approximation of the Šidák correction is called the Bonferroni correction:

$$\alpha_{id} = \frac{\alpha_{fw}}{m}$$

which, in this case, would simply equal 0.05/100 = 0.0005. This closely approximates the Šidák correction.

This mathematical solution is seemingly very effective, but there are two general problems, and as we shall see later, some problems that are specific to economics. **A first general problem** is that while these corrections keep the false rejections constant, it greatly increases the probability of false non-rejections (type 2 errors). In other words, it significantly reduces the **power**. Say someone in class is actually psychic. In order to find a p-value at an individual 0.00005 level will be difficult. It is thus much more likely that we will not reject, even if the null hypothesis is actually false.

A second general problem is that we need to know what the family of test is. How many tests are actually relevant for the comparison at hand? This, it turns out, is much more difficult that it may at first sight appear. We will discuss this issue below

3 • THE LOOK ELSEWHERE EFFECT AND POWER

Before I go on to examine the philosophical problems with multiple testing in the context of economics, we will discuss one important example from physics. In 2008, the Large Hadron Collider in Geneva began operating with an important mission: finding a particle that was hypothesized to exist by what is called **the standard model**. The alternative model was called the **Higgsless model.**

A large Hadron Collider is generating data from particles that are accelerated with great energy. This data is statistical in nature (see Figure 6.2). The LHC began gathering data by running experiments at different energy intensities. On some level of intensity, they would expect the existence of the Higgs particle to create a significant deviation from the Higgsless model (the red dotted line). So, the aim was to find a statistically significant deviation from the Higgsless model on some level of energy intensity. This meant that a large number of tests needed to be conducted, at different energy intensities (the X-axis).

Because so many tests were conducted at different energy intensities, the research teams working on discovering the Higgs boson corrected for **multiple testing**. They called the problem of multiple testing **the look elsewhere effect**. The reason for this is that someone finds a deviation somewhere, but then discovers that other researchers have

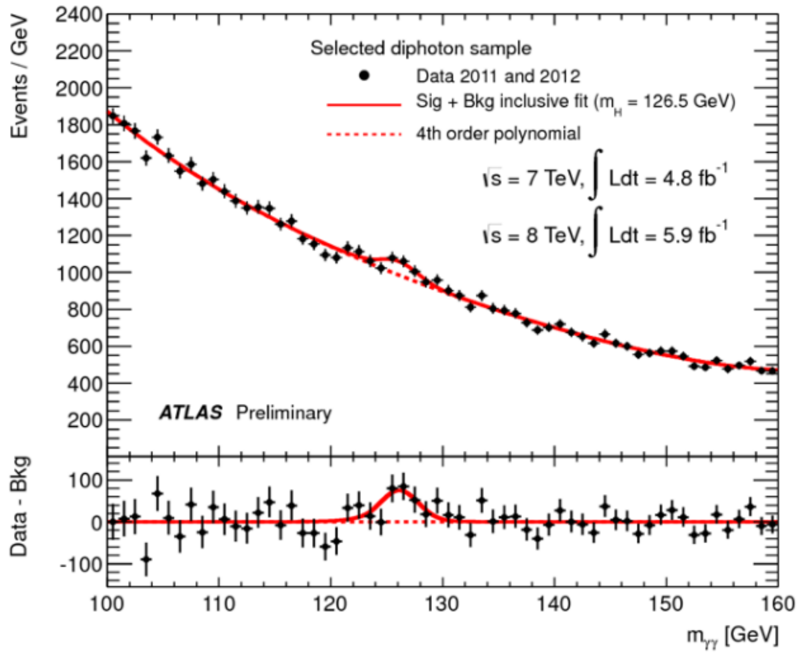


FIGURE 6.2 Higgs boson discovery.

been looking for deviations elsewhere (say, at a p-value of 0.03), a significant deviation may suddenly not be significant anymore, if a corrected significance level is used. So, *looking elsewhere may make your significant finding insignificant*, because we know that running more tests should require a correction, and a correction will decrease the individual cutoff value.

This problem, for the particle physicists, was not only a theoretical problem. Louis Lyons, one of the statisticians at the project writes: "we have all too often seen interesting effects at the 3σ or 4σ level go away as more data are collected" (Lyons 2008, 904). Because of the search conducted in the project, Lyons writes: "Thus the chance of a 5% fluctuation occurring somewhere in the data is much larger than might at first appear."

Ultimately, the researchers at the LHC went for a 5-sigma significance level (equal to a p-value of 0.000003), and found evidence for the Higgs boson in April 2012.

4 • TWO QUESTIONS ABOUT MULTIPLE TESTING IN SPECIFIC FOR ECONOMICS

The discovery of the Higgs Boson highlights some interesting but troubling features of multiple testing in science. Multiple testing is not limited to individual researchers, but may occur in **a group of researchers**, even researchers that are **not familiar** with each other's searches.

In economics, things are different from physics, and this makes it even more difficult to correct for multiple testing for two reasons. This brings us to two problems related to multiple testing that are **specific to economics**. It is perhaps clear that things work differently in economics, but just to highlight why the statistical rigor used in the Higgs boson example is not really possible in economics, we can run through why this does could not really work.

For one, discoveries, and hypotheses tests, are **not coordinated** in economics. Probably one significant reason for this is that while it is extremely costly and difficult to do searches for evidence in particle physics, the opposite is true for macroeconomics. Running a regression on the data is very simple. And it would be very difficult for the science of economics to keep track of all the individuals who are running regressions and interpreting p-values. But, it would seem, that if we want to be serious about multiple testing, this is what we need to know. The question thus is: should economics coordinate its searches in the data, so that it can take account of multiple testing?

A second issue is that macroeconomics is based on **a limited dataset**: the economic data of the world as it is actually occurring in the world. As we discussed in Chapter 5, macroeconomics cannot base itself on designed experiments. We can only generate new macroeconomic data as fast as time goes (only 1 data point per annual data set per year). At the same time, many economists use the same data set (data from OECD countries, the United States economy, etc.). In light of multiple testing this is problematic, because as we control for more and more tests, the significance level required to keep the familywise error rate at a constant level goes down, and more data is required to keep power at a sufficiently high level. If we use small datasets and a significance level that is very low, it is very unlikely that we will ever reject anything.

5 • WHAT IS A FAMILY OF TEST EXACTLY?

You may be surprised to learn that there is no clear answer to the question what defines a family of tests. However, we do know two important things:

- **A family of tests is not limited to a single researcher**. Some families of tests include tests that different researchers run. This is exactly where the look elsewhere effect gets its name from. Exactly because multiple testing is not limited to individual researchers that are testing hypotheses, a search conducted by another researcher may sometimes influence the significance level that you should use. Think, for example, of our main example that we discussed above: if instead of one researcher, 100 researchers would conduct 100 tests, we would be expected to find 5 significant findings.
- **A family of tests is not limited to a single dataset**. Strictly speaking, the data set in *100 psychic tests* above are all different datasets. In search of the Higgs boson, the researchers combined many different datasets. Datasets thus also do not properly limit the family of tests.

This means that any test that is run that is run to answer a specific research question can be part of the family of a test. As long as we are answering the same research question, all the tests that are run on this research question, by us or others that we do not know about, we should take this into account in our family of tests.

Take our example *100 psychic tests*, if we would be asking different questions every time: “Is student A psychic?”, “Is student B psychic?”, etc., these are all different research questions. And, for all these individual questions, the error rate will be 5% when we use a significance level of 5%. But, when we ask the more general question: “Are any of the students in class psychic?”, we should use a familywise error rate, because we have to take account of the fact that we are expected to find false rejections, even if all null hypotheses are true.

This illustrates how difficult it is in practice to properly account for multiple testing. One of the reasons for this is related to the way that science is organized. Typically, in most statistical sciences, **only the significant results** get published. So, even if 95 of our significance test are insignificant, other researchers will only see our 5 significance tests that are significant that possibly get published. So, even if all the null hypotheses are true, we expect that some significant tests will be published, even though these are false rejections. We expect this, even when 100 researchers are all investigating their single hypotheses. Unbeknownst to any of these 100 researchers, the tests that they run are therefore part of a family of tests, and the 5 significant tests are not an indication that something truly significant has occurred. This is called the **file drawer problem**.

6 • PUBLICATION BIAS AND THE FILE DRAWER PROBLEM

In a paper from 1979, statisticians Robert Rosenthal discussed a problem in the evaluation of statistical research: in order to be publishable, a finding generally has to be statistically significant. But what happens to the findings that were not significant? They disappear in metaphorical (or literal) file drawers. This means that the economics findings that end up in scientific papers are a biased subset of all the findings of economists. For our example above, an unrealistic extreme example, this would mean that we would be able to publish 5 papers with significant findings about the psychic abilities of econometrics students. But this would give an awfully bad reflection of what the evidence really is. The effect of the file drawer problem is that we will only be able to see the significant results, but without knowing what the positive results are, we have a biased perspective on the evidence. This bias is called **publication bias**.

To summarize: the problem of multiple testing shows us that we don't only need to know if, and how many, significant tests we find, but also how many non-significant tests we find. However, due to file drawer problem, we do not know what the amount of non-significant tests is, and this selective presentation of evidence is called publication bias.

How bad is the problem of publication bias? This is very difficult to know. It depends on (1) how many tests are run, and (2) what is the power of the tests? Or, to put this alternatively: how many of the null hypotheses are actually true?

Imagine our main example, but in this case, half of the students actually *are* psychic. If the power of our test is 0.8, we would find

False reject: $50 * 0.05 = 2.5$

False non-reject: $50 * 0.2 = 10$

Correct reject: $50 - 10 = 40$

Correct non-reject: $50 - 2.5 = 47.5$

So we would in total observe 42.5 rejections, and 57.5 non-rejections. Out of these rejections, only 2.5 are false rejections, so out of all the individuals that the test would identify as psychic, only $2.5\%/42\% = 6.25\%$ are incorrectly identified as psychic. That is not that bad. In the initial case, all (100%) of the rejections were false rejection! The extent of the problem depends on how many of the hypotheses we are testing are actually true. Or, in other words, it depends on the prior probability of the hypotheses.

This may sound like Bayesians have an advantage here, but the prior probability that a random hypothesis that is being tested in an econometric study is true or false is an uncertain probability. In other words, there is no statistical basis for having a justified belief about whether a hypothesis is true, before we test that hypothesis.[8] So, regardless of whether you are Bayesian or not, it is difficult to know how many non-significant findings lie behind every significant one. Bayesians may take account of it in their priors, but this will be based on a guess of how many hypotheses are actually true, or false.

It is good to note that the file drawer problem comes up, even when economists have the best of intentions. No fraud, or deliberate attempts to manipulate the results are necessary to get at the expectation that a non-trivial number of findings (p<.05) will be statistical artefacts. But, we also know that researchers do face significant pressure to publish, and publishing typically requires significant results. We typically assume that fraud does not happen in scientific inquiry. Researchers are supposed to be interested in the truth, and fraud means lying, the deliberate obfuscation of truth. Even the Reinhart & Rogoff controversy, described in the introduction, was apparently based on an honest mistake. But fraud does happen. In 2021, some researchers found that a 2012 study co-authored by world renown economist Dan Ariely on honest (I do not make this up), was based on fraudulent data (Shu et al. 2012 for the now retracted study; see Stern 2023 for a journalistic account; and see Ritchie 2020 for an excellent book on this general problem in science). But even without outright fraud, the pressure to publish statistically significant results can affect the family-wise error rate of the whole field in other ways. In a survey from 2014 among academic economists, 32% said that they "[p]resented empirical findings selectively so that they confirm one's argument",

[8] In the 100 psychic test case, a Bayesian would say that we have good reasons to believe that none of the hypotheses are true. After all, we have never found evidence for any psychic ability in anyone, so it would be rather odd if we would find it now: our prior would have to be low. However, even so, for any individual hypothesis, we are expected to observe some particularly low values of $(E|H)$ if we conduct many tests. So, while the Bayesian can say that there is, after the test, still a small probability of any of the individuals being psychic, they would have to accept that for at least some of the individuals, the posterior probability of the hypothesis that the individual is psychic would be much higher than the priors. So, this is a problem for both theories of statistical evidence.

and 38% agreed that they had "[s]topped statistical analysis when you had a desired result" (Necker 2014; discussed in Ritchie 2020, 97). If we take this into account, the likelihood that our findings (P<.05) represent statistical artefacts only increases further. Moreover, if we follow such strategies, false rejections (and false non-rejections) will be particularly biased in a way that fits with the interest, or bias, of the researcher conducting them.

7 • SUM-UP SO FAR

Multiple testing is a difficult problem in econometrics. While it is relatively easy to adjust your α in case you know what your family of tests is, this has two significant disadvantages. For one, it will significantly reduce the power of your test: because the required p-value is so low, it will be less likely to discover false hypotheses. Second, it is very difficult to determine what the familywise error is for specific estimations. The familywise error rate requires us to know how many tests were run in total on a specific research question, but the file drawer problem makes this practically impossible. Without adjusting the error rates for our family of tests, it is difficult to know how reliable our statistical findings really are. If we take into account that there is significant pressure on academic economists to find statistically significant findings, we should be all the more concerned about the significant p-values that this results in.

8 • A NEW HOPE: META-ANALYSIS AND FUNNEL GRAPHS

Is there nothing that science can do about this problem? There is something, and it is called **meta-analysis**: the evaluation of sets of different statistical findings about the same research question. Meta-analysis refers to a wide-ranging set of methods to evaluate the following question: different published results find different estimates when they are estimating the same coefficient or other parameter in a statistical study. Sometimes they are significant, sometimes not. They may even sometimes be positive, and other times negative. What should we conclude from all these different studies?

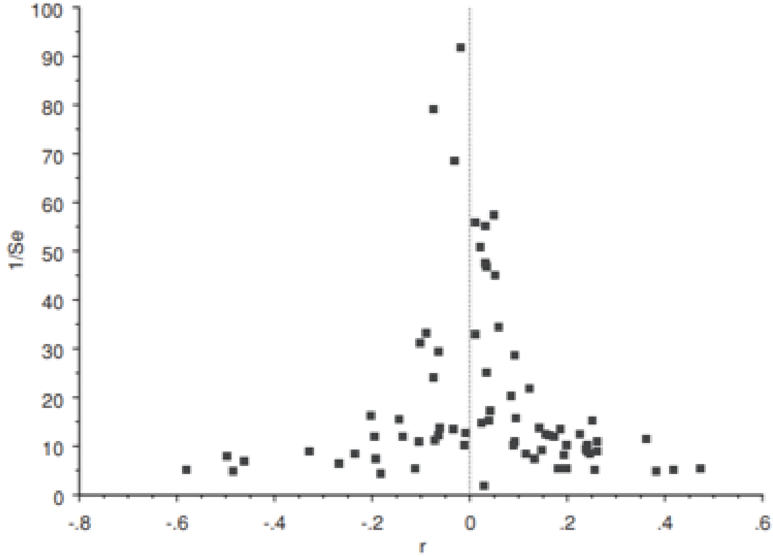


FIGURE 6.3 Union productivity partial correlations estimates and their standard errors (from Stanley and Doucouliagos 2010).

Meta-analysis can actually help with accounting for publication bias in two ways. It can estimate, or illustrate, whether there is publication bias, and how bad the problem is. It can also, at least in part, correct for it.

Meta-analysis requires a set of estimates from other studies. This allows us to make **funnel graphs**. Funnel graphs picture a set of findings: their estimates on the x-axis, and a measure of precision on the y-axis. This measure of precision is generally 1 divided by the standard error of the estimation, or $1/SE$. The higher this number, the more precise we would expect our estimate to be. If there would be no publication bias, we would expect a funnel graph to look like the first figure (Figure 6.3): the more precise an estimate is, the more closely we would expect it to lie to the true value, and less precise estimates we would expect to be distributed around this value. In case there is publication bias as a result of the fact that only significant results get published, we would expect imprecise estimates, so estimates at the lower end of this graph, to drop out. This, however, is not a problem for the picture as a whole, if the unpublished results are not biased in a particular direction. In case of unbiased publication bias, we should expect that the true estimate will lie somewhere in the center point of a symmetric funnel graph.

However, in some cases, publication bias may skew the results. If so, we see something else. Figure 6.4 is an example of this. There is big gap in the literature where we would expect some findings, just to the right of the zero-line. This graph is depiction of estimates about the price elasticity of water. The most precise estimates are, as we would expect, quite close to 0, but slightly negative. However, there are almost no positive estimates, even though we do see quite a few much more negative, less precise, estimates. Why do we not see any slightly positive estimates? Probably because researchers would be highly surprised to find that a positive price elasticity for water demand: when water costs more, would people consume more of it?(!)

Consequently, results that find such a surprising finding will be unlikely to publish their results, and if they try, are more likely to get rejected. This, however, creates a clear publication bias for negative estimates. If we would only look at the average of the estimates as a whole, however, not taking into account the precision of the estimates, we would highly exaggerate the negative effect of water prices on water demand.

The graph in figure 6.4 is a clear depiction of publication bias. It shows that some results simply do not get published, even though they are there: there simply is no statistical explanation of why a graph would have such a gap as it has in figure 6.4. In many cases, things will be more subtle. See the final graph: figure 6.5. This is a depiction of research estimates that are published about the relationship between minimum wage and unemployment. As you can see, the number of studies is high. We can also see that the reliable estimates center quite closely around 0. Nevertheless, the largest body of the data are negative estimates. There are also some positive estimates, but the negative ones

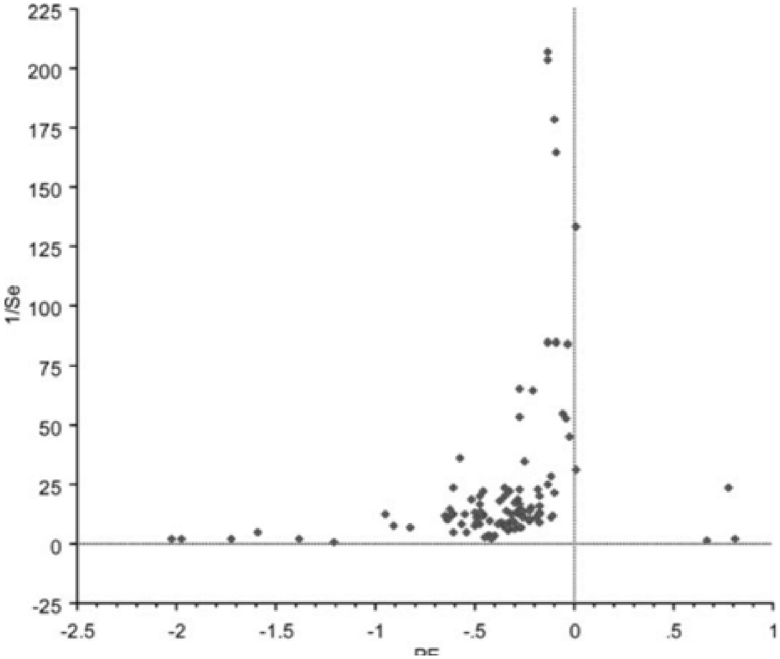


FIGURE 6.4 Estimates of price elasticities for water demand and their standard errors (from Stanley and Doucouliagos 2010).

are highly outnumbered by the positive estimates. However, the graph is clearly skewed. There is no explanation for this skewed relationship except for publication bias. In other words, positive estimates are less likely to be published.

What this discussion shows is that funnel graph can actually help us detect publication bias. And it seems that the problems are indeed there. Non-significant results may be less likely to get published, but we can also see that results that do not fit with particular theoretical assertions are less likely to get published. The combination of these two observations may lead to a significant bias in the published econometric results.

However, funnel graphs also show that estimates of coefficients can be gathered from focusing on the most reliable estimates. The more

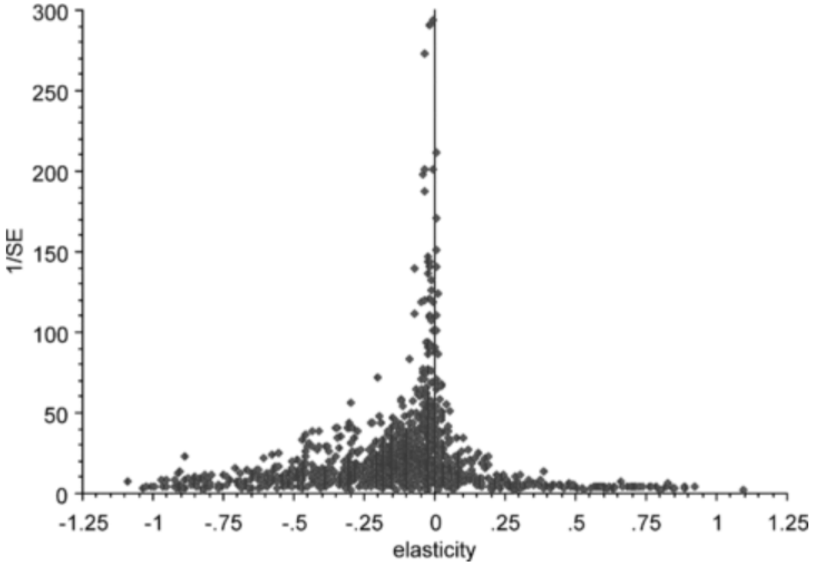


FIGURE 6.5 Estimates of the effect of minimum wage of unemployment and their standard errors (from Stanley and Doucouliagos 2010)

reliable they are, the more they center around a specific mean, that seems to be a good indication of a true mean.

Funnel graphs are a good indication, but they also have a significant drawback. They can only be made in fields of research in which a lot of data is available. If we are interested in relatively unexplored topics, it is more difficult to estimate what the impact of multiple testing is, or will be. This is particularly a problem for fields of economics in which the data availability is quite low. In all these examples of the funnel graphs, a lot of studies were available. However, in some fields of study, the data availability is much smaller.

9 • CONCLUSION

Running multiple tests at the same time has an important influence on the reliability of a statistical inference, and when this happens, we have to interpret the results differently. When we have clearly defined family of test, this is fairly straightforward. We can simply use a familywise error rate. However, in economic practice, it is both difficult to interpret what a family of test should be, and to have the information available that is required for calculating the family of test.

Funnel graphs give an indication of whether there is a clear bias in the published literature, at least in case a sufficient amount of results have been published. As it turns out, publication bias can be a significant factor, although funnel graph can help us identify where biases lie. The file drawer problem shows how important meta-analyses are.

This leaves a reduced, but significant set of econometric tests to be liable to issues of multiple testing and publication bias: cases in which the samples are limited. This seems particularly problematic in fields of macroeconomic in which the same datasets are used repeatedly. The example of empirical studies of long-run economic growth is one such example. In these cases, even if there are different researchers that calculate different coefficients, these coefficients are all based on the same or overlapping datasets, and will thus be dependent on each other. Coming up with an estimate of publication bias will be much more difficult. In these cases, the interpretation of p-values, and the specification of a proper significance level, leaves much room for debate.

LEARNING GOALS FOR THIS CHAPTER

After studying this chapter, you should be able to:

1. Explain the concept of multiple testing, and why it is a threat to statistical inference.
2. Explain how we can correct for multiple testing, and explain the concept of a family of tests.
3. Explain the look elsewhere effect and its relation to multiple testing.
4. Explain the concept of the file drawer problem, and publication bias, and explain how meta-analyses, and funnel graphs can detect biases.

7 Causality

1 • CAUSALITY: A CHALLENGE FOR ECONOMETRICIANS

You may remember from our discussion about Keynes in Chapter 5 that Keynes took the goal of economics to be the discovery of the causal relationships. To cause something is to do something that has an effect. Textbook writer Jeffrey Wooldridge agrees:

> "In most tests of economic theory, and certainly for evaluating public policy, the economist's goal is to infer that one variable (such as education) has a causal effect on another variable (such as worker productivity)." (Wooldridge 2009, 12)

For example, we want to:

- estimate financial returns, in order to increase profits;
- estimate the effect of minimum wage on unemployment in order to implement the minimum wage level that is most desirable;
- establish which variables determine economic growth, in order to increase economic growth.

In all these cases, our ultimate interest is to obtain knowledge of **causal relationships.**

As we get taught in our introduction to statistics classes: **correlation does not imply causality**. At the same time, in order to be useful, in order to have economic significance, statistical findings need to tell us something about causality. Ultimately, statistical findings in economics derive their importance *only* from the way they help us **change** the world. In order to change the world in predictable ways, we need to have causal knowledge.

Think of an example in which there is a correlation but no causality. The price of bread and gross GDP are correlated over time in almost all countries, because both are upward trending. However, it is not true that increasing the price of bread will increase GDP in important ways. This correlation, then, is not useful: if we want to increase GDP, this finding does not have any policy implications. Without knowledge about causality, correlations do not give policy makers any useful information for guiding policies.

Because correlation does not imply causality, the official line is that econometricians should be careful with deriving causality from the observed correlations and model fit: in principle, doing so is not possible and should be avoided. However, there is nevertheless much language use in econometrics that betrays that the ultimate aim is to learn about causality in the world. For example, the notion of an "effect size" clearly hints at the idea that the coefficient does not merely indicate a conditional correlation coefficient, but that there is an **effect**, or, in other words, a causal relation. The same is true when econometricians say that one variable "affects", or "influences" another. In fact, the whole notion of a **dependent variable** derives from the idea that this variable **depends** on the independent variables. All of these notions are instances of what we may call **causal language**.

This causal language is only a symptom of a more general tension that econometricians have to deal with. We can call this the econometrician's **causality tension**:

1. What we can observe in economic data are correlations.
2. Correlation does not imply causality.
3. Ultimately, we want to know what the causal relationships are in economics.

In other words, correlations do not get us causation, but correlations without causation is not useful. In this final chapter, we will analyze this tension, and see if there is anything we can do to alleviate it. We have already discussed Point 3. What we now turn to Point 1: what is causality, and why is it so difficult to observe? In doing so, we look at the philosophical problem of causality. As we will see, in some cases, we *can* infer causality from the data, even though this is not often true for economic data in general. Secondly, we will discuss Point 2: does correlation really not imply causality? Perhaps our statistics 101 teachers have been a bit too quick here. Finally, we will discuss how this relates to econometric practice. How good are econometricians at observing causality?

2 • THE PROBLEM OF CAUSALITY

One of the most influential thinkers shaping our understanding of causality was Scottish philosopher, and good friend of economist Adam Smith, David Hume (1711-1776). David Hume was troubled by the concept of causality. Take for instance the following question: when I hit a pool ball with my pool stick, will it move? The simple answer to these questions, of course, is yes. But why are we so certain of this? We know that statements like these are true, because of the regularity we observe in the world. Whenever a round ball, that is not glued or stuck to the floor, is hit with some force, it moves. The force on the ball *causes* it to move. However, Hume observed, we never *really* observe this causality. What we do observe is a sequence of events that always occur after each other:

Event 1: the ball is pushed

Event 2: the ball moves

What we do not see is causality itself. This problem has come to be known as Hume's **problem of causality**. We only see events that are what Hume calls **spatio-temporally contiguous**: they happen right after each other, the effect following the cause, at the same location.

However, not everything that always happens at the same time and place has a direct causal link. For example, hemlock is a poisonous plant that has a peculiar taste. If you eat it, you will experience this taste, after which you die. This taste then always comes before the poisonous reaction. However, this taste does not *cause* you to die. It is a mere byproduct of the plant, whose poison kills you. So, spatio-termporal contiguity is not itself sufficient for causality. But how *do* we know that something causes something else?

What causality means then, according to Hume, is just following (Reiss 2013):

> X is a cause of Y if and only if:
>
> - X is universally associated with Y;
> - Y follows X in time;
> - X and Y are spatio-temporally contiguous, i.e. there are no time-wise or space-wise gaps between X and Y.

This also shows that there is a problem, because we cannot observe "universal association". The **problem of causality** is that we often see things happening after each other, but that we cannot see causality itself. This problem should be familiar to you: it is a particular variant of **the problem of induction**. When we say that a push on a ball forces it to move, is to say that in the same conditions, a push on a ball will *always* result this ball to move. In other words, a causal claim is always a **generalization**. And, as the problem of induction shows, we can never derive a generalization from particular observations. As causal claims are types of generalizations, it seems that we cannot derive causality from observations.

Hume was an **empiricist**: someone who believed that empirical observations are the only reliable tool for learning about the world. But empiricism runs into trouble if it wants to establish causality. If we want to base ourselves on observations, we can never truly observe causality, the thing that ultimately matters.

3 • THE PRACTICAL PROBLEM OF CAUSALITY

This problem so far seems purely philosophical: will the ball move if I push it? Of course! Only philosophers can take such question seriously. However, there is also a practical problem related to causality that are the result of the fact that we cannot observe causality itself. This practical problem is that we can observe correlations, but correlations do not provide sufficient information to determine causal relationships by themselves.

Think about our now familiar example of minimum wage and unemployment: if we increase the minimum wage, will unemployment increase, and if so, by how much? Even if minimum wage and unemployment move together, we cannot tell that this is because minimum wage increases unemployment. While we may think that a higher minimum wage causes there to be more unemployment, there may be other causal effects occurring that are mitigating or amplifying this causal effect in the data. For example (with a "–" indicating a mitigation, a "+" an amplification):

- whenever unemployment is low, governments implement higher minimum wages (-);
- whenever there are progressive governments, minimum wages go up, but progressive governments are more likely to be elected in economic downturns (+);
- minimum wages increases total consumption, stimulating economic activity, and decreasing unemployment (-).

These other causal factors may make it more difficult to detect the causal effect of interest in the data. After all, which part of the correlation, if there is any correlation, can be ascribed the causal effect of interest, and which part cannot?

4 • APRIORISM

Some economists and some philosophers think that we just have to accept that we cannot observe causality, and observations will not help us identify causal relationships. Rather, our judgments about causality should not come from observing the world, but from our understanding of economic theory. This view is called apriorism, from the latin "a priori", which means "before the evidence". Theories, apriorists claim, can be derived without observations. One example of how this type of reasoning can help us determine causal relationship is the example of economic theories of rational behavior. It is difficult to determine how much a price increase will affect demand, but we know from theory, i.e. before making any observations, that it will have a negative effect on demand. After all, we can know, before the evidence, that it is rational to we prefer having more money rather than less. If people are indeed somewhat rational, an increase in price should lead to a decrease in demand. These thinkers claim that all reliable causal knowledge comes from theory.

This solution is less than satisfying. After all, when econometricians are advising policymakers on whether to increase minimum wage or not, we want our answer to be "evidence-based": justified by the evidence. It is notable that some theory-minded economists were extremely skeptical of Card and Krueger's minimum wage study, because they figured that it is an a priori fact that higher wages should lead to a lower demand for workers. Nobel prize winning economist James Buchanan wrote in the Wall Street Journal:

"The inverse relationship between quantity demanded and price is the core proposition in economic science, which embodies the presupposition that human choice behavior is sufficiently rational to allow predictions to be made. Just as no physicist would claim that "water runs uphill," no self-respecting economist would claim that increases in the minimum wage increase employment." (*Wall Street Journal* 1996)

The problem with such a strong faith in economic theory is that economic theory now becomes scientifically instructinable by evidence. Even physics should be open to the possibility that someone discovers water that runs uphill, and that we should consequently revise our theories. Even if theory plays a role in reasoning about causality, we should acknowledge that theory is fallible.

5 • JOHN STUART MILL'S SOLUTION: EXPERIMENTATION

If we cannot observe causality, how can we ever be sure that one thing causes anther? 19th century philosopher-economomist John Stuart Mill (1806-1873) presents an important solution to this problem. Mill called his solution the method of difference:

> **Method of difference:** If two cases are exactly the same, except for 1 thing, C, and two different outcomes result from these two cases, than we know the difference between these two cases was **caused** by C.

Mill's method of difference is essentially what has come to be the underlying inferential method of experiments. For example, in medical experiments, experimenters generally give two randomly assigned treatment groups two different treatments: one placebo, and one new medication. If the group that got the new medication did better, this must be *caused* by the new medication. After all, the groups are exactly the same, except for 1 factor: the pill. The different outcome **must logically be caused** by this one factor.

Mill's method of difference is remarkable. Even though it does not solve Hume's problem entirely, after all, a staunch sceptic will say that it still does not allow us to observe causes, it helps us capture them in an resourceful way. The method of differences leaves no other explanation except for the fact that the difference must be due to the one factor that is different.

The experimental method lies at the basis of much scientific methodology, but, we have also seen that in econometrics, doing experiments is difficult. The limitation of the method of difference is that it requires us to be sure there really is no other factor that we have not considered, that could be driving the different outcomes? We call such factors **confounding factors:**

> **Confounding factor:** If two cases are the same, except for 2 things, C and **Confounding Factor**, and two different outcomes result from these two cases, than we do not know whether the different result was due to C or due to Confounding Factor.

The problem with confounding factors in econometrics is that there are always some possible difference between two groups , and it is difficult to establish that two cases are exactly the same except for one factor. For example, if we do an experiment by assigning different companies different minimum wages to implement, a skeptic can always say that the companies were on some level different. For example, one could claim that all the companies that accepted the invitation in the experiment were the companies that had a particularly friendly managing style, biasing the results. Or we can imagine that the randomization did not equalize all the relevant differences: the companies that got assigned a particular minimum wage were more often in the east of the country, where the costumers are wealthier, biasing the results.

In macroeconomics, we rarely encounter cases that fit perfectly with Mill's method of difference. However, some econometricians believe that we should still try to emulate this model as much as possible. Below, we will explore to what extent Mill's method can help us identify causes in economics.

6 • CORRELATION AND CAUSATION: THE REICHENBACH PRINCIPLE

Mill's method of difference alleviates some philosophical skepticism about causality. Is there any other way in which we can identify causes in economics? One way to do so is to return to the relationship between correlation and causation. They are not the same, but can we really not learn anything about causality from correlations?

Why does correlation not imply causation? One common example to illustrate this is the following:

> Ice cream sales correlates with forest fires (see figure 7.1

below). Therefore, forest fires are caused by ice cream consumption.

Clearly, reducing ice cream consumption will not affect forest fires in any way: the causal relationship is not in any way genuine. However, it is still too simple to say that there is *no causal* relationship here. The causality, however, comes from another, third, factor: temperature. Is that what is going on in all cases in which there is a correlation that is not causal in the expected way? There is simply a third factor? Correlation then *does* imply causation, but not always the one that you may think of at first.

German philosopher Hans Reichenbach formulated the following principle to connect correlations and causation:

Reichenbach principle: Any two variables A and B are correlated if and only if either (i) A causes B, (ii) B causes A, (iii) a common cause C causes both A and B, or (iv) any combination of (i)–(iii).

If Reichenbach is right, correlations *do* imply causation, just not always the ones that you expect. Let's go over these options in turn.

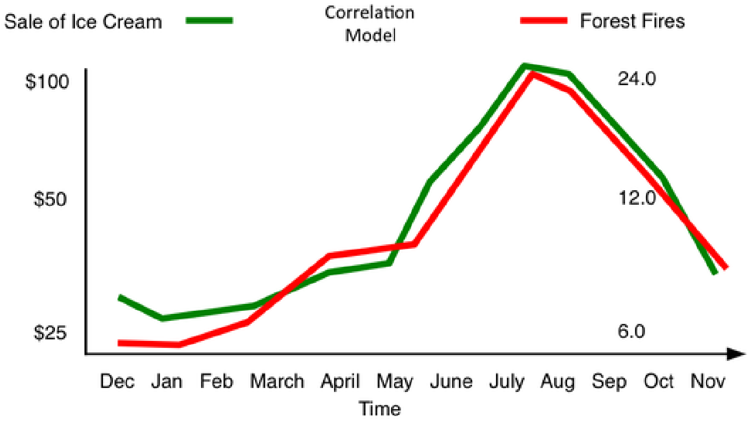


FIGURE 7.1 Forest fires and the sale of ice cream through time. Retrieved from https://www.decisionskills.com/blog/how-ice-cream-kills-understanding-cause-and-effect.

Consider our example here: variable A is ice cream, variable B is forest fires. As we know, they are correlated. The possibilities are:

1. forest fires cause ice cream consumption;
2. ice cream consumption cause forest fires;
3. a common cause (a third factor) causes them both;
4. a combination of these three.

Clearly, 2 is false, and 1 as well. However, indeed, a common cause, namely temperature, affects them both. This principlean thus help us identify causation in the world. Even though we cannot derive causation from correlations directly, it can identify a small range of possible causal relationships that must be there.

There are, however, **three exceptions** to the Reichenbach principle. **First**, see the following graph of the US GDP and GNP. They are, to put it mildly, highly correlated.

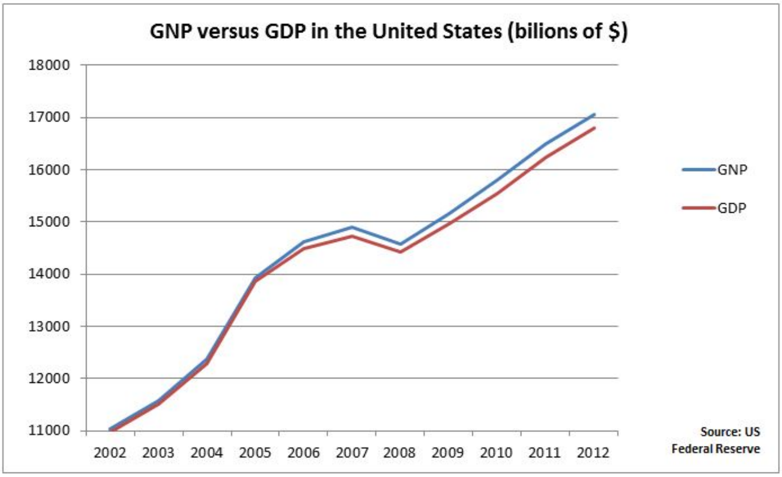


FIGURE 7.2 The correlation between GDP and GNP.

Nevertheless, it would be very odd to say that GDP and GNP are causally related. The reason for that is that they are simply highly overlapping by definition. Their relationship is close, but it is not a causal relationship, but a **constitutive relationship**: they are defined in almost the same way. A similar relationship exists between number of individuals unemployed in Germany, and number of individuals unemployed in the EU. This relationship will correlate, though much less than GNP and GDP, but not because they are causally related. The first is simply partly made up of the same information as the second.

Second, consider the correlation that we have discussed between the movements of Paul the Octopus, and the country winning in matches in the 2010 World Cup. While the correlation was perfect, and statistically significant, we realized that we do not have much trust in the results. Another example is the graph below that illustrates the correlation between age of Miss America and murders by steam, hot vapours, and hot objects. If the Reichbach principle would apply here, either the age of Miss America would cause these murders, these murders would cause the age of Miss America, there would be a common cause, or a combination of these. None of these seem particularly plausible. So, what is going wrong?

These examples, are examples of **spurious correlations**.

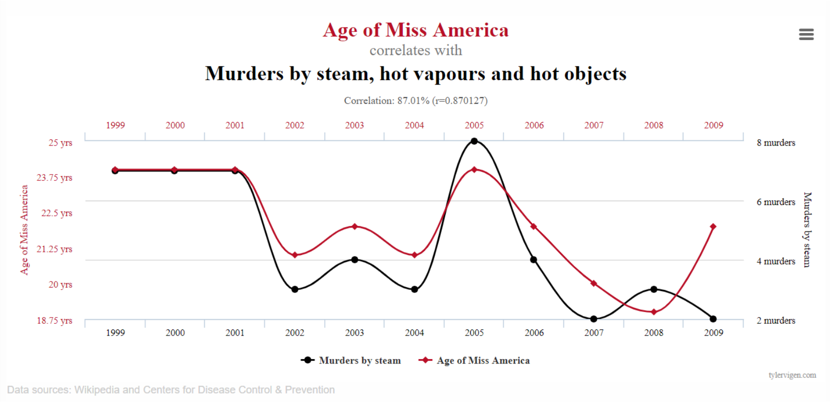


FIGURE 7.3 A spurious correlation. Retrieved from: https://tylervigen.com/view_correlation?id=2948.

> **Spurious correlations:** correlations that are solely the result of coincidental concurrences.

How can I be sure that this is the case? Identifying spurious correlations is difficult, but they have a couple of features. Spurious correlations tend to be based on small samples, are derived from multiple correlations (called "data dredging", or "p-hacking"), or, even more commonly, both. The larger a sample is, the less likely it is that we would observe a spurious correlation. Spurious correlations are also coincidental, so, if a correlation is spurious, the relationship is expected to be absent in new incoming data. This is one motivation for the **use novelty requirement** from Chapter 5.

A **third** category of exceptions to the Reichenbach principle may look similar to the second. For example, take a look at the following relationship:

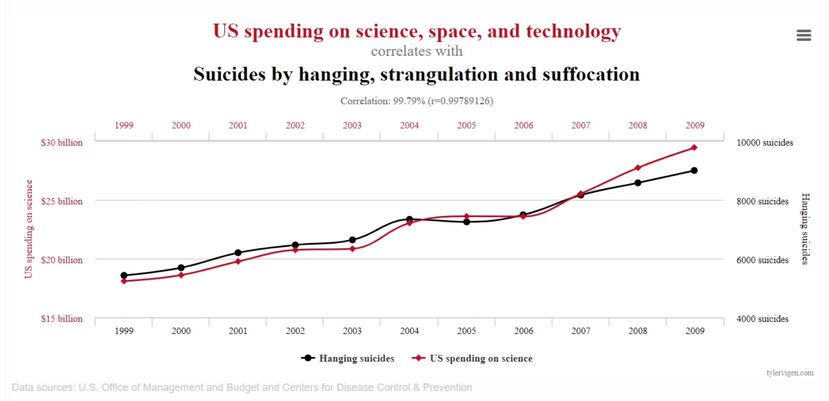


FIGURE 7.4 Two time-trended variables. Retrieved from: https://tylervigen.com/view_correlation?id=1597.

Clearly these two variables are not causally related in any meaningful sense. Another influential case similar to this one is the correlation between the sea levels in Venice, that are increasing over time, and the bread prices in the UK, which are also increasing over time. What is different in these cases compared to the second exception? If more data comes in, this relationship is unlikely to disappear. The United States is increasingly spending more money on science, space, and technology. And, unfortunately, suicide rates are also up in the Unites States. The bread price is expected to keep increasing, just as the sea levels in Venice. Nevertheless, their relationship is still clearly not causal in any meaningful sense. The problem is that both these variables simply have **a time-trend**. Time-trends are very likely to have a relationship to each other, even if there is no causal relationship between them whatsoever.[9]

There is a way to avoid this category of exceptions. We can note that these two variables – US spending on science and technology, and suicides by hanging, etc. – are not really correlated in a meaningful way, because they are not **co-integrated**. Co-integration means that if we remove the time trend, the correlations persist. In these examples, this is not likely. The price of bread may particularly increase as a result of a good harvest season, but the sea levels in Venice will not respond to the same shifts. While this is a good way to test if a trended variable is meaningfully correlated with another trended variable, the Reichenbach principle is formulated about correlations in general, and therefore, the third exception remains an exception to this principle.

To sum up, correlations imply causation in some sense: if A and B are correlated, they are causally related, or there is a common cause. But there are three exceptions to this rule: definitional relationships, spurious correlations , and time-trended data

7 • ECONOMETRICS AND CAUSALITY: NATURAL EXPERIMENTS AND INSTRUMENTAL VARIABLES

As helpful as the Reichenbach principle may appear, it still leaves too

[9] Someone may object and say that this is not actually an exception to the Reichenbach principle: time is the common cause in this case. Time is the cause of suicide rates going up, as well as the spending on technology and science going up. However, time itself is not generally seen as a cause itself: rather, it is a dimension in which causes operate, just like "space" is not a cause itself, it is simply background needed for causes to operate in.

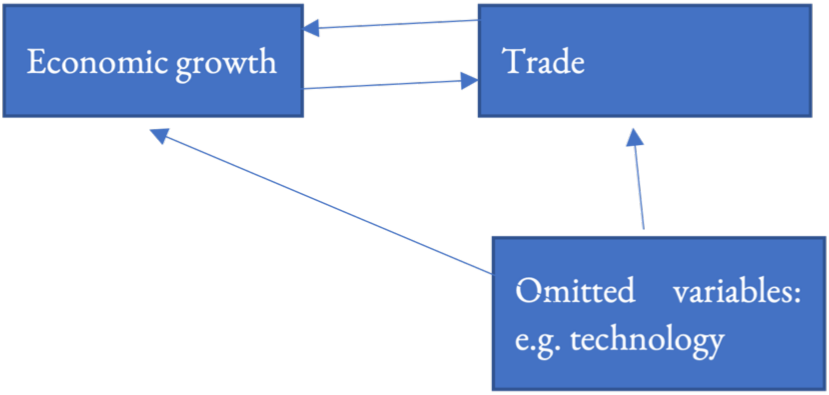


FIGURE 7.5 A model of economic growth.

many options open in order to be result in economically useful causal knowledge. Consider, for example, the relationship between trade and growth in international economics. There is a strong correlation between economies that trade and that grow. However, the correlation cannot easily be interpreted as a causal effect of trade on growth, because trade does not only positively affect growth, but economies that grow, may also trade more. In other words, there is **reverse causality**. Moreover, there are also other variables that are a common cause, such as technology: more technology is good for economic growth, but may also stimulate economic trade relations, for example, through making it easier to import goods over the internet. Even if the Reichenbach principle applies, we still want to know which particular relationship lies behind this correlation (see figure 7.5).

What can econometricians do? As we have already seen, Mill's method of difference is difficult to apply to macroeconomics on a large scale. However, in some rare occasions, **natural experiments** occur. Natural experiments are non-artificial cases that closely resemble experimental set-ups. In the language of Mill: cases in which, by coincidence, we have two cases that are similar in all respects, except 1.

One example of this the closing of the Suez Canal, a politically motivated event, in which Egypt, as a result of a conflict with Israel and the Western Nations, closed the Suez Canal for 8 years (from 1967 to 1975; Feyrer 2009). This created a situation in which for about half of the world, the main trading routes suddenly became a lot longer for 8 years, for the other, the main trading routes were not affected (see figure 7.6 below).

Economist James Feyrer thought that this would make an excellent natural experiment for assessing the impact of trade on economic growth. The impact difference in economic growth between these two countries over the years can then, perhaps, be described as a causal effect of the extending trade routes:

> Group 1: countries whose trade routes were affected by the closing of the Suez Canal;
>
> Group 2: countries whose trade routes were not affected by the closing of the Suez Canal;
>
> There are no further relevant differences between group 1 and 2.

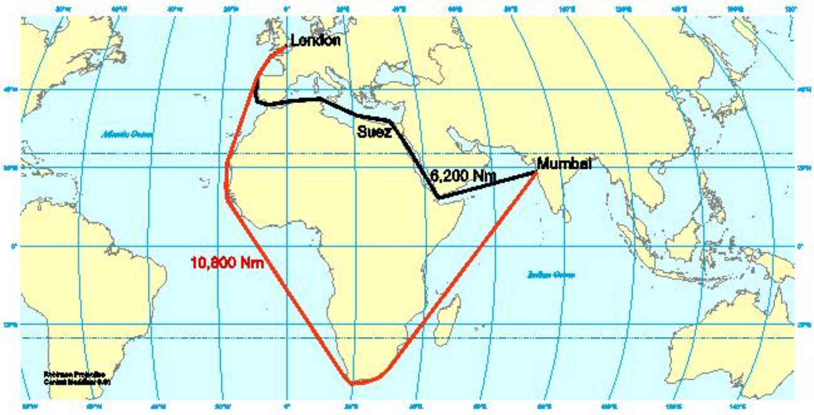


FIGURE 7.6 An example of increased travel distance between two cities as a result of the closing of the Suez Canal. Retrieved from: https://cepr.org/voxeu/columns/1967-75-suez-canal-closure-lessons-trade-and-trade-income-link.

The difference in economic growth between these two groups of countries can thus be ascribed to the causal effect of the extension of trading routes. Two caveats apply. First, in reality, these two groups of countries will, in many ways, be different. However, because the differences between these countries are **random**, their differences can be argued to not affect this analysis in an important way. So, while the experiment is not perfect, this natural experiment, closely resembles Mill's method of difference, and consequently, is likely to identify an effect that is causal.

Second, in reality, the difference between group 1 and group 2 is not strict, but differs by degree. However, for the identification of causality, this does not matter. Natural experiments create what econometricians call **exogenous variation**. Exogenous variation is the type of variation that is independent of anything else in the model. In our example, independent of economic growth, and technology (and other factors that may affect both growth and trade). This feature helps us estimate the causal effect of trade on growth. After all, the closing of the Suez Canal affected trade routes, the trades routes affect growth, and there is no reverse causality or another common cause (see Figure 7.7). So, the correlation that we observe here is genuinely causal.

We call this method of identifying causality through exogenous variation this way the instrumental variables approach, and we can call the variable that creates the exogenous variation the **instrumental**

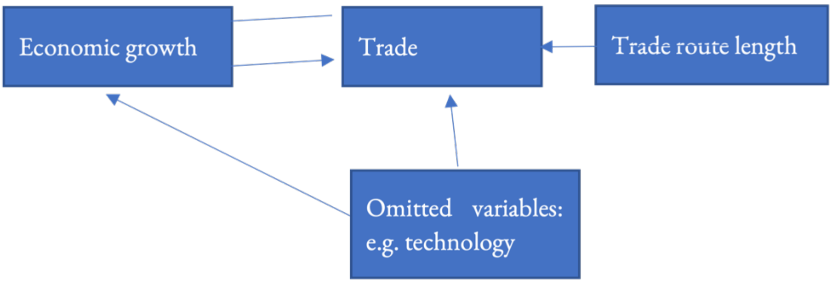


FIGURE 7.7 The relationship between trade route length, trade, economic growth, and omitted variables.

variable, or IV. Instrumental variables essentially apply the same logic as (natural) experiments, but rather than looking for events that bring about a difference in the world, they look at variables that bring about a specific effect in the world, namely, an **exogenous variation**. Exogenous variation is always present in (natural) experiments, but it may also exist outside of setting that is properly described as an experiment. We can define instrumental variables formally as follows: a variable Z is an instrumental variable if and only if:

1. Z causes the independent variable (in this case, trade);
2. Z affects the dependent variable, if at all, only through the independent variable;
3. Z is not itself caused by the dependent variable or by a factor that also affects the dependent variable.

The logic behind IV's is as follows: if variable Z is correlated with both the independent variable of interest and the dependent variable, then, there is no other explanation except for the fact that it is causal. After all, Condition C rules out that there this correlation is explained in another way.

Instrumental variables are not easy to come by, few variables are exogenous in the right way, and they often are quite ingenuous. Here is another example.

Education affects earnings, but earnings and education also have a common cause: ability. So, the correlation between earnings and education may be a reflection of the fact that those who have a high ability at certain (intellectual) tasks are just more likely to make a higher earning later. It turns out that students who are born in January, February, and March do better in school, because they are generally admitted in a class with younger students. However, the month in which you are born is not related to your general innate ability. Thus, being born in fall is an instrumental variable for estimating the effect of education on earnings (this study was conducted by Angrist and Keueger 1991).

When such exogenous variation can be found, a strong argument can be made for causality. However, if the assumptions are not satisfied (Conditions A-C), the argument fails. One of the most focal critics of IV's is Nobel Prize winner Angus Deaton. He argues that most IV's only *appear* to satisfy Conditions A-C, but in fact, do not (Deaton 2010). For example, while being born in fall seems to be a good IV, its effect on income may not only be caused through educational success: relatively older children may also learn other skills better than younger ones, for example, leadership skills and assertiveness. If that is so, being born in fall is not, in fact, a good IV for learning about the effect educational achievement has on income.

8 • CONCLUSION AND SUMMARY

We started this chapter with the problem of causality tension:

1. What we can observe in economic data are correlations.
2. Correlation does not imply causality.
3. Ultimately, we want to know what the causal relationships are in economics.

This seemed troublesome, but we have now seen how this can be solved. First, if we can find exogenous variation, we sometimes *can* infer causality from correlations. Second, we have also seen that correlation does imply causality under normal conditions. However, the direction of causality is not always clear.

Given that it is easy to misinterpret the direction of causality and difficult to find exogenous variation, it is appropriate that econometricians have some restraint if it comes to causality. However, we should not forget that the ultimate goal of econometrics should be to establish causal relations. Therefore, thinking about what your econometric results tell us about causality is of crucial importance.

LEARNING GOALS FOR THIS CHAPTER

After studying this chapter, you should be able to:

1. explain the theoretical problem of causality, the practical problem of causality in economics, and the relationship between the two.
2. explain how Mill's method of difference can address the problem of causality, and the relation this method has to experimentation in science.
3. explain the relationship between correlations and causality, and in particular, the Reichenbach principle.
4. Explain the concept of natural experiments, the concept of exogenous variation, and the way the instrumental variables approach addresses the practical problem of causality in economics.

References

Ali, Karim, Tanweer Azher, Mahin Baqi, Alexandra Binnie, Sergio Borgia, François M. Carrier, Yiorgos Alexandroa Cavayas, Nicolas Chagnon, Matthew P. Cheng, and John Conly. 2022. "Remdesivir for the Treatment of Patients in Hospital with COVID-19 in Canada: A Randomized Controlled Trial." *CMAJ* 194 (7): E242–51.

Angrist, Joshua D., and Alan B. Keueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106 (4): 979–1014.

Arcà, Emanuele, Francesco Principe, and Eddy Van Doorslaer. 2020. "Death by Austerity? The Impact of Cost Containment on Avoidable Mortality in Italy." *Health Economics* 29 (12): 1500–1516.

Becker, Gary S., Michael Grossman, and Kevin M. Murphy. 1990. "An Empirical Analysis of Cigarette Addiction." National Bureau of Economic Research.

Bem, Daryl J. 2011. "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect." *Journal of Personality and Social Psychology* 100 (3): 407.

Besselink, Marc GH, Hjalmar C van Santvoort, Erik Buskens, Marja A Boermeester, Harry van Goor, Harro M Timmerman, Vincent B Nieuwenhuijs, et al. 2008. "Probiotic Prophylaxis in Predicted Severe Acute Pancreatitis: A Randomised, Double-Blind, Placebo-Controlled Trial." *The Lancet* 371 (9613): 651–59. https://doi.org/10.1016/S0140-6736(08)60207-X.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, and Taizan Chan. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (6280): 1433–36.

Card, David, and Alan Krueger. 1994. “Minimum Wages and Employment: A Case Study of the New Jersey and Pennsylvania Fast Food Industries.” *American Economic Review* 84 (4): 772–93.

Deaton, Angus. 2010. “Instruments, Randomization, and Learning about Development.” *Journal of Economic Literature* 48 (2): 424–55.

Dickson, Michael, and Davis Baird. 2011. “Significance Testing.” In *Philosophy of Statistics*, 199–229. Elsevier.

Fisher, Ronald A. 1935. “The Design of Experiments,(1960, New York. Hafner.).”

Greenberg, Edward. 2012. *Introduction to Bayesian Econometrics*. Cambridge University Press.

Herndon, Thomas, Michael Ash, and Robert Pollin. 2014. “Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff.” *Cambridge Journal of Economics* 38 (2): 257–79.

Hoover, Kevin D., and Mark V. Siegler. 2008. “Sound and Fury: McCloskey and Significance Testing in Economics.” *Journal of Economic Methodology* 15 (1): 1–37.

Ioannidis, John, and Chris Doucouliagos. 2013. “What’s to Know About the Credibility of Empirical Economics?” *Journal of Economic Surveys* 27 (5): 997–1004. https://doi.org/10.1111/joes.12032.

Keynes, John Maynard. 1936. “The General Theory of Employment, Interest and Money (London, 1936).” *KeynesThe General Theory of Employment, Interest and Money1936*.

———. 1939. “Professor Tinbergen’s Method.” *The Economic Journal* 49 (195): 558–77.

———. 1940. “On a Method of Statistical Business-Cycle Research. A Comment.” *The Economic Journal*, 154–56.

Lyons, Louis. 2008. “Open Statistical Issues in Particle Physics.” *The Annals of Applied Statistics* 2 (3): 887–915.

Magnus, Jan R., and Mary S. Morgan. 1999. “Methodology and Tacit Knowledge: Two Experiments in Econometrics.”

Mayo, Deborah G. 2010. “An Ad Hoc Save of a Theory of Adhocness? Exchanges with John Worrall.” *Mayo and Spanos* 2010: 155–69.

McCloskey, Deirdre N., and Stephen T. Ziliak. 1996. “The Standard Error of Regressions.” *Journal of Economic Literature* 34 (1): 97–114.

Necker, Sarah. 2014. “Scientific Misbehavior in Economics.” *Research Policy* 43 (10): 1747–59. https://doi.org/10.1016/j.respol.2014.05.002.

Neyman, Jerzy, and Egon Sharpe Pearson. 1933. "IX. On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231 (694–706): 289–337.

Ogrodnik, Irene. 2014. "Couple Finds Photo of Themselves Together 11 Years before They Met - National | Globalnews.Ca." Global News. 2014. https://globalnews.ca/news/1484569/couple-finds-photo-of-themselves-together-11-years-before-they-met/.

Philipse, Herman. 2015. "Probability Arguments in Criminal Law - Illustrated by the Case of Lucia de Berk." *Utrecht Law Review* 11 (1): 19–32. https://doi.org/10.18352/ulr.310.

Reinhart, Carmen M., and Kenneth S. Rogoff. 2010. "Growth in a Time of Debt." *American Economic Review* 100 (2): 573–78.

Reiss, Julian. 2013. *Philosophy of Economics: A Contemporary Introduction*. Routledge.

Ritchie, Stuart. 2020. *Science Fictions: Exposing Fraud, Bias, Negligence and Hype in Science*. Random House.

Sala-I-Martin, Xavier, Gernot Doppelhofer, and Ronald I. Miller. 2004. "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach." *American Economic Review* 94 (4): 813–35.

Sala-i-Martin, Xavier X. 1997. "I Just Ran Two Million Regressions." *The American Economic Review* 87 (2): 178.

Shu, Lisa L., Nina Mazar, Francesca Gino, Dan Ariely, and Max H. Bazerman. 2012. "Signing at the Beginning Makes Ethics Salient and Decreases Dishonest Self-Reports in Comparison to Signing at the End." *Proceedings of the National Academy of Sciences* 109 (38): 15197–200. https://doi.org/10.1073/pnas.1209746109.

Sober, Elliott. 2001. "Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause." *The British Journal for the Philosophy of Science* 52 (2): 331–46.

Stanley, T.d., and Hristos Doucouliagos. 2010. "Picture This: A Simple Graph That Reveals Much Ado About Research." *Journal of Economic Surveys* 24 (1): 170–91. https://doi.org/10.1111/j.1467-6419.2009.00593.x.

Stern, Jacob. 2023. "An Unsettling Hint at How Much Fraud Could Exist in Science." The Atlantic. August 2, 2023. https://www.the-atlantic.com/science/archive/2023/08/gino-ariely-data-fraud-allegations/674891/.

Tinbergen, Jan. 1939a. "Statistical Testing of Business Cycle Theories: Part i: A Method and Its Application to Investment Activity."

———. 1939b. "Statistical Testing of Business Cycle Theories: Part II: Business Cycles in the United States of America, 1919-1932."

———. 1940. "On a Method of Statistical Business-Cycle Research. A Reply." *The Economic Journal*, 141–54.

Tullock, Gordon. 2001. "A Comment on Daniel Klein's" A Plea to Economists Who Favor Liberty"." *Eastern Economic Journal* 27 (2): 203–7.

Wald, Abraham. 1939. "Contributions to the Theory of Statistical Estimation and Testing Hypotheses." *The Annals of Mathematical Statistics* 10 (4): 299–326.

Wall Street Journal. 1996. "Minimum Wage vs. Supply and Demand," April 25, 1996.

Wang, Yeming, Dingyu Zhang, Guanhua Du, Ronghui Du, Jianping Zhao, Yang Jin, Shouzhi Fu, Ling Gao, Zhenshun Cheng, and Qiaofa Lu. 2020. "Remdesivir in Adults with Severe COVID-19: A Randomised, Double-Blind, Placebo-Controlled, Multicentre Trial." *The Lancet*.

Woodbury, Stephen A., and Robert G. Spiegelman. 1987. "Bonuses to Workers and Employers to Reduce Unemployment: Randomized Trials in Illinois." *The American Economic Review*, 513–30.

Wooldridge, Jeffrey M. 2009. *Introductory Econometrics: A Modern Approach*. 5th ed. Nelson Education.

Ziliak, Stephen T., and Deirdre N. McCloskey. 2004. "Size Matters: The Standard Error of Regressions in the American Economic Review." *The Journal of Socio-Economics* 33 (5): 527–46.

———. 2008. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. University of Michigan Press.