Fact or Fluke?

Fact or Fluke?

A Critical Look at Statistical Evidence

Ronald Meester and Klaas Slooten

Amsterdam University Press

Also published in Dutch as: Kan dat geen toeval zijn? Een kritische blik op statistische bewijsvoering, Ronald Meester en Klaas Slooten. Amsterdam University Press, 2022.
DOI 10.5117/9789463725088

Translation: Reinie Erné

Cover design: Gijs Mathijs Klunder
Lay-out: Crius Group, Hulshout

ISBN 978 94 6372 349 7
e-ISBN 978 90 4855 744 8
DOI 10.5117/9789463723497
NUR 916

Table of Contents

Preface

This book is intended for anyone who is in some way interested in statistical evidence: scientific researchers, students, teachers, mathematicians, philosophers, lawyers, managers, and probably many others. It is not an ordinary book about statistics, and it certainly is not a recipe book that tells the reader which test goes with, or which software can be used, for which problem. So what kind of book is it? We will try to explain that now.

In this book we explain how statistical argumentation works. We will see what types of questions are asked, and what logic underlies how we try to answer those questions. Broadly speaking, there are two ways of doing statistical argumentation, both of which have a long history. One way is to consider a research hypothesis proved if the obtained data fit very poorly with the research hypothesis not being true. The other way consists of looking for the best explanation for the data one wants to interpret. These forms of statistics are known as, respectively, the classical (or frequentist) approach and the Bayesian approach. Contrary to what the names suggest, the Bayesian approach has the oldest foundations: it goes back to Thomas Bayes (1702–1761), whereas the frequentist approach was developed mainly from the beginning of the twentieth century onward.

There are many textbooks in statistics that explain how, in the classical approach, to interpret data, formulate hypotheses, design experiments, and decide whether to reject a hypothesis. In this book, we want to show that certain aspects of this classical interpretation of evidence are problematic. As a result, reported claims may be unfounded. These problems with statistical evidence are partly responsible for what is being called the *replication crisis*, the apparent inability to reproduce research results. In other words, it is sometimes impossible to confirm the conclusion derived from a first experiment through a second experiment. What was considered "proved" scientifically may, on closer inspection, not be so.

This is of course a serious problem. The reliability and status of statistical evidence are at stake. This may have far-reaching consequences, not only for science but also for society as a whole, for example for health care or the administration of justice. After all, data are gathered not only to investigate whether a particular hypothesis is correct, but also to draw conclusions or attach consequences from its correctness (or incorrectness). This can happen in all sorts of subject areas, because the data evaluated statistically can relate to anything. A statistical interpretation is needed whenever these data are (partly) determined by chance. This may concern the effectiveness

of a medicine, the relationship between diet and health, opinion polls, or elementary particles colliding in a particle accelerator; in each case, the collected data are determined partly by chance.

In this book, we will explain, hopefully in a readable and understandable way, how it is possible that statistics is fraught with so many problems. We will have to consider many things in doing so. What is statistical evidence anyway? This may seem like a simple question, but we will see that it is much more complex than one might think at first glance. We will therefore try to go back to the essence, namely the way of thinking and the reasoning behind statistics, because they show us what we can and cannot expect from statistics. Using this approach, we will explain that the classical way of testing hypotheses *does not* answer the question whether the data provide evidence for or against a hypothesis and is therefore an inadequate method for drawing conclusions about it.

Next, we will try to show how thinking about the nature of statistical evidence can put us on the right track to reason in a different way and arrive at conclusions that are logically justified.

We are of course not the first to criticize the current statistical practice, and some of the problems identified in this book have already been recognized by colleagues. Let us mention a few. Biostatistician Richard Royall already voiced strong criticism of classical statistics in 1996 [27]. In 2005, epidemiologist John P.A. Ioannidis published an article in which he stated that most published scientific research results are false [18]. Recently, statistician William Briggs, for example, also rejected the use of the classical *p*-values [6]. There are also journals that have banned the use of the classical *p*-values, such as *Basic and Applied Psychology*, *Epidemiology*, and *Political Analysis*; see, for example, [10], [15], and [30]. That little has changed so far despite these—and many similar—publications is remarkable, especially when one considers that solutions have also been proposed, about which we will say (much) more later.

Our contribution differs from the previously mentioned publications in that we do not focus primarily on mathematics but mainly explain how classical statistics works and why it works that way, why and when it sometimes fails, and what one could do instead. We want to show that statistics as we advocate it is very much in line with our *intuition* about what we can expect from statistical research. We explain what we can or may, and also what we cannot or may not, expect from statistical evidence, to what extent objectivity can be found or pursued, where objectivity ends, and whether it is an advantage or disadvantage for a method to be objective. Each time, we explain the reason for our approach to the reader and try

to point out the problems. This book therefore is both philosophical and applied. That makes it quite unique: we have not found any other texts in the literature with this approach, even though we believe that it is the best way to understand the nature of statistical reasoning. We therefore hope that this book will contribute to a more realistic view of statistical evidence.

The book is structured as follows. In Part I, we discuss classical statistics. The first chapter deals with significance testing. We first conclude that it is possible that many reported claims may be false, and we give a chilling example from legal practice of what can go wrong. In Chapter 2, *p*-values are discussed in detail. We explain what a *p*-value is and why it cannot be adequately used to measure the strength of statistical evidence. If, as often happens, *p*-values are interpreted as such, this can lead to problems and errors. Chapter 3 deals with confidence intervals. These are closely related to *p*-values and are subject to similar interpretation problems, which we also discuss extensively.

In Part II, we discuss a more Bayesian approach. We begin, in Chapter 4, by explaining what statistical evidence truly is, and we introduce an instrument to express evidential strength: the likelihood ratio. We conclude that statistical evidence can seldom be absolute. Instead, it is usually relative to an alternative explanation. In Chapter 5, we give several concrete examples of likelihood ratios in action and discuss the critical difference between evidence and belief. Next, in Chapter 6, we consider the extent to which data can be misleading. After all, it is unreasonable to expect every shred of data to lead to the truth about how the data came to be, but one can ask how often and to what degree it can go wrong. Fortunately, there are limits to this. We also show how the likelihood ratio can be used to design experimental setups. This concludes Part II.

Part III is devoted to statistics in practice. In Chapter 7, we present a comprehensive statistical analysis of two concrete situations. In Chapter 8, we explain that there are situations in which *p*-values can be used pragmatically despite the methodological problems associated with them.

Mathematically, the book does not surpass the level of a first course in applied statistics. The most complicated mathematics consists of calculating some integrals, using conditional probabilities, and applying the quadratic formula.

However, the mathematical content is often not the biggest problem in understanding statistics. Statistics is more than mathematics. The difficulty lies mainly in the *meaning* of what we do. So although we try to present the material in the most accessible way possible, some effort is required of the reader. There may be sections that the reader may need to review

several times, and it may be necessary to use pen and paper to go over our calculations. We think the reward is more than worth it.

The reader familiar with mathematical statistics may be surprised to learn that the concepts propagated in this book have long been accepted within mathematical statistics. What we hope to achieve, however, is that a reflection on what statistics can or cannot do will almost automatically lead to a more or less Bayesian approach without losing sight of the valuable aspects of the more classical "frequentist" statistics.

Some sections in this book are a bit more technical and are labeled with an asterisk (*). This means that the content of these sections is a bit more complex and can safely be skipped—though we do of course recommend them to enthusiasts. We hope the reader enjoys this book and that reading it will be worth the effort.

Several people have helped us immensely with all sorts of comments on earlier versions of this manuscript. Our thanks go to (in alphabetical order) Marc Jacobs, Wouter Kager, Boukje Meester, Luit Jan Slooten, Robert van der Toorn, Wessel van Wieringen, and Harry van Zanten.

Ronald Meester and Klaas Slooten

Prologue

In December 1996, young Christopher, the first son of the British solicitor Sally Clark and her husband, was found lifeless in his bed. At the time of his death, Christopher, not yet three months old, was home alone with his mother. His death was not found suspicious and was classified as sudden infant death syndrome (SIDS).

In January 1998, something similar happened to Sally Clark's second son Harry. While alone with his mother, Harry, who was almost two months old, suddenly and inexplicably became unwell and died shortly after. Doubts now arose about the cause of his death as well as Christopher's. Shortly thereafter, Sally Clark was charged with the murder of both of her sons.

An important role in the trial was played by a witness called by the prosecution, Professor of Pediatrics Sir Roy Meadow. According to Meadow, the probability of two consecutive children within a family both dying of SIDS was about 1 in 73 million. This number was questionable: it was obtained by squaring the postulated (and in itself already questionable) probability of an individual SIDS death, 1 in 8,543, because it concerned two children. This assumes that occurrences of SIDS in different children from the same family are independent of one another. This is questionable because (common) genetic aspects may play a role, as can certain aspects of the children's care.

But the fact that the number of 1 in 73 million was questionable is not our main concern; undoubtedly, the probability of two SIDS deaths in two consecutive children is small. A much greater problem is in the following reasoning, which became tempting for the jury: if Sally Clark were innocent, something extremely unusual would have to have occurred, namely an event with probability 1 in 73 million. Sally Clark was found guilty on November 19, 1999 by a 10-2 majority verdict of the jury and sentenced to life imprisonment for two counts of murder. The conviction was upheld at appeal in 2000. Meanwhile, statisticians began to get involved in the case: in 2001, the Royal Statistical Society raised concerns about both the basis for Meadow's number and its application in the case against Sally Clark. In 2003, the conviction was overturned, and Sally Clark was released from prison.

How could this happen? How can something that first seemed to be strong statistical evidence turn out not to be so? Why is Meadow's number not *proof* that Sally Clark killed her children? And how does one actually determine what statistical evidence is and how strong it is? These are all questions we address in this book.

Part I

Classical Statistics

1. Significance Testing

In 2005, the article "Why Most Published Research Findings Are False," written by epidemiologist John Ioannidis, appeared in *PLOS Medicine* [15]. In this article the author argued that it should not be surprising that most statistically proved connections do not truly exist. This is not so much due to errors in the statistical calculations, but rather to a misunderstanding of what conclusions can be drawn from the results of those statistical analyses. The article hit like a bombshell, judging by the many thousands of times it has been cited. For it immediately provides an explanation for what has become known as the "replication crisis," namely that a second group of researchers often fails to obtain the same research result as a first group. This of course raises the question whether that result was correct. In some disciplines, this fate struck so many of the published research results that there was a genuine crisis: somehow, a significant portion of the seemingly proven results seemed (or turned out) to not be true (see, for example, [4] for the results of a survey on this by *Nature*). This is explained precisely by Ioannidis' article. His argumentation provides a simplified representation of reality, yet he manages to expose a problem that is indeed real using elementary arguments. That is why Ionnadis' approach is a good beginning for this book.

1.1 Testing the Null Hypothesis and Statistical Significance

In his article, Ioannidis assumes that conclusions are drawn based on so-called significance testing, often referred to as NHST (null hypothesis significance testing). This type of testing follows a fixed pattern, which, at first glance, seems entirely reasonable and logical. In this chapter, we explain in detail what this type of testing involves, why it is designed as it is, and what disadvantages it has when widely applied to a variety of hypotheses. Simply put, Ioannidis describes what would happen if careless researchers applied this decision scheme on a large scale and made decisions based solely on the resulting statistical analysis. Of course, this is not always how things are in real life: the quality of a statistical analysis and the subsequent decision-making involve much more than what Ioannidis takes into account. But as a global, qualitative analysis, Ioannidis' reasoning is absolutely valuable. The testing protocol he describes is indeed one of the most widely used in statistics. This approach was designed over a century ago