CAMBRIDGE

Cambridge University Press & Assessment
978-1-108-42356-4 — Statistical Hypothesis Testing in Context
Michael P. Fay , Erica H. Brittain
Excerpt
More Information

1

Introduction

1.1 A Cautionary Tale

On March 20, 2020, an online version of Gautret et al. (2020) about the treatment of COVID-19 with hydroxychloroquine was posted. This was the beginning of the pandemic, and we (the authors) were just beginning the start of over a year of working from home. The world was anxious for a treatment for this deadly disease, and the article (to some) looked like a promising treatment. The first results table of the paper (Table 2), showed that by Day 3 post inclusion, $10/20$ (50%) of patients treated with hydroxychloroquine had negative nasopharyngeal tests for the SARS-CoV-2 virus, compared to only $1/16$ (6.3%) of control patients. A chi-squared test reveals a significant difference (with a two-sided p-value[1] of $p = 0.005$), suggesting the results are much more different between arms than would be expected to occur if there was no true treatment effect. Some interpreted that result and similar ones from that paper as implying that it is strong evidence that hydroxychloroquine helps clear the body of the virus that causes COVID-19; however, we and many others could see from reading Gautret et al. (2020) that there were many potential problems with the presented results.

We begin with this cautionary tale, because we do not want the users of this book to use it to compute p-values such as those in Gautret et al. (2020). It would be very easy for a user to think: "I have a two-sample study" (e.g., a new treatment and a standard of care) "and my responses are binary" (e.g., presence of virus three days after treatment) "so I will just look into Chapter 7 to find the appropriate test, and to find how to calculate my p-value." Unfortunately, the user could use the book that way, but if they do that, they are not assured that they have done an appropriate statistical and scientific analysis. There is much more to a statistical analysis than computing a p-value.

Before we get into the concerns about Gautret et al. (2020), we want to emphasize that we are not implying any intention to mislead from the authors. Remember, at this time the world was desperate for a treatment of this new and deadly disease. The authors may have thought that they were doing good science by looking at the data many ways until a pattern with a very reasonable explanation became very clear to them. Scientists and other researchers are all susceptible to deceiving themselves that their explanations of data patterns are supported

[1] We define the p-value formally in Chapter 2. Briefly, the p-value, p, is a statistic used for statistical hypothesis testing with possible values from 0 to 1, with lower values indicating more evidence to reject the null hypothesis, and 0.05 being the traditional boundary (in many disciplines) for deciding between the null hypothesis ($p > 0.05$) or the alternative ($p \leq 0.05$).

by the data, especially when the need is so great. That is why the methods and principles presented in this book are important, as they will help avoid such self-deception.

We review some of the weaknesses in Gautret et al. (2020), many of which are outlined in Rosendaal (2020). First, the study was a nonrandomized clinical trial and there may have been systematic ways in which the two groups differed. In this study, some of the control group were people that declined the treatment, and others in the control group were from a different treatment center. So there may have been critical differences between the two groups besides their treatment that caused the apparent treatment effect. In general, researchers should be very careful when comparing groups of individuals, in those scenarios where individuals (or their doctors) choose for themselves which group they are in. We may even get a difference effect between the two groups that is repeatable if the repeat study allows the groups to be chosen in the same way, but that does not mean that the repeatable difference can be interpreted such that the treatment caused the different outcomes (see Section 3.2 and Chapter 15).

A second problem is that not everyone that started out in the two arms of the study is analyzed in Table 2 of Gautret et al. (2020). Six people in the treatment arm were excluded from those analyses, and importantly, some of those that were excluded had worsening outcomes suggesting that the virus was not controlled in their bodies (one died, and three were transferred to the intensive care unit of the hospital). In other words, the reason that those individuals were missing from the analysis is very likely related to the outcome we are trying to measure. Such missing data can lead to very misleading results (see Section 3.5.2 and Chapter 17).

A third problem is that the prespecified primary endpoint of the study is listed as SARS-CoV-2 virus presence at Day 1, 4, 7, and 14; however, the endpoints listed in Table 2 are Days 3, 4, 5, 6. Similar results to the Day 3 results appear for Day 4 (12/20 vs 4/16, $p = 0.04$), Day 5 (13/20 vs 3/16, $p = 0.006$), and Day 6 (14/20 vs 2/16, $p = 0.001$). It is concerning that Day 4 is the only prespecified day that was ultimately presented, and that is the one with the largest p-value of the presented results. Prespecifying a primary endpoint is a way of avoiding problems of multiple testing due to looking at the data many ways (see Chapter 13).

A fourth problem is that the statistical test used in the table was Pearson's chi-squared test, but the statistical methods states that "statistical differences were evaluated by Pearson's chi-squared or Fisher's exact tests as categorical variables, as appropriate." In our view, Fisher's exact test is more appropriate than Pearson's chi-squared test, because Fisher's exact test is valid even for the small sample size, but Pearson's chi-squared test is an approximation (see Chapter 7). Further, for every case in Table 2, Pearson's chi-squared has a lower p-value than the p-value from Fisher's exact test, so that has the appearance of choosing the test with the lowest p-value (see Example 13.5, page 238).

We will not go through all the concerns with the publication, but all the flaws are important because this Gautret et al. (2020) paper is a critical part of the history of hydroxychloroquine treatment for COVID-19. Gould and Norris (2021) review that history, including the Food and Drug Administration emergency use authorization (EUA) (March 28, 2020) and the EUA retraction (June 15, 2020). By February 2021, published meta-analyses of several randomized clinical trials estimated that hydroxychloroquine treatment for COVID-19 actually leads to a lower rate of negative PCR tests at Day 7 than standard care (rate ratio 0.86 [95% confidence interval: 0.68, 1.09], Analysis 1.4), and more importantly it shows that hydroxychloroquine treatment "does not reduce deaths from COVID-19, and probably does not reduce the number of people needing mechanical ventilation" (Singh et al., 2021, p. 3). This is a cautionary tale because hydroxychloroquine treatment for COVID-19 turned out not to be helpful for the treatment of COVID-19, and resources were used to find out that information that may have been used more efficiently on other potential treatments.

This is just one example of a study that could have been improved by following the principles and methods discussed in this book. But each study may have its own set of challenges that make scientific inferences difficult. For example, how do we estimate the effect of vaccines? We want to get enough scientific information about the vaccine efficacy to know whether it is a viable public health response, without wasting too much time collecting more information than is needed (see, e.g., Section 7.8 and Chapter 18). Before going into those details, we start with the more basic question: what is science?

1.2 Science, Reproducibility, and Statistical Inferences

Science is about developing theories that both explain reality and reproducibly predict outcomes from studies. If a theory does not predict something reliably, and it is not possible to disprove it with data, then it is not a scientific theory. This book is about the most common school of statistical thought in making scientific claims from data, the frequentist school of statistics, whose basic building block is the statistical hypothesis test.[2] This type of hypothesis testing uses data to decide between two sets of probability models (or two hypotheses): a null hypothesis and an alternative hypothesis. The fundamental idea of these methods is that if we repeatedly apply a test using a specific method then we will rarely make the mistake of deciding in favor of the alternative hypothesis when in fact the null hypothesis is true. We set up the null hypothesis as something we are trying to disprove by our study, so that the interesting results from a study occur when we decide in favor of the alternative hypothesis. Typically, such interesting results are called *statistically significant effects* (see Note N1). Restating, the core fundamental idea of frequentist statistical hypothesis testing is about keeping our error rate of concluding interesting results low *for any specific study*.

In the last decade there has been renewed awareness that a substantial proportion of scientific discoveries published in peer reviewed scientific journals have later turned out to not be reproducible. For example, The Open Science Collaboration (2015) conducted replications of 100 experiments in psychological science published in 3 important psychology journals. Of the 97 studies that originally found a significant effect (a p-value with $p \leq 0.05$), only 35 found a significant effect when the same study design was replicated. Given this perceived crisis in reproducibility, one might think that there would be a clamoring for more books and research about statistical hypothesis testing, a method whose fundamental core is about reproducibility. In fact, it seems like the opposite is happening, and there appears to be an increase in voices calling for replacements of frequentist statistical hypothesis testing.

[2] One can test statistical hypotheses using Bayesian statistics, the other main school, but this book is not about that. We compare the two schools' approaches to statistical hypothesis testing in Chapter 21.

The thinking seems to be: if most scientific studies use these hypothesis tests and define significance as p-values less than 0.05, and we have a reproducibility crisis, then we need to find a new way of doing science that does not use $p \leq 0.05$ to define "significant" effects. We disagree, but not totally.

First, we question whether the reproducibility crisis is truly a full-blown crisis, as it may partly be a problem of incorrect interpretation of published scientific discoveries. For example, if of all hypotheses tested: (1) there is a high proportion where the null hypothesis is true; (2) of the small proportion of null hypotheses that are false the average power to detect that falsity is low; and (3) if only significant effects are published, then it is likely that most published research findings will be false (Problem P2). So part of the reproducibility "crisis" may be an unrealistic expectation that almost all published scientific discoveries will be true. Another part of the explanation of the low reproducibility rate is that many researchers may not be properly using p-values. In this book, we emphasize the tools of frequentist statistics that address the latter problem. While the causes of lack of reproducibility are varied and complex, the proper use of these tools should enhance reproducibility. We still find it acceptable to use a threshold such as $p \leq 0.05$ to define significant effects, but we find it helpful to use other frequentist tools as well. For example, we recommend reporting the results of your test with not just the binary decision, significant at the 0.05 level or not, but with the p-value, which allows the reader to determine if the null hypothesis would have been rejected at other significance levels besides 0.05. It is important, whenever feasible, to accompany the p-value with an estimate and a confidence interval. Whenever possible, you should prespecify your primary hypothesis test before beginning your study, but when that has not been done, you should adjust for all the ways you selected which tests to do, and how many of those tests you performed. If you are interested in results which allow an interpretation of causality (e.g., this drug caused that effect), then you need to pay attention to how you design your study.

Why are the frequentist tools used so much, and how can they be used better? Consider the case of the Wilcoxon–Mann–Whitney test. This is a popular test, and this book presents it in a different way to increase its proper applicability to science. Its popularity comes from its simplicity and lack of strong assumptions. Suppose you want to compare two groups of individuals, where one group is given one drug and the other group is given a placebo, and the response is measured on a numeric scale. The Wilcoxon–Mann–Whitney test can be applied by only assuming independence of responses and a null hypothesis probability model. The null model does not need to assume a specific distributional form for the responses (like a normal distribution or a Poisson distribution), it only needs to assume that the drug does not affect the response, so that the distribution of the responses for individuals given the drug is the same as the distribution for those given the placebo. The p-value is calculated by ranking all responses and permuting the treatment labels to find the distribution of the difference in the means of the ranks in the two groups, assuming no treatment difference. The permutation is conceptually simple, and the computer can handle the details. Thus, the Wilcoxon–Mann–Whitney test can be easily applied to many situations with minimal assumptions. Our book is different, as illustrated in how we take the Wilcoxon–Mann–Whitney test inferences to the next level. We present methods for estimating a parameter from the test, discuss the causal interpretation of that parameter, explain how to calculate the confidence interval on that parameter, and discuss the additional assumptions needed to ensure validity of the confidence interval. Further, because the confidence intervals are new (Fay and Malinovsky, 2018), we provide an R package asht that has a function to do all of these calculations. That function has exact versions suitable for small sample sizes, and approximations suitable for larger sample sizes.

This book focuses on applying frequentist statistical methods, which requires both knowledge of the mathematics and of scientific ideas. Mathematical ideas relate to how to define probability models and find valid p-values and confidence intervals given the probability models. Scientific ideas relate to deciding on the appropriate probability model and properly interpreting the result. The interpretation will include determining what kind of causal statements we can make about the result. Loosely speaking, applying frequentist statistical methods requires asking what are we trying to learn about, how can we design a study from which we can learn that, and how do we analyze our data and present the results to properly and fairly convey what we have learned from the study. Proper application requires many tools to meet each specific situation, hence a book is appropriate.

1.3 Using the Book

Here is an outline of the rest of the book. We give the mathematical theory of statistical hypothesis tests in Chapter 2 and go over some of the scientific theory in Chapter 3. Many of the subsequent chapters are based on specific types of data, with increasingly complicated data coming later (Chapters 4, 5, 6, 7, 9, 11, 12, and 16). Chapter 8 addresses robustness in testing, and is presented just before the chapter on two-sample tests (Chapter 9), so that in that chapter we can discuss, for example, how the two-sample t-test does not necessarily require normally distributed data to be approximately valid. Chapter 10 gives a brief introduction to general methods for calculating p-values or confidence intervals (e.g., asympototic likelihood-based methods, bootstrap, permutation tests). Chapter 11 covers k-sample tests including multiple comparison adjustments for subsequent tests made after the k-sample test, while Chapter 13 covers more general multiple comparison adjustments (e.g., Holm's adjustment or false discovery rate). Chapter 14 provides a brief overview of some very useful standard models (e.g., generalized linear models) and how testing may be done using those models. Chapter 15 gives a brief introduction to causality. Chapter 16 discusses censored data, which is common with time-to-event responses. Chapter 17 covers missing data. Chapter 18 addresses group sequential testing. Chapter 19 covers situations where what we want to show is not the typical alternative hypothesis. For example, even though we want to show that the data fit a model well, the goodness-of-fit test defines the null hypothesis as a correctly specified model, so in rejecting the null hypothesis we can only show lack-of-fit by that test. Also, if we want to show two parameters are equal, we must reformulate the problem so that the alternative is that the parameters are close (to within a prespecified margin). We also study the related noninferiority testing, when we want to show one treatment is not inferior to another within a prespecified margin. Chapter 20 deals with power and sample size calculations. Chapter 21 reviews some Bayesian approaches to hypothesis testing.

At the end of most chapters we have a summary of the main points, a section with extra references, a section on R packages, and Notes and Problems. The extra references are not meant to be comprehensive, nor to point to the first one that developed an idea, but to be a useful place to find needed extra details. In some cases (e.g., Chapter 15 on causality), the references given will be for entire books about the subject of the chapter. The lists of R packages are not comprehensive either, especially in the latter part of the book (e.g., Chapter 14 on testing from models), where it would be difficult to list all the relevant packages. In fact, there is an extreme bias in the R package lists, since one of us (Fay) maintains several packages (e.g., asht, bpcp, exact2x2) to calculate some of the methods emphasized in this book. For packages that we do not maintain, we try to recommend only packages that have existed for a while and with a reputation for quality. The Notes section provides extra details, and the Problems may be used in teaching a course, but sometimes also contain valuable extra information as well.

To find a particular topic, first turn to the Concept Index. The **bold** entry for a topic is often the most relevant, such as its definition. If you know an important reference for a topic, then you can look at the end of the Bibliography entry for that reference and it will list the page numbers where that reference was cited. Notation is defined as it is introduced. Because there are so many concepts, most letters (Greek or Roman) will be used to represent multiple different types of values; however, as a general rule, notation for any letter does not change within a chapter. If you have trouble finding the definition of any notation, first look in the previous paragraph, then go to the Notation Index (see page 420).

1.4 Notes

N1 **Statistically significant:** The summary article of the American Statistical Association's special issue on "Moving to a World Beyond '$p < 0.05$'" states that "it is time to stop using the term 'statistically significant' entirely." (Wasserstein et al., 2019, Section 2). The main reason is that "using bright-line rules for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making" (see Problem P1). That summary recommends using a p-value and not dichotomizing it into significant or nonsignificant. We agree that it is generally better not to dichotomize p-values, but we do not recommend never doing that.

1.5 Problems

P1 Suppose two studies explored the same effect, and one study found a significant effect and the other did not, where significant effect was defined for both as rejecting the null hypothesis that the risk ratio equals 1 at the 5% significance level.

(a) Do these studies contradict each other?

(b) What if we additionally told you that the estimates and 95% confidence intervals on the risk ratios were 1.20 (95% CI: 1.09, 1.33) for one study and 1.20 (95% CI: 0.97, 1.48) for the other. Now do the studies contradict each other? Explain. (This is a real example from the medical literature, see Greenland (2017, p. 640) for references and other examples.)

P2 Suppose we have a scientific program where we test many hypotheses (null versus alternative). Let p_A be the proportion of hypotheses where the alternative is true.

Cambridge University Press & Assessment
978-1-108-42356-4 — Statistical Hypothesis Testing in Context
Michael P. Fay , Erica H. Brittain
Excerpt
More Information

For hypotheses where the alternative hypothesis is true, let $1 - \beta$ be the probability that we reject the null hypothesis. (We can interpret $1 - \beta$ as the average power from a random selection of studies with different powers, among studies where the alternative is true.) For hypotheses where the null hypothesis is true, let α be the probability that we mistakenly reject the null hypothesis. In this ideal scenario, if we test many hypotheses and only publish the significant ones (those that reject the null hypothesis), then show that most of those published significance claims will be false if $p_A(1 - \beta) < \alpha(1 - p_A)$ (Ioannidis, 2005).