The Limitations of *P*-values: an Appeal for Alternatives

Research paper Business Analytics

Hans Nuijt (j.nuijt@student.vu.nl)

Supervisor: prof. dr. R.W.J. Meester

Vrije Universiteit, February 28, 2019

Abstract

In this paper we posit several arguments against the use of p-values. We show that they cannot discover cause and cannot be used to falsify any statement. Furthermore we explain some wrong uses of p-values that have appeared in literature. Also, we show in what sense the frequentist nature of p-values can be problematic. We give several directions for a solution, both philosophical and practical, by respectively pleading for a logical view on probability and advocating a focus on predictive ability of models and predictive probabilities.

Keywords— Probability, p-values, predictive probabilities, logical probability, regression

Contents

1 Introduction

For years statistics have made up a large part in science. From the so called 'hard sciences' to the 'soft sciences' it has supported much research. In these statistical methods p-values often play a significant role, sometimes even that large that the conclusion of a scientific paper is, in the end, based on those p-values. Typical concluding statements we then find are like 'Men and women are different with respect to this measurement' or 'X is linked to Y'. However, for these conclusions to be plausible the concept of p-values must be based on firm grounds. This is something that is not self-evident. Many authors have posed objections against p-values, which range from very practical (see a.o. Halsey et al. [22]) to more philosophical (like Briggs [7]). In this debate the view on probability one holds can influence ones opinion about p-values, since p-values originate from a specific view on probability, namely frequentism. So, the question to what extent p-values are useful is not a stand-alone issue, but is part of a broader question, namely what probability is. In this paper we attempt to give a view on p-values, focusing on the question whether p-values add knowledge and to what extent they provide knowledge that we could not obtain without using them. Specifically we give several arguments why we think p-values are often unneeded. Furthermore, flowing from our arguments, we posit that probability models in general should be used to make predictions and hence should have verified predictive ability.

Before we can come to our arguments, we need some building blocks. Firstly, we must give some background about views on probability and the view we think to be the most complete. We realize that this is already a heavy debate between the different schools. Therefore, to focus on what we really need for our arguments, we expound our view on probability and only give a brief explanation of other views and what could be objected against those. Then, in the arguments against p-values, some arguments make explicitly use of the view on probability we hold, whereas other arguments are also relevant for those who hold a different view.

The second building block is a historical overview of p-values. We give some details on the foundation of and the idea behind them. By giving this overview we will get a better understanding of the original purpose of them. In our arguments, subsequently, we are able to distinguish between on the one hand the use of p-values as it was originally meant to use, and bad practices on the other hand.

The rest of this paper is structured as follows. In Section 2 we explain our view on probability. Then, in Section 3 we review the history of p-values, their origin and purpose. Using these building blocks we develop our arguments in Section 4, with a direction for a solution in Section 5. In Section 6 we give a practical example of the use of two p-value alternatives and finally we conclude in Section 7.

2 What probability is

In the past several different views on probability have been developed. Hájek [21] distinguishes six interpretations: classical, logical and subjective probability and the frequency, propensity and best-system interpretations. Of these views we find ourselves most in agreement with the logical interpretation. In the next subsection we will give some context around this interpretation, with a review of some important authors. Furthermore, we explain why we find ourselves most in agreement with this view. Then, in the subsections that follow, we give a brief background on two other important views, the subjective and frequentist. Given the limited space we cannot give a complete overview together with all the arguments that were given in the past. Hence we will at some places refer to reference works for a further discussion.

2.1 A logical view

Logic studies the relation between propositions and thus has to do with arguments. Keynes [27] points out that probability is also concerned with the part of knowledge we obtain by arguments, as opposed to the knowledge we obtain directly. In deductive logic the premises logically entail the conclusion. Hence, the truth of the premises guarantees the truth of the conclusions. Every argument is either deductively valid or invalid. However, sometimes premises only partially entail a conclusion. Logical probability defines probability in these terms of partial entailment: as a

measure of the degree of partial entailment. So, it treats of different degrees in which results obtained are conclusive or inconclusive. This makes probability closely connected to logic, being 'concerned with the degree of belief which it is rational to entertain in given conditions' [27]. Furthermore, under the assumption that truth exists, we can state that propositions are either true or false. Hence, terms like certain and probable express a degree of belief about a proposition and express the relationship between the proposition and a corpus of knowledge. Since a proposition is either true or false, 'an argument is true or false is to speak of the relation and not strictly of the propositions' [4].

There are several authors who contributed to the logical view -sometimes called objective Bayesianism- and who further formalized it. Two of those are Cox [9] and Jaynes [24]. Cox [9] made an attempt to justify the logical interpretation of probability from a set of axioms. He wanted his system to meet the following requirements. It had to be comparable (the plausibility of a proposition must be a real number), it must agree with common sense and it had to be consistent in the sense that several derivations of the plausibility of a proposition have to be equal. Using Boolean algebra and the following two axioms he derived a system of probability.

1. The probability of an inference on given evidence determines the probability of its contradictory on the same evidence.
2. The probability on given evidence that both of two inferences are true is determined by their separate probabilities, one on the given evidence, the other on this evidence with the additional assumption that the first inference is true.

Jaynes [24], at his turn, built on the foundation laid down by Cox and some other authors, including Jeffreys [25], Shannon [39] and Polya [35]. He used the theory developed by Cox to show how probability can be seen as extended logic. He starts by expounding how deductive reasoning differs from what he calls plausible reasoning. A strong (deductive) syllogism is e.g.

$$\left.\begin{array}{l}\text{If A is true then B is true}\\ \text{A is true}\end{array}\right\} \quad \text{Therefore B is true} \tag{1}$$

In contrast to that, plausible reasoning is the reasoning in which we fall back to weaker syllogisms like

$$\left.\begin{array}{l}\text{If A is true then B is true}\\ \text{B is true}\end{array}\right\} \quad \text{Therefore A becomes more plausible} \tag{2}$$

Jaynes gives a nice clarifying example of this. Take A='It will start to rain by 10 AM at latest.' and B='The sky will become cloudy before 10 AM.' Then A does not logically follow from B, however our common sense will possibly (if the clouds are dark enough e.g.) act as if it logically follows. In the remaining part of his book, Jaynes focuses on the fact that mathematical rules of probability theory are more than only calculating frequencies of random variables; they are also consistent rules to conduct inference (i.e. plausible reasoning). Hence, the logical view seeks to include both frequentist calculations (calculating frequencies to express uncertainty about an observable), and Bayesian calculations, where prior information can be taken into account.

At this point it is good to reflect on what these fundamental ideas about probability mean for the notion of probability. First, as probability speaks about the relation between propositions, it is not ontic, meaning that it does not exist physical or factual. It rather expresses the knowledge about a proposition given a corpus of knowledge. Hence, probability is always conditional on the evidence. This deserves some explanation. We started with the initial position that probability is an argument, like logic. Given that logic is the study of the relation between propositions, the same is true for probability. From this it follows that the probability of a proposition cannot be known before we know the given premises that form the evidence. Hence, all probability is conditional. This means, that for a logical probabilist, it is not possible to assign a single probability to the statement Pr(the sun will rise tomorrow). This statement can only be assigned a numerical probability whenever a corpus of evidence is given, notated X. Thus, Pr(the sun will

rise tomorrow | X) for evidence X can be assigned a probability. The conditionality restriction of probability posits the question what is the probability of the following

$$Pr(\text{X is in state } j|\text{X is an n-sided object that must be in one of the n states}), \quad j \in 1, ..., n \quad (3)$$

This probability is $1/n$ and arises from the proportional syllogism. This is an important assumption for logical probability. The proportional syllogism in general takes the form

$$\left.\begin{array}{l}\text{X proportion of F are G}\\ \text{I is an F}\end{array}\right\} \quad \text{Therefore the probability that I is G is X} \quad (4)$$

Different premises can lead to different probabilities. As such, there can be varying degrees of belief about a proposition, depending on the evidence. In the logical view, this makes probability not subjective though. That is, it is not subject to human caprice. What this means is, that when the premises are fixed, the probability is also fixed. Aumann [2] stated it as follows: "If two people have the same priors, and their posteriors for an event A are common knowledge, then these posteriors are equal." People may decide to choose different premises, i.e. different priors, and in this sense probability may seem subjective.

In the Bayesian framework, we can hence write probabilities like

$$Pr(Y|WX) = \frac{Pr(W|YX)Pr(Y|X)}{Pr(W|X)} \quad (5)$$

Notice that we conditioned every probability on premises X. Furthermore, after we have learned W, we can compute Pr(Y|WX). In this framework Pr(W|X) must be deduced from the premises X in order to be regarded as a logical probability. What this means is that for logical probabilists putting priors on W, not deduced from X is regarded as decision and not probability. Inventing a prior that cannot be deduced from the premises is thus adding to the given premises [4].

Furthermore probability as logic implies that it does not speak about causal but logical relationships. In the example above this was clear. It is obvious that the rain at 10 AM does in a causal sense not imply that the sky becomes cloudy before 10 AM. In a logical sense however, rain starting at 10 AM implies that the sky has become cloudy before 10 AM. This distinction between logical and causal implications is crucial. Taking logical implication as if it were causal implication would lead to wrong conclusions as we will discuss later on. One last remark with respect to logical probability is that, in the broadest sense, it does not have to be a number. Take for example the propositions A='Most Dutch people like ice skating.' and B='Jan is a Dutch person.' Then the probability of B given A is not determined by a number as long the word 'most' is not quantified by a number. In reality it can often be very difficult to assign a numerical probability to a proposition given a corpus of evidence. Of course, when dealing with probability in the mathematical sense, we speak about numerical probability. It is good, though, to realize that it is not always possible to quantify a probability by a number.

2.2 Frequentist interpretation

Typical introductory statistical textbooks constitute a different view on probability than we just described. For example, Ross [38] explicitly starts with a frequentist view on probability and he very quickly moves on to Kolmogorovs axioms of probability. So, the philosophical part is almost entirely skipped and only the mathematical part is dealt with. The underlying perspective is, or at least seems to be, that probability *is* limiting relative frequency. Well-known mathematicians and statisticians as Kolmogorov, Fisher and Pearson also held this view.

The frequentist view comes in two variants. The first is finite frequentism, which originates from Venn [42] and states that 'the probability of an attribute A in a finite reference class B is the relative frequency of actual occurrences of A within B.' This definition seems straightforward and intuitive, however, to define what probability *is*, it falls short. For events that are unrepeatable, finite frequentism has troubles to define probabilities. The frequentist needs a sequence of events to be able to specify any probability. That means that, if we regard the proposition 'Thierry Baudet wins the next Dutch elections', we should embed this proposition in a sequence. One could think of the sequence of all right wing politicians, the sequence of all populist politicians, the sequence

of all euro-sceptic European politicians and so forth. Hence, for such single cases, frequentism has troubles to attach probabilities to events. This is one of the main objections as pointed out by Hájek [19; 21].

This problem generalizes to the second variant of frequentism, which is hypothetical frequentism. This variant defines probability as the relative frequency of events in the limit, i.e. an infinite number of trials is needed to define the probability of an event. Since these infinite sequences do not exist in reality, the infinite sequence has to be imagined and probabilities, then, are the relative frequencies if the sequence would be infinite. This idea was made popular by among others Reichenbach [37] and Von Mises [43]. Besides the aforementioned objection, there are a lot of objections to this hypothetical frequentism. Hájek [20] and, following him, Briggs [4] mention some. Since an infinite sequence is to be observed in order to know the probability of a proposition, the main objection is that probabilities can never be known, because infinite sequences cannot be observed. The difficulty lies in the fact that, how large a sequence might be, it is infinitely smaller than an infinite sequence.

Another important problem with hypothetical frequentism is its circularity in the definition. If hypothetical frequentism is true, every finite frequency is an approximation to a probability. Then, in order to know how good this approximation is, the error of the measurements has to be expressed. However, it can only be expressed as a probability, which is exactly the concept which we then are trying to define [15]. These are some of the problems of using frequentism to define what probability is. Literature offers more extensive enumerations of the problems [19; 20].

Although frequentism falls short to define what probability *is*, it is not true that frequencies are of no value. Instead, a large part of statistics has to do with finite frequencies and in many cases finite frequentism is perfectly in line with logical probability. Consider the following example. Given the evidence 'the observed finite relative frequency of A is p', the finite frequentist will come to the conclusion that Pr(a new event is A | data)=p. Logical probability would come to exactly the same conclusion. Hence, frequencies can often constitute evidence to calculate probabilities. As such, frequencies can be very valuable.

2.3 Subjective interpretation

In the twentieth century several mathematicians, like De Finetti [10] and Ramsey [36], independently came to an interpretation of probability that states that probability is a subjective degree of belief. This view is radically different from frequentism and most of its justifications have been made by showing that rational choices in a betting setting require assigning subjective probabilities to random events. When compared to the logical view, the main difference between the two is that the logical view sees probability as an objective degree of belief, whereas the subjective interpretation views it as a personal degree of belief based on the evidence available to you. The subjective view has placed some limitations to this subjective degree of belief, such that the degrees of belief must cohere in the sense that they follow the mathematical laws of probability. This view on probability comes with practical advantages, such as being able to numerically express ones learning from experiences. On the other hand, the question arises how to assess a subjective degree of belief and closely connected to that, we may wonder how to assess the prior distributions that people put on a problem [16]. Although subjectivism from a logical probabilists point of view yields the problem of obtaining probabilities that cannot directly be seen as evidence based on a corpus of knowledge, subjectivism has some similarities with the logical view. The statement that a probability is a degree of belief, and hence is in the mind, is something we can endorse. In practice the main difference between logical probabilists and subjectivists will mainly be in putting slightly different priors on a problem.

3 On the origin and use of p-values

In the 1700s Arbuthnot [1] wanted to show that the yearly births of males exceeding the births of females was not the effect of chance but divine providence. To do so, he calculated what would be the chance, if for every birth it is equally probable that it is a male or a female, that the birth records as obtained in London in those years, would occur. In fact, this can be seen as an early application of statistical significance tests. When we speak about statistical significance tests, we

enter the field of p-values. To get a better understanding of the place of these within statistics, we will put them in historical context.

After the work done by Arbuthnot [1], Laplace also investigated a similar question [29]. However, further formalization and much more interest came in the 1900s, starting with Pearson [34]. This paper was 'on the criterion that a given system of deviations from the probable, in the case of a correlated system of variables, is such that it can be reasonably supposed to have arisen from random sampling' or, put differently, 'whether the sample may be reasonably considered to represent a random system of deviations from the theoretical frequency distribution of the general population [...]'. In this paper p-values were given for the chi-squared distribution. However, the use of significance tests and p-values became only commonplace when Ronald Fisher extensively introduced them in the handbook Statistical Methods for Research Workers [13]. In this handbook the idea of significance tests is described and many examples are given.

Fisher had a clear frequentist idea of probability. His assumption was that

> Even in the simplest cases the values (or sets of values) before us are interpreted as a random sample of a hypothetical infinite population of such values as might have arisen in the same circumstances. The distribution of this population will be capable of some kind of mathematical specification, involving a certain number, usually few, of parameters, or "constants" entering into the mathematical formula. These parameters are the characters of the population.

According to Fisher this infinite population is even fundamental to all statistical work. Subsequently, using a sample one could get an idea of the infinite hypothetical population from which the sample is drawn. Furthermore, if the expectation of the hypothetical population inferred from a second sample is different from that from the first sample, it is said that the second sample is drawn from a different population. This conclusion is drawn from the so called significance tests. The steps for these significance or hypothesis tests are as follows. First a probability model for the observations is formed. Then, a null hypothesis is posited. The third step is to calculate a statistic, which is 'a value calculated from an observed sample with a view to characterising the population from which it is drawn' [13]. When this statistic is calculated, the agreement between the observations and the hypothesis is tested. This is done using the p-value, which expresses the probability of seeing a larger test statistic, given the null hypothesis, given the probability model and given the data, assuming that we could repeat the experiment an infinite number of times. Hence, the lower the p-value, the higher the significance. So, put simply, the p-value is a probability that a function of the data will exceed some value when a specified probability model holds and it quantifies the statistical significance of evidence. Fisher proposed to use a p-value of 0.05 as a cutoff-value: whenever the p-value is lower than this value, the null hypothesis is rejected. This cutoff-value or alpha level is the probability of a type I error: the probability of rejecting the null hypothesis when it is in fact true. The type II error is the failure to reject a false null hypothesis. It should be stated that the p-value is *not* the probability that the null hypothesis is true given the data. That would be a Bayesian way of thinking. The p-value is also not the probability of the data given the null hypothesis.

How did Fisher come to this particular idea of 'rejecting null hypotheses' and 'failing to reject the null hypothesis'? Although the idea of a null hypothesis did not appear in Fishers Statistical Methods for Research workers, it did appear in his later work The Design of Experiments [14]. The experimental setup in which a null hypothesis is to be rejected, has some similarities with the idea of falsifiability, which became popular in the twentieth century. This concept evolved from the criticism on inductive reasoning. Several philosophers developed a very skeptical attitude towards inductive reasoning. David Hume even went so far that he stated that inductive arguments are unreasonable. This skepticism towards induction led Karl Popper to the definition that statements and theories must be falsifiable in order to be scientific. These ideas might have influenced Fisher, although it must noticed that Fisher already published about significance tests in 1925 whereas Poppers important work Logik der Forschung appeared in 1934. However, one could argue that Fisher certainly was influenced by this falsifiability concept as he speaks in his Design of Experiments (1935) explicitly about falsification of the null hypothesis whenever a significant value of the test statistic is obtained.

4 Arguments against the use of *p*-values

4.1 *P*-values cannot discover cause

Imagine a medical experiment involving two equal sized groups of people, suffering all from the same disease. The first group, consisting of hundred people, gets a medicine during a predetermined period, whereas the other group receives a placebo. When the period of the experiment has ended, thirty out of the hundred are healed in the first group and seven out of the hundred are healed in the second group. What is the probability that more people in the first group are healed than in the second group? Given the information that thirty people in the first group were healed and seven in the second group, the only right answer is obviously one. We are certain that more people in the first group are healed than in the second group. Secondly, what is the cause of the fact that more people in the first group were healed than in the second group? This is a far more difficult question. Intuitively it feels logical to say that the medicine was the cause that those people were healed. This could very well be, but can we *discover* this causality by probability models and their accompanying p-values? We argue that this is not possible.

The first observation we could do based on our hypothetical experiment above is that there is at least one cause different from the medicine that caused people to heal. This is clear, since there are people in the group who did not get the medicine who were healed though. Based on this experiment we do not know what caused these people to heal from their disease. The second observation is as follows. If the medicine is a cause for the healing of the disease, there is a blocking mechanism which prevents the medicine from causing the disease in all the people to heal. From this it follows that the relationship between the medicine and the disease at least involves other factors, that are, based on the experiment and the data, unknown.

One could obtain a p-value to test in the above mentioned experiment whether the difference in proportions is significant. The p-value for this is significant, but does that give us any information about the causality? Recall the definition of the p-value to see that it only states what the probability is, conditional on the probability model etc., that a larger test statistic would be obtained when the experiment could be repeated an infinite number of times. This does clearly say nothing about the causal nature of any relationship. The statistician would, based on the p-value, state that there is a significant difference between the medicine and placebo group with respect to the number of healings of the disease. In this experiment the only difference measured between the groups is whether or not they got the medicine. There are many measures that can be made on these people, so that it may very well be that the two groups also differ with respect to any of those measurements. As a result, the conclusion based on the significance test would be, then, that the two groups significantly differ with respect to that measurement. If the p-value that was obtained in the experiment would say anything about causality, so would the p-value that expresses that the difference is significant with respect to *any* measurement [5; 7].

Statisticians often say that 'correlation does not imply causation' - which we wholeheartedly endorse - but in practice this is often forgotten. If a probability model and a resulting p-value does, by itself, not say anything about a causal relationship, then they should not be used to *discover* cause. Regularly results are published in which 'variable X is shown to be linked to Y'. These results are often interpreted as if X causes Y, but this cannot be said to be generally true. When we again think of our hypothetical medical experiment, we may ask how we can know whether the medicine is a 'good' (whatever that may be) medicine. In the medical world a medicine is often developed based on knowledge about the interaction between elements, cells and so forth. This knowledge stems from an understanding *how* things in the body work, which is basically knowledge obtained by physics. Without going deep into how we can know any cause, this gives a sense of how this knowledge may be obtained. The question that arises is what the added value of probability models and statistics is, then. We are sure that these have added value. Suppose that the medicine is developed based on knowledge of the physics of the body and the disease. The medical experiment then can still be used to see whether the causal relationship is approved by the experiment. When not many people in the medicine group are healed, the researchers may wonder whether there is a blocking mechanism that they did not known till that moment.

We are aware that by carefully designing an experiment, two groups can be constituted that are on average not different with respect to many measurements. This is the idea of randomized trials, which, when performed with large groups of people, indeed ensure that with respect to a lot of measurements the groups on average do not differ. Hence, this section does not call for an abolition of statistics, but rather a reappraisal for it, by recognizing that it by itself cannot discover cause, but that it can help researchers to find directions in which they have to search in order to comprehend causal mechanisms.

That causal mechanisms are sometimes actually assumed whenever statistically significant results are obtained, can be seen in several examples. The first example comes from the field of epidemiology. As many epidemiological studies showed that women taking hormone replacement theory (HRT) had lower-than-average prevalence of coronary heart disease (CHD), researchers concluded that HRT protects against CHD. So, a conclusion of causal nature was made. However, follow-up randomized clinical trial studies showed an increased risk of CHD. Further studies showed that women undertaking HRT most of time are of higher socio-economic groups, of which it is known that in those groups diets are better than in other groups [30]. This shows that, although probability models cannot discover cause, they can be helpful in order to comprehend causal mechanisms. Another example is the following. In a study investigating the correlations between unexpected outcomes of football games and U.S. juvenile court decisions, the authors described their findings as follows:

> Employing the universe of juvenile court decisions in a U.S. state between 1996 and 2012, we analyze the effects of emotional shocks associated with unexpected outcomes of football games played by a prominent college team in the state. We investigate the behavior of judges, the conduct of whom should, by law, be free of personal biases and emotions. We find that unexpected losses increase disposition (sentence) lengths assigned by judges during the week following the game. Unexpected wins, or losses that were expected to be close contests ex-ante, have no impact [12].

The terminology in this abstract has quite a causal nature as they state 'that unexpected losses increase disposition lengths', whereas the evidence for these causal claims are based on significant p-values in a regression model. Unfortunately, when the results of scientific research are translated into newspaper articles, correlation dealt with as causation is even more abounding. Many more examples can be found in which p-value based results are interpreted being causal. These signs call for a return to a better understanding of the nature of statistics.

4.2 *P*-values cannot be used to falsify anything

We argued that p-values cannot discover cause, but we also warned that this does not mean that probability models are not useful. To what extent they are useful and whether p-values should be used, is still an open question. In this subsection we argue that p-values at least cannot do what some think it can, namely falsifying statements. An example of someone who *did* state that p-values falsify a statement is Fisher, who explicitly mentioned that the null hypothesis is falsified whenever a significant value of the test statistic is obtained: 'If [...] a significant value of z has been obtained the null hypothesis has been falsified, and may therefore be set aside.'[14]

Let us first recall, when do we actually falsify a statement? Consider a model M and a variable X. Suppose that M states $P(X > 0) = 0$. When we observe an $X > 0$, this contradicts the model M and hence the model is falsified. The observation that was inconsistent falsified the model. When we however would have a model M_1 stating that $P(X > 0) = \epsilon$ for a certain ϵ small, the situation would be different. If we then observe an $X > 0$, this does not contradict the model. Given the model M_1 it is improbable that an $X > 0$ is observed, but it is not impossible. Therefore, observing $X > 0$ is not inconsistent with the model and hence the model is not falsified. So, a probability model that makes a probability statement between 0 and 1 - the open interval without the boundary values - cannot be falsified as no observation is strictly inconsistent with the model.

Now let us turn to p-values. Under the null hypothesis the p-value has a uniform distribution. Hence every value is equally likely to be observed. Specifically, we can write this like: model M implies that the p-value in $(0, 1)$. Any p-value that we observe then, is not strictly inconsistent

with M. This means that it is impossible to falsify any statement with any p-value. Whenever a statistician states that he has falsified the null hypothesis when a low p-value is attained, he either makes a mistake or he has a very weak definition of falsification - practical or nearly falsification. This definition is however not useful as falsification in the Popperian sense and hence it is *not* falsification.

Besides that it is not possible to use p-values to falsify a statement, it is not consistent with falsification as meant by Popper. The design of a significance test involves a null hypothesis and an alternative hypothesis. When we reject the null hypothesis, we accept its counterpart, the alternative hypothesis. This is true, since the null and alternative hypothesis are mutually exclusive. Accepting a theory or statement is not in line with the Popperian idea of falsifiability, in which you only reject theories and not verify them. The attempt to use p-values to falsify any statement, is hence a violation of the idea that statements cannot be verified. P-values therefore should not be used to falsify any statement.

4.3 The nature of p-values is frequentistic

In the previous two arguments we explained for which purposes p-values should not be used. Considering that p-values cannot discover cause and cannot be used to falsify any statement is already very important for any statistician. There is however more to say against the use of p-values, at a more fundamental level. P-values are namely creatures with a clear frequentistic nature. This can be problematic in several ways.

As we already saw in the historical context of p-values, their nature and explanation is fully frequentistic. The p-value is the probability that a test statistic obtains at least a certain value, *assuming* that the experiment can be repeated an infinite number of times. This has two implications. First, when using p-values the statistician has to agree upon the validity of this frequentist interpretation of probability in his experiment. Besides, p-values should have a philosophical justification in frequentist probability. With respect to the first implication, it seems that p-values are often used without realizing that this means that one silently agrees with a frequentist mindset. Now, we do not say that this by itself is an argument against the use of p-values. It only means that before using a statistical procedure like p-values, one should have to agree upon the interpretation of probability. In reality this often seems to be lacking. Hence, the frequentistic nature of p-values is an appeal to everyone who considers using it, to think about the nature of probability.

Even if one holds a frequentist view on probability, several objections against the philosophical justification can be brought in. In the later editions of Fishers Statistical Methods for Research Workers the following argument for p-values occurs:

> Belief in the [null] hypothesis as an accurate representation of the population sampled is confronted by the logical disjunction: Either the hypothesis is untrue, or the value of χ^2 [the test statistic] has attained by chance an exceptionally high value.

Briggs [7] points out that this argument does not contain a logical disjunction. We agree with him in that respect but we do not completely follow him in the rest of his argument. The disjunction in the argument is not logical as the first part relates to the null hypothesis and the second part relates to the value of the test statistic. A logical disjunction would be of the type 'either it is raining or it is not raining.' Briggs, then, proposes a fix of the disjunction to make it a logical disjunction and he tries to show with this that a small p-value has no bearing on any hypothesis unrelated to the p-value itself. His fixed version of the argument is 'either the hypothesis is untrue and the test statistic has attained a high value or the hypothesis is true and the test statistic has attained a high value.' This, then, boils down to 'either the hypothesis is untrue or it true and the test statistic has attained a high value'. The tautology adds no information and therefore we are left with 'the test statistic has attained a high value'. Briggs' argument however ignores how probable or improbable it is that the test statistic attains a high value when the alternative hypothesis -which is mutually exclusive to the null hypothesis- is true. Therefore, the disjunction should be fixed in a way like 'Either the alternative hypothesis is true and the test statistic has attained a likely high value, or the null hypothesis is true and the test statistic has attained a unlikely high value.' The original disjunction would be logically valid if it stated 'Either the null is

false or something impossible has occurred. It is certainly not true that something impossible has occurred. Therefore it is certainly true that the null is false.' In the real argument it goes wrong as 'improbable' is used as an approximation of 'impossible'.

Another objection against the justification is closely related to the inability of p-values to falsify a statement. Frequentist theory states that under the null hypothesis each value is equally likely to be observed, i.e. the p-value is uniformly distributed over the interval (0,1). This means that every p-value in (0,1) supports the null hypothesis. Logically speaking, to state that the null hypothesis is false whenever a low p-value is observed, is hence a non-sequitur. As significance tests involve both a null hypothesis and an alternative hypothesis with their respective p-value distribution, it becomes crucial to consider the power of the test. If the power of the test, that is, the probability that a test rejects the null hypothesis when the null hypothesis is false, is low, the test becomes less reliable. Halsey et al. [22] state it in clear terms:

> Many scientists who are not statisticians do not realize that the power of a test is equally relevant when considering statistically significant results, that is, when the null hypothesis appears to be untenable. This is because the statistical power of the test dramatically affects our capacity to interpret the p-value and thus the test result.

A little further they expound this, saying:

> If statistical power is limited, regardless of whether the p-value returned from a statistical test is low or high, a repeat of the same experiment will likely result in a substantially different p-value and thus suggest a very different level of evidence against the null hypothesis.

So, this is again a reason to be very careful in using p-values. The authors of the aforementioned article even show that the power of the test must be very high, higher than in most experiments, to give results that are more than only tentative.

4.4 Without predictive ability, models and p-values are not verified

In some researches there seems to be a tendency to see p-values as the endpoint and final result of the research. Whenever a link between certain variables is found, i.e. a low or 'significant' p-value is found, the result is regarded as publishable. Fortunately, there are already calls from the statistical community that criticize this state of affairs (see among others [44]). Our point is that p-values sometimes too easily are accepted as the endpoint of a research and that in contrast to this, researchers should seek to find additional evidence in the form of predictive ability of the proposed models. Predictive ability is important as a main goal of models is often to make actual predictions.

When p-values are regarded as the endpoint of a research, often a few mistakes are made. First, p-values can easily be seen as *the* evidence that the null hypothesis is false, which is, as we discussed already, not right. Furthermore, such a p-value based conclusion can convey the impression that the model is good, in the sense that a link between variables is said to be found. However, after establishing a model, the model should be verified in order to judge the veracity of the links that are found. There are two steps that are crucial in the verification of a model. First, additional evidence besides p-values is important, as Trafimow et al. [41] among others points out. For example, the use of likelihood ratios has been advocated (see e.g. Mogie [33]). Secondly, once models have been established, they should be verified using data that has been unseen up till that moment. Testing on outside data indicates to what extent the model generalizes and hence it shows what the predictive ability is of the model. Since the goal of many models is to make predictions, the model should be good at making predictions, and how could we better know this than by verifying the predictive ability? Fortunately, this approach in which models must be tested on outside data has become the standard in the machine learning community. There is only one caveat. When models are built and subsequently tested on outside data, the model may never be adapted based on this outside data. When one would adapt the model again, it should again be validated and thus the predictive ability must still be proven. In other fields of work however, like sociology and psychology, such approaches are less common. This could result in conclusions that are too certain.

4.5 *P*-values are often used wrongly

It is clear that p-values have a valid interpretation, if one sticks to frequentism. We mentioned some arguments against frequentism, which must make people aware that adopting the frequentist view is a bit controversial. Still there are many statisticians who can be regarded as frequentist. So, assuming that this is not an unreasonable philosophical position, p-values can be used, be it in the right way. Unfortunately, here it goes wrong too often. The use and interpretation of p-values is often flawed. This might indicate that the interpretation of them is more difficult than often thought and that we must search for alternatives that are conceptually easier. The misuse of p-values has attained a lot of interest in the past years [18; 28; 44]. We here mention some of the most important misuses.

The wrong use of p-values starts with a wrong conception about p-values. The definition of p-values is often not remembered, which results in researchers applying a technique of which they do not exactly know what it means. Although this is a questionable practice, this is not the main problem. Kim and Bang [28] mention three common misuses. The first they mention is the misconception that a large p-value automatically means that there is 'no difference' (in the case of comparing the means of two groups). To better understand that this is a misconception, recall that the p-value is a function of the sample size, under the alternative hypothesis. Hence, p-values are likely to be smaller when the sample size increases. The second misuse is closely connected to the first. A smaller p-value does not necessarily imply a higher 'significance'. Thus, p-values from experiments with different sample sizes could not be compared directly. This even raises the philosophical question whether 'a large size of the difference of parameters from a small sample' or 'a small size of the difference in parameters from a large sample' provides more evidence. Kim and Bang [28], following many statisticians, call these differences in parameters 'effect sizes', but as this may wrongly create the impression that we are speaking about causal effects, we hesitate to agree with this term. The aforementioned question especially makes clear that sole reliance on p-values is not a good idea.

Observed	**Event**	**No event**	**Sum**
Group A	35	25	60
Group B	25	35	60
Sum	60	60	120

Observed	**Event**	**No event**	**Sum**
Group A	70	50	120
Group B	50	70	120
Sum	120	120	240

Table 1: Two examples where the event rates are similar, but p-values are different due to different sample sizes. In the left table the p-value of the Pearson χ^2 test is 0.0679 and in the right table it is 0.0098. This shows that p-values of different experiments cannot be compared directly.

The third misuse with respect to p-values is how multiple testing is handled. When multiple tests are performed, one has to consider that the probability of obtaining false positives increases. Multiple testing can occur in different forms. It is clear that it happens in a study with multiple outcomes or groups. However, it is also used when for example different categorizations of a variable are considered. Solutions like the Bonferroni correction have been proposed to overcome the problems with multiple testing. Unfortunately, as Briggs [7] points out, these adjustments are not unique. Besides the Bonferroni correction there are also other corrections for multiple testing, each with its own advantages. This gamut of corrections makes the use of p-values prone to inconsistent use and this at its turn could lead to flawed results.

That multiple testing is sometimes indeed handled wrongly, can be seen in the following example. In a large study Keetley et al. [26] investigated the performance of a group volunteers on some neuropsychological tests during real or sham exposure to a digital mobile phone set to maximum permissible radiofrequency power output. The study provided statistical evidence that there was a cognitive difference in performance. This claim was based on the result of 18 hypothesis tests, of which seven gave a p-value lower than the alpha level of 0.05. However, Lewis [31] points out that if a standard Bonferroni correction was applied, the alpha level would decrease by a factor 18 and as a result of this, none of the tests reached the normally accepted level of significance.

There is a fourth important misuse with respect to p-values. Briggs [7] reports a serious case that appeared in the Wall Street Journal [45]. Boston Scientific (BS) introduced a new stent, the

Taxus Liberte. Within their experiments they used the Wald test statistic, which gave the desired 'significant' results. However, a competitor claimed that the Wald statistic was not the right statistic to use. To further investigate this, they computed the p-values of several test statistics and concluded that those did not indicate 'significance'. The problem revealed in this story is that test statistics are often not unique. For example, in the analysis of contingency tables, there are several test statistics that could be used, like Fishers exact test and Barnards exact test. With several test statistics being available, the risk increases that researchers search for the statistic that gives the desired result. This could result in conclusions that are far too certain and hence yield over-certainty.

5 A direction for a solution

In the previous section we discussed a few problems with p-values. In fact, the list of problems we mentioned is far from complete, although we think that some main problems are covered. Authors who covered even more problems are for example Berger and Sellke [3]; Gigerenzer [17], and Harrell [23]. We believe that these problems are part of the replicability crisis, and that the misuse of p-values can cause over-certainty in the conclusions. Hence, these vital problems ask for a solution. As the problems are deeply ingrained, it might be difficult to progress towards a solution. In this section we raise some considerations which might be helpful in the strive towards a better use of probability and statistics.

5.1 A logical view on probability helps

A part of the solution is, we believe, to be found on the philosophical side. We do not pretend to give a definite answer in the philosophical probability debate, but we show some aspects in which a logical view can help statisticians. The logical interpretation of probability helps to acknowledge that probability models are in essence not causal. Probability models quantify the uncertainty that we have about an observable. The logical view explicitly states that probability tells something about a degree of belief about a proposition with respect to a corpus of knowledge. When we relate this to the use of p-values to select which variables to include in a model, we can conclude the following. P-values are, among others, used for universal decisions in variable selection. However, this does not state that the variables that are included in the final model are the only right ones to include. This is true because in the logical view all probability is conditional and hence every set of variables gives a different model and different probabilities [6]. In fact, this is a vital point. It means that, given a dataset, there is not *one* probability model that is true. Instead, every probability model gives true probabilities, each with different premises. Furthermore, given that probabilities say something about epistemic knowledge, it becomes clear that we are usually not falsifying anything with probability models. Adopting the logical view (but also adopting subjective Bayesianism) and hence acknowledging the serious limitations to the frequentist definition of probability, will also pull one away from the clear frequentist methodologies, like p-values. This opens the door for other methods, that in a more explicit way quantify the strength of evidence, as the aforementioned likelihood ratios. In fact, to overcome the difficulties which accompany p-values, we believe that a shift from frequentist statistics to Bayesianism, in either its subjective or objective form, could be very helpful. At least, more attention should be given to it, also in student text books. There are actually introductory textbooks in probability which put more emphasis on Bayesianism, like Tijms [40], which, we think, is very positive.

5.2 Using plain probabilities

We already saw that the interpretation of p-values is somewhat difficult. The definition of p-values is often not remembered and p-values are sometimes used wrongly. Furthermore, hypothesis tests pretend that there is a binary outcome based on the data: either the null hypothesis is rejected or it is not. This does however not fully reflect the uncertainty that is in the data. When instead plain predictive probabilities are used, this problem can be overcome. This method as substitute for classical hypothesis testing appeared in literature very recently [6; 8]. To explain what we mean with predictive probabilities, let us consider a linear regression. Traditionally an important part of linear regression is the variable selection: which variables are good to include in the model? To define what 'good' is, p-values are usually employed. Whenever the p-value is below the

threshold of 0.05 (or any other threshold), the variable is considered significant and is included in the model. Hence, this definition of 'good' bases all decisions on the p-values. When we use predictive probabilities, we start with considering the relevance of variables. We say that for a model M, a variable X_n is relevant if for any observable $Y = y$ and variables $X_1, ..., X_n$ it is true that

$$Pr(Y = y|X_1, ..., X_{n-1}, M) = d \tag{6}$$
$$Pr(Y = y|X_1, ..., X_n, M) = d + \epsilon. \tag{7}$$

In other words, if the conditional probability of y changes when variable X_n is added to the model, the variable is considered relevant. The next step is that we want to know how the relevant variables are correlated (in the plain English sense) with the Y. That is, we want to know for different values y

$$Pr(Y = y|X, M). \tag{8}$$

If these probabilities are known, there is not one answer to the question if a variable should be included in the model. It depends on the goal of the decision maker if he then wants to include the variable. This is because we view probability as logical and conditional, meaning that there is not one true probability for y, but only conditional probabilities, where each probability expresses the uncertainty about the values for Y given the premises. When one changes the premises, the probability changes.

The frequentist toolbox does not provide the tools to directly compute these probabilities. What we need for these, is an expression for the uncertainty that is in the model. Bayesian regression provides tools for this, by accompanying the model with parameter distributions. Using this parameter distributions, we can simulate the model for several values y to obtain the predictive probabilities.

5.3 Models should have predictive ability

We can build a model and give conclusions about the variables that are included in the model, but in the end it is important whether the model has predictive ability. Therefore, instead of only reporting to what extent we were able to fit the data with a model, we should focus on the ability to predict new, unseen data well. With this verification we still do not know exact causal relationships, but it gives an idea of the usefulness of a model. Fortunately, such practices have become standard practice in the machine learning world (which is in essence building statistical models), where models are built using training data and are tested on completely unseen test data. Whenever models are not verified in this way, it should be made clear that the models are not verified.

The predictive ability of a model can also be made central in the process of building a model. In the previous subsection we already mentioned the use of the predictive posterior distribution in the setup of a Bayesian regression model. This predictive distribution can be used to decide upon the variables one wants to include in the model. Alternative approaches have been proposed, which offer an alternative to p-values in regression settings. Lu and Ishwaran [32] propose a good alternative, whose origins can be found in machine learning. They use leave-one-out bootstrap to obtain out-of-bag (OOB) prediction errors. Repeating the bootstrap procedure and averaging of the errors yields the OOB error, which is a cross-validated estimate of the accuracy of the model. The next step they propose is to use OOB errors to obtain a variable importance (VIMP) index. This is done as follows. In every step of the bootstrap procedure a regression model is fitted and the OOB prediction error is calculated. Then, still in the same step of the bootstrap, the coefficient relating to one variable is set to zero and the OOB prediction is again calculated. The first prediction error is subtracted from the latter to get a prediction error difference. After all the bootstrap rounds the prediction error differences are averaged and this yields the VIMP. Variables with a VIMP around zero or negative could be considered to be dropped from the model, as they are not important for the predictive ability of the model. VIMP measures how much a variable contributes to the prediction precision of a model. This measure looks purely at the predictive ability of the model and is very easy to interpret and is therefore a worthy alternative for p-values in the regression setting.

6 Practical example

In this paper we gave several arguments why p-values are less valuable than many think. We argued for a proper understanding of p-values and their limitations. The main limitations we discussed are the inability of p-values to discover cause and to falsify any statement. Furthermore we stated that the frequentistic nature of p-values can be problematic. We also pointed out that the use of p-value has resulted in bad practices in which models are not verified using new data and that p-values are too often wrongly used. A solution for these problems is not very simple, but we mentioned three considerations which might be helpful in the search for a solution. First, we argued for a logical view on probability. Besides we stated that predictive probabilities could be a good alternative to p-values and lastly that models should always be verified using new data. These considerations surely will not solve all problems. For example, good practices cannot prevent people from misusing them. However, the problems and solutions that we mentioned can still be translated in some practical applications. In this section we show an example of how predictive probabilities and VIMP can help to analyze the results of a linear regression model.

The data set we use for our example, is the Auto MPG data set from the UCI Machine Learning Repository [11]. The data set contains of 392 cars for which several characteristics, like the number of cylinders, horse power and acceleration time, are given. The goal is to model the miles per gallon for each car, based on these characteristics. As a reminder, this model does not necessarily have to capture (partial) causal mechanisms. The model is built to be predictive and to express the uncertainty that we have about the observable miles per gallon.

In a traditional frequentist, p-value based regression the model is often built using backward elimination: first a model is built with all variables included and then the 'non-significant' variables are removed. In our example, building the regression model with all variables included, yields the following result.

Listing 1: R output of regression

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.218435   4.644294  -3.707  0.00024 ***
cylinders    -0.493376   0.323282  -1.526  0.12780
displacement  0.019896   0.007515   2.647  0.00844 **
horsepower   -0.016951   0.013787  -1.230  0.21963
weight       -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration  0.080576   0.098845   0.815  0.41548
modelyear     0.750773   0.050973  14.729  < 2e-16 ***
origin        1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

In the regression, the method of backward elimination would eliminate the variables cylinders, horsepower and acceleration. However, this mechanism, which has automated a part of the modeling process, can lead to serious misconceptions about the model. When the conclusion would be that the aforementioned variables have 'no effect' and that the variables that were significant do have an effect, two mistakes are made. First, the word choice is poor, as it suggests that the relationships that were found are causal. Secondly, in this particular model much of the variance can already be captured using the other variables, this does however not mean that the variables that had a p-value higher than the alpha level should necessarily be excluded from the model. The p-values in the former regression are highly influenced by the multicollinearity that exists in the data. As Lu and Ishwaran [32] point out, p-values are thus nonrobust. To overcome this, we can approach the problem in several ways. Above we mentioned some solutions, two of which are the use of predictive probabilities and the use of variable importance (VIMP) to assess whether to

include variables in the model based on the predictive ability. We here show how these directions would help in the analysis of the data described above.

6.1 Predictive probabilities

To obtain predictive probabilities, we would have to adopt a Bayesian framework. This means that we have to put a prior on the parameters. When we use for example the package MCMCPack in R, by default a conjugate prior is put on the parameters. This prior hence, is an explicit premise, which is added to the original problem. When we subsequently perform the regression we can use the posterior parameter distributions to obtain a predictive distribution via simulation. With these predictive distributions the relevance of variables can be shown. A variable is relevant when the predictive probabilities change when the variable is added to the model. The next step is then to find a way to express *how* relevant the variables are. Briggs et al. [8] propose to build 'predictive ANOVA' tables. In these type of tables the central estimate for the condition noted is shown, with all other measurements fixed at their median or base levels. Next to that, these tables give the approximated probability that an instance from the data set holding these values for the variables would have a higher y-variable -in our case miles per gallon- than a base level instance. The tables can be modified to the specific questions that are of interest for decision makers. In Table 2 such a predictive ANOVA is shown for our regression problem. This way of regression modelling can give valuable insights, but it requires much and hard work. It does explicitly not give one answer to the question which variables should be included in a model, as this is assumed to be dependent on the goal of the decision maker. The advantage of the predictive approach is that it gives probabilities, which are easy to understand and can be used very well for decisions.

Variable	**Level**	**Central MPG**	**Pr(MPG>base level)**
cylinders	3	24.3	0.56
	4	23.8	0.5
	5	23.3	0.44
	6	22.8	0.39
	8	21.9	0.29
Displacement	105	22.9	0.39
	151	23.8	0.5
	276	26.3	0.76
horsepower	75	24.1	0.54
	94	23.8	0.5
	126	23.3	0.43
weight	2225	27.6	0.87
	2804	23.8	0.5
	3615	18.6	0.06
acceleration	14	23.7	0.48
	16	23.8	0.5
	17	23.9	0.51
modelyear	73	21.6	0.25
	76	23.8	0.5
	79	26.1	0.75
origin	1	23.8	0.5
	2	25.3	0.66
	3	26.7	0.8

Table 2: Table with predictive probabilities. The column Central MPG shows the estimate of the miles per gallon when all variables are kept at their base levels (median) and the variable in the left column takes the value given in the column Level. For numeric variables the first, second and third quantile are shown here as levels, but same calculations can be done for other values.

6.2 Selection of variables based on predictive ability

When the goal of a regression is to build a predictive model, we want to know which variables we should include in the model. In the approach above the relevance of variables was expressed in

probabilities. Another very intuitive way to assess which variables are important in the regression model, without using p-values, is by assessing the variable importance (VIMP) in the way described in the previous section. In Table 3 we have given the VIMPs for all the variables in the model, where we used the RMSE as the prediction error. The order of the variables, ordered by VIMP, shows similarities with the order based on p-values. One important difference is that the predictive variable importance of the variable origin is only the fifth variable in order of importance, whereas it was third in order of p-values. In the column Error step we have given the difference in prediction error between a model with only variables that are more important than the variable in the row and a model where the variable in that row is added to the more important variables. In the last column the marginal VIMP is given, which expresses the difference in VIMP between a model fitted with the variable included and a model without that variable included. The marginal VIMP differs from the VIMP in the sense that the latter expresses the difference in prediction error between a model where the variable is included and the same model with the coefficient corresponding to that variable set to zero. With the approach using the VIMPs it is directly clear which variables are important in terms of predictive ability and furthermore, the output is easy to understand.

Variable	**VIMP**	**Error step**	**Marginal VIMP**
Modelyear	53.67	6.39	0.85
Weight	17.06	3.44	0.4
Displacement	2.29	3.45	0.02
Cylinders	1.22	3.47	0
Origin	0.83	3.39	0.11
Horsepower	0.69	3.38	0
Acceleration	0.61	3.39	-0.01

Table 3: Table with results of VIMP

7 Conclusion

In this paper we laid down several arguments against the use of p-values. Starting with a philosophical view on probability and the historical context of p-values we gave some background with respect to our view on probability and the concept we are speaking of. Furthermore we gave several misinterpretations and misuses of p-values. The fact that p-values are accompanied by many problems is explained and admitted by many authors. Also, the risky use of p-values is one of the causes of the replicability crisis. Therefore, we came up with what we think are some directions in which we have to think in order to solve the p-value related problems. These directions did not give a final answer though. Specifically, the solution direction of adopting Bayesianism and using this to accompany models with predictive probabilities, comes with some problems. One of the problems is how priors of models should be chosen and how they should be interpreted. Furthermore, there are fields of work where p-values play a different role than in for example regression modelling. When in legal sciences the strength of evidence has to be quantified, different approaches may be needed, on which we have not focused in this paper. Future research hence has to focus mainly on alternatives for p-values. In different branches of science the alternatives will differ, depending on the particular goals. Furthermore, the philosophy of probability, which underlies these probability-related topics, is vital and hence should get much attention, both from scientists and in university courses.

References

[1] J. Arbuthnot. An argument for divine providence. *Philosophical Transactions*, 27:186–190, 1710.

[2] R. J. Aumann. Agreeing to disagree. *The annals of statistics*, pages 1236–1239, 1976.

[3] J. O. Berger and T. Sellke. Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American statistical Association*, 82(397):112–122, 1987.

[4] W. Briggs. *Uncertainty: the soul of modeling, probability & statistics*. Springer, 2016.

[5] W. M. Briggs. The crisis of evidence: Why probability and statistics cannot discover cause. *arXiv preprint arXiv:1507.07244*, 2015.

[6] W. M. Briggs. The substitute for p-values. *Journal of the American Statistical Association*, 112(519):897–898, 2017.

[7] W. M. Briggs. Everything wrong with p-values under one roof. In *International Econometric Conference of Vietnam*, pages 22–44. Springer, 2019.

[8] W. M. Briggs, H. T. Nguyen, and D. Trafimow. The replacement for hypothesis testing. In *International Conference of the Thailand Econometrics Society*, pages 3–17. Springer, 2019.

[9] R. T. Cox. *Algebra of Probable Inference*. JHU Press, 1961.

[10] B. De Finetti. Foresight: its logical laws, its subjective sources. In *Breakthroughs in statistics*, pages 134–174. Springer, 1992.

[11] D. Dua and E. Karra Taniskidou. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[12] O. Eren and N. Mocan. Emotional judges and unlucky juveniles. *American Economic Journal: Applied Economics*, 10(3):171–205, 2018.

[13] R. A. Fisher. *Statistical Methods for Research Workers*, volume 13. Oliver and Boyd, Edinburgh, 1925.

[14] R. A. Fisher. *The design of experiments*. Oliver And Boyd; Edinburgh; London, 1937.

[15] D. A. Freedman and P. B. Stark. What is the chance of an earthquake. *NATO Science Series IV: Earth and Environmental Sciences*, 32:201–213, 2003.

[16] A. Gelman et al. Objections to bayesian statistics. *Bayesian Analysis*, 3(3):445–449, 2008.

[17] G. Gigerenzer. Mindless statistics. *The Journal of Socio-Economics*, 33(5):587–606, 2004.

[18] S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, 31(4):337–350, 2016.

[19] A. Hájek. "mises redux"—redux: Fifteen arguments against finite frequentism. In *Probability, Dynamics and Causality*, pages 69–87. Springer, 1997.

[20] A. Hájek. Fifteen arguments against hypothetical frequentism. *Erkenntnis*, 70(2):211–235, 2009.

[21] A. Hájek. Interpretations of probability. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2012 edition, 2012.

[22] L. G. Halsey, D. Curran-Everett, S. L. Vowler, and G. B. Drummond. The fickle p value generates irreproducible results. *Nature methods*, 12(3):179, 2015.

[23] F. Harrell. A litany of problems with p-values, Aug 2018. URL `http://www.fharrell.com/post/pval-litany/`.

[24] E. T. Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

[25] H. Jeffreys. *Theory of probability Clarendon Press.* Oxford,, 1939.

[26] V. Keetley, A. W. Wood, J. Spong, and C. Stough. Neuropsychological sequelae of digital mobile phone exposure in humans. *Neuropsychologia*, 44(10):1843–1848, 2006.

[27] J. M. Keynes. A treatise on probability. vol. 8 of collected writings (1973 ed.), 1921.

[28] J. Kim and H. Bang. Three common misuses of p values. *Dental hypotheses*, 7(3):73, 2016.

[29] P.-S. Laplace. *Pierre-Simon Laplace Philosophical Essay on Probabilities: Translated from the fifth French edition of 1825 With Notes by the Translator*, volume 13. Springer Science & Business Media, 2012.

[30] D. A. Lawlor, G. Davey Smith, and S. Ebrahim. Commentary: The hormone replacement–coronary heart disease conundrum: is this the death of observational epidemiology? *International Journal of Epidemiology*, 33(3):464–467, 05 2004.

[31] M. B. Lewis. Mobile phones are good for you, p$<$ 0.36! observations on keetley, wood, spong and stough (2006). *Neuropsychologia*, 45(7):1580–1581, 2007.

[32] M. Lu and H. Ishwaran. A prediction-based alternative to p values in regression models. *The Journal of thoracic and cardiovascular surgery*, 155(3):1130–1136, 2018.

[33] M. Mogie. In support of null hypothesis significance testing. *Proceedings of the Royal Society of London B: Biological Sciences*, 271(Suppl 3):S82–S84, 2004.

[34] K. Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[35] G. Polya. Mathematics and plausible reasoning: Vol. ii: Patterns of plausible inference. *Princeton, NJ, Princeton Univ. Press.(2nd Ed.. 1968)*, 1954.

[36] F. Ramsey. Truth and probability. ramsey (1931) the foundations of mathematics and other logical essays, 1926.

[37] H. Reichenbach. *The theory of probability.* Univ of California Press, 1971.

[38] S. Ross. *A First Course in Probability.* Pearson Education Limited, 2014.

[39] C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27 (3):379–423, 1948.

[40] H. Tijms. *Probability: A Lively Introduction.* Cambridge University Press, 2017.

[41] D. Trafimow et al. Manipulating the alpha level cannot cure significance testing. *Frontiers in Psychology*, 9:699, 2018.

[42] J. Venn. *The logic of chance: an essay on the foundations and province of the theory of probability, with especial reference to its logical bearings and its application to moral and social science, and to statistics.* Macmillan, 1888.

[43] R. Von Mises. *Probability, Statistics, and Truth: 2d Rev. English Ed. Prepared by Hilda Geiringer.* Allen and Unwin, 1957.

[44] R. L. Wasserstein, N. A. Lazar, et al. The asa's statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.

[45] K. J. Winstein. Boston scientific stent study flawed. *Wall Street Journal*, Aug 2008. URL `https://www.wsj.com/articles/SB121867148093738861`.