Stop and Think About p-Value Statistics: Fisher, Neyman, and E. Pearson Revisited

Ryota Morimoto*

Abstract

Good statistical practice is an integral part of modern science. Null hypothesis significance testing (NHST) is the most widely used statistical method. Nevertheless, misuse and misinterpretation of NHST are widespread, and severe criticisms have been levelled against NHST. In this study, I revisit the primary documents of Fisher, Neyman, and E. Pearson relating to statistical testing. I compare Fisher’s significance testing with Neyman-Pearson hypothesis testing and clarify their thoughts on statistical testing. I hope this study will guide researchers in stopping and thinking about p-value statistics before abandoning it.

Key words: NHST, significance testing, hypothesis testing, Fisher, Neyman, Pearson, p-value

1. Criticism of NHST practice

Statistical testing is an integral part of modern science. Null hypothesis significance testing (NHST) is the most widely used statistical testing method and has supported remarkable developments in numerous fields of science. The NHST procedure can be summarized in three steps (Gigerenzer 2004, p.588):

1. Set up a statistical null hypothesis of “no mean difference” or “zero correlation.” Don’t specify the predictions of your research hypothesis or of any alternative substantive hypotheses.
2. Use 5% as a convention for rejecting the null. If significant, accept your research hypothesis. Report the result as $p < 0.05$, $p < 0.01$, or $p < 0.001$ (whichever comes next to the obtained p-value).
3. Always perform this procedure.

Gigerenzer (2004) called this procedure the null ritual that “became institutionalized in curricula, editorials, and professional associations” (ibid., p.589).

* Health Sciences University of Hokkaido
E-mail: ryota@hoku-iryo-u.ac.jp
Nevertheless, many NHST studies have failed to replicate empirical findings. One such study was conducted by social psychologist Daryl Bem, in which he reported evidence of extrasensory perception (Bem 2011). However, he was criticized for misuse of the NHST procedure (Wagenmakers, Wetzels, Borsboom and Maas 2011). Another such criticism involved social psychologist Diederik Stapel (Strobe, Postmes and Spears 2012). Stapel admitted to fabricating data and has retracted 58 papers according to Retraction Watch¹. The Reproducibility Project tried to replicate 100 psychology studies, but only 36% of the replications had statistically significant results within a 95% confidence interval (Open Science Collaboration 2015).

There has been a lot of criticism against NHST. In 2015, the editors of Basic and Applied Social Psychology banned the use of the NHST procedure and announced that the journal would no longer publish papers containing p-values (Trafimow and Marks 2015). In 2016, the American Statistical Association (ASA) released a statement regarding NHST. Wasserstein and Lazar note, “While the p-value can be a useful statistical measure, it is commonly misused and misinterpreted. This has led to some scientific journals discouraging the use of p-values, and some scientists and statisticians recommending their abandonment, with some arguments essentially unchanged since p-values were first introduced. In this context, the American Statistical Association believes that the scientific community could benefit from a formal statement clarifying several widely agreed upon principles underlying the proper use and interpretation of the p-value” (Wasserstein and Lazar 2016, p.131). This statement urges researchers to address the misuse and misinterpretation of NHST. Three years later, the ASA published a special issue “Moving to a World Beyond “p < 0.05”” consisting of 43 papers. “We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term “statistically significant” entirely (Wasserstein, Schirm and Lazar 2019, p.2). Similar radical statements can be found in, for example, “The Insignificance of Statistical Significance Testing” (Johnson 1999), “Abandon Statistical Significance” (McShane et al. 2019), and “Retire Statistical Significance” (Amrhein et al. 2019). The last study was supported by 854 signatories.

In fact, such criticism has been levelled against NHST for a long time. Morrison and Henkel edited a collection of papers titled The Significant Test Controversy in 1970. In the preface, they state, “Even their [significant tests’] strongest proponents agree that there is much misuse, misinterpretation, and meaningless use of the tests.” Spielman points out, “Some of the tests used in these areas violate strict Fisherian theory, and many young researchers in the social sciences tend to use a hybrid theory of testing” (Spielman 1974, p.211). There is an endless list of criticisms and warnings about the misuse of NHST.

¹ https://retractionwatch.com/2015/12/08/diederik-stapel-now-has-58-retractions/
The misuse and misinterpretation of NHST stems from an incoherent hybrid of two different testing theories: significance testing and hypothesis testing. The former was invented by Ronald Fisher and the latter by Jerzy Neyman and Egon Pearson. Unfortunately, not many researchers know that NHST is an incoherent hybrid, and if they have heard of it, they are not aware of the difference between the two theories. Huberty reviewed statistical textbooks from 1910–1992 in detail and stated that “[i]t is not statistical testing itself that is at fault; rather, some of the textbook presentation, teaching practices, and journal editorial reviewing may be questioned” (Huberty 1993, p. 317). Gigerenzer visited a distinguished statistical textbook author and identified “three culprits: his fellow researchers, the university administration, and his publisher” (Gigerenzer 2004, p. 588). He argued that most researchers are not interested in statistical thinking but only in publishing their papers. The university administration is interested in the number of publications and reinforces researchers’ attitude to publish as many papers as possible. The publisher demands a single-recipe cookbook, that is, the null ritual. And that’s why I think many researchers cannot break out of this vicious cycle and go on to misuse NHST.

Some of the reviews warning against the misuse of statistical testing do not seem to have examined the primary documents related to statistical testing. Such reviews may lead to further misuse or misinterpretation of NHST. Before abandoning or retiring NHST, we must learn about statistical testing from Fisher, Neyman, and Pearson. In this study, I revisit their primary documents on statistical testing and conduct a historical survey to clarify what Fisher, Pearson, and Neyman thought of statistical testing. Through the analysis of these documents, I identify the source of differences between significance testing and hypothesis testing, and conclude that there are many precious lessons that modern researchers have forgotten that can help them avoid misuse or misinterpretation of statistical testing.

2. Fisher’s significance testing

Ronald Fisher is known as the founder of modern statistics. His contributions to statistics include conceptualizing a statistical model, parameter, likelihood, analysis of variance, and the theory of experimental design. The foundation of significance testing is one of his greatest accomplishments. His testing theory precedes NHST\(^2\) as well as Neyman-Pearson hypothesis testing, all three of which are quite different. To resolve any misunderstanding related to these three, it is meaningful to revisit Fisher’s primary documents on significance testing.

Fisher introduced significance testing in a 1921 paper. At the end of this paper, he summarized his theory:

\(^2\) The origin of statistical testing is often traced to a John Arbuthnot paper published in 1710.
We may discuss the probability of occurrence of quantities which can be observed or deduced from observations, in relation to any hypotheses which may be suggested to explain these observations. We can know nothing of the probability of hypotheses or hypothetical quantities. On the other hand we may ascertain the likelihood of hypotheses and hypothetical quantities by calculation from observations: while to speak of the likelihood (as here defined) of an observable quantity has no meaning. (Fisher 1921, p. 25)

Fisher explains the p-value as the probability of data under a hypothesis or likelihood of the hypothesis. This is one of the core concepts of significance testing. He invented the concept of likelihood and defined it as “a quantity proportional to the probability that, from a population having that particular value of [parameter] \( \rho \), a sample having the observed value \( r \), should be obtained” (ibid., p. 24). Notice that he carefully distinguishes p-value—the probability of data under a hypothesis—from the probability of a hypothesis. He emphasizes, “probability and likelihood are quantities of an entirely different nature” (ibid., p. 24). To confuse the two quantities is a common misuse of NHST (Wasserstein and Lazar 2016). Fisher calculated the p-value and concluded that its value “is now much more significantly apparent” (Fisher 1921, p. 23).

In 1925, Fisher published the book *Statistical Methods for Research Workers* to supply researchers with significance testing. He wrote, “The prime object of this book is to put into the hands of research workers, and especially of biologists, the means of applying statistical tests accurately to numerical data accumulated in their own laboratories or available in the literature” (Fisher 1925, p. 17). This book deals with tests of goodness of fit, independence, significance of means, and other statistical methods that appear in current standard textbooks on statistics. Fisher invented the p-value not for verifying but for disproving a hypothesis:

The term Goodness of Fit has caused some to fall into the fallacy of believing that the higher the value of P the more satisfactorily is the hypothesis verified. Values over .999 have sometimes been reported which, if the hypothesis were true, would only occur once in a thousand trials. Generally, such cases have proved to be due to the use of inaccurate formulae, but occasionally small values of \( \chi^2 \) beyond the expected range do occur, as in Ex. 4 with the colony numbers obtained in the plating method of bacterial counting. In these cases, the hypothesis considered is as definitely disproved as if P had been .001. (ibid., p. 81)

This conservative stance is ubiquitous in Fisher’s significance testing. He invented null hypothesis with the same mindset.

The null hypothesis is another core concept of significance testing. It was first
introduced in the 1935 book *The Design of Experiments*. In the section titled "The Null Hypothesis," Fisher explains the term as follows:

[T]he hypothesis that the judgments given are in no way influenced by the order in which the ingredients have been added. . . . This hypothesis, which may or may not be impugned by the result of an experiment, is again characteristic of all experimentation. Much confusion would often be avoided if it were explicitly formulated when the experiment is designed. In relation to any experiment we may speak of this hypothesis as the “null hypothesis,” and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis. (Fisher 1935a, pp. 18–19)

According to Fisher, the null hypothesis is never proved but can be disproved, and this asymmetrical property is essential to the null hypothesis. Fisher’s conception is different from that of NHST and hypothesis testing. In NHST, the null hypothesis is symmetrical with respect to proof and disproof; that is, the null hypothesis can be disproved as well as proved. Similar is the case with hypothesis testing.

This asymmetrical property of the null hypothesis is closely related to the aim of Fisher’s significance testing and his philosophy:

It would, therefore, add greatly to the clarity with which the tests of significance are regarded if it were generally understood that tests of significance, when used accurately, are capable of rejecting or invalidating hypotheses, in so far as these are contradicted by the data; but that they are never capable of establishing them as certainly true. (Fisher 1935c, p. 474)

Fisher emphasizes that the aim of significance testing was not to accept but to reject hypotheses. Distinguishing between rejection and acceptance of hypotheses is important for understanding not only significance testing but also Neyman-Pearson hypothesis testing. This will be discussed in detail later.

Fisher required exactness of the null hypothesis. He writes, “[N]ull hypothesis must be exact, that is free from vagueness and ambiguity, because it must supply the basis of the ‘problem of distribution,’ of which the test of significance is the solution” (Fisher 1935a, p. 19). He did not approve as null hypothesis a vague hypothesis where the subject could discriminate between two sorts of objects. It is a matter of degree to make discrimination; therefore, such an expression is inexact. However, Fisher raised an exact hypothesis that death rates of two groups of animals were equal as an example of the null hypothesis. He said, “In such cases it is evidently the equality rather than any particular values of the death-rates that the experiment is designed to test, and possibly to disprove” (*ibid.*, p. 20). Exactness is necessary for the null
hypothesis to be disproved. If it were not exact, it could not be disproved.

To set up an exact null hypothesis, researchers must specify statistics and referent distribution. In *Statistical Methods for Research Workers*, Fisher introduces statistics and three principal distributions: “A statistic is a value calculated from an observed sample with a view to characterising the population from which it is drawn” (Fisher 1925, pp. 44–45). Average and variance are examples of statistics, and the character of the population is a parameter that researchers want to know. Fisher then discusses three principal distributions: the normal distribution, the Poisson series, and the binomial distribution. He says, “It is important to have a general knowledge of these three distributions, the mathematical formulae by which they are represented, the experimental conditions upon which they occur, and the statistical methods of recognizing their occurrence” (*ibid.*, pp. 45–46). Research workers must specify statistics and related distributions depending on their experimental conditions. Only after setting up the null hypothesis exactly can they begin collecting the data.

It is noteworthy that Fisher set up a single hypothesis to test significance, not two. He points out some misunderstandings regarding significance testing.

On the whole the ideas … that the purpose of the test is to discriminate or “decide” between two or more hypotheses, have greatly obscured their understanding, when taken not as contingent possibilities but as elements essential to their logic. The appreciation of such more complex cases will be much aided by a clear view of the nature of a test of significance applied to a single hypothesis by a unique body of observations. (Fisher 1959, p. 42)

Although modern research workers often devise an alternative hypothesis, significance testing allows for only one hypothesis. I will compare this property of significance testing with Neyman-Pearson hypothesis testing later.

How can a \(p\)-value be used to reject the null hypothesis? Fisher introduces a criterion named the level of significance or the significance level. After explaining standard deviation for a normal distribution in *Statistical Method for Research Workers*, Fisher refers to the significance level:

The value for which \(P = .05\), or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion, we should be led to follow up a negative result only once in 22 trials, even if the statistics were the only guide available. Small effects would still escape notice if the data were insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty. (Fisher 1925, pp. 47–48)
In 1926, Fisher raised the question, “When is a Result Significant?” (p. 503):

If one in twenty does not seem high enough odds, we may, if we prefer it, draw
the line at one in fifty (the 2 per cent. point), or one in a hundred (the 1 per cent.
point). Personally, the writer prefers to set a low standard of significance at the
5 per cent. point, and ignore entirely all results which fail to reach this level. A
scientific fact should be regarded as experimentally established only if a properly
designed experiment rarely fails to give this level of significance. (Fisher 1926,
p. 504, emphasis in the original)

The significance level is used to judge whether the null hypothesis is significant or
not. If the actual p-value is below a certain level of significance, the null hypothesis
is considered significant. This is the rule of significance testing. Fisher interprets the
p-value as the standard for rejecting a null hypothesis. He suggests 0.05 as the value
of significance level and claims that lowering this standard has little effect in a small
sample. Furthermore, he says that 0.05 is specified by convenience or preference. It
does not have a statistical foundation.

Fisher believed that significance testing, especially a judgment on the null hy-
pothesis, is justified by randomization, which is one of the three principles of the
experimental design he invented. Randomization is the process of random alloca-
tion of treatments to different experimental units. In The Design of Experiments,
he points out, “[I]t would be impossible to present an exhaustive list of such possible dif-
fences appropriate to any one kind of experiment, because the uncontrolled causes
which may influence the result are always strictly innumerable” (Fisher 1935a, p. 21).
He then suggests the process of randomization to solve this issue:

[W]ith satisfactory randomisation, its validity is, indeed, wholly unimpaired, let
us imagine all causes of disturbance—the strength of the infusion, the quantity
of milk, the temperature at which it is tasted, etc.—to be predetermined for
each cup. (ibid., p. 23)

He adds, “It may be said that the simple precaution of randomization will suffice to
guarantee the validity of the test of significance, by which the result of the experiment
is to be judged” (ibid., p. 24). In Statistical Methods for Research Workers, he wrote:

For our test of significance to be valid the difference in fertility between plots
chosen as parallels must be truly representative of the differences between plots
with different treatment; and we cannot assume that this is the case if our plots
have been chosen in any way according to a pre-arranged system; for the system-
atic arrangement of our plots may have, and tests with the results of uniformity
trials show that it often does have, features in common with the systematic vari-
ation of fertility, and thus the test of significance is wholly vitiated. (Fisher 1925,
Fisher repeatedly emphasizes the importance of randomization for significance testing. Do researchers allow randomization in their experiments when using significance testing? If not, such a test would not be justified.

Thus far, we have discussed Fisher’s significance testing based on his original writings. The significance testing procedure is summarized below.

1. Set up the null hypothesis \( (H_0) \).
2. Specify test statistics and referent distribution.
3. Fix the significance level.
4. Set up the experimental design.
5. Collect the data, and calculate the statistics.
6. Calculate the \( p \)-value, that is, the probability of data under the null hypothesis.
7. Reject \( H_0 \) if the \( p \)-value is small and otherwise retain \( H_0 \).

However, Fisher later changed his conception of \( p \)-values and significance testing. For Fisher, the purpose of significance testing was to judge whether the null hypothesis should be rejected, and the \( p \)-value was interpreted as the basis for this judgment. According to Lehmann, Fisher believed in this thought from the first edition of *Statistical Methods for Research Workers* in 1925 up to its twelfth edition in 1954. However, in the thirteenth edition of 1958, Fisher changed his mind:

In preparing this table we have borne in mind that in practice we do not want to know the exact value of \( P \) for any observed \( X^2 \), but, in the first place, whether or not the observed value is open to suspicion. If \( P \) is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. Belief in the hypothesis as an accurate representation of the population sampled is confronted by the logical disjunction: Either the hypothesis is untrue, or the value of \( X^2 \) has attained by chance an exceptionally high value. The actual value of \( P \) obtainable from the table by interpolation indicates the strength of the evidence against the hypothesis. A value of \( X^2 \) exceeding the 5 per cent. point is seldom to be disregarded. (Fisher 1958, p. 80)

Fisher interprets the \( p \)-value as the strength of evidence against the null hypothesis. This conception is obviously different from Fisher’s earlier thoughts. In 1956, Fisher published *Statistical Methods and Scientific Inference* and wrote in the same vein:

In general tests of significance are based on hypothetical probabilities calculated from their null hypotheses. They do not generally lead to any probability statements about the real world, but to a rational and well-defined measure of reluctance to the acceptance of the hypotheses they test. (Fisher 1956, p. 44, emphasis
Later, Fisher regarded the $p$-value as the measure or strength of reluctance to accept the null hypothesis, indicating that the numerical value of the $p$-value is in itself meaningful. In his earlier conception of the $p$-value, it was crucial to determine whether the value was below the significance level. The numerical value did not indicate any measure of reluctance.

Sober explains the difference between these two interpretations with an example\textsuperscript{3}. Suppose you want to test the null hypothesis that a coin is fair and specify the significance level at 0.05. You toss the coin 20 times and obtain 4 heads. The $p$-value of this outcome is calculated to be 0.012. Under Fisher’s earlier conception, the null hypothesis should be rejected. If a different outcome, such as 6 heads in 20 tosses, is obtained, the $p$-value is 0.115, and the hypothesis should not be rejected. Because the significance level is specified by convention or preference, judgment differs with the value of the outcome. However, according to Fisher’s later conception, the lower the $p$-value, the stronger the evidence against the null hypothesis (Sober 2008, p. 54). Although there are many review papers on statistical testing, only a few studies have analyzed this change of conception. Most studies have stereotyped significance testing based on Fisher’s later conception.

3. Neyman-Pearson hypothesis testing

In 1926, Egon Pearson sent some notes on hypothesis testing to Jerzy Neyman, which marked the beginning of their collaboration (Lehmann 2011). Neyman and Pearson published 10 joint papers from 1928–1938. They modified Fisher’s significance testing and invented hypothesis testing.

The first joint paper by Neyman and Pearson was published in 1928. In the last sentence of “Introductory,” they argue, “It is indeed obvious, upon a little consideration, that the mere fact that a particular sample may be expected to occur very rarely in sampling from [the sample] $\Sigma$ would not in itself justify the rejection of the hypothesis that it had been so drawn, if there were no other more probable hypotheses conceivable” (Neyman and Pearson 1928, p. 178). Neyman and Pearson claim that Fisher’s significance testing could not justify rejection of a hypothesis. Instead, they invented hypothesis testing.

Neyman and Pearson state the purpose of hypothesis testing as follows: “In general the method of procedure is to apply certain tests or criteria, the results of which will enable the investigator to decide with a greater or less degree of confidence

\textsuperscript{3} While Sober analyzes the two versions of Fisher’s significance testing, he criticizes both and does not accept either of them. He also criticizes Neyman-Pearson hypothesis testing (Sober 2008).
whether to accept or reject Hypothesis A, or, as is often the case, will show him that further data are required before a decision can be reached” (ibid., p.175). Hypothesis testing focuses on whether to accept or reject a hypothesis. They mention the purpose of hypothesis testing more clearly in a 1933 paper:

Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong. Here, for example, would be such a “rule of behaviour”: to decide whether a hypothesis, H, of a given type be rejected or not, calculate a specified character, x, of the observed facts; if \( x > x_0 \) reject H, if \( x \approx x_0 \) accept H. Such a rule tells us nothing as to whether in a particular case H is true when \( x \approx x_0 \) or false when \( x > x_0 \). But it may often be proved that if we behave according to such a rule, then in the long run we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false. (Neyman and Pearson 1933a, p.291)

They express hypothesis testing as a rule of behavior or decision making in the long run. A rule of behavior is crucial for their theory, which will be discussed later.

Neyman and Pearson admitted the behavior of accepting a hypothesis in statistical testing. This characteristic is essential for hypothesis testing, but Fisher never admitted it in his significance testing. Neyman explains this in detail:

The terms “accepting” and “rejecting” a statistical hypothesis are very convenient and are well established. It is important, however, to keep their exact meaning in mind and to discard various additional implications which may be suggested by intuition. Thus, to accept a hypothesis \( H \) means only to decide to take action \( A \) rather than action \( B \). This does not mean that we necessarily believe that the hypothesis \( H \) is true. Also, if the application of a rule of inductive behavior “rejects” \( H \), this means only that the rule prescribes action \( B \) and does not imply that we believe that \( H \) is false. (Neyman 1950, pp.259–260)

Neyman distinguishes between accepting and rejecting a hypothesis and believing a hypothesis. He uses the word “believing” in its Bayesian context. Acceptance and rejection are matters of neither belief nor truth. Accepting or rejecting a hypothesis is frequently confused with the truth or falsehood of the hypothesis (Wasserstein and Lazar 2016, McShane et al. 2019). Neyman and Pearson kept this distinction in mind, and so did Fisher. They use these terms in a technical sense. Acceptance and rejection refer to decision making to take some action.

There is an important but seldom mentioned point regarding the types of decision options. Neyman and Pearson do not set dichotomous decisions, such as rejection
or acceptance of the hypothesis. They admit to another kind of behavior.

A statistical test is therefore equivalent to a rule of the following type,

(a) Reject $H_0$ if $\Sigma$ falls into a region $w$.
(b) Accept $H_0$ if $\Sigma$ falls into another region $w'$.
(c) Remain in doubt if $\Sigma$ falls into a third region $w''$. (Neyman and Pearson 1933b, p. 493)

Neyman and Pearson approved the behavior to “remain in doubt” in hypothesis testing. If the third option was absent, there would not be exhaustive alternatives in decision making. Thus, hypothesis testing covers all logical possibilities exhaustively. Fisher’s significance testing either to reject a hypothesis or not also covers all possibilities exhaustively. However, NHST admits only rejection or acceptance of the hypothesis, which does not cover all possibilities. In this sense, Neyman-Pearson hypothesis testing and Fisher’s significance testing are more modest than the modern NHST.

There is another key characteristic of hypothesis testing. Neyman and Pearson’s first joint paper begins by discussing the problem of testing a single hypothesis and then deals with the likelihood comparison of two hypotheses:

Suppose there to be two hypotheses regarding the population from which a given sample has been drawn; there will be one density filed for Hypothesis A corresponding to [the sample] $\Sigma$, and another for Hypothesis A’ corresponding to $\Sigma'$. By surrounding the sample point ... we have in the ratio of the “weights” of the elements a measure of the relative frequency of occurrence, on the two hypotheses, of samples lying within the prescribed limits. (Neyman and Pearson 1928, pp. 184–185, emphasis in the original).

As mentioned earlier, probability of data under a hypothesis is the likelihood of the hypothesis. A likelihood comparison is essential for hypothesis testing. It usually deals with two or more hypotheses, although Fisher’s significance testing deals with a single hypothesis. Neyman and Pearson introduced the likelihood criterion expressed as the ratio of likelihood of hypotheses. They state:

Without claiming that this method is necessarily the “best” to adopt, we suggest that the use of this contour system ... provides at any rate one clearly defined method of discriminating between samples for which Hypothesis A is more probable and those for which it is less probable. It is a method which takes into account the likelihood of alternative hypotheses. (Neyman and Pearson 1928, p. 188)

They claim that likelihood ratio testing is a clearly defined method. Five years later, in their seminal paper, “On the Problem of the Most Efficient Tests of Statistical
Hypothesis," they prove that it is the best method for hypothesis testing. Notice that hypothesis testing admits more than one alternative hypothesis, whereas NHST usually admits one null hypothesis and one alternative hypothesis\(^4\).

The introduction of alternative hypotheses and acceptance procedures in hypothesis testing leads to one more type of error than that found in Fisher's significance testing:

We may accept or we may reject a hypothesis with varying degrees of confidence; or we may decide to remain in doubt. But whatever conclusion is reached the following position must be recognized. If we reject H\(_0\), we may reject it when it is true; if we accept H\(_0\), we may be accepting it when it is false, that is to say, when really some alternative H\(_i\) is true. These two sources of error can rarely be eliminated completely. (Neyman and Pearson 1933a, pp. 295–296)

These types of errors are called Type I and Type II errors. A Type I error occurs when we reject a true hypothesis, and a Type II error occurs when we accept a false hypothesis.

Hypothesis testing prioritizes two types of errors. Neyman used a marketing toxic drug as an example. In this case, a Type I error occurs when a toxic drug declared nontoxic or harmless is put on the market, and a Type II error occurs when nontoxic drugs declared toxic are returned to the manufacturer. A Type I error is unusually more unpleasant than a Type II error. Neyman states, “There may be financial loss to the manufacturer and an increase in the price of the drug” (Neyman 1950, p. 263). The death of experimental animals is also unpleasant. Neyman explains:

As already mentioned, the situation where the consequences of the two kinds of errors are of unequal importance is of a very general occurrence. It is true that in many cases the relative importance of the errors is a subject matter. . . . However, this subjective element lies outside of the theory of statistics. The essential point to notice is that, in most cases, the person applying a test of a statistical hypothesis considers one of the possible errors more important to avoid than the other.

Postulating this to be the ordinary case we will use the expression *error of the first kind* to describe that particular error in testing hypotheses which is considered more important to avoid. The less important error will be called the *error of the second kind*. In the rare cases where the two kinds of error are of exactly the same importance, it is immaterial which of them is called error of the first kind and which the error of the second kind. (Neyman 1950, p. 263, emphasis in the original)

---

\(^4\) Neyman and Pearson seldom used the term null hypothesis invented by Fisher. Instead, they used the term “hypothesis tested.”
As an example of rare cases in which these two types of errors are of equal importance, they raise the hypothesis of betting against double six when the probability of double six is 0.6, and the hypothesis of betting on double six when the probability is 0.4. Except for such rare cases, a Type I error is usually more serious than a Type II error. This is due to the associated financial loss and death of experimental animals. These reasons lie outside the scope of statistics.

It is desirable that the probabilities of both types of errors are minimized simultaneously, but this is impossible. Neyman and Pearson consider that first the probability of Type I error should be minimized at a fixed level, and then the probability of Type II error should be made as small as possible. This fixed level is the significance level and is written as \( \alpha \).

We shall now consider how to find the best critical region for \( H_0 \) with regard to a single alternative \( H_1 \); this will be the [critical] region \( w_0 \) for which \( P_1(w_0) \) is a maximum subject to the condition that \( P_0(w_0) = \varepsilon \). (Neyman and Pearson 1933a, p. 298)

Under this condition, they prove that the likelihood ratio test is the most powerful among all possible tests. The consequence of this proof is called the Neyman-Pearson lemma, which is the greatest accomplishment of their collaboration. The term “power” is used here in a technical sense. It means the rejection probability as a function of the alternative, which is expressed as \( 1 - \beta \).

The likelihood ratio test is defined as \( (L(x) = P(x|H_1)/P(x|H_0), a) \), where \( x \) is data, and \( a \) is some constant such that \( P(L(x) > a|H_0) = \alpha \). If \( L(x) > a \), then \( L(x) \) rejects \( H_0 \), and if \( L(x) \leq a \), then \( L(x) \) accepts \( H_1 \). An arbitrary statistical test is defined as \( (M(x), b) \), where \( b \) is a constant such that \( P(M(x) > b|H_0) = \alpha \). If \( M(x) > b \), then \( M(x) \) rejects \( H_0 \), and if \( M(x) \leq b \), then \( M(x) \) accepts \( H_1 \). Thus, \( P(L(x) > a|H_1) > aP(L(x) > a|H_0) \) if \( L(x) > a \), and \( P(L(x) \leq a|H_1) \leq aP(L(x) \leq a|H_0) \) when \( L(x) \leq a \).

Suppose that the significance level \( \alpha \) is fixed at a certain value under a statistical test, that is, \( P(L(x) > a|H_0) = P(M(x) > b|H_0) \). Under this condition, Neyman and Pearson proved that the likelihood ratio test is the most powerful among all statistical tests, that is, \( P(L(x) > a|H_1) > P(M(x) > b|H_1) \). Therefore, the likelihood ratio ratio
test is the most powerful test\(^6\).

The procedure of Neyman–Pearson hypothesis testing can be summarized as follows:

1. Set up the tested hypothesis \((H_0)\) and the alternative hypotheses \((H_i)\).
2. Specify test statistics and referent distribution.
3. Specify the significance level \(\alpha\) and determine a critical region.
4. Maximize the power \((1 - \beta)\) under the fixed value \(\alpha\).
5. Collect data and calculate statistical values.
6. Reject \(H_0\) if data fall into the critical region; accept \(H_0\) if data fall into another region; and remain in doubt if data fall into a third region.

4. The source of differences between significance and hypothesis testing

We have discussed Fisher’s significance testing and Neyman-Pearson hypothesis testing. Both theories are self-consistent. Therefore, the problem is not about which of these two theories is true. Rather, the issue is which theory is better or more appropriate. Both sides clearly distinguish between the truth of a hypothesis and the appropriateness of a judgment. They mainly argue about the latter and do not deal with the former in statistical testing. In this section, I analyze the differences between these two theories.

I think that the crucial issue is the type of procedure that should be admitted. While both sides admitted the rejection procedure, only Neyman and Pearson admit the acceptance procedure. As mentioned previously, Neyman and Pearson consider hypothesis testing as a theory of decision making and thought it was natural to introduce the acceptance procedure in statistical testing. Fisher denied it and criticized Neyman and Pearson, as well as Wald. Abraham Wald had developed hypothesis testing as a decision-making theory.

Much ingenuity and skill must be exercised in making an acceptance procedure a really effectual and economical one. It is not therefore at all in disdain of an artifice of proved value, in commerce and technology, that I shall emphasize some of the various ways in which this operation differs from that by which improved theoretical knowledge is sought in experimental research. This emphasis is primarily necessary because the needs and purposes of workers in the experimental sciences have been so badly misunderstood and misinterpreted. It is, of course,

\(^6\) More precisely, this proof is made under simple hypotheses, where the distribution of the population is uniquely specified. For composite hypotheses where the distribution is not uniquely specified, there are limited situations in which the likelihood ratio test is the most powerful test.
also to be suspected that those authors, such as Neyman and Wald, who have treated these tests with little regard to their purpose in the natural sciences, may not have been more successful in the application of their ideas to the needs of acceptance procedures. It is, however, to the evident advantage of both kinds of application that the theories developed and taught to mathematicians should not confuse their several requirements. (Fisher 1956, pp. 76–77)

According to Fisher, statistical testing should be considered a scientific methodology. He regards significance testing as a tool for experimental sciences but rejects Neyman-Pearson hypothesis testing and Wald’s decision theory as approaches not suitable for science.

There is another reason Fisher denied the acceptance procedure. If he admitted it, he would inevitably admit a Type II error, that is, an error of accepting a false hypothesis.

For the logical fallacy of believing that a hypothesis has been proved to be true, merely because it is not contradicted by the available facts, has no more right to insinuate itself in statistical than in other kinds of scientific reasoning. Yet it does so only too frequently. Indeed, the “error of accepting an hypothesis when it is false” has been specially named by some writers “errors of the second.” … In fact that “errors of the second kind” are committed only by those who misunderstand the nature and application of tests of significance. (Fisher 1935c, p. 474)

Fisher regards a Type II error as a logical fallacy. It should be noted that there is asymmetry between verification and falsification, as Popper (1959) emphasizes. Falsification is a logically valid inference modus tollens. However, verification is an invalid inference. It is logically or deductively impossible that a hypothesis is proved to be true from the data. Fisher is correct in this sense.

Andrews and Huss analyzed Type I and Type II errors and made fairly fine-grained distinctions. They divided Type II errors into two types of errors.

(Horn 1) Type-II error = failure to reject a null hypothesis when it is in fact false.

(Horn 2) Type-II error = acceptance of a null hypothesis when it is in fact false. (Andrews and Huss 2014, p. 720)

They rightly claim that (Horn 1) is not a fundamental error, while (Horn 2) is an error. (Horn 1) describes attitudes consistent with agnosticism and puts a clear decision on hold. If you do not know a fact, it is not an error to put it off. (Horn 2) is obviously erroneous because it is a fundamental error to accept a false hypothesis. This may be what Fisher wished to say. Fisher would admit (Horn 1) definition of
a Type II error, whereas Neyman and Pearson would presume (Horn 2) definition a
Type II error.
Incidentally, as previously stated, significance testing was applied to a single
hypothesis, and Fisher denied setting alternative hypotheses:

It was only when the relation between a test of significance and its correspond-
ing null hypothesis was confused with an acceptance procedure that it seemed
suitable to distinguish errors in which the hypothesis is rejected wrongly, from
errors in which it is “accepted wrongly” as the phrase does. The frequency of
the first class, relative to the frequency with which the hypothesis is true, is
calculable, and therefore controllable simply from the specification of the null
hypothesis. The frequency of the second kind must depend not only on the fre-
quency with which rival hypotheses are in fact true, but also greatly on how
closely they resemble the null hypothesis. Such errors are therefore incalculable
both in frequency and in magnitude merely from the specification of the null
hypothesis, and would never have come into consideration in the theory only of
tests of significance, had the logic of such tests not been confused with that of
acceptance procedures. (Fisher 1955, p.73).

Fisher says that the frequency of a Type II error could not be calculated, and this
is the reason he rejected alternative hypotheses. However, Neyman and Pearson in-
vented a power function that gave the probabilities of a Type II error. If an alternative
hypothesis is specified, the frequency of a Type II error can be calculated. Thus, the
odds are in favor of Neyman and Pearson’s introduction of alternative hypotheses.

There is another difference between significance testing and hypothesis testing.
Neyman says, “The subject of the dispute may be symbolized by the opposing terms
“inductive reasoning” and “inductive behavior”. Professor Fisher is a known propo-
nent of inductive reasoning” (Neyman 1961, p.148). Both sides share the idea that
statistical testing is inductive. The difference lies in whether it is a reasoning or a
rule of behavior.

As Neyman says, Fisher regarded significance testing as inductive inference or
reasoning. Fisher explains inductive inference as follows:

I have called my paper “The Logic of Inductive Inference.” It might just as well
have been called “On making sense of figures.” For everyone who does habitually
attempt the difficult task of making sense of figures is, in fact, essaying a logical
process of the kind we call inductive, in that he is attempting to draw inferences
from the particular to the general; or, as we may usually say in statistics, from
the sample to the population. (Fisher 1935b, p.39)

This kind of inductive inference—that is, drawing from the particular to the general
or from the sample to the population—is an empirical generalization or an enumer-
ative induction. In addition, Fisher thought that inductive reasoning is the process of approaching the truth:

In inductive reasoning we are performing part of the process by which new knowledge is created. The conclusions normally grow more and more accurate as more data are included. It should never be true, though it is still often said, that the conclusions are no more accurate than the data on which they are based. Statistical data are always erroneous, in greater or less degree. The study of inductive reasoning is the study of the embryology of knowledge, of the processes by means of which truth is extracted from its native ore in which it is fused with much error. (Fisher 1935b, p. 54, emphasis in the original)

Fisher states that inductive reasoning could extract the truth from erroneous data. He compares this process with the embryology of knowledge.

In contrast, Neyman states that hypothesis testing is a rule of inductive behavior:

The term “rule of inductive behavior” was introduced with reference to situations where the desirability of the several actions contemplated depends on the nature of the frequency function of some observable random variables. This term was used to describe any rule for choosing an action in accordance with the particular values of these random variables determined by observation. It follows that a test of a statistical hypothesis is a rule of inductive behavior. (Neyman 1950, p. 258, emphasis in the original)

Neyman explains the rule of inductive behavior using an example: “The term “inductive behavior” means simply the habit of humans and of other animals (Pavlov’s dog, etc.) to adjust their actions to noticed frequencies of events, so as to avoid undesirable consequences” (Neyman 1961, p. 148). The metaphor of Pavlov’s dog makes it easy to understand inductive behavior. Neyman explains inductive behavior in more detail as follows:

With many phenomena certain permanencies appear quite stable. This created the habit of regulating our actions in regard to some observed events by referring to the permanencies which at the particular moment seem to be established. This is what we call inductive behavior.

Early in human history it was established that rain or snow storms follow the appearance of heavy clouds. This is one of many permanencies noted. . . . human beings and also some animals tend to take cover whenever dark clouds appear in the sky. This is an example of inductive behavior. (Neyman 1950, pp. 1–2)

Inductive behavior refers to the habit of regulating actions by referring to permanencies. Neyman claims that the acceptance or rejection of hypotheses by hypothesis
testing is a rule of inductive behavior. His conception is quite different from that of Fisher.

Again, the crucial issue is whether the acceptance procedure should be admissible in statistical testing. Fisher aimed to invent significance testing as a tool for natural sciences. This is the reason he never approved of the acceptance procedure and criticized hypothesis testing for introducing it:

The “Theory of Testing Hypotheses” was a later attempt, by authors who had taken no part in the development of these tests, or in their scientific application, to reinterpret them in terms of an imagined process of acceptance sampling, such as was beginning to be used in commerce; although such processes have a logical basis very different from those of a scientist engaged in gaining from his observations an improved understanding of reality. (Fisher 1956, pp. 4–5)

Fisher blamed Neyman and Pearson for making significance testing unscientific.

The conclusions drawn from such tests constitute the steps by which the research worker gains a better understanding of his experimental material, and of the problems which it presents.

It is noteworthy, too, that the men who felt the need for these tests, who first conceived them, or later made them mathematically precise, were all actively concerned with researches in the natural sciences. More recently, indeed, a considerable body of doctrine has attempted to explain, or rather to reinterpret, these tests on the basis of quite a different model, namely as means to making decisions in an acceptance procedure. (ibid., pp. 75–76)

As mentioned in Section 3, Fisher says that the null hypothesis is never proved but possibly disproved. He hypothesized that a subject possesses no sensory discrimination between two different sorts of objects as a null hypothesis and says, “[T]his hypothesis could be disproved by a single failure, but could never be proved by any finite amount of experimentation” (Fisher 1935a, p. 19). Fisher claims the norm of science: what science should be.

However, Neyman admits the acceptance procedure and says, “The problem of testing a statistical hypothesis occurs when circumstances force us to make a choice between two courses of action” (Neyman 1950, p. 258). He explains his view of science as follows:

[W]e ... examine the concluding phase of scientific research with the object of determining the nature of its constituent mental processes. This is just the phase frequently described as induction. I hope to show that its constituent processes fall under three headings: (i) visualization of several possible sets of hypotheses relevant to the phenomena studied, (ii) deductions from these sets of hypotheses,
and (iii) an act of will or decision to take a particular action, perhaps to assume a particular attitude towards the various sets of hypotheses mentioned under (i). ... These processes are certainly not any sort of “reasoning”, at least not in the sense in which this word is used in other instances; they are acts of will. (Neyman 1957, pp. 10–11).

Neyman analyzed scientific practices rather than the norm of science. He explains scientists’ mental processes, which included acts of will.

Ultimately, the issue of whether acceptance of a hypothesis is admissible stems from the different views of science between Fisher and Neyman. Fisher’s view of science resembles that of Popper in that acceptance of a hypothesis is prohibited. Their attitude toward science is conservative. Nevertheless, there is an important difference between the two. Fisher admits inductive inference in science, whereas Popper never admitted it as a scientific inference. In contrast, Neyman’s view of science includes scientists’ mental processes.

When Fisher argues for inductive reasoning, he focuses on the inferential aspect of statistical testing. In contrast, when Neyman discusses inductive behavior, he treats scientists’ decision making in statistical testing. These are different aspects of scientific activities. Fisher developed scientific inference, but Neyman analyzed the behavior of scientists. Theory and people are different. This difference seems to affect their conception of testing theory: scientific inference and scientists’ decision making.

Interestingly, Pearson distanced himself from this issue. In replying to Fisher’s criticisms in “Statistical Methods and Scientific Induction,” Pearson states, “Professor Fisher’s final criticism concerns the term “inductive behaviour”; this is Professor Neyman’s field rather than mine” (Pearson 1955, p. 207). In addition, he wrote:

[W]e were not speaking of the final acceptance or rejection of a scientific hypothesis on the basis of statistical analysis. We speak of accepting or rejecting a hypothesis with a “greater or less degree of confidence”. Further, we were very far from suggesting that statistical methods should force an irreversible acceptance procedure upon the experimenter. Indeed, from the start we shared Professor Fisher’s view that in scientific enquiry, a statistical test is “a means of learning”. ... I would agree that some of our wording may have been chosen inadequately, but I do not think that our position in some respects was or is so very different from that which Professor Fisher himself has now reached. (ibid., pp. 206–207, emphasis in the original)

It seems that Pearson hesitated to enter a controversy with Fisher.
5. Lessons learned from primary documents on statistical testing

We can learn several lessons from Fisher’s, Neyman’s, and Pearson’s primary documents on statistical testing. As mentioned above, Fisher never admitted alternative hypotheses and the acceptance procedure because of his conservative view of science. He claimed that significance testing, especially a judgment on the null hypothesis, is not justified without setting up an experimental design. Do you perform randomization in your experiment when you criticize significance testing? In addition, Fisher later changed his conception of \( p \)-value and significance testing and gave two kinds of interpretations. Although most review papers explain the \( p \)-value of Fisher’s significance theory as the strength of evidence against the null hypothesis, Fisher had earlier regarded it as the basis for judgment, which is similar to NHST.

However, Neyman and Pearson approved the behavior not only to accept a hypothesis but also to remain in doubt. Their theory covers all possibilities exhaustively, unlike the NHST. Neyman and Pearson also admitted alternative hypotheses and included a Type II error. To deal with this type of error carefully, they introduced a power function and proved the so-called “Neyman-Pearson lemma.” According to this lemma, if under the fixed significance level \( \alpha \), power function \( 1 - \beta \) should be maximized, then the likelihood test is the most powerful among all possible tests. This lemma is the greatest achievement of their joint work and one of the core results in modern statistical testing.

Ultimately, the crucial issue is whether the acceptance procedure should be allowed in statistical testing. This difference stems from Fisher’s and Neyman’s different views of science, which cannot be learned in statistical textbooks. Considering scientific practice, Fisher’s theory is too narrow. However, this does not mean that we should admit the acceptance procedure easily in science. This way of thinking is too wide and may lead to an anything goes approach. At this point, the Neyman-Pearson lemma is a useful and important result. Either way, researchers should learn the philosophy of science to think about their views of science.

Many reviewers have warned against the misuse of NHST. While I agree with them, I suggest revisiting the primary documents on statistical testing and learning from them. This will help recognize the differences among significance testing, hypothesis testing, and NHST. As mentioned previously, some reviewers have proposed abandoning or retiring NHST. However, misuse of NHST should not be confused with whether it is true. Note again that people and theories are different. The misuse of a theory should not be blamed for the theory itself; rather, it is the researchers who are at blame. Therefore, misuse is not an excuse for abandoning or retiring the NHST.

In fact, there are several problems with significance testing, hypothesis testing, and NHST, including, for example, the sample size problem and the stopping rule problem. The former problem was addressed by Lindley (1957). The \( p \)-value statis-
tics are sensitive to sample size. As the sample size increases, the null hypothesis becomes easy to reject. The latter problem was addressed by Howson and Urbach (2006). The stopping rule determines when an experiment should be stopped. Howson and Urbach gave the example of a coin flip. In testing the fair coin hypothesis, one researcher stopped the experiment after the coin was flipped 20 times, and another stopped the experiment when six heads appeared. Howson and Urbach illustrated that different stopping rules led to different decisions regarding whether the hypothesis should be rejected. These problems are logical fallacies of the $p$-value statistics. Therefore, such problems concern the theories, not the researchers.

Remember that the logic of the $p$-value statistics is not deduction. Not only Fisher, but also Neyman and Pearson, recognized that statistical testing is based on induction. Moreover, statistics in general is not deductive. Every statistical theory has some side effects. $P$-value statistics have both advantages and disadvantages. This is similar to Bayesian statistics. It is often recommended to abandon the $p$-value statistics and use Bayesian statistics. I do not think this is a sensible approach for two reasons. First, all statistical theories have both advantages and disadvantages. Second, there are many other options besides these two theories. For example, Akaike information criterion (AIC) and Bayesian information criterion (BIC) are well-known model selection approaches. Likelihood approach is another option. Simple dichotomization—that is, to choose either $p$-value statistics or Bayesian statistics—can sometimes lead to poor judgment.

Revisiting the primary documents on statistical testing can help avoid the misuse or misinterpretation of the $p$-value statistics. If Fisher, Neyman, and Pearson were alive, what would they have thought about the current situation? Gigerenzer rightly answers, “Readers beware. Each of these eminent statisticians would have rejected the null ritual as bad statistics” (Gigerenzer 2004, p. 589). Thus, it is better to stop and think about $p$-value statistics.

References
Amrhein, V., Greenland S. and McShane B. (2019). Retire Statistical Significance. *Nature* 567: 305–307.
Andrews, K. and Huss, B. (2014). Anthropomorphism, Anthropectomy, and the Null Hypothesis. *Biology and Philosophy* 29(5): 711–729.
Bem, D. J. (2011). Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect. *Journal of Personality and Social Psychology* 100: 1–19.
Barnett, V. (1999). *Comparative Statistical Inference*. Third Edition. John Wiley & Sons.
Fisher, R. (1921). On the “Probable Error” of a Coefficient of Correlation Deduced from a Small Sample. *Metron* 1: 3–32.
Fisher, R. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical
Fisher, R. (1925). *Statistical Methods for Research Workers*. First Edition. Oliver & Boyd. (Twelfth Edition in 1954, Thirteenth Edition in 1958).

Fisher, R. (1926). The Arrangement of Field Experiments. *Journal of the Ministry of Agriculture of Great Britain* 33: 503–513.

Fisher, R. (1929). The Statistical Method in Psychical Research. *Proceedings of the Society for Psychical Research* 39: 189–192.

Fisher, R. (1935a). *The Design of Experiments*. Oliver and Boyd, Edinberg.

Fisher, R. (1935b). The Logic of Inductive Inference. *Journal of the Royal Statistical Society* 98: 39–82.

Fisher, R. (1935c). Statistical Tests. *Nature* 136: 474.

Fisher, R. (1955). Statistical Methods and Scientific Induction. *Journal of the Royal Statistical Society* B, 17: 69–78.

Fisher, R. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd.

Gigerenzer, G. (2004). Mindless Statistics. *the Journal of Socio-Economics* 33: 587–606.

Howson, C. and Urbach, P. (2006). *Scientific Reasoning: The Bayesian Approach*. Third Edition. Open Court.

Huberty, C. J. (1993). Historical Origins of Statistical Testing Practices: The Treatment of Fisher Versus Neyman-Pearson Views in Textbooks. *The Journal of Experimental Education* 61(4): 317–333.

Johnson, D. (1999). The Insignificance of Statistical Significance Testing. *The Journal of Wildlife Management* 63(3): 763–772.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? *Journal of the American Statistical Association* 88: 1242–1249.

Lehmann, E. L. (2011). *Fisher, Neyman, and the Creation of Classical Statistics*. Springer.

Lindley, D. V. (1957). A Statistical Paradox. *Biometrika* 44: 187–192.

McShane, B. B., Gal, D., Gelman, A., Robert C. and Tackett, J. L. (2019). Abandon Statistical Significance. *The American Statistician* 73(S1): 235–245.

Morrison, D. E. and Henkel, R. E. (1970). *The Significant Test Controversy: A Reader*. New Brunswick.

Neyman, J. (1950). *The First Course in Probability and Statistics*. Holt.

Neyman, J. (1957). “Inductive Behavior” as a Basic Concept of Philosophy of Science. *Revue de l’Institut International de Statistique* 25: 7–22.

Neyman, J. (1961). Silver Jubilee of My Dispute with Fisher. *Journal of the Operations Research Society of Japan* 3: 145–154.

Neyman, J. and Pearson, E. (1928). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I. *Biometrika* 20A: 175–240.

Neyman, J. and Pearson, E. (1933a). On the Problem of the Most Efficient Tests of Statistical Hypothesis. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231: 289–337.

Neyman, J. and Pearson, E. (1933b). The Testing of Statistical Hypotheses in Relation to Probabilities a Priori. *Mathematical Proceedings of the Cambridge Philosophical Society* 24: 492–510.

Open Science Collaboration (2015). Estimating the Reproducibility of Psychological Science.
Pearson, E. (1955). Statistical Concepts in the Relation to Reality. *Journal of the Royal Statistical Society* B17: 204–207.
Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A Tutorial for Teaching Data Testing. *Frontiers in Psychology* 6: 1–11.
Popper, K. (1959). *The Logic of Scientific Discovery*. Hutchinson.
Sober, E. (2008). *Evidence and Evolution*. Cambridge University Press.
Spielman, S. (1974). The Logic of Tests of Significance, *Philosophy of Science* 4(3): 211–226.
Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific Misconduct and the Myth of Self-Correction in Science. *Perspectives on Psychological Science* 7(6): 670–688.
Trafimow, D. and Marks, M. (2015). Editorial, *Basic and Applied Social Psychology* 37: 1–2.
Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011). *Journal of Personality and Social Psychology* 100: 426–432.
Wasserstein, R. and Lazar, N. (2016). The ASA’s Statement on p-Values: Context, Process and Purpose. *The American Statistician* 70(2): 129–133.
Wasserstein, R., Schirn, A., and Lazar, N. (2019). Editorial. Moving to a World Beyond “p < 0.05.” *The American Statistician* 73(S1): 1–19.

(Received 2021.2.10; Revised 2021.7.4; Accepted 2021.7.10)