Filomat 39:34 (2025), 12327–12343
https://doi.org/10.2298/FIL2534327A

Published by Faculty of Sciences and Mathematics,
University of Niš, Serbia
Available at: http://www.pmf.ni.ac.rs/filomat

Reevaluating p-value and its impact on conventional results display and interpretation

Atif Avdović[a,*], **Zoran Vidović**[b]

[a] *Department of natural sciences and mathematics, State university of Novi Pazar, Serbia*
[b] *Faculty of Education, University of Belgrade, Serbia*

Abstract. There are numerous instances of misunderstanding, misinterpreting, and misusing p-values, which have raised concerns about the reliability of statistical conclusions based on them. This paper provides a comprehensive reevaluation of the mathematical foundations of the p-value, discusses common misconceptions in its teaching and application, and explores alternatives such as confidence intervals, effect size, and Bayes' factor. Additionally, it highlights scenarios where the p-value remains valuable beyond hypothesis testing. The goal is to emphasize the continued relevance of the p-value while acknowledging potential pitfalls in its misuse. Given that most of the discussion is rooted in applied statistics, this paper aims to offer clarity to researchers across various disciplines.

1. Introduction

Hypothesis testing remains one of the most widely used statistical techniques in research across diverse disciplines [34]. Since the early 20th century, p-values have been a fundamental part of statistical inference, serving as a measure of how compatible observed data are with a given null hypothesis. By quantifying the likelihood that observed results could occur under the assumption of no real effect, p-values play a crucial role in decision-making and empirical analysis [13].

Initially introduced by Ronald Fisher, the p-value was conceptualized as the probability that an observed effect is due to chance. Fisher also proposed 0.05 as a conventional threshold for statistical significance [13]. While originally intended for specific experimental settings, this threshold has been generalized across disciplines, often without sufficient critical evaluation [31, pp. 279-280], [52, p. 8], [57].

There are many researches about p-value from the perspective of medicine and psychology rather than general or any other perspective [13, 22, 24, 34, 45]. Mathematicians do not seem to address this as a problem. Even in few sources that deal with p-value mathematically, some of them are in other disciplines [63]. This is probably due to clarity of the conclusion obtaining that p-value gives theoretically ($p > \alpha$ retain H_0, and reject otherwise). When it comes to nature, context and the sensibility needed in interpreting

2020 *Mathematics Subject Classification.* Primary 62-00; Secondary 62C05, 62C12, 62F03, 62G10.
Keywords. p-value, significance, effect size, confidence intervals, Bayes' factor.
Received: 18 June 2025; Accepted: 22 October 2025
Communicated by Aleksandar Nastić

* Corresponding author: Atif Avdović
Email addresses: aavdovic@np.ac.rs (Atif Avdović), zoran.vidovic@uf.bg.ac.rs (Zoran Vidović)
ORCID iDs: https://orcid.org/0000-0001-8672-1607 (Atif Avdović), https://orcid.org/0000-0002-6076-7073 (Zoran Vidović)

p-value, that might not be the case [8, 25]. Many authors even discuss the problems of the information sufficiency or result reliability when it is acquired using p-value [24].

One major challenge is that p-values are frequently misunderstood [60]. A common misconception is equating statistical significance with practical significance, leading researchers to overemphasize results based on arbitrary cutoffs [43]. This dichotomous approach to inference—where results are deemed either significant or not based solely on whether $p < 0.05$—can obscure the nuances of statistical reasoning.

Moreover, concerns have been raised regarding the misuse of p-values in scientific research. Practices such as p-hacking and selective reporting, where researchers manipulate data or conduct multiple tests until significant results are obtained, compromise the integrity of statistical conclusions [33, 34]. Note that these practices do not constitute p-value issue, but the publication bias issue (nonsignificant results are usually not published) [54]. Misinterpretations of p-values can have real-world consequences, particularly in fields like medicine and psychology, where research findings directly influence clinical decisions and policy-making [25].

On the other hand, misunderstanding p-values seems to be identified with its utility and interpretability [22]. This has led to p-value even being discarded as relevant inference method and being replaced by some of its possible alternatives [26, 34, 54]. Despite critique and suggestions that p-value should be replaced with alternative methods, modern papers still use p-value as if issues have not been raised.

Some papers [8, 25] indicate that p-value is a valuable information source and is not difficult to interpret once understood correctly. Also, misunderstanding or misinterpreting p-value is not always the reason for wrong results. P-value, as any other statistical method, offers conclusions as correct as the sample observed is representative of population, or as the test used is powerful [10, 15, 34, 55]. Some argue that p-value inherits perks and flaws of the test it refers to [54].

Though there have been many suggestions on emerged problems remedies – many of which have been given in majority of articles and other publications this paper refers to – it seems that not much success has been accomplished in terms of application [23, 29, 44]. Despite these challenges, p-value remains an essential tool in statistical inference. This paper seeks to clarify the role of the p-value, address common misconceptions, and explore how it can be effectively integrated with alternative statistical techniques.

In section 2 there is elaboration on mathematics of p-value that might help proper understanding of p-value and improve on its teaching when necessary. In section 3, important teaching and methodology problems that are not usually found in other researches are addressed with guidelines. In section 4, usually suggested alternatives of p-value are addressed by implying on their advantages and disadvantages. Finally, in section 5 other (than hypothesis testing) applications of p-value are given as additional motivation for p-value not to be dismissed but properly worked on. Concluding remarks are also given.

2. Mathematics behind the p-value

The null hypothesis is often introduced via following formulations.

Test statistic makes an estimate of an underlying variable's parameter. For this reason, some authors refer to a parameter in the parameters space when using hypothesis generalization [64, p. 149]. Suppose that the parameter space Θ is divided into two disjoint sets Θ_0 and Θ_1, and that statement $H_0 : \theta \in \Theta_0$ is to be tested versus $H_1 : \theta \in \Theta_1$. H_0 is referred to as the null hypothesis and H_1 as the alternative hypothesis [31, p. 267], [64, p. 149].

Let $X_1, \ldots, X_n$ be the variables for which the hypothesis H_0: "Variables $X_1, \ldots, X_n$ satisfy the property $S(X_1, \ldots, X_n)$." is to be tested. Let

$$\begin{aligned}
&(X_{11}, X_{12}, ..., X_{1m_1}) = \vec{X}_1(m_1)\\
&(X_{21}, X_{22}, ..., X_{2m_2}) = \vec{X}_2(m_2)\\
&\vdots\\
&(X_{n1}, X_{n2}, ..., X_{nm_n}) = \vec{X}_n(m_n)
\end{aligned} \tag{1}$$

be the sample the testing procedure is to be performed on, T the function that determines test statistic of chosen test and

$$W = \left\{ T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), ..., \vec{X}_n(m_n)\right) \mid S(X_1, ..., X_n) \text{ does not hold} \right\} \tag{2}$$

the critical region. If W is such that

$$\alpha = P_{S(X_1,...,X_n)}\left(T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), ..., \vec{X}_n(m_n)\right) \in W \right) \tag{3}$$

for given level of significance α, then denote $W = W_\alpha$.

Definition 2.1. *P-value of the test determined by T is given with*

$$p = \inf\left\{ \alpha \in (0,1) \mid T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), ..., \vec{X}_n(m_n)\right) \in W_\alpha \right\}. \tag{4}$$

Definition 2.1 can be formulated as follows.

"p-value is the lowest level of significance for which the test determined by T rejects the null hypothesis". (5)

Definition 2.1 implies that for any $\alpha \geq p$ the test will reject the null hypothesis, and retain it otherwise. This is usually used "rule of thumb" for interpreting the p-value when concluding the test results. While using it, it is important to emphasize that the obtained conclusion is given for the level of significance α. In other words, to emphasize that the conclusion does not necessarily hold for sample (1) in general. Another implication from definition 2.1 is that p-value is proportionate to the quantity of the evidence supporting the null hypothesis.

It is very useful for authors and teachers that deal with statistics to illustrate the phenomenon of (4) with the Figure 1 as proposed by Wasserman [64, p. 157].

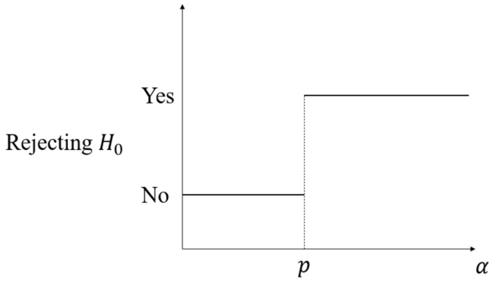

Figure 1. p-value definition illustration.

Based on Definition 2.1 (or (4)) and Figure 1, p-value definition can be formulated as follows.

Definition 2.2. *P-value of the test determined by T is given with*

$$p = \sup\left\{ \alpha \in (0,1) \mid T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), ..., \vec{X}_n(m_n)\right) \notin W_\alpha \right\}. \tag{6}$$

Definition 2.2 can be formulated as follows.

"p-value is the highest level of significance for which the test determined by T retains the null hypothesis". (7)

This understanding of p-value comes from the following preliminary explanation, which is followed by the theorem that will be discussed later.

For statistic T and given level of significance α, critical region W given with (2) can have one of the following forms depending on the alternative hypothesis.

CR_1. $W_\alpha = (-\infty, c]$, $c \in \mathbb{R}$;

CR_2. $W_\alpha = [c, +\infty)$, $c \in \mathbb{R}$;

CR_3. $W_\alpha = (-\infty, c_1] \cup [c_2, +\infty)$, $c_1, c_2 \in \mathbb{R}$;

CR_4. $W_\alpha = (-\infty, c] \cup [c, +\infty)$, $c \in \mathbb{R}$, when the distribution of T is symmetrical.

When $|T|$ is bounded from above, then the critical region bounds are all finite numbers replacing infinities in CR_1-CR_4 where needed. These cases are not treated separately due to the p-value being calculated in the same way.

Theorem 2.3. *P-value for referent critical region is calculated with following formulas.*

$$\begin{aligned}
p_{CR_1} &= \sup_{S(X_1,\ldots,X_n)} P\left(T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), \ldots, \vec{X}_n(m_n)\right) \le T\left(\vec{x}_1(m_1), \vec{x}_2(m_2), \ldots, \vec{x}_n(m_n)\right)\right);\\
p_{CR_2} &= \sup_{S(X_1,\ldots,X_n)} P\left(T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), \ldots, \vec{X}_n(m_n)\right) \ge T\left(\vec{x}_1(m_1), \vec{x}_2(m_2), \ldots, \vec{x}_n(m_n)\right)\right);\\
p_{CR_3} &= 2\min\left\{\sup_{S(X_1,\ldots,X_n)} P\left(T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), \ldots, \vec{X}_n(m_n)\right) \underbrace{\le}_{\ge} T\left(\vec{x}_1(m_1), \vec{x}_2(m_2), \ldots, \vec{x}_n(m_n)\right)\right)\right\};\\
p_{CR_4} &= \sup_{S(X_1,\ldots,X_n)} P\left(T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), \ldots, \vec{X}_n(m_n)\right) \ge \left|T\left(\vec{x}_1(m_1), \vec{x}_2(m_2), \ldots, \vec{x}_n(m_n)\right)\right|\right).
\end{aligned}$$

This theorem can be stated as:
"p-value is calculated as the highest probability that the sample $\vec{X}_1(m_1), \vec{X}_2(m_2), \ldots, \vec{X}_n(m_n)$ differs from the null hypothesis state in proportion to the evidence supporting its realization $\vec{x}_1(m_1), \vec{x}_2(m_2), \ldots, \vec{x}_n(m_n)$".

Proof. The proof is given for the case CR_1. $W_\alpha = (-\infty, c]$, $c \in \mathbb{R}$. For other cases it is done analogously.

Based on the Definition 2.1, the goal is to calculate the smallest α for which the null hypothesis will be rejected, given the assumption "H_0 is rejected if and only if $T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), \ldots, \vec{X}_n(m_n)\right) \le c$". Since

$$\alpha = \sup_{S(X_1,\ldots,X_n)} P\left(T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), \ldots, \vec{X}_n(m_n)\right) \le c\right)$$

is decreasing function for c as a function variable, the problem is equivalent to calculating the highest c for which the null hypothesis will be rejected, given the same assumption. Obviously, the highest such a value is $c = T(\vec{x}_1(m_1), \vec{x}_2(m_2), \ldots, \vec{x}_n(m_n))$. Hence, required level of significance is

$$\alpha = \sup_{S(X_1,\ldots,X_n)} P\left(T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), \ldots, \vec{X}_n(m_n)\right) \le T\left(\vec{x}_1(m_1), \vec{x}_2(m_2), \ldots, \vec{x}_n(m_n)\right)\right).$$

Then,

$$\begin{aligned}
p &= \inf\left\{\alpha \in (0,1) \mid T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), \ldots, \vec{X}_n(m_n)\right) \in W_\alpha\right\}\\
&= \sup_{S(X_1,\ldots,X_n)} P\left(T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), \ldots, \vec{X}_n(m_n)\right) \le T\left(\vec{x}_1(m_1), \vec{x}_2(m_2), \ldots, \vec{x}_n(m_n)\right)\right).
\end{aligned}$$

□

Several other interpretations of the p-value implied by Theorem 2.3 can be found in the literature, even though (8) (given later in the text) is the interpretation that is most frequently used [24].

As argued by Neumann and Pearson, Theorem 2.3 clearly cannot be used to derive the real p-value [52, p. 8]. This is because the obtained effect size does not necessarily have to be the most extreme one. Therefore, Theorem 2.3 is only a method for estimating the p-value for the acquired sample $\vec{x}_1(m_1), \vec{x}_2(m_2), ..., \vec{x}_n(m_n)$. The formulas that follow show that estimate

$$p_{CR_1} \approx P\left(T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), ..., \vec{X}_n(m_n)\right) \leq T\left(\vec{x}_1(m_1), \vec{x}_2(m_2), ..., \vec{x}_n(m_n)\right)\right);$$
$$p_{CR_2} \approx P\left(T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), ..., \vec{X}_n(m_n)\right) \geq T\left(\vec{x}_1(m_1), \vec{x}_2(m_2), ..., \vec{x}_n(m_n)\right)\right);$$
$$p_{CR_3} \approx 2\min\left\{P\left(T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), ..., \vec{X}_n(m_n)\right) \underbrace{\leq}_{\geq} T\left(\vec{x}_1(m_1), \vec{x}_2(m_2), ..., \vec{x}_n(m_n)\right)\right)\right\};$$
$$p_{CR_4} \approx P\left(T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), ..., \vec{X}_n(m_n)\right) \geq \left|T\left(\vec{x}_1(m_1), \vec{x}_2(m_2), ..., \vec{x}_n(m_n)\right)\right|\right).$$

As the p-value obtained using those formulas is merely an approximation, the conclusion that follows is dependent only on how reliable the approximation is.

Theorem 2.4. *Let the distribution of statistic T be continuous one. If H_0 is true, then the p-value has a uniform distribution on interval* $(0,1)$.

Proof. P-value is the lowest level of significance for which the test determined by T rejects null hypothesis. That definition, when H_0 holds and T is continuous, implies

$$p = 1 - F_T\left(T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), ..., \vec{X}_n(m_n)\right)\right),$$

where F_T denotes the cumulative distribution function of T. Since $F_T\left(T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), ..., \vec{X}_n(m_n)\right)\right)$ is uniformly distributed on $(0,1)$ so is p [63]. □

A consequence of this theorem is that if the alternative hypothesis is true, then p-value tends to concentrate closer to 0 [64, p. 158].

Remark 2.5. *Theorem 2.4 requires continuous distribution of T. However, if that is not the case, distribution of p could be considered as discrete one where the probabilities of all modalities are equal. Therefore, it does not affect this approach in applicability of p-value.*

3. Teaching and methodology of p-value – Problems and guidelines

Greenland et al. [24] have produced a very impactful report that conveniently lists many errors and inaccurate definitions utilized and mentioned across textbooks and research journals. This paper deals with some errors and misconceptions, or problems concerning p-value in general, that have so far not been dealt with in published studies. Though there were many papers published that aimed to remedy problems that arose, few of them have been actively applied since. Also, ASA (American Statistical Association) published statement regarding proper understanding and interpretation of p-value [65]. This statement has mostly been received well since it is simply formulated, does not discard p-values and addresses most of the raised issues. Principles included in ASA's statement are:

1. "P-values can indicate how incompatible the data are with a specified statistical model"
2. "P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone"
3. "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold"
4. "Proper inference requires full reporting and transparency"

5. "A p-value, or statistical significance, does not measure the size of an effect or the importance of a result"

6. "By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis".

There have however been critiques of that statement, one of which is that it contains guidelines on what not to do rather than what to do [12]. Wellek [67] strongly criticizes the 6th principle of the statement and questions the consensus on the precise meaning of concept of evidence, i.e. what is a good measure of evidence and what is a bad one.

3.1. *Using parameter in conceptualizing hypothesis*

Several authors supplement the introduction of null hypothesis by considering a family of distributions that can be assigned to the test variables [51]. While each variable has a distribution, the introduction of the null hypothesis does not require information on that distribution. The fact that some underlying variable distributions are assumed just for parametric hypotheses may potentially cause confusion among students. The confusion is such that they fail to understand the distinction between the parametric and nonparametric hypothesis. This is especially the case for students in subjects unrelated to mathematics, such as social sciences and humanities [19, 24, 36, 61].

3.2. *Improper defining of the p-value*

A widespread issue in statistical education is the inaccurate definition of the p-value. Many textbooks and instructional materials describe it with the following statement [24, 40, 54]; [9, p. 76], [16, p. 140], [52, p. 8], [58, p. 200].

> *"p-value is probability of obtaining the test statistic value that is as extreme, or more extreme, in terms of alternative hypothesis, than its realized value."* (8)

Statement (8) is alternative formulation of Theorem 2.3. It is even found in the ASA statement [65], though with hedging using word "informally".

Both definitions include the verbal formulation given in the statements (5) or (7), as well as the mathematical formulation indicated by (4) or (6). Both formulations are necessary for statisticians and mathematicians to be proficient in instructional writing or teaching [49]. Since they are better at mathematics, this method works well in related subjects including engineering, computer science, physics, and some other sciences [27, 53].

According to Gibbs [20], statistical analysis in the social sciences has not consistently shown to be effective. This highlights the need for an understanding and use of mathematics in the model. Consequently, a preferable teaching style is to combine written instruction with vocal instruction [14]. The main conclusions based on data used in social science and related disciplines are obtained through p-value interpretation, which makes this important [34].

3.3. *Statistical significance*

Another pedagogical issue is the rigid application of significance thresholds. The widespread use of $\alpha = 0.05$ as a universal cutoff has led to oversimplified interpretation of statistical results. A more nuanced approach should be encouraged, where researchers consider effect sizes, confidence intervals, and study context when drawing conclusions [24].

When testing at the significance level of 0.05, Wasserman [64, p. 156] suggests applying the following rule of interpretation to make sure the conclusion is as correct as possible given the information that is available based on the estimated p-value. Note that this holds only in case when researcher is certain that this threshold satisfies the context and sensitivity of the topic. There have been many instances of lower level of significance (e.g. $\alpha = 0.005$) being recommended [54].

Instead of treating statistical significance as a binary outcome, instructors should emphasize the concept of *statistical compatibility*, i.e. the degree to which observed data align with the null hypothesis. This

Table 3.1. Interpretation of p-value.

p-value	Evidence for rejecting H_0	Evidence for retaining H_0
$p < 0.01$	Very strong	Weak or no evidence
$0.01 \leq p < 0.05$	Strong	Weak
$0.05 \leq p < 0.1$	Weak	Strong
$p \geq 0.1$	Weak or no evidence	Very strong

approach shifts the focus from arbitrary cutoffs to a broader understanding of uncertainty in statistical inference [50].

In general, one should not unambiguously conclude that the null hypothesis is rejected. It is appropriate to use the term ″the effect is statistically significant″ rather than "the null hypothesis is rejected". The reason for that is the fact that test-based conclusion depends on data quality, data size, power of the test and chosen level of significance.

Substantial effects are detected only with large data sets. Power of chosen test partially depends on sample size [4]. That indicates that interpretation based on the p-value should take the sample size into account as well [10, pp. 6-7], [55].

The estimate of p-value is calculated based on the test statistic properties. Hence, it will yield a conclusion whose reliability depends on the power of the test. Though there is no direct relation between the p-value and power calculation, the relation between them can be manifested so that, for instance, power can be approximately calculated using p-values [69]. This is additionally elaborated on in section 6.

3.4. *Context of the research*

The effect that is statistically significant does not always indicate that the effect is considered noteworthy or meaningful. Even when an effect is genuinely statistically significant, authors debate whether it matters for the referent research if many other preconditions are not verified ([8]; [16, pp. 139-140]). Context of the research is very important to be taken into account [8], [38]. Cabrera and McDougall [9, pp. 206-207] claim that significance in contingence tables analysis testing is properly interpreted if the following issues are addressed.

1. When can the effect be considered meaningful?
2. What are the confounding factors of the analysis?
3. Are all the variables essential for research included in the analysis?
4. Is data censored?
5. What is the type of study?
6. Is data size equal across all eventual categories?
7. How sensitive is the analysis to small changes in sample elements?
8. If re-coding occurred, how is it done?

This approach can be generalized for all tests in order to properly consider the context of the research.

Stating that detected effect is (not) likely to be significant or that it is plausible for that (not) to be the case is an example of hedging when drawing conclusions based on hypothesis testing for a particular level of significance, with or without the p-value [40]. This way, when error occurs, the researcher has indicated possibility for that to happen.

3.5. Misuse of examples in teaching

It is common knowledge that using examples in instruction and textbook writing helps students grasp the material better [66]. Furthermore, students frequently find it difficult to understand theoretical notions unless they are applied or demonstrated through an example, especially students in disciplines that are unrelated to mathematics [68]. If there is additional time during instruction or room in the textbook for examples to be given, it can be quite helpful [11].

The purpose of the examples should be result illustration and student's feedback. Examples should thus be used as an additional teaching method. If used to formally introduce concepts, it should be along with proper theoretical introduction. This is due to examples illustrating only a finite number (usually one) of cases [68]. As such, introducing some new concept loses its generality. However, p-value is frequently defined using examples without theoretical definition.

The most often case of using an example instead of formal definition of p-value is the one testing the hypothesis of equal means of two dependent samples [31, pp. 279-280], [52, p. 8], [57]. Such an introducing usually resembles (8) or the verbal interpretation of Theorem 2.3. In that case, the variety of hypotheses that might have different formulation are neglected. For instance, goodness of fit, independence, outlier discrepancy, sample adequacy, sphericity, etc. Examples usually used are good for motivational approach to Theorem 2.3. It should be noted that Theorem 2.3 should also be stated in its general form, at least by (8).

As mentioned before, the p-value obtained when testing hypothesis based on some data is only estimate of real p-value calculated based on Theorem 2.3. As such, there is no proper example that would illustrate factual definition of p-value. However, that should not be an issue since definition of p-value given with (4), and especially with (5), is rather clear and easy to understand.

3.6. Lack of graphical aids

Visualizing statistical concepts can significantly enhance understanding [7], particularly for students in social sciences and biomedical fields [5]. As will be further discussed, the relationship between the p-value and the observed significance of the discovered effect is more complex and calls for a more sensitive methodology [8, 25, 32].

When introducing the p-value, educators should supplement theoretical explanations with graphical representations, such as:

I Figure 1: A visual illustration of the p-value in the context of hypothesis testing.

II Figure 2: Depictions of confidence intervals demonstrating strong vs. weak evidence.

III Figure 3: The relationship between effect size, critical regions, and p-value.

By integrating these visual tools into teaching materials, instructors can reduce common misunderstandings and promote a more intuitive grasp of statistical significance. Note that we have found no textbooks in statistics that has treated p-value introduction and further explanation using graphical aids.

Figures 2 and 3 are displayed in the following sections.

3.7. Publication bias, p-hacking, reproducibility and dance of the p-values

These issues are usually treated as separate problems [34, 54, 67]. However, we shall argue that these are problems that are connected and sometimes caused by each other.

In modern research endeavors journals with impact factor, especially high one, are priority for researchers trying to publish their work. Most common reason for that are conditions set by state laws, universities and other research institutions, that require for researcher to publish one or more papers in SCI list journals in order to being able to keep their job or to get promoted.

As they should, journals encourage researchers to make their work as high quality as possible. That usually means high sample sizes, thorough reviews of existing related literature, expensive equipment and materials etc., all of which require long time to obtain. Occasionally, such a research process ends in failure known as "insignificant results", i.e. results that do not support research hypotheses, are not lucrative, don't motivate new research etc. Though publishing such results can be useful – for instance to save time of other researches that might try do get to the same discovery – most of the journals reject publishing them. That leads researchers to resort to unethical practices such as p-hacking causing their research not to be reproducible. There are many examples of this in academia [30, 33].

These problems are often associated with the property of p-value given with Theorem 2.4, often referred to as "dance of the p-values" [54]. Randomness (uniformity of distribution) of CDF – thus the randomness of sampling (precondition for representative sample) – causes p-values to "dance" in interval $[0, 1]$. However, in that case only $\alpha \cdot 100\%$ of p-values will be lower than or equal to α (thus report type I error) if the null hypothesis is true. Also, the sample that causes p-value to report type I error will likely cause any other method to report it as well. For example, any randomness test – as well as any other statistical tool – will report sample $9, 9, 9, 9, 9, 9, 9$ as non-random, but it's theoretically possible for random number generator to generate this sample. Then the p-value of some randomness test will be lower than α, though the null hypothesis might be true.

The information based on hypothesis testing is reliable if the sample represents population (it does not depend on p-value). The same holds for other methods. To ensure that conclusion is good, the same result should be replicated several times without discarding replications that do not support what researcher "needs". Research performed this way will almost certainly be reproducible. This is important because if the error occurs it is usually the fault of the sample rather than of p-value.

3.8. *General guidelines*

Some recommendations for solving these issues are proper teaching, longer presentation on p-value that might take a repetition to be successful along with several examples and detailed studying material with proper content as given in this paper. These steps will induce students gaining capability of adjusting the definition to any example and circumstance, perform and illustrate good calculation of it using Theorem 2.3 and thus understand the difference between the calculation procedure and the definition. Such an approach can result in correct and precise interpretation of result indicated by p-value with regards to context and research type that is not "pattern like" [8, 25, 32]).

Stating "statistical" significance (or its absence) with the level of significance when interpreting testing results can be essential for proper conclusion. Namely, significance (or its absence) being statistical indicates its dependence on data quality, power of the test and sample size [55]. Level of significance compared to p-value indicates how strong the evidence for conclusion is.

P-value used to quantify the discrepancy from the null hypothesis property can be clearer for researchers or students who are not mathematics majors. Unnecessary display of extra results of statistical analysis does not always contribute to information required nor does it improve the quality of results presentation, and thus the research. What it does contribute to is expanding the text concerning results and discussion in articles and making it appear more complex and informative.

P-value deficiencies impact can be reduced by programs like open data sharing and preregistration of study protocols, which are meant to increase transparency and rigor in scientific reporting.

4. Alternative to p-value

4.1. *Confidence intervals*

Determining the algorithm for confidence interval (CI) bounds estimation is not always simple, or even possible [28]. In that cases CI bounds are obtained via trimming Monte Carlo or bootstrap modelled sample, or are put in tables that can be used [17].

Let $\theta = \theta(X) \in \mathbb{R}$ be the parameter estimated by test statistic T. The best case is that

$$\lim_{n, m_i \to +\infty} T\left(\vec{X}_1(m_1), \vec{X}_2(m_2), ..., \vec{X}_n(m_n)\right) = \theta.$$

Let $[\theta_L(n, m_1, \ldots, m_n, 1-\alpha), \theta_R(n, m_1, \ldots, m_n, 1-\alpha)]$ be the $(1-\alpha) \cdot 100\%$ CI for θ. The hypothesis H_0: "Variables $X_1, \ldots, X_n$ satisfy the property $S(X_1, \ldots, X_n)$." is equivalent to $H_0 : \theta = \theta_0$.

Confidence levels (significance levels) are correlated with CIs through their width. Consequently, one needs to obtain separate CIs for each α. The p-value calculation methodology is usually the most accessible and user-friendly approach. This is among the factors contributing to p-value applications' dominance. P-values are calculated only once and allow for simple comparison to draw conclusions for all α.

It is discussed that the information obtained by CI is equal or superior to that obtained by p-value [41]. This kind of critical approach to p-value is rather incomplete and frequently incorrect. Namely, while both p-value and confidence interval result in conclusions that are equivalent, there are some that are characteristic for each one and not the other.

P-value and CIs for the parameter that establishes the null hypothesis can both show strong evidence when it comes to the indication of retaining the null hypothesis significance [32, 40]. If $p \geq 0.1$, then the CI bounds are approximately equally distanced from the null hypothesis parameter value. In that case, sample information gives strong evidence that the null hypothesis should be retained. Moreover, there is an advantage in using a p-values in this case. This is because the $p \geq 0.1$ guarantees that the evidence for retaining the null hypothesis is strong, whilst that does not have to be the case when the CI bounds are on the opposite sides of the null hypothesis parameter value. Such bounds of CIs directly guarantee $p > \alpha$, but the interpretation on whether the evidence for retaining is strong or not can be relative. Additionally, the condition "CI bounds are approximately equally distanced from the null hypothesis parameter value" can be both correct or incorrect, depending on data. That condition gives conclusion that can be ambiguous when bounds of CI have small absolute values.

The following figure illustrates the Table 3.1 cases when the conclusion is obtained via CIs.

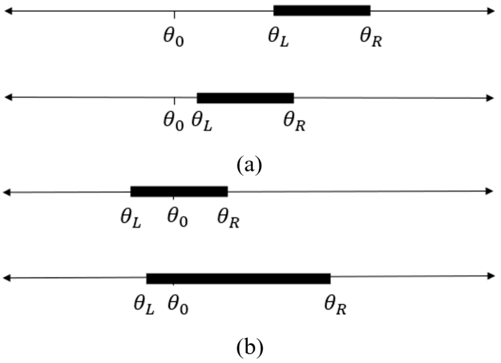

Figure 2. Strong and weak evidence for (a) rejecting null hypothesis based on CI; (b) retaining null hypothesis based on CI.

4.2. Effect size

Effect size (ES) is the realized value of the test statistic. Nearly all studies presenting results from testing hypotheses show the ES, s-value, p-value, and, when relevant, the CIs [25, 50].

The obvious uses of ES are to calculate the p-value and compare the discrepancy from the null hypothesis property S for more samples obtained under various conditions, or even when the underlying variables are measured differently [1]. If hypothesis is tested only for one sample, is it still necessary to present the ES in addition to the p-value? This seems to be an unanswered question or one that the solution to is not fully understood.

It is explained that a proficient statistician can determine the degree of discrepancy with the null hypothesis or the conclusion certainty by analyzing merely the ES [42]. Cohen's d as an effect size is useful as both supplementary to p-value, or displayed by itself. It even provides good information on whether to accept or reject hypothesis and with what certainty to do so, but is only applicable in several parametric tests used [10, 42]. Therefore, it is claimed that the combination of the ES and p-value can reveal more information than either one by itself [57].

Let $t = T\,(\vec{x}_1(m_1), \vec{x}_2(m_2), ..., \vec{x}_n(m_n))$ be the ES. The following figures illustrate in what relation are p-value and critical region in various cases.

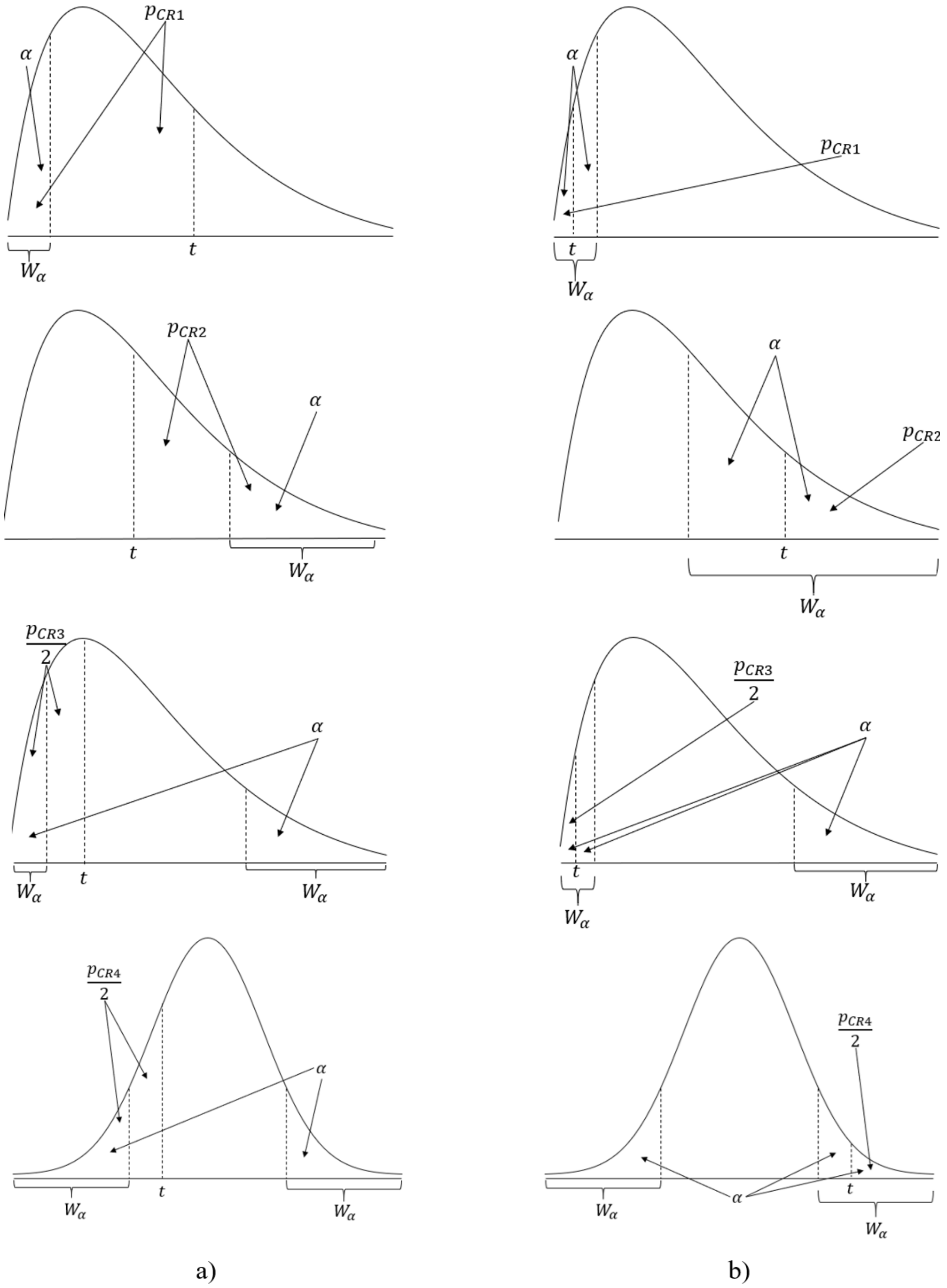

Figure 3. p-value and CR in all discussed cases; a) retaining H_0; b) rejecting H_0.

However, p-value gives more definite conclusion making available by comparison with the level of significance. In case of the ES, critical region (CR) bounds, and thereby more complex calculations, are required. Additionally, the CR depends on the level of significance. Hence, conclusion based on the ES is not definitive for other level of significance.

Since p-value depends on the ES, it can be informative about ES as well. Referring to the p-value as compatibility measure is as effective when properly interpreted [50]. This is possible due to p-value being the highest level of significance for which the test retains the null hypothesis, i.e. for which the sample is considered not to deviate significantly from the null hypothesis property.

Graphical illustrations are known to be very important educational tool [2]. In case of p-value or concluding about the observed significance, no more illustrations than those given with figures resembling 1 and 2 are found. ES determines p-value and CR determines the conclusion based on the ES. Thus, both ES and CR are essential for understanding p-value.

Note that graphics given by Figure 3 are useful for illustrating purposes and can improve learner's understanding of all concepts it refers to, thus the p-value as well. However, though it might be helpful even in interpreting some results, it would not be more informative than the actual values and is thus unnecessary in that case.

Graphs given with Figure 3 are of great importance for understanding Theorem 2.3. These graphs clearly display how does T affect the CR. More precisely, the figure itself illustrates why does larger discrepancy causes lower p-value. Along with the definition, it also explains why is the comparison to α important.

In Figure 3, the curve by which all the hypothesis testing concepts are considered is probability density function of test statistic distribution. In the unlikely event that the distribution is discrete [4], a similar method can be used with probability mass function.

4.3. *Bayes' factor*

Many researchers are prone to replacing p-values with other statistical methods that could possibly be as informative, such as Bayes' factor (BF) [22], introduced by Jeffreys [35].

Let D be the random event "observed data (1) is obtained from variables $X_1, \dots, X_n$". BF for hypothesis H_0: "Variables $X_1, \dots, X_n$ satisfy the property $S(X_1, \dots, X_n)$." is given by the following definition.

Definition 4.1. *BF based on data (1) is calculated by*

$$BF = \frac{P(D|H_0)}{P(D|H_1)}. \tag{9}$$

Essentially, BF also provides with dichotomous rule of accepting or rejecting the null hypothesis. BF equal to 1 indicates that there is no certain evidence for retaining H_0 compared to H_1. BF greater than 1 indicates evidence for retaining H_0 compared to H_1. BF lower than 1 indicates evidence against H_0 compared to H_1. Similar as with the ES, that rule is based on some acceptance/rejection region. Some alternative more sophisticated rules for interpretation can be used for both p-value and BF [48]. Using the analogy with the Table 3.1 for p-value, following table, Table 4.1 is obtained. Similar, though a bit more complicated is suggested by Kass and Raftery [39].

Table 4.1. Interpretation of BF.

BF	Evidence for rejecting H_0	Evidence for retaining H_0
$BF < 0.5$	Very strong	Weak or no evidence
$0.5 \leq BF < 1$	Strong	Weak
$1 \leq BF < 2$	Weak	Strong
$BF \geq 2$	Weak or no evidence	Very strong

Depending on the context of the research the criteria can be made more sophisticated in terms of more intervals for BF interpretation. BF given with (9) is often denoted with BF_{12}. When $H_1 = H_0^{\complement}$, BF is denoted as BF_{01}. Value $BF_{10} = \frac{1}{BF_{01}}$ is often used and referred to as BF. This BF variant is interpreted in the way opposite to one for BF_{01} [18]. The same holds for $BF_{21} = \frac{1}{BF_{12}}$.

Though there are certain methods for calculation of BF such as BF functions, these calculations are restricted only to some conventionally used test statistics [37]. As can be seen, the calculation of the Bayes factor can be very difficult or even impossible for many hypotheses [39].

During BF calculation, both null and alternative hypothesis are considered as well. During p-value calculation only null hypothesis is considered. This indicates that when both are available, BF gives conclusion closer to true information. However, there are several issues of BF that can make researcher consider it inferior to p-value in terms of choosing one over other when testing hypotheses.

Null, as well as alternative hypothesis can be one of many simple and compound hypotheses. Generally, cases where null and alternative hypotheses are such that both $P(D|H_0)$ and $P(D|H_1)$ can directly be calculated are extremely rare [48]. Hypothesis can be such that calculating $P(D|H_i)$, $i = 1, 2$ is possible only using simulations [39]. Simulations used are usually Monte Carlo simulations, but in some cases bootstrap or other type of simulations must be used [6].

As an advantage of BF relative to p-value is that BF does not need to be determined by some test chosen. However, in many cases calculating $P(D|H_0)$ or $P(D|H_1)$ is complicated or impossible [39]. Namely, for that to be done the information on the distribution of X is required. Also, it would probably be easier (or possible) only to approximate BF by Monte Carlo simulations.

If D is to be approximated using some statistic T, then it can be more accessible. For BF to be used in hypothesis testing, it needs to be applied via test, say the one determined by T. Event D is to be formulated so that, if possible, it is completely determined by calculations using T and its properties. Since the ES t is realization of T for data in D as argument, D can be stated as $D : T = t$. However, the distribution of T is often continuous, so for $D : T = t$, both probabilities figuring in BF are 0. The alternate formulation $D : T \leq t$ or $D : T \geq t$, supporting the H_0, is thus better. This does not guarantee properly calculated BF. D formulated by T means only that observed data results in $T = t$ which can be case for other data sets as well. In other words, D cannot always be completely determined by calculations using T and its properties.

Statisticians usually use Bayes' factor while ignoring the context-dependent specifications of research [56]. It has also been shown that p-values tend to be more sensitive relative to BF, though conclusions are mainly the same [3]. One way of improving hypothesis testing conclusions – if one's aim is to minimize type I error occurring – is thus to use p-values for small and moderate size samples, while BF might be better choice for large size samples when calculated directly, i.e. not depending on T [39]. For large sample sizes, test statistics are usually more sensitive to small effects making both p-values and BFs calculated by them also more sensitive to small effects. Then, researcher should be aware that type II error is also a possibility [62]. Namely, significance level should not be taken arbitrarily since it can cause test power to be reduced [10, p. 5].

5. Other applications of p-value

5.1. *Quality control*

Theorem 2.4 indicates possibility of application in quality control. Namely, quality control represents statistical analysis method that tends to determine whether the sequence of small samples has certain quality. Those samples are drawn from the variable that quantifies process's quality of interest. That quality can be whatever is important for certain process that is controlled. Usually, quality is certain distribution that underlying variable is to have. The most often it is normal distribution [47, pp. 88-95]. Let (1) be such a sequence of samples, F the cumulative distribution function of underlying variable X, G chosen goodness of fit test statistic and $p_1, p_2, \ldots, p_n$ p-values for testing the hypothesis $H_0 : X \sim F(x)$ for given sequence samples via G. If process is in control state, then the p-value should be uniformly distributed on $(0, 1)$ and testing $p_1, p_2, \ldots, p_n$ for goodness of fit with uniform distribution on $(0, 1)$ should result in retaining H_0. Also, control chart with lower control limit $y = 0$ and upper control limit $y = 1$ can be used, where points (j, p_j), $j = 1, \ldots, n$ connect to a control line that should cover the band $(0, 1)$ uniformly.

Remark 5.1. *Null hypothesis can be generalized in the form H_0: "X satisfies quality Q" as long as there is a statistic that can be used to test that hypothesis.*

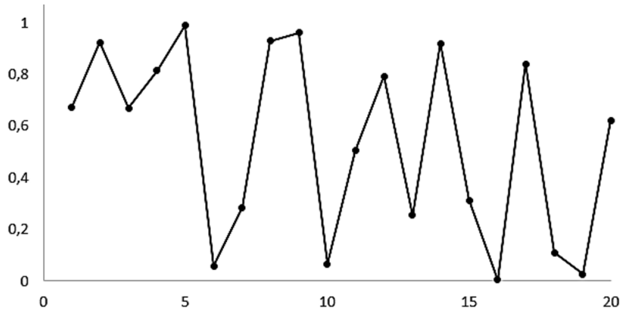

Figure 4. P-value control chart.

This chart may be beneficial if a simple test statistic G with acceptable power can be found, even if many tests have statistics that make calculating the p-value for a large n difficult [4].

5.2. *Power of tests*

The power of test is proportionate to the frequency of p-values that are less than the selected level of significance for repeated testing. P-value can thus be applied in approximate calculating the power of a chosen test. Theorem 2.4 suggests that p-value is non-uniformly distributed under alternative hypothesis. This implication can result in algorithms for fast and precise approximation of power, especially in multiple testing procedures [69].

5.3. *Efficiency of tests*

P-value is known to be part of Bahadur slope $c_{T_n} = -\lim_{n\to+\infty} \frac{2}{n} \ln p(T_n|H_0)$ (p being the p-value) of test determined by statistic T_n. Bahadur slope is used to calculate empirical relative or absolute test Bahadur efficiency [46]. If theoretical calculations of Bahadur slope are too complex or impossible, one can use p-value of test to approximate Bahadur slope. When simulating samples under alternative hypothesis and calculating p-value of each sample, for $n = 1, 2, 3, \dots$ until $n_1 \in \mathbb{N}$ and $n_2 = n_1 + 1$ such that

$$\left|\frac{\ln p(T_{n_1}|H_0)}{n_1} - \frac{\ln p(T_{n_2}|H_0)}{n_2}\right| \leq \varepsilon,$$

are reached for arbitrary chosen $\varepsilon > 0$, then $c_{T_{n_2}} = -\frac{2}{n_2} \ln p(T_{n_2}|H_0)$ can be taken as approximate value of Bahadur slope.

5.4. *Pseudorandom generators and random variable modelling*

Another use of p-value could be comparative analysis of pseudorandom numbers generators or random variable modelling algorithms. If several algorithms are given and for each a sequence of variates is generated, using chosen randomness or goodness of fit test and calculating p-value for each sample gives easily comparable numbers. Larger the p for algorithm tested, the required fit or randomness of generated numbers is better. This procedure can be repeated for various lengths of sequence. If several sequences for each algorithm are generated than empirical means of obtained p-values could be compared. Some researches deal with similar approaches [59].

Very often used application of p-value is one in verifying that novel distributions describe some variables by testing goodness of fit for real data [21].

6. Conclusion

While p-values have faced criticism for their misuse and misinterpretation, they remain a vital component of statistical inference. The key challenge lies not in the inherent limitations of p-values but in how they are applied and understood. By addressing misconceptions, refining teaching methods, and integrating complementary statistical tools, researchers can make more informed decisions and improve the reproducibility of scientific findings.

Rather than advocating for the outright replacement of p-values, this paper argues for their responsible use in conjunction with other statistical measures. Confidence intervals, effect sizes, and Bayesian methods each offer unique advantages that can strengthen statistical conclusions when used appropriately. Moreover, enhancing statistical education through clearer definitions, visual aids, and nuanced interpretations can mitigate many common errors in p-value application.

Future research should explore methods for improving the reliability of statistical inference, particularly in the context of multiple testing and power analysis. A deeper investigation into the relationship between p-values, sample size, and effect size could provide valuable insights into optimizing statistical decision-making. Also, comparing the quality of conclusions obtained via p-value versus the one obtained via Bayes' factor is a possible study topic with developing proper theoretical or empirical (in terms of possible simulations) methodology for that to be performed.

References

[1] H. Aguinis, J. C. Beaty, R. J. Boik, C. A. Pierce *Effect Size and Power in Assessing Moderating Effects of Categorical Variables Using Multiple Regression: A 30-Year Review*, J Appl. Psych. **90** (2005), 94–107.

[2] Q. U. Ain, M. A. Chatti, K. G. C. Bakar, S. Joarder, R. Alatrash *Automatic Construction of Educational Knowledge Graphs: A Word Embedding-Based Approach*, Information. **14** (2023), 526.

[3] A. G. Assaf, M. Tsionas, *Bayes factors vs. P-values*, Tourism Management **67** (2018), 17–31.

[4] A. Avdović, V. Jevremović, *Discrete Parameter-Free Zone Distribution and Its Application in Normality Testing*, Axioms **12** (2023), 1087.

[5] F. S. Azevedo, M. J. Mann, *The Mathematics in the Social Studies Textbook: A Critical Content Analysis and Implications for Students' Reasoning*, Creative Education **10** (2019), 1–25.

[6] G. J. Babu, C. R. Rao, *Goodness-of-Fit Tests When Parameters Are Estimated*, Sankhyā: The Indian Journal of Statistics **66** (2004), 63–74.

[7] J. R. Beniger, D. L. Robyn, *Quantitative Graphics in Statistics: A Brief History*, The American Statistician **32** (1978), 1–11.

[8] R. A. Betensky, *The p-Value Requires Context, Not a Threshold*, The American Statistician **73** (2019), 115–117.

[9] J. Cabrera, A. McDougall, *Statistical Consulting*, Springer, 2002.

[10] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, Routledge Academic, New York, 1988.

[11] V. Collinson, T. Fedoruk Cook, *"I don't have enough time" - Teachers' interpretations of time as a key to learning and school change*, Journal of Educational Administration **39** (2001), 266–281.

[12] D. Colquhoun, *The reproducibility of research and the misinterpretation of p-values*, Royal Society Open Science **4** (2017), 171085.

[13] T. Dahiru, *P-value, a true test of statistical significance? A cautionary note*, Annals of Ibadan Postgraduate Medicine **6** (2008), 21–26.

[14] B. Easton, *Teaching Mathematics to Social Sciences Undergraduates*, International Journal of Mathematical Education in Science and Technology **2** (1971), 83–89.

[15] J. Faber, L. M. Fonseca, *How sample size influences research outcomes*, Dental Press Journal of Orthodontics **19** (2014), 27–29.

[16] T. Fischetti, *Data Analysis with R: A Comprehensive Guide to Manipulating, Analyzing, and Visualizing Data in R*, (2nd edition), Packt Publishing, 2018.

[17] R. S. Flowers-Cano, R. Ortiz-Gómez, J. E. León-Jiménez, R. López Rivera, L. A. Perera Cruz, *Comparison of Bootstrap Confidence Intervals Using Monte Carlo Simulations*, Water **10** (2018), 166.

[18] C. Garnett, S. Michie, R. West, J. Brown, *Updating the evidence on the effectiveness of the alcohol reduction app, Drink Less: using Bayes factors to analyse trial datasets supplemented with extended recruitment*, F1000Research **8** (2019), 114.

[19] B. Gerald, T. F. Patson, *Parametric and Nonparametric Tests: A Brief Review*, International Journal of Statistical Distributions and Applications **7** (2021), 78–82.

[20] G. R. Gibbs, *Mathematics and Statistics Skills in the Social Sciences*, in C. M. Marr, M. J. Grove (eds.), Responding to the Mathematics Problem: The Implementation of Institutional Support Mechanisms, The Maths, Stats and OR Network, 2010, 44–50.

[21] J. Gillariose, O. S. Balogun, E. M. Almetwally, R. A. K. Sherwani, F. Jamal, J. Joseph, *On the Discrete Weibull Marshall–Olkin Family of Distributions: Properties, Characterizations, and Applications*, Axioms **10** (2021), 287.

[22] S. Goodman, *A dirty dozen: twelve p-value misconceptions*, Seminars in Hematology **45** (2008), 135–140.

[23] H. Goel, D. Raheja, S. K. Nadar, *Evidence-based medicine or statistically manipulated medicine? Are we slaves to the P-value?*, Postgraduate Medical Journal **100** (2024), 451–460.

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

[25] S. Greenland, *Valid P-Values Behave Exactly as They Should: Some Misleading Criticisms of P-Values and Their Resolution With S-Values*, The American Statistician **73** (2019), 106–114.

[26] L. G. Halsey, *The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum?*, Biology Letters **15** (2019), 20190174.

[27] D. Harris, L. Black, P. Hernandez-Martinez, B. Pepin, J. Williams, *Mathematics and its value for engineering students: what are the implications for teaching?*, International Journal of Mathematical Education in Science and Technology **46** (2015), 321–336.

[28] A. Hazra, *Using the confidence interval confidently*, Journal of Thoracic Disease **9** (2017), 4124–4129.

[29] T. Heckelei, S. Hüttel, M. Odening, J. Rommel, *The p-Value Debate and Statistical (Mal)practice—Implications for the Agricultural and Food Economics Community*, German Journal of Agricultural Economics **72** (2023), 47–67.

[30] M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, M. D. Jennions, *The extent and consequences of p-hacking in science*, PLoS Biology **13** (2015), e1002106.

[31] R. V. Hogg, J. W. McKean, A. T. Craig, *Introduction to Mathematical Statistics*, Pearson, 2019.

[32] D. Infanger, A. Schmidt-Trucksäss, *P value functions: An underused method to present research results and to promote quantitative reasoning*, Statistics in Medicine **38** (2019), 4189–4197.

[33] J. P. Ioannidis, *Why most published research findings are false*, PLoS Medicine **2** (2005), e124.

[34] J. P. A. Ioannidis, *What Have We (Not) Learnt from Millions of Scientific Papers with P Values?*, The American Statistician **73** (2019), 20–25.

[35] H. Jeffreys, *Some Tests of Significance, Treated by the Theory of Probability*, Mathematical Proceedings of the Cambridge Philosophical Society **31** (1935), 203–222.

[36] S. Jhade, M. Muthusamy, A. Singh, *Parametric and Non-Parametric Analysis*, in S. K. Mahapatra (ed.), Advances in Agricultural Research Methodology, S P Publishing, 2023.

[37] V. E. Johnson, S. Pramanik, R. Shudde, *Bayes factor functions for reporting outcomes of hypothesis tests*, Proceedings of the National Academy of Sciences of the United States of America **120** (2023), e2217331120.

[38] M. Kafi, M. Ansari-Lari, *"A statistically non-significant difference": Do we have to change the rules or our way of thinking?*, Iranian Journal of Veterinary Research **23** (2022), 300–301.

[39] R. E. Kass, A. E. Raftery, *Bayes Factors*, Journal of the American Statistical Association **90** (1995), 773–795.

[40] T. Konold, X. Fan, *Hypothesis Testing and Confidence Intervals*, in P. Peterson, E. Baker, B. McGaw (eds.), International Encyclopedia of Education (3rd edition), Elsevier Ltd, 2010, 216–222.

[41] A. Kühberger, A. Fritz, E. Lermer, T. Scherndl, *The significance fallacy in inferential statistics*, BMC Research Notes **8** (2015), 84.

[42] D. Lakens, *Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs*, Frontiers in Psychology **4** (2013), 863.

[43] D. Lakens, *The practical alternative to the p value is the correctly used p value*, Perspectives on Psychological Science **16** (2021), 639–648.

[44] P. Lytsy, M. Hartman, R. Pingel, *Misinterpretations of P-values and statistical tests persist among researchers and professionals working with statistics and epidemiology*, Upsala Journal of Medical Sciences **127** (2022).

[45] B. B. McShane, D. Gal, A. Gelman, C. Robert, J. L. Tackett, *Abandon statistical significance*, The American Statistician **73** (2019), 235–245.

[46] I. I. Nikitin, *Asymptotic Efficiency of Nonparametric Tests*, Cambridge University Press, 1995.

[47] J. S. Oakland, *Statistical Process Control*, (5th edition), Butterworth-Heinemann, 2003.

[48] T. O'Hagan, *Bayes Factors*, Significance **3** (2006), 184–186.

[49] R. G. Pourdavood, P. Wachira, *Importance of mathematical communication and discourse in secondary classrooms*, Global Journal of Science Frontier Research F: Mathematics and Decision Sciences **15** (2015).

[50] Z. Rafi, S. Greenland, *Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise*, BMC Medical Research Methodology **20** (2020), 244.

[51] N. Reid, *Significance, Tests of*, in N. J. Smelser, P. B. Baltes (eds.), International Encyclopedia of the Social & Behavioral Sciences, Pergamon, 2001, 14085–14091.

[52] A. Reinhart, *Statistics Done Wrong – The Woefully Complete Guide*, No Starch Press, 2015.

[53] H. Santos, J. Vicencio, S. de Ocampo, *Mathematics Competency and Proficiency of Engineering Freshmen of Wesleyan University-Philippines*, Open Journal of Social Sciences **10** (2022), 31–40.

[54] R. L. Sapra, S. Nundy, *Why the p-value is under fire?*, Current Medicine Research and Practice **8** (2018), 222–229.

[55] J. P. Simmons, L. D. Nelson, U. Simonsohn, *Life after p-hacking*, in Meeting of the Society for Personality and Social Psychology, New Orleans, LA, 2013, 17–19.

[56] A. M. Stefan, Q. F. Gronau, F. D. Schönbrodt, E. J. Wagenmakers, *A tutorial on Bayes Factor Design Analysis using an informed prior*, Behavior Research Methods **51** (2019), 1042–1058.

[57] G. M. Sullivan, R. Feinn, *Using Effect Size—or Why the P Value Is Not Enough*, Journal of Graduate Medical Education **4** (2012), 279–282.

[58] G. Upton, I. Cook, *Oxford Dictionary of Statistics*, Oxford University Press, 2006.

[59] I. Vattulainen, K. Kankaala, J. Saarinen, T. Ala-Nissila, *A Comparative Study of Some Pseudorandom Number Generators*, Computer Physics Communications **86** (1993), 209–226.

[60] B. Vidgen, T. Yasseri, *P-Values: Misunderstood and Misused*, Frontiers in Physics **4** (2016), 6.

[61] C. M. Vrbin, *Parametric or nonparametric statistical tests: Considerations when choosing the most appropriate option for your data*, Cytopathology **33** (2022), 663–667.

[62] R. J. Walley, A. P. Grieve, *Optimising the trade-off between type I and II error rates in the Bayesian context*, Pharmaceutical Statistics **20** (2021), 710–720.

[63] B. Wang, Z. Zhou, H. Wang, X. M. Tu, C. Feng, *The p-value and model specification in statistics*, General Psychiatry **32** (2019), e100081.

[64] L. Wasserman, *All of Statistics – A Concise Course in Statistical Inference*, Springer, 2005.

[65] R. L. Wasserstein, N. A. Lazar, *The ASA statement on p-values: context, process, and purpose*, The American Statistician **70** (2016), 129–133.

[66] A. Watson, J. Mason, *Seeing an Exercise as a Single Mathematical Object: Using Variation to Structure Sense-Making*, Mathematical Thinking and Learning **8** (2006), 91–111.

[67] S. Wellek, *A critical evaluation of the current "p-value controversy"*, Biometrical Journal **59** (2017), 854–872.

[68] G. Wilson, *Ten quick tips for creating an effective lesson*, PLoS Computational Biology **15** (2019), e1006915.

[69] F. Zhang, J. Gou, *A P-value model for theoretical power analysis and its applications in multiple testing procedures*, BMC Medical Research Methodology **16** (2016), 1–7.