Received: 14 May 2025 | Revised: 15 January 2026 | Accepted: 22 January 2026

DOI: 10.1002/test.70033

ORIGINAL ARTICLE

WILEY

Analogies and differences between the logic behind statistical hypothesis testing and proofs by contradiction: What can we learn from them?

Maria Cristina Amoretti[1] | **Daniele De Martini**[2]

[1]DAFIST - Dipartimento di Antichità, Filosofia e Storia, Università di Genova, Genova, Italy

[2]DIEC - Dipartimento di Economia, Università di Genova, Genova, Italy

Correspondence
Daniele De Martini, DIEC - Dipartimento di Economia, Università di Genova, Genova, Italy.
Email: daniele.demartini@unige.it

Abstract

Statistical hypothesis testing (SHT) is widely employed across numerous scientific disciplines, and a clear understanding of its underlying logic is essential for the broader scientific community. Here, drawing upon both epistemological and statistical perspectives, we aim to clarify—primarily for educational purposes—the logical relationship between proof by contradiction (PBC) and SHT. We begin by outlining the logics of SHT and PBC, followed by a discussion of their key similarities and differences. We then explore the pedagogical value of these analogies and distinctions through illustrative examples. Our objective is to help prevent common misinterpretations in the application of SHT, particularly in the interpretation of its outcomes. By elucidating the conceptual parallels between SHT and PBC, we aim to make the logical structure of SHT more transparent. Finally, we highlight and discuss several critical aspects relevant to teaching, with the intention of enhancing the pedagogical effectiveness of instruction in this area.

KEYWORDS

epistemology of statistics, proof by contradiction, scientific reasoning, statistical hypothesis testing, teaching statistics

1 | INTRODUCTION

Statistical tests were originally conceived as a technique to support, albeit stochastically, the alternative hypothesis. In fact, when Sir Ronald Fisher conducted the famous experiment with the eight cups of tea, his goal was to provide his acquaintance, Lady Muriel Bristol, who claimed to be able to discern whether tea or milk was poured first, with an opportunity to demonstrate her ability (see the *Lady Tasting Tea* experiment; Fisher, 1935).

Today, the use of statistical hypothesis testing (SHT) is widespread; indeed, many papers published in major scientific journals rely on these methods. Furthermore, SHT is applied across a wide range of scientific disciplines, from medicine to engineering, and from physics to psychology and economics.

For example, about 70% of the papers on the web page of Nature Communications (as of August 29, 2025) adopt SHT; applications are more frequent in Biological Sciences and Health Sciences (almost 100%), and in the Scientific Community and Society sections. We note that some of the most relevant recent results in Physical Sciences (e.g., those concerning the Higgs boson [Aad et al., 2012] and gravitational waves [Abbott et al., 2016]),

a section of Nature Communications where SHT is applied less frequently, are also based on SHT. As a further example, Böschen (2023) reports that about 85% of papers in psychology adopted SHT, and Chavalarias et al. (2016) reports that the vast majority of papers in the biomedical literature reported SHT results.

Consequently, a thorough understanding of the underlying logic and general characteristics of SHT is of paramount importance for the entire scientific community.

It is important to note that many research articles, across different scientific areas, focus on the *p*-value (or even on an asterisk-based scale indicating different levels of statistical significance). Misinterpretations of *p*-values, as well as the overlooking of confidence intervals and effect sizes, are issues extensively discussed in the literature (e.g., Goodman, 2008). In this paper, we do not directly discuss these topics; however, *p*-values inevitably enter our analysis insofar as they are part of the standard machinery of SHT, for instance, in connection with non-rejection decisions, sample size dependence, and the assessment of statistical significance. Our focus, instead, remains on the logical structure of SHT and on the interpretation of its outcomes, rather than on a systematic treatment of *p*-values themselves. We argue that the very logic underlying SHT may be often misunderstood by those who employ it. In some cases (based on the personal experience of one of the authors and confirmed by an anonymous referee), even by those who teach it.

The aim of this paper is to elucidate the logical structure of SHT by drawing on our combined epistemological and statistical expertise. We adopt the standard pedagogical stance that the research question specifies H1, while the *p*-value is computed under the null hypothesis H0 (the roles and meanings of H1 and H0 will be introduced explicitly in the next section); we thus aim at clarifying the logical rationale behind this routine practice through the analogy between the logic of proof by contradiction (PBC) in mathematics and that of SHT. Given the inherent randomness of SHT results, we also argue, as a corollary, for the adoption of appropriate reliability indicators, such as the Repeatability Probability estimate (RP), which will be discussed later in the paper and in Appendix 1. Accordingly, the paper targets instructors and students across disciplines and applied researchers who routinely use SHT, not only statistics or mathematics students. Moreover, we are also addressing a broader audience, for example, those researchers whose *motto* is "semper discendum est" ("one must always keep learning").

The paper is organized as follows: in Sections 2 and 3, we present the two distinct procedures, SHT and PBC, introducing their respective logical structures; then, in Sections 4 and 5, we highlight their similarities and differences; subsequently, in Section 6, we illustrate what we believe to be the pedagogical benefits and value of the results of this comparison; finally, in Section 7, we conclude by discussing this topic in the context of statistics education.

2 | LOGICAL STRUCTURE OF SHT

SHT is typically regarded as a method to assess whether a predefined hypothesis (H0, the null hypothesis) aligns with the observed data. However, historically, the primary goal of statistical tests has been to support the hypothesis complementary to H0, known as H1. This perspective is reflected in Sir Ronald Fisher's statement: "Every experiment may be said to exist in order only to give the facts a chance of disproving the null hypothesis" (Fisher, 1935). Since H0 and H1 are complementary (Lehmann & Romano, 2022), this statement is equivalent to saying: "Every experiment may be said to exist in order only to give the facts a chance of supporting the alternative hypothesis."

The essence of statistical testing can be summarized as follows:

1. To support a hypothesis (H1), referred to as the alternative hypothesis (e.g., "there is a difference between groups," "a relationship exists between variables," or "a micro-particle not included in the basic model exists"), start by assuming the complementary hypothesis H0, known as the null hypothesis, to be true. The null hypothesis represents a baseline or *default* position (e.g., "there is no difference between groups," "no relationship exists between variables," or "a micro-particle not included in the basic model does not exist").
2. When performing a statistical test, a function of the data that reflects the truthfulness of the hypotheses under consideration, referred to as the test statistic T, is defined (e.g., T is the difference between sample means appropriately standardized, the sample correlation, or the count of collisions between micro-particles).
3. Next, a set R of possible values for T is established (which are extreme values for T) "in the direction of H1." R is called the rejection region and must have small probability under H0. The complementary set NR (NR $= R^c$) represents the non-rejection region and has high probability under H0. The probabilities associated with these regions must be specified before running the experiment. The small probability of observing values in R under H0 is denoted by α, and in practice defines the rejection region which, indeed, is usually labeled R_α. Notably, the probability of observing values in R_α under H1 (i.e., the power of the test, denoted by π) is (usually) greater than α (i.e., $\pi > \alpha$).

(Typically, NR is also called an *acceptance region*, but this habit is, in fact, misleading, since failure to reject H0, that is, the fact that the data do not provide sufficient evidence in favor of H1, does not amount to establishing that H0 is true, nor that it is supported by the data. More precisely, labeling NR as an *acceptance region* may naturally suggest that, whenever the test statistic falls into NR, the null hypothesis is accepted *as true*, or at least *as empirically corroborated*. However, this interpretation conflicts with the logic of SHT, where non-rejection of H0 reflects only the lack of sufficiently strong evidence against it, not positive evidence in its favor.)

4. Finally, after observing the sample, the value of the test statistic T is calculated. If the observed value of T falls within R_α we are faced with the following situation: on one hand, H0 is assumed to hold; on the other hand, the observed value of T lies within a set of highly unlikely outcomes under H0 (a situation that may seem almost paradoxical). Furthermore, since this value of T lies "in the direction of H1" we are compelled to reject H0. This provides experimental evidence, essentially stochastic support, for H1.

(This reasoning is often illustrated by the court of law analogy: just as a defendant is presumed innocent unless the evidence is sufficiently strong to minimize the probability of a wrong conviction, H0 is assumed to hold unless the evidence against it is sufficiently strong, and the probability of wrongly rejecting H0 is controlled and kept small. However, this analogy should not be pushed too far. Unlike legal reasoning, where a defendant who is not convicted is commonly regarded as innocent, the non-rejection of H0 in SHT does not amount to establishing its truth, nor to providing positive evidence in its favor.)

It is important to note that H1 does not always merely represent the presence of an effect or a correlation (e.g., a non-zero effect or non-zero correlation). More generally, it may represent an effect or correlation that exceeds a predefined threshold θ_0, with $|\theta_0| > 0$, where θ_0 is set on the basis of scientific relevance. Thus, H0 does not necessarily represent the absence of an effect or correlation (the so-called *nil* hypothesis), but, more broadly, an effect or correlation that is at most θ_0.

2.1 | Example

Suppose we want to test whether a new treatment (t) has a clinically meaningful effect compared to placebo (p). Given that the test concerns the difference between population means, that is, $\theta = \mu_t - \mu_p$, the maximum clinical non-relevant difference is $\theta_0 > 0$. The test is as follows.

1. Define the hypotheses. H1: $\theta > \theta_0$ (i.e., the new treatment has a clinically relevant effect), and H0: $\theta = \theta_0$ (i.e., the difference is not clinically relevant). Assume H0 is true.
2. Consider a sample of size n run under the new treatment, and a second independent sample under placebo, still of size n, providing the sample means X_t and X_p. So, $T = (X_t - X_p - \theta_0)/\sqrt{(2/n)}$.
3. Assuming that the variance in both populations is 1, the distribution of T under H0 is, at least approximately, $N(0,1)$. Extreme values for T "in the direction of H1" are those in the interval $(t, +\infty) = R$, where t is sufficiently greater than 0. Then, setting $\alpha = 1\%$ we obtain $t = 2.33$, and $R = (2.33, +\infty)$. For example, if $\theta - \theta_0 = 0.5$ and $n = 72$, then $\pi = 0.75$.
4. Given the sample data, if the resulting $T > 2.33$, then it is usually said that H0 is rejected. In practice, experimental data stochastically support H1.

3 | LOGICAL STRUCTURE OF PBC

The mathematical PBC is the mathematical reasoning used to demonstrate the truth of a statement by showing that assuming the opposite leads to a logical contradiction. Here is a clearer explanation:

1. If the goal is to prove that proposition P is true, assume the opposite or the negation of this proposition, that is suppose ¬P is true. The assumption of ¬P is the starting point for the proof.
2. Based on the assumption ¬P, apply legitimate mathematical operations, known facts, and logical rules to derive conclusions.
3. Reach a contradiction. This could be:
   a. A contradiction of a known fact (e.g., $2 = 1$, or the number x is simultaneously even and odd).
   b. A contradiction within the logic of the problem (e.g., assuming $A = B$ leads to a derivation that $A \neq B$, which cannot both be true).
4. Since the assumption ¬P leads to a contradiction, which is an impossible situation, it cannot be true (it is false). Therefore, P (the original proposition) must be true.

Proving the truth of a statement by rejecting its negation is thus the logic of PBC.

3.1 | Examples

A classic example of a theorem whose proof relies on PBC is a result from Euclidean geometry, namely the

Converse of the Alternate Interior Angles Theorem: if a transversal intersects two lines and the alternate interior angles formed are congruent, then the two lines are parallel. In a PBC, one assumes that the two lines are not parallel and instead intersect at a point C, together with their intersections with the transversal, A and B, forming triangle ABC. Then, by applying the Exterior Angle Inequality Theorem, a contradiction is obtained.

As a further example, consider the proof that $\sqrt{2}$ is irrational, that is that it cannot be expressed as a fraction of two integers. In this case, assume that $\sqrt{2}$ is rational and can be expressed as p/q, where p and q are integers with no common divisors (that is, their greatest common divisor is 1).

If $\sqrt{2} = p/q$, then $2q^2 = p^2$. This implies p^2 is even. If p^2 is even, then p must also be even. Let $p = 2k$ for some integer k. This implies $2q^2 = 4k^2$, that is, $q^2 = 2k^2$. This implies q^2 is even. If q^2 is even, then q must also be even. If both p and q are even, they are both divisible by 2. This contradicts the assumption that p and q have no common divisors. As the assumption that $\sqrt{2}$ is rational leads to a contradiction, it cannot be true. Therefore, $\sqrt{2}$ must be irrational.

4 | ANALOGIES BETWEEN SHT AND PBC

There are some notable analogies between SHT in statistics and PBC in mathematics (National Academy of Sciences—Engineering—Medicine, 2019; Batanero, 2000; Batanero & Díaz, 2006; Castro Sotos et al., 2007; Falk & Greenbaum, 1995; Liu & Thompson, 2005).

First, both methods commence with an assumption that functions as a negation.

In SHT, one starts by assuming the null hypothesis (H0), which can be interpreted as the negation of the alternative hypothesis (H1), such that $\mathrm{H0} = \neg\mathrm{H1}$. For example, if the objective is to test for the hypothesis regarding the existence of an effect, difference, or correlation between variables, H0 encodes the absence of such an effect, difference, or correlation ($\neg$H1). Within this framework, SHT is understood as beginning with the assumption of $\neg$H1 in order to ultimately support H1.

Similarly, in PBC, one assumes the negation of the target statement. More precisely, one begins with $\neg$P with the ultimate goal of establishing P.

Second, both methods share a similar logical structure, reaching their conclusions by defeating the initial assumption; that is by identifying a specific type of contradiction.

In the context of SHT, the validity of the null hypothesis H0 is questioned by data that are highly unlikely under the assumption that H0 is true. Specifically, one assumes H0 and examines a test statistic T, identifying a contradiction when the data (particularly the value of the test statistic) fall within a region "in the direction of H1" for an experimental result (i.e., R_α), which is of small probability (i.e., the predefined α) if H0 is true. Since the observed data are real and valid, this improbability leads to the conclusion that H0 is false and must therefore be rejected, supporting the alternative hypothesis H1. Symbolically:

$$\mathrm{H0} \rightarrow T \text{ in } R_\alpha \Longrightarrow \mathrm{H1}.$$

For clarity, α represents the "probabilistic threshold of absurdity." It is worth noting that this threshold is conventional and context-sensitive, being fixed by the research design.

In PBC, the validity of the assumption $\neg$P is challenged by statements that contradict either a known fact (e.g., an axiom) or the initial assumption itself. Specifically, one assumes $\neg$P, derives a contradiction with a known fact or within the logical framework of the problem (ctr), and concludes that $\neg$P must be false, thereby proving P. Symbolically:

$$\neg\mathrm{P} \rightarrow \mathrm{ctr} \Longrightarrow \mathrm{P}.$$

Third, both methods establish their target only indirectly.

Neither SHT nor PBC directly stochastically confirms or logically proves the respective target claim (the alternative hypothesis H1 or P). Rather, each proceeds by defeating its negation: SHT rejects the null hypothesis H0 on the basis of highly improbable data under H0 (i.e., $T \in R_\alpha$), while PBC derives a logical contradiction from $\neg$P and thus concludes that $\neg$P is false (Antonini & Mariotti, 2008).

5 | KEY DIFFERENCES BETWEEN SHT AND PBC

Despite the above analogies, there are notable distinctions between SHT and PBC (Otani, 2019).

First, a significant difference lies in the nature of the "contradiction" identified in the two methods.

In SHT, the "contradiction" is probabilistic rather than strictly logical, and is better characterized as a state "analogous to contradiction." Such a state arises when the observed data fall within a region that is highly improbable under the null hypothesis. Therefore, rejecting H0 reflects statistical improbability, not a formal logical inconsistency. In contrast, in PBC, the contradiction

is a strict logical inconsistency within classical logic (e.g., P∧¬ P). To put it differently, the contradiction that defeats ¬P holds only relatively to a stable background of axioms, rules of inference, and ontological assumptions. This means that the contradiction arises within a specific deductive framework.

Second, the relationship between premises and conclusions differs significantly between the two methods.

In SHT, even if the premises (which, moreover, generate the collected data) are true, and the data lead to the rejection of H0, in conclusion, H1 is not logically proven but only stochastically supported.

In stark contrast, in PBC, if the premises are true, P is logically proven and must be necessarily and certainly true. In other words, while probability and uncertainty characterize hypothesis testing, necessity and certainty define PBC. For this reason, SHT has been aptly described as a "stochastic proof by contradiction" (Rubin, 2004).

A third difference regards the uniqueness of the conclusion.

In SHT, the rejection of H0, due to the statistical improbability of observed data, supports an H1 which can be quite generic and can be induced by other alternative hypotheses H2, ..., H*k*. In fact, some of the latter may be wrong, and some others might be stochastically confirmed as well and may even be more reasonable, explanatory, and fruitful than H1 (for instance, if H1 is "there is a correlation between violence and video games," this correlation could be implied by different hypotheses, such as H2: "video games cause violence," H3: "violent individuals are more likely to play video games," or H4: "a third variable increases both violence and video game use").

Quite differently, in PBC the falsity of ¬P logically entails the truth of P because this method is based on specific ontological assumptions and inference rules, such as the law of excluded middle, double negative elimination, and so on. Thus, PBC's force constitutively depends on adopting such assumptions and principles. This makes explicit that PBC is conditional on an axiomatic framework (definitions, axioms, and inference rules). In other words, no other statement other than P can be inferred by the falsity of ¬P.

A fourth and particularly significant distinction pertains to the nature of the premises.

In the context of SHT, premises correspond to the collected data, which are predominantly non-mathematical in nature and, crucially, are random variables, as they represent a random sample (Otani, 2019). Moreover, in SHT what "constrains" the inference (namely the sample size n, the pre-specified significance level α, the test statistic, and other modeling assumptions) does not come from logical axioms but from those practical design choices. Consequently, reproducing or replicating a statistical test does not necessarily yield identical outcomes. For instance, if a test T_1 results in the rejection of H0, there is no assurance that subsequent, identical tests T_2, T_3, ..., T_n will produce the same conclusion. Therefore, reliability indicators of the result, such as the RP (see Section 6.2), become essential.

Conversely, in PBC, the premises consist of mathematical statements (e.g., "if p^2 is even, then p must also be even"), which are invariant and universally assumed to be true. This ensures that once a contradiction is identified, it holds universally and permanently. In other words, replicating the proof will invariably lead to the same conclusion. As a result, reliability indicators for such proofs are unnecessary.

A fifth difference, implicit in the foregoing but worth making explicit, concerns what the "bounding" depends on.

In SHT the decision to reject H0 and support H1 crucially depends on contingent design choices and resources that vary across studies (sample size n, the pre-specified significance level α, the test statistic, and modeling assumptions).

By contrast, in PBC, whether a proof succeeds does not depend on how it is written or on the number of cases examined; it depends solely on the accepted axioms, definitions, and rules of inference.

In short, PBC is constrained by accepted logical axioms, definitions, and rules of inference, whereas SHT is constrained by contingent and often varying design choices.

6 | USEFULNESS OF ANALOGIES AND VALUE OF DIFFERENCES

First, note that we draw examples from clinical trials, quality control, and psychology simply because their adoption is standard in statistics instruction. In fact, the examples reported in this section are purely illustrative: the theoretical and pedagogical points we make are domain general and do not rely on any particular area of application.

6.1 | Experimental data never support H0

To highlight the usefulness of analogy between SHT and PBC, we address one of the most frequent misconceptions arising in the application of statistical tests: the tendency among practitioners (ranging from biologists to

engineers) to regard the null hypothesis (H0) as true when it is not rejected. The analogy provides a conceptual framework for practitioners to recognize the fallacy of considering that the null hypothesis is true, or simply stochastically supported by data, merely because it has not been rejected (Reeves and Brewer, 1980).

In mathematics, the absence of a contradictory statement derived from an assumption does not lead to the conclusion that the negation of the assumption (¬P) is true. Similarly, in statistical inference, failing to find data that are sufficiently improbable under the assumption that H0 is true neither justifies concluding that H0 is true, nor implies that H0 has been stochastically confirmed.

6.1.1 | Example

In the statistical process control (SPC), particularly in control charts (CCs), it is usual to define H0 as representing the process being under control (e.g., the mean weight of a package of pasta is 500 g), while H1 represents the process being out of control (e.g., the mean weight deviates from 500 g in a way that triggers an out-of-control signal in the CC). Accordingly, the CC is planned, and in particular the sample size, such that deviations of, for instance, 5 g have a 95% probability of being detected. However, in case of a non-significant test result (i.e., no rejection of H0), this does not imply that the process is under control, or that the mean weight is precisely 500 g, as it is often misinterpreted to mean. Conversely, to give data the chance of supporting that the process is under control, a test of equivalence should be adopted, where H1 represents the mean lying within 500 ± 5 g, that is, H1: $495 < \mu < 505$.

6.2 | Results of SHT are stochastic, not definitive, and therefore require reliability indicators

Differences between SHT and PBC mainly remark that in SHT H1 is not logically proven, as in PBC, but merely stochastically confirmed. This distinction is critical because it serves as a reminder to not overemphasize an eventual statistical significance. In fact, overreliance on it can lead to the mistaken belief that a statistically significant result (T in R_α) constitutes definitive "proof" of H1, thereby misinterpreting the findings and exaggerating the strength of the evidence. Hence, reliability indicators for SHT results are, on the one hand, needed, and on the other hand, they should help in recalling that a significant result is not definitive.

Usually, for this purpose, the p-value is adopted, but it is often misinterpreted (Goodman, 2008) and, more importantly, it does not directly quantify the reliability of SHT results (since it reports the value of α that would have to be set in order for the observed test statistic to fall on the boundary of R_α; this is also rather unhandy and awkward).

To the contrary, since the outcome of an SHT is inherently random, any statistically significant result provides only *stochastic* support for H1. For this reason, the focus should be on assessing its randomness. Indeed, in SHT, the implicit randomness of gathered data means that replicating a statistical test does not always yield the same result. Given that statistical inferences are based on sampled data, replication is crucial for verifying whether an observed effect is consistent across different samples from subsequent experiments.

In general, reproducibility and replicability are an absolute necessity in experimental sciences, as Galileo's scientific method emphasized the repeatability of experiments as a truth indicator. Recall, for example, Galileo's falling bodies experiment, repeatedly reproduced by others, that was crucial in demonstrating that reliable knowledge arises from replicable results, thereby consolidating the foundations of modern physics. Hence, since experiments involving SHT inherently include randomness, reproducibility and replicability become even more critical in such contexts.

However, achieving reproduction and replication is not always straightforward. Practical constraints, such as scarce resources, time, and ethical considerations, can make reproduction and replication infeasible or economically unsustainable. Moreover, in some cases, limited or no reproducibility may stem from intrinsic features of the phenomenon under investigation, such as context sensitivity, inherent variability, or when the experiment is tied to historically situated and unique events. In such cases, the initial conditions of the experiment cannot be reconstructed, and the reproduction of the empirical event in its specificity is not feasible. Nevertheless, when the observed data exhibit a stochastic component and are analyzed by means of SHT, the inferential outcome, although referring to a unique empirical event, can still be legitimately assessed in terms of reliability indicators. This assessment relies on the assumption that the data-generating process can be reasonably described by the same probabilistic law, independently of the impossibility of concretely repeating the specific historical event. In this sense, non-replicability concerns the empirical event in its uniqueness, but not the statistical model adopted to describe it, nor the inferential procedure applied to the data.

6.2.1 | Repeatability Probability estimate as a reliability indicator

The considerations outlined above, together with rigorous logic, lead us, both in interpreting results and in adopting reliability indicators, to favor an approach different from that based on the p-value, namely the adoption of the RP estimate (also referred to as Reproducibility Probability).

RP is the probability of finding stochastic support for H1 (i.e., statistical significance) in an identical replication of the experiment, that is, a further experiment performed under identical research protocol and experimental conditions (De Martini, 2008). This definition is context-independent and applies regardless of the application domain. The value of RP is unknown, since it depends on the unknown population distribution from which the data are sampled; however, RP can be estimated (see Appendix 1, for a brief introduction to RP estimation). Accordingly, RP estimates quantify the randomness of statistical test-based findings. In other words, RP serves as an indicator to measure the stability of the experimental findings.

Notably, RP does not quantify the probability of repeating the observed event, but rather the probability of obtaining the same inferential outcome in an ideal replication of the statistical procedure, conducted under the same probabilistic assumptions. For this reason, assessing reliability in terms of RP remains conceptually appropriate even in the presence of experiments that are not strictly replicable, provided that the adopted statistical is explicitly stated and justified.

Moreover, some theorems claim that the outcome of a statistical test can be considered significant if, and only if, the RP estimate is greater than ½ (De Martini, 2008; De Capitani & De Martini, 2015, 2021). This means that the outcome estimated to be the most repeatable (among significance/non-significance) is the result.

In practice, when the RP estimate is only slightly above ½, the test result is significant, but its stability remains extremely dubious, and further testing is necessary to support H1. On the contrary, if the RP estimate is substantially higher than ½, the result is significant and can be considered sufficiently stable, justifying the avoidance of additional studies and experiments.

We noticed that, although the issue of reproducibility has been addressed for several decades (e.g., Goodman, 1992) and the so-called reproducibility crisis gained prominence in 2014 (see Anonymous, 2014; McNutt, 2014), in recent years scholarly attention to this matter appears to have declined (possibly as a consequence of the COVID-19 emergency), and the adoption of the RP has not yet become widespread. In our view, the 2014 crisis can be attributed primarily to a lack of "statistical literacy" (cf. Gal, 2000, 2003). At the same time, statistical education, being an integral component of statistical literacy (Gal, 2000), requires reconsideration, with a stronger emphasis on the logic underlying SHT (to which this work seeks to contribute) and on the implicit randomness of SHT results, which ultimately motivates the adoption of RP estimates.

6.2.2 | Example

The clinical trial study NCT01848054 (clinicaltrials.gov) reports the results of a comparison between two frequencies, 113/128 versus 122/128. The (approximated) Z-test was significant, as the test statistic was 2.07, exceeding the critical value of 1.96. The simplest RP estimator gives $RP = 54.4\% > ½$, while more precise estimators provide 53.0% and 52.6%. It is immediately clear that the observed significance is not stable, as the probability of repeating a significant finding in a subsequent identical experiment is estimated to be close to ½. The result resembles that of tossing a fair coin. The p-value of 0.038, which is not very close to 0.05, might have been misleading regarding the stability of the significance.

6.3 | Flexibility in different research areas

Another useful difference between SHT and PBC is that the former are flexible in their severity depending on the field of application.

Rejecting H0 is (data-)driven by statistical improbability, instead of logical inconsistency, and the criterion defining statistical improbability is not invariable. In fact, such a criterion depends on the value of a very key parameter: the Type I error probability α (see Section 2, point 3). This probability can differ, and it actually does across studies and disciplines.

(It is worth noting that probability computation of observed data also depends on the chosen test statistic to analyze experimental data, but in this general context we omit to develop this technical aspect).

To put it differently, while PBC relies on strict logical necessity, SHT involves probabilistic reasoning shaped by methodological choices made by researchers. This variability highlights the unavoidable involvement of non-epistemic values, practical interests, and contextual scopes in SHT, in contrast to PBC (Douglas, 2000; Wilholt, 2009).

In particular, we refer to the assessment of so-called "inductive risk" (i.e., the risk associated with the

possibility of making a wrong inference when there is uncertainty, particularly in contexts where rejecting or failing to reject a hypothesis could lead to morally, socially, or practically significant consequences) when determining statistical significance. As (Douglas, 2000) convincingly emphasizes, the choice of α is not dictated by purely factual or epistemic criteria, but often involves value judgments and practical trade-offs, such as balancing the risks given by false positives (viz. type I errors) and false negatives (viz. type II errors) results, in a given research context.

Now, it is worth noting that the choice of α also affects β (i.e., the probability of a type II error, $\beta = 1 - \pi$), given that the other components of the statistical test (e.g., test statistic, sample size, and experimental settings) remain the same. Specifically, as α decreases, the power function decreases, leading to an increase of β.

6.3.1 | Some examples in different research contexts

In phase III clinical trials, the value of α is relatively high (e.g., 5%), as it reflects a regulatory and statistical balance—shaped by considerations of inductive risk—between controlling the risk of false-positive conclusions and avoiding the loss of truly effective treatments. A false-positive result can lead to the approval of a treatment that, although safe, provides no or less-than-expected clinical benefit, thereby exposing patients to ineffective therapies and generating ethical concerns and societal costs. At the same time, maintaining a relatively high α helps limit the risk of false negatives, thereby reducing the probability that effective treatments are withheld from patients who may benefit from them.

In SPC, particularly in CCs, α is set at a much lower level (approximately 0.27%, as in 3-sigma CCs). In this context, a false positive would result in an unnecessary interruption of the production process, leading to a loss of time and financial resources that should ideally be minimized. Conversely, false negatives often correspond to the production of non-conforming parts, that is, errors that can sometimes be corrected at a later stage and that, due to tolerance limits, may incur lower costs (i.e., losses) and risks.

By contrast, in fields such as particle physics, where the discovery of a new fundamental particle requires extraordinary evidence, an extremely stringent α threshold is adopted (e.g., 10^{-7}) to minimize the risk of false positives. In this domain, false negatives can be addressed through further, more precise, and more powerful research efforts over time.

In conclusion, the choice of α is not solely a matter of logic or statistical convention but depends on the specific context and the consequences of errors. In particular, it is influenced by the associated costs (economic, scientific, ethical, and social) and the corresponding potential losses.

6.4 | Interpreting correlations: alternative hypotheses in SHT

The non-uniqueness of the conclusion in SHT, that is, the fact that H1 encompasses several alternative hypotheses, creates opportunities for further research, potentially leading to deeper, more nuanced, and precise findings. In other words, alternative hypotheses H2, H3, ..., H_k, which account for potential confounding variables and contextual factors influencing the results, may also receive stochastic confirmation.

Specifically, when data support the presence of a correlation between two variables (i.e., H1: $\rho \neq 0$), different causal interpretations remain possible. On the one hand, if H2 posits a direct causal relationship between the two variables, H2 may still be false despite rejecting H0. On the other hand, the observed correlation may instead suggest the existence of a third variable that causally influences both, leading to an alternative hypothesis H3.

6.4.1 | Example

Consider a study examining the correlation between playing violent video games and higher scores on aggression scales. The hypotheses tested are:

> H0. There is no correlation between playing violent video games and aggression scores ($\rho = 0$).
>
> H1. A correlation exists between playing violent video games and aggression scores ($\rho \neq 0$).

Suppose that H0 is rejected on the base of statistical significance. This result provides stochastic support for the existence of a correlation, but it does not determine its causal interpretation. In particular, rejecting H0 does not warrant the conclusion that playing violent video games *causes* increased aggression scores (i.e., H2). Several alternative hypotheses are compatible with the same statistically significant result. For instance, one may consider the following ones:

> H2. Playing violent video games causes increased aggression scores (direct causal hypothesis).

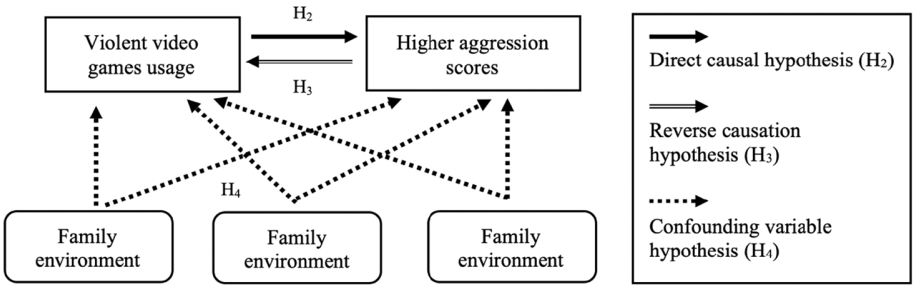

FIGURE 1 Alternative causal structures compatible with a statistically significant correlation between violent video games usage and higher aggression scores. H2 represents the direct causal hypothesis, H3 the reverse causation hypothesis, and H4 different confounding variable hypotheses, in which a third factor influences both variables. All three structures are consistent with rejection of H0, illustrating that statistical significance alone does not determine causal interpretation.

H3. Individuals with higher aggression scores tend to engage more frequently with violent video games (reverse causation hypothesis: higher aggression scores cause increased violent video game usage).

H4. One or more third variables, such as family environment, socioeconomic status, or peer influences, cause both increased violent video game usage and higher aggression scores (confounding variable hypothesis).

Importantly, the mere rejection of H0 may be consistent with any of these hypotheses. Thus, while the statistical test stochastically supports the generic alternative hypothesis H1, it does not discriminate among competing causal explanations, some of which may involve confounding variables influencing both measured outcomes (see Figure 1).

7 | CONCLUSIONS: TEACHING STATISTICS

In general, the utility and impact of this article can be primarily envisioned in relation to the teaching of statistics. Our intended audience includes instructors and students of statistics and research methods across disciplines, as well as applied researchers.

Based on our experience, the logical relationship between SHT and PBC is not yet part of the standard conceptual framework typically used in statistics education. Over the past 2 years, the authors have had several formal opportunities, such as seminars and book presentations, to discuss this relationship with colleagues. In many of these contexts, the proposed analogy was met with skepticism or regarded as peripheral, which may help explain why these ideas are not always systematically conveyed to students.

Nevertheless, as was the case with (Kuhn, 1970), there is reason to hope that future generations of educators will contribute to integrating these perspectives into statistics education, thereby enhancing conceptual understanding of the analogies and differences between SHT and PBC.

Furthermore, we believe that the integration of epistemological-philosophical and technical-statistical components, which constitutes the foundation of this work, represents a rather original contribution (we did not find any contribution on this topic that integrated both aspects). We thus hope that it may enhance the clarity of the subject under discussion.

In particular, the analogies and differences discussed provide valuable insights into the nature of statistical tests, particularly in two key aspects: (1) the primary objective is to provide evidence supporting H1, rather than to "generically test" the hypotheses; and (2) any such evidence is inherently stochastic in nature. Consequently, incorporating these reflections into the teaching of statistics would offer significant pedagogical benefits, and this document has been written, at least in part, with that objective in mind.

To date, such considerations have not been systematically included in statistical textbooks, whether at the advanced or introductory level, likely for the following reasons.

Mainly, there are notable differences in disciplinary and historical contexts: mathematical logic and statistics have distinct goals and traditions. PBC belongs to a purely deductive framework, whereas statistical tests operate within an inductive domain characterized by uncertainty. Statistics, as a discipline, has largely been developed empirically and with an applied focus, whereas mathematical logic has followed a more theoretical and axiomatic trajectory. Consequently, the two fields have evolved in parallel with little interaction, which may explain why the analogy has not become part of the standard narrative.

Moreover, perhaps due to the differences in disciplinary and historical contexts mentioned above, it seems that statistics instructors are largely unaware of the analogies (and differences) between SHT and PBC, particularly regarding the logical and philosophical foundations of statistical tests (a suitable scientific survey might be conducted in the near future to support these impressions).

Finally, even if statistics instructors were aware of the aforementioned logical relationships, they might find them ill-suited for teaching purposes. In fact, statistical textbooks are often oriented toward practical applications rather than the underlying philosophy. The primary goal is to teach the use of statistical tools rather than to delve into the logical foundations of the methodology. For many students, statistics is already a challenging subject, and introducing comparisons with concepts from formal logic may have been deemed unhelpful or overly complex. We, on the other hand, believe that these aspects can contribute to a better understanding of statistical tests and their use and, indeed, consider them indispensable.

ACKNOWLEDGMENTS

Open access publishing facilitated by Universita degli Studi di Genova, as part of the Wiley - CRUI-CARE agreement.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest. The authors have no financial, personal, or professional relationships that could influence the work they have submitted.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Maria Cristina Amoretti https://orcid.org/0000-0003-2892-577X
Daniele De Martini https://orcid.org/0000-0002-6937-5287

REFERENCES

G. Aad, T. Abajyan, B. Abbott, et al., *Observation of a new particle in the search for the standard model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. B **716** (2012), 1–29.

B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, V. B. Adya, C. Affeldt, M. Agathos, K. Agatsuma, N. Aggarwal, O. D. Aguiar, L. Aiello, A. Ain, P. Ajith, B. Allen, A. Allocca, P. A. Altin, S. B. Anderson, W. G. Anderson, K. Arai, M. A. Arain, M. C. Araya, C. C. Arceneaux, J. S. Areeda, N. Arnaud, K. G. Arun, S. Ascenzi, G. Ashton, M. Ast, S. M. Aston, P. Astone, P. Aufmuth, C. Aulbert, S. Babak, P. Bacon, M. K. Bader, P. T. Baker, F. Baldaccini, G. Ballardin, S. W. Ballmer, J. C. Barayoga, S. E. Barclay, B. C. Barish, D. Barker, F. Barone, B. Barr, L. Barsotti, M. Barsuglia, D. Barta, J. Bartlett, M. A. Barton, I. Bartos, R. Bassiri, A. Basti, J. C. Batch, C. Baune, V. Bavigadda, M. Bazzan, B. Behnke, M. Bejger, C. Belczynski, A. S. Bell, C. J. Bell, B. K. Berger, J. Bergman, G. Bergmann, C. P. Berry, D. Bersanetti, A. Bertolini, J. Betzwieser, S. Bhagwat, R. Bhandare, I. A. Bilenko, G. Billingsley, J. Birch, R. Birney, O. Birnholtz, S. Biscans, A. Bisht, M. Bitossi, C. Biwer, M. A. Bizouard, J. K. Blackburn, C. D. Blair, D. G. Blair, R. M. Blair, S. Bloemen, O. Bock, T. P. Bodiya, M. Boer, G. Bogaert, C. Bogan, A. Bohe, P. Bojtos, C. Bond, F. Bondu, R. Bonnand, B. A. Boom, R. Bork, V. Boschi, S. Bose, Y. Bouffanais, A. Bozzi, C. Bradaschia, P. R. Brady, V. B. Braginsky, M. Branchesi, J. E. Brau, T. Briant, A. Brillet, M. Brinkmann, V. Brisson, P. Brockill, A. F. Brooks, D. A. Brown, D. D. Brown, N. M. Brown, C. C. Buchanan, A. Buikema, T. Bulik, H. J. Bulten, A. Buonanno, D. Buskulic, C. Buy, R. L. Byer, M. Cabero, L. Cadonati, G. Cagnoli, C. Cahillane, J. Calderón Bustillo, T. Callister, E. Calloni, J. B. Camp, K. C. Cannon, J. Cao, C. D. Capano, E. Capocasa, F. Carbognani, S. Caride, J. Casanueva Diaz, C. Casentini, S. Caudill, M. Cavaglià, F. Cavalier, R. Cavalieri, G. Cella, C. B. Cepeda, L. Cerboni Baiardi, G. Cerretani, E. Cesarini, R. Chakraborty, T. Chalermsongsak, S. J. Chamberlin, M. Chan, S. Chao, P. Charlton, E. Chassande-Mottin, H. Y. Chen, Y. Chen, C. Cheng, A. Chincarini, A. Chiummo, H. S. Cho, M. Cho, J. H. Chow, N. Christensen, Q. Chu, S. Chua, S. Chung, G. Ciani, F. Clara, J. A. Clark, F. Cleva, E. Coccia, P. F. Cohadon, A. Colla, C. G. Collette, L. Cominsky, M. Constancio, A. Conte, L. Conti, D. Cook, T. R. Corbitt, N. Cornish, A. Corsi, S. Cortese, C. A. Costa, M. W. Coughlin, S. B. Coughlin, J. P. Coulon, S. T. Countryman, P. Couvares, E. E. Cowan, D. M. Coward, M. J. Cowart, D. C. Coyne, R. Coyne, K. Craig, J. D. Creighton, T. D. Creighton, J. Cripe, S. G. Crowder, A. M. Cruise, A. Cumming, L. Cunningham, E. Cuoco, T. Dal Canton, S. L. Danilishin, S. D'Antonio, K. Danzmann, N. S. Darman, C. F. Da Silva Costa, V. Dattilo, I. Dave, H. P. Daveloza, M. Davier, G. S. Davies, E. J. Daw, R. Day, S. De, D. DeBra, G. Debreczeni, J. Degallaix, M. De Laurentis, S. Deléglise, W. Del Pozzo, T. Denker, T. Dent, H. Dereli, V. Dergachev, R. T. DeRosa, R. De Rosa, R. DeSalvo, S. Dhurandhar, M. C. Díaz, L. Di Fiore, M. Di Giovanni, A. Di Lieto, S. Di Pace, I. Di Palma, A. Di Virgilio, G. Dojcinoski, V. Dolique, F. Donovan, K. L. Dooley, S. Doravari, R. Douglas, T. P. Downes, M. Drago, R. W. Drever, J. C. Driggers, Z. Du, M. Ducrot, S. E. Dwyer, T. B. Edo, M. C. Edwards, A. Effler, H. B. Eggenstein, P. Ehrens, J. Eichholz, S. S. Eikenberry, W. Engels, R. C. Essick, T. Etzel, M. Evans, T. M. Evans, R. Everett, M. Factourovich, V. Fafone, H. Fair, S. Fairhurst, X. Fan, Q. Fang, S. Farinon, B. Farr, W. M. Farr, M. Favata, M. Fays, H. Fehrmann, M. M. Fejer, D. Feldbaum, I. Ferrante, E. C. Ferreira, F. Ferrini, F. Fidecaro, L. S. Finn, I. Fiori, D. Fiorucci, R. P. Fisher, R. Flaminio, M. Fletcher, H. Fong, J. D. Fournier, S. Franco, S. Frasca, F. Frasconi, M. Frede, Z. Frei, A. Freise, R. Frey, V. Frey, T. T. Fricke, P. Fritschel, V. V. Frolov, P. Fulda, M. Fyffe, H. A. Gabbard, J. R.

Gair, L. Gammaitoni, S. G. Gaonkar, F. Garufi, A. Gatto, G. Gaur, N. Gehrels, G. Gemme, B. Gendre, E. Genin, A. Gennai, J. George, L. Gergely, V. Germain, A. Ghosh, S. Ghosh, J. A. Giaime, K. D. Giardina, A. Giazotto, K. Gill, A. Glaefke, J. R. Gleason, E. Goetz, R. Goetz, L. Gondan, G. González, J. M. Gonzalez Castro, A. Gopakumar, N. A. Gordon, M. L. Gorodetsky, S. E. Gossan, M. Gosselin, R. Gouaty, C. Graef, P. B. Graff, M. Granata, A. Grant, S. Gras, C. Gray, G. Greco, A. C. Green, R. J. Greenhalgh, P. Groot, H. Grote, S. Grunewald, G. M. Guidi, X. Guo, A. Gupta, M. K. Gupta, K. E. Gushwa, E. K. Gustafson, R. Gustafson, J. J. Hacker, B. R. Hall, E. D. Hall, G. Hammond, M. Haney, M. M. Hanke, J. Hanks, C. Hanna, M. D. Hannam, J. Hanson, T. Hardwick, J. Harms, G. M. Harry, I. W. Harry, M. J. Hart, M. T. Hartman, C. J. Haster, K. Haughian, J. Healy, J. Heefner, A. Heidmann, M. C. Heintze, G. Heinzel, H. Heitmann, P. Hello, G. Hemming, M. Hendry, I. S. Heng, J. Hennig, A. W. Heptonstall, M. Heurs, S. Hild, D. Hoak, K. A. Hodge, D. Hofman, S. E. Hollitt, K. Holt, D. E. Holz, P. Hopkins, D. J. Hosken, J. Hough, E. A. Houston, E. J. Howell, Y. M. Hu, S. Huang, E. A. Huerta, D. Huet, B. Hughey, S. Husa, S. H. Huttner, T. Huynh-Dinh, A. Idrisy, N. Indik, D. R. Ingram, R. Inta, H. N. Isa, J. M. Isac, M. Isi, G. Islas, T. Isogai, B. R. Iyer, K. Izumi, M. B. Jacobson, T. Jacqmin, H. Jang, K. Jani, P. Jaranowski, S. Jawahar, F. Jiménez-Forteza, W. W. Johnson, N. K. Johnson-McDaniel, D. I. Jones, R. Jones, R. J. Jonker, L. Ju, K. Haris, C. V. Kalaghatgi, V. Kalogera, S. Kandhasamy, G. Kang, J. B. Kanner, S. Karki, M. Kasprzack, E. Katsavounidis, W. Katzman, S. Kaufer, T. Kaur, K. Kawabe, F. Kawazoe, F. Kéfélian, M. S. Kehl, D. Keitel, D. B. Kelley, W. Kells, R. Kennedy, D. G. Keppel, J. S. Key, A. Khalaidovski, F. Y. Khalili, I. Khan, S. Khan, Z. Khan, E. A. Khazanov, N. Kijbunchoo, C. Kim, J. Kim, K. Kim, N. G. Kim, N. Kim, Y. M. Kim, E. J. King, P. J. King, D. L. Kinzel, J. S. Kissel, L. Kleybolte, S. Klimenko, S. M. Koehlenbeck, K. Kokeyama, S. Koley, V. Kondrashov, A. Kontos, S. Koranda, M. Korobko, W. Z. Korth, I. Kowalska, D. B. Kozak, V. Kringel, B. Krishnan, A. Królak, C. Krueger, G. Kuehn, P. Kumar, R. Kumar, L. Kuo, A. Kutynia, P. Kwee, B. D. Lackey, M. Landry, J. Lange, B. Lantz, P. D. Lasky, A. Lazzarini, C. Lazzaro, P. Leaci, S. Leavey, E. O. Lebigot, C. H. Lee, H. K. Lee, H. M. Lee, K. Lee, A. Lenon, M. Leonardi, J. R. Leong, N. Leroy, N. Letendre, Y. Levin, B. M. Levine, T. G. Li, A. Libson, T. B. Littenberg, N. A. Lockerbie, J. Logue, A. L. Lombardi, L. T. London, J. E. Lord, M. Lorenzini, V. Loriette, M. Lormand, G. Losurdo, J. D. Lough, C. O. Lousto, G. Lovelace, H. Lück, A. P. Lundgren, J. Luo, R. Lynch, Y. Ma, T. MacDonald, B. Machenschalk, M. MacInnis, D. M. Macleod, F. Magaña-Sandoval, R. M. Magee, M. Mageswaran, E. Majorana, I. Maksimovic, V. Malvezzi, N. Man, I. Mandel, V. Mandic, V. Mangano, G. L. Mansell, M. Manske, M. Mantovani, F. Marchesoni, F. Marion, S. Márka, Z. Márka, A. S. Markosyan, E. Maros, F. Martelli, L. Martellini, I. W. Martin, R. M. Martin, D. V. Martynov, J. N. Marx, K. Mason, A. Masserot, T. J. Massinger, M. Masso-Reid, F. Matichard, L. Matone, N. Mavalvala, N. Mazumder, G. Mazzolo, R. McCarthy, D. E. McClelland, S. McCormick, S. C. McGuire, G. McIntyre, J. McIver, D. J. McManus, S. T. McWilliams, D. Meacher, G. D. Meadors, J. Meidam, A. Melatos, G. Mendell, D. Mendoza-Gandara, R. A. Mercer, E. Merilh, M. Merzougui, S. Meshkov, C. Messenger, C. Messick, P. M. Meyers, F. Mezzani, H. Miao, C. Michel, H. Middleton, E. E. Mikhailov, L. Milano, J. Miller, M. Millhouse, Y. Minenkov, J. Ming, S. Mirshekari, C. Mishra, S. Mitra, V. P. Mitrofanov, G. Mitselmakher, R. Mittleman, A. Moggi, M. Mohan, S. R. Mohapatra, M. Montani, B. C. Moore, C. J. Moore, D. Moraru, G. Moreno, S. R. Morriss, K. Mossavi, B. Mours, C. M. Mow-Lowry, C. L. Mueller, G. Mueller, A. W. Muir, A. Mukherjee, D. Mukherjee, S. Mukherjee, N. Mukund, A. Mullavey, J. Munch, D. J. Murphy, P. G. Murray, A. Mytidis, I. Nardecchia, L. Naticchioni, R. K. Nayak, V. Necula, K. Nedkova, G. Nelemans, M. Neri, A. Neunzert, G. Newton, T. T. Nguyen, A. B. Nielsen, S. Nissanke, A. Nitz, F. Nocera, D. Nolting, M. E. Normandin, L. K. Nuttall, J. Oberling, E. Ochsner, J. O'Dell, E. Oelker, G. H. Ogin, J. J. Oh, S. H. Oh, F. Ohme, M. Oliver, P. Oppermann, R. J. Oram, B. O'Reilly, R. O'Shaughnessy, C. D. Ott, D. J. Ottaway, R. S. Ottens, H. Overmier, B. J. Owen, A. Pai, S. A. Pai, J. R. Palamos, O. Palashov, C. Palomba, A. Pal-Singh, H. Pan, Y. Pan, C. Pankow, F. Pannarale, B. C. Pant, F. Paoletti, A. Paoli, M. A. Papa, H. R. Paris, W. Parker, D. Pascucci, A. Pasqualetti, R. Passaquieti, D. Passuello, B. Patricelli, Z. Patrick, B. L. Pearlstone, M. Pedraza, R. Pedurand, L. Pekowsky, A. Pele, S. Penn, A. Perreca, H. P. Pfeiffer, M. Phelps, O. Piccinni, M. Pichot, M. Pickenpack, F. Piergiovanni, V. Pierro, G. Pillant, L. Pinard, I. M. Pinto, M. Pitkin, J. H. Poeld, R. Poggiani, P. Popolizio, A. Post, J. Powell, J. Prasad, V. Predoi, S. S. Premachandra, T. Prestegard, L. R. Price, M. Prijatelj, M. Principe, S. Privitera, R. Prix, G. A. Prodi, L. Prokhorov, O. Puncken, M. Punturo, P. Puppo, M. Pürrer, H. Qi, J. Qin, V. Quetschke, E. A. Quintero, R. Quitzow-James, F. J. Raab, D. S. Rabeling, H. Radkins, P. Raffai, S. Raja, M. Rakhmanov, C. R. Ramet, P. Rapagnani, V. Raymond, M. Razzano, V. Re, J. Read, C. M. Reed, T. Regimbau, L. Rei, S. Reid, D. H. Reitze, H. Rew, S. D. Reyes, F. Ricci, K. Riles, N. A. Robertson, R. Robie, F. Robinet, A. Rocchi, L. Rolland, J. G. Rollins, V. J. Roma, J. D. Romano, R. Romano, G. Romanov, J. H. Romie, D. Rosińska, S. Rowan, A. Rüdiger, P. Ruggi, K. Ryan, S. Sachdev, T. Sadecki, L. Sadeghian, L. Salconi, M. Saleem, F. Salemi, A. Samajdar, L. Sammut, L. M. Sampson, E. J. Sanchez, V. Sandberg, B. Sandeen, G. H. Sanders, J. R. Sanders, B. Sassolas, B. S. Sathyaprakash, P. R. Saulson, O. Sauter, R. L. Savage, A. Sawadsky, P. Schale, R. Schilling, J. Schmidt, P. Schmidt, R. Schnabel, R. M. Schofield, A. Schönbeck, E. Schreiber, D. Schuette, B. F. Schutz, J. Scott, S. M. Scott, D. Sellers, A. S. Sengupta, D. Sentenac, V. Sequino, A. Sergeev, G. Serna, Y. Setyawati, A. Sevigny, D. A. Shaddock, T. Shaffer, S. Shah, M. S. Shahriar, M. Shaltev, Z. Shao, B. Shapiro, P. Shawhan, A. Sheperd, D. H. Shoemaker, D. M. Shoemaker, K. Siellez, X. Siemens, D. Sigg, A. D. Silva, D. Simakov, A. Singer, L. P. Singer, A. Singh, R. Singh, A. Singhal, A. M. Sintes, B. J. Slagmolen, J. R. Smith, M. R. Smith, N. D. Smith, R. J. Smith, E. J. Son, B. Sorazu, F. Sorrentino, T. Souradeep, A. K. Srivastava, A. Staley, M. Steinke, J. Steinlechner, S. Steinlechner, D. Steinmeyer, B. C. Stephens, S. P. Stevenson, R. Stone, K. A. Strain, N. Straniero, G. Stratta, N. A. Strauss, S. Strigin, R. Sturani, A. L. Stuver, T. Z. Summerscales, L. Sun, P. J. Sutton, B. L. Swinkels, M. J. Szczepańczyk, M. Tacca, D. Talukder, D. B. Tanner, M. Tápai, S. P. Tarabrin, A. Taracchini,

R. Taylor, T. Theeg, M. P. Thirugnanasambandam, E. G. Thomas, M. Thomas, P. Thomas, K. A. Thorne, K. S. Thorne, E. Thrane, S. Tiwari, V. Tiwari, K. V. Tokmakov, C. Tomlinson, M. Tonelli, C. V. Torres, C. I. Torrie, D. Töyrä, F. Travasso, G. Traylor, D. Trifirò, M. C. Tringali, L. Trozzo, M. Tse, M. Turconi, D. Tuyenbayev, D. Ugolini, C. S. Unnikrishnan, A. L. Urban, S. A. Usman, H. Vahlbruch, G. Vajente, G. Valdes, M. Vallisneri, N. van Bakel, M. van Beuzekom, J. F. den Brand, C. Van Den Broeck, D. C. Vander-Hyde, L. van der Schaaf, J. V. van Heijningen, A. A. Veggel, M. Vardaro, S. Vass, M. Vasúth, R. Vaulin, A. Vecchio, G. Vedovato, J. Veitch, P. J. Veitch, K. Venkateswara, D. Verkindt, F. Vetrano, A. Viceré, S. Vinciguerra, D. J. Vine, J. Y. Vinet, S. Vitale, T. Vo, H. Vocca, C. Vorvick, D. Voss, W. D. Vousden, S. P. Vyatchanin, A. R. Wade, L. E. Wade, M. Wade, S. J. Waldman, M. Walker, L. Wallace, S. Walsh, G. Wang, H. Wang, M. Wang, X. Wang, Y. Wang, H. Ward, R. L. Ward, J. Warner, M. Was, B. Weaver, L. W. Wei, M. Weinert, A. J. Weinstein, R. Weiss, T. Welborn, L. Wen, P. Weßels, T. Westphal, K. Wette, J. T. Whelan, S. E. Whitcomb, D. J. White, B. F. Whiting, K. Wiesner, C. Wilkinson, P. A. Willems, L. Williams, R. D. Williams, A. R. Williamson, J. L. Willis, B. Willke, M. H. Wimmer, L. Winkelmann, W. Winkler, C. C. Wipf, A. G. Wiseman, H. Wittel, G. Woan, J. Worden, J. L. Wright, G. Wu, J. Yablon, I. Yakushin, W. Yam, H. Yamamoto, C. C. Yancey, M. J. Yap, H. Yu, M. Yvert, A. Zadrożny, L. Zangrando, M. Zanolin, J. P. Zendri, M. Zevin, F. Zhang, L. Zhang, M. Zhang, Y. Zhang, C. Zhao, M. Zhou, Z. Zhou, X. J. Zhu, M. E. Zucker, S. E. Zuraw, and J. Zweizig, *Observation of gravitational waves from a binary black hole merger*, Phys. Rev. Lett. **116** (2016), 061102.

Anonymous, *Journals unite for reproducibility*, Nature **515** (2014), 7.

S. Antonini and M. A. Mariotti, *Indirect proof: what is specific to this way of proving?* ZDM **40** (2008), 401–412.

C. Batanero, *Controversies around significance tests*, Math. Think. Learn. **2** (2000), 75–98.

C. Batanero and C. Díaz, *Methodological and didactical controversies around statistical inference, Actes du 36ièmes Journées de la Société Française de Statistique*, Société Française de Statistique, Paris, 2006.

I. Böschen, *Changes in methodological study characteristics in psychology between 2010-2021*, PLoS One **18** (2023), e0283353.

A. E. Castro Sotos, S. Vanhoof, W. van den Noortgate, and P. Onghena, *Students' misconceptions of statistical inference: a review of the empirical evidence from research on statistics education*, Educ. Res. Rev. **2** (2007), 98–113.

D. Chavalarias, J. D. Wallach, A. H. T. Li, and J. P. A. Ioannidis, *Evolution of reporting P values in the biomedical literature, 1990-2015*, Jama **315** (2016), 1141–1148.

L. De Capitani and D. De Martini, *On stochastic orderings of the Wilcoxon rank sum test statistic, with applications to reproducibility probability estimation testing*, Stat. Probab. Lett. **81** (2011), no. 8, 937–946.

L. De Capitani and D. De Martini, *Reproducibility probability estimation and testing for the Wilcoxon rank-sum test*, J. Stat. Comput. Simul. **85** (2015), 468–493.

L. De Capitani and D. De Martini, *Reproducibility probability estimation and RP testing for some nonparametric tests*, Entropy **18** (2016), no. 4, 142.

L. De Capitani and D. De Martini, *Improving reproducibility probability estimation and preserving RP-testing*, Stat. Methods Appl. **30** (2021), 49–77.

D. De Martini, *Reproducibility probability estimation for testing statistical hypotheses*, Stat. Probab. Lett. **78** (2008), 1056–1061.

H. E. Douglas, *Inductive risk and values in science*, Philos. Sci. **67** (2000), 559–579.

R. Falk and C. W. Greenbaum, *Significance tests die hard – the amazing persistence of a probabilistic misconception*, Theory Psychol. **5** (1995), 75–98.

R. Fisher, *The design of experiments*, Oliver and Boyd, Edinburgh, London, 1935.

I. Gal, "*Statistical literacy: conceptual and instructional issues*," *Perspectives on adults learning mathematics: Theory and practice*, Kluwer Academic Publishers, New York, Boston, Dordrecht, London, Moscow, 2000.

I. Gal, *Teaching for statistical literacy and services of statistics agencies*, Am. Stat. **57** (2003), 80–84.

S. N. Goodman, *A comment on replication, p-values and evidence*, Stat. Med. **11** (1992), 875–879.

S. Goodman, *A dirty dozen: twelve P-value misconceptions*, Semin. Hematol. **45** (2008), 135–140.

T. S. Kuhn, *The structure of scientific revolutions*, University of Chicago Press, Chicago, IL, 1970.

E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*, Springer, Cham, 2022.

Y. Liu and P. W. Thompson, "*Teachers' understanding of hypothesis testing*," *Proceedings of the Twenty-Seventh Annual Meeting of the International Group for the Psychology of Mathematics Education, Roanoke, VA*, S. Wilson (ed.), Virginia Tech, Vicksburg, VA, 2005 [online]. http://pat-thompson.net/PDFversions/2005PMENA%20HypTest.pdf.

M. McNutt, *Journals unite for reproducibility*, Science **346** (2014), 679.

National Academy of Sciences—Engineering—Medicine, *Consensus Study Report, Reproducibility and Replicability in Science*, The National Academies Press, Washington, DC, 2019.

H. Otani, *Comparing structures of SHT with proof by contradiction in terms of argument*, Hiroshima J. Math. Educ. **12** (2019), 1–12.

C. A. Reeves and J. K. Brewer, *Hypothesis testing and proof by contradiction: an analogy*, Teach. Stat. **2** (1980), 57–59.

D. B. Rubin, *Teaching statistical inference for causal effects in experiments and observational studies*, J. Educ. Behav. Stat. **29** (2004), 343–367.

T. Wilholt, *Bias and values in scientific research*, Stud. Hist. Philos. Sci. A **40** (2009), 92–101.

How to cite this article: M. C. Amoretti and D. De Martini, *Analogies and differences between the logic behind statistical hypothesis testing and proofs by contradiction: What can we learn from them?* Teach. Stat. (2026), 1–13, DOI 10.1002/test.70033.

APPENDIX 1: REPEATABILITY PROBABILITY ESTIMATION

The Repeatability Probability (RP) is the so-called "true power," that is the power function (or functional) π computed at the true, albeit unknown, value of the parameter (or distribution) under test (De Martini, 2008). Referring to Example 1.1, for given values of α and n, π is a function of the parameter under test θ, that is, $\pi(\alpha, n, \theta)$, and the true power is $\pi(\alpha, n, \theta_t)$, where θ_t denotes the unknown (true) value of θ. Hence, the true power can be viewed as the probability of observing a statistical significant result in an identical replication of the experiment, that is: $RP = \pi(\alpha, n, \theta_t)$. Note that RP is unknown, but it can be estimated. A naïve estimator of RP is obtained by plugging a point estimate of θ_t (i.e., $\theta\hat{}$) into the power function: $RP\hat{} = \pi(\alpha, n, \theta\hat{})$. A more informative and practically useful estimator is obtained by plugging-in the median-unbiased estimator of θ, that is $\theta^{0.5}$ such that $P(\theta^{0.5} \leq \theta) = 0.5$; in practice $\theta^{0.5}$ is the 0.5 lower bound of a one-sided confidence interval for θ. In this way, we obtain $RP^{0.5} = \pi(\alpha, n, \theta^{0.5})$. It is worth noting that $RP^{0.5}$ fulfills the so-called RP-testing theorem: A statistical test is significant if and only if $RP^{0.5} > 1/2$. In a general parametric context (see De Martini, 2008) θ represents the noncentrality parameter of the distribution of the test statistic T, introduced in point 2 of Section 2.

For nonparametric tests, the power is a functional of the population distribution F, so that $RP = \pi(\alpha, n, F_t)$. Given the empirical distribution function $F\hat{}$, based on sample data, a bootstrap estimator analogous to $RP^{0.5}$, say $RP^{0.5,BO}$, is defined as the median of the distribution of $\pi(\alpha, n, F^*)$, where F^* denotes the empirical distribution function of a sample of size n drawn with replacement from $F\hat{}$. Moreover, some useful semiparametric RP estimators have been proposed for specific nonparametric tests (De Capitani & De Martini, 2015, 2016, 2011).

Recently, a class of improved RP estimators has been developed that consistently reduces the mean squared error (De Capitani & De Martini, 2021). One such estimator is the Average Conservative estimator, $RP^{AC} = \int_{(0,1)} \pi(\alpha, n, \theta^{\gamma})\, d\gamma$, where θ^{γ} is the γ-conservative estimator of θ, that is, the γ-lower bound of a one-sided confidence interval.