A Generalization of the Zero-Probability Theorem

Miodrag M. LOVRIC

mlovric@radford.edu

Radford University, VA, USA

ABSTRACT: *In this paper, a generalized form of the Zero-Probability Theorem (initially called a paradox) is established. Originally, it was proved for the standard hypothesized form that contains a single number. New version demonstrates that when testing a mean of the normal population using a point null hypothesis which is formulated as a set of all countably infinite algebraic numbers, the probability of such hypothesis is zero within the set of all real numbers. This result shows that point-null hypothesis paradigm is based on flimsy foundation and points to one of the most important root causes of the current reproducibility crisis in science. We explore implications of this theorem on the Fisherian significance testing, Neyman-Pearson hypothesis testing, and Bayesian testing based upon the Bayes factor. We recommend that point-null hypotheses of a normal mean should be immediately abandoned. For decades, their testing has produced countless criticisms and recently even methodological crises in many fields of science and has done serious damage to the image of statistics and statisticians. The new model should be based on the negligible null hypotheses accompanied by the practically relevant alternative hypotheses. We regard this simply achieved modification as a new paradigm in the Kuhn's sense. Enough arguments are provided to confirm that this will not only eradicate most of the objections against frequentist testing and breathe a new life into them, but also considerably reconcile communication in inference between frequentist and Bayesian approaches. In conclusion, no frequentist nor Bayesian should test point-null hypotheses for continuous parameters*.

1. Introduction

> "*We will all be Bayesians in 2020, and then we can be a united profession*." D.V. Lindley's 1995 interview with A.F.M Smith, Statistical Science.
>
> "*I have lamented that Bayesian statisticians do not stick closely enough to the pattern laid down by Bayes himself: if they would only do as he did and publish posthumously we should all be saved a lot of trouble*." [M. Kendall, On the Future of Statistics, JRSS(A), (1968), 131, 182-204].

The first formal significance test was conducted by Arbuthnott [4]. From today's perspective, it can be argued that this was an auspicious event in statistics history. However, at the same time, Arbuthnott opened Pandora's box foreshadowing the controversies about the role of statistical tests. From one standpoint, he correctly analyzed data on the yearly number of male and female christenings in London from 1629 to 1710 and demonstrated that males were born at a greater rate than females. Notwithstanding this distinguished achievement, this is also the first recorded case of confusing statistical with scientific hypotheses. That is, Arbuthnott paralleled mere rejection of a null hypothesis with an irrefutable statistical argument for divine providence. Moreover, his approach in testing was without delay challenged by several scientists, including W.J. 'sGravesande, Nicholas Bernoulli and de Moivre (see [45] pp. 275-285.).

Broadly speaking present-day hypothesis testing can be conducted in two methodologically different ways, using frequentist concepts and using Bayesian perceptions. Frequentist concepts are based on the idea that probability is a limiting relative frequency and Bayesians rely on the notion that probability is a degree of personal belief that some event will occur. The pivotal indicator in modern frequentist testing is a p-value (see, for example, [19]), and Bayesian testing is founded on

posterior probabilities and Bayes factor ([55], [56], [58], [41], [48]). Frequentist statisticians are currently still in the majority regardless of the famous prediction made by Lindley [95] given above.

One of the most fundamental problems in Statistical science is that we can analyze the same dataset using those two different concepts and reach divergent results. More importantly, as shown for example in Lovric [66], as the sample size increases frequentist tests will tend to reject point (sharp) null hypothesis and Bayesian testing will incline to support the same null hypothesis. This is the essence of the famous Jeffreys-Lindley paradox. For the past sixty years many statisticians have been trying to find a cure and reunite Bayesian and frequentist inference but without success. Recently, Gelman and Shalizi [32] concluded that the Jeffreys-Lindley paradox is really a problem without a solution.

Frequentist statistical tests are usually regarded as an anonymous hybrid of two divergent classical statistical paradigms. Fisherian *significance testing* is founded on a single null hypothesis, p values, inductive reasoning, and drawing conclusions. By contrast, Neyman-Pearson *hypothesis testing* is established on two hypotheses: null and an alternative, two types of errors, fixed-level significance statements, making decisions, and deductive reasoning but inductive behavior. These opposing views about the proper manner to conduct a test were never reconciled by their authors, nonetheless, been amalgamated by contemporary authors of statistics textbooks. As pointed out by Berger [10, p. 4] "disagreement between Fisher and Neyman has had a significantly deleterious effect on the practice of statistics in science, essentially because it has led to widespread confusion and inappropriate use of testing methodology in the scientific community." In contrast, Lehmann [60] argues that the differences between Fisherian and Neyman-Pearson testing approaches to testing are largely rhetorical rather than substantial and that two theories are complementary rather than contradictory.

Frequentist statistical testing has become extensively accepted by almost all researchers as the most common statistical inferential approach in almost all fields of science. However, over the past 80 years, numerous objections and severe criticisms have been raised against their usefulness to the point that they should be banned. Many critics also emphasize that statistical tests are very often misinterpreted and misused. Specifically, there is almost universal confusion over the interpretation of the p-values and significance levels. P-value is one of the most pervasive and at the same time, misapprehended statistical indicators in scientific research (for a review of twelve typical misconceptions, see Goodman [42]). We will illustrate this statement with the following two examples.

(1) In the recent paper written by seven eminent world statisticians (Greenland, Senn, Rothman, Carlin, Poole, Goodman, and Altman 2016) with the title "Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations", authors have provided an explanatory list of 25 misinterpretations of p values. To educate their readers they have given their "correct" definition (p. 339): "The p value is then the probability that the chosen test statistic would have been at least as large as its observed value if every model assumption were correct, including the test hypothesis". However, this definition is very wrong since it includes only right-tailed tests and should be included in their list. Dr. Greenland admitted that they made a mistake, but regretfully that paper was cited by 1594 authors and downloaded 169,000 times, apparently nobody noticed this vast error.
(2) Haller and Krauss [46] asked participants from psychology departments in 6 German universities to fill out a short questionnaire in order to assess their functional knowledge of p-values. They discovered that 100% of psychology students lack this understanding, 97% of

academic psychologists don't understand p-values, and the most alarming finding was that 80% of statistics instructors misunderstand p-values.

The remaining of this paper is structured as follows. In the next section, we provide a brief overview of the main objections against point null-hypothesis testing. Section 3 is the heart of the paper. We focus on the inferential aspects to the problem of testing a point-null hypothesis of a normal mean. In that section, we prove the Generalized Zero probability theorem. This theorem states that in testing a mean of the normal population, even when a null hypothesis is formulated as a set of all countable infinite algebraic numbers, it has zero probability. In light of this, in Section 4 we briefly explore the consequences of the Zero theorem on Fisherian, Neyman-Pearson, and in the Bayesian approach in testing. We show that the dismissal of the famous Jeffreys-Lindley paradox is one of the direct implications of the theorem. In the last section, we propose a paradigm shift in the foundation of statistical science. We give arguments to confirm our thesis that this is a paradigm shift in Kuhn's sense. It will not only eliminate most of the objections against frequentist testing but also considerably reconcile communication in inference between frequentist and Bayesian testing methodologies.

It is, however, important to note that this paper focuses only on testing of a parameter in the case of one sample, based upon the model that assumes absolutely continuous distribution with respect to the Lebesgue measure λ.

1. Statistical crisis in science

Since Buchanan-Wollaston [15] raised the first criticism against significance testing, this foundational field of statistics has drawn increasingly active and stronger opposition in numerous fields of science, including the prohibition of statistical inference [99], that the term "statistical significance" [102] should be banned, and p-values abandoned [11]. We regard that majority[1] of researchers and statisticians are incognizant of the dimension and intensity of this problem. Therefore, we consider that it is essential to provide a short compendium of selected references (from the voluminous literature, enormous in breadth and details) in which notable critics have raised their voices against significance testing, from the following scientific disciplines:

> *Accounting* [65], *Atmospheric research* [78], *climate science* [3], *clinical medicine and epidemiology* ([84], [40], [97]), *consumer research* [53], *criminology* [16], *ecology and evolutionary biology* [98], *economics* ([69], [70]), *ergonomics* [100], *forecasting* [5], *forest sciences* [26], *management science* [91], *marine ecology* [8], *marketing* [51], *neuroscience* [49], *organizational science* [92], *political sciences* [36], *psychology and education* ([86], [6], [17], [72], [80], [35], [22], [27], [89], [46], [67]), *road safety research* [47], *scieontometrics* [90], *sociology* [74], and *wildlife field* [57].

Some of the main arguments against significance testing can be found in the following list. We argue that almost all of these objections can be ruled out with the wide approval of a new proposal to make a shift to the interval hypothesis testing paradigm.

1. The dichotomous reject/fail to reject criteria is arbitrary and contributes to black-and-white thinking.

[1] For example, Spanos [96] p. 645) claims that "the use, abuse, interpretations and reinterpretations of the notion of a P value has been a hot topic of controversy since the 1950s in statistics and *several* [italicized by the author] applied fields, including psychology, sociology, ecology, medicine, and economics."

2. A specific null hypothesis is almost always nil null (that is specified as a point null).
3. The null hypothesis cannot be literally true.
4. Traditional testing is a "trivial exercise," because the null hypothesis can always be rejected, given a large enough sample size.
5. Apparent validity of findings depends on researchers' efforts to obtain enough data.
6. Traditional testing obscure important findings.
7. The vast majority of null hypotheses are false and scientifically irrelevant.
8. "$p < .05$" does not actually refer to anything very interesting.
9. Traditional significance testing ignores effect size.
10. It can, at best, only test statistical (not substantive) null hypotheses.
11. Traditional significance testing leads to misinterpretation of results.
12. Focus on $\alpha=0.05$ ignores or leads to low power.
13. Procedure does not tell researchers what they want to know, that is the probability that the null hypothesis is true given that we have obtained a set of data.
14. A common misconception involves interpreting statistical significance as theoretical or practical significance.
15. Traditional testing highlights trivial findings.
16. Increases type-I error rate in published papers.
17. Contributes substantially to publication bias.
18. Promotes arbitrary data dredging ("p-value fishing").
19. Leads to erosion of researchers' devotion since repeated and very public misuse of testing s creates cynicism and confusion.
20. Misinterpretation of statistical non-significant result as evidence that the null hypothesis is true.

Many non-statisticians advocate the reform of statistical inference and statistics education. They claim that less emphasis should be placed upon the reporting of *p* values and more on effect size, confidence intervals, use of information-theoretic approaches, and Bayesian inference. Nevertheless, they are aware that neither reliance on confidence intervals nor Bayesian inference will suffice to preclude injudicious statistical conclusions. As Abelson ([2] p. 13) warned, "Under the Law of Diffusion of Idiocy, every foolish application of significance testing will beget a corresponding foolish practice for confidence limits."

By contrast, there are also some examples of papers that support null hypothesis statistical testing, including Fleiss [29], Frick [30], Abelson [1], Hagen [44], Cortina and Dunlap [24], Chow [20], Hagen [44], Mulaik, Raju, and Harshman [74], Wainer and Robinson [101], Dennis [25], Mogie [73], Sawilowsky [88], and Murtaugh [76]. Some of the defenders assert that a great deal of the criticism concerning these problems "is related to the misapplication and misunderstanding of NHST [Null Hypothesis Significance Testing] and, specifically, the use of *p* values by researchers, editors, and consumers of research" (LeMire [61]). Senn [93] thinks that *p* values can have some limited usage, but that "Bayesians in particular, find them ridiculous" (p. 193) and that "*p*-values are a practical success but a critical failure." Gibson [34] argues that the reproducibility crisis in modern science is not the fault of p-values, but rather mainly a consequence of unwarranted optimism. Begleya and Snapinn [7] agree with Gibson and point out that p-values are not the cause of the problem but a tool that is frequently used inappropriately. In contrast, the substantial support to frequentist reasoning is maintained by the error statistics paradigm accompanied by severity testing established by Mayo and Spanos [68].

We regard that we have rendered enough arguments to ascertain that there is a *profound statistical crisis in* science. The most obvious reflection of this is the so-called reproducibility (replicability) crisis in modern science. This was initiated by Ioannidis [54] in his famous paper titled "Why most published research findings are false". He pointed out that published laboratory research findings were found not to being repeatable when researchers tried to follow them up. Recently, there has been an explosion of papers about the origins of the reproducibility crisis and many tend to blame the widespread use of p-values (for example, Branch [14]).

We further illustrate our thesis of statistical crisis with the following three quotes:

> "It's science's dirtiest secret: The 'scientific method' of testing hypotheses by statistical analysis stands on a flimsy foundation… As a result, countless conclusions in the scientific literature are erroneous." (Siegfried, Science News [94])

> "Despite the …cautions about *p* values not being Type I error rates, it is sobering to note that even well-known statisticians such as Barnard (1985), Gibbons and Pratt (1975) and Hinkley (1987) nevertheless make the mistake of equating them… Lehmann (1993) similarly fails to distinguish between measures of evidence versus error." (Hubbard [50], p. 307).

> "*Since professional statisticians are among those who do not understand these tests* [italicized by the author], no one should be surprised to discover widespread confusion about NHST, in the public and by the people who have studied statistics…this combination of inherent limitations [addressed on the previous page] and inappropriate applications of NHST impedes the accumulation of knowledge" (Schwab and Starbuck [91] p. 31).

We argue that this crisis and almost universal misapprehension of hypothesis testing and especially *p* values produced a very negative image of statistics and statisticians. It is therefore imperative to provide appropriate answers to the highly unsatisfied scientific researchers. In principle, there are four possible actions based on four different type of thinking.

(1) *Status quo ante*: turn a blind eye on the problem;

(2) *Draconian*: Proclaim that traditional tests are deficient and throw them away. Furthermore, abandon all other frequentist methods, as being advocated by Lindley ([63] p. 112), "What shall we do with sampling theory statistics, with significance tests, with confidence intervals; with all those methods that violate the likelihood principle? The answer is, let them die… I see no excuse for wasting our time on them except in a course on the history of our subject… How about a moratorium on research for two years? In the first of these we will all read de Finetti's first volume: the next year will do for the second."

(3) *Integrative*: provide the scientific world with a unified frequentist/Bayesian testing methodology (for example, Berger et al. [9], Berger [10]; and

(4) *Lateral*: Present a solution that appears as "obvious" in hindsight; modify traditional tests in such a way to eliminate as many criticisms as possible and upswing their public stature.

The aim of this paper is to propose lateral thinking. We will show that proof of a seemingly impossible theorem logically leads towards the conclusion that point-null hypotheses are meaningless in real-life research. This theorem implicitly suggests that the main sources of problems with significance tests are caused by the point-null hypotheses. We argue that point-null hypotheses should

be abandoned in science, not the significance testing. We, therefore, propose that point-null hypotheses have to be replaced by interval nulls, or negligible nulls. Paradoxically, in turn, this will not only eradicate most of the arguments against frequentist testing and breathe a new life into them, but also considerably reconcile communication in inference between frequentist and Bayesian approaches. We regard this as a direct answer to the challenge posed by Schwab ([92] p. 1117) "No one has proposed changes to NHSTs that purport to correct the main problems".

2. General zero probability theorem

In this section, we prove that in real-world research, the probability of an exact formulation ("guessing") of a mean of a normally distributed population is zero. We show that even if we state a null hypothesis of the normal mean not as a single rational number but as a collection of all countable infinite set of all rationals, the probability of such a null hypothesis is still zero. The General Zero Probability theorem is an extension of the Zero Probability paradox enunciated by C.R. Rao and Lovric in 2016. In this paper, authors have observed a null hypothesis that is made of a single number. Here, we will extend the null by including all rational numbers. In other words, the null hypothesis will state that the population means can be any rational number.

Certainly, this kind of infinitely integrated null hypothesis cannot be regarded as an orthodox point null. Nevertheless, we provide this generalized proof since intuition could suggest that a null hypothesis that is composed of all point null hypotheses that have been formulated so far by researchers "significantly" increases the odds for the null in comparison to the alternative. Once we have proven this generalized case, by straightforward deduction, it is easy to see that a point null that includes just a single hypothesized value also has probability zero to be true. In addition, we focus our attention only on the inferential aspect of the problem, not on the decision-making approach.

Many statisticians and non-statisticians stated the claim that in reality, point-null hypotheses are almost always false (Good [39], Savage [87], Nunnally [79], Meehl [71], Cohen [21], Ghosh et al. [33], Berger and Delampady [11], Krueger [59]). However, they supported this statement only by intuitive arguments and common sense.

To fully understand the proof, we recommend a prior reading of an excellent paper "Types of infinity" by Cook and Bossé [23].

General Zero Probability theorem. Suppose, that a random sample of size n, $X = (X_1, X_2, \ldots, X_n)$, is selected from the normal population $N(\theta, \sigma^2)$, where θ is an unknown mean that can assume values in a parameter space $\Theta \subset \mathbb{R}^1$. Without loss of generality, suppose that the variance, $\sigma^2 > 0$ is known. Let $\mathbb{Q}$ be the set of all rational real numbers and $\mathbb{Z}$ the set of all integers. Split parameter space into two disjoints sets $\Theta_{\mathbb{Q}}$ and $\Theta_{\mathbb{R}\setminus\mathbb{Q}}$ that are mutually exclusive $(\Theta_{\mathbb{Q}} \cap \Theta_{\mathbb{R}\setminus\mathbb{Q}} = \varnothing)$ and exhaustive $(\Theta = \Theta_{\mathbb{Q}} \cup \Theta_{\mathbb{R}\setminus\mathbb{Q}})$. From this we see that the set $\Theta_{\mathbb{Q}}$ represents the set of all population means expressed as rational numbers $\mathbb{Q}$ and $\Theta_{\mathbb{R}\setminus\mathbb{Q}}$ set of all the means that are irrational numbers $\mathbb{R}\setminus\mathbb{Q}$.

It is required to test the composite null hypothesis $H_0 : \theta \in \Theta_{\mathbb{Q}}$ versus an unspecified alternative hypothesis $H_1 : \theta \notin \Theta_{\mathbb{Q}}$, that is $H_1 : \theta \in \Theta_{\mathbb{R}\setminus\mathbb{Q}}$. Then, the probability of this composite null hypothesis (that encompasses all normal means expressed as rational numbers) is equal to zero:

$$\mathrm{P}(\{H_0 \mid \forall \theta \in \Theta_{\mathbb{Q}}\})=0$$

This is equivalent to saying that probability of the null hypothesis

$$P(H_0 : \theta \in \Theta_{\mathbb{Q}})= 0, \text{ or } P(H_0 : \theta \text{ is any rational number}) = 0 \text{ , and}$$

$$P(H_1 : \theta \in \Theta_{R\setminus\mathbb{Q}})= 1, \text{ or } P(H_1 : \theta \text{ is any irrational number}) = 1$$

As stated previously we regard rational numbers on the number line as pointers of the means of the matched normal populations that have rational numbers as their means.

Proof:

In real-life research any point null hypothesis in the standard normal model is almost always stated as a single rational number. This statement can be further extended by including any algebraic number (like all roots of integers). Cantor proved that there is in one-one correspondence between rational numbers and natural numbers. In other words, rational numbers are countable. Therefore, we may enumerate them as a sequence $\{q_i\}$, or $\mathbb{Q} = \bigcup_{i=1}^{\infty}\{q_i\}$. In hypothesis testing this is tantamount to saying that the set of all point null hypothesized values, expressed as rational numbers, $\Theta_{\mathbb{Q}}$, is also countable.

Since every countable set has Lebesgue measure zero, Lebesgue measure of the set of all rationals is also zero, that is $\lambda(\mathbb{Q}) = \lambda(\bigcup_{i=1}^{\infty}\{q_i\}) = \sum_{i=1}^{\infty}\lambda(\{q_i\} = 0.$ Therefore, Lebesgue measure of the set of all point null hypothesis stated as rationals is also zero because this set is countable, $\lambda(\Theta_{\mathbb{Q}}) = 0.$

It is well-known that the Normal distribution is absolutely continuous with respect to the Lebesgue measure λ. This implies that all sets which have zero Lebesgue measure must also have zero probability under probability measure. Since for an absolutely continuous distribution, a countably infinite set of all rational numbers has Lebesgue measure zero we conclude that its probability measure is also zero.

Finally, the probability measure for a composite null hypothesis that is composed of all rational numbers in testing a normal mean is zero, $\mathrm{P}(\{H_0 \mid \forall \theta \in \Theta_{\mathbb{Q}}\})=0$.

This clearly amounts to the deduction that any single-point null hypothesis about the normal mean has also probability zero, that is,

P(Point null hypothesis formulated as a rational number | distribution is absolutely continuous) = 0,

since a singleton has Lebesgue measure zero.

The comprehensiveness of the General Zero Probability theorem can be further extended by observing an even more general set of all point null hypotheses stated as real algebraic numbers, that is, the roots of single-variable polynomial equations whose coefficients are all integers. This set includes rational numbers, Gaussian integers, golden ratio, constructible numbers, some irrational numbers like $\sqrt{2}$, etc. Since this set is countable, it has Lebesgue measure zero and therefore under Normal distribution, its probability is zero. The cardinality (the number of elements in the set) of the

algebraic numbers is $\aleph_0$ (aleph-naught), the same as the cardinality of natural numbers and rational numbers,

$$card(A) = card(\mathbb{Z}) = card(\mathbb{Q}) = \aleph_0,$$

while the cardinality of the set of transcendental numbers is the same as that of the set of real numbers, i.e. the cardinality of the continuum.

3. Implications of the General zero probability theorem

Let us first restate the General Zero probability theorem in the following way: In practice, when testing a mean of the normal distribution using a point null hypothesis, the probability of that hypothesis is zero. This is based upon the stipulation that researchers, in reality, will almost surely state their null as an algebraic number.

It is important to stress out that this theorem applies both in the case when the population variance is known and unknown. Likewise, it is germane for testing parameters for all distributions that are absolutely continuous with respect to the Lebesgue measure λ including the beta, the Cauchy, the gamma, the uniform, chi-square, Student, exponential, etc. A typical example is in testing population variance using the classical chi-square test. Probability of any point null hypothesis that specifies a rational value of the population variance is always zero.

We now briefly discuss implications of the Zero probability theorem for the Fisherian significance tests, Neyman-Pearson tests, Bayesian analysis. Discussion is limited to the one-sample testing of a normal mean.

4.1 Fisherian significance testing. Frequentist significance tests (*Z* and *t*-test) are consistent, and as the sample size increase, they will detect even the smallest disagreement from the hypothesized (almost surely false) null hypothesis. This indicates that in the real-world testing, any point-null hypothesis of the normal mean will be eventually, almost surely, rejected with a large enough sample size. This is sometimes called a "large sample" problem, but this is not a problem at all, just the natural consequence of the false null hypotheses. Therefore, before conducting any significance test, we know that the resulting *p* value will be less than any preselected level of significance (in Fisher's sense) and that the point null hypothesis will be rejected. The only condition to reach this conclusion is to collect enough data.

Fisher ([28] p. 18) argued that "every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis." Zero probability theorem demonstrates that this is purposeless under the current paradigm: almost every point null hypothesis of a normal mean is false, and could be refuted a priori, without wasting time in performing any experimental study.

4.2 Neyman-Pearson hypothesis testing approach. Let us now consider the Neyman-Pearson paradigm, as a "procedure for choosing between two hypotheses" ([85] p. 132). The practical consequences of the Zero probability theorem applied on testing a sharp null hypothesis under the model that presumes absolute continuous distribution in respect to the Lebesgue measure are as follows.

First, type I errors (rejecting the null hypothesis when it is true) in the normal mean testing are extremely unlikely to be committed since point nulls are almost always false. Thus, the probability of a Type I error is almost surely equal to zero. Furthermore, Type II error (failure to reject a null hypothesis when it is false) could almost never happen as long as a researcher adopts a "rule of behavior" to reject any point null hypothesis, without even seeing her data! On almost every occasion, she will make the correct choice. By following this kind of paradoxically uniform "inductive behavior", a researcher will also ensure that the power of her test (probability of correctly rejecting the null hypothesis when it is false) is almost always equal to one, regardless of the sample size.

Neyman and Pearson ([77] p. 291) argued that "as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis." However, we do not need the test, but the probabilistic logic exposed in the Zero probability theorem to confirm that as far as a particular point hypothesis of the normal mean is concerned, we have valuable evidence of the falsehood of that hypothesis.

We acknowledge that classical Neyman-Pearson lemma does have significant importance from a conceptual viewpoint. However, it considers only two point hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$. From the *inferential* perspective, its practical relevance in real-life parameters testing within the models that specify absolutely continuous distributions is highly limited. The reason is simple: the probability of both simple hypotheses stated as rational (or algebraic) numbers is zero.

3.3 Bayesian testing of a point-null hypothesis.

According to one of the leading proponents of the "objective" Bayesian statistics, Bernardo, the Bayesian approach provides a complete coherent paradigm for both statistical inference and decision making under uncertainty; it constitutes a "scientific revolution in Kuhn's sense" (Bernardo [12] p. 108). Notwithstanding Gelman's ([31] p. 445) claim that "Bayesian inference is one of the more controversial approaches to statistics", Bayes theorem provides an easy, elegant, and extremely powerful way to show the full repercussions of the Zero probability theorem.

A typical criticism of the Bayesian approach is embedded in the question: where do the priors come from? In our analysis, it is easy to dismiss it; priors come from the knowledge base, that is from the Zero probability theorem. The assignment of the prior probability to the point null hypothesis is straightforward: it has to be zero, that is $P^{\pi}(\theta = \theta_0) = 0$. As a consequence, *the posterior probability for any data is equal to zero, that is* $\pi(H_0 \mid x) = 0$*, and the point null hypothesis will always be rejected*.

Thus, the Bayesian theorem undoubtedly confirms that testing a point null hypothesis of a continuous parameter is an irrational task. In that illogical procedure, for decades, researchers have been assessing a probability of picking a set that consists of one single point on the real line (singleton) out from the "number of elements in the uncountably infinite set of all real numbers" (the cardinality of continuum). This probability, as has been shown, is less than a probability of picking by chance a particular atom in the entire multiverse. Even if the alternative hypothesis is substituted by a more reasonable open interval that does not contain θ_0, that interval is still equinumerous with cardinality of continuum. The controversial point null paradigm has been sometimes justified by the philosophical stance that "we wish to learn (via significance tests) 'how false' the null is." ([68] p. 120). However, we do not believe that millions of researchers would agree to waste their careers in measuring a *degree of the falsehood of the false point null hypotheses*.

The above approach is in accordance with Bernardo's view ([13] p. 57), that "$\pi(\theta_0) = 0$ is not in violation of Cromwell's rule[2], but a simple consequence of the fact that H_0 is a measure zero set in this setting [point null hypothesis when θ is a continuous parameter]." He further correctly argues that it is justifiable to assign positive prior under H_0 only when the parameter space is finite.

However, the adherents of the Cromwell's rule would most probably argue that the account of the last paragraph is crucially mistaken. For example, Pericchi ([82] p. 26) strongly points out that assigning null hypothesis a zero probability is "a case of pure dogmatism or a violation of the Cromwell rule described by Lindley" (Pericchi [82] p. 26).

4. Concluding remarks (what should be done)

During the past 80 years, too many disturbing anomalies have accumulated in statistical testing of a point null hypothesis within both frequentist and Bayesian framework and reached the point so serious that have created the statistical crisis in science. Overcoming accumulated inconsistencies is always a crucial method in science. As pointed out by Good ([37] p. 107) "The resolution of inconsistencies will always be an essential method in science" and Good ([38] p. 489) "a Bayes/non-Bayes compromise or synthesis is necessary for human reasoning".

We regard that it is imperative to abandon the obsolete paradigm. As stated before, we limit our argument to a single sample point-null hypothesis testing of a normal mean. Before discussing the advantages of the proposed paradigm, it is important to emphasize that, in principle, there is no collision between frequentist and Bayesian approaches in testing one-sided null hypothesis and between frequentist confidence intervals and Bayesian credible intervals based on certain "flat" priors.

(1) In one-sided testing, for many classes of reasonable prior distributions the infimum of the Bayesian posterior probability of H_0 is equal to the *p*-value, or even strictly lower bound on the *p*-value (Casella and Berger 1987). Particularly, for one-sided testing of a normal mean, frequentist measure yields essentially the same answer, as does Bayesian analysis with a noninformative prior.
(2) Two-sided frequentist confidence intervals and Bayesian HPD ((Highest Posterior Density) intervals are numerically equivalent for certain classes of noninformative ("flat") priors. These two intervals are also approximately equivalent under conjugate normal priors, and for large samples nearly identical.

Hence, harmonization must be achieved in two-sided testing since as the sample size increases, by virtue of the Jeffreys-Lindley paradox the number of opposite conclusions between frequentist and Bayesian testing will start to upsurge. Thus, with the point-null paradigm frequentist and Bayesian concepts will show irreconcilable differences. We argue that this reconciliation will be accomplished if we abolish point-nulls.

We recommend a paradigm shift in statistical inference. Instead of looking for the statistical significance, we propose searching for the practical significance. In other words, as an alternative to

[2] Lindley advocated that zero prior probabilities should be used only for logically false statements. He ([64] p. 104) advised Bayesians to "leave a little probability for the moon being made of green cheese; it can be as small as 1 in a million."

the obsolete and almost always false point-null hypotheses, as a natural replacement, we advocate testing a negligible null hypothesis:

$H_0 : |\theta - \theta_0| \le \delta$ (Effect size is unimportant) against

$H_1 : |\theta - \theta_0| > \delta$ (Effect size is practically meaningful), result will be practically significant.

We argue that this configuration corresponds to the goal of the majority of researchers. It has many-sided advantages in comparison to the traditional one, as being supported by the following arguments.

(1) Mathematical statisticians are responsible to provide researchers in other sciences with non-conflicting, coherent, and consistent concepts of testing the statistical hypotheses. Otherwise, statistical tests will harm progress in science. Researchers and scientists will feel confused and deceived by statistics and statisticians.

In sharp contrast to the current point-nulls model, it is possible to harmonize inferential results of frequentist and Bayesian testing within the new framework. In other words, frequentist and Bayesian inference will become, in principle, compatible and would (or at least could) lead to similar conclusions in (a) one-sided testing, (b) two-sided testing, and (c) interval estimation.

This easily achievable reform in statistical testing is the greatest positive effect;

(2) One of the main objections to significance statistical testing, that all the nulls are false from the outset, vanishes in the air since the probability that a realized value from the interval null is larger than zero.

(3) All logical inconsistencies of Fisherian significance testing (considered in 4.1) and Neyman-Pearson hypothesis testing (discussed in 4.2) will be eradicated;

(4) Famous Berkson's large sample significance paradox will be eliminated. This paradox states that with increasing sample size frequentist tests will reject point-null hypothesis. This is called paradox since a researcher who is aware of it does not have to perform any test since they know in advance that with sufficiently large sample null hypothesis will be rejected. Surely, this logic cannot be applied to the interval null hypotheses that contain an infinite number of values.

(5) In Bayesian testing there would be no longer a necessity to assign highly polemic and unnatural point mass on the null value;

(6) Jeffreys-Lindley paradox, the point of the irreconcilable divergence between the frequentist and Bayesian inference will be annihilated. As confirmed by Berger and Delampady ([11] p. 322), testing interval null hypothesis "will often result in $P(H_0|\overline{x}_n) \to \alpha$ as $n \to \infty$ in marked contrast to Jeffreys's paradox".

(7) A well-known limitation of frequentist hypothesis tests is their inability to distinguish between statistical significance and practical significance. The negligible-null setup eradicates this. From the frequentist perspective, rejection of the negligible null hypothesis implies that the *p*-value indicates at least the existence of the pre-specified effect size δ (certainly under the condition that the Type I error was not committed). In other words, the observed effect bears practical meaning. In contrast, a significant *p*-value in the traditional setup only can, at most, suggests that the effect size is non-zero. Such test outcome is trivial and almost absolutely noninformative. With very large samples, impressive-looking *p* values can simply signify that the magnitude of the effect is less than, say, 10^{-10}, which does not have any real-world

importance. This is one of the most widely recognized limitations of frequentist tests. As Abelson ([1] p. 121) remarks, "Typically, mere difference from zero is totally uninteresting";

(8) The real scientific interest is in testing negligible nulls, that is, in hypotheses that the parameter value is relatively close to θ_0. As Levine et al. ([62] p. 181) remark, "researchers are interested in making substantive claims, and statistical analyses are only meaningful to the extent they are informative about the viability of substantive hypotheses.";

(9) In frequentist testing, an embarrassing situation in which every study will eventually produce a statistically significant effect when researchers collect enough data is eliminated. As we have seen, this large sample problem is generated by the falsehood of the point nulls. By specifying a demarcation point (which is a major challenge) between negligible and practically meaningful effects a crucial step towards building a tool for massive data set is achieved;

(10) Some important misconceptions that are shared among many researchers are inherently eradicated by the nature of the construct of the rival hypotheses, including the following: (a) a small *p*-value means a treatment effect has large magnitude, (b) a statistically significant finding is practically important and (c) that a *p*-value is a numerical index of the magnitude of the effect.

(11) One of the most common criticisms of the point null hypothesis statistical testing is that it "does not tell us what we really want to know…What we want to know is whether the null hypothesis is true given the data" (Orlitzky [81] pp. 200-201). Within the current paradigm, in the classic normal model based on a point null hypothesis, the answer to this criticism is simple: the null hypothesis is almost never true. This researchers' objective could be ultimately achieved via Bayesian analysis with the negligible null model and priors freed from the unrealistic point masses;

(12) Frequentist tests have been criticized to achieve significance too easily. With the proposed model, it is more difficult to attain a significant result. This can, in turn, increase the prospects of reproducibility of research findings as a cornerstone of scientific progress and circumvent "reproducibility crisis" (like recently in cancer research and psychology). It is essential here to reiterate one of the most important, but almost forgotten, principles of the experimental design, following Fisher ([28] pp. 13-14).

> "We thereby admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon… In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result.";

It is clear that Fisher insisted on repetitions of an experiment. Therefore, the assertion expressed by Hubbard et al. ([52] p. 173) that "Fisher claimed that his significance tests were applicable to single experiments" is not persuasive;

(13) Perhaps the most crucial advantage of the negligible null model is that it implicitly requires raising standards for the research. As a result, the credibility of published results will increase. Moreover, we recommend that researchers, whenever possible, should verify internal consistencies of their results using *p* values, posterior probabilities of H_0, Bayes factors, confidence intervals, and HPD intervals.

Finally, more studies should be done in the future with the intention to extend notion of the General Zero Probability theorem to the problems of two and more samples.

6. References

[1] Abelson, R. P. (1997b). A retrospective on the significance test ban of 1999 (if there were no significance tests, they would be invented). In: L. L. Harlow, S.A. Mulaik, and J. H. Steiger (Eds.), What if there were no significance tests? Hillsdale, NJ: Erlbaum. pp. 117-141.

[2] Abelson, R.P. (1997a). On the surprising longevity of flogged horses: Why there is a case for the significance test. Psychological Science. 8(1). pp. 12–15.

[3] Ambaum, M.H.P. (2010). Significance tests in climate science. Journal of Climate. 23. pp. 5927–5932.

[4] Arbuthnott, J. (1710). An Argument for Divine Providence, taken from the constant Regularity observ'd in the Births of both Sexes. Philosophical Transactions of the Royal Society of London. Vol. 27 No. 325–336. pp. 186–190.

[5] Armstrong, J. S. (2007). Significance tests harm progress in forecasting. International Journal of Forecasting. 23(2). pp. 321-327.

[6] Bakan, D. (1966). The test of significance in psychological research. Psychological Bulletin. 66(6). pp. 423-437.

[7] Begley, C. G. and S. M. Snapinn (2021) Comment on "The Role of p-Values in Judging the Strength of Evidence and Realistic Replication Expectations", Statistics in Biopharmaceutical Research, 13:1, 40-42, DOI: 10.1080/19466315.2020.1782259.

[8] Beninger, P. G., Boldina, I. and Katsanevakis, S. (2012). Strengthening statistical usage in marine ecology. Journal of Experimental Marine Biology and Ecology. 426-426. pp. 97-108.

[9] Berger, J. 0., Boukai, B. and Wang, Y. (1997). Unified Frequentist and Bayesian Testing of a Precise Hypothesis. Statistical Science. Vol. 12. No. 3. pp. 133-160.

[10] Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman Have Agreed on Testing? Statistical Science. Vol. 18, No. 1. pp. 1–32.

[11] Berger, J. O. and Delampady, M. (1987). Testing Precise Hypotheses. Statistical Science, Vol. 2, No. 3, pp. 317-335.

[12] Bernardo, J. M. (2010). Bayesian Statistics. In: M. Lovric (Editor). International Encyclopedia of Statistical Science. Springer, pp. 107-133.

[13] Bernardo, J.M. (2011). Integrated objective Bayesian estimation and hypothesis testing. Bayesian Statistics. 9. (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press. pp. 1-68.

[14] Branch, M. N. (2019). The "reproducibility crisis:" Might the methods used frequently in behavior-analysis research help? Perspectives on Behavior Science, 42(1), 77–89. https://doi.org/10.1007/s40614-018-0158-5.

[15] Buchanan-Wollaston, H. J. (1935). Statistical Tests. Nature, 136, 182-183.

[16] Bushway, S., Sweeten, G. and Wilson, D. B. (2006). Size matters: Standard errors in the application of null hypothesis significance testing in criminology and criminal justice. Journal of Experimental Criminology. 2. pp. 1–22.

[17] Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review. 48(3). pp. 378-399.

[18] Casella, G. and Berger, R.L. (1987). Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem. Journal of the American Statistical Association. Vol. 82, No. 397. pp. 106-111.

[19] Chandrakantha, L. (2020). "Visualizing the p-value and Understanding Hypothesis Testing Concepts Using Simulation in R", The Electronic Journal of Mathematics and Technology, Volume 14, Number 3.

[20] Chow, S. L. (1998). The Null-hypothesis significance-test procedure is still warranted. Behavioral and Brain Sciences. 21(2). pp. 228-235.

[21] Cohen, J. (1990). Things I have learned (so far). American Psychologist. 45(12). pp. 1304-1312.

[22] Cohen, J. (1994). The earth is round ($p<.05$). American Psychologist. 49(12). pp. 1007-1003.

[23] Cook, W. J. and Bossé, M. (2021). “Types of infinity”. The Electronic Journal of Mathematics and Technology, Volume 15, Number 2.
[24] Cortina, J. M. and Dunlap, W. P. (1997). On the logic and purpose of significance testing. Psychological Methods, 2(2). pp. 161-172.
[25] Dennis, B. (2004). Statistics and the scientific method in ecology (with commentary). In: Taper, M. L. and Lele, S. R. (Eds.). The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations. The University of Chicago Press. pp. 327-378.
[26] Estay, S.A. and Naulin, P. I. (2011). Data analysis in forest sciences: why do we continue using null hypothesis significance tests? Scientific Electronic Library Online - Chile. Retrieved at December 10, 2014, from the website Open Educational Resources (OER) Portal at http://www.temoa.info/node/596993
[27] Falk, R. and Greenbaum, C. W. (1995). Significance Tests Die Hard: The Amazing Persistence of a Probabilistic Misconception. Theory & Psychology, 5(1). pp. 75-98.
[28] Fisher, R. A. (1935). The Design of Experiments. Edinburgh: Oliver and Boyd.
[29] Fleiss, J. L. (1986). Significance tests do have a role in epidemiological research: Reactions to A. A. Walker. American Journal of Public Health. 76(5). pp. 559–560.
[30] Frick, R. W. (1996). The appropriate use of null hypothesis testing. Psychological Methods, 1, 379-390.
[31] Gelman, A. (2008). Objections to Bayesian statistics. Bayesian Analysis. 3. Number 3. pp. 445–450.
[32] Gelman, A. and Shalizi, C. (2013). Philosophy and the practice of Bayesian statistics. British Journal of Mathematical and Statistical Psychology. 66. pp. 8–38.
[33] Ghosh, J. K, Delampady M. and Tapas, S. (2006). An Introduction to Bayesian Analysis: Theory and Methods. Springer, 2006. ISBN-10: 0-387-40084-2.
[34] Gibson, E.W. (2020), “The Role of p-Values in Judging Strength of Evidence and Realistic Replication Expectations,” Statistics in Biopharmaceutical Research, 13. DOI: 10.1080/19466315.2020.1724560.
[35] Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In: G. Keren & C. Lewis (Eds.), A handbook for data analysis in the behavioral sciences: methodological issues. Hillsdale. NJ: Erlbaum. pp.311-339.
[36] Gill, I.J. (1999). The Insignificance of Null Hypothesis Significance Testing. Political Research Quarterly. Vol. 52. No. 3. pp. 647-674.
[37] Good, I. J. (1952). Rational Decisions. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 14, No. 1, pp. 107-114.
[38] Good, I. J. (1980). Some history of the hierarchical Bayesian methodology. Trabajos de Estadistica Y de Investigacion Operativa. Vol. 31, Issue 1, pp 489-519.
[39] Good, I.J. (1956). Which comes first, probability or statistics? Journal of the Institute of Actuaries. 82, p. 249-255.
[40] Goodman, S. N. (1993). P values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. American Journal of Epidemiology. 137(5). pp. 485-496.
[41] Goodman, S. N. (1999), “Towards Evidence-Based Medical Statistics, II: The Bayes Factor,” Annals of Internal Medicine, 130, 1005–1013.
[42] Goodman, S. N. (2008). A Dirty Dozen: Twelve P-Value Misconceptions. Seminars in Hematology. Vol. 45, Issue 3. pp. 135–140.
[43] Greenland, S., Senn, S.J., Rothman, K.J. et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 31, 337–350 (2016). https://doi.org/10.1007/s10654-016-0149-3.
[44] Hagen, R.L. (1997). In Praise of the Null Hypothesis Statistical Test. American Psychologist. 52, No. 1. pp. 15-24.
[45] Hald, A. (2003). A History of Probability and Statistics and Their Applications before 1750. John Wiley & Sons.
[46] Haller, H. and Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? Methods of Psychological Research, 7(1). pp. 1-20.

[47] Hauer, E. (1983). Reflections on methods of statistical inference in research on the effect of safety countermeasures. Accident Analysis & Prevention. 15. pp. 275–286.
[48] Held, L., and Ott, M. (2018), "On p-Values and Bayes Factors," Annual Review of Statistics and Its Application, 5, 393–419.
[49] Hentschke, H. and Stuüttgen, M. C. (2011). Computation of measures of effect size for neuroscience data sets. European Journal of Neuroscience, Vol. 34, pp. 1887–1894.
[50] Hubbard, R. (2004). Blurring the Distinctions Between p's and α's in Psychological Research. Theory Psychology. Vol. 14 no. 3. pp. 295-327.
[51] Hubbard, R. and Armstrong, J. S. (2006). Why we don't really know what statistical significance means: Implications for educators. Journal of Marketing Education. 28(2). pp. 114-120.
[52] Hubbard, R., Bayarri, M. J., Berk, K. N. and Carlton, M. A. (2003). Confusion over Measures of Evidence (p's) versus Errors (α's) in Classical Statistical Testing. The American Statistician, Vol. 57, No. 3 (Aug., 2003), pp. 171-182.
[53] Iacobucci, D. (2005). On p-values. Journal of Consumer Research 32 (1). pp. 6-11.
[54] Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. PLoS Medicine. 2(8): e124. doi:10.1371/journal.pmed.0020124.
[55] Jeffreys, H. (1935), "Some Tests of Significance, Treated by the Theory of Probability," Mathematical Proceedings of the Cambridge Philosophical Society, 31, 203–222.
[56] Jeffreys, H. (1961), Theory of Probability (3rd ed.), Oxford, UK: Oxford University Press.
[57] Johnson, D. H. (1999). The Insignificance of Statistical Significance Testing. Journal of Wildlife Management 63(3). pp. 763-772.
[58] Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," Journal of the American Statistical Association, 430, 773–795.
[59] Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. American Psychologist, Vol 56(1), 16-26.
[60] Lehmann, E. L. (1993). The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? Journal of the American Statistical Association, Vol. 88, No. 424. pp. 1242-1249.
[61] LeMire, S. D. (2010). An Argument Framework for the Application of Null Hypothesis Statistical Testing in Support of Research. Journal of Statistics Education, Volume 18, Number 2.
[62] Levine, T. R., Weber, R., Hullett, C., Park, H. S., and Lindsey, L. L. M. (2008). A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research. Human Communication Research. 34(2). 171–187.
[63] Lindley, D. V. (1975). The Future of Statistics: A Bayesian 21st Century. Advances in Applied Probability, Vol. 7, Supplement: Proceedings of the Conference on Directions for Mathematical Statistics (Sep., 1975), pp. 106-115.
[64] Lindley, D. V. (1991). Making Decisions. John Wiley & Sons. 2nd ed.
[65] Lindsay, R.M. (1995). Reconsidering the status of tests of significance: An alternative criterion of adequacy. Accounting, Organizations and Society. 20(1). pp. 35–53.
[66] Lovric, M. M. (2019). On the Authentic Notion, Relevance, and Solution of the Jeffreys-Lindley Paradox in the Zettabyte Era. Journal of Modern Applied Statistical Methods, 18(1), eP3249. doi: 10.22237/jmasm/1556670180.
[67] Masicampo, E.J. & Lalande, Daniel. (2012). A Peculiar Prevalence of p Values Just Below .05. Quarterly journal of experimental psychology (2006). 65. 2271-9. 10.1080/17470218.2012.711335.
[68] Mayo, D. (2005). Evidence as Passing Severe Tests: Highly Probable versus Highly Probed Hypotheses. In: Achinstein, P. (Ed.). Scientific Evidence: Philosophical Theories and Applications. Johns Hopkins University Press. Baltimore. pp. 95-127.
[69] McCloskey, D. N. (1985). The loss function has been mislaid: the rhetoric of significance tests. American Economic Review. 75(2). pp. 201-205.
[70] McCloskey, D. N. and Ziliak, S. T. (1996). The standard error of regression. Journal of Economic Literature. 34(3). pp. 97-114.

[71] Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. Philosophy of Science. 34. pp. 103-115. Reprinted in The Significance Test Controversy - A Reader, Eds. D. E. Morrison and R. E. Henkel, 1970, Aldine Publishing Company (Butterworth Group).
[72] Meehl, P. E. (1978). Theoretical risks and tabular asterisk: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Counseling and Clinical Psychology. 46. pp. 806-834.
[73] Mogie, M. (2004). In support of null hypothesis significance testing. Proceedings of the Royal Society of London, Series B, Biology Letters, 271, pp. S82–S84.
[74] Morrison, D.E. and Henkel, R.E. (1969). Significance tests reconsidered. American Sociologist. 4. pp. 131-140.
[75] Mulaik, S. A., Raju, N. S. and Harshman, R. A. (1997). There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.). What if there were no significance tests? (pp. 65–116). Mahwah, NJ: Erlbaum.
[76] Murtaugh, P. A. (2014). In defense of P values. Ecology. 95(3). pp. 611–617.
[77] Neyman, J. and Pearson, E. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. Philosophical Transactions of the Royal Society of London. Series A, Vol. 231. pp. 289-337.
[78] Nicholls, N. (2001). The insignificance of significance testing. Bulletin of the American Meteorological Society. 82(5). pp. 981–986.
[79] Nunnally, J. (1960). The place of statistics in psychology. Educational and Psychological Measurement, 20, 641—650.
[80] Oakes, M. (1986). Statistical Inference: A Commentary for the Social and Behavioral Sciences. New York: Wiley.
[81] Orlitzky, M. (2012). How Can Significance Tests Be Deinstitutionalized? Organizational Research Methods. 15(2). pp. 199-228.
[82] Pericchi, L. (2011). [Integrated objective Bayesian estimation and hypothesis testing]. Discussion. In: J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds. Bayesian Statistics 9. Oxford: University Press. pp. 25-30.
[83] Rao, C.R. and Lovric. M. (2016). Testing Point Null Hypothesis of a Normal Mean and the Truth: 21st Century Perspective. Journal of Modern Applied Statistical Methods. Vol. 15(2).
[84] Rothman, K. J. (1986). Significance questing. Annals of Internal Medicine. 105(3). pp. 445-447.
[85] Royall, R. (2004). The likelihood paradigm for statistical evidence. In: Taper, M.L. and Lele, S.R. (editors). The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations, Chicago: University of Chicago Press.
[86] Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. Psychological Bulletin, 57(5). pp. 416- 428.
[87] Savage, I. R. (1957). Nonparametric Statistics. Journal of the American Statistical Association, 52, 331-344.
[88] Sawilowsky, S. (2010). Frequentist Hypothesis Testing: a Defense. In: M. Lovric (Ed.) International Encyclopedia of Statistical Science. Springer-Verlag. pp. 547-550
[89] Schmidt, F. L. and Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In: L. L. Harlow, S.A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? Hillsdale. NJ: Erlbaum. pp. 37-64.
[90] Schneider, J.W. (2013). Caveats for using statistical significance tests in research assessments. Journal of Informetrics. Vol. 7, Issue 1, pp. 50–62.
[91] Schwab, A. and Starbuck, W.H. (2009). Null-hypothesis significance tests in behavioral and management research: We can do better. In: Bergh, D. B. and Ketchen, D. J. (Eds.) Research Methodology in Strategy and Management. Vol. 5. Emerald. pp. 30-54.
[92] Schwab, A., Abrahamson, E., Starbuck, W.H. and Fidler, F. (2011): Perspective—Researchers Should Make Thoughtful Assessments Instead of Null-Hypothesis Significance Tests. Organization Science. 22 (4). pp. 1105-1120.
[93] Senn, S. (2001). Two cheers for P-values? Journal of Epidemiology and Biostatistics. Vol. 6. No. 2. pp. 193–204.

[94] Siegfried, T. (2010). Odds are, it's wrong: Science fails to face the shortcomings of statistics. Science News.
[95] Smith, A. (1995). A Conversation with Dennis Lindley." Statistical Science, vol. 10, no. 3, 1995, pp. 305–319.
[96] Spanos, A. (2014). Recurring controversies about P values and confidence intervals revisited. Ecology. 95(3). pp. 645–651.
[97] Stang, A., Poole, C. and Kuss, O. (2010). The ongoing tyranny of statistical significance testing in biomedical research. European Journal of Epidemiology. 25(4). pp. 225-230.
[98] Stephens, P. A., Buskirk, S. W. and del Rio, C. M. (2006). Inference in ecology and evolution. Trends in Ecology and Evolution. Vol.22. No.4. pp. 192-197.
[99] Trafimow, D. and M. Marks (2015) Editorial, Basic and Applied Social Psychology, 37:1, 1-2, DOI: 10.1080/01973533.2015.1012991
[100] Vicente, K.J. and Torenvliet, G.L. (2000). The earth is spherical (P<.05): alternative methods of statistical inference. Theoretical Issues in Ergonomics Science. 1(3), pp. 248–271.
[101] Wainer, H., & Robinson, D. H. (2003). Shaping up the practice of null hypothesis significance testing. Educational Researcher. 32. pp. 22-30.
[102] Wasserstein, R. L., Schirm, A. L. & N. A. Lazar (2019) "Moving to a World Beyond 'p < 0.05', The American Statistician, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913.