Journal of Modern Applied Statistical Methods

Volume 15 | Issue 2 Article 3

11-1-2016

Testing Point Null Hypothesis of a Normal Mean and the Truth: 21st Century Perspective

Calyampudi Radhakrishna Rao
Penn State University, crr1@psu.edu

Miodrag M. Lovric
University of Kragujevac, mlovric@kg.ac.rs

בס"ד
Journal of Modern Applied Statistical Methods
November 2016, Vol. 15, No. 2, 2-21.
doi: 10.22237/jmasm/1478001660

ISSN 1538 − 9472

Testing Point Null Hypothesis of a Normal Mean and the Truth: 21st Century Perspective

Calyampudi Radhakrishna Rao
Penn State University
State College, PA

Miodrag M. Lovric
University of Kragujevac
Kragujevac, Serbia

Testing a point (sharp) null hypothesis is arguably the most widely used statistical inferential procedure in many fields of scientific research, nevertheless, the most controversial, and misapprehended. Since 1935 when Buchanan-Wollaston raised the first criticism against hypothesis testing, this foundational field of statistics has drawn increasingly active and stronger opposition, including draconian suggestions that statistical significance testing should be abandoned or even banned. Statisticians should stop ignoring these accumulated and significant anomalies within the current point-null hypotheses paradigm and rebuild healthy foundations of statistical science. The foundation for a paradigm shift in testing statistical hypotheses is suggested, which is testing interval null hypotheses based on implications of the Zero probability paradox. It states that in a real-world research point-null hypothesis of a normal mean has zero probability. This implies that formulated point-null hypothesis of a mean in the context of the simple normal model is almost surely false. Thus, Zero probability paradox points to the root cause of so-called large *n* problem in significance testing. It discloses that there is no point in searching for a cure under the current point-null paradigm.

Keywords: zero-probability paradox, point null hypothesis, Lebesgue measure, rational numbers, algebraic numbers, almost sure false null hypothesis, inexactification, paradigm shift in testing statistical hypotheses.

> *"It cannot be denied that, during the recent rapid development of practical methods, fundamental problems have been ignored and fundamental paradoxes left unresolved"*
>
> *Fisher (1922)*

Dr. Rao is the Eberly Professor Emeritus of Statistics and the Director of Center for Multivariate Analysis. Email him at crr1@psu.edu. Dr. Lovric is a Professor in the Faculty of Economics. Email him at mlovric@kg.ac.rs.

Introduction

Following Fisher's foundational contribution to significance tests, and Neyman and Pearson to hypothesis tests, statistical testing has become widely adopted by researchers as the most common statistical inferential approach in almost all different branches of science. However, there has been a steadily growing dissatisfaction in the scientific community with traditional tests of the point (sharp, precise) null hypothesis. Since Buchanan-Wollaston (1935) raised the first criticism against significance testing, their application has been debated extensively, and numerous objections and severe complaints have been leveled against their utility. Critics also accentuated statistical tests are not only overused, but are often misunderstood and misused. Nickerson (2000) provided a summary of common misconceptions, and criticisms as well as arguments in support of null hypothesis testing, from a non-statistician viewpoint.

The most trenchant critics requested significance tests should be abandoned, banned or deinstitutionalized (e.g., Lindley, 1975; Hunter, 1997; Armstrong, 2007; Orlitzky, 2012). The editors of the *American Journal of Public Health* imposed a ban, although it only lasted two years. Similarly, in 1997 the officers of the American Psychological Association (APA) created a task force to make recommendations about appropriate statistical practice and to consider banning significance testing. The proposal was regarded as too extreme and was rejected (Wilkinson, 1999). More recently, in 2015, the editors of *Basic and Applied Social Psychology* journal enforced a ban on significance testing (as well as confidence intervals). On behalf of the ASA Board of Directors, Wasserstein & Lazar (2016) formulated six principles regarding the usage of p-values, hoping that the ASA statement would open a fresh discussion with regards to the use of statistical inference.

The ASA's statement should be praised as the first organized reply from statistics community to the abovementioned issues. However, it did not address the fundamental problems and did not provide a new perspective on statistical testing.

Critics advocated reform of statistical inference and statistics education. They recommended less emphasis should be placed on reporting of *p* values, cynically termed "harvest of asterisks" (Cohen, 1990). The reformers, mainly non-statisticians, argued attention should be shifted to effect size, point estimation, confidence interval, information theoretic approaches (e.g., Akaike Information Criterion), graphical methods, and progressively more on the communication of results using Bayesian inference.

Consider two of the most important criticisms of significance testing: (1) point null hypotheses are unlikely to be true, and (2) a statistical significant result is always obtainable with a sufficiently large sample. The scope of this paper is limited to the problem of testing the mean of a normal distribution, although this problem is of substantial importance because of its widespread application in statistical theory and practice. The primary objective is to prove that in the real-world research when testing the mean of a normal distribution using a point-null hypothesis, the probability of that hypothesis is zero. We call this result the Zero probability paradox. This paradox undoubtedly reveals logical deficiency of a point-null hypothesis of a normal mean: in reality, its testing is actually a procedure that unequivocally will lead (with sufficiently large sample) to a foregone conclusion that formulated null hypothesis is almost surely false. The logical name for this procedure in which a sharp null hypothesis is ultimately being rejected should be "inexactification," rather than testing (Good, 1994, p. 241).

The Existence of Point Null Hypothesis: History and Overview

Testing a point null hypothesis is arguably the most widely used and at the same time the most controversial, misapprehended and severely criticized statistical procedure in many fields of scientific research. Focus on one of the most common criticisms, that point null hypotheses are not realistic. The Zero probability paradox, presented here, evolved as a result of persuasive and accumulated ideas of statisticians, and non-statisticians referred to in this section.

There is a vast amount of references in statistics and non-statistics literature with the claim that, in reality, point null hypotheses are almost always false. Critics, however, supported this statement only by intuitive arguments, empirical evidence, and common sense. One of the early critics, L. J. Savage (1954, p. 254), disproved the validity of tests "in which the null hypothesis is such that it would not really be accepted by anyone." I. R. Savage, (1957, p. 332-333) asserted the "null hypotheses of no difference are usually known to be false before the data are collected…when they are, their rejection or acceptance simply reflects the size of the sample and the power of the test, and is not a contribution to science." Nunnally (1960, p. 642) expressed a similar assertion, but admitted he agreed although he cannot prove it directly. However, he argued it is supported both by common sense and by practical experience. Likewise, Meehl, (1967, p. 108) pointed out there is "universal agreement that the old point-null hypothesis…is

[quasi-] always false in biological and social science." His opinion was based on the result that in "psychological and sociological investigations involving very large numbers of subjects, it is regularly found that almost all correlations or differences between means are statistically significant" (p. 109). Meehl illustrated this by providing an example of a large sample of over 55,000 Minnesota high school seniors that revealed 91% significant associations among a collection of 45 variables.

In the same way, Cohen (1990, p. 1308) stated the null hypothesis "taken literally (and that's the only way you can take it in formal hypothesis testing), is always false in the real world. It can only be true in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false). If it is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection. So if the null is always false, what's the big deal about rejecting it?"

There is near consensus in the literature that exactly true point null hypotheses are extremely rare in reality. This is exemplified by the following by Kadane (1987, p. 347): "For the last 15 or so years I have been looking for applied cases in which I might have some serious belief in a null hypothesis. In that time I found only one [testing an astrologer claim that on the bases of peoples birthdays it is possible to predict who is likely to have a drug problem]... I do not expect to test a precise hypothesis as a serious statistical calculation."

In a similar manner, there was a quest for an existence of a realistic case for which a null hypothesis cannot be regarded beforehand as false. As a result of this pursuit, a commonly given example is found, that there is no extrasensory effect in a parapsychological experiment. Good (1994, p. 241) argued there is at least one example of a precisely sharp null hypothesis: precognition is impossible. Similarly, Ghosh et al. (2006, p. 45) suggested astrology cannot predict the future. Berger and Delampady (1987, p. 320), although admitting that it is perhaps impossible to have a null hypothesis that can be exactly modeled as $\theta = \theta_0$, noted talking to plants has no effect on their growth. Nevertheless, they admitted minor biases in the design of the experiments may produce statistical significance. They also argued that point null hypotheses are reasonable approximations to fuzzy precise (small interval) nulls. However, as pointed out by Bernardo (1999, p. 102) "this approximation always breaks down for sufficiently large samples." Likewise, Rousseau (2007) showed for large samples the Bayes factor associated with point null hypotheses is a poor approximation of Bayes factors of interval null hypotheses unless the intervals are extremely small.

In contrast, Zellner (1987, p. 339) emphasized many realistic examples of point null hypotheses can be given in testing well-formulated physical laws, such as $s =. 5gt^2$ and $E = mc^2$. Kass and Raftery (1995, p. 788) argued although "one rarely believes a scientific law in an absolute sense, it is a great convenience to speak and to act as if laws are valid. When one says that a certain theory is correct, one means that deviations from it are sufficiently minor to be irrelevant for all practical purposes at hand."

Based on the above arguments, a natural question arises: why are we testing point null hypotheses at all, when it is known in advance they are almost never exactly true in the real world? Sprenger (2013) argued these hypotheses often give useful idealization of reality. He considered this originated in the Popperian philosophy of science: "only a highly testable or improbable theory is worth testing and is actually (and not only potentially) satisfactory if it withstands severe tests." (Popper, 1963, p. 219–220)

According to Cox (2006, p. 31) null hypothesis refers to a probability model, and this implies idealization. He argued it would be absurd to think that a mathematical model could be an exact representation of a real system. Thus, null hypotheses are postulated within a system that is untrue.

Good (1956, p. 254) remarked a null hypothesis is tested, although it is known in advance it cannot be exactly true, because "we wish to test whether the hypothesis is in some sense approximately true, or whether it is rejectable on the sort of size of sample that we intend to take." Kruskal (1968) indicated the need is to test whether the mean is near μ_0, meaning as near as makes no substantive difference. He stated this will be achieved as long as the sample sizes and significance levels are reasonable and the power is at least moderately large for alternatives interestingly different from the null hypothesis.

Edwards, et al. (1963) presented a Bayesian view on the sharp null hypothesis problem. They acknowledged in usual applications the null hypothesis is known to be false from the outset, because realistically the null hypothesis cannot be infinitely sharp. From a Bayesian perspective, a sharp null hypothesis is likely to be appropriate only when it deserves special initial credence. They also highlighted in Bayesian analysis the null hypothesis is "*a hazily defined small region rather than a point* [italicized by authors]" (p. 235).

Finally, consider Krueger's (2001) attempt to explain why *all* null hypotheses are false. He started from the premise that in statistics populations are mathematical abstractions that contain infinite possible observations. "This implies an infinite number of possible states of the population, and each of these states may be a distinct hypothesis. With an infinite number of hypotheses, no

individual hypothesis can be true with any calculable probability" (p. 17). It is, however, clear that his arguments on the survival of the flawed significance testing are themselves flawed. It is erroneous to claim that one-sided and interval null hypotheses are always false.

It can be concluded existing literature does not offer proof of the extraordinary statement that all point null hypotheses are false.

The Nature of a Point Null Hypothesis

Before exposing the Zero probability paradox, it is of fundamental importance to clarify some misconceptions about the nature of the point null hypothesis.

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 assuming values in a parameter space $\Theta \subset \mathbb{R}^1$. Suppose also that the variance, $\sigma^2 > 0$ is known. It is required to test the null hypothesis $H_0 : \theta = \theta_0$ versus an unspecified alternative hypothesis $H_1 : \theta \neq \theta_0$. Regard this sharp or point null hypothesis as a *numerically exact* statement, that is free of vagueness and ambiguity, namely as an assertion that exactly specifies a single value of a parameter θ_0. In other words, it is obvious that θ_0 as a crisp number, not a fuzzy number.

It is well known that to every real number there corresponds a unique point on the number line and vice versa. Obviously, point hypothetical value θ_0 corresponds to a distinctive point on the real number line, not to an interval. As Euclid gave an intuitive definition in the first sentence from his Elements book 1, "a point is that which has no part, or which has no magnitude." In the contemporary notion, this is tantamount to saying that a point is a dimensionless entity that has only a location. It also naturally implies that "every point is unextended" (Playfair, 1819, p. 289).

Claims that there are different kinds of sharp hypotheses, some fuzzy sharp and some infinitely sharp, in other words, that equal sign can be perceived in infinitely different ways, are unconvincing. If testing "hazily defined small region" is considered a null hypothesis in a scientific, non-subjective way, then it is a sine qua non to formulate that hypothesis accurately, for example, as $H_0 : |\theta - \theta_0| \leq \delta$ or using fuzzy set theory as $H_0 : \tilde{\theta} = \tilde{\theta}_0$, where $\tilde{\theta}$ is the unknown fuzzy parameter and $\tilde{\theta}_0$ a known fuzzy number. However, in the traditional point null hypothesis $H_0 : \theta = \theta_0$, in practice, (since the pioneering work of Arbuthnott (1710)) θ_0 has always been formulated as a crisp rational number, never as a fuzzy number $\tilde{\theta}$.

A fuzzy number, $\tilde{\theta}$, in contrast, is a distinctly different entity. It is defined as a fuzzy set in $\mathbb{R}$ with a normal, fuzzy convex and a continuous membership function of bounded support. Note also that, in the fuzzy set framework, the possible values of the parameter of interest are expressed as linguistic variables, and that the data are observations of a normal fuzzy random sample. In conclusion, $\theta_0 = \tilde{\theta}_0$, that is, (Crisp number = Fuzzy number), is nothing else but a self-deception.

Zero Probability Paradox

In a real-world research, the probability of an exact point-null hypothesis of the mean of a normally distributed population is zero. Let $\mathbb{Q}$ be the set of all rational real numbers, that is $\mathbb{Q} = \{m/n; m, n \in \mathbb{Z}, n \neq 0\}$, where $\mathbb{Z}$ stands for the set of all integers. Suppose, as in the previous section, 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 assuming values in a parameter space $\Theta \subset \mathbb{R}^1$. Divide 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 ($\Theta = \Theta_{\mathbb{Q}} \cup \Theta_{\mathbb{R}\setminus\mathbb{Q}}$). Suppose further that the set $\Theta_{\mathbb{Q}}$ is equivalent to the set of all rational numbers $\mathbb{Q}$ and that $\Theta_{\mathbb{R}\setminus\mathbb{Q}}$ is equivalent to the set of all irrational numbers $\mathbb{R}\setminus\mathbb{Q}$.

It is desired to test the traditional null hypothesis

$$H_0 : \theta = \theta_0 \tag{1}$$

versus an unspecified alternative hypothesis

$$H_1 : \theta \neq \theta_0,$$

where θ_0 is a rational number, i.e. $\theta_0 \in \Theta_{\mathbb{Q}}$

.

Point-null zero probability paradox (Zero Probability paradox). *Probability of the null hypothesis (1) is equal to zero*:

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

This is tantamount to saying that probability of the null hypothesis

$$P\left(H_0 \mid \theta \in \{\text{All rational numbers}\}\right) = 0, \text{ and}$$
$$P\left(H_1 \mid \theta \in \{\text{All irrational numbers}\}\right) = 1.$$

Here, regard rationals on the number line as indicators of the means of corresponding normal distributions that have rational numbers as their means.

Proof:

A) In scientific research and statistical practice, any point null hypothesis of the normal population is almost always stated as a single rational number.

B) As proved by Cantor in 1873, rational numbers are countable—that is, there is in one-one correspondence between the rational numbers and the natural numbers (see, for example, Calkin and Wilf, 2000, for a binary tree argument). Because the rational numbers, q_i, are countable, enumerate them as a sequence $\{q_i\}$, or $\mathbb{Q} = \bigcup_{i=1}^{\infty}\{q_i\}$. Hence, the set of all hypothetical null values of the point-null hypotheses that could be expressed using rational numbers, $\Theta_{\mathbb{Q}}$, is also countable. In other words, this set has a bijective correspondence to the set of rational numbers.

C) The Lebesgue measure of any singleton set, $\{x\}$, is zero (where singleton means the smallest possible nonempty set). Every countable set has Lebesgue measure zero (see, for example, Adams and Guillemin, 1996, p. 9). Therefore, Lebesgue measure of the set of all rational numbers is also zero, that is $\lambda(\mathbb{Q}) = \lambda\left(\bigcup_{i=1}^{\infty}\{q_i\}\right) = \sum_{i=1}^{\infty}\lambda\left(\{q_i\} = 0\right.$.

In light of this fact, Lebesgue measure of the set of all hypothetical null values of the point-null hypotheses that could be expressed using rational numbers ($H_0 : \theta \in \Theta_{\mathbb{Q}}$) is also zero because this set is countable, $\lambda(\Theta_{\mathbb{Q}}) = 0$.

D) Normal distribution is absolutely continuous with respect to the Lebesgue measure λ. This signifies that all sets which have zero Lebesgue measure must also have zero probability under probability

measure; i.e., for all events $A \in R$ such that $\mu(A) = 0 \rightarrow P_X(A) = 0$. As Borovkov (2013, p. 39) has nicely exemplified "for an absolutely continuous distribution, the probability of hitting a set of zero Lebesgue measure is zero."

E) Because for an absolutely continuous distribution, a countably infinite set of all rational numbers has Lebesgue measure zero, conclude their probability measure is also zero.

F) Therefore, probability measure of a set of all possible hypothetical null rational values of the point-null hypotheses in testing a normal mean is also zero, $P\left(\left\{H_0 \mid \forall \theta \in \Theta_{\mathbb{Q}}\right\}\right) = 0$. This unequivocally 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 | Normal distribution) = 0.

Quod erat demonstrandum.

Subsequently, the probability of point null formulated as an irrational number is one. Figuratively speaking, rationals occupy zero length on a real line and the set of irrationals is uncountably infinite.

The scope of the Zero probability paradox can be further extended to the even more general set of all point null hypotheses asserted 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 such as $\sqrt{3}$, etc. Because this set is countable, as also proved by Cantor in 1874, (see, for example, Kaplansky, 2001, Paradox 4, p. 23) it has Lebesgue measure zero and therefore under Gaussian distribution its probability is zero. The cardinality (a measure of the "number of elements of the set") of the algebraic numbers is $\aleph_0$ (aleph-naught), the same as the natural numbers and rational numbers. However, the cardinality of the set of transcendental numbers is the same as that of the set of real numbers $|\mathbb{R}|$, the cardinality of the continuum. Almost all real numbers are transcendental, but we are familiar with almost none of them (except, for example, π, e, Liouville numbers, Champernowne constant, etc.).

It is important to emphasize that the Zero probability paradox applies both in the case when population variance is known and unknown.

It might be objected that a point-null hypothesis that the mean of the errors made in astronomical observations is equal to zero is reasonable and that its probability could be larger than zero. Karl Pearson (1935a, p. 296) replied, "I have never found a normal curve fit anything if there are enough observations! The astronomical data provided to prove that errors of observation follow normal curves are pitiably scanty, and if proper tests are applied usually show that they do not!"

Conclusion

Zero Probability and Impossibility.

Before discussing some of the implications of the Zero probability paradox, it is of considerable interest to clarify the difference between zero probability and impossibility. The most common and persistent misconception in the literature about probability is the interpretation that zero probability implies that an event is impossible. This is equally shared by many applied statistics textbooks writers (for example, Everitt, 1999, p. 14; de Muth, 2014, p. 20; Burns & Burns, 2008, p. 164; Sharma, 2010, p. 191) and non-statisticians (for example, Poole & Mackworth, 2010, p. 296; Finlayson & McMahon, 2004, p. 360; Yoe, 2012, p. 305; Quinn and Keough, 2002, p. 7). This does not come as a surprise since many notable scholars held the same false impression in the past.

As reported by Finetti (2008, p. 49), Borel used to say "let us consider the probabilities 10^{-3}, 10^{-10}, 10^{-100}, 10^{-1000}. A probability of 10^{-1000} is roughly equal to the probability of picking by chance a particular atom in the entire universe." Indeed, Borel (1962, p. 3), one of the founding fathers of measure theory, proposed in a book for the non-scientists published in 1943 "the single law of chance," or Borel's law. It states "Events with a sufficiently small probability never occur; or at least, we must act, in all circumstances, as if they were *impossible*." Similar interpretations were given by many other eminent scientists who tried to relate probabilities to the physical world. For example, Bernoulli (1713, pp. 211-212) stated in the first chapter of Part IV of his *Ars Conjectandi* that "if one thing is considered morally certain which has 999/1000 certainty, another thing will be morally impossible, which has only 1/1000 certainty." Cournot (1843, p. 78) also tried to build a bridge from probability theory to the physical world by stating that "*a physically impossible event is one whose probability is infinitely small*." Likewise, Popper (2002, p. 195) pointed out that

"the rule that extreme improbabilities have to be neglected…agrees with the demand for *scientific objectivity*."

However, today, there is an almost general agreement among statisticians that probability zero means "almost surely impossible" or extremely unlikely. In other words, an event of zero probability will almost never happen but there may be exceptions. For example, Kolmogorov (1956, p. 5) emphasized that "P(A) = 0 does not imply the impossibility of A…all we can assert is that…event A is practically impossible." According to Hand (2014, p. 6), "extremely improbable events are commonplace. It's a consequence of more fundamental laws, which all tie together to lead inevitably and inexorably to the occurrence of such extraordinarily unlikely events." Although we approve of Hand's position that events of vanishingly small probability will ultimately happen, we strongly disagree with establishing statistical tests on point-null hypotheses and expecting for coincidences and miracles to happen.

In light of the previous discussion, we restate the Zero probability paradox in the following, more comprehensible way: in practice, when testing a mean of the normal distribution using a point-null hypothesis, the probability of that hypothesis is zero. This does not imply that it is "absolutely" impossible to state a true point-null hypothesis, but that formulated point-nulls in the context of the simple normal model are almost surely false.

Some Implications of the Zero Probability Paradox.

Fisher's illuminating words (1922) are more relevant today than in 1922:

> It cannot be denied that, during the recent rapid development of practical methods, fundamental problems have been ignored and fundamental paradoxes left unresolved…This anomalous state of statistical science…the obscurity which envelops the theoretical bases of statistical methods may perhaps be ascribed to two considerations. In the first place, it appears to be widely thought, or rather felt, that in a subject in which all results are liable to greater or smaller errors, precise definition of ideas or concepts is, if not impossible, at least not a practical necessity. In the second place, it has happened that in statistics a purely verbal confusion has hindered the distinct formulation of statistical problems. (p. 311-312)

We argue that the Zero probability paradox has a specific power to shed new light on some fundamental problems in the foundations of statistical science that have been ignored, and help us to resolve some accumulated anomalies related to the point-null hypothesis testing, including so-called large *n* problem in significance testing, and the Jeffreys-Lindley paradox. It can also elucidate the notion of the Bayes factor, mixed prior distribution advocated by Jeffreys, "irreconcilability of *p*-values and evidence" (Berger & Sellke, 1987), and Cromwell's rule (Lindley, 1991, p. 104), among others.

However, detailed consideration of the implications of the Zero probability paradox for the Fisherian significance testing, Neyman-Pearson hypothesis testing, and Bayesian testing are beyond the scope of this paper. We confine ourselves, therefore, only to some general implications. Berkson (1938) was the first to notice dependence of significance testing on the sample size. He objected that it is possible to obtain a statistically significant chi-square test merely by increasing sample size:

> I believe that an observant statistician who has had any considerable experience with applying the chi-square test repeatedly will agree with my statement that, as a matter of observation, when the numbers in the data are quite large, the P's tend to come out small… we have something here that is apt to trouble the conscience of a reflective statistician using the chi-square test. For I suppose it would be agreed by statisticians that a large sample is always better than a small sample. If, then, we know in advance the P that will result from an application of a chi-square test to a large sample there would seem to be no use in doing it on a smaller one. *But since the result of the former test is known, it is no test at all*!" [italicized for emphasis]

Berkson failed to recognize that the same deficiency (sensitivity to sample size) is also shared by other significance tests based on point-null hypotheses and continuous data. Today this is well known as the large *n* problem. As argued by Mayo (2006, p. 809): "for any discrepancy from the null, however small, one can find a sample size such as there is a high probability (as high as one likes) that the test will yield a statistically significant result (for any *p*-value one wishes)." She claims that the large *n* problem is the basis for the famous Jeffreys-Lindley paradox (Lindley, 1957), probably the most quoted divergence between the frequentist and Bayesian approaches to inference. A number of suggestions have been proposed to alleviate this problem, including adjustment of *p*-values to a

fixed sample size (Good, 1988, p. 391), rules of thumb for decreasing α as *n* increases, and indicated effect size.

Karl Pearson (1935b, p. 550) opined "there is only one case in which an hypothesis can be definitely rejected, namely when its probability is zero." Relating this to the Zero probability paradox leads to the following conclusion. Focusing on the inferential aspects of the problem (not on the decision-making approach) permits rejecting the point-null hypothesis a priori, before seeing data. To paraphrase Berkson, because the result of the significance tests are known, they are no test at all. Term testing is a misnomer in this case and should be replaced by inexactification. These tests are merely procedures that ask researchers to waste their time and financial resources, to collect enough data, and when ultimately reject their point nulls to confirm what they knew beforehand, that their point nulls were almost surely false.

Zero probability paradox points to the root cause of the large *n* problem and discloses that there is no cure for it under the current point-null paradigm. Because classical significance tests (*Z* and *t*-test) are consistent, as the sample size increase, they will become extremely sensitive and therefore, detect even the tiniest discrepancy from the crisp hypothetical (almost surely false) null hypothesis. In other words, classical test statistic converges almost surely to ∞ and therefore, gives the asymptotically correct result (see, for example, DasGupta, 2008, p. 337, or Lehman and Romano, 2005, p. 462). Again, this means that in the real world testing any sharp null hypothesis of the normal mean will be ultimately, almost surely, rejected with large enough sample size.

This significant logical inconsistency of the significance testing was not an overwhelming issue in the first half of the past century when Gosset was "'naughtily' playing about with absurdly small numbers" (Eagon Pearson, 1939, p. 217). However, if Efron's view (2010, p. VII) is embraced that in the 21st century, statisticians will deal with large data sets and complex questions, it is clear that the current point-null paradigm is inadequate. Van der Laan and Rose (2010), for example, indicated that next generation of statisticians must construct new tools for massive data sets since the current ones are severely limited. Similarly, Hand (1998, p. 113) insisted in data mining instead of "statistical significance, consider more carefully substantive significance: is the effect important or valuable or not?"

To rephrase Box (1979): the only question of interest is "Is the normal model based on point-null hypothesis illuminating and useful?" The answer must be "No".

So, what should we do? This article is an initial contribution to making a paradigm shift in testing statistical hypotheses. Instead of testing highly

problematic and almost surely false point null hypotheses, as a natural replacement, test a negligible null hypothesis:

$H_0 : |\theta - \theta_0| \leq \delta$ (Effect size is negligible) against
$H_1 : |\theta - \theta_0| > \delta$ (Effect size is practically meaningful).

We propose naming this avant-garde proposal the "Hodges-Lehmann paradigm". Hodges and Lehmann (1954) were among first statisticians who had noted deficiencies of the point null hypothesis and formulated testing of "material significance" in their path-breaking paper "Testing the approximate validity of statistical hypotheses". We do not regard the Hodges-Lehmann paradigm as deus ex machine, nor as a magic alternative to the traditional point-null testing. However, we argue that it will substantially improve scientific research based on statistical testing. The argument that point nulls are mathematically more tractable is obsolete and belongs to the pre-MCMC era.

We regard statistics as the grammar of science. Thus, we are responsible for providing unambiguous rules of that grammar. We should not feel proud if non-statisticians are trying to make reform in statistical inference and statistics education. We, statisticians, are accountable to provide researchers in other sciences non-conflicting, coherent, and consistent concepts of testing the statistical hypotheses. Otherwise, significance tests "can actually impede scientific progress." (Kirk, 2003, p. 100) and even harm "development of scientific knowledge" (Armstrong, 2007, p. 321). Researchers and scientists will feel confused and deceived by statistics and statisticians. As pointed out by Cousins (2014, p. 35): "More than a half century after Lindley drew attention to the different dependence of *p*-values and Bayes factors on sample size *n* (described two decades previously by Jeffreys), there is still no consensus on how best to communicate results of testing scientific hypotheses."

Presumably, we all agree on the point that overcoming of accumulated inconsistencies is always a crucial method in science. As pointed out by Good (1982, p. 489), "a Bayes/non-Bayes compromise or synthesis is necessary for human reasoning." We argue that this compromise is impossible to reach within the point null-hypothesis testing paradigm, as Jeffreys-Lindley paradox evidently testifies.

In sharp contrast to the current point-nulls model, we argue that 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 the similar conclusions in (a) one-sided testing, (b) two-sided testing, and (c) interval estimation.

However, to make this proposal fully justifiable it is necessary to obtain a proof that point nulls are also almost always false in the case of two samples. The initial clue is given by Tukey (1991, p. 100):

> "Statisticians classically asked the wrong questions—and were willing to answer with a lie, one that was often downright lie. They asked "Are the effects of A and B different?" and they were willing to answer "no." All we know about the world teaches us that the effects of A and B are always different—in some decimal place—for any A and B. Thus asking 'Are the effects different' is foolish."

Only then, we can set as one of the fundamental rules of the 21st century Statistical Science Decalogue: *Hypotheses exactas non fingo*!

References

Adams, M. & Guillemin, V. (1996). *Measure theory and probability*. Basel, Switzerland: Birkhäuser.

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*, *27*(325–336), 186–190.

Armstrong, J. S. (2007). Significance tests harm progress in forecasting. *International Journal of Forecasting*, *23*(2), 321-327. doi: 10.1016/j.ijforecast.2007.03.004

Berger, J. O. & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*(3), 317-335. doi: 10.1214/ss/1177013238

Berger, J. O. & Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p values and evidence, *Journal of the American Statistical Association*, *82*(397), 112-122. doi: 10.1080/01621459.1987.10478397

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, *33*(203), 526-542. doi: 10.1080/01621459.1938.10502329

Bernardo, J. M. (1999). Nested hypothesis testing: the Bayesian reference criterion. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, Eds.

Bayesian Statistics (6th Ed.). Oxford: Oxford University Press, 101-130 (with discussion).

Bernoulli, J. (1713). *Ars conjectandi*. (Opus posthumum). Impensis Thurnisiorum, Fratrum.

Borel, E. (1962). *Probabilities and life*. Mineola, New York: Dover Publications.

Borovkov, A. (2013). *Probability theory*. NY: Springer.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building, In R. L. Launer and G. N. Wilkinson, Eds. *Robustness in Statistics*. Cambridge, MA: Academic Press, pp. 201–236.

Buchanan-Wollaston, H. J. (1935). Statistical tests. *Nature*, *136*(3431), 182-183. doi: 10.1038/136182b0

Burns, R. P. & Burns, R. (2008). *Business research methods and statistics using SPSS*. London: Sage Publications.

Calkin, N. & Wilf, H. (2000). Recounting the rationals. *American Mathematical Monthly*, *107*(4), 360–363. doi: 10.2307/2589182

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304-1312. doi: 10.1037/0003-066x.45.12.1304

Cousins, R. D. (2014). The Jeffreys–Lindley paradox and discovery criteria in high energy physics. *Synthese*, 1-38. doi: 10.1007/s11229-014-0525-z

Cournot, A. A. (1843). *Exposition de la théorie des chances et des probabilities*. Paris: Hachette.

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge, UK: Cambridge University Press.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. NY: Springer.

de Finetti, B. (2008). *Philosophical lectures on probability: collected, edited, and annotated by Alberto Mura*. NY: Springer.

De Muth, J. E. (2014). *Basic statistics and pharmaceutical statistical applications* (3rd Ed.). Boca Raton, FL: CRC Press.

Edwards, W., Lindman, H. & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193-242. doi: 10.1037/h0044139

Efron, B. (2010). The Future of Statistics. In M. Lovric (Ed.). *International Encyclopedia of Statistical Science*. NY: Springer, pp. VII-X.

Everitt, B. (1999) *Chance rules: an informal guide to probability, risk and statistics*. NY: Springer-Verlag.

Finlayson, B. L. & McMahon, T. A. (2004). *Stream hydrology: an introduction for ecologists*. Chichester, UK: John Wiley & Sons.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A, 222*(594-604), 309-368. doi: 10.1098/rsta.1922.0009

Ghosh, J. K, Delampady M. & Tapas, S. (2006). *An introduction to Bayesian analysis: theory and methods*. NY: Springer.

Good, I. J. (1994). The existence of sharp null hypothesis. *Journal of Statistical Computation and Simulation*, *49*(3-4), 241-242. doi: 10.1080/00949659408811587

Good, I. J. (1956). Which comes first, probability or statistics? *Journal of the Institute of Actuaries*, *82*(2), 249-255. doi: 10.1017/S0020268100046448

Good. I. J. (1988). The interface between statistics and philosophy of science. *Statistical Science, 3*(4), 386-397. doi: 10.1214/ss/1177012754

Hand, D. (1998). Data mining: statistics and more? *The American Statistician*, *52*(2), 112-118. doi: 10.1080/00031305.1998.10480549

Hand, D. J. (2014). *The improbability principle: why coincidences, miracles, and rare events happen every day*. NY: Scientific American / Farrar, Straus and Giroux.

Hodges, J. L. & Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society*. *Series B, 16*(2), 262–268.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8*(1), 3-7. doi: 10.1111/j.1467-9280.1997.tb00534.x

Kadane, J. B. (1987). [Testing precise hypotheses]: Comment. *Statistical Science, 2*(3), 347–348. doi: 10.1214/ss/1177013244

Kaplansky, I. (2001). *Set Theory and Metric Spaces* (2nd Ed.). Providence, RI: AMS Chelsea Publishing.

Kass, R. E. & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773-795. doi: 10.1080/01621459.1995.10476572

Kirk, R. E. (2003). The importance of effect magnitude. In S. F. Davis (Ed.), *Handbook of research methods in experimental psychology*. Malden, MA: Blackwell, pp. 83–105

Kolmogorov, A. N. (1956). *Foundations of the theory of probability* (2nd English Ed.). Providence, RI: AMS Chelsea Publishing.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, *56*(1), 16-26. doi: 10.1037//0003-066x.56.1.16

Kruskal, W. H. (1968). Significance, tests of. In: *The international encyclopedia of the social sciences*, 14. NY: The Macmillan Co. and the Free Press, pp. 238–250.

Lehmann, E. L. & Romano, J. P. (2005). *Testing statistical hypotheses*, 3rd Ed. NY: Springer.

Lindley, D. V. (1957) A statistical paradox. *Biometrika*, *44*(1-2), 187-192. doi: 10.1093/biomet/44.1-2.187

Lindley, D. V. (1975). The future of statistics: a Bayesian 21st century. *Advances in Applied Probability*, *7*(Supplement: Proceedings of the Conference on Directions for Mathematical Statistics), 106-115. doi: 10.2307/1426315

Lindley, D. V. (1991). *Making decisions* (2nd Ed.). NY: John Wiley & Sons.

Mayo, D. 2006. Philosophy of statistics. In: S. Sarkar & J. Pfeifer, Eds. *The philosophy of science: an encyclopedia*. London: Routledge, pp. 802–815.

Meehl, P. E. (1967). Theory testing in psychology and physics: a methodological paradox. *Philosophy of Science*, *34*(2), 103-115 (doi: 10.1086/288135). Reprinted in *The significance test controversy - a reader*, D. E. Morrison and R. E. Henkel, Eds. 1970. London: Aldine Publishing Company (Butterworth Group).

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241-301. doi: 10.1037//1082-989x.5.2.241

Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, *20*(4), 641-650. doi: 10.1177/001316446002000401

Orlitzky, M. (2012). How can significance tests be deinstitutionalized? *Organizational Research Methods*, *15*(2), 199-228. doi: 10.1177/1094428111428356

Pearson, E. S. (1939). "Student" as statistician. *Biometrika*, *30*(3-4), 210-250. doi: 10.2307/2332648

Pearson, K. (1935a). Statistical tests. *Nature*, *136*(3434), 296-297. doi: 10.1038/136296a0

Pearson, K. (1935b). Statistical tests. *Nature*, *136*(3440), 550. doi: 10.1038/136550a0

Playfair, J. (1819). *Elements of geometry: containing the first six books of Euclid, with a supplement on the quadrature of the circle and the geometry of solids; to which are added, Elements of plane and spherical trigonometry*. NY: G. Long.

Poole, D. L. & Mackworth, A. K. (2010). *Artificial intelligence: foundations of computational agents*. Cambridge, UK: Cambridge University Press.

Popper, K. R. (2002).*The logic of scientific discovery*, 2nd Ed. NY: Routledge.

Popper, K. R. (1963). *Conjectures and refutations: the growth of scientific knowledge*. New York: Harper.

Quinn, G. P. & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge, UK: Cambridge University Press.

Rousseau, J. (2007). Approximating interval hypothesis: p-values and Bayes factors. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, Eds. *Bayesian Statistics* (8th Ed.). Oxford: University Press, pp. 417-452.

Savage, I. R. (1957). Nonparametric Statistics. *Journal of the American Statistical Association*, *52*(279), 331-344. doi: 10.1080/01621459.1957.10501392

Savage, L. J. (1954). *The foundations of statistics*. New York: John Wiley and Sons.

Sharma, J. K. (2010). *Fundamentals of business statistics*. Noida, India: Dorling Kindersley.

Sprenger, J. (2013). Testing a precise null hypothesis: the case of Lindley's paradox. *Philosophy of Science*, *80*(5), 733–744. doi: 10.1086/673730

Trafimow, D. & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1-2. doi: 10.1080/01973533.2015.1012991.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, *6*(1), 100-116. doi: 10.1214/ss/1177011945

van der Laan, M. J. & Rose, S. (2010). Statistics ready for a revolution: next generation of statisticians must build tools for massive data sets. *Amstat News*, *399*, 38-39.

Wasserstein, R. L. & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *American Statistician*. *70*(2), 129-133. doi: 10.1080/00031305.2016.1154108

Wilkinson, L. (1999). Statistical methods in psychology journals: guidelines and explanations. *American Psychologist*. *54*(8), 594-604. doi: 10.1037/0003-066x.54.8.594

Yoe, C. (2012). *Principles of risk analysis: decision making under uncertainty*. Boca Raton, FL: CRC Press.

Zellner, A. (1987). [Testing precise hypotheses]: Comment. *Statistical Science*, *2*(3), 339-341. doi: 10.1214/ss/1177013241