Logical Defects in Null Hypothesis Testing and p-Value Calculation

Authors: Jiang Hongbing, Gao Lin, Xiang Yulin, Zhang Chenxi, Jiang Hongbing

Date: 2024-05-26T00:00:00+00:00

Abstract

Misuse and abuse of null hypothesis significance testing (NHST) and p-values have become quite serious in scientific research. NHST is a hybrid of Fisher's significance test and the Neyman-Pearson hypothesis test, yet how exactly this hybridization is formed and how it manifests in computational steps remains unclear. Where do the logical flaws in NHST and p-value calculation lie? These questions lack detailed and accessible explanations. Clearly explicating the procedures of Fisher's significance test, the Neyman-Pearson hypothesis test, and NHST, analyzing and comparing them, and supplementing this with typical examples to analyze the logical flaws in NHST and p-value calculation can provide valuable insights for empirical researchers who are not deeply specialized in statistics.

Full Text

Preamble

Logical Defects in Null Hypothesis Significance Testing and p-Value Calculation

Jiang Hongbing*, Gao Lin, Xiang Yulin, Zhang Chenxi
(School of Management, Zhengzhou University, Zhengzhou 450001, China)

Abstract

The misuse and abuse of null hypothesis significance testing (NHST) and p-values have become quite serious in scientific research. NHST represents a hybrid of Fisher's significance test and the Neyman-Pearson (N-P) hypothesis test, yet detailed and accessible explanations of how this hybridization occurs and how it manifests in computational procedures remain scarce. Similarly, the

precise location of logical flaws in NHST and p-value calculations has not been thoroughly addressed in plain terms. This paper provides a clear exposition, analysis, and comparison of the procedures underlying Fisher's significance test, the N-P hypothesis test, and NHST, supplemented by typical examples to analyze the logical defects in NHST and p-value calculations. Our aim is to offer insights for empirical researchers who have not specialized deeply in statistics.

Keywords: p-value; null hypothesis significance testing; Neyman-Pearson hypothesis test; Fisher significance test

Main Text

The dominant research strategy in empirical studies—the hypothetico-deductive method—typically begins with a research question, reviews relevant literature and theory, discusses theoretical hypotheses that help answer the research question, and then derives research hypotheses from these theoretical hypotheses. This is followed by research design (selecting or developing construct measurement tools, designing data collection and analysis methods), actual data collection, preprocessing, analysis, interpretation, and finally summarization and report writing [1]. A core component of this process is null hypothesis significance testing [2, 3]. Generally, research hypotheses appear as alternative hypotheses in NHST, where researchers hope to obtain p-values below certain thresholds (e.g., 0.05, 0.001) to reject the null hypothesis and support their research hypotheses. However, by mistakenly treating p-values as measures of evidence strength and pursuing publication, researchers often neglect the substantive validity of their work, leading to publication bias. Consequently, some journals have explicitly prohibited using p-values as the sole criterion for research validity. For example, the *American Journal of Public Health* has required authors to remove all p-values since 1983, otherwise directing them to submit elsewhere; *Epidemiology* declared at its founding in 1990 that "ignoring significance tests would improve the likelihood of manuscript acceptance… we simply do not adopt this method" [4]. In 2016, the American Statistical Association issued a statement on six principles regarding p-values [5]. While these principles represent longstanding concerns among statisticians, this marked the first time such an influential international statistical organization formally addressed p-value issues. Shortly thereafter, the top American political science journal *Political Analysis* announced in 2018 that it would ban p-values on the grounds that "p-values themselves cannot provide evidence supporting relevant patterns or hypotheses." Clearly, the misuse and abuse of null hypothesis significance testing and p-values have become quite severe [6].

Lü Xiaokang (2014) [4] argues that NHST is a hybrid of Fisher's significance test and the Neyman-Pearson hypothesis test, representing neither purely Fisher's approach nor purely N-P's approach, but rather a pragmatic compromise between utility and mathematical elegance. However, how exactly are Fisher's significance test and the N-P hypothesis test hybridized? How does this hybridization manifest in computational steps? Lü Xiaokang (2014) does not pro-

vide detailed answers. Additionally, p-value application remains controversial in NHST: are p-values merely misused, or do their calculations contain inherent logical flaws? Hao Li et al. (2016) [7] argue that p-values are misused for two reasons: first, simplifying p-values as "the probability that the null hypothesis is true," effectively treating them as equivalent to α values in N-P testing [4]; second, most research follows the logic of "inferring from 'null hypothesis is true' to 'alternative hypothesis is false,' then extending 'p-value is the probability the null hypothesis is true' to 'p-value is the probability the alternative hypothesis is false.'" In other words, while p-values only measure evidence against the null hypothesis, most scholars believe p-values provide sufficient evidence to evaluate research hypotheses (alternative hypotheses). Other scholars contend that p-value calculations themselves contain logical flaws; for instance, Lindley (1993) [8] notes that p-values may differ depending on the conceived experimental design.

In summary, the unreasonable hybridization of Fisher's significance test and N-P hypothesis testing creates logical defects in NHST, which also explains why p-values are misinterpreted or misused. Additionally, p-values themselves suffer from computational logical flaws. This paper employs literature analysis to compare the testing procedures of Fisher's significance test, N-P hypothesis test, and NHST, analyzing in detail how NHST constitutes a hybrid of the former two. Simultaneously, we analyze the logical defects in NHST and p-value calculations through specific computational examples.

We then address why NHST and p-values continue to dominate despite repeated criticism. This paper's contribution lies not in creating new knowledge about hypothesis testing, but in clarifying issues and providing accessible explanations for empirical researchers without deep statistical training, enabling them to better understand the logical defects in NHST and p-value calculations.

2.1 Fisher Significance Test

NHST is essentially a hybrid of Fisher's and N-P's ideas about hypothesis testing, containing various internal contradictions [9, 10]. In Fisher's framework, the p-value represents the probability of obtaining the observed or more extreme data, assuming hypothesis H0 is true and other relevant assumptions hold. A p-value below a given threshold only indicates that H0 is wrong or that a rare event has occurred [11], but typically one cannot determine whether to reject H0 based on a single trial. Fisher argued that H0 could only be reasonably rejected when, after multiple trials with no important design flaws, statistically significant results constitute an overwhelming majority [12]. Therefore, a statistically significant result from a single trial only provides suggestive evidence that the result warrants attention and requires further investigation. To test another hypothesis requires designing a separate testing procedure rather than rejecting one hypothesis and simultaneously accepting another in a single test. Consequently, Fisher considered the introduction of alternative hypotheses entirely unnecessary [13].

Fisher's significance test procedure can be summarized as follows [3, 14, 15]: 1. Specify the statistical hypothesis H0. 2. Select an appropriate test statistic T and determine its distribution under the assumption that H0 is true. 3. Calculate the value t of the test statistic T based on current experimental data. 4. Determine the significance level p corresponding to t according to T's distribution (under H0). 5. If the obtained p is smaller than the expected value, then either H0 is not true or a rare event has occurred.

2.2 Neyman-Pearson Hypothesis Testing

[Figure 1: see original paper]

The N-P hypothesis test employs the form of null hypothesis versus alternative hypothesis. Its essence is to minimize the probability of Type II error β while constraining the probability of Type I error not to exceed significance level α [16]. N-P hypothesis testing theory is based on repeated sampling and cannot guarantee whether decisions to accept or reject based on a single sample's observations are correct or incorrect [13]. Moreover, the N-P framework makes no mention of p-values; instead, they use rejection regions to judge hypotheses.

N-P hypothesis testing originated from quality control needs in industrial production [17]. For example, a factory producing screw caps receives an order requiring diameters of 2 ± 0.01 cm. In this context, the minimum effect size needed to detect defective products is clearly 0.01 cm. Sample size is easily controlled, as drawing large samples from thousands of screw caps presents no difficulty. On one hand, if the factory cannot effectively detect caps that are too large (>2.001 cm) or too small (<1.999 cm)—that is, if the probability of Type II error is too large—orders may be canceled. On the other hand, if caps meeting the diameter standard (2 ± 0.01 cm) are frequently judged as defective—meaning the probability of Type I error is too large—the factory suffers losses and unnecessarily increases production costs. In such situations, the benefits and costs of reducing or increasing these two types of errors can be measured in monetary terms through appropriate conversion. This means we can use mathematical optimization methods to determine the most appropriate α and β values that minimize loss or maximize benefit.

However, in most actual research, we cannot easily control sample size, nor can we readily determine minimum effect sizes or reasonable α and β values. For example, when testing whether two groups of people differ significantly in intelligence, we often do not know the required minimum effect size. Although some rules of thumb exist [18-20] telling us what constitutes large, medium, and small effect sizes, judging effect size magnitude depends heavily on the research question. For instance, when developing a new drug, even a small effect size improving cure rates may be acceptable.

The N-P hypothesis testing procedure can be summarized as follows [15, 21, 22]: 1. Specify two statistical hypotheses : null hypothesis HM and alternative hypothesis HA. 2. Select an appropriate test statistic T and determine its

distribution under HM. 3. Specify the maximum acceptable probability α of committing a Type I error. 4. Based on (1), (2), (3), and specified statistical power , minimum effect size, etc., calculate the minimum sample size. 5. Based on the Neyman-Pearson lemma and its extensions, calculate the rejection region C. 6. Calculate the value t of the test statistic T based on current experimental data. 7. If t falls in rejection region C, reject the null hypothesis and accept the alternative hypothesis; otherwise, accept the null hypothesis.

2.3 Null Hypothesis Significance Testing (NHST)

[Figure 2: see original paper]

NHST is a hybrid of Fisher's significance test and N-P hypothesis test. Its typical procedure is as follows [4, 15, 23]: 1. Specify two statistical hypotheses: null hypothesis HM and alternative hypothesis HA. 2. Select an appropriate test statistic T and determine its distribution under HM. 3. Specify the maximum acceptable probability α of committing a Type I error. 4. Calculate the value t of the test statistic T based on current experimental data. 5. Determine the significance level p corresponding to t according to T's distribution. 6. If $p \leq \alpha$, reject HA and accept HM; if $p > \alpha$, accept HA and reject HM.

[Figure 3: see original paper]

Here, hypotheses refer to simple hypotheses. If a statistical hypothesis can completely determine the population distribution, it is called a simple statistical hypothesis; otherwise, it is called a composite hypothesis. When both null and alternative hypotheses are simple hypotheses, the Neyman-Pearson lemma can determine the form of the most powerful test. Generally, the most powerful rejection region for testing H0 vs. H1 may differ from that for testing H0 vs. H2. Therefore, when the alternative hypothesis is composite (e.g., H0: $= 0$ vs. H1: > 0), Neyman-Pearson did not completely solve how to determine its most powerful rejection region. Karlin and Rubin extended the Neyman-Pearson lemma, proposing the Karlin-Rubin theorem, which can derive uniformly most powerful tests for certain composite hypothesis testing problems.

When the alternative hypothesis is composite, calculating statistical power becomes complex. If this composite hypothesis contains simple hypotheses H1, H2, …, Hn, …, each simple hypothesis will have a statistical power, generally different: 1-β(H1|α), 1-β(H2|α), …, 1-β(Hn|α), … In this case, for an alternative hypothesis that is composite, the term "power function" is more appropriate than "power." Both power and effect size can be specified with standard levels a priori, then actual levels calculated from samples.

The hybrid nature of NHST manifests as follows: Steps (1), (2), (3), and (4) in NHST correspond to steps (1), (2), (3), and (6) in N-P hypothesis testing, while steps (4) and (5) correspond to steps (3) and (4) in Fisher's significance test. However, step (6) in NHST lacks information about statistical power, minimum effect size, and rejection regions found in N-P hypothesis testing, mistak-

enly treating Fisher's p-value as "equivalent" to N-P's α value. Neither Fisher nor Neyman-Pearson would endorse NHST's computational process. At first glance, Fisher's hypothesis H0 appears identical to N-P's null hypothesis HM, with the latter merely adding an alternative hypothesis HA. In reality, the two frameworks differ substantially. This becomes clear when we alter the form of statistical hypotheses in N-P testing [22]: HM: M1 - M2 = 0 ± MES (minimum effect size); HA: M1 - M2 ≠ 0 ± MES. Here, M1 is a parameter of the probability distribution determined by HM, and M2 is a parameter determined by HA. If the research design does not utilize information provided by the alternative hypothesis HA (i.e., minimum effect size and Type II error probability), then N-P hypothesis testing degenerates into Fisher's significance test mode. Common statistical software like SPSS primarily follows Fisher's statistical testing philosophy [22], meaning that in most cases, empirical research papers we read actually employ Fisher's significance test, rendering the research hypothesis—as an alternative hypothesis in hypothesis testing—essentially decorative.

In summary, NHST selectively adopts attractive features from both Fisher's and N-P's frameworks while neglecting the preconditions necessary for each to function properly. Specifically, NHST only exploits the convenience of Fisher's p-value in measuring the degree of support experimental data provides for a hypothesis H, and N-P's ease of decision-making. However, it ignores the fact that effective decision-making in N-P hypothesis testing requires additional information such as β values, power values, minimum effect sizes, and alternative hypothesis HA, leading in practice to the mistaken belief that "not HM" is equivalent to alternative hypothesis HA [24-26].

3.1 Frequently Ignoring Information Provided by the Alternative Hypothesis

Directional hypotheses in social sciences do not utilize information from the alternative hypothesis and are therefore essentially Fisher significance tests. As long as the null hypothesis is rejected, whatever the alternative hypothesis may be, it will ultimately be accepted. This represents a serious misuse of statistics. The Sally Clark case, described by journalist Geoffrey Wansell as "the greatest miscarriage of justice in modern British legal history," powerfully illustrates this point. A woman named Sally Clark lost her first child shortly after birth; doctors found no other cause and diagnosed it as SIDS (Sudden Infant Death Syndrome). Unfortunately, her second child also died shortly after birth, leading police to suspect Clark had murdered both children. Pediatric expert Roy Meadow persuaded the jury with statistical evidence and reasoning, resulting in Clark's conviction for murder and a life sentence. Meadow's argument was that in a family like Clark's, the probability of one infant dying from SIDS was only 1/8,543, making the probability of two infants dying from SIDS $1/8{,}543^2$ 1/73,000,000. Therefore, the probability of Clark's innocence was only 1/73,000,000; a rare event occurred, so Clark must be guilty. In NHST terms: HM: both infants died from SIDS; HA: Sally Clark murdered

both infants. The probability of both infants dying from SIDS was too small, so HM was rejected. Rejecting the null hypothesis meant accepting alternative hypothesis HA, therefore Clark was guilty. In this reasoning, HA served merely as decoration, playing no substantive role. In fact, the probability of a mother murdering her own infant is lower than the probability of SIDS. One estimate places the probability of a mother killing her infant at approximately 1/92,000 [27], which is smaller than the 1/8,543 probability of dying from SIDS. In other words, if forced to choose between HA and HM, the rational choice should be HA: both infants died from SIDS.

3.2 Confusing the Meanings of p-values and α Values

If ignoring information from the alternative hypothesis is not inherent to NHST itself, then confusing p-values with α values represents a fundamental flaw for which NHST cannot escape responsibility [28]. The p-value, proposed by Fisher, is the probability of obtaining the current or more extreme data D, assuming hypothesis H and other relevant assumptions are true [15], denoted as pr(D|H). The p-value is a property of the current experimental data, measuring the degree to which current and more extreme data oppose hypothesis H. Each trial has one p-value; in other words, the p-value is a random variable—conduct one trial and calculate one p-value. The α value, proposed by Neyman-Pearson, expresses the meaning that under the condition that null hypothesis HM is true, conducting N (sufficiently many) trials will result in rejecting HM no more than N$\times\alpha$ times. α is a property of the test, not a property of experimental data. The main differences between p-values and α are: (1) p-values result from a single trial; α results from N trials; (2) p-values are properties of data; α is a property of the test; (3) p-value thresholds (e.g., 0.05, 0.001) can be determined either before or after trials; α thresholds must be determined before trials because rejection region calculation in N-P hypothesis testing requires α values; (4) p-values allow continuous strength-of-evidence judgments about data's support for H; α does not—it only permits trichotomous decisions: support, reject, or require further investigation.

Confusing p and α is the root of NHST's logical confusion. The p-value is a random variable; calculating a p-value from a single trial yields the same result whether the threshold of 0.05 or 0.001 is determined before or after the trial. α differs fundamentally—it is a predetermined value that experimenters use, along with other specifications, to design experiments. When designing experiments, regardless of whether a particular trial produces significant results, the long-run probability of committing a Type I error is guaranteed to be less than α. A concrete example illustrates this: setting $p = 0.05$ in Fisher's framework and $\alpha = 0.05$ in the N-P framework, then conducting 10,000 trials (assuming this is sufficiently many). Under Fisher's framework, the number of trials with $p < 0.05$ is uncertain—it could be 20, 50, 100, 500, etc. Under the N-P framework, however, the number of Type I error decisions (according to probability theory) would not exceed 250 (assuming HM is true in 10,000 trials).

3.3 Failing to Obtain the Results We Want

What we most want to obtain is this: based solely on a single trial's results, we want to determine the probability that null hypothesis H0 or alternative hypothesis H1 is true when results are significant. However, in actual calculations, we often cannot obtain such desired results. What most attracts us in Fisher's hypothesis testing is the p-value, which in NHST is equated with the α value. We typically mistakenly believe α is the probability of error if we reject the null hypothesis when a trial yields significant results. For example, when a study obtains $p < \alpha = 0.05$, it is often interpreted as: if we reject the null hypothesis, we would be wrong no more than 5 times out of 100 trials. This is a highly desirable interpretation, but the reality is far more complex. Such a judgment requires a premise: that the null hypothesis in the hypothesis test is actually true. If this premise does not hold, the above interpretation is invalid. As shown in [Figure 4: see original paper], although $\alpha = 0.05$, 36% of the 125 significant results could be erroneous, far exceeding the believed 5% error rate. In actual research, a true:false ratio of 9:1 for null hypotheses is quite normal [17].

It is also important to note that α and β are probabilities under long-run trial conditions, not obtainable under single-trial conditions. What constitutes long-run trials? Under the frequency school's definition of probability [29]: in N repeated trials where event A occurs KN times, the frequency of event A is PN(A) = KN/N = number of occurrences of A/number of repeated trials. Long-run trials relate to N; as the number of repetitions N increases, frequency stabilizes around a constant. This constant, independent of N, is the probability P(A) of event A occurring. Fisher, Neyman, and Pearson were all major representatives of the frequency school. Neyman and Pearson's hypothesis testing theory is built upon the frequency interpretation of probability [30]. However, in the real world, we cannot repeat a trial infinitely, making it difficult to obtain P(A). What we can often do is repeat trials sufficiently many times and use PN(A) to approximate P(A).

4.1 How p-Values Are Calculated

It was a summer afternoon in Cambridge in the late 1920s. A group of university gentlemen and their ladies, along with visitors, were sitting around an outdoor table enjoying afternoon tea. During the tea tasting, a lady insisted that adding tea to milk versus adding milk to tea would produce different tastes. This claim aroused everyone's interest, and someone suggested testing the proposition. Fisher detailed in Chapter 2 of *The Design of Experiments* how to design various schemes to test this lady's claim [12]. However, Fisher's presentation was overly complex. Here we use a scheme improved by Lindley that retains Fisher's essential ideas [8] to explain how p-values are actually calculated.

The improved experimental design presents the lady with two cups of tea each time, telling her that one cup has milk added first then tea, while the other

has tea added first then milk. The lady must identify which is which. If her judgment is correct, record R; otherwise, record W. The trial is repeated six times. Assuming the result is RRRRRW (only the last judgment is wrong), Fisher's analysis proceeds as follows.

First, assume the lady cannot distinguish between the two preparation methods (the null hypothesis). This means each judgment is random, with correct and incorrect outcomes each having probability 1/2, and each trial is independent of the others. The probability of observing RRRRRW would then be $1/2^6 = 1/64$. Fisher would conclude: either the null hypothesis is correct and a rare event occurred, or the null hypothesis is incorrect—that is, the lady can distinguish between the two preparation methods.

The result of 1/64 constitutes a rare event. According to Fisher's analysis, we might be inclined to think the null hypothesis is false. Fisher immediately realized this analysis was flawed because in this experimental arrangement of six trials, any possible outcome has probability 1/64, making it absurd to determine a hypothesis's truth based on such reasoning.

To avoid this absurdity, Fisher claimed that any result with five correct and one incorrect judgments (regardless of which trial was wrong) carries equal evidential weight. Since W could appear in any of six positions, the probability would be $6/2^6 = 6/64 = 0.094$, which is not significant at the 5% level. This avoids the absurd result where any observed outcome would be significant.

However, Fisher soon realized this still did not solve the problem. For example, if the lady made 300 judgments with 150 correct and 150 incorrect, the probability would be $\times(1/2^{300}) = 0.046$. This is the most likely outcome in 300 trials; any other outcome has a smaller probability. The same problem re-emerges: any outcome would be significant at the 5% level, which remains unreasonable. How to solve this? Let us temporarily return to the six-trial result RRRRRW. The brilliant Fisher realized that if five correct and one incorrect judgments can reject the null hypothesis, then six correct judgments should certainly reject it even more strongly. Therefore, the probability of six correct judgments should also be included: $0.094 + 1/64 = 0.109$. Applying the same logic to 300 trials, the probability of correct judgments exceeding $)\times(1/2^{300}) = 0.523$.

In summary, the p-value actually consists of three probability components: (1) the probability of the current experimental result; (2) the probabilities of possible results equally extreme as the current result; and (3) the probabilities of possible results more extreme than the current result. [Figure 5: see original paper] illustrates how p-values are calculated using the lady tasting tea example.

4.2 Logical Defects in p-Value Calculation

What event's probability does the p-value actually represent? In fact, it is not a single event but the sum of probabilities of multiple events. It involves the current experimental data plus data equally extreme and more extreme than the

current data [31]. As [Figure 5: see original paper] shows, the p-value combines probabilities of observed and unobserved portions. The logical problem with p-values primarily lies in how to calculate probabilities for the unobserved portion [8, 32-34]. How should we define possible results equally extreme as the current result? How should we define possible results more extreme than the current result? This is where p-values face logical dilemmas.

Using the same lady tasting tea example, assume we have two experimental designs. The first, already described, has the lady make six judgments, recording each set of six judgments as one trial result. The second design has the lady make judgments continuously until the first incorrect judgment appears, recording these judgments as one trial result.

Assume we obtain the trial result RRRRRW. Let us calculate the p-value. Under the first design, as previously analyzed, the p-value equals 0.109, and the null hypothesis cannot be rejected at the 5% significance level. Under the second design, what is the p-value? We calculate it using the same logic: (1) the probability of the current result is 1/64; (2) the probability of results equally extreme as the current result is 0; (3) results more extreme than the current result are RRRRRRW, RRRRRRRW, …, so their probability sum is $1/2^7$ + … + $1/2\hat{}\infty = (1/2^7) \div (1\text{-}1/2) = 1/64$. Thus, the p-value under the second design is: 1/64 + 0 + 1/64 = 0.031. At the 5% significance level, the null hypothesis would be rejected. This creates a contradiction: we provide the lady with the same information (the experimental design exists only in our conception, not communicated to her), the lady makes honest judgments, yet the same result (RRRRRW) yields contradictory conclusions. This is clearly unreasonable.

Another logical problem with p-values is whether identical p-values imply identical evidential strength. Assume we have the same null hypothesis. Experiment 1 has sample size 20 and yields p = 0.042; Experiment 2 has sample size 100 and also yields p = 0.042. Do these two results have the same evidential strength? If not, which result provides stronger evidence against the null hypothesis? Academic debate on this question is substantial. Some scholars argue Experiment 1 provides stronger evidence [35-37], others argue Experiment 2 does [38], while still others, particularly Fisher himself, hold that identical p-values have identical evidential strength as long as calculated correctly [39].

5. Why NHST and p-Values Remain Popular

Despite numerous scholars' questions and criticisms, NHST and p-values continue to be widely endorsed for several reasons:

Practicality: As an experienced applied statistician, Fisher understood that statistical tools should emphasize practical utility in work, while Neyman and Pearson, as pure mathematicians, pursued mathematical precision and perfection in statistical tools, leading to serious disagreements. As a hybrid, NHST absorbs Fisher's p-value for measuring experimental data's support for hypothesis H and N-P's decision-making convenience [40]. Yes-or-no decisions are

unavoidable facts in scientific research and daily life, and researchers are deeply wary of unsubstantiated yes-or-no decisions (especially those lacking numerical support). NHST and p-values greatly satisfy this psychological need, providing fertile ground for their survival.

Journal Orientation: The exemplary effect of top-tier journals in social sciences on publication practices has amplified p-value usage , evolving into a standardized empirical research procedure and methodological requirement: all statistical inferences must conduct hypothesis tests, and all hypothesis tests must report p-values. This requirement is repeatedly demonstrated in statistical textbooks across disciplines and ultimately becomes fully institutionalized throughout the field [5].

Anxiety about Scientific Status: According to Lü Xiaokang (2014) [4], "Regardless of discipline, social sciences face this pressure… they must constantly 'prove' to outsiders that they are scientific enterprises while internally integrating disciplinary systems and development directions. One way to address this pressure is to establish an integrated analytical framework that defines the discipline's basic theoretical paradigm while introducing a series of mathematical tools." NHST and p-values have served this function to some extent.

6. Research Conclusions

Building on Lü Xiaokang (2014) [4] and Hao Li et al. (2016) [7], this paper addresses three research questions: (1) Where lie the logical defects of null hypothesis significance testing (NHST)? (2) Where lie the logical defects of p-value calculation? (3) Why do these flawed methods remain dominant?

For question (1), we argue that NHST's key logical defect lies in confusing p-values with α values, thereby ignoring information provided by the alternative hypothesis and failing to obtain desired trial results. For question (2), we argue that p-value calculation suffers from logical defects: Different conceived experimental designs may yield different p-values for the same experimental result; Whether identical p-values imply identical evidential strength remains undetermined. For question (3), the answers are: They greatly reduce researchers' wariness of uncertain, data-less decisions; Top-tier journals' exemplary effects promote widespread NHST and p-value usage; Various social science disciplines need to prove their scientific status to outsiders, requiring an integrated analytical framework that NHST and p-values satisfy.

This paper does not deny the role of NHST and p-values in psychology, sociology, and other disciplines, consistent with views expressed by Lü Xiaokang (2014) [4] and Jiang Hongbing (2017) [1]. Although NHST and p-values face repeated criticism, they remain scholars' first choice for empirical research largely due to inertia in research tool usage, and changing this inertia requires sufficient time [13]. However, we must clearly recognize that both Fisher's significance test and N-P hypothesis testing require long-run trials to draw conclusions. We often lack such patience, always hoping to obtain definitive conclusions from a

single study. In this regard, NHST and p-values will certainly disappoint. Yet completely banning them is unnecessary [41]; truly understanding their logical defects and using them judiciously is the more reasonable approach.

For example, in 1991 *American Sociological Review* established a new publication requirement explicitly prohibiting significance levels above 0.05 and mandating the use of "*", "**", "***" to indicate $p<0.05$, $p<0.01$, and $p<0.001$ respectively.

References

[1] Jiang Hongbing. The Enlightenment of Scientific Perspectivism for Management Research[J]. Foreign Economics & Management, 2017, 39(03): 99-113.
[2] Feng Xiaotian. Twenty Years of Sociological Methods: Application and Research[J]. Sociological Studies, 2000(01): 1-11.
[3] Jiao Can, Zhang Minqiang. Lost Boundaries: Exploring Null Hypothesis Testing Methods in Psychology[J]. Social Sciences in China, 2014(02): 148-163+207.
[4] Lü Xiaokang. From Tool to Paradigm: A Sociological Reflection on Hypothesis Testing Controversies[J]. Society, 2014, 34(06): 216-236.
[5] Wasserstein R L, Lazar N A. The ASA Statement on P-values: Context, Process, and Purpose[J]. The American Statistician, 2016, 70(2): 129-133.
[6] Baker M. Statisticians Issue Warning over Misuse of p Values[J]. Nature News, 2016, 531(7593): 151.
[7] Hao Li, Liu Leping, Shen Yafei. Statistical Significance: A Misinterpreted P-value—Based on the ASA Statement[J]. Statistics & Information Forum, 2016, 31(12): 3-10.
[8] Lindley D V. The Analysis of Experimental Data: The Appreciation of Tea and Wine[J]. Teaching Statistics, 1993, 15(1): 22-25.
[9] Nuzzo R. Scientific Method: Statistical Errors[J]. Nature News, 2014, 506(7487): 150.
[10] Newman M C. "What Exactly Are You Inferring?" A Closer Look at Hypothesis Testing[J]. Environmental Toxicology and Chemistry: An International Journal, 2008, 27(5): 1013-1019.
[11] Hubbard R, Bayarri M J. Confusion Over Measures of Evidence (p's) Versus Errors (α's) in Classical Statistical Testing[J]. The American Statistician, 2003, 57(3): 171-178.
[12] Fisher R A. The Design of Experiments[M]. London: Oliver and Boyd, 1935.
[13] Lü Xiaokang. The Dispute Between Fisher and Neyman-Pearson and Hypothesis Testing Controversies in Psychological Statistics[J]. Psychological Science, 2012, 35(06):
[14] Gill, J. The Insignificance of Null Hypothesis Significance Testing[J]. Political Research Quarterly, 1999, 52(3): 647-674.
[15] Gigerenzer G. Mindless statistics[J]. The Journal of Socio-Economics, 2004, 33(5): 587-606.

[16] Neyman J. Frequentist Probability and Frequentist Statistics[J]. Synthese, 1977: 97-131.
[17] Szucs D, Ioannidis J. When Null Hypothesis Significance Testing is Unsuitable for Research: A Reassessment[J]. Frontiers in Human Neuroscience, 2017, 11: 390.
[18] Quan Chaolu. The Meaning and Measurement Methods of Effect Size[J]. Psychological Exploration, 2003(02): 39-44.
[19] Cohen J. The Statistical Power of Abnormal-social Psychological Research: A Review[J]. The Journal of Abnormal and Social Psychology, 1962, 65(3): 145.
[20] Sedlmeier P, Gigerenzer G. Do Studies of Statistical Power Have an Effect on the Power of Studies?[J]. Psychological Bulletin, 1989, 105(2): 309-316.
[21] Perezgonzalez J D. The Meaning of Significance in Data Testing[J]. Frontiers in Psychology, 2015, 6: 1293.
[22] Perezgonzalez, Jose D. Fisher, Neyman-Pearson or NHST? A Tutorial for Teaching Data Testing[J]. Front Psychol, 2015, 6: 223.
[23] Wu Xizhi. Statistics: From Data to Conclusions[M]. 4th Edition. Beijing: China Statistics Press, 2013.
[24] Patriota A G. On Some Assumptions of the Null Hypothesis Statistical Testing[J]. Educational & Psychological Measurement, 2016, 77(3): 507–528.
[25] Chang M. What Constitutes Science and Scientific Evidence: Roles of Null Hypothesis Testing[J]. Educational and Psychological Measurement, 2016, 77(3): 475–488.
[26] Häggström O. The Need for Nuance in the Null Hypothesis Significance Testing Debate[J]. Educational and Psychological Measurement, 2016, 77(4): 616–630.
[27] Sesardic N. Sudden Infant Death or Murder? A Royal Confusion About Probabilities[J]. The British Journal for the Philosophy of Science, 2007, 58(2): 299–329.
[28] Hubbard R, Bayarri M J. Confusion Over Measures of Evidence (p's) Versus Errors (α's) in Classical Statistical Testing[J]. The American Statistician, 2003, 57(3): 171–178.
[29] Mao Shisong, Zhou Jixiang. Probability Theory and Mathematical Statistics[M]. 1st Edition. China Statistics Press, 1996.
[30] Chen Xiru. Course in Mathematical Statistics[M]. 1st Edition. University of Science and Technology of China Press, 2009.
[31] Pollard P, Richardson J T. On the Probability of Making Type I Errors[J]. Psychological Bulletin, 1987, 102(1): 159–163.
[32] Wagenmakers E J. A Practical Solution to the Pervasive Problems of P Values[J]. Psychonomic Bulletin & Review, 2007, 14(5): 779-804.
[33] Schneider, Jesper W. Null Hypothesis Significance Tests. A Mix-up of Two Different Theories: the Basis for Widespread Confusion and Numerous Misinterpretations[J]. Scientometrics, 2015, 102(1): 411-432.
[34] Jeffreys H. Theory of probability[M]. Oxford: Oxford University Press, 1961.
[35] Rosenthal R, Gaito J. The Interpretation of Levels of Significance By

Psychological Researchers[J]. The Journal of Psychology, 1963, 55(1): 33–38.
[36] Nelson N, Rosenthal R, Rosnow R L. Interpretation of Significance Levels and Effect Sizes By Psychological Researchers[J]. American Psychologist, 1986, 41(11): 1299–1301.
[37] Peto R, Pike M C, Armitage P, et al. Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient. I. Introduction and Design[J]. British Journal of Cancer, 1976, 34(6):
[38] Bakan D. The Test of Significance in Psychological Research[J]. Psychological Bulletin, 1966, 66(6):
[39] Fisher R A, Statistical Methods for Research Workers[M]. London: Oliver And Boyd, 1950.
[40] Krueger J. Null Hypothesis Significance Testing. On The Survival of A Flawed Method[J]. Am Psychol, 2001, 56(1): 16-26.
[41] Zhong Xiaobo. The Controversy Over Hypothesis Testing: Clarification and Resolution of Issues[J]. Advances in Psychological Science, 2016, 24(10): 1670-1676.

Note: Figure translations are in progress. See original paper for figures.