Relevance of Null Hypothesis Significance Testing (NHST) in biomedical sciences: sociological approach

Relevancia de las pruebas de significación de hipótesis nulas (NHST) en ciencias biológicas: enfoque sociológico

Olha Sobetska[1]

[1] *Faculty of Social Sciences and Philosophy, Institute of Sociology, University of Leipzig, Beethoven Street 15, 04107, Leipzig, Germany*
[1] *Systemic Modeling and Applications, The Center Leo Apostel (CLEA), Vrije Universiteit Brussel (VUB), Krijgskundestraat 33, 1160, Brussels, Belgium*

Reception date of the manuscript: 24/03/2023
Acceptance date of the manuscript: 17/04/2023
Publication date: 28/04/2023

Abstract—Significance tests play a very important role in the scientific community, and the biomedical research community is not an exception. This is due, on the one hand, to the widespread use of the test in scientific methodology and the corresponding frequency of its application in research, and, on the other hand, to the general misinterpretation of the results obtained using this method. Misunderstanding of significance testing in academia and erroneous conclusions in research, regardless of the scientific field, are at the root of the distrust of this statistical method. This article aims to give insight into the relevance of this kind of method in the biomedical field and find a theoretical explanation for this phenomenon, and subsequently regulate the correct interpretation of the null hypothesis significance test (NHST), as well as consider alternative statistical methods. In addition, some relevant empirical studies from a geographical and multidisciplinary perspective will be presented to determine the real extent of misspecification at the academic level. In this way, both practical and theoretical arguments will be applied to address the problem of NHST at multiple levels.

Keywords—Bioinformatics, Biometric, p-value, Significance tests, Statistical testing, Misinterpretation

Resumen— Las pruebas de significación desempeñan un papel muy importante en la comunidad científica, y la comunidad de investigación biomédica no es una excepción. Esto se debe, por un lado, al uso generalizado de la prueba en la metodología científica y a la correspondiente frecuencia de su aplicación en la investigación y, por otro, a la mala interpretación generalizada de los resultados obtenidos con este método. La incomprensión de las pruebas de significación en el mundo académico y las conclusiones erróneas en la investigación, independientemente del ámbito científico, están en el origen de la desconfianza hacia este método estadístico. Este artículo pretende dar a conocer la relevancia de este tipo de método en el ámbito biomédico y encontrar una explicación teórica a este fenómeno, para posteriormente regular la correcta interpretación de la prueba de significación de hipótesis nula (NHST), así como considerar métodos estadísticos alternativos. Además, se presentarán algunos estudios empíricos relevantes desde una perspectiva geográfica y multidisciplinar para determinar el alcance real de la mala especificación a nivel académico. De este modo, se aplicarán argumentos tanto prácticos como teóricos para abordar el problema de la NHST a múltiples niveles.

Palabras clave— Bioinformática, Biometría, p valor, Pruebas de significación, Pruebas estadísticas, Interpretación errónea

Introduction

Debates, warnings, prohibitions, and precautions against null hypothesis significance testing (NHST) have become commonplace in the scientific community. Moreover, criticisms of this type of statistical testing can be found regardless of the scientific field. The misinterpretation and misuse of NHST results are widely practised in medicine, biology, social sciences, and many other areas. Therefore, there are many recommendations to use alternative methods of statistical analysis, or even guidance on how to NOT interpret the results of significance tests (Wasserstein and Lazar, 2016). Moreover, there are even guidelines if one still prefers to use the p-value. All these measurements are still not

able to solve the “p-value issue”. The importance of this test for biological and clinical research is not difficult to determine, as it is the gold standard for the most commonly used clinical trials (Kelter, 2020). On the one hand, the prevalence of use, and on the other hand, the prevalence of misinterpretation, prompts a more in-depth analysis of this problem. So, why do we still use this test? Firstly, most statistical software uses this test as an apriori method and secondly, just because we still learn it at universities. Thus, students are still studying NHST at the university, despite the prohibitions and restrictions of scientific and statistical societies. If for the first argument, we could still use modern software (e.g., a programming language where it is possible to perform any required analysis) and develop and promote the dissemination of machine learning algorithms in methodology with its two-stage nature: cross-validation and algorithm (Bzdok and Meyer-Lindenberg, 2018), then what about the second argument and why is it so important? In the following, we will find out the theoretical arguments for the emergence of misinterpretation. An attempt will also be made to identify whether it is related to geographical location and field of activity. We will also propose a research design that could test such theoretical arguments. The final part will discuss some alternatives to the p-value test, their limitations and if machine learning can help by neutralizing misinterpretation in research.

NHST IN BIOMEDICAL RESEARCH

The problem of not understanding the concept of p-value is dramatic because it affects directly reproducibility of scientific research and raises growing concerns about the credibility of claims of new findings based on ’statistically significant’ results (Benjamin et al., 2017, p. 5). Therefore, Szucs and Ioannidis (2017) investigated the replication success and reported poor replication rates in psychology, and added that we may expect even lower replication rates in cognitive neuroscience. Such ’findings’ can be found in studies that determine the effect of a drug being tested by comparing this effect between control and treatment groups. They are also used in finding the association of spillover patterns with the disease of interest. These are just some examples of the types of studies in which the NHST is generally accepted and possibly used right now.

According to Gao (2020), reproducibility is not the only harmful consequence. He added, that the p-value problem can also impact treatment choices in medical practice and model specification in empirical analysis (Gao, 2020, p. 1). Moreover, Ioannidis (2019) in his paper the titled "What Have We (Not) Learnt from Millions of Scientific Papers with P Values?"provides more details about publication issues and why many studies (including biomedical literature) are debatable.

So, we just defined why exactly the p-value problem is relevant for the biomedical field. The next relevant point is how actually biomedical scientists are aware of this problem and how competent they are in knowing the main concepts of NHST. Unfortunately, empirical research does not yield positive results on this issue. Therefore, Windish et al. (2007) demonstrated in their multiprogram survey that medicine residents gave a total overall of 41.4 per cent correct answers in a proposed program on statistical interpretation, and the rate for correct interpretation of p-value is 58.8 per cent (53.0-64.6) (Windish et al., 2007, p. 1014). These 58.8 per cent answered correctly (according to the design of the survey) on the question about the interpretation of $p > 0.05$. There were four possible answers (Windish et al., 2007):

a. The chances are greater than 1 in 20 that a difference would be found again if the study were repeated.
b. The probability is less than 1 in 20 that a difference this large could occur by chance alone.
c. The probability is greater than 1 in 20 that a difference this large could occur by chance alone.
d. The chance is 95 per cent that the study is correct.

At this point, it is necessary to add the following quote:

"58.8 per cent of the residents selected choice c which was designated by the authors as the correct answer. The irony is that choice c is not correct either. In fact, none of the four choices is correct. So, not only were the residents who picked choice c wrong but also the authors as well. Keep in mind, the paper was peer-reviewed and published by one of the most prestigious medical journals in the world."(Gao, 2020, p. 12)

This study was carried out 15 years ago. It is logical to assume that the situation in the biomedical field has changed due to different recommendations, guidelines and bans in some journals. A recent Swedish study very clearly refutes this. (Lytsy et al., 2022) conducted a study on understanding the concept of significant testing among the PhD students with medical and statistical and/or epidemiological backgrounds. Results: correct answers to addressing both questions, that no statistically significant result can be derived either as proof or as a measure of hypothesis probability, were given by 10.7 per cent of doctoral students and 12.5 per cent of statisticians/epidemiologists (Lytsy et al., 2022).

Thus, the problem is clearly a global one. For a more detailed analysis, it is necessary to establish how global the problem is, i.e. whether it goes beyond the borders of Sweden and beyond the borders of statisticians and medics.

GEOGRAPHICAL AND MULTIDISCIPLINARY DIMENSIONS

Unfortunately, Sweden is not alone in the widespread misunderstanding of the NHST concept at the academic level. The following empirical results support this argument. Therefore, they demonstrate that the problem of understanding and interpreting the significant test is not local and, furthermore, has little to do with the academic specialization of the respondents. Surveys in Germany, China, Spain, Italy and Chile have shown that not only students but also academic teachers with sufficient experience in university teaching commit errors in carrying out a significant test. In addition, misinterpretation extends beyond the biological and medical sciences to other fields. The following will briefly discuss the results of these studies.

German survey

Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means t-test and your result is ($t = 2.7$, d.f. $= 18$, $p = 0.01$). Please mark each of the statements below as "true" or "false". "False" means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.

1) You have absolutely disproved the null hypothesis (that is, there is no difference between the population means). [ ] true / false [ ]

2) You have found the probability of the null hypothesis being true. [ ] true / false [ ]

3) You have absolutely proved your experimental hypothesis (that there is a difference between the population means). [ ] true / false [ ]

4) You can deduce the probability of the experimental hypothesis being true. [ ] true / false [ ]

5) You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision. [ ] true / false [ ]

6) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions. [ ] true / false [ ]

Figure 1: The Questionnaire of the survey of Haller and Kraus (Haller and Kraus, 2002, p.5)

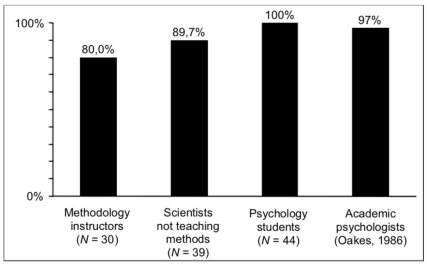

Figure 2: Percentages of participants in each group who made at least one mistake, in comparison to Oakes' original study (1986) (Haller and Kraus, 2002, p.7)

The main purpose of Haller and Kraus' study was to determine whether psychology students and teachers can interpret significance tests correctly and give corresponding pedagogical guidelines for the correct use of NHST Haller and Kraus (2002). They offered in their paper a discussion about possible reasons for misinterpretation. As result, they determined that statistical textbooks and statistical instructors could be the cause. The researchers presented a questionnaire, in which they collected questions related to the NHST and this served as a sufficient measure to check the level of knowledge in the application of the NHST by students and teachers of psychology. In addition, they divided the participants into three groups: teachers of methodology, like professors with NHST accreditation, research psychologists (who do not teach), and psychology students. The questionnaire is presented in Figure 1. As the correct answer to all the questions is 'False', they also added an explanation of why this is false for each of the questions. Furthermore, they compare the results of their survey with those of the Oakes survey, which is also very important for the discussion on the relevance of the NHST. The Haller and Kraus study shows that 16 years after the Oakes study, which shows the highest rate of misinterpretation of significance tests by academic psychologists, the problem is still relevant (see Figure 2).

Chinese survey

The second perceptive study is also one of the most recently published studies in this field (Lyu et al., 2020). Compared to the study by Haller and Kraus (2002), this study has several improvements. They examined how students interpret the significance test and the confidence interval, the main alternative to the significance test. Their study showed that most respondents, even those with academic degrees, regardless of the field of study and career stage, were not able to interpret P values and confidence intervals accurately. Another empirical advantage is the measure of respondents' confidence in their judgement, and how confident they are in their answers. Unfortunately, there is no good result here either. Most respondents are confident in their answers, which makes the problem of misinterpreting the P-value even more dangerous.
1479 respondents took part in this survey. The questions for the NHST are similar to the questionnaire by Haller and Kraus. The questionnaire for the CI is relatively new and was taken from their previous studies. The structure and questions of this questionnaire are very plausible and can provide a good measure for interpreting the CI. They also added a second version of the questionnaire: one scenario with a significant outcome, p >0.05, and a second scenario without a significant outcome, p <0.05. One additional question measuring respondent confidence was also added for each question in both parts of the questionnaire. However, they did not explore why respondents were confident in their decisions. The results showed that 89 per cent of respondents made at least one error when interpreting the P-value and 93 per cent of respondents made at least one error when interpreting the CI. This may mean that they do not pay enough attention to CI, but rather to NHST, which most of them also interpret incorrectly. This study is also crucial for assessing the 'magnitude of the tragedy', as it not only presents results from psychologists (and social scientists) but also from respondents from other fields (see Figure 3 and Figure 4). At this point, we can conclude that the misinterpretation of NHST and CIs is a multidisciplinary problem. Moreover, even mathematicians and statisticians have not shown acceptable results for either the non-significant scenario or the insignificant scenario. This fact should motivate the whole scientific community to continue the 'p-war', especially at the academic level.

Spanish survey

The third survey Badenes-Ribera et al. (2015) presents the results of a survey of Spanish academic psychologists and methodology teachers (n = 418). Interestingly, the respondents' average length of service as university professors is 14.16 years. Thus, their conclusion is based on the competence of psychology teachers rather than on the knowledge level of the students. This study tested whether Spanish academic psychologists have a proper understanding of the concept of p-value. The researchers divided their questionnaire into four parts. Each part includes questions concerning a particular type of delusion: inverse probability, replication, effect size and clinical significance fallacies. Moreover, they concluded that the first type, reverse probability fallacy, was the most common among respondents. The fact that academic psychologists have such a high percentage of errors in the questionnaire showed that the high percentage of misinterpretations may lie not only

	Statement (significant scenario)	Science N = 133 (9%)	Eng/Agr. N = 72 (5%)	Medicine N = 69 (5%)	Economics N = 93 (6%)	Management N = 51 (3%)	Psychology N = 125 (8%)	Social Science N = 111 (8%)	Math/Statistics N = 105 (7%)	Average N = 759 (51%)
p value (significant)	(a) You have absolutely disproved the null hypothesis.	53%	53%	49%	60%	63%	50%	59%	44%	53%
	(b) You have found the probability of the null hypothesis being true.	58%	62%	52%	44%	55%	59%	45%	32%	51%
	(c) You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.	53%	62%	51%	67%	71%	77%	67%	70%	65%
	(d) You have a reliable experimental finding in the sense that if, hypothetically, the experiment was repeated a great number of times, you would obtain a significant result on 99% of occasions.	62%	54%	64%	63%	53%	42%	59%	48%	55%
	Total (endorsed at least one statement)	93%	90%	90%	92%	94%	95%	95%	88%	92%
CI (significant)	(a) There is a 95% probability that the true mean lies between .1 and .4.	56%	53%	52%	60%	63%	66%	67%	33%	56%
	(b) If we were to repeat the experiment over and over, then 95% of the time the true mean falls between .1 to .4.	59%	56%	54%	54%	51%	54%	59%	48%	55%
	(c) If the null hypothesis is that there is no difference between the mean of experimental group and control group, the experiment has disproved the null hypothesis.	57%	53%	49%	53%	59%	31%	48%	40%	48%
	(d) The null hypothesis is that there is no difference between the mean of experimental group and control group. If you decide to reject the null hypothesis, the probability that you are making the wrong decision is 5%.	62%	53%	48%	66%	63%	70%	56%	58%	60%
	Total (endorsed at least one statement)	97%	93%	93%	96%	98%	94%	94%	88%	94%

Figure 3: Percentage of misinterpretation of p values and CIs for each statement (significant scenario) (Lyu et al., 2020, p.5)

	Statement (nonsignificant scenario)	Science N = 114 (8%)	Eng/Agr. N = 79 (5%)	Medicine N = 61 (4%)	Economics N = 71 (5%)	Management N = 44 (3%)	Psychology N = 147 (10%)	Social Science N = 106 (7%)	Math/Statistics N = 98 (7%)	Average N = 720 (49%)
p value (non-significant)	(a) You have absolutely proved the null hypothesis.	63%	57%	48%	48%	55%	54%	53%	43%	53%
	(b) You have found the probability of the alternative hypothesis being true.	57%	43%	54%	42%	48%	40%	49%	34%	45%
	(c) You know, if you decide not to reject the null hypothesis, the probability that you are making the wrong decision.	54%	56%	64%	65%	70%	63%	59%	55%	60%
	(d) You have an unreliable experimental finding in the sense that if, hypothetically, the experiment was repeated a great number of times, you would obtain a significant result on 21% of occasions.	61%	48%	43%	42%	43%	29%	45%	32%	42%
	Total (endorsed at least one statement)	87%	9%	82%	90%	93%	84%	87%	78%	86%
CI (non-significant)	(a) There is a 95% probability that the true mean lies between -.1 and .4.	62%	54%	62%	61%	55%	69%	63%	33%	58%
	(b) If we were to repeat the experiment over and over, then 95% of the time the true mean falls between -.1 to .4.	53%	49%	52%	56%	61%	48%	60%	53%	53%
	(c) If the null hypothesis is that there is no difference between the mean of experimental group and control group, the experiment has proved the null hypothesis.	54%	44%	61%	46%	43%	46%	50%	37%	48%
	(d) The null hypothesis is that there is no difference between the mean of experimental group and control group. If you decide not to reject the null hypothesis, the probability that you are making the wrong decision is 5%.	52%	58%	51%	51%	68%	53%	63%	45%	54%
	Total (endorsed at least one statement)	95%	92%	92%	89%	98%	89%	93%	85%	91%

Figure 4: Percentage of misinterpretation of p values and CIs for each statement (non-significant scenario) (Lyu et al., 2020, p.6)

in the educational program but also arise from the lack of competence of academics in the application of statistical methods.

Chilean-Italian study survey

Another study Badenes-Ribera et al. (2016), which is a replication of the previous one, has not shown any satisfactory results as well. Only two parameters were changed: the geography of the survey (Italy and Chile instead of Spain) and the response scale (correct answer scale instead of true/false). The questionnaire also includes questions on inverse probability fallacy, replication fallacy, effect size fallacy, and clinical significance fallacy. Respondents in this survey reported more correct answers compared to the previous study (Badenes-Ribera et al., 2015), what they attempted to explain by differences between countries. However, they did not provide any strong arguments to prove this and in any case, the level of misinterpretation is still high, especially for academics. This way, both studies force us to personalise the educational strategies and competencies of academics (in the field of statistical methods) in their conclusions.

The previously mentioned studies are essential for the discussion about the failure of significant tests in biomedical sciences. Moreover, the results of the second study showed that this problem should be addressed by medical scientists as well as other scientists (natural science, mathematics/statistics, management, sociologists, etc.). The third and fourth studies concluded that this problem has little or no correlation with the location of the survey. All of the countries described above did not have satisfactory survey results: USA, Germany, China, Spain, Chile, and Italy. Thus, the misinterpretation of the p-value is not local but global. However, none of these studies provides a

theoretical explanation for this phenomenon. That is why a new empirical study is warranted. These studies (and others in the field) explain the roots of the misunderstanding of NHST from a statistical or methodological perspective, but not from a cognitive perspective. Analyzing the studies cited above, it becomes clear at what stage of interpretation an error occurs methodologically and even how to categorize such misinterpretation, but there is still no understanding of why this happens so systematically. Understandably, the type of misinterpretation is important for further analysis and work on preventive measures, but it is still not enough to reduce the number of misinterpretations and misuses of the p-value. In the next section, we will propose some theoretical arguments that may explain this phenomenon and encourage new research in this area by applying a cognitive perspective to the problem.

Source theory as theoretical argumentation

The preceding empirical outcomes motivate the idea that there is something behind methodological errors. Even if we have found the roots of the errors, learned how to classify them, developed guidelines for avoiding these errors, and achieved bans at a high scientific level, because of the 'p-war' in the scientific community, NHST is still used in research and in university courses. Moreover, even if we 'win' this war, will the problem be solved, or will other misunderstandings of statistical methods arise? To try to explore this, we need to go back to the starting point by applying a sociological approach. It is logical to assume that if we are going to use statistical methods, we first need to learn them. The process of learning these methods in biology, medicine, and other scientific fields begins at universities if one considers science. Consequently, empirical evidence shows that often professors cannot provide evidence of 100 per cent understanding of concepts related to p-value resulting in some of them teaching their students about NHST without conveying a fully correct interpretation of these concepts. Potentially, we can find the results of research conducted by scientists without a complete understanding of NHST interpretation in medical trials, therapeutic surveys, and other high-impactful studies. Now, is there anything special about this transit between professor and student?

Such an explanation should contain a theoretical mechanism that describes the influence of the teacher (the object that provides the primary information about the ST) on the student (the object that perceives this information). In this relationship, we must determine the origin of the error. So, why don't students check the validity of the learning material? It is logical to assume that teachers have a high degree of trust and authority on the part of students and that the information and knowledge they provide are perceived as reliable. We assume that this problem arises when transferring knowledge from teacher to student. This means that the mistake and the root of misinterpretation arise at this very moment. Thus, students receive information about significant tests without having to verify it and use this information in their research or teaching other students in their careers. This is a potential scenario that can lead to such global misinterpretation in the application of statistical methods.

Thus, the focus of this section is on source theory, a theory that derives expectations based on characteristics and competence. Source theory belongs partly to the Expectation-States Theory, which is not a concrete theory or a paradigm, but rather a research program that units many other theories to explain the relationship between performance expectations and social influence (Kalkhoff and Thye, 2006).

The student trusts their teacher and has no doubts about their competence. The main question of this interaction process is: why do students trust their teachers in the first place? To answer this question, we use a source theory approach, where we point to the teacher as a source of evaluation for the student. Savage and Webster (1972) explained source theory as a combination of two theoretical concepts. The first concept of the 'Mirror Self' from Cooley and Mead postulates:

"Evaluations from a significant other, to use Sullivan's term, will predictably be accepted by the individual, whereas the opinions of others (with uncertain characteristics) are likely to be ignored" (Savage and Webster, 1972, p. 317)

This statement should be accompanied by an explanation of the concept of the 'significant other':

"Cooley located such significant others primarily in families and peers but there are individuals who have the right to evaluate the performance of others in many kinds of more formal settings as well; employers have the right to evaluate employees, teachers the right to evaluate students." (Berger et al., 1983, p. 24)

According to Savage and Webster (1972), the second theoretical concept is a crucial claim of expectation states theory, which argues that 'many regularly reported observable behaviours among the members of problem-solving groups, such as an unequal number of chances to perform, evaluations of performances, the likelihood of performing, and rejection of influence, may be explained if one postulates the existence of expectation states or cognitive beliefs about the ability of each member of the group' (Savage and Webster, 1972, p. 318). In sum, they make a statement of the Source theory such as: 'an individual whom a high ability evaluator evaluates will often believe him and form an expectation state based on those evaluations, while an individual evaluated by a low ability evaluator will usually ignore him; and an individual who holds high self-expectations will be more likely to reject influence than an individual who holds low self-expectations (Savage and Webster, 1972, p. 318). In addition to the claim that our phenomenon is perfectly explained by resource theory, it should be added that Berger et al. (1983) also note the importance of having rules in this interaction:

"The more rules there are, the less likely incongruence is to arise in the first place." (Berger et al., 1983, p. 36)

Thus, it is more likely to notice incongruence between the expectation state and authority in informal interaction than informal. For example, the interaction between teacher and student is full of formality, e.g., university regulations, official appeals, subordination, etc.

Since we have discussed all the essential components of the Source theory, we can then describe the teacher-student interaction in terms of this theory. In a possible situation, the student has high expectations of their teacher, based on the teacher's status, and so the teacher acts as a 'significant other' to the student. Therefore, as a significant other, the teacher can legitimately judge the student's performance and the student accepts this. Having rules in this interaction makes it less likely that the teacher will lose their authority from the student's perspective. In addition, the teacher influences the student and the student accepts this influence and applies it to the learning process. Thus, if a teacher transmits incorrect knowledge about NHST to a student, the student is more likely to believe them because of their high expectations and less likely to test the validity of the learning material. However, this possible theoretical scenario excludes cases in which a teacher provides the correct concept of NHST to their students. It also excludes cases where students are less motivated to learn and make no significant effort to understand the learning material. In this theoretical scenario, we focus only on cases where students are motivated to understand the concept of NHST and teachers provide incorrect knowledge.

Moreover, this explanation is not intended to discredit professors. On the contrary, this theoretical argumentation makes it clearer how powerful real pedagogical influence is and what critical consequences it can lead to. In addition, it can help to understand the fundamental role of critical thinking skills in an academic environment.

PROPOSAL OF RESEARCH DESIGN AND SURVEY

The theories described above have a sufficient number of experimental tests (Thye and Kalkhoff, 2009) and therefore we will not operationalize the theory in our study and will use it as a given. To derive the hypothesis, we need to formulate two premises. First, in line with empirical research, we recognise that teachers can teach students an inaccurate conception of NHST. [assumption 1]. Secondly, given this, we also believe that if teachers provide misinformation, some students may verify it from other sources due to low trust, in which case we expect better results from these students. Thus, students' critical perception is the second precondition for the hypothesis [assumption 2]. On this basis, we expect worse test results from students who trust their teachers more (according to the theory). Based on the theoretical discussion and the two premises, the following hypothesis will be derived:

Given that the level of competence of an instructor in NHST is low, the more a student will trust this instructor in statistics, the more this student will commit errors in NHST.

To determine the association between the effect of teacher status and the level of conceptual understanding of NHST, we derive two variables: a dependent variable and an independent variable. The dependent variable refers to the number of correct answers given by the student, and the independent variable is the student's confidence in the competence of their teacher in statistical knowledge. To investigate the level of underestimation of NHST, we propose six questions contained in our questionnaire (see appendix). The sum of the correct answers (min = 0, max = 6) is the measure of the dependent variable. To measure the independent variable, trust, we first need to define the meaning of trust in the context of our study. 'Trust' means that the student has full confidence in the competence of their teacher and does not check the verification of the learning material. It will also ask whether the teacher (if they are the source of knowledge about the NHST for the student) influences the student's level of knowledge. After each test question, this question will be asked to track which questions students make the most mistakes in. Also, before the test questions, we will ask them how they assess the competence of their statistics instructor who teached them the basics of NHST. This 'trust' will serve as the independent variable. Logically, we will accept the hypothesis if we find a sufficient number of students who score well on the significant test questionnaire and most of their questions were not influenced by their teacher (Figure 5, quadrant II). We will also accept it if we get a sufficient number of students who have poor scores on the NHST test, but most of their questions were influenced by their teacher (Figure 5, quadrant IV). The case when we reject the hypothesis is when the survey shows a large number of students who did not commit errors in the NHST test and most of their questions were influenced by the teacher, and the reverse (Figure 5, quadrants I and III).

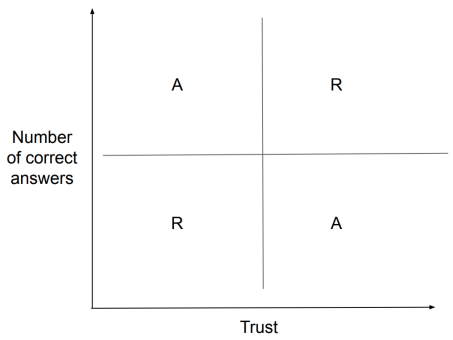

Figure 5: Accept or reject the hypothesis

To test the hypothesis and get a generalizable knowledge about the causes of a misconception of NHST in the educational system, we suggest the following study design. Since the previous studies provide a questionnaire to measure the level of knowledge in NHST, we will continue to apply a comparable design (see appendix). The questionnaire for NHST is taken from the study Haller and Kraus (2002). The confidence intervals questionnaire was compiled by us (see appendix, sections 4-5). Before each of both test-questionnaire starts, we will ask students how they rate the competence of their lecturer in statistics who teaches them the basics of significance testing/confidence intervals,

using the Likert scale. Very competent means that they fully believe in the plausibility of the material taught, incompetent means, that they do not believe in the plausibility of the material taught and check it with other sources. In addition, for each question of the test questionnaire they have to answer whether their answer is influenced by the knowledge they have acquired with the help of the teacher. We also accept the advice from Badenes-Ribera et al. (2016) to use the three-response-format «True/False/Don't know»:

"By not asking to explicitly classify statements as either true or false, it is not possible to differentiate omissions from items identified as false. A three-response format (True/False/Don't know) would have been far more informative since this would have also allowed identifying omissions as such." (Badenes-Ribera et al., 2016, p. 8)

and we have added it to our questionnaire in form of 'True/False/Neither/nor'. Our survey is designed for students, professors, and tutors to check the level of knowledge of each of these categories. However, in order to test the hypothesis, its acceptance or rejection, we will take the results only of those who indicated their student status in the first part of the questionnaire and chose lecture as a source of information for ST. Thus, for a valid analysis, the group of students must be matched to their statistics teacher. This means that we should first measure the level of NHST knowledge of the students (who choose lecture as a source of NHST knowledge), their assessment of the teacher's competence, and then compare it with the level of NHST knowledge of this teacher.

The results of this analysis can potentially be compared with the results of previous studies to produce a complete conclusion.

Discussion: Practical alternatives to NHST

According to Ioannidis (2019), there is a great call for resolute action not only on the part of statisticians but also on the part of the entire scientific community. The theoretical arguments presented above are not intended to change the relationship between teachers and students, but rather to demonstrate how critically students perceive information from their professors and to draw attention to the widespread misinterpretation of NHST from a cognitive and pedagogical perspective. Now, methodologically speaking, the main mistake many researchers make is to use the NHST as a method to draw definitive conclusions. In fact, this test serves as a filter tool. Moreover, a non-significant result does not mean that the study is meaningless (MacGillivray, 2019). If a drug study shows a p >0.05, it may still have medical relevance. In other words, this study may contribute to the development of many studies in other areas of medicine. Thus, the P-value is not a measure of success or failure:

"P values are neither objective nor credible measures of evidence in statistical significance testing. Moreover, the authenticity of many published studies with p <0.05 findings must be called into question." (Hubbard and Lindsay, 2008, p. 81)

If the results are not properly analyzed, the research provides incorrect or incomplete information, and this starts a chain of incorrect data that grows like a snowball. This is why proper interpretation is vital for every field of research. In our discussion, we rely on the following statement about the correct interpretation of the NHST for further analysis:

"The p-value is not the probability of the null hypothesis; rejecting the null hypothesis does not prove that the alternative hypothesis is true; not rejecting the null hypothesis does not prove that the alternative hypothesis is false; and the p-value does not give any indication of the effect size. Furthermore, the p-value does not indicate the replicability of results. Therefore, NHST only tells us about the probability of obtaining data which are equally or more discrepant than those obtained in the event that H0 is true (...)" (Badenes-Ribera et al., 2016, p. 7)

Nevertheless, it is also possible to choose the right strategy for teaching statistical methods. For example, Haller and Kraus (2002) proposed in their paper four steps to avoid misunderstanding by students. The main idea is to present NHST with terms for conditional probabilities, such as Bayes' rule (Haller and Kraus, 2002, p. 10):

"Teach students the underlying idea of the Bayesian inference approach: Considering p(H | D) To find out the probability of a hypothesis (H) given data (D), we can apply Bayes' rule: Considering p(H | D):

$$p(H|D) = \big((p(D|H)p(H))/(p(D|H)p(H)+p(D|H)p(H)\big)"$$

Moreover, (Kelter, 2020) proposed in his paper to use of special software (JASP) as an alternative to NHST, which is based on Bayesian inference and makes it more clear for scientists to rehearse.

On the one hand, this strategy could definitely increase understanding of the concept of NHST and positively influence the quality of the application. On the other hand, a deductive approach is needed not only for NHST but also for teaching other statistical methods.

Thus, a second important risk factor is the lack of critical thinking among students (Santos, 2017). The idea that critical thinking should be an important aspect of science education is widely recognised (Tanti et al., 2020) For example, the National Science Education Standards sets as one of its goals the promotion of science as a research activity (National Academy of Sciences, 1996). This goal includes numerous items that focus on critical thinking, such as 'identifying assumptions, applying critical and logical thinking, and considering alternative explanations; analysing events and phenomena firsthand and critically examining secondary sources; testing the reliability of the knowledge they generate; and critical skills in analysing arguments by reviewing current scientific understanding, weighing evidence, and examining logic to decide which explanations and models are best' (Bailin, 2002, p. 361)

As already mentioned, there are also other alternatives to the significance test: confidence intervals, effect sizes and al-

gorithms of machine learning. We will briefly discuss their advantages below.

Compared to the p-value, the CI provides a reasonable estimate of the size of the effect in the population, indicates the precision or reliability of the estimate by the width of the interval, and the CI focuses on whether the results are meaningful to the population rather than aiming for a statistically significant value (Hubbard and Lindsay, 2008, p. 81). Furthermore, p-values do not provide information about the size of the effect. A statistical test will almost always show a significant difference in a larger sample unless there is no effect at all, i.e. the effect size is zero (Sullivan and Feinn, 2012, pp. 279–280). Sullivan and Feinn (2012) concluded that an estimation of the effect size is required before the start of the study to calculate the number of subjects needed to avoid a type II error (Sullivan and Feinn, 2012, p. 279). Thus, large but insignificant effects may lead to further searches with a higher power, while trivially small effects that are significant due to large sample sizes may warn researchers to potentially overestimate the observed effect (Fritz et al., 2011, p. 2). For a detailed explanation of alternative techniques for different types of studies, such as vitro and animal studies, genetic studies, equivalence and noninferiority trials and also descriptive and diagnostic studies, see Schmidt et al. (2018)

However, despite numerous recommendations and advice on the use of alternatives, this does not actually seem to be useful or resultant. According to Ioannidis (2019), empirical data suggest that across the biomedical literature (1990–2015), when abstracts use P values 96 per cent of them have P values of 0.05 or less. The same percentage (96 per cent) applies to full-text articles. Among 100 articles in PubMed, 55 report P values, while only 4 present confidence intervals for all the reported effect sizes, none use Bayesian methods and none use a false-discovery rate. (Ioannidis, 2019, p. 20)

Moreover, there are great studies and literature, which recommend decent interpretations of the p-value: J.Benjamin and Berger (2019); Amrheinm et al. (2017); Colquhoun (2017); Benjamin et al. (2017); Gliner et al. (2001); Zhang and Wu (2022); Greenland et al. (2016), if one argues that alternative methods or alternative teaching strategies are quite complex.

The final argument is machine learning and its prominent role in the debate about alternatives to the significance test. Bzdok and Meyer-Lindenberg (2018) describe in their paper how machine learning procedures and approaches can eliminate the inductive problems associated with NHST. Such approaches allow multiple outcomes at once, which is very important when investigating the effects of drugs or disease-induced patterns; they also offer solutions on how to treat multi-class predictions instead of one isolated model; as already mentioned, machine learning algorithms include a two-step procedure, which means that the original data set is split into subsets for testing and learning to see the accuracy of the chosen model, this split can be applied to further research (Bzdok and Meyer-Lindenberg, 2018) or generally help build more accurate models by tuning and comparing different models simultaneously. Moreover, machine learning is particularly advanced in predicting modelling. Predictive modelling is very relevant to biomedical research, as it can help health professionals by making predictions for patients at the individual risk of developing a disease (or disorders) and further assisting them with diagnostic tasks (Steyerberg et al., 2018). In addition, they can help healthcare professionals write prescriptions, better assess patients' conditions, and improve their lifestyles.

Indeed, machine learning has many chances to correct inductive problems by understanding and applying statistical methods in scientific research. The biggest drawback here is the (seeming) complexity of algorithms and programming languages for many scientists. However, establishing machine learning algorithms could lead to an increase in the overall quality of research due to their more accurate interpretation of results

Conclusion

The problem of the scandalous p-value does not rest on the rock of faith: it is not likely or less likely. It has become a 'physical law' in the scientific community. Consequently, it also strongly influences the development and quality of research in the biological and medical fields. Furthermore, we can observe that despite the guidelines, recommendations, bans, and other measures against p-value as the only 'source of truth', the trend towards a decline in its use is not expected. In this regard, we accepted the challenge to investigate the roots of this problem not only from a statistical or methodological perspective but also to apply an approach from other sciences, in our case, a sociological approach. We have analytically derived some arguments that point to the lack of a unified pedagogical strategy for teaching NHST to students of various disciplines. We also proposed a potential research design that could test the validity of the sources logically derived from the theory. Of course, we would not see $p < 0.05$ as an indicator of the success or failure of such a study. If such arguments hold true, this may encourage the scientific community to take action to regulate the limits of p-value strength not only at the publication level, but also at the academic level. We also believe that combining scientific approaches from different disciplines (biomedical and sociological in our case) can bring much more benefit and power to regulating problems in the scientific community, be they of p-value or otherwise.

Acknowledgements

I would like to express my gratitude to Dr. Stephan Poppe for fruitful discussions and advices.

References

[1] Amrheinm, V., Korner-Nievergelt, F., and Roth, T. (2017). "The earth is flat (p>0.05): significance thresholds and the crisis of unreplicable research". *PeerJ*, 5:e3544.

[2] Badenes-Ribera, L., Frias-Navarro, D., i Bort, H. M., and Pascual-Soler, M. (2015). "Interpretation of the p value: A national survey study in academic psychologists from spain". *Psicothema*, 27:290–295.

[3] Badenes-Ribera, L., Frias-Navarro, D., Iotti, D., BonillaCampos, A., and Longobardi, C. (2016). "Misconceptions of the p-value among chilean and italian academic psychologists". *Frontiers in Psychology*, 7:1247.

[4] Bailin, S. (2002). "Critical thinking and science education". *Science & Education*, 11:361–375.

[5] Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., Boeck, P. D., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A., Morgan, S. L., Munafo, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Zandt, T. V., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J., and Johnson, V. E. (2017). "Redefine statistical significance". *Nature Human Behaviour*, 2(1):6–10.

[6] Berger, J., Wagner, D. G., and Zelditch, M. J. (1983). "Expectation states theory: The status of a research program". Technical report Nº 90, *Stanford University* .

[7] Bzdok, D. and Meyer-Lindenberg, A. (2018). "Machine learning for precision psychiatry: Opportunities and challenges". *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(3):223–230.

[8] Colquhoun, D. (2017). "The reproducibility of research and the misinterpretation of p-values". *Royal society open science*, 4(12):171085.

[9] Fritz, C., Morris, P., and Richler, J. (2011). "Effect size estimates: Current use, calculations, and interpretation". *ournal of experimental psychology. General*, 141:2–18.

[10] Gao, J. (2020). "P-values–a chronic conundrum". *BMC Medical Research Methodology*, 20:1–8.

[11] Gliner, J. A., Morgan, G. A., Leech, N. L., and Harmon, R. J. (2001). "Problems with null hypothesis significance testing". *Journal of the American Academy of Child and Adolescent Psychiatry*, 40(2):250–252.

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

[13] Haller, H. and Kraus, S. (2002). "Misinterpretations of significance: A problem students share with their teachers?" *Methods of Psychological Research*, 7(1):1–20.

[14] Hubbard, R. and Lindsay, R. (2008). "Why p values are not a useful measure of evidence in statistical significance testing". *Theory & Psychology*, 18:69–88.

[15] Ioannidis, J. P. (2019). "What have we (not) learnt from millions of scientific papers with p values?" *The American Statistician*, 73(sup1):20–25.

[16] J.Benjamin, D. and Berger, J. O. (2019). "Three recommendations for improving the use of p-values". *The American Statistician*, 73(sup1):186–191.

[17] Kalkhoff, W. and Thye, S. R. (2006). "Expectation states theory and research: New observations from meta-analysis". *Sociological Methods & Research*, 35(2):219–249.

[18] Kelter, R. (2020). "Bayesian alternatives to null hypothesis significance testing in biomedical research: a non-technical introduction to bayesian inference with jasp". *BMC Medical Research Methodology*, 20(1):1–12.

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

[20] Lyu, X. K., Xu, Y., Zhao, X. F., Zuo, X. N., and Hu, C. P. (2020). "Beyond psychology: Prevalence of p value and confidence interval misinterpretation across different fields". *Journal of Pacific Rim Psychology*, 14(e6):1–8.

[21] MacGillivray, B. H. (2019). "Null hypothesis testing scientific inference: A critique of the shaky premise at the heart of the science and values debate, and a defense of value-neutral risk assessment". *Risk Analysis*, 39(7):1520–1532.

[22] National Academy of Sciences (1996). *National Science Education Standards*. National Academy Press, Washington, DC.

[23] Santos, L. F. (2017). "The role of critical thinking in science education". *Online Submission*, 8(20):60–173.

[24] Savage, I. R. and Webster, M. J. (1972). "Source of evaluations reformulated and analyzed". *The Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 4, edited by L. M. LeCam, J. Neyman, and E. L. Scott. Berkeley(3-4):317–327.

[25] Schmidt, S. A., Lo, S., and Hollestein, L. M. (2018). "Research techniques made simple: sample size estimation and power calculation". *Journal of Investigative Dermatology*, 138(8):1678–1682.

[26] Steyerberg, E. W., Uno, H., Ioannidis, J. P., Calster, B. V., C. Ukaegbu, T. D., Syngal, S., and Kastrinos, F. (2018). "Poor performance of clinical prediction models: the harm of commonly applied methods". *Journal of clinical epidemiology*, 98:133–143.

[27] Sullivan, G. M. and Feinn, R. (2012). "Using effect size—or why the p value is not enough". *Journal of Graduate Medical Education*, 4(3):279–282.

[28] Szucs, D. and Ioannidis, J. (2017). "Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology". *PLoS biology*, 15(3):e2000797.

[29] Tanti, T., Kurniawan, D. A., Kuswanto, K., Utami, W., and Wardhana, I. (2020). "Science process skills and critical thinking in science: Urban and rural disparity". *Jurnal Pendidikan IPA Indonesia*, 9(4):489–498.

[30] Thye, S. R. and Kalkhoff, W. (2009). "Seeing the forest through the trees: An updated meta-analysis of expectation states research". *Current Research in Social Psychology*, 15(1).

[31] Wasserstein, R. L. and Lazar, N. (2016). "The asa statement on p-values: context, process, and purpose". *The American Statistician*, 70(2):129–133.

[32] Windish, D. M., Huot, S. J., and Green, M. L. (2007). "Medicine residents' understanding of the biostatistics and results in the medical literature". *Jama*, 289(9):1010–1022.

[33] Zhang, H. and Wu, Z. (2022). "The generalized fisher's combination and accurate p-value calculation under dependence". *Biometrics*.

Proposal of the questionnaire

1.1. Select your status:

- Student
- Academic teacher
- Tutor
- None above

1.2. Select your discipline:

- Medicine
- Biology
- Psychology
- Social science
- Other ()

1.3. Select your university:

- University ()
- Other ()

1.4. Have you participated in statistical courses (lectures, seminars or similar)?

- Yes
- No

1.5. How long ago was the last course?*
*(triggered by Q1.4: yes)

- less than 1 year
- between 1 and 2 years
- more than 3 years

1.6. Have you acquired statistical knowledge yourself (outside of courses)?*
*(triggered by Q1.1: Academic teacher)

- Yes
- No

2.1. How would you rate your knowledge of significance testing?

✹ very good ✹ good ✹ partially ✹ not good ✹ no knowledge

2.2. Where did you acquire your knowledge of the basics of significance testing?

- Classes (lectures)
- Internet sources (internet courses)
- Other (please specify here)

2.3. How would you rate your knowledge of significance testing?*

(Very competent means that you fully believe in the plausibility of the contents taught,
no competence means that you do not believe in the plausibility of the contents taught and that you check them on your own)
*triggered by Q2.2: Classes(lectures))

✹ very competent ✹ competent ✹ poor competent ✹ no competence at all

2.4. Do you know about any alternatives to the significance test for testing hypotheses?

- Yes
- No

2.5. What would you use to show an effect in your data? (Multiple answers possible)

- Null hypothesis significance testing (NHST)
- Confidence Interval
- Other alternatives (effect sizes, power test)
- Other()

Suppose you apply a simple t-test for independent samples to examine a mean difference between an experimental and a control group. The difference between the groups is significant at the 1% level (more precisely: t = 2.7, df = 18 degrees of freedom, p = 0.01). Please mark each of the following statements as "true", "false" or "neither/nor". "Neither/nor" means that the statement does not follow strictly logically from the above premises.

3.1. It is clearly proven that the null hypothesis (that there is no difference between the population targets) is false.

3.2. The probability of the null hypothesis being true has been found.

3.3. It has been proven that your alternative hypothesis (that there is a difference between the population targets) is true.

3.4. The probability that the alternative hypothesis is correct can now be derived.

3.5. If one now decides to reject the null hypothesis, then one now knows the probability that this decision could be wrong.

3.6. The experimental finding is reliable in the sense that you would get a significant result in 99% of the cases if you repeated the experiment very often.

- True
- False
- Neither/nor

(asked after each question):
Does the knowledge you have acquired with the help of your teacher(s) influence your answer?

- Yes
- No
- I had no teacher

4.1. How would you rate your knowledge of confidence interval?

✹ very good ✹ good ✹ partially ✹ not good ✹ no knowledge

4.2. Where did you acquire your knowledge of the basics of confidence interval?

- Classes (lectures)
- Internet sources (internet courses)
- Other (please specify here)

4.3. How would you rate your knowledge of confidence interval?*

(Very competent means that you fully believe in the plausibility of the contents taught,
no competence means that you do not believe in the plausibility of the contents taught and that you check them on your own.)
*(triggered by Q4.2: Classes(lectures))

✹ very competent ✹ competent ✹ poor competent ✹ no competence at all

Suppose you are working at a research institute and you are dealing with the issue of gender pay gap. You sample 100 people to estimate the pay gap.
They observe an average difference in hourly earnings of 4.5€, with a symmetric 95% confidence interval of [3.8, 9.4].

Appendix

5.1. It is clearly proven that the null hypothesis (that there is no difference between the population targets) is false.

Appendix 1 (see bellow)

- True
- False

5.2. The probability of the null hypothesis being true has been found..

- wider
- narrower

5.3. It has been proven that your alternative hypothesis (that there is a difference between the population targets) is true.

- wider
- narrower

5.4. The probability that the alternative hypothesis is correct can now be derived.

- True
- False

5.5. If one now decides to reject the null hypothesis, then one now knows the probability that this decision could be wrong.

- True
- False

(asked after each question):
Does the knowledge you have acquired with the help of your teacher(s) influence your answer?

- Yes
- No
- I had no teacher