A Primer on p-Value Thresholds and α-Levels – Two Different Kettles of Fish

Norbert Hirschauer and Sven Grüner
Martin Luther University Halle-Wittenberg

Oliver Mußhoff
Georg-August-University of Goettingen

Claudia Becker
Martin Luther University Halle-Wittenberg

Abstract

It has often been noted that the "null-hypothesis-significance-testing" (NHST) framework is an inconsistent hybrid of Neyman-Pearson's "hypothesis testing" and Fisher's "significance testing" that almost inevitably causes misinterpretations. To facilitate a realistic assessment of the potential and the limits of statistical inference, we briefly recall widespread inferential errors and outline the two original approaches of these famous statisticians. Based on the understanding of their irreconcilable perspectives, we propose "going back to the roots" and using the initial evidence in the data in terms of the size and the uncertainty of the estimate for the purpose of statistical inference. Finally, we make six propositions that hopefully contribute to improving the quality of inferences in future research.

Keywords

statistical decision theory; Neyman; Pearson; inductive inference; Fisher; α-level; p-value threshold; null-hypothesis-significance-testing; random error

HURLBERT and LOMBARDI (2009: 318):
"[…] we all would like a quick, objective and automatic way to evaluate our results, but there is none that also meets the additional requirement of logical and 'useful'. We must simply apply the same sort of nuanced thinking and nuanced language we use in other contexts involving gradations in the strength of evidence."

1 Introduction

Following a longstanding debate concerned with inferential errors, the *American Statistical Association* (ASA) issued an unprecedented methodological warning in 2016 that stressed that the *p*-value can neither be used to determine whether a hypothesis is true nor whether a finding is important (WASSERSTEIN and LAZAR, 2016). Three years later, *The American Statistician* published a special issue "Statistical Inference in the 21st Century: A World Beyond p < 0.05." Summing up the reform suggestions, the issue's editors state that it is time to abandon statistical significance testing (WASSERSTEIN et al., 2019). Almost simultaneously, a widely supported *p*-value petition to "Retire statistical significance" was published in *Nature* (AMRHEIN et al., 2019). In the same year, the NATIONAL ACADEMIES OF SCIENCES (2019) took up the criticisms in its "Consensus Report on Reproducibility and Replicability in Science".

Inferential errors in the form of overconfident conclusions following statistical significance declarations are a major problem in the social sciences. This includes agricultural economists who, as other empirically working scientists, believe that data can be used to gain information regarding scientific propositions (hypotheses) about the real world. We often engage ourselves in the scientific exercise of generalization and want to draw conclusions from the particular (e.g. a sample) to the general (a real-world state of interest). We denote this as "inductive reasoning" throughout this paper.[1] For example, we might analyze data

[1] There is often prior knowledge that enables researchers to formulate hypotheses. A prior hypothesis may either be a vague directional hypothesis or a more concrete proposition regarding the magnitude of an effect. Starting off with a proposition regarding a real-world state or relationship of interest is often considered to be the "deductive" because we conclude something from "theory." Being aimed at the accumulation of knowledge and the refinement (and, if necessary, the adaptation) of theory, the process of research consists of a sequence of deductive and inductive steps. After the formulation of a hypothesis that reflects the state of knowledge at a given point in time, we probe its prediction with our data from which we obtain an effect size estimate associated with a certain uncertainty. Estimation is inherently inductive because we ask the question of what we can learn from the particular (the sample) regarding the general (the real-world state or relationship of interest). At the same

DOI: 10.30430/70.2021.2.123-133

because we want to find out whether or by how much different levels of exposure to the herbicide Glyphosate increase the probability of cancer, or how various plant protection agents affect earthworm populations. In the attempt of arriving at a rational inductive belief regarding a real-world state or relationship of interest, we should be aware of the fundamental limitation of inductive reasoning that we cannot obtain certain knowledge from a particular set of data such as a random sample. Hence, we must embrace any remaining uncertainty and be cautious in our judgment of what we should most reasonably believe. Nonetheless, we are regularly interested in obtaining epistemic probabilities for scientific propositions. That is, we want to talk, for example, about the probability that the null hypothesis or some other hypothesis is true. Epistemic probabilities reflect the degree of rational belief that we can and should have given all prior knowledge (e.g. from previous studies) *and* the incremental information that was extracted from a particular study's data.

Unfortunately, the prevalent "null hypothesis significance testing" (NHST) framework of statistical inference does not help answer the question of interest. This is largely because NHST is an inconsistent amalgamation between the "*hypothesis* testing" approach of NEYMAN and PEARSON (1933a, 1933b), on the one hand, and the "*significance* testing" approach of FISHER (1925), on the other (cf. LEHMANN, 1993; ZILIAK and MCCLOSKEY, 2008; HURLBERT and LOMBARDI, 2009; KENNEDY-SHAFFER, 2019). In a seminal paper titled "Mindless Statistics," GIGERENZER (2004) investigates the two original approaches and describes the cognitive biases that result from statistical practices based on their inconsistent hybrid NHST. In line with ZILIAK and MCCLOSKEY (2008), he calls them "statistical rituals" and claims that they have largely eliminated critical thinking in the social sciences. Gigerenzer deplores collective delusions and widely internalized flawed practices that are believed to facilitate automatic inferences (see also GIGERENZER, 2018). In line with this assessment, a large body of literature (cf. e.g., HALLER and KRAUSS, 2002; GIGERENZER et al., 2004; KRÄMER, 2011) decries that misconceptions and misapplications have been entrenched and passed on for decades through inadequate teaching and even best-selling statistics textbooks.

Inferential errors associated with NHST seem to be caused to a large extent by a lacking familiarity with the differing but irreconcilable original perspectives: Fisher focuses on inductive reasoning and on forming rational scientific beliefs from a given set of data. In contrast, statistical decision theory according to Neyman and Pearson aims at providing behavioral rules across repeated decisions under consideration of error costs. We believe that getting acquainted with the basic arguments of these very different perspectives will clarify the potential and the limits of statistical inference.[2] This paper therefore looks back in history, scrutinizes the two original approaches, and shows that the NHST-framework is an ill-understood amalgamation that virtually invites inferential misconceptions. The paper is thus part of the vast body of critical literature regarding NHST that has been accumulating for many decades. But above all, it is motivated by the fact that misinterpretations of NHST continue to be an alarmingly "normal" practice despite the publicly disclosed errors. With a few exceptions, including the prominent publications by ZILIAK and MCCLOSKEY (2008) and KRÄMER (2011), the acknowledgement of the problems associated with NHST seems to be low in economics including agricultural economics. Therefore, we believe that agricultural economists as relevant audience of this journal will benefit from this historical methodological perspective.

After briefly recalling the problems associated with *p*-values and NHST in Section 2, we provide a primer of the two approaches in Section 3. Based on the understanding of the fundamental differences between Neyman-Pearson's and Fisher's perspective, we argue in Section 4 that *moving forward* in the field of statistical inference actually requires *going back* to the roots and using the standard error as measure of the

time, we ask whether the evidence in the data is consistent with the original proposition or whether we should change this proposition in the light of the new evidence. And then, the sequence of deductive and inductive steps starts again.

[2] This paper focuses on the inferential meaning that data-derived statistics can have *if* there was a probabilistic data generation process such as independent random sampling. It is beyond this paper's scope to discuss violations of this fundamental precondition of *statistical* inference and resulting problems such as sample selection bias. (cf. e.g., HIRSCHAUER et al., 2020b, or ROSENTHAL and ROSNOW, 2009). Unless adequately corrected for, selection bias precludes statistical inference because sample members may be systematically different from other members of the population, and because standard errors cannot be correctly estimated. Convenience samples do not meet the precondition of statistical inference. Given their widespread use, the violation of essential assumptions is therefore a major issue in practical research.

uncertainty of sample-based estimates. We finally conclude with six propositions that are hopefully useful for the revision of journal guidelines aimed at improving statistical inferences in empirical research in the social sciences and agricultural economics (Section 5).

2 A Brief Look at Inferential Errors

To avoid misinterpretations of the *p*-value, one must realize five fundamental facts (HIRSCHAUER et al., 2018, 2019): first, the use of *p*-values is concerned with the uncertainty of estimates resulting from random error as expressed through the standard error, which is the statistical label attached to the standard deviation of the sampling distribution. Using *p*-values presupposes a probabilistic data generation mechanism such as random sampling.[3] Second, estimates of quantities such as effect sizes (no causal implication intended), standard errors, and finally *p*-values can exhibit considerable sample-to-sample variability. Ignoring this variability will cause an overestimation of the *p*-value's per se limited inferential content. Third, unbiased estimators estimate correctly on average. Therefore, we need all estimates from statistical replications – irrespective of their *p*-values and their being large or small – to obtain an appropriate idea of the size of a population quantity.[4] Forth, it is a serious mistake to believe that a *p*-value of let's say 0.05 indicates that the probability of the null is 5 %. This is logically impossible since *p*-values are computed under the assumption that the null is true.[5] Fifth, a *p*-value indicates how (in)compatible a set of data (random sample) is with a specified statistical model including the null hypothesis, but *p*-values are not the epistemic probabilities that reflect the degree of rational belief regarding scientific hypotheses that we can have given the available evidence.

Being inherently based on the assumption that the null is true, *p*-values cannot be used to test hypotheses in terms of determining whether a null hypothesis or an alternative proposition is true (cf. WASSERSTEIN and LAZAR, 2016). It is even logically impossible to derive probabilities for hypotheses from the *p*-value without prior knowledge. In any ordinary sense of the word, a *p*-value can therefore neither "test" nor "confirm" a hypothesis. This holds even though the delusive terminology of the NHST framework does label findings as either "positive" (statistically significant) or "negative" (statistically non-significant) and indeed speaks of "hypothesis *testing*" and "*confirmatory* analysis."

After decades of critical debate and several years of high-level institutional attempts to reform statistical practice, we know that *p*-values and statistical significance tests have much less inferential content than what has been widely believed in the past. Whether or not statistical significance testing is reasonable and helpful at all is now prominently disputed in the literature. When researchers continue to use *p*-values and statistical significance tests as inferential tools, they should justify their approach and answer critical questions such as "Has there been a probabilistic process of data generation?" or "Which inductive inferences can be drawn from the result of a statistical significance test?" Irrespective of the auxiliary tools that are used to help make inferences, researchers should first describe the empirical evidence they found in their specific data set and then, in a separate step, tackle inference and the question of validity. When assessing the validity of findings, we should first specify the "inferential leap" that we want to make (To which

[3] While randomization represents another probabilistic data generation mechanism that facilitates the use of *p*-values, we focus on observational data and skip the discussion of causal inference in randomized controlled trials (cf. e.g., HIRSCHAUER et al., 2020a).

[4] TRAFIMOW et al. (2018) note that it is the *abnormally* large sample effect sizes that produce "highly significant" *p*-values. Considering only significant findings (i.e. one tail of the sampling distribution) would necessarily introduce bias. This has implications for what we have to understand by replicability. A meaningful definition needs to consider that each properly implemented study provides an incremental piece of evidence. Consequently, meta-analysis would need to obtain the weighted average of all estimates from replications, possibly even ones with opposite signs, to obtain the "best" summary of the accumulated evidence. But in the replication debate, a study with a "significant" finding followed-up by a study with a very high *p*-value or even an opposite sign of the estimate is often perceived as a replication failure. This applies even though popular replication projects such as CAMERER et al. (2016, 2018) use not only the statistical significance criterion but also other criteria to assess the replicability of original studies.

[5] COHEN (1994: 997) coined the term "inverse probability error" for this fallacy and noted: "[a *p*-value] does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does! What we want to know is 'given these data, what is the probability that H0 is true?' But […], what it tells us is 'given that H0 is true, what is the probability of these (or more extreme) data?'"

real-world population and context do we want to generalize?). We should then state our inferences (What do we claim to have learned from the analyzed set of data?) and explain how we arrived at these inferences. In general, the validity of findings increases the smaller the inferential leap from the idiosyncrasies of a particular data set to the social group and real-life setting of interest.

We may sum up that, contrary to beliefs that are often associated with the conventional routine of statistical significance declarations, inductive inferences do not flow from data automatically. Sample quantities such as standard errors, *p*-values (and confidence intervals for that matter) are summary statistics of certain properties of the particular set of data under study. If these statistics are correctly interpreted and if their probabilistic application requirements (in particular, random sampling) are met, they may help make inferences, but they remain auxiliary tools and cannot substitute scientific reasoning.

3 Confusion of α-Levels and p-Value Thresholds as a Source of Misunderstandings

Even when we succeed in identifying the most rational inductive belief given the available evidence, advisable courses of action cannot automatically be derived from such beliefs. As BERRY (2017: 895) pointedly noted: "[…] believing or advertising something as true and acting as though it is true are very different kettles of fish." That is, we need to distinguish two different intellectual tasks: judging what we should most reasonably believe regarding a real-world state of interest as opposed to judging what we should most reasonably do after additionally considering the costs of wrong decisions. For the sake of clarification, let us look at the Glyphosate example: even when we succeed in forming the most rational but still uncertain belief regarding the health impact of Glyphosate exposure given all available evidence, we run the risk of giving bad advice if we neglect the costs that different decisions – banning or not banning Glyphosate – entail when in the end they turn out to be wrong. Since the amalgamation of Fisher's approach and Neyman-Pearson's approach into the hybrid NHST ignores the crucial distinction between the above-mentioned two "kettles of fish," we now look into the two original approaches.

3.1 Statistical Decision Theory and the α-Level in Hypothesis Testing

In statistical decision theory according to Neyman and Pearson, the world ("parameter space") is divided into two mutually exclusive states – represented by the null and the alternative hypothesis. In addition, a constant choice structure across many repeated decisions is assumed. This has several implications:

1. One must not only specify a null hypothesis H_0 (implying that the alternative hypothesis is a vague "non-null" proposition) but also a concrete alternative hypothesis H_A. Moreover, H_0 and H_A together have to represent all possible states of the world.
2. A choice has to be made between these two hypotheses based on a statistical test. This test represents the decision rule ("rule of behavior") for decisions that are made many times – each based on a new random sample obtained from an identical sampling design (e.g. in industrial quality control).
3. For each random sample, a test score (e.g. a *z*- or *t*-value) is computed based on the effect size and the standard error (estimated variability of the sampling distribution).
4. A test result that induces the decision associated with H_A (=rejection of H_0) or the decision associated with H_0 (=non-rejection of H_0) is not linked to an inductive belief that either H_A or H_0 is true or more likely in a particular instance. Instead, accepting a hypothesis means to act as if it were true in the light of the consequences associated with either choice.
5. Different choices are fraught with different types of errors: type I errors arise if one acts as if H_A were true when in fact H_0 is true (false rejection of H_0). In contrast, type II errors arise if one acts as if H_0 were true when in fact H_A is true (false non-rejection of H_0).
6. The rule of behavior is based on an a priori fixed level of the type I error rate (false positive rate), which is usually designated by α. When the test results in $p \leq \alpha$, one is to act as if H_A were true ("accept H_A"). When the test results in $p > \alpha$, one is act as if H_0 were true ("accept H_0").
7. Since the test is a "rule of behavior" aimed at guiding decisions that are made many times under constant conditions, the particular value of p in a particular test is completely irrelevant. The only relevant information is whether p falls into the rejection region or not.

8. In subsequent samples and tests, *p*-values will be different. A *p*-value found in a particular sample is therefore not the type I error rate over many replications. But consistently following the rule of rejecting H_0 when $p \leq \alpha$ guarantees that, in the long run, the type I error rate will be α. This also follows from the fact that the *p*-value is uniformly distributed under the null.
9. Decreasing the test's type I error rate α will decrease its power $1-\beta$, i.e., the long-term rate of acting as if H_A were true when it is true (true positive rate).[6] Consequently, there is a tradeoff when setting the level of α: decreasing the type I error rate α (false positive rate) across repeated decisions will increase the type II error rate β (false negative rate) across these decisions.
10. While using $\alpha = 0.05$ is often seen as a general default, Neyman and Pearson explicitly warned against a standard level for all decision contexts. Importantly, the magnitude of type I and type II error costs in a particular context must be considered when setting the level of α.

Statistical decision theory in the Neyman-Pearson tradition rejects the idea of making inductive inferences about some real-world state of interest. Instead, statistical decision theory uses "hypothesis testing" to identify rules of behavior in the light of a "loss function" that considers the magnitude of, and the relationship between, type I and type II error costs. NEYMAN and PEARSON (1933a: 296; 1933b: 497) note that it "must be left to the investigator" to set an appropriate α that strikes the balance between the two types of errors "to meet the type of the problem before us." Along the same lines, ZILIAK and MCCLOSKEY (2008: 8-9) note that "without a loss function a test of statistical significance is meaningless […]." Many vivid examples have been used to underpin the importance of considering type I and type II error costs when determining the decision rule α. A recent example is HARVEY (2017: 1408) who uses the comparison between "a jet engine failing" vs. "a water heating failing" to illustrate how different the problems before us can be: "In the case of the jet engine, we are willing to accept a lot of false positives (incorrectly label a part defective) to minimize chances of false negatives (miss detecting a defective part), so α is set to a higher level. The particular situation therefore dictates not only how low α will be set but also the Type II error rate."

While being labeled „statistical *decision* theory", it must be noted that the approach by Neyman and Pearson remains a conditional-probability concept: α is the long-term type I error rate *when* H_0 is true ($P(type\ I\ error|H_0)$), and the corresponding β is the long-term type II error rate *when* H_A is true ($P(type\ II\ error|H_A)$). Since no scientific propositions regarding the probabilities of H_0 or H_A are provided, an important piece of information is missing that is indispensable when we want to obtain a normative rule from a decision theoretic point of view. For illustration sake, let's assume that we test the jet engine 10,000 times and that, after setting the type I error rate to $\alpha = 0.05$, the type II error rate is $\beta = 0.2$. Imagine that the costs of making a type I or a type II error are known to be 1,000 € and 10,000 €, respectively. Imagine also that, from prior evidence, we expect to see a defective part very often and correspondingly assume probabilities $P(H_0 = no\ defective\ part) = 0.01$ and $P(H_A = defective\ part) = 0.99$. In this case, we would incorrectly label a part defective in only 5 of the 10,000 test instances. The costs of these 5 false alarms would amount to a total of 5,000 €. At the same time, we would miss detecting a defective part in 1,980 of the 10,000 test instances. Incurred costs would total 19.8 mill. €. In this context, it would obviously be rational to use a decision rule α much larger than 0.05 in order to reduce β and thus the costs of missing defective parts. Things are very different if, based on prior evidence, it is reasonable to assume $P(H_0) = 0.99$. In this case, we can expect to make 495 type I errors and 20 type II errors over the 10,000 tests. That is, the costs of false alarms would total 495,000 € and the costs of missing a detective part would total 200,000 €. We should consequently lower the decision rule α and accept an increase of β from a long-term cost perspective. This illustrates that a decision rule that minimizes expected costs needs to take into account *how often* we can expect to commit which type of error in the long run of testing.[7] Besides

[6] Power is the zeroth order lower partial moment of the *p*-value distribution over replications under the alternative hypothesis H_A for the value of the test statistic associated with a particular α. This partial moment is sufficient in a dichotomous rejection/non-rejection context, i.e. power quantifies the repeatability of $p \leq \alpha$ when H_A is true.

[7] For the sake of easy traceability, our numerical illustration of the "jet engine" vs. "water heating" example is based on the assumption of a given sample size. In this case, considering how lowering α (and thus type I error costs) increases β (and thus type II error costs) suffices to minimize long-term costs. In contrast, including the

the conditional type I and type II error rates, this depends on the prior probabilities of the null and the alternative hypothesis, a concept beyond the conditional probability approach by Neyman-Pearson.[8]

HURLBERT and LOMBARDI (2009: 311) claim that "[t]he original Neyman-Pearson framework has no utility outside quality control type applications." In the light of the jet engine example above, we might add that even in this context, its utility is limited from a decision-theoretic point of view – unless prior probabilities are implicitly considered when setting the decision rule α. Regardless of this, basic and applied empirical research is usually not aimed at providing practical decision support in terms of assessing the costs, benefits, and risks associated with particular choices in particular decision environments. Rather, its objective is to gain (incremental) knowledge regarding a real-world feature by studying a particular set of data and weighing the evidence. Thresholds and dichotomies are not only superfluous but seriously misleading in such contexts.

3.2 Inductive Reasoning and the *p*-Value in Statistical Significance Testing

NHST, i.e., using *p*-values accompanied by ostensibly self-explaining significance declarations that advertise something as true/existent or not, is still the statistical methodology in dominant use in econometric studies. Even though NHST has borrowed the term "hypothesis testing" from statistical decision theory, it is closer in spirit to Fisher's view that the *p*-value represents a helpful tool in the difficult exercise of making inductive inferences; and most econometricians would probably agree with FISHER's (1935: 39) statement of what this exercise is about: "[…] everyone who does habitually attempt the difficult task of making sense of figures is, in fact, essaying a logical process of the kind we call inductive, in that he is attempting to draw inferences from the particular to the general." In other words, despite its hypothesis testing terminology, NHST has not adopted the Neyman-Pearson perspective of providing a behavioral rule for repeated decisions given their error costs. Instead, it shares Fisher's interest of drawing inductive inferences from a given set of data. Consequently, there are also common grounds with regard to what is to be understood by replication. Whereas statistical decision theory focuses on tests as behavioral rules for repeated choices under constant conditions, replication from the viewpoint of inductive reasoning is first of all a mind experiment based on which a sampling distribution can be conceived of. That is, we study one sample but envisage what would happen if we drew many other equal-sized random samples from the same population and applied the same econometric model to these samples (hypothetical statistical replications). We then ask the question of how much the sample quantity such as a regression coefficient would vary across these statistical replications. The standard error, which is itself an estimate that varies from sample to sample, indicates the dispersion (standard deviation) of the sampling distribution (e.g. the standard deviation of regression coefficients across replications) to the best of our knowledge.

The practical approach of NHST in econometrics is also more similar to Fisher's than to Neyman-Pearson's perspective in that there is usually no well-specified alternative hypothesis. But being a misleading hybrid, many NHST-based empirical studies focus on "asterisks" and make inductive inferences in a quasi-automated way depending on whether the *p*-value is below or above some arbitrary threshold. While Fisher emphasized that inductive inferences always remain uncertain, he agreed with labeling results associated with $p \leq 0.05$ as "statistically significant" and even claimed that "non-significant" results can be ignored. This seems to have played an important role in the dissemination of dichotomous language and thinking in inductive inference, which, in turn, has contributed to the confusion with the decision rule α (often also set to a default of 0.05) used by Neyman and Pearson. Even Fisher's clarification that low *p*-values simply *signify* "worth a second look" and a later warning that, beyond "convenient" significance statements, "exact" *p*-values indicating the strength of evidence against the null should be used as an aid to judgment in inductive inference (FISHER, 1960: 25) did not prevent the confusion – inherent to NHST – between *p*-value thresholds used for "convenient" significance declarations in inductive inference and α used as a behavioral rule in statistical decision theory.

sample size into the set of decision variables requires considering additionally that a costly increase of sample size n will ceteris paribus decrease β (and thus type II error costs). Another label for this would be to say that an increase of sample size increases power. In this more complex decision context, minimizing long term costs would imply finding the optimal combination of n and α.

[8] NEYMAN and PEARSON (1933a: 291) may have triggered off overvaluations by claiming: "Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behavior with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong."

4 Back to the Roots: Using the Standard Error as Measure of the Uncertainty of Estimates

Understanding the potential and the limits of statistical inference is straightforward when we remember that it is conceptually based on the sampling distribution and its standard error. This implies keeping with the following step-by-step approach: (1) We collect a random sample from a defined parent population. (2) We observe a sample quantity ("effect size"), which can be a difference between two groups or an association between two variables such as a regression coefficient. (3) Aiming to make generalizing inferences, we use the observed sample quantity as an estimate for the population quantity of interest. (4) When assessing the validity of the estimate, we consider that inductive inferences are inherently uncertain and that they do not flow from a summary statistic of the data such as a *p*-value automatically. (5) While the validity of findings beyond the narrow confines of an idiosyncratic study depends on more than just random sampling error, we realize that statistical inference is limited to considering the uncertainty of estimates resulting from random error. (6) Accounting for the specific data collection design (simple random sampling, stratified sampling, cluster sampling etc.), we estimate the standard error to quantify the uncertainty of the estimate.[9] (7) We now possess information that can be understood as "signal" and "noise", with the sample effect size representing the signal, and the standard error representing the noise from random sampling. (8) We use these two intelligible pieces of information – the estimate and its uncertainty – in a comprehensive scientific reasoning that makes reasonable inferences from the idiosyncratic sample towards the real-world state of interest in the light of all available information. If convenient, we aggregate signal and noise into a signal-to noise ratio such as a *t*- or *z*-score.

Table 1 illustrates the potential and the limits of statistical inference by using an idealized example: our presumed research interest is to gain knowledge regarding the magnitude of the income gap between men and women in Berlin. We assume that two independent research teams tackle the issue but use different data collection designs. The first one draws a simple random sample of size $n = 600$ from the real-world population of interest (i.e., the residents of Berlin). For convenience sake, we assume that exactly 300 women and 300 men happen to be in this sample. Having less resources, the second research team draws a simple random sample of size $n = 60$. For convenience sake, we assume again that the random draw results in equal-sized groups. Imagine now that, by coincidence, both research teams find the same group difference $D = 100$ as well as identical within-group standard deviations $s = 745$ in the male and female groups.[10]

Now, both teams must "make sense of their figures" and address the inferential task of generalizing from the finding in the particular sample towards the population of interest. While the identified size of the income difference is 100 in both research designs, the variability of differences that would be found across many repeatedly drawn random samples would be much smaller in design 1 ($SE = 60.83$) than in design 2 ($SE = 192.36$). Despite identical effect sizes, the researchers using design 2 would face much more remaining uncertainty and would have to be much more cautious in their reasoning what they can claim to have learned from their small sample regarding the income differences between women and men in Berlin. This is because, due to the law of large numbers, the inferential leap is wider from a small sample to the parent population than from a large sample to that population.

With some loss of information, the evidence from the data in terms of the size of the group difference found in the sample (estimate or "signal") and its uncertainty (standard error or "noise") can also be expressed as a signal-to-noise ratio, which is 1.64 in design 1 and 0.52 in design 2. When computing the *p*-value, we use the signal-to-noise ratio and combine it with the assumption of no-effect in the parent population (point null hypothesis). The *p*-value then answers the question: What is the conditional probability of

[9] Saying that the uncertainty of estimates is used to assess the validity of generalizations towards a parent population (population validity) is at odds with the terminology of measurement theory which would distinguish between *precision* (or reliability or certainty) and *accuracy* (or validity). For example, we might find nearly identical sample effect sizes in statistical replications (precise/reliable/certain "measurement"), but they might be systematically biased estimates of the population effect size (inaccurate/invalid "measurement"). The terminological conflict would be that, in statistical inference, we use a measure of (un)certainty to assess the population validity, which would not be termed "validity" in measurement theory.

[10] It should be noted that we use stylized numbers for the sake of easy traceability and understanding of the fundamental methodological issues. They are not chosen to reflect empirical facts.

Table 1. Inferential statistics for a group difference

	Study design 1		Study design 2	
	Group 1 (men)	Group 2 (women)	Group 1 (men)	Group 2 (women)
Sample size *n*	300	300	30	30
Standard deviation	745	745	745	745
Group mean	2 100	2 000	2 100	2 000
Difference in group means (signal): *D*	100		100	
Standard error of group difference (noise): *SE*	60.83		192.36	
Signal-to-noise ratio: *z*	1.64		0.52	
p-value (one-sided)	0.05		0.30	

Source: own calculation

finding this data (or more precise: of finding the observed signal-to-noise ratio or even a larger one) in random replications *if* we assumed the point null hypothesis (here: no income difference between women and men) to be true in the parent population.

Expressing the signal and noise information as a *p*-value in conjunction with the dichotomous significance declarations and the delusive terminology of NHST, which speaks of hypothesis testing and of rejecting or confirming hypotheses depending on whether *p* is above or below some specified "significance" threshold, has apparently led to much confusion. Despite NHST's delusive invitation to interpret a *p*-value below or above some arbitrary threshold as a rule of what to believe, it would be a gross mistake to advertise the *p*-value of 0.30 in study 2 as an indication of no difference and a failure to replicate the finding of study 1. There is not even the slightest indication pointing in this way: after all, we did find a difference of 100 in the small sample and this is completely consistent with what we found in the large sample. It would also be a gross mistake to interpret the "statistically significant" *p*-value in design 1 as a confirmation (of an ex post invented hypothesis) of a real difference of 100. In other words, conventional "statistical significance" is neither sufficient nor necessary to conclude that there is a substantial effect. The only thing that from the perspective of the single study can be said is that $p = 0.05$ represents stronger evidence against the point null hypothesis than $p = 0.30$ because small *p*-values occur more often if there is an effect compared to no effect. Doing so, one should recognize that, contrary to a signal-to-noise ratio such as a *z*-or *t*-value, the *p*-value is a non-linear statistic in that a difference between, let's say, a *p*-value of 0.30 and 0.29 does not indicate the same increase of the strength of evidence against the point null as a difference between 0.05 and 0.04. Another way of expressing the meaning of a *p*-value would be to say that lower *p*-values indicate a lower compatibility of the data with the point null hypothesis.

Effect size and standard error – or "estimate and uncertainty" or "signal and noise" – and their derivatives such as *p*-values can help make inductive inferences from a particular set of data (random sample) towards a parent population. All the rest that we are confronted with in the NHST framework, i.e., hypothesis testing terminology and dichotomous significance declarations associated with arbitrary *p*-value thresholds, propagate cognitive biases and seduce researchers to make logically inconsistent and overconfident inferences, both when *p* is below and when it is above the "significance" threshold. These misinterpretations are rooted in the very fabric of NHST as an inconsistent hybrid of the "hypothesis testing" approach by Neyman and Pearson and the ill-termed "significance testing" approach by Fisher.

The amalgamation of the two approaches into NHST is inconsistent because it pretends that we can use a critical signal-to-noise ratio (e.g. 1.645) to obtain a dichotomous inductive rule of what to believe, while such a dichotomy would at best make sense as a behavioral rule of what to do in the light of a presumably known loss function. As a consequence, it invites the misunderstanding that inferences are just a matter of statistics and that they flow from data automatically. Unfortunately, statistical practitioners seem to have widely succumbed to NHST's misleading invitation and forgotten that the standard error, upon which all inferential statistics are based, is no more (and no less) than a continuous estimate of the variability (standard deviation) of the sampling distribution.[11] Because of

[11] While we do not focus on randomized experiments in this paper, it seems worthwhile noting that many experimental economists consider power analysis as a means to improve the quality of experimental research. This implies that, contrary to most of econometric research, they identify a well-specified alternative hypothesis. But while this resembles the approach by Neyman and Pearson at first view, experimental economists are interested in making inductive inferences from a particular set of data and not in obtaining a behavioral rule for concrete decisions fraught with specific error costs. Using the concept of power in the exercise of inductive reasoning presupposes that one keeps with dichotomous significance declarations and resulting yes/no conclusions be-

dichotomous significance declarations, which downgrade the intelligible and continuous uncertainty information to a misleading binary variable, researchers all too often ignore that what they have is a signal and a noise information. Consequently, common interpretations in the widely used NHST framework deviate substantially from Fisher's modest "worth a second look" interpretation of the *p*-value. While at the very best we can reduce uncertainty (through larger samples) and assess uncertainty (by correctly estimating the standard error), we must embrace the remaining uncertainty caused by random error as long as we study samples instead of full populations.

5 Conclusion

We have not been digging very deep into the history of statistical science, but we hope that this primer on the differing perspectives of inductive reasoning (Fisher) as opposed to statistical decision theory (Neyman and Pearson) helps prevent inferential errors that are largely due to the delusive NHST-amalgamation of these two irreconcilable approaches. While we believe that dichotomous approaches are not helpful in inductive inference, we realize that not all empirical researchers agree with the calls to abandon significance testing as those made, for example, by AMRHEIN et al. (2019) and WASSERSTEIN et al. (2019). In view of the ongoing debate, we hope that the following propositions contribute to reaching a consensus that improves the quality of scientific communication and inductive inference in the future:

1. For the sake of clarity, *first describe the empirical evidence you found in your specific data set* and then, in a subsequent step, tackle inductive inference.
2. When you use inferential statistics (i.e., standard-error-based summary statistics of your data) as auxiliary tools for drawing inferences, clearly communicate *which probabilistic data generation mechanism* (e.g. simple random sampling, stratified sampling or cluster sampling) you applied.
3. Transparently communicate, for *which kind of inference* you want to use inferential statistics as an aid to judgment: generalizing inference, causal inference (not addressed here), or both.
4. In the case of generalizing inference, clearly state from which parent population the random sample was drawn and therefore *to which parent population you want to generalize* with the help of inferential statistics.
5. When keeping with NHST, explain *how your inference is better supported by dichotomous significance declarations* than by exact *p*-values that represent a graded measure of the strength of evidence against a point null hypothesis.
6. When using *p*-values as an aid to judgment, explain *how your inferences gain from reporting a non-linear measure of the strength of evidence against a point null hypothesis* instead of reporting the "original" evidence you found in the data in terms of the size and the uncertainty of the estimate.

We believe that an efficient way of supporting a quick change for better inferences in the 21st century (cf. WASSERSTEIN et al., 2019) would be to revise journal guidelines in the spirit of these hopefully consensual proposals. Such revisions would have to include corresponding propositions for randomized controlled trials where inferences are first of all causal in nature ("treatment effects") and where the standard error indicates the dispersion of the randomization distribution. Besides the hopefully immediate benefits of appropriate guideline revisions, support for better inferences should also come from professional associations and, in particular, from statistics teaching. It should provide students not only with a deep understanding but also an internalized intuitive grasp of the concept of the sampling distribution, a grasp that immunizes them against misinterpretations. Such misinterpretations seem to be invited by the fact that statistical terminology often contradicts the primal meanings from natural language, which, however, are more deeply rooted in people's mind. Confusing the term "statistically significant" with "large" or "scientifically proven" is but one example.

We hope that the propositions above contribute to an overall better understanding of the fact that the scientific enterprise represents an ongoing process of accumulating evidence and knowledge. In this process, each appropriately implemented single study that transparently communicates the data and the analytical methods plays an important role, but by itself a single study cannot provide probabilities, let alone definite conclusions for scientific propositions. The knowledge contribution of the single study must therefore always be critically assessed, and we must take into account both the potential and the limits

cause, assuming the conventional threshold, power indicates the repeatability of $p < 0.05$ under the alternative hypothesis. Endorsing power therefore implies endorsing NHST – in opposition to present calls to abandon statistical significance testing.

of inferential statistics in this assessment. In some research contexts, for example when an explicit alternative hypothesis is meaningful, it might be worth to get granular on Bayesian procedures. Even when we are reluctant to specify prior odds, the Bayes factor of the single study is informative because it can be used in a what-if-analysis that shows how much *various* prior odds *would* be changed through the study under consideration. In other instances, for example when there are many structurally similar studies in the field of interest, meta-analytical approaches might be useful to assess what is the most plausible estimate of a population quantity – and its remaining uncertainty – in the light of the evidence across the summarized studies. The advantage of meta-analytical approaches is that they leave behind rash interpretations of the single study. Instead, meta-analysis synthesizes, with adequate weights, the informational content of the included studies irrespective of their respective *p*-values.

Literature

AMRHEIN, V., S. GREENLAND. and B. MCSHANE (2019): Retire statistical significance. In: Nature 567: 305-307.

BERRY, D. (2017): A p-Value to Die For. In: Journal of the American Statistical Association 112 (519): 895-897.

CAMERER, C.F., ..., H. WU (2018): Evaluating the Replicability of Social Science Experiments in Nature and Science Between 2010 and 2015. In: Nature Human Behaviour 2: 637-644.

CAMERER, C.F., ..., H. WU (2016): Evaluating replicability of laboratory experiments in economics. In: Science 351 (6280): 1433-1436.

COHEN, J. (1994): The earth is round ($p < 0.05$). In: American Psychologist 49 (12): 997-1003.

FISHER, R.A. (1925): Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.

FISHER, R.A. (1935): The Logic of Inductive Inference. In: Journal of the Royal Statistical Society 98: 39-54.

FISHER, R.A. (1960): The Design of Experiments. 7th ed. Oliver and Boyd, Edinburgh.

GIGERENZER, G. (2004): Mindless statistics. In: The Journal of Socio-Economics 33: 587-606.

GIGERENZER, G. (2018): Statistical Rituals: The Replication Delusion and How We Got There. In: Advances in Methods and Practices in Psychological Science 1 (2): 198-218.

HALLER, H. and S. KRAUSS (2002): Misinterpretations of Significance: A Problem Students Share with Their Teachers? In: Methods of Psychological Research Online 7 (1): 1-20.

HARVEY, C.R. (2017): Presidential Address: The Scientific Outlook in Financial Economics. In: The Journal of Finance LXXII (4): 1399-1440.

HIRSCHAUER, N., S. GRÜNER, O. MUßHOFF and C. BECKER (2018): Pitfalls of significance testing and p-value variability: An econometrics perspective. In: Statistics Surveys 12 (2018): 136-172.

HIRSCHAUER, N., S. GRÜNER, O. MUßHOFF and C. BECKER (2019): Twenty steps towards an adequate inferential interpretation of p-values in econometrics. In: Journal of Economics and Statistics 239 (4): 703-721.

HIRSCHAUER, N., S. GRÜNER, O. MUßHOFF and C. BECKER (2020a): Inference in economic experiments. In: Economics. The Open-Access, Open-Assessment E-Journal 14 (2020-7): 1-14.

HIRSCHAUER, N., S. GRÜNER, O. MUßHOFF, C. BECKER and A. JANTSCH (2020b): Can p-values be meaningfully interpreted without random sampling? In: Statistics Surveys 14 (2020): 71-91.

HURLBERT, S.H. and C.M. LOMBARDI (2009): Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. In: Annales Zoologici Fennici 46: 311-349.

KENNEDY-SHAFFER, L. (2019): Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize p-Values and Significance Testing. In: The American Statistician 73 (sup1): 82-90.

KRÄMER, W. (2011): The Cult of Statistical Significance – What Economists Should and Should Not Do to Make their Data Talk. In: Schmollers Jahrbuch 131 (3): 455-468.

LEHMAN, E.L. (1993): The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? In: Journal of the American Statistical Association 88: 1242-1249.

NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE (2019): Reproducibility and Replicability in Science. Consensus Study Report. The National Academies Press, Washington, DC. https://doi.org/10.17226/25303.

NEYMAN, J. and E.S. PEARSON (1933a): On the problem of the most efficient tests of statistical hypotheses. In: Philosophical Transactions of the Royal Society of London A 231: 289-337.

NEYMAN, J. and E.S. PEARSON (1933b): The testing of statistical hypotheses in relation to probabilities a priori. In: Proceedings of the Cambridge Philosophical Society 29: 492-510.

ROSENTHAL, R. and R.L. ROSNOW (2009): Artifacts in Behavioral Research. Oxford University Press, Oxford.

TRAFIMOW, D., ..., F. MARMOLEJO-RAMOS (2018): Manipulating the alpha level cannot cure significance testing. In: Frontiers in Psychology 9. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00699/full.

WASSERSTEIN, R.L. and N.A. LAZAR (2016): The ASA's statement on p-values: context, process, and purpose. In: The American Statistician 70 (2): 129-133.

WASSERSTEIN, R.L., A.L. SCHIRM and N.A. LAZAR (2019): Moving to a World Beyond "$p < 0.05$". In: The American Statistician 73 (sup1): 1-19.

ZILIAK, S.T. and D.N. MCCLOSKEY (2008): The Cult of Statistical Significance. How the Standard Error Costs Us Jobs, Justice, and Lives. The University of Michigan Press, Michigan.

Acknowledgments

We would like to thank the German Research Foundation for financial support.

Contact author:
PROF. DR. NORBERT HIRSCHAUER
Martin-Luther-Universität Halle-Wittenberg
Institut für Agrar- und Ernährungswissenschaften
06120 Halle (Saale), Karl-Freiherr-von-Fritsch-Str. 4
e-mail: norbert.hirschauer@landw.uni-halle.de