Institute for Food and Resource Economics

University of Bonn

Discussion Paper 2021:2

The replicability crisis and the p-value debate – what are the consequences for the agricultural and food economics community?

Thomas Heckelei, Silke Hüttel, Martin Odening, Jens Rommel

Editor: Thomas Heckelei

Institute for Food and Resource Economics

University of Bonn

Nußallee 21

53115 Bonn, Germany

Phone: +49-228-732332

Fax: +49-228-734693

E-mail: thomas.heckelei@ilr.uni-bonn.de

Agricultural and Resource Economics, Discussion Paper 2021:2

The replicability crisis and the p-value debate – what are the consequences for the agricultural and food economics community?[1]

Thomas Heckelei, Silke Hüttel, Martin Odening, Jens Rommel

Abstract

A vivid debate is ongoing in the scientific community about statistical malpractice and the related publication bias. No general consensus exists on the consequences and this is reflected in heterogeneous rules defined by scientific journals on the use and reporting of statistical inference. This paper aims at discussing how the debate is perceived by the agricultural economics community and implications for our roles as researchers, contributors to the scientific publication process, and teachers. We start by summarizing the current state of the p-value debate and the replication crisis, and commonly applied statistical practices in our community. This is followed by motivation, design, results and discussion of a survey on statistical knowledge and practice among the researchers in the agricultural economics community in Austria, Germany and Switzerland. We conclude that beyond short-term measures like changing rules of reporting in publications, a cultural change regarding empirical scientific practices is needed that stretches across all our roles in the scientific process. Acceptance of scientific work should largely be based on the theoretical and methodological rigor and where the perceived relevance arises from the questions asked, the methodology employed, and the data used but not from the results generated. Revised and clear journal guidelines, the creation of resources for teaching and research, and public recognition of good practice are suggested measures to move forward.

Keywords: Statistical inference, p-hacking, pre-registration, publication bias, replication crisis
JEL classification: C10, C18, Q00

[1] We gratefully acknowledge very valuable input from participants of an organized session and a pre-conference workshop at the annual meetings of the GEWISOLA in 2020 and 2021, respectively. In addition, several contributors to these events shall be mentioned here whose presentations and, in parts, comments to an earlier version of this paper helped deriving and formulating our views on the matter: Carola Grebitus, Norbert Hirschauer, Carl-Johan Lagerkvist, Martin Petrick, and Pallavi Shukla.

1 Introduction

Replicability of research results is at the core of scientific credibility. The discussion of a "replication crisis" in science has intensified over the last years (Loken and Gelman, 2017; Schooler, 2014) and also reached the community of environmental and resource economics (Ferraro and Shukla, 2020). Practices like selective reporting of results, incentives to find "significant" effects in statistical analysis and the underrepresentation of null results (Mervis, 2014) are discussed as core issues in the debate.

A more specific but strongly related issue is the use and interpretation of p-values and "p-hacking" in the context of statistical hypothesis tests. "Mindless statistics" (Gigerenzer, 2004) and "The cult of statistical significance" (Ziliak and McCloskey, 2008) are terms to describe the widespread misuse and misinterpretation of p-values and statistical significance in reporting results of statistical and econometric analyses. The American Statistical Association has published a statement (Wasserstein and Lazar, 2016), and several researchers signed a call to "retire statistical significance" (Amrhein, Greenland and McShane, 2019). However, this is countered by others who acknowledge existing problems but nevertheless defend p-values, basically saying that nothing is wrong with p-values if they are used correctly (Imbens, 2021). Currently, no consensus across the scientific community exists on the consequences of publication bias and malpractices, and this is reflected in heterogeneous rules defined by scientific journals on the use and reporting of statistical inference.

The agricultural economics community in Germany joined the debate by the fundamental work of Hirschauer *et al.* (2019) who suggest changes of rules for using p-values and statistical inference. After the first discussion in an organized session at the annual meeting of the German agricultural economics association (GEWISOLA) in 2019, the association created a working group with the task to assess how "p-hacking" and the misuse of statistical hypothesis tests in our scientific publications can be best avoided. In addition to the discussion of specific rules and best practices, the incentives leading to p-hacking and misinterpretations in the publication process were of interest. Ultimately, the working group targets at giving recommendations to the members of the association on how we can improve upon the current practice by changing relevant aspects of teaching, research and the scientific publication process.

This paper presents results of the working group and discusses implications of the debate on p-values and statistical inference for our roles as researchers, contributors to the scientific publication process, and teachers, as well as for a needed cultural change. To arrive at this end, we first offer some background knowledge generated through the working group's activities on the current state of the p-value debate and statistical practices more generally in the literature. This is followed by motivation, design, results, and discussion of a survey on statistical knowledge and practice among the researchers in the agricultural economics community in Austria, Germany and Switzerland. Based on this knowledge and additional input from external experts and participants of two GEWISOLA events in 2020 and 2021, implications for the community are developed.

2 The p-value debate and related statistical practice

The "p-value debate" has many facets. We argue that it is useful to distinguish two main problem areas: first, unintentional misinterpretations and wrong conclusions from statistical inference, particularly significance tests and p-values. Second, malpractices when applying statistical test procedures, such as p-hacking or HARKing (**H**ypothesizing **A**fter the **R**esults are **K**nown). This distinction is useful, as we believe each calls for distinct responses from the community.

2.1 *Misunderstanding and common flaws when applying p-values and statistical hypothesis testing*

2.1.1 Wrong interpretations of p-values and significance tests

Before we turn to common misinterpretations of p-values and statistical hypothesis testing, we briefly reiterate their meaning. The purpose of a statistical test is to infer how compatible observed data D are with a null hypothesis H_0, which is specified in the framework of a statistical model, e.g. a regression model. The null hypothesis can be a statement about the size of a model parameter, e.g. the assumption that an unknown regression coefficient belonging to an economic variable has the value zero.[2] A statistical test requires (i) the derivation of a test statistic T, e.g. a z-score, a t-value or an F-value, for which the probability distribution is known, when the null hypothesis is true and some other distribution when the null hypothesis is false, and given that the set of model assumptions A are true, e.g. independence of the model's error terms; and (ii) a rejection rule, such if the value of the test statistic is an extreme one that would rarely be encountered by chance under the null hypothesis, then the test provides evidence against the null hypothesis.

In this setting, Fisher (1925) defines the p-value as the conditional probability of the test distribution that refers to the observed value of the test statistic t, i.e. $Prob\ (T < t \mid H_0, A)$ for a one-sided test. Since it is often desired to arrive at a decision about the presence of an economic effect, the observed p-value is compared with a predetermined cut off-rate α, the "significance level", usually 0.05. If the observed p-value is smaller than the significance level, the null hypothesis is rejected, otherwise not.[3] The significance level reflects the type-I-error, i.e. $Prob\ (reject\ H_0 \mid H_0\ is\ true)$. The p-values are also

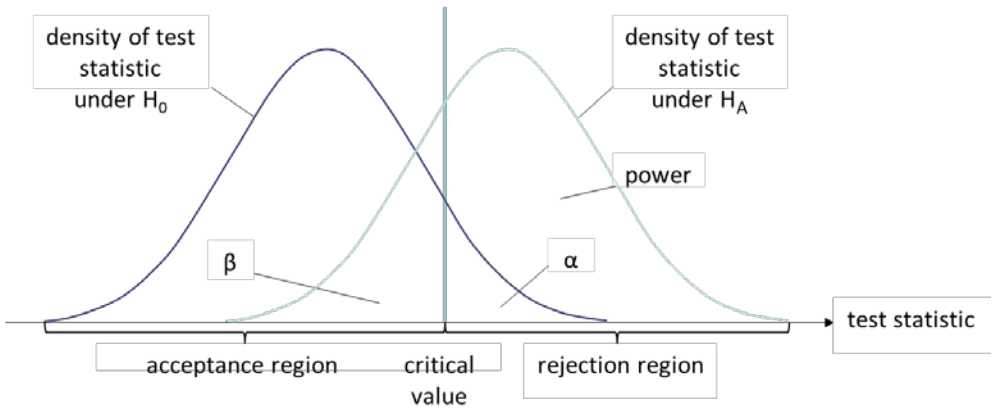
[2] It is important to note that H_0 need not to be a "nil hypothesis", as it is of the case in economic applications. In fact, the choice of a meaningless null hypothesis as a "strawman hypothesis", that can easily be rejected, has be blamed by Ziliak and McCloskey (2008) as being part of the "cult of null hypothesis significance testing" (NHST).

[3] Some authors prefer to speak of a "non-rejection" and avoid "acceptance" following the approach of falsification, and also to avoid the wrong conclusion that H_0 is actually true.

called marginal significance levels as it relates to the respective test statistic's greatest level for which the test based on the test statistic fails to reject the null hypothesis.

The complete decision-theoretic framework as proposed by Neyman and Pearson (1933) further involves the definition of an alternative hypothesis H_A and the determination of the test statistic's distribution under H_A. The distribution of the test statistic under H_A is used to determine the type-II-error $\beta = Prob\ (accept\ H_0 \mid H_A\ is\ true)$ and the power of the test $1 - \beta = Prob\ (reject\ H_0 \mid H_A\ is\ true)$ (see figure 1)[4]. In econometric applications, however, alternative hypotheses are often not explicitly spelled out, which renders the determination of β-errors and power calculations impossible.

Figure 1: Statistical Hypothesis Testing



Source: Neyman and Pearson (1933)

Even stern critics of the concept of statistical hypothesis testing do not deny that p-values contain some useful information. Loosely speaking, the p-value informs how compatible data are with a null hypothesis (Wasserstein and Lazar, 2016). Thus, they are a quantitative tool to challenge our initial belief and can be considered as a "first defense line against being fooled by randomness" (Benjamini, 2016). However, one should not get confused by this statement. From the above definition of a p-value it follows that they are derived from the sample data and thus observed p-values are random themselves. They vary from sample to sample, a characteristic, that is sometimes labelled as "p-value dance" (Greenland, 2019).

[4] Hirschauer *et al.* (2021) emphasize that the concepts suggested by Fisher (1925) and Neyman and Pearson (1933) are "two different kettle of fish". While Fisher aimed at inductive inference, i.e., identifying the most rational belief given the available data, Neyman and Pearson's statistical decision theory provides behavioral rules for repeated decisions.

Another characteristic of p-values is that they merge information regarding the size of an effect (the difference between the estimate and the hypothesized value) and the precision of the estimate (the standard error of the estimate). This "confounding" of information is per se not a problem (Greenland, 2019), but it facilitates a common confusion of statistical significance and economic importance (Gelman and Carlin, 2017). If enough data are available, the standard error of the estimate becomes small and in turn, even a small difference between the estimated model parameter and its hypothesized value is classified as "significant" regardless of its practical relevance. Conversely, large effects may not become statistically significant in small samples. In response to this potential confusion, some authors suggest not to use the term "significant" in empirical applications any more (Hirschauer *et al.*, 2019; Wasserstein, Schirm and Lazar, 2019).

A common misunderstanding that has been deplored in the p-value debate, applies to the interpretation of the outcome of statistical tests as a proof that either the null hypothesis or the alternative hypothesis are true or wrong (Greenland *et al.*, 2016). According to Gigerenzer (2018) researchers are driven by the desire to provide empirical evidence for or against a hypothesis and hence p-values are erroneously interpreted as $Prob\ (H_0 \mid data)$. P-values are related to this conditional probability via Bayes theorem, i.e. $Prob\ (H_0 \mid Data) \sim Prob\ (Data \mid H_0) * Prob\ (H_0)$, where $Prob\ (H_0)$ denotes the a-priori probability of the null hypothesis. Nevertheless, these probabilities are different entities and equating them would constitute a "fallacy of reverse inference" (Krueger and Heck, 2019). Thus, it would be incorrect to conclude from a p-value larger than 0.05 (or any other pre-defined threshold) that an economic effect is absent or in other words: "absence of evidence is not evidence of absence" (Altman and Bland, 1995). In real-world applications, this is especially relevant when considering very rare but very impactful events. Likewise, it would be wrong to infer from a small p-value that a specific alternate hypothesis is true. A small p-value merely reflects a misfit of the null hypothesis (under maintained model assumptions) to the data. A small p-value is compatible with many alternative hypotheses and might also be caused by a violation of other model assumptions.

A related problem is the interpretation of p-vales or significance levels as false discovery rates (FDR) (Hirschauer *et al.*, 2016). A FDR defines the probability of rejecting the null hypothesis though it is true. It is an unconditional probability that depends on the significance level, the probabilities of H_0 and H_A being true as well as the power of the test (Colquhoun, 2014). Apparently, the significance level α captures only a part of the FDR, because it is the conditional probability of rejecting the null under the assumption that H_0 is true.

Finally, it has been stressed in the literature that $1 - p$ does not measure the probability of replicating an observed result. Gigerenzer (2018) provides a simple example to illustrate this "replication fallacy". If H_0 and H_A reflect the hypotheses that a dice is fair or loaded, respectively and two times "six" is observed, one would reject H_0, because the probability of this event under H_0 is $\frac{1}{36} =$

$0.03 < 0.05$. However, this does not imply that one can expect to observe two sixes in 97% of all future dice throws.

2.1.2 Erroneous applications of significance tests

Even if the notion of p-values is well understood by applied econometricians, several problems prevail that may invalidate the calculation of p-values and undermine conclusions that are derived from a statistical test. Here we focus on two issues that are highlighted in the current p-value debate, namely inference with data that do not constitute a (random) sample and multiple testing.

Multiple testing becomes an issue if several individual hypotheses are tested with the same data set (Romano, Shaikh and Wolf, 2010). If α is the desired significance level and m hypotheses H_i are tested, then the probability of getting at least one significant result by chance is:

$$Prob(\text{at least one significant result}) = 1 - Prob(\text{no significant result}) = 1 - (1-\alpha)^m \quad (1)$$

This probability, which depicts the familywise error rate (FWER), exceeds the significance level α considerably if m is large. Several proposals have been made to address this accumulation of type-I-error. These correction procedures control either the FWER (e.g. Bonferroni correction) or the FDR (e.g. Benjamini-Hochberg method). While a correction of significance levels is standard in biostatistics, particularly in genomic applications, it is often ignored in socioeconomics. This begs the question how relevant the consideration of multiple testing issues is in economic applications. Hirschauer, Mußhoff and Grüner (2018:p.137) argue that "*multiple testing is evident in multiple regression analysis whenever researchers independently perform and interpret more than one test on one data set*". Multiple testing can definitely lead to an inflation of "significant" results in explorative studies, where regression models are fed ad hoc with available data and p-values are scanned a-posteriori. If, however, the specification of multiple regression models is guided by economic theory, which is reflected by a set of predetermined hypotheses about the sign and the size of specific model parameters, no adjustment of significance levels is required. This holds a fortiori in situations where a single hypothesis is of particular interest and the inclusion of covariates is motivated by mitigating an omitted variable bias. Adjusting the significance level of the variable of interest would unnecessarily deteriorate statistical power (Albers, 2019).

A fundamental objection against statistical inference is raised by Hirschauer *et al.* (2020) in case of full population surveys. They argue that displaying p-values does not make sense, because there is nothing to infer, and sampling error does not exist. Obvious examples are studies that search for relationships among variables using data from all existing entities (e.g., individuals, states, countries) in a predefined population. However, it is not that clear to which situations this "urn model" applies and to which not. For example, in price analyses often data of all (available) transactions can be accessed that occurred in a specific market in a certain time period. Is it inappropriate to conduct statistical inference and hypothesis testing regarding price determinants using a full sample? The answer is "no",

at least if one can think of observed prices as an outcome of a data generating process. More data will be generated by this process in the future and even in the past more transactions could have been potentially observed. That means, the true population size is unknown and the "full sample" is still a random sample.

When inferring from observed realizations to the properties of the unknown data generation process by means of a statistical model, one has, of course, to consider selectivity issues and the fact that the data generating process can change over time – even though this can be rather challenging given the uncertainties involved. A related issue is the use of non-random samples for inferential reasoning. Non-random sampling techniques include convenience sampling, quota sampling or snowball sampling. These techniques became increasingly important and are nowadays quite common in survey-based social science. Several potential problems arise with non-random samples (cf. Elliott and Valliant, 2017). Selection bias occurs if the sample differs from the non-sample part of the population such that the sample cannot be projected to the population of interest. Another problem is attrition, i.e. the systematic drop-out of participants in a panel. There is a controversial discussion whether or not non-random samples should be used for inferential statistics. Hirschauer *et al.* (2019) argue that convenience sampling precludes the use of p-values because researchers run the risk of misestimating coefficients and standard errors, at least if selection bias is not adequately considered. In contrast, Smith (1983) and Elliott and Valliant (2017) show how quasi-randomization and superpopulation modeling can mitigate potential biases and under what assumptions non-random samples still can be used for statistical inference.

P-hacking or HARKing describe intentional or unintentional practices by researchers to adjust test procedures/model specifications, variables, data, or narratives to present statistically significant results with generally lower p-values. Researchers could only present tests or model specifications that have yielded statistically significant results, while not disclosing other tests or models they have used. The same applies to the transformation of dependent or independent variables or the removal of influential observations. Researchers could also explore the data and then retrofit theories, hypotheses, and narratives to findings after the results are known (HARKing).

A large share of researchers in environmental economics has admitted questionable research practices in a recent survey (Ferraro and Shukla, 2020), and the economic literature in major general interest journals appears biased towards false positive findings, as indicated by an unusual hump in the distribution of p-values around the common p-value threshold of 0.05 (Brodeur *et al.*, 2016). O'Boyle, Banks and Gonzalez-Mulé (2017) study PhD dissertations and subsequent research papers published from those dissertations and note that the "*the ratio of supported to unsupported hypotheses more than doubled*". While this may indicate p-hacking or HARKing, Huntington-Klein *et al.* (2021) further demonstrate a large variation in results if different teams analyze the same data.

2.2 *Proposed remedies*

While broad consensus about potential flaws of p-values seem to exist, opponents and proponents often disagree about remedies. In fact, proposals range from a complete ban of statistical hypotheses testing to a maintenance of current practice due to the lack of superior alternatives. In what follows, we structure these proposals and discuss their pros and cons.

Banning of significance testing and p-values

In view of the aforementioned concerns some authors suggest not to display p-values or asterisks (Hirschauer *et al.*, 2019) or even to completely retire the concept of statistical significance (Amrhein, Greenland and McShane, 2019; Gigerenzer, 2004), and some scientific journals followed these suggestions. This critical view, however, is also challenged: Verhulst (2016), Gelman (2016) and Benjamini (2016) demur that most concerns about p-values also apply to alternative methods. Fricker *et al.* (2019) try to assess the implications of a p-value ban empirically by analyzing the quality of 31 empirical papers published in "Basic and Applied Social Psychology" after this journal prohibited the use of the null hypothesis significance testing procedure (including p-values and statements about significance) in 2015. In their conclusions, the authors state "*we found multiple instances of results seemingly overstated beyond what data would support if p-values* […] *had been used. Thus, the ban seems to be allowing authors to make less substantiated claims* […]". At the time being, it appears unlikely that this radical approach will be copied by many scientific journals.

Emphasizing economic significance and relevance with a clear distinction from statistical significance

In two empirical studies investigating the statistical practice in the American Economic Review in the 1980s (McCloskey and Ziliak, 1996) and the 1990s (Ziliak and McCloskey, 2004), Deirdre McCloskey and Stephen Ziliak highlight the importance of interpreting research results in light of their real-world substance. They argue that economists do a poor job in distinguishing statistical significance (the uncertainty of an estimate) and economic significance (the size of an estimate). Among other things, they propose to use confidence intervals to gauge the plausibility of an estimate and to use simulations to explore a range of possible economic outcomes. In addition, they emphasize the role of power analysis and considering the implications of type II errors rather than solely focusing on type I error. Although the two authors witness some improvements over time (Ziliak and McCloskey, 2004) many problems prevail. As discussed by Rommel and Weltin (2021), similar problems are present in major agricultural economics journals.

Replacing p-values and use of Bayesian methods

The desire of researchers "*to turn a p-value into a statement about the truth of a null hypothesis*" (Wasserstein and Lazar, 2016) has prompted the promotion of a Bayesian approach, which, in principle,

is capable to combine a data likelihood and a prior probability to derive a posterior probability. This Bayesian posterior inference offers the intuitive interpretation of a probability that a parameter of interest falls into a certain range (conditional on model assumptions) alleviating the troubles with interpreting p-values and confidence intervals under the frequentist paradigm.[5] It also provides the possibility to leave the dichotomous world of classical hypothesis testing with all its problems laid out above by rather comparing hypotheses in a probabilistic manner[6] (Bendtsen, 2018).

There are probably two main reasons why the Bayesian approach has not yet overtaken the frequentist statistical inference despite an increasing use in recent times (Geweke, Koop and van Dijk, 2011). First, the derivation of posterior distributions of model parameters has long been a tedious and case-specific challenge requiring to derive posteriors via probability calculus and/or simulation-based analysis. Recent advances in automated Bayesian inference ("probabilistic programming", see van de Meent *et al.* (2018) and Bingham *et al.* (2019)) may offer a general solution in the medium-term for conventional and "big data" but this will also require a change in educating applied (ag-) economists. The second reason is the need to specify a prior distribution for all hypotheses and many scientists are reluctant to do so (Krueger and Heck, 2019) even though one could argue that frequentist approaches do this implicitly (e.g. Bendtsen, 2018). Against this backdrop, Harvey (2017) suggests the use of the minimum Bayes factor as a compromise that takes advantage of the Bayesian paradigm but bypasses the need to specify a particular alternative hypothesis and a full prior distribution. The Bayes factor is the ratio of the likelihood under H_0 and H_A, respectively. The minimum Bayes factor utilizes a special choice for the likelihood under H_A, namely the maximum likelihood given the data. If one is willing to express prior information as an odds ratio of the two hypotheses, one can derive a posterior odds ratio using Bayes' theorem:

$$\underbrace{\frac{p(H_0|Data)}{p(H_A|Data)}}_{\substack{posterior\\ odds\ ratio}} = \underbrace{\frac{p(Data|H_0)}{p(Data|H_A)}}_{\substack{(minimum)\\ Bayes\ factor}} \cdot \underbrace{\frac{p(H_0)}{p(H_A)}}_{\substack{prior\\ odds\ ratio}} \quad (2)$$

Based on this expression, Goodman (2001) and Harvey (2017) show how to derive "Bayesianized" p-values from the minimum Bayes factor, which provide the desired probability that a hypothesis is true.

[5] It is interesting to note that Ionides *et al.* (2017) interpret the ASA statement on p-values (Wasserstein and Lazar (2016) as an attempt to advocate the Bayesian paradigm and to discourage researchers from using frequentist inference and deductive reasoning.

[6] For issues debated among those using Bayesian statistical inference see Aczel *et al.* (2020).

Complementing p-values by additional information

Instead of banning or replacing p-values, a couple of proposals have been made to complement them while maintaining the general framework of statistical significance testing. This is in line with Amrhein, Korner-Nievergelt and Roth (2017), who conclude that "*apparently, bashing or banning p-values does not work. We need a smaller incremental step*…". Greenland *et al.* (2016) emphasize that statistical test should be interpreted carefully by examining effect sizes and confidence intervals instead of focusing just on p-values. Confidence intervals have the advantage of disentangling the size and the precision of an estimate that are merged in a p-value. Moreover, Gigerenzer (2018) reminds us that the design of insightful economic experiments requires sufficient statistical power. While power, effect sizes, loss functions, and type-II-errors are an integral part of the Neyman-Pearson theory, they are typically ignored in the NHST ritual. Button *et al.* (2013) show that in low powered studies the replicability of significant results is low. Furthermore, the positive predictive value (PPV), i.e. the probability that a "positive" research finding reflects a true effect, is positively linked to the statistical power of the study. Unfortunately, a meta-analysis conducted by Ioannidis, Stanley and Doucouliagos (2017) reveals that empirical economics research is often severely underpowered. However, at least in some research areas, particularly in experimental economics, power calculations started becoming a norm.

Multiverse analysis

Different research teams can come up with fundamentally different conclusions even when they are working with the same data and research questions (Huntington-Klein *et al.*, 2021). Another problem is that researchers may strategically report robustness tests if they support a preferred narrative (Young and Holsteen, 2017). Specification curves acknowledge this problem by running a wide range of plausible models that could for instance include different sets of covariates (Steegen *et al.*, 2016). The outcome is not a single p-value linked to a single estimate, but a distribution of plausible estimates and p-values that define a distribution of plausible results for a reasonable set of models (see chapter 7 of Christensen, Freese and Miguel, 2019 for more details).

Replication studies and meta-analysis

Statistically significant research results may be the outcome of chance. To detect false positive findings, researchers have advocated replication studies. Replication can take different forms (see chapter 9 of Christensen, Freese and Miguel, 2019). It may involve reanalyzing the original data of a study with the same (verification) or different (reanalysis) methods. It can also involve new data collection applying the same methods (direct replication) or different methods (extension). Direct replications of economic experiments show that the rate of false positives is substantially higher than expected from pure chance alone (e.g. Camerer *et al.*, 2016). Replication studies and the aggregation of studies for meta-analysis can increase the confidence in research findings, but a recent study has shown that studies that did not

replicate are more widely cited than those that replicate (Serra-Garcia and Gneezy, 2021), calling into question the use of citations as an indicator of scientific quality and the power of the research community to self-correct more generally.

Pre-registration, registered reports, and results-blind review

Other remedies target the scientific publication process. Pre-analysis plans are a written commitment to a specific data analysis before the data are obtained or collected (see Olken, 2015 for a detailed discussion). In a pre-registration, researchers also submit this plan to a repository, thereby increasing the commitment by making it publicly available and referring to the pre-registration in publications. Although these two instruments substantially limit researcher degrees of freedoms and may successfully safeguard against p-hacking, they only address the producers of research findings, whereas editors and reviewers could still exhibit bias against non-significant findings. Registered reports or results-blind review have been proposed as a solution to this problem. In a registered report, a study design and analysis plan are submitted to a journal and reviewed by peers in a "first stage report" before data collection. If the authors pass this stage, the journal and publisher commit to a publication irrespective of the results. Results-blind review mimics this process, by suppressing results from the manuscript, hence allowing reviewers and editors to focus on research questions and methodological rigor.

3 Views of the community

We conducted a survey among agricultural economists and social scientists in Germany, Austria, and Switzerland to explore the views of the German-speaking agricultural economics community in 2020. The survey was administered in English to address non-German speakers in the three countries. The general objective of the survey was to get an overview on the problem perceptions, knowledge, practices, and attitudes regarding econometrics and statistics. The survey started with a short introduction, data use, and contact information. Consent to participate was obtained. The first part of the survey covered perceptions of the debate on statistical practices and knowledge on the topic. The second part dealt with practices and preferred remedies. Finally, respondents had to provide some personal details. We refer to the appendix/online appendix to see the full survey, data, and code.

3.1 *Survey design and respondent characteristics*

The survey was distributed in the summer and fall 2020 to all members of the German Association of Agricultural Economists (GEWISOLA), members of the Swiss and Austrian associations of agricultural economists, an e-mail list of early career researchers in agricultural economics in Germany, and doctoral students enrolled in the Doctoral Certificate Program in Agricultural Economics. There is some overlap between these groups. We estimate that approximately one thousand people have been invited to participate in the survey by mail.

Different distribution links for these channels indicate that approximately 34% have entered the survey from the GEWISOLA invitation, 31% from the doctoral certificate program and 25% from the early career researchers e-mail lists. The remaining respondents came from the Austrian and Swiss societies or other sources.

In total, 305 respondents opened the link, but there was a high drop out on the first screens. We removed one response due to highly inconsistent responses. For the analysis, we use all respondents who completed the survey at least until the second last screen, leaving us with a total of 108 respondents. Note that there are still missing observations for some of the variables which could lead to a lower number of observations for some of the recorded items, as we did not force answers on any of the questions (i.e. all responses were voluntary). All presented analysis is descriptive and must be viewed as explorative, as it stems from a self-selected sample.

The median response time in the survey was approximately 14 minutes. Most respondents either had a PhD (55%) or were in the process of obtaining one (39%). Participants indicated their gender as male (61%), female (35%), or did not indicate a gender (4%). The average age was 38 years (with a range from 25 to 72 and a median of 34; SD = 10.4). Approximately 39% stated that they were permanently employed. More than half of the respondents had five years or less of research experience. A little more than half of the respondents stated that they had published three or less research peer-reviewed research articles over the past five years.

3.2 *Problem perception and knowledge*

The survey started with several general questions on the perception of the problem. We openly asked whether generally speaking respondents "think there are problems with the way the scientific community represented by the Austrian, German, Swiss, and European associations of agricultural economists (GEWISOLA, ÖGA, SGA, and EAAE) deals with statistics and econometrics in research and teaching?" Respondents were asked to use a ten-point scale to differentiate their responses (1 = no problems at all to 10 = a lot of problems). The mean response was 5.13 (SD = 2.14; median = 5). We also asked people to assess their own statistical and econometric knowledge on a ten-point scale where higher values indicate better knowledge (mean = 6.38; SD = 1.59; median = 7). Finally, we asked for an assessment in which percentile respondents would place themselves in terms of knowledge relative to the target community. The median respondent placed themselves in the top 50%.

We used the six survey items developed by Oaks (1986) to get an overview on knowledge of the correct interpretation of a p-value. These items have been widely applied to different samples of researchers (see Gigerenzer, 2018 for an overview of studies in different academic communities). Respondents were presented with the following scenario:

"Suppose you have an exogenous variation that you suspect may alter the outcome you are interested in for a certain task or behavior in a given population. You compare the means of your control and treatment groups (say 20 randomly selected 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."

Table 1 presents the six statements and displays responses. All of the statements are false and represent different delusions regarding the meaning of a p-value (Gigerenzer, 2018). Hence, the percentage of respondents endorsing a statement as true may be viewed as an indicator of knowledge. Approximately 80% of respondents who have responded to all six items endorse at least one of the delusions. Note that the number of correctly answered statements only weakly correlated with the item of self-assessed knowledge above (Spearman's rho = 0.1; n = 75).

Table 1: Overview on endorsement of statements

Statement	*Percentage of respondents wrongly endorsing the statement as true*
You have absolutely disproved the null hypothesis that there is no difference between the population means. (Illusion of certainty)	26.3% (n = 95)
You have found the probability of the null hypothesis being true. (Bayesian wishful thinking)	21.4% (n = 98)
You have absolutely proved your alternative hypothesis that there is a difference between the population means. (Illusion of certainty)	18.3% (n = 93)
You can deduce the probability of the alternative hypothesis being true. (Bayesian wishful thinking)	48.4% (n = 91)
You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision. (Bayesian wishful thinking)	57.6% (n = 92)
You have a reliable finding in the sense that if, hypothetically, the study was repeated a great number of times, you would obtain a significant result on 99% of occasions. (Replication fallacy)	51.2% (n = 86)
Percentage of respondents wrongly endorsing at least one statement (among those who responded to all statements)	81.3% (n = 75)

Notes: Adapted from Gigerenzer, 2018.

Source: Own calculations.

We also asked about knowledge of and experience with some of the remedies/practices outlined above (Table 2).

Table 2: Knowledge and applications of remedies

	Yes, and I have used one/it.	*Yes, but I have never used one/it.*	*I have heard about it, but I am not entirely sure.*	*Never heard about it before.*
Pre-analysis plan (n = 107)	20.6%	40.2%	23.4%	15.9%
Power analysis (n = 107)	20.6%	32.7%	19.6%	27.1%
Minimum Bayes factor (n = 107)	5.6%	18.7%	39.3%	36.5%
Standardized coefficient (n = 107)	64.5%	20.6%	9.4%	5.6%

Source: Own calculations.

3.3 *Practices and attitudes*

We asked people for their agreement with several survey items to understand attitudes and practices regarding statistics and econometrics (Table 3). There were high levels of agreement with the importance of economic significance and data sharing practices. At the same time, respondents stated that they feel pressured to produce statistically significant findings. Many stated they have committed or witnessed p-hacking.

Table 3: Overview on attitudes and practices

Statement	*Mean (SD)*	*1 = Strongly disagree*	*2*	*3*	*4*	*5 = Strongly agree*
The economic significance of an estimated effect is more important than the statistical significance. (n = 102)	3.5 (1.2)	10.8%	5.9%	28.4%	32.4%	22.6%
Only statistically significant findings should be published. (n = 106)	1.7 (1.1)	64.2%	18.9%	9.4%	0.9%	6.6%
I have often witnessed colleagues/other researchers to search for an econometric specification that produces statistically significant findings. (n = 102)	3.6 (1.2)	5.9%	11.8%	21.6%	35.3%	25.5%
I have often searched for an econometric specification that produces statistically significant findings. (n = 104)	3.0 (1.2)	15.4%	18.3%	25.0%	30.8%	10.6%
Statistically insignificant findings should not be discussed in publications. (n = 105)	1.7 (1.0)	59.1%	24.8%	10.5%	1.0%	4.8%
One should conduct many different analyses to find statistically significant findings. (n = 101)	2.4 (1.2)	25.8%	32.7%	21.8%	10.9%	8.9%
I feel pressured to produce statistically significant findings when I want to publish. (n = 103)	3.8 (1.3)	9.7%	5.8%	18.5%	30.1%	35.9%
If possible, research data for a publication should always be shared online by the authors. (n = 104)	3.7 (1.3)	6.7%	13.5%	19.2%	27.9%	32.7%

Source: Own calculations.

3.4 *Views on remedies and suggested fields of future action*

We asked about an assessment of how useful remedies were perceived (Table 4). There was a high perceived usefulness of confidence intervals, display of effect sizes and standardized effect size, as well

as summary statistics. There was little perceived usefulness in the overall ban of p-values or asterisks/stars from research results or publications.

Table 4: Attitudes on remedies

Remedy	*Mean (SD)*	*1 = Not at all useful*	*2 = Somewhat useful*	*3 = Fairly useful*	*4 = Very useful*
Abandon p-values (n = 90)	1.7 (0.9)	50%	33.3%	12.2%	4.4%
Abandon stars/asterisks (n = 93)	2.2 (1.0)	32.3%	25.8%	29.0%	12.9%
Display confidence intervals (n = 97)	3.1 (0.8)	2.1%	20.6%	38.1%	39.2%
Display effect sizes (n = 93)	3.4 (0.7)	1.1%	8.6%	41.9%	48.4%
Display standardized effect sizes (n = 78)	3.3 (0.8)	1.3%	16.7%	37.2%	44.9%
Power analysis before data analysis (n = 62)	2.9 (0.9)	4.8%	32.3%	35.5%	27.4%
Pre-registration (n = 93)	3 (0.9)	6.5%	23.7%	33.3%	36.6%
Minimum Bayes factor (n = 32)	2.2 (0.8)	15.6%	56.3%	18.8%	9.4%
Full Bayesian analysis (n = 33)	2.4 (0.9)	15.2%	39.4%	36.4%	9.1%
Mandatory code sharing (n = 95)	2.9 (1.1)	10.5%	28.4%	20.0%	41.1%
Mandatory data sharing (n = 98)	2.8 (1.1)	13.3%	28.6%	23.5%	34.7%
Mandatory summary statistics (n = 101)	3.3 (0.8)	3.0%	13.9%	28.7%	54.5%

Source: Own calculations.

To identify target areas and fields of action we asked respondents to state who would have the largest impact on one's statistical and econometric practice (Table 5). Respondents assigned high importance to colleagues, teachers and educators, as well as reviewers as drivers for their own statistical practice.

Table 5: Who affects statistical practice the most

Group/person	*Mean (SD)*	*1 = no impact at all*	*2*	*3*	*4*	*5 = a lot of impact*
Editors (n = 92)	2.8 (1.3)	16.3%	32.6%	15.2%	22.8%	13.0%
Reviewers (n = 96)	3.8 (1.1)	2.1%	12.5%	17.7%	38.5%	29.2%
Teachers/educators (n = 100)	3.8 (1.2)	6.0%	8.0%	17.0%	34.0%	35.0%
Universities/research organizations (n = 99)	3.1 (1.3)	11.1%	26.3%	22.2%	25.3%	15.2%
Colleagues/other researchers (n = 100)	4.0 (1.0)	2.9%	5.8%	17.5%	40.8%	33.0%

Source: Own calculations.

3.5 *Discussion of the survey results*

The survey results reveal that knowledge on the interpretation of statistical hypothesis testing and p-values, and the potential remedies of current malpractices may still need a substantial educational push at various levels. At the same time, the community feels fairly strong about not abandoning p-values altogether (50% consider this remedy "not at all useful"). The dichotomous nature of the current practice in hypothesis testing is seen somewhat more critical (a clear majority considers abandoning the use of stars/asterisks at least "somewhat useful").

At least a 70% majority of respondents view certain practices offering information beyond the pure outcome of hypotheses tests and that are not yet widely applied at least as "fairly useful". These include those that allow better understanding or visualizing uncertainty of statistical results (display of confidence intervals) and understanding better the (relative) economic importance of the determinants considered (standardized coefficients and economic effect sizes).

To the extent that these remedies are known, respondents consider power analysis (n = 62) and pre-registration plans (n = 93) at least "fairly useful" with a majority larger than 60%. Hidden behind these responses might be a differentiated view on the question for what type of analysis such remedies are useful. They are discussed and implemented in the context of controlled experiments, where sample size and treatments are often part of the deductive analytical design decided upon before the data collection. Also, in these cases good priors are often available. Pre-registration may in principle also be considered for observational or even explorative studies to prevent that the research design is driven by initial results in a not fully reflected empiricist manner (cf. Haven and van Grootel, 2019; Olken, 2015).

Only a rather small share of survey participants feels comfortable to judge Bayesian remedies (Minimum Bayes Factor and full Bayesian analysis with n = 32 and 33, respectively), but those who do are moderately positive about them (above 70% consider them somewhat or fairly useful). Whereas the

low level of participation in these questions may indicate a limited amount of training and experience with Bayesian analysis, the moderation in viewing the positive contribution could be explained with the discussed effort still needed to develop case-specific Bayesian approaches and their limitations in providing a full alternative to classical hypothesis testing—at least as long as one considers dichotomous test outcomes as relevant.

The community feels quite strongly about the usefulness of mandatory data sharing, code sharing and summary statistics with majorities larger than 80% considering them at least "somewhat useful". More than a third consider data and code sharing and more than half summary statistics "very useful". The comparatively moderate responses regarding the data sharing may reflect the not uncommon situation that confidentiality requirements of individual firm and consumer level data often restrict the possibilities to share data. The reason why data summary statistics do not even have a stronger support may lie in the view that it alone does not help to solve the statistical inference issues of the framing even if more respondents may view it as a key ingredient to a transparent and "data-aware" empirical economic analysis. Suggested remedies in this area probably require some context-dependent qualifications and a connection to "mindful statistics" (see below) that go beyond the specific solution of the debated issues in inference practices.

Respondents clearly consider teachers/educators to have the largest impact on statistical practices (almost 70% in the top two impact categories). This coincides with the in parts limited knowledge on some statistical misinterpretations and malpractices found above and points at the longer-term effort needed to fundamentally change practices through a revision of curricula and teaching methods.

It is quite interesting to note that respondents view reviewers as affecting statistical practices more (almost 60% in the top two impact categories) compared to editors (less than 40%). It raises the question if editors of the journals relevant for the community are rather passive with respect to setting and guarding editorial policies on statistical practices and/or often shy away from evaluating/adjusting/weighing/complementing reviewer comments with respect to the editorial standards in this respect. Editors could have a crucial role in changing statistical inference practices if they took an active stance on it.

4 Implications for the community

4.1 *Research and researchers*

The most important implication resulting from the p-value debate from the viewpoint of researchers is to avoid what Gigerenzer (2004) blamed as "mindless statistics". Many fallacies may arise when applying the statistical hypothesis testing framework; we argue that the merit of the p-value debate is to recall (at least some) potential fallacies to researchers' minds. Being aware of these problems is in fact the most undisputable implication. Yet we are reluctant to recommend a ban of p-values or the use of

asterisks in general. This view is shared by the majority of our survey participants. Correctly calculated and interpreted p-values contain useful information about the underlying statistical hypotheses that otherwise would be neglected. In a recent paper, Nobel Prize laureate Guido Imbens characterizes economic applications where p-values are dispensable and where they contain relevant information (Imbens, 2021). Testing a null hypothesis versus an alternative hypothesis is meaningful in some situations and examples include the efficient market hypothesis, market integration or the existence of speculative bubbles. Moreover, it is often necessary to test whether data show certain statistical properties, such as stationarity, variance homogeneity or spatial and temporal independence. In these situations, a decision shall be made based on a statistical decision rule. This then necessarily includes a threshold determining what the decision will be.

In many economic applications, however, testing against a null hypothesis of "no effect" is not of particular interest. For example, it is not exciting to test whether farmers' education increases farm income or not, whether a gender pay gap exists or not or whether investment aid stimulates investment demand or not. Here the magnitude of the (treatment) effect is what matters and the causal mechanism, e.g. how investment aid stimulates investments. We believe that in situations, where no specific decision on a hypothesis has to be made, it suffices to display standard errors or to interpret p-values as indicators of the general compatibility of the data with the corresponding hypothesis. In these cases, specific thresholds have no defendable meaning beyond a long-practiced ritual. Given the documented publication bias around these thresholds (e.g. Brodeur *et al.*, 2016), avoiding the use of asterisks potentially reduces incentives for p-hacking. Whether with or without specific significance levels, the important point is, however, that we as researchers derive hypotheses based on logical thinking and theories, and apply statistical analysis "mindfully" in light of an underlying theoretical concept, and to avoid extreme forms of empiricism.

In situations where statistical hypothesis testing makes sense, the following aspects deserve attention when designing, conducting and interpreting statistical tests. Perhaps the most basic question is whether observed data can be considered as a random sample, i.e. as an outcome of a random data generating process, because this is a prerequisite for inferential statistics. If, in contrast, data fully describe the entire population, there is no need for statistical testing. If in this instance, inferential reasoning is based on the notion of a superpopulation, this should be clearly labelled and defined. Moreover, if data come from convenience sample, any source of potential bias regarding estimates of regression coefficients and standard errors should be carefully considered and discussed.

From the stated objectives of an empirical analysis it should be clear whether a study is explorative or whether it aims at testing of hypotheses that are derived from theory. This distinction is important, because in explorative studies that try to identify potential relationships among dependent and explanatory variables, a multiple testing problem is immanent that calls for an adjustment of significance levels to avoid false rejections of null hypotheses. Unfortunately, this distinction is not always

straightforward in applied economics, because theoretical predictions do not cover all aspects of econometric model specification. That is, even if theory suggests a positive or negative relationship among economic variables, it might be necessary to "explore" the appropriate functional form in a regression model or the number of lags in a time series model (Olken, 2015). We do not consider this search for a data fitting model specification per se as "p-hacking". The crucial point is to describe this process in a transparent manner and to report the results of alternative model specifications instead of presenting only selected results. In these instances, careful documentation of data and code, as well as tools such as multiverse analysis, may address selective reporting more appropriately (Steegen *et al.*, 2016).

The need for flexibility during the model specification process limits the scope of instruments that have been proposed to prevent p-hacking in some instances, e.g. pre-registration. Recent studies show that researchers who use pre-registration rarely specify pre-analysis in sufficient detail (Bakker *et al.*, 2020). In other instances, there may be a risk that pre-analysis plans limit the reporting of relevant findings (Banerjee *et al.*, 2020). The effectiveness of pre-registration is also sensitive to the platform used (Bakker *et al.*, 2020). Yet, perceived benefits from pre-registration outweigh the costs in many instances, and major benefits emerge from thinking about analysis before the data are collected (Logg and Dorison, 2021). Therefore, pre-analysis plans need to enter PhD- and third-party funded project plans and output/performance measures, as pre-registration/pre-analysis plans are resource-consuming. Current schemes of performance measures of universities and researchers seem not to value such efforts and at first glance, per se not to outweigh additional costs and efforts related to pre-analysis/pre-registration. In conclusion, pre-registration and pre-analysis plans can be a useful tool in many fields that involve primary data collection, while the risk that pre-registration becomes ritualized and a form of virtue signaling if not complemented by more fundamental cultural change (Buck, 2021) exists.

Another important insight of the discussion about verification, re-analysis, and aggregation of scientific research is the need to pay more attention to adequate power of statistical tests. This is important to avoid "false negatives" but also to ensure a high positive predictive value, i.e. the likelihood that a claimed relationship is actually true (Christensen and Miguel, 2018). Researchers have at least two options to control statistical power. First, via sample size which can be determined for a desired power level in an a priori power analysis, given that information about the effect size is available, e.g. from pilot projects or similar studies (Ioannidis, Stanley and Doucouliagos, 2017). Computational software is available that supports this calculation for many research designs, e.g. G*Power (Faul *et al.*, 2007). The second option is the choice of the statistical test. In time series analyses, for example, the use of panel unit root tests can help improving power compared with standard unit roots tests, which are known to have low power.

Regarding the interpretation of statistical test results, two recommendations appear unchallenged. First, presentation of statistical results should include effect sizes, and the interpretation should involve

the economic relevance of variables rather than focusing solely on their statistical significance. Coefficient plots along with marginal effects discussion may for instance support this way. Second, p-values should be interpreted as what they are, the likelihood for observed data given a null hypothesis, though it is tempting to consider them incorrectly as evidence in favor of or against a hypothesis. Hirschauer *et al.* (2016) provide an illustrative example of how the use of sloppy language turns a statistically correct statement into a wrong one. We thus strongly recommend to use precise wording when interpreting the results of statistical hypotheses tests, along with careful documentation of the test procedure. Our survey showed strong support for confidence intervals and descriptive statistics, and authors and journals may consider them even more. Although confidence intervals are easily calculated from standard errors and coefficient estimates, displaying them may change the reader's perspective and provide an additional incentive to leave the purely dichotomous interpretation of results within a cultural change.

4.2 *Editors, referees, journals and the publication process*

Journals can achieve a lot through submission guidelines, which should be up to date and enforced. For instance, check lists on how to report statistics and results of statistical testing may be useful and can have an impact (see Giofrè *et al.*, 2017); some authors even call for "statistical co-editors" (Wehrden, Schultner and Abson, 2015). A prerequisite for any change to the better is, however, that all involved stakeholders are clear in their communication, reach their audience effectively and editors take responsibility to moderate reviews carefully and decide according to clearly communicated rules.

Our survey showed support for open data and methods, which could be supported by making code and data sharing mandatory. Data and code sharing do not only increase transparency of results, they also make it easier to discover data manipulations and in turn, researchers will become more reluctant to violate good research practice. This in turn, will improve quality and reproducibility of the results. Clearly, relevant journals in a field should pursue similar policies in this regard to avoid a selection of authors into journals with less restrictive policies. In some cases, sharing of raw data may be hampered by data protection regulations. This applies to farm level data, such as data from the Farm Accountancy Data Network (FADN); here other ways of reproducing the results need to be made available, for instance, by remote access. Moreover, if data are bought from and owned by third parties, researchers cannot easily share them, yet also here, replicability can be made available by remote solutions together with third parties; owners or providers of data sets are expected to be interested in most reliable results produced with the data. These additional efforts are again resource-consuming and could be alleviated if raw data collected by the public (e.g. FADN data) would by generally accessible in anonymized form

for scientific research institutions. In turn, all scientific institutions should commit to FAIR principles[7] for research data, and universities should collaborate for efficient research data management processes which would benefit the whole community.

Researchers have highlighted problems with the direct replicability of research results especially in experimental economics and business economics' studies, and the sensitivity of research results to context (Camerer *et al.*, 2016; Rahwan, Yoeli and Fasolo, 2019). When engaging in a replication, authors bear major publication risks when editors predominately select manuscripts on novelty. New publication formats could lower these risks. In a recent call for papers in the journal Ap*plied Economic Perspectives and Policy* (AAEA, 2021), the editors invite replications in a two-stage format. Replication protocols are reviewed *before* the bulk of the work is done, and the journal and editors commit to a conditional acceptance for publication for the selected proposals (or reject proposals). Adopting this format on a regular basis either in the form of special issues or a new publication format could give rise to more replication attempts, as authors can substantially lower their risks of engaging in replication.

Registered reports—a two-stage publication format where the study design is reviewed *before* the data collection (see Lemken, 2021 for a recent example of a first stage report in agribusiness consumer research)—could be embraced by more journals in the agricultural and food economics domain. Whereas a pre-registration only involves the authors, a registered report is integrated with the peer review and journal publication process. Hence, with a registered report, several important steps of the research and publication process are front-loaded, potentially reducing risks for authors and the research community in several important ways. Authors will benefit from feedback on their work already in the design stage. Other researchers become aware of what others are working on earlier, facilitating collaboration and innovation. Editors and reviewers evaluate studies on novelty and a sound research design, rather than results. In the future, research funders may even condition grants and research funds on the acceptance of registered reports for studies that involve primary data collection. As of today, only a few journals in which agricultural economists publish offer register reports (PLOS ONE, Nature Human Behavior, Journal of Development Economics, Q Open), and more journals and editors may want to consider opening up for the format. We encourage editorial boards and scholarly associations to discuss this option. Pre-registration can also be applied to some types of explorative and qualitative research, but it will be critical to adjust platforms such as the open science framework to the specific needs of the qualitative research community (Haven and van Grootel, 2019).

[7] FAIR principles for research data target at a sustainable data collection, processing and use. F stands for findable, where meta data should be made available, A indicates accessible, where meta data must be available, I stands for interoperable, i.e. clearly documented and applicable language, and lastly, R means re-usable, i.e. a clear data use agreement/license is required. For Germany, more details can be found for instance here: https://www.forschungsdaten.org

4.3 *Teachers and teaching*

As the solution to the crisis includes mindful use and practice of hypothesis testing and other statistical methods to gain knowledge, and to contribute event to a "regime shift" or "cultural change", this implies taking a long-term perspective and to go beyond the above discussed remedies. To make these suggestions the new norm, we argue that the p-value debate offers several lessons in the field of applied agricultural, resource and food economics for teaching research methods. We see teaching at all levels as the key to educate the next generation of researchers. This in turn calls for open mindedness of all active researchers as teachers for the need to change education and teaching methods to reconsider current ways to teaching statistics and empirical research methods in agricultural economics.

Higher education in agricultural economics typically rests on an interdisciplinary curriculum with specific modules covering methods for empirical research and scientific working. Against the debate, we see a specific strand of the curriculum to impart a sound understanding of empirical research, including hypothesis testing, as a qualification for higher education covering all levels: Bachelor, Master and PhD.

Teaching methods for empirical research at Bachelor- and Master studies must impart sustainable knowledge on research methods, must qualify students that they are able to apply, and critically reflect existing methods/practices such that they are prepared for their theses and PhD studies. To achieve these objectives, first, we suggest to offer modules for quantitative methods that provide a clear understanding of empirical methods and different ways of hypothesis testing, including statistical inference. Here we see working with simulated data sets (Bekkerman, 2015) and calculating test statistics "by hand" as core to understand the idea of statistical inference.

Second, we suggest to present examples strongly related to topics and research in the agricultural economics domain, i.e. to go beyond a "plain" method lecture. This could be achieved directly in the method-modules or in other modules that rely on empirical findings. For instance, discussions of research designs, data sets, methods for data collection, empirical hypothesis testing based on the research question and the empirical model using specific and interesting examples can raise attention and stimulate critical reflection. Thereby, we suggest to enhance lectures and offer modules where application and practice of methods by students have a strong emphasis. Beyond pure assignments, we suggest (poster) presentations and short reports about the data work; at higher levels (advanced Master studies and PhD), these may include critical reflection of existing research and presenting best practice examples.

Third, linked to the modules covering empirical methods, we see modules about scientific practices and scholarship as another pillar in higher education. These modules must include how to work transparently, ethics, research design, data collection and documentation, differences between theoretical/behavioral models and empirical models, empirical identification strategies, and how to

distinguish between mindful and not so mindful empirical work based on a sound understanding of philosophy of science (falsification). To foster best storage in memory and train students' behavior for empirical research, we see experiential learning as a fruitful guide.

For instance, do's and don'ts in a sense of a checklist for orientation for the theses could be part of student work in these modules with continuous update and monitoring by teachers. Material for standards for empirical work, data handling/management and ethics must be provided, while students discuss the material and prepare/develop and update their checklists. At higher levels, the student work may include contrasting examples based on empirical papers as well as critically assessing and discussing the procedures. Recently established asynchronous teaching could increase contact time with students at their home university with a focus on specific problem sets, critical reflections and presentation of own work.

We see the goal to enable students to strengthen their ability to critically reflect on their choice of method not only for their research and as authors but also as future reviewers and editors as core. This calls for interactive modules about empirical methods supported by Wikis and forums. Again, we recommend to link these modules to modules that cover philosophy of science, scientific working and writing as the good scientific practice and "standards" including research data management, pre-registration issues and ethics need to take up more room in the curriculum.

4.4 *Cultural change is needed*

Changing statistical practices is a challenge as they develop in a complex dynamic interplay of what we have been taught and what experiences we make interacting with our peers in publication processes and collaborative research as we build our career. Developed rituals are not easily changed, and such change requires a new consciousness to slowly penetrate all our academic activities. A long-term, cultural change of knowledge and norms is needed with complementary changes in teaching, research, and publishing activities that go beyond the definition of rules for use and interpretation of specific statistical tools.

New rules and recommendations to use some statistical tools and not others will not alone ensure that research is conducted and papers written to primarily generate replicable scientific knowledge. The currently observed misuse relates to a considerable extent to the explicit or implicit expectation that certain findings are more interesting than others. What is required is a culture of acceptance of scientific work that is largely based on the theoretical and methodological rigor and where the perceived relevance arises from the questions asked, the methodology employed and the data used but not from the results generated.

The quality of statistical analysis in economics falls and rises with the careful argumentation backed up by theories from the field of economics, social sciences and psychology, and related subjects that

govern agents' decisions and respective results. A discussion of most likely mechanisms underlying the data generation process guards against pure empiricist interpretations of statistical results and the confusion of correlation in the data with "true" effects or causation (Angrist and Pischke, 2008). With a sound theoretical foundation, the conditionality of statistical results on the model employed in the analysis becomes transparent and thereby creates an inherent caution with respect to the interpretation of results. Some even argue that "both statistical foundations and basic statistics can and should be taught using formal causal models" (Greenland, 2020). Thinking carefully about "what matters" for economics actors will also help in recognizing that a dichotomous world of hypothesis testing is not sufficient to derive meaningful implications. The size of effects of policies or other determinants of economic behavior matter for stakeholders and should receive at least as much attention as the question if there is an effect or not.

5 Concluding remarks

We like to conclude our paper with a few brief ideas on what could be done at the "policy level" to improve the situation in the short-term and to foster a cultural change of statistical inference and research practice in the long-term. Here we suggest a set of "top-down" measures that have some promise in bringing about the needed change jointly with the desirable "bottom-up" developments at the individual scientists' level.

Communication on best practices can clearly move forward right away. Here, scientific journals and connected learned societies can work together. Recent discussions and activities seem to lead towards a closer relationship between the GEWISOLA and the German Journal of Agricultural Economics (GJAE). A joint activity between the association and the journal can lead to setting standards of reporting statistical inference in journal articles that are then clearly communicated with the instructions to authors for the preparation of manuscripts and by the association to its members moving the community of reviewers. Hirschauer (2021) suggests guidelines that might serve as a starting point for the discussion on the formulation of such standards.

Better recognition of the effort reviewers put into the publication process may go some way in alerting to the value of this resource scarce in quantity and quality. Some journals like the European Review of Agricultural Economics already have a best referee award, which also could be picked up by the association in collaboration with the GJAE. Choosing criteria for awarding these prices wisely and making them transparent may offer another piece to making community members more aware of some remedies for the current replicability crisis. One could additionally consider awards to authors for outstanding transparency and excellent communication regarding data and statistical analysis.

To allow for better statistical inference from a sample to a population, researchers should be put in a position where they can draw random samples from a population. Often this is not a simple task, especially if farmers are involved. Making registry data more widely and more openly available for

research purposes would be an important task for the future. Alternatively, a farmer panel, similar to the socio-economic panel, could be maintained as a critical research infrastructure in the GEWISOLA field.

Support by the community for revising the teaching curricula and methods could be to establish a central pool of teaching examples for experiential learning and assignments in the domain of the community. Associations such as the GEWISOLA or the EAAE could provide the infrastructure and incentivize investments of teachers into such modules to foster sharing materials that offers clear guidance on good scientific practices, including hypothesis testing and mindful statistics, transparency in data, code and writing. Replication studies could be incentivized also for teaching purposes by journals and publishers to overall foster a longer-term change of the social norms governing our practices.

The ideas mentioned here are certainly not exhaustive and may be complemented as we go along this process of change. Perhaps it would be helpful to have one agenda element on the issue of statistical and/or scientific practice in each annual meeting of the GEWISOLA in the coming years, actively solicited by those responsible for the program and nudged by the association. They can have different formats – presentation on current developments, workshop, organized session, best practice updates – depending on what currently concerns the members or more generally the scientific community. Perhaps a future stronger liaison between the GJAE and the association can help to identify a person responsible to keep this on the agenda.

References

AAEA (2021). Call for Papers for a Special Issue on 'Replications in Agricultural Economics' in Applied Economic Perspectives and Policy. http://blog.aaea.org/2020/09/call-for-papers-for-special-issue-on.html, Accessed October 28, 2021.

Aczel, B., Hoekstra, R., Gelman, A., Wagenmakers, E.-J., Klugkist, I. G., Rouder, J. N., Vandekerckhove, J., Lee, M. D., Morey, R. D., Vanpaemel, W., Dienes, Z. and van Ravenzwaaij, D. (2020). Discussion points for Bayesian inference. *Nature Human Behaviour* 4(6): 561–563.

Albers, C. (2019). The problem with unadjusted multiple and sequential statistical testing. *Nature Communications* 10(1): 1921.

Altman, D. G. and Bland, J. M. (1995). Absence of evidence is not evidence of absence. *BMJ* 311(7003): 485.

Amrhein, V., Greenland, S. and McShane, B. (2019). Scientists rise up against statistical significance. *Nature* 567(7748): 305–307.

Amrhein, 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.

Angrist, J. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*.

Bakker, M., Veldkamp, C. L. S., van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D. and Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology* 18(12): e3000937.

Banerjee, A., Duflo, E., Finkelstein, A., Katz, L., Olken, B. and Sautmann, A. (2020). In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plans for RCTs in Economics. *National Bureau of Economic Research* (Working Paper 26993).

Bekkerman, A. (2015). The role of simulations in econometrics pedagogy. *Wiley Interdisciplinary Reviews: Computational Statistics* 7(2): 160–165.

Bendtsen, M. (2018). A Gentle Introduction to the Comparison Between Null Hypothesis Testing and Bayesian Analysis: Reanalysis of Two Randomized Controlled Trials. *Journal of Medical Internet Research* 20(10): e10873.

Benjamini, Y. (2016). It's not the p-values' fault. *The American Statistician, Online Discussion* 70: 1–2.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P. and Goodman, N. D. (2019). Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research* 20(1): 973–978.

Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y. (2016). Star Wars: The Empirics Strike Back. *American Economic Journal: Applied Economics* 8(1): 1–32.

Buck, S. (2021). Beware performative reproducibility. *Nature* 595(7866): 151.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J. and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14(5): 365–376.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M. and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science* 351(6280): 1433–1436.

Christensen, G., Freese, J. and Miguel, E. (2019). *Transparent and Reproducible Social Science Research: How to Do Open Science*. University of California Press.

Christensen, G. and Miguel, E. (2018). Transparency, Reproducibility, and the Credibility of Economics Research. *Journal of Economic Literature* 56(3): 920–980.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science* 1(3): 140216.

Elliott, M. R. and Valliant, R. (2017). Inference for Nonprobability Samples. *Statistical Science* 32(2): 249–264.

Faul, F., Erdfelder, E., Lang, A.-G. and Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39(2): 175–191.

Ferraro, P. J. and Shukla, P. (2020). Feature—Is a Replicability Crisis on the Horizon for Environmental and Resource Economics? *Review of Environmental Economics and Policy* 14(2): 339–351.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fricker, R. D., Burke, K., Han, X. and Woodall, W. H. (2019). Assessing the Statistical Analyses Used in Basic and Applied Social Psychology After Their p -Value Ban. *The American Statistician* 73(sup1): 374–384.

Gelman, A. (2016). The Problems With P-Values are not Just With P-Values. *The American Statistician, Online Discussion*.

Gelman, A. and Carlin, J. (2017). Some Natural Solutions to the p -Value Communication Problem—and Why They Won't Work. *Journal of the American Statistical Association* 112(519): 899–901.

Geweke, J., Koop, G. and van Dijk, H. (2011). Introduction. In *The Oxford Handbook of Bayesian Econometrics*, 1–8.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics* 33(5): 587–606.

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

Giofrè, D., Cumming, G., Fresc, L., Boedker, I. and Tressoldi, P. (2017). The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices. *PLOS ONE* 12(4): e0175583.

Goodman, S. N. (2001). Of P-Values and Bayes: A Modest Proposal. *Epidemiology* 12(3): 295.

Greenland, S. (2019). Valid P -Values Behave Exactly as They Should: Some Misleading Criticisms of P -Values and Their Resolution With S -Values. *The American Statistician* 73(sup1): 106–114.

Greenland, S. (2021). The causal foundations of applied probability and statistics. In: Dechter, R., Halpern, J., and Geffner, H., eds. Probabilistic and Causal Inference: The Works of Judea Pearl. ACM books.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., 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(4): 337–350.

Harvey, C. (2017). Presidential Address: The Scientific Outlook in Financial Economics. *The Journal of Finance* 72(4): 1399–1440.

Haven, T. L. and van Grootel, L. (2019). Preregistering qualitative research. *Accountability in Research* 26(3): 229–244.

Hirschauer, N. (2021). *The debate on p-values and statistical inference: What are the consequences for our community? Problems and solutions in statistical practice*. http://www.ilr.uni-bonn.de/agpo/publ/dispap/download/Hirschauer_2021.pdf

Hirschauer, N., Grüner, S., Mußhoff, O., Becker, C. (2018): Pitfalls of significance testing and p-value variability: An econometrics perspective. Statistics Surveys 12: 136-172

Hirschauer, N., Grüner, S., Mußhoff, O., Becker, C. (2019): Twenty steps towards an adequate inferential interpretation of p-values in econometrics. Journal of Economics and Statistics 239(4): 703-721

Hirschauer, N., Grüner, S., Mußhoff, O. and Becker, C. (2021). A Primer on p-Value Thresholds and α-Levels – Two Different Kettles of Fish. *German Journal of Agricultural Economics* 70(2): 123–133.

Hirschauer, N., Grüner, S., Mußhoff, O., Becker, C. and Jantsch, A. (2020). Can p-values be meaningfully interpreted without random sampling? *Statistics Surveys* 14: 71–91.

Hirschauer, N., Sven, G., Musshoff, O., Ulrich, F., Insa, T. and Peter, W. (2016). Die Interpretation des p-Wertes – Grundsätzliche Missverständnisse. *Journal of Economics and Statistics (Jahrbuecher fuer Nationaloekonomie und Statistik)* 236(5): 557–575.

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J. R., Burli, P., Chen, N., Grieco, P., Ekpe, G., Pugatch, T., Saavedra, M. and Stopnitzky, Y. (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry* 59(3): 944–960.

Imbens, G. W. (2021). Statistical Significance, p -Values, and the Reporting of Uncertainty. *Journal of Economic Perspectives* 35(3): 157–174.

Ioannidis, J. P. A., Stanley, T. D. and Doucouliagos, H. (2017). The Power of Bias in Economics Research. *The Economic Journal* 127(605): F236-F265.

Ionides, E. L., Giessing, A., Ritov, Y. and Page, S. E. (2017). Response to the ASA's Statement on p -Values: Context, Process, and Purpose. *The American Statistician* 71(1): 88–89.

Krueger, J. I. and Heck, P. R. (2019). Putting the P -Value in its Place. *The American Statistician* 73(sup1): 122–128.

Lemken, D. (2021). The price penalty for red meat substitutes in popular dishes and the diversity in substitution. *PLOS ONE* 16(6): e0252675.

Logg, J. M. and Dorison, C. A. (2021). Pre-registration: Weighing costs and benefits for researchers. *Organizational Behavior and Human Decision Processes* 167: 18–27.

Loken, E. and Gelman, A. (2017). Measurement error and the replication crisis. *Science* 355(6325): 584–585.

McCloskey, D. N. and Ziliak, S. T. (1996). The Standard Error of Regressions. *Journal of Economic Literature* 34(1): 97–114.

Mervis, J. (2014). Research Transparency. Why null results rarely see the light of day. *Science* 345(6200): 992.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231(694-706): 289–337.

O'Boyle, E. H., Banks, G. C. and Gonzalez-Mulé, E. (2017). The Chrysalis Effect. *Journal of Management* 43(2): 376–399.

Oaks, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

Olken, B. A. (2015). Promises and Perils of Pre-Analysis Plans. *Journal of Economic Perspectives* 29(3): 61–80.

Rahwan, Z., Yoeli, E. and Fasolo, B. (2019). Heterogeneity in banker culture and its influence on dishonesty. *Nature* 575(7782): 345–349.

Romano, J. P., Shaikh, A. M. and Wolf, M. (2010). Multiple Testing. *The New Palgrave Dictionary of Economics* 4.

Rommel, J. and Weltin, M. (2021). Is There a Cult of Statistical Significance in Agricultural Economics? *Applied Economic Perspectives and Policy* 43(3): 1176–1191.

Schooler, J. W. (2014). Metascience could rescue the 'replication crisis'. *Nature* 515(7525): 9.

Serra-Garcia, M. and Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances* 7(21).

Smith, T. M. F. (1983). On the Validity of Inferences from Non-random Sample. *Journal of the Royal Statistical Society. Series A (General)* 146(4): 394.

Steegen, S., Tuerlinckx, F., Gelman, A. and Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on psychological science : a journal of the Association for Psychological Science* 11(5): 702–712.

van de Meent, J.-W., Paige, B., Yang, H. and Wood, F. (2018). An Introduction to Probabilistic Programming.

Verhulst, B. (2016). In Defense of P Values. *AANA journal* 84(5): 305–308.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. *The American Statistician* 70(2): 129–133.

Wasserstein, R. L., Schirm, A. L. and Lazar, N. A. (2019). Moving to a World Beyond " p < 0.05". *The American Statistician* 73(sup1): 1–19.

Wehrden, H. von, Schultner, J. and Abson, D. J. (2015). A call for statistical editors in ecology. *Trends in Ecology and Evolution* 30(6): 293–294.

Young, C. and Holsteen, K. (2017). Model Uncertainty and Robustness. *Sociological Methods & Research* 46(1): 3–40.

Ziliak, S. and McCloskey, D. (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor, MI: University of Michigan Press.

Ziliak, S. T. and McCloskey, D. N. (2004). Size matters: the standard error of regressions in the American Economic Review. *The Journal of Socio-Economics* 33(5): 527–546.