Beyond P-Value-Obsession: When are Statistical Hypothesis Tests Required and Appropriate?

Anne Margarian
Thünen Institute of Market Analysis, Braunschweig

Abstract

Complementing more specific "p-value discussions", this paper presents fundamental arguments for when null hypothesis statistical significance tests (NHST) are required and appropriate. The arguments, which are paradigmatic rather than technical, are operationalised and broken down to the extent that their logic can be mapped into a decision tree for the use of NHST. We derive a perspective that does not ban p-values but proposes to minimize their use. P-values will become rather rare in (agricultural) economics if they are not applied in any cases, where the conditions for their proper use are violated or where their use is not appropriate or required in order to answer the questions asked of the data. The accompanying shift from prioritising inferential statistics to recognising the value of descriptive statistics requires not only a change in entrenched habits of thought. This shift also has the potential to trigger changes in the research processes and in the evaluation of new approaches within the disciplines.

Keywords

p-value discussion; statistical inference; null hypothesis statistical significance tests; descriptive statistics; causal inference

1 Introduction

The perspective on what counts as rigorous scientific analysis has changed in agricultural economics as in many other disciplines in the last three decades or so. In the assessment, empirical evidence is given increasingly more weight over theoretical arguments (COLANDER, 2019). This explains also the high weight attributed nowadays to statistical inference. Its purpose is to assess, on a primarily empirical basis and in the face of probabilistic data characterised by random variations, whether the information we obtain from random samples may substantially challenge certain reasonable expectations we hold for the population. The interest that the so called p-value discussion has evoked in agricultural economics lately (see for example HECKELEI et al., 2022, or HIRSCHAUER et al., 2021a) has to be seen against this background because (frequentist) statistical inference, NHST and p-values are intimately linked to each other. This synthesizing review intends to complement the discussion by summarising basic issues of NHST beyond technical details and putting them into perspective.

The controversial discussion of NHST has a long history among statisticians; it started, in fact, with disputes among their modern "founding fathers": FISHER (1992) on the one side and NEYMAN and PEARSON (1928) on the other side. Even though these founding fathers have felt that their two approaches lack compatibility, NHST theory has been shown to represent a "mix-up" of both (SCHNEIDER, 2015; HIRSCHAUER et al., 2021a) and the critical discussion of the approach has never ceased. The "p-value discussion" has started to gain more recognition outside from the profession of statisticians when the "ASA Statement on Statistical Significance and P-Values" was published (WASSERSTEIN and LAZAR, 2016).[1] It was then increasingly discussed that the strict conditions for the use of p-values might often not apply in socio-economic fields of research like agricultural economics. These fields rely heavily on observational data, where the required random variance cannot be guaranteed. However, not only are the requirements for the use of p-values rarely met; possibly even more consequential is the fact that p-values are rarely interpreted correctly. GOODMAN (2008) discusses twelve common "P-Value misconceptions" and deplores the concomitant tendency to undervalue external evidence and "the plausibility of the underlying mechanism".

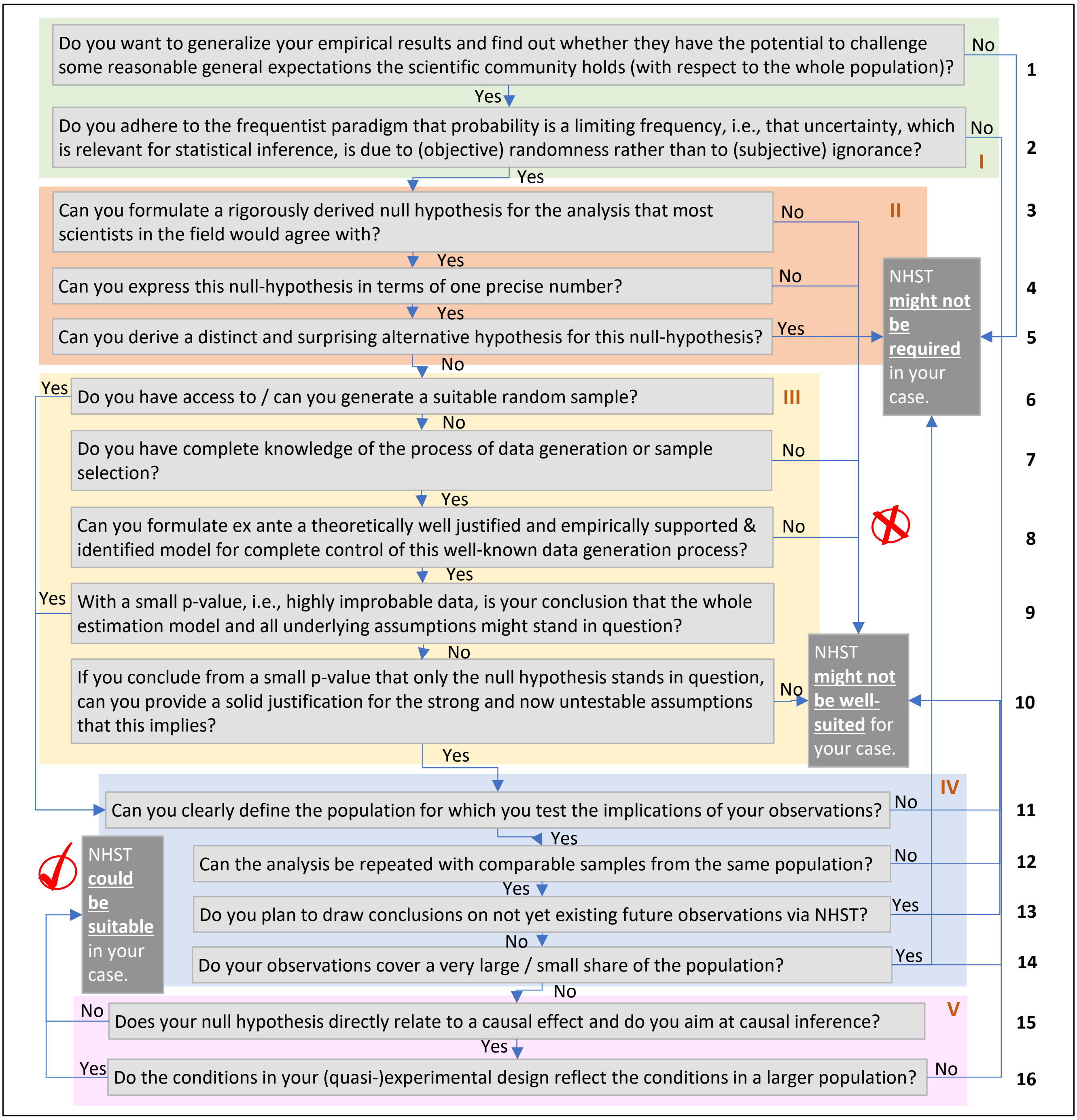
Together with the specific incentive systems of the scientific and publication sectors, these p-value misconceptions have encouraged the proliferation of problematic uses of NHST. For example, where only results with small p-values (below a certain threshold) are judged to be of relevance, only these results tend to be published (HIRSCHAUER et al., 2018). Such "p-hacking" and other malpractices can have severe

[1] For a brief review of this most recent discussion see HIRSCHAUER (2022).

DOI: 10.30430/gjae.2022.0283

consequences because with "selective dissemination of findings", public knowledge will be "based on a biased sample of the studies conducted" (MARKS-ANGLIN and CHEN, 2020: 725). Against this background, a working group of the German agricultural economics association (GEWISOLA) was attributed "the task to assess how 'p-hacking' and the misuse of statistical hypothesis tests in our scientific publications can be best avoided" (HECKELEI et al., 2022: 2). This assignment, however, might not do full justice to the problem. While it is extremely important to discuss how p-values are (mis)used, it is also important to discuss the alternatives to NHST and to understand its value and limitations in the process of scientific progress. The obsession with p-values may indicate that some researchers confuse statistical inference with scientific inference, even though the actual contribution of the former to the latter is rather small (HUBBARD et al., 2019). A principal alternative or indispensable complement to empirical generalisation using statistical inference is theoretical generalisation, which requires the development of strong hypotheses in the preparation of empirical analyses.

Figure 1. Decision tree on use of statistical hypothesis tests with p-values



1 Do you want to generalize your empirical results and find out whether they have the potential to challenge some reasonable general expectations the scientific community holds (with respect to the whole population)?

2 Do you adhere to the frequentist paradigm that probability is a limiting frequency, i.e., that uncertainty, which is relevant for statistical inference, is due to (objective) randomness rather than to (subjective) ignorance?

3 Can you formulate a rigorously derived null hypothesis for the analysis that most scientists in the field would agree with?

4 Can you express this null-hypothesis in terms of one precise number?

5 Can you derive a distinct and surprising alternative hypothesis for this null-hypothesis?

6 Do you have access to / can you generate a suitable random sample?

7 Do you have complete knowledge of the process of data generation or sample selection?

8 Can you formulate ex ante a theoretically well justified and empirically supported & identified model for complete control of this well-known data generation process?

9 With a small p-value, i.e., highly improbable data, is your conclusion that the whole estimation model and all underlying assumptions might stand in question?

10 If you conclude from a small p-value that only the null hypothesis stands in question, can you provide a solid justification for the strong and now untestable assumptions that this implies?

11 Can you clearly define the population for which you test the implications of your observations?

12 Can the analysis be repeated with comparable samples from the same population?

13 Do you plan to draw conclusions on not yet existing future observations via NHST?

14 Do your observations cover a very large / small share of the population?

15 Does your null hypothesis directly relate to a causal effect and do you aim at causal inference?

16 Do the conditions in your (quasi-)experimental design reflect the conditions in a larger population?

NHST **might not be required** in your case.

NHST **might not be well-suited** for your case.

NHST **could be suitable** in your case.

Source: own figure

We arrive at this insight in the course of the systematic discussion of the fundamental and not the technical tasks and conditions of NHST. We summarise the results of this systematisation of arguments in a decision tree (Figure 1). Our overview is necessarily brief and cursory. HIRSCHAUER et al. (2022) provide a more complete accessible up-to-date discussion of statistical inference itself. With respect to more concrete and practical suggestions, e.g. for a suitable alternative representation of regression results, we refer readers to WASSERSTEIN et al. (2019a) for a first summary of important points in this regard. An intensive discussion of many of the critical points summarised in our paper can be found in the 44 articles in the special issue of "The American Statistician" edited by WASSERSTEIN et al. (2019b). Our arguments are close to those summarised in BERNER and AMRHEIN (2022) and much of the literature cited there.

The decision tree also serves as a guide through the article, in which the guiding questions are discussed one after the other. The guiding questions are divided into five blocks that appear as chapters in the article. The first block (Chapter 2) discusses aim and scope of NHST, the second (Chapter 3) elaborates on the fundamental role of different kinds of hypotheses for a proper understanding of NHST. The third block explains the relevance of random samples as in contrast to observational data for NHST (Chapter 4). The fourth block hints at the specific relevance of well-defined populations and of acceptable sample sizes (Chapter 5). The fifth and last block (Chapter 6) briefly touches upon the additional challenge of causal inference as far as it is of direct relevance for empirical generalization and NHST. Chapter 7 concludes.

2 Aim and Scope of NHST (I)

NHST and p-values belong to the field of inferential statistics. Statistical inference is applied to assess whether the results obtained from a sample might have an impact on existing knowledge about the population. Descriptive statistics, in contrast, summarize reliably the available information if observations cover the entire population of interest and do not suffer from systematic measurement errors. Descriptive statistics of observed parameter values then describe "true effects". NHST, in contrast, rests on many assumptions and the preconditions in terms of prior knowledge, data quality, and mastery of the research environment are high: "In descriptive statistics the modeler begins with a set of data in search of a model that conveniently summarizes the information in these data. […] Statistical inference reverses the order by postulating a statistical model a priori and interpreting the data in its context" (SPANOS, 2000: 562).

Statistical inference does not put into question what has been observed within a given sample (LUDWIG, 2005) and "a non-significant effect is not the same thing as a non-existent effect" (HERRERA-BENNETT, 2019: 134). Statistical inference only helps us to assess in how far we could expect to observe this effect within other samples from the same population as well; making additional assumptions to derive test statistics for statistical inference is pointless if interest concentrates on conditions and relationships within a given sample. In that case, NHST should not be applied.

→ Guiding Question No. 1

NHST serves the generalization from sample to population in the context of the frequentist paradigm. The frequentist paradigm generally assumes that probability is an objective "property of the external world", which describes "the limiting relative frequency of the occurrence of an event as the number of suitably defined trials goes to infinity" (POIRIER, 1988: 122). Given the implied objective existence of random deviations, researchers cannot know without additional tests or information whether deviations from expected values observed in samples are random, i.e. within the range of the "normal", or whether they really challenge expectations. Fortunately, the random deviations follow certain regular patterns as described by the central limit theorem: When the data generation or sampling process is repeated many times, the sampling distribution converges to a normal distribution where the mean is equal to the population mean. This enables statistical hypothesis testing and the calculation of the p-value (see Box 1).

Exploitation of knowledge "on the sampling distribution of the test (or other) statistic (i.e., its distribution in hypothetical repetition)" is characteristic for the frequentist approach (COX and MAYO, 2010: 281). Uncertainty that is due to ignorance, in contrast, is not considered in frequentist statistical inference (WAGENMAKERS et al., 2008). Quite to the contrary, in NHST, ignorance has to be ruled out as it could lead to model-misspecifications and biased estimates. With bias, deviations from expected mean values are not any more purely random in nature (see Box 1) and the central limit theorem does not hold any more.

Box 1: Standard errors, p-values and their derivation

Given their random distribution, observations are usually statistically described by a minimum of two moments: the mean and the variance or standard deviation; in the context of regression models, the two moments are the point estimator and the standard error. The standard error is the standard deviation of the sampling distribution of means and equals the standard deviation of a parameter divided by the square-root of the number of observations (BIAU, 2011). Since the standard deviation of the estimated parameter is not known, the standard error is estimated on the basis of the observed estimation error, taking into account variances and covariances of all exogenous data in the analysis (see for example CLARKE, 2005). If the condition of randomness of the error terms is violated, i.e., if the estimation is biased, the estimated standard error of a parameter is not any more consistent with the parameter's true standard deviation. Consequently, "[t]he estimated standard errors of coefficients tell us something about the observed fit of the regression to the data, but they do not reflect uncertainty about the 'true parameters'" (WARD et al., 2010: 372).

The p-value is derived from a test statistic, which is the ratio of the point estimate ("signal") to its standard error ("noise"). This (random) test statistic provides a direct link to the p-value if it follows the standard normal distribution when sampling is repeated a sufficient number of times in accordance with the central limit theorem. The p-value describes a conditional probability for the case if the whole statistical model with all its assumptions, including random data generation and the null hypothesis, is true. Under these conditions, the p-value expresses the probability that a signal to noise ratio (the test statistic) of the size in the original sample or larger could be observed in the population mean, i.e., if a sufficiently high number of repeated random samples were drawn from the population (HIRSCHAUER et al., 2021a), while repetition is required since p-values are random variables themselves (WANG et al., 2019).

Formally, this can be expressed as

$$\mathbb{P}(\tau(\mathbf{X}) \geq \tau(\mathbf{x}); H_0 \text{ is valid}) = p$$

where τ ($\mathbf{x}$) denotes the value of the test statistic τ ($\mathbf{X}$), given the particular sample realization $\mathbf{X} = \mathbf{x}$ (SPANOS, 2000: 691).

The p-value depends on the validity of the central limit theorem; it therefore loses its validity and all meaning if non-random influences are not reliably controlled. Standard errors, in contrast, can then still be interpreted as indicators of the general uncertainty of estimates. Consequently, it means a massive loss of information if the presentation of the results is limited to the point estimator and the p-value (HIRSCHAUER et al., 2021b).

Consequently, NHST does not provide the answer to "interesting questions" (COHEN, 1994) like that on the probability of H0 given the data (D) because "[t]he probability p(D|H0) is not the same as p(H0|D)" (GIGERENZER, 2004: 595). If (and only if) the estimation model and all accompanying assumption are true, a small p-value (in the mean of sufficiently many samples from the same population) tells us that the observed data are unlikely if the null hypothesis is true (KENNEDY-SHAFFER, 2019). If the estimation model, H0 or any accompanying assumption are false, the p-value is invalid irrespective of its size. Small p-values can therefore be considered as reliable evidence against a model (including the hypothesis expressed by it). High p-values, on the other hand, do not tell us much: the p-value could be high because the assumptions and therefore the p-value and the null hypothesis are invalid; or the p-value is high because the data are consistent with the model and the null hypothesis (and, potentially, with many other hypotheses). From the appropriate (in the context of NHST) frequentist perspective, high p-values are therefore not very informative.

The alternative Bayesian paradigm, in contrast, deals with ignorance. In Bayesian approaches, the (subjective) probability of hypotheses is explicitly considered in so called priors. ABADIE (2020) applies (subjective) prior probabilities from the Bayesian paradigm to argue for an interpretation of high p-values as support for the null hypothesis. This irresolvable contradiction with the argument against an interpretation of high p-values developed from the frequentist internal view illustrates how erroneous conclusions can occur when concepts from different paradigmatic approaches are mixed. Since NHST follows the frequentist paradigm, it should not be applied in the context of analyses that follow other paradigms such as Bayesian ("subjectivist", POIRIER, 1988) statistics. Even if this could be justified from a pragmatic "tool-kit" perspective (BANDYOPADHYAY and FORSTER, 2011), it would require intensive and careful reflection and interpretation.

→ Guiding Question No. 2

3 Hypotheses (II)

In the context of NHST, "the null [hypothesis] must be specified as to represent the most established prior scientific belief" (HIRSCHAUER, 2022: 43). In science in general, in contrast, hypotheses are rather understood as "happy guesses" that "organize our thinking about what *might* be true, based on what we've observed so far" (MILNER, 2018, emphasis added). These "happy guesses" are not null- but rather alternative hypotheses. Progressive science is characterised by the establishment of these potentially surprising hypotheses and by the frequently creative empirical analyses that put them to a rigorous test. These tests of hypotheses are not null hypothesis tests in the sense of NHST, which cannot assess evidence for any specific alternative given the condition of the true null hypothesis (WILKINSON, 2013). We cannot test alternative hypotheses with NHST because the sampling distribution of the test statistic is only known for H0 (HAGEN, 1997: 17), while the alternative model's validity has still to be proven. NHST tests established knowledge and in the best case creates a more or less comprehensive negative model of what is not (any longer) valid, which leaves much room for a large variety of possible alternative models and hypotheses. Consequently, "the delusive terminology of NHST, which speaks of hypothesis testing […] has apparently led to much confusion" (HIRSCHAUER et al., 2021a: 130). (Null-)Hypotheses for NHST have to reflect a generally agreed upon state of knowledge and have to be derived from models that reflect a generally agreed upon state of knowledge.

→ Guiding Question No. 3

For NHST, the null hypothesis must be expressed in terms of a precise value against which the data can be tested. In the vast majority of all applications of NHST in (agricultural) economics and other social sciences, this supposedly expected, precise value is zero. However, the hypothesis that a coefficient has exactly the value "zero" is in many cases arbitrary and often theoretically difficult to justify. The problem is therefore not so much, as often stated, that with increasing sample size even the smallest deviations "attain significance" (The "p-value problem"; see for example LIN et al., 2013). The problem is rather that an exact value rarely corresponds to our real, often much less precise, expectation. If we have to expect the rejection of a null hypothesis with sufficiently large samples, then the sharp null hypothesis must be questioned from the outset (IMBENS, 2021).

The case is even worse when the null hypothesis is deliberately chosen contrary to actual expectations; it then represents a straw man that can be torn down with demonstrative simplicity (GELMAN, 2016). The statistical test is then not only of little informative value. Moreover, for a given sample size, the probability that the null hypothesis of a zero effect is true decreases with increasing actually expected effect size (SULLIVAN and FEINN, 2012), so that the p-value loses validity. High p-values can then also not, as suggested by ABADIE (2020), be evaluated as (unexpected) support for a straw man hypothesis and as evidence against the alternative hypothesis. Therefore, if a null hypothesis, as expressed in one precise number, does not express our true expectation given all model assumptions, NHST will not deliver informative results.

→ Guiding Question No. 4

In many applications of NHST, the alternative hypothesis is not explicitly stated. The alternative hypothesis is then simply the "non-null", which merely expects an unspecified deviation from the null hypothesis. In order to test it, statistical hypothesis tests are indispensable in the presence of random deviations if no full sample is available. In contrast to the weak "non-null", strong (alternative) hypotheses deviate considerably from the "normal" expectation expressed in the null hypothesis. They are, in other words, surprising. P-values by themselves do not have much to say about the consistency of the data at hand with the alternative hypothesis as observed data are principally consistent with a large number of imaginable hypotheses (BERNER and AMRHEIN, 2022). Strong hypotheses, however, can increase the statistical power of NHST in a given case and can eventually render it irrelevant.

Statistical power determines the probability of rejecting the null hypothesis if it is false. Power increases with sample size but also with decreasing expected precision of a null hypothesis, respectively, of an estimator (see equations in SERDAR et al., 2021): With decreasing expected precision, H0 is only declared false if the observed deviation from the expected value is strong. As the bandwidth for a "true" H0 becomes larger, statistical power, i.e., the probability to correctly identify cases, where H0 should be declared false according to the chosen precision increases.

The required precision could be determined at hand of the alternative hypothesis. As the effect size expected by the alternative hypothesis increases, an increasingly smaller precision of the estimator is required in order to correctly identify non-null results that are relevant in the light of the alternative hypothesis. With power analyses, one can additionally determine, how many observations would be required for a reliable assessment of H0 (SERDAR et al., 2021). If the effect expected by the alternative hypothesis is sufficiently large, i.e. if the required precision of the estimate is sufficiently low, a single observation is eventually sufficient to reject the null hypothesis of a zero effect rightly and with almost 100 per cent certainty, irrespective of the standard error.

Then, the alternative hypothesis itself still needs to be confirmed. What is important here is that with clear-cut differences in effects, formal assessment of errors becomes relatively unimportant (COX and MAYO, 2010: 278) and NHST turns ultimately obsolete. Their ability to formulate very clear and discriminating null- and alternative hypotheses explains, why in the context of modern physics and chemistry "[t]he data from experiments […] do not usually require statistical analysis" (SPANOS, 2000: 571). Instead, a scientifically convincing theoretical explanation predicts under which conditions certain surprising observations can be expected to (re)occur. The prediction can then be assessed at hand of few observations with an adequate experimental or observational design and without NHST, at hand of descriptive statistics alone (GIGERENZER, 2004).

→ Guiding Question No. 5

Since theoretical justifications always play a major role in validating surprising alternative hypotheses, this process could be called "theoretical generalisation" in contrast to empirical generalisation via NHST.

4 Random Samples and Observational Data (III)

Despite this high potential relevance of descriptive statistics in the appropriate research setup, NHST is given a very high weight in (agricultural) economics. However, given the conditionality of the probability expressed by the p-value (see Chapter 2), NHST can only develop informative power if the analyst is fully aware of the data generation process and controls it. As long as the estimators are not further (causally) interpreted (see Chapter 6), it is sufficient that the sample selection mechanism is fully known and controlled; control of what happens within the sample is then not required in order to apply NHST.

The sample selection mechanism determines how one happens to observe certain observations rather than some other observations, respectively, if certain data are missing within the sample (HECKMAN, 1979). In a linear regression model, for example, uncontrolled selection yields "biased and inconsistent estimates of the effects of the independent variables […] when data on the dependent variable are missing nonrandomly conditional on the independent variables" (WINSHIP and MARE, 1992: 328). In order to ensure that a sufficient number of samples reflects in its mean the characteristics of the population, it must be ensured that the samples have not been selected in a biased way, either consciously or unconsciously. This explains the demand for random samples (see for example HIRSCHAUER et al., 2020b, or WHITE and GORARD, 2021), which guarantee that sample observations only differ by random deviations from the observations in the population (HIRSCHAUER et al., 2021b).

Random samples can be generated by the means of probability sampling. With probability sampling, each member of the population must have a known, non-zero chance of selection set by the sampling procedure (GOODMAN and KISH, 1950: 350). Probability sampling not only generates a sample. It also generates knowledge about the sampling mechanism. It thereby blocks or helps to block uncontrolled environmental influences from affecting sample selection. Random sampling is a special case in probability sampling; it permits every single population member to have an equal chance of presence in the sample. This is obviously not always easy to secure.

In many cases, certain sampling designs will have to be applied in order to ensure, for example, that inaccessible members of a population have the same selection probability like easily accessible members, or that extremely rare traits from a large population be represented adequately in relatively small samples. If the differences in selection probabilities are known as for example in stratified sampling (HIRSCHAUER et al., 2021b), adequate ex post corrections can still preclude systematic deviations of the sample from the population (for a more detailed discussion see HIRSCHAUER et al., 2020b). However, each attempt to "control" random sampling brings with it the thread of introducing unnoticed bias (GOODMAN and KISH, 1950). Uncorrected differences in selection probabilities create a selection bias in subsequent analyses with NHST.

→ Guiding Question No. 6

While controlled (probability) sampling provides at least in principal perfect control of sample selection, it occurs to be rather rare in scientific practice (GIGERENZER, 2004; HAGER, 2013). In (agricultural) economics, analyses are regularly based on observational data, i.e., on data that have not been chosen for the analysis in a deliberate, controlled process. Then, both observable and non-observable non-random influences (VELLA, 1998) on data generation must be controlled ex post to ensure that the total variance in the observed data is reduced to its irreducible random core (GREENLAND, 1990). However, in socio-economic analyses with observational data, models necessarily describe only a small subset of the potentially influential environment of an observation. The need to control sample selection then challenges any system boundary of models. Failure in the effective control of sample selection, however, "will invalidate any statistical inference results built upon the premises of the postulated model" (SPANOS, 2000: 190) since p-values then become meaningless (HIRSCHAUER et al., 2020b; BERK et al., 2010).

That has inconvenient consequences: "Statistical tests in designed studies attempt to answer the question, 'Given random sampling, what are the chances of this result?' In an observational study we can only ask, 'Given the data (acquired without randomization), what are the chances that it is random?'" (LUDWIG, 2005: 678). In order to draw any further conclusions from NHST, the analyst needs complete knowledge and effective controls for the sampling mechanism from the outset, even though in socio-economic studies, "[s]electivity is not only a source of bias in research, but also the subject of substantive research" (WINSHIP and MARE, 1992: 328).

→ Guiding Question No. 7

This knowledge must then be translated into an estimable model to control ex post for structural influences on data generation or sample selection. Overviews over modelling approaches like that by VELLA (1998) or WINSHIP and MARE (1992) show that ex post control of sample selection is always technically demanding, loaded with many assumptions and requires good subject knowledge of the field of study as well as complete data (see also HIRSCHAUER et al., 2021b: 23).

There are no technical panaceas to compensate for a lack of complete knowledge or data. Taking all available control variables into account in the estimation model in order to avoid an omitted variable bias, for example, does not necessarily lead to a minimisation of the risk of sample selection bias. PEARL (2010) shows in his seminal work on causal graphs that with incomplete control, controlling the wrong variables might open new "back-door paths" for confounding influences (see also LUCA et al., 2015, and CLARKE, 2005).

Empirical model optimisation or selection cannot compensate for a lack of knowledge either. In models designed by data driven adjustments, we often observe an inflation of "significant" results, which is a sign of "overfitting" (TONG, 2019). Overfitting means that a model is fitted to the existing data to such an extent that the target population (to be discussed in Chapter 5) becomes identical with the sample. The model thereby loses external validity and NHST, whose aim it is to generalise results, becomes invalid or meaningless as demonstrated empirically by WARD et al. (2010). Model specification testing can contribute to the development of models to some extent, but it is a separate step in the research process and must be done independently from the generalisation of results through NHST (SPANOS and MCGUIRK, 2001; TONG, 2019; BERK et al., 2010). In summary, validity of NHSTs with observational data requires complete knowledge about sample selection, which must be transferred into reliable estimable models.

→ Guiding Question No. 8

Understanding the sample selection process with observational data is often itself essential for understanding the field of enquiry (WINSHIP and MARE, 1992). Models that control sample selection may therefore reflect accumulated knowledge of a discipline on the subject in question. Within them, however, the null hypothesis alone serves to test the expectations, and among potentially numerous estimates "[i]t must be made unequivocally clear which of these p-values are used as a basis for making significance statements, and which were computed only as descriptive summary measures of parts of the data" (WELLEK, 2017: 859). Then, if this one p-value is sufficiently small, it indicates nothing less and nothing more than that the null-hypotheses, the model or both could be flawed. The single p-value puts the whole knowledge embodied in the model into question.

→ Guiding Question No. 9

In order to interpret the p-value as indicator of the fit between data and null hypothesis, researchers have to be completely sure that their model on data generation respectively sample selection is valid (KENNEDY-SHAFFER, 2019). Only then could a small p-value be interpreted as an indication of a misfit of the data specifically to the null hypothesis. In this way, the model itself, which reproduces how the data in the analysis came to be, is excluded from empirical scrutiny. Justification of the model and its assumptions must come from sources other than the estimation itself.

→ Guiding Question No. 10

5 Populations and Sample Size (IV)

Samples are only one side of the coin. For inferential statistics to be meaningful, it must also be clearly defined for which *population* generalised conclusions can be drawn (HIRSCHAUER et al., 2020b: 72; GIGERENZER, 2004: 599). Only with regard to this population can the analysis claim "external validity" (see also HIRSCHAUER et al., 2022: 11-13). The adequate population for generalization is the (parent) population from that a sample is drawn, while the maximum extent of the relevant population (the "target population", BRACHT and GLASS, 1968; THACKER, 2020) depends on the generality and scope that a researcher can defend credibly for a model and a hypothesis (FINDLEY et al., 2021). Practically, the population for inference via NHST is restricted to the "accessible population" (THACKER, 2020) that can serve as parent populations from which samples can actually be drawn (BRACHT and GLASS, 1968).What then really constitutes the parent population of a sample for an analysis additionally depends on "design" decisions (FINDLEY et al., 2021) that might among other things be guided by pragmatic cost-benefit considerations (SERDAR et al., 2021). According to HUBBARD et al. (2019: 93) "assessing the external validity (generalizability) of an investigation's results demands the sampling of settings, treatments, and observations as well as people." Therefore, it might be wise to strive for inference to a rather small, homogenous population initially in order to avoid, for example, the need to sample on settings.

With observational data, determining the corresponding population to which the results can be generalised using NHST is far more difficult if not impossible. Sometimes it is claimed that one should determine ex post whether a sample is "representative" of a population (see for example SEDDON and SCHEEPERS, 2012). However, it remains unclear when "representativeness" is achieved, since random deviations prevent us from expecting the mean values of the variables in any sample to be truly identical to those of the population.

→ Guiding Question No. 11

That statistical correspondence between a single (random) sample and a population is never guaranteed (DEATON and CARTWRIGHT, 2018) is in fact another reason for why NHST demands clearly defined populations. Only from clearly defined populations can comparable samples be drawn repeatedly, and repetition of the analysis with comparable samples from the identical population is required in order to make reliable statements on effect sizes and other estimates. One problem that might prevent researchers from drawing comparable samples repeatedly from identical populations is that populations are dynamic themselves or are affected by changing environmental conditions.

→ Guiding Question No. 12

A related problem arises when NHST is used to make statements about future conditions. Unspoken, this is very often the case. Without further information and assumptions, however, it is inadmissible in most contexts. In processes of non-evolving systems that are situated in completely controlled environments the well-defined process of data generation alone determines the population characteristics. Here, we may conclude through NHST from present observations on still unobservable future observations as the very same process can be expected to generate more data with identical characteristics and distributions in the future. This idea has also inspired the notion of abstract "super-populations", which could be described at hand of data that are generated in simulation runs of a model. In the social sciences, however, population characteristics usually depend largely on exogenous influences that are not under model control. Then, NHST "provides no guide to the length of time over which the initial observations remain valid" (SUMMERFIELD, 1983: 145) for further inference and invoking hypothetical populations would resemble pulling out "a 'get out of jail free card' on external validity", which unfortunately does not exist (FINDLEY et al., 2021: 379).

From a slightly different perspective, the wish to generalize results into the future reflects an attempt for generalizing results beyond the population that has served as reference for an analysis. This attempt is covered by the idea of "transportability" (FINDLEY et al., 2021). Transportability, however cannot be assessed by means of NHST because it can never be guaranteed that a (random) sample from one population differs from another population only by random deviations. To put it the other way around, to "be confident in making broad generalizations necessitates sampling from a 'super-population' composed of every circumstance imaginable which may impact the result" (HUBBARD et al., 2019: 94). This demand is insurmountable, and statistical inference from the current sample on a future population or generally on other populations than that from which the sample has been drawn is invalid.

→ Guiding Question No. 13

The notion of a super-population is sometimes also invoked to justify the use of NHST under the conditions of a full-sample analysis. However, in full samples, estimates describe only what can be observed in the population under the assumptions of the estimation model. There is no larger population in which estimates could deviate at random from the current estimates and from which further samples could be drawn. Thus, if an estimate from the full sample deviates from the expectation expressed in the null hypothesis allowing for the expected precision, the model and/or the null hypothesis must be rejected without the need for further test statistics. NHST is pointless if analyses are conducted with data on the complete population of interest (HIRSCHAUER et al., 2020a; LAKENS, 2022).

In order to generalize the discussion on full samples to the question of sample size, we can refer to the concept of statistical power again (see also Chapter 3). Since standard errors decrease with sample size, large random samples provide us with high statistical power, i.e., the difference between random and systematic deviations can be identified reliably even for small effects and high noise. The "finite population correction factor" (fpc) has been developed in order to correct for the fact that the confidence interval width depends on relative rather than on absolute sample size (LAKENS, 2022: 3). There is only a minor source of random variance left in the close to full-sample case and none in the full-sample case. The fcp and with it the corrected standard error thereby approach zero with an increase of the relative sample size (LAKENS, 2022; HIRSCHAUER et al., 2020a). So, it is precisely in situations of high statistical power that the relevance of NHST is low (SCHNEIDER, 2015).

In contrast, "a small sample taken from a population is unlikely to reliably reflect the features of that population" (HALSEY et al., 2015: 180). Consequently, random variability is high and the statistical power of inference is low in this case. Resulting problems with Type I and Type II Errors (see for example SERDAR et al., 2021) could (and should) of course be mitigated if sampling and the analysis were sufficiently often repeated. From a practical viewpoint, however, this often proves specifically difficult exactly in those situations, where researchers find themselves restricted to small sample sizes. Consequently, NHST can mainly be of potential value if samples are neither very small nor very large.

→ Guiding Question No. 14

6 Causal Inference and its Empirical Generalization (V)

Causal inference is concerned with the empirical identification of causal relationships. There seems to be some confusion (GREENLAND, 1990), but for empirical causal effect identification, what is applied is random *assignment* into a treatment group (COHEN, 2011) but not random *sampling* (DEATON and CARTWRIGHT, 2018). NHST and p-values have no role to play in causal effect identification itself (HIRSCHAUER et al., 2020a).[2] An empirical assessment of population-wide implications of identified causal effects via NHST requires the combined use of random sampling and steps for causal effect identification (THOMAS et

[2] This does not necessarily also apply to the reverse case. With observational data, causal knowledge is often required to develop a model that reliably controls sample selection (see Chapter 4). If structural models are then used to control sample selection, causal determinants should be identified. This need not be the case with reduced form models.

al., 2017; ACKERMAN et al., 2019). Results from causal inference may not generalize to the population if it is conducted on non-representative samples.

The identification of causal effects itself generally requires certain far-reaching assumptions, which cannot be tested statistically but are a pre-condition for the statistical identification of causal effects (PEARL, 2009). Consequently, and in contrast to purely descriptive statistics, identified causal effects could be false. If we identify a causal effect with a model that lacks internal validity (see HIRSCHAUER et al., 2022) and then generalize the result with respect to (an adequately defined) population, we might thereby generalize what we only mistakenly believe to be a causal relationship (HIRSCHAUER et al., 2020a).

→ Guiding Question No. 15

Moreover, the very specific conditions that serve the identification of causal effects can make empirical generalisations of causal effects untenable. The causal effect is identified by an implicit or explicit comparison of a situation with intervention with the contrafactual, necessarily hypothetical, identical situation without intervention (HECKMAN, 2005). Just as random sampling is the "gold standard" in order to control sample selection, randomized control trials (RCTs) are the gold standard for the control of treatment assignment (RUBIN, 2008). However, exactly the specificity of the conditions in RCTs that allows for causal inference is also the ultimate reason for why RCTs' external validity is often questioned (BRACHT and GLASS, 1968; PETERS et al., 2018). DEATON and CARTWRIGHT (2018: 2) go as far as to conclude that "[d]emanding 'external validity' is unhelpful because it expects too much of an RCT while undervaluing its potential contribution".

With observational data, it is notoriously difficult to conduct RCTs for a multitude of reasons (RUBIN, 2008; PETERS et al., 2018). Generally, "without randomization, it is assumptions that will identify the causal effects. These assumptions will be untestable in general and require subject-matter knowledge to justify" (BLACKWELL, 2013; see also GRIER, 2022). RUBIN (2008) proposes to "design" models such that they resemble experiments. These "quasi-experimental" designs identify regions of overlap between comparable observations in the groups of the treated and in the group of the untreated (KUANG et al., 2020). All experimental and quasi-experimental approaches represent so-called "black-box models" (PEARL, 2009). They control the specific individual, historical and spatial conditions of the observed effect but they do not illuminate them (GRIER, 2022). With black-box models, results hold only under the very specific conditions of overlapping observations within one historical situation, but under which conditions the identified effect applies to whom to what extent stays in the dark, and the target population (compare Chapter 5), if any, remains unknown.

In fact, just like certain RCTs, black-box models might not be very "useful" in so far as their results may not apply to "broader population dimensions" or to "a broader set of cases" (FINDLEY et al., 2021: 377) and the pre-conditions for using NHST might not be fulfilled due to a lack in external validity (GREENLAND, 1990).

→ Guiding Question No. 16

Structural models represent the alternative to experimental and quasi-experimental causal effect identification. As they are unreliable from a purely empirical perspective, they are not considered in the decision tree. The advantage of structural models is that they support a true understanding of causal relationships (SIGNORINO and YILMAZ, 2003; DEATON and CARTWRIGHT, 2018): knowledge of treatment selection mechanisms makes us understand in which environments or populations causes are (how) effective (HONG and RAUDENBUSH, 2013) and knowledge of causal mechanisms explains the different affectedness of heterogeneous observations by a treatment (BRAND and THOMAS, 2013). The models themselves thus give us an idea of the populations to which the results can be generalised.

Structural models for causal inference, like those for sample selection control (see Chapter 4), reflect accumulated knowledge of a scientific field. Causal effect identification then requires an extension of valid and comprehensive existing models. It has consequently to be assumed that the new causal parameter complements the "old" model only in an additive sense, i.e., that "the functional relationship between the regressors and the dependent variable is unconditionally monotonic" (SIGNORINO and YILMAZ, 2003: 563). Otherwise, its earlier exclusion could have contributed to omitted variable bias. This structural approach to the generation of scientific knowledge with its manifold strong and untestable assumptions therefore reflects the piecemeal puzzle-solving process that is characteristic for "normal science" (KUHN, 2009). According to KUHN (2009), it is going to go on until the inner contradictions of the model become so disturbing that somebody proposes an alternative.

An accumulation of inner contradiction is to be expected, for example, because "[u]nfortunately, in many areas unconditional monotonicity may be the exception, rather than the rule" (SIGNORINO and YILMAZ, 2003: 564). The violation of additivity assumptions and other internal contradictions of the model are not going to be discovered by NHST, however, because "[i]t does not include devising or modifying the model" (MURPHY et al., 1986: 334). The accumulated contradictions could instead be revealed, for example, within the separate analytical step of model specification, which tries to identify via "a thorough probing of the probabilistic assumptions" potential "misspecification vis-à-vis the information contained in the data" (SPANOS and MCGUIRK, 2001).

7 Conclusions

Today, "normal science" (KUHN, 2009) seems to bind itself to the direct statistical identification of causal effects and to their generalization via inferential statistics. Given all the requirements for a proper conduct of NHST this process does not serve large scientific break-throughs or "scientific revolutions" (KUHN, 2009). At the same time, NHST is so presuppositional that it rarely delivers trustable and robust results with observational data. In agricultural economics and most other disciplines, an honest assessment of analyses at hand of the requirements that are summarized in the decision tree would probably confirm that NHST should not be applied in many studies.

The second big problem in the use of p-values, next to a lack in validity, is their common misinterpretation (LUDWIG, 2005) and their frequent application to questions that they cannot answer (IMBENS, 2021). Since "[p]-values can [only] indicate how incompatible the data are with a specified statistical model" (WASSERSTEIN and LAZAR, 2016: 131), WHITE and GORARD (2021: 58) conclude that "when these outputs are interpreted correctly, they produce information that is at best irrelevant and at worst misleading." We will principally have to recognize that "scientific generalization from a single study is unwarranted" (AMRHEIN et al., 2019: 266). Even the so-called "replication crisis" may be at least partly due to the undervalued fact that estimation results always depend on case-specific conditions and assumptions (AMRHEIN et al., 2019) and that results, including the p-value, vary between samples.

The solution to the problems associated with the use of NHST is simple: avoidance of the use of inferential statistics as far as possible and reasonable. This implies a shift from prioritising inferential statistics to recognising the value of descriptive statistics. Even p-values can be used in descriptive contexts (AMRHEIN et al., 2019). The p-value would then be interpreted "as a statistical summary of the compatibility between the observed data and what we would predict or expect to see if we knew the entire statistical model (all the assumptions used to compute the P value) were correct" (GREENLAND et al., 2016: 339). We, thereby, come to similar conclusions with respect to econometric models like DEATON and CARTWRIGHT (2018: 3) with respect to RCTs: econometric models might be "oversold" because extrapolating or generalizing their results "requires a great deal of additional information", but "under-sold", because they "can serve many more purposes than predicting that results obtained in a trial population will hold elsewhere."

Putting much stronger emphasis on descriptive statistics not only demands a change in routines and in entrenched habits of thought (GIGERENZER, 2004). The shift also has the potential to trigger changes in the research process and in the evaluation of approaches within the disciplines. There could be a renewed awareness that the general validity of hypotheses, even within a carefully defined environment, can never be confirmed on the basis of data alone. More attention in the conception and review of research could then again be paid to how convincingly a theory answers open questions and how informative ("surprising") hypotheses derived from it are, given the current state of knowledge. As a consequence, it could be recognised that case studies, descriptive analyses and the documentation of a few, remarkable observations can potentially make at least as great a contribution to scientific progress as inferential statistics.

Literature

ABADIE, A. (2020): Statistical Nonsignificance in Empirical Economics. In: American Economic Review: Insights 2 (2): 193-208.

ACKERMAN, B., I. SCHMID, K.E. RUDOLPH, M.J. SEAMANS, R. SUSUKIDA, R. MOJTABAI and E. A. STUART (2019): Implementing statistical methods for generalizing randomized trial findings to a target population. In: Addictive behaviors 94: 124-132.

AMRHEIN, V., D. TRAFIMOW and S. GREENLAND (2019): Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication. In: The American Statistician 73 (sup1): 262-270.

BANDYOPADHYAY, P.S. and M.R. FORSTER (2011): Philosophy of Statistics. In: Philosophy of Statistics. Elsevier: 1-50.

BERK, R., L. BROWN and L. ZHAO (2010): Statistical Inference After Model Selection. In: Journal of Quantitative Criminology 26 (2): 217-236.

BERNER, D. and V. AMRHEIN (2022): Why and how we should join the shift from significance testing to estimation. In: Journal of evolutionary biology 35 (6): 777-787.

BIAU, D.J. (2011): In brief: Standard deviation and standard error. In: Clinical orthopaedics and related research 469 (9): 2661-2664.

BLACKWELL, M. (2013): Observational Studies and Confounding. PSC 504. https://www.mattblackwell.org/teaching/psc504/.

BRACHT, G.H. and G.V. GLASS (1968): The External Validity of Experiments. In: American Educational Research Journal 5 (4): 437.

BRAND, J.E. and J.S. THOMAS (2013): Causal Effect Heterogeneity. In: Morgan, S.L. (ed.): Handbook of Causal Analysis for Social Research. Handbooks of Sociology and Social Research. Springer Netherlands, Dordrecht: 189-213.

CLARKE, K.A. (2005): The Phantom Menace: Omitted Variable Bias in Econometric Research. In: Conflict Management and Peace Science 22 (4): 341-352.

COHEN, H.W. (2011): P values: use and misuse in medical literature. In: American journal of hypertension 24 (1): 18-23.

COHEN, J. (1994): The earth is round (p < .05). In: American Psychologist 49 (12): 997-1003.

COLANDER, D. (2019): Introduction to symposium on teaching undergraduate econometrics. In: The Journal of Economic Education 50 (4): 337-342.

COX, D. and D.G. MAYO (2010): Objectivity and Conditionality in Frequentist Inference. In: Mayo, D.G. and A. Spanos (eds.): Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science. Cambridge University Press, Cambridge: 276.

DEATON, A. and N. CARTWRIGHT (2018): Understanding and misunderstanding randomized controlled trials. In: Social science & medicine 210 (8): 2-21.

FINDLEY, M.G., K. KIKUTA and M. DENLY (2021): External Validity. In: Annual Review of Political Science 24 (1): 365-393.

FISHER, R.A. (1992): Statistical Methods for Research Workers. In: Kotz, S. and N. L. Johnson (eds.): Breakthroughs in Statistics. Springer Series in Statistics. Springer New York, New York, NY: 66-70.

GELMAN, A. (2016): The Problems With p-Values Are Not Just With p-Values. supplemental material to the ASA statement on p-values and statistical significance. In: The American Statistician 70 (10).

GIGERENZER, G. (2004): Mindless statistics. In: The Journal of Socio-Economics 33 (5): 587-606.

GOODMAN, R. and L. KISH (1950): Controlled Selection—A Technique in Probability Sampling. In: Journal of the American Statistical Association 45 (251): 350-372.

GOODMAN, S. (2008): A dirty dozen: twelve p-value misconceptions. In: Seminars in hematology 45 (3): 135-140.

GREENLAND, S. (1990): Randomization, Statistics, and Causal Inference. In: Epidemiology 1 (6): 421-429.

GREENLAND, S., S.J. SENN, K.J. ROTHMAN, J.B. CARLIN, C. POOLE, S.N. GOODMAN and D.G. ALTMAN (2016): Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. In: European journal of epidemiology 31 (4): 337-350.

GRIER, K. (2022): Causal Inference and Austrian Economics. In: D'Amico, D.J. and A.G. Martin (eds.): Contemporary Methods and Austrian Economics. Advances in Austrian Economics. Emerald Publishing Limited, Bingley, UK: 105-114.

HAGEN, R.L. (1997): In praise of the null hypothesis statistical test. In: American Psychologist 52 (1): 15-24.

HAGER, W. (2013): The statistical theories of Fisher and of Neyman and Pearson: A methodological perspective. In: Theory & Psychology 23 (2): 251-270.

HALSEY, L.G., D. CURRAN-EVERETT, S.L. VOWLER and G.B. DRUMMOND (2015): The fickle P value generates irreproducible results. In: Nature methods 12 (3): 179-185.

HECKELEI, T., S. HÜTTEL, J. ROMMEL and M. ODENING (2022): The Replicability Crisis and the p-Value Debate - what Are the Consequences for the Agricultural and Food Economics Community? Preprints. https://www.preprints.org/manuscript/202201.0311/v1, call: 28.4.2022.

HECKMAN, J.J. (1979): Sample Selection Bias as a Specification Error. In: Econometrica 47 (1): 153.

HECKMAN, J.J. (2005): 1. The Scientific Model of Causality. In: Sociological Methodology 35 (1): 1-97.

HERRERA-BENNETT, A. (2019): How do researchers evaluate statistical evidence when drawing inferences from data? Dissertation zum Erwerb des Doctor of Philosophy (Ph.D.). Munich Center of the Learning Sciences, Ludwig-Maximilians-Universität München, Munich.

HIRSCHAUER, N. (2022): Unanswered questions in the p -value debate. In: Significance 19 (3): 42-44.

HIRSCHAUER, N., S. GRÜNER and O. MUßHOFF (2022): Fundamentals of statistical inference. What is the meaning of random error? Springer Nature, Cham.

HIRSCHAUER, N., S. GRÜNER, O. MUßHOFF and C. BECKER (2021a): A Primer on p-Value Thresholds and α-Levels - Two Different Kettles of Fish. In: German Journal of Agricultural Economics 70 (2): 123-133.

HIRSCHAUER, N., S. GRÜNER, O. MUßHOFF, C. BECKER and A. JANTSCH (2021b): Inference using non-random samples? Stop right there! In: Significance 18 (5): 20-24.

HIRSCHAUER, N., S. GRÜNER, O. MUßHOFF and C. BECKER (2020a): Inference in economic experiments. In: Economics 14 (1).

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.

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.

HONG, G. and S.W. RAUDENBUSH (2013): Heterogeneous Agents, Social Interactions, and Causal Inference. In: Morgan, S.L. (ed.): Handbook of Causal Analysis for Social Research. Handbooks of Sociology and Social Research. Springer Netherlands, Dordrecht: 331-352.

HUBBARD, R., B.D. HAIG and R.A. PARSA (2019): The Limited Role of Formal Statistical Inference in Scientific Inference. In: The American Statistician 73 (sup1): 91-98.

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

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 (Suppl 1): 82-90.

KUANG, K., L. LI, Z. GENG, L. XU, K. ZHANG, B. LIAO, H. HUANG, P. DING, W. MIAO and Z. JIANG (2020): Causal Inference. In: Engineering 6 (3): 253-263.

KUHN, T.S. (2009): The structure of scientific revolutions. University of Chicago Press, Chicago.

LAKENS, D. (2022): Sample Size Justification. In: Collabra: Psychology 8 (1).

LIN, M., H.C. LUCAS and G. SHMUELI (2013): Too Big to Fail: Large Samples and the p-Value Problem. In: Information Systems Research 24 (4): 906-917.

LUCA, G. de, J.R. MAGNUS and F. PERACCHI (2015): On the Ambiguous Consequences of Omitting Variables. In: SSRN Electronic Journal.

LUDWIG, D.A. (2005): Use and Misuse of p-Values in Designed and Observational Studies. Guide for Researchers and Reviewers. In: Aviation, Space, and Environmental Medicine 76 (7): 675-680.

MARKS-ANGLIN, A. and Y. CHEN (2020): A historical review of publication bias. In: Research synthesis methods 11 (6): 725-742.

MILNER, S. (2018): Newton didn't frame hypotheses. Why should we? In: Physics Today 24. April 2018. DOI: 10.1063/PT.6.3.20180424a.

MURPHY, E.A., E.M. ROSELL and M.I. ROSELL (1986): Deduction, inference and illation. In: Theoretical medicine 7 (3): 329-353.

NEYMAN, J. and E.S. PEARSON (1928): On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I. In: Biometrika 20A (1/2): 175.

PEARL, J. (2009): Causal inference in statistics: An overview. In: Statistics Surveys 3: 96-146.

PEARL, J. (2010): An introduction to causal inference. In: The international journal of biostatistics 6 (2): Article 7.

PETERS, J., J. LANGBEIN and G. ROBERTS (2018): Generalization in the Tropics - Development Policy, Randomized Controlled Trials, and External Validity. In: The World Bank Research Observer 33 (1): 34-64.

POIRIER, D.J. (1988): Frequentist and Subjectivist Perspectives on the Problems of Model Building in Economics. In: Journal of Economic Perspectives 2 (1): 121-144.

RUBIN, D.B. (2008): For objective causal inference, design trumps analysis. In: The Annals of Applied Statistics 2 (3).

SCHNEIDER, J.W. (2015): Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. In: Scientometrics 102 (1): 411-432.

SEDDON, P.B. and R. SCHEEPERS (2012): Towards the improved treatment of generalization of knowledge claims in IS research: drawing general conclusions from samples. In: European Journal of Information Systems 21 (1): 6-21.

SERDAR, C.C., M. CIHAN, D. YÜCEL and M.A. SERDAR (2021): Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. In: Biochemia medica 31 (1): 10502.

SIGNORINO, C.S. and K. YILMAZ (2003): Strategic Misspecification in Regression Models. In: American Journal of Political Science 47 (3): 551-566.

SPANOS, A. (2000): Probability theory and statistical inference. Econometric modelling with observational data. Cambridge University Press, Cambridge.

SPANOS, A. and A. MCGUIRK (2001): The Model Specification Problem from a Probabilistic Reduction Perspective. In: American Journal of Agricultural Economics 83 (5): 1168-1176.

SULLIVAN, G.M. and R. FEINN (2012): Using Effect Size-or Why the P Value Is Not Enough. In: Journal of graduate medical education 4 (3): 279-282.

SUMMERFIELD, M.A. (1983): Populations, Samples and Statistical Inference in Geography. In: The Professional Geographer 35 (2): 143-149.

THACKER, L.R. (2020): What Is the Big Deal About Populations in Research? In: Progress in transplantation (Aliso Viejo, Calif.) 30 (1): 3.

THOMAS, R.L., P.R. BARACH, J.D. WILKINSON, A.A. FAROOQI and S.E. LIPSHULTZ (2017): The danger of relying on the interpretation of p-values in single studies: Irreproducibility of results from clinical studies. In: Progress in Pediatric Cardiology 44 (3): 57-61.

TONG, C. (2019): Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science. In: The American Statistician 73 (sup1): 246-261.

VELLA, F. (1998): Estimating Models with Sample Selection Bias: A Survey. In: The Journal of Human Resources 33 (1): 127.

WAGENMAKERS, E.-J., M. LEE, T. LODEWYCKX and G.J. IVERSON (2008): Bayesian Versus Frequentist Inference. In: Hoijtink, H., I. Klugkist and P.A. Boelen (eds.): Bayesian Evaluation of Informative Hypotheses. Springer New York, New York, NY: 181-207.

WANG, B., Z. ZHOU, H. WANG, X.M. TU and C. FENG (2019): The p-value and model specification in statistics. In: General psychiatry 32 (3): e100081.

WARD, M.D., B.D. GREENHILL and K.M. BAKKE (2010): The perils of policy by p-value: Predicting civil conflicts. In: Journal of Peace Research 47 (4): 363-375.

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 (2019a): Moving to a World Beyond "p < 0.05". In: The American Statistician 73 (sup1): 1-19.

WASSERSTEIN, R.L., A.L. SCHIRM and N.A. LAZAR (2019b): Statistical Inference in the 21st Century: A

World Beyond p < 0.05. In: The American Statistician 73 (sup1).

WELLEK, S. (2017): A critical evaluation of the current "p-value controversy". In: Biometrical journal. Biometrische Zeitschrift 59 (5): 854-872.

WHITE, P. and S. GORARD (2021): Against Inferential Statistics. How and why current statistics teaching gets it wrong. In: Statistics Education Research Journal 16 (1): 55-65.

WILKINSON, M. (2013): Testing the null hypothesis: the forgotten legacy of Karl Popper? In: Journal of sports sciences 31 (9): 919-920.

WINSHIP, C. and R.D. MARE (1992): Models for Sample Selection Bias. In: Annual Review of Sociology 18 (8): 327-350.

DR. ANNE MARGARIAN
Thünen Institute of Market Analysis
Bundesallee 63, 38116 Braunschweig
e-mail: anne.margarian@thuenen.de