Indian Journal of Agricultural Economics 81: 1(2026):148-158
DOI:10.63040/25827510.2026.01.010

RESEARCH NOTE

The Fallacy of Null Hypothesis Significance Testing: A discourse on the Mis(use) of P Values in Agricultural Social Sciences

Aditya K.S.[1], Neha W. Qureshi[2], and Girish K. Jha[3]

ABSTRACT

Null Hypothesis Significance Testing (NHST), once considered a gold standard in empirical research, is increasingly under scrutiny for its misuse and widespread misinterpretation, also in agricultural social sciences. This review revisits foundational concepts such as P-values, confidence intervals, and levels of significance, and critically examines their limitations when used as rigid inferential tools. Drawing from literature and illustrative examples, we explore how overreliance on p-values facilitates binary thinking, misrepresents uncertainty, and incentivizes practices like P-hacking, selective reporting, and hindsight bias. We challenge the false equivalence between statistical and economic significance, advocating instead a shift toward transparency, contextual interpretation, and a focus on effect sizes and confidence intervals. While not rejecting NHST outright, we call for more nuanced and responsible use of inferential statistics in social science research. Our goal is to stimulate reflection and dialogue among researchers about the assumptions underlying their analyses and the robustness of their conclusions. This change coupled with open science practices like open access publication, sharing data and replication codes can help to overcome the replication crisis that is plaguing our community.

Keywords: Null Hypothesis Significance Testing (NHST), P-values and Confidence Intervals, Statistical Misinterpretation, p-hacking

JEL Codes: A20, B41, C12, C18

I

INTRODUCTION

Statistical inference remains a cornerstone of empirical research, particularly in the agricultural and social sciences, where data-driven decision-making informs policy, practice, and theory. Among the most widely used inferential tools are null hypothesis significance testing (NHST) and P-values, often accompanied by confidence intervals and specified levels of significance (commonly 0.05). While these tools were designed to aid researchers in quantifying uncertainty and testing the plausibility of hypotheses, their widespread and often mechanical use has sparked growing concern across disciplines (Wasserstein et al., 2019; Haaf et al., 2019; Heckelei et al., 2021). This is particularly true in case of social sciences, where NHST assumes even greater importance in explorative research using observational data.

[1]ICAR- National Institute of Agricultural Economics and Policy Research, New Delhi
[2]Fisheries Economics, Extension and Statistics Division, ICAR-Central Institute of Fisheries Education, Mumbai
[3]Division of Agricultural Bioinformatics, ICAR- Indian Agricultural Statistics Research Institute, New Delhi

The null hypothesis, typically denoted as $H_{0,}$ posits that there is no effect or no difference between groups or variables under study. Through NHST, researchers compute a test statistic from sample data and derive a P-value, which indicates the probability of observing the given data (or more extreme) assuming that H_0 is true. A result is often deemed "statistically significant" if this P-value falls below a pre-defined alpha level (e.g., 0.05), suggesting sufficient evidence to reject H_0. The common practice is to first estimate a model to obtain the coefficients, and then go for NHST. Only the coefficients that are 'statistically significant' are interpreted.

However, over time, the P-value has become more of a gatekeeper than a guide, often misused as a binary indicator of truth or importance (Imbens, 2021; Krueger & Heck, 2019; Betensky, 2019). This misuse is compounded by a general lack of understanding about what the P-value actually represents. Contrary to popular belief, the P-value is not the probability that the null hypothesis is true, nor does it reflect the size or practical relevance of an effect. Rather, it is a conditional probability: the probability of obtaining the observed data, or something more extreme, assuming the null hypothesis is true. This distinction is subtle but critical. A low P-value does not prove that the null hypothesis is false; it merely suggests that the observed data would be unlikely under the assumption of no effect. Likewise, a high p-value does not confirm that the null hypothesis is true; it may simply indicate that the data are insufficiently informative or that the study lacks statistical power.

Confidence intervals (CIs), on the other hand, provide a range of plausible values for an estimated parameter and serve as a complementary, or even preferable, alternative to NHST. A 95% confidence interval, for example, suggests that if the same study were repeated many times, approximately 95% of the computed intervals would contain the true population parameter. Yet even confidence intervals are often misinterpreted or misused when treated as indirect tests of statistical significance.

The level of significance (α\alpha), usually set at 0.05, plays a critical role in hypothesis testing. While intended as a conventional threshold to control Type I error (the probability of falsely rejecting a true null hypothesis), its universal application has led to arbitrary dichotomization of results into "statistically significant" and "statistically non-significant," often overshadowing the broader scientific context, study design, or real-world implications. For instance, P value is a decreasing function of sample size (Brodeur et al., 2016; Hirschauer et al., 2018; Imbens, 2021); P value tends to be lower for larger sample sizes, and thus, the universality of having 0.05 as a threshold does not make sense. If one looks at the published literature using nationally representative datasets, most coefficients will have P-values less than 0.05.

In the context of agricultural social sciences, where research often intersects with complex systems, human behaviour, and diverse socio-economic settings, a blind adherence to P-values and significance thresholds can lead to misleading conclusions, misplaced policy recommendations, and publication biases. In a way,

the overemphasis on NHST and P values is also a contributing factor to the widespread replication crisis in the social sciences (Heckelei et al., 2021; Haaf et al., 2019; Wasserstein et al., 2019).

This paper critically examines the theoretical underpinnings, practical applications, and common misinterpretations of P-values and related statistical tools in the field. It advocates for a more nuanced and transparent approach to statistical inference; one that emphasizes effect sizes, uncertainty, and the substantive context of findings over rigid thresholds. We provide recommendations that also contribute to increased transparency, rigour, and open-science approaches in the social sciences (Ankel-Peters et al., 2025; Finger et al., 2024).

The paper contributes to the agricultural social science community in the following ways. 1. Draws attention of the researchers towards routine and mechanical statistical practice of hypothesis testing and its implications for sound science. 2. Provide a brief overview of the 'P value debate' and provide practical suggestions that can be implemented, both at the level of the researcher and the journals. The paper is structured to present the issues with current hypothesis-testing practices and then discuss possible steps we can take to improve rigour and transparency in science.

II

NHST IN PRACTICE

We would like to begin with quoting Leamer 1983 on the work of the quantitative social scientist. So, the quantitative social scientist has a difficult problem of dealing with observational data due to lack of randomization. We measure the values for the variables and then try to examine for the association between them, without having control over variables which allow experimental manipulation (Leamer, 1983; Tong, 2019; Imbens, 2021). Furthermore, we aim to estimate the parameter values from the sample. The Null Hypothesis Significance Testing (NHST) is an integral part of such inference. However, the overemphasis on statistical hypothesis testing and consequent problems has received considerable attention in academia worldwide. In this paper, we reassess the validity and interpretive scope of this commonly used metric. Our aim is not to criticize or to offer solutions to this issue – instead, we want to bring it to the table so that readers can introspect on their own work in light of this debate.

> "Econometricians would like to project the image of agricultural experimenters who divide a farm into a set of smaller plots of land and who select randomly the level of fertilizer to be used on each plot. If some plots are assigned a certain amount of fertilizer while others are assigned none, then the difference between the mean yield of the fertilized plots and the mean yield of the unfertilized plots is a measure of the effect of fertilizer on agricultural yields. The econometrician's humble job is only to determine if that difference is large enough to suggest a real effect of fertilizer, or is so small that it is more likely due to random variation.
>
> This image of the applied econometrician's art is grossly misleading. I would like to suggest a more accurate one. The applied econometrician is like a farmer who notices that the yield is somewhat higher under trees where birds roost, and he uses this as evidence that bird droppings increase yields. However, when he presents this finding at the annual meeting of the American Ecological Association, another farmer in the audience objects that he used the same data but came up with the conclusion that moderate amounts of shade increase yields. A bright chap in the back of the room then observes that these two hypotheses are indistinguishable, given the available data. He mentions the phrase "identification problem," which, though no one knows quite what he means, is said with such authority that it is totally convincing. The meeting reconvenes in the halls and in the bars, with heated discussion whether this is the kind of work that merits promotion from Associate to Full Farmer; the Luminists strongly opposed to promotion and the Aviophiles equally strong in favour. Leamer(1983).

Let us start with an example. Assume that we want to examine if the farmer's fertiliser use depends on if the farmer received some formal training. The usual procedure for an observational study would be to select some 'n' number of farmers randomly, out of which some farmers have received the training. Then using a structured survey, measure the quantity of fertilizer used along with other socio-economic variables. Estimate a regression with fertilizer use as a dependent variable and dummy variable to indicate if farmer has received the training, along with other controls like age, education, soil type and land holding. The estimate we are interested, let us denote it as β, is the regression coefficient for training on quantity of fertilizer used. The question that we would like to answer is, with certain level of confidence, can we say that this value is a good approximation of the population parameter (or true value)? That is where NHST comes in. We start with a Null Hypothesis (NH) that training has no role on quantity of fertilizer used, and hence $\beta=0$. Suppose that from our regression we get a value of $\beta = 45$ and a P-value of 0.04.

With the help of P value, we can say that if the NH of zero impact is true and if the same study is repeated a large number of times, each time drawing a random sample from the population, then only 4% of the time the effect can be equal or greater than 45. In other words, if the NH were to be true, the effect size of 45 is less likely to be obtained (Betensky, 2019; Tong, 2019). So, given this result, we can reject the NH that training has no effect on quantity of fertilizer used. We can say that farmers, on average use 45kg more fertilizer per acre conditional on mean, if they had received formal training, at 5% level of significance. The P values, though useful, are often misinterpreted and misused (mostly unintentionally).

The usual rule, though much criticized recently, is to compartmentalise result into statistically significant and non-significant (Wasserstein et al., 2019; Krueger & Heck, 2019; Betensky, 2019). If the P value of the coefficient is less than 0.05, then the coefficient is considered as statistically significant, and interpreted. This dichotomization of the hypothesis testing has led to some problems. The statistically insignificance is used as evidence for absence of relationship, which is again problematic. Also, statistical significance is used as the sole metrics to argue that the results of the two studies are different, which is not true as we explore in the next section.

Also, many researchers have highlighted frequent misinterpretation of the P values, though benign. A p-value below 0.05 is often mistakenly taken as proof that the null hypothesis is false, when it only reflects the probability of observing results at least as extreme as those obtained, assuming the null hypothesis holds. Also, many papers simply use asterisk marks (*) to indicate 10, 5 and 1 % levels of statistical significance (LoS). However, P value depends on the sample size and for studies with larger sample sizes ends up getting lower P values and the arbitrary cut-off like 5% LoS may not be appropriate. Another common mistake in hypothesis testing is, to claim the absence of evidence as evidence of absence, as noted before. For instance, in our earlier example, if the P value for the training variable turns out be 0.11, we can only say that we could not find any empirical evidence that suggests formal training affects the quantity of fertilizer used. But, we cannot say that since the variable is not statistically significant, it is proved that training doesn't affect the fertiliser use. One should remember that the P value depends on the null hypothesis and the observed data. Apart from the misinterpretation of P values, there are other concerns with their current use in the scientific literature, which we have tried to highlight here.

1. **Statistical significance as an indicator of relevance:** The significance and non-significance are considered as proof for saying whether the variable is important or not. However, the P-value is not a measure of confidence in the estimate, rather it is the tail probability of observing data as extreme as those obtained, assuming null hypothesis is true. Non-significance does not

indicate the absence of an effect, but rather the absence of sufficient evidence to detect it, often based on an arbitrary threshold.

2. **Violations of assumptions of statistical inference:** To be able to draw inference on the population from the sample, there are a few prerequisites- the population should be defined, sampling frame should be defined clearly, sample should be drawn randomly, and sample size should be sufficient based on power calculations. Most studies don't even define the population or the sampling frame. Sampling frame i.e., list of sampling units in the population are seldom available, which is needed for randomization. In this case, the P-value can be misleading. Also, the inference of the P value is always contingent on the null hypothesis, which is seldom stated explicitly.
3. **Reproducibility Crisis and P-value:** Reproducibility-the ability to replicate a study's findings using the same methods- is essential to credible science. However, a growing reproducibility crisis has emerged, with many studies failing to replicate. A key cause is the overreliance on P-values, often treated as definitive proof of significance if below the arbitrary 0.05 threshold. This value is frequently misunderstood, with small differences (e.g., 0.049 vs. 0.051) being given undue weight, despite offering little real-world distinction. Why P-value doesn't ensure reproducibility? The P value is estimated from the data comprising of random variables and is itself a random variable (Hung et al,1997 and Murdoch et al. 2008). Despite its popularity, the P-value is rarely reported with a standard error or confidence interval, unlike other statistics. This is because it reflects a property of the sample, not an underlying population value—there's no "true" p-value. Its random nature and high variability across samples are often overlooked, a phenomenon referred to as the "dance of the P-values." (Cummings, 2014). This variability is a major reason for poor reproducibility (Sapra and Nundi, 2018)
4. **Researchers' degrees of freedom (or method selection bias):** In data collection and analysis, a series of decisions have to be made by researchers, each of which can influence the P value. For instance, a researcher might try different model specifications and different transformations of variables and report the results of the model which is either more attractive or easier to interpret. This overfitting is called by many names – researchers' degrees of freedom or model selection bias or P hacking or File drawer problem. There are various sources of this – selection of variables, various functional forms, data cleaning procedures, transformations etc. There is also a case of publication bias – journals are reluctant to publish studies that report statistically non-significant findings. Hence, not only a researcher is incentivized for 'P hacking', it leads to unnecessary duplication in chase of non-existent effects (Mervis, 2014; Brodeur et al., 2016; Heckelei et al., 2021). For example, research has indicated that most of the published papers

have P values which are very close to 0.05, and just statistically significant. The distribution of the test statistics of the published papers has a bi-modal distribution (as against a uni-modal distribution if there is no P hacking), which is a shred of evidence that researchers try different models and settle for the one which produces statistically significant results.

5. **Selective reporting of P Values:** It is found that authors are more likely to selectively mention and interpret those variables where the coefficient has the lowest P value. This is again problematic, as mentioned earlier, a lower P value doesn't indicate strength of evidence.
6. **Significant sameness vs significant difference paradigm**: Significant difference paradigm is most commonly used in social science research. In these methods, the focus is usually on designing unique or novel studies, theory building, and generalization based on small number of studies. These studies heavily rely on the significance of testing procedures. The problem is that any finding based on individual datasets can never be generalized for the entire population. There will always be some or the other boundary conditions (or contextual factors) which limits the generalization. This is particularly an issue in case of social science research where despite ample amount of empirical work, there is a lack of stubborn facts or universally applicable theories. Referring to marketing studies, Leone and Schultz (1980, p. 11) remarked that marketing's knowledge base is "more marsh than bedrock". Hence, the need of the hour is for a shift towards the significant sameness paradigm, where the studies build on existing knowledge, replicated using different datasets, and identifying the similarities and stubborn facts, and then building a theory to explain these phenomena. Such an approach doesn't rely on point estimates and NHST but focuses more on overlapping confidence intervals.
7. **Multiple hypothesis testing:** When many hypotheses are tested simultaneously, like in case of Randomised Control Trials (RCT's) with many treatments, the false error rates increase. So, the P values unadjusted for multiple hypothesis testing are biased.
8. **Sign econometrics:** Many studies simply use the P value threshold to identify the independent variables that influence the dependent variables. The interpretation is limited to which variables positively influence or negatively influence the dependent variable based on just the P values and the sign of the coefficient. Using our earlier example, if we say that training is positively related to the quantity of fertilizer used, without mentioning by how much, then it is a case of sign econometrics. Discussion on effect size (which is also known as 'economic significance') is important and often more interesting (Leone & Schultz, 1980; Imbens, 2021; Betensky, 2019: Lindena and Hess, 2022).

9. **Extreme empiricism and hindsight bias:** This is also known as HARKing – Hypothesizing After the Results are Known. Since the overreliance on point estimates and NHST, many studies are not grounded on strong theoretical basis. Researchers often don't think about the expected size and sign of the coefficient for the variables under consideration, nor do they have a sound theoretical basis for selection of variables. In some instances, hypotheses are formulated in response to empirical results, particularly the statistical significance of variables. It is often observed that regression models include a wide array of potential explanatory variables, with attention subsequently focused on those that are statistically significant. Interpretations are then constructed based on the signs and significance of the coefficients, sometimes leading to ex post theoretical justification of the findings. When presented without adequate clarification of the research design, this approach may create the impression that the hypotheses were specified a priori and subsequently supported by the data.

Concerns about the overemphasis on P-value thresholds and the routine application of statistical procedures have led some researchers to propose reducing or even discontinuing the use of P-values in certain research contexts. A few others are calling for a revision of the 0.05 threshold, reducing it to 0.005 in light of access to big data. A few others, like Imbens, are calling for more careful use of P values (Wasserstein et al., 2019; Imbens, 2021; Haaf et al., 2019). American Statistical Association has published a special session on concerns surrounding NHST. Based on a brief review, we summarise these suggestions to some actionable points.

III

WHAT CAN WE DO?

Though there is no agreement on an alternative to NHST or how to prevent the 'P hacking', literature in this area suggests the following broad directions that one can take.

1. **Understand the difference between exploratory and confirmatory research:** For exploratory research, scientific inference is best suited. Careful observations, qualitative study designs, descriptive statistics, and other nonparametric methods are best suited for this. If we use the NHST here, there is no way to avoid researcher bias (or model selection bias). Confirmatory analysis, on the other hand, is done when we already know some basic information about the topic. The literature review is strong enough to guide the selection of variables and the choice of model. In such cases, a common method is to submit the data analysis plan to a journal or board for approval. Such things can minimise bias due to overfitting.
2. **Use the confidence interval:** Confidence intervals are more useful than point estimates in most cases. If we repeat the study following the same

method and random sampling, then 95% confidence intervals are likely to contain the population mean. A 95% confidence interval means that if we were to repeatedly draw samples and construct confidence intervals in the same way, 95% of those intervals would contain the true population mean. Further, we can do away with the dichotomy between statistically significant and non-significant results by using confidence intervals. Suppose the impact of a new technology on food security is -0.03 to 0.20 (assuming standardized food security score to range between 0 and 1). Under the NHST, this would simply be interpreted as not statistically significant. But using confidence intervals, we can say that the results are more consistent with an increase in the food security score due to technology adoption.

3. **Adjust P values for multiple hypothesis testing:** As mentioned earlier, P values should be adjusted in most applications. The Bonferroni-adjusted P-value is commonly used to account for this (Korn & Graubard, 1990).
4. **Don't use the term significance arbitrarily:** It is important that in the text, one should avoid the arbitrary use of the word significant without making the distinction between economic significance and statistical significance. For example, Aurbacher et al. (2024) emphasize the importance of distinguishing between effect size and statistical significance, as well as the need for unbiased assessments of empirical findings—including those that are statistically insignificant. If you say the variable 'X' is significant in explaining 'Y', the statement is arbitrary. One can say that the coefficient of X on Y is statistically significant with a P value of ---- and the effect size is ---- (With 95% CI of -----). Such statements are more accurate and useful. For instance, Lindena and Hess (2022) demonstrate that while herd size is statistically significant in their regression models, the effect size is small, underscoring the importance of interpreting results in context rather than relying solely on statistical significance.
5. **Transparent reporting:** When reporting the results of a regression or NHST, also provide details of the theoretical basis for model selection and the alternative models considered. This provides assurance that the results provided are not from P-hacking! Going back to Leamer 1983, 'mapping the assumptions to the inference' can be a useful tool. This is also called as 'robustness checks' in modern analogy. Provide results for alternate model specifications, what happens if some assumptions are relaxed, what happens if some variables or added or removed. If the inferences are stable under these assumptions, then they are more reliable. Transparently reporting the data and the codes are also encouraged.
6. **Pre-registrations:** Hüttel and Hess (2024) argue that merely increasing openness in science is insufficient; structural reforms are required to address the root causes of ithe replication crises and ritualised statistical practices. To avoid the problem of HARKing or P hacking, it is suggested to pre-register

the research design, so that researcher apriori states the hypothesis which they are testing, how the variables will be measured and how the data will be analysed.

V

CONCLUSION

We don't think that the debate surrounding P values has reached the stage of providing stylized solutions to the problem highlighted. At this point, it is important to be aware of the pitfalls of statistical hypothesis testing and to follow the recommendations. It's high time to sensitize this problem, especially the random behaviour of P-Values (Referred as "Dance of P-Values") that happens to be the inferential base of most of our research findings through conferences, workshops, publications, etc. (Sapra and Nundy, 2018). Over time, more concrete solutions will emerge collectively. Aurbacher et al. (2024) and Hüttel and Hess (2024) both stress the need for transparency in data, code, and analysis, as well as the integration of statistical thinking and open science practices into training and research. As a community, we also need to embrace these changes and move towards open science practices.

Acknowledgement: We thank the anonymous reviewers and editors for their time and constructive feedback on earlier versions of this manuscript. Their comments and suggestions helped improve the clarity and presentation of this work.

Received January 2026 *Revisions Accepted February 2026*

REFERENCES

Ankel-Peters, J., Brodeur, A., Dreber, A., Johannesson, M., Neubauer, F., & Rose, J. (2025). A protocol for structured robustness reproductions and replicability assessments. Q Open, 5, qoaf004. https://doi.org/10.1093/qopen/qoaf004

Aurbacher, J., Bahrs, E., Banse, M., Hess, S., Hirsch, S., Hüttel, S., Latacz-Lohmann, U., Mußhoff, O., Odening, M., & Teuber, R. (2024). Comments on the p-Value Debate and Good Statistical Practice. German Journal of Agricultural Economics, 73(1), 1–3. https://doi.org/10.52825/gjae.v73i1.9881

Betensky Rebecca A (2019) The p-Value Requires Context, Not a Threshold, The American Statistician, 73:sup1, 115-117, DOI: 10.1080/00031305.2018.1529624

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2016). Star wars: The empirics strike back. American Economic Journal: Applied Economics, 8(1), 1-32.

Cumming G. New statistics: why and how. Psychol Sci. 2014;25:7-29.

Finger, R., Henningsen, A., Höhler, J., Huber, R., Rommel, J., & Grebitus, C. (2024). Open science in agricultural economics. Q Open, 5, qoae029. https://doi.org/10.1093/qopen/qoae029

Haaf, J. M., Ly, A., & Wagenmakers, E. J. (2019). Retire significance, but still test hypotheses. Nature, 567(7749), 461-462.

Heckelei, T., Hüttel, S., Odening, M., & Rommel, J. (2021). The replicability crisis and the p-value debate–what are the consequences for the agricultural and food economics community? (No. 1548-2021-3222).

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

Hirschauer, N., Grüner, S., Mußhoff, O., & Becker, C. (2019). Twenty steps towards an adequate inferential interpretation of p-values in econometrics. Jahrbücher für Nationalökonomie und Statistik, 239(4), 703-721.

Hüttel, S., & Hess, S. (2024). Are lessons being learnt from the replication crisis or will the revolution devour its children? Open Q science from the editor's perspective. Q Open, 5, qoae019. https://doi.org/10.1093/qopen/qoae0192

Hung HMJ, ONeill Bauer, Kohne P. The behavior of the P-value when the alternative hypothesis is true. Biometerics. 1997;53:11-12.

Imbens, G. W. (2021). Statistical significance, p-values, and the reporting of uncertainty. Journal of Economic Perspectives, 35(3), 157-74.

Korn, E. L., & Graubard, B. I. (1990). Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni t statistics. The American Statistician, 44(4), 270-276.

Krueger, J. I., & Heck, P. R. (2019). Putting the p-value in its place. The American Statistician, 73(sup1), 122-128.

Leamer, E. E. (1983). Let's take the con out of econometrics. The American Economic Review, 73(1), 31-43.

Leone, R. P., & Schultz, R. L. (1980). A study of marketing generalizations. Journal of Marketing, 44(1), 10-18.

Lindena, T., & Hess, S. (2022). Is animal welfare better on smaller dairy farms? Evidence from 3,085 dairy farms in Germany. Journal of Dairy Science, 105(11), 8924–8945. https://doi.org/10.3168/jds.2022-21906

Mervis, J. (2014). Why null results rarely see the light of day. 992-992.

Murdoch DJ, Yu-Ling Tsai Y, Adcock J. P-values are random variables. Am Statistician. 2008;62:242-243.

Sapra, R. L., & Nundy, S. (2018). Why the p-value is under fire?. Current Medicine Research and Practice, 8(6), 222-229.

Tong Christopher (2019) Statistical Inference Enables Bad Science;Statistical Thinking Enables Good Science, The American Statistician, 73:sup1, 246-261, DOI: 10.1080/00031305.2018.1518264

Verhulst B. In Defense of P Values. AANA J. 2016 Oct;84(5):305-308. PMID: 28366961; PMCID: PMC5375179.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p< 0.05". The American Statistician, 73(sup1), 1-19.