IS THERE A CULT OF STATISTICAL SIGNIFICANCE IN AGRICULTURAL ECONOMICS?

Jens Rommel and Meike Weltin

jens.rommel@zalf.de

Leibniz Centre for Agricultural Landscape Research

2017

Vortrag anlässlich der 57. Jahrestagung der GEWISOLA
(Gesellschaft für Wirtschafts- und Sozialwissenschaften des Landbaues e.V.)
und der 27. Jahrestagung der ÖGA
(Österreichische Gesellschaft für Agrarökonomie)
„Agrar- und Ernährungswirtschaft zwischen Ressourceneffizienz und gesellschaftlichen Erwartungen“
Weihenstephan, 13. bis 15. September 2017

Copyright 2017 by authors. All rights reserved. Readers may make verbatim copies of this document for non-commercial purposes by any means, provided that this copyright notice appears on all such copies.

IS THERE A CULT OF STATISTICAL SIGNIFICANCE IN AGRICULTURAL ECONOMICS?

Abstract

In an analysis of articles published in ten years of the American Economic Review, Deirdre McCloskey and Stephen Ziliak have shown that economists often fail to adequately distinguish economic and statistical significance. In this paper, we briefly review their arguments and develop a ten-item questionnaire on the statistical practice in the Agricultural Economics community. We apply our questionnaire to the 2015 volumes of the American Journal of Agricultural Economics, the European Review of Agricultural Economics, the Journal of Agricultural Economics, and the American Economic Review. We specifically focus on the "sizeless stare" and the negligence of economic significance. Our initial results indicate that there is room of improvement in statistical practice. Empirical papers rarely consider the power of statistical tests or run simulations. The economic consequences of estimation results are often not adequately addressed. We discuss the implications of our findings for the publication process and teaching in Agricultural Economics.

Keywords

Econometrics, Economic significance, Methodology, p-hacking, Regression, Statistics.

1 Introduction

In their seminal contributions, MCCLOSKEY AND ZILIAK (1996) and ZILIAK AND MCCLOSKEY (2004, 2008) present a critique of the abuse of the concept of statistical significance in empirical Economics. Based on data from the 1980s (MCCLOSKEY AND ZILIAK, 1996) and 1990s (ZILIAK AND MCCLOSKEY, 2004) volumes of the American Economic Review, they show that economists frequently fail to adequately distinguish statistical and economic significance. In a regression, a statistically significant coefficient may be small in magnitude and will, thus, have only a small effect on an outcome variable of interest. It may also be large and statistically not significant. Empirically working economists should primarily be concerned with the economic consequences of their estimates. An overemphasis of statistical significance may obstruct scientific knowledge production by introducing publication biases (FRANCO ET AL., 2014), a lock-in and path dependence of journals publishing only statistically significant findings (ALTMAN, 2004), and the manipulation of specifications in favor of "significant findings" by researchers, commonly known as p-hacking (cf. VERESOGLOU, 2015; BRODEUR ET AL., 2016).

The misinterpretation of p-values has recently received notable attention from economists (e.g., KRÄMER, 2011; HIRSCHAUER ET AL., 2016a,b, 2017) and scholars from other disciplines (e.g., GREENLAND ET AL., 2016; WASSERSTEIN AND LAZAR, 2016). However, little is known about the scope of the problem in Agricultural Economics. Agricultural economists frequently use econometric methods and take some pride in applying Econom(etr)ics to real-world problems and in informing agricultural policy. It is the aim of this paper to present an explorative overview on the statistical state of affairs in Agricultural Economics. Specifically, we focus on the "sizeless stare" (cf. HIRSCHAUER ET AL., 2017) at coefficient estimates.

Inspired by the seminal work of MCCLOSKEY AND ZILIAK (1996) and ZILIAK AND MCCLOSKEY (2004), we develop a questionnaire on the use of the concepts of statistical and economic significance. We apply this questionnaire to the 2015 volume of three leading journals of the Agricultural Economics community, namely the American Journal of

Agricultural Economics (AJAE), the European Review of Agricultural Economics (ERAE), and the Journal of Agricultural Economics (JAE). We also compare the data against the benchmark of the American Economic Review (AER) which allows us to assess as to how far things are better or worse in field journals.

2 Empirical approach

Our initial data were all published articles in the 2015 volumes of the AER, the AJAE, the ERAE, and the JAE which are among the leading journals in Agricultural Economics (cf. HERRMANN ET AL., 2012; RIGBY ET AL., 2015). In a first step, we downloaded all articles, using the Scopus data base. Articles were then screened for empirical papers which use econometric methods such as regression analysis. The final sample that met these criteria consisted of 151 articles (48 in the AER, 52 in the AJAE, 20 in the ERAE, and 31 in the JAE). Bibliographic data for these papers were collected, and one half of the sample was assigned to each of the two authors for coding. We used random assignment to articles in order to avoid correlation of the coding author with journals, publishing dates, etc.

Our questionnaire is a simplified version of the original questionnaire (MCCLOSKEY AND ZILIAK, 1996) which consisted of 19 yes or no-questions, e.g., "Avoid choosing variables for inclusion solely on the basis of statistical significance?" (question 14). Questions in the original study are constructed in a way that a "yes" response indicates the desirable practice which allows for the calculation of aggregate scores of "good econometric practice" for which a higher value indicates better quality. We follow this approach, but limit our questionnaire to ten questions. In addition, we survey the minimum and maximum number of observations used for regression models, and we compile a more detailed overview on the use of summary statistics. All questions were tested on a small sub-sample by the authors, discussed, and adjusted for their practicality. After all articles were coded, we tested for differences between the two coding authors and discussed, reassessed, and recoded questions until we reached convergence in the data.

Most questions are coded based on a simple keyword search. We mainly focus on the abstract, results, discussion, and conclusion sections of the papers, as well as regression tables and descriptive statistics. Our first questions address statistical background information on the paper, such as number of observations and the use of summary statistics. In four sub-questions on summary statistics, we ask if the paper provides mean and standard deviation, the number of observations per variable, the range (maximum and minimum values), and the median or percentiles of explanatory variables which allows readers to assess distributional properties of the data and potential issues with selection bias. In our data, one notable "best practice case" in this regard is the article by TADELIS AND ZETTELMEYER (2015), as it provides very detailed summary statistics.

Question 3 addresses the awareness of the authors for understanding the ambiguous meanings of the word "significance" by asking: "at first use, do the authors consider statistical significance to be one among other criteria of importance?" We searched for the term "signific*" and considered the first match of significance being used in a statistical sense. Then, we reviewed the two sentences before and after the term for any other criteria of importance, such as the size of an effect or the economic or political consequences of an estimate. One positive example was the study of HACKMAN ET AL. (2015) who write "[t]he estimate in the first column implies that enrolment in the individual market increased by 26.5 percentage points. This is both statistically and economically significant."

Question 4 is concerned with the power of the statistical tests. A careful assessment of the statistical significance of a result should also include a discussion of the power of the tests used. We used the search terms "power," "beta error" and "type II error" and assessed if the

authors have touched upon the issue. One positive example in our data is the article by BONANNO ET AL. (2015).

Question 5 relates to poor statistical practice, such as the ranking of coefficients in regression tables by their p-values or including only statistically significant variables in regression models, often termed forward or backward selection in the literature. We checked the regression tables. Articles which avoid such practices received a positive coding.

Questions 6 and 7 address the magnitudes of regression coefficients and their discussion. We examined whether the authors discussed the relative size of coefficients, for instance, by referring to them as small, large, etc., using keyword search. We also screened the text for numbers to investigate if the exact size of coefficients is discussed in the text. In both cases, one positive example was sufficient to meet our criterion and to receive a positive score. For instance, the statement "the correlation between TseTse suitability and intensive cultivation is negative: a one standard deviation increase in the TSI decreases intensive agriculture by 9 percentage points, which is about one third of the sample mean" (ALSAN, 2015) received a positive assessment.

Question 8 asks whether the paper uses a simulation to assess the plausibility of model estimates. We used the search term "simulat*" to identify respective articles. For example, CASTRIOTA ET AL. (2015) use a simulation based on their estimated model to identify the optimal number of producers in a coalition.

Questions 9 and 10 refer to the relationship and the representation of economic and statistical significance in two key parts of the paper – the conclusion and abstract. Abstracts and conclusions were screened. For articles that combined discussion and conclusion, we referred to the last two paragraphs of the section. Conclusion and abstract were assessed positively if they referred to the exact size of an effect (e.g., in terms of a number) or if they argued for the relative size of an effect (e.g., by using small, large, etc.) in relation to welfare, political or economic implications, trade-offs, or any practical implications. The abstract was also positively assessed if at least the intention to assess welfare effects and economic consequences was mentioned. One positive example reads as follows: "Our results can be employed to approximate the economic gains brought about by 4-H. As an illustration, consider the present value of lifetime earnings for a fourth-grader from an increase of one standard deviation in standardized test scores calculated by Kane and Staiger (2002) at between $90,000 and $210,000. Using a simple back-of-the-envelope calculation, these numbers imply that the present value of lifetime earnings from 4-H participation, according to our preferred model's results for the FCAT reading subtest, is between $6,300 and $14,700, on average" (FLORES-LAGUNO ET AL., 2015). A list of all questions is provided in the appendix.

3 Results

Table 1 displays the main results by question and journal.

Table 1: Main findings by journals

Question	Sample	AER	AJAE	ERAE	JAE
1a) Maximum Number of observations	152,069.63	99,785.02	60,000.98	71,396.37	14,804.21
(mean, median, std. dev.)	2,950	10,356	2,177	1,828	2,807.5
	737,228.1	334,079.4	32,8003.1	234,531.3	47,123.04
1b) Minimum Number of observations	46,121.79	368,417.6	26,372.11	16,999.95	13,295.64
(mean, median, std.		1,140.5	924	1,089	1,392

dev.)	1,103 (211907.8)	1,224,719	149,763.7	37,108.98	47,435.83
			Proportion of papers with positive evaluation (in %)		
2a) Mean	58.28	39.58	71.15	65.00	61.29
2b) Observations per variable	39.07	41.67	48.08	25.00	29.03
2c) Range (Minimum and Maximum)	23.18	10.42	30.77	45.00	16.13
2d) Median or percentiles	16.56	16.67	23.08	5.00	12.90
3) Significance at first use	38.41	52.08	40.38	15.00	29.03
4) Power of the tests	10.60	14.58	11.54	10.00	3.23
5) Inclusion of coefficients	98.01	97.92	100.00	95.00	96.77
6) Relative size of coefficients	80.13	89.58	80.77	75.00	67.74
7) Exact size of coefficients	91.39	100.00	92.31	90.00	77.42
8) Simulation	26.49	31.25	30.77	15.00	19.35
9) Conclusion	56.29	70.83	55.77	40.00	45.16
10) Abstract	31.13	41.67	34.62	5.00	25.81
N	151	48	52	20	31

Surveyed articles, on average, use a relatively high number of observations. Although a majority of the surveyed articles presents mean values for their data, most articles do not report more detailed descriptive statistics. Generally, most articles score high on several of the questions. On a positive note, the inclusion of independent variables based solely on their statistical significance is hardly practiced anymore. Notable omissions exist for the consideration of the power of tests, simulations, an unambiguous use of the word significance, and references to economic significance in a paper's abstract.

To compare journals, we have formed a composite indicator that is defined as the sum of the mean of questions 2a to 2d (not to give too much weight to summary statistics) and questions 3 to 10. The scores are displayed in Table 2.

Table 2: Composite Indicator by Journals

	Mean	**Std. dev.**	**Median**	**Min**	**Max**
Sample	4.77	1.58	4.75	1	8.75
AER	5.31	1.36	5.75	2	8.75
AJAE	5.01	1.49	5	1.5	8

ERAE	3.95	1.32	3.75	1.75	6.75
JAE	4.06	1.79	4	1	7

We cannot reject the null hypothesis of a normal distribution for this indicator (skewness and kurtosis tests for normality; n = 151; χ^2 [d.f. = 2] = 2.05; p = 0.3580). Thus, we use a one-way ANOVA and two-sided two-sample t-Tests to test for differences in means between journals. Differences between journals are jointly statistically significant (one-way ANOVA; F [d.f. = 3] = 6.84; p = 0.0002). We report pair-wise t-Tests in Table 3.

Table 3: Two-sided t-Tests of differences in means

Pair-wise Comparison	Difference in Means	Std. Error	95% CI Lower Bound	95% CI Upper Bound	p-value
AER – AJAE	0.30	0.29	-0.27	0.86	0.3073
AER – ERAE	1.36	0.36	0.64	2.07	0.0003
AER – JAE	1.24	0.35	0.54	1.95	0.0008
AJAE – ERAE	1.06	0.38	0.31	1.82	0.0066
AJAE – JAE	0.95	0.36	0.23	1.67	0.0108
ERAE – JAE	-0.11	0.47	-1.05	0.82	0.8068

One can see from the tables that there are two pairs of journals with similar scores. The AER and the AJAE score approximately one point higher than the ERAE and the JAE. Table 1 suggests that these differences are largely driven by a different treatment of the concept of economic significance vs. statistical significance, because the AER and the AJAE score notably higher in questions 3, 9, and 10 which are concerned with an unambiguous use of the term significance and references to economic significance at first use in the paper, in the conclusion, and in the abstract, respectively.

4 Discussion and concluding remarks

Our initial analysis of three leading journals in Agricultural Economics and the AER has shown that in all journals there is scope for improvement in terms of introducing datasets (descriptive statistics) and a more precise distinction between economic and statistical significance. Statistically significant coefficients may not necessarily have economic relevance, and large – albeit non-significant coefficients – may point towards important economic relationships for which more efficient estimates would be desirable. Furthermore, we find that simulations and power calculations are rarely used. We also find some indication that differences exist between journals, with the AER and AJAE scoring higher than the ERAE and the JAE. Apparently, our dataset is limited to only one volume, and it would be instructive to see the results for a larger dataset or to see developments over time. However,

we believe that our results can provide some initial diagnosis of problems in the misinterpretation of statistical significance. Furthermore, our questionnaire is relatively easy to use and might be used to extend the analysis to other journals and fields in Economics.

We believe that the topic of this paper deserves more attention among empirically working agricultural economists. In the scientific publication process this concerns especially an increase in awareness from authors, editors, and reviewers. For instance, in SAUER AND LATACZ-LOHMANN (2015, fn. 9, p. 160) the authors explicitly acknowledge the role of a reviewer in pointing out the trade-off between a large number of observations and statistical significance: "We are grateful to a reviewer for pointing out the statistical trade-off implied by large datasets: small differences or correlations become statistically significant although those differences/connections are minor in practice."

Journals may ensure best practice for instance by appointing statistical editors who screen papers for methodical weaknesses, as suggested by VON WEHRDEN ET AL. (2015). Journals may also explicitly encourage publication of null results to combat publication bias and p-hacking, as it is common for example in some Experimental Economics journals, such as the Journal of the Economic Science Association or the Journal of Behavioral and Experimental Economics. More generally, corrections for multiple hypotheses tests should become more common in empirical work (cf. LIST ET AL., 2016). Another promising approach is the encouragement of officially registered pre-analysis plans *before* any data are collected (cf. HIRSCHAUER ET AL., 2017; MUNAFÒ ET AL., 2017). Professional statisticians employed by universities and research institutes may also play an important role in supporting empirically working scholars in the study design and analysis phases (MUNAFÒ ET AL., 2017).

An obvious field for action is the training of the next generation of agricultural economists. Various accessible material exists on the misinterpretation of p-values and the malpractices in Economics (e.g., HIRSCHAUER ET AL., 2016a,b) which should be used more widely in teaching. In our view, basic statistical knowledge should be given a higher priority than it is now the case. Teaching on differences between Type I and Type II errors or basic knowledge of Bayesian inference could increase the quality of empirical work. For instance, students should understand that sampling error and subsequently statistical significance play no role in the analysis of population or census data. For the design of their own research, students should also understand that standard errors cannot only be reduced by increasing the sample size but also by reducing noise in the data. Properly designed survey questions that reduce random variation in the data will often have a greater positive impact on statistical power than an increase sample size. In our experience, these trade-offs between data quality and interference deserve far more attention.

References

ALSAN, M. (2015): The effect of the tsetse fly on African development. In: American Economic Review 105(1): 382-410.

ALTMAN, M. (2004): Statistical significance, path dependency, and the culture of journal publication. In: The Journal of Socio-Economics 33(5): 651-663.

BONANNO, A., HUANG, R., & LIU, Y. (2015): Simulating welfare effects of the European nutrition and health claims' regulation: the Italian yogurt market. In: European Review of Agricultural Economics 42(3): 499-533.

BRODEUR, A., Lé, M., SANGNIER, M., & ZYLBERBERG, Y. (2016): Star wars: The empirics strike back. In: American Economic Journal: Applied Economics 8(1): 1-32.

CASTRIOTA, S., & DELMASTRO, M. (2015): The economics of collective reputation: Evidence from the wine industry. In: American Journal of Agricultural Economics 97(2): 469-489.

FLORES-LAGUNES, A., & TIMKO, T. (2015): Does Participation in 4-H Improve Schooling Outcomes? Evidence from Florida. In: American Journal of Agricultural Economics 97(2): 414-434.

FRANCO, A., MALHOTRA, N., & SIMONOVITS, G. (2014): Publication bias in the social sciences: Unlocking the file drawer. In: Science 345(6203): 1502-1505.

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

HACKMANN, M. B., KOLSTAD, J. T., & KOWALSKI, A. E. (2015): Adverse selection and an individual mandate: When theory meets practice. In: American Economic Review 105(3): 1030-1066.

HERRMANN, R., BERG, E., DABBERT, S., PÖCHTRAGER, S., & SALHOFER, K. (2011): Going Beyond Impact Factors: A Survey-based Journal Ranking by Agricultural Economists. In: Journal of Agricultural Economics 62(3): 710-732.

HIRSCHAUER, N., GRÜNER, S., MUßHOFF, O., & BECKER, C. (2017): Pitfalls of significance testing and p-value variability: Implications for statistical inference. The Centre for Statistics working paper series (ZfS Working papers) of the Georg August University Goettingen, No. 18.

HIRSCHAUER, N., MUßHOFF, O., GRÜNER, S., FREY, U., THEESFELD, I., & WAGNER, P. (2016a): Die Interpretation des p-Wertes: Grundsätzliche Missverständnisse. In: Jahrbücher für Nationalökonomie und Statistik 236(5): 557-575.

HIRSCHAUER, N., MUßHOFF, O., GRÜNER, S., FREY, U., THEESFELD, I., & WAGNER, P. (2016b): Grundsätzliche Missverständnisse bei der Interpretation des p-Werts. In: WiSt-Wirtschaftswissenschaftliches Studium 45(8): 407-412.

KRÄMER, W. (2011): The cult of statistical significance: What economists should and should not do to make their data talk. In: Schmollers Jahrbuch 131(3): 455-468.

LIST, J. A., SHAIKH, A. M., & XU, Y. (2016): Multiple hypothesis testing in experimental economics (No. w21875). National Bureau of Economic Research.

MCCLOSKEY, D. N., & ZILIAK, S. T. (1996): The standard error of regressions. In: Journal of Economic Literature 34(1): 97-114.

MUNAFÒ, M. R., NOSEK, B. A., BISHOP, D. V., BUTTON, K. S., CHAMBERS, C. D., DU SERT, N. P., ... & IOANNIDIS, J. P. (2017): A manifesto for reproducible science. In: Nature Human Behaviour 1: 0021.

RIGBY, D., BURTON, M., & LUSK, J. L. (2014): Journals, preferences, and publishing in agricultural and environmental economics. In: American Journal of Agricultural Economics 97(2): 490-509.

SAUER, J., & LATACZ-LOHMANN, U. (2015): Investment, technical change and efficiency: empirical evidence from German dairy production. In: European Review of Agricultural Economics 42(1): 151-175.

TADELIS, S., & ZETTELMEYER, F. (2015): Information disclosure as a matching mechanism: Theory and evidence from a field experiment. In: American Economic Review 105(2): 886-905.

VERESOGLOU, S. D. (2015): P hacking in biology: An open secret. In: Proceedings of the National Academy of Sciences 112(37): E5112-E5113.

VON WEHRDEN, H., SCHULTNER, J., & ABSON, D. J. (2015): A call for statistical editors in ecology. In: Trends in Ecology and Evolution 30(6): 293-294.

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

ZILIAK, S. T., & MCCLOSKEY, D. N. (2004): Size matters: The standard error of regressions in the American Economic Review. In: The Journal of Socio-Economics 33(5): 527-546.

ZILIAK, S. T., & MCCLOSKEY, D. N. (2008): The cult of statistical significance: How the standard error costs us jobs, justice, and lives. University of Michigan Press.

Appendix – List of Questions

Q1: Use a small number of observations, such that statistically significant differences are not found at the conventional levels merely by choosing a large number of observations? (cf. McCloskey and Ziliak, 1996, Question 1)

Q2: Report descriptive statistics for regression variables? (cf. McCloskey and Ziliak, 1996, Question 2)

Q3: At its first use, consider statistical significance to be one among other criteria of importance? (cf. McCloskey and Ziliak, 1996, Question 7)

Q4: Consider the power of the tests? (cf. McCloskey and Ziliak, 1996, Question 9)

Q5: Eschew "sign econometrics," that is, remarking on the sign but not the size of the coefficients? (cf. McCloskey and Ziliak, 1996, Questions 11, 12)

Q6: Discuss the size of the coefficients? (cf. McCloskey and Ziliak, 1996, Question 11, 12)

Q7: Avoid choosing variables for inclusion solely on the basis of statistical significance? (cf. McCloskey and Ziliak, 1996, Question 14)

Q8: Do a simulation to determine whether the coefficients are reasonable? (cf. McCloskey and Ziliak, 1996, Question 17)

Q9: In the conclusions, distinguish between statistical and economic significance? (cf. McCloskey and Ziliak, 1996, Question 18)

Q10: Does the abstract refer to the magnitude/relevance of an effect not only statistical significance and sign?