Dear Professor “Professor’s name”:

We contact you as the first author of the attached article “Article title”.

We are currently preparing a draft version of a paper that analyzes the prevalence of reporting errors in top economics journals. The aim of this project is to improve the reliability of published empirical findings.

We define a reporting error as a mismatch between the significance level reported by means of an eye-catcher (usually asterisks) and the significance level implied by a reported coefficient with respective standard error (or t-value or p-value). We reviewed all articles published in AER, JPE and QJE between 2005 and 2011 that include at least one empirical test. We find that about half of the articles in our sample contain at least one reporting error. However, our analysis does not find evidence for scientific misconduct.

We rather suspect honest mistakes to be the main reason for the diagnosed errors in reporting and we would like to kindly ask for your help to shed light on the potential causes. The aim is to develop recommendations how reporting errors can be reduced in future research. Please find attached a survey as Excel (or open document) spreadsheet that lists diagnosed reporting errors in the abovementioned article as identified by our algorithm. This survey contains two questions:

(1) The reporting error is due to… (e.g. “error in eye-catcher”, “error in coefficient”, “I don’t know”, etc.)
(2) Why is there a reporting error? (e.g. “Error occurred in typesetting by the publisher and remained undetected in proofreading”, “Error occurred while transferring results from statistical software to Word/Latex”, etc.).

We would highly appreciate if you could review the potential reporting errors and indicate the suspected causes. Your answers will be treated confidentially. Thank you very much for your feedback and please do not hesitate to contact us if you have trouble with opening the attached spreadsheet or any other open questions.

In the following, we provide information on our algorithm and the information in the attached spreadsheet:

Statistical values are rounded and usually reported with low precision. For example, a reported coefficient of 0.020 is consistent with any value from the interval [0.0195, 0.0204(9)] (where (9) denotes ‘9 repeating’) and a reported standard error of 0.010 is consistent with any value from the interval [0.0095, 0.0104(9)]. Hence, any t-value from the interval [1.857143, 2.157895] is consistent with this reported coefficient and respective standard error. Using critical values from the standard

normal distribution and, as example, a two-sided test, this t-value interval corresponds to the p-value interval [0.030936, 0.063291].

A reported significance level by means of an eye-catcher also implies an interval of p-values. For example, two and three asterisks usually denote the 0.05 level and 0.01 level of statistical significance, respectively. If the reported coefficient of 0.020 is then labeled with two asterisks, the p-value should lie in the interval [0.01, 0.05]. Please note that our algorithm uses information from the table notes to link eye-catchers with significance levels and can cope with alternative ways of labelling as well.

We diagnose reporting errors by evaluating whether the p-value interval implied by eye-catchers is disjoint from the p-value interval implied by reported coefficients and respective standard errors (or t-values or p-values). In this example, the intervals [0.030936, 0.063291] and [0.01, 0.05] overlap and we do not diagnose a reporting error. If, however, the coefficient of 0.020 with a standard error of 0.010 is reported to be statistically significant at the 0.01 level, the two p-value intervals are disjoint and a reporting error is diagnosed.

Please note that our algorithm uses critical values from the standard normal distribution rather than from the t-distribution with the exact degrees of freedom used for the respective estimate. Therefore, our algorithm may falsely diagnose a reporting error if the degrees of freedom are small. Our analysis of sample sizes suggests that the vast majority of analyzed tests is based on a sufficiently large number of degrees of freedom. However, in some cases low degrees of freedom may be present. Question 1 and 2 of the survey offer answers to deal with this case: ‘There is no reporting error’ and ‘Reporting error is falsely diagnosed due to low degrees of freedom of the corresponding test’, respectively.

Sincerely,

Peter Pütz and Stephan Bruns

--
Peter Pütz
Chair of Statistics
University of Göttingen
Phone.: +49 (0)551 39 24687
Email: ppuetz@uni-goettingen.de