How many researchers use 'boilerplate' statistical analysis sections?

An observational study of papers published in *PLOS ONE*.

Nicole White, Richi Nayak, Adrian Barnett

21/05/2020

An ideal statistical analysis will use appropriate methods to create insights from the data and inform the research questions. Unfortunately many current statistical analyses are far from ideal, with many researchers using the wrong methods, misinterpreting the results, or failing to adequately check their assumptions (Goodman 2008; Leek et al. 2017). Some researchers take a "mechanistic" approach to statistics, copying the few methods they know regardless of their appropriateness, and then going through the motions of the analysis (Stark and Saltelli 2018).

Many researchers may not have received adequate training in research methods, and statistics is something they do with trepidation and even ignorance (Altman 1994; King et al. 2019). However, using the wrong statistical methods can cause real harm (Altman 1994) and bad statistical practices are being to used abet weak science (Stark and Saltelli 2018). Statistical mistakes are a key source of waste in research and partly explain the current reproducibility crisis in science (Allison et al. 2016). Even when the correct methods are used, many researchers fail to describe them adequately, making it difficult to reproduce the results (Ernst and Albers 2017; Zhou and Skidmore 2018).

The International Committee of Medical Journal Editors recommend that researchers should: "Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to judge its appropriateness for the study and to verify the reported results" (ICJME 2019). Although the general lack of statistical understanding from both authors and reviewers means this recommendation may not be checked. A recent survey of editors found that only 23% of health and medical journals used expert statistical review for all articles (Hardwicke and Goodman 2020), which was little different from a survey from 22 years ago (Goodman, Altman, and George 1998).

As statisticians we have heard researchers admit that they sometimes copy-and-paste their statistical methods sections from other papers, regardless of whether they are appropriate. The aim of this paper is to use text-mining methods to estimate the extent that researchers are using 'boilerplate' statistical methods sections. Use of these methods sections indicates that little thought has gone into the statistical analysis.

Methods

Data source

We chose *PLOS ONE* because it is a large multi-disciplinary open access journal, and because the journal laudibly make the full text of papers easily available for analysis via an API. This enabled us to isolate the statistical analysis section of the paper, and also collect descriptive data on the paper (e.g., field of research).

The criteria for publication include, "Experiments, statistics, and other analyses are performed to a high technical standard and are described in sufficient detail". *PLOS ONE* does not use a separate statistical peer reviewer. Articles are handled by an academic editor who selects peer reviewers. *PLOS ONE* expect authors to use reporting guidelines from the EQUATOR network, which were designed in part because of the long history of poor statistical application and reporting in health and medical journals (Altman and Simera 2016).

We used the ‘rplos’ package to read the data from the API into R (Chamberlain, Boettiger, and Ram 2020).

Questions:

- What questions are asked of PLOS ONE Reviewers?
- Is the manuscript technically sound, and do the data support the conclusions?
- Has the statistical analysis been performed appropriately and rigorously?
- Does the manuscript adhere to the PLOS Data Policy?
- Is the manuscript presented in an intelligible fashioin and written in standard English?
- Reference updates to stats reporting guidelines; e.g see:
- https://everyone.plos.org/2019/09/26/new-plos-one-statistical-reporting-guidelines/.
- https://web.archive.org/web/20190607174803/https://journals.plos.org/plosone/s/submission-guidelines
- https://web.archive.org/web/20150507175314/https://journals.plos.org/plosone/s/submission-guidelines

Search strategy

Our search strategy comprised targeted searches of the PLOS API and filtering of full-text articles based on section headings. All full-text Research Articles published in *PLOS ONE* since 2006 were eligible for inclusion.

Initial searches of the PLOS API were completed using the ‘rplos’ package (Chamberlain, Boettiger, and Ram 2020). Submitted search queries focussed on the presence of analysis-related terms anywhere within a full-text article. Individual terms combined the words “data” or “statistical”, with one of: “analysis”, “analyses”, “method”, “methodology” or “model(l)ing”. Results terms were intended to be broad and, by allowing terms to appear anywhere within the article, we accounted for the placement of relevant text in different sections, for example, in the Material and Methods section versus Supplementary Information. Search results were indexed by Digital Object Identifier (DOI), which allowed for removal of duplicate records. Metadata collected for each DOI included journal volume, PLOS subject classification(s) and number of article views since publication.

Full-text XML data were then extracted for all search results for further processing. Data were extracted by DOI and organised by individual section and subsection headings. In order to consistently identify statistical methods sections, partial matching on commonly used sections headings was done to filter results. [TODO add more details of headings/partial matching here. strings were: ‘statistical analysis’,‘statistical analyses’,‘statistical method’,‘statistics’,‘data analysis’, ‘data analyses’].

Text cleaning focussed on standardising notation and references to statistical terminology, whilst minimising major changes to article style and formatting. Details of text cleaning steps, including R syntax, are provided in Supplementary File 1.

Mathematical notation including Greek letters was converted from Unicode characters to plain text (Supplementary File 2). For example, instances of the code ‘<U+03B8>’, representing the Greek letter θ, were replaced with ‘theta’. Common symbols not captured by Unicode blocks were also converted into plain text, using functions available in the ‘textclean’ package (e.g. % to ‘percent’, < to ‘less-than’) (Rinker 2018). General punctuation, in-text references (e.g. [23]) centred equations and remaining non-ascii characters were removed. Text contained inside brackets was retained in the dataset to maximise content for analysis, with brackets removed.

We compiled an extensive list of common methods and related terminology as a way of standardising descriptions of statistical analyses undertaken. An initial list was compiled by calculating individual word frequencies in the data and identifying relevant terms that appeared at least 100 times. Additional terms were added based on index searches of statistical textbooks [TODO need to mention this here? If so, add references] and spell checks (e.g. 'fischer' instead of 'fisher'). Using the final list, we performed multiple searches to identify possible variants of each term and replaced these with a standardised spelling. For example, instances 'chi square' and 'chi squared' were replaced with 'chi-square'. Other variants of terminology considered plurals (e.g. chi squares), combinations of characters and numbers (e.g. chi 2) and combined text (e.g. chisquare).

As a final step, common stop words including pronouns, contractions and selected prepositions were removed from the data. Stop words were identified from lists available from the 'tm' R package (Supplementary File 2). We chose to keep some prepositions included in the available list to minimise potential changes to the context of statistical method(s) being described (e.g. 'between', 'against').

Clustering algorithm

Discussion

The first line in many statistical analysis sections was the software used, implying that the software is the most important detail. As Doug Altman said, "Many people think that all you need to do statistics is a computer and appropriate software" (Altman 1994). This is not the case, and whilst it is important for researchers to mention the software and version used for reproducibility purposes, it is a relatively minor detail compared with detailing what methods were used and why.

Despite the extensive array of tests available, many authors are reporting the same few methods.

Limitations

We did not check whether papers used the correct methods, and for some simple studies a 'boilerplate' statistical methods section would be fine.

We examined papers where there was a statistics section, and we missed papers that used statistical analysis but did not include a statistical analysis section.

We only examined one journal and hence our results may not be generalisable to all journals, especially those that use a statistical reviewer for all papers.

References

Allison, David B., Andrew W. Brown, Brandon J. George, and Kathryn A. Kaiser. 2016. "Reproducibility: A Tragedy of Errors." *Nature* 530 (7588): 27–29. https://doi.org/10.1038/530027a.

Altman, D G. 1994. "The Scandal of Poor Medical Research." *BMJ* 308 (6924): 283–84. https://doi.org/10.1136/bmj.308.6924.283.

Altman, Douglas G, and Iveta Simera. 2016. "A History of the Evolution of Guidelines for Reporting Medical Research: The Long Road to the EQUATOR Network." *Journal of the Royal Society of Medicine* 109 (2): 67–77. https://doi.org/10.1177/0141076815625599.

Chamberlain, Scott, Carl Boettiger, and Karthik Ram. 2020. *Rplos: Interface to the Search Api for 'Plos' Journals*. https://CRAN.R-project.org/package=rplos.

Ernst, Anja F., and Casper J. Albers. 2017. "Regression Assumptions in Clinical Psychology Research Practicea Systematic Review of Common Misconceptions." *PeerJ* 5: e3323. https://doi.org/10.7717/peerj.3323.

Goodman, Steven. 2008. “A Dirty Dozen: Twelve P-Value Misconceptions.” *Seminars in Hematology* 45 (3): 135–40. https://doi.org/10.1053/j.seminhematol.2008.04.003.

Goodman, Steven N., Douglas G. Altman, and Stephen L. George. 1998. “Statistical Reviewing Policies of Medical Journals.” *Journal of General Internal Medicine* 13 (11): 753–56. https://doi.org/10.1046/j.1525-1497.1998.00227.x.

Hardwicke, Tom E, and Steve Goodman. 2020. “How Often Do Leading Biomedical Journalsuse Statistical Experts to Evaluate Statistical Methods? The Results of a Survey.” MetaArXiv. https://doi.org/10.31222/osf.io/z27u4.

ICJME. 2019. “Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals.” http://www.icmje.org/icmje-recommendations.pdf.

King, Kevin M., Michael D. Pullmann, Aaron R. Lyon, Shannon Dorsey, and Cara C. Lewis. 2019. “Using Implementation Science to Close the Gap Between the Optimal and Typical Practice of Quantitative Methods in Clinical Science.” *Journal of Abnormal Psychology* 128 (6): 547–62. https://doi.org/10.1037/abn0000417.

Leek, Jeff, Blakeley B. McShane, Andrew Gelman, David Colquhoun, Michèle B. Nuijten, and Steven N. Goodman. 2017. “Five Ways to Fix Statistics.” *Nature* 551 (7682): 557–59. https://doi.org/10.1038/d41586-017-07522-z.

Rinker, Tyler W. 2018. *textclean: Text Cleaning Tools*. Buffalo, New York. https://github.com/trinker/textclean.

Stark, Philip B., and Andrea Saltelli. 2018. “Cargo-Cult Statistics and Scientific Crisis.” *Significance* 15 (4): 40–43. https://doi.org/10.1111/j.1740-9713.2018.01174.x.

Zhou, Yuanyuan, and Susan Skidmore. 2018. “A Reassessment of Anova Reporting Practices: A Review of Three Apa Journals.” *Journal of Methods and Measurement in the Social Sciences* 8 (1): 3–19. https://doi.org/10.2458/v8i1.22019.