Building Bridges with Transparent Statistics

Shion Guha
Marquette University
shion.guha@marquette.edu

Matthew Kay
University of Michigan
mjskay@umich.edu

Steve Haroz
Sorbonne University
stats@steveharoz.com

Pierre Dragicevic
Inria, France
pierre.dragicevic@inria.fr

Abstract

Transparent statistics is a philosophy of statistical reporting whose purpose is scientific advancement rather than persuasion. This position paper proposes that doing better and more transparent statistics can act as an aspect of the bridge that we propose to build to bridge AIS SIGHCI and ACM SIGCHI. Over the past few years, both communities have developed disparate norms of statistics and statistical reporting that often serve to divide, rather than unite them. Recognizing limitations and accepting tradeoffs between these norms can address all points of the three horned problem of (a) precision and scientific rigor, (b) realism of context and (c) generalizability of findings. Our recently developed philosophy of transparent statistics can play an integral role in this effort.

Author Keywords

Statistics; methodology; user studies.

ACM Classification Keywords

H.5.m [Information interfaces and presentation (e.g., HCI)]

submitted to Workshop on Bridging HCI at CHI'18

Motivation

Empirical studies in HCI typically consist of solitary experiments analyzed through null hypothesis significance testing (NHST). However, this traditional approach is under growing criticism at both communities, SIGCHI and SIGHCI, [11, 3, 7, 12] and has been strongly criticized for more than 50 years in other fields [13, 5, 4].

Both communities argue that there are massive problems with current practices that include [7, 14, 11, 12]:

- The use of statistical constructs (e.g, *p*-values) that most researchers have trouble grasping intuitively
- Overemphasis on conveying evidence and numbers rather than useful information and generalizable conclusions, leading to tedious *p*-cluttered reports
- Dichotomous thinking, i.e., thinking of hypotheses as either true or false, and of effects and evidence as either existing or not existing
- Undisclosed flexibility in data analyses, yielding cherry-picked results or *p*-hacking (even if unintentional)
- Simplistic criteria for paper acceptance (e.g., looking at whether results are "significant") leading to positive results bias, and thus an incomplete and distorted literature
- A lack of focus on research as a cumulative and collective enterprise, including a lack of incentives for sharing experimental data and study materials, a lack of replication,

and virtually no meta-analysis.

Problems with statistics in HCI extend beyond mere procedural mistakes committed by researchers who might need more statistical training. We believe these are deeper issues worthy of a conversation about how to reform the prevalent methods in both communities that can act as a buttress to bridge them. This is especially true for each prong of the three horned problem. Precision and scientific rigor suffers if appropriate analysis is not conducted while a lack of understanding, training and education is a severe impediment to study contexts realistically. Finally, this leads to over or under claiming of results that undermine confidence in the generalizability of results.

What is Transparent Statistics?

Our use of the term *transparent statistics* is not meant to imply that statistical reports at HCI are necessarily opaque. Instead, it aims to emphasize transparency in reporting. More specifically, we propose to refer to transparent statistics as *a philosophy of statistical reporting whose purpose is to advance scientific knowledge rather than to persuade*. Although transparent statistics recognizes that rhetoric plays a major role in scientific writing [1], it dictates that when persuasion is at odds with the dissemination of clear and complete knowledge, the latter should prevail. For example, when empirical data provides incomplete or mixed evidence, a transparent investigator should refrain from drawing definitive conclusions and instead communicate all relevant information *"in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions"* [8]. Transparent statistics puts clarity before messiness, and messiness before false clarity—study results are often disappointingly complex, but in transparent statistics the quest for scientific truth prevails over *"aesthetic criteria of novelty, narrative facility, and perfection"* [9].

Acknowledging the messiness of results is often at odds with our desire to make strong, definitive statements (*"technique A outperforms technique B"*). But conveying uncertainty more faithfully represents our results and even makes them more useful: practitioners do not want to know if p is less than .05; they want to know by how much does technique A improve over technique B (plus-or-minus some error) so that they can perform a cost-benefit analysis and decide whether to adopt it. Besides advancing clarity within our field, transparent statistics can help address another existential crisis for HCI—impact on real-world systems—by expressing our results in statistical language that is amenable to assessing practical significance.

How to Move Towards Transparent Statistics?

Reporting Transparent Statistics

Transparent statistics are about both *what* we report and *how* we report it. While methodologists have been discussing *what* to report to maximize transparency (e.g, communicating simple/standardized effect sizes with frequentist/Bayesian interval estimates, clearly distinguishing between planned and unplanned analyses), HCI can advance guidelines for *how* to report transparent statistics in a user-friendly manner. For instance, clear, straightforward graphical communication of effects can be written into modern reporting guidelines [7]. These approaches could become both the standard within HCI and the standard we aspire to create through new statistical tools—what if the output of any procedure in a statistical package was an annotated, self-explanatory visualization, rather than a cryptic table? This approach may make some uncomfortable, as guidelines already exist that insist upon many orthodox practices that can be harmful to transparent statistical communication. These older standards lead to ubiquitous impenetrable results sections that are peppered with numerical statistical results. We plan to discuss how authors can educate reviewers when writing results that do not follow old norms. This includes amassing a set of citations that lend credence to (currently) unorthodox approaches; e.g., essays by advocates of estimation [5, 7] and of Bayesian methods [12].

Having more papers in the field using these methods can also help. Done well, these methods could speak for themselves. Clearer communication (with relevant citations) can be enough to convince reviewers simply through the deeper understanding they gain from the work. However, some rethinking is still necessary: a wide confidence interval that just overlaps 0 in a small-n study is more honest than a *p* value just above .05 (and better informs future meta- or Bayesian analysis), but might feel like a lackluster result to a reviewer used to thinking in binary rejection criteria.

Emphasizing Practical Significance over Testing

In contrast to a focus on binary testing (is A better than B?), transparent statistics emphasize effect size (how much better?) and uncertainty (what are the upper and lower bounds on the difference?). These inform us on practical significance: is the difference large enough, and are we certain enough to act on it? Given an estimated difference between two conditions, a practitioner could apply a cost function to decide whether the increase in performance is worth the cost of switching to a new interface or technique. Cost/benefit analysis, not statistical significance, is the language of industry, and therefore one way for results from HCI to make it out of the lab and into real-world systems.

Training and Education

Training and education is an important part of this debate. Many HCI researchers learn statistics in one of two ways: through an applied statistics course (for non-statisticians) taught by statisticians, or through a course (or part of a course) taught by an HCI or computer science professor in their home departments. The latter approach can perpetuate old norms in the field which, as we have argued, need to be reexamined and reformed. How can we better integrate transparent statistics education into HCI curricula (as is becoming more common in other fields)? This is especially pertinent given that the home departments of AIS SIGHCI and ACM SIGCHI researchers are quite different i.e. the former mostly being in schools of business and/or departments of information science and the latter in computer science and some information science departments.

Open Data and Replications

While clear communication of statistical analyses is critical, publishing the underlying data allows those analyses to be verified. Open data allows readers to answer questions about aspects of analysis that may be missing from the text. It also allows subsequent researchers to analyze facets of the data that the original researchers did not examine, perform meta analyses on multiple publications, and more easily use existing data to form priors for future Bayesian analyses. Science is a cumulative and collective enterprise.

Nevertheless, questions have arisen regarding the costs and merits of open data. Documenting and anonymizing data takes time. There are also limits to its error-correcting ability. While reexamination of an experiment's data can help detect mistakes, problems can occur in any stage of an experiment, including incorrect stimulus presentation, incorrect response recording, and the possibility of a statistical fluke. Furthermore, reusing materials can propagate these mistakes across multiple publications. Overcoming these problems requires complete experiment replication [14], not just reproduction of the analysis.

Transparent Conclusions

While our focus is on reporting and analysis, transparent statistics necessarily go hand-in-hand with well-designed and implemented experiments with reasonable conclusions. The conclusions should be nuanced and not convey more certainty than the results [7]. Overgeneralizing results should also be avoided. If a technique is beneficial in one implementation or task [10], how can theory be used to make conclusions that extend beyond the narrow scope of the experiment? How we write about generalizability typically follows uncodified conventions that depend on whether the research took a hypothesis-driven or data-driven approach—themselves direct successors of deductive and inductive reasoning [6]. Failure to differentiate the two often results in overclaiming about the external validity or generalizability of human-centered research [2]. Transparency is increased if research projects describe (1) how they connect to and build off of existing theories and (2) why or if the conclusions are externally valid.

Conclusion

We propose that emphasizing the philosophy of transparent statistics at this workshop and can lead to the clarity, reliability, and impact of quantitative results in both communities.

References

[1] Robert P Abelson. 2012. *Statistics as principled argument*. Psychology Press.

[2] Michael S Bernstein, Mark S Ackerman, Ed H Chi, and Robert C Miller. 2011. The trouble with social computing systems research. In *CHI'11 Extended Abstracts*. ACM, 389–398.

[3] Paul Cairns. 2007. HCI... not as it should be: inferential statistics in HCI research. In *People and Computers: HCI... but not as we know it*, Vol. 1. 195–201.

[4] Open Science Collaboration and others. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251, aac4716.

[5] Geoff Cumming. 2013. The new statistics why and how. *Psychological science*.

[6] Andrew Dillon and Charles Watson. 1996. User analysis in HCI–the historical lessons from individual differences research. *Int J Human-Comp Studies* 45,6.

[7] Pierre Dragicevic. 2016. Fair Statistical Communication in HCI. In *Modern Statistical Methods for HCI*, J. Robertson and M.C. Kaptein (Eds.). Springer. tinyurl.com/fairstats-author In press.

[8] Ronald Fisher. 1955. Statistical methods and scientific induction. *Journal of the Royal Statistical Society. Series B (Methodological)*, 69–78.

[9] Roger Giner-Sorolla. 2012. Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science* 7, 6, 562–571.

[10] Steve Haroz and David Whitney. 2012. How capacity limits of attention influence information visualization effectiveness. *IEEE TVCG* 18, 12, 2402–2410.

[11] Maurits Kaptein and Judy Robertson. 2012. Rethinking statistical analysis methods for CHI. In *CHI 2012*.

[12] Matthew Kay, Gregory Nelson, and Eric Hekler. 2016. Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. In *CHI 2016*.

[13] Rex B Kline, American Psychological Association, and others. 2004. Beyond significance testing: Reforming data analysis methods in behavioral research.

[14] Max Wilson, Wendy Mackay, Ed Chi, Michael Bernstein, and Jeffrey Nichols. 2012. RepliCHI SIG: From a panel to a new submission venue for replication. In *CHI'12 Extended Abstracts*.