Harvard Data Science Review • Issue 3.2, Spring 2021

Response to Andrew Fowlie's Comments

Thomas Junk¹ Louis Lyons²

¹Fermilab, Fermi National Accelerator Laboratory, Fermi Research Alliance, United States
Department of Energy Office of Science, Batavia, Illinois, United States of America,

²Department of Physics, University of Oxford, Oxford, England, United Kingdom

Published on: Apr 30, 2021

DOI: <https://doi.org/10.1162/99608f92.ebd73e8c>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Keywords: reliability, reproducibility, replication, particle physics

We thank Andrew Fowlie for his thoughtful comments on [our article](#). Since our paper is primarily about reproducibility and replication, our explanation of the procedures used for making a discovery in particle physics are somewhat abbreviated. In particular, although the 5σ requirement on a z -score is a primary criterion, this is usually supplemented by other information, such as the goodness of fit for the alternative hypothesis. (See [Section 3.2](#), the fifth bullet of the ‘Hypothesis Testing’ subsection).

We address his criticisms below:

A) The usual interpretation of choosing $p_0 < \alpha = 2.87 \times 10^{-7}$ in experimental particle physics is the error-theoretic one and not Fisher’s evidential interpretation. We do however thank Fowlie for making us realize that the subsection label ‘Using p Values to Quantify Discovery Significance’ could be better written as ‘Using the p Value as a Tool for Discovery,’ which is a better match for the text that follows and which matches our logic for testing alternative hypotheses as well.

B) Fowlie says that p values overemphasize the significance of possible new phenomena, and quotes articles pointing out that numerically they tend to be smaller than the corresponding likelihood ratios or Bayesian posterior odds ratios. Clearly these values are going to differ; apart from all else, the likelihood ratio involves the alternative hypothesis, while p_0 is just for the null. This does not invalidate p values. Furthermore, Bayesian methods for discovery also introduce dependence on the choice of prior, which is more pronounced for hypothesis testing than for parameter estimation. We thus consider it unlikely that Fowlie’s final suggestion that p values be replaced will be adopted for hypothesis testing in experimental particle physics.

C) We are accused of sometimes equating p values with the probability of the hypothesis being true. We clearly state that this is not so (See our [Section 3.4](#) and footnote 4). We think that some ambiguity may be created by our statement that effects with p values corresponding to more than 3σ constitutes ‘evidence’ against the null hypotheses. We are not using the word ‘evidence’ in a technical Bayesian sense, but merely to distinguish it from the stronger ‘discovery’ claim for a $> 5\sigma$ effect.

D) We acknowledge that p values can be constructed in ways that are incoherent, though the examples Fowlie points to generally do not arise in particle physics. For example, when testing an interval hypothesis, the customary strategy is to test each point hypothesis within the interval, such as is the case when excluding a range of masses for a hypothetical particle. More commonly, tests are one-sided such as those on production rates, and we are unaware of instances of incoherence in the procedures generally used.

E) Fowlie points out that there are several observations over the years of possible new effects at p values corresponding to z -scores between three and five. As Fowlie himself says, this could well be due to the large

number of searches for new physics carried out in particle physics. We mention the relevance of this ‘Look Elsewhere Effect’ in [Section 3.5](#). These random effects are part of the reason we use such a stringent cut of $p_0 < \alpha = 2.87 \times 10^{-7}$ for discovery claims. We also prefer keeping the Type-1 error rate well defined, even as the sample size changes. Otherwise, we would have to adjust published p values and limits when new results of any sort are made available. If a data dredger selects some results and not others from among those that are published, or that have been derived from published datasets, then the appropriate Look-Elsewhere Effect correction needs to be applied at that stage. Experimental particle physicists set a good example by publishing results regardless of the experimental outcome.

F) We believe it is vital to preserve the data and software of our analyses. This will enable experimental data being available for further study and comparison with future data and theories, primarily to members of the collaboration originally responsible for the data, but also for other experimental and theoretical particle physicists, and for the general public. As Fowlie points out, it will be necessary to judge cautiously any anomalous effects uncovered by non-blind trawling through the data. Misunderstandings of systematic effects by downstream consumers of the data may be a bigger issue than the statistical ones, especially given the complexity of the experimental apparatus and the physics processes, and the associated systematic uncertainties.

G) Imposing a very high standard on the use of the word ‘discovery’ reduces the false-discovery rate but does not make it zero. High-significance ‘discoveries’ that have not been replicated almost always are caused by poorly-understood systematic effects, a consequence of the high statistical significance threshold. Systematic uncertainties are difficult to estimate properly, and we hope that our article explains some of the things particle physicists have learned over the years in dealing with their challenges.

H) The 750 GeV di-photon excess in the ATLAS and CMS data was ultimately explained as a statistical fluctuation, which was ascertained by collecting more data and by looking for systematic effects. It is not very satisfying, but it sometimes happens that a conflict between the results from different datasets from the same detector, or from different experiments, has no obvious systematic explanation and ‘statistical fluctuation’ becomes the only possibility.

We believe that effort expended on following up on possible hints of new physics is well spent. Many of the explanations proposed after the fact do resemble HARKing, though they are often proposed in the context of earlier models that have not been committed to a file drawer. They also serve to remind us that if a signal is seen for a new particle or interaction, many explanations may be possible and further experimental work must be done in order to distinguish among the possibilities.

Acknowledgments

This document was prepared by Thomas R. Junk and Louis Lyons using the resources of the Fermi National Accelerator Laboratory (Fermilab), a U.S. Department of Energy, Office of Science, HEP User Facility. Fermilab is managed by Fermi Research Alliance, LLC (FRA), acting under Contract No. DE-AC02-07CH11359.

Disclosure Statement

Thomas Junk and Louis Lyons have no financial or non-financial disclosures to share for this article.

©2021 Thomas Junk and Louis Lyons. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.