18-Mar-2014

RE: TAS-13-205, "Statistical Detection of Potentially Fabricated Numerical Data: A Case Study"

Dear Dr Hill: Although I am sorry for the delay in evaluating your paper, it is my experience as both a writer and editor that such delays are often a consequence of what the paper contains.

In statistics we have a long history of distinguishing between "applied" papers and methodological/theoretical" papers. Despite your subtitle, the text of your paper seems to fit into the latter category. In fact, the paper falls in the former. This is relevant because the standards for judging the papers are somewhat different. In the latter case the criteria revolve around whether the proposed methods provide reasonable tools for statisticians to use. In the former case the criteria include whether the tools used have led to well justified conclusions. The initial evaluations by the AE and referees understandably mistook the paper to be methodological.

My biggest technical concern is that you are treating your tests as hypothesis tests rather than significance tests in a situation for which that is inappropriate. The mere fact that you can reject the null model of randomness does not imply an alternative model of intentional cheating. There are always a variety of things that could go wrong with the null model and it is a matter of serious scientific investigation to tease out what has, in fact, gone wrong in any particular application.

My second biggest technical concern is that you appear to have assumed that cheating existed and looked for test statistics that corroborated your prior opinion. The situation is similar to a one-way ANOVA. When comparing the means for two groups, a simple t test suffices. However, if you intentionally choose to compare the largest and smallest observed means, the usual null distribution is no longer appropriate and the t distribution must be replaced by the studentized range. If you select a test statistic because you think it is likely to show significance in the data at hand, that fact invalidates use of the random digit model as an appropriate null hypothesis distribution.

I also have concerns that the tone of the article is inappropriate for The American Statistician. This began when I realized that it would be possible to identify the "culprit" from the information given in your article. The American Statistician is interested in discussions of good and bad scientific methods. It is not interested in exposing moral turpitude by individuals.

I might also mention that this evaluation was developed while I was waiting for Dr. Speed's (attached) evaluation and is, therefore, independent of it. I consider the fact that it agrees in substance with Dr. Speed's to be significant. For these

reasons I have concluded that your paper is inappropriate for The American Statistician.

Nonetheless, we thank you for submitting your manuscript for possible publication in The American Statistician (TAS). The reviewer comments are included at the bottom of this letter or as attachments.

I realize that this decision will be disappointing for you. Whatever you decide to do with the paper, I wish you the best in finding it a proper venue.

Sincerely,

Ronald Christensen, Editor
The American Statistician
fletcher@stat.unm.edu

Reviewer(s)' Comments to Author:

Reviewer: 1

Comments to the Author
TAS-13-205

Overall Comments

This is an interesting paper and overall it shows an interesting approach, particularly to analysis of triplicate counts.

I am not a good enough statistician to be sure that the details of the methods described in the appendix and in the text are correct but the overall argument is clearly correct.

It is a clear paper, though it might be able to shortened for the American Statistician readership.

Specific Comments which might merit a response

1 The references to the statistical methods that might be used do not cover the full range, nor do they give some of the older work. It seems immodest to suggest some of the following should be cited but tha range of statistical methods given in these sources is rather greater than implied by these authors, but they may be ignored-

Evans, Stephen, Statistical aspects of the detection of fraud. in Fraud & Misconduct in Medical Research, Eds. Lock, S. and Wells, F. BMJ (1993) pp 61-74.

Evans SJW. The Detection of Fraud. In Encyclopedia of Biostatistics. Eds. Armitage, P & Colton T. 1998; Wiley, Chichester.

Buyse M, George SL, Evans S, Geller NL Ranstam J, Scherrer B, Lesaffre E, Murray G, Edler L, Hutton J, Colton T, Lachenbruch P & Verma BL. The role of Biostatistics in the Prevention, Detection and Treatment of Fraud in Clinical Trials. Statist. Med. 1999; 18:3435-51.

Buyse M & Evans SJW. Fraud in Biostatistics in Clinical Trials. Eds Redmond C & Colton T. Chichester, Wiley. 2001.

Evans, S. Statistical aspects of the detection of fraud. in Fraud & Misconduct in Medical Research, 3rd Edn. Eds. Wells, F, Lock, S. and Farthing M. BMJ (2001) {This book as a whole has much relevant material}

Marc Buyse has commercial software of considerable sophistication for such detection but it may not be acceptable to refer to it.

2 The second issue is whether even simpler tests might be used fir screening triples. Did the authors simply look at the within triple variance and compare the mean variance across the different investigators. If that does not show anything (or if it does), it is worth reporting. (I am inclined to use the data supplied to check it for interest).

3 A third issue is whether the symmetry within the triple is worth looking at. Has a measure of skewness (I think kurtosis may not be reliable) within the triples been looked at?

4 The reason for suggesting these methods is that if crooked investigators learn that using the mean as one of the values can be detected, they will alter the middle value to be slightly different to the mean, but these other methods are less dependent on the exact value of the mean or mid-point being used.

5 Have the authors looked at any of the patterns that might be detected by looking at successive values. The ordering of the data are recorded and if independent there will be certain paterns that will differ from genuine data.

6 Al Marzouki et al used last digit preference between randomised groups, and this is actually somewhat different from looking at last digits in a single group. At least some warning should be given about genuine data having digit preference – See Evans (1993).

Reviewer: 2

Comments to the Author
This is a well-written paper that addresses an important problem: the apparently increasing prevalence of falsified or fabricated data in scientific publications. The authors prevent a convincing case study that the data from one researcher in a laboratory appear to have been fabricated. Their novel methods of testing (1) whether the rounded mean is included among triplicate values or (2) whether the mid-ratio of triplicates lies in a small interval around 0.5, have wide potential applicability even beyond the colony assays and Coulter counts that they present, since many biological experiments use triplicates for their basic measurements.

Minor Concerns:
[1] The authors do not discuss the implications of applying their test routinely to many studies. In the case study that they present, the p-values are so small that any reasonable (or even unreasonable, as they point out) cutoff would identify the data in question as suspect. The difficulty arises in the other direction, with potential "false positive" identification of other suspect data. I don't think that even the firmest believers in the magic frequentist significance value cutoff of $p < 0.05$ would want that value applied to test whether their own data were suspect. Have the authors thought about practical ways to implement their procedure widely without generating false positives that could easily damage researchers careers?

[2] Table 1 needs reformatting. The vertical line at its current position is worse than having no line at all. It might be better to have vertical lies after every two columns, probably extending into the header. In addition, there is an obvious typo in the sentence "lambda continues to decrease after lambda = 4".

Associate Editor
Comments to the Author:

[Ed. Note: The following is the AE's INITIAL evaluation. Following subsequent discussion and evaluation, the AE concurs with my decision.]

This is a well-written paper on an interesting and particularly relevant topic. There have been news stories recently on the public's lack of trust in scientists, and I can't imagine that some of the highly visible reports on data fraud have helped this perception.

Both reviewers have provided useful comments that should be addressed in any revision of this paper.

Some general comments, in order of appearance in the text:

1) p. 7, top: I wondered, as did the first reviewer, if it might also be worthwhile to

examine if some measure of spread in the targeted data sets is consistent with the underlying probability model for the experiment. I would suggest at a minimum that other possibilities (such as those mentioned by the first reviewer) be acknowledged in the discussion section of the paper.

2) p. 12, paragraph starting on line 8: In Section 5.5 you mention that you are restricting your test to triples that contain a gap of at least two or more (see line 47). I assume that this restriction was also applied to calculate the probabilities in the MidProb table, but nowhere in this paragraph is that mentioned. Clearly, you would want the gap of two or more restriction applied to the MidProb calculation as well.

3) p. 13, Section 5.7: As pointed out by the authors, the lambdas are estimated, not known, in the application of this proposed procedure. There is therefore uncertainty in the p-value obtained from the procedure which has not been quantified. As The American Statistician is a statistics journal, I would suggest addressing the issue of quantifying the uncertainty in these p-values. Could a nonparametric or parametric bootstrap approach be applied? Is there a rigorous bounding method?

4) p. 19, Section 5.10, line 53: Use of the word "calculate" here suggests you have an analytic procedure for calculating the probability that the mid-ratio of a triple falls within [0.4,0.6], as you did for calculating the probability that a triple includes its own rounded mean value (Appendix A). Is this the case, or did you do a simulation study to estimate these probabilities in the mid-ratio scenario? If it was a simulation study, you should state what you did to control the error in your estimates. If it is analytic, provide details in an Appendix B (or provide a reference, if such a result has already been published).

5) p. 26, Section 10.3: Please provide all code used to perform calculations in this paper as supplementary material (rather than "available by request"). This will be a useful contribution to other investigators.

6) p. 26, line 31: "Dr Pitt"? This is supposed to be a blinded submission.

7) Appendix: I was curious as to why gaps of size 0 or 1 (for which triplicates automatically contain their rounded mean) were excluded. Does including them in your test for mean-containing triples impact any of the conclusions in this paper (I don't see how it would, but I'm curious)?

8) p. 28: Rigorously prove the statement at the bottom of this page, that you can obtain a value of P(A) accurate to 5 decimal places by choosing N as indicated here.

9) p. 29: Clarify the last sentence. When you say you performed "bootstrap calculations," do you mean you ran a simulation study of 200,000 trials for

various values of lambda to check the resulting estimated probability of mean-containing triples against the analytical calculation? The use of "bootstrap" here could be confused with the resampling procedure of that name, which is not what I think you meant to refer to here.

10) The PDF version of this paper that I downloaded for review contained 152 pages. Clearly not all of these can be published in The American Statistician. All pages after the references (i.e. p. 32 and beyond) should be removed from the paper itself and submitted as supplementary material. You might also think about tightening up the text in the main body of the paper as well, in an attempt to bring the total page count closer to 20-25 pages.

Minor typos:

1) p. 17, line 53: "contain" should be "containing"

2) p. 18, line 5: Insert "Sections" before "5.7"

3) p. 18, line 13: Insert "Section" before "5.6"

4) p. 18, line 21: "sections" should be "Sections"

5) p. 18, line 45: "contain" should be "containing"

6) p. 21, line 23: Delete "of" after "further"

7) p. 22, line 41: Delete "is" after "it"

8) p. 26, line 55: Delete the comma after "is"

9) p. 27, line 13: Should be "Hence P(A_j) = …" and the "A_{jk}" summand on line 16 should be "A_{j,k}"

10) p. 27, line 31: Insert "rounded" between "its" and "mean"

<Report-on-Pitt-Hill.pdf>

Earlier email exchanges

3/18/14 from Christensen

Dear Dr. Hill: I have finally heard back from the reviewer that I refereed to below and have forwarded that evaluation back to the AE to look it over and consult with me before I make my final decision. It should not be very long now. Sincerely, Ronald Christensen.

On Fri, March 14, 2014 2:34 pm, Helene Hill wrote:
> It is now MORE than 5 months since we first submitted our paper. This is a very long time to wait especially if your response is going to be negative. I do hope that you can give us your answer soon.
> Helene
>
> On 3/3/2014 12:07 PM, Ronald Christensen wrote:
> Dear Dr. Hill: When the paper came back on my desk after being evaluated by the AE, I decided that I needed additional perspective so I sent the paper to a prominent statistician who has experience with the corresponding legal issues and not merely the narrow statistical issues. That was about the time I sent you my previous email. I was hoping the person would have responded by now, but I do not think it has been so long that I need to press them for a response. I can, however, use your email as an excuse for contacting the person -- and I will do that.
> Sincerely,
> Ronald Christensen
> TAS Editor
>
> On Mon, March 3, 2014 9:18 am, Helene Hill wrote:
> Dear Dr Christensen,
>
> It is now nearly 5 months since we submitted our MS "Statistical Detection of Potentially Fabricated Numerical Data: A Case Study" to the American Statistician. We hope that you will let us know your decision about it soon.
> Thank you.
>
> Helene Z Hill, PhD
> Professor of Radiology
> Rutgers NJ Medical School
> Newark, NJ 07101-1709