https://dx.doi.org/10.22606/pra.2019.14004

Using Statistics from Binary Variables to Detect Data Anomalies, Even Possibly Fraudulent Research

Walter R. Schumm[1*], Duane W. Crawford[1], and Lorenza Lockett[2]

[1] School of Family Studies and Human Services, College of Health and Human Sciences, Kansas State University, Manhattan, KS 66506, USA
[2] Department of Sociology, Anthropology, and Social Work, College of Arts and Sciences, Kansas State University, Manhattan, KS 66506, USA
Email: schumm@ksu.edu

Abstract. Heathers and his colleagues have proposed a variety of tests to detect inconsistencies in research data, including the GRIM, SPRITE, DEBIT, and RIVETS tests. Here we focus on relatively simple ways of examining binary data for results that are impossible or results that feature inconsistencies, using binomial tests to evaluate whether anomalous results could be explained as random typographical errors. Hypothetical data are used to illustrate our suggested procedures. Advantages and limitations of the approaches are discussed.

Keywords: Research methods; detecting fraudulent research; binary tests; binomial tests; RIVETS, DEBIT, SPRITE, GRIM, and GRIMMER tests

1 Introduction

There appears to be increasing pressure on academic scholars to publish more often, even at lower ranks (Warren, 2019). Such pressure may account, in part, for an increase in the number of scholarly articles that have been retracted on account of scientific misconduct in recent decades; some academics have had dozens of their articles retracted (Fanelli, 2009, 2013; Steen, Casadevall, & Fang, 2013; Stroebe, Postmes, & Spears, 2012) in spite of the serious consequences when scientific misconduct is exposed (Stern, Casadevall, Steen, & Fang, 2014). In many cases, published data have been outright fabricated. What are editors, reviewers, and scholars to do? As we have noted elsewhere, Econ Watch Journal has been providing a valuable service to the social sciences by providing a platform for scholars to critique research post-publication and, in some cases, point out instances of fraudulent research (Schumm, 2015, p. 3). However, other journals should probably take up the same cause, performing a watchdog role for the field of social science, checking on potentially fraudulent research. Editors might need to be careful because some fraudulent research might be authored by racial, sexual, ethnic, gender, or other minorities and could subject them to charges of racism, misogyny, transphobia, or homophobia if their journal helped expose fraudulent research.

1.1 General Tests for Data Anomalies

The GRIM Test. One suggested possibility for detecting fraudulent data is the GRIM test (Brown & Heathers, 2017) in which a mean score can be multiplied by the sample size to obtain (or not) an integer value. The theory is that if the scores are integer values then mean scores should be able to be multiplied by the sample size to attain integer values. For example, suppose a study has 31 cases and a mean score is 2.90; then multiplying 2.90 x 31 will yield 89.9, which is close to the integer of 90, which would suggest a genuine value. However, if you had 31 cases and a mean score of 2.63, then multiplying by 31 would yield 81.5, which is not a likely genuine value (not an integer). While the GRIM test is helpful, its effectiveness decreases with larger samples and it cannot protect against raw data that has been fabricated. One could fabricate raw data such that as long as it fit the range of the data for a variable, it might pass the GRIM test; the GRIM test is better at detecting mean scores that have been fabricated. Aside from the GRIM test, some data might pass such statistical tests but still be suspect. For example, an author might report results for a measure whose scores ran from 0 to 10; if the only scores reported were 0 and 10, with no scores from 1 to 9, the whole point of using an eleven point scale would have been useless. Even if such data appeared to be valid statistically, it would not make much theoretical sense.

The SPRITE Test. Another test for false data is the SPRITE test (Heathers, Anaya, van der Zee, & Brown, 2018; see also the GRIMMER test, Anaya, 2016) which checks both means and standard deviations to see if both are possible for a given sample size. The SPRITE test can help to determine if no possible set of data could yield a particular combination of mean and standard deviation and also performs a GRIM test by itself. However, it can also occur that a particular mean and standard deviation could be calculated from multiple possibilities, some of which might, again, be fabrications of raw data. For ordinal or interval data, the SPRITE test (Heathers et al., 2018) may be useful, although it suffers limitations when sample sizes are larger (e.g. more than 200).

The RIVETS Test. Perhaps the most recent test developed by Brown and Heathers (2019) is the RIVETS test. We tested the idea ourselves by running a regression equation from some available dissertation data, predicting a dependent variable from ten independent variables using SPSS software. In no case did the resulting t-test for each parameter equal exactly the unstandardized regression coefficient divided by the standard error, using what we could see for data up to three decimal points. In other words had we calculated the t-values by hand, they would have looked perfect but would have disagreed with what the computer software obtains by using a deeper level of decimal points and rounding the results. The RIVETS test considers the chances of data in a published article being hand calculated versus software calculated. Since few scholars calculate their data by hand after the advent of inexpensive computer software, there is a risk that "hand calculated" data are actually made up data.

1.2 Tests for Binary Data Anomalies

In 2018, we pointed toward another test for checking for valid data, checking whether standard deviations from binary data fit what would have been predicted by their mean scores (Schumm, Crawford, Higgins, Lockett, AlRashed, & Ateeq, 2018, p. 786). We noted that standard deviations for binary variables should seldom exceed 0.55, so if an article reported a standard deviation of 0.75, it would have to be an error, either a typographical error or possibly falsified data. We included a formula for predicting the standard deviation from the mean, for binary variables (e.g., 0 and 1 being the only possible values).

More recently, Heathers & Brown (2019) have proposed a DEBIT test along the same lines. They report the same formula as the square root of [N/(N-1) times m(1 − m)]. Data that do not fit the expected pattern might indicate rounding errors, unreported missing data, or as Heathers and Brown (2019) call it, altered data. It is possible that means were reported incorrectly or that standard deviations were reported incorrectly, or both. It is possible that sample sizes were reported incorrectly. They noted that standard deviations from grouped data might not fit the mean/SD pattern for the whole sample. Thus, issues are raised with respect to the analysis of multi-level data which includes individual level variables as well as group level variables. Until further research is done with respect to such group-level data, the best data for checking binary patterns would be that data reported for entire samples at the individual level.

2 Objectives/Methods

Our objective is to take some hypothetical binary variable data and show how merely visualizing the data could reveal inconsistencies that might suggest serious problems, even fabrication of data. We will also consider how binary tests can be combined with binomial tests or examinations of internal consistency as ways of detecting data anomalies. We hope that other scholars will take our suggestions as simple ways for checking the legitimacy of reported results in social science. There will always be more complicated ways, but we try to find simpler approaches useful for a wider range of scholars (Schumm, 2018; Schumm & Crawford, 2019).

3 Results

Figure 1 shows a plot of means and standard deviations for a binary variable when N = 100 (A plot for N = 10 is available elsewhere; see Schumm, Crawford, and Lockett, 2019). It is clear from Figure 1 that standard deviations are predicted perfectly from mean scores for binary variables. Table 1 lists some possible data from group and individual level analyses across six hypothetical studies, each over 1,000 cases. Because Heathers and Brown (2019) suggested that standard deviations might differ between grouped and individual data, we compared the mean scores for standard deviations as a function of level of data. A *t*-test ($df = 33$) = 7.80 ($p < .001$) from means of 0.155 ($SD = 0.073$) and 0.363 ($SD = 0.087$) for grouped and individual data, respectively. A similar result was obtained from a nonparametric Mann-Whitney U test, with $z = 4.68$ ($p < .001$).

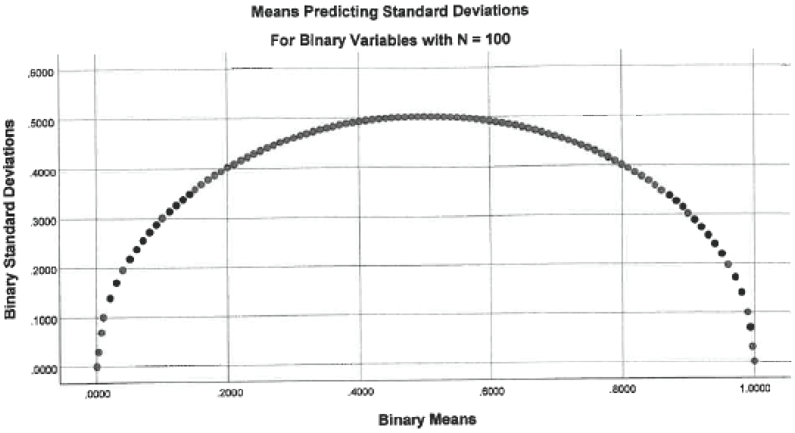

Figure 1. Standard deviations predicted from mean scores for binary variables with N=100

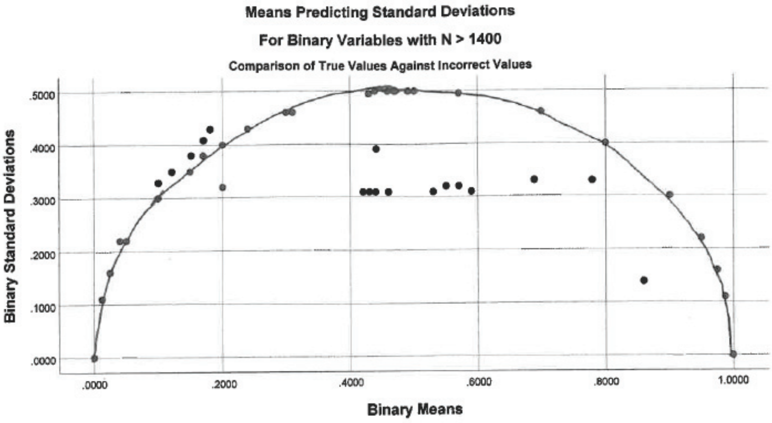

Figure 2. Comparing true values (on curve) against impossible values (off curve) when predicting standard deviations from mean scores of binary variables with N>1400

Table 1. Hypothetical examples of reported binary variable means and standard deviations from six sources, STUDY1 through STUDY6

Binary Variables/N	STUDY1 N > 10,000	STUDY2 N > 1,000	STUDY3 N > 1,000	STUDY4 N > 1,000	STUDY5 N > 2,000	STUDY6 N > 1,000
Group Level						
A		.12/.11	.16/.14	----	.18/.14	----
B		.53/.13	.53/.13	.51/.13	.49/.12	----
C		.10/.14	.19/.15	----	.20/.13	.15/.09
D	.57/.22					
E	.59/.39	----	----	.55/.20	----	.54/.12
F	.40/.11					
Individual Level						
G		.31/.46	N/A	N/A	N/A	N/A
H		.86/.14	100%	100%	100%	100%
I		.10/.33	0%	0%	0%	0%
J		.04/.22	0%	0%	0%	0%
K		.47/.50*	.47/.50**	.44/.50*	.44/.50**	.57/?
L		.59/.31	.59/.31**	.53/.31	.53/.31**	.52/?
M		.42/.31	.42/.31**	.44/.31	.46/.31	.48/?
N		.46/.50*	.46/.50**	.49/.50*	.49/.50**	.52/?
O		.43/.31	.43/.31**	.57/.32	.51/.30	.57/?
P		.78/.33	.78/.33**	.69/.33	.69/.33**	.66/?
Q		.17/.41	.17/.41**	----	.20/.32	----
R		.15/.35	.15/.35**	----	.12/.35	----
S		.24/.43	.24/.43**	----	.18/.43	----
T		.17/.38	.17/.38**	----	.15/.38	----
U		.44/.39	.44/.39**	----	.55/.32	----

? indicates that standard deviations were not reported

* Plausible values

** Duplicated values, not counted as separate values

Because the standard formula for the binary standard deviation as a function of sample size and mean might not work for grouped data, we limited our analysis to the 26 instances associated with individual data, of which four had plausible results as notated in Table 1. Figure 2 plots the 26 data points against the expected values for binary means and standard deviations. Figure 2 illustrates the wide differences in Table 1 for the individual data versus the expected results for binary variables with N > 1,000.

There is always a chance that an error in a mean or standard deviation might be the result of a typographical error, possibly a type-setting error by the publisher rather than the author(s). For example, we noted that two of Regnerus's (2012, pp. 757-758) 40 binary variables had impossible values (in the range of 0.70-0.75) (Schumm et al., 2018, p. 786). How likely would that have occurred by error? If the base error rate for typographical errors was .005, there would be, from a binomial test calculation, a 16.4% chance of one error and a 1.6% chance of a second error. If the error rate was .001, those results would change to 3.8% and less than one tenth of one percent. If the error rate was .01, the results would be 27.0% and 5.3%. Thus, having two or fewer errors out of 40 seems not unlikely in the case of Regnerus's results, if only as a result of typographical errors.

In the hypothetical case, 2 errors out of 26 would be less likely for an error rate of .001 (2.5%, .03%), .005 (11.5%, .7%), or .01 (20.2%, 2.6%) than in the case of Regnerus, but might nevertheless be explained by infrequent typographical errors. However, in Table 1 for individual data there were 22 errors out of 26; regardless of whether the underlying typographical error rate was .001 to 0.10, the chances of 22 or more errors remained less than .00000001. Thus, an error rate of 22/26 was impossibly unlikely by chance errors. Even if the typographical error rate was 50%, the chance of 22 or more out of 26 errors would only be less than .00025. Thus, it would appear that having 22 of 26 cases in error by random typographical errors would be extremely unlikely, regardless of any reasonable probability of those errors. Even if the typographical error rate was .10 and one were to argue that only half of the 26 data points were far off from their expected values, the probability of half (13) or more of the data points being noticeably distant from their expected values would remain very remote, less than .0000003.

Other Approaches Including Checks for Inconsistencies. There would be at least three indications of errors in the six studies. First, when the standard deviations are plotted as a function of the mean scores, most points do not fall on the only possible curve, as shown in Figure 2. Thus, there are a majority of data points that are not possible, mathematically and therefore, must be incorrect. If there were only one or two points in error, then we might graciously assume there were some typesetting errors by the publisher or the author. However, when a majority of the points are not even possible, much less reasonable, then the likelihood of the studies being fraudulent is much greater.

Second, within each study the scores are internally inconsistent, given that as the mean scores increase to about .50, the standard deviations should increase monotonically in some proportion to the change in the mean score. In STUDY1, even though the mean scores of .57 and .59 are very close, the standard deviations are very different (.22 vs. .39), a difference that seems large even if group effects were involved in creating differences in the standard deviations. Since STUDY1 did not report results for individual data and because group results might not fit the pattern in Figure 1, the weakness in STUDY1 is more an issue of internal inconsistency rather than an inconsistency between binary means and standard deviations related to the pattern in Figure 1. In STUDY2, there are several inconsistencies. A mean of .12 has nearly the same standard deviation (.11) as the mean of .53 (SD = .13); at the individual level, means of .43, .44, and .46 have quite different standard deviations (.31, .39, and .50); means of .17 at the individual level have standard deviations of .38 and .41. In STUDY3, means of .16 and .53 have nearly the same standard deviations (.14 and .13); a mean of .15 has a higher standard deviation (.35) than a mean of .42 (SD = .35). In STUDY4, at the individual level, means of .44 and .53 have the same standard deviation of .31; means of .57 and .69 have nearly the same standard deviations (.32 and .33). In STUDY5, means of .46 and .53 have the same standard deviation (.31); at the group level, means of .18 to .49 have very small range of standard deviations (.12 to .14). In STUDY6, at the group level, a mean of .15 and a mean of .54 have nearly the same standard deviations (.09 and .12). In STUDY6, many of the standard deviations were not reported, removing the possibility of noticing any inconsistencies. Even if a scholar did not know about the limited number of possible scores for binary means and standard deviations, the numerous inconsistencies within each report should catch one's notice. Again, these might be typographical errors, but with so many being internally inconsistent, one might wonder how carefully the author(s), reviewers, or editors were checking the original manuscript if they did not notice any of the many internal inconsistencies in the studies.

Third, there are discrepancies between studies. Studies two through six share the same values for variables H, I, and J; even though the standard deviations would be zero for variables H, I, and J, we did not count them among the possibilities because despite those similarities, the studies had very different values on most of the other variables. For example, studies two and three have differences on variables A, C, H, I, and J and yet have identical values for variables K through U. Our guess is that the values for the latter variables might be typical when results were copied over from study two into study three rather than having been recalculated for a sample which was of a substantially different composition as suggested by the differences on the former set of variables. There are also five duplicated values between studies four and five, even though all of the other variables differ.

4 Discussion

4.1 Limitations

Binary testing will not catch fraud in which a researcher merely doubles or triples the number of cases in order to create a larger sample size. Astute cheaters might revise their binary standard deviations to make them more reasonable, even though that would take some time. Fraudulent researchers may choose to only report mean scores or basic percentages without reporting standard deviations, in order to not permit anyone to detect problems with standard deviations. Other approaches to testing for fraudulent data are possible but not the focus of this report. For example, checking the randomness of last digits in any results might be another approach since last digits are probably, in essence, random digits. Thus, any last digit should be about as likely as any other; 0's should be as likely as 1's, for example. Binomial tests could be used to compare such results across two digits (e.g., 0 versus 1) or chi-square tests could be used across the spectrum of digits (0 to 9). A further approach might be noticing any uncommon patterns in standard deviations or standard errors (Simonsohn, 2013). In other words, if one were to notice that in a series of models predicting a dependent variable from 20 independent variables, with different selections of those 20 variables, the standard errors were all identical for each variable regardless of the model used, one might question the likelihood of that possibility and wonder if the results had been made up (it would be easier to replicate standard errors rather than to take the time to make up unique ones in each instance). Another anomaly might be observed when multiple t-tests were being reported and most of the standard deviations were identical across the two groups.

4.2 Strengths

Binary testing is not limited by sample size. As sample size increases, the formula approaches the square root of [m(1–m)]. Data points can be plotted easily to see if they conform to the expected curve of means versus standard deviations. The plots can be examined to see if similar mean scores feature widely different standard deviations even though similar mean scores should feature similar standard deviations. Heathers and Brown (2019) have proposed more specific ways to test each data point against its expected value in the binary plot. Binary tests aside, inconsistencies among means and standard deviations are likely to occur, we think, if data are being falsified at the published level because avoiding discrepancies would take a great deal of time, probably the very thing an author might be trying to avoid by fabricating data. In other words, if you fix one inconsistency, especially across studies, you might be creating another one, maybe more than one. It might take a very long time to clean up internal and across-study inconsistencies.

The dilemma reminds me of what the senior author's mother used to tell me: "Don't lie. Once you lie, you have to keep lying to cover up the previous lies and sooner or later you forget what you said was true and what you lied about, so even cover up lies become impossible to maintain with any consistency. Sooner or later lying will be caught." In the case of scholars creating results out of their imagination rather than from real data, the similar challenge will be, as noted above, that when they "fix" one standard deviation, they may create further inconsistencies within an article or across related articles, as well as violating the mean/standard deviation pattern. Since that mean/standard deviation relationship or pattern changes with each sample size, it will not be easy to guess the correct results, without genuine raw data. An easier "solution" for those scholars not interested in doing more real scholarly work would be to simply *not report* binary standard deviations as was done for individual data in STUDY6 in Table 1, thereby eliminating any possibility for the detection of binary data inconsistencies. In a real situation, one might wonder if "the heat was on" and the study author(s) had decided to resolve challenges to the validity of their data by that relatively simple and easy method, of not reporting enough data to allow for checks of inconsistency. However, absence of important data should alert reviewers and editors to not accept manuscripts for publication unless all important results are presented and presented accurately.

5 Conclusion

While the DEBIT test may offer greater precision in detecting whether individual data points are different enough from their expected values, we think a visual inspection combined with binomial tests may suffice in most cases to distinguish typographical or other random errors from systematic indications of bias or fraud. Of course, if raw data are manufactured, this procedure will not work any better than the DEBIT approach. However, we suspect the whole point of fabricating data is to save time and present appealing results; if raw data are manufactured, then the results may not become appealing without multiple revisions to the raw data, which undercuts the idea of saving time. Therefore, we think it would appeal more to most fabricators to save time and fabricate at the published level rather than at the level of raw data, which might take multiple revisions to create results that "fit" the proposed hypotheses in the desired directions. Thus, the temptation will be to fabricate results at the manuscript level, even without any data, in order to minimize time requirements while getting maximum publications for minimal effort. Our proposal to combine visual inspections of the binary variable mean/standard deviation relationship along with binomial tests could become one standard way, that does not rely upon complex multivariate statistics, of trying to detect scientific fraud or related accuracy problems.

References

1. Anaya, J. (2016). The GRIMMER test: A method for testing the validity of repeated measures of variability. *PeerJ Preprints*, *4*, e2400v1.
2. Brown, N. J., & Heathers, J. A. (2017). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, *8*(4), 363-369.
3. Brown, N. J., & Heathers, J. A. (2019). Rounded Input Variables, Exact Test Statistics (RIVETS): A technique for detecting hand-calculated results in published research. Unpublished paper. Bouve College of Health Sciences, Northeastern University, 360 Huntington Avenue, Boston, MA, USA 02115.
4. Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *Plos One*, *4*(5), e5738.
5. Heathers, J., Anaya, J., van der Zee, T., & Brown, N. J. L. (2018). Recovering data from summary statistics: Sample parameter reconstruction via iterative techniques (SPRITE). *Peerj Preprints*. doi.org/10.7287/peerj.preprints.26968v1.
6. Heathers, J. A., & Brown, N. J. (2019). DEBIT: A simple consistency test for binary data. Unpublished paper. Bouve College of Health Sciences, Northeastern University, 360 Huntington Avenue, Boston, MA. USA 02115.
7. Regnerus, M. (2012). How different are the adult children of parents who have same-sex relationships? Findings from the New Family Structures Study. *Social Science Research*, *41*, 752-770.
8. Schumm, W. R. (2015). Navigating treacherous waters – One researcher's 40 years of experience with controversial scientific research. *Comprehensive Psychology*, *4*, 24, 1-40.
9. Schumm, W. R. (2018). Making sense of probability: What formulas to use under different conditions. *Biomedical Journal of Scientific and Technical Research*, *3*(1), 2983-2984.
10. Schumm, W. R., Crawford, D. W., & Lockett, L. (2019). Patterns of means and standard deviations with binary variables: A key to detecting fraudulent research. *Biomedical Journal of Scientific and Technical Research,* in press.
11. Schumm, W. R., & Crawford, D. W. (2019). Scientific consensus on whether LGBTQ parents are more likely (or not) to have LGBTQ children: An analysis of 72 social science reviews of the literature published between 2001 and 2017. *Journal of International Women's Studies*, *20*(7), 1-12.
12. Schumm, W. R., Crawford, D. W., Higgins, M., Lockett, L., AlRashed, A., & Ateeq, A. B. (2018). Estimating the standard deviation from the range: A replication of analysis of demographic data reported in *Marriage & Family Review*, 2016-2017. *Marriage & Family Review*, *54*, 777-792.
13. Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*(10), 1875-1888.
14. Steen, R. G., Casadevall, A., & Fang, F. C. (2013). Why has the number of scientific retractions increased? *PloS One*, *8*(7), e68397.
15. Stern, A. M., Casadevall, A., Steen, R. G., & Fang, F. C. (2014). Financial costs and personal consequences of research misconduct resulting in retracted publications. *Elife*, *3*, e02956.
16. Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science*, *7*(6), 670-688.
17. Warren, J. R. (2019). How much do you have to publish to get a job in a top sociology department? Or to get tenure? Trends over a generation. *Sociological Science*, *6*, 172-196.