Statistical Whistleblower: Am I Brave Enough? Am I Dumb Enough?

Laurence D. Robinson
Department of Mathematics and Statistics, Ohio Northern University, Ada, OH 45810

Abstract
At my university I have seen numerous examples of the misuse of statistics. As the senior statistician at the university, I feel that perhaps I have a certain responsibility for addressing this problem. Will I? Should I? Do I want to be hated that much?

Key Words: Misuse of statistics

1. Some Very Important Initial Points That Must Be Made

With regard to the misuse of statistics, I do not believe that my university is in any way exceptional. I believe the misuse of statistics is ubiquitous, occurring at universities and colleges across the country, as well as in the private sector.

In general I believe that my university offers a high quality education to its students. This includes those faculty members who on occasion may be involved in the misuse of statistics.

2. A Profound Quote

"The Greatest Obstacle to Discovery Is Not Ignorance: It Is The Illusion Of Knowledge."

Daniel Joseph Boorstin (1914 to 2004)

Daniel Boorstin was an American historian, writer, professor, attorney, and served as the twelfth Librarian of the United States Congress from 1975 to 1987.

Is there anywhere in the world of science where Daniel Boorstin's quote is more true than with regard to the use (and misuse) of statistics? I don't think so.

Why is this the case, you ask? There are several reasons, including the following: It is very difficult to truly understand statistics, but very easy to fool oneself into thinking that you do. It is very difficult to use statistical methods to analyze data correctly, but very easy (particularly with today's statistical software) to use statistical methods to analyze data incorrectly. Furthermore, it is very easy to misuse statistical methods without this misuse being detected by others. This can be attributed to referees having an inadequate knowledge and understanding of statistics, and to the fact that papers and reports often omit crucial details.

3. Example 1

I received the following letter by email from a top administrator at my university:

Dear Colleagues,

As you may know, ONU is administering the CAAP test to assess the progress of our students throughout their time at ONU. Thus far, only 28 A&S students have taken the test. In order for the results to be statistically significant, we need 89 A&S seniors to take the CAAP test. The test will be offered four more times before the end of March. If you are teaching a course that enrolls seniors, please encourage them to sign up for the CAAP assessment. Some faculty in the other colleges have offered extra credit to students who take the CAAP test, and you may wish to consider such an incentive to encourage A&S students to participate. Thank you for your assistance …

The Collegiate Assessment of Academic Proficiency (CAAP) is the standardized, nationally normed assessment program from ACT that enables postsecondary institutions to assess, evaluate, and enhance student learning outcomes and general education program outcomes. (http://www.act.org/caap/)

The administrator's misuse of statistics in this instance is clear. The sampling method being used here is based on the use of volunteers, which clearly creates the possibility for a biased assessment. In fact, I think it is quite likely that the students who volunteer will tend to be stronger students than those who do not, and thus the measure of performance on the test will be biased high.

4. Example 2

In my building on the beautiful Ohio Northern University campus, the walls are filled with posters pertaining to research being conducted by faculty and students. On one such poster, with the first author being a professor, a scatterplot of a response variable Y vs. predictor variable X was given, and it was stated that the correlation between the two variables was $r = .75$. It was concluded that the correlation was very strong.

The statistical errors contained in this poster include the following:

- I question whether an r of .75 constitutes a very strong correlation. Further, the correlation did not appear (to me) to be that strong.

- The response variable Y was plotted on the horizontal axis, the predictor variable X on the vertical axis, contradicting the standard practice of plotting response variables on the vertical axis, predictor variables on the horizontal axis.

- There was a notable curve to the regression function, so that summarizing the relationship with only a correlation coefficient was inadequate.

- The correlation coefficient was referred to as the "product *movement* correlation coefficient", rather the correct term "product *moment* correlation coefficient".

It should be noted that, in a personal conversation, the chair of the department where this poster was located had previously described the professor who was listed as first author as being one of that department's most knowledgeable faculty members with regard to the use of statistics.

5. Example 3

On another poster, on another wall, on the beautiful Ohio Northern University campus, with the first author being a professor, and with numerous student names following, the results of a study were given. In this study a sample of $n = 56$ subjects were asked 11 questions pertaining to a particular subject, and each answer was determined to be correct or incorrect. The subjects then attended an educational session, after which they were asked the same 11 questions, again with each answer determined to be correct or incorrect. For each of the 11 questions the poster provided certain information. The following table gives the information provided for Question 7:

	Number (Percent) correct
Pre-educational session	52 (93)
Post-educational session	55 (98)

For each question a bar chart representing the pre- and post-educational session data was given. A P-value for the test of the equality of the proportions of correct answers (pre- vs. post-educational session) was given for each question. For Question 7, the P-value given was .08. The poster stated that the hypothesis tests conducted were "paired sample *t*-tests".

At first glance one statistical error stands out immediately. The outcomes for each question were dichotomous (rather than quantitative), hence "*t*-tests" could play no role. From calculations I performed, I determined that the tests conducted are those referred to as "McNemar's test", which in fact are equivalent to "paired sample *z*-tests".

It should now be noted that the representation of the data given in the poster (in both the tables and bar charts) was inadequate. This not only makes interpreting the study results more difficult for the intended audience, it also made it more difficult for me to determine which test was actually used.

A complete representation of the data for Question 7 is as follows:

		Post- Correct	Incorrect	Total
Pre-	Correct	52	0	52
	Incorrect	3	1	4
	Total	55	1	56

Note that given only the marginal totals, which is the information given on the poster, a second possible table can be constructed:

		Post- Correct	Incorrect	Total
Pre-	Correct	51	1	52
	Incorrect	4	0	4
	Total	55	1	56

How do we know the first table was the actual result? Conducting McNemar's test for the first table we obtain $X^2 = (3 - 0)^2 / (3 + 0) = 3$, resulting in P-value = $\Pr(\chi^2 > 3 \mid df = 1) = .083 \approx .08$, the P-value given in the poster. Applying McNemar's test to the second table gives $X^2 = (4 - 1)^2 / (4 + 1) = 1.8$, resulting in P-value = $\Pr(\chi^2 > 1.8 \mid df = 1) = .180$.

Of course, at this point it may still be possible that the data in the second table was actually obtained, with a different test being conducted. It was through checking the results of many of the 11 survey questions that I was able to determine that it was McNemar's test that had been used, and simultaneously to determine the complete data results for each question, in spite of the fact that only the marginal totals were given in the poster.

It is well known that the validity of McNemar's test for the study design used, *i.e.* the paired design with dichotomous responses, requires sufficiently large samples, and that the approximate test can (usually) be improved through the use of a continuity correction. A common (and very liberal) guideline, used to determine whether the sample size is sufficiently large for McNemar's test to be used, is that the number of discordant pairs of observations (pairs where the pre- and post- response changed from incorrect to correct, or vice versa) should be 10 or more.

For Question 7 (and other questions on the survey) this guideline is not met. Specifically, the number of discordant pairs is only 3 + 0 = 3, which is substantially less than 10. Furthermore, the P-value that was calculated (for Question 7, and the other questions as well) did not use a continuity correction. Use of a continuity correction for the Question 7 data gives the following results: $X^2 = (|3 - 0| - 1)^2 / (3 + 0) = 1.333$, resulting in P-value = $\Pr(\chi^2 > 1.333 \mid df = 1) = .248$. This suggests that the exact P-value is much larger than the value of .08 reported in the poster.

Calculation of the exact P-value (for all questions) is very easy. The exact test is simply a two-sided alternative test that the parameter p of a binomial distribution equals .5, with the sample size being the number of discordant pairs. For Question 7, we compute the P-value as $2 \Pr[Y = 3 \mid Y \sim \text{binomial}\ (p = .5, n = 3)] = 2(.5^3) = .25$. Note this is (somewhat surprisingly, given such a small sample size) extremely close to the value of .248, obtained by McNemar's test using the continuity correction.

Thus, for Question 7, the P-value reported by the authors was much lower than the exact P-value, *i.e.* the authors were claiming greater significance than actually existed. This was generally the case for the other questions on the survey.

How is it that the professor (and students) allowed themselves to use a χ^2 approximation with such a small sample size, *i.e.* $n = 3$, and without using a continuity correction? I suspect that they did not recognize that the relevant sample size for the test is not the total

number of matched pairs, which was 56, but rather the number of discordant pairs, which was only 3. Also, it is my perception that most introductory statistics courses and text-books tend to downplay the use of continuity corrections.

6. Final Comments

On another poster, on another wall, on the beautiful Ohio Northern University campus, the results of another study were given (authored by a student and two professors). I had many questions and doubts regarding the statistical analysis conducted. Upon seeing one of the professors in the hallway near the poster, I asked the professor if I could ask some questions regarding the poster. The professor was clearly pleased that I was interested in the study. A few moments later, upon realizing that I had questions and doubts regarding the statistical analysis, the professor was clearly not pleased.

Statistical Whistleblower: Am I Brave Enough? (This remains to be seen.) Am I Dumb Enough? (Without a doubt.)

References

1. McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12 (2): 153–157.

2. Edwards, A. L. (1948). Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. Psychometrika 13: 185–187.

3. Fleiss, J. L. (1981). Statistical methods for rates and proportions (2nd ed.). New York: John Wiley & Sons. p. 114.