Inaccurate rounding of statistics in medical papers

Emili García-Berthou*, Carles Alcaraz

Department of Environmental Sciences, University of Girona, E-17071 Girona, Spain

* Corresponding author

Tel.: +34 972 418 467

Fax: +34 972 418 150

E-mail: emili.garcia@udg.es

Abstract

Background
Given an observed test statistic and its degrees of freedom, one may compute the observed *P* value with most statistical packages.

Methods
We checked the accuracy of statistical figures reported in all the papers of volumes 409-412 of *Nature* (2001) and 63 figures from volumes 322-323 of *BMJ* (2001).

Results
11.6% (21 of 181) and 11.1% (7 of 63) of the statistical figures published in *Nature* and *BMJ* respectively during 2001 were inconsistent, probably mostly due to incorrect rounding as confirmed by digit preference. One or more of these errors appeared in 38% and 25% of the papers. In 12% of the cases, the significance level might change one or more orders of magnitude.

Conclusions
This frequent misuse is another example that statistical practice is generally poor, even in the most renowned scientific journals, and that quality of papers should be more controlled and valued.

Background

Statistics is a difficult topic to teach and learn and there is ample evidence that its application is often faulty in medicine [1-6] as well as in many other scientific disciplines. Errors include aspects of design, analysis, and reporting and interpretation. Although there has recently been considerable effort in the improving and standardisation of reporting (e.g., the CONSORT statement for randomised controlled trials [7], there is apparently no literature proving the incorrect computation or reporting of figures beyond general deficiencies of computer packages [8,9]. Beyond deficiencies of software, such numerical errors may later originate in the transcription of results from computer outputs to reports and manuscripts, wrong rounding of results, or uncorrected typesetting errors. We investigated this question by checking the statistical figures reported in all the papers of volumes 409-412 of *Nature* (2001) and some papers in vol. 322-323 of *BMJ* (2001). We show that the occurrence of errors is very high and we review ways to improve current practice.

Methods

Given an observed test statistic and its degrees of freedom, one may compute the observed P value or significance level (or vice versa) with most statistical packages. We are thus able to check the consistency and accuracy of figures consisting of the test statistic, df and a precise P value. We cannot check figures consisting only of a P value or with no precise P value (e.g. $P < 0.05$ instead of $P = 0.023$) and therefore these figures were not considered in our review. Note that the latter are bad practices and reporting both the observed test statistic and the "exact" *P* value has been recommended [3]. We did not check the accuracy of confidence intervals and other statistics because it would be generaly impossible without access to the raw data.

We checked all the statistical figures (consisting of the test statistic, df and a precise *P* value) reported in all the papers of volumes 409-412 of *Nature* (2001) and 12 randomly selected papers from vol. 322-323 of *BMJ* (2001). We checked the figures with three different packages: *SPSS for Windows 10.1*, *STATISTICA '98 for Windows*, and the freeware *NCSS Probability Calculator (*www.ncss.com*)*. The results of the three statistical packages were identical at least up to the 4th decimal. All the errors detected are detailed in the additional files.

We only considered an error when it was not possibly due to rounding in the original paper. For instance, the result of "χ^2 = 1.7, df = 1, *P* = 0.30" in vol. 322, *p. 769-770* of *BMJ* cannot be due to correct rounding of the test statistic and *P* value, given the following precise results: χ^2 = 1.65, df = 1, *P* = 0.199; χ^2 = 1.70, df = 1, *P* = 0.192; χ^2 = 1.75, df = 1, *P* = 0.186. If the statistic was really χ^2 = 1.7, then the *P* value should have been much lower than 0.3. In fact, a χ^2 of 1.07 with 1 df yields a *P* value of 0.3, suggesting a transcription error.

In contrast, the result "χ^2 = 1.2, df = 2, *P* = 0.54" in vol. 322, *p. 336-342* is consistent with the following precise results after rounding: χ^2 = 1.15, df = 2, *P* = 0.563; χ^2 = 1.20, df = 2, *P* = 0.549; χ^2 = 1.25, df = 2, *P* = 0.535.

Results and discussion

We found that a surprising 11.6% (21 of 181) of the computations in *Nature* were inconsistent. A less exhaustive check in *BMJ* resulted in a very similar percentage (11.1%, 7 of 63). One or more of these errors appeared in 38% (12 of 32) and 25% (3 of 12) of the papers of *Nature* and *BMJ* respectively, indicating that they are widespread and not concentrated in a few papers. For instance, in vol. 411, *p*. 88 of *Nature* "$F_{2, 14}$ = 10.89, *P* = 0.014" was reported while the consistent *P* value is 0.0014, suggesting a transcription error. Another transcription error is "$F_{7, 79}$ = 7.09, *P* = 0.0094" in vol. 412, *p*. 74, in which the *P* value corresponds to an *F* with 1 and 79 degrees of freedom.

Many errors are probably due to incorrect rounding, e.g. "*r* = 0.30, *N* = 21, *P* = 0.20" (consistent *P* = 0.186) in vol. 411, *p*. 297 of *Nature* or "χ^2 = 0.01, df = 1, *P* = 1.00" in vol. 322, *p*. 336-342 of *BMJ*. Some authors state *P* = 0.001, when they should state *P* < 0.001 or *P* << 0.001.

These inconsistencies are probably due to inaccurate rounding or transcription. Software deficiencies are usually orders of magnitude less important [8,9], and would be restricted to specific papers using a certain statistical package, contrary to our findings of over 25% of the papers with errors. Most typesetting errors are probably detected by authors' corrections and errors in previous steps of edition are probably more frequent and difficult to detect. Interestingly, independent evidence of rounding misuse stems from digit preference [10,11]. We collected 610 test statistics from the same *Nature* volumes and counted the frequencies of the last digit reported (Fig. 1). The counts significantly deviate from the expected uniform distribution (Kolmogorov-Smirnov test, *z* = 2.7, *P* < 0.0005) and show that authors tend to round more frequently, inconsistently and sometimes wrongly, when the last digit is high (as expected for psychological reasons) and when it is 4, 6 or 9. The counts of the last digit of *P* values also significantly deviate from the uniform distribution (z = 1.4, P = 0.043), and 0, 4, and 9 are less common than expected.

The estimate of 11-12% of inconsistent figures is a conservative one since some cases were not considered errors because they might have been caused by rounding. It is not possible to be certain of the real importance of these errors because without access to the raw data, we do not know the correct result. Apparently, the conclusion would change from significant to nonsignificant in only about 4% (1/27) of the errors using the arbitrary 5% level. However, the median of the relative bias (absolute difference between the reported and consistent *P* values, divided by the consistent *P* value) was 38% and in 12% of the cases the relative bias was larger than 10%, showing that the significance level might change one or more orders of magnitude.

These kinds of errors are probably relatively unimportant, while other errors during the study might be much more harmful. Our concern is that these kinds of errors are probably present in all numerical figures (e.g., means, percentages, confidence intervals) and all steps of scientific research, with potentially important practical consequences. Moreover, poor presentation provides clues that there may be serious errors elsewhere [12].Our findings confirm that the quality of research and scientific papers needs improvement and should be more carefully checked and evaluated in these days of high publication pressure [13-15].

Recommendations

Several detailed guidelines on the practice and reporting of statistics in medical papers are available [3,7,16,17]. There is considerable consensus on the most desirable practices, and some of their suggestions are:

1) In medical research, confidence intervals are often more appropriate than hypothesis testing. If hypothesis testing is used, it is desirable to report not only the *P* values but also the observed values of test statistics and the degrees of freedom.

2) Exact *P* values (to no more than two significant figures) should be given rather than reporting $P > 0.05$ or $P < 0.01$. It is unnecessary to specify levels of *P* lower than 0.0001.

3) Spurious precision adds no value to a paper and even detracts from its readability and credibility. Results need to be rounded [18-20].

To this we need to add that:

1) Numerical results should be <u>correctly</u> rounded. The problem of introducing bias by rounding digits ending in five [21] is a trivial one compared to the misuses reported in our paper.
2) The preparation and edition of manuscripts should be more carefully checked. Increasing the use in medical journals of statistical reviewers [1,12] and of unlimited publication of correspondence on the web [2] may help to improve the quality of papers.
3) In principle, authors of research papers (including systematic reviews) should make the raw data freely available on the Internet and journals should implement and estimulate this practice. The benefits of this recent practice mainly involve: further analyses analyses not directly addressed by the primary researchers are

possible [22,23], including effective systematic review and meta-analysis [24] or the estimation of adequate sample sizes (power analysis) [25]; other researchers can check whether the results are correct and the conclusions justified [24,25]; fraud and sloppiness may be more easily detected and is thus discouraged [22].

Among others, Altman and coauthors give details of many other ways to improve the practice and reporting of statistics in medicine and their suggestions are widely applicable to other research fields [1,3,5,12].

References

1. DG Altman: **Statistics in medical journals.** *Stat Med* 1982, **1:** 59-71.
2. DG Altman: **Poor-quality medical research: what can journals do?** *JAMA* 2002, **287:** 2765-2767.
3. DG Altman, SM Gore, MJ Gardner, SJ Pocock: **Statistical guidelines for contributors to medical journals.** *Br Med J* 1983, **286:** 1489-1493.
4. JR O'Fallon, SD Dubey, DS Salsburg, JH Edmonson, A Soffer, T Colton: **Should there be statistical guidelines for medical research papers?** *Biometrics* 1978, **34:** 687-695.
5. DG Altman, JM Bland: **Improving doctors' understanding of statistics.** *J R Statist Soc A* 1991, **154:** 223-267.
6. DG Altman: **Statistics in medical journals: developments in the 1980s.** *Stat Med* 1991, **10:** 1897-1913.
7. DG Altman, KF Schulz, D Moher, M Egger, F Davidoff, D Elbourne, PC Gøtzsche, T Lang, for the CONSORT group: **The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration.** *Annals of Internal Medicine* 2001, **134:** 663-694.
8. BD McCullough: **Assessing the Reliability of Statistical Software: Part II.** *Am Stat* 1999, **53:** 149-159.
9. BD McCullough: **On the accuracy of statistical procedures in Microsoft Excel 97.** *Comput Statist Data Anal* 1999, **31:** 27-37.
10. L Edouard, A Senthilselvan: **Observer error and birthweight: digit preference in recording.** *Public Health* 1997, **111:** 77-79.
11. T Clemons, M Pagano: **Are Babies Normal?** *Am Stat* 1999, **53:** 298-302.

12. DG Altman: **Statistical reviewing for medical journals.** *Stat Med* 1998, **17:** 2661-2674.

13. BA Hawkins: **More haste, less science?** *Nature* 1999, **400:** 498.

14. DG Altman: **Statistics in medical journals: some recent trends.** *Stat Med* 2000, **19:** 3275-3289.

15. DG Altman: **The scandal of poor medical research.** *Br Med J* 1994, **308:** 283-284.

16. JC Bailar III, F Mosteller: **Guidelines for statistical reporting in articles for medical journals. Amplifications and explanations.** *Annals of Internal Medicine* 1988, **108:** 266-273.

17. DG Altman, JM Bland: **Presentation of numerical data.** *Br Med J* 1996, **312:** 572.

18. ASC Ehrenberg: **The problem of numeracy.** *Am Stat* 1981, **35:** 67-71.

19. ASC Ehrenberg: **Rudiments of numeracy.** *J R Statist Soc A* 1977, **140:** 277-297.

20. DJ Finney, JL Harper: **Editorial code for presentation of statistical analyses.** *Proc R Soc Lond , Ser B: Biol Sci* 1993, **254:** 287-288.

21. D Green: **Eliminating bias due to rounding.** *Teaching Statistics* 1990, **12:** 69.

22. GD Smith: **Increasing the accessibility of data.** *Br Med J* 1994, **308:** 1519-1520.

23. DC Hoaglin, DF Andrews: **The reporting of computation-based results in statistics.** *Am Stat* 1975, **29:** 122-126.

24. DJR Hutchon: **Infopoints: Publishing raw data and real time statistical analysis on e-journals.** *Br Med J* 2001, **322:** 530.

25. T Delamothe: **Whose data are they anyway?** *Br Med J* 1996, **312:** 1241-1242.

Figure 1

Histogram of the last digit of 610 test statistics in volumes 409-412 of *Nature*. The reference line corresponds to the mean count (61).

Figure 2

Histogram of the last digit of 181 *P* values in volumes 409-412 of *Nature*. The reference line corresponds to the mean count (18.1).

Additional files

Additional file 1 – BMJ.xls

Excel file with 7 errors (rows) detected in vol. 322-323 of *BMJ*

Additional file 2 – Nature.xls

Excel file with the 181 errors detected in *Nature*

Additional file 3 – columns.txt

ASCII file explaining the variables (11 columns) in the two previous additional files

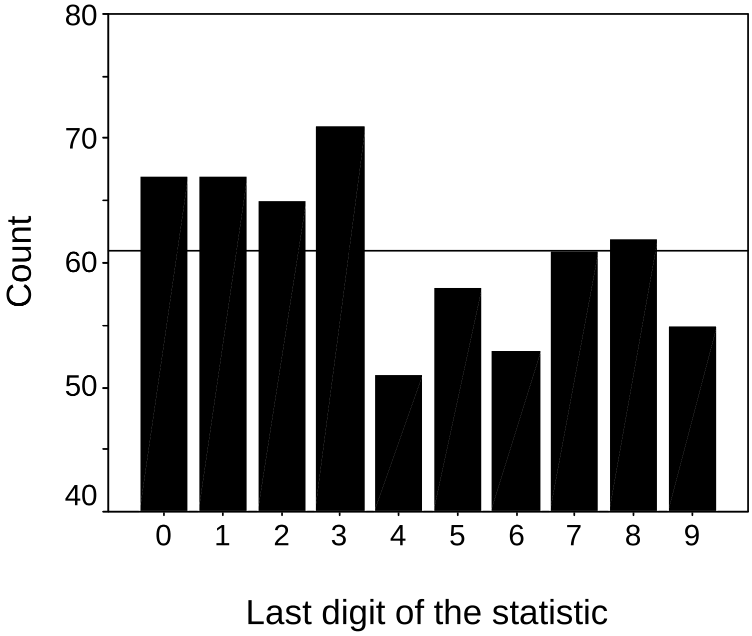

Figure 1

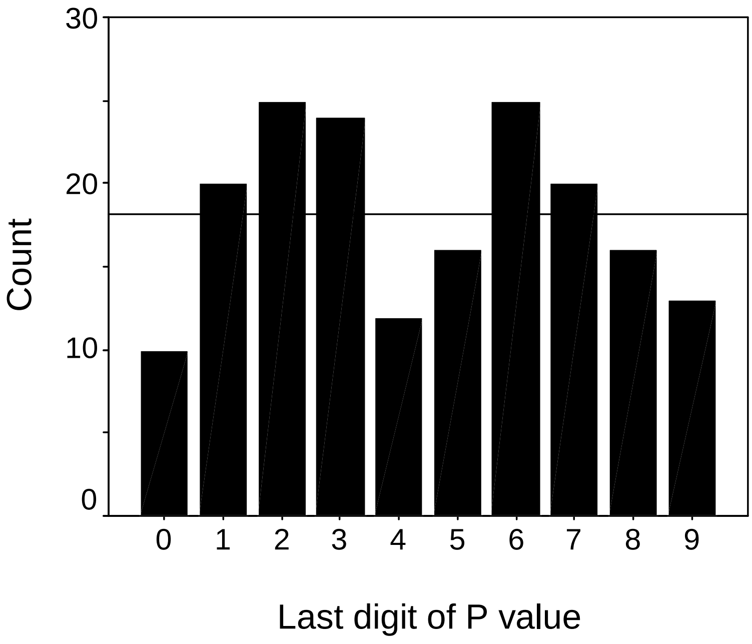

Figure 2

Additional files provided with this submission:

Additional file 1: BMJ.xls : 2KB
http://www.biomedcentral.com/imedia/1407769967280553/sup1.xls

Additional file 2: Nature.xls : 32KB
http://www.biomedcentral.com/imedia/1106475093280558/sup2.xls

Additional file 3: columns.txt : 0KB
http://www.biomedcentral.com/imedia/2134696819280559/sup3.txt