

Author's response to reviews

Title: Incongruence between test statistics and P values in medical papers

Authors:

Dr Emili Garcia-Berthou (emili.garcia@udg.es)

Carles Alcaraz (carles.alcaraz@udg.es)

Version: 2 Date: 30 Mar 2004

PDF covering letter

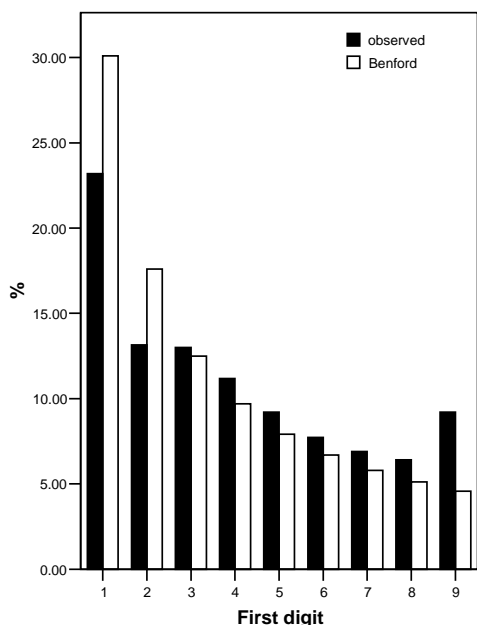
We followed most of the helpful suggestions of referees. Specific comments follow.

Dr. Harris review

- 1) We changed the title and adopted the “incongruence” term and removed other terms throughout the MS
- 2) We added more explanation on the digit preference topic

Dr. McCullough review

- 3) We added reference to Diaconis & Freedman and clarification on the digit preference topic. We did not mention Benford’s law in our previous MS. We added clarification on this in the MS, namely that Benford’s law applies to the first LEADING digits but vanishes very soon (see e.g. the table in <http://www.usfca.edu/fac-staff/huxleys/Benford.html>).



The above Figure shows that for the FIRST digit (not analysed in the MS) of the 610 test statistics, the distribution is similar to Benford’s law but deviates significantly ($P \ll 0.0005$) with observed frequencies more uniform than expected. According to Benford’s law, for the second digit the difference should be of only 12.0% for 0 and 8.5% for 9 and the above Figure suggests that the difference should be even smaller for the observed data.

The test statistics were generally reported to 3-4 significant figures (so we should not expect Benford’s law but in principle a rather uniform distribution for the LAST digit). You are right that there might be a problem for P values (where significant figures are usually less than 4) but Fig. 2 of the MS shows that Benford’s law

does not apply either as there is no decrease at all in frequencies. For these reasons, it is customary for LAST digits to test for equally probable numbers; I added several references that exemplify that.

The test that you suggested yields that the proportion of even and odd digits are not significantly different ($\chi_1 = 0.32$, $P = 0.57$). However, I think that this test is not very informative because, as explained in the MS and the literature cited, digits 4 (even) and 9 (odd) are more often rounded and such test would not detect that. I think it is best to test for a uniform distribution.

4) I corrected the two discretionary revisions.

Dr. Ferraris review

- 5) The conclusion would change from significant to nonsignificant in only about 4% (1/27) of the ERRORS using the arbitrary 5% level. There were a total of 28 errors (21 in *Nature* and 7 in *BMJ*). I added to the revised MS “(1 error reporting “1,9” df for a *t* statistic was not considered)”; we cannot test whether that result would change or not because we do not know whether the df is 1 or 9 (the *P* value was not congruent with either). I apologise for having omitted that detail in the previous MS.
- 6) As suggested by Dr. McCullough I changed “one or more of the errors” with “at least one such error”, probably my English was not good enough. In the MS I state:

“We found that a surprising 11.6% (21 of 181) of the computations in *Nature* were incongruent. A less exhaustive check in *BMJ* resulted in a very similar percentage (11.1%, 7 of 63). At least one such error appeared in 38% (12 of 32) and 25% (3 of 12) of the papers of *Nature* and *BMJ* respectively, [...]”

I hope that now it is clear that ca. 12% of the computations (computation = set of test statistic, df and P value) were incongruent and more than 25% of the papers contained at least one such error. The difference is due to that a paper often contains several such computations.