Review of: "Intersections of Statistical Significance and Substantive Significance: Pearson's Correlation Coefficients Under a Known True Null Hypothesis"

William M. Goodman[1]

1 University of Ontario Institute of Technology

Potential competing interests: No potential competing interests to declare.

The key points in Komaroff's simulation-based paper are valid and important. He views his paper as extending, in a more accessible and "non-technical" format, the findings he published earlier in Komaroff (2020). That simplification effort is worthwhile. However, in the process, important details and explanations were omitted from the paper's method section, which makes it hard to follow. Revisions in that section may also require adjustments in other sections.

The paper warns against articles and journals that rely on effect size measures alone to support claims of substantive significance for sample results. That approach can err by ignoring the important issue of Type-I error risk. That is, if researchers avoid the use of p-values (with or without using the terminology "statistical significance"), it can easily happen, particularly for smaller samples, that there is really no effect—yet based just on a deemed-large-enough observed effect size, researchers may claim or conclude that they have observed an effect. Performing a hypothesis test and obtaining a low p-value is not a foolproof indicator—and it does not address, at all, Type-II error risk (which is concluding no effect when in fact there is one); but it can provide a useful screen against false positives based solely on using measures for effect size.

Komaroff's points about p-values are intended as general; but the present focus is on one particular class of cases where one might employ p-values: He simulates sampling to find evidence of linear correlations among pairs of variables. In the simulations, all the variable values in the dataset are randomized, so any apparent findings of correlations would be false positives.

If researchers analyzed such (simulated) data and relied only on the magnitudes of r values observed, they might conclude they'd found a substantive correlation if an r value reached an arbitrary threshold like $|r| \geq 0.1$ But had they tested, as well, for a low p-value for H_0: $\boldsymbol{\rho} = 0$, that would substantially limit the risk of reaching false positive conclusions.

Graphs like the paper's Figures 10 and 15 consistently support Komaroff's goal: to provide a "simple, intuitive … understanding of statistical significance for [students and others] who are not statisticians". Those figures display clearly how often observed effect sizes can be misleading if viewed as evidence, on their own, for true effects, and how p-values could screen off many such errors.

Where I have concerns with the paper is primarily in its Method section. This needs revision and expansion to be

accessible to its intended audience, and changes there may impact other sections. Komaroff acknowledges that his newer paper draws on a prior publication of his (2020); so, possibly, omissions from the earlier paper's explanations have arisen during simplifying. Regardless, there are some non-intuitive gaps, and for some details, I had to look to the older paper for clues about what was intended.

Most critically, A simulation-based paper geared to non-specialists should clarify how its model's features relate to relevant features of the process under study—in this case, sampling to find evidence of bivariate correlations. But the one sentence in the paper that may have pertained to this, I could not understand: The paper, it says, simulates "empirical sampling distributions comprised of 4950 bivariate correlations computed with 100 iid random variables with … instructive sample sizes: n = 4, 30, 100, 1000, 2000". This summary begs the questions: Where do these correlations come from? And why 4950 of them? And how are they organized?

Komaroff's earlier paper (2020), on the other hand, provided an analogy that would have been very helpful to include and expand on in the new paper. The gist of what was probably intended, which should have been explicit, is something like this: Suppose a sample of respondents, say n = 30, answers a survey, consisting of, say, 15 questions (items) with answers that are numeric (such as values on a numeric scale.) Such a sample's results could all be collected in a spreadsheet with one row for each respondent's answers; and one column for each question (variable), i.e., containing all respondents' respective answers to that question. The "bivariate correlations" under discussion, in that case, would refer to correlations between the different variables (columns) in the sample.

Further, the Method should distinguish between the different *stages* in the simulation exercise. The heart of this simulation is populating *all* cells for *all* answers with random numbers. This simulates the scenario where no true correlations would be expected, when comparing any two—or even *every* two—columns of data. However, the SAS hypothesis testing procedures for correlation, mentioned in the paper's Methods section, refers to a separate, *next* step in the simulation: *First,* illustrative data sets must be generated (for different sample sizes, and so on); *then* these data sets can be examined, analyzed, and compared.

In my view, the results of the simulation-runs might seem more surprising, and so more illuminating, to readers if the intuitive background described above were provided. The paper documents that, in spite of the simulated data for no-correlations-whatsoever, we nonetheless observe *many* magnitudes for the correlation coefficient *r* that are *not* zero at all, nor even "negligible". This would support needing an extra tool, like p-values, to screen results.

The Methods section is also confusing for its brief mention of "Fisher's r to z transformation (zr) … to test $\rho = 0$ for statistical significance". A reader seeking to understand those transformations would have to look elsewhere. The text leaves out procedural, and interpretive, steps.

Clearly, the paper's intention is simply to bring in the *p*-values—for comparing observed *r* magnitudes from simulation runs with the p-values that correspond. But by calling on *z* values for that purpose, a not-intuitive *double* transform has to occur: (1) from *r* to *z*r, which yields only the normal *shape*; and then (not mentioned in the paper) (2) normalizing *z*r, in turn, to *z* (i.e., dividing each *z*r by the standard deviation of the *z*r distribution). I assume Komaroff is doing this so he can

address non-specialists, whom he may assume would at least have heard of z-values, and of the rule of thumb, for normal distributions, that “when |z-value| ≥ 1.96, *p-value* ≤ 0.05”.

To go that route, the paper requires more work to explain the steps and calculations to be performed, and so on. However, I’d recommend simply dispensing with *z*, instead:

Why not include in the paper, like in old stats texts, a small table of critical values for ***r***, for p-value = 0.05 and a handful of specific sample sizes? In that way, for example, Figure 6 in the paper, which shows the simulated r distribution when n = 30, could simply add asterisks near r = -0.35 and r = +0.35 on the x axis, as the corresponding critical values for *r* for *p*-value = 0.05. This would dramatically—and without needing to keep the *z*-based figures like Figure 7—show how many of the simulated outcomes would be false positives, if left unscreened by considering the p-value.

Overall, this paper makes a valuable contribution to the p-value literature, if revised for improved clarity.