A constructive and efficient co-pilot approach to data analysis

Scientific advisory committee, dept. Psychology, UvA

Hilde Huizenga, Disa Sauter, Heleen Slagter, Astrid Homan & Eric-Jan Wagenmakers

June, 20, 2017

A few months ago, one of the authors of this paper (hh) wanted to conduct a regular ANCOVA. She had her results checked by a co-author, and the co-author indicated that she had mistakenly included interactions between the covariate and other independent variables. She was relieved that the co-author spotted this error -- if the co-author hadn't, reviewers might have rejected the manuscript or the mistake could have found its way into the published literature.

Errors in statistical analyses are common (Nuijten, Hartgerink, Assen, Epskamp, & Wicherts, 2015) and are of a varied nature. For example, analyses may not fit the type of data (e.g., an ANOVA executed for a nominal dependent variable), intended analyses may have not been carried out adequately (as in the ANCOVA example above), results may not have been reported adequately (e.g., degrees of freedom are incorrect), or results may not have been interpreted adequately. Such errors may be less likely if four, instead of two, eyes are involved in analysis and reporting, the so-called co-pilot approach to data analysis (Veldkamp, Nuijten, Dominguez-Alvarez, Van Assen, & Wicherts, 2014)

The purpose of this document is to highly recommend a co-pilot approach at the UvA Psychology department. A co-pilot approach should increase the likelihood that conclusions are supported by the data, thereby facilitating scientific progress.

Two potential concerns should be addressed when implementing a co-pilot approach. First, a careless approach could lead to an atmosphere of suspicion in which authors and "data police" have conflicting interests. We therefore propose that pilot and co-pilot act together in a constructive team. A second concern is that routine independent reanalysis and reporting is too time consuming, thereby sacrificing too much quantity for quality. We therefore propose to act in a stepwise manner that increases efficiency.

1. A constructive and efficient co-pilot approach

Below we outline a proposal for a constructive and efficient co-pilot approach consisting of the following steps:

a) When the study is designed, it is decided who will analyze the data (i.e., the analyzer), who will screen the data analysis (i.e., the screener) and who will, if necessary, re-analyze the data (i.e., the re-analyzer). The screener and re-analyzer might be the same person. All of the involved individuals agree on the data analytic plan[1,2]. It is the responsibility of the first author of the paper to organize

[1] If authors wish to pre-register the study, this is the moment to do so.

these roles. If the first author is supervised by another, e.g. in case of (PhD) students, the supervisor may help in organizing these roles.
b) The analyzer analyzes the data according to the data-analytic plan and writes up the methods and results section (or another overview of results).
c) The screener checks data-analysis by asking screening questions (examples below).
d) If any errors or conflicting results are detected, the data are re-analyzed again by the analyzer and the results section is rewritten accordingly.
e) After re-analysis and revision, the screener screens the results again. If no errors or conflicting results are detected, the authors proceed with writing up the remainder of the manuscript. If errors are detected, go back to c, or the analyzer provides all code and the re-analyzer reanalyzes with an updated code.

2. Screening questions

Screening questions are central to the constructive and efficient co-pilot approach, below we provide a list of questions which are already asked frequently in our department. The list below is meant to grow over time, as UvA Psychology members contribute additional screening questions. Please feel free to add these questions by sending an updated version of this document to Joost van der Meer: J.J.W.vanderMeer@uva.nl. Twice a year an updated version of this document will be made available.

2.1 General screening questions:

Coding:

1. Are missing values properly coded? Always inspect scatterplot and histograms of data to check whether, for instance, 999s are treated as missing values and not as regular data.
2. If data are transformed, is this done correctly? Examples include: text to numerical values, zeroes to non-zero values, merging variables into another variable, dots to semicolons, recoding of NA's, specification of conditions under which a variable should be transformed, etc. Sometimes it helps to check a few cases together.

Removal of data:

3. In case of multiple observations within participants, e.g. reaction times to multiple stimuli: Is it stated how many observations were removed from the analysis, for what reason, and according to which criteria?
4. Is it stated how many participants were removed from the analysis, and according to which criteria (e.g. outliers, failed manipulation checks, etc.)? The remaining sample size should be congruent with the error degrees of freedom in the analyses.

Required significance:

5. Is it specified whether testing was one-sided or two-sided? In case of one-sided: is a motivation given?
6. Is it specified whether there is a correction for multiple comparisons? If present: indicate which type of correction.

---

[2] In the following we assume that the data analytic plan is correct. For example, that an ANOVA is not used for nominal outcome variables, and that a two-sample *t*-test is not used for one sample with two repeated measures.

7. Is the required level of significance specified? If this level deviates from the commonly used .05, is this then stated and motivated?

Status of variables:

8. Is it clearly indicated which variables are independent and which ones dependent variables? Is it clearly indicated which variables act as moderators, mediators, and/or covariates? Does this match the theory and hypotheses put forward in the introduction?

(Interpretation of) results:

9. Are test results correctly copied (i.e., without typos) from the output into the main text / tables of the manuscript? Mistakes are often made because of copy-pasting the formatting of other reported results.
10. Are results in text, figures and tables consistent? For instance, if it is stated in a table that an effect is significant (e.g., by adding an asterisk), is it then also treated as significant in the text, and vice versa? Another example: Is the interpretation of the direction of an interaction effect consistent with how this interaction effect is depicted in a figure? And a final example: Is the discussion of a difference between conditions supported by the reported means and SDs in the text or a table?
11. Interpretation of rescored variables. Sometime variables are rescored, for example such that high values indicate good performance. Is interpretation of results consistent with this rescoring?
12. Are effect sizes (e.g. Cohen's d or a correlation) of a magnitude and in a direction which is consistent with the literature?

2.2 Specific: General linear model: regression, correlation, ANOVA, ANCOVA, t-test

Assumptions

13. Are assumptions tested?

Degrees of freedom

14. Are degrees of freedom correct? The hypothesis degrees of freedom denote the number of parameters to be tested in a specific effect, whereas the error degrees of freedom denote the total sample size minus the total number of parameters to be estimated.
15. If Greenhouse-Geisser of Huyn-Feldt correction or unequal variances t-test: Degrees of freedom are often non-integers, is this the case?
16. ANCOVA: is the interaction between the covariate and factor of interest omitted from the model (this can be checked via the error degrees of freedom, as omitting it should decrease the degrees of freedom by at least one)?

(Interpretation of) results

17. Are statistics, degrees of freedom, and p-values congruent? The program statcheck can check this: http://statcheck.io https://mbnuijten.com/statcheck/
18. Are all main and interaction effects listed, not only the significant ones?
19. Are the proper results used for interpretation? For example, if an interaction is predicted, is the conclusion based on a test of an interaction and not on the main effect? Similarly, if a hypothesis requires a specific contrast, are the results of a test on this contrast reported and interpreted?
20. For a regression analysis: Are dummy codings for categorical variables (e.g. gender, condition) clearly stated and do the results and interpretations of these results correspond with the coding

used? For example, if gender is coded as 0=male; 1=female, is a positive coefficient correctly interpreted as females scoring higher than males on the dependent variable?

2.3 Specific: Multilevel analysis for continuous dependent variables

21. Is the level one random effect covariance structure the one you intend to fit? Note that independent variables at level one can never be associated with a random effect.
22. Is the level 2 covariance structure the one you intend to fit?
23. Does the model converge? If not, a more restricted model should be specified.
24. Is there an error indicating a non-positive definite Hessian? If so, a more restricted model should be specified.
25. Are some random effects estimated as zero? If so, a more restricted model should be specified.

3. References

Nuijten, M. B., Hartgerink, C. H. J., Assen, M. A. L. M. van, Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, (2011), 1–22. https://doi.org/10.3758/s13428-015-0664-2

Veldkamp, C. L. S., Nuijten, M. B., Dominguez-Alvarez, L., Van Assen, M. A. L. M., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PLoS ONE*, *9*(12), 1–19. https://doi.org/10.1371/journal.pone.0114876