Column: Tips on Research and Publication

COMMON ERRORS IN REPORTING OF STATISTICAL ANALYSES

Samir Kumar Praharaj[1] Shahul Ameen[2]

[1]Professor & Head, Department of Psychiatry, Kasturba Medical College, Manipal, Manipal Academy of Higher Education, Manipal, Karnataka, India.

[2]Consultant Psychiatrist, St. Thomas Hospital, Changanacherry, Kerala

Correspondence: Department of Psychiatry, Kasturba Medical College, Manipal, Manipal Academy of Higher Education, Manipal, Karnataka, India. Email: samirpsyche@yahoo.co.in

First submitted: 19/04/2019; Published online on 26/4/2019

INTRODUCTION

The statistical analysis section in a manuscript is often poorly written. Such problems mostly arise out of statistical naivety: The authors seek statistical consultations, but they do not ask the details and often fail to incorporate all the necessary information about the tests performed, in the manuscripts. Many times, there is a discrepancy between what is described under statistical analysis section and the tests reported under the results section.

The statistical methods should be described in sufficient detail so that (1) given the data, another researcher can replicate the analysis, and (2) the results can be incorporated into other analyses like a meta-analysis. How much detail is to be actually written also depends on the word limit for the manuscript. General statements such as "descriptive statistics were used to summarize data" can be omitted.

RESEARCH DESIGN AND INSTRUMENTS

The research design should be adequately and correctly described. Do not try to cloak a study in one guise to try to give it the assumed reputation of another. For studies that have multiple goals, define and prioritize those goals. Operational definitions of all explanatory and response variables should be provided. If a questionnaire has been used to collect data, summarize the psychometric properties (like validity and reliability) of its scores with specific regard to the way it was used in a population. Any unusual statistical tests need to be described adequately and referenced. Also, the details of the statistical software with version should be specified.

Please cite this article as: Praharaj S K, Ameen S. Common errors in reporting of statistical analyses. Kerala Journal of Psychiatry 2018; 31(1): 46-50. doi: 10.30834/KJP.31.1.2019.153

PRELIMINARY ANALYSES

Mention the statistical procedures, if any, used to modify the raw data before the actual analysis. Examples include mathematically transforming continuous measurements to make distributions closer to the normal distribution, combining categories, etc. If any data transformations were carried out, keep this in mind while discussing the results. Also, the transformed values should be reported throughout the manuscript, including the tables and figures (e.g. "log x" instead of "x").

DESCRIBING CONTINUOUS DATA

Continuous data are presented as mean (standard deviation, SD) if normally distributed, and as median (interquartile range, IQR) for non-normal data. SD values more than half of the mean should alert the presence of possible non-normal data. Report the upper and lower boundaries of interpercentile ranges and the minimum and maximum values of ranges, not just the size of the range. The data should be plotted to visualize the distribution, and if necessary, appropriate statistical tests (e.g., Shapiro- Wilk test) can be used to confirm data normality. If in doubt, it is appropriate to report both mean (SD) and median (IQR).

Some authors report standard error of the mean (SEM) instead of SD; this does not give information on the dispersion of data. It is not appropriate to report mean and median values without SD and IQR. Sometimes, SD values are mentioned as ± [e.g., mean age 23.2 (±1.4) years]. To avoid confusion, identify the meaning of the interval at the first use. [e.g., "The mean ± SD was 23.2±1.4 years" or "The treatment resistance group had a statistically significantly lower mean (±SD) age at onset: (18.6 (±4.9) years)"].

Sometimes, the summary statistics are reported with higher precision than what was measured. For example, "mean age 34.628 (SD 5.24) years" does not reflect the actual data collected. Instead, it should be reported as "34.6 (S D 5.2)". Many suggest using only one decimal place more than the precision used to measure the variable. Also, consistency in reporting throughout the manuscript is to be ensured.

DESCRIBING CATEGORICAL DATA

Categorical data are presented as n (%). However, reporting only n and percentage, without mentioning the size of the sample for which the analysis was done [e.g., "14 patients (23.33%) had no improvement"], may be misleading if any data is missing. About writing the percentage, some guidelines suggest using one decimal point when n is more than 100, to use no decimal point when n is between 20 and 100, and to not mention the percentage at all if n is less than 20.

Categories are sometimes arbitrarily created from continuous data (e.g., the median split of a scale, or age categories). This reduces statistical power and should be avoided. Also, the choice of the cut-off could influence the results, especially if it is done after visualizing the data.

CHOICE OF STATISTICAL TEST

It is imperative to report how the underlying assumptions for the statistical tests were tested.

Parametric tests are carried out only when all their assumptions are fulfilled. If a distribution is non-normal, nonparametric tests should be reported. If there are paired data (e.g., pre- and post- data), then appropriate tests for such data should be used. It is not uncommon to see independent sample t-tests reported for paired data; in such situations, a paired t-test is

appropriate. When assumptions for the chi-square test are not met, the alternative Fisher's exact test should be reported. In published papers, many times, chi-square is reported when there are cells with zero values, which is wrong. Also, it is not uncommon to see a statistic value reported for Fisher's exact test, which requires reporting of P value only.

Multivariate tests are more powerful and adjust for confounding variables in observational studies. These should be reported whenever possible, but only if the underlying assumptions are valid. Time to event data should be analyzed using appropriate survival analysis method. Using only the event or only the time to event in those experiencing the event leads to loss of information and should be avoided.

It is advisable to choose the statistical tests *a priori* and to specify that in the protocol. Data-driven tests are not recommended. (For e.g., choosing one test over another as the P becomes 0.047 with the former in contrast to 0.051 with the latter.) Sometimes, it may be necessary to plan some analyses after looking at the initial results. Identify such post-hoc analyses, including unplanned subgroup analyses, as exploratory.

HOW MISSING DATA WAS HANDLED

It is not uncommon to have missing data in research. What was done for the missing data should be clearly stated. One strategy is to omit participants with missing values and report a complete case analysis. This can be problematic if a large proportion of data has missing values. Another strategy is to impute values for the missing data. Some of the imputation methods (e.g., mean values, last observation carried forward) can increase type I error and are best avoided. Multiple imputation and regression-based methods are less likely to introduce errors and are preferred. If the results change markedly based on whether complete case analysis or imputed data analysis was carried out, it is appropriate to report both the results.

REPORTING P VALUES

P values are reported when a hypothesis is tested. Usually, hypotheses are formulated for the primary outcomes of interest. The cutoff value of alpha (usually $P < 0.05$) should be specified, along with whether the hypothesis testing is one-tailed or two- tailed. Most often than not it is two-tailed, which is written as: "$P < 0.05$ (two-tailed)". It is not uncommon to see P values reported for sample characteristics, which is usually not necessary. Simple descriptive statistics is enough to describe the sample. If any characteristics appear unbalanced and could possibly influence the outcome, either stratified analysis or adjusted analysis could be reported.

Mentioning the exact P value (e.g., "$P = 0.02$") is more meaningful and a superior way of reporting than writing "$P < 0.05$". Similarly, non-significant values (e.g., $P = 0.068$) should not be mentioned as "$P = NS$" or "$P > 0.05$" as these are less precise. Also, using only *or ** for $P < 0.05$ or $P < 0.01$ should be avoided and full P values are to be mentioned.

It is unnecessary to report P values more than three digits after the decimal point (e.g., $P = 0.0147$ could be written as $P = 0.015$). It is not uncommon to see "P=0.000" mentioned in the published manuscripts. P value can never be zero; it is mistakenly written so for very small P values (e.g., $P = 0.00000025$). In such situations, it should be reported as "$P < 0.001$".

For P values between 0.05 and 0.1, many authors use terms such as "approaching significance" or "trends towards significance," which is better avoided. However, even in such situations,

comments can be made about the clinical importance, if any, of those results. Also avoid reporting only the P values in tables, without mentioning the statistical tests that were carried out.

CORRECTION FOR MULTIPLE HYPOTHESIS TESTING

Multiple comparisons increase the possibility of type I error. One way to reduce this problem is to adjust the P value for multiple testing. Bonferroni correction is a simple method for correction, in which the cut-off P value is taken as 0.05/n, where n is the number of comparisons. This has been criticized as being overly conservative, and alternate methods can be used (e.g., Holm or Hochberg procedure). Some authors use a lower cut off of $P < 0.01$, rather than using any specific procedure. This may be an acceptable alternative if there are not too many tests. The best way to avoid this problem, however, is to have only a few comparisons which are the most important, and those should be specified prior to the study.

REPORTING EFFECT SIZES

Measures of effect sizes should be reported along with P values. This gives an impression of the magnitude of the findings. Sometimes, effect sizes are reported without a mention of the actual measure, which can create confusion between small, medium and large. For example, r can have a maximum value of 1, whereas, Cohen's d can be more than 1. Any r value more than 0.5 is usually considered large, in contrast to 0.8 cut off for Cohen's d.

However, the interpretation of the obtained r value also depends on the nature of the study. For e.g., an r of 0.7 between birth weight and marital satisfaction would be unusually high because the relationship between these variables is obviously much more complex than suggested. On the other hand, an r of 0.7 between the results obtained by two different laboratories for the same test on the same blood sample may be considered low.

For measures of effect size other than the common ones (e.g. eta or omega), it would be useful for the readers to provide cut-off values for small, medium and large, with appropriate references. While reporting, errors such as confusing "partial eta" with "eta" are not uncommon and are better avoided.

INTERPRETING TESTS

Correlation is carried out for associations between variables. It is not uncommon to see correlation equated with causation, which is incorrect. Although the temporal sequence of events is evident in some situations (e.g., childhood adversity and current psychopathology), findings from correlation should be interpreted as associations only. Similarly, regression analysis carried out in cross-sectional studies is reported and interpreted as "predictors", which should be used only in longitudinal studies.

MISCELLANEOUS ASPECTS

All sample size calculations are done *a priori*. If it was not carried out before the study, do not bother the readers with post hoc power calculations. It can be clearly stated under limitations that no sample size calculations were carried out.

Outliers can be present and are not uncommon in data. If it is not a wrong entry during data capturing or while entering in statistical software, it could be real observed value. It is inappropriate to remove such outliers during data analysis. It can lead to spurious results.

Instead, sensitivity analysis can be carried out running the tests with and without the outliers to give the readers an estimate of their effect.

For primary comparisons, specify the degrees of freedom (df) of the test, if applicable. Student's t test, ANOVA, and the chi-square test all use the concept of degrees of freedom. This is specifically important in multivariate tests when some correction to the df value may be necessary (e.g., Greenhouse-Geisser correction in repeated measures ANOVA).

Similarly, reporting of 95% confidence intervals (CI) is more meaningful than the conventional P values. CI should be provided for all primary comparisons, regardless of whether the results are statistically significant or not. CI can be provided for differences between group means, mean changes in the same group over time, proportions, odds ratios, risk ratios, survival rates, etc. Trailing zeros is a common mistake observed in manuscripts. For example, 12.30 and 56.00 can be rather written as 12.3 and 56. This is especially common in figures, as the default settings in statistical software produce values up to two digits after decimal points. Editing such values improves readability.

CONCLUSIONS

Statistical testing is one of the most important aspects of quantitative studies. Hypotheses are accepted or refuted based on the tests applied. The correct interpretation of the tests will depend on transparent reporting of what was planned and what was done. Therefore, it is imperative that these tests are conducted and reported judiciously

SUGGESTED READING

1. Lang T, Altman D. Basic statistical reporting for articles published in clinical medical journals: the SAMPL Guidelines. In: Smart P, Maisonneuve H, Polderman A (eds). Science Editors' Handbook, European Association of Science Editors, 2013.
2. Lang TA, Secic M, Lang T. How to report statistics in medicine: annotated guidelines for authors, editors, and reviewers. ACP Press; 2006.
3. Wilkinson L and the Task Force on Statistical Inference. Statistical methods in psychology journals. Guidelines and explanations. Am Psychologist 1999; 54:594- 604.
4. Andrade C. Describing research design. Indian J Psychol Med 2019; 41:201-2.
5. Habibzadeh F. Statistical data editing in scientific articles. J Korean Med Sci 2017; 32:1072-6.

Source of Support: None Conflict of interest: None declared