Letter

Lowering the P-Value from 0.05 to 0.005 Conflicts with the 3R Rules – an Advocacy for Alternatives to Hypothesis Testing with the P-Value Approach

Konradin Metze[1], Fernanda Aparecida Borges da Silva[1] and Irene Lorand-Metze[2]

[1]Department of Pathology, Faculty of Medical Sciences, University of Campinas, Campinas, Brazil and National Institute of Science and Technology on Photonics Applied to Cell Biology (INFABIC), University of Campinas, Campinas, Brazil; [2]Department of Internal Medicine, Faculty of Medical Sciences, University of Campinas, Campinas, Brazil

There is increasing current discussion in the literature concerning the lack of reproducibility in science (Wasserstein and Lazar, 2016; Ioannidis, 2018; Smith, 2018; Trafimow et al., 2018; Van Calster et al., 2018), especially in preclinical animal research (Voelkl et al., 2018). Some authors claim that this might be due to low statistical standards and suggested that the significance level should be set at $p < 0.005$ in order to lower the number of false positive results (Benjamin et al., 2018; Ioannidis, 2018). But this suggestion has drawbacks, since larger sample sizes are required in order to maintain the statistical test power. We performed 400 computer simulations for Pearson correlations and one-way analyses of variance (ANOVA) applying the programs Primer of Biostatistics and Winstat 3.1 to calculate the increase of the necessary sample sizes at a test-power of 80% after lowering the threshold from $p < 0.05$ to $p < 0.005$.

For Pearson's correlation coefficient "r", the percentage of additional cases could be approximated by the formula: increase [%] = (0.77 -0.326 x r) x 100. For $r = 0.7$ the increase was about 50%, for $r = 0.5$ about 60% and for $r = 0.2$ approximately 69% of the initial sample size. When further lowering the threshold to $p < 0.001$ an estimate was given by: increase [%] = (1.32-0.64 x r) x 100. Then, for a correlation coefficient of $r = 0.7$ we have to add about 86%, for $r = 0.5$ about 100% and for $r = 0.2$ even about 116% of the original sample size.

Regarding ANOVA, we defined "q" as the quotient between the expected standard deviation of residuals and the minimal detectable difference, with $0.5 \leq q \leq 2$ in our simulations. Reducing the significance threshold to $p < 0.005$, we got the following results: for 2 groups, the amount of additional cases varied between 50% and 75%. With a higher number of groups this value dropped to 57-67% ($n = 3$ or 4) and converged to values between 50 and 55% for $n = 7$.

These simulations showed that a new threshold of $p = 0.005$ would make investigations far more expensive. In experimental studies, the number of animals needed would increase considerably. This would be incompatible with the aim of minimizing the number of animals used per experiment according to the 3R concept.

Fitts (2011) pointed out the problems concerning stopping rules: Defining 0.005 as significance level in experiments with the fixed-stopping rule may lead to a considerable waste of animals in case of a very small or nonexistent hypothesized effect. Moreover, in some exploratory data analyses, even variables with $p > 0.05$ can be interesting. For example, in multivariate regression models all variables with $p < 0.1$ in the univariate analyses are usually included for the calculation of the multivariate final model so as not to miss relevant variables (Ferro et al., 2011). Reducing the α-level may considerably increase the overestimation of significant effects by up to 320% (Van Calster et al., 2018). Trafimow et al. (2018) considered the significance level at $p = 0.005$ "deleterious for the finding of new discoveries."

According to the American Statistical Association, a p-value estimates the "incompatibility between a particular set of data and a proposed model for these data". Smaller p-values decrease the statistical compatibility of the data with the null hypothesis. P-values are not estimates of the effect size or the biological relevance (Wasserstein and Lazar, 2016). Usually p-values are used for simple accept/reject decisions of scientific hypotheses. Yet, dichotomization of continuous variables has shown itself to be a potentially misleading procedure in statistics due to the loss of information (Metze, 2011a,b), and this also applies to the p-value. It is difficult to find any fundamental difference between $p = 0.0505$ and $p = 0.0495$, although only the latter is usually considered to be "significant".

According to the American Statistical Association, "scientific conclusions should not be based only on whether a p-value passes a specific threshold" (Wasserstein and Lazar, 2016). So what to do? It is recommended to calculate the effect size which expresses the relevance of a biological phenomenon. The effect size cannot be estimated by the p-value, because small and irrelevant

Received July 9, 2018; Accepted July 13, 2018;

ALTEX 35(4), 516-517. doi:10.14573/altex.1807091

Correspondence: Konradin Metze, PhD, Universidade Estadual de Campinas, Faculdade de Ciências Médicas da UNICAMP, CP 6111, Barão Geraldo, 13081-970 - Campinas, SP - Brasil
(kmetze@fcm.unicamp.br)

differences may have a significant p-value when the sample size is large and *vice versa* (Smith, 2018). Furthermore, it is necessary to add the 95% confidence interval (CI95%). The CI95% is defined in the following way: repeating the investigation many times with other samples, 95% of the CI95%s would include the "true" effect size of the whole population. The width of the CI95% is an estimate of the uncertainty of the effect, even when a result is statistically significant, and can thus indicate whether larger studies are necessary. Overlapping confidence intervals indicate areas of agreement between studies, even if in one of them the p-value for hypothesis testing was not significant (Smith, 2018). Furthermore, graphical representations of the data would be helpful, since they help to identify outliers which may hamper the results (Smith, 2018; van Calster et al., 2018).

In summary, hypothesis testing based on the p-value approach has disadvantages. Lowering the p-value to $p = 0.005$ conflicts with the 3R principles. The calculation of the effect size together with its CI95% and graphical demonstrations are good alternatives.

References

Benjamin, D. J., Berger, J. O., Johnson, V. E. et al. (2018). Redefine statistical significance. *Nat Hum Behav 2*, 6-10. doi:10.1038/s41562-017-0189-z

Ferro, D. P., Falconi, M. A., Adam, R. L. et al. (2011). Fractal characteristics of May-Grünwald-Giemsa stained chromatin are independent prognostic factors for survival in multiple myeloma. *PLoS One 6*, e20706. doi:10.1371/journal.pone.0020706

Fitts, D. A. (2011). Ethics and animal numbers: Informal analyses, uncertain sample sizes, inefficient replications, and type I errors. *J Am Assoc Lab Anim Sci 50*, 445-453.

Ioannidis, J. P. A. (2018). The proposal to lower p value thresholds to .005. *JAMA 319*, 1429-1430. doi:10.1001/jama.2018.1536

Metze, K. (2011a). Pitfalls in the assessment of prognostic factors. *Lancet Oncol 12*, 1095-1096. doi:10.1016/S1470-2045(11)70309-6

Metze. K. (2011b). Dichotomizing continuous prognostic factors can cause paradoxical results in survival models. *J Am Coll Surg 212*, 132-134. doi:10.1016/j.jamcollsurg.2010.10.004

Smith, J. R. (2018). The continuing misuse of null hypothesis significance testing in biological anthropology. *Am J Phys Anthropol 166*, 236-245. doi:10.1002/ajpa.23399

Trafimow, D., Amrhein. V., Areshenkoff, C. N. et al. (2018). Manipulating the alpha level cannot cure significance testing. *Front Psychol 9*, 699. doi:10.3389/fpsyg.2018.00699

Van Calster, B. Steyerberg, E. W., Collins, G. S. and Smits, T. (2018). Consequences of relying on statistical significance: Some illustrations. *Eur J Clin Invest 48*, e12912. doi:10.1111/eci.12912

Voelkl, B., Vogt, L., Sena, E. S. and Würbel, H. (2018). Reproducibility of preclinical animal research improves with heterogeneity of study samples. *PLoS Biol 16*, e2003693. doi:10.1371/journal.pbio.2003693

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *Am Stat 70*, 129-133. doi:10.1080/00031305.2016.1154108