Science and Statistics

My first contact person within medical science in the field of experimentation was the head of an Institute of Physiology. He, coming from the Göttingen University, told me: "*If I am familiar enough within my particular area of physiology, and if I have a scientific issue, I should develop a hypothesis. In order to test the hypothesis, I need to develop, define and test an experimental model and then perform an experiment. If the data from the experiment are confirmative, I will publish my issue*". This advice reaches back almost 50 years.

In the meantime, we had quite a few scientific issues, and we performed more than one experiment. We used descriptive statistics to present our data, having in mind that averaging was data murder. Still, we presented pre- and post-means ± SDs and graphs, and the journals were happy. (NB: At that time, our work was being paid by the publisher). Times changed and the situation has become more complex. The effects of different interventions were to be compared at different points in time, and we understood that our multiple *t*-testing was witches' brew. So, we learned to differentiate terms as ANOVA from MANOVA. If these analyses were not in the focus of our scientific activities, we needed to contact a statistical ambulance. Anyway, it took time prolonging the project.

In parallel, the statisticians' influence grew, such that the journals demanded the Methods section to be expanded by a statistics paragraph. We are now exhorted to include how the sample size was determined, why blinding and random assignment was warranted or not warranted, whether or not the groups were matched, and how the nature of the data distribution was tested. Finally, the climax of statistics – which test should the authors employ to determine whether the differences were significant with the *P*-value being ≤ 0.05.

Some journal's statisticians even wanted to read: the *P*-value was 0.034. How much does that contribute to a better understanding the effect of an intervention? To exemplify my displeasure: One study might compare the effects of air versus oxygen-enriched air (Nx) on the minute ventilation while intense fin swimming. The result: ventilation of air is higher over Nx, the difference being $0.3\ L{\cdot}min^{-1}$, i.e., has no clinical importance. How can this difference be statistically significant? Because of the 850 participants.

To remember: Researchers want to answer reasonable questions using reasonable experimental models. To do so, the researchers need to be creative, but also firmly founded in scientific reasoning. Statisticians at journals sometimes seem to misunderstand their role. They are important adjuncts, but they are not the protagonists. Nevertheless, often enough statistics became the Cerberus refusing admittance to the publication world.

Hope comes from the '*P*-value statement' of the American Statistical Society.[1] Ron Wasserstein (ASA's executive director) is to be admired in this context: the *P*-value was never intended to be a substitute for scientific reasoning. And he continues: Well-reasoned statistical arguments contain much more than the value of a single number with an arbitrary threshold. ASA is intended to steer research into a 'post $P < 0.05$ era'. As one result: The editors of *Basic and Applied Social Psychology* decided radically and banned *P*-values.

Expectedly, not all journals will react so radically. Maybe, the coming generation of statisticians will become ASA followers. A Nature article titled "*Scientific Method: Statistical Errors*"[2] might be helpful to step into the new era.

References

1 Wasserstein RL, Lazar NA. The ASA Statement on *P*-values: Context, process, and purpose. Am Stat. 2016;70:129–33. doi: 10.1080/00031305.2016.1154108.

2 Nuzzo R. Scientific method: statistical errors. Nature. 2014;506:150–2. doi: 10.1038/506150a.

Submitted: 03 February 2021
Accepted: 19 March 2021

doi: 10.28920/dhm51.2.230. PMID: 34157744.

Jochen D Schipke
Research Group, Experimental Surgery, University Hospital Dusseldorf, Germany
j.schipke@gmx.org

Key words
Effect size; *P*-value; Scientific reasoning; Statistic; Letters (to the Editor)