Rianne de Heide
Faculty of EEMCS
University of Twente
r.deheide@utwente.nl

Column Rianne takes her chance

A plea for a new statistical paradigm

Rianne de Heide regularly writes a column on everyday statistical topics in this magazine.

By the second half of the nineteenth century, the predominant view of science was that of a *clockwork universe*: a small set of mathematical formulas (like Newton's laws) would suffice to describe and predict the deterministic world around us. Discrepancies between the theory and the observations were attributed to measurement imperfections or human error, Laplace captured all these in an *error function*, and it was believed that by the advancement of measuring methods, this function would decrease. It did not. By the beginning of the twentieth century, science had shifted to the new paradigm of an inherently stochastic world.

In 1935, Sir Ronald Fisher wrote a book titled *The Design of Experiments*, which includes the famous second chapter about 'A Lady Tasting Tea'. It was an important contribution to the quantitative revolution in all fields of science in the first half of the twentieth century. (The Lady Tasting Tea was a randomised experiment, devised by Fisher, to test whether a lady could identify by tasting whether the milk or the tea was poured first into a cup. It was his original exposition of the notion of a *null hypothesis*.) The standard statistical framework for testing hypotheses, designed by Ronald Fisher for the analysis of small-scale agricultural experiments, is now taught to students in almost every scientific discipline from ecology to economics, and used in a very large proportion of scientific papers. As a field, statistics has grown, matured, and diversified.

However, at a fundamental level, the current state-of-the-art in statistical methodology still reflects the use-context of Fisher's day, where researchers would engage in well-controlled, well-planned, single experiments with a prespecified research question and a set end date. The standard statistical methodology, a combination of Fisher's p-values and Neyman and Pearson's null hypothesis significance testing, still assumes that the researcher makes all important analysis choices a priori, before gathering the data. This includes sample size, the research question(s) and the statistical model. There exist modern, more sophisticated designs, e.g. with *interim analyses* (you may have heard of *alpha spending*, *group sequential trials*, and the like), where multiple looks are allowed, but these still have to be pre-planned. The underlying mathematics for all these *classical* methods relies heavily on the assumption that analysis choices are independent of the data. Important properties, such as error guarantees, are lost when this assumption is violated, leading to an excessive risk of false positive findings.

Classical statistics has become a straitjacket for researchers: it imposes a normative model for researcher behaviour. By only offering methods for the classical use-context—which was fine for the small-scale twentieth-century agricultural experiments—researchers are forced to practise science in a way the statistical methods impose. We find ourselves, however, in the age of modern data science with endless computational possibilities. Data often comes in gradually, and it makes sense for the researcher to interact with it. Many measurements are gathered simultaneously and it is often natural to pursue research questions that were not of obvious interest a priori. Data are often observational, and statistical models may not be easy to specify. In all these cases, the classical paradigm feels rigid, inflexible, inefficient, and overly restrictive.

We, as statisticians, accuse practitioners of *sloppy science* when they look at their data during the analysis. Talk after talk we explain the underlying mathematics to non-mathematicians, to show why their false positive rates blow up when they do *optional stopping*. But why should we fight this losing battle, telling everyone how they shouldn't do their research, because the statistical methodology cannot handle it? And then proclaim a *reproducibility crisis* if it turns out that everyone *did* look at their data and our scientific edifice crumbles on the false positives.

Rather than trying to adapt researchers to the assumptions of classical statistical methods, we should develop the right mathematical framework to accommodate actual researcher behaviour. They should be able to look at data as they come in, end the experiment as soon as results become clear, and refocus their analysis on the striking results they happen to find. This all makes perfectly sense from the researcher's viewpoint, and we, mathematical statisticians, should develop the theory to accommodate for this: a novel statistical paradigm enabling interactivity while retaining strong error control guarantees. This is my mission.

Acknowledgements
Many thanks to Jelle Goeman for a lot of interesting conversations about this topic. And to Peter Grünwald for inspiring and coaching a whole new generation of statisticians to develop this new paradigm.