**The .05 level of significance and the economy of research**

By: Cornelis Menke

The 5%, or .05, level of significance in "Null Hypothesis Significance Testing" (NHST) has become one of most general standards within science. It is enforced by publication policies of more than 1,000 academic journals that adopted the so-called APA Style of the American Psychological Association. At the same time, the adoption of the .05 level of significance poses a puzzle, historically and systematically: historically, it was adopted while no statistician recommended it; systematically, the adoption of a fixed 5% standard is foreign to all major statistical approaches, frequentist an Bayesian alike.

There are two parts to the paper. In the first part, I shall consider the origin of the .05 level in the practice of agricultural field experiments in the 1920s, and argue that its meaning and function has been misunderstood: it does not express the frequency of an error of the first kind, but the probability, and thus costs, of being mislead by experiments in future research. In the second part, I shall argue that this interpretation allows for a better analysis of some aspects of the crisis of confidence in recent statistical practises. While NHST has often been criticised for its logic of inference (e.g., that there is 'no evidence for the null' in NHST), I shall argue that at least some forms of Questionable Research Practices (QRPs) like p hacking are bests understood as being rooted in the economy, not the logic, of research: p hacking is an externality (in the economicsts' sense). This analysis allows for a (partial) remedy, too.

1. It is known that the choice of $P < .05$ originated from the work of Ronald A.
Fisher on tests of significance in the 1920s. While Fisher proposed it as a "convenient standard of significance", he never explicitly stated his reasons. According to the received view (due to antagonist statistician Jerzy Neyman), the level of significance of a test expresses a frequency, viz., the frequency of 'falsely rejecting the null', or error of the first kind. I shall consider the origin of the .05 standard within Fisher's work on agricultural field experiments at Rothamsted Experimental Station and argue that the choice is best understood as being based not on general considerations relating to a standard of evidence but to considerations concerning the "economy of research" which are typical for the field experiments pursued at Rothamsted.

The origin of .05 level At the center of Fisher's considerations, or so I shall argue, is a trade-off between two types of costs connected with field experiments: costs of future possibly fruitless experiments on the one hand, and the costs of field experiments themselves, on the other. Choosing a loose standard of significance considerably reduces the latter; at the same time, it increases the risk of being "misled" by experimental results that are only seemingly promising. Within field experiments, a level of significance balancing the two types of costs has to take into account features peculiar to the design of field experiments: large fields allow for a more demanding standard of significance, but do not necessarily reduce the overall experimental error, since, for instance, the larger field experiments are designed, the larger the effects of soil heterogeneity are. Only when interpreted as an error frequency is the .05 standard an arbitrary convention; in the practice of field experiments, it is a meaningful standard for the design of experiments based on considerations of the economy of research.

2. Interestingly, considerations of the economy of research – the costs and benefits of statistical practises – re-emerged in the context of the crisis of confidence of statistics. Since its development in mid-nineteenth century, NHST has been charged with logical flaws and problems. But the problems of

NHST could fruitfully be understood in terms of costs and benefits, too. False positive results are not just 'falsely accepted', but lead to further costs for future research. Moreover, false positive results are resilient, methodologically as well as socially: methodologically, since there is "no evidence against the null" – no systematic method to disproof a null hypothesis but by attempted replications – socially, since failed replications are not easy to interpret unambiguously, and furthermore hard to publish.

That there is no evidence for the null in NHST is traditionally considered a (methodo)logical flaw. However, with reference to the economy of research, the problem might equally be conceived of in terms of the economy of research, viz., externalities: it is only since the costs of fruitless further research do not have to be payed by researchers themselves, that there are incentives for scientists to engage in Questionable Research Practices like p hacking.

This economic approach to the replication crisis inspires a new possible remedy. Usually, the answer to the replication crisis is either sought in abandoning frequentist statistics in favour of Bayesian approaches, or in measures to prohibit QRPs by, for instance, pre-registration of statistical studies. The final part of the talk will briefly indicate a different possible answer: from an economic point of view, part of the problem is the combination of (i) cheap and (ii) externalised costs of QRPs. I shall summarise results from computer simulations which show that the costs of QRPs depend heavily on the exact level of significance chosen; by choosing a lower level (say, .01), the costs of QRPs become so high that in normal experimental designs there is no incentive to pursue these practices.