Some people present a statistical significance like a trophy. It is possible that no other statistical result is misinterpreted as often as the p-value. The two most important rules for a decent interpretation of p-values are: (1) “Never interpret a p-value without the appropriate descriptive statistic.” and (2) “Distinguish between explorative and confirmative testing.” This presentation deals with the latter. There are two steps to experimental research: hypothesis generation and hypothesis confirmation. A p-value can therefore be explorative, i.e. it produces a new hypothesis or it is confirmative, i.e. it provides “statistical proof.” The distinction between explorative and confirmative statistics is not linked to the type of the outcome variable (qualitative or quantitative), the number of groups (1, 2 or more groups), the design of the study (parallel groups or cross-over settings), nor to the statistical test applied (chi-square test, Wilcoxon test, analysis of variance, etc.). In order to allow for a confirmative interpretation of p-values, two prerequisites must be fulfilled: Firstly, a precise hypothesis has to be established beforehand independent from the data now used for the statistical test at that point. Secondly, only one single statistical test has been computed or the p-values have been adjusted for multiple testing. Let us do an experiment in our minds in order to understand the meaning of the p-value. Let us assume that we have data coming from a random number generator – in modern times this is not a dice but by a computer program – and that there is definitely no effect. Here, the probability to obtain a false significant result is the p-value. If 300 tests are done with random numbers, for example, each of them with a level of significance of 5%, then we expect 300 * 0.05 = 15 false significant results. A reasonable scientist will not compute all of these 300 tests. If he can deduct from the means or the frequencies that there is no chance of a significant result, he will save unnecessary work. But formally, all of these 300 tests are done. The purpose of toxicological screening programs is to identify all kinds of interesting things within the considered field. Therefore, they are explorative. The words “to screen” and “to explore” already have a similar meaning. Let us now look at the US National Toxicology Program’s Technical Report TR 578 on Ginkgo biloba as an example. In this report, a 3-month study on rats, a 3-month study on mice, a 2-year study on rats, and a 2-year study on mice are being reported. In this investigation, around 1000 statistical tests are possible. If the data stemmed from a random number generator and if a level of significance of 5% was selected, then we expect 1000 * 0.05 = 50 false significant results. The number of reported significant findings is roughly the same as the number of expected false significant tests. Gaus (2014) points out that all significant findings in TR 578 are explorative and that they generate new hypotheses. Kissling et al (2014) are more or less involved in the NTP and have rejected it passionately. This is an indication that the distinction of explorative and confirmative testing is important, but not widespread. Furthermore, we recommend to not only look at p-values, but to also consider a wide variety of information sources in a cross-matching approach (Heinonen 2015).

Biography
Wilhelm Gaus, PhD was head of the department of Medical Documentation and Biometry of the Medical Faculty of the University Ulm, Germany for more than three decades. In 2004 he retired but is still active. Totally he published about 250 papers. He was especially active in proving efficacy and safety of herbal medicines. In 2014 he published a textbook on Medical Statistics.

wilhelm.gaus@uni-ulm.de