Peter Grünwald (CWI and Leiden University)

Safe Testing

A large fraction (some claim > 1/2) of published research in top journals in applied sciences such as medicine and psychology is irreproduceable. In light of this 'replicability crisis', standard p-value based hypothesis testing has come under intense scrutiny. One of its many problems is the following: if our test result is promising but nonconclusive (say, p = 0.07) we cannot simply decide to gather a few more data points. While this practice is ubiquitous in science, it invalidates p-values and error guarantees.

Here we propose an alternative hypothesis testing methodology based on supermartingales - it has both a gambling and a data compression interpretation. This method allows us to consider additional data and freely combine results from different tests by multiplication (which would be a mortal sin for p-values!), and avoids many other pitfalls of traditional testing as well. If the null hypothesis is simple (a singleton), it also has a Bayesian interpretation, and essentially coincides with a proposal by Vovk (1993), and is similar to a proposal by Berger, Brown and Wolpert (1994). We work out the case of composite null hypotheses, which allows us to formulate safe, nonasymptotic versions of the most popular tests such as the t-test and the chi square tests. Safe tests for composite H0 are *not* always Bayesian, but rather based on the 'reverse information projection', an elegant concept with roots in information theory rather than statistics.