Safe Testing: online, anytime valid hypothesis tests

Rosanne J. Turner[1,2], Alexander Ly[1,3], Judith ter Schure[1] and Peter D. Grünwald[1]
1: Machine Learning Group, Centrum Wiskunde & Informatica*
2: Brain Center, UMC Utrecht
3: Psychological Methods, University of Amsterdam
Email addresses: rosanne@cwi.nl, a.ly@uva.nl, schure@cwi.nl, pdg@cwi.nl

Keywords (max 5)
Online learning, hypothesis testing, Bayes factors, sequential testing

Purpose and background

We present implementations of safe testing: a new kind of hypothesis testing (e.g., A/B testing), which can be used in the online setting with anytime valid error guarantees and facilitates effortless combining of multiple experiments. The currently available classical hypothesis tests, such as the two-sample *t*-test (for comparing means of two strategies) or the chi-squared test (for comparing proportions of two strategies) only provide guarantees on the chance of making a *wrong decision if the number of samples for each experiment and the number of experiments are fixed in advance*. This means these tests should not be used in the *online* learning setting. Furthermore, they even do not provide guarantees when experiments are conducted *sequentially*, when the decision to start a new experiment is based on previous results (ter Schure & Grünwald, 2019).

This lack of flexible hypothesis tests greatly restricts the possibilities of dynamic, on-the-fly learning. However, since feasible alternatives have been lacking so far, researchers have been using the classical tests in flexible settings. As the chance of wrongly deciding that the alternative hypothesis is true (*the type I error*, e.g., wrongly deciding that the bigger advert attracts more clicks, or that the new drug works better than the placebo) is not controlled for in this situation, this faulty use of methodology probably is one of the major causes of the current *replication crisis* in science. A large part of “new findings” in scientific papers cannot be replicated by subsequent studies, suggesting the initial finding was *false*, and large amounts of resources were wasted on expensive follow-up studies and new strategies.

Safe tests

Safe tests do allow for such flexibility. With safe tests, S-values are used as a notion of evidence for the alternative hypothesis in the experiment. The prime interpretation of S-values is very intuitive, in terms of *investing*, with each S-value corresponding to the profit or loss resulting from an *investment in the alternative hypothesis.* The higher the S-value, the more evidence the experiment reveals for the alternative hypothesis. When an S-value exceeds a certain threshold based on the required type I error guarantee, the null hypothesis can be rejected. When needed, an S-value can also be converted into a p-value, which allows interpretation within the classical hypothesis-testing framework. Surprisingly, an overall S-value can be computed by multiplying S-values of individual experiments, and this combined S-value still has the same type I error guarantee.

It has been shown that in theory S-values exist for completely general testing problems, for any composition of the hypotheses one wants to compare (Grünwald, de Heide, & Koolen, 2019). To ensure that the S-values quickly yield evidence when the alternative hypothesis is true, “GROW” S-values can be used, which lead to fastest *growth* of our *investment*. These are fully characterized by the joint information projection (JIPr) between the set of all Bayes marginal distributions on the null and alternative hypotheses. Thus, optimal S-values also have an interpretation as Bayes factors, with priors given by the JIPr. These

*Postal address: P.O. Box 94079, 1090 GB Amsterdam, Telephone: +31205929333

S-values often turn out to have a special form, for example, some are nonnegative supermartingales, for which it is known that they can be used in the online learning setting while retaining error guarantees.

Results and short discussion

GROW S-values were developed to provide safe alternatives for *t*-tests (one-sample, two-sample, paired, and two- and one-sided), for correlation tests, and for tests of two proportions (Fisher's exact test or the chi-squared test). It turns out that (GROW) S-values for the *t*-tests and tests of two proportions can be composed by adopting discrete, 2- or 1-point priors on the parameters from the null- and alternative hypotheses, which means that *computation of S-values is straightforward, and can be executed efficiently in the online setting.*

Performance of safe tests was compared to the performance of their classical equivalents. It turns out that, when using safe tests in the online setting, the decision to stop testing can often be made before we would have stopped data collection with a classical test setup, whereas classical tests need to be performed as they were planned. This means that with safe testing, we can *save resources and that we obtain answers to our research questions faster*. For example, when one wants to test that proportions in two groups, treated with two different strategies, differ at least 0.2, with Fisher's exact test, one would need to collect 220 samples to yield a test with a power of 0.80. With the safe test for two proportions, we would collect on average 194 samples in the online setting, and in 65 percent of the experiments we would need fewer samples than in the classical setup.

In conclusion, safe testing provides exciting new possibilities for testing in the online setting. Resources can be saved, as one can decide on the fly that enough evidence for a hypothesis has been collected, and experiments can be stopped early. In comparison to previously designed sequential tests, such as Wald's likelihood ratio test (Wald, 1941), S-values are remarkably more flexible; with Wald, one can only stop testing *once enough evidence has been collected for either hypothesis A or B*. So, when no evidence is collected for either of the hypotheses, possibly because the truth lies in the middle of A and B, one would have to keep collecting data infinitely. With safe testing, a stopping time (or, sample size) can be determined in advance, and one either stops because enough evidence for the alternative hypothesis is collected, or the experiment goes on as planned. After the stopping time, one could then decide to start a second experiment, perhaps even with a *different primary outcome measure*, and combine the S-values of the experiments to try to gain enough evidence for the alternative.

References

Grünwald, P. D., de Heide, R., & Koolen, W. (2019). Safe Testing. *arXiv:1906.07801*.

ter Schure, J., & Grünwald, P. D. (2019). Accumulation Bias in meta-analysis: the need to consider time in error control. *F1000 Research*.

Wald, A. (1941). Asymptotically most powerful tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 12:1-19.