Safe tests for 2 x 2 contingency tables and the Cochran-Mantel-Haenszel test*

Rosanne J. Turner[1,2**]
Thesis supervisor: Prof. Peter. D. Grünwald[1,3]

[1] Machine Learning Group, Centrum Wiskunde & Informatica, Netherlands
[2] Brain Center, University Medical Center Utrecht, Netherlands
[3] Mathematical Institute, Leiden University, Netherlands

Keywords: Online learning · Bayesian learning · Information theory.

Rationale In my thesis, I developed implementations of *safe statistics*: a new framework for collecting evidence for hypotheses, particularly suitable for online and sequential learning [1]. Currently, p-values and (frequentist) confidence intervals are the most widely-used methods for collecting evidence for hypotheses. However, with these methods, error bounds are only guaranteed if the number of samples for each experiment *and* the number of experiments are fixed in advance. This means these statistics should not be used in an online setting (a prototypical example is A/B testing); would one do this anyway, the probability of obtaining false "significant" results would approximate 1 as the number of data points collected grows. Since feasible, easily implementable methods that are robust under online use have not been available to the research community, classical methods have been used anyway, with many expensive false-positive findings as a consequence.

Similarly, standard statistics also do not provide guarantees in the common situation that experiments (e.g. randomised trials) are conducted sequentially, when the decision to start a new experiment is based on previous results [3]. It directly follows that meta-analysis results, and even combined evidence from multiple experiments performed within the same research group can be misleading. The safe statistics framework provides methods that *can* be used to analyse data in real-time, and to effortlessly combine statistics from sequential experiments.

Safe statistics Within the safe statistics framework, random variables called E-variables[4] are used to represent the *evidence* for a hypothesis in the data. By definition, an E-variable is a nonnegative random variable that has an expected value of at most 1 under the *null hypothesis* $\mathcal{H}_0$. The higher an E-value, the more evidence there is in the data in favour of the *alternative hypothesis* $\mathcal{H}_1$. From

* Two-page abstract of the master thesis written by Rosanne. J. Turner at Leiden University for the Master Statistical Science for the Life and Behavioural Sciences, defended September 23, 2019, see [4].

** Corresponding author: Rosanne J. Turner, rosanne@cwi.nl

[4] called S-variables in the previous versions of the framework and my master thesis

the definition of E-variables, it can straightforwardly be derived that when we use the rule that we reject $\mathcal{H}_0$ when the E-value exceeds $\frac{1}{\alpha}$ for some $\alpha \in [0, 1]$, we have a test where the probability of falsely rejecting the null is bounded by α. The definition also implies that all E-variables can be used in the sequential setting simply by multiplying them. It also turns out that a special subset of E-variables can be used in the online testing setting [1].

To optimise the amount of evidence collected, an information-theoretic criterion for *good* E-variables was defined: GROW, which stands for *Growth Rate Optimal in the Worst case* [1]. GROW E-variables tend to grow fastest for some alternative hypothesis $\mathcal{H}_{1,\delta} : \{P_{\theta_1} : \theta_1 \in \Theta_1(\delta)\}$ defined by a distance metric δ, even *in the worst case* scenario where data are generated by a distribution in $\mathcal{H}_{1,\delta}$ that yields little evidence. It turns out that these GROW E-variables have the form of *Bayes factors* and can be derived for any pair of hypotheses $\mathcal{H}_1$ and $\mathcal{H}_0$ [1], but the corresponding prior distributions are sometimes completely different from what Bayesian machine learners or statisticians would normally use.

Results and short discussion For this thesis, I developed GROW E-variables equivalent to two classical frequentist hypothesis tests: the two-by-two contingency table test and its stratified version, the Cochran-Mantel-Haenszel test. Two versions of the E-variable were developed. For the first version, $\mathcal{H}_{1,\delta}$ was defined with δ the Kullback-Leibler divergence. This E-variable could be useful when one wants to design a test optimised for distributions that would yield a certain minimal growth rate if they would generate the data. For the second version, $\mathcal{H}_{1,\delta}$ was defined with δ the absolute difference between the proportions. Such an E-variable is useful when one has more clear ideas about the applied goal of the experiment and wants to detect a *minimal* difference between two groups.

For the 'minimal absolute difference' version, the GROW E-variable was derived analytically. I showed that when using this E-variable in an online, real-time fashion, the expected sample size needed to achieve a desired power can be lower than when using its classical equivalent, Fisher's exact test. No analytic expression could be found for the Kullback-Leibler version: this GROW E-variable has to be found through numerical optimisation. Nevertheless, the Kullback-Leibler version could still be preferred in some cases: it was shown to gain higher power for certain data-generating distributions compared to the absolute difference E-variable.

Both E-variables were implemented in the Safestats R package, a collaborative project with other machine learning researchers from Amsterdam [2]. The work in this thesis gave rise to some interesting follow-up questions, such as the development of 'most powerful' GROW E-variables, safe confidence sequences for proportions, and applications of E-variables for healthcare research, and is continued in my current PhD project.

References

1. Grünwald, P., de Heide, R., Koolen, W.: Safe testing. arXiv preprint arXiv:1906.07801 (2019)
2. Ly, A., Turner, R.J.: Safestats: an R package for safe, anytime-valid inference. `https://github.com/AlexanderLyNL/safestats`, accessed: 2020-08-28
3. Ter Schure, J., Grünwald, P.: Accumulation bias in meta-analysis: the need to consider time in error control. F1000Research **8** (2019)
4. Turner, R.J.: Safe tests for 2x2 contingency tables and the Cochran-Mantel-Haenszel test. Master thesis. `https://www.universiteitleiden.nl/binaries/content/assets/science/mi/scripties/statscience/2019-2020/thesis_rjturner_for_publication.pdf` (2019), accessed: 2020-10-22