Rianne de Heide
Faculty of EEMCS
University of Twente
r.deheide@utwente.nl

Research

The e-value

Recently, the e-value has been featured in several Dutch national newspapers: *Trouw*, *De Volkskrant*, *NRC*, *The New Scientist*, and others, published articles and interviews, in connection with two events. Firstly, in January, Peter Grünwald, Rianne de Heide and Wouter Koolen presented their pioneering paper titled 'Safe Testing' [2] at the Royal Statistical Society in London. Secondly, early April it was announced that Peter Grünwald receives an ERC Advanced grant, totaling 2.5 million euros, for research into e-values. In the last few years the research into this new theory of hypothesis testing with e-values grows almost exponentially, software becomes available, and the first applications get off the ground. A good moment for NAW to ask Rianne de Heide: what is this e-value exactly, and what is the Safe Testing paper about?

A trial

When my child turned four months old in February, we moved on to size 2 of the reusable nappies. For the night one can add an extra absorption layer, called a *booster*, and when I ordered some, the nappy store asked me whether I would like to test a new kind of boosters. I agreed and decided to make a trial out of it. We have two groups:

- Group A: the ordinary boosters.
- Group B: the new boosters.

We collect the data sequentially: after each night we obtain a new data point. The outcome is binary: no leakage (encoded with a 0), or leakage (encoded with a 1). We assume that the data are independently and identically distributed according to a Bernoulli distribution with a certain parameter — the probability of leakage — $\theta_j \in [0,1], j \in \{a,b\}$. We thus have two data streams $Y_{1,a}, Y_{2,a}, \ldots$ i.i.d. $\sim P_{\theta_a}$ and $Y_{1,b}, Y_{2,b}, \ldots$ i.i.d. $\sim P_{\theta_b}$, and we wonder whether there is a significant difference between θ_a and θ_b, which tells us that there is a difference in the probability of leakage.

	1	2	3	4	5	6
ordinary boosters	0	0	1	0	0	0
new boosters	1	1	1	1	1	1

Table 1 The data: one measurement contains two observations: one with the ordinary booster and one with the new one. A zero indicates no leakage, a one indicates leakage.

We take as significance level $\alpha = 0.05$. The null hypothesis and the alternative hypothesis are as follows:

- H_0: $\theta_a = \theta_b$,
- H_1: $\theta_a \neq \theta_b$.

We collect the data in pairs: we alternate the nights with the ordinary and the new boosters. After each pair we can calculate the *e*-value: the conclusions and Type I error guarantees remain valid if we do so, and we decide on the basis of the results so far, whether we continue to collect data, or whether we stop the trial. As I will explain later, the usage of e-values makes this easily doable. In [3] it can be found how to exactly construct the e-value for the present trial; there is an R-package too: [4], which we will use here. I use the following R code:

```
safe.prop.test(ya=ya, yb=yb, pilot=T)
```

The vector `ya` contains the data stream from group A, and the vector `yb` contains the data stream from group B. I used the argument '`pilot=True`', because I have no idea about the expected effect size, and I have no prior idea about how long I would like to continue the experiment. I may stop for whatever reason, and the e-value I then report, can be interpreted as evidence against the null hypothesis, and comes with a Type I error guarantee. I particular, I can stop whenever the e-value is larger than $1/\alpha = 20$, because that means *enough* evidence against the null hypothesis at significance level $\alpha = 0.05$. After every two nights, I plot the e-value in Figure 1.

The result is overwhelming in a direction I did not expect: the new boosters leaked every single night; the ordinary ones only one out of six nights. After 12 nights, 6 with either booster, the e-value became 21.088, and thus enough to stop the trial, reject the null hypothesis, and conclude that the new boosters were convincingly worse than the ordinary ones.

How to do this with p-values?

In a classical experiment with p-values, I would have had to make an estimate of the *effect size*, and calculate with a *power analysis* what the *sample size* would have to be. Because if the whole *sampling plan* would not be fixed upfront, p-values and their error guarantees break down. But I had not the faintest idea about the effect size, since there was no prior data on the new boosters, I even had no idea about the direction of the effect size (though I would have guessed that the new booster would be *better* than the ordinary ones, not worse) ...

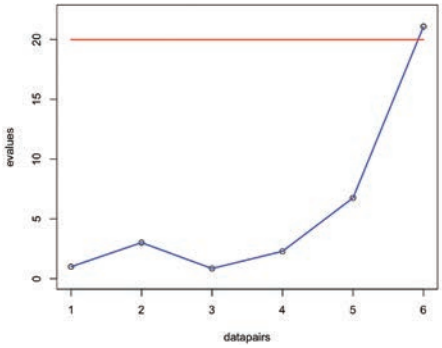

Figure 1 The blue line connects the e-values after each addition of a data pair: one observation with the ordinary boosters and one observation with the new boosters. The red line marks the boundary 20, and we stop the trial as soon as we obtain an e-value that crosses it.

So, I asked some statistician friends for help: how should I design this study?

They sent me long e-mails which I summarize as follows. First I need to do a pilot study to estimate the effect. For this they advised me to obtain 12 measurements from either group. Then I need to discard that data, and calculate the sample size for the *real* trial on the basis of the estimated effect size.

Now suppose that the second trial would take 12 nights too to reach a conclusion, then the whole study would take at least 18 nights with leakage: it seems to me that I would have stopped the experiment early for ethical reasons, and on top of that we would not even be able to report a p-value on the basis of the data collected up until that point, since intermediate stopping invalidates p-values and their error guarantees. Working with e-values thus seems much more natural in this situation.

The e-value

In the Safe Testing paper we lay the foundation for a new theory of hypothesis testing. Our theory is based on e-values. If we want to test hypotheses, we can regard the null hypothesis $\mathcal{H}_0$ and the alternative hypothesis $\mathcal{H}_1$ as sets of probability distributions. An e-value S is simply a non-negative random variable for which it holds that under every distribution $P \in \mathcal{H}_0$, the expectation under P is at most 1: $\mathbb{E}_P[S] \leq 1$. That means that, if the null hypothesis is true, you don't expect the e-value to become (much) larger than 1. For a significance level $\alpha \in (0,1)$, we can define an hypothesis test that rejects $\mathcal{H}_0$ if the e-value is larger than $1/\alpha$. Via Markov's inequality, we obtain a Type I error guarantee $P(\text{reject}\,\mathcal{H}_0) \leq \alpha$, as follows:

$$P(S \geq 1/\alpha) \leq \alpha \mathbb{E}_P[S] \leq \alpha.$$

(Markov's inequality states that if X is a non-negative random variable and $c > 0$, then the probability that X is at least c is at most the expectation of X divided by c: $P(X \geq c) \leq \mathbb{E}[X]/c$.)

Benefits of e-values

Hypothesis testing with e-values brings several advantages over the classical framework with p-values. Here are some important ones:

- e-values behave (they retain error guarantees and remain interpretable) under *optional continuation*: when you decide on basis of the results so far whether or not to add some more data to the experiment, like in the diaper study, or in a meta-analysis.
- They have a simple interpretation as evidence against the null hypothesis, which remains in place if you reject the whole concept of significance [1].
- They are flexible: one can construct e-values on the basis of Bayesian prior information, on basis of preliminary data, but also based on minimax guarantees, and in all cases they retain the Type I error guarantees. You can easily combine e-values stemming from different paradigms (often it is as easy as multiplicating them!): finally the Bayesians and the Frequentists have a common coin!
- For many applications, e-values turn out to have amazing properties. For example, in multiple testing, the e-value analogue of the Benjamini–Hochberg procedure, called e-BH, provides FDR control under arbitrary dependence [5].

The mathematical justification for these, and more, can be found in the Safe Testing paper.

Mathematical contribution

The most important mathematical contribution of the Safe Testing paper is a general way to construct *good* e-values for hypothesis testing problems with a *composite null* (i.e. the set of distributions comprising $\mathcal{H}_0$ contains more than one element). e-values and the related test martingales existed for longer, but until the appearance of the Safe Testing paper on arXiv (and half a year later the Universal inference paper [6]), it was unknown how to construct e-values for problems with a composite null. Most practical problems have such a composite null, think of the t-test and the chi-square test.

Besides that, an important question is: What is a *good* e-value? If you look at the definition, you see that the e-value that is always 1, irrespective of the data, is a valid e-value. Of course we want an e-value that gets large as fast as possible when the null hypothesis is not true. In the Safe Testing paper we formalize this with a criterion called GRO: Growth-Rate Optimal. You can view this as an alternative for power in the sequential setting (here, e-value tests have power 1), the setting where you can add data or stop whenever you like. In the paper we prove that our general way to construct e-values generates e-values that are GRO. We also define a worst-case version (GROW) and a relative version (REGROW). In the paper we furthermore provide an overview of related work on sequential testing (among which Wald's work), which is much more restrictive than our framework.

The future

The future of e-values looks bright if you ask me! There is so much fundamental research to be done, that gets picked up by many great groups all over the world, mainly in mathematical statistics and probability theory. In the mean time we see e-values pop up everywhere: for example in the multiple testing community, or at tech companies where they perform large-scale experimentation with them. Especially good news is the ERC Advanced Grant with which Peter Grünwald will appoint PhD students and postdocs to conduct e-value research the coming years.

This article appeared earlier in Dutch in *STATOR* 2024-2. Rianne de Heide is assistant professor in mathematical statistics and statistical learning theory at the University of Twente, https://riannedeheide.github.io. Currently she is working on her Veni project 'e-values for multiple testing'.

References

1 Valentin Amrhein, Sander Greenland and Blake McShane, Scientists rise up against statistical significance, *Nature* 567(7748) (2019), 305–307.

2 P. Grünwald, R. de Heide and W.M. Koolen, Safe Testing, *Journal of the Royal Statistical Society, Series B: Statistical Methodology* (2024), forthcoming.

3 Rosanne Turner, Alexander Ly and Peter Grünwald, Peter Generic e-variables for exact sequential k-sample tests that allow for optional stopping (2022), arXiv:2106.02693.

4 Rosanne Turner, Alexander Ly, Muriel Felipe Pérez-Ortiz, Judith ter Schure and Peter Grünwald, R-package safestats, *R package version 0.8.7*, 2022.

5 Ruodu Wang and Aaditya Ramdas, False discovery rate control with e-values, *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 84(3) (2022), 822–852.

6 Larry Wasserman, Aaditya Ramdas and Sivaraman Balakrishnan, Universal inference, *Proceedings of the National Academy of Sciences* 117(29) (2020), 16880–16890.