Journal of Statistical Planning and Inference 230 (2024) 106116

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Generic E-variables for exact sequential k-sample tests that allow for optional stopping

Rosanne J. Turner [a,b,*], Alexander Ly [a,c], Peter D. Grünwald [a,d]

[a] CWI, part of NWO-I, Amsterdam, The Netherlands
[b] University Medical Center Utrecht, Brain Center, Utrecht, The Netherlands
[c] University of Amsterdam, Department of Psychology, Amsterdam, The Netherlands
[d] Leiden University, Department of Mathematics, Leiden, The Netherlands

ARTICLE INFO

Keywords:
E-variables
Hypothesis testing
Sequential test
Type-I error control
Composite hypothesis
Test martingale

ABSTRACT

We develop E-variables for testing whether two or more data streams come from the same source or not, and more generally, whether the difference between the sources is larger than some minimal effect size. These E-variables lead to exact, nonasymptotic tests that remain safe, i.e., keep their type-I error guarantees, under flexible sampling scenarios such as optional stopping and continuation. In special cases our E-variables also have an optimal 'growth' property under the alternative. While the construction is generic, we illustrate it through the special case of $k \times 2$ contingency tables, i.e. k Bernoulli streams, allowing for the incorporation of different restrictions on the composite alternative. Comparison to p-value analysis in simulations and a real-world 2×2 contingency table example show that E-variables, through their flexibility, often allow for early stopping of data collection — thereby retaining similar power as classical methods — while also retaining the option of extending or combining data afterwards.

1. Introduction

We develop hypothesis tests that remain statistically valid under flexible sampling scenarios, where one is allowed to engage in optional continuation and optional stopping. We focus on the setting with data coming from several groups (often: treatment(s) versus control), with the goal of testing whether the underlying distributions are all the same. We design a family of tests for this scenario based on E-variables and test martingales that preserve type-I error guarantees under optional stopping. Hence, if the level α-test is performed and the null hypothesis holds true, the probability that the null will *ever* be rejected is bounded by α. Our tests can be implemented, and are exact, for composite null and alternative hypotheses, arbitrary distributions and in combination with arbitrary divergence measures. While our E-variable construction works for general parametric models, in the practical part of this paper we restrict ourselves to sequential categorical data, i.e. Bernoulli streams, for which we provide explicit implementation details and test scenarios.

Relevance. Even in this age of big data and huge models, simple tests for comparing two populations are still used as heavily as ever in clinical trials, psychological studies and so on — areas heavily plagued by the *reproducibility crisis* (Pace and Salvan, 2020). In a by-now notorious questionnaire (John et al., 2012), more than 55% of the interviewed psychologists admitted to the practice of 'adding data until the results look good'. While classical methods lose their type-I error guarantee if one does this

* Correspondence to: National Research Institute for Mathematics and Computer Science in The Netherlands (CWI), Science Park 123, 1098 XG Amsterdam, The Netherlands.
E-mail address: Rosanne.Turner@cwi.nl (R.J. Turner).

https://doi.org/10.1016/j.jspi.2023.106116
Received 22 June 2022; Received in revised form 18 September 2023; Accepted 24 October 2023
Available online 26 October 2023

(an example of this is provided in Appendix S4 of the Supplementary Material), E-variable based tests allow for it, while, due to the option of stopping early, remaining competitive in terms of sample sizes needed to obtain a desired power. We illustrate the practical advantage of our test in Section 7 using the recent real-world example of the SWEPIS trial which was stopped early for harm (Wennerholm et al., 2019). Their analysis being based on a p-value (by definition designed for fixed sampling plan), the question whether there was indeed sufficient evidence available to stop early is very hard to answer, since the sampling plan was not followed, and consequently the p-value based on which they stopped the study was by definition incorrectly calculated. This also makes it very difficult to combine the test results with results from earlier or future data while keeping anything like error control. We show that with our E-variable based methodology we would have obtained sufficient evidence to stop for harm after the same number of events had occurred, because we are allowed to perform an interim analysis each time one pair of treatment and control samples has been collected. Additionally, this E-variable, even though based on a stopped trial, can be effortlessly combined with E-variables from other trials while retaining error guarantees. Also, our results are of interest beyond mere testing: the E-variables we develop in this paper can be used to obtain *anytime-valid confidence intervals* (Howard et al., 2021) that also remain valid under optional stopping (Turner and Grünwald, 2023).

In Sections 4 and 5 we refine our generic test to the 2×2 and $k \times 2$ model. An advantage of focusing on this simple setting is that it is arguably the simplest and clearest example in which there is a nuisance parameter (the proportion under the null) that does not admit a group invariance. Nuisance parameters that satisfy such an invariance (such as the variance in the t-test, or the grand mean in the two-sample t-test) are quite straightforward to turn into E-variables and test martingales via the method of maximal invariants, as explained by Grünwald et al. (2024) and already put into practice by e.g. Robbins (1970) and Lai (1976). The present paper shows that the proportion under the null can also be handled in a clean and simple manner. As explained below, the resulting instantiated 2×2 test appears to be quite different from existing sequential and Bayesian approaches. Thus, more than 85 years after *the lady tasting tea*, we are able to still say something quite new about the age-old problem of contingency table testing.

Related work. A sequential test for the 2×2 setting has been suggested as early as 1947 by Wald (1947). Wald's test statistic can be viewed as a product of E-variables and hence his test can be modified so as to remain valid under optional stopping. Yet, as explained in Section 8.2, in the 2×2 setting, Wald's E-variables lack the optimality property of the ones we introduce here, and they cannot be generalized to arbitrary models or effect size notions. Other earlier approaches (e.g. Siegmund, 2013, Section V.2 and Johari et al., 2022) are based on asymptotic approximations, or consider a somewhat different problem in which the null is simple (Lindon and Malek, 2022) (and then standard likelihood ratio tests Royall, 1997 can be used). In contrast, our E-variable based tests are exact and nonasymptotic, meaning they are valid in (even the smallest) finite samples, and hold for general composite null and alternative hypotheses. E-variables also offer a lot more flexibility than traditional α-spending and group sequential methods: although these methods allow for interim looks at the data, most often at pre-specified moments, a maximum sample size still needs to be set in advance, which does not truly allow for optional stopping and optional continuation (a more elaborate comparison of the two methods can be found in Ter Schure et al., 2020, Section 1).

In fact our tests are more closely related to, yet still different from, Bayes factor tests: in the case of simple null hypotheses, E-variable based tests coincide with Bayes factors (Grünwald et al., 2024). However, in the 2×2 setting the null is not simple, and while the Bayes factor is a ratio of two Bayes marginal likelihoods, our E-variables are ratios of more general, 'prequential' (Dawid, 1984) likelihood ratios. In some special cases, the numerator is still a Bayes marginal likelihood, but the denominator, in the 2×2 setting, almost never is (Section 3.2). Thus, while similar in 'look', our approach is in the end quite different from the default Bayes factors for tests of two proportions that were proposed by Kass and Vaidyanathan (1992) and by Jamil et al. (2017), the latter based on early work by Gunel and Dickey (1974). To illustrate, in Appendix S3 (Supplementary Material) we show that none of the variants of the Gunel–Dickey Bayes factor that are applicable in our set-up yield valid E-variables (are anytime-valid).

Another recent approach that bears some similarity to ours are the two-sample tests from Manole and Ramdas (2023) and Shekhar and Ramdas (2021). They focus on a nonparametric setting and their test martingales satisfy optimality properties as the sample size gets large. Instead, we focus on the parametric case and, for this case, manage to derive E-variables that are equal to or closely approximate to "optimal" (see Section 2.2) E-variables, thus optimizing for the small-sample case (in principle, our tests could be used in a nonparametric setting as well, but since they rely on using a prior on the alternative, the test martingales of Manole and Ramdas (2023) and Shekhar and Ramdas (2021) might be easier to use in that case). Another general nonparametric two-sample approach with a sequential flavor, but without optional stopping error guarantees, is Lhéritier and Cazals (2018).

Contents. In Section 2 we formally introduce the notation used throughout this paper and restate the concepts of E-variables, optional stopping and the Growth Rate Optimality (GRO) criterion, GRO being the analogue of 'optimal power' in our optional continuation setting. In Section 3 we propose our generic E-variable for tests of two streams in general and investigate when it has the GRO property. In Sections 4 and 5 we specifically show how these general E-variables can be applied in the setting of a test of two proportions, with and without restrictions on the alternative hypothesis. In Sections 6 and 7 we provide, through simulations and a real-world example, comparisons of various E-variables and Fisher's exact test with respect to GRO and power. In Section 8 we compare our generic approach to other E-variables one might define for this problem, including the ones based on Wald's test. We end with a conclusion. All proofs are in Appendix S1 in the Supplementary Material.

2. Setup, notation and preliminaries

In this section we describe our setup and notation in detail, and cover the necessary preliminaries from the theory of safe anytime-valid inference with E-variables. We refer to Ramdas et al. (2022), Grünwald et al. (2024) and Shafer et al. (2021), respectively, for an extensive introduction to this theory, to the use of E-variables in 'optional continuation' over several studies in particular, and to their enlightening betting interpretation.

2.1. Setup

Suppose we collect samples from two distinct groups, denoted a and b. In both groups, data are i.i.d. and come in sequentially — even though, as explained underneath (2.2) below, our approach can also be fruitfully used in the fixed design case. 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. $\sim P_{\theta_b}$ with $\theta_a, \theta_b \in \Theta$, $\{P_\theta : \theta \in \Theta\}$ representing some parameterized underlying family of distributions, all assumed to have a probability density or mass function denoted by p_θ on some outcome space $\mathcal{Y}$. We will use notation $P_{(\theta_a,\theta_b)}$ (density $p_{(\theta_a,\theta_b)}$) to represent the joint distribution of both streams. Since it considerably simplifies notation and treatment, we focus on two-sample tests throughout the paper, pointing out at the relevant places how to extend our results to the k-sample setting for $k > 2$. We further assume that all streams are mutually fully independent, so that (returning to $k = 2$), the (marginal) probability of the first $t = t_a + t_b$ outcomes, given that t_a of these are in group a and t_b in group b, and writing $y^t = (y_1, \ldots, y_t)$, is given by the probability density (or mass function)

$$p_{\theta_a,\theta_b}(y_a^{t_a}, y_b^{t_b}) := p_{\theta_a}(y_a^{t_a}) p_{\theta_b}(y_b^{t_b}) = \prod_{t=1}^{t_a} p_{\theta_a}(y_{t,a}) \prod_{t=1}^{t_b} p_{\theta_b}(y_{t,b}). \tag{2.1}$$

To indicate that random vector $(Y_a^{t_a}, Y_b^{t_b}) := (Y_{1,a} \ldots, Y_{t_a,a}, Y_{1,b}, \ldots, Y_{t_b,b})$ has a distribution represented by (2.1) we write '$Y_a^{t_a}, Y_b^{t_b} \sim P_{\theta_a,\theta_b}$'. According to the *null hypothesis* $\mathcal{H}_0 = \{P_{\theta_a,\theta_b} : (\theta_a, \theta_b) \in \Theta_0\}$, $\Theta_0 = \{(\theta, \theta) : \theta \in \Theta\}$, both processes coincide. Thus, we have that $\theta_a = \theta_b = \theta_0$ for some $\theta_0 \in \Theta$ and then the density of data $y_a^{t_a}, y_b^{t_b}$ is given by $p_{\theta_0}(y_{1,a}, \ldots, y_{t_a,a}, y_{1,b}, \ldots, y_{t_b,b})$. The alternative $\mathcal{H}_1$ expresses that $d(\theta_a, \theta_b) > \delta$ for some divergence measure d and some effect size $\delta \geq 0$.

To enable sequential application of our E-variables, we define a block $Y_{(j)}$ as a set of data consisting of n_a outcomes in group a and n_b outcomes in group b, for some pre-specified n_a and n_b. The n_a and n_b used for the jth block $Y_{(j)}$ are allowed to depend on past data, but they must be fixed before the first observation in block j occurs (this rule can be loosened to some extent, see Section 3.1 and Appendix S5). A classical paired one-sample test corresponds to the special case with $n_a = n_b = 1$ and data coming in the order $a, b, a, b, \ldots$.

2.2. E-variables and test martingales

While to some extent going back as far as Darling and Robbins (1967), interest in E-variables has exploded only very recently (Howard et al., 2021; Ramdas et al., 2020; Vovk and Wang, 2021; Shafer et al., 2021; Grünwald et al., 2024; Pace and Salvan, 2020; Manole and Ramdas, 2023; Henzi and Ziegel, 2022). In its simplest form, an E-*variable* is a nonnegative random variable S such that under all distributions P in the null hypothesis,

$$\mathbf{E}_P[S] \leq 1. \tag{2.2}$$

We use the term E-value for the realized value of S, analogously to its classical counterpart, the p-value. Our test works by first designing E-variables for a *single block* of data, and then later extending these to sequences of blocks $Y_{(1)}, Y_{(2)}, \ldots$ by multiplication. At each point in time, the running product of block E-values observed so far is itself an E-variable, and the random process of the products is known as a *test martingale*:

Definition 1. Let $\{Y_{(j)}\}_{j \in \mathbf{N}}$, with all $Y_{(j)}$ taking values in some set $\mathcal{Y}$, represent a discrete-time random process. Let $\mathcal{H}_0$ be a collection of distributions for the process $\{Y_{(j)}\}_{j \in \mathbf{N}}$. For all $j \in \mathbf{N}$, let $S_{(j)}$ be a non-negative random variable that is adapted to $\sigma(Y^{(j)})$, with $Y^{(j)} = (Y_{(1)}, \ldots, Y_{(j)})$, i.e. there exists a function s such that $S_{(j)} = s(Y^{(j)})$.

1. We say that $S_{(j)}$ is an E-variable for $Y_{(j)}$ conditionally on $Y^{(j-1)}$ if for all $P \in \mathcal{H}_0$,

$$\mathbf{E}_P\left[S_{(j)} \mid Y_{(1)}, \ldots, Y_{(j-1)}\right] \leq 1. \tag{2.3}$$

   That is, for each $y^{(j-1)} \in \mathcal{Y}^{j-1}$, all $P_0 \in \mathcal{H}_0$, (2.2) holds with $S = s(y_{(1)}, \ldots, y_{(j-1)}, Y_{(j)})$ and P set to $P_0 \mid Y^{(j-1)} = y^{(j-1)}$.
2. If, for each j, $S_{(j)}$ is an E-variable conditional on $Y_{(1)}, \ldots, Y_{(j-1)}$, then we call the process $\{S_{(j)}\}_{j \in \mathbf{N}}$ a *sequential* E-*variable process* relative to the given $\mathcal{H}_0$ and $\{Y_{(j)}\}_{j \in \mathbf{N}}$ and we call $\{S^{(m)}\}_{m \in \mathbf{N}}$ with $S^{(m)} = \prod_{j=1}^{m} S_{(j)}$ the corresponding *test martingale*.

Henceforth, we omit the phrase 'relative to $\mathcal{H}_0$ and $\{Y_{(j)}\}_{j \in \mathbf{N}}$' whenever it is clear from the context. By the tower property of conditional expectation, one verifies that for any process of conditional E-variables $\{S_{(j)}\}_{j \in \mathbf{N}}$, we have for all m that the product $S^{(m)}$ is itself an 'unconditional' E-variable as in (2.2), i.e. $\mathbf{E}_P[S^{(m)}] \leq 1$ for all $P \in \mathcal{H}_0$. Definition 1 adapts and slightly modifies terminology from Ramdas et al. (2022) and Shafer et al. (2011).

Safety. The interest in E-variables and test martingales derives from the fact that we have type-I error control irrespective of the stopping rule used: for any test martingale $\{S^{(j)}\}_{j \in \mathbf{N}}$, Ville's inequality (Shafer et al., 2021) tells us that, for all $0 < \alpha \leq 1$, $P \in \mathcal{H}_0$,

$$P(\text{there exists } j \text{ such that } S^{(j)} \geq 1/\alpha) \leq \alpha. \tag{2.4}$$

Thus, if we measure evidence against the null hypothesis after observing j data units by $S^{(j)}$, and we reject the null hypothesis if $S^{(j)} \geq 1/\alpha$, then our type-I error will be bounded by α, no matter what stopping rule we used for determining j. We thus have type-I error control even if we use the most aggressive stopping rule compatible with this scenario, where we stop at the first j at which $S^{(j)} \geq 1/\alpha$ (or we run out of data, or money to generate new data). We also have type-I error control if the actual stopping rule

is unknown to us, or determined by external factors independent of the data $Y_{(j)}$. We will call any test based on $\{S^{(j)}\}_{j\in\mathbf{N}}$ and a (potentially unknown) stopping time τ that, after stopping, rejects iff $S^{(\tau)} \geq 1/\alpha$ a *level α-test that is safe under optional stopping*, or simply a *safe test*.

GRO-optimality, simple $\mathcal{H}_1$. Grünwald et al. (2024) (in the first version of their paper put on arXiv in 2019) introduced a definition of E-variable optimality that has by now become standard. To explain it, first consider a simple $\mathcal{H}_1 = \{Q\}$ and consider

$$\mathbf{E}_Q[\log S_{(j)}] \quad ; \quad \mathbf{E}_Q[\log S^{(m)}] \tag{2.5}$$

where $S_{(j)}$ and $S^{(m)}$ are E–variables (i.e. non-negative random variables satisfying (2.2)) that, respectively, can be written as a function of $Y_{(j)}$ and $Y^{(m)} = (Y_{(1)}, \ldots, Y_{(m)})$. The E-variable which maximizes the quantity on the left among all E-variables that can be written as a function of $Y_{(j)}$, assuming it exists, is called the *Growth Rate Optimal* E-variable for $Y_{(j)}$ relative to Q, or simply 'Q-GRO for $Y_{(j)}$', and denoted as $S_{\mathrm{GRO}(Q),(j)}$. Similarly, the E-variable maximizing the quantity on the right, among all E-variables that can be written as function of $Y^{(m)}$, is called Q-GRO for $Y^{(m)}$. Grünwald et al. (2024), Shafer et al. (2021) and Ramdas et al. (2022) explain why the logarithm is the appropriate function to use here.

In 'nice' cases, the Q-GRO E-variable for m outcomes can be obtained by multiplying the individual Q-GRO E-variables:

Proposition 1. *Let $\mathcal{H}_1 = \{Q\}$ be simple and $\mathcal{H}_0$ be potentially composite, and 'nondegenerate' in the sense that for some $P \in \mathcal{H}_0$, $D(Q \parallel P) < \infty$, $D(\cdot \parallel \cdot)$ denoting the KL divergence. We define the following condition, with q, p the density of Q and P, respectively:*

$$\textit{There exists a } P \in \mathcal{H}_0 \textit{ such that } S_{(1)} = q(Y_{(1)})/p(Y_{(1)}) \textit{ is an E-variable.} \tag{2.6}$$

When this condition holds, $S_{(1)} = S_{\mathrm{GRO}(Q),(1)}$ is the Q-GRO E-variable for $Y_{(1)}$. An E-variable of this form automatically exists if $\mathcal{H}_0$ is simple. If we further assume that $Y_{(1)}, Y_{(2)}, \ldots$ are i.i.d. according to all distributions in $\mathcal{H}_0 \cup \mathcal{H}_1$, then $S^{(m)}_{\mathrm{GRO}(Q)} = \prod_{j=1}^m S_{\mathrm{GRO}(Q),(j)}$.

If Condition (2.6) holds and $Y_{(1)}, Y_{(2)}, \ldots$ are i.i.d. according to all distributions in $\mathcal{H}_0 \cup \mathcal{H}_1$, it thus makes sense to define the *Q-GRO test martingale* to be the test martingale $(S^{(j)}_{\mathrm{GRO}(Q)})_{j\in\mathbf{N}}$. We will then have that $S_{\mathrm{GRO}(Q),(j)} = s_Q(Y_{(j)})$ for a fixed function $s_Q : \mathcal{Y} \to \mathbf{R}_0^+$.

In Section 3 (Theorem 1) we develop functions s_Q (denoted $s(\cdot; n_a, n_b, \theta_a^*, \theta_b^*)$ there) for simple $\mathcal{H}_1 = \{Q\}$ so that $S_{Q,(1)} = s_Q(Y_{(1)})$ is an E–variable even though (2.6) does not necessarily hold, so that Proposition 1 does not apply. Since we invariably assume the $Y_{(j)}$ are i.i.d., $S_{Q,(j)} := s_Q(Y_{(j)})$ is an E–variable as well and with $S_Q^{(m)} := \prod_{j=1}^m S_{Q,(j)}$, $(S_Q^{(m)})_{m\in\mathbf{N}}$ is a test martingale. The construction works for the general setting of two data streams discussed in the introduction, and for some special $\mathcal{H}_0$ (even though composite), (2.6) does hold, and then the $S_{Q,(j)}$ will in fact be Q-GRO and $(S_Q^{(m)})_{m\in\mathbf{N}}$ will be the Q-GRO test martingale. These include the $\mathcal{H}_0$ that arise in the 2×2 setting, our main application. For other $\mathcal{H}_0$, the E-variables $S_{Q,(j)}$ will not necessarily have the Q-GRO-property; they are designed to have (2.5) large, but it may be even larger for other E-variables.

2.3. From simple to composite setting: choice of the E-variable and optimality

In case $\mathcal{H}_1$ is composite, no direct analogue of the GRO-criterion for designing E-variables exists, since it is not clear under what distribution $Q \in \mathcal{H}_1$ we should maximize (2.5). In this paper, we deal with this situation by *learning* Q from the data in a Bayesian fashion. It is now convenient to write $\mathcal{H}_1 = \{P_\theta : \theta \in \Theta_1\}$ in a parameterized manner (accordingly, henceforth we shall write θ_1-GRO E-variable instead of P_{θ_1}-GRO E–variable and $S_{\mathrm{GRO}(\theta),(j)}$ instead of $S_{\mathrm{GRO}(P_\theta),(j)}$). We will assume i.i.d. data, thus, if $\mathcal{H}_1$ were true, then data would be i.i.d. $\sim P_{\theta_1^*}$ for some $\theta_1^* \in \Theta_1$. Starting with a distribution W on Θ_1, i.e. a prior, at each point in time j, we determine the Bayesian posterior $W \mid Y^{(j-1)}$ and use the Bayes predictive $P_{W|Y^{(j-1)}} := \int_{\Theta_1} P_\theta dW(\theta \mid Y^{(j-1)})$ as an estimate for the 'true' P_{θ^*}. As is well-known, under conditions on W and $\mathcal{H}_1$ (which, if $\mathcal{H}_1$ is finite-dimensional parametric, are very mild), the posterior will concentrate around θ^* and hence $P_{W|Y^{(j-1)}}$ will resemble $P_{\theta_1^*}$ more and more, with very high probability, as more data becomes available.

At each point in time j, we use our current estimate $P_{W|Y^{(j-1)}}$ to design a conditional E-variable $S_{(j)}$. Note that even though our test depends on the choice of a prior distribution on the alternative, the choice of prior does not affect the type-I error safety guarantee, hence it is fine, even from a frequentist point of view, if such a prior is chosen based on vague prior knowledge. On an informal level, as long as $P_{W|Y^{(j-1)}}$ converges to the 'true' $P_{\theta_1^*}$, the $S_{(j)}$ will in fact also start to more and more resemble the E–variables $S_{\mathrm{GRO}(\theta_1^*),(j)}$ we designed for $\mathcal{H}_1 = \{P_{\theta_1^*}\}$ and which were designed to have a large expected growth under the 'true' $P_{\theta_1^*}$. If we had known the true $P_{\theta_1^*}$ all along, the best test martingale we could have used is $S^{(m)}_{\mathrm{GRO}(\theta_1^*)} = \prod_{j=1}^m S_{\mathrm{GRO}(\theta_1^*),(j)}$, which maximizes $\mathbf{E}_{Y^{(m)}\sim P_{\theta_1^*}}[\log S]$ over all E-variables S for $Y^{(m)}$. Assuming the convergence happens fast, we expect the following quantity to be small:

$$\mathbf{E}_{Y^{(m)}\sim P_{\theta_1^*}}\left[\log S^{(m)}_{\mathrm{GRO}(\theta_1^*)} - \log \prod_{j=1}^m S_{(j)}\right], \tag{2.7}$$

i.e., we may expect that the test martingale $\prod_{j=1}^m S_{(j)}$ grows not much slower than $S^{(m)}_{\mathrm{GRO}(\theta_1^*)}$. We note that (2.7) is an instance of what is called *regret* in the statistical and machine learning theory literature, measuring how much worse our $S_{(j)}$ performs compared to the e-variable that is optimal given additional knowledge, namely θ_1^*; Grünwald et al. (2024) explore this connection further.

3. Two-stream safe tests

3.1. A generic E-variable for 2-stream–blocks

We first consider the case in which the alternative hypothesis is simple: $\Theta_1 = \{\theta_1\}$ for some fixed $\theta_1 = (\theta_a^*, \theta_b^*) \in \Theta^2$. Consider a fixed sample size of size n, and assume that we will observe a block of n_a outcomes in group a and n_b outcomes in group b. In this case, we can define an E-variable as the likelihood ratio between $p_{\theta_a^*,\theta_b^*}$ and a carefully chosen distribution that is a product of mixtures of distributions from Θ_0: for $n_a, n_b \in \mathbf{N}$, $n := n_a + n_b$ and $y_a^{n_a} = (y_{1,a}, \ldots, y_{n_a,a}) \in \mathcal{Y}^{n_a}$ and $y_b^{n_b} = (y_{1,b}, \ldots, y_{n_b,b}) \in \mathcal{Y}^{n_b}$, we define:

$$s(y_a^{n_a}, y_b^{n_b}; n_a, n_b, \theta_a^*, \theta_b^*) := \frac{p_{\theta_a^*}(y_a^{n_a})}{\prod_{i=1}^{n_a} \left(\frac{n_a}{n} p_{\theta_a^*}(y_{i,a}) + \frac{n_b}{n} p_{\theta_b^*}(y_{i,a}) \right)} \cdot \frac{p_{\theta_b^*}(y_b^{n_b})}{\prod_{i=1}^{n_b} \left(\frac{n_a}{n} p_{\theta_a^*}(y_{i,b}) + \frac{n_b}{n} p_{\theta_b^*}(y_{i,b}) \right)}. \tag{3.1}$$

Theorem 1. *The random variable* $S_{[n_a,n_b,\theta_a^*,\theta_b^*]} := s(Y_a^{n_a}, Y_b^{n_b}; n_a, n_b, \theta_a^*, \theta_b^*)$ *is an* E*-variable, i.e. we have:*

$$\sup_{\theta \in \Theta} \mathbf{E}_{V^n \sim P_\theta} \left[s(V^n; n_a, n_b, \theta_a^*, \theta_b^*) \right] \leq 1.$$

Moreover, if $\{P_\theta : \theta \in \Theta\}$ *is a convex set of distributions, then* $S_{[n_a,n_b,\theta_a^*,\theta_b^*]}$ *is the* (θ_a^*, θ_b^*)*-GRO* E*-variable: for any non-negative function* s' *on* $\mathcal{Y}^{n_a+n_b}$ *satisfying* $\sup_{\theta \in \Theta} \mathbf{E}_{V^n \sim P_\theta} \left[s'(V^n) \right] \leq 1$*, we have:*

$$\mathbf{E}_{Y_a^{n_a}, Y_b^{n_b} \sim P_{\theta_a^*,\theta_b^*}} [\log s(Y_a^{n_a}, Y_b^{n_b}; n_a, n_b, \theta_a^*, \theta_b^*)] \geq \mathbf{E}_{Y_a^{n_a}, Y_b^{n_b} \sim P_{\theta_a^*,\theta_b^*}} [\log s'(Y_a^{n_a}, Y_b^{n_b})].$$

Crucially, in the second part of the theorem, we do not require convexity of $\mathcal{H}_0$, a set of distributions over $\mathcal{Y}^{n_a+n_b}$ (if $\mathcal{H}_0$ were convex, the GRO property would already follow automatically Koolen and Grünwald, 2022), but instead of $\{P_\theta : \theta \in \Theta\}$, a set of distributions on $\mathcal{Y}$. In the 2×2 case $\mathcal{H}_0$ is not convex, since the set of i.i.d. Bernoulli distributions over $n_a + n_b > 1$ outcomes is not convex. Nevertheless, $\{P_\theta : \theta \in \Theta\}$ is just the Bernoulli model on one outcome, which is convex, so in this setting, we get the GRO E-variable.

To illustrate, consider the basic case in which data comes in fixed batches $Y_{(1)}, Y_{(2)}, \ldots$, with each batch $Y_{(j)} = ((Y_{(j-1)n_a+1,a}, Y_{(j-1)n_a+2,a}, \ldots, Y_{jn_a,a}), (Y_{(j-1)n_b+1,b}, Y_{(j-1)n_b+2,b}, \ldots, Y_{jn_b,b}))$, having exactly n_a outcomes in group a and n_b outcomes in group b, and let $n = n_a + n_b$. This case would obtain, for example, in a sequential clinical trial in which patients come in one by one, each odd patient is given the treatment and each even patient is given the placebo. Then $n = 2$, $n_a = n_b = 1$. We may then measure the evidence against the null hypothesis by the product E variable

$$S^{(m)}_{[n_a,n_b,\theta_a^*,\theta_b^*]} := \prod_{j=1}^{m} S_{(j),[n_a,n_b,\theta_a^*,\theta_b^*]} \quad ; \quad S_{(j),[n_a,n_b,\theta_a^*,\theta_b^*]} := s(Y_{(j)}; n_a, n_b, \theta_a^*, \theta_b^*). \tag{3.2}$$

By Ville's inequality (2.4), the probability under any distribution in the null that there is an m with $S^{(m)}_{[n_a,n_b,\theta_a^*,\theta_b^*]}$ larger than $1/\alpha$, is bounded by α, hence, type-I error guarantees are preserved under optional stopping if we perform the test based on $\{S^{(m)}_{[n_a,n_b,\theta_a^*,\theta_b^*]}\}_{m \in \mathbf{N}}$ as defined underneath (2.4), as long as we stop between and not 'within' batches (if we stop within a batch, the E-variable $S^{(m)}_{[n_a,n_b,\theta_a^*,\theta_b^*]}$ is undefined).

If the data do not come in batches of equal size, we may proceed as follows. First, we need to fix some $n_a \geq 1$ and $n_b \geq 1$ of our own choice. The treatment below will give valid E-variables irrespective of our choice of n_a and n_b, but it will be seen that some choices are much more reasonable (will lead to much more evidence against the null, if the null is false) than others.

Thus, fix n_a and n_b, set $n = n_a + n_b$. At each time t, we will have observed, so far, some number t_a of outcomes in group a, and t_b in group b. Now let m_t be the largest m such that $mn_a \leq t_a$ and $mn_b \leq t_b$. Now, for $m = 1, 2, \ldots$, define $Y_{(m)}$ as above. At any given time t, $Y_{(1)}, Y_{(2)}, \ldots, Y_{(m_t)}$ will have been observed, and there may be a number n'_j remaining observations in group $j \in \{a, b\}$ so that either $n'_a < n_a$ or $n'_b < n_b$ or both. Since the $\{Y_{(j)}\}_{j \in \mathbf{N}}$ determine a test martingale in the sense of Definition 1, optional stopping while preserving type-I error guarantees is then possible at any point in time t, as long as the E-variable is calculated as (3.2) above for $m = m_t$, thus ignoring the final $n'_a + n'_b$ outcomes.

How should n_a and n_b be chosen in practice? For example, consider a variation of the clinical trial setting above in which the treatment-control assignment is randomized: for each incoming patient, a fair coin is flipped to decide treatment (a) or placebo (b). Then at any given time the number of patients in group a and b will not be precisely equal, but if we choose $n_a = n_b = 1$ as above it is highly unlikely that the amount of data we have to ignore at any given time t is very large. Similarly, if G_t, the group membership of the tth observation, is itself i.i.d. according to some distribution P^*, we might have some idea of the probability $p^*(a)$ assigned to group a; if $p^*(a) = 2/5$ (say), we would choose $n_a = 2, n_b = 3$.

We can add a significant amount of extra flexibility by allowing for variable group sizes, i.e., the chosen n_a and n_b may depend on the past. Appendix S5 in the supplementary material describes how to do this. In this way, one can in principle *learn* $p^*(a)$ from the data, changing group sizes n_a and n_b flexibly as data come in. For simplicity, we have not followed this approach here, but all our results readily extend to this case.

Extension to k-sample streams. It is entirely straightforward to extend (3.1) to the scenario where we do not compare 2, but k i.i.d. data streams. Indeed, in the supplementary material we state and prove the generalization of Theorem 1 to k data streams. We again consider some fixed $\vec{\theta} = (\theta_a, \theta_b, \ldots, \theta_k) \in \Theta^k$. The probability of the first $t = \sum_{g=1}^k t_g$ outcomes is now given by the density or mass function $p_{\vec{\theta}} := p_{\theta_a}(y_a^{t_a}) p_{\theta_a}(y_b^{t_b}) \ldots p_{\theta_k}(y_k^{t_k})$. We now need to fix the k group outcome numbers $\vec{n} := (n_a, n_b, \ldots, n_k)$ in advance, which allows us to define the extended E-variable as a function of the data $\vec{y}^n = (y_a^{n_a}, y_b^{n_b}, \ldots, y_k^{n_k})$, with $n = \sum_{g=1}^k n_g$ for testing the null where $\theta_a = \theta_b = \cdots = \theta_k$:

$$s(\vec{y}^n; \vec{n}, \vec{\theta}^*) := \prod_{g=1}^k \frac{p_{\theta_g^*}(y_g^{n_g})}{\prod_{i=1}^{n_g} \left(\sum_{g'=1}^k \frac{n_{g'}}{n} p_{\theta_{g'}^*}(y_{i,g}) \right)}. \tag{3.3}$$

This E-variable is again GRO if $\{P_\theta : \theta \in \Theta\}$ is convex. To keep notation as clear as possible, we now return to the simpler 2-sample case except for a short example of an application of this extension as a flexible and exact (non-asymptotic) alternative to the chi-square test in Section 6.

3.2. The generic E-variable with Bayesian alternative

Now fix some prior W_1 with density w_1 on the alternative $\Theta_1 \subseteq \Theta^2$. We can trivially extend the definition of our generic E–variable relative to singleton (θ_a^*, θ_b^*) to an E–variable relative to arbitrary prior W_1 on (θ_a^*, θ_b^*): define $p_{W_1,a}(y) := \int p_{\theta_a}(y) dW_1(\theta_a)$, the integration being over the marginal prior distribution over θ_a, and similarly, $p_{W_1,b}(y) := \int p_{\theta_b}(y) dW_1(\theta_b)$. Then, as a corollary of Theorem 1, the following is also an E–variable:

$$s(y_a^{n_a}, y_b^{n_b}; n_a, n_b, W_1) := \frac{\prod_{i=1}^{n_a} p_{W_1,a}(y_{i,a})}{\prod_{i=1}^{n_a} \left(\frac{n_a}{n} p_{W_1,a}(y_{i,a}) + \frac{n_b}{n} p_{W_1,b}(y_{i,a}) \right)} \cdot \frac{\prod_{i=1}^{n_b} p_{W_1,b}(y_{i,b})}{\prod_{i=1}^{n_b} \left(\frac{n_a}{n} p_{W_1,a}(y_{i,b}) + \frac{n_b}{n} p_{W_1,b}(y_{i,b}) \right)}. \tag{3.4}$$

This follows from applying Theorem 1 with a 'meta'-set of distributions, which is possible since we made no assumptions at all on the set Θ in Theorem 1: we replace Θ by $\mathcal{W}(\Theta)$, the set of distributions on Θ; we replace the background set of distributions $\{p_\theta : \theta \in \Theta\}$ by the set of distributions $\{p_W : W \in \mathcal{W}(\Theta)\}$; we replace the simple $\mathcal{H}_1 = \{P_{\theta_a^*, \theta_b^*}\}$ by a 'simple' $\mathcal{H}_1' = \{P_{W_a, W_b}\}$ for some distributions W_a and W_b on Θ. Such W_1-based generic E–variables can be used to *learn* the parameters θ_a^*, θ_b^* as more data in both streams come in, and this is how we will use them in a sequential context with optional stopping. Thus, assume again that data comes in batches $Y_{(1)}, Y_{(2)}, \ldots$ with each $Y_{(j)}$ consisting of n_a outcomes in group a and n_b outcomes in group b (generalization to flexible group sizes changing in time and depending on the past as described at the end of Section 3.1 is straightforward). We start with some prior W_1 for the first batch $Y_{(1)}$ but we now use, for the jth batch $Y_{(j)}$, the *Bayesian posterior* $W_1 \mid Y^{(j-1)}$ as prior to define the jth E–variable with:

$$S^{(m)}_{[n_a,n_b,W_1]} := \prod_{j=1}^m S_{(j),[n_a,n_b,W_1]} \quad ; \quad S_{(j),[n_a,n_b,W_1]} := s(Y_{(j)}; n_a, n_b, W_1 | Y^{(j-1)}). \tag{3.5}$$

Again, $\{S_{(j),[n_a,n_b,W_1]}\}_{j \in \mathbf{N}}$ is a sequential E–variable process, so testing based on the corresponding test martingale is safe under optional stopping by (2.4). If data are sampled from some alternative hypothesis (θ_a^*, θ_b^*), then as data accumulates, the posterior W_1 will, with high probability, concentrate narrowly around (θ_a^*, θ_b^*) and so $S_{(j),[n_a,n_b,W_1]}$ will behave more and more similarly to the 'best' (θ_a^*, θ_b^*) E-variable. Still, with the exception of a special case we indicate below, in general we cannot expect it to be the W_1-GRO E-variable. But we are not particularly concerned by this: our experiments in Section 6 indicate that, at least in the 2×2 table setting, it behaves quite well in terms of power, which is often the main practical interest.

Simplification when $\{P_\theta : \theta \in \Theta\}$ is convex and $\mathcal{Y}$ is finite. Denoting $W_{1,g} | Y^{(m)}$ as the marginal posterior for θ_g, for $g \in \{a, b\}$, we can rewrite (3.5) as

$$\begin{aligned} S^{(m)}_{[n_a,n_b,W_1]} &= \prod_{j=1}^m \frac{\prod_{i=1}^{n_a} p_{W_{1,a}|Y^{(j-1)}}(Y_{(j-1)n_a+i,a}) \prod_{i=1}^{n_b} p_{W_{1,b}|Y^{(j-1)}}(Y_{(j-1)n_b+i,b})}{\prod_{g \in \{a,b\}} \prod_{i=1}^{n_g} \left(\frac{n_a}{n} p_{W_{1,a}|Y^{(j-1)}}(Y_{(j-1)n_g+i,g}) + \frac{n_b}{n} p_{W_{1,b}|Y^{(j-1)}}(Y_{(j-1)n_g+i,g}) \right)} \\ &\overset{\text{if } \{P_\theta : \theta \in \Theta\} \text{ convex, } \mathcal{Y} \text{ finite}}{=} \prod_{j=1}^m \prod_{i=1}^{n_a} \frac{p_{W_{1,a}|Y^{(j-1)}}(Y_{(j-1)n_a+i,a})}{p_{\breve{\theta}_0|Y^{(j-1)}}(Y_{(j-1)n_a+i,a})} \prod_{i=1}^{n_b} \frac{p_{W_{1,b}|Y^{(j-1)}}(Y_{(j-1)n_b+i,b})}{p_{\breve{\theta}_0|Y^{(j-1)}}(Y_{(j-1)n_b+i,b})}. \end{aligned} \tag{3.6}$$

Here we define $\breve{\theta}_0 | Y^{(j-1)} \in \Theta$ s.t. $p_{\breve{\theta}_0|Y^{(j-1)}} = (n_a/n) p_{W_{1,a}|Y^{(j-1)}} + (n_b/n) p_{W_{1,b}|Y^{(j-1)}}$, the existence of $\breve{\theta}_0 | Y^{(j-1)}$ being guaranteed if $\{P_\theta : \theta \in \Theta\}$ is convex and the sample space is finite (for then, by Carathéodory's Theorem, Eckhoff, 1993, for any distribution W on Θ there is a distribution W' on Θ with finite support such that $p_W = p_{W'}$, and by convexity, there is θ° such that $p_{W'} = p_{\theta^\circ}$). This rewrite will enable several additional results for such Θ.

Connection to Bayes factors. Consider W_1 such that θ_a and θ_b are independent under W_1 with marginal distributions W_a and W_b, and now further take $n_a = n_b = 1$. By basic telescoping, and using that if θ_a and θ_b are independent under the prior, they must also be independent under the posterior, we can then further rewrite (3.5) as

$$\frac{\int p_{\theta_a}(Y_a^m)dW_a(\theta_a)\int p_{\theta_b}(Y_b^m)dW_b(\theta_b)}{\prod_{j=1}^m \prod_{g\in\{a,b\}}\left(\frac{1}{2}p_{W_{1,a}|Y^{(j-1)}}(Y_{j,g})+\frac{1}{2}p_{W_{1,b}|Y^{(j-1)}}(Y_{j,g})\right)} \overset{\text{if } \{P_\theta : \theta\in\Theta\} \text{ convex}}{=} \tag{3.7}$$

$$\frac{\int p_{\theta_a}(Y_a^m)dW_a(\theta_a)\int p_{\theta_b}(Y_b^m)dW_b(\theta_b)}{\prod_{j=1}^m \prod_{g\in\{a,b\}} p_{\check{\theta}_0|Y^{(j-1)}}(Y_{j,g})}. \tag{3.8}$$

The equality holds if $\{P_\theta : \theta \in \Theta_0\}$ is convex and $\mathcal{Y}$ is finite so that (3.6) holds. As seen from (3.7), even without finiteness or convexity, the numerator of the generic product E-variable is now equal to the Bayesian marginal likelihood of the data based on prior W_1. Thus, in this special case (i.e. $n_a = n_b = 1$, prior independence; the derivation breaks down if these do not hold), if the denominator could also be written as a Bayes marginal likelihood, then our E-variable would really be a Bayes factor. Yet, even if $\{P_\theta : \theta \in \Theta\}$ is convex, it cannot be written in this way, though it is very 'close': each of the m factors in the denominator in (3.8) is the product density function of two identical distributions for one outcome, and Proposition 2 below shows that, in the special case of the 2×2 model with W_a and W_b independent beta priors, this distribution may itself be the Bayes predictive distribution obtained by equipping Θ_0 with another beta prior. Still, for a real Bayes factor corresponding to $\mathcal{H}_0$, for each j, the two outcomes $Y_{j,a}, Y_{j,b}$ in the jth block would not be independent given $Y^{(j-1)}$, whereas in (3.8) they are, so we may conclude that in general, our e-variables are not equivalent to any Bayes factor.

4. Safe tests for two proportions

We assume the setting above and, for now, assume that both streams are Bernoulli. This will substantially simplify the formulae. Thus, $\Theta = [0, 1]$ and (2.1) now specializes to

$$p_{\theta_a,\theta_b}(y_a^{t_a}, y_b^{t_b}) := p_{\theta_a}(y_{1,a},\dots,y_{t_a,a})p_{\theta_b}(y_{1,b},\dots,y_{t_b,b}) = \theta_a^{t_{a1}}(1-\theta_a)^{t_a-t_{a1}}\theta_b^{t_{b1}}(1-\theta_b)^{t_b-t_{b1}}. \tag{4.1}$$

t_{a1} represents the number of outcomes 1 in stream a among the first t_a ones, and t_{b1} the number of outcomes 1 in stream b among the first t_b ones. According to the null hypothesis, we have that $\theta_a^* = \theta_b^* = \theta_0$ for some $\theta_0 \in \Theta = [0, 1]$. (4.1) now simplifies to:

$$p_{\theta_0}(y_a^{t_a}, y_b^{t_b}) := \theta_0^{t_1}(1-\theta_0)^{t_0}.$$

t_1 represents the number of ones in the sequence $y^{t_a+t_b} = y_1,\dots,y_{t_a+t_b}$, and similarly for t_0.

We now run through the results of the previous section for this instantiation of our test. Again, we start with the case of a simple $\mathcal{H}_1 = \{P_{\theta_a^*,\theta_b^*}\}$. (3.1) can now be written as:

$$s(y_a^{n_a}, y_b^{n_b}; n_a, n_b, \theta_a^*, \theta_b^*) := \frac{p_{\theta_a^*}(y_a^{n_a})}{p_{\theta_0}(y_a^{n_a})}\cdot\frac{p_{\theta_b^*}(y_b^{n_b})}{p_{\theta_0}(y_b^{n_b})} \quad ; \quad \theta_0 = \frac{n_a}{n}\theta_a^* + \frac{n_b}{n}\theta_b^*. \tag{4.2}$$

Theorem 1 tells us that this is an E-variable. Since $\{P_\theta : \theta \in \Theta\}$, the Bernoulli model, is convex, the theorem also tells us that in this case the generic E-variable with simple alternative is always (θ_a^*, θ_b^*)-GRO.

We now turn to the generic E–variable relative to arbitrary prior W_1. For the Bernoulli model the Bayes posterior predictive distribution is itself a Bernoulli distribution, with its parameter equal to the posterior mean. Therefore, while the generic E–variable relative to prior W_1 is still given by (3.4), this now simplifies to:

$$s(y_a^{n_a}, y_b^{n_b}; n_a, n_b, W_1) = s(y_a^{n_a}, y_b^{n_b}; n_a, n_b, \theta_a^*, \theta_b^*) \quad ; \quad \theta_g^* = \mathbf{E}_{\theta_g\sim W_1}[\theta_g],\ g \in \{a, b\}. \tag{4.3}$$

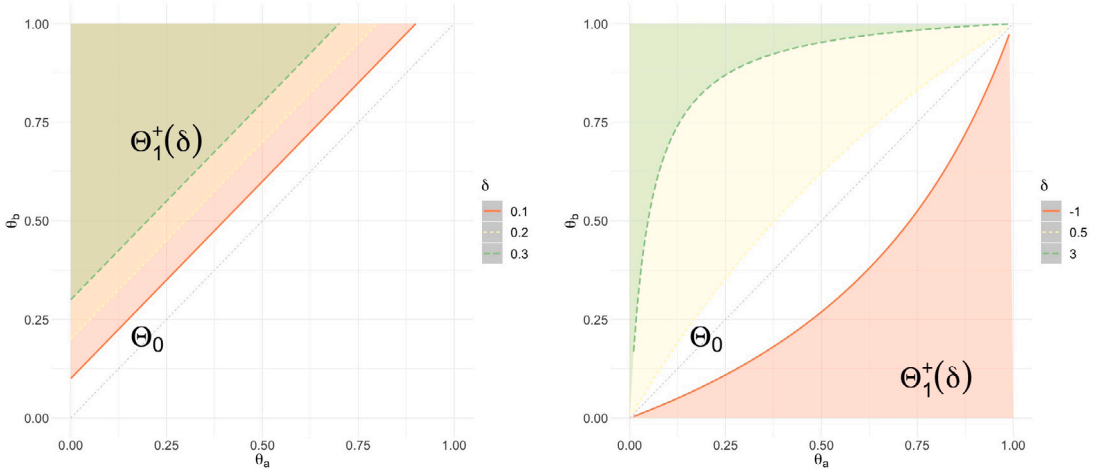
Combining this with (3.6) we infer that

$$S^{(m)}_{[n_a,n_b,W_1]} = \prod_{j=1}^m \prod_{i=1}^{n_a}\frac{p_{\check{\theta}_a|Y^{(j-1)}}(Y_{(j-1)n_a+i,a})}{p_{\check{\theta}_0|Y^{(j-1)}}(Y_{(j-1)n_a+i,a})}\prod_{i=1}^{n_b}\frac{p_{\check{\theta}_b|Y^{(j-1)}}(Y_{(j-1)n_b+i,b})}{p_{\check{\theta}_0|Y^{(j-1)}}(Y_{(j-1)n_b+i,b})} \tag{4.4}$$

where $\check{\theta}_a|Y^{(j-1)} = \mathbf{E}_{\theta_a\sim W|Y^{(j-1)}}[\theta_a]$ and $\check{\theta}_b|Y^{(j-1)} = \mathbf{E}_{\theta_b\sim W|Y^{(j-1)}}[\theta_b]$ and $\check{\theta}_0|Y^{(j-1)} = (n_a/n)\check{\theta}_a \mid Y^{(j-1)} + (n_b/n)\check{\theta}_b \mid Y^{(j-1)}$.

Simplified calculations with independent beta priors. Now take the special case in which θ_a and θ_b are independent under the prior W_1 with marginals W_a and W_b. In this case, θ_a and θ_b are also independent under the posterior, and we can simplify $\check{\theta}_a|Y^{(j-1)} = \mathbf{E}_{\theta_a\sim W_a|Y_a^{(j-1)n_a}}[\theta_a]$, the expectation of θ_a under the posterior W_a given all data so far in group a, and similarly for group b. Using beta priors, this expectation is easy to calculate and we get:

Proposition 2. *Let θ_a, θ_b be independent under W_1, with marginals W_a and W_b respectively. Suppose that these are beta priors with parameters (α_a, β_a) and (α_b, β_b) respectively. Then, upon defining $U_a = \sum_{i=1}^{(j-1)n_a} Y_{i,a}, U_b = \sum_{i=1}^{(j-1)n_b} Y_{i,b}, U = \sum_{i=1}^{(j-1)n}(Y_{i,a} + Y_{i,b})$ we have that $\check{\theta}_a, \check{\theta}_b, \check{\theta}_0$ as above satisfy: $\check{\theta}_a|Y^{(j-1)} = (U_a + \alpha_a)/((j-1)n_a + \alpha_a + \beta_a)$, $\check{\theta}_b|Y^{(j-1)} = (U_b + \alpha_b)/((j-1)n_b + \alpha_b + \beta_b)$ respectively, and $\check{\theta}_0|Y^{(j-1)}$ is as further above. In the special case that we fix the prior parameters in the groups proportional to the group size fraction $\kappa := n_b/n_a$, i.e we fix $\alpha_b = \kappa\alpha_a$, $\beta_b = \kappa\beta_a$, the expression for $\check{\theta}_0$ simplifies to $\check{\theta}_0|Y^{(j-1)} = (U + (1+\kappa)\alpha_a)/((j-1)n + (1+\kappa)\alpha_a + (1+\kappa)\beta_a)$.*



(a) $d((\theta_a, \theta_b)) = \theta_b - \theta_a$ (b) $d((\theta_b, \theta_a)) = \log[\theta_b/(1-\theta_b) \times (1-\theta_a)/\theta_a]$

Fig. 1. Examples of restricted alternative hypothesis parameter spaces for several values of two divergence measures; the difference between group means and the log odds ratio. Θ_0 denotes the null hypothesis parameter space; $\Theta_1^+(\delta)$ the restricted alternative hypothesis parameter space.

5. (Un)restricted composite $\mathcal{H}_1$ in the 2 × 2 setting

In this section we describe the main instantiations of the 2 × 2 stream testing scenario that are relevant in practice. These differ in the choice of $\mathcal{H}_1$: the choice can be fully unrestricted (we simply want to find whether there is any discrepancy from $\mathcal{H}_0$ at all); restricted in terms of effect size; or restricted because we have prior knowledge about either θ_a^* or θ_b^*. We consider each in turn, the second and third scenario in a separate subsection. Section 6 provides extensive numerical simulations for all three scenarios.

In the first scenario, a researcher wants to perform a *two-sided test*; they simply aim to find any discrepancy from $\mathcal{H}_0$ if it exists, with no restrictions are placed on $\mathcal{H}_1$. In this case, if we choose W_1 as independent beta priors on θ_a and θ_b, we can simply proceed as described in Proposition 2 above, taking a beta prior for simplicity. We will develop a reasonable 'default' choice for the hyper parameters by experiment in Section 6.

5.1. Dealing with effect sizes

In the second scenario we really want to test $\mathcal{H}_0$ against a restricted $\mathcal{H}_1$ consisting of those hypotheses that have a certain minimal *effect size* δ. This would then be a one-sided test. For example, a researcher might know that a new treatment must cure at least a certain number of patients more compared to a control treatment to provide *a clinically relevant treatment effect* δ. In this case, $\mathcal{H}_1$ could be restricted to either of the sets $\Theta(\delta)$ or $\Theta^+(\delta)$, where

$$\Theta(\delta) = \left\{\theta \in [0,1]^2 : d(\theta) = \delta\right\} \quad ; \quad \Theta^+(\delta) = \begin{cases} \left\{\theta \in [0,1]^2 : d(\theta) \geq \delta\right\} & \text{if } \delta > 0 \\ \left\{\theta \in [0,1]^2 : d(\theta) \leq \delta\right\} & \text{if } \delta < 0, \end{cases} \tag{5.1}$$

where we set $d((\theta_a, \theta_b)) = \theta_b - \theta_a$. A second notion of effect size that often will be applicable in this sort of research is the *log odds ratio* between θ_b and θ_a, with restricted parameter space again given by (5.1) but d set to

$$d((\theta_a, \theta_b)) = \log\left(\frac{\theta_b}{1-\theta_b} \cdot \frac{1-\theta_a}{\theta_a}\right). \tag{5.2}$$

These are the two effect size notions that will feature in our experiments. An illustration of both divergence measures and the resulting restricted parameter spaces is given in Fig. 1. A third popular notion of effect size, the relative risk, behaves, for small θ_a and $\delta > 0$, very similarly to the odds ratio, and will therefore not be separately considered in our experiments.

If we pick $\mathcal{H}_1$ restricted to $\Theta(\delta')$, then we could simply use the beta prior mentioned before with support conditioned on this set. What about the more realistic case of a $\mathcal{H}_1$ with $\delta \in \Theta^+(\delta')$? A first, intuitive (and certainly defensible) approach would be to use a prior W_1' that is spread out over $\Theta^+(\delta')$, e.g. (if $\delta' > 0$) the beta prior as above conditioned on $\delta \geq \delta'$. However, in terms of the GRO criterion, there are good reasons to still use a prior W_1^* that puts all prior mass on $\Theta(\delta')$, the boundary of the real parameter space $\Theta(\delta^+)$. Namely, for the resulting E-variable process $S^{(1)}_{[n_a,n_b,W_1^*]}, S^{(2)}_{[n_a,n_b,W_1^*]}, \ldots$, it holds for every m that

$$\text{for all } (\theta_a, \theta_b) \text{ with } d((\theta_a, \theta_b)) > \delta', \ \mathbf{E}_{Y^{(m)} \sim P_{(\theta_a,\theta_b)}}[\log S^{(m)}_{[n_a,n_b,W_1^*]}] \geq$$
$$\min_{\theta \in \Theta(\delta')} \mathbf{E}_{Y^{(m)} \sim P_\theta}[\log S^{(m)}_{[n_a,n_b,W_1^*]}]. \tag{5.3}$$

Thus, we might want to use the prior W_1^* also if δ can be more extreme than δ', since if δ is actually more extreme, the expected (log-) evidence against $\mathcal{H}_0$ using W_1^* (even though designed for δ') will actually get larger anyway.

The advantage of the first approach is that it will lead to a much higher growth rate ($\mathbf{E}_{P_{(\theta_a,\theta_b)}}[\log S^{(m)}_{[n_a,n_b,W_1']}]$ much larger than $\mathbf{E}_{P_{(\theta_a,\theta_b)}}[\log S^{(m)}_{[n_a,n_b,W_1^*]}]$) if we are 'lucky' and $|d(\theta_a,\theta_b)| \gg |\delta'|$. The price to pay is that it will lead to somewhat smaller growth if $d((\theta_a,\theta_b))$ is (still larger than but) close to δ' (experiments omitted). It is easy to see why: the prior W_1' must spread out its mass over a much larger subset of $[0,1]^2$ than W_1^*. Therefore, the E-variables based on W_1' will perform somewhat worse than those based on W_1^* if the data are sampled from a point (θ_a^*,θ_b^*) in the support of W_1^*, simply because W_1^* gives much larger prior support in a neighborhood of (θ_a^*,θ_b^*). For this reason, and also because it is computationally a lot simpler, we decided to focus our experiments on the second approach rather than the first.

Calculating the prior and posterior for restricted $\mathcal{H}_1$. For both notions of effect size, θ_a and θ_b can no longer be independent for any prior on $\Theta(\delta)$. Hence, the prior and posterior do not longer admit the composition in terms of beta densities as in Proposition 2. For example, when putting a prior on $\Theta(\delta)$ with the additive effect size notion, we know the new domain of θ_a would be $[0, 1-\delta]$. θ_b is completely determined by θ_a and δ in this case. We will still use a beta prior on $\Theta(\delta)$ and calculate posteriors by a numerical approach, explained in Appendix S2 in the Supplementary Material.

5.2. Working with restrictions on event rate

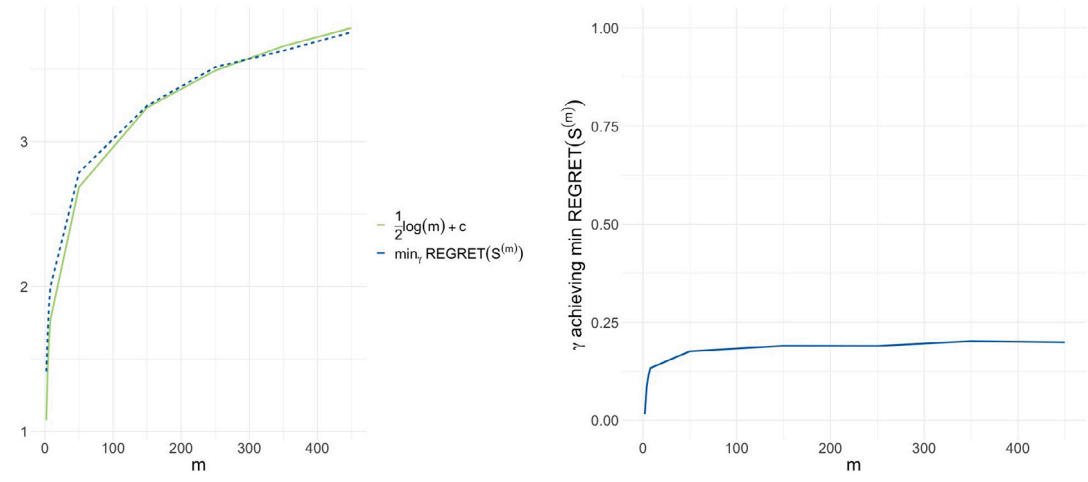
In practice, researchers often already have estimates of the occurrence rate of events in the control group in their experiments; for example, estimates of the proportion of patients that recover from a disease under standard care are known, and researchers investigate whether the proportion of recovered patients is higher in a group receiving an experimental treatment. This restriction on θ_a can be incorporated in the E-variable. This incorporation becomes especially easy if $\mathcal{H}_1$ is already restricted to a set $\Theta^+(\delta')$ with minimal relevant effect size δ'. For then $\Theta(\delta')$ contains just one point (θ_a^*,θ_b^*) (in the case of the linear effect size, this is $(\theta_a,\theta_a+\delta)$), and the E–variable constructed according to the guidelines of the previous subsection, which puts all its mass on δ' even though we allow $\delta \geq \delta'$, would be the generic E–variable corresponding to putting prior mass 1 on (θ_a^*,θ_b^*).

6. Illustration via simulated data

In this section, we illustrate properties of our E-variables for 2×2 application through simulated data, generated with our software package (Ly et al., 2022). First, we determine a reasonable choice of beta prior hyper-parameter to use in (4.4) in terms of the GRO-criterion. Thereafter, we show by more simulations that our proposal for the beta prior hyper-parameter based on GRO also performs well in terms of power. Finally, we compare the power of our E-variable with this default prior choice and different restrictions on $\mathcal{H}_1$ to Fisher's exact test.

REGROW. For simplicity, in all our experiments we will invariably set the beta prior hyper-parameters to $\alpha_a = \alpha_b = \beta_a = \beta_b = \gamma$ for some $\gamma > 0$ (recall that any such choice leads to a valid E-variable). We will aim for the γ that minimizes (2.7) in the worst-case over all $\theta_1^* \in [0,1]^2$, thereby following the REGROW *(relative growth-rate optimality in worst-case)* criterion of Grünwald et al. (2024), who give a minimax regret motivation for this choice. In essence, the prior minimizing, among all distributions over $[0,1]^2$, the maximum of (2.7) over all θ_1^* can be viewed as the prior that allows us to learn θ_1^* as fast as possible (based on a minimal sample) in the worst-case. Here we are contented to adopt a sub-optimal but computationally convenient prior by restricting the minimum to be over a 1-dimensional family of beta priors with hyper parameter γ. We find the minimizing γ through experiments: results are depicted in Fig. 2. It depends on the number of data blocks m, which is unknown in advance, but for large m, in the setting with $n_a = n_b = 1$, it converges to $\gamma \approx 0.18$, and this is the value we will take as our default choice — our experiments below indicate that it remains a good choice, also when our main concern is power, and also under restrictions on $\mathcal{H}_1$.

Power. Whereas growth rate is the natural performance measure in experiments that may always be continued at some point in the future, traditionally oriented researchers may be more interested in power. The question is then whether the optimal asymptotic choice $\gamma \approx 0.18$ in terms of the relative GRO property for unrestricted $\mathcal{H}_1$ is also the optimal choice in terms of power (which is usually considered in combination with some minimal effect size, i.e. a restricted $\mathcal{H}_1$). The following experiment shows that by and large it is. For simplicity we only illustrate the case $n_a = n_b = 1$ and a desired power of 0.8. For various effect sizes δ, and various values of γ, we first determined the smallest sample size (number of blocks) m such that, under optional stopping up until and including m, the power is ≥ 0.8 in the worst case over all (θ_a,θ_b) with $\delta = \theta_b - \theta_a$. Here by 'optional stopping up until and including m', we mean 'we stop and reject the null iff $S^{(m')}_{[n_a,n_b,W_{[\gamma]}]} > \alpha^{-1}$ for some $m' \in \{1,2,\ldots,m\}$, and we stop and accept the null if this is not the case (so m is the maximal sample size we consider)'. We call this m the *worst-case* sample size needed for 80% power at effect size δ with prior parameter γ. The reason for calling it worst-case is that in practice, by engaging in optional stopping with a fixed maximal sample size, the *expected sample size* of this procedure is smaller: if, for $m' < m$, we already have $S^{(m')}_{[n_a,n_b,W_{[\gamma]}]} > \alpha^{-1}$ then we stop and reject early; if not, we go on until we have seen m blocks and then stop (and reject iff $S^{(m)}_{[n_a,n_b,W_{[\gamma]}]} > \alpha^{-1}$). We thus performed two simulation experiments: first, to estimate the worst-case sample size (at $\alpha = 0.05$), and second, to estimate the expected sample size. Again, the estimates were obtained by re-simulating a sequence of data blocks K times for a large number of K, making sure the bias and variance of the estimates were sufficiently small.



(a) $\min_\gamma \text{REGRET} S^{(m)}$ (b) $\arg\min_\gamma \text{REGRET} S^{(m)}$

Fig. 2. Minimized regret w.r.t. Beta prior hyperparameter γ for the two-sample stream E-variable for two proportions (4.3). Relative growth rate (see (2.7)) was estimated through 10000 simulations and REGRET was calculated as the maximum over θ_1^*.

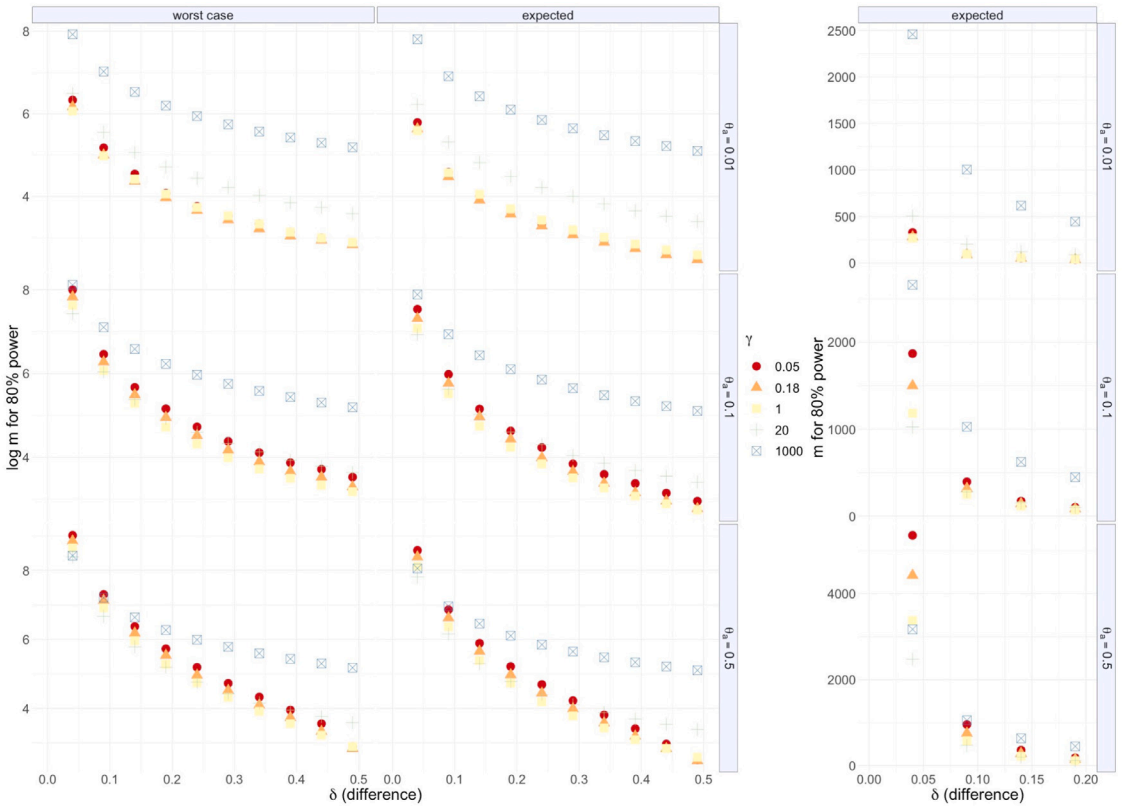


Fig. 3. In 2000 simulations the natural logarithm, left, or identity, right, of the number of data blocks m ("sample sizes") needed for achieving 80% power while testing at $\alpha = 0.05$ for distributions with varying group means and varying differences between group means were estimated for different beta prior parameter values.

In Fig. 3 results of these experiments are depicted. We make two observations: first, almost no difference in sample sizes to plan for between $\gamma = 0.18$ and $\gamma = 0.05$ was observed for distributions with small expected sample sizes (represented by the triangles and the dots, which overlap for most data points), and other values of γ obtained smaller power, indicating that the relative growth-optimal $\gamma = 0.18$ could in practice be used as a default setting for our E-variable — and as a consequence, we recommend it as such.

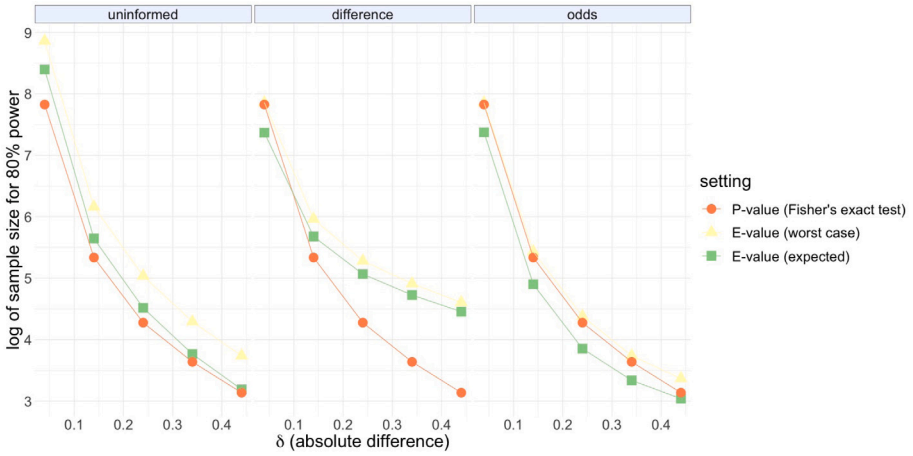


Fig. 4. Estimates from 1000 simulations of worst-case and expected sample sizes for achieving 80% power estimated for three types of E-variables with different restrictions on $\mathcal{H}_1$, and the sample size to plan for with Fisher's exact test. Hypothesized effect sizes were 0.04 for the E-variables with prior information on the absolute difference and were converted equivalently for the log odds ratio prior information case, and we set $\gamma = 0.18$ for the beta priors.

Second, in the rightmost panel we see that for distributions with *very* small relative differences between θ_a and θ_b, e.g. $P_{0.5,0.58}$, values of γ higher than 0.18 yielded a higher power, whereas for such δ, the relative GROW criterion was optimized for $\gamma = 0.18$ for the corresponding (very large) stopping times in our simulation experiments. This is not surprising given what is known for simple $\mathcal{H}_0 = \{P_{\theta_0}\}$: when testing a point null θ_0 with a 1-dimensional exponential family alternative, safe tests based on Bayes factors with standard Bayesian (e.g. Gaussian or conjugate) priors do not obtain optimal power in an asymptotic sense: they reject if $|\hat{\theta} - \theta_0|^2 \gtrsim (\log n)/n$ (with $\hat{\theta}$ denoting the MLE; see the example on Z-tests by Grünwald et al., 2024) whereas based on nonstandard 'switching' (van der Pas and Grünwald, 2018) or 'stitching' methods (Howard et al., 2021), corresponding to special priors with densities going to infinity as effect size goes to 0, one can get rejection if $|\hat{\theta} - \theta_0|^2 \gtrsim (\log \log n)/n$. However, there is a significant price to pay in terms of the constants hidden in the asymptotics, and in practice, 'standard' priors may very well perform better at all but very large sample sizes (Maillard, 2019). Given that the higher γ, the more the beta prior behaves like a switch prior, we conjecture that what we see in Fig. 3 on the right at very small δ is a version of the switching/stitching phenomenon with a composite null; since it only kicks in at very large sample sizes, we prefer $\gamma = 0.18$ as the default choice after all.

Finally, we compared the performance of our E-variables with the "default" beta priors with $\gamma = 0.18$ with their classical counterpart, Fisher's exact test. We show that with Fisher's exact test, type-I error probability guarantee is lost, whereas with the E-variables it remains bounded — since these results are exactly as would be expected from the theory they have been placed in the supplementary material (Fig. S4.1 in the Supplementary Material). In the main text below, we compare worst-case and expected stopping times of the E-variables with- and without restrictions on $\mathcal{H}_1$ for sample sizes one would need to plan for when analyzing experiment results with Fisher's exact test; see Fig. 4. We noticed that the expected sample sizes achieved under optional stopping with the E-variable with unrestricted $\mathcal{H}_1$ were very similar to the sample sizes needed to plan for with Fisher's exact test. When using a correctly specified restriction on $\mathcal{H}_1$ (the leftmost data points in the second and third subfigures), this expected number of samples is even considerably lower than the sample size to plan for with Fisher's exact test. However, under misspecification, when the difference or log odds ratio used in the design of the E-variable turns out to be a lot smaller than the real difference present in the data generating machinery, one should expect to collect more samples (the data points towards the right in the second subfigure). This effect would disappear if we were to put a prior on the full $\Theta^+(\delta)$ rather than the boundary $\Theta(\delta)$, at the price of slightly worse behavior in the well-specified case when data is sampled from $\Theta(\delta)$. Note that in Fig. 4 we used the default beta prior parameters $\gamma = 0.18$ found optimal for the unrestricted case for the restricted cases as well; some first experiments revealed that changing the prior parameter values did not lead to significant changes in power for the restricted E-variables (results not shown). We do however offer the possibility in our software package (Ly et al., 2022) to run similar experiments for users to determine the optimal prior parameter γ for a given expected sample size and $\Theta^{(+)}(\delta')$.

Beyond two-stream data: safe tests for k proportions. We also compared the performance of the extended version of our E-variable for k Bernoulli data streams to the corresponding classical, nonsequential counterpart, the chi-squared test (McHugh, 2013). In this setting, we have a $k \times 2$ contingency table test, where we test whether k Bernoulli data streams come from the same source. The extension of (4.4) to k data streams analogously to (3.3) is straightforward. In simulation experiments, it was observed that our E-variable with uniform priors significantly outperforms the chi-square test for small sample sizes and large effect sizes (see Fig. 5). For absolute differences of at least $\delta_{\max} = 0.45$, the expected sample size becomes significantly smaller than the fixed sample size needed for the chi-squared test. This is probably partially explained by the fact that the statistic used for the chi-squared test only asymptotically follows a chi-squared distribution, in contrast to our E-variable test, which is exact, valid under finite sample sizes. This means that for expected cell counts smaller than 5 the chi-square test should not be used, reflected in an increased number of samples needed for similar power (McHugh, 2013).

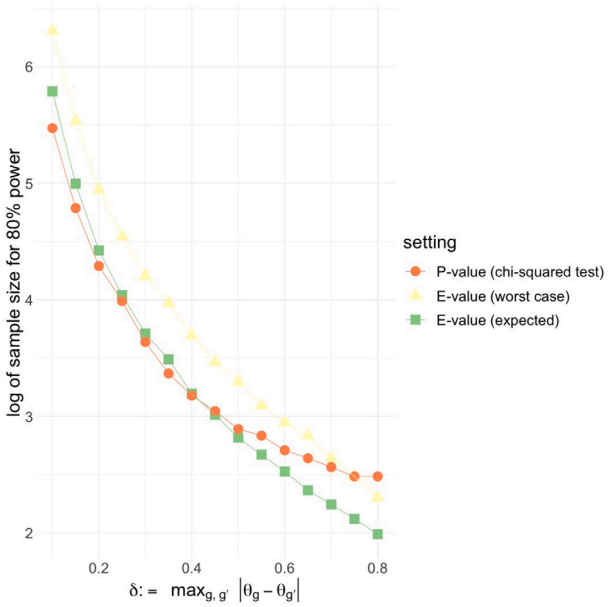


Fig. 5. Estimates from 1000 simulations of worst-case and expected sample sizes for achieving 80% power estimated for testing with the k-stream E-variable, and the sample size to plan for with the chi-square test. Data were simulated with balanced data blocks, $\vec{n} = (1, 1, 1, 1)$ and $\vec{\theta}$ was set as an equally spaced grid from $\theta_a = 0.1$ to $\theta_k = \theta_a + \delta_{\max}$. We set $\gamma = 1$ for the beta priors.

7. Illustration via real world data

We will now demonstrate the approach through a real-world example: the SWEPIS study on labor induction (Wennerholm et al., 2019). Wagenmakers and Ly (2020) have used this example before to illustrate how using single p-values to make decisions can hide valuable information in research data.

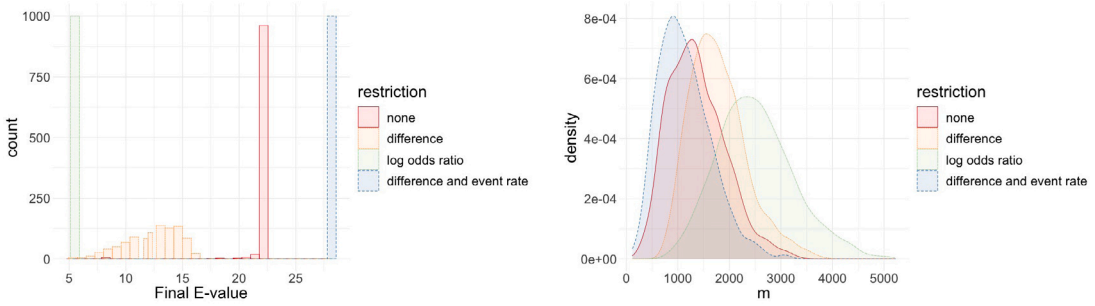
In the SWEPIS study, two groups of pregnant women were followed. In the first group labor was induced at 41 weeks, and in the second labor was induced after 42 weeks. The study was stopped early, as 6 cases of stillbirth were observed in the 42-weeks group (at $n_b = 1379$), as compared to 0 in the 41-weeks group (at $n_a = 1381$). These data yield a significant Fisher's exact test, $\text{P} \approx 0.015$, for testing that the number of stillbirths in the 42-weeks group is higher, when (wrongly) assuming that n_a and n_b were fixed in advance to the above values.

If we had used E-variables for continuously analyzing this data, would we then have found evidence for superiority of the 41 weeks approach, and would we have stopped the study earlier? As the E-variables we propose are not exchangeable, i.e., their values change under permutations of the data sequences, a direct comparison to the results of the SWEPIS study is not possible as the exact data stream is not available. To simulate a "real-time" scenario equivalent to the SWEPIS study, we assume we collect a total of 1380 data blocks, with $n_a = n_b = 1$, with a total of 2760 observations. We already know that in group a, 0 events are observed. In group b, 6 events are observed, of which we know that the last event was observed in data block 1380, directly before the study was stopped. Hence, we can simulate the "real-time" data by permuting the indices of the observations in group b in the 1379 first data blocks.

Four different approaches for analyzing the data with E-variables were explored: without any restriction on $\mathcal{H}_1$, with a restriction based on the additive divergence measure (the minimal difference between the groups), with a restriction based on the log odds ratio, and with a restriction on the event rate in the control group *and* on the minimal difference. The minimal difference, log odds ratio and event rate used were chosen based on a large recent meta-analysis on stillbirths (Muglu et al., 2019); we used $\delta = 0.00318$ as a restriction on the difference between the groups, $\log(2)$ for the log odds ratio and 0.0001 as the event rate. For all E-variables, the default beta prior hyperparameters with $\gamma = 0.18$ as earlier were used.

In Fig. 6 the spread of the evidence collected with the four types of E-variables in 1000 simulations analogous to the SWEPIS setting is depicted. Because the observed effect size was higher than expected, E-values obtained with the (too low) restriction on the effect size were lower than the E-values obtained with the E-variable without restrictions. Adding the restriction on the event rate increased the E-values, and in all 1000 simulations, the SWEPIS study would have been stopped before the occurrence of the sixth stillbirth. Fig. 6 also depicts results of a second simulation experiment, where we sampled 1000 data streams from $P_{0.6/1380}$ and recorded the stopping times while analyzing the streams with the four E-variables with different restrictions on $\mathcal{H}_1$. With the E-variables without restriction, or with a restriction on the event rate and difference between the groups, we would have often stopped data collection earlier than in the SWEPIS setting.

Wagenmakers and Ly (2020) with their method also found evidence for the existence of a difference between the two groups, but not nearly of the same degree: they reported Bayes factors that varied, depending on the choice of the prior, between 1 and



(a) Simulated E-values in SWEPIS setting, stopping at $m = 1380$ or when $E \geq 20$

(b) Simulated stopping times in setting with continuing until $E \geq 20$

Fig. 6. Spread of E-values and stopping times observed with safe analysis of 1000 simulations of data streams analogous to the SWEPIS scenario, with four different types of restrictions on $\mathcal{H}_1$.

5.4 (note that whenever we reject, our product of E-values, which like a Bayes factor can be thought of as a prequential likelihood ratio, must be ≥ 20). A possible explanation for this difference could be that the Bayes factors used for collecting evidence in their study are not designed for analyzing stream data. As we also saw in our experiments, choosing the wrong prior or restriction on $\mathcal{H}_1$ can make a large difference for the evidence collected.

We can thus conclude that, would the monitoring of the study have been performed with E-variables instead of p-values, first of all we would have collected *correct* evidence for a higher proportion of stillbirths in the 42-weeks group, and second, the degree of evidence is quite similar to that collected with the (incorrectly determined) p-value: both are significant at the 0.05 level. The study design with E-variables could effortlessly follow the classical flow of clinical trial design: before the start of the trial, a power analysis could be carried out to determine the minimum sample sizes that one needs to arrange resources for under the desired sampling scheme (balanced or unbalanced, see Ly et al., 2022, Vignettes). In collaboration with experts, a restriction could be put on the event rate or difference between the groups to potentially improve the power. During the study, because the SWEPIS design is balanced, an E-value is calculated each time a new patient has come in the control and treatment groups, and the researchers and data safety monitoring boards are allowed to look at the results and decide to stop the study at any time, not affecting Type-I error probability guarantees. After the study or in case the study is stopped early because of reasons beyond rejecting the null hypothesis, because E-values were used, one can always continue a study later or combine E-values across multiple studies in an anytime-valid meta-analysis (Ter Schure and Grünwald, 2022)

8. Other E-variables for two data streams

8.1. The GRO E-variable for some exponential and location families

The simplification (4.2) shows that in the Bernoulli case with simple $\Theta_1 = \{(\theta_a^*, \theta_b^*)\}$, we can take in our denominator p_{θ_0} with $\theta_0 = \frac{n_a}{n}\theta_a^* + \frac{n_b}{n}\theta_b^*$ — which can also be interpreted as the distribution in the null corresponding to a mixture of the means, rather than the mixture of two distributions in the null. The Bernoulli model is a special case of 1-parameter exponential families which can all be parameterized in terms of their means so that $\Theta \subset \mathbf{R}$ and $\mathbf{E}_{P_\theta}[Y] = \theta$; this is also possible for some location families that are not of exponential form. This suggests that, for all such models, instead of (3.1) we might also consider the likelihood ratio (4.2). For the Bernoulli model, both definitions will coincide, but for general 1-parameter exponential families they do not since their corresponding set of densities is not convex. The question is now whether (4.2) defines an E-variable for general exponential families. It turns out that the answer is *no* in general, but *yes* in some special cases. For a negative example, consider the case with $\Theta = \mathbf{R}^+$ representing the family of exponential distributions in their mean-value parameterization, i.e. $p_\theta(y) = \lambda \exp(-\lambda y)$ with $\lambda = 1/\theta$ and take $n_a = n_b = 1$. A simple calculation shows that for any $\theta_a^* \neq \theta_b^* \in \Theta$, we have $\lim_{\theta \to \infty} \mathbf{E}_{Y_a, Y_b \text{ i.i.d.} \sim P_\theta}[p_{\theta_a^*}(Y_a) p_{\theta_b^*}(Y_b) / p_{(\theta_a^* + \theta_b^*)/2}(Y_a, Y_b)] = \infty$. The negative binomial families provide, by a similar calculation, another negative example. For a positive example, consider the case with $\Theta = \mathbf{R}$ representing the Gaussian location family with fixed variance 1 and again take $n_a = n_b = 1$. A simple calculation shows that (4.2) is equal to the likelihood ratio for testing whether the difference $Z = Y_a - Y_b$ is a Gaussian with variance $\sqrt{2}$ with either mean 0 or mean $\theta_b - \theta_a$. This is in fact the standard paired-sample Z-test that would normally be advised in this situation. In fact it is the GRO E-variable for this situation:

Proposition 3. *Let $\{P_\theta : \theta \in \Theta\}$ represent a family of probability distributions with densities p_θ, with Θ a convex set in $\mathbf{R}^k$ for some $k \geq 1$. For any $\theta_a^*, \theta_b^* \in \Theta$ we have: if (4.2) is an E-variable for $\Theta_1 = \{(\theta_a^*, \theta_b^*)\}$ then it is the GRO E-variable for $\Theta_1 = \{(\theta_a^*, \theta_b^*)\}$.*

The proof is immediate from Proposition 1. The proposition implies that in the special cases in which (4.2) does provide an E-variable, it is to be preferred (achieves better growth) above our original construction (3.1). (3.1) has the advantage that it provides an E-variable relative to arbitrary models. We plan to study the cases in which (4.2) can be used instead in future work.

8.2. The conditional E-variable for tests of two proportions

Wald (1947) proposed a 2-sample sequential probability ratio test (SPRT) for the 2×2 setting. Since SPRTs can be written in terms of products of E-variables (although products of E-variables often do not give SPRTs; see the discussion by Grünwald et al., 2024), let us see what E-variables Wald's test corresponds to. The setting is restricted to size-2 blocks with $n_a = n_b = 1$. We measure effect size with d the log-odds ratio (5.2) and consider an alternative with a $d(\theta_a, \theta_b)$ that is at least some given δ. Using that, for all $(\theta_a, \theta_b) \in (0,1)^2$, $z \in \{0,1,2\}$, the conditional probability mass function $p_{\theta_a,\theta_b}(Y_a, Y_b \mid \sum Y_a + Y_b = z)$ only depends on the log-odds ratio, we can write it, as $q_\delta(y_a, y_b|z)$ where q_δ is a probability mass function whose definition depends on (θ_a, θ_b) only via $\delta = d((\theta_a, \theta_b))$. We then take as our E-variable $S_{\text{COND},\delta} := q_\delta(Y_a, Y_b \mid Y_a + Y_b)/q_0(Y_a, Y_b \mid Y_a + Y_b)$. Since the conditional distribution $q_0(Y_a, Y_b \mid Z)$ is the same for all distributions in the null, this conditional likelihood gives an E-variable and can be used instead of our generic E-variable. Since for this Bernoulli case, our E-variable is in fact GRO, we would expect this new conditional E-variable to perform worse in terms of GRO (and for the reasons given in Section 2 also in terms of the amount of data needed before one can reject at a desired power), and experiments (not reported here) confirm that it indeed performs slightly worse for δ close to 0, and substantially worse for larger δ. This is already suggested by the fact that, unlike the GRO E-variable, $S_{\text{COND},\delta}$ takes on value 1 whenever $y_a = y_b$, effectively ignoring data blocks in which both outcomes are the same. Another disadvantage is that it can only be used in combination with effect size given by the odds ratio or any monotonic transformation thereof; whereas the GRO E-variable can also be combined with the difference $\theta_b - \theta_a$ or any other desirable notion of effect size.

9. Conclusion

We have established E-variables and test martingales for the general i.i.d.-data streams problem. We have demonstrated, using theory, simulations and a real-world example that, for tests of two proportions, by choosing an appropriate prior on Θ_1, the method can be made competitive with classical methods that do not allow for optional stopping. Whereas in this paper, we have focused on testing, our E-variables can also be extended to get *anytime-valid confidence sequences* (Howard et al., 2021; Lai, 1976), i.e. confidence sequences for effect sizes that are valid even under optional stopping. This requires us to first extend the testing to scenarios with $\delta \geq \delta_1$ vs. $\delta \leq \delta_0$ for $\delta_0 \neq 0$, that is, null hypotheses with $\theta_a \neq \theta_b$. We have reported on this extension in Turner and Grünwald (2023). Our work also suggests a question for future work that is practically relevant, easy to state but hard to answer: to what extent do our findings generalize to logistic regression?

Acknowledgments

The authors gratefully acknowledge Reuben Adams, Rianne de Heide, Wouter Koolen, Muriel Perez, Judith ter Schure and Akshay Balsubramani for useful conversations and in particular Adams and De Heide for performing experiments that inspired the E-variables presented here. This work is part of the Enabling Personalized Interventions (EPI) project, which is supported by the Dutch Research Council (NWO) in the Commit2 - Data –Data2Person program under contract 628.011.028.

Appendix A. Supplementary data

The online Supplementary Material consists of five appendices. Appendix S1 contains detailed proofs. Appendix S2 contains a detailed description of the numerical approach to calculating E-variables for restricted H1. Appendix S3 contains a detailed description of Gunel–Dickey Bayes factors. Appendix S4 contains optional stopping experiments, and, finally, Appendix S5 describes how to 'learn' appropriate block group sizes n_a and n_b based on past data.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jspi.2023.106116.

References

Darling, D.A., Robbins, H., 1967. Confidence sequences for mean, variance, and median. Proc. Natl. Acad. Sci. USA 58 (1), 66.

Dawid, A.P., 1984. Present position and potential developments: Some personal views, statistical theory, the prequential approach. J. Roy. Statist. Soc. Ser. A 147 (2), 278–292.

Eckhoff, J., 1993. Helly, radon, and Carathéodory type theorems. In: Handbook of Convex Geometry, vol. A,B. Elsevier, pp. 389–448.

Grünwald, Peter, de Heide, Rianne, Koolen, Wouter, 2024. Safe testing. J. R. Statist. Soc. Ser. B Stat. Methodol. (To appear, with discussion, in 2024. Available as preprint as arXiv:1906.07801 [math.ST]).

Gunel, E., Dickey, J., 1974. Bayes factors for independence in contingency tables. Biometrika 61 (3), 545–557.

Henzi, Alexander, Ziegel, Johanna F., 2022. Valid sequential inference on probability forecast performance. Biometrika.

Howard, Steven R., Ramdas, Aaditya, McAuliffe, Jon, Sekhon, Jasjeet, 2021. Time-uniform, nonparametric, non-asymptotic confidence sequences. Ann. Statist..

Jamil, Tahira, Ly, Alexander, Morey, Richard D., Love, Jonathon, Marsman, Maarten, Wagenmakers, Eric-Jan, 2017. Default "Gunel and Dickey" Bayes factors for contingency tables. Behav. Res. Methods 49 (2), 638–652.

Johari, Ramesh, Koomen, Pete, Pekelis, Leonid, Walsh, David, 2022. Always valid inference: Continuous monitoring of a/b tests. Oper. Res. 70 (3), 1806–1821.

John, Leslie K., Loewenstein, George, Prelec, Drazen, 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychol. Sci. 23 (5), 524–532.

Kass, Robert E., Vaidyanathan, Suresh K., 1992. Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. J. R. Stat. Soc. Ser. B Stat. Methodol. 54 (1), 129–144.

Koolen, Wouter M., Grünwald, Peter, 2022. Log-optimal anytime-valid e-values. Internat. J. Approx. Reason. 141, 69–82.

Lai, Tze Leung, 1976. On confidence sequences. Ann. Statist. 4 (2), 265–280.

Lhéritier, Alix, Cazals, Frédéric, 2018. A sequential non-parametric multivariate two-sample test. IEEE Trans. Inform. Theory 64 (5), 3361–3370.

Lindon, Michael, Malek, Alan, 2022. Anytime-valid inference for multinomial count data. arXiv preprint arXiv:2011.03567.
Ly, Alexander, Turner, Rosanne, Schure, Judith Ter, 2022. R-package `safestats`. CRAN.
Maillard, Odalric-Ambrym, 2019. Mathematics of statistical sequential decision making. Thèse de Habilitation.
Manole, Tudor, Ramdas, Aaditya, 2023. Martingale methods for sequential estimation of convex functionals and divergences. IEEE Trans. Inform. Theory.
McHugh, Mary L., 2013. The chi-square test of independence. Biochem. Medica 23 (2), 143–149.
Muglu, Javaid, Rather, Henna, Arroyo-Manzano, David, Bhattacharya, Sohinee, Balchin, Imelda, Khalil, Asma, Thilaganathan, Basky, Khan, Khalid S., Zamora, Javier, Thangaratinam, Shakila, 2019. Risks of stillbirth and neonatal death with advancing gestation at term: A systematic review and meta-analysis of cohort studies of 15 million pregnancies. PLoS Med. 16 (7), e1002838.
Pace, Luigi, Salvan, Alessandra, 2020. Likelihood, replicability and Robbins' confidence sequences. Internat. Statist. Rev. 88 (3), 599–615.
van der Pas, S., Grünwald, P., 2018. Almost the best of three worlds: Risk, consistency and optional stopping for the switch criterion in nested model selection. Statist. Sinica 28 (1), 229–255.
Ramdas, Aaditya, Grünwald, Peter, Vovk, Vladimir, Shafer, Glenn, 2022. Game-theoretic statistics and safe anytime-valid inference. arXiv preprint arXiv:2210.01948.
Ramdas, Aaditya, Ruf, Johannes, Larsson, Martin, Koolen, Wouter, 2020. Admissible anytime-valid sequential inference must rely on nonnegative martingales. arXiv preprint arXiv:2009.03167.
Robbins, Herbert, 1970. Statistical methods related to the law of the iterated logarithm. Ann. Math. Stat. 41 (5), 1397–1409.
Royall, Richard, 1997. Statistical Evidence: A Likelihood Paradigm. Chapman and Hall.
Shafer, Glenn, Shen, Alexander, Vereshchagin, Nikolai, Vovk, Vladimir, 2011. Test martingales, Bayes factors and p-values. Statist. Sci. 84–101.
Shafer, Glenn, et al., 2021. Testing by betting: A strategy for statistical and scientific communication. J. Roy. Statist. Soc. Ser. A 184 (2), 407–431.
Shekhar, S., Ramdas, A., 2021. Nonparametric two-sample testing by betting. arXiv preprint arXiv:2112.09162.
Siegmund, David, 2013. Sequential Analysis: Tests and Confidence Intervals. Springer Science & Business Media.
Ter Schure, Judith, Grünwald, Peter, 2022. ALL-IN meta-analysis: Breathing life into living systematic reviews. F1000Research 11.
Ter Schure, Judith, Pérez-Ortiz, Muriel F., Ly, Alexander, Grunwald, P., 2020. The safe logrank test: Error control under continuous monitoring with unlimited horizon. arXiv preprint arXiv:2011.06931.
Turner, Rosanne J., Grünwald, Peter D., 2023. Exact anytime-valid confidence intervals for contingency tables and beyond. Statist. Probab. Lett. 109835.
Vovk, Vladimir, Wang, Ruodu, 2021. E-values: Calibration, combination, and applications. Ann. Statist..
Wagenmakers, Eric-Jan, Ly, Alexander, 2020. Bayesian scepsis about SWEPIS: Quantifying the evidence that early induction of labour prevents perinatal deaths. URL psyarxiv.com/5ydpb.
Wald, Abraham, 1947. Sequential Analysis. Wiley.
Wennerholm, Ulla-Britt, Saltvedt, Sissel, Wessberg, Anna, Alkmark, Mårten, Bergh, Christina, Wendel, Sophia Brismar, Fadl, Helena, Jonsson, Maria, Ladfors, Lars, Sengpiel, Verena, et al., 2019. Induction of labour at 41 weeks versus expectant management and induction of labour at 42 weeks (SWEdish post-term induction study, SWEPIS): Multicentre, open label, randomised, superiority trial. Br. Med. J. 367.