Generic E-Variables for Exact Sequential k-Sample Tests that allow for Optional Stopping

Rosanne Turner[a,b]*, Alexander Ly[a,c] and Peter Grünwald[a,d]

June 23, 2022

*[a]CWI, Amsterdam, part of NWO-I [b]University Medical Center Utrecht, Brain Center, [c]University of Amsterdam, Department of Psychology and [d]Leiden University, Department of Mathematics
Declarations of interest: none. *Corresponding author: Rosanne Turner (`Rosanne.Turner@cwi.nl`) National Research Institute for mathematics and computer science in the Netherlands (CWI), Science Park 123, 1098 XG Amsterdam, The Netherlands

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, where we also allow for the incorporation of different restrictions on a composite alternative. Comparison to p-value analysis in simulations and a real-world 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 are robust under flexible sampling scenarios, in which one is allowed to engage in optional continuation and optional stopping. We focus on the setting with data coming from several groups, the goal being to test whether the underlying distributions are all the same or not. 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 $k > 2$. Our methods are based on E–variables and test martingales. While to some extent going back as far as Darling and Robbins (1967), interest in these concepts has exploded only very recently, in part in relation to the ongoing replicability crisis in the applied sciences (Howard et al., 2021; Ramdas et al., 2020; Vovk and Wang, 2021; Shafer, 2021; Grünwald et al., 2022; Pace and Salvan, 2019; Manole and Ramdas, 2021; Henzi and Ziegel, 2021).

Thus, 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 (1.1) 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. We consider a testing scenario, in which the null hypothesis $\mathcal{H}_0$ expresses that $\theta_a = \theta_b$ and the alternative $\mathcal{H}_1$ expresses that $d(\theta_a, \theta_b) > \delta$ for some divergence measure d and some effect size $\delta \geq 0$. We design a family of tests for this scenario 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 arbitrary $\{P_\theta : \theta \in \Theta\}$ and in combination with arbitrary divergence measures d. To our knowledge such a general construction is entirely new. For purposes of illustration and insight we choose to apply it to a very simple, classical problem: 2×2 contingency tables, with, in Section 5, an extension to $k \times 2$ tables. As is well-known (for completeness we provide simulations demonstrating this in the supporting information), if a standard fixed-design method for this scenario, the p-value resulting from Fisher's exact test, is (ab)used with optional stopping, the type-I error blows up. In contrast, our tests retain type-I error guarantee while, due to the optional stopping, having power competitive with Fisher's p-value. In fact, in the $k \times 2$ application (but not in general) our test has a GRO[1] *(growth-rate optimal)* property, GRO being the analogue of 'optimal power' in our optional continuation setting.

Our test depends on the choice of a prior distribution on the alternative $\mathcal{H}_1 = \{P_{(\theta_a, \theta_b)} : (\theta_a, \theta_b) \in \Theta_1\}$ with $\Theta_1 \subset \{(\theta_a, \theta_b) : \theta_a, \theta_b \in \Theta\}$. 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. Still, the prior affects how fast one will tend to reject the null if it is indeed false. For the case that no clear prior knowledge is available, one may use the prior

---

[1]Nonstandard abbreviations: GRO: growth-rate optimal; REGROW: relative growth-rate optimality in worst-case

that is optimal in terms of the relative GRO criterion; again the resulting test also has good power properties.

E–Variable Perspective; Block–wise Approach; Optional Continuation In its simplest form, an E-*variable* is a nonnegative random variable S such that under all distributions P in the null hypothesis,

$$\mathbf{E}_P[S] \leq 1. \tag{1.1}$$

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. A block is 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 j-th 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 2.1).

At each point in time, the running product of block E-variables observed so far is itself an E-variable, and the random process of the products is known as a *test martingale*. An E-variable-based test at level α is then a test with, in combination with any stopping rule τ, reports 'reject' if and only if the product of E-values corresponding to all blocks that were observed so far and have already been completed, is larger than $1/\alpha$. The full definition of τ may, and often will, be unknown to the user — the user only needs to get the signal to stop and can then report the product E-variable. A classical paired one-sample test corresponds to the special case with $n_a = n_b = 1$ and data coming in in the order $a, b, a, b, \ldots$.

We can combine E-variables from different trials that share a common null (but may be defined relative to a different alternative) by multiplication, and still retain type-I error control. If we used p-values rather than E-variables we would have to resort to e.g. Fisher's method for combining p-values, which, in contrast to multiplication of e-values, is invalid if there is a dependency between the (decision to perform) tests. With E-variables, such dependencies pose no problems for error control. Thus, in our setting, even if the design (i.e. n_a and n_b) is fixed in advance and optional stopping plays no role, we might still want to use the E-variable based tests described in this paper rather than a classic p-value based approach, since it allows us to do optional continuation over many experiments/studies (essentially, doing a meta-analysis (Ter Schure et al., 2021)) while keeping type-I error control.

E-variables and test martingales are explained in more detail in Section 1.1 below, but we refer to Grünwald et al. (2022); Shafer (2021) for an extensive introduction to E-variables, their use in 'optional continuation' over several studies, and their enlightening *betting* interpretation. The general story that emerges from these papers as well as, for example, (Vovk and Wang, 2021; Ramdas et al., 2020) is that E-variables and test martingales are the 'right' generalization of likelihood ratios to the case that both $\mathcal{H}_0$ and $\mathcal{H}_1$ can be composite and combination of data from several trials may be required.

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, 2019). 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 (Figure S3.1 in Appendix S3 in the Supporting Material), E–value 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 6 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 so that the p-value that led them to stop 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–value based methodology we would have obtained sufficient evidence to stop for harm after the same number of events had occurred. Additionally, this E–value, even though based on a stopped trial, can be effortlessly combined with E–values 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. We will report on this extension elsewhere.

SWEPIS summarized its data as a 2×2 contingency table. In Section 3 and 4 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. (2022) and already put into practice by e.g. Robbins (1970); 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 in his seminal (Wald, 1947) . Wald's test can be turned into a product of E-variables and would then be safe to use under optional stopping. Yet, as explained in Section 7.2, in the 2×2 setting the resulting E-variables do not grow as fast as the ones introduced here, and the underlying idea does not generalize to arbitrary models or effect size notions. Other earlier approaches (e.g. (Siegmund, 2013, Section V.2)) are based on asymptotic approximations. In contrast, our E–variable based

tests are exact and nonasymptotic. 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., 2022). 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 2.2) . 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 S2 (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.

Another, very recent, approach that bears some similarity to ours are the two-sample tests from Manole 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 the optimal (as measured according to the GRO criterion) 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 (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 the remainder of this introductory section, we formally introduce E-variables, optional stopping and the concept of GRO-optimality. In Section 2 we propose our generic E-variable for tests of two streams in general and investigate when it has the GRO property. In Sections 3 and 4 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 5 and 6 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 7 we compare our generic approach to other E-variables one might define for this problem, including the ones based on Wald's section test. We end with a conclusion. All proofs are in the appendix.

1.1 E-Variables and Test Martingales, Safety and Optimality

We first need to extend the notion of E-variable to random processes:

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 $P \in \mathcal{H}_0$,*

$$\mathbf{E}_P \left[S_{(j)} \mid Y_{(1)}, \ldots, Y_{(j-1)} \right] \leq 1. \tag{1.2}$$

   That is, for each $y^{(j-1)} \in \mathcal{Y}^{j-1}$, all $P_0 \in \mathcal{H}_0$, (1.1) 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 (1.1), i.e. $\mathbf{E}_P[S^{(m)}] \leq 1$ for all $P \in \mathcal{H}_0$. Definition 1 adapts and slightly modifies terminology from (Shafer et al., 2011). As follows from that paper, in standard martingale terminology, what we call a test martingale is a non-negative supermartingale relative to the filtration induced by $\{Y_{(j)}\}_{j \in \mathbf{N}}$, with starting value 1.

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, 2021) tells us that, for all $0 < \alpha \leq 1$, $P \in \mathcal{H}_0$,

$$P(\text{there exists } j \text{ such that } S^{(j)} \geq 1/\alpha) \leq \alpha. \tag{1.3}$$

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*.

Example 1. Let P_0 and Q be any two distributions for the process $Y_{(1)}, Y_{(2)}, \ldots$, and let $\mathcal{H}_0 = \{P_0\}$ represent a simple null. Let $S^{(m)}$ denote the likelihood ratio for m outcomes and $S_{(j)}$ its constituent factors, i.e.

$$S^{(m)} = \frac{q(Y^{(m)})}{p_0(Y^{(m)})} = \prod_{j=1}^{m} S_{(j)} \text{ with } S_{(j)} = \frac{q(Y_{(j)} \mid Y^{(j-1)})}{p_0(Y_{(j)} \mid Y^{(j-1)})} \tag{1.4}$$

7

where $q(y_{(m)} \mid y^{(m-1)})$ denotes the conditional density corresponding to Q and $p_0(y_{(m)} \mid y^{(m-1)})$ the one corresponding to P_0 with respect to a common underlying measure. Then the likelihood ratio process $\{S^{(m)}\}_{m \in \mathbf{N}}$ constitutes a test martingale, and the process of past-conditional likelihoods $\{S_{(j)}\}$ is a sequential E-variable process relative to $\mathcal{H}_0$. This can be immediately verified by directly calculating the conditional expectation of $S_{(j)}$ given $Y^{(j-1)}$, noticing that the densities $p_0(Y_{(j)}|Y^{(j-1)})$ cancel in the calculation.

GRO-Optimality, Simple $\mathcal{H}_1$ Just like for p-values, the definition of E-variables only requires explicit specification of $\mathcal{H}_0$, not of an alternative hypothesis $\mathcal{H}_1$. $\mathcal{H}_1$ becomes crucial once we distinguish between 'good' and 'bad' E-variables: E-variables have been designed to remain small, with high probability, under the null $\mathcal{H}_0$. But if $\mathcal{H}_1$ rather than $\mathcal{H}_0$ is true, then 'good' E-variables should produce evidence (grow — because the larger the E-variable, the closer we are to rejecting the null) against $\mathcal{H}_0$ as fast as possible. To make this precise, first consider simple (singleton) $\mathcal{H}_1 = \{Q\}$. We start with the one-outcome setting of (1.1), i.e. we look at a single E-variable $S_{(j)}$ in isolation for a single outcome $Y_{(j)}$. Its optimality is measured in terms of

$$\mathbf{E}_Q[\log S_{(j)}], \tag{1.5}$$

and the E-variable which maximizes this quantity among all E-variables that can be written as functions of $Y_{(j)}$ (i.e. non-negative random variables satisfying (1.1)), 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 $S_{\mathrm{GRO}(Q),(j)}$ More generally, E-variable $S^{(m)}$ is called *growth rate optimal* relative to Q for $Y^{(m)}$, or simply Q-GRO for $Y^{(m)}$, if, among all (unconditional) E–variables that can be written as a function of $Y^{(m)}$, it maximizes

$$\mathbf{E}_Q[\log S^{(m)}]. \tag{1.6}$$

We will denote this E–variable, if it exists, by $S^{(m)}_{\mathrm{GRO}(Q)}$. The idea to maximize (1.6) goes back to Kelly (1956); the GRO-terminology is from Grünwald et al. (2022). The larger an E–variable or test martingale tends to be under the alternative, the better it scores in the GRO sense. Of course, the same would still hold if we were to replace the logarithm by another strictly increasing function. But there are various compelling reasons for why one should take a logarithm here — see Grünwald et al. (2022); Shafer (2021). One interesting reason, not explicitly covered by these two papers, was already given by Breiman (1961) and is explained in detail by (Ter Schure et al., 2021, Appendix B.1): the Q-GRO test martingale is also the test martingale which minimizes the expected number of data points needed before the null can be rejected if we use the test with the aggressive stopping rule described before (reject at the smallest j such that $S^{(j)} \geq 1/\alpha$). Thus, using the Q-GRO test martingale is quite analogous to employing a test that maximizes power. One can also directly see that both notions must be connected by noting that GRO implies optimizing the expectation of $\log S^{(j)}$ whereas power at fixed sample size j is the probability that $\log S^{(j)}$ is larger than $-\log \alpha$. Note that we cannot directly use power in designing tests, since the notion of power requires a

fixed sampling plan, which we will usually not have: we may not want or not be able to stop at the first j such that we can reject — for example, we might want to stop early for harm (Section 6), or we might want to lower α if the first few outcomes look very promising. So we will measure optimality in terms of GRO instead, but for practical usefulness we *do* hope that, in cases where we do follow the sampling plan above (stop as soon as $S^{(j)} \geq 1/\alpha)$), our power remains reasonable. This is suggested by Breiman's observation above, but we want to check it nevertheless. Such a check is done successfully for the 2×2 model in Section 5.

In 'nice' cases, the Q-GRO E–variable (1.6) 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 $P \in \mathcal{H}_0$, $D(Q\|P) < \infty$, $D(\cdot\|\cdot)$ denoting the KL divergence. Suppose the following condition holds (with q, p the density of Q and P, respectively):*

$$\text{There exists a } P \in \mathcal{H}_0 \text{ such that } S_{(1)} = q(Y_{(1)})/p(Y_{(1)}) \text{ is an E–variable.} \tag{1.7}$$

Then $S_{(1)} = S_{\mathrm{GRO}(Q),(1)}$ 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 $S_{\mathrm{GRO}(Q)}^{(m)} = \prod_{j=1}^{m} S_{\mathrm{GRO}(Q),(j)}$, i.e. the Q-GRO optimal (unconditional) E-variable for $Y^{(m)}$ is the product of the individual Q-GRO optimal E-variables.

If Condition (1.7) 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 $(S_{\mathrm{GRO}(Q)}^{(j)})_{j \in \mathbf{N}}$. We will then have that $S_{\mathrm{GRO}(Q),(j)} = s_Q(Y_{(j)})$ for a fixed function $s_Q : \mathcal{Y} \to \mathbf{R}_0^+$.

Example 2. [Simple $\mathcal{H}_1$ and Simple $\mathcal{H}_0$] Consider $\mathcal{H}_1 = \{Q\}$ and simple $\mathcal{H}_0 = \{P_0\}$ and arbitrary Q' such that the $Y_{(j)}$ are i.i.d. according to P, Q and Q'. Then $S_{(j)} = q'(Y_{(j)})/p_0(Y_{(j)})$ is an E-variable for $Y_{(j)}$, irrespective of the definition of Q', by the same argument as in Example 1. By the Proposition above, the Q-GRO E-variable for $Y_{(j)}$ is given by setting $q' = q$. Then $\mathbf{E}_Q[S_{\mathrm{GRO}(Q),(j)}] = \mathbf{E}_{Y_{(j)} \sim Q}[\log q(Y_{(j)})/p_0(Y_{(j)})]$ also coincides with the KL divergence between Q and P_0.

In Section 2 (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 $\mathcal{H}_0$ is composite and not convex, 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 and nonconvex), 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 (1.6) large, but it may be even larger for other E-variables.

9

GRO and Composite $\mathcal{H}_1$ 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 (1.6). 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' $P_{\theta_1^*}$. 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 $S_{(j)}$. 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' $P_{\theta_1^*}$. Assuming the convergence happens fast, we have that

$$\mathbf{E}_{Y^{(m)} \sim P_{\theta_1^*}} \left[\log S_{\mathrm{GRO}(\theta_1^*)}^{(m)} - \log \prod_{j=1}^{m} S_{(j)} \right] \tag{1.8}$$

is small, i.e. we may expect that the test martingale $\prod_{j=1}^{m} S_{(j)}$ grows not much slower than $S_{\mathrm{GRO}(\theta_1^*)}^{(m)} = \prod_{j=1}^{m} S_{\mathrm{GRO}(\theta_1^*),(j)}$, the best test martingale (maximizing $\mathbf{E}_{Y^{(m)} \sim P_{\theta_1^*}}[\log S]$ over all E-variables S for $Y^{(m)}$) we could have used if we had known the true $P_{\theta_1^*}$ all along.

2 Two-Stream Safe Tests

Consider the two-stream setting introduced in the beginning of the paper. To formalize it further, we introduce *calendar time* $t = 1, 2, \ldots$ and corresponding random variables V_t and G_t: at each t, we obtain an outcome V_t in $\mathcal{Y}$ in group $G_t \in \{a, b\}$. Importantly though, at this point we make no assumptions about the relative ordering of outcomes from the two groups. At time t, we have that t_a, the number of a's that are observed so far, and t_b, the number of b's observed so far, satisfy $t_a + t_b = t$, but subject to this constraint we allow them coming in any order, e.g. first all a's, or first all b's, or interleaved. For example, with $t_a = 3$ and $t_b = 2$, we might have $V_1 = Y_{1,a}, V_2 = Y_{2,a}, V_3 = Y_{3,a}, V_4 = Y_{1,b}, V_5 = Y_{2,b}$ (all as come first, $G_1 = G_2 = G_3 = a, G_4 = G_5 = b$) but also, for example $V_1 = Y_{1,a}, V_2 = Y_{1,b}, V_3 = Y_{2,a}, V_4 = Y_{3,a}, V_5 = Y_{2,b}$.

We thus have that 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)

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

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 $p_{\theta_0}(y_{1,a}, \ldots, y_{t_a,a}, y_{1,b}, \ldots, y_{t_b,b})$.

2.1 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:

$$s(y_a^{n_a}, y_b^{n_b}; n_a, n_b, \theta_a^*, \theta_b^*) :=$$

$$\frac{p_{\theta_a^*}(y_a^{n_a})}{\prod_{i=1}^{n_a} \left(\frac{n_a}{n} p_{\theta_a^*}(y_{i,a}) + \frac{n_b}{n} p_{\theta_b^*}(y_{i,a}) \right)} \cdot \frac{p_{\theta_b^*}(y_b^{n_b})}{\prod_{i=1}^{n_b} \left(\frac{n_a}{n} p_{\theta_a^*}(y_{i,b}) + \frac{n_b}{n} p_{\theta_b^*}(y_{i,b}) \right)}. \qquad (2.2)$$

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* $\sup_{\theta \in \Theta} \mathbf{E}_{V^n \sim P_\theta} [s'(V^n)] \leq 1$, *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, 2021)), 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; but $\{P_\theta : \theta \in \Theta\}$ is just the Bernoulli model on one outcome, which is convex, so that in this setting, we get the GRO E-variable.

To illustrate, consider the basic case in which data comes in 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-value

$$S^{(m)}_{[n_a,n_b,\theta_a^*,\theta_b^*]} := \prod_{j=1}^{m} S_{(j),[n_a,n_b,\theta_a^*,\theta_b^*]} \quad ; \quad S_{(j),[n_a,n_b,\theta_a^*,\theta_b^*]} := s(Y_{(j)}; n_a, n_b, \theta_a^*, \theta_b^*). \tag{2.3}$$

By Ville's inequality (1.3), 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 (1.3), 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 (2.3) 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 t-th 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. For this, we introduce a function $f : \bigcup_{t\geq 0} \mathcal{Y}^t \times \{0,1\}^t \to \{\text{STOP-BLOCK}, \text{CONTINUE}\}$ that, at each point in time t, decides whether the current

block should end ($f(V^t, G^t) =$ STOP-BLOCK) or not ($f(V^t, G^t) =$ CONTINUE). As long as the value of this function does not depend on the actual outcomes V_t observed after the last block that was completed, all requirements for having a test martingale and thus for safe optional stopping are met. For example, suppose that on data $V_1, G_1, V_2, G_2, \ldots, V_t, G_t$ observed so-far, f has output STOP-BLOCK at m occasions, the last time at $t' = t - k$ for some $k > 0$. Then $f(t)$ is allowed to depend on $Y^{(m)}$ and G^t, but for any fixed $Y^{(m)} = y^{(m)}, G^t = g^t$, for all $y^k, y'^k \in \mathcal{Y}^k$, we must have $f((y^{(m)}, y^k), g^t) = f((y^{(m)}, y'^k), g^t)$. 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 (2.2) to the scenario where we do not compare 2, but k i.i.d. data streams. Indeed, in the appendix 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 $p_{\vec{\theta}} := p_{\theta_a}(y_a^{t_a}) p_{\theta_b}(y_b^{t_b}) \ldots p_{\theta_k}(y_k^{t_k})$. 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 $n = \sum_{g=1}^{k} n_g$:

$$s(\vec{y}^n; \vec{n}, \vec{\theta}^*) := \prod_{g=1}^{k} \frac{p_{\theta_g^*}(y_g^{n_g})}{\prod_{i=1}^{n_g} \left(\sum_{g'=1}^{k} \frac{n_{g'}}{n} p_{\theta_{g'}^*}(y_{i,g}) \right)}, \tag{2.4}$$

for testing the null where $\theta_a = \theta_b = \ldots = \theta_k$; it is again GRO if $\{P_\theta : \theta \in \Theta\}$ is convex. We now return to the notationally simpler 2-sample case except for a short example of an application of this extension as a flexible and exact alternative to the chi-square test in section 5.

2.2 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,

$$s(y_a^{n_a}, y_b^{n_b}; n_a, n_b, W_1) :=$$

$$\frac{\prod_{i=1}^{n_a} p_{W_1, a}(y_{i,a})}{\prod_{i=1}^{n_a} \left(\frac{n_a}{n} p_{W_1, a}(y_{i,a}) + \frac{n_b}{n} p_{W_1, b}(y_{i,a}) \right)} \cdot \frac{\prod_{i=1}^{n_b} p_{W_1, b}(y_{i,b})}{\prod_{i=1}^{n_b} \left(\frac{n_a}{n} p_{W_1, a}(y_{i,b}) + \frac{n_b}{n} p_{W_1, b}(y_{i,b}) \right)}. \tag{2.5}$$

is itself also an E–variable, as 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

13

$\mathcal{W}(\Theta)$, 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 2.1 is straightforward). We start with some prior W_1 for the first batch $Y_{(1)}$ but we now use, for the j-th batch $Y_{(j)}$, the *Bayesian posterior* $W_1 \mid Y^{(j-1)}$ as prior to define the j-th E–variable with:

$$S_{[n_a, n_b, W_1]}^{(m)} := \prod_{j=1}^{m} S_{(j),[n_a,n_b,W_1]} \quad ; \quad S_{(j),[n_a,n_b,W_1]} := s(Y_{(j)}; n_a, n_b, W_1 | Y^{(j-1)}). \tag{2.6}$$

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 (1.3). 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 5 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 (2.6) as

$$S_{[n_a, n_b, W_1]}^{(m)} = \prod_{j=1}^{m} \frac{\prod_{i=1}^{n_a} p_{W_{1,a}|Y^{(j-1)}}(Y_{(j-1)n_a+i,a}) \prod_{i=1}^{n_b} p_{W_{1,b}|Y^{(j-1)}}(Y_{(j-1)n_b+i,b})}{\prod_{g \in \{a,b\}} \prod_{i=1}^{n_g} \left(\frac{n_a}{n} p_{W_{1,a}|Y^{(j-1)}}(Y_{(j-1)n_g+i,g}) + \frac{n_b}{n} p_{W_{1,b}|Y^{(j-1)}}(Y_{(j-1)n_g+i,g}) \right)}$$

$$\underset{=}{\text{if } \{P_\theta : \theta \in \Theta\} \text{ convex, } \mathcal{Y} \text{ finite}} \prod_{j=1}^{m} \prod_{i=1}^{n_a} \frac{p_{W_{1,a}|Y^{(j-1)}}(Y_{(j-1)n_a+i,a})}{p_{\breve{\theta}_0|Y^{(j-1)}}(Y_{(j-1)n_a+i,a})} \prod_{i=1}^{n_b} \frac{p_{W_{1,b}|Y^{(j-1)}}(Y_{(j-1)n_b+i,b})}{p_{\breve{\theta}_0|Y^{(j-1)}}(Y_{(j-1)n_b+i,b})} \tag{2.7}$$

with $\breve{\theta}_0|Y^{(j-1)} \in \Theta$ 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, 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,

14

and using that, if independent under the prior, θ_a and θ_b must also be independent under the posterior, we can then further rewrite (2.6) as

$$\frac{\int p_{\theta_a}(Y_a^m)dW_a(\theta_a)\int p_{\theta_b}(Y_b^m)dW_b(\theta_b)}{\prod_{j=1}^m \prod_{g\in\{a,b\}}\left(\frac{1}{2}p_{W_{1,a}|Y^{(j-1)}}(Y_{j,g})+\frac{1}{2}p_{W_{1,b}|Y^{(j-1)}}(Y_{j,g})\right)} \quad \underset{\text{if } \{P_\theta:\theta\in\Theta\}\text{ convex}}{=} \quad (2.8)$$

$$\frac{\int p_{\theta_a}(Y_a^m)dW_a(\theta_a)\int p_{\theta_b}(Y_b^m)dW_b(\theta_b)}{\prod_{j=1}^m \prod_{g\in\{a,b\}}p_{\breve\theta_0|Y^{(j-1)}}(Y_{j,g})} \quad (2.9)$$

where the equality holds if $\{P_\theta:\theta\in\Theta_0\}$ is convex and $\mathcal{Y}$ is finite so that (2.7) holds. As seen from (2.8), even without finiteness or convexity, the numerator of the generic product E–value 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 (2.9) is the product density function of two identical distributions for one outcome, and Proposition 2 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 j-th block would not be independent given $Y^{(j-1)}$, whereas in (2.9) they are, so we may conclude that in general, our e-variables are not equivalent to any Bayes factor.

3 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

$$p_{\theta_a,\theta_b}(y_a^{t_a},y_b^{t_b}) := p_{\theta_a}(y_{1,a},\ldots,y_{t_a,a})p_{\theta_b}(y_{1,b},\ldots,y_{t_b,b}) = \theta_a^{t_{a1}}(1-\theta_a)^{t_a-t_{a1}}\theta_b^{t_{b1}}(1-\theta_b)^{t_b-t_{b1}}. \quad (3.1)$$

with t_{a1} 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 $\theta_0 \in \Theta = [0,1]$. (3.1) now simplifies to:

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

with t_1 the number of ones in the sequence $y^{t_a+t_b}=y_1,\ldots,y_{t_a+t_b}$, 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 $\mathcal{H}_1=\{P_{\theta_a^*,\theta_b^*}\}$. (2.2) can now be written as:

$$s(y_a^{n_a},y_b^{n_b};n_a,n_b,\theta_a^*,\theta_b^*) := \frac{p_{\theta_a^*}(y_a^{n_a})}{p_{\theta_0}(y_a^{n_a})}\cdot\frac{p_{\theta_b^*}(y_b^{n_b})}{p_{\theta_0}(y_b^{n_b})}, \quad\text{where}\quad \theta_0=\frac{n_a}{n}\theta_a^*+\frac{n_b}{n}\theta_b^*. \quad (3.2)$$

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 (2.5), this now simplifies to:

$$s(y_a^{n_a}, y_b^{n_b}; n_a, n_b, W_1) = s(y_a^{n_a}, y_b^{n_b}; n_a, n_b, \theta_a^*, \theta_b^*) \text{ for } \theta_g^* = \mathbf{E}_{\theta_g \sim W_1}[\theta_g], g \in \{a, b\}. \tag{3.3}$$

Combining this with (2.7) we infer that

$$S_{[n_a, n_b, W_1]}^{(m)} = \prod_{j=1}^{m} \prod_{i=1}^{n_a} \frac{p_{\breve{\theta}_a | Y^{(j-1)}}(Y_{(j-1)n_a+i,a})}{p_{\breve{\theta}_0 | Y^{(j-1)}}(Y_{(j-1)n_a+i,a})} \prod_{i=1}^{n_b} \frac{p_{\breve{\theta}_b | Y^{(j-1)}}(Y_{(j-1)n_b+i,b})}{p_{\breve{\theta}_0 | Y^{(j-1)}}(Y_{(j-1)n_b+i,b})} \tag{3.4}$$

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

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 $\breve{\theta}_a | Y^{(j-1)} = \mathbf{E}_{\theta_a \sim W_a | Y_a^{(j-1)n_a}}[\theta_a]$, 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 $\breve{\theta}_a, \breve{\theta}_b, \breve{\theta}_0$ as above satisfy: $\breve{\theta}_a | Y^{(j-1)} = (U_a + \alpha_a)/((j-1)n_a + \alpha_a + \beta_a), \breve{\theta}_b | Y^{(j-1)} = (U_b + \alpha_b)/((j-1)n_b + \alpha_b + \beta_b)$ respectively, and $\breve{\theta}_0 | Y^{(j-1)}$ 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 $\breve{\theta}_0$ simplifies to $\breve{\theta}_0 | Y^{(j-1)} = (U + (1 + \kappa)\alpha_a)/((j-1)n + (1 + \kappa)\alpha_a + (1 + \kappa)\beta_a)$.*

4   (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 5 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 5.

4.1 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

$$\Theta(\delta) = \left\{ \theta \in [0,1]^2 : d(\theta) = \delta \right\} \quad ; \quad \Theta^+(\delta) = \begin{cases} \left\{ \theta \in [0,1]^2 : d(\theta) \geq \delta \right\} & \text{if } \delta > 0 \\ \left\{ \theta \in [0,1]^2 : d(\theta) \leq \delta \right\} & \text{if } \delta < 0, \end{cases} \quad (4.1)$$

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 (4.1) but d set to

$$d((\theta_a, \theta_b)) = \log \left(\frac{\theta_b}{1 - \theta_b} \cdot \frac{1 - \theta_a}{\theta_a} \right). \quad (4.2)$$

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 Figure 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 $\mathcal{H}_1$ restrict 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

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

$$\min_{\theta \in \Theta(\delta')} \mathbf{E}_{Y^{(m)} \sim P_\theta}[\log S^{(m)}_{[n_a,n_b,W_1^*]}]. \quad (4.3)$$

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.

(a) $d((\theta_a, \theta_b)) = \theta_b - \theta_a$ (b) $d((\theta_b, \theta_a)) = \log\left[\frac{\theta_b}{1-\theta_b}\frac{1-\theta_a}{\theta_a}\right]$
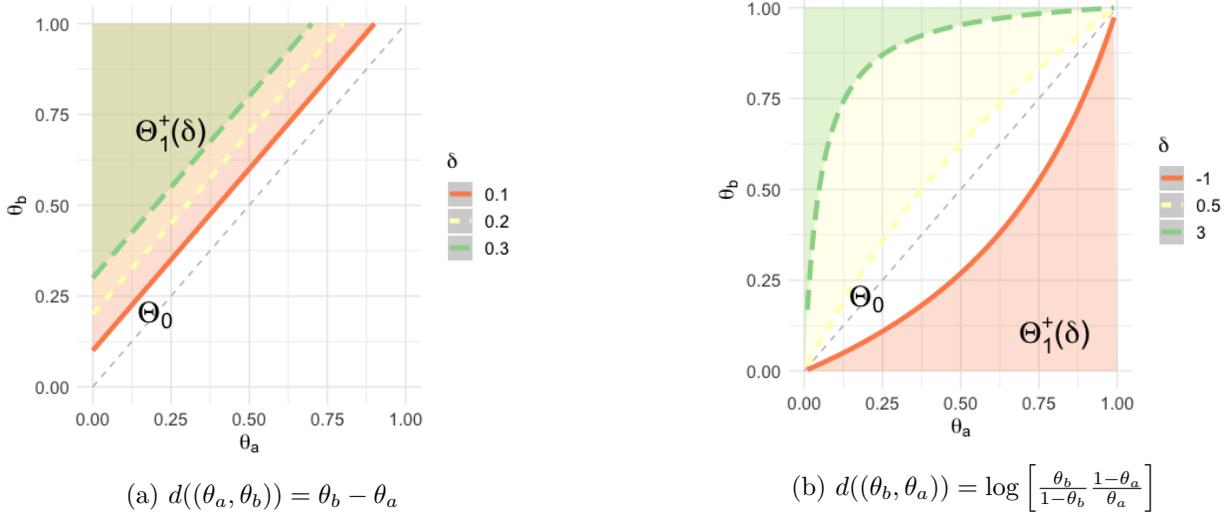
Figure 1: Examples of restricted alternative hypothesis parameter spaces for several values of two divergence measures; the difference between group means and the log odds ratio. Θ_0 denotes the null hypothesis parameter space; $\Theta_1^+(\delta)$ the restricted alternative hypothesis parameter space.

The advantage of the first approach is that it will lead to much higher GROwth ($\mathbf{E}_{P_{(\theta_a, \theta_b)}}[\log S_{[n_a, n_b, W_1']}^{(m)}]$ much larger than $\mathbf{E}_{P_{(\theta_a, \theta_b)}}[\log S_{[n_a, n_b, W_1^*]}^{(m)}]$) 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 arger 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 S1 of the Supplementary Material.

4.2 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^*).

5 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 publicly available through Github (Ly et al., 2020). First, we determine a reasonable choice of beta prior hyper-parameter to use in (3.4) 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 (recall from Section 1.1 that while we cannot optimize for power directly, we do want procedures with reasonable 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 (1.8) 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. (2022), who give a minimax regret motivation for this choice. In essence, the prior minimizing, among all distributions over $[0,1]^2$, the maximum of (1.8) 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 by experiment: results are depicted in Figure 2. It depends on 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 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
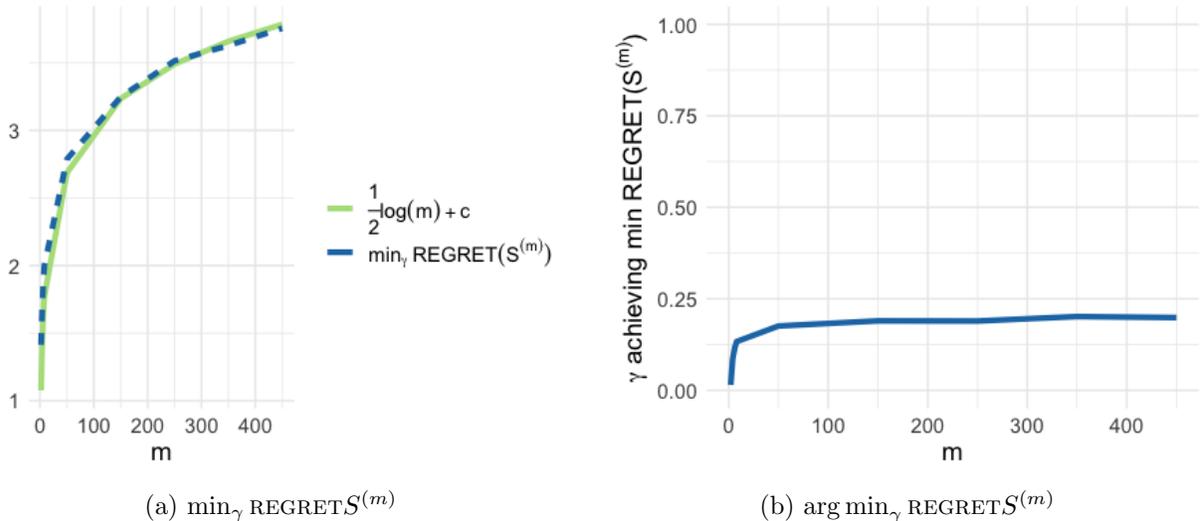
(a) $\min_\gamma$ REGRET$S^{(m)}$ (b) $\arg\min_\gamma$ REGRET$S^{(m)}$

Figure 2: Minimized regret w.r.t. Beta prior hyperparameter γ for the two-sample stream E-variable for two proporions (3.3). Relative growth rate (see (1.8)) was estimated through 10000 simulations and REGRET was calculated as the maximum over θ_1^*.

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. In Figure 3 results of these experiments are depicted. We make two observations: first, almost no difference in sample sizes to plan for between
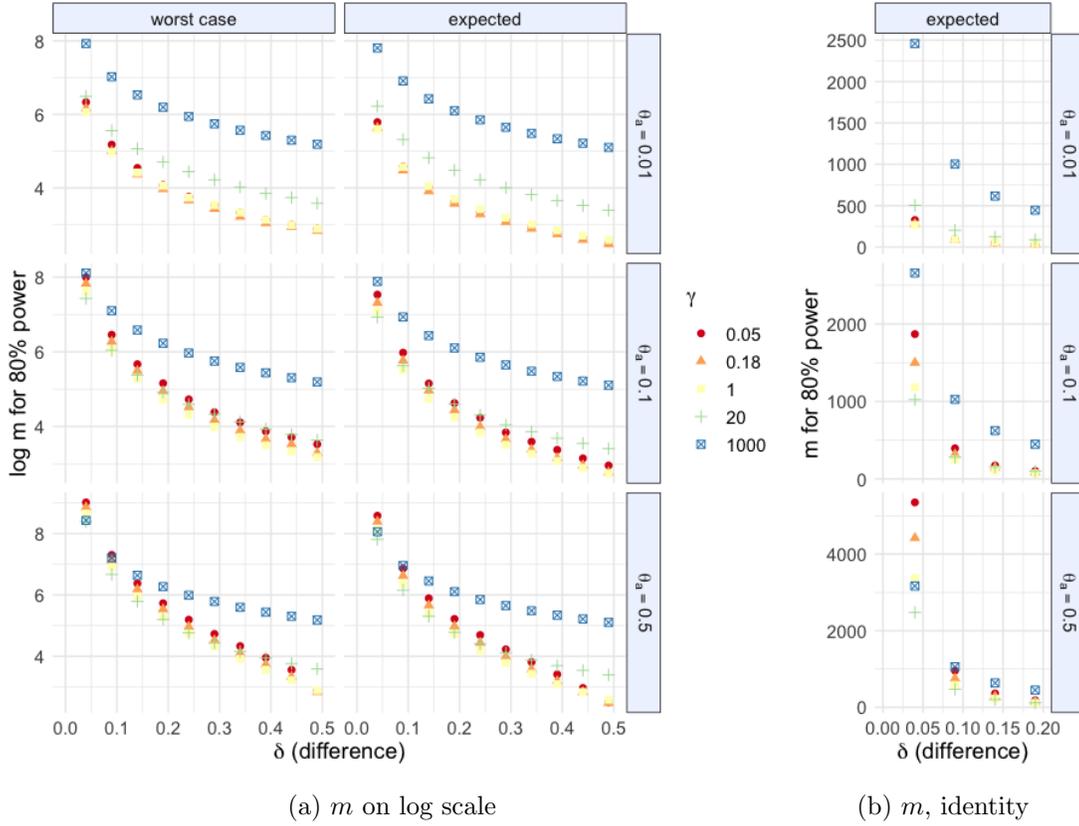
20

(a) m on log scale

(b) m, identity

Figure 3: In 2000 simulations, the (natural logarithm 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.

$\gamma = 0.18$ 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. 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

21

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. (2022)) 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 Figure 3(b) 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 (Figure S3.1 in Appendix S3 in the Supporting 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 Figure 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 behaviour in the well-specified case when data is sampled from $\Theta(\delta)$.

Note that in Figure 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., 2020) 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-square 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 (3.4) to k data streams analogously to (2.4) is straightforward. Our E-variable
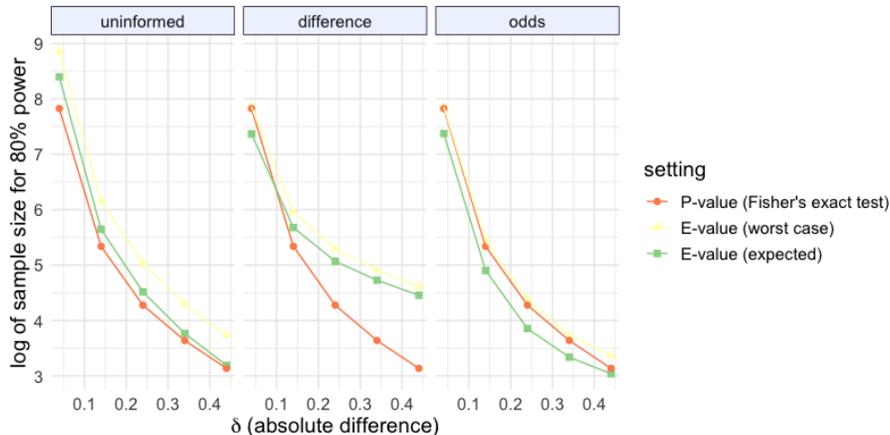
Figure 4: Estimates from 1000 simulations of worst-case and expected sample sizes for achieving 80% power estimated for three types of E-variables with different restrictions on $\mathcal{H}_1$, and the sample size to plan for with Fisher's exact test. Hypothesized effect sizes were 0.04 for the E-variables 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.

with uniform priors significantly outperforms the chi-square test for small sample sizes and large effect sizes (see Figure 5), probably explained by the fact that the chi-square test is not exact, but the E-variable is. 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).

6 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, $P \approx 0.015$, 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?
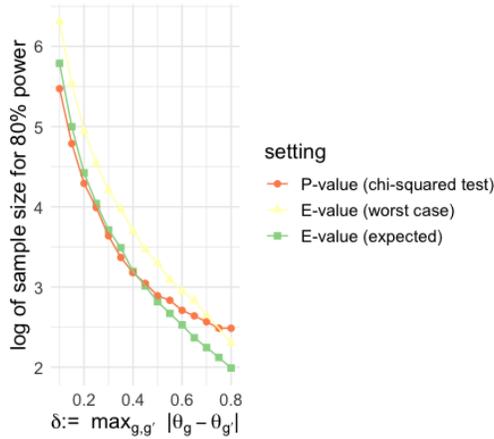
Figure 5: Estimates from 1000 simulations of worst-case and expected sample sizes for achieving 80% power estimated for testing with the k-stream E-variable, and the sample size to plan for with the chi-square test. Data were simulated with balanced data blocks, $\vec{n} = (1, 1, 1, 1)$ and $\vec{\theta}$ 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. We see that for large enough $\delta_{\max}$, the expected sample size becomes significantly smaller than the fixed sample size needed for the chi-squared test, overtaking it at approximately $\delta_{\max} = 0.45$, which really means that one third of the times the effect size is 0.15, one third it is 0.3 and one third 0.45. At these δ, the expected sample size for our 2-stream E-variable is still larger than the fixed number needed for Fisher's exact test.

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.

24

In Figure 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. Figure 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.

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. Wagemakers and Ly 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 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. These results show that when planning a prospective study, using E-variables for analysis could, through their flexibility, contribute to earlier evidence collection compared to existing methods.
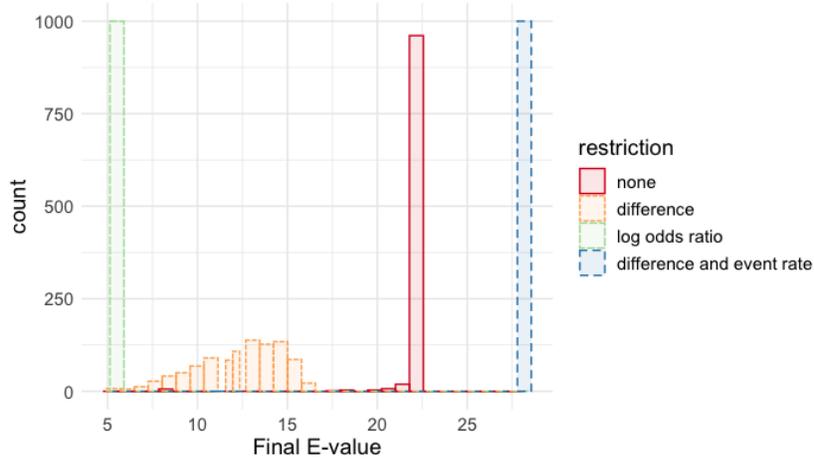
(a) Simulated E-values in SWEPIS setting, stopping at $m = 1380$ or when $E \geq 20$
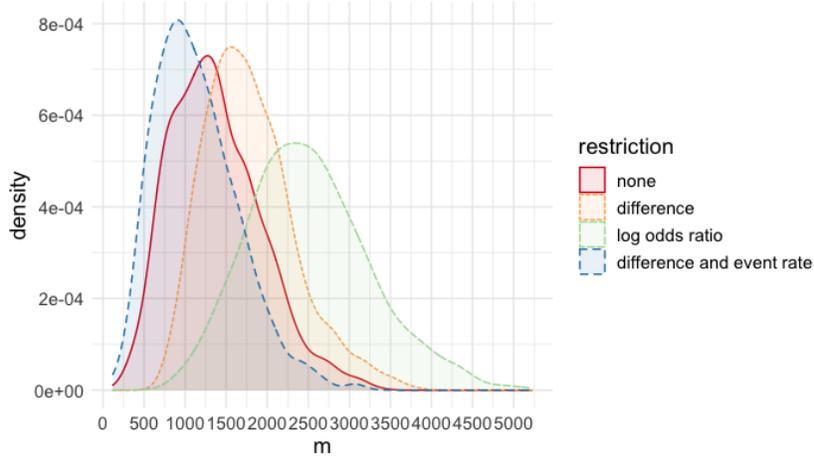
(b) Simulated stopping times in setting with continuing until $E \geq 20$

Figure 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 $\mathcal{H}_1$.

7 Other E-Variables for Two Data Streams

7.1 The GRO E-variable for some Exponential and Location Families

The simplification (3.2) 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

26

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 (2.2) we might also consider the likelihood ratio (3.2). 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 (3.2) 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 (3.2) 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 (3.2) 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 (3.2) does provide an E-variable, it is to be preferred (achieves better growth) above our original construction (2.2). (2.2) has the advantage that it provides an E-variable relative to arbitrary models. We plan to study the cases in which (3.2) can be used instead in future work.

7.2 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. (2022)), 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 (4.2) 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 1.1 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, $S_{\mathrm{COND},\delta}$ 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.

8   Conclusion

We have established E-variables and test martingales for the general two-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 will report on this extension elsewhere. 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?

Acknowledgements

Supplementary material

- Appendix S1: detailed description of numerical approach to calculating E-variables for restricted $\mathcal{H}_1$

- Appendix S2: detailed description of Gunel-Dickey Bayes factors

- Appendix S3: optional stopping experiments

References

Breiman, L. (1961). Optimal gambling systems for favorable games. *Fourth Berkeley Symposium*.

Darling, D. and Robbins, H. (1967). Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences of the United States of America*, 58(1):66.

Dawid, A. (1984). Present position and potential developments: Some personal views, statistical theory, the prequential approach. *Journal of the Royal Statistical Society, Series A*, 147(2):278–292.

Grünwald, P., de Heide, R., and Koolen, W. (2022). Safe testing. *accepted, pending minor revision, for publication in Journal of the Royal Statistical Society: Series B*.

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

Henzi, A. and Ziegel, J. F. (2021). Valid sequential inference on probability forecast performance. *Biometrika*.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Uniform, nonparametric, non-asymptotic confidence sequences. *Annals of Statistics*.

Jamil, T., Ly, A., Morey, R. D., Love, J., Marsman, M., and Wagenmakers, E.-J. (2017). Default "Gunel and Dickey" Bayes factors for contingency tables. *Behavior Research Methods*, 49(2):638–652.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 23(5):524–532.

Kass, R. E. and Vaidyanathan, S. K. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):129–144.

Kelly, J. (1956). A new interpretation of information rate. *Bell System Technical Journal*, pages 917–926.

Koolen, W. and Grünwald, P. (2021). Anytime GROW E-values. *International Journal of Approximate Reasoning*. Special issue to celebrate G. Shafer's 75th Birthday. Accepted pending major revision.

Lai, T. L. (1976). On confidence sequences. *The Annals of Statistics*, 4(2):265–280.

Lhéritier, A. and Cazals, F. (2018). A sequential non-parametric multivariate two-sample test. *IEEE Transactions on Information Theory*, 64(5):3361–3370.

Ly, A., Turner, R., and Ter Schure, J. (2020). R-package `safestats`. install in R by `devtools::install_github("AlexanderLyNL/safestats", ref = "logrank", build_vignettes = TRUE)`.

Maillard, O.-A. (2019). Mathematics of statistical sequential decision making. Thèse de Habilitation.

Manole, T. and Ramdas, A. (2021). Sequential estimation of convex divergences using reverse submartingales and exchangeable filtrations. *arXiv preprint arXiv:2103.09267*.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica*, 23(2):143–149.

Muglu, J., Rather, H., Arroyo-Manzano, D., Bhattacharya, S., Balchin, I., Khalil, A., Thilaganathan, B., Khan, K. S., Zamora, J., and Thangaratinam, S. (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 medicine*, 16(7):e1002838.

Pace, L. and Salvan, A. (2019). Likelihood, replicability and Robbins' confidence sequences. *International Statistical Review*.

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409.

Shafer, G. (2021). The language of betting as a strategy for statistical and scientific communication. *Journal of the Royal Statistical Society, Series A*.

Shafer, G., Shen, A., Vereshchagin, N., and Vovk, V. (2011). Test martingales, Bayes factors and p-values. *Statistical Science*, pages 84–101.

Siegmund, D. (2013). *Sequential analysis: tests and confidence intervals*. Springer Science & Business Media.

Ter Schure, J., Perez-Ortiz, M. F., Ly, A., and Grünwald, P. (2021). The safe log rank test: Error control under continuous monitoring with unlimited horizon. *arXiv preprint arXiv:1906.07801*.

van der Pas, S. and Grünwald, P. (2018). Almost the best of three worlds: Risk, consistency and optional stopping for the switch criterion in nested model selection. *Statistica Sinica*, 28(1):229–255.

Vovk, V. and Wang, R. (2021). E-values: Calibration, combination, and applications. *Annals of Statistics*.

Wagenmakers, E.-J. and Ly, A. (2020). Bayesian scepsis about swepis: Quantifying the evidence that early induction of labour prevents perinatal deaths.

Wald, A. (1947). *Sequential Analysis*. Wiley.

Wennerholm, U.-B., Saltvedt, S., Wessberg, A., Alkmark, M., Bergh, C., Wendel, S. B., Fadl, H., Jonsson, M., Ladfors, L., Sengpiel, V., 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. *British Medical Journal*, 367.

Appendix: Proofs

The proofs below repeatedly use Theorem 1 of Grünwald et al. (2022) and a direct corollary (called Corollary 2 by Grünwald et al. (2022)), which we re-state here, for convenience, combined as a single statement. We use the notation adopted later in the paper: for $\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\}$ and, for W a distribution on Θ_0, we write $P_W = \int P_\theta dW(\theta)$.

Theorem (Theorem 1 of Grünwald et al. (2022)) Let Y be a random variable taking values in a set $\mathcal{Y}$. Suppose Q is a probability distribution for Y with density q that is strictly positive on all of $\mathcal{Y}$ and let $\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\}$ be a set of distributions for Y where each P_θ has density p_θ. Let $\mathcal{W}_0$ be the set all distributions on Θ_0. Assume $\inf_{W_0 \in \mathcal{W}_0(\Theta_0)} D(Q\|P_{W_0}) < \infty$. Then (a) there exists a (potentially sub-) distribution P_0^* with density p_0^* such that

$$S^* := \frac{q(Y)}{p_0^*(Y)}$$

is an E-variable (p_0^* is called the *Reverse Information Projection (RIPr) of q onto* $\{p_W : W \in \mathcal{W}_0\}$ (Li, 1999; Li and Barron, 2000; Grünwald et al., 2019)). Moreover, (b), S^* satisfies

$$\sup_{S \in \mathcal{E}(\Theta_0)} \mathbf{E}_{Y \sim Q}[\log S] = \mathbf{E}_{Y \sim Q}[\log S^*] = \inf_{W_0 \in \mathcal{W}_0(\Theta_0)} D(Q\|P_{W_0}) = D(Q\|P_0^*). \tag{8.1}$$

and is thus the Q-GRO E-variable for Y. If the minimum is achieved by some W_0^*, i.e. $D(Q\|P_0^*) = D(Q\|P_{W_0^*})$, then $P_0^* = P_{W_0^*}$. Moreover, (c), if there exists an E-variable S of the form $q(Y)/p_{W_0}(Y)$ for some $W_0 \in \mathcal{W}_0$ then W_0 must achieve the infimum in (8.1) and S must be essentially equal to S^* in the sense that for all $P \in \mathcal{H}_0 \cup \{Q\}$, $P(S^* = q(Y)/p_{W_0}(Y)) = 1$. Similarly (d), if there exists a $W_0^* \in \mathcal{W}_0$ that achieves the infimum in (8.1) then $S = q(Y)/p_{W_0^*}(Y)$ is an E-variable and S is again essentially equal to S^*.

8.1 Proof of Propositions

Proof of Proposition 1 Below we state and prove a slight generalization of Proposition 1.

Proposition 4. *Let $\mathcal{H}_1 = \{Q\}$ be a singleton and let $\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\}$ be such that for some distribution W on Θ_0, $D(Q\|P_W) < \infty$. For general $\theta \in \Theta_0$ and distributions W on Θ_0, define $S_{\theta,(j)} := q(Y_{(j)})/p_\theta(Y_{(j)})$ and $S_{W,(j)} = q(Y_{(j)})/p_W(Y_{(j)})$. We have:*

1. *Suppose there exists a distribution W on Θ_0 such that $S_{W,(1)}$ is an E-variable. Then $S_{W,(1)}$ is the Q-GRO E-variable for $Y_{(1)}$. In particular, if W puts mass 1 on a particular $\theta^\circ \in \Theta_0$, then $S_{W,(1)} = S_{\theta^\circ,(1)}$ is the Q-GRO E-variable.*

2. *If $\Theta_0 = \{\theta_0\}$ is simple then, with the prior W_0 putting mass 1 on θ_0, $S_{W_0,(1)} = S_{\theta_0,(1)}$ is an E-variable and hence, by the above, also the Q-GRO E-variable.*

3. *If, for some $\theta^\circ \in \Theta_0$, $S_{\theta^\circ,(1)}$ is an E-variable and 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 $S_{\mathrm{GRO}(Q)}^{(m)} = \prod_{j=1}^m S_{\theta^\circ,(j)}$; that is, the Q-GRO optimal (unconditional) E-variable for $Y^{(m)}$ is the product of the individual Q-GRO optimal E-variables.*

Proof. Part 1 The theorem above, part (b), implies, with $Y = Y_{(1)}$, that some Q-GRO E-variable S^* for $Y_{(1)}$ exists. Part (c) then implies that we can take S^* to be equal to $S_{W,(1)}$. This implies the statement.

Part 2 is immediate. *Part 3* We assume that $S_{\theta^\circ,(1)}$ is an E–variable. Then the i.i.d. assumption implies that $S_{\theta^\circ}^{(m)} := \prod_{j=1}^m S_{\theta^\circ,(j)} = \prod q(Y_{(j)})/p_{\theta^\circ}(Y_{(j)})$ is also an E-variable. But (Grünwald et al., 2019, Theorem 1), part (c) as stated above implies (by taking a distribution W putting mass 1 on θ) that for $\mathcal{H}_0$ for which data are i.i.d., for each $m \geq 1$, that if a $\theta \in \Theta_0$ exists such that $S_\theta^{(m)}$ is an E-variable, then $S_\theta^{(m)}$ must be the Q-GRO E-variable for $Y^{(m)}$. This proves the statement. $\square$

Proof of Proposition 2 The formulae for $\breve{\theta}_a | Y^{(j-1)}$ and $\breve{\theta}_b | Y^{(j-1)}$ are standard expressions for the Bayes predictive distribution based on the given beta priors; we omit further details. As to the expression for $\breve{\theta}_0 | Y^{(j-1)}$ in terms of $\kappa = n_b/n_a$: Straightforward rewriting gives, for general $\alpha_a, \alpha_b, \beta_a, \beta_b$:

$$\breve{\theta}_0 | Y^{(j-1)} = \frac{1}{1+\kappa} \breve{\theta}_a | Y^{(j-1)} + \frac{\kappa}{1+\kappa} \breve{\theta}_b | Y^{(j-1)}. \tag{8.2}$$

If we plug in the expressions for $\breve{\theta}_a|Y^{(j-1)}, \breve{\theta}_b|Y^{(j-1)}$ and we instantiate to $\alpha_b = \kappa\alpha_a$, and $\beta_b = \kappa\beta_a$, this becomes

$$\breve{\theta}_0|Y^{(j-1)} = \frac{1}{1+\kappa}\frac{U_a + \alpha_a}{n_a(j-1) + \alpha_a + \beta_a} + \frac{\kappa}{1+\kappa}\frac{U_b + \alpha_b}{\kappa(n_a(j-1) + \alpha_a + \beta_a)}$$

$$= \frac{1}{1+\kappa}\frac{U_a + U_b + (1+\kappa)\alpha_a}{n_a(j-1) + \alpha_a + \beta_a} = \frac{U + (1+\kappa)\alpha_a}{n(j-1) + (1+\kappa)\alpha_a + (1+\kappa)\beta_a},$$

which is what we had to prove.

8.2 Proof of Theorem 1

We first restate Theorem 1 in its extended version that holds for $k \geq 2$ data streams. Let $\vec{n} = (n_1, \ldots, n_k), n = \sum_{g=1}^k n_g, \vec{\theta} = (\theta_a, \ldots, \theta_k) \in \Theta^k$ and $\vec{y}^n$ be as above (2.4). We use '$\vec{Y}^n \sim P_{\vec{\theta}*}$' as an abbreviation for '$Y_1^{n_1} \sim P_{\theta_1^*}; \ldots; Y_k^{n_k} \sim P_{\theta_k^*}$'.

Theorem 1. *Let*

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

The random variable $S_{[\vec{n},\vec{\theta}^*]} := s(\vec{Y}^n; \vec{n}, \vec{\theta}^*)$ *is an* E-*variable, i.e. we have:*

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

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

$$\mathbf{E}_{\vec{Y}^n\sim P_{\theta^*}}[\log s(\vec{Y}^n; \vec{n}, \vec{\theta}^*)] \geq \mathbf{E}_{\vec{Y}^n\sim P_{\theta^*}}[\log s'(\vec{Y}^n)].$$

Proof of Theorem 1 The following fact plays a central role in the proof:

Fact For $g \in (1, ..., k)$, let $n_g \in \mathbf{N}, n := \sum_{g=1}^k n_g$ and let $u_g \in \mathbf{R}^+$. Suppose that $\sum_{g=1}^k n_g u_g \leq n$. Then $\prod_{g=1}^k u_g^{n_g} \leq 1$.

This result follows from the following standard generalization of Young's inequality to k numbers: for any k numbers $u_1, \ldots, u_k \in \mathbf{R}_0^+$ and any k nonnegative numbers $p_1, \ldots, p_k$ with $\sum_{g=1}^k p_g = 1$, we have $\prod_{g=1}^k u_g^{p_g} \leq \sum_{g=1}^k p_g u_g$. Applying this with $p_g = n_g/n$ to u_g and n_g as above, we get $\prod_{g=1}^k u_g^{n_g/n} \leq \sum_{g=1}^k (n_g u_g)/n \leq 1$, and the result follows by exponentiating to the power n.

Part 1 For $y \in \mathcal{Y}$, set set $p^\circ(y) := \sum_{g=1}^{k}(n_g/n)p_{\theta_g^*}(y)$ and $p^\circ(y^m) = \prod_{i=1}^{m} p^\circ(y_i)$. For all $\theta \in \Theta$ we have:

$$\mathbf{E}_{V^n \sim P_\theta}\left[s(V^n; \vec{n}, \vec{\theta}^*)\right] = \prod_{g=1}^{k} \mathbf{E}_{Y_g^{n_g} \sim P_\theta}\left[\frac{p_{\theta_g^*}(Y_g^{n_g})}{p^\circ(Y_g^{n_g})}\right] = \prod_{g=1}^{k}\left(\mathbf{E}_{Y \sim P_\theta}\left[\frac{p_{\theta_g^*}(Y)}{p^\circ(Y)}\right]\right)^{n_g}. \tag{8.3}$$

We also have

$$\sum_{g=1}^{k}\frac{n_g}{n}\mathbf{E}_{Y \sim P_\theta}\left[\frac{p_{\theta_g^*}(Y)}{p^\circ(Y)}\right] = \mathbf{E}_{Y \sim P_\theta}\left[\sum_{g=1}^{k}\frac{n_g}{n}\cdot\frac{p_{\theta_g^*}(Y)}{\sum_{g'=1}^{k}\frac{n_{g'}}{n}p_{\theta_{g'}^*}(Y)}\right] = 1. \tag{8.4}$$

The result now follows by combining (8.3) with (8.4) using the Fact further above.

Part 2 By convexity of $\{P_\theta : \theta \in \Theta\}$, there exists $\theta^\circ \in \Theta$ such that $p_{\theta^\circ} = \sum_{g=1}^{k}(n_g/n)p_{\theta_g^*}$ and then the numerator in (2.2) can we rewritten as $p_{\theta^\circ}(\vec{y})$. The GRO-property is now an immediate consequence of Proposition 4, Part 1.

Supporting Information

Appendix S1　Numerical approach to calculating E-variables for restricted $\mathcal{H}_1$

In this subsection we describe how we propose to approximate the beta prior and posterior on the restricted $\mathcal{H}_1$ with parameter space $\Theta(\delta)$, as defined in (4.1). Note that we limit ourselves to $\delta > 0$ in this detailed description; for $\delta < 0$ one can apply an entirely equivalent approach, with an extra term in the reparameterization. We define

$$
\zeta = \begin{cases} \delta \text{ if } d((\theta_a, \theta_b)) = \theta_b - \theta_a, \\ 0 \text{ if } d((\theta_a, \theta_b)) = \text{log-odds-ratio}(\theta_a, \theta_b), \end{cases}
$$

such that we have $\theta_a \in (0, 1 - \zeta)$ and in both cases, θ_b is completely determined by θ_a: $\theta_b = d^{-1}(\delta; \theta_a)$. Hence, our density estimation problem now becomes one-dimensional, which enables us to put a discretized prior on the restricted parameter space.

First, we discretize the parameter space Θ_a to a grid (a vector) with precision $K, K \in (0, 1 - \zeta)$ and $1/K \in \mathbb{N}^+$: $\bar{\boldsymbol{\theta}}_{\boldsymbol{a}} = (K, 2K, 3K, \ldots, 1 - \zeta)$. Then, we reparameterize $\theta_a = (1 - \zeta)\rho$, with $\rho \in (0, 1)$. Then, we have $\bar{\boldsymbol{\rho}} = (K/(1 - \zeta), 2K/(1 - \zeta), \ldots, 1)$. For the discretized grid $\bar{\boldsymbol{\rho}}$, we compute the prior $W = \text{Beta}(\alpha, \beta)$ densities and normalize them, which also gives us the discretized densities for each $\theta_a^i \in \bar{\boldsymbol{\theta}}_{\boldsymbol{a}}$ (with $i \in (1, 2, \ldots, 1/K)$):

$$
\pi_{\alpha, \beta, \zeta}(\theta_a^i) = \frac{\text{Beta}(\frac{\theta_a^i}{1 - \zeta}; \alpha, \beta)}{\sum_{k=1}^{\frac{1}{K}} \text{Beta}(\frac{\theta_a^k}{1 - \zeta}; \alpha, \beta)}.
$$

For all elements of $\bar{\boldsymbol{\theta}}_{\boldsymbol{a}}$, the corresponding θ_b is retrieved and the likelihood of incoming data points $p_{\theta_a, \theta_b}(Y^{(j-1)})$ is calculated. We can then estimate the posterior density of $\theta_a^i \in \bar{\boldsymbol{\theta}}_{\boldsymbol{a}}$:

$$
p(\theta_a^i | Y^{(j-1)}) = \frac{\pi_{\alpha, \beta, \zeta}(\theta_a^i) p_{\theta_a^i, \theta_b^i}(Y^{(j-1)})}{\sum_{k=1}^{\frac{1}{K}} \pi_{\alpha, \beta, \zeta}(\theta_a^k) p_{\theta_a^k, \theta_b^k}(Y^{(j-1)})}.
$$

We can then estimate $\breve{\theta}_a | Y^{(j-1)} = \mathbf{E}_{\theta_a \sim W | Y^{(j-1)}}[\theta_a]$ as $\sum_{i=1}^{\frac{1}{K}} p(\theta_a^i | Y^{(j-1)}) \theta_a^i$, and $\breve{\theta}_b | Y^{(j-1)} = d^{-1}(\delta; \theta_a | Y^{(j-1)})$.

Appendix S2   The Gunel-Dickey Bayes Factors do not give rise to E–variables

Sampling scheme	Fixed parameters	Bayes factor (10) for 2x2 table
Poisson	none	$\frac{8(n+1)(n_1+1)}{(n+4)(n+2)}\left[\frac{n_{a1}!n_{b1}!n_{a0}!n_{b0}!n!}{(n_1+1)!n_0!n_a!n_b!}\right]$
Joint multinomial	n	$\frac{6(n+1)(n_1+1)}{(n+3)(n+2)}\left[\frac{n_{a1}!n_{b1}!n_{a0}!n_{b0}!n!}{(n_1+1)!n_0!n_a!n_b!}\right]$
Independent multinomial	n_a, n_b	$\frac{\binom{n}{n_1}}{\binom{n_a}{n_{a1}}\binom{n_b}{n_{b1}}}\frac{(n+1)}{(n_a+1)(n_b+1)}$
Hypergeometric	n_a, n_b, n_1	$\frac{n_{a1}!n_{b1}!n_{a0}!n_{b0}!n!}{\prod_{i\in\{a,b,0,1\}}(n_i+\mathbb{I}_{n_i=min(n_a,n_b,n_0,n_1)})!}$

Table 1: Overview of (objective) Bayes factors for contingency table testing provided by Gunel and Dickey (1974) and Jamil et al. (2017).

We will not consider the hypergeometric and joint multinomial scenarios for this paper, where the number of successes n_1 is fixed, as they do not match the block-wise data design in this paper. The Bayes factor for the Poisson sampling scheme is not an E-variable, as the expectation under the null hypothesis with Poisson distributions on individual cell counts exceeds 1 for rates $\lambda \geq 1$:

$$\mathbb{E}_{n_{rc}\sim\text{Poisson}(\lambda_{rc})}\left[BF_{10}(N_{a1}, N_{b1}, N_{a0}, N_{b0})\right] =$$

$$\sum_{n_{a1}=0}^{\infty}\ldots\sum_{n_{b0}=0}^{\infty}\pi_{\lambda_{a1}}(n_{a1})\ldots\pi_{\lambda_{b0}}(n_{b0})BF_{10}(n_{a1}, n_{b1}, n_{a0}, n_{b0}) =$$

$$\frac{8}{\exp(\lambda_{a1}+\ldots+\lambda_{b0})}\sum_{n_{a1}=0}^{\infty}\ldots\sum_{n_{b0}=0}^{\infty}\lambda_{a1}^{n_{a1}}\ldots\lambda_{b0}^{n_{b0}}\frac{(n+1)(n_1+1)}{(n+4)(n+2)}\frac{n!}{(n_1+1)!n_0!n_a!n_b!},$$
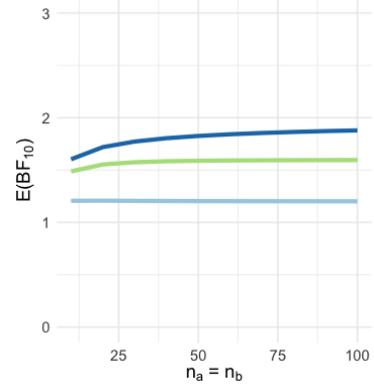
as illustrated numerically in Figure S2.1 for increasing limits for the sums $\sum_{n_{rc}=1}^{\max n_{rc}}$.
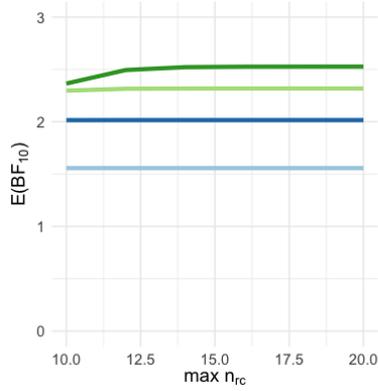
(a) The Gunel-Dickey Bayes factor for the Poisson sampling scheme is not an E-variable: $\sum_{n_{a1}=0}^{\max n_{rc}} \cdots \sum_{n_{b0}=0}^{\max n_{rc}} \pi_{\lambda_{a1}}(n_{a1}) \ldots \pi_{\lambda_{b0}}(n_{b0}) BF_{10}(n_{a1}, n_{b1}, n_{a0}, n_{b0})$ for various $\max n_{rc}$ and λ_{rc}.

(b) The Gunel-Dickey Bayes factor for the independent multinominal sampling scheme is not an E-variable: $\mathbb{E}_{N_{a1}, N_{b1} \sim \text{Binomial}(\theta)} \left[BF_{10}(N_{a1}, N_{b1} | n_a, n_b) \right]$ for various choices of θ and n_g.

Figure S2.1: GD

For the independent multinomial sampling scheme, let, without loss of generality, $n_a < n_b$. We get, with $n_0 = n - n_1$,

$$\mathbb{E}_{N_{a1}, N_{b1} \sim \text{Binomial}(\theta)} \left[BF_{10}(N_{a1}, N_{b1} | n_a, n_b) \right] =$$

$$\sum_{n_{a1}=0}^{n_a} \sum_{n_{b1}=0}^{n_b} \binom{n_a}{n_{a1}} \binom{n_b}{n_{b1}} \theta^{n_1} (1-\theta)^{n_0} \frac{\binom{n}{n_1}}{\binom{n_a}{n_{a1}} \binom{n_b}{n_{b1}}} \frac{(n+1)}{(n_a+1)(n_b+1)} =$$

$$\frac{(n+1)}{(n_a+1)(n_b+1)} \sum_{n_{a1}=0}^{n_a} \sum_{n_{b1}=0}^{n_b} \binom{n}{n_1} \theta^{n_1} (1-\theta)^{n_0}$$

Numerical simulations show that, for a range of choices for n, n_a and θ this exceeds 1; see Figure S2.1.

Appendix S3   Type-I error guarantee under optional stopping

Type-I Error   In Figure S3.1 type-I error rates of several E-variables and Fisher's exact test estimated through a simulation experiment are depicted. 2000 samples of length 1000 were drawn according to a Bernoulli(0.1) distribution to represent 1000 data streams in two groups. After each complete block $m \in \{1, \ldots, 1000\}$ an E-value or p-value was calculated and the proportion of rejected experiments up until m with each test type was recorded. As the stream lengths increase, the type-I error rate under (incorrectly applied) optional stopping with Fisher's exact test increases quickly. The type-I error rate of the E-variables remains bounded.
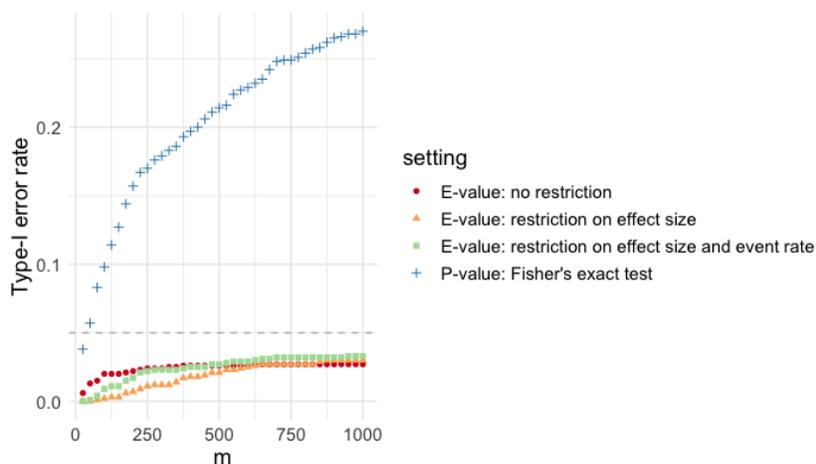

Figure S3.1: Type-I error rates for various E-variables and Fisher's exact test under optional stopping estimated with 1000 simulations of two Bernoulli(0.1) data streams of length 1000, with $n_a = n_b = 1$. Significance level $\alpha = 0.05$ was used (grey dashed line). For the safe tests, beta prior parameter values used were $\gamma = \alpha_a = \beta_a = \alpha_b = \beta_b = 1/2$ ($\gamma = 0.18$ gave comparable results). For the E-variables with restrictions on $\mathcal{H}_1$, we used $\delta = 0.05$ and $\theta_a = 0.1$.

References

Grünwald, P., de Heide, R., and Koolen, W. (2019). Safe testing. *arXiv preprint arXiv:1906.07801*.

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

Posner, E. (1975). Random coding strategies for minimum entropy. *IEEE Transactions on Information Theory*, 21(4):388–391.

Li, J. (1999). *Estimation of Mixture Models*. PhD thesis, Yale University, New Haven, CT.

Li, J. and Barron, A. (2000). Mixture density estimation. In Solla, S., Leen, T., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems*, volume 12, pages 279–285, Cambridge, MA. MIT Press.

Jamil, T., Ly, A., Morey, R. D., Love, J., Marsman, M., and Wagenmakers, E.-J. (2017). Default "Gunel and Dickey" Bayes factors for contingency tables. *Behavior Research Methods*, 49(2):638–652.