Learning under constraints: inference, selection and validation with maximum entropy models

Giuffrida, F.

CHAPTER 4

E-values for contingency tables

This chapter is based on:
Y. Wang, S. Arnold, F. Giuffrida, P. Grünwald, and T. Dickhaus. E-values for Contingency Tables. In preparation.

Abstract

We empirically investigate several e-values and e-processes for testing in 2×2 contingency tables. While the e-process recently introduced by Turner et al. exhibits theoretical growth-optimality properties for paired data sequences, we find that, when applied to a single large table (the *batch setting*), its performance deteriorates, with the other standard e-process based on universal inference performing even worse. By contrast, *conditional* e-variables tend to perform better, both in terms of growth rate and expected stopping times, with *uniformly most powerful* e-variables showing the best overall performance. When batches arrive sequentially, the comparison becomes more nuanced: Turner's e-variable is generally optimal in the asymptotic regime, whereas the *normalized maximum likelihood* e-variable is often, though not always, preferable at practically relevant sample sizes.

4.1 Introduction

Over the last few years, e-variables have attracted increasing interest in the statistical community as new, safe methods for assessing statistical evidence; see, e.g., [24, 23, 22]. In a nutshell, an *e-variable*, also *e-statistic* or *e-value*, is a non-negative random variable S with expectation at most one under the null, and large realizations of e-variables give evidence against the null. The latter is easily seen, since, for every distribution P_0 in the null hypothesis, we have $P_0(S \geqslant 1/\alpha) \leqslant \alpha$, for any $\alpha \in (0, 1)$, as a direct consequence of Markov's inequality. E-variables allow for a monetary interpretation as bets against the null [27], and may be combined by taking averages or products. This simple behavior under combinations makes them especially attractive for real-world problems where we want to combine evidence across different studies or outcomes, and accumulate statistical evidence sequentially over time. For a comprehensive introduction to the theory of e-values, we refer to the recent monograph by [23].

In this paper, we study and propose various e-variables and e-processes for analyzing 2×2 contingency tables, a problem that dates back as far as the field of statistics itself. Such a table is defined by two group sizes, n_a and n_b, counts $Y_g \in \{0, \ldots, n_g\}$ for $g \in \{a, b\}$, where we can write $Y_g = \sum_{i=1}^{n_g} X_i^g$. According to the null hypothesis, the outcomes in both groups are i.i.d. Bernoulli with the same parameter $P(X_i^g = 1) = \theta$, where θ is an unknown nuisance parameter, making the null composite. According to the alternative, the parameter varies across the groups: outcomes within each group g are i.i.d. Bernoulli with parameter θ_g. Surprisingly, despite e-values being a rapidly growing and developing research field, much remains unknown about the design of optimal e-variables and e-processes for this simple yet fundamental problem. At first, this statement may seem to contradict the findings of [85, 30], who have derived the GRO e-variable for a given contingency table and simple given alternative (θ_a, θ_b), and then use it to derive an e-process for data arriving sequentially in small batches for composite alternatives; for the case that data arrive in pairs, they provide encouraging empirical results. Here, GRO means 'growth-optimal,' the standard and best-researched optimality criterion for e-variables. However, when one attempts to use the *Turner e-variable* for analysis of a particular table taken from [86], the results, which we reproduce in Table 4.1 below, are very disappointing. The crucial differences in the setting in which Turner performed well were that the table was *large*, with *highly different* group sizes, and that it was observed *one shot*, rather than as a sequence of small batches, such as pairs. The 43 other tables considered by [86] yielded qualitatively similar results to those in Table 4.1; for brevity, we omit them.

In this example, the p-value from Fisher's exact test is 0.00282, strongly suggesting that the outcomes depend on group membership. Ideally, we want to draw the same conclusion if we base our analysis on e-values. However, a naive calculation of Turner's e-variable for the same data gives an uninformative 1 (one), whereas the non-naive mini-batching approach suggested by [30] (and explained in detail in the next sections) yields a non-conclusive small value of 4.11 (usually we perceive e-values larger than 10 or 20 as conclusive). Upon closer analysis, we find that the problem is that the Turner approach to deal with composite alternatives, although quite a standard one,

	a	b
1	192	674
0	506	2370

Table 4.1: Data provided by [86] for $n_a = 698$ participants in group a, and $n_b = 3044$ participants in group b. The groups a and b indicate a certain difference in the genotype of the participants, and the outcomes indicate whether some studied phenotype was present (1) or absent (0).

Method	E-value
Turner (naive)	1
Turner (split)	4.11
UI (mixture)	0.0687
Fisher f-calibrated	10.087
Fisher g-calibrated	17.83
Conditional (mixture)	7.83
Conditional (UMP-2,α =0.05)	37.33
Conditional (UMP-2,α =0.01)	42.92
Conditional (NML)	2.30

Table 4.2: Different e-values for the data in Table 4.1.

does not combine well at all with the highly skewed group sizes (group b is more than four times larger than group a) of the given table, as we will explain in Section 4.3. Besides Turner's e-value, there are two further well-known available approaches to compute e-variables for the given problem: The *universal inference (UI) e-variable* proposed by [25], and the e-variable obtained by calibrating Fisher's exact p-value into an e-value by a so-called *p-to-e-calibrator*, which maps any p-value to an e-value, see [26]. The UI e-variable is easy to compute and very broadly applicable, but also known to be overly conservative, as it also turns out to be the case in our example, where we obtain a useless value of 0.0687 (see Table 4.2). There exist many valid p-to-e calibrators, and there are no guidelines as to which one to use in which situation, so we consider just two standard ones: the functions $g(p) := 1/\sqrt{p} - 1$ and $f(p) = (1 - p + p \log p)/(p(-\log p)^2$. $g(p)$ turns Fisher's exact $p = 0.00282$ into an e-value of 17.83, which seems quite reasonable. However, $g(p)$ has the property that it could vanish to 0 with non-zero probability, and as such, we do not advocate its use. To see why, recall that a central motivation for using e-values instead of p-values lies in their behaviors for sequential problems where we can multiply e-values over time to accumulate evidence from various outcomes and studies. Suppose, for example, that the study underlying Table 4.1 was carried out in several stages, each leading to a batch of data summarized by its own sub-table, and that the decision to continue with the study would have depended on the strength of evidence in the batches seen so far. Then it would not have been allowed to reuse all data at the final step to compute Fisher's exact p-value based on the entire table, since part of the data had

already been used for the intermediate inference. In contrast, providing a final e-value as the product of the e-values for all individual batches and rejecting if it exceeds the significance level α does provide a valid test. However, if one of the batches had been small, then with non-negligible probability the calibrator $g(p)$ would have given us an e-variable of 0 for that batch. Then the final product would have become 0 as well, rendering g useless (or at best: highly risky) in such a standard optional continuation scenario.

For that reason, we prefer the other standard calibrator f (which cannot reach 0), which provides an e-value of $f(0.00282) = 10.087$. The value of 10.087 serves as a benchmark in this example: for at least two reasons, we aim to find better (more powerful) e-values than the one obtained by calibration. The first reason is that the calibration operation is 'blind' and does not make use of the underlying setting — one would hope that knowledge of the setting/model should allow one to design improved e-variables. The second reason is that, while useful, we can predict it will be far from optimal in the optional continuation scenario described above. There, one can, of course, compute the p-values $p_1, \ldots, p_T$ for each batch, calibrate each into an e-value $f(p_1), \ldots, f(p_T)$, and report their product as the total summary statistic. However, such an approach is undesirable, since we do not include any potentially valuable information from the first $j-1$ studies in the computation of an e-value for the j-th study. This stands in contrast to the approach proposed by [85], which, as is possible for most e-variables designed for specific problems, incorporates implicit *learning* of the parameters, allowing us to leverage valuable information from past batches in the e-variable designed for the present one. However, as we already remarked, Turner's method does not yield satisfactory results in the table above either.

4.1.1 Novel methods considered in this paper

Here, we provide *conditional* e-values for analyzing contingency tables that are easy to implement and powerful. By 'powerful', we mean that the e-values should (with high probability) be larger than the calibrated p-value based on Fisher's exact test if the alternative hypothesis of association is true. By 'easy to compute,' we mean that resource-intensive operations, such as a loop over all contingency tables with given marginal counts, should be avoided, a requirement that applies when many e-values must be computed simultaneously, as in genome-wide association studies. Recently, [28] and [29] have studied optimal e-values in situations where the null and alternative hypotheses are given by exponential families sharing the same sufficient statistic(s). Recall that, by definition, conditioned on the sufficient statistic, the distribution of the data does not depend on the unknown parameter of interest anymore; that is, all relevant information about the parameter is contained in the sufficient statistic. The authors use this characterization of sufficiency to study *conditional e-values*, in which they condition on the sufficient statistic to convert the composite null to a simple hypothesis. Conditional e-values are attractive alternatives to GRO e-values since they are easy to compute and asymptotically optimal in the sense that the difference in expected growth in comparison to the theoretically optimal GRO e-variable grows sub-linearly (logarithmically). This paper builds on the fact that the null hypothesis

of no association may be represented as a one-dimensional exponential family with the total number of successes as a sufficient statistic, and that, conditioned on the number of ones, the null becomes simple.

Conditional e-variables are generalizations of conditional likelihood ratios; specifically, instantiated to our contingency table model, they generalize the conditional likelihood ratio for sequences of paired data as used already by Wald in his sequential probability ratio test for two proportions [87] and the conditional Bayes factors used by, for example, [88]. Indeed, what we call the conditional (mixture) e-variable in Table 4.2 is just such a conditional Bayes factor (though we extend it in a non-standard way to the sequential setting). The other versions of our conditional e-variables have, to the best of our knowledge, never been considered before. An initial study of a similar type of conditional e-variable, slightly different from the ones we consider here but that served as an inspiration for the present one, was applied, under the name *microcanonical e-variable* in Chapter 3.

Table 4.2 presents the conditional e-values in three versions, as discussed in what follows, for the data example in comparison with the previously discussed e-values. We see that one particular version, *UMP*-2, performs remarkably well on our table. UMP-2 stands for 'the 2-sided extension of the uniformly most powerful conditional Bayes factor', an e-variable construction that will be explained in the next section. We have calculated the e-variable types listed in Table 4.2 on various other tables as well, and consistently find that UMP-2 shows the best behavior, which is in accord with the heuristic analysis we give in Section 4.3. Thus, we conclude that it is the method of choice for the one-shot scenario, where we receive a single table 'thrown at us' and must analyze it.

Our findings are more intricate in the sequential setting, where batches (tables, potentially of different sizes and group size ratios) from the same source arrive sequentially. In this setting, by its very construction, UMP-2 will suffer from the same defect as the p-to-e calibrators: when data comes in batches, it does not allow one to improve the e-variable for the next batch based on previous batches. The NML conditional e-variables do allow for this behavior, and soon start to outperform UMP as the number of batches increases. Since they are harder to implement in the sequential setting, in this preliminary study, we have not experimented with the mixture conditional e-variables; we expect them to behave very similarly to the NML ones. At the same time, though, we find that the Turner e-variables generally perform better as the number of batches increases — like the NML and mixture e-variables, they adapt to the data-generating distribution as more data comes in. In a nutshell, we find that *asymptotically*, in all our experiments, the Turner e-variable outperforms the conditional NML one; but for reasonable sample sizes and reasonable targets (i.e., rejection thresholds $1/\alpha$), the NML one either is substantially better, or it is slightly (but never a lot) worse.

Thus, based on this preliminary study, our final recommendation for the case of sequentially arriving batches will be to handle the first batch (if it is not too small) using conditional UMP, and subsequent batches using the conditional NML approach, and return as evidence the product of all these e-variables. We hasten to add that the preliminary status of these findings means they should be interpreted cautiously: since

the results are not entirely clear-cut, with the Turner e-variable sometimes slightly outperforming the NML one, we aim to perform several further experiments before we submit the present manuscript to a journal.

We also want to stress that we focus on the case where (a) under the null, $\theta_a = \theta_b$, and (b) the alternative is the entire set of $(\theta_a, \theta_b) \in (0,1)^2$ with $\theta_a \neq \theta_b$ — thus there is no minimum relevant effect size, or a priori knowledge that the discrepancy between θ_a and θ_b has a certain minimum size. This is in contrast to [85], who extend the Turner e-variables to deal with nulls in which $\theta_a \neq \theta_b$ and minimum effect sizes; one can certainly extend the conditional e-variables to this setting as well, but we do not consider these extensions in this paper.

The remainder of this paper is structured as follows: in section 4.2, we introduce notation and all e-values of interest in a rigorous mathematical manner. In Section 4.3, we explain the results underlying Table 4.2, and more generally, the one-shot case, in detail. In Section 4.4, we compare the performance of the proposed e-values in simulations when batches (sub-tables) arrive sequentially. We end with a conclusion.

4.2 Preliminaries and notation

We consider binary data from two independent groups, denoted by a and b, and suppose that, in each stream, data are i.i.d. Bernoulli with parameters θ_a and θ_b, respectively, i.e.

$$X_1^a, X_2^a, \ldots \overset{\text{i.i.d.}}{\sim} \text{Ber}(\theta_a) \quad \text{and} \quad X_1^b, X_2^b, \ldots \overset{\text{i.i.d.}}{\sim} \text{Ber}(\theta_b), \tag{4.1}$$

for $(\theta_a, \theta_b) \in \Theta = (0,1)^2$. We are interested in testing the null hypothesis of no association

$$\mathcal{P} : \theta_a = \theta_b = \theta_0, \quad \theta_0 \in (0,1). \tag{4.2}$$

We abbreviate the Bernoulli distribution with parameter θ_a by P_{θ_a}, and let $P_{(\theta_a,\theta_b)}$ be the distribution of a bivariate Bernoulli vector with independent components and parameters θ_a, θ_b. We denote by p_{θ_a} and $p_{(\theta_a,\theta_b)}$ the corresponding densities, and use the same notation for multiple outcomes, naturally extended by the i.i.d. assumption. We write the null hypothesis at (4.2) equivalently as $\mathcal{P} = \{P_{(\theta_0,\theta_0)} \mid \theta_0 \in (0,1)\}$.

4.2.1 E-variables for a single batch/table

Recall that an *e-variable* for $\mathcal{P}$ is a non-negative test statistic S such that $\mathbb{E}_P(S) \leqslant 1$, for all $P \in \mathcal{P}$. In the following, we will introduce various e-variables for the testing problem at hand. For the sake of clarity, we assume, first, that we have a single batch of data available and explain in the next subsection how to derive e-processes and compute e-values if data arrives in batches over time.

Assume then that we observe a batch of data consisting of observations $X_1^k, \ldots, X_{n_k}^k \in \{0,1\}$, for $n_k \geqslant 1$ and $k = a, b$. We denote by $Y_a = \sum_{i=1}^{n_a} X_i^a$ and $Y_b = \sum_{i=1}^{n_b} X_i^b$ the number of ones (successes) in group a and b, respectively, and let $n = n_a + n_b$, and $N_1 = Y_a + Y_b$.

The Turner E-Variables For arbitrary fixed $(\theta_a, \theta_b) \in \Theta$, [30] show that

$$S_{\mathrm{GRO}(\theta_a,\theta_b)} = \frac{p_{\theta_a}((X_i^a)_{i=1}^{n_a})}{p_{\theta_0}((X_i^a)_{i=1}^{n_a})} \cdot \frac{p_{\theta_b}((X_i^b)_{i=1}^{n_b})}{p_{\theta_0}((X_i^b)_{i=1}^{n_b})}, \quad \text{for } \theta_0 = \frac{n_a}{n}\theta_a + \frac{n_b}{n}\theta_b, \tag{4.3}$$

is the GRO e-variable with respect to the simple alternative $Q = \{P_{(\theta_a,\theta_b)}\}$, see [30, Theorem 1]. Note that

$$S_{\mathrm{GRO}(\theta_a,\theta_b)} = s_{\mathrm{GRO}(\theta_a,\theta_b)}(Y_a, Y_b) \tag{4.4}$$

for some fixed function s, i.e. the e-value depends on the original outcomes only through their counts. However, often we do not want to test $\mathcal{P}$ against the simple alternative $Q = \{P_{(\theta_a,\theta_b)}\}$ but rather against the full alternative

$$\mathcal{Q} = \{P_{(\theta_a,\theta_b)} \mid (\theta_a, \theta_b) \in \Theta, \theta_a \neq \theta_b\}. \tag{4.5}$$

For this case, [30] propose to consider a prior W over Θ under which θ_a and θ_b are independent. Thus, W decomposes into sub-priors W_a, W_b such that $\theta_a \sim W_a$ and $\theta_b \sim W_b$ are independent. We can then write $P_{W_g} = \int P_{\theta'_g} \, \mathrm{d}W_g(\theta'_g)$ for $g \in \{a, b\}$. By convexity of the Bernoulli model, we have $P_{W_g} = P_{\bar{\theta}_g}$ for some $\bar{\theta}_g$, for $g \in \{a, b\}$. [30] then calculate the GRO e-variable relative to this simple alternative $(\bar{\theta}_a, \bar{\theta}_b)$. We denote the corresponding *Turner e-variable* by $S_{\mathrm{TUR}(W)}$. In case $n_a = n_b = 1$, this coincides with the GRO e-variable relative to W, but in all other cases, as long as W is non-degenerate, it does not, since if $n_g > 1$ then $P_{W_g}(X_1^g, \ldots, X_{n_g}^{n_g}) \neq P_{\bar{\theta}_g}(X_1^g, \ldots, X_{n_g}^{n_g})$; see [30] for more extensive discussion. [30] suggests using a Beta prior for W, and we use the uniform Beta prior unless otherwise stated.

Universal Inference In the given problem, a standard version of the *Universal Inference (UI) e-variable* proposed by [25] with respect to the fixed alternative $\{P_{(\theta_a,\theta_b)}\}$ is given by

$$S_{\mathrm{UI}(\theta_a,\theta_b)} = \frac{\theta_a^{Y_a}(1-\theta_a)^{n_a - Y_a}\theta_b^{Y_b}(1-\theta_b)^{n_b - Y_b}}{\hat{\theta}^{N_1}(1-\hat{\theta})^{N_1 - y}},$$

where $\hat{\theta}$ denotes the MLE for the given data with respect to the binomial distribution with sample size n_1. For the general alternative $\mathcal{Q}$ and general prior W (in this case, θ_a and θ_b do not need to be independent under W), we report the mixture e-value $S_{\mathrm{UI}(W_1)} = \int S_{\mathrm{UI}(\theta'_a,\theta'_b)} \, \mathrm{d}W(\theta'_a, \theta'_b)$.

Conditional E-Variables As mentioned in the introduction, the Turner e-variable as well as the UI e-variable may perform poorly (i.e., remain small) for some tables that intuitively provide strong evidence against the null. Recently, [28] and [29] have introduced *conditional e-values* as an attractive alternative to the theoretically optimal GRO e-value for general exponential families. Clearly, the hypothesis $\mathcal{P}$ at (4.2) forms a one-dimensional exponential family with parameter θ_0, and the hypothesis $\mathcal{Q}$ at (4.5) forms a two-dimensional exponential family with parameters (θ_a, θ_b). For the given data, the summary statistic $N_1 = Y_a + Y_b$ is sufficient for $\mathcal{P}$, and the vector

(N_1, Y_a) is sufficient for $\mathcal{Q}$. Moreover, under the null, N_1 is Binomially distributed with parameters n and θ_0, and, under the alternative, Y_k is binomially distributed with parameters n_k and θ_k, for $k = a, b$. We denote the probability mass functions of the sufficient statistic Y_a by f_{Y_a}, and refer to [66] for an extensive study of exponential families. Next, we will condition on $N_1 = n_1$, for some $n_1 \leqslant n_a + n_b$. It is well-known, and may be verified by standard calculation, that, given $N_1 = n_1$, Y_a follows a hypergeometric distribution under the null, that is

$$f_{Y_a}(y \mid N_1 = n_1) = \frac{\binom{n_a}{y}\binom{n_b}{n_1 - y}}{\binom{n}{n_1}}, \quad y \in \{\gamma^-, \dots, \gamma^+\},$$

for $\gamma^- = \max\{0, n_1 - n_b\}$ and $\gamma^+ = \min\{n_1, n_a\}$. Crucially, the density $f_{Y_a}(\cdot \mid N_1 = n_1)$ is independent of θ_0 by sufficiency. Similarly, for general $(\theta_a, \theta_b) \in \Theta$, Y_a, given $N_1 = n_1$, follows a *Fisher's noncentral hypergeometric distribution*, that is

$$f_{Y_a}(y \mid N_1 = n_1) = \frac{\binom{n_a}{y}\binom{n_b}{n_1 - y} e^{\delta y}}{\sum_{k=\gamma^-}^{\gamma^+} \binom{n_a}{k}\binom{n_b}{n_1 - k} e^{\delta k}}, \quad y \in \{\gamma^-, \dots, \gamma^+\}, \tag{4.6}$$

where δ denotes the log odds-ratio,

$$\delta = \log\left(\frac{\theta_a}{1 - \theta_a} \cdot \frac{1 - \theta_b}{\theta_b}\right) \in \mathbb{R}. \tag{4.7}$$

Note that the hypergeometric distribution results as a special case for $\delta = 0$, which holds if, and only if, $\theta_a = \theta_b$. For fixed $n_a, n_b \geqslant 1$, $n_1 \leqslant n_a + n_b$, and $\delta \in \mathbb{R}$, we denote the density of Fisher's non-central hypergeometric distribution with parameters n_a, n_b, n_1 and δ, by $f_{\mathrm{FN}(n_a, n_b, n_1, \delta)}$. Since n_a and n_b, as well as the number of ones n_1 (by conditioning) are assumed to be given, we abbreviate $f_{\mathrm{FN}(n_a, n_b, n_1, \delta)}$ simply by $f_{\mathrm{FN}(\delta)}$ whenever n_a, n_b, and n_1 are clear from the context. As seen from (4.6), for fixed n_a, n_b, n_1, the set of distributions corresponding to $f_{\mathrm{FN}(\delta)}$ themselves constitute a 1-dimensional exponential family with canonical parameter δ. The *conditional e-value* is defined as

$$S_{\mathrm{c}(\delta)}(Y_a) = \frac{f_{\mathrm{FN}(n_a, n_b, n_1, \delta)}(Y_a)}{f_{\mathrm{FN}(n_a, n_b, n_1, 0)}(Y_a)} = \frac{f_{\mathrm{FN}(\delta)}(Y_a)}{f_{\mathrm{FN}(0)}(Y_a)}, \quad \delta \in \mathbb{R}. \tag{4.8}$$

Lemma 4.2.1.
For any $\delta \in \mathbb{R}$, the conditional e-value $S_{\mathrm{c}(\delta)}$ at (4.8) is an e-value for $\mathcal{P}$.

Proof. For any $P \in \mathcal{P}$, Y_a, given $N_1 = n_1$, is hypergeometrically distributed, and thus

$$\mathbb{E}_P(S_{\mathrm{c}(\delta)} \mid N_1 = n_1) = \sum_{y=\gamma^-}^{\gamma^+} f_{\mathrm{FN}(0)}(y) S_{\mathrm{c}(\delta)}(y) = \sum_{y=\gamma^-}^{\gamma^+} f_{\mathrm{FN}(\delta)}(y) = 1.$$

The claim follows from the tower property and the fact that $S_{\mathrm{c}(\delta)}$ is non-negative.

By conditioning, we have reduced the dimensionality of the alternative parameter space, and $S_{\mathrm{c}(\delta)}$ is an e-value for any $\delta \in \mathbb{R}$. Analogously to the unconditional case, we usually do not know a priori against which particular δ we should test $\mathcal{P}$. Thus, we discuss three possible choices in the sequel.

1. *Mixture conditional e-variable.* Given a prior W on $\mathbb{R}$, it is natural to consider the *mixture conditional e-variable*

$$S_{\mathrm{c}(W)}(Y_a) = \int S_{\mathrm{c}(\delta)}(Y_a)\,\mathrm{d}W(\delta). \tag{4.9}$$

   Note that this e-variable can be thought of as a 'conditional' Bayes factor, and as such it has appeared before in the Bayesian literature: [88] and [89] give closed form expressions for the conjugate prior on δ for Fisher's non-central hypergeometric distribution. Other natural choices for W include some Gaussian prior or Jeffreys' prior.

2. *NML conditional e-variable.* Ideally, we would like to choose δ such that the enumerator in (4.8) is as large as possible, which is achieved by the MLE, denoted by $\hat{\delta}(Y_a)$, for the given data. However, $S_{\mathrm{c}(\hat{\delta}(Y_a))}$ is not an e-variable in general, since usually $\mathbb{E}_P(S_{\mathrm{c}(\hat{\delta}(Y_a))}) \gg 1$ for $P \in \mathcal{P}$. Nevertheless, by appropriately scaling, we obtain the *normalized maximum likelihood (NML)* conditional e-variable

$$S_{\mathrm{c(NML)}}(Y_a) = S_{\mathrm{c}(\hat{\delta}(Y_a))}(Y_a) \cdot \frac{1}{\sum_{k=\gamma^-}^{\gamma^+} f_{\mathrm{FN}(\hat{\delta}(k))}(k)}, \tag{4.10}$$

   where $\hat{\delta}(k)$ denotes the MLE for $f_{\mathrm{FN}(n_a,n_b,n_1,\delta)}$ with outcome $k \in \{\gamma^-, \ldots, \gamma^+\}$. The normalized maximum likelihood approach has been derived within the *Minimum Description Length (MDL)* approach to statistical model selection and prediction [18, 17] and often combined with some *luckiness function*, where we replace $\hat{\delta}(k)$ and $\hat{\delta}(Y_a)$, with the MLEs with respect to the modified data where $n'_a = n_a + k_a, n'_b = n_b + k_b$, $n'_1 = n_1 + k_1$, for some $k_a, k_b, k_1 \geqslant 0$ with the default choice $k_a = k_b = k_1 = 1$. Classic results in MDL theory for unconditional settings with simple nulls indicate that, for large enough sample sizes, and under regularity conditions, which are met in our case, the log Bayes marginal likelihood based on Jeffreys' prior will become equal, up to $o(1)$, to the log NML e-variable. Although these classic results do not cover the present conditional setting, they suggest that NML and Bayesian methods will behave similarly.

3. *UMP e-variable.* Finally, since $f_{\mathrm{FN}(\delta)}(\cdot)$ constitutes a 1-dimensional exponential family, we can use a result of [90] which states that, for every pre-specified significance level $\alpha \in (0,1)$, among all tests of our usual null $\delta = 0$ against the *one-sided* alternative $H_1 : \delta > 0$ that take the form $\mathbf{1}\{S_{\mathrm{c}(\delta)} \geqslant 1/\alpha\}$ for some $\delta \in \mathbb{R}$, the choice $\delta = \delta^+_{\mathrm{UMP}}$ is *uniformly most powerful*, where $\delta^+_{\mathrm{UMP}} > 0$ is given by

$$D_{\mathrm{KL}}\Big(f_{\mathrm{FN}(\delta^+_{\mathrm{UMP}})} \,\Big\|\, f_{\mathrm{FN}(0)}\Big) = \log(1/\alpha), \tag{4.11}$$

   and the results hold as long as $\delta^+_{\mathrm{UMP}} > 0$ satisfying the above display exists; it will then also be unique. Existence is guaranteed if $\min\{n_a, n_b\}$ is larger

than some minimal n^* [90, 91]. Even though there is no closed-form expression for δ^+_{UMP} available, we may solve (4.11) efficiently numerically by convexity in δ. Similarly, if (4.11) holds with δ^+_{UMP} replaced by some δ^-_{UMP}, then the test $\mathbf{1}\{S_{\mathrm{c}(\delta^-_{\mathrm{UMP}})} \geqslant 1/\alpha\}$ is UMP among all tests of above form against alternative $H_1 : \delta < 0$. If one is interested, like we are (since we always consider *every* distribution $(\theta_a, \theta_b) \in (0,1)^2$ with $\theta_a \neq \theta_b$ as potential alternative), in a two-sided test, this suggests to take, in analogy to the standard construction of a two-sided test, the e-variable $S_{\mathrm{UMP}-2} = (1/2)S_{\mathrm{c}(\delta^-_{\mathrm{UMP}})} + (1/2)S_{\mathrm{c}(\delta^+_{\mathrm{UMP}})}$. We call this the UMP-2 e-variable, even though it formally does not have UMP status anymore.

4.2.2 Sequential (conditional) e-variables

We now assume that data arrives in batches $(B_t)_{t\in\mathbb{N}}$ over time, and that, at each time point $t \in \mathbb{N}$, we observe batch B_t consisting of $n_{a,t}$ observations in group a and $n_{b,t}$ observations in group b. We denote by $Y_{a,t}$ and $Y_{b,t}$ the number of ones at $t \in \mathbb{N}$ in group a and b, respectively, and let $n_t = n_{a,t} + n_{b,t}$, and $N_{1,t} = Y_{a,t} + Y_{b,t}$. Similarly, we let $n^t_k = \sum_{\ell=1}^t n_{k,\ell}$, and $Y^t_k = \sum_{\ell=1}^t Y_{k,\ell}$, for $k = a, b$, and $n^t = n^t_a + n^t_b$, and $N^t_1 = Y^t_a + Y^t_b$. For $t \in \mathbb{N}$, we summarize the data B_t *at* t, and the total data $B^t = (B_1, \ldots, B_t)$ *until* t in tables as given in Tables 4.3 and 4.4.

outcome	a	b	
1	$Y_{a,t}$	$Y_{b,t}$	$n_{1,t}$
0	$n_{a,t} - Y_{a,t}$	$n_{b,t} - Y_{b,t}$	$n_{0,t}$
	$n_{a,t}$	$n_{b,t}$	n_t

Table 4.3: Data B_t at $t \in \mathbb{N}$.

outcome	a	b	
1	Y^t_a	Y^t_b	n^t_1
0	$n^t_a - Y^t_a$	$n^t_b - Y^t_b$	n^t_0
	n^t_a	n^t_b	n^t

Table 4.4: Data B^t until $t \in \mathbb{N}$.

Throughout the paper, we consider the natural filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$, where $\mathcal{F}_t = \sigma(B^t)$ for $t \in \mathbb{N}$. It was argued in the introduction that e-values are particularly useful in sequential analysis where we want to combine evidence from different studies (tables, batches) over time. We call a sequence $(S_t)_{t\in\mathbb{N}}$ of e-variables for $\mathcal{P}$ *sequential* (with respect to the filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$), if $(S_t)_{t\in\mathbb{N}}$ is adapted and, for all $P \in \mathcal{P}$,

$$\mathbb{E}_P(S_t \mid \mathcal{F}_{t-1}) \leqslant 1, \quad t \in \mathbb{N}. \tag{4.12}$$

Given some sequence of sequential e-values $S_1, S_2, \ldots$, we consider their cumulative products

$$M_t = \prod_{\ell=1}^t S_\ell, \quad t \in \mathbb{N}. \tag{4.13}$$

It follows directly that $(M_t)_{t\in\mathbb{N}}$ is a nonnegative super martingale under any $P \in \mathcal{P}$, and that it satisfies thus *Ville's inequality* [92], claiming that, for any $P \in \mathcal{P}$,

$$P(\exists t : M_t \geqslant 1/\alpha) \leqslant \alpha, \quad \text{for all } \alpha \in (0,1). \tag{4.14}$$

Ville's inequality lays the foundation for safe anytime-valid inference [22]. It may be seen as a time-uniform generalization of Markov's inequality, and is stronger than $P(M_t \geqslant 1/\alpha) \leqslant \alpha$ for all $t \in \mathbb{N}$, see e.g. [93] for a discussion. Any sequence of sequential e-values thus suggests the natural sequential tests $\Psi_t = \mathbf{1}\{M_t \geqslant 1/\alpha\}$, $t \in \mathbb{N}$, where we reject the null $\mathcal{P}$ at level $\alpha \in (0,1)$ as soon as M_t exceeds the threshold $1/\alpha$, and, by (4.14), one may upper bound the probability of ever doing a type-I error by α.

Sequential Turner - Simple Alternative For the simple alternative (θ_a, θ_b), the GRO e-variable for each block is given by (4.3), which takes the form of a likelihood ratio between the alternative and a particular element of the null. It then easily follows [94] that the GRO e-variable for any sequence of blocks is given by the product of the individual e-variables (4.3) for each block, i.e. $M_t = \prod_{\ell=1}^{t} s(Y_{a,\ell}, Y_{b,\ell})$, with s given by (4.4), is the GRO e-variable for data B^t.

Sequential Turner - Composite Alternative For the full composite alternative, we may turn the Turner e-variable into a sequential sequence of e-variables as follows: we start with a prior W_1 for the first batch and calculate $S_{\mathrm{TUR}(W_1)}$. We then, for the t-th batch, use the Turner e-variable with prior W_t set equal to the posterior $W_1 \mid B_1, \ldots, B_{t-1}$ based on all previously observed data.

We note a few properties of this procedure:

1. *No Within-Batch Learning.* The posterior is updated only after a new batch arrives. Therefore, if batch B_t has large size, the Turner e-value for that batch is a likelihood ratio in which all outcomes in group a within that batch are predicted by the same Bernoulli distribution, i.e., for $g \in \{a, b\}$, every such outcome x_i^g appearing in B_t appears as a factor $p_{\bar{\theta}}(x_i^g)$ in the likelihood for the same $\bar{\theta}$ that only depends on outcomes in past batches $B_1 \ldots, B_{t-1}$.

2. *Sequential GRO in Special Cases.* If all batches B_t are pairs, $n_{a,t} = n_{b,t} = 1$, and the prior W_1 renders θ_a and θ_b independent, then so does, for each t, the posterior W_t that is used as prior in batch B_t. As a result, following the reasoning of the previous section, the t-th sequential Turner e-variable $S_{\mathrm{TUR}(W_t)}$ is the W_t-GRO e-variable and the resulting martingale $(M_t)_{t\in\mathbb{N}}$ is *sequential* W_1-*GRO*. [29] shows that, for general exponential families, the growth rate of this sequential version of GRO differs from the ('overall' optimal) GRO by $c \log n$ for some small ($< 1/2$) positive constant c.

Sequential conditional We suggest using *sequential conditional e-values* defined as

$$S_{\mathrm{c}(\delta_t),t}(Y_{a,t}) = \frac{f_{\mathrm{FN}(n_{a,t}, n_{a,t}, N_{1,t}, \delta_t)}(Y_{a,t})}{f_{\mathrm{FN}(n_{a,t}, n_{b,t}, N_{1,t}, 0)}(Y_{a,t})}, \quad t \in \mathbb{N}, \delta_t \in \mathbb{R}, \tag{4.15}$$

for some predictable sequence $(\delta_t)_{t\in\mathbb{N}} \subseteq \mathbb{R}$. Importantly, the conditional requirement at (4.12) allows us to learn the true log-odds ratio δ adaptively, and thereby increasing

power in case that the alternative is true. For the mixture approach, this suggests considering

$$S_{\mathrm{c}(W_t),t} = \int S_{\mathrm{c}(\delta),t}(Y_{a,t})\,\mathrm{d}W_t(\delta), \tag{4.16}$$

for some predictable sequence $(W_t)_{t\in\mathbb{N}}$ of priors over $\mathbb{R}$, where the natural choice is to choose W_t as the posterior for δ after having seen the data B^{t-1}. Note that, while the individual e-variables have a conditional Bayes factor interpretation, this is not so clear for their product: the conditional Bayes factor for data consisting of a series of batches $B_1, \ldots, B_T$ is not equal to the product of conditional Bayes factors for each separate batch B_t with prior W_t set to the posterior based on B^{t-1} (such a coherence results only holds for *unconditional* Bayes factors). Also, for the NML approach, we want to predictably update the MLE, which suggests using

$$S_{\mathrm{c(NML)},t} = S_{\mathrm{c}(\hat{\delta}(Y_a^t)),t}(Y_{a,t}) \cdot \frac{1}{\sum_{k=\gamma_t^-}^{\gamma_t^+} f_{\mathrm{FN}(n_{a,t},n_{b,t},N_{1,t},\hat{\delta}(k))}(k)}, \tag{4.17}$$

where $\gamma_t^- = \max\{0, N_{1,t} - n_{b,t}\}$ and $\gamma_t^+ = \min\{N_{1,t}, n_{a,t}\}$, and $\hat{\delta}(Y_a^t)$, and $\hat{\delta}(k)$ denote the MLE for the full data B^t with respect to $f_{\mathrm{FN}(n_a^t,n_b^t,N_1^t,\delta)}$ for Y_a^t and $Y_a^{t-1} + k$, respectively.

Finally, we note that the UMP-based approach presented in the previous subsection is not well-suited for a sequential application, since the optimal log-odds ratio would be computed for each batch independently, and there is no learning of the alternative between batches. That is, we actually would obtain independent e-variables rather than sequential e-variables, which are expected to perform worse in the long run.

Notation	Description
$S_{\mathrm{GRO}(\theta_a,\theta_b)}$	GRO e-variable wrt $\mathcal{Q} = \{P_{\theta_a,\theta_b}\}$
$S_{\mathrm{TUR}(W)}$	Turner e-variable with prior W
$S_{\mathrm{c}(\delta)}$, $S_{\mathrm{c}(W)}$, $S_{\mathrm{c(NML)}}$	Cond. e-value for $\delta \in \mathbb{R}$, with prior W, and cond. NML e-value
B_t (B^t)	Data entering at (until) time t
$n_{a,t}, n_{b,t}$ (n_a^t, n_b^t)	Group sizes at (until) time t
$Y_{a,t}, Y_{b,t}$ (Y_a^t, Y_b^t)	Total number of ones in group a and b at (until) time t
$N_{1,t}$ (N_1^t)	Total number of ones in both groups at (until) time t
$S_{\mathrm{c}(\delta),t}$ $(S_{\mathrm{c}(\delta)}^t)$	Conditional e-value for $\delta \in \mathbb{R}$ at (until) time t
$S_{\mathrm{c}(W),t}$ $(S_{\mathrm{c}(W)}^t)$	Conditional e-value wrt some prior W on $\delta \in \mathbb{R}$ at (until) time t
$S_{\mathrm{c(NML)},t}$ $(S_{\mathrm{c(NML)}}^t)$	NML conditional e-value at (until) time t

Table 4.5: Notation used in the paper.

4.3 Calculating and analyzing e-variables for a single table

In this section, we provide details underlying the calculations that gave rise to Table 4.2 and we analyze the dismal behavior of the Turner e-variable on the data underlying

Table 4.1. To start with the latter: if we naively apply the Turner e-variable to the whole table, considered as a single batch, we get an e-variable of 1. This is a direct consequence of the fact we alluded to before: the Turner e-variable does not facilitate parameter learning *within* a batch. In the notation of Section 4.2.1, [30] use priors $W_{1,a}$ and $W_{1,b}$ that are identical, which implies that we get $\bar{\theta} = \bar{\theta}_a = \bar{\theta}_b$ for some θ, resulting in a likelihood ratio with the alternative represented by $(\theta_a, \theta_b) = (\bar{\theta}, \bar{\theta})$. The corresponding distribution is an element of the null, so the e-variable equals 1. This was, in fact, noticed by [85], and for the case of one-shot large tables B, they implicitly suggested to decompose the table into *mini-batches*, i.e. several small tables $B_1, \ldots, B_t$ whose entries add up to those of B. We tried this for a decomposition into (approximately 700) mini-batches with $n_{a,t} = 1, n_{b,t} = 4$, and that gave rise to the value 4.11 in Table 4.1. In more detail, we computed Turner's sequential e-variable for these individual batches (with priors in each batch equal to the posteriors from the previous batch), and report their product. Additionally, we averaged the results over 5000 runs of random sample splitting (with a standard deviation of 1.90 when drawing without replacement) to obtain a more objective result, as the resulting e-value strongly depends on the particular splitting sample chosen.

The main reason why this e-variable is much inferior to the ones discussed next seems to be that this procedure forces us to throw away data points — at some point, one of the two groups is 'exhausted,' and we are left with a remaining batch of data points that all fall in the same group. These data points (252 data points in group b, for our table) cannot be used (similarly, we could have tried batches with $n_{a,t} = 1, n_{b,t} = 5$ but this forces us to throw away points in group a rather than b and the resulting e-variable still remains small). We tried several other combinations of n_a and n_b, but none of them yielded significantly better e-values. An additional issue is that, in practice, it would not be clear which decomposition to use. Another reason for the bad behavior is that, since in all reasonable decompositions $n_{a,t} \neq n_{b,t}$, the e-values we use do not have GRO status.

The *UI* e-variable behaves even worse than the Turner one. While surprising at first, this finding is consistent with the asymptotic results presented in [29]. Their Theorem 4 (parts 1 and 3) roughly states that, if we were to pick a prior W that puts all its mass on the 1-dimensional curve of all (θ_a, θ_b) that all share the same odds ratio, i.e. that satisfy (4.7) for the same δ, then the expected logarithmic difference between the UI e-variable applied to fixed sample size n and the Turner e-variable, for a total of n batches, is of order $c \log n + O(1)$ for a constant c close to $1/2$ (in the language of their Theorem 4, what we call a 'batch' is a single data point; what we call the Turner e-variable is their *sequential-RIPr* e-variable). We would then expect a ratio between both e-variables of about $\exp(c \log n + O(1)) = c_n \sqrt{n}$, where n is the number of batches, which, due to our setting $n_{a,t} = 1, n_{b,t} = 4$ (see above) was approximately 700 (accounting for a factor of $\sqrt{700} \approx 26$), and c_n tends to a positive unknown constant. The fact that this asymptotic analysis holds for *every* fixed δ, suggests that something similar still holds if, as in our implementation of UI and Turner, we use a prior W spread out over all (θ_a, θ_b) rather than just those corresponding to a particular δ. This independently suggests, but of course does not prove, that (at least this version of) UI is non-competitive.

As regards the *e-to-p calibrators*, the underlying p-value — based on Fisher's exact test — is obtained by conditioning on n_1 and modeling Y_a as a hypergeometric random variable, in close analogy with our conditional e-variables. From this perspective, some degree of similarity in their behavior is to be expected. Nevertheless, the specific choice of the calibrators f and g is not derived from, nor directly connected to, the hypergeometric model. For this reason, we found it somewhat surprising that they exhibit such good performance, at least when compared to the Turner e-variable.

Turning to the *conditional* e-variables, we observe remarkably strong performance of the UMP-2 e-variable. Importantly, although the construction requires fixing an initial significance level α to derive a (powerful) choice of δ, the resulting UMP-2 e-variable remains valid for any alternative significance level $\alpha^* \neq \alpha$. That is, under the null hypothesis, it satisfies

$$P_0\left(S_{\text{UMP-2}} > \frac{1}{\alpha^*}\right) \leqslant \alpha^* \qquad \text{for all } \alpha^* \in (0,1).$$

When $\alpha^* \neq \alpha$, the resulting test may no longer be optimal in terms of power, but it still retains the Type-I error control at level α^*. Moreover, our experiments indicate that the performance of the UMP-2 e-variable is relatively robust with respect to the choice of the initial significance level, provided that α lies within a reasonable range (e.g. $[0.01, 0.1]$).

The two remaining conditional e-variables, although grounded in standard approaches for handling composite alternatives, do not perform particularly well. This is essentially due to the fact that both are designed to achieve an expected logarithmic growth rate that, regardless of the true parameter pair (θ_a, θ_b) generating the data, remains close to that of the optimal *oracle* conditional e-variable, defined in terms of the true value of δ associated with (θ_a, θ_b) via (4.7).

This design choice entails a non-negligible overhead. In the case of the mixture conditional e-variable, the overhead arises because the prior $W(\delta)$ distributes its mass over the entire parameter space. In the conditional NML case, it is induced by the normalization factor appearing in (4.10). In the unconditional setting, the resulting loss in expected logarithmic growth relative to the oracle amounts to $(1/2)\log n + O(1)$, both for NML and mixture e-variables [17] (see also Chapter 3), and we expect a similar behavior to hold in the conditional case.

By contrast, the UMP conditional e-variables circumvent this cost by concentrating essentially all prior mass on values of δ that allow rejection with high probability for the given sample sizes and significance levels α. This advantage, however, comes at the cost of markedly poor performance when α or sample sizes differ substantially.

4.4 Comparison of Turner's and conditional e-variables in sequential settings

We assess performance in a sequential setting from three complementary perspectives: asymptotic optimality, practical evidence accumulation, and stopping-time behavior. To this end, we consider independently distributed data streams for groups a and b,

generated by two Bernoulli processes with fixed oracle parameters θ_a and θ_b, respectively. We evaluate the performance of various sequential e-variables S_t by simulation.

We present three figures, Figures 4.1–4.3, each comprising 16 panels corresponding to simulation settings that differ in the data-generating parameters (θ_a, θ_b), the total batch size $n_a + n_b$, and the group-size ratio n_a/n_b. For simplicity, we assume that all batches within a given simulation have identical sizes and ratios; that is, for each batch B_t, we have $n_{a,t} = n_a$ and $n_{b,t} = n_b$ for some fixed n_a and n_b.

We selected two configurations of the *effect size* δ for the data generating distribution: $\delta = 2\log 19$ (very large effect, rejecting the null is easy, first two rows of the figures) and $\delta = 2\log(3/2)$ (small-to-moderate effect, more difficult to discover, third and fourth row of the figures), which can be spotted by different δs' *distance* to the null, the diagonal line in the parameter space in Figure 4.5. For both these effect sizes, we chose two concrete instantiations: one on the counter diagonal, i.e. $(\theta_a, \theta_b) = (0.95, 0.05)$ corresponding to $\delta = 2\log 19$ (first row of the figures) and $(\theta_a, \theta_b) = (0.6, 0.4)$ corresponding to $\delta = 2\log(3/2)$ (third row of the figures); and one more towards the corner, i.e. $(\theta_a, \theta_b) = (0.78, 0.01)$ corresponding to $\delta = 2\log 19$ (second row of the figures) and $(\theta_a, \theta_b) = (0.16, 0.08)$ corresponding to $\delta = 2\log(3/2)$ (fourth row of the figures). Oracle parameters are visualized as individual dots in Figure 4.5. We also considered small and medium batches with group size ratio 1 (first two columns of the graphs) and with a group size ratio far from 1 (third and fourth columns of the graphs). In this way, each figure consists of 16 graphs, each depicting a separate data-generating mechanism.

Figure 4.1 measures the deviation from the (θ_a, θ_b)-GRO e-variable. This metric is defined as the expected difference

$$\mathbb{E}_{P_{(\theta_a,\theta_b)}}\left[\log \prod_{\ell=1}^{t} s_{\mathrm{GRO}(\theta_a,\theta_b)}(Y_{a,\ell}, Y_{b,\ell}) - \log \prod_{\ell=1}^{t} S_\ell\right] = \\ = \mathbb{E}_{P_{(\theta_a,\theta_b)}}\left[\sum_{\ell=1}^{t} \log \frac{s_{\mathrm{GRO}(\theta_a,\theta_b)}(Y_{a,\ell}, Y_{b,\ell})}{S_\ell}\right], \quad (4.18)$$

with the explicit formula for the GRO given by (4.3) and (4.4), and where S_ℓ denote the sequential e-values of interest. Figure 4.1 provides a comparison for S_ℓ either the sequential Turner $S_{\mathrm{TUR}(W_\ell)}$ or the sequential NML e-variable $S_{\mathrm{c(NML)},\ell}$.

Secondly, we report the expected accumulated evidence over time as a measure of performance. Figure 4.2 depicts the averaged cumulative products of the sequential Turner and sequential NML e-values with increasing number of batches. Here, the red horizontal line indicates the threshold $\log(100) \approx 4.60$, corresponding to rejecting the null at level $\alpha = 0.01$. Intuitively, we expect to reject the null after observing roughly the number of batches in which the expected evidence exceeds this threshold. However, since this reasoning can be misleading in general, we report also the stopping time distribution as a third measure of performance: In Figure 4.3, we compare the stopping times $\tau = \min\left\{t \geqslant 1 \mid \prod_{\ell=1}^{t} S_\ell \geqslant 1/\alpha\right\}$ at level $\alpha = 0.01$ for the sequential Turner and sequential NML e-values. All plots are generated by averaging over $N = 1000$ runs of each simulation.

Conclusions Figure 4.1 clearly illustrates that, asymptotically, the Turner e-variable always wins in the sense of achieving the fastest growth rate. The only panel where this behavior is not immediately apparent is the one in the third row and fourth column; however, we expect that, upon adding a sufficiently large number of additional batches, the Turner e-variable would eventually overtake the conditional NML e-variable, as observed in the other panels.

This behavior can be explained as follows. For each batch, the expected growth of the conditional NML e-variable — even if it were equipped with the oracle value of δ corresponding to the true parameter pair (θ_a, θ_b) — falls short of the expected growth of the optimal (oracle) e-variable by a small amount $\epsilon > 0$. As t increases, the maximum likelihood estimator $\hat{\delta}$ used in the NML procedure at batch t converges to this true value of δ, implying that the relative growth in (4.18) converges, to first order, to $n\epsilon$, for some small $\epsilon > 0$. As shown in [29], ϵ vanishes as the batch size increases; nevertheless, for any fixed batch size, the conditional NML e-variable remains linearly suboptimal relative to the oracle e-variable as the number of batches tends to infinity.

By contrast, the behavior of the Turner e-variable is, in a sense, dual. Consider each batch B_t itself as a sequence of observations $\{X_i^a\}$ and $\{X_i^b\}$. As already discussed in Section 4.2.2, the Turner e-process does not perform any *within-batch learning*: the approximation $(\bar{\theta}_a, \bar{\theta}_b)$ to the true, unknown parameter pair (θ_a, θ_b) used to evaluate the e-variable remains fixed throughout the batch, thereby incurring a constant approximation error for all outcomes *within* that batch.

When a new batch becomes available, however, this approximation is updated and converges to the true data-generating parameters (θ_a, θ_b). Consequently, in contrast to the conditional NML case, the approximation error — corresponding to each term in (4.18) — vanishes as the number of batches ℓ grows. Within any given batch, the approximation error remains non-zero but constant. As a result, for the Turner e-variable, smaller batch sizes are advantageous, a behavior that is consistent with what is observed in the plots.

Figure 4.2 shows that, for large effect sizes (top two rows), the conditional NML e-variable outperforms the Turner e-variable for practically relevant numbers of batches, that is, for total number of observations at which the accumulated evidence is not much larger than $\log(1/\alpha)$ for reasonable choices of $\alpha \in (0, 1)$. Note that the horizontal scale in Figure 4.2 is therefore substantially smaller than that in Figure 4.1, where our focus was on asymptotic behavior.

When the effect size is small (bottom two rows), corresponding to a more challenging testing problem, the comparison becomes less clear-cut, and the two e-variables exhibit very similar performance. Moreover, for large effect sizes, the non-centrality parameter appears to have little influence on performance, as the first two rows look essentially the same. Finally, we observe that the crossover point at which the Turner e-variable overtakes the conditional NML e-variable occurs particularly early for balanced data and small batch sizes, as illustrated by the top two rows of the first column in Figure 4.2. This observation is consistent with our earlier findings that small and balanced batch sizes are favorable for the Turner e-variable.

From Figure 4.3 we can conclude that, for large effect sizes, rejection typically

occurs slightly faster with the conditional NML e-variable than with the Turner e-variable. The difference in the required number of batches, however, is very small, since both methods lead to rejection extremely quickly: across all eight simulations in the first two rows, no more than four batches are ever needed. For smaller effect sizes (bottom two rows), the stopping times associated with the conditional NML e-variable appear to be somewhat more dispersed than those of the Turner e-variable.

Figure 4.4 zooms in on one of the sixteen cases shown in Figure 4.3 and additionally displays the individual trajectories of the conditional NML and Turner e-variables across all simulation runs. The visualization of simulated trajectories together with stopping-time distributions is inspired by the informative plots presented in [95]. From these trajectories, we can clearly observe the potential advantage of the conditional NML e-variable over the Turner e-variable already in the first batch, as well as the fact that, for certain data realizations, the evidence accumulated by the Turner method may drop to very low levels before eventually reaching significance.

4.5 Conclusion

This paper investigates e-values and e-processes for testing 2×2 contingency tables, with a particular focus on the sequential performance of Turner's e-variable and a newly proposed conditional NML e-variable. Our simulation study suggests that, while the Turner e-variable dominates asymptotically in terms of growth rate, the conditional NML e-variable performs competitively at practically relevant sample sizes and target significance levels. In these regimes, the NML-based approach is either substantially better than Turner or only slightly worse. We emphasize that these findings should be interpreted as preliminary. They provide initial evidence on the relative strengths of the methods, but do not yet offer a complete characterization of their performance. A more systematic exploration through additional experiments remains an important direction for future work.

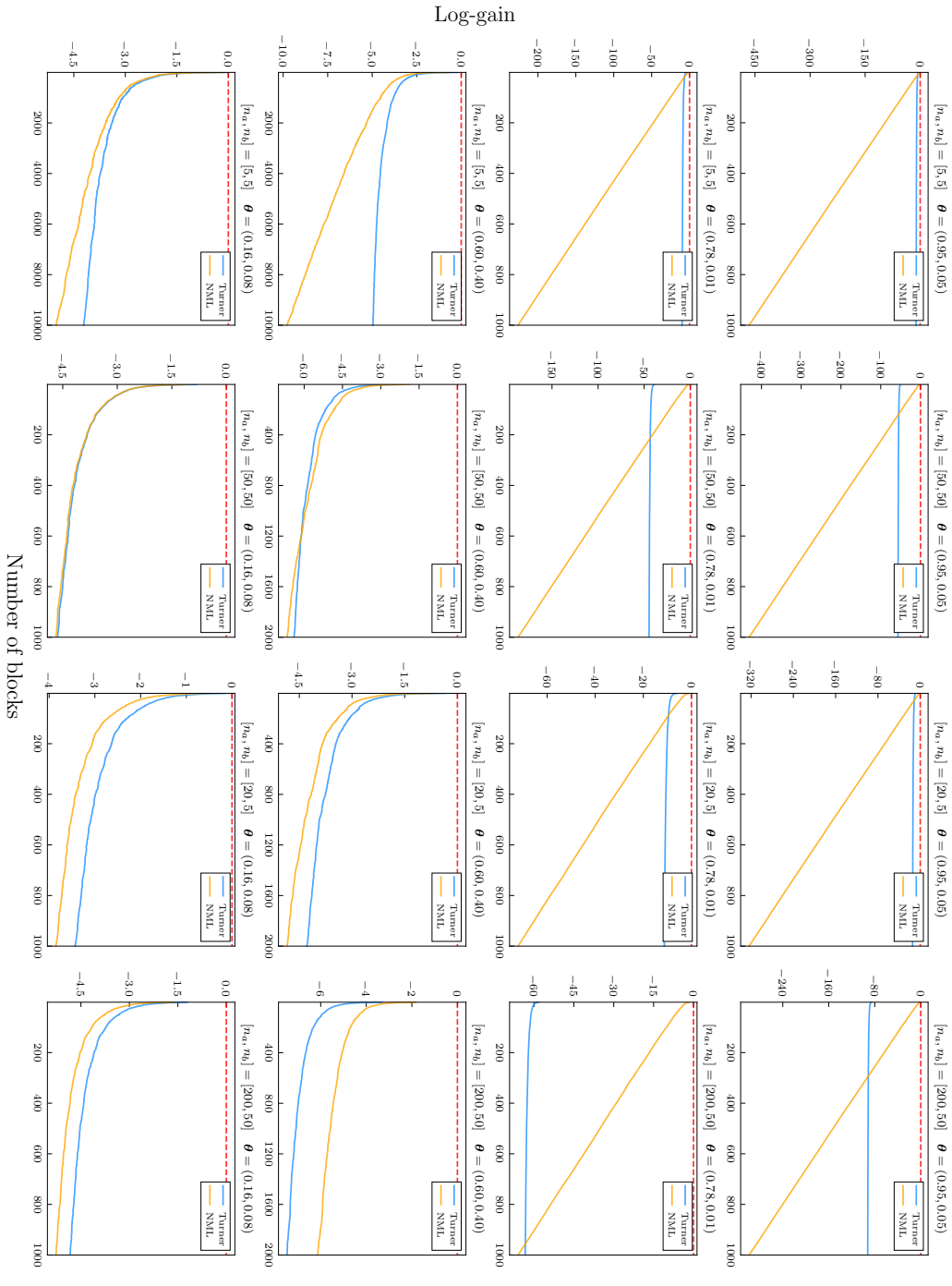

Figure 4.1: Empirical expected (log) gain (as given in (4.18)*) of theoretically optimal GRO in comparison to sequential Turner (blue) and sequential conditional NML (orange).*

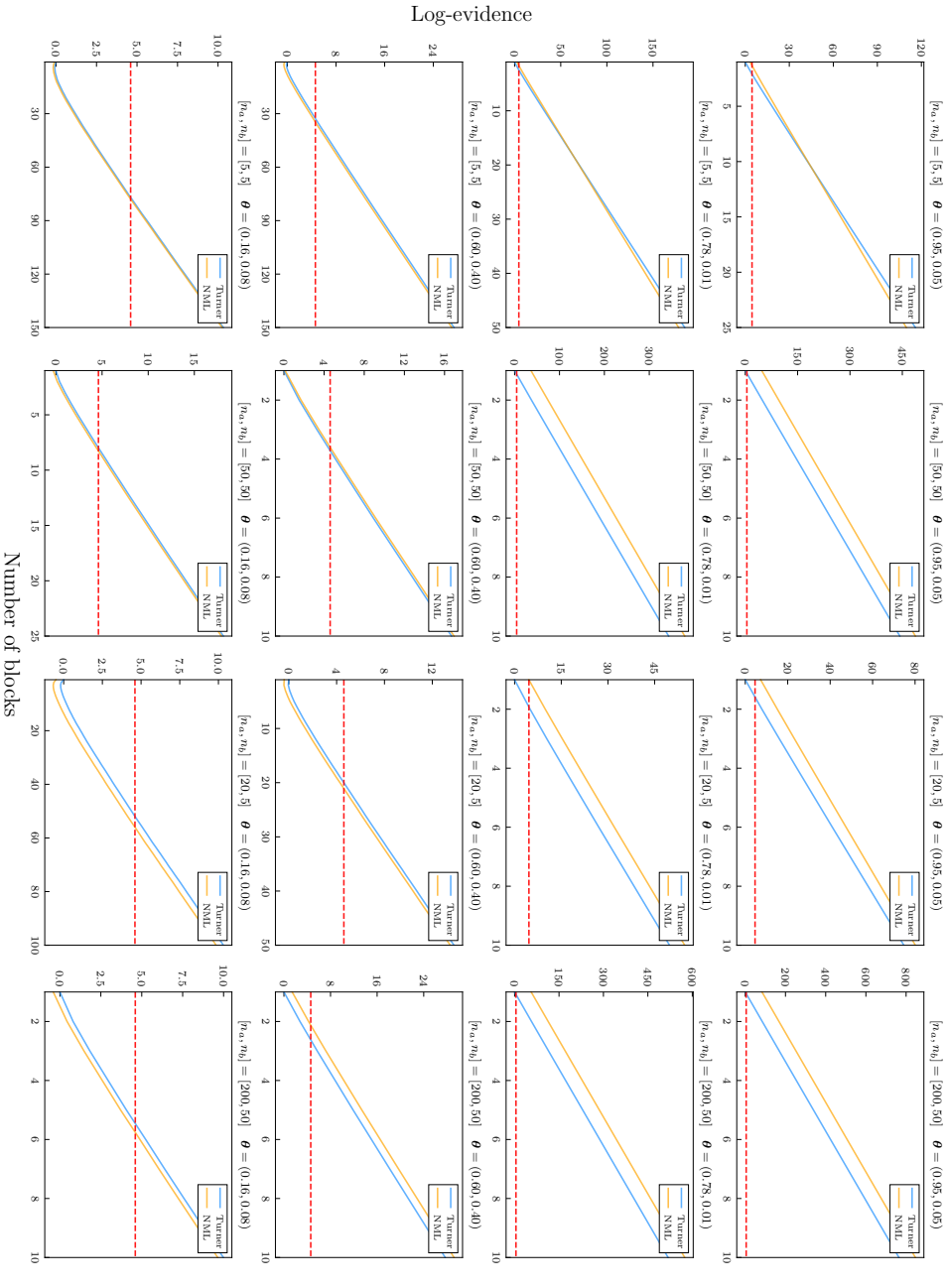

Figure 4.2: Empirical expected (log) evidence of sequential Turner (blue) and sequential conditional NML (orange). The horizontal line indicates the threshold corresponding to rejection at level $\alpha = 0.01$.

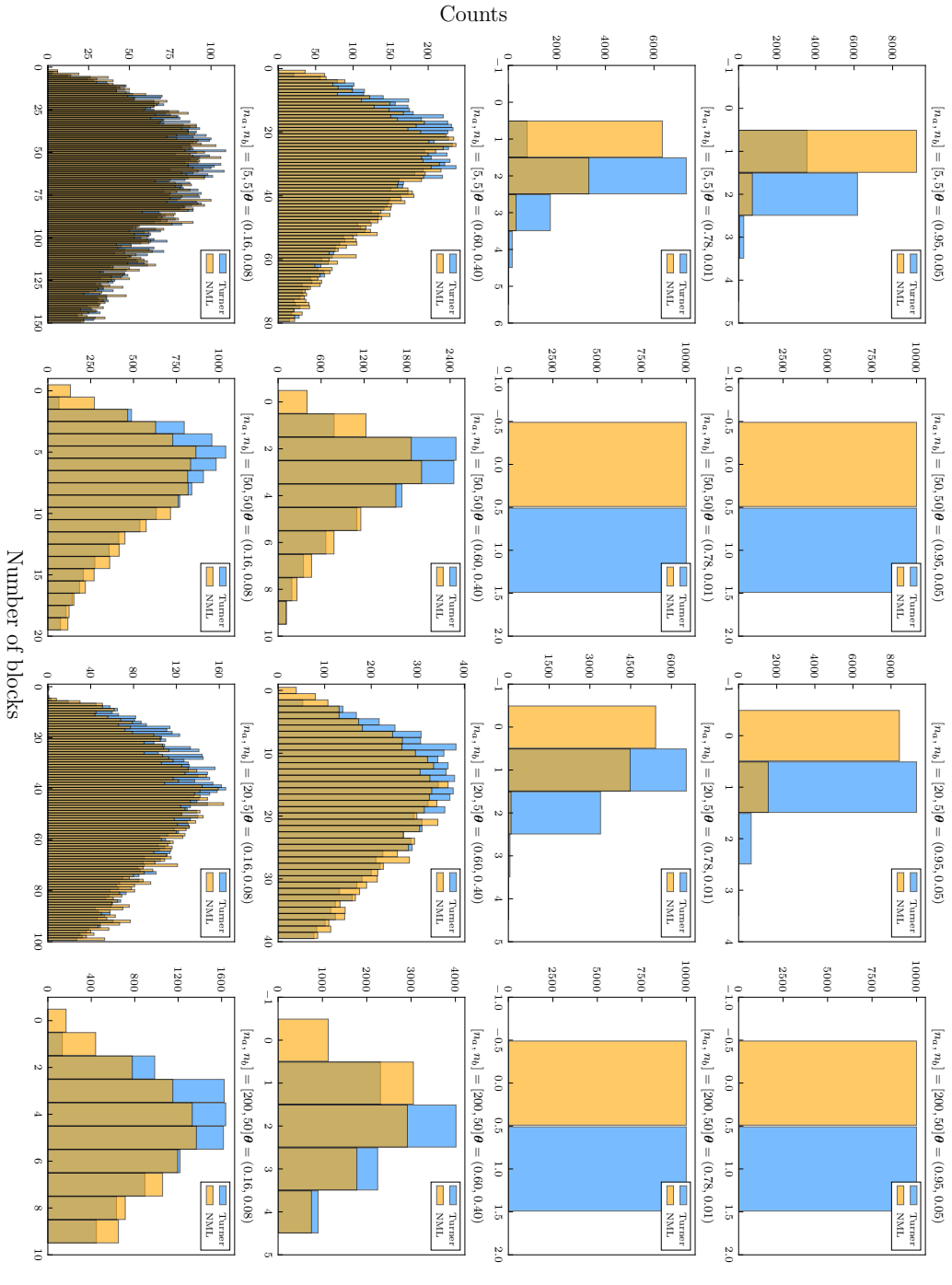

Figure 4.3: Empirical distributions of stopping times for sequential Turner (blue) and sequential conditional NML (orange) for confidence level $\alpha = 0.01$.

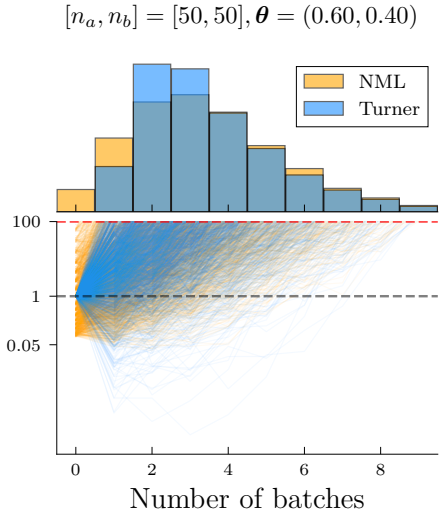

Figure 4.4: For one simulation configuration ($(n_a, n_b) = (50, 50)$, $(\theta_a, \theta_b) = (0.6, 0.4)$), the stopping time distribution for $\alpha = 0.01$ is plotted jointly with trajectories of the sequential Turner (blue) and sequential conditional NML (orange) accumulated evidence.

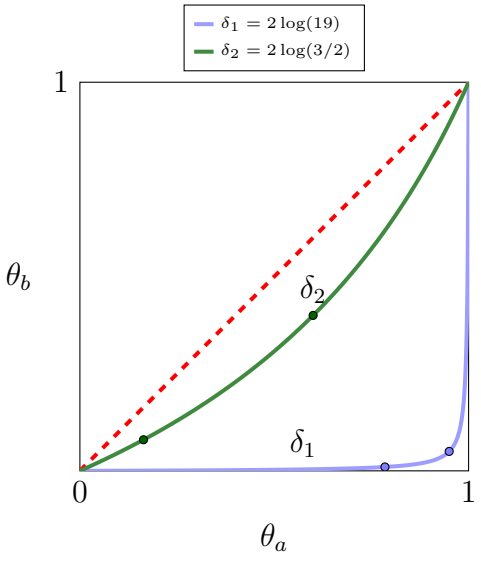

Figure 4.5: Parameter space with oracle parameters used for data generation.