Abstract

Sequential decision making significantly speeds up research and is more cost-effective compared to fixed-n methods. We present a method for sequential decision making for stratified count data that retains Type-I error guarantee or false discovery rate under optional stopping, using e-variables. We invert the method to construct stratified anytime-valid confidence sequences, where cross-talk between subpopulations in the data can be allowed during data collection to improve power. Finally, we combine information collected in separate subpopulations through pseudo-Bayesian averaging and switching to create effective estimates for the minimal, mean and maximal treatment effects in the subpopulations.

1 INTRODUCTION

Fixed-n hypothesis tests and confidence intervals limit research opportunities and quick decision making, as they rely on static research designs where data are only evaluated at one time point. We aim to develop hypothesis tests for conditional independence and anytime-valid confidence sequences for stratified treatment effects in subpopulations that retain a guarantee on the probability of falsely rejecting the null hypothesis and coverage of the true effect under continuous monitoring of data. To this end we use e-values, tools for constructing tests that keep the type-I error rate (or false positive rate) controlled under sequential testing with optional stopping. Over the last four years, e-values have become the standard tools (essentially, the appropriate alternative for p-values) for dealing with such settings. Below we summarize the essentials: for much more background on the budding field of e-processes (also known as ‘testing by betting’ and ‘safe testing’) see the recent overview (Ramdas et al., 2022) and specifically for details on e-values refer to Grünwald et al. (2022); Vovk and Wang (2021). In this paper, we develop e-processes for stratified $2 \times 2$ tables, enabling, in Section 2, anytime-valid (i.e. valid under optional stopping) conditional independence (CI) tests for Bernoulli streams for two groups $a$ and $b$ (e.g. $a$ is control, $b$ is treatment), where the test is conditional on a third variable, the stratum. Based on these CI tests, we then, in Section 3, develop anytime-valid confidence sequences (henceforth just called ‘confidence sequences’) for a notion of effect size representing divergence from CI. The importance of our tests is ubiquitous in e.g. medical statistics — we can think of the CI test in Section 2 as an an anytime-valid sequential version of the Cochran-Mantel-Haenszel test, a work-horse in the field of epidemiology. Our e-processes are generalizations of those designed for $2 \times 2$ tables (same setting as ours, but with just a single stratum) by Turner et al. (2021); Turner and Grünwald (2022). To achieve the generalization, we employ tools from the theoretical machine learning literature, most notably the literature on prediction with expert advice (Cesa-Bianchi and Lugosi, 2006), which extends Bayesian learning techniques with ideas such as ‘sleeping’, ‘switching’ and the like. Moreover, inspired by these ideas, we develop the novel notion of cross-talk between strata, which allows us to make confidence intervals narrower if outcomes in various strata are interrelated, while nevertheless remaining valid even if they are not. While for many statistical models, anytime-valid tests need more data to reach a desired conclusion than fixed $n$ methods and anytime-valid confidence intervals are somewhat wider than standard ones (Ramdas et al., 2022; Grünwald et al., 2022), we find in this paper that we can partially counteract this difference by employing the cross-talk strategy (which is not available for fixed-$n$ methods), as is illustrated by comparing our confidence sequences to fixed-$n$ confidence intervals for Mantel-Haenszel risk differences in Section 3.

E-Processes Consider a random process $Y_1, Y_2, \ldots$ and let $\mathcal{H}_0$, the null hypothesis, be a set of distributions for this process. An e-variable for $Y_j, Y_{j+1}, \ldots, Y_m$ conditional on $Y^{(j-1)} = (Y_1, \ldots, Y_{j-1})$ for testing $\mathcal{H}_0$ is any nonnegative random variable $S$ that can be written as function of $Y^{(m)} = (Y_1, \ldots, Y_m)$ such that

$$\forall P \in \mathcal{H}_0 : \mathbb{E}_P[S | S^{(j-1)}] \leq 1; \quad (1)$$

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).
for \( j = 1 \) we set \( \mathbb{E}_\mathcal{P}[S \mid S^{(0)}] := \mathbb{E}_\mathcal{P}[S] \) and call \( S \) an unconditional e-variable; an e-value is the value an e-variable takes on a realized sample. It is easily shown that for any sequence \( S_1, S_2, \ldots \) where \( S_j \) is an e-variable for \( Y_j \) conditional on \( Y_j \), the product \( E^{(m)} := \prod_{j=1}^{m} S_j \) is an unconditional e-variable for \( Y^{(m)} \). \( E^{(1)}, E^{(2)}, \ldots \) is called a test martingale or e-process (see Ramdas et al. (2022) on how e-processes strictly generalize test martingales). Via Ville’s inequality, it is shown that e-processes have the remarkable property that, for any \( 0 < \alpha < 1 \), the probability that there exists an \( m \) such that \( E^{(m)} \geq 1/\alpha \) is bounded by \( \alpha \). As a consequence, if we look at the data at some time \( m \) and reject if \( E^{(m)} \geq 1/\alpha \), the probability under the null of falsely rejecting the null is at most \( \alpha \) no matter how we chose \( m \); it may be determined by external circumstances (do we have money to experiment further?) or by aggressive stopping rules such as ‘keep sampling until you can reject the null’, or even by peeking into the future. Tests with this property are called safe under optional stopping and Ramdas et al. (2020) show that, in essence, all reasonable such tests should be based on e-processes. Just like p-values can be converted into confidence intervals, e-process can be converted into anytime-valid confidence tests, also known as confidence sequences — we will explore these in Section 3.

**Setting** We consider the stratified contingency table setting/model. Under the global null hypothesis (we consider more complicated nulls later), outcomes \( Y \in \{0, 1\} \) are independent of groups \( X \in \{a, b\} \) (e.g. representing interventions) given their stratum \( k \in [K] := \{1, \ldots, K\} \). We formalize this by measuring time in terms of blocks: we assume that at each time \( j = 1, 2, \ldots \), we are given a stratum indicator \( k_j \in [K] \) and we observe a block of \( n = n_a + n_b \) outcomes, with \( n_a \) outcomes in group \( a \) and \( n_b \) in group \( b \), all in the same stratum \( k_j \). We write \( Y^{(m)} = (Y_1, \ldots, Y_m) \) with \( Y_j \) the data vector corresponding to the \( j \)-th block arriving. Hence \( Y = (Y_{j,a,1}, \ldots, Y_j,a,n_a, Y_j,b,1, \ldots, Y_j,b,n_b) \) is a vector in \( \{0, 1\}^n \) denoting \( n = n_a + n_b \) outcomes in \( k_j \). Under both null and alternative, all blocks are assumed independent, with each outcome in group \( x \) in stratum \( k \) independently ~ Bernoulli(\( \theta_{x,k} \)). Formally, the null hypothesis then expresses that

\[
\mathcal{H}_0: \theta_{a,k} = \theta_{b,k} \text{ for all } k. \tag{2}
\]

We will assume \( n_a = n_b = 1 \) for all strata in simulation examples in this paper, but these can be chosen freely in practice and can even be adapted in between data blocks — as long as they are set at or before the beginning of a data block, they are allowed to depend on the past. Of course, in practice, we often deal with \( 2K \) i.i.d. streams of data, one for each group-stratum combination, with data not necessarily coming in at the same rate for different strata/groups. While superficially different, we can still recast this setting in terms of blocks: for example, participant may sequentially enter a study and are each independently randomized with probability \( 1/2 \) to receive ‘treatment’ (group \( b \)) or ‘control/placebo’ (\( a \)). We then wait until the first time \( t_1 \) that we have seen \( n_a \) outcomes in group \( a \) and \( n_b \) outcomes in group \( b \) in the same stratum; we call this stratum \( k_1 \), denote these \( n \) outcomes \( Y_1 \), and proceed observing outcomes in the various streams until the first time \( t_2 \) that there is another stratum \( k_2 \) (potentially \( k_2 = k_1 \)) so that we have seen \( n_a \) outcomes in group \( a \), \( n_b \) in group \( b \) in stratum \( k_2 \); we denote these \( n \) outcomes \( Y_2 \), and so on. If we want to stop at any time \( t \), we take as data all blocks that have been completed so far, and ignore all started-yet-unfinished blocks.

**Related Work** The first paper to use e-processes for conditional independence testing is (Lindon and Malek, 2020), but their tests are very different from ours and involve a simple null hypothesis, allowing them to use Bayes factors for their e-processes. Further, Turner et al. (2021); Turner and Grünwald (2022) develop independence tests and confidence sequence for \( 2 \times 2 \) tables; our paper is a direct extension of theirs, extending their techniques to the stratum-conditional case. Very recently four other related papers have appeared: (Pandeva et al., 2022; Grünwald et al., 2022; Shaer et al., 2022; Duan et al., 2022): these papers all differ from ours in that they assume data are jointly i.i.d. (i.e. one observes a single i.i.d. stream \( (X_1, Y_1, K_1), (X_2, Y_2, K_2), \ldots \)). The latter three also make the so-called Model-X assumption (the distribution of \( X_i \mid K_i \) is assumed known). Our paper is complementary: we do not need the i.i.d. or Model-X assumption and as explained above, our setting does not just capture data in blocks (such as paired data) but also data in the form of \( 2K \) i.i.d. streams, one for each group in each stratum, with no stochastic assumptions about what group or what stratum arrives at what time. The price we have to pay is that we can only deal with a small number of strata and with finite sets of outcomes and number of groups (in this paper we focus on 2 but extension to the finite case is straightforward); aforementioned references can deal with arbitrary covariate and outcome random variables \( K_i \) and \( Y_i \). Nevertheless, small-strata-count-studies are highly common in the medical statistics world, and we show here how to construct efficient sequential tests for them.

The code used for experiments in this paper will initially be placed on the repository linked to this publication (Turner, 2023), and will later be integrated in the safestats R package (Ly et al., 2022).

## 2 E-VARIABLES FOR TESTING THE GLOBAL NULL

We first consider the case where there is only one stratum, \( k_j = k^* \) for each each \( j \). The problem is then reduced to
testing whether two Bernoulli data streams come from the same source. Turner et al. (2021) showed that in this case, for arbitrary estimators \( \hat{\theta}_a | Y_{j-1}, \hat{\theta}_b | Y_{j-1} \), the following is an e-variable for \( Y_j \) conditional on \( Y_{j-1} \), i.e. (1) holds with \( S := S_j \) given by

\[
S_j = \prod_{i=1}^{n_a} p_{\theta_a}(Y_{j,a,i}) \prod_{i=1}^{n_b} p_{\theta_b}(Y_{j,b,i}).
\]

where \( p_{\theta}(Y) = \theta^Y (1 - \theta)^{1-Y} \) denotes the Bernoulli(\( \theta \)) probability of \( Y \in \{0, 1\} \), as long as we pick \( \hat{\theta}_a \in \Theta_0 = [0, 1] \) as follows:

\[
\hat{\theta}_0 = \hat{\theta}_0 | Y_{j-1} := \arg \min_{\theta \in [0, 1]} D(D_{\theta_a, \theta_b} \| P_{\theta, \theta})
\]

where \( D_{\theta_a, \theta_b} \) represents the distribution on \( \theta \in \Theta_0 \) independent binary outcomes with the first \( n_a \) outcomes \~ Bernoulli(\( \theta_a \)) and the subsequent \( n_b \) outcomes \~ Bernoulli(\( \theta_b \)), i.e. the distribution of outcomes in a single block according to \( \{ \theta_a, \theta_b \} \), and \( D(D_{\theta_a, \theta_b} \| P_{\theta, \theta}) \) abbreviates the KL divergence between two such distributions. Equality (a) follows by simple calculus.

Importantly, in (3), \( (\hat{\theta}_a, \hat{\theta}_b) \in \Theta_1 = [0, 1]^2 \) can be chosen as a function of past data anyway we like, not affecting the Type-I error guarantee. Nevertheless, if we were given the true probabilities \( \theta^*_a \) and \( \theta^*_b \) of the two groups in block \( j \), then we could set \( \hat{\theta}_a = \theta^*_a \) and \( \hat{\theta}_b = \theta^*_b \) and this choice is special: the e-variable (3) then has, among all e-variables, the largest expected logarithm under the true alternative \( P_{\theta^*_a, \theta^*_b} \). We then say it is growth-rate optimal (GRO) for collecting evidence against the null hypothesis (Grünewald et al., 2022). Formally, we define

\[
\text{GRO}(\theta^*_a, \theta^*_b) := \sup_S E_{Y_j \sim P_{\theta^*_a, \theta^*_b}} [\log S]
\]

where the supremum is over all random variables \( S \) that are e-variables for \( Y_j \) under \( \mathcal{H}_0 \). It directly follows from (Grünewald et al., 2022, Theorem 1) that, if we plug in \( \hat{\theta}_a = \theta^*_a \) and \( \hat{\theta}_b = \theta^*_b \) into (3), then the resulting \( S_j \) is GRO and its growth rate is equal to the KL divergence, i.e.

\[
\text{GRO}(\theta^*_a, \theta^*_b) = \sup_S E_{Y_j \sim P_{\theta^*_a, \theta^*_b}} [\log S] = \text{GRO}(\theta^*_a, \theta^*_b) = D(P_{\theta^*_a, \theta^*_b} \| P_{\theta, \theta}),
\]

where \( \theta = (n_a/n)\theta^*_a + (n_b/n)\theta^*_b \). Growth-rate optimality is the analogue of statistical power in the sequential setting: if we plug in these ‘true’ \( \hat{\theta}_a = \theta^*_a \) and \( \hat{\theta}_b = \theta^*_b \), we expect the product \( E^{(m)} \) to increase as fast as possible in \( m \), enabling us to reach \( 1/\alpha \) and reject the null hypothesis as fast as possible, compared with all other possible e-processes. In practice though, \( \theta^*_a \) and \( \theta^*_b \) are unknown, but to get near-grow-optimal e-variables, we can estimate \( \hat{\theta}_a \) and \( \hat{\theta}_b \) based on all data seen before data block \( j \) — then \( \hat{\theta}_a \) and \( \hat{\theta}_b \) converge to \( \theta^*_a \) and \( \theta^*_b \) and our e-variables \( S_j \) get better and better in the GRO sense. We follow Turner et al. (2021) who successfully chose to place a beta prior on the parameter space and took the Bayesian posterior mean as an estimate.

In treatment/control test settings, there often exists prior knowledge of a minimal clinically relevant or expected odds ratio \( \text{OR}(\theta_a, \theta_b) := (\theta_b/(1 - \theta_b))/((1 - \theta_a)/\theta_a) \), i.e. it is known that \( \text{OR}(\theta_a, \theta_b) = \phi \) for some given \( \phi \). In that case, one can restrict estimating \( \hat{\theta}_a \) and \( \hat{\theta}_b \) to \( \Theta_1(\phi) = \{(\theta_a, \theta_b): \text{OR}(\theta_a, \theta_b) = \phi\} \), possibly improving power and growth-rate of the test (Turner et al., 2021). Both search spaces are illustrated in Figure 1.

**Figure 1:** Parameter space \( \hat{\theta}_a | Y_{j-1} \) and \( \hat{\theta}_b | Y_{j-1} \) are estimated in, in \( 2 \times 2 \) table without strata; either through placing a beta prior on the entire unit square (in light orange) and calculating the posterior mean with all data up to and including time \( j - 1 \) or through restricting the posterior estimation to a particular odds ratio value \( \phi \) and placing a beta prior on all pairs \( (\theta_a, \theta_b) \) corresponding to this odds ratio value (for example the red curve, for \( \phi = 2 \)).

**Combining e-variables from individual strata** We can use the e-variable in (3) to calculate e-process values \( E^{(m),k} \) for each stratum \( k \) separately. To be precise, we set \( S_j^k \) to the equivalent of (3) if \( k = k_j \),

\[
S_j^k = \prod_{i=1}^{n_k} p_{\theta_{k,i} Y_{j-1}}(Y_{j,a,i}) \prod_{i=1}^{n_k} p_{\theta_{k,i} Y_{j-1}}(Y_{j,b,i}),
\]

and \( S_j^k = 1 \) otherwise, i.e. if \( k_j \neq k \), and \( E^{(m),k} := \prod_{j=1}^{n_k} S_j^k \) — note that at each ‘time’ \( j \), the product e-variable only changes for the \( k \) such that \( j \)-th block was a block of outcomes in stratum \( k \).

We now need to combine the e-processes-per-stratum into a single e-process for (2) to measure evidence against \( \mathcal{H}_0 \) and allowing tests with type-I error probability guarantee on (2), the global null hypothesis that the odds ratio of
the success probabilities equals 1 in each stratum. There are several ways to do this. The first and most straightforward option is to multiply the individual e-values across the strata:

\[ E^{(m)} = \prod_{j=1}^{m} S_j^k = \prod_{j=1}^{m} \prod_{k=1}^{K} S_j^k. \]  

\[(8)\]

To see that \(E^{(1)}, E^{(2)}, \ldots\) is an e-process, simply note that each \(S_j^k\) is a conditional e-variable (i.e. it satisfies (1) with \(S = S_j^k\)) since, given that \(S_j\) in (3) is a conditional e-variable, \(S_j^k\) must be an e-variable as well. When \(\theta_{a,k} = \theta_{b,k}\) in a few of the strata, this might be a data-inefficient approach, as one would need to collect a lot of extra evidence in the strata where the success probabilities are substantially different to counteract the expected small e-values in the other strata. A second option that possibly better handles these cases is to create a convex combination, i.e. a mixture, of e-values at each time point \(j\) (any convex combination of e-variables is also an e-variable (Yok and Wang, 2021)): A simple first option is to pick some prior distribution on the strata \(\pi(k)\), and to use that distribution for calculating the mixture after each batch comes in:

\[ S_j := \sum_{k=1}^{K} \pi(k) S_j^k; \quad E^{(m)} = \prod_{j=1}^{m} S_j \quad \text{so that also} \]

\[ E^{(m)} = \prod_{k=1}^{K} E^{(m),k} \quad \text{with} \quad E^{(m),k} = \prod_{j=1}^{m} S_j^k. \]

\[(9)\]

Extending the simple averaging above, we could replace the prior \(\pi(k)\) in (9) with a distribution \(\pi(k|y^{(j-1)})\) that depends on previous data \(y^{(j-1)}\), since, since we assume the data itself in each block are independent, dependency of \(\pi\) on past data will not affect guarantee (1). Such an approach is called the method of mixtures in the anytime-valid testing literature (Ramdas et al., 2022). Thus, any distribution on \([K]\) that depends on the past is allowed here, but an intuitive choice is a pseudo-Bayesian posterior

\[ \pi(k|y^{(j-1)}) := \frac{\pi(k)(E^{(j-1),k})^\eta}{\sum_{k'} \pi(k')(E^{(j-1),k'})^\eta}, \]

\[(10)\]

where by definition, \(E^{(0)} = 1\) and we pick \(\eta\) beforehand as a learning rate: if we set it to a higher value, we will focus on strata with higher e-values more quickly; with \(\eta = 1\), (10) becomes similar to a Bayesian posterior. Just as the beta-posterior used to determine \(\theta_{x,k}\) in (7) allows us to learn \(\theta^*_{x,k}\), this new posterior allows us to learn which strata can help us most to reject the null. However, even for \(\eta = 1\) the analogy to Bayes only goes so far — for example, at each \(j\), only the e-variable \(S_j^k\) for stratum \(k_j\) changes; the other \(S_j^k\) ‘sleep’ (Koolen and van Erven, 2010) and thus \(E^{(j-1),k}\) behaves differently from a likelihood. This more general past-determined updating originates in the area of machine learning called prediction with expert advice where many other such ‘posterior’-updates have been considered (Herbster and Warmuth, 1998; van Erven et al., 2007; Koolen and de Rooij, 2013). These include the more extreme approach called switching. With this approach, we calculate (9) with \(\pi(k)\) replaced by any distribution we like (the choice is again allowed to depend on \(Y^{(j-1)}\) up to and including a particular batch \(j^*\). Thereafter, for \(j \geq j^*\), we set

\[ \pi^*(k|y^{(j)}) = \begin{cases} 1 & \text{if } k = k^* \text{ with } k^* = \arg \max_k E^{(j^*),k} \\ 0 & \text{otherwise} \end{cases} \]

\[(11)\]

creating a new E-process \(E^{(1)}_{[j^*]} , E^{(2)}_{[j^*]} , \ldots\) such that, for \(m \leq j^*\),

\[ E^{(m)}_{[j^*]} = E^{(m)} \quad \text{and, for } m > j^*, \]

\[ E^{(m)}_{[j^*]} = E^{(j^*)} \cdot \prod_{j=j^*+1}^{m} E^{(j),k^*} \]

\[(12)\]

\(j^*\) could arbitrarily be picked prior to the study, or we could also place a prior on the moment of switching and take a weighted average over (12) for various values of \(j^*\) for each \(k\), thereby obtaining yet another e-process with \(j^*\) ‘integrated out’ (see Figure S2 in the supplementary material for a more elaborate comparison of switch priors in a simulation experiment for confidence sequences).

In Figure 2, the three different methods for combining e-variables for testing \(\mathcal{H}_0\) are compared with respect to power: the expected probability of rejecting \(\mathcal{H}_0\) under some fixed data generating distribution. For Figure 2, data were sampled from a distribution where risk differences and control group rates all differed between strata. It can be observed that all methods that took the stratification into account outperformed the unstratified approach, where just one sequential e-variable was calculated for all strata combined. The three different methods will be re-compared for confidence sequences in Figure 6.

**Cross-talk between strata** To further improve power of the hypothesis test, we will allow for cross-talk between strata while estimating \(\theta_{a,k}\) and \(\theta_{b,k}\) based on data seen so far. In the current simple setting of testing the global null, ‘cross-talk’ simply amounts to design \(S_j\) that grow faster (allowing for faster rejecting of the null) if the alternative satisfies certain constraints. For example, if one expects treatment effects (say, measured as odds ratios) to be stable (identical) throughout different strata, but control group recovery rates to vary, one would like cross-talk about the odds ratios between strata. Practically, this means that to arrive at the estimates \(\hat{\theta}_{a,k} \mid Y^{(j-1)}\), we first limit the parameter space to \(\Theta_1(\hat{\phi}^{(j-1)})\), i.e. all vectors \(\theta_{x,k}\) with odds ratio \(\phi^{(j-1)}\), set to be the maximum likelihood odds ratio based on all previous data in all strata, i.e. calculated by ignoring strata. We then calculate \(\hat{\theta}_{x,k} \mid Y^{(j-1)}\) as posterior means using beta priors conditioned on the parameters being in \(\Theta_1(\hat{\phi}^{(j-1)})\). Similarly, when one expects control group recovery rates to be stable, but the treatment effects
Figure 2: Power for rejecting the null at level $\alpha = 0.05$ that the odds ratio in all strata equals 1 estimated with 1000 repeated experiments for various e-variable combination methods. 40 batches were collected in each of three strata (so maximum sample size was $n = 120$) and sampling was stopped as soon as $E_P(m) > \alpha$. Real control group success rates were 0.1, 0.2, 0.8 and real risk differences were 0.05, 0.4, −0.6. Pseudo-Bayesian approaches were implemented with learning rates (LR) 1 and 2. Switch approaches were implemented for switching at point $j^* = 10$, or with a uniform prior on switch times $j = 5$ until $m - 5$.

control rate or odds ratio improves power in the respective scenarios. The cross-talk mixture performs comparably to the optimal cross-talk options in both cases. Cross-talk can be expected to improve power even if, in truth, under the alternative, the odds-ratio resp. control-group rate is just similar, but not exactly the same under all groups; and the confidence sequences of the next section remain valid (but will get wider) even if the odds-ratios resp. control-group rates happen to be completely different. Thus, the method described here cannot really be viewed as a constraint on the model, and we chose to call it cross-talk instead: data in one stratum informs, ‘talks to’ estimates for other strata.

Figure 3: Power for rejecting the null hypothesis at level $\alpha = 0.05$ that the odds ratio in all strata equals 1 estimated with 100 repeated experiments for various types of cross-talk. 40 batches were collected in each stratum and sampling was stopped as soon as $E_P(m) > \alpha$. On the left, real control group success rates were 0.49, 0.5 and 0.51 in each stratum; risk differences were $-0.09, -0.49, 0.39$. On the right, real odds ratios were 4, 4.01, 2.95.

A GRO-Sanity Check While the simulations above and below show encouraging empirical results regarding the power of our methods, it is still useful to have some theoretical assurance that, no matter the ‘true’ alternative generating the data, all methods we consider produce e-values that grow fast (i.e. achieve good power) under this alternative. We now provide a simple theorem to this end. As usual in the e-value and safe-testing literature, and for reasons explained by Grünwald et al. (2022), we concentrate on GRO (5) rather than power.

**Theorem 2.1.** Suppose that we observe $m = m_1 + \ldots + m_K$ blocks, with $m_k$ blocks lying in stratum $k$, each such block sampled independently from $P_{\theta_{a,k}, \theta_{\theta,k}}$. Then, with $E$ denoting expectation under this distribution, the e-process $E_P(m)$ defined by multiplication as in (8) and the MIX e-process $E_P(m)$ as above with constituent e-processes defined multiplicatively as in (8) both achieve:

$$\sum_{k=1}^{K} m_k \text{GRO}(\theta_{a,k}, \theta_{\theta,k}) = E[\log E_P(m)] + O(\log m). \tag{13}$$

To interpret the result, note that, if an oracle were to sup-
Safe Sequential Testing and Effect Estimation in Stratified Count Data

Figure 4: Examples of 95% stratified confidence intervals ((a), (b) and (c)) and mean confidence interval widths estimated over 100 runs ((d), (e) and (f)) with different types of cross-talk. In (a), (b) and (c) the true risk difference of the data generating distribution in each stratum is indicated by a dashed line. For (a) and (d), the data were generated by distributions with different control group success rates (0.1, 0.2 and 0.8) and risk differences (0.05, 0.4 and −0.6) in each stratum. For (b) and (e), strata sizes were unbalanced: as can be seen for stratum 1, the red points, data collection stopped after 10 batches. Control group success rates were all 0.5 and risk differences were different (−0.49, −0.25 and 0.1). For (c) and (f), strata sizes were unbalanced as well, and now odds ratios were the same in each stratum (2), but control group rates differed again (0.2, 0.25 and 0.85).

(3) in the 2 × 2-table (single stratum) can be generalized, to test null hypotheses $\mathcal{H}_0 := \{P(\theta_a, \theta_b); (\theta_a, \theta_b) \in \Theta_0\}$ beyond ‘$\theta_a = \theta_b$’:

$$S_{j, [\Theta_0]} = \prod_{i=1}^{n_a} P_{\theta_a, Y(j-1)}(Y_{j,a,i}) \prod_{i=1}^{n_b} P_{\theta_b, Y(j-1)}(Y_{j,b,i})$$

is an e-variable, as long as $\Theta_0 \subset [0, 1]^2$ is convex and

3 EXTENSION TO CONFIDENCE SEQUENCES

Turner and Grünwald (2022) showed that (3) in the 2 × 2-table (single stratum) can be generalized, to test null hypotheses $\mathcal{H}_0 := \{P(\theta_a, \theta_b); (\theta_a, \theta_b) \in \Theta_0\}$ beyond ‘$\theta_a = \theta_b$’:
closed. Here \( \hat{\theta}_{a}^{\circ} | Y^{(j-1)} \) is defined to minimize KL divergence, i.e. the pair \((\theta_{a}, \theta_{b}) \in \Theta_{0} \) that minimizes, over \( \Theta_{0} \),
\[
D(P_{\theta_{a}} Y^{(j-1)} \delta_{b} Y^{(j-1)} Y_{a}^{m} Y_{b}^{m} \mid P_{\theta_{a}} Y_{a}^{m} Y_{b}^{m}).
\]
(3) is a special case since with \( \Theta_{0} = \{ (\theta, \theta) : \theta \in [0, 1] \} \), this KL divergence is minimized by \( (\hat{\theta}_{a}^{\circ}, \hat{\theta}_{b}^{\circ}) \) with \( \hat{\theta}^{\circ} \) as defined underneath (3). Again, \( \hat{\theta}_{a} \) and \( \hat{\theta}_{b} \) are estimated based on past data \( Y^{(j-1)} \) as in (3). Based on (14) one can construct an exact (nonasymptotic) confidence sequence (CS)
\[
CS_{a,(m)}(\delta) = \left\{ \delta : E_{[\Theta_{0}]}(\delta) \leq \frac{1}{\alpha} \right\},
\]
with \( \Theta_{0}(\delta) \in [0, 1]^{2} \) a null hypothesis determined by a divergence measure. By construction, such a confidence sequence is always-valid (Ramdas et al., 2022) in the sense that for any \( \delta \), any \( \theta \in \Theta_{0}(\delta) \), the \( P_{\theta} \)-probability that there will ever be an \( m \) such that \( \delta \notin CS_{a,(m)} \) is at most \( \alpha \). This means that we can take the running intersection of the confidence sequence while retaining coverage, which will be used throughout the simulation experiments in this paper. In this paper, we are going to construct confidence sequences for risk differences as examples, where we are going to test hypotheses of the form \( \Theta_{0}(\delta) := \{ (\theta_{a}, \theta_{b}) \in [0, 1]^{2} : \theta_{a} - \theta_{b} = \delta \} \) — below we extend this to the case that differentiates in terms of the strata. Still, everything could also easily be adapted to construct confidence intervals for other divergence measures, such as odds and risk ratios (Turner and Grünwald, 2022).

### 3.1 One CS per stratum

If we expect the effect size values to differ between the strata, one could decide to report a separate confidence sequence for each stratum using (15) above. To reach a better estimate sooner, we could however still allow cross-talk on control group success rates or odds ratios between subpopulations, as described in section 2 above. In this setup, we would end up with a collection of \( k \) confidence sequences:
\[
CS_{a,(m)}^{k}(\delta) = \left\{ \delta : E_{[\Theta_{0}]}(\delta) \leq \frac{1}{\alpha} \right\},
\]
with \( \hat{\theta}_{a} \) and \( \hat{\theta}_{b} \) in \( E^{(m),k} \) estimated based on data seen up to time \( m \) and \( E^{(m),k} \) defined as in (9) with \( S^{k}_{j} \) replaced by \( S^{k}_{j[\Theta_{0}]} \) as in (14), calculated for stratum \( k \). Illustrations of confidence intervals over time with the three options for cross-talk are depicted in Figure 4. As can be observed there, not allowing cross-talk gives the best results when the true data generating distributions in the strata have different control group success rates and odds ratios (see the circle-shaped points in Figure 4d, especially in the third stratum, where the effect size has a different sign). However, when control group rates or odds ratios are similar across strata, allowing cross-talk improves results. See for example Figure 4e, where interval width decreases much faster in the smaller stratum 1 while allowing cross-talk about the control group rate. Similar experiments for comparing confidence sequences with and without the mixture of cross-talk methods can be found in the supplementary material, Figure S1.

### 3.2 CS for the minimum or maximum

In some scenarios, for example when we do not have the means to collect a large data sample, or when data is very unbalanced in one or more strata, it could be more informative to create one CS for the minimum or maximum effect size value over all strata. To achieve this, we introduce two new forms of null hypotheses and corresponding e-variables that will subsequently be inverted to create two one-sided confidence sequences, for lower and upper bounds on the minimum or maximum.

#### One-sided CS: upper bound

We will first illustrate how to estimate an upper bound on some minimal effect size value over strata\(^1\). To this end, we consider a null hypothesis of the form \( \mathcal{H}_{0,\delta} : \forall k : \theta_{k} \in \Theta_{0}(\geq \delta) \) (i.e. for risk difference effect size, \( \Theta_{0}(\geq \delta) = \{ (\theta_{a}, \theta_{b}) \in [0, 1]^{2} : \theta_{a} - \theta_{b} \geq \delta \} \) and aim to design e-variables to test it. E.g. in the example depicted in Figure 5a, we aim to design an e-variable that will reject \( \mathcal{H}_{0,\delta} \) at any batch \( j \) with probability less than \( \alpha \) (i.e., that offers type-I error guarantee), when the data in the strata are in reality generated by \( \theta_{a,1}, \theta_{b,1} \) and \( \theta_{a,2}, \theta_{b,2} \). We do eventually want to reject \( \mathcal{H}_{0,\delta} \) as \( \delta(\theta_{a,2}, \theta_{b,2}) < \delta' \). As we collect more and more data, we can reject null hypotheses corresponding to values of \( \delta' \) for which \( \delta' - \delta(\theta_{a,2}, \theta_{b,2}) \) gets closer and closer to 0.

Let us denote the e-process consisting of the e-variables for testing \( \theta_{k} \in \Theta_{0}(\geq \delta) \) in each stratum combined, using any of the methods described above in Section 2, as \( E^{(m)} \). The one-sided confidence interval for the minimum effect size can be defined as:
\[
CS_{a,(m)}^{+} = \left[ -1, \min \left\{ \delta : E^{(m)} \geq \frac{1}{\alpha} \right\} \right].
\]
All possible approaches for combining e-variables from separate strata, as described in Section 2 above, to find an upper bound for the minimal effect size value are compared in the confidence intervals in the paragraph below.

#### One-sided CS: lower bound

We now also aim to estimate a lower bound for the minimal effect size value (or, analogously, an upper bound for the maximal effect size value). To achieve this, we now consider a null hypothesis of the form \( \mathcal{H}_{0,\delta} : \exists k : \theta_{k} \in \Theta_{0}(\leq \delta) \). Looking at Figure 5b as an example, where data are generated by \( \theta_{a,1}, \theta_{b,1} \) and \( \theta_{a,2}, \theta_{b,2} \), we aim to design an e-variable that will reject \( \mathcal{H}_{0,\delta} \) at any batch \( j \) with probability less than \( \alpha \) (i.e.,
\(^1\)Analogously, with this method a lower bound on some maximal effect size value can be estimated by reversing all signs.)
we again want type-I error guarantee if \( H_{0, \delta} \) is true, as \( \delta((\theta_{a,2}, \theta_{b,2})) < \delta' \). We do want to reject as quickly as possible \( H_{0, \delta''} \), as \( \forall k : \delta((k^{(i)})) > \delta'' \). As we collect more data, we can reject null hypotheses with values of \( \delta'' \) for which \( \delta((\theta_{a,2}, \theta_{b,2})) - \delta'' \) gets closer and closer to 0.

To test \( H_{0, \delta} \) we will use the minimum of \( E_{\alpha, (m), (j)}^{(i)}(\theta_{(i)}) \) over all \( k \), which provides an e-variable for \( H_{0, \delta} \). To see this, let us assume \( H_{0, \delta} \) is true and that for some \( k^{*} \), \( \theta_{k^{*}} \in \Theta_{0}(\leq \delta) \); the other data generating distributions might or might not come from \( \Theta_{0}(\leq \delta) \). Then: \( \mathbb{E}(\min_k S^k) \leq \min_k \mathbb{E}(S^k) \leq \mathbb{E}(S^{k^{*}}) \leq 1 \).

Combining into confidence interval We now combine the lower bound and upper bound estimation methods established above to build confidence intervals for the minimal effect size value. This can be achieved through taking the intersection of the one-sided confidence sequences introduced above: \( CS_{\alpha, (m)} := CS_{\alpha, (m), (1)} \cap CS_{\alpha, (m), (2)} \). Results from an experiment where in one of the strata the treatment effect was substantially smaller than in the others are depicted in Figure 6 (with average interval widths in the supplementary material, S3). In early phases of data collection, multiplication gives the quickest convergence, but as more data is collected, the “sequential learning” methods converge quicker. When risk differences where about the same across all strata, multiplication converged the quickest (see Figure S4 in the Supplementary material).

3.3 CS for the mean effect

In addition to estimating the minimum or maximum effect in one of the strata, one might be interested in estimating the mean effect an intervention will have on an entire population, given the existence of subpopulations. For example, one might want to estimate the effect a vaccination will have on the probability of people being contaminated with a disease, taking into account that a certain proportion of the population concerns elderly or immunocompromised citizens.

Assuming we have a trustworthy estimate of the proportion of subjects belonging to each stratum \( k \) in the population of interest, \( \pi_k \), we aim to estimate the mean risk difference (mean expected effect of the intervention) \( \delta^{*} := \Sigma_k \pi_k \delta_k \). We can build a confidence sequence for \( \delta^{*} \) by construct-
ing an e-variable for the set of all possible success probability distributions satisfying this \( \delta^* \), \( \mathcal{H}_{0, \delta^*} : \{ P_{\theta} : d(\theta) = \sum_k \pi_k d((\theta_{a,k}, \theta_{b,k})) = \delta^* \} \). It is not directly clear what an optimal e-variable could look like; one option that offers both the type-I error guarantee with potentially good power is to combine the growth-rate optimal e-variable (3) for a specific \( \delta_k \) in each stratum with the universal inference (Wasserman et al., 2020) method for designing e-processes. Based on this strategy, we look at the set of all fixed-n designs. This is a strong assumption, and we do not make it ourselves; but we can adapt (get narrow) if the risk difference is indeed homogeneous. One recent fixed-n approach for this setting was described and implemented by Klingenberg (2014). In Figure 7, our confidence sequence for the mean effect is compared to the Miettinen-Nuninen (MN) confidence interval from Klingenberg (2014) at fixed time 50 in a setting where risk differences were homogeneous. The MN-interval is slightly narrower, but because we are allowed to continuously monitor the confidence interval while retaining coverage with the confidence sequence, we can exclude 0 from the CS considerably earlier than with the fixed-n method — which is remarkable because unlike the MN fixed-n confidence interval, our anytime-valid confidence sequences are also valid if in fact risk differences are not homogeneous.

4 CONCLUSION AND FUTURE WORK

We have introduced a new method for global null hypothesis testing and constructing exact anytime-valid confidence sequences in stratified count data. Our method is complementary to previously proposed methods for similar settings as we need no stochastic assumptions about the arrival times of the subgroups or strata, and no Model-X assumptions. We have shown that our tests and estimates are efficient in terms of power, and that precise effect size estimations can be reached with less strong model assumptions compared to pre-existing fixed-n methods, while retaining coverage guarantees and allowing sequential decision making. We have also shown that we can improve the traditional model of global null testing in the CMH-setting through incorporating ideas from machine-learning: allowing for cross-talk between strata, and incorporating pseudo-Bayesian learning and switching between strata for learning compound effect measures.

Our work extends that of Turner et al. (2021) and Turner and Grünwald (2022) to incorporate strata for count data. Their methods, however, are generally implementable for any convex null hypothesis, and future work should explore if they also can feasibly be extended to stratified sequential effect estimation for continuous outcome variables.

4.1 Acknowledgements

This work is part of the Enabling Personalized Interventions (EPI) project, which is supported by the Dutch Research Council (NWO) in the Commit2Data Data2Person program, contract 628.011.028.

References


Duan, B., Ramdas, A., and Wasserman, L. (2022). Inter-


S1 PROOFS

Proof. (of theorem 2.1). First consider the basic case with \( E^{(m)} \) as in (8). As we show below, we have, with \( E \equiv E_{P^*_\theta} \),

\[
E \left[ \log E^{(m)} \right] = E \left[ \sum_{j=1}^{m} \log S_j \right] = E \left[ \sum_{j=1}^{m} \sum_{x \in (a,b)} \sum_{i=1..n_x} \log \frac{p_{\tilde{\theta}_{x,k,j} | Y(j-1)}(Y_j,x,i)}{p_{\tilde{\theta}_{0,k,j} | Y(j-1)}(Y_j,x,i)} \right] \geq \sum_{j=1..m} \sum_{x \in (a,b)} \sum_{i=1..n_x} \log \frac{p_{\tilde{\theta}_{x,k,j} | Y(j-1)}(Y_j,x,i)}{p_{\tilde{\theta}_{0,k,j} | Y(j-1)}(Y_j,x,i)} + O(1) = \sum_{k=1..K} m_k \cdot D(P_{\tilde{\theta}_{k,k}^*, \theta^*_{k,k}} \| P_{\tilde{\theta}_{0,k,k}}) + O(\log m) \tag{18}
\]

where we use notation \( D(P_{\theta^*_{a,k}} \| P_{\theta_0}) \) as in (4); and \( \tilde{\theta}_{0,k} \) is defined as arg \( \min_{\tilde{\theta}_0} D(P_{\tilde{\theta}_{a,k}} \| P_{\theta_0}) \) which by the same calculation as the one leading up to (4), is given by \( \tilde{\theta}_{0,k} = (n_a/n)\theta^*_{a,k} + (n_b/n)\theta^*_{b,k} \), and \( m_k \) denotes the number of times that an instance of block \( k \) was observed in the first \( m \) blocks, and we remind the reader that \( +O(\log m) \) may also indicate a negative difference of order \( \log m \). (18) immediately implies the result, using (6).

The first two equalities in (18) are immediate. The first inequality follows because \( P_{\theta_{0,k}, \tilde{\theta}_{0,k}} \) minimizes KL divergence to \( P_{\tilde{\theta}_{a,k}, \tilde{\theta}_{b,k}} \) among all \( \theta \in [0,1] \), within each block \( j \). The final equality follows by independence and basic calculus.

It remains to show the second inequality. This one follows because we use a prior \( W(\theta_{a,k}, \theta_{b,k}) \) under which \( \theta_a \) and \( \theta_b \) are independently beta distributed with strictly positive densities on (0,1). We can then use a standard Laplace approximation of the Bayesian marginal likelihood to obtain, for each fixed \( k \in \{1, \ldots, K\} \), where the expectation \( E \) is over \( Y^{(1)}, \ldots, Y^{(m')} \sim P_{\tilde{\theta}_{a,k}, \tilde{\theta}_{b,k}} \):

\[
E \left[ -\log \prod_{j=1}^{m'} \prod_{x \in (a,b)} \prod_{i=1}^{n_x} p_{\tilde{\theta}_{x,k} | Y(j)}(Y_j,x,i) \right] = E \left[ -\log \int_{\theta_a} p_{\theta_{a,k}}(Y_{j,x,i}) \, d\theta \right] \leq E \left[ \sum_{j=1}^{m'} -\log p_{\theta_{a,k}, \theta_{b,k}}(Y_{j}) + O(1) \right] + \log(n_a + n_b)m' + O(1).
\]

Here the equality is standard telescoping of the Bayesian marginal likelihood, and the inequality is the Laplace approximation, i.e. the same calculation as the one leading up to the \( (d/2) \log n \) BIC approximation of Bayesian marginal likelihood for a \( d \)-parameter exponential family; here \( d = 2 \) since we have two free parameters, \( \theta^*_{a,k} \) and \( \theta^*_{b,k} \); see (Gr"unwald, 2007, Chapter 8) for proof and detailed explanation).

This shows the result for the basic case that \( E^{(m)} \) is arrived at by multiplication, (8). The case for \( E^{(m)}_{\text{MIX}} \) follows similarly by noting that, by construction, \( E^{(m)}_{\text{MIX}} \geq E^{(m)}_{\text{NONE}}/3 \), where \( E^{(m)}_{\text{NONE}} \) denotes the standard e-process with multiplication and without cross-talk, for which we have already (just) shown the result.
S2 ADDITIONAL EXPERIMENTS

Figure S1: Examples of 95% stratified confidence intervals ((a), (b) and (c)) and mean confidence interval widths estimated over 100 runs ((d), (e) and (f)) with different types of cross-talk, including mixing different types of cross-talk. In (a), (b) and (c) the true risk difference of the data generating distribution in each stratum is indicated by a dashed line. For (a) and (d), the data were generated by distributions with different control group success rates (0.1, 0.2 and 0.8) and risk differences (0.05, 0.4 and −0.6) in each stratum. For (b) and (e), strata sizes were unbalanced: as can be seen for stratum 1, the red points, data collection stopped after 10 batches. Control group success rates were all 0.5 and risk differences were different (−0.49, −0.25 and 0.1). For (c) and (f), strata sizes were unbalanced as well, and now odds ratios were the same in each stratum (2), but control group rates differed again (0.2, 0.25 and 0.85).
Figure S2: Example of a confidence sequence and average difference from upper bound to true minimal effect size value through 100 simulations, for different switch priors on $j^\ast$. 30 observations were made in each stratum, and the real differences were 0.5, 0.4 and 0.05. For the priors on early switch times, all prior mass was distributed between batch numbers 5 up to 10. $\alpha$ was set to 0.05.

Figure S3: Average interval width (upper bound for the respective methods minus lower bound estimated with the minimum method) of confidence sequences for the lower- (LB) and upper (UB) bounds of the minimum effect and estimated through 100 simulations. 30 observations were made in each stratum, and the real differences were 0.5, 0.4 and 0.05. With the switch method, a uniform prior ranging from $j^\ast = 5$ until 30 was applied. With the pseudo-Bayesian approach, the learning rate $\eta$ was set to 1 and 2. $\alpha$ was set to 0.05.
Figure S4: Example of confidence sequences for the lower- (LB) and upper (UB) bounds of the minimum effect, and average interval width (upper bound for the respective methods minus lower bound estimated with the minimum method). 30 observations were made in each stratum, and the real differences were 0.4, 0.4 and 0.5. With the switch method, a uniform prior ranging from $m_{\text{switch}} = 5$ until 30 was applied. With the pseudo-Bayesian approach, the learning rate $\eta$ was set to 1 and 2. $\alpha$ was set to 0.05.

Figure S5: Simulated example of a confidence sequence for the mean effect across subpopulations. 25 observations were made in each stratum, and the real risk differences were 0.2 and 0.5. The confidence sequence for the mean difference is plotted alongside the confidence sequence for the minimum of the differences, estimated with pseudo-Bayesian averaging and a uniform switch prior. $\alpha$ was set to 0.05.