Safe anytime-valid inference: from theory to implementation in psychiatry research

Turner, R.J.

Chapter 3

Exact Anytime-valid Confidence Intervals for Contingency Tables and Beyond

Rosanne J. Turner[1,2], Peter D. Grünwald[1,3]

1: CWI, Machine Learning group, Netherlands
2: University Medical Center Utrecht, Brain Center, Netherlands
3: Leiden University, Department of Mathematics, Netherlands

Abstract

E-variables are tools for retaining type-I error guarantee with optional stopping. We extend E-variables for sequential two-sample tests to general null hypotheses and anytime-valid confidence sequences. We provide implementations for estimating risk difference, relative risk and odds-ratios in contingency tables.

3.1 Introduction

We consider a setting where we collect samples from two distinct groups, denoted a and b. In both groups, data come in sequentially and are i.i.d. We thus have two data streams, $Y_{1,a}, Y_{2,a}, \ldots$ i.i.d. $\sim P_{\theta_a}$ and $Y_{1,b}, Y_{2,b}, \ldots$ i.i.d. $\sim P_{\theta_b}$ where we assume that $\theta_a, \theta_b \in \Theta$, $\{P_\theta : \theta \in \Theta\}$ representing some parameterized underlying family of distributions, all assumed to have a probability density or mass function denoted by p_θ on some outcome space $\mathcal{Y}$.

e-variables [Grünwald et al., 2022a, Vovk and Wang, 2021] are a tool for constructing tests that keep their Type-I error control under optional stopping and continuation. Previously, Turner et al. [2021] developed e-variables for testing equality of both data streams, i.e. with null hypothesis $\vec{\Theta}_0 := \{(\theta_a, \theta_b) \in \Theta^2 : \theta_a = \theta_b\}$. Here we first generalize these e-variables to more general null hypotheses in which we may have $\theta_a \neq \theta_b$. We then use these generalized e-variables to construct *anytime-valid* confidence sequences; these provide confidence sets that remain valid under optional stopping [Darling and Robbins, 1967, Howard et al., 2021].

As in [Turner et al., 2021], we first design e-variables for a *single block* of data $(Y_a^{n_a}, Y_b^{n_b})$, where a block is a set of data consisting of n_a outcomes $Y_a^{n_a} = (Y_{a,1}, \ldots, Y_{a,n_a})$ in group a and n_b outcomes $Y_b^{n_b} = (Y_{b,1}, \ldots, Y_{b,n_b})$ in group b, for some pre-specified n_a and n_b. An *e-variable* is then, by definition, any nonnegative random variable $S = s'(Y_a^{n_a}, Y_b^{n_b})$ such that

$$\sup_{(\theta_a,\theta_b)\in\vec{\Theta}_0} \mathbf{E}_{Y_a^{n_a}\sim P_{\theta_a}, Y_b^{n_b}\sim P_{\theta_b}} [s'(Y_a^{n_a}, Y_b^{n_b})] \leq 1. \tag{3.1}$$

Turner et al. [2021] first defined such an e-variable for $\vec{\Theta}_0 = \{(\in \Theta^2 : \theta_a = \theta_b\}$ so that it would tend to have high power against a given simple alternative $\vec{\Theta}_1 = \{(\theta_a^*, \theta_b^*)\}$. Their e-variable is of the following simple form (with $n = n_a + n_b$):

$$s'(Y_a^{n_a}, Y_b^{n_b}) = \frac{p_{\theta_a^*}(Y_a^{n_a})}{\prod_{i=1}^{n_a}(\frac{n_a}{n} p_{\theta_a^*}(Y_{a,i}) + \frac{n_b}{n} p_{\theta_b^*}(Y_{a,i}))} \cdot \frac{p_{\theta_b^*}(Y_b^{n_b})}{\prod_{i=1}^{n_b}(\frac{n_a}{n} p_{\theta_a^*}(Y_{b,i}) + \frac{n_b}{n} p_{\theta_b^*}(Y_{b,i}))}. \tag{3.2}$$

These e-variables can be extended to sequences of blocks $Y_{(1)}, Y_{(2)}, \ldots$ by multiplication, and can be extended to composite alternatives by sequentially learning (θ_a^*, θ_b^*) from the data, for example via a Bayesian prior on $\vec{\Theta}_1$. The n_a and n_b used for the j-th block $Y_{(j)}$ are allowed to depend on past data, but they must be fixed before the first observation in block j occurs. For simplicity, in this note we only consider the case with n_a and n_b that remain fixed throughout; extension to the general case is straightforward.

By a general property of e-variables, at each point in time, the running product of block e-variables observed so far is itself an e-variable, and the random process of the products is known as a *test martingale* [Grünwald et al., 2022a, Shafer et al., 2021]. An e-variable-based test at level α is a test which, in combination with any stopping rule τ, reports 'reject' if and only if the product of e-values correspond-

ing to all blocks that were observed at the stopping time and have already been completed, is larger than $1/\alpha$. Such a test has a type-I error probability bounded by α irrespective of the stopping time τ that was used; see the aforementioned references for much more detailed introductions and, for example [Henzi and Ziegel, 2022], for a practical application.

In case $\{P_\theta : \theta \in \Theta\}$ is convex, the e-variable (3.2) has the so-called GRO- (*growth-rate-optimality*) property: it maximizes, over all e-variables (i.e. over all nonnegative random variables $S = s'(Y_a^{n_a}, Y_b^{n_b})$ satisfying (3.1)) the logarithmic growth rate

$$\mathbf{E}_{Y_a^{n_a} \sim P_{\theta_a^*}, Y_b^{n_b} \sim P_{\theta_b^*}} [\log S], \tag{3.3}$$

which implies that, under (θ_a^*, θ_b^*), the expected number of data points before the null can be rejected is minimized [Grünwald et al., 2022a].

Below, in Theorem 3.1 in section 3.2, which generalizes Theorem 1 in Turner et al. [2021], we extend (3.2) to the case of general null hypotheses, $\vec{\Theta}_0 \subset \Theta^2$, allowing for the case that the elements of $\vec{\Theta}_0$ have two different components, and provide a condition under which it has the GRO property. From then onwards we focus on what we call 'the 2×2 contingency table setting' in which both streams are Bernoulli, θ_j denoting the probability of 1 in group j. For this case, Theorem 3.2 gives a simplified expression for the e-variable and shows that the GRO property holds if $\vec{\Theta}_0 \subset [0,1]^2$ is convex. Then we will extend this e-variable to deal with composite $\vec{\Theta}_1$ and use this to define anytime-valid confidence sequences. We illustrate these through simulations. All proofs are in Appendix S3.A.

3.2 General Null Hypotheses

In this section, we first construct an e-variable for general null hypotheses that generalizes (3.2). We then instantiate the new result to the 2×2 case. The following development and results require $\{P_\theta : \theta \in \Theta\}$ to be 'nondegenerate' in the sense that there exists $\theta \in \Theta$ such that for all $\theta' \in \Theta$, $D(P_\theta \| P_{\theta'}) < \infty$. This mild condition holds, for example, for exponential families; we tacitly assume nondegeneracy from now on.

Our goal is thus to define an e-variable for a block of $n = n_a + n_b$ data points with n_g points in group g, $g \in \{a, b\}$. For notational convenience we define, for $\theta_a, \theta_b \in \Theta$, P_{θ_a,θ_b} as the joint distribution of $Y_a^{n_a} \sim P_{\theta_a}$ and $Y_b^{n_b} \sim P_{\theta_b}$, so that $p_{\theta_a,\theta_b}(y_a^{n_a}, y_b^{n_b}) = \prod_{i=1}^{n_a} p_{\theta_a}(y_{a,i}) \prod_{i=1}^{n_b} p_{\theta_b}(y_{b,i})$ so that we can write the null hypothesis as $\mathcal{H}_0 := \{P_{\theta_a,\theta_b} : (\theta_a, \theta_b) \in \vec{\Theta}_0\}$. Our strategy will be to first develop an e–variable for a *modified* setting in which there is only a single outcome, falling with probability n_a/n in group a and n_b/n in group b. To this end, for $\vec{\theta} = (\theta_a, \theta_b)$, we define $p'_{\vec{\theta}}(Y|a) := p_{\theta_a}(y), p'_{\vec{\theta}}(Y|b) := p_{\theta_b}(y)$, all distributions with a $'$ refering to the modified setting with just one outcome. We let $\mathcal{W}^\circ(\vec{\Theta}_0)$ be the set of all distributions on $\vec{\Theta}_0$ with finite support. For $W \in \mathcal{W}^\circ(\vec{\Theta}_0)$, we define $p'_W(Y|g) = \int p'_{\vec{\theta}}(Y|g) dW(\vec{\theta})$. We set $p'_W(y^k|g) := \prod_{i=1}^k p'_W(y_i|g)$. We further define, for given alternative $\vec{\Theta}_1 = \{(\theta_a^*, \theta_b^*)\}$, $p^\circ(\cdot|g)$, $g \in \{a, b\}$ to be, if it exists,

the conditional probability density satisfying

$$\mathbf{E}_{G\sim Q'}\mathbf{E}_{Y\sim P_{\theta^*_G}}[-\log p^\circ(Y \mid G)] = \inf_{W\in\mathcal{W}^\circ(\vec{\Theta}_0)} \mathbf{E}_{G\sim Q'}\mathbf{E}_{Y\sim P_{\theta^*_G}}[-\log p'_W(Y \mid G)] \tag{3.4}$$

with $Q'(G)$ the distribution for $G \in \{a, b\}$ with $Q'(G = a) = n_a/n$. Clearly we can rephrase (3.4) equivalently as:

$$D(Q'(G,Y)\|P^\circ(G,Y)) = \inf_{W\in\mathcal{W}^\circ(\vec{\Theta}_0)} D(Q'(G,Y)\|P'_W(G,Y)), \tag{3.5}$$

where D is the KL divergence. Here we extended the conditional distributions $P'_W(Y|G)$ and $P^\circ(Y|G)$ (corresponding to densities $p'_W(Y|G)$ and $p^\circ(Y|G)$) to a joint distribution by setting $P'_W(G,Y) := Q'(G)P'_W(Y|G)$ (and similarly for P°) and we extended $Q'(G,Y) := Q'(G)P_{\theta^*_G}(Y)$. We have now constructed a modified null hypothesis $\mathcal{H}'_0 = \{P'_{\vec{\theta}}(G,Y) : \vec{\theta} \in \vec{\Theta}_0\}$ of joint distributions for a single 'group' outcome $G \in \{a, b\}$ and 'data' outcome $Y \in \mathcal{Y}$. We let $\bar{\mathcal{H}}'_0 = \{P_W(G,Y) : W \in \mathcal{W}^\circ(\vec{\Theta}_0)\}$ be the convex hull of $\mathcal{H}'_0$.

The p° satisfying (3.5) is commonly called the *reverse information projection* of Q' onto $\bar{\mathcal{H}}'_0$. Li [1999] shows that p° always exists under our nondegeneracy condition, though in some cases it may represent a sub-distribution (integrating to strictly less than one); see [Grünwald et al., 2022a, Theorem 1] (re-stated for convenience in the supplementary material) who, building on Li's work, established a general relation between reverse information projection and e–variables. Part 1 of that theorem establishes that if the minimum in (3.4) (or (3.5)) is achieved by some $W^\circ \in \mathcal{W}^\circ$ then $p^\circ(\cdot|\cdot) = p'_{W^\circ}(\cdot|\cdot)$ and, with $\vec{\theta}^* = (\theta^*_a, \theta^*_b)$, for all $\vec{\theta} \in \vec{\Theta}_0$,

$$\mathbf{E}_{G\sim Q'}\mathbf{E}_{Y\sim P'_{\vec{\theta}}|G}\left[\frac{p'_{\vec{\theta}^*}(Y|G)}{p^\circ(Y|G)}\right] = \mathbf{E}_{G\sim Q'}\mathbf{E}_{Y\sim P'_{\vec{\theta}}|G}\left[\frac{p'_{\vec{\theta}^*}(G,Y)}{p^\circ(G,Y)}\right] \le 1. \tag{3.6}$$

This expresses that $p'_{\vec{\theta}^*}(Y|G)/p^\circ(Y|G)$ is an e-variable for our modified problem, in which within a single block we observe a single outcome in group g, with g chosen with probability n_g/n. If we were to interpret the e–variable of the modified problem as in (3.6) as a likelihood ratio for a single outcome, its corresponding likelihood ratio for a single block of data in our original problem with n_g outcomes in group g would be:

$$\begin{aligned} s(y_a^{n_a}, y_b^{n_b}; n_a, n_b, (\theta^*_a, \theta^*_b); \vec{\Theta}_0) &:= \frac{p'_{(\theta^*_a,\theta^*_b)}(y_a^{n_a}|a)p'_{(\theta^*_a,\theta^*_b)}(y_b^{n_b}|b)}{p^\circ(y_a^{n_a}|a)p^\circ(y_b^{n_b}|b)} \\ &= \frac{p_{\theta^*_a}(y_a^{n_a})p_{\theta^*_b}(y_b^{n_b})}{p^\circ(y_a^{n_a}|a)p^\circ(y_b^{n_b}|b)}. \end{aligned} \tag{3.7}$$

The following theorem expresses that this 'extension' of the e-variable in the modified problem gives us an e-variable in our original problem:

Theorem 3.1. $S_{[n_a,n_b,\theta^*_a,\theta^*_b;\vec{\Theta}_0]} := s(Y_a^{n_a}, Y_b^{n_b}; n_a, n_b, (\theta^*_a, \theta^*_b); \vec{\Theta}_0)$ as in (3.7) is an

E-variable, i.e. with $s'(\cdot) = s(\cdot; n_a, n_b, (\theta_a^*, \theta_b^*); \vec{\Theta}_0)$, we have (3.1). Moreover, if $\mathcal{H}_0' = \{P_{\vec{\theta}}' : \vec{\theta} \in \vec{\Theta}_0\}$ (the null hypothesis for the *modified* problem) is a convex set of distributions and $\mathcal{Y}$ is finite (so that $\mathcal{H}_0' = \bar{\mathcal{H}}_0'$) and furthermore $\mathcal{H}_0'$ is compact in the weak topology, then (a) $p^\circ(\cdot|\cdot) = p'_{\vec{\theta}}(\cdot|\cdot)$ for some $\vec{\theta} \in \vec{\Theta}_0$ and (b) $S_{[n_a, n_b, \theta_a^*, \theta_b^*; \vec{\Theta}_0]}$ is the (θ_a^*, θ_b^*)-GRO e-variable for the *original* problem, maximizing (3.3) among all e-variables.

In the case that $\mathcal{H}_0'$ is not convex and compact, we do not have a simple expression for p° in general, and we may have to find it numerically by minimizing (3.4). In the 2×2 table (Bernoulli Θ) case though, there are interesting $\mathcal{H}_0$ for which the corresponding $\mathcal{H}_0'$ is convex, and we shall now see that this leads to major simplifications.

3.2.1 General Convex $\vec{\Theta}_0$ for the 2×2 contingency table

In this subsection and the next, $\{P_{\theta_a, \theta_b}\}$ refers to the 2×2 model again, with $\mathcal{Y} = \{0, 1\}$ and θ denoting the probability of 1. We now let $\vec{\Theta}_0$ be any closed convex subset of $[0, 1]^2$ that contains a point in the interior of $[0, 1]^2$. Again, note that the corresponding $\mathcal{H}_0 = \{P_{\vec{\theta}} : \vec{\theta} \in \vec{\Theta}_0\}$ need not be convex; still, $\mathcal{H}_0'$, the null hypothesis for the modified problem as defined above, must be convex if $\vec{\Theta}_0$ is convex, and this will allow us to design e-variables for such $\vec{\Theta}_0$. Let $\mathcal{H}_1 = \{P_{\theta_a^*, \theta_b^*}\}$ with (θ_a^*, θ_b^*) in the interior of $[0, 1]^2$, and let

$$\mathrm{KL}(\theta_a, \theta_b) := D(P_{\theta_a^*, \theta_b^*}(Y_a^{n_a}, Y_b^{n_b}) \| P_{\theta_a, \theta_b}(Y_a^{n_a}, Y_b^{n_b})) = \sum_{y_a^{n_a} \in \{0,1\}^{n_a}, y_b^{n_b} \in \{0,1\}^{n_b}} p_{\theta_a^*}(y_a^{n_a}) p_{\theta_b^*}(y_b^{n_b}) \log \frac{p_{\theta_a^*}(y_a^{n_a}) p_{\theta_b^*}(y_b^{n_b})}{p_{\theta_a}(y_a^{n_a}) p_{\theta_b}(y_b^{n_b})} \quad (3.8)$$

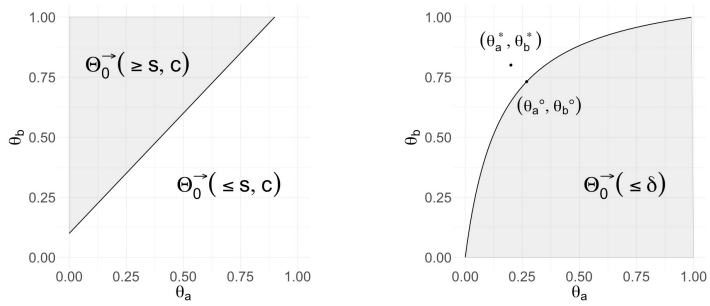
stand for the KL divergence between $P_{\theta_a^*, \theta_b^*}$ and P_{θ_a, θ_b} restricted to a single block (note that in the previous subsection, KL divergence was defined for a single outcome Y). The following result builds on Theorem 3.1:

Theorem 3.2. $\min_{(\theta_a, \theta_b) \in \vec{\Theta}_0} \mathrm{KL}(\theta_a, \theta_b)$ is uniquely achieved by some $(\theta_a^\circ, \theta_b^\circ)$. If $(\theta_a^*, \theta_b^*) \in \vec{\Theta}_0$, then $(\theta_a^\circ, \theta_b^\circ) = (\theta_a^*, \theta_b^*)$. Otherwise, $(\theta_a^\circ, \theta_b^\circ)$ lies on the boundary of $\vec{\Theta}_0$, but not on the boundary of $[0, 1]^2$. The e–variable (3.7) is given by the distribution W that puts all its mass on $(\theta_a^\circ, \theta_b^\circ)$, i.e.

$$s(y_a^{n_a}, y_b^{n_b}; n_a, n_b, (\theta_a^*, \theta_b^*); \vec{\Theta}_0) = \frac{p_{\theta_a^*}(y_a^{n_a}) p_{\theta_b^*}(y_b^{n_b})}{p_{\theta_a^\circ}(y_a^{n_a}) p_{\theta_b^\circ}(y_b^{n_b})} \quad (3.9)$$

is an e-variable. Moreover, this is the (θ_a^*, θ_b^*)-GRO e-variable relative to $\vec{\Theta}_0$.

We can extend this e-variable to the case of a composite $\mathcal{H}_1 = \{P_{\theta_a, \theta_b} : (\theta_a, \theta_b) \in \vec{\Theta}_1\}$ by *learning* the true $(\theta_a^*, \theta_b^*) \in \vec{\Theta}_1$ from the data [Turner et al., 2021]. We thus replace, for each $j = 1, 2, \ldots$, for the block $Y_{(j)}$ consisting of n_a points $Y_{(j),a,1}, \ldots, Y_{(j),a,n_a}$ in group a and n_b points $Y_{(j),b,1}, \ldots, Y_{(j),b,n_b}$ in group



(a) linear boundary (b) odds boundary

Figure 3.1: Examples of null hypothesis parameter spaces for two types of boundaries.

b, the 'true' θ_g^* for $g \in \{a, b\}$ by an estimate $\breve{\theta}_g \mid Y^{(j-1)}$ based on the previous $j-1$ data blocks. The e-variable corresponding to m blocks of data then becomes

$$S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0]} = \prod_{j=1}^{m} \prod_{i=1}^{n_a} \frac{p_{\breve{\theta}_a|Y^{(j-1)}}(Y_{(j),a,i})}{p_{\breve{\theta}_a^\circ|Y^{(j-1)}}(Y_{(j),a,i})} \prod_{i=1}^{n_b} \frac{p_{\breve{\theta}_b|Y^{(j-1)}}(Y_{(j),b,i})}{p_{\breve{\theta}_b^\circ|Y^{(j-1)}}(Y_{(j),b,i})} \tag{3.10}$$

where, for $g \in \{a, b\}$, $\breve{\theta}_g|Y^{(j-1)}$ can be an arbitrary estimator (function from $Y^{(j-1)}$ to θ_g) and $(\breve{\theta}_a^\circ \mid Y^{(j-1)}, \breve{\theta}_b^\circ \mid Y^{(j-1)})$ is defined to achieve $\min_{(\theta_a,\theta_b)\in\vec{\Theta}_0} D(P_{\breve{\theta}_a|Y^{(j-1)},\breve{\theta}_b|Y^{(j-1)}}(Y_a^{n_a}, Y_b^{n_b}) \| P_{\theta_a,\theta_b}(Y_a^{n_a}, Y_b^{n_b}))$.
No matter what estimator we choose, (3.10) gives us an e-variable. In Section 3.3, as in [Turner et al., 2021], we implement this estimator by fixing a prior W and using the Bayes posterior mean, $\breve{\theta}_g|Y^{(j-1)} := \mathbf{E}_{\theta_g \sim W|Y^{(j-1)}}[\theta_g]$. Let us now illustrate Theorem 3.2 for two choices of $\vec{\Theta}_0$.

$\vec{\Theta}_0$ with linear boundary First, we let $\vec{\Theta}_0(s, c)$, for $s \in \mathbf{R}, c \in \mathbf{R}$, stand for any straight line through $[0,1]^2$: $\vec{\Theta}_0(s, c) := \{(\,) \in [0,1]^2 : \theta_b = s + c\theta_a\}$. This can be extended to $\vec{\Theta}_0(\leq s, c) := \bigcup_{s' \leq s} \vec{\Theta}_0(s', c)$ and similarly to $\vec{\Theta}_0(\geq s, c) := \bigcup_{s' \geq s} \vec{\Theta}_0(s', c)$. For example, we could take $\vec{\Theta}_0 = \vec{\Theta}_0(s, c)$ to be the solid line in Figure 3.1(a) (which would correspond to $s = 0.1, c = 1$), or the whole area underneath the line ($\vec{\Theta}_0(\leq s, c)$) including the line itself, or the whole area above it including the line itself ($\vec{\Theta}_0(\geq s, c)$). Now consider a $\vec{\Theta}_0(s, c)$ that has nonempty intersection with the interior of $[0,1]^2$ and that is separated from the point alternative (θ_a^*, θ_b^*), i.e. $\min_{(\theta_a,\theta_b)\in\vec{\Theta}_0} \mathrm{KL}(\theta_a, \theta_b) > 0$. Utilizing the independence of the observations, we can rewrite (3.8) as follows:

$$\mathrm{KL}(\theta_a, \theta_b) := n_a \mathbf{E}_{Y \sim p_{\theta_a^*}} \left[\log \frac{p_{\theta_a^*}(Y)}{p_{\theta_a}(Y)} \right] + n_b \mathbf{E}_{Y \sim p_{\theta_b^*}} \left[\log \frac{p_{\theta_b^*}(Y)}{p_{\theta_b}(Y)} \right].$$

As we defined θ_b to be completely determined as $\theta_b = s + c\theta_a$, substituting and combining with simple differentiation w.r.t. θ_a gives that the minimum is achieved by the unique $(\theta_a^\circ, \theta_b^\circ) \in \vec{\Theta}_0$ satisfying:

$$n_a \left(-\frac{\theta_a^*}{\theta_a^\circ} + \frac{1-\theta_a^*}{1-\theta_a^\circ} \right) + n_b \cdot c \cdot \left(-\frac{\theta_b^*}{\theta_b^\circ} + \frac{1-\theta_b^*}{1-\theta_b^\circ} \right) = 0. \tag{3.11}$$

This can now be plugged into the e-variable (3.9) if the alternative is the simple alternative, or otherwise into its sequential form (3.10). In the basic case in which $\vec{\Theta}_0 = \{(^) \in [0,1]^2 : \theta_a = \theta_b\}$, the solution to (3.11) reduces to the familiar $\theta_a^\circ = \theta_b^\circ = (n_a\theta_a^* + n_b\theta_b^*)/n$ from Turner et al. [2021].

If (θ_a^*, θ_b^*) lies above the line $\vec{\Theta}_0(s,c)$, then by Theorem 3.2, $\min_{(\theta_a,\theta_b)\in\vec{\Theta}_0(\leq s,c)} \mathrm{KL}(\theta_a, \theta_b)$ must lie on $\vec{\Theta}_0(s,c)$. Theorem 3.2 gives that it must be achieved by the $(\theta_a^\circ, \theta_b^\circ)$ satisfying (3.11). Similarly, if (θ_a^*, θ_b^*) lies below the line $\vec{\Theta}_0(s,c)$, then $\min_{(\theta_a,\theta_b)\in\vec{\Theta}_0(\geq s,c)} \mathrm{KL}(\theta_a, \theta_b)$ is again achieved by the $(\theta_a^\circ, \theta_b^\circ)$ satisfying (3.11).

$\vec{\Theta}_0$ with log odds ratio boundary Similarly, we can consider $\vec{\Theta}_0(\delta)$, $\vec{\Theta}_0(\leq\delta)$, $\vec{\Theta}_0(\geq\delta)$ that correspond to a given log odds effect size δ. That is, we now take

$$\begin{aligned}
\vec{\Theta}_0(\delta) &:= \left\{ (\theta_a, \theta_b) \in [0,1]^2 : \log \frac{\theta_b(1-\theta_a)}{(1-\theta_b)\theta_a} = \delta \right\} \\
\vec{\Theta}_0(\leq\delta) &:= \left\{ (\theta_a, \theta_b) \in [0,1]^2 : \log \frac{\theta_b(1-\theta_a)}{(1-\theta_b)\theta_a} \leq \delta \right\} \\
\vec{\Theta}_0(\geq\delta) &:= \left\{ (\theta_a, \theta_b) \in [0,1]^2 : \log \frac{\theta_b(1-\theta_a)}{(1-\theta_b)\theta_a} \geq \delta \right\}.
\end{aligned}$$

For example, we could now take $\vec{\Theta}_0 = \vec{\Theta}_0(\leq\delta)$ to be the area under the curve (including the curve boundary itself) in Figure 3.1(b), which would correspond to $\delta = 2$. Now let δ and point alternative (θ_a^*, θ_b^*) be such that $\delta > 0$ and $\vec{\Theta}_0(\leq\delta)$ is separated from (θ_a^*, θ_b^*), i.e. $\min_{(\theta_a,\theta_b)\in\vec{\Theta}_0(\leq\delta)} \mathrm{KL}(\theta_a, \theta_b) > 0$. Let $(\theta_a^\circ, \theta_b^\circ) := \arg\min_{(\theta_a,\theta_b)\in\vec{\Theta}_0(\delta)} \mathrm{KL}(\theta_a, \theta_b)$. As Figure 3.1(b) suggests, $\vec{\Theta}_0(\leq\delta)$ is convex. Theorem 3.2 now tells us that $\min_{(\theta_a,\theta_b)\in\vec{\Theta}_0(\leq\delta)} \mathrm{KL}(\theta_a, \theta_b)$ is achieved by $(\theta_a^\circ, \theta_b^\circ)$. Plugging these into (3.9) thus gives us an e-variable. $(\theta_a^\circ, \theta_b^\circ)$ can easily be determined numerically. Similarly, if $\delta < 0$, $\vec{\Theta}_0(\geq\delta)$ is convex and closed and if (θ_a^*, θ_b^*) is separated from $\vec{\Theta}_0(\geq\delta)$, the $(\theta_a^\circ, \theta_b^\circ)$ minimizing KL on $\vec{\Theta}_0(\delta)$ gives an e-variable relative to $\vec{\Theta}_0(\geq\delta)$.

3.3 Anytime-valid confidence sequences for the 2×2 case

We will now use the e-variables defined above to construct anytime-valid confidence sequences. Let $\delta = \delta(\theta_a, \theta_b)$ be a notion of effect size such as the log odds ratio (see above) or absolute risk $\theta_b - \theta_a$ or relative risk θ_b/θ_a. A $(1-\alpha)$-*anytime-valid*

(AV) confidence sequence [Darling and Robbins, 1967, Howard et al., 2021] is a sequence of random (i.e. determined by data) subsets $\mathrm{CS}_{\alpha,(1)}, \mathrm{CS}_{\alpha,(2)}, \ldots$ of Γ, with $\mathrm{CS}_{\alpha,(m)}$ being a function of the first m data blocks $Y^{(m)}$, such that for all $(\theta_a, \theta_b) \in [0,1]^2$,

$$P_{\theta_a,\theta_b}\left(\exists m \in \mathbf{N} : \delta(\theta_a,\theta_b) \notin \mathrm{CS}_{\alpha,(m)}\right) \leq \alpha.$$

We first consider the case in which for all values $\gamma \in \Gamma$ that δ can take, $\vec{\Theta}_0(\gamma) := \{(\theta_a,\theta_b) \in [0,1]^2 : \delta(\theta_a,\theta_b) = \gamma\}$ is a convex set, as it will be for absolute and relative risk. Fix a prior W_1 on $[0,1]^2$. Based on (3.10) we can make an *exact* (nonasymptotic) AV confidence sequence

$$\mathrm{CS}_{\alpha,(m)} = \left\{\delta : S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0(\delta)]} \leq \frac{1}{\alpha}\right\} \tag{3.12}$$

where $S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0(\delta)]}$ is defined as in (3.10) and is a valid e-variable by Theorem 3.2. To see that $(\mathrm{CS}_{\alpha,(m)})_{m\in\mathbf{N}}$ really is an AV confidence sequence, note that, by definition of the $\mathrm{CS}_{\alpha,(m)}$, we have
$P_{\theta_a,\theta_b}\left(\exists m \in \mathbf{N} : \delta(\theta_a,\theta_b) \notin \mathrm{CS}_{\alpha,(m)}\right)$ is given by

$$P_{\theta_a,\theta_b}\left(\exists m \in \mathbf{N} : S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0(\delta)]} \geq \frac{1}{\alpha}\right) \leq \alpha,$$

by Ville's inequality [Grünwald et al., 2022a, Turner et al., 2021]. Here the $\mathrm{CS}_{\alpha,(m)}$ are not necessarily intervals, but, potentially losing some information, we can make a AV confidence sequence consisting of intervals by defining $\mathrm{CI}_{\alpha,(m)}$ to be the smallest interval containing $\mathrm{CS}_{\alpha,(m)}$. We can also turn any confidence sequences $(\mathrm{CS}_{\alpha,(m)})_{m\in\mathbf{N}}$ into an alternative AV confidence sequence with sets $\mathrm{CS}'_{\alpha,(m)}$ that are always a subset of $\mathrm{CS}_{\alpha,(m)}$ by taking the *running intersection*

$$\mathrm{CS}'_{\alpha,(m)} := \bigcap_{j=1..m} \mathrm{CS}_{\alpha,(j)}.$$

In this form, the confidence sequences $\mathrm{CS}'_{\alpha,(m)}$ can be interpreted as *the set of δ's that have not yet been rejected* in a setting in which, for each null hypothesis $\vec{\Theta}_0(\delta)$ we stop and reject as soon as the corresponding e-variable exceeds $1/\alpha$. The running intersection can also be applied to the intervals $(\mathrm{CI}_{\alpha,(m)})_{m\in\mathbf{N}}$. To simplify calculations, it is useful to take W_1 a prior under which θ_a and θ_b have independent beta distributions with parameters $\alpha_a, \beta_a, \alpha_b, \beta_b$. We can, if we want, infuse some prior knowledge or hopes by setting these parameters to certain values — our confidence sequences will be valid irrespective of our choice [Howard et al., 2021]. In case no such knowledge can be formulated (as in the simulations below), we advocate the prior, which, among all priors of the simple form asymptotically achieves the REGROW criterion (a criterion related to minimax log-loss regret, see [Grünwald et al., 2022a]), i.e for the case $n_a = n_b = 1$ we set W_1 to an independent beta prior on θ_a and θ_b with $\gamma = 0.18$ as was empirically found to be the 'best'

value [Turner et al., 2021].

Log Odds Ratio Effect Size The situation is slightly trickier if we take the log odds ratio as effect size, for $\vec{\Theta}_0(\delta)$ is then not convex. Without convexity, Theorem 3.2 cannot be used and hence the validity of AV confidence sequences as constructed above breaks down. We can get nonasymptotic anytime-valid confidence sequences after all as follows. First, we consider a one-sided AV confidence sequence for the submodel of positive effect sizes $\{(\theta_a, \theta_b) : \delta(\theta_a, \theta_b) \geq 0\}$, defining

$$\mathrm{CS}^+_{\alpha,(m)} = \{\delta \geq 0 : S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0(\leq\delta)]} \leq \alpha^{-1}, \}$$

where we note that $\vec{\Theta}_0(\leq \delta)$ is convex (since $\delta \geq 0$) and also contains (θ_a, θ_b) with $\delta(\theta_a, \theta_b) < 0$. This confidence sequence can give a lower bound on δ. Analogously, we consider a one-sided AV confidence sequence for the submodel $\{(\theta_a, \theta_b) : \delta(\theta_a, \theta_b) \leq 0\}$, defining

$$\mathrm{CS}^-_{\alpha,(m)} = \{\delta \leq 0 : S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0(\geq\delta)]} \leq \alpha^{-1}\},$$

and derive an upper bound on δ. By Theorem 3.2, both sequences $(\mathrm{CS}^+_{\alpha,(m)})_{m=1,2,\ldots}$ and $(\mathrm{CS}^-_{\alpha,(m)})_{m=1,2,\ldots}$ are AV confidence sequences for the submodels with $\delta \geq 0$ and $\delta \leq 0$ respectively. Defining $\mathrm{CS}_{\alpha,(m)} = \mathrm{CS}^+_{\alpha,(m)} \cup \mathrm{CS}^-_{\alpha,(m)}$, we find, for (θ_a, θ_b) with $\delta(\theta_a, \theta_b) > 0$,

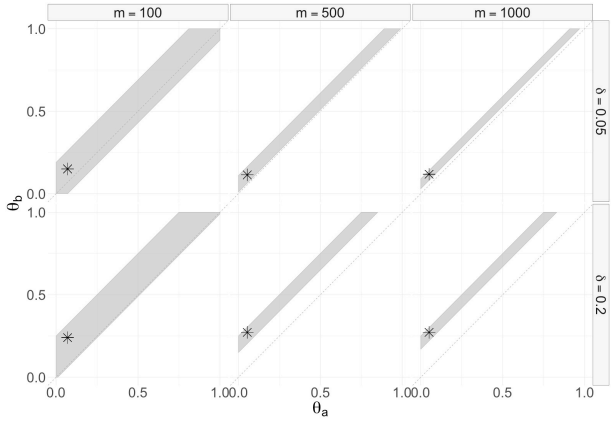
$$P_{\theta_a,\theta_b}\left(\exists m \in \mathbf{N} : \delta(\theta_a, \theta_b) \notin \mathrm{CS}_{\alpha,(m)}\right) = P_{\theta_a,\theta_b}\left(\exists m \in \mathbf{N} : \delta(\theta_a, \theta_b) \notin \mathrm{CS}^+_{\alpha,(m)}\right) \leq \alpha,$$

and analogously for (θ_a, θ_b) with $\delta(\theta_a, \theta_b) < 0$. We have thus arrived at a confidence sequence that works for all δ, positive or negative.

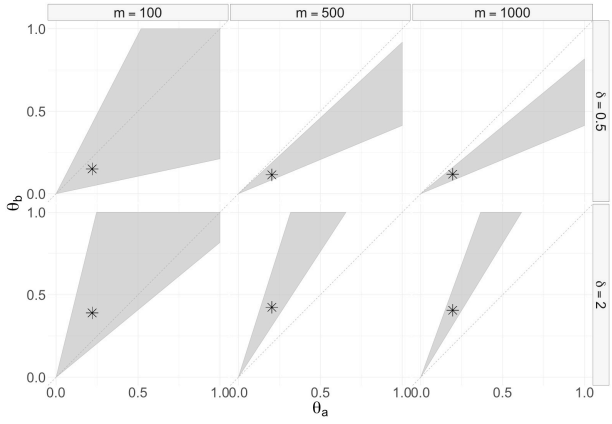
3.3.1 Simulations

In this section some numerical examples of confidence sequences for the two types of effect sizes are given. All simulations were run with code available in our software package [Ly et al., 2022].

Risk difference Risk difference is defined as the difference between success probabilities in the two streams: $\delta = \theta_b - \theta_a$. Figure 3.2 shows running intersections of confidence sequences with δ as the risk difference for simulations for various distributions and stream lengths. These sequences are constructed by testing null hypotheses based on $\vec{\Theta}_0(s, c)$, with $c = 1$ and $s = \delta$. $\mathrm{CI}_{\alpha,(m)}$ for the risk difference on $\vec{\Theta}_0$ is an interval, corresponding to the 'beam' of $(\theta_a, \theta_b) \in [0, 1]^2$ bounded by the lines $\theta_b = \theta_a + \delta_{\mathrm{L}}$ and $\theta_b = \theta_a + \delta_{\mathrm{R}}$ with $\delta_{\mathrm{L}} > \delta_{\mathrm{R}}$ being values such that $S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0(\delta_{\mathrm{L}})]} = S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0(\delta_{\mathrm{R}})]} = 1/\alpha$. In Appendix S3.B we illustrate the



(a) Risk difference



(b) Relative risk

Figure 3.2: Depiction of parameter space with running intersection of confidence sequence for data generated under various effect sizes, at different time points m in a data stream. The asterisks indicate the maximum likelihood estimator at that time point. The significance threshold was set to 0.05. The design was balanced, with data block sizes $n_a = 1$ and $n_b = 1$.

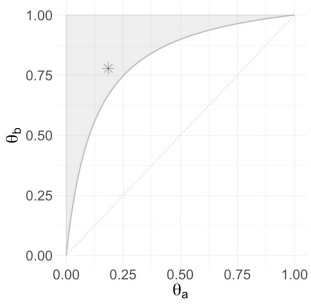
calculations leading to Figure 3.2. Figure S3.1 in the Appendix illustrates that the running intersection indeed improves the confidence sequence, albeit slightly.

Relative risk Relative risk is defined as the ratio between the success probabilities in group b and a: $\delta = \theta_b/\theta_a$. Hence, confidence sequences for this effect size measure can again be constructed using the linear boundary form $\vec{\Theta}_0(s, c)$ again, but now with $s = 0$ and $c = \delta$. Figure 3.2 shows running intersections of confidence sequences with δ as the relative risk.

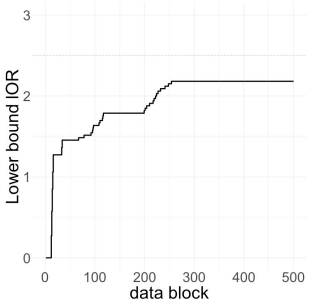
Log odds ratio boundary If the maximum likelihood estimate based on $Y^{(m)}$ lies in the upper left corner as in Figure 3.3(a), the confidence sets $\mathrm{CS}_{(m)}$ we get at time m have a one-sided shape such as the shaded region, or the shaded region in Figure 3.3(c), if the estimate lies in the lower right corner. Again, we can improve these confidence sequences by taking the running intersection; running intersections over time are illustrated in Figures 3.3(b) and 3.3(d).

3.4 Conclusion

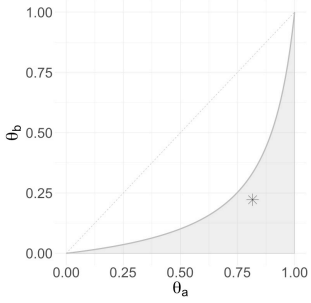
We have shown how e-variables for data streams can be extended to general null hypotheses and non-asymptotic always-valid confidence sequences. We specifically implemented the confidence sequences for the 2×2 contingency tables setting; the resulting confidence sequences are efficiently computed and show quick convergence in simulations. For estimating risk differences or relative risk ratios between proportions in two groups, to our knowledge, such exact confidence sequences did not yet exist. For the log odds ratio we could also have used the sequential probability ratio (SPR) in Wald's SPR test [Wald, 1945] test, which can be re-interpreted as a (product of) e-variables [Grünwald et al., 2022a]. However, the SPR does not satisfy the GRO property making it sub-optimal (see also [Adams, 2020]); moreover, as should be clear from the development, our method for constructing confidence sequences can be implemented for any effect size notion with convex rejection sets $\vec{\Theta}_0(\leq \delta)$ and $\vec{\Theta}_0(\geq \delta)$, not just the log odds ratio. A main goal for future work is to use Theorem 3.2 to provide such sequences for sequential two-sample settings that go beyond the 2×2 table.



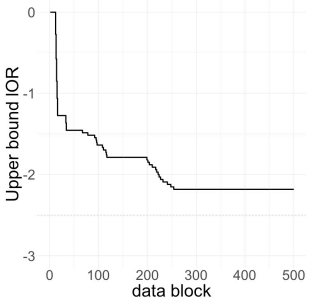
(a) CS^+ at $n = 500$, true lOR 2.5



(b) Running lower bound CS^+, true lOR 2.5



(c) CS^- at $n = 500$, true lOR -2.5



(d) Running upper bound CS^-, true lOR -2.5

Figure 3.3: One-sided confidence sequences for odds ratios. 500 data blocks were generated under P_{θ_a,θ_b} with $\theta_a = 0.2$ and log of the odds ratio (lOR) 2.5 for figures a and b, and $\theta_a = 0.8$ and lOR -2.5 for figures c and d. The asterisks indicate the maximum likelihood estimator at $n = 500$. The significance threshold was set to 0.05. The design was balanced, with data block sizes $n_a = 1$ and $n_b = 1$. Note that CS^- is empty for (a) and (b) and CS^+ for (c) and (d) in these confidence sequences.