A Tutorial on Safe Anytime-Valid Inference: Practical Maximally Flexible Sampling Designs for Experiments Based on e-Values

Alexander Ly[1], Udo Boehm[1,2], Peter Grünwald[1,3], Aaditya Ramdas[4], and and Don van Ravenzwaaij[5]

[1]Centrum Wiskunde & Informatica
[2]Tilburg University
[3]Leiden University
[4]Carnegie Mellon
[5]University of Groningen

Abstract

We demonstrate how e-values simplify both experimental design and the inference process. With e-values researchers can perform anytime-valid tests and construct confidence intervals that maintain type I error control *regardless of the sample size*. This enables real-time monitoring of evidence as data are collected, permitting early termination of experiments without intolerably inflating the risk of false discoveries. Early stopping not only conserves resources, but also mitigates risk for participants in clinical settings. Anytime-valid tests allow for optional continuation, that is, the extension of an experiment, for instance if more funds become available, or even if the evidence looks promising and the funding agency, a reviewer, or an editor urges the experimenter to collect more data. Analogously, a researcher can be assured that a 95% anytime-valid confidence interval will, with at least 95% probability, cover the true effect size regardless of how, or even if, data collection is stopped. We use the free and open-source software package `safestats` implemented in `R` to illustrate the practical benefits of this novel inference framework.

Keywords: adaptive sampling designs; evidence; reproducible science; research waste reduction; sequential analysis

Reproducible science is a demanding endeavour: To minimise reporting, publication, and hindsight biases, researchers must clearly define their theories and hypotheses in a pre-registration document before data acquisition. Such a pre-registration document typically also requires the pre-specification of a (sampling) rule for terminating data collection. This

Correspondence concerning this article should be addressed to Udo Boehm, Tilburg University, Department Methodology and Statistics, PO Box 90153, 5000 LE Tilburg, The Netherlands, Email: U.Bohm@tilburguniversity.edu

serves to prevent the commonly used classical p-value test from becoming invalid, thus, unreliable. Without a predefined plan, repeatedly checking whether $p < \alpha$ can make any working hypothesis appear "significant", even when it truly is not. Simmons, Nelson, and Simonsohn (2011, p. 4) therefore emphasised the need for transparency and adherence to these sampling rules, urging reviewers to enforce compliance. But this solution ultimately relies on eliminating flexibility by confining researchers to a rigid sampling protocol, all in service of an old-fashioned statistical inference tool developed back in the 1930s. In practice, however, even the most carefully planned studies face setbacks such as slow recruitment, budget shortfalls, or unforeseen crises like pandemics. Being left with invalid inferences after managing such challenges can be both discouraging and counterproductive.

Instead of tying researchers to rigid classical tools that demand strict sampling plans, we introduce anytime-valid tests and e-value based confidence intervals. These methods allow researchers to dynamically decide whether to continue or stop an experiment, all without breaking statistical guarantees. Reviewers benefit too, since they no longer need to verify adherence to rigid sampling protocols.

Our primary contribution lies in offering insightful yet *non-binding* recommendations for maximally flexible sample sizes, providing clear guidance on how to allocate resources efficiently when designing and conducting experiments. In a nutshell (all jargon will be explained further below), we provide planned sample sizes n_{plan} at which a desired power at a guessed (or minimally clinically relevant) effect size is guaranteed; non-binding means that both the e-value based tests and confidence intervals remain valid irrespective of whether the planned sample size is actually realised, or whether the guessed effect size is even approximately correct — we may always stop early or sample new data points at will.

These planned sample sizes can easily be obtained with the `safestats` `R` software package as elaborated on in Section 2.3. Before doing so, we first address the challenges of monitoring a $p < \alpha$ test in Section 1.1 and demonstrate how e-values resolve these issues in Section 1.2. More generally:

Organisation of this Paper.

- Section 1 focuses on *type I* errors, specifically addressing scenarios where a treatment effect is truly absent. This section is more detailed, as it establishes the notation and provides the main motivation for e-values.
- Section 2 shifts the focus to *type II* errors, equivalently, power analyses, and investigates the behaviour of e-values under alternatives (when an effect is truly present).
- Section 3 provides *practical recommendations* on the choice of e-value procedures in the t-test setting.
- Section 4 demonstrates how a planned e-value analysis enables reliable conclusions using *approximately 58% and 70% fewer data points* in two real-world examples.
- Before digging into details, the hurried reader may want to consult the concluding Section 5 which summarises *advantages* but also lists potential *disadvantages* of using e-value based inference methods, as well as further important (mostly advantageous) aspects of e-values that we could not cover here in detail.

This introduction to e-values is intended for practitioners who wish to make reliable inferences with minimal statistical hurdles. It focuses on accessibility for an applied audience, leaving rigorous mathematical treatment to specialised works such as those by Grünwald, De Heide, and Koolen (2024), Howard, Ramdas, McAuliffe, and Sekhon (2021), Ramdas, Grünwald, Vovk, and Shafer (2023), and Shafer (2021).

1 An e-value quantifies the evidence against the null as indicated by the data

To make the exposition concrete, we consider the experimental setup where two groups, say, treatment and control, are being compared. As usual, individuals in each group are assumed to have normally distributed responses with unknown population means μ_1, μ_2, respectively, and unknown population standard deviation σ. To argue for the effectiveness of the treatment, we test the null hypothesis that states $\mathcal{H}_0 : \varphi := \mu_1 - \mu_2 = \varphi_0$, where φ_0 is a postulated (known) null value, typically, $\varphi_0 = 0$. Under the null hypothesis, any discrepancy between the postulated null value φ_0 and the observed sample mean difference $\bar{x}_1 - \bar{x}_2$ is attributable solely to random variation.

The conventional approach uses a two-sample t-test conducted at a pre-registered – thus, data-blind – sample size or "time point" n. At n, data collection stops, the test is conducted, and a decision is made to either reject, if $p < \alpha$, or to not reject the null hypothesis, if $p \geq \alpha$. As usual, α represents the tolerable type I error, the probability (i.e. relative frequency) with which a testing procedure is allowed to produce a false null rejection. To ease exposition, we think of time n as the sample size(s) by assuming $n = n_1 = n_2$.[1] Determining n at which data collection stops and the $p < \alpha$ test is conducted is challenging: if the sample size is set too low, we risk losing statistical power, therefore, missing a true treatment effect with too high probability; conversely, setting n too high wastes valuable resources.

To circumvent this problem, we propose dynamically determining the sample sizes by tracking a sequence of e-values, where e stands for the evidence against the null hypothesis indicated by the accumulating data. An e-value of 0 means no evidence, 1 means neutral evidence, and values above 1 indicate evidence against the null. The larger the e-value, the stronger the evidence, which, in principle, can grow without bound to represent irrefutable evidence against the null. As data accumulate, we observe a sequence of e-values written as $e := (e_1, e_2, \ldots, e_n, \ldots, e_{n_{\text{plan}}})$, where n_{plan} is some time horizon (not a fixed sample size), which we arbitrarily set to $n_{\text{plan}} = 200$ for the moment. At each time point n the e-value e_n is a non-negative number forming, for example, the sequence $e = (1, 0.83, \ldots, 27)$. As soon as e_n passes $1/\alpha$, we can reliably stop an experiment in favour of a treatment effect. Importantly, when there truly is no effect, there is at most α (e.g. $\alpha = 0.05$) probability that we will *ever* observe misleading evidence against the null hypothesis larger than $1/\alpha = 20$. This result is known as *Ville's inequality*, which we use to motivate the formal definition of e-values and E-processes below.

[1] The highlighted properties of p and e-values discussed here extend naturally to cases with $n_1 \neq n_2$ at some or each time point, but this would require bookkeeping that muddles the presentation.

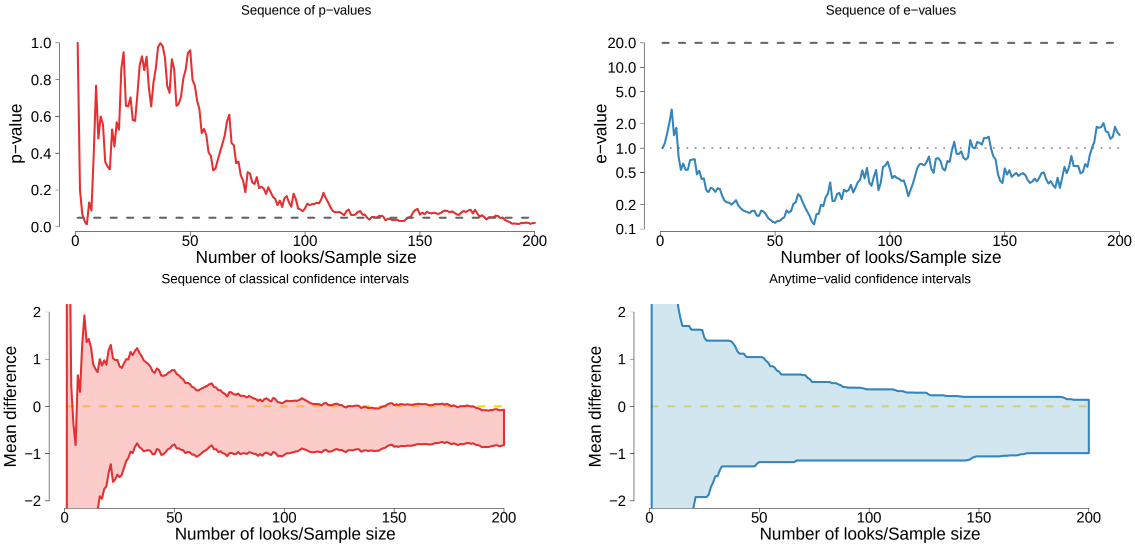


Figure 1

Top left panel: Given a sufficiently large time horizon, here $n_{\text{plan}} = 200$, a p-value will always dip below any α-threshold (here $\alpha = 0.05$), even if the null holds true. Bottom left panel: The classical 95% confidence intervals will not cover the true effect at all moments in time. Top right panel: For the same data that were generated under the null, the sequence of e-values remains under $1/\alpha$. Bottom right panel: The 95 % anytime-valid confidence intervals cover the true effect size (here 0) at all moments in time.

1.1 Continuously monitoring a $p < \alpha$ test guarantees the eventual rejection of the null hypothesis, even when it is true

To appreciate Ville's inequality and to motivate e-values, we contrast it to the repeated $p < \alpha$ testing procedure, which leads to the (well-known) phenomenon of sampling to a foregone conclusion (Kotz, 2006):[2] If the null hypothesis is true, then repeatedly collecting data will inevitably cause any sequence of p-values to fall below a significance threshold (such as $\alpha = 0.05$) resulting in false rejections of the null and misplaced confidence in ineffective treatments.

An instance is shown in the top left panel of Fig. 1 for which we simulated normally distributed data under the null hypothesis of no effect with $\mu_1 = \mu_2 = 4$ and standard deviation $\sigma = 2$. At each time $n = 1, 2, \ldots, 200$ a p-value is computed yielding the sequence $p := (p_1, p_2, \ldots, p_{200})$. The p-value dips below $\alpha = 0.05$ on 31 occasions, namely, at $n = 4$, 5, 127, 128, 131, 135, 144, 179, 185, and between $n = 187$ and $n_{\text{plan}} = 200$. Hence, monitoring the p-value and stopping at the first instance $p < \alpha$ will, for this data set, yield a false null rejection.

The problem is not alleviated if we track the confidence intervals across time instead. The bottom left panel of Fig. 1 shows that we arrive at similar incorrect conclusions about

[2] We hope that the code provides some additional insights to this phenomenon for those who are unfamiliar with it. All `R` code used for the simulations and plots can be found on the OSF repository `https://osf.io/mdbqe/files/osfstorage`.

the true population mean difference, here $\varphi = \mu_1 - \mu_2 = 0$, if we monitor the confidence intervals and stop as soon as zero falls out of the interval. This is due to classical $1 - \alpha$ confidence intervals being the inversions of the $p < \alpha$ tests (e.g. Bickel & Doksum, 2015, Section 4.5, and Appendix A). The red curves in Fig. 2 illustrate the performance of the procedures that monitors the $p < 0.05$ test and the 95% confidence interval under repeated use. The simulation study is based on 5000 data sets generated under the null with $\mu_1 = \mu_2 = 4$ and $\sigma = 2$ as before. Though, we get the same graph if each replicating data set is generated with a different, but common $\mu_1 = \mu_2$ and a different σ. The red curve in the left panel represents the *realised type I error* (also known as the False Positive Rate, henceforth, FPR) as a function of time n.[3] At $n_{\text{plan}} = 200$ 2158 out of 5000 p-values sequences (43.16%) had (at least one) time point with $p < \alpha$. Hence, the situation represented in the left column of Fig. 1 is common. The dramatic increase in FPR of the repeated $p < \alpha$ test is caused by the number of times a researcher checks for a significant p-value, not by the test being performed after each pair of observations. For instance, if the $p < \alpha$ test were conducted after, say, 38, an additional 20, and another 40 more participants, then the FPR would be approximately 5%, 8.64%, and 10.80% as in Fig. 2 at times $n = 1, 2, 3$. By extending the time horizon, here, $n_{\text{plan}} = 200$, indefinitely, we end up at an FPR of 100% resulting in a certain $p < \alpha$ rejection, which makes a treatment appear "significant", despite the null being true. More formally, for any data-generating distribution $\mathbb{P}$ from the null model $\mathcal{M}_0$ consisting of $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$ with $\mu_1 = \mu_2$, we have

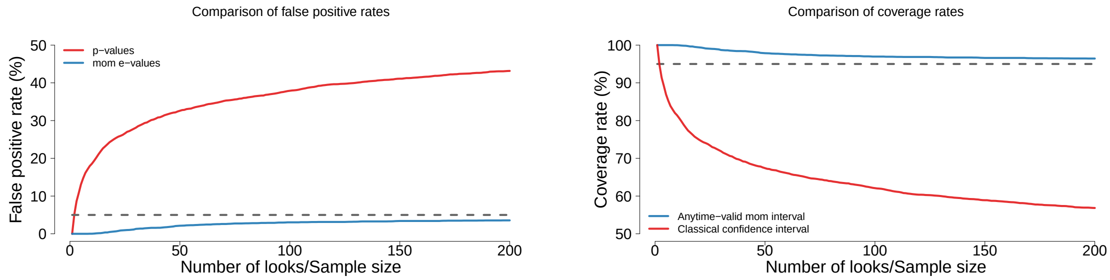


Figure 2

Left panel: The false positive rates of monitoring the $p < \alpha$ *test (red) increase well beyond the tolerable* $\alpha = 0.05$*. On the other hand, the FPR of the* $e \geq 1/\alpha$ *test (blue) remains below the tolerable* α *at all moments in time. Right panel: The coverage rates of the classical confidence interval (red) dips below the nominal 95% level, if it is used sequentially. The coverage rate of the anytime-valid confidence interval remains above the nominal 95% as promised.*

$$\lim_{n_{\text{plan}} \to \infty} \mathbb{P}(\text{There exists a time } n \leq n_{\text{plan}} \text{ such that } p_n < \alpha) = 1. \tag{1}$$

[3] The value of the curve at n stems from tallying the number of p-value sequences that, up to time n, led to a false rejection of the null hypothesis. The number of false positives is then divided by the total number of p-value sequences leading to the relative frequency of interest. Similarly, the red curve in the right panel of Fig. 2 represents the fraction of replicating data sets that included the true mean difference of $\mu_1 - \mu_2 = 0$ at all times up to $n = n_1 = n_2$.

This probability is already larger than α when $n_{\text{plan}} = 2$, which implies that p-values become more unreliable with more data. Eq. (1) occurs for all classical p-value testing scenarios, not just t-tests.

This shows that the event $p < \alpha$ can only be legitimately interpreted as rare – occurring with probability at most α – under the null, if the test is performed once, and only once. Any deviation from the sampling plan undermines the validity of this interpretation, potentially misleading us into erroneous conclusions.[4]

1.2 Stopping times and the definition of E-processes

The desire for a measure of evidence that is resilient to sampling to a foregone conclusion can be seen as the main motivation for e-values. Its definition rests on one key requirement: controlling the risk of false evidence — that is, bounding the probability of observing misleadingly large evidence against a true null hypothesis. Meeting this requirement depends on two technical concepts. The first is *stopping times*, which describe when data collection ends, whether by design or external factors. The second is *Markov's inequality*, a technique that uses the expected value to bound the probability of extreme outcomes. Together, they provide the backbone for the robustness we want, as we explain in more detail below.

In brief, we model a sequence of e-values as an outcome of a random processes $E := (E_1, E_2, \ldots, E_n, \ldots)$. This process is designed so that, at any stopping time, its expected value never exceeds one. We denote by E_n the anticipated evidence at time n, written in uppercase to emphasise its dependence on future data. Once the data at time n are observed, the random variable E_n realises as a fixed value (e.g. $e_n = 7$) and loses its randomness. Hence, a random process E formalises the sequential procedure that converts accumulating data into a running sequence of e-values that evolves over time.

We next elaborate on the practical relevance of stopping times through examples, then demonstrate how knowledge of a random process's expectation yields the desired control.

1.2.1 Stopping times

A stopping time is a, possibly data-dependent, time N at which data collection will be halted, as long as this decision does not depend on future events. Here are two (familiar) examples:

(a) The **deterministic stopping rule** to terminate data collection at a fixed time, say, $N = 200$, after observing $n_1 = n_2 = 200$ samples is a stopping time. Halting is not influenced by any observation, particularly those after $N = 200$.

[4] There are group-sequential and α-spending p-value methods that compensate for FPR inflation with different thresholds α_n. The choice of α_n, and therefore the interpretation of $p_n < \alpha_n$ as rare, depends on: (i) a pre-determined final time point/sample size $n_{\max}$, (ii) the total number of times the test is conducted, (iii) when (relative to $n_{\max}$) the tests are conducted, and (iv) an α-spending function. For these methods type I error control is only guaranteed up to $n_{\max}$ and requires practitioners to adhere to the stopping rule $p_n < \alpha_n$, which is not the case for e-value based methods that are therefore more flexible, see the discussion on stopping times below, and Ter Schure, Perez-Ortiz, Ly, and Grünwald (2024) for a comparison.

(b) Stopping at the **first-passage time** N, the first time when $e_N \geq 1/\alpha$ or $p_N < \alpha$, uses only the data collected up to time N.

Adapting stopping rules (a) and (b) to the available data also leads to valid stopping times:

(c) **Delayed first-passage time**: Assume that the first-passage time N occurs early on. To increase our confidence in the finding, we require the e-value to remain above $1/\alpha = 20$ for, say, an additional 6 time points before actually stopping.

(d) **Forced continuation**: We observed $e = 12$, but a reviewer, who is convinced that the evidence should be larger than 20, demands that we test another 37 observations before concluding the experiment. Halting after the additional 37 observations is a stopping time.

(e) **Optimistic continuation**: A statistician tells us to stop data collection for the null the first instance $e_N \leq 0.21$, and to stop for a treatment effect as soon as $e_N \geq 16$. During data acquisition $e_n \leq 0.21$ is observed, but the optimistic principal investigator insists on continuing sampling until $e_N \geq 16$. The first time N at which $e_N \geq 16$ occurs after first observing $e_n \leq 0.21$ is a stopping time.

Note that the last two stopping times incorporate hopeful intentions. Sometimes we are forced to stop due to external causes, which are also examples of stopping times:

(f) The **resource-exhausting stopping time** is the moment when data collection is halted due to depleted funds, a malfunctioning measurement instrument such as an fMRI scanner, a catastrophic event, or mandatory shutdowns due to a pandemic.

(g) **Frustrated stopping time**: The moment the study ground to a halt because the principal investigator, bested by a stubbornly malevolent printer, resigned in exasperation.

(h) **Convenient stopping time**: The long-awaited moment when a meeting is finally held with the funding agent, governing board, and ethics committee, allowing the data analytic results to be reviewed and a decision to be made to stop all experiments (e.g. Ter Schure et al., 2022).

An illustration of what does not qualify as a stopping time is provided in Section 1.3. The moment of stopping might be based on whichever of the above conditions occurred first, which is also a stopping time. More formally, the minimum between two stopping times N and M is also a stopping time. For instance, N could be the first-passage time $E_N \geq 1/\alpha$, and M some planned sample size n_{plan} derived from a power analysis. By additionally taking the minimum with the point at which resources (e.g. funding, time, or energy) run out, we obtain a model of a stopping time that is well suited to real-world applications.

1.2.2 Bounding the exceedance probability of a non-negative random variable with its expectation

Every process eventually stops and we write N_0 for the (potentially unknown in advance) stopping time at which data collection will be halted. Despite not knowing the stopping time N_0, we aim to base our inference on a random process E that, upon stopping, produces misleadingly large evidence against the null hypothesis with only a small probability. To achieve this, we employ Markov's inequality (see Appendix B for a proof), which bounds the probability of any non-negative random variable X exceeding $c > 0$ by $1/c$ times its expectation. That is,

$$\mathbb{P}(X \geq c) \leq \mathbb{E}_{\mathbb{P}}[X]/c. \tag{2}$$

where $\mathbb{E}_{\mathbb{P}}$ is the expectation under the data-governing distribution $\mathbb{P}$.[5] Since the anticipated evidence E_n at every n is non-negative, it is also non-negative at the stopping time N_0. Applying Eq. (2) with $c = 1/\alpha$, e.g., $c = 20$ for $\alpha = 0.05$, to the random variable E_{N_0} yields

$$\mathbb{P}(E_{N_0} \geq 1/\alpha) \leq \alpha \mathbb{E}_{\mathbb{P}}[E_{N_0}]. \tag{3}$$

For the left-hand side to be bounded by α for *any* $\mathbb{P}$ from the null, and to compensate for not knowing which stopping time N_0 underlies the sampling process, we crudely *require* the process E to have $\mathbb{E}_{\mathbb{P}}[E_N] \leq 1$ for *all* $\mathbb{P}$ from the null and *all* possible stopping times N simultaneously.

1.2.3 The definition of E-processes

We call the random process $E := (E_1, \ldots, E_n, \ldots)$ an *E-process for the null model* $\mathcal{M}_0$ if it satisfies the following:

(i) It has to quantify neutral evidence at the start, that is, $E_1 = 1$.

(ii) At each time n the anticipated e-value E_n may take on values between 0 and ∞ representing absolutely no evidence and irrefutable evidence against the null, respectively.

(iii) But under *any* data-generating distribution $\mathbb{P}$ from the null, and all possible stopping times N, we expect E_N to convey at most neutral evidence:

$$\text{For } \textit{any} \text{ stopping time } N \text{ and } \textit{all } \mathbb{P} \in \mathcal{M}_0 \text{ we require } \mathbb{E}_{\mathbb{P}}[E_N] \leq 1. \tag{4}$$

The random variable E_n, the E-process at time n, is referred to as an *E-variable* if Eq. (4) holds for the fixed time $N = n$.

The null model (the set of data-generating distributions consistent with the null hypothesis) only appears in the definition of an E-process through requirement Eq. (4). This requirement is least restrictive when the null model is simple. That is, when it contains only a single distribution, such as the standard normal distribution $\mathcal{N}(0, 1)$ in the one-sample z-test setting.

An E-process for a simple $\mathcal{M}_0$ is characterised by a likelihood ratio with the denominator corresponding to the null model. More specifically, let $p(x^n)$ denote the density function (e.g. Ly, Marsman, Verhagen, Grasman, & Wagenmakers, 2017) of the simple $\mathcal{M}_0$ at outcomes $x^n := (x_1, \ldots, x_n)$ that are realised from $X^n := (X_1, \ldots, X_n)$, and $q(x^n)$ *any* density function that integrates to one. In Appendix E we verify that the likelihood ratio

[5] Recall that the expectation is a weighted sum/integral of the outcomes x of X, where the weights p are defined by the data-governing distribution $\mathbb{P}$, i.e. $\mathbb{E}_{\mathbb{P}}[X] = \int_0^\infty x p(x) \mathrm{d}x$.

InfoBox 1: The betting interpretation of e-values

We may always interpret E-processes as a bettor's potential wealth in an ongoing multi-round game, in which no money is expected to be gained if the null hypothesis holds true. We explain the precise rules of the game in the simplest setting, where the E-process takes the form of a likelihood ratio process. In this setting the bettor starts with initial capital $e_1 = 1$, say, dollar, and the pay-off depends on $o(x)$, where x is an outcome taking values in a finite set, say, $\{\text{RED}, \text{BLACK}\}$. If the null hypothesis corresponds to $p(\text{RED}) = p(\text{BLACK}) = 1/2$, as for a roulette table without the green 0 outcome, then the game maker (i.e. the casino/nature) can set $o(x) = 1/p(x) = 2$.

At each round n, before seeing the outcome of X_n, the bettor determines a strategy for allocating her current wealth e_{n-1} across all possible outcomes; for example, $q(\text{RED}) = 2/3, q(\text{BLACK}) = 1/3$. She receives e_{n-1} multiplied by (i.e. the likelihood ratio) $l_n(x_n) := q(x_n) \times o(x_n) = q(x_n)/p(x_n)$; in our example, if $x_2 = \text{RED}$ then $e_2 = l_1(x_1) \times e_1 = 4/3$; the money put on BLACK is lost. Thus, at round n, after observing x_n, the bettor's accumulated wealth becomes $e_n = \prod_{i=1}^{n} l_i(x_i) = l_1(x_1) \times l_2(x_2) \times \cdots \times l_n(x_n)$. For example, after observing a sequence $(\text{RED}, \text{RED}, \text{BLACK})$ the bettor above would have accumulated $(4/3)^2(2/3) = 1.19$ dollar.

If (as in the example above, with the null a 'real casino') the $o(x)$ is set such that the bettor is not expected to gain money if the null hypothesis is true, then (and only then) is the potential wealth process $E = (E_1, E_2, \ldots)$ an E-process. It further is a *good* E-process if the bettor is expected to get rich fast when the alternative is true. Betting with strategy $q(\text{RED}) = 2/3$ at each round n makes sense if the bettor thinks that there is a substantially larger probability of observing RED compared to BLACK. If this is truly the case, then the bettor's wealth E_n will grow exponentially fast. The higher this wealth, the more evidence is acquired against the null.

By refining the game, the betting analogy can be extended to composite nulls and continuous outcomes as is the case for the T-statistic: every E-process has a sequential betting interpretation, which can be used to gain intuition. For instance, the fact that E-processes preserve type I error control under optional stopping is the *same* mathematical phenomenon as the fact that there is at most $1/c$ probability of you multiply your initial capital in a real casino by a factor c – no matter what betting strategy you use or what rule is used to stop betting. The fact that a valid stopping time may depend on past data but not on the future corresponds to the truism that your decision to stop betting in a real casino may depend on what you have seen so far but not on what you will see in the future.

Further elaboration on this betting interpretation is unfortunately beyond the scope of this paper, and we refer to Grünwald et al. (2024); Ramdas et al. (2023); Shafer (2021); Waudby-Smith and Ramdas (2024) and at an introductory level to (Ter Schure, 2023, Chapter 1) for further details. ⋄

process $E^q := (E_1^q, E_2^q, \ldots, E_n^q, \ldots)$, where $E_n^q := \frac{q(X^n)}{p(X^n)}$, is indeed an E-process for this simple $\mathcal{M}_0$. The arguments apply to *any* density function q, which makes the collection $\mathcal{E}(\mathcal{M}_0)$ of all E-processes for $\mathcal{M}_0$ vast. Box 1 provides a concrete example of likelihood ratios and their betting interpretation. We elaborate on E-processes for the t-test setting after discussing Ville's inequality.

1.2.4 Safe anytime-valid tests

For any E-process E from the vast collection $\mathcal{E}(\mathcal{M}_0)$, its definition (i.e. Eq. (4)) and the discussion surrounding Eq. (3) imply the following: Under any distribution from the null, there is at most α probability that an E-process produces misleading evidence exceeding $1/\alpha$ upon stopping, no matter which (possibly unknown) stopping time is used to terminate data collection:

$$\text{For all data generating } \mathbb{P} \text{ from } \mathcal{M}_0 \text{ and any stopping time } N : \mathbb{P}(E_N \geq 1/\alpha) \leq \alpha. \tag{5}$$

Eq. (5) forms the basis for safe anytime-valid (henceforth savi) tests and confidence intervals. Since Eq. (5) holds for all stopping times, it does so for the ones mentioned in Section 1.2.1. Applying Eq. (5) to the first time N at which E_n exceeds $1/\alpha$ (see Howard, Ramdas, McAuliffe, and Sekhon (2020), Ruf, Larsson, Koolen, and Ramdas (2023) and Appendix C for details) yields Ville's inequality:

$$\text{For all } \mathbb{P} \text{ from } \mathcal{M}_0 : \mathbb{P}(\text{There exist a time } n \text{ such that } E_n \geq 1/\alpha) \leq \alpha. \tag{6}$$

In contrast to Eq. (1) for p-values, Ville's inequality ensures that additional data do not mislead an E-process into favouring ineffective treatments with probability above α. This robustness holds even if extra data are gathered at the insistence of a hopeful reviewer or investigator, or if data collection continues beyond the moment $e_n \geq 1/\alpha$, since Eq. (5) still holds. For these reasons, we refer to the $e_n \geq 1/\alpha$ test as a savi test of level α.

The right panel of Fig. 1 shows an e-value sequence (blue) that is realised from a so-called `mom` type E-process with tuning parameter set to $g_{\texttt{mom}} = \frac{0.5^2}{2}$ in the t-test setting, which we introduce in the next section. For the same data as in the top left panel, which resulted in $p < \alpha = 0.05$ on 31 occasions, we see that the e-values correctly remained below $1/\alpha = 20$ at all $n = 200$ time points.

1.2.5 E-processes for t-tests

E-processes for composite $\mathcal{M}_0$ (comprising more than one distribution, as in the case of t-tests) are more complicated than when $\mathcal{M}_0$ is simple. The increase in complexity stems from requirement Eq. (4), which is more restrictive for a larger class of null-consistent probability distributions. There are, fortunately, various general approaches to construct E-processes for composite null models: (a) the universal inference construction that works for any $\mathcal{M}_0$ that admits a (type of) maximum likelihood estimator (Wasserman, Ramdas, & Balakrishnan, 2020), (b) the reverse information projection approach (e.g. Grünwald et al., 2024; Larsson, Ramdas, & Ruf, 2025) that is guaranteed to lead to powerful E-variables by optimisation, (c) the p-to-e-calibrator approach that uses specific functions to convert any p-value into an e-value (e.g. Shafer, Shen, Vereshchagin, & Vovk, 2011, Vovk & Wang, 2021), as well as constructions based on, for instance, the betting interpretation (Waudby-Smith & Ramdas, 2024), and the Doob martingale (Koning & van Meer, 2025). The first two constructions lead to expressions that can be immediately recognised as variations of likelihood ratios.[6]

This is also the case for the problem at hand, as was shown by Pérez-Ortiz, Lardy, de Heide, and Grünwald (2024) and Hendriksen, de Heide, and Grünwald (2021), in a generalised setting. More precisely, in the t-test setting, any savi test defining tuning parameter δ_s leading to the process $E^{\delta_s} = (E_1^{\delta_s}, E_2^{\delta_s}, \ldots, E_n^{\delta_s}, \ldots)$ is an E-process comprised

[6] These variations are not the same as the classical concept of generalised likelihood ratios, which do not form E-processes. Furthermore, they do not fully align with the notion of Bayes factors, as not all Bayes factors are E-variables, see Appendix H for a (counter)example, and not all E-variables are Bayes factors.

of E-variables $E_n^{\delta_s}$ given by[7]

$$E_n^{\delta_s} := \frac{f_\nu(T \,|\, \sqrt{n_{\text{eff}}}\delta_s)}{f_\nu(T)}, \tag{7}$$

where in the denominator $f_\nu(t)$ is the likelihood of the (standard) T-distribution with ν degrees of freedom, and where in the numerator $f_\nu(t \,|\, \sqrt{n_{\text{eff}}}\delta_s)$ is the T-likelihood with non-centrality parameter $\sqrt{n_{\text{eff}}}\delta_s$ for outcome t. The outcome t refers to a realisation of the usual T-statistic, where

$$T := \sqrt{n_{\text{eff}}}\frac{\bar{X}_1 - \bar{X}_2 - \varphi_0}{S_p}, \text{ where } S_p := \sqrt{\frac{1}{\nu}\Big(\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 + \sum_{j=1}^{n_2}(X_{2j} - \bar{X}_2)^2\Big)}, \tag{8}$$

is the pooled sample deviation based on $\nu = n_1 + n_2 - 2$ degrees of freedom, and where $\bar{X}_k := \frac{1}{n_k}\sum_{i=1}^{n_k} X_{ki}$ is the sample mean of population $k = 1, 2$, $n_{\text{eff}} = \frac{n_1 n_2}{n_1 + n_2}$ the effective sample size, and n_k the sample size of group $k = 1, 2$.

In sum, every choice of δ_s in Eq. (7) leads to an E-process in the t-test setting. We are, however, not limited to fully committing to a single value, such as $\delta_s = 0.5$. Instead, we can spread our bet by taking a weighted average over the tuning parameter δ_s and construct what is known as a mixture E-process with respect to a prior $\pi(\delta_s)$. For instance, using the point prior that puts all mass at $\delta_s = 0.5$ recovers the aforementioned choice for a single value. The prior is selected before data observation as in Bayesian statistics. However, we do not interpret π as prior belief about δ_s. Instead, we treat it as a a tuning distribution, which we optimise over (as elaborated on in Section 2 below). The distinction between this approach and a Bayesian prior belief is further discussed in the concluding Section 5.

Regardless of the chosen prior π, as long as the non-negative weights integrate or sum to one, that is, $\int \pi(\delta_s)\mathrm{d}\delta_s = 1$, we get a mixture E-process comprised of E_n^π at time point n defined as

$$E_n^\pi := \int E_n^{\delta_s}\pi(\delta_s)\mathrm{d}\delta_s. \tag{9}$$

Since Eq. (4) holds for each $E_n^{\delta_s}$, it also holds for the mixture. To see this, let N_0 be an arbitrary stopping time, and $\mathbb{P}_0$ any distribution from the null. Swapping the order of taking the expectation with respect to T, and the mixture/integral with respect to δ_s yields

$$\mathbb{E}_{\mathbb{P}_0}[E_{N_0}^\pi] = \int \mathbb{E}_{\mathbb{P}_0}[E_{N_0}^{\delta_s}]\pi(\delta_s)\mathrm{d}\delta_s \overset{Eq.\ (4)}{\leq} \int 1\pi(\delta_s)\mathrm{d}\delta_s = 1. \tag{10}$$

As this holds for any N_0 and $\mathbb{P}_0$ from $\mathcal{M}_0$, we conclude that the mixture is indeed an E-process. Note that this argument does not depend on $E_N^{\delta_s}$ being a likelihood ratio. The general statement is that weighted averages of E-processes remain E-processes. This further hints at the candidate set of E-processes for $\mathcal{M}_0$ being vast.

[7] Eq. (7) makes intuitive sense in light of the result for simple $\mathcal{M}_0$. The major difference, however, is that the likelihood ratio for simple $\mathcal{M}_0$ are likelihood functions at the raw data x^n. For the t-test $\mathcal{M}_0$ is composite, and the likelihood functions are at the reduced (and only marginally sufficient) T-statistic in both the numerator and denominator. This reduction leads to some challenges that makes the rigorous proof of this result much more involved, see Pérez-Ortiz et al. (2024) and Lindon, Ham, Tingley, and Bojinov (2025).

All illustrations and practical applications of anytime-valid inference discussed here are based on the `safestats` `R` package (Ly et al., 2024). This package includes the following four types of mixture E-processes for the t-test setting when comparing $\mathcal{H}_0 : \mu_1 - \mu_2 = \varphi_0$ against $\mathcal{H}_1 : \mu_1 - \mu_2 \neq \varphi_0$. For the sake of completeness (only), we also provide the formulas:

- `grow`: This mixture uses two point priors for the tuning parameter in Eq. (7), assigning equal weight (50%) to each of the values, say, $\delta_s = -0.5$ and $\delta_s = 0.5$. This prior is visualised as the purple arrows in Fig. 3 and yields the mixture:

$$E_n^{\texttt{grow}} = \frac{1}{2}\frac{f_\nu(T \mid -\sqrt{n_{\text{eff}}}\delta_s)}{f_\nu(T)} + \frac{1}{2}\frac{f_\nu(T \mid \sqrt{n_{\text{eff}}}\delta_s)}{f_\nu(T)}. \tag{11}$$

- `eGauss`: This set-up uses a zero-centred Gaussian prior $\mathcal{N}(0, g)$ on δ_s (e.g. Gönen, Johnson, Lu, & Westfall, 2005) resulting in

$$E_n^{\texttt{eGauss}} := \frac{1}{\sqrt{1+n_{\text{eff}}g}}\left(\frac{(1+n_{\text{eff}}g)(\nu+T^2)}{\nu(1+n_{\text{eff}}g)+T^2}\right)^{\frac{\nu+1}{2}}. \tag{12}$$

  The dashed yellow curve in Fig. 3 illustrates the prior with $g = 0.5^2$.

- `mom`: This mixture is based on a so-called non-local moment prior on δ_s (e.g. Johnson & Rossell, 2010; Pramanik & Johnson, 2022). The corresponding two-bump/camel prior is plotted as the blue curve in Fig. 3, and the positions of the bumps take on the role of the tuning parameter controlled by $g_{\texttt{mom}}$. Taking $g_{\texttt{mom}} = \frac{0.5^2}{2}$ mimics the `grow` choice at ±0.5. Appendix F shows that the `mom` type E-variable at n equals:

$$E_n^{\texttt{mom}} := \left(\sqrt{1+n_{\text{eff}}g_{\texttt{mom}}}\right)^{-\frac{3}{2}}\left(\frac{(1+n_{\text{eff}}g_{\texttt{mom}})(\nu+T^2)}{\nu(1+n_{\text{eff}}g_{\texttt{mom}})+T^2}\right)^{1+\frac{\nu+1}{2}}\left(\frac{1+\nu(1+n_{\text{eff}}g_{\texttt{mom}})+n_{\text{eff}}g_{\texttt{mom}}(\nu+1)T^2}{(1+n_{\text{eff}}g_{\texttt{mom}})(\nu+T^2)}\right). \tag{13}$$

- `eCauchy`: This mixes δ_s with respect to a Cauchy prior $\delta_s \sim \text{Cauchy}(0, \kappa^2)$ (e.g. Jeffreys, 1961; Rouder, Speckman, Sun, Morey, & Iverson, 2009) with scale parameter $\kappa > 0$ represented by the dashed brown curve in Fig. 3, where $\kappa = 0.5$. The corresponding mixture E-variable can be accurately computed numerically as:

$$E_n^{\texttt{eCauchy}} := \frac{\kappa}{\sqrt{2\pi}}\int_0^\infty \frac{1}{\sqrt{1+n_{\text{eff}}g}}\left(\frac{(1+n_{\text{eff}}g)(\nu+T^2)}{\nu(1+n_{\text{eff}}g)+T^2}\right)^{\frac{\nu+1}{2}} g^{-\frac{3}{2}} e^{-\frac{\kappa^2}{2g}}\,\mathrm{d}g \tag{14}$$

All four types of E-processes have in some form been introduced in the literature as Bayes factors (Jeffreys, 1961, Ly, Verhagen, & Wagenmakers, 2016a, 2016b) with specifically chosen priors on the nuisance parameters, i.e. the global mean $\mu_g = \mu_1 = \mu_2$ and σ (e.g. Gronau, Ly, & Wagenmakers, 2020). It is not well known that these Bayes factors are also E-variables for which Ville's inequality holds. This latter property suggests a frequentist-based application of these Bayes factors, which departs from their standard use; see Appendix I for a more detailed discussion.[8] The left panel of Fig. 2 shows in blue the (frequentist) guarantee of the procedure that tracks the savi `mom` type t-test which rejects the null as soon as

[8] The e-value based testing procedure also differs from the approach based on Wald's sequential probability ratio test, see for instance Schnuerch, Heck, and Erdfelder (2022), Steinhilber, Schnuerch, and Schubert (2024), and also Fischer and Ramdas (2024) for an improvement of Wald's sequential test.

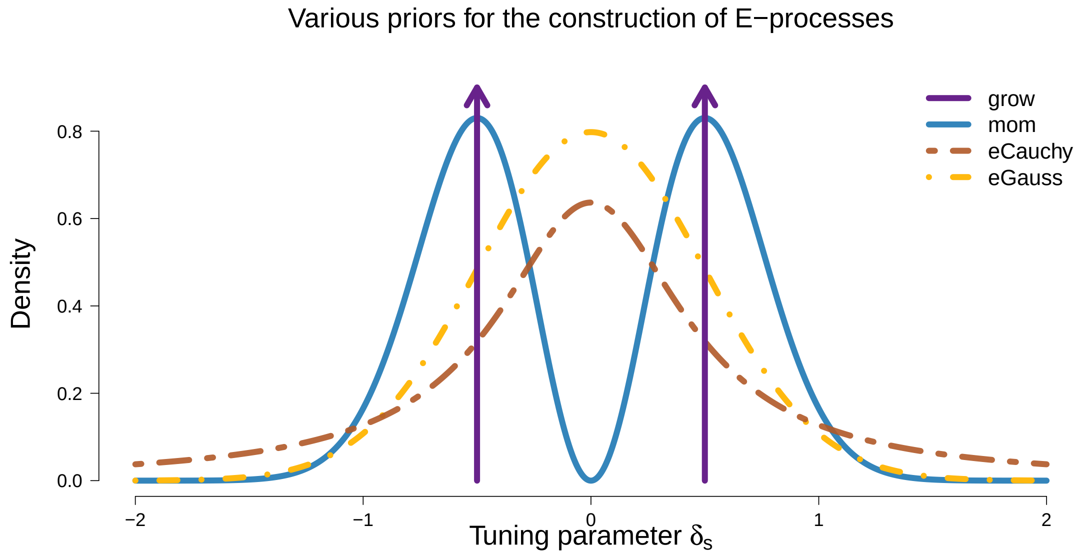


Figure 3

The various types of E-processes for the two-sided anytime-valid t-tests can be identified by the prior used to mix the T-likelihood ratio: `mom` *is based on the non-local moment prior (blue),* `grow` *is based on the two point priors (purple arrows),* `eGauss` *is based on the Gaussian prior (dashed yellow), and* `eCauchy` *is based on the Cauchy prior (dashed brown) on δ_s. All priors are optimised to a minimal clinically relevant effect size $\delta_{\text{min}} = 0.5$.*

$e_n \geq 1/\alpha = 20$. Only 182 out of the 5000 e-value sequences (3.64%) contained a time point at which the e-value exceeded $1/\alpha = 20$. As the number of checks $e_n \geq 1/\alpha$ increases, so will the FPR, but only slightly. Ville's inequality guarantees that the FPR of the savi test will not exceed the tolerable $\alpha = 5\%$, even if the time horizon is increased indefinitely.

1.2.6 Anytime-valid confidence intervals

In other words, at least $1 - \alpha$ proportion, e.g. 95%, of the e-value sequences will *forever* remain below $1/\alpha = 20$. This restatement of Ville's inequality can be derived by inverting Eq. (6), see Appendix D. This requires negating a "there exists" clause within the probability statement, which converts it into a "for all" clause, and vice versa, yielding:[9]

$$\text{For any possible data-generating } \mathbb{P} \in \mathcal{M}_0 : \mathbb{P}(\text{For all times } n : E_n < 1/\alpha) \geq 1 - \alpha. \qquad (15)$$

Eq. (15) is also used to construct anytime-valid confidence intervals, that is, a *sequence of confidence intervals* that, with probability of a least $1 - \alpha$ will forever cover, say, the mean difference parameter $\varphi = \mu_1 - \mu_2$. Prior to experimentation any value φ_0 on the real line is a potential candidate for the data-governing effect size. Each φ_0 can serve as a null value in the T-statistic Eq. (8), which in turn affects the e-value at time n. By inverting $E_n < 1/\alpha$, we collect all those effect sizes φ_0 for which the associated e-values up to n have not (yet)

[9] For example, the existence of just one black swan is enough to disprove the statement "all swans are white".

exceeded $1/\alpha$. The resulting "running intersection" (Howard et al., 2021) $1-\alpha$ confidence sequence, thus, sequentially removes unlikely values φ_0 for φ that the data provide strong evidence against. Since the inversion of $E_n < \alpha$ is carried out inside the probability statement of Eq. (15), the resulting intervals will forever cover the true parameter φ with probability at least $1-\alpha$. These anytime-valid confidence intervals are therefore also attractive to those who suggest to eliminate the concept of statistical significance in favour of interval inference (Amrhein, Greenland, & McShane, 2019).

The bottom right panel of Fig. 1 shows a 95% confidence sequence. The horizontal yellow line depicts the true data generating $\varphi = 0$, which is within the intervals at all moments in time. While the sample mean difference estimator (centre of the intervals) may exhibit bias in the sequential setting, the intervals reliably include the true φ such as $\varphi = 0$, or $\varphi = 1$ in Fig. 9.

The blue curve in the right-panel Fig. 2 shows that 4818 out of 5000 (96.36%) sequences of confidence intervals encapsulated the true data generating φ at all 200 looks. Ville's inequality Eq. (15) guarantees that the coverage probability remains above 95%, even if we extend the time horizon n_{plan} indefinitely.

1.3 Cheating is still not allowed with *e*-values

Ville's and Markov's inequality, Eq. (5), do not protect against data dredging practices, such as selectively reporting results from only $n = 40$ data points, while omitting 50 "outliers", because they led to a low e-value. This deceptive reporting strategy is equivalent to a retroactive stop at $n = 40$ that depends on the full data set of size $n = 90$. As this is not a valid stopping time, we cannot guarantee that data dredging yields e-values larger than $1/\alpha$ with a probability of at most α.

With E-processes we are allowed to choose which group to sample from, before seeing the outcome at time n, based on data up to time $n-1$. However, it is not allowed to fraudulently change the group label after acquiring the observation at time n. Cheating is still prohibited within this framework of inference, and it is not anticipated to be permitted within any reasonable framework of inference (Steinhilber, Schnuerch, & Schubert, 2025).

Cherry-picking an E-process from the vast collection $\mathcal{E}(\mathcal{M}_0)$ after data observation can also invalidate the interpretation of $e_n \geq 1/\alpha$ as rare under the null. To avoid this trap, we should select (and pre-register) an E-process prior to data collection. Below, we demonstrate how the `safestats` software package supports good inferential practices based on e-values.

1.4 Using Ville's inequality to dynamically determine the sample sizes under the null

Once an E-process is selected, Ville's and Markov's inequality Eq. (5) ensure that inference remains safe and anytime-valid. This is achieved by bounding the probability of an E-process producing misleading evidence, even under the most aggressive first-passage time rule, which rejects the null as soon as $e_n \geq 1/\alpha$, while continually extending the time horizon to maximise the potential for misdirection.

The first occurrence of $e_n \geq 1/\alpha$ does not mandate stopping or the rejection of the null hypothesis. Instead, it serves as the earliest signal to consider halting data collection.

```
# Pseudo code: This code does NOT run
n ← 1
eValueAtTime[1] ← currentEValue ← 1

while (currentEValue < 1/alpha && n <= nPlan) {
  currentEValue ← saviTTest(x[1:n], y[1:n], designObj=designObj)
  eValueAtTime[n] ← currentEValue

  if (currentEValue >= 1/alpha) {
    stop() #if we want to
  } else {
    "Increase sample size and test again
          at the start of the while loop"
    n ← n + 1
  }
}
```

Protocol 1.1: The e-value testing procedure in pseudo `R` code with time horizon n_{plan}.

Sampling can safely continue, allowing further data to potentially lower the e-value and revisit any preliminary hard conclusions regarding the null hypothesis.

Thus, the condition $e_n \geq 1/\alpha$ is best interpreted as an (initial) indicator to stop collecting data, rather than a strict rejection criterion. Additional data always enhance inference, for example by tightening the `mom` type confidence intervals. In the end, the realised e-value is most naturally understood as a continuous measure of evidence against the null (e.g. Koning, 2024). In sum, instead of fixing the sample size in advance, we propose to use the data-driven procedure as described in Protocol 1.1 in pseudo `R` code.

2 Power and the fastest evidence accumulating E-process under alternatives

For an E-process to quantify evidence effectively, it should not only generate e-values that remain forever small when the null holds true, but also produce sequences that grow large when the null is false. As more data are collected under the alternative, the evidence against the null should become increasingly pronounced, allowing the procedure to potentially achieve 100% statistical power.

Fig. 4 illustrates this contrasting behaviour by applying Protocol 1.1 with the `mom` type E-process, where $g_{\texttt{mom}} = \frac{0.5^2}{2}$ as introduced earlier, and the arbitrarily set time horizon $n_{\text{plan}} = 200$. The left panel depicts sequences of e-values under the null, whereas the right panel shows sequences under the alternative with standardised treatment effect $\delta = \frac{\mu_1 - \mu_2 - \varphi_0}{\sigma}$ set to $\delta = 0.5$. Under the alternative, most of the e-value sequences (yellow curves) increase rapidly. By quickly accumulating evidence against the null, we gain persuasive support (i.e. $e_n \geq 1/\alpha = 20$) for a treatment effect sooner. In the full simulation 4921 out of 5000 e-value sequences pass $1/\alpha = 20$ by $n_{\text{plan}} = 200$. Stopping as soon as $e_n \geq 1/\alpha = 20$ leads to an average stopping time of $n_{\text{mean}} = 67.19$. Thus, by time $n_{\text{plan}} = 200$, this method correctly indicated persuasive evidence in favour of an effect in 98.42% of cases, demonstrating its high statistical power. An E-process with 100% power implies that the remaining e-value

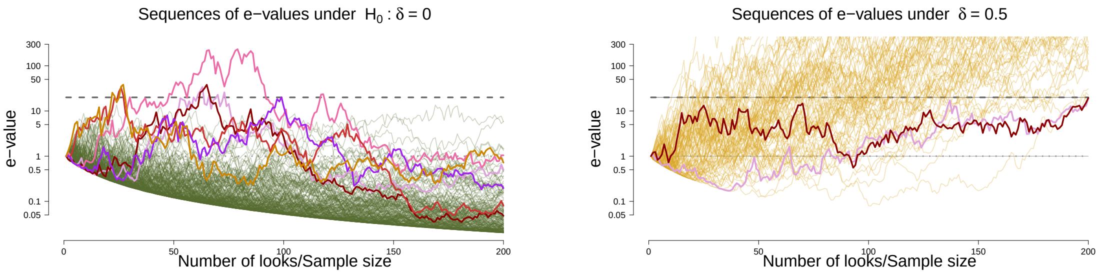


Figure 4

Left panel: The `mom` *type E-process starts at one and under the null yields sequences that slowly drift towards zero (green). Out of the 200 runs shown, six e-value sequences (red/pink/orange/brown/purple/lilac) produced misleading evidence. Right panel: Under the alternative* $\delta = 0.5$*, the e-value sequences increase rapidly (golden yellow). Out of the 200 runs shown, two e-value sequences (red/pink) remained below* $1/\alpha = 20$*.*

sequences will also cross $1/\alpha$, if we continue sampling. The contrasting behaviour depicted by the two panels shows that a sequentially observed $e_n \geq 1/\alpha$ can genuinely be viewed as more likely under the alternative than under the null.

Ideally, we use the fastest evidence accumulating E-process, as it will provide us with the earliest option to stop data collection and potentially maximise resource conservation. We use this optimisation principle to select E-processes in Section 2.1 below. Section 2.3 focuses on how n_{plan} in Protocol 1.1 can be chosen based on a power analysis. As with a classical power analysis, we assume to be given a desired power, e.g. $1 - \beta = 80\%$ therefore a tolerable type II error $\beta = 0.2$, and an expected, or minimum clinically relevant, effect size δ_{min}. Such a δ_{min} is useful when we are not interested in tiny effects of (relatively) costly new treatments. In two-sided settings we refer to data-governing effect sizes δ with $|\delta| \geq \delta_{\text{min}} > 0$ as clinically relevant.

2.1 Determining the fastest evidence accumulating E-variable under $\mathbb{P}_{\delta_{\text{min}}}$

For efficient inference, we favour the procedure that produces larger e-values on average, when an effect δ exists. Suppose that $E_n^{s_1}$ and $E_n^{s_2}$ are two E-variables with savi test defining tuning parameters s_1, s_2, respectively. Under an alternative $\mathbb{P}_\delta$, we prefer $E_n^{s_1}$ over $E_n^{s_2}$, if we expect $E_n^{s_1}$ to, say, triple, whereas $E_n^{s_2}$ to only double the evidence against the null at n. In general, we favour inference based on $E_n^{s_1}$ over $E_n^{s_2}$, if the rate of growth of $E_n^{s_1}$ exceeds that of $E_n^{s_2}$, which we capture using the logarithmic function. That is, if $\mathbb{E}_{\mathbb{P}_\delta}[\log(E_n^{s_1})] \geq \mathbb{E}_{\mathbb{P}_\delta}[\log(E_n^{s_2})]$. If the data are indeed generated under $\mathbb{P}_\delta$, then the most preferable E_n from $\mathcal{E}_n(\mathcal{M}_0)$, the collection of all E-variables for $\mathcal{M}_0$ at time n, is the one that achieves the following maximum (Grünwald et al., 2024; Koolen & Grünwald, 2022):[10]

$$\max_{E_n \in \mathcal{E}_n(\mathcal{M}_0)} \mathbb{E}_{\mathbb{P}_\delta}[\log(E_n)]. \tag{16}$$

[10] To ease exposition, we write maximum, but it should in fact be a supremum (the smallest upper bound). Similarly, below we write a minimum, which should actually be an infimum (the largest lower bound). The differences between maximum/minimum and supremum/infimum matters substantially on a technical level, see Grünwald et al. (2024) and Larsson et al. (2025) for further details.

The E-variable that attains this maximum is referred to as the growth-rate optimal E-variable under $\mathbb{P}_\delta$. Intuitively, an optimal choice maximises the rate at which the E-variable accumulates evidence, thereby *minimising the average sample size required to achieve persuasive support* under an effect of size δ (Ter Schure et al., 2024).

The optimal E-variable that solves Eq. (16), however, typically depends on the data-governing δ. Due to not knowing $\mathbb{P}_\delta$ in practice, we cannot specify the growth-rate optimal E-variable. One workaround involves a minimal clinically relevant effect size $\delta_{\min}$ and the adoption of a conservative approach. The so-called `grow` (the growth-rate optimal in worst case) E-variable then solves the following maximin problem

$$\max_{E_n \in \mathcal{E}_n(\mathcal{M}_0)} \min_{|\delta_{\min}| \geq \delta} \mathbb{E}_{\mathbb{P}_\delta}[\log(E_n)]. \tag{17}$$

As before, the maximum is over the vast collection $\mathcal{E}_n(\mathcal{M}_0)$, and the minimum (the worst-case part) is over all (alternative) distributions $\mathbb{P}_\delta$ that are clinically relevant. For models that have the so-called monotone likelihood property, the problem gets easier as the data-governing δ is further away from the null. For these models the inner minimum of Eq. (17) is attained by $\delta = \delta_{\min}$ and simplifies to the problem Eq. (16). This is the case for t-tests.

2.2 Fastest evidence accumulating E-processes in the t-test setting

Pérez-Ortiz et al. (2024) showed that for t-tests, the *one-sided* `grow` E-variable (alternative $\delta \geq \delta_{\min} > 0$) is given by δ_s in Eq. (7) set equal to $\delta_{\min}$. It follows that the *two-sided* `grow` E-variable is the t-likelihood ratio Eq. (11) based on the two-point mixture that puts half its mass at $\delta_s = -\delta_{\min}$ and at $\delta_s = \delta_{\min}$, see the purple arrows in Fig. 3 in case $\delta_{\min} = 0.5$. Importantly, this `grow` result is not a consequence of modelling subjective belief, but emerges from the desire to use the fastest evidence accumulating procedure.

Instead of optimising over the full collection $\mathcal{E}_n(\mathcal{M}_0)$, we can restrict the candidate set of E-variables to those belonging to the `mom`, `eGauss`, or `eCauchy` types E-variables in Eq. (17) resulting in optimal parameter values of $g_{\texttt{mom}} = \frac{\delta_{\min}^2}{2}$, $g = \delta_{\min}^2$, and $\kappa = |\delta_{\min}|$ respectively.[11] The right panel of Fig. 5 depicts in yellow the sampling distribution of the `mom` E-variables with $g = \frac{\delta_{\min}^2}{2}$ as a function of n for data generated under $\delta = \delta_{\min} = 0.5$. The 95% and 5% quantiles are depicted as golden yellow curves, and the mean of the logarithm of the E-variables as the black curve, which all continue to increase as n grows. For completeness, the left panel shows the sampling distribution under the null. This stark contrast in the behaviour of the E-variable under the alternative and null suggests that e-value sequences have the potential to remain self-consistent and non-contradictory over time under both hypothesis.

The selected `mom` E-process being optimal for the given $\delta_{\min}$ implies that the slope of the black curve in the right panel of Fig. 5 is the steepest achievable amongst all `mom` E-variables under $\mathbb{P}_{\delta_{\min}}$ with $\delta_{\min} = 0.5$. As such, it is guaranteed to cross any evidence threshold $1/\alpha$ (e.g. dashed horizontal grey line) under $\mathbb{P}_{\delta_{\min}}$ the soonest. Provided with β we can also determine when this crossing occurs.

[11] These optimal solutions are derived by Taylor expanding the logarithm of the E-variables for ν large, integrating with respect to t, followed by differentiating with respect to the tuning parameter, i.e. $g_{\texttt{mom}}, g, \kappa$.

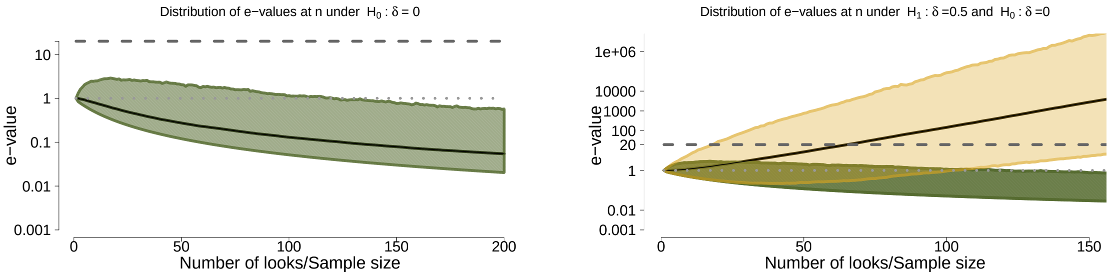


Figure 5

Left panel: Under the null, the sampling distribution of E-variables does not increase and typically drift towards zero. Right panel: Under the alternative, here with $\delta = \delta_{\text{min}} = 0.5$, the distribution of `mom` *e-values increases rapidly. Since* `mom` *is tuned to δ_{min} and thus accumulates the evidence fastest, we know that the depicted increase is the steepest amongst the* `mom` *E-variables.*

2.3 Determining n_{plan} for a desired power $1 - \beta$, δ_{min} and tolerable α

The intuition behind the derivation of n_{plan} in a power analysis is to horizontally cut the yellow sampling distribution depicted in the right panel of Fig. 5 at $1/\alpha$. This is equivalent to studying the distribution of the first times at which the E-process passes $1/\alpha$, which we access by sampling.

For this reason we developed routines in the `safestats` package that efficiently sample the first-passage time distribution. Specifically, using (essentially) just Line 3 of `R` Code 2.1, we simulate (by default) `nSim` = 1000 e-value sequences under $\mathbb{P}_{\delta_{\text{min}}}$, all in under 2.4 seconds on a 2021 iMac M1. [12] Line 4 plots the first-passage time distribution as in Fig. 6. The first 100 e-value sequences until they hit $1/\alpha$ are shown in yellow. The histogram of the `nSim` = 1000 first-passage times N at which $e_N \geq 1/\alpha$ occurred is depicted

```
1 library(safestats)
2
3 designObj ← designSaviT(deltaMin=0.5, beta=0.2, testType="twoSample")
4 plot(designObj)
5
6 result ← saviTTest(x, y, designObj=designObj)
7 plot(result)
8 plot(result, wantConfSeqPlot=TRUE)
```

`R` **Code** 2.1: All the code needed to design, perform and visualise a savi test. Inference based on data vectors x and y occurs on Line 6. Line 7 plots the e-value sequence, and Line 8 illustrates the anytime-valid confidence interval as a function of the sample sizes.

[12] For those interested to code along, we recommend installing version 0.8.8 or higher of the `safestats` package, which, if not available on the Comprehensive R Archive Network (CRAN), can be installed by running the command `remotes::install_github("AlexanderLyNL/safestats", ref="088")` in `R`.

in blue.[13] The top panel of Fig. 6 illustrates the distinct roles of $\alpha, \delta_{\min}, \beta$ for this power analysis. Firstly, α defines the minimum level of evidence for persuasive support, e.g. $1/\alpha = 20$, which is depicted as the horizontal black line. Secondly, $\delta_{\min}$ defines the optimal `mom` E-process which is expressed by the steepness of the average upward drift of the e-value sequences. Lastly, the blue histogram shows that after $n_1 = n_2 = 100$ observations, a wee few more than 800 out of the `nSim` = 1000 e-value sequences correctly resulted in a first option to stop in favour of the presence of a treatment effect. Hence, by monitoring $e_n \geq 1/\alpha$ up to $n_{\text{plan}} = 100$ enables a true treatment effect size of $\delta_{\min} = 0.5$ to be detected with at least $1 - \beta = 80\%$ power. To acknowledge that n_{plan} is derived from simulations, the design object (Fig. 7) also reports twice the bootstrap standard error of 5.69. It also shows an average sample size of $n_{\text{mean}} = 59$ for both groups, which is the average between the first times $N < n_{\text{plan}}$ at which $e_N \geq 1/\alpha$ occurs, and $N = n_{\text{plan}}$ for those e-value sequences that continued until n_{plan}. In comparison, a classical $p < 0.05$ test with the same $\alpha, \beta, \delta_{\min}$ should be performed at $n_1 = n_2 = 64$. Hence, under $\mathbb{P}_{\delta_{\min}}$, the `mom` E-process will require on average 5 fewer (but in the worst-case, 36 more) participants in both groups compared to the classical $p < 0.05$ test. The flexibility of e-value based tests comes at the price of a larger sample size to *plan for*, but Fig. 6 shows that in return there is about 59% probability to realise a stopped experiment before $n = 64$, if $\delta = \delta_{\min}$.

2.4 The behaviour of n_{plan} under other alternatives

Monitoring the test is even more beneficial when the data-governing effect size is larger than $\delta_{\min}$. The middle panel of Fig. 6 shows that, under $\mathbb{P}_\delta$ with $\delta = 0.6$, the evidence against the null accumulates even faster, and the whole first-passage time distribution is shifted to the left. The average sample size at which persuasive support is gathered is then at $n_{\text{mean}} = 46$. This gain in efficiency goes unnoticed for the classical $p < 0.05$ test that for type I error control should only be performed at $n = 64$.

On the other hand, if the data-governing effect size is smaller than expected, then more samples are needed to observe $e_n \geq 1/\alpha$ with 80% power. The bottom panel of Fig. 6 shows that under $\mathbb{P}_\delta$ with $\delta = 0.45$ about 69.5% of the e-value sequences led to $e_n \geq 1/\alpha$ at $n = 100$.

Hence, although the n_{plan} was tuned to the case $\delta = \delta_{\min}$, it also suffices for all data-governing δ larger than $\delta_{\min}$, as one would intuit for a minimal clinically relevant effect size. Upon reaching n_{plan} without the e-value sequence crossing $1/\alpha$, we can halt data collection, maintain the null hypothesis as the status quo, and assert that the effect is not clinically relevant, with a permissible type II error rate of at most β. Alternatively, we can also continue sampling if the e-value looks promising and we are keen to investigate a smaller effect. Similar conclusions can be reached by exploring the anytime-valid confidence intervals. The role of the planned sample size is to guarantee at least $1 - \beta$ probability to gain persuasive support for a treatment effect under continuous monitoring.

[13] The first-passage time of an e-value sequence that did not yet pass $1/\alpha$ is set to ∞, as is customary.

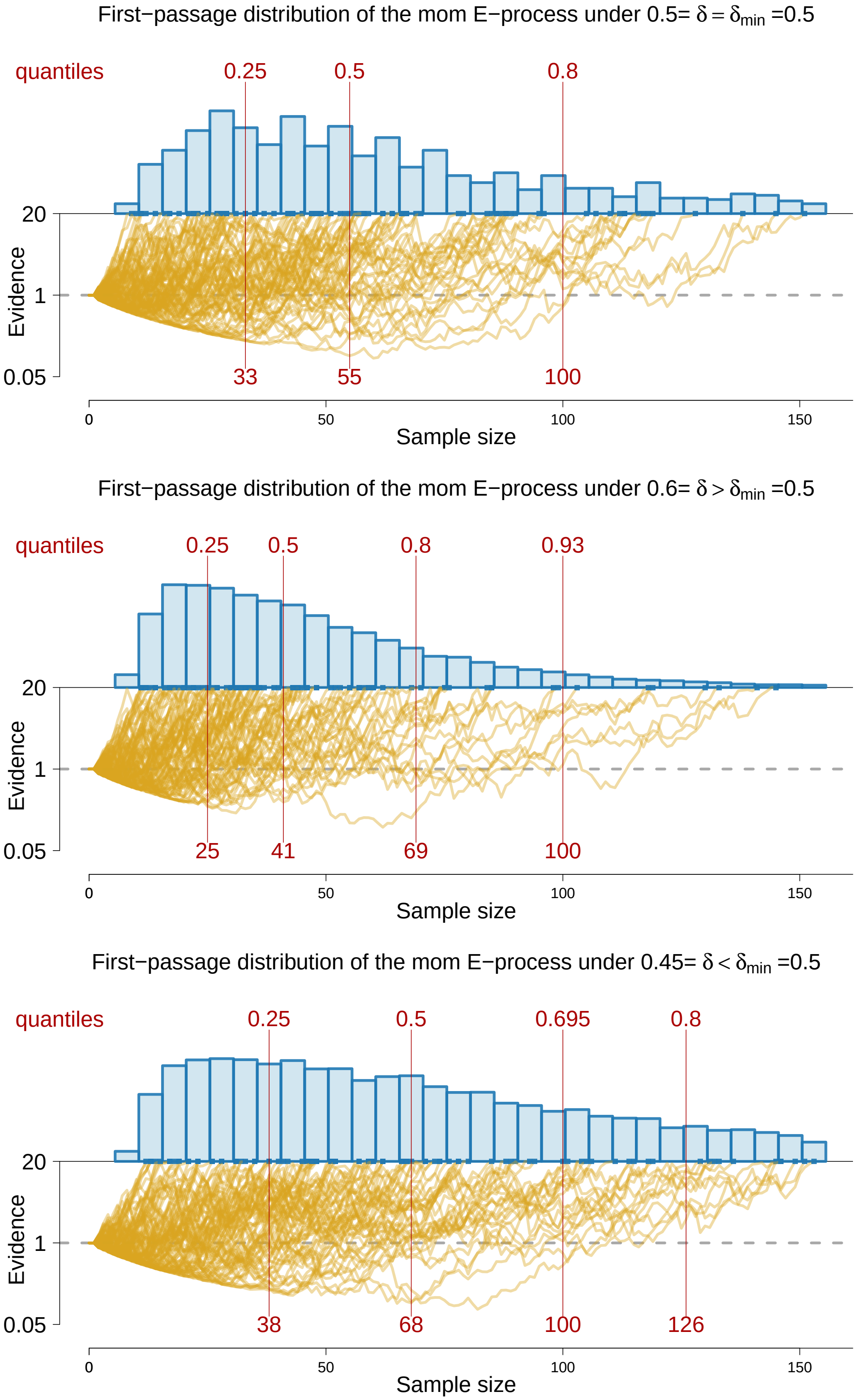
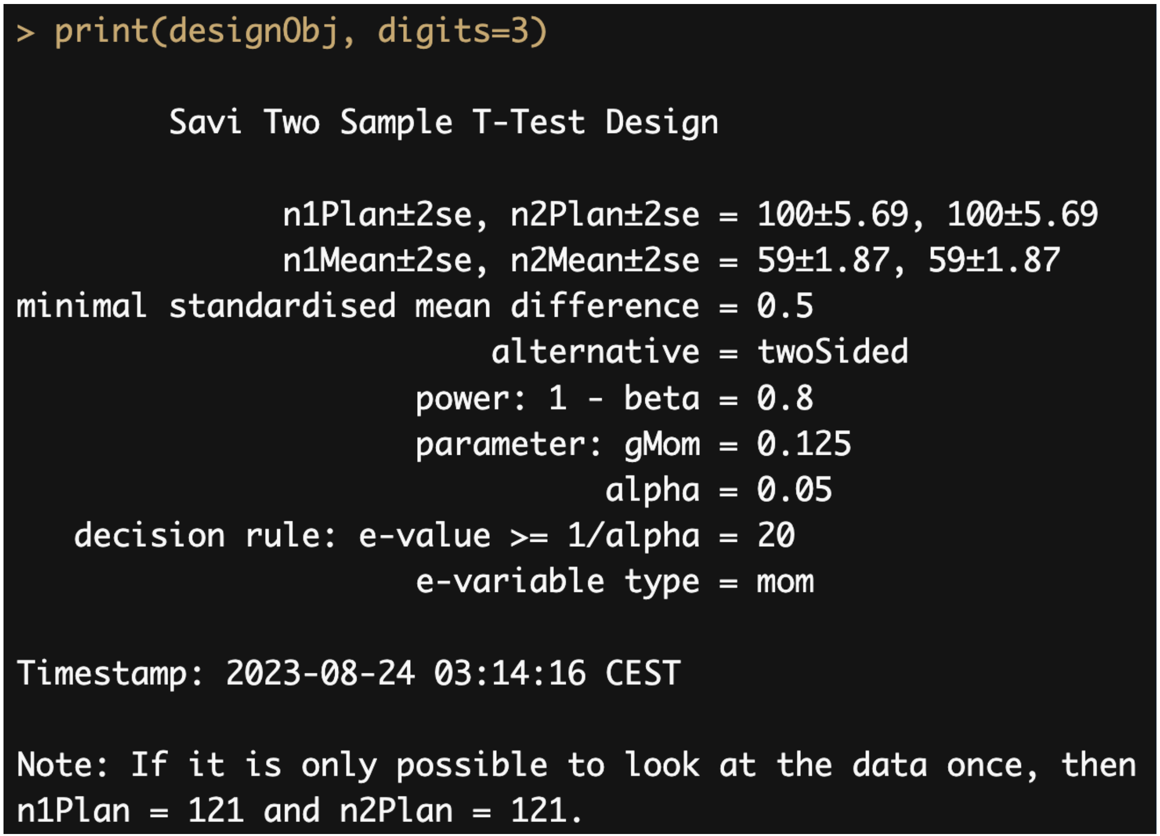


Figure 6

Regardless of the actual value of the data-governing δ, it can (eventually) be detected with a savi test. From top to bottom: Distribution of the first-passage distribution when the data-governing $\delta = \delta_{\min} = 0.5$, $\delta = 0.6 > \delta_{\min}$, and $\delta = 0.45 < \delta_{\min}$, respectively.



```
> print(designObj, digits=3)

        Savi Two Sample T-Test Design

              n1Plan±2se, n2Plan±2se = 100±5.69, 100±5.69
              n1Mean±2se, n2Mean±2se = 59±1.87, 59±1.87
minimal standardised mean difference = 0.5
                         alternative = twoSided
                     power: 1 - beta = 0.8
                     parameter: gMom = 0.125
                               alpha = 0.05
   decision rule: e-value >= 1/alpha = 20
                     e-variable type = mom

Timestamp: 2023-08-24 03:14:16 CEST

Note: If it is only possible to look at the data once, then
n1Plan = 121 and n2Plan = 121.
```

Figure 7

Screenshot of the design object created on Line 3 in `R` *Code 2.1.*

2.5 Other power analysis scenarios: Determining the power $1 - \beta$, or the minimal detectable effect size $\delta_{\min}$

There are circumstances where the available budget forms the bottleneck of our investigation, yielding a restriction on the sample sizes, say, at most $n_1 = n_2 = 40$. Before running Protocol 1.1 with $n_{\text{plan}} = 40$, we might want to get an indication of the δ that we can detect with, say, 80% power. By providing the `designSaviT` function with $\alpha, n_{\text{plan}}, \beta$, e.g. Lines 3 and 4 of `R` Code 2.2, we see that the smallest effect that we can detect with 80% probability is about $\delta_{\min} = 0.882$. Subject experts might claim that an effect size of $\delta = 0.7$ is more realistic. Such an effect can still be detected if we sample up to $n_1 = n_2 = 40$, but with smaller probability. Providing `designSaviT` with $\alpha, n_{\text{plan}}, \delta_{\min}$, e.g. Lines 5 and 6 of `R` Code 2.2, we see that under $\mathbb{P}_\delta$ with $\delta = 0.7$ we have 64.2 % power to observe $e_n \geq 20$ by monitoring up to $n_{\text{plan}} = 40$.

If any of these prospective analyses show that (a) the planned sample size is too high, (b) the smallest detectable effect size is unrealistically large, or (c) the power is too low, then we can either request more funds to invite the derived additional number of participants to the study, or decide, in advance, that it is futile to conduct this experiment, and spend our time and efforts on different endeavours instead.

```r
1 designNPlan ← designSaviT(deltaMin=0.5, beta=0.2, seed=13,
2                            alpha=0.05, testType="twoSample")
3 designDeltaMin ← designSaviT(nPlan=c(40, 40), beta=0.2, seed=3,
4                               alpha=0.05, testType="twoSample")
5 designBeta ← designSaviT(deltaMin=0.7, nPlan=c(40, 40),
6                           alpha=0.05, testType="twoSample", seed=2)
7 designNarrowestInterval ← designSaviT(nPlan=c(40, 40), alpha=0.05,
8                                        testType="twoSample")
```

`R` **Code** 2.2: The design function `designSaviT()` takes as input α and any two of the three quantities $\beta, \delta_{\text{min}}, n_{\text{plan}}$ to yield an indication of the remaining quantity and the optimal parameter as output. Lines 1 and 2 specify as input $\beta, \delta_{\text{min}}$ and output n_{plan}, Lines 3 and 4 take as input β, n_{plan} and output an indication of δ_{min}, Lines 5 and 6 take as input $\delta_{\text{min}}, n_{\text{plan}}$ and yield β, thus, power. Lastly, Lines 7 and 8 take as input n_{plan} and output the parameter that at n_{plan} has the narrowest confidence interval.

Lastly, if neither δ_{min} nor a desired power $1 - \beta$ is known in advance, we can run `designSaviT` with only α, n_{plan}, e.g. Lines 7 and 8 of `R` Code 2.2. The underlying code then finds the parameter value that *at the specified* n_{plan} has the narrowest confidence interval. For $n_{\text{plan}} = n_1 = n_2 = 40$ this yields $g_{\texttt{mom}} = 0.134$. It is worth noting that the associated `mom` confidence interval becomes even narrower if we continue sampling. For instance, at $n_{\text{plan}} = 40$ the narrowest `mom` confidence width is 1.315 attained by $g_{\texttt{mom}} = 0.134$, if $s^2 = 1$. But at $n_1 = n_2 = 100$ the interval width is shrunk to 0.85. Running Lines 7 and 8 of `R` Code 2.2 with $n_{\text{plan}} = n_1 = n_2 = 100$ reveals that the narrowest `mom` interval is attained by $g_{\texttt{mom}} = 0.05193$, resulting in a width of 0.82. In all cases, the design function optimises the requested E-process, which can then be used for subsequent inference as exemplified below (Section 4).

The key point is that e-value power analyses in t-test settings can be derived with just a few lines of code. The resulting `designObj` specifies the type of E-process to use, the tuning parameter for inference, and the testing duration required to detect δ_{min} with $1 - \beta$ power. To prevent concerns about cherry-picking, these specifications, thus, `designObj`, can be saved and submitted to an online pre-registration repository before conducting the experiment. Importantly, the planned sample sizes are indicative only and do not represent a binding commitment or guarantee regarding the number of samples to be collected. The realised sample size at which the experiment can be stopped can be smaller(larger) than planned, when the true effect is larger(smaller) than minimal clinically relevant. Regardless of the actual value of the data-governing δ, we can safely adapt our experiments to the available evidence, as quantified by the e-value.

3 Practical guidance for selecting amongst various types of E-processes

In Section 1.2.5 we listed four types of E-processes in the t-test setting. Each type provides type I error control over time, though some are better suited for certain purposes than others. In this section, we provide insights that guide the choice of the E-process type.

3.1 Default choice: The mom E-process

We recommend the mom E-process (Johnson & Rossell, 2010) as a default procedure, because the resulting savi test provides the best balance between efficiency and robustness to the choice of δ_{min}. It also yields well-concentrated anytime-valid confidence intervals (blue curves in Fig. 9). Table 1 shows n_{plan} for the four types of E-processes when $\alpha = 0.05$ and $\beta = 0.2$ in the two-sample t-test setting with $n = n_1 = n_2$ for $\delta_{\text{min}} = 0.5$ and $\delta_{\text{min}} = 0.2$. The ordering of the types of E-processes in terms of the lowest n_{plan} remains the same for various values of α, β, δ_{min}, e.g. Fig. 8. For a fair comparison between the four types of E-processes, we chose the fastest evidence accumulating E-process within each type (e.g. Section 1.2.5). The average sample sizes can be found under the columns with header n_{mean}. For instance, under $\delta = \delta_{\text{min}} = 0.5$, the first two rows show that the procedure, which monitors the grow (or mom) e-value up to n_{plan} and stops as soon as $e_n \geq 1/\alpha$, will, on average, stop after $n_{\text{mean}} = n_1 = n_2 = 56$ (or $n_1 = n_2 = 59$) participants. For the same α, β and δ_{min}, the classical p-value test should always be performed at $n_1 = n_2 = 64$ – even if $\delta > \delta_{\text{min}}$. In contrast, all E-process can adapt to the advantageous situation, allowing experiments to be be stopped even earlier, which further increases resource savings.

Table 1

The mom E-process provides the best balance between efficiency and robustness to the choice of δ_{min}. The table shows the planned and average sample sizes of $n_1 = n_2$ based on $\alpha = 0.05$ and $\beta = 0.2$. To acknowledge that these planned and average sample sizes are found by simulation we also included two times the bootstrap standard errors. The grow E-process has the lowest n_{plan} for $\delta \geq \delta_{\text{min}}$, but requires an enormous number of samples to yield high evidence with $1 - \beta$ probability under $|\delta| < \delta_{\text{min}}$. The mom E-process is more robust to the choice of δ_{min} and yields competitive n_{plan} close to those of the grow E-process when $|\delta| \geq \delta_{\text{min}}$. The average sample sizes of grow and mom are lower than the sample sizes needed for a classical fixed sample size p-value test.

Tuning	$\delta_{\text{min}} = 0.5$				$\mathring{\delta}_{\text{min}} = 0.2$	
True	$\delta = \delta_{\text{min}}$		$\delta = 0.8 > \delta_{\text{min}}$	$\delta = 0.2 < \delta_{\text{min}}$	$\delta = \mathring{\delta}_{\text{min}}$	
	n_{plan}	n_{mean}	n_{plan}	n_{plan}	n_{plan}	n_{mean}
grow	89 ± 1.71	56 ± 0.48	40 ± 0.99	2612 ± 25.93	531 ± 11.83	323 ± 2.96
mom	100 ± 1.77	59 ± 0.59	40 ± 0.97	814 ± 14.46	596 ± 10.27	344 ± 3.61
eGauss	106 ± 1.75	66 ± 0.61	45 ± 0.77	702 ± 13.38	643 ± 12.87	385 ± 3.79
eCauchy	114 ± 1.88	70 ± 0.68	47 ± 0.90	741 ± 13.58	686 ± 11.51	407 ± 4.21

Under $\delta = \mathring{\delta}_{\text{min}} = 0.2$, we get average sample sizes of $n_{\text{mean}} = 323$ and 344 participants in both groups for the grow and mom E-processes, respectively, whereas the classical p-value test should always be performed at $n_1 = n_2 = 394$. Hence, both the grow and mom E-process will on average outperform the classical p-value test.

The fact that the grow E-process yields the lowest n_{plan} is due to it being the fastest evidence accumulating procedure amongst all E-processes. However, in case the grow t-test was optimised for $\delta_{\text{min}} = 0.5$, but the data-governing effect size is actually $\delta = 0.2$, then the grow E-process requires many more samples to yield $e_n \geq 1/\alpha$ with $1-\beta = 80\%$ power. If the `grow` E-process were tuned to $\mathring{\delta}_{\text{min}} = 0.2$ from the beginning, then it only needs $n_{\text{plan}} = 531$ instead of $n_{\text{plan}} = 2612$, which equates to a relative increase of 392%. The relative increase is much less for `mom` (36.6%), `eGauss` (9.2%), and `eCauchy` (8.0%), though, the latter two types have higher baselines (643 and 686 compared to 596 respectively), see also Fig. 8.

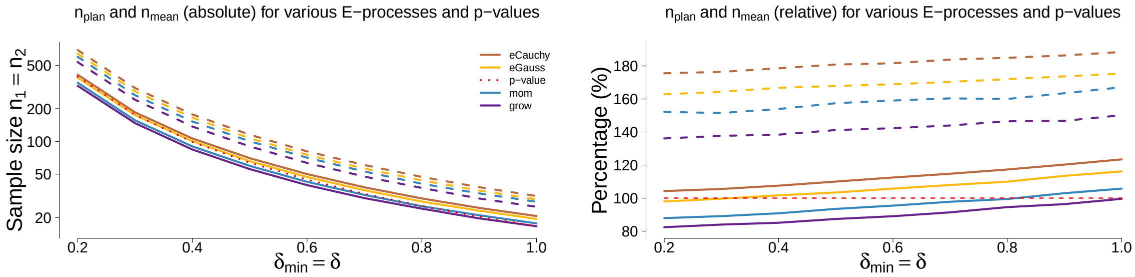


Figure 8

To produce persuasive support for δ with 80% power, the savi tests require higher (worst-case) planned sample sizes, compared to the classical fixed sample p-value test under $\delta = \delta_{\text{min}}$. On average the `grow` and `mom` E-process tests outperform the classical test, and the gain is higher when the data-governing $\delta = \delta_{\text{min}}$ is smaller. Left panel: The dotted thin red line represents the number of samples needed for the classical p-value test to reject the null with 80% under δ shown on the x-axis. The four dotted lines at the top represent the worst-case n_{plan} of `eCauchy` (brown), `eGauss` (yellow), `mom` (blue) and `grow` (purple). The solid lines represent the average sample sizes. Right panel: The same information as in the left panel is shown, but scaled so the p-value sample sizes are set to 100% representing the baseline. Roughly speaking, the additional data in the worst-case planning stage required for `eCauchy` (brown) is 82%, for `eGauss` (yellow) is 69%, for `mom` (blue) is 58%, and for `grow` (purple) is 42%. Roughly speaking, we require for `eCauchy` (brown) 12% and for `eGauss` (yellow) 6% more data on average. For `mom` (blue) we require 4%, and for `grow` (purple) 10% less data.

3.2 Fast detection: The `grow` E-process

If the focus is on sequentially detecting an effect as quickly as possible, then there is a case to be made for the `grow` E-process. It has the theoretical desirable property of being the fastest evidence accumulating E-process in the worst case amongst *all* E-processes for $\mathcal{M}_0$. Table 1 suggests the `grow` choice when effects smaller than δ_{min} are truly uninteresting or impossible to measure due to the limits of our measurement instruments. Unfortunately, this `grow` E-process comes with the additional caveat that its associated anytime-valid confidence interval stops shrinking after a certain sample size. If it can be guaranteed that sampling will not exceed a certain sample size, then `grow` confidence intervals can still be reasonable. This, however, might be hard to guarantee in practice.

Figure 9

The confidence interval associated to the `grow` (purple) E-process stops shrinking after a certain sample size. The other confidence intervals shrink and are hard to distinguish from each other. Listed in order of widest to narrowest at $n = 1024$ we have: `grow` in purple, `mom` in blue, `eCauchy` in brown, and `eGauss` in yellow. The data were generated with a true mean difference of $\varphi = 1$.

3.3 Eventually narrowest anytime-valid confidence interval: The `eGauss` E-process

One of the major advantage of the `eGauss` E-process is that both the two-sided e-value and the anytime-valid confidence interval have an explicit form, which makes them a tad less hard to analyse (H. Wang & Ramdas, 2025). If the goal is to eventually get the most precise inference regarding the magnitude of the effect, then the `eGauss` E-process can be recommended. Listed in order of widest to narrowest at $n = 1024$, Fig. 9 shows the confidence intervals associated with `grow` in purple, `mom` in blue, `eCauchy` in brown, and `eGauss` in yellow. The differences between `mom`, `eCauchy` and `eGauss` are hardly visible. This (eventually) narrower confidence width comes at a cost in terms of somewhat larger planned sample sizes for the test compared to those of `grow` and `mom`. Note that this choice for `eGauss` relies on long term gains, as it does not yield the narrowest interval at all times. For instance, the `grow` interval is the narrowest between $n = 16$ and $n = 64$.[14]

3.4 Information consistent inference: The `eCauchy` E-processes

Of the listed E-processes, only the `eCauchy` E-process is information consistent. It will therefore yield irrefutable evidence against the null in case the data are overwhelmingly

[14] Hence, in terms of confidence intervals width there is no procedure that outperforms another at all time points simultaneously.

informative. For the case at hand, overwhelmingly informative data correspond to observing a non-zero sample mean difference without any sampling variability, see Jeffreys (1961), and Ly et al. (2016a, 2016b) for more details. When it comes to n_{plan} `eCauchy` performs relatively poorly, while its confidence interval width is close to that of `eGauss`. Another reason to choose `eCauchy` would be its overall robustness to the specification of δ_{min}.

Table 1 shows that choosing δ_{min} close to the data generating δ can result in fast detection of $|\delta| \geq \delta_{\text{min}}$, but a relatively harsh penalty in terms of n_{plan} whenever $|\delta| < \delta_{\text{min}}$. This is not a reason to choose δ_{min} as small as possible, as a smaller δ_{min} yields a larger n_{plan}, which in turn leads to wasteful testing whenever the null holds true, see Fig. 4. The increase in n_{plan} for $|\delta| < \delta_{\text{min}}$ should not pose a problem if δ_{min} truly represents the minimal clinically relevant effect size. Conventions in the field such as those posed by Cohen (1988) can help select δ_{min}. A meta-analysis can also help establish δ_{min}, or it can be informed by an original finding that we aim to replicate.

4 Two real-world examples illustrating e-value based inference in action

We show how e-values can speed up inference by examining two replication attempts from the ManyLabs2 Project (Klein et al., 2018), which investigated variations across samples and settings in the replicability of 28 classic psychological findings.

4.1 Example 1: Moral typecasting (Gray & Wegner, 2009, Study 1a)

Moral typecasting is the process where a moral agent (doer of right or wrong actions) is less likely to be perceived as a receiver of that action, and vice versa. Gray and Wegner (2009) argued that age shapes perceptions of morality and postulated that children are perceived as less responsible for their actions.[15]

In the original experiment 69 participants read a story about either an adult man (high in moral agency) harming a baby, or a baby (low in moral agency) harming an adult man by knocking over a tray of glasses. Participants rated the responsibility of the offender on a 7-point scale from 1 (low) to 7 (high). On average, participants who read the story of the offending adult man rated him as more responsible ($\bar{x}_1 = 5.29, s_1 = 1.86$) compared to participants who were presented with the offending baby ($\bar{x}_2 = 3.86, s_2 = 1.64$). It was concluded that the observed mean difference was significant due to observing $t(68) = 3.32$ and $p = 0.001 < \alpha$.

Out of 61 ManyLabs2 replication attempts, a total of 58 (95.08%) led to a significant p-value. The pre-registration required each replication attempt to contain at least 80 participants, say, $n_1 = n_2 = 40$ in each group. To ensure that the $p < \alpha$ tests remain valid, data collection across all attempts were first stopped before the tests were conducted. The replication results were based on $n_1 = 4028$ and $n_2 = 3974$ participants in the two groups.

A reasonable estimate of the underlying standardised effect size based on the original findings (Gray & Wegner, 2009, Study 1a) is within $(0.769, 0.872)$.[16] We err on the side

[15] They also postulated that children are viewed as having less intention of doing right or wrong, and more likely to be perceived as receivers of moral actions compared to adults. The main focus here, as in ManyLabs2, is on perceived responsibility.

[16] The lower (upper) bound equals the observed mean difference 5.29 − 3.86 divided by the largest (smallest) sample standard deviation. Typical estimates divide the mean difference by some type of average between the two standard deviations.

of caution by using the lower bound for our power analysis. Furthermore, we can argue for a one-sided test (Grünwald & Koolen, 2025), as the working hypothesis deals with the offending adult man being perceived as more, not less, responsible compared to the offending baby. A classical power analysis shows that for $\delta_{\text{min}} = 0.769$ and $1 - \beta = 80\%$, the one-sided p-value test should be performed after gathering data from $n_1 = n_2 = 22$ participants in each group.

Lines 1 to 3 of R Code 4.1 show that with e-values we should plan for $n_{\text{plan}} = n_1 =$

```r
1 deltaMin ← (5.29-3.86)/1.86
2 designObj ← designSaviT(deltaMin=deltaMin, beta=0.2, seed=1,
3                          testType="twoSample", alternative="greater")
4 plot(designObj)
5
6 result ← saviTTest(x, y, designObj=designObj, sequential=TRUE)
7 plot(result)
8 plot(result, wantConfSeqPlot=TRUE)
```

R **Code** 4.1: Code adapted from R Code 2.1 for the ManyLabs study, see the OSF repository `https://osf.io/mdbqe/files/osfstorage` for full details.

$n_2 = 37$ participants in each group to observe evidence against the null larger than 20 with 80% probability under $\delta_{\text{min}} = 0.769$.[17] The procedure that samples until this n_{plan} or the first time $e_n \geq 1/\alpha$ will then on average stop after $n_{\text{mean}} = 22$ participants in each group, if $\delta = \delta_{\text{min}}$.

We compare the two designs side by side. For the classical analysis we pretend that for each replication attempt, we sample up to $n_1 = n_2 = 22$, stop the experiment, and then compute the one-sided p-value.[18] By reducing the sample sizes, we are going to have less power to reject the null. Indeed, this procedure results in 48 null rejections at $p < 0.05$ out of a total of 61 replication attempts (78.7%). For this conclusion, we used data from a total of $n_1 = 1315$ and $n_2 = 1324$ participants, thus, on average $n_1 = 21.56$ and $n_2 = 21.71$ in each replication attempt. Hence, this classical design required 2713 and 2650 fewer participants.

The one-sided e-value test allows for informed conclusions with even less data. Tracking the e-value up to $n_{\text{plan}} = 37$ or stopping as soon as $e_n \geq 20$ yields 51 times persuasive evidence for the effect, out of a total of 61 replication attempts (83.6%). This evidence is gathered with a total of $n_1 = 1031$ and $n_2 = 1062$ participants, corresponding to an average of $n_1 = 16.90$ and $n_2 = 17.41$ per replication attempt. Compared to the original set-up, stopping as soon as $e_n \geq 1/\alpha = 20$ reduced the number of participants by 2997 in one group and 2912 in the other, resulting in roughly 73% resource savings.

This demonstration highlights the benefits of planned analyses in general, and efficiency of e-value based tests in particular. The general conclusions remains the same if we tuned the E-process to the upper bound for δ_{min} instead. Hence, the precise specification

[17] For a two-sided e-value test we require $n_{\text{plan}} = 45$ and the procedure is then expected to stop after $n_{\text{mean}} = 28$ in each group on average.

[18] For trials that gathered fewer than $n_1 = n_2 = 22$ data points the p-value test is done at the end of the trial. This was, for instance, the case for the "tanzaniaon" data set, which had only $n_1 = 3$ and $n_2 = 13$ valid responses.

of δ_{min} is not necessary to gain efficiency and conserve resources.

The left panel of Fig. 10 shows in yellow the evolution of the e-value sequences that hit $1/\alpha = 20$, and in green the sequences that did not reach this threshold before $n_{\text{plan}} = 37$.

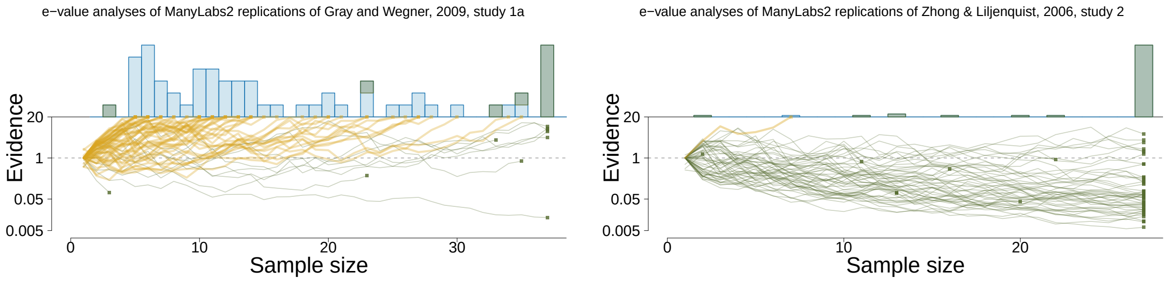


Figure 10

Left panel: 83.6% of the e-value sequences (yellow) of the Gray and Wegner (2009) Many-Labs2 replication attempt reached $1/\alpha = 20$ *before* $n_{\text{plan}} = 37$*, whereas the green sequences did not. The green part of the stopping histogram corresponds to an e-value sequence that did not reach* $1/\alpha = 20$*. Right panel: 98.2% of the e-value sequences (green) of the Zhong and Liljenquist (2006) ManyLabs2 replication attempt did not reach* $1/\alpha = 20$*. There was one sequence that hit* $1/\alpha = 20$*.*

4.2 Example 2: The Macbeth effect - Moral violations and desire for cleansing (Zhong & Liljenquist, 2006, Study 2)

Zhong and Liljenquist (2006) hypothesised that a threat to one's moral purity induces the need to (physically) cleanse oneself, which they referred to as the "Macbeth effect". In Study 2 a total of 27 participants copied a first-person account of an ethical act (helping a co-worker), or an unethical act (sabotaging a co-worker). Afterwards, the participants rated the desirability of five cleaning products and five non-cleaning products on a scale from 1 (not at all) to 7 (very much). Participants who copied the unethical story ($\bar{x}_1 = 4.95, s_1 = 0.84$) found the cleaning products more desirable compared to participants who copied the ethical story ($\bar{x}_2 = 3.75, s_2 = 1.32$). The null hypothesis of no effect was rejected based on $t(25) = 2.64$ and $p = 0.01 < \alpha = 0.05$.

There were 57 ManyLabs2 replication attempts that fulfilled the inclusion criteria, of which 3 (5.26%) yielded a significant p-value less than $\alpha = 0.05$. These conclusions were based on $n_1 = 3439$ and $n_2 = 3508$ participants.[19] A reasonable estimate of the data-governing standardised effect size based on the original findings in Zhong and Liljenquist (2006, Study 2) is within $(0.909, 1.429)$. As before, we err on the side of caution by using the lower bound for our design. A classical power analysis shows that for $\delta_{\text{min}} = 0.909$ and $1 - \beta = 80\%$ power, the one-sided p-value test should be conducted after gathering data from $n_1 = n_2 = 16$ participants in each group. Lines 2 to 3 of `R` Code 4.1 with `deltaMin <-`

[19] Unfortunately, we were unable to retrieve the correct "bogota" data from the data sets uploaded at `https://osf.io/8cd4r/`, which is why our results differ slightly compared to what is reported in Klein et al. (2018).

`(4.95-3.75)/1.32` shows that with the correspondingly tuned `mom` E-process, we should plan for $n_{\text{plan}} = n_1 = n_2 = 27$ participants in each group to observe $e_n \geq 20$ with 80% power under $\delta_{\min} = 0.909$. If $\delta = \delta_{\min}$ the procedure is expected to stop after testing $n_{\text{mean}} = 17$ participants in each group.[20] For most replication attempts the one-sided e-value test hit n_{plan}. We acquired $e_n \geq 1/\alpha = 20$ in one out of a total of 57 replication attempts (1.75%) based on a total of $n_1 = 1427$ and $n_2 = 1441$ participants per group, thus, $n_1 = 25.04$ and $n_2 = 25.28$ on average. If the magnitude of the data governing δ was indeed at least $\delta_{\min}$, then we have provided such δ ample chance to yield persuasive evidence. Since this did not occur for most replication attempts, we infer that the postulate $\delta \geq \delta_{\min}$ is unlikely with high probability. Moreover, we come to this conclusion with 2012 and 2067 fewer participants per group, which is a reduction of 58.5% and 58.9% from the total sample sizes used in the original analysis.

We would like to reiterate the point that with e-values we do not have to stop at the first-passage time $e_n \geq 1/\alpha$, nor at n_{plan}. Due to E-processes being robust to any stopping time, there is no need to discard newly available data once the test is conducted or the confidence interval is computed, as is the case with a classical analysis.

5 Summary, discussion and concluding remarks

Determining the appropriate sample size at which a classical p-value test or confidence interval should be computed is difficult. This is especially true before experimentation has started when no data are present. The fact that the valid use of classical p-value tests and confidence intervals is limited to them being performed once – and only once – puts undue pressure on the well-intentioned researcher dedicated to upholding the highest standards of research practice through pre-registering their confirmatory analyses.

This problem is circumvented with e-value based methods, which can be used flexibly, allowing us to adapt the experiment to new information as they become available. Only a few lines of code suffice, e.g. `R` Code 2.1, to derive a power analysis with a *non-binding* planned sample size n_{plan} based on an optimal (by default) `mom` type t-test E-process. Analogous code can be used to construct optimal savi z-tests (i.e. `designSaviZ()`), savi tests for two proportions (Turner, Ly, & Grünwald, 2024) (i.e. `designSaviTwoProportions()`), savi logrank test (Ter Schure et al., 2024) (i.e. `designSaviLogrank()`), and many more are scheduled to be implemented into the `safestats` `R` package (Ly et al., 2024).

Simulations with the aggressive first-passage time illustrated that the probability of misleading evidence larger than $1/\alpha$ remains bounded by α, despite flexible use of the optimal E-process. At the same time, under the alternative hypothesis, the method demonstrated higher average power compared to classical approaches. This increase in power translates into smaller expected sample sizes, which further highlights the non-binding nature of n_{plan}. The ability to derive more reliable conclusions with less data allows us to save time, money and effort that can be effectively allocated to other research endeavours. For one set of replication attempts from the ManyLabs2 project, the use of the optimal E-process allows the number of participants to be reduced by more than 70% (Section 4.1), whereas it can result in a reduction of 58% in another set (Section 4.2). These conclusions

[20] For a two-sided e-value test we require $n_{\text{plan}} = 33$ and the procedure stops after $n_{\text{mean}} = 21$ in each group on average.

should not be viewed as criticism on the ManyLabs2 project, which did not exclusively focus on replicating a particular effect efficiently, but also aimed to examine the variation in replicability across samples and settings. Once the data are collected, it is best to use them all, yielding narrower anytime-valid confidence intervals, thus, more precise inference of the effects of interest.

5.1 Summary of (dis-)advantages of e-values

While anytime-valid inference methods provide clear advantages, they also introduce certain costs, most of which are manageable but worth acknowledging. We already discussed some of these points in the preceding text; for clarity, we summarise them here:

1. **Sometimes more data are needed after all** In settings where the evidence cannot be monitored and the test is designed for a single application, sequential methods may be less powerful compared to conventional p-value approaches (e.g. Fig. 8). The phenomenon we observed for the t-test holds for many other models (e.g. the logrank test (Ter Schure et al., 2024) and contingency tables (Turner et al., 2024)) as well: to obtain a desired power, less data than in a classical test is needed *on average*, but more (typically between 40 and 70%) is needed in *in the worst-case*, namely, n_{plan}.[21]

2. **Biased estimation** If a data-dependent stopping time is used, then standard *point estimators* of effect sizes may become biased for small sample sizes (Section 1.2.6). Unbiasedness is not promised by e-value methods, but they do promise to yield intervals that forever contain the true value with the advertised probability. For large sample sizes, these intervals are guaranteed to yield unbiased estimates, as long as they continue to shrink over time, as is the case for the `mom` type E-process. The development of estimators that account for the bias due to using a specific data-dependent stopping time is still a subject of ongoing research.

3. **Choice of E-process/prior** When using e-value based inference, choices must be made regarding both the type of E-process to rely on (Section 3), and the tuning parameter within that type. These choices thus introduce a level of subjectivity to the analysis, which might be viewed as undesirable: suppose a research group performs a psychological experiment and reports both the raw data and the corresponding evidence in terms of e-values; then a second group re-analyses the same data. The second group gets a different result, if they use a different tuning parameter. In simple settings, such as with t-tests, there is essentially only one reasonable p-value procedure to choose from, so both groups arrive at the same result. While this may seem preferable, we should add that the classical t-test is not without a similarly subjective choice either: instead of choosing an E-process type and tuning parameter, we then have to choose the fixed sample size n at which the analyses are performed (based on a power analysis, e.g. van Zwet et al., 2023). Two different research groups, each responsible for both the experimental design and analysis, will most likely arrive at quite a different sample size n, and hence a different results. Since only one group

[21] There also exist so-called *all-or-nothing E-variables* which achieve exactly the same power as classical methods, but these are not useful when, at a later stage, new data come in after all (Shafer, 2021).

gets to design the experiment, the subjective component remains invisible, but it is there nevertheless.

It is also important to note that while the tuning parameter in the selection of the E-process takes the form of a prior distribution, the subjectivity it introduces differs fundamentally from that in Bayesian methods. Specifically, the type I error guarantees of e-value based methods remain valid *regardless of the choice of E-variable or prior*: altering the prior affects only the statistical power, not the validity of the results. In the context of confidence intervals, the choice of prior affects their width but does not compromise their validity. By contrast, in Bayesian statistics, poorly informed priors lead to unreliable posteriors and, hence, invalid credible intervals (Grünwald, 2023).

Point 3 suggests that the methodology may be vulnerable to cherry-picking biases. To ensure reliable inference with e-values, it is therefore crucial to pre-register the chosen E-process, its tuning parameter, and the confirmatory hypotheses in advance. What is not required, however, is strict adherence to fixed sample sizes or even the specified significance thresholds. We elaborate on this latter advantage below (cf. point ii) in the list of additional benefits of e-values.

i. Choice of E-process/prior The choice of priors can be viewed as an advantage, as it enables the integration of existing data to inform the design of E-processes. Once chosen, the ensuing savi tests and anytime-valid confidence intervals will remain statistically valid, and the better the prior was informed, the narrower the confidence interval will tend to be (Grünwald, 2023).

ii. The e-value as a continuous measure of evidence; Roving α Similar to Bayesian approaches, e-values circumvent the paradoxes stemming from counterfactual reasoning that underlies p-values and classical confidence intervals (Wagenmakers, 2007). In fact, we can even set $\alpha = 1/e$, where e is the observed e-value. The resulting data-adapted α then no longer functions as a type I error bound, but can be meaningfully interpreted in terms of an *expected loss* (Grünwald, 2024). We do not necessarily advocate this practice; our point is mainly to illustrate that a large e-value has inferential meaning *in its own right*, independently of a pre-set α. Doing so within the classical context, i.e. $\alpha := p$, where p is the observed p-value, however, destroys all guarantees of the inferred procedure (Goodman, 1999; Hubbard & Bayarri, 2003).

iii. E-value based multiple testing/nonparametrics Waudby-Smith and Ramdas (2024) showed that E-variables require far fewer samples to reach the same statistical power as existing Bayesian or classical methods in certain nonparametric settings, where distributional assumptions are minimal or absent. Their impact is equally transformative in multiple testing scenarios, where e-values are driving what can be described as a paradigm shift in the field (R. Wang & Ramdas, 2022; Xu et al., 2025).

iv. Meta-analysis While e-value-based methods might require larger sample sizes during the planning phase of a *single* study, their advantages become most evident in reproducibility studies such as ManyLabs2, where studies are replicated repeatedly. In these contexts, the *law of large numbers* ensures that the average outcome, rather than the worst-case scenario, prevails: across multiple studies, e-value methods almost surely require fewer data points than classical tests. This principle also extends to meta-analyses, where e-values prove especially valuable (Grünwald et al., 2024; Ter Schure et al., 2022). In such settings, a meta-analytical e-value can be constructed to robustly synthesise evidence as data accumulate across multiple studies, while maintaining type I error control, even when individual studies vary in sample sizes or trigger replication attempts.

The list above shows that e-values are far more than a tool for anytime-valid testing — they represent a new paradigm of inference, not just "yet another method". Their key property is flexibility: they adapt not only to different sampling plans, but also enable the incorporation of existing data, and to data-dependent adjustments of significance thresholds. This flexibility simplifies statistical practice. By contrast, classical approaches impose rigid constraints that often obstruct rather than advance scientific progress. Reproducible and generalisable science certainly requires effort, but that effort should not be squandered on upholding outdated, rigid sampling plans that e-values render unnecessary.

Acknowledgements

We express our deepest gratitude to Judith ter Schure, whose initial efforts laid the foundation for this manuscript. We also acknowledge the discussions with Wouter Koolen, Rianne de Heide, and the participants of the SAVI workshop at the 56th Annual Meeting of the Society for Mathematical Psychology in Amsterdam in 2023, which greatly contributed to shaping the content of this work.

Declarations

Funding

This research was partly financed by (a) the Dutch Research Council NWO, as part of the VENI projects "Increasing Scientific Efficiency with Sequential Methods" (VI.Veni.211G.040, Alexander Ly) and "Efficient Models of Decision-Making for Assessing Cognitive Processing States" (VI.Veni.201G.045, Udo Boehm), and (b). the European Union via ERC Advanced Project 101142168 FLEX. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency.

Competing interests

The authors declare no competing interests.

Availability of data and materials

In Section 4 we reanalyse data from ManyLabs2 (Klein et al., 2018). The extracted data are stored on our OSF repository.

Code availability

All the code for the simulation examples is available at `https://osf.io/mdbqe/files/osfstorage`.

6 References

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305–307.

Bickel, P. J., & Doksum, K. A. (2015). *Mathematical statistics: Basic ideas and selected topics Volume I* (2nd ed. ed.). Chapman and Hall/CRC.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed. ed.). Lawrence Erlbaum Associates.

De Heide, R., & Grünwald, P. D. (2021). Why optional stopping can be a problem for Bayesians. *Psychonomic Bulletin & Review*, *28*(3), 795–812.

Fischer, L., & Ramdas, A. (2024). Improving the (approximate) sequential probability ratio test by avoiding overshoot. *arXiv e-prints*, arXiv–2410.

Ghosal, S., & van der Vaart, A. W. (2017). *Fundamentals of nonparametric Bayesian inference* (Vol. 44). Cambridge University Press.

Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). The Bayesian two-sample t test. *The American Statistician*, *59*(3), 252–257.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine*, *130*(12), 995–1004.

Gradshteyn, I. S., & Ryzhik, I. M. (2007). *Table of integrals, series, and products* (7th ed.). Academic Press.

Gray, K., & Wegner, D. M. (2009). Moral typecasting: divergent perceptions of moral agents and moral patients. *Journal of personality and social psychology*, *96*(3), 505.

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian *t*-tests. *The American Statistician*, *74*(2), 137–143.

Grünwald, P. D. (2023). The e-posterior. *Philosophical Transactions of the Royal Society A*, *381*(2247), 20220146.

Grünwald, P. D. (2024). Beyond Neyman-Pearson: e-values enable hypothesis testing with a data-driven alpha. *Proceedings of the National Academy of Sciences*, *121*(39).

Grünwald, P. D., De Heide, R., & Koolen, W. (2024). Safe testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *86*(5), 1091–1128. (With discussion)

Grünwald, P. D., & Koolen, W. M. (2025). Supermartingales for one-sided tests: Sufficient monotone likelihood ratios are sufficient. *arXiv preprint arXiv:2502.04208*.

Hendriksen, A., de Heide, R., & Grünwald, P. D. (2021). Optional stopping with Bayes factors: A categorization and extension of folklore results, with an application to invariant situations. *Bayesian Analysis*, *16*(3), 961–989.

Howard, S. R., Ramdas, A., McAuliffe, J., & Sekhon, J. (2020). Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, *17*, 257–317.

Howard, S. R., Ramdas, A., McAuliffe, J., & Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, *49*(2), 1055–1080.

Hubbard, R., & Bayarri, M. J. (2003). P values are not error probabilities. *Institute of Statistics and Decision Sciences, Working Paper*(03-26), 27708–0251.

Jeffreys, H. (1961). *Theory of probability* (3rd ed. ed.). Oxford, UK: Oxford University Press.

Johnson, V. E., & Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*, 143–170.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., . . . Nosek, B. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490.

Koning, N. W. (2024). Continuous testing: Unifying tests and e-values. *arXiv preprint arXiv:2409.05654*.

Koning, N. W., & van Meer, S. (2025). Free anytime validity by sequentializing a test and optional continuation with tests as future significance levels. *arXiv preprint arXiv:2501.03982*.

Koolen, W. M., & Grünwald, P. D. (2022). Log-optimal anytime-valid e-values. *International Journal of Approximate Reasoning*, *141*, 69–82.

Kotz, S. (2006). Sampling to a foregone conclusion. In S. Kotz, N. Balakrishnan, C. B. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (2nd ed., Vol. 11, pp. 7428–7429). New Jersey: John Wiley & Sons.

Larsson, M., Ramdas, A., & Ruf, J. (2025). The numeraire e-variable and reverse information projection. *The Annals of Statistics*, *53*(3), 1015 – 1043. Retrieved from `https://doi.org/10.1214/24-AOS2487` doi: 10.1214/24-AOS2487

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course.* Cambridge: Cambridge University Press.

Lindon, M., Ham, D. W., Tingley, M., & Bojinov, I. (2025). Anytime-valid linear models and regression adjusted causal inference in randomized experiments. *arXiv preprint arXiv:2210.08589v5*.

Ly, A., Marsman, M., Verhagen, A. J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). A tutorial on Fisher information. *Journal of Mathematical Psychology*, *80*, 40–55.

Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Analytic posteriors for Pearson's correlation coefficient. *Statistica Neerlandica*, *72*(1), 4–13. doi: 10.1111/stan.12111

Ly, A., Turner, R. J., Pérez-Ortiz, M. F., Boehm, U., Ter Schure, J., & Grünwald, P. D. (2024). `safestats`: Safe anytime-valid inference [Computer software manual]. (R package version 0.8.8, `https://cran.r-project.org/package=safestats`)

Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016a). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32.

Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016b). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, *72*, 43–55.

Pandeva, T., Bakker, T., Naesseth, C. A., & Forré, P. (2024). E-valuating classifier two-sample tests. *Transactions on Machine Learning Research.* Retrieved from `https://openreview.net/forum?id=dwFRov8xhr`

Pawel, S., Ly, A., & Wagenmakers, E.-J. (2024). Evidential calibration of confidence intervals. *The American Statistician*, *78*(1), 47-57. doi: 10.1080/00031305.2023.2216239

Pérez-Ortiz, M. F., Lardy, T., de Heide, R., & Grünwald, P. D. (2024). E-statistics, group invariance and anytime valid testing. *Annals of Statistics*, *52*(4), 1410–1432.

Pramanik, S., & Johnson, V. E. (2022). Efficient alternatives for Bayesian hypothesis tests in psychology. *Psychological Methods.*

Ramdas, A., Grünwald, P. D., Vovk, V., & Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, *38*(4), 576–601.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

Ruf, J., Larsson, M., Koolen, W. M., & Ramdas, A. (2023). A composite generalization of Ville's martingale theorem using e-processes. *Electronic Journal of Probability*, *28*, 1–21.

Schnuerch, M., Heck, D. W., & Erdfelder, E. (2022). Waldian t tests: Sequential Bayesian t tests with controlled error probabilities. *Psychological Methods.*

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological methods*, *22*(2), 322–339. doi: 10.1037/met0000061

Shafer, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *184*(2), 407–431.

Shafer, G., Shen, A., Vereshchagin, N., & Vovk, V. (2011). Test Martingales, Bayes Factors and p-Values. *Statistical Science*, *26*(1), 84 – 101. Retrieved from `https://doi.org/10.1214/10-STS347` doi: 10.1214/10-STS347

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359–1366.

Steinhilber, M., Schnuerch, M., & Schubert, A.-L. (2024). Sequential analysis of variance: Increasing efficiency of hypothesis testing. *Psychological Methods*.

Steinhilber, M., Schnuerch, M., & Schubert, A.-L. (2025). The dark side of sequential testing: A simulation study on questionable research practices. *PsyArXiv*.

Ter Schure, J. (2023). *ALL-IN meta-analysis* (Unpublished doctoral dissertation). Leiden University.

Ter Schure, J., Ly, A., Belin, L., Benn, C. S., Bonten, M. J., Cirillo, J. D., ... van Werkhoven, C. H. (2022). Bacillus Calmette-Guérin vaccine to reduce COVID-19 infections and hospitalisations in healthcare workers–a living systematic review and prospective ALL-IN meta-analysis of individual participant data from randomised controlled trials. *medRxiv*, 2022–12.

Ter Schure, J., Perez-Ortiz, M., Ly, A., & Grünwald, P. D. (2024). The safe logrank test: Error control under continuous monitoring with unlimited horizon. *The New England Journal of Statistics in Data Science*, *2*(2), 190–214.

Turner, R. J., Ly, A., & Grünwald, P. D. (2024). Generic e-variables for exact sequential k-sample tests that allow for optional stopping. *Journal of Statistical Planning and Inference*, *230*, 106116. Retrieved from `https://www.sciencedirect.com/science/article/pii/S037837582300085X` doi: https://doi.org/10.1016/j.jspi.2023.106116

van Zwet, E., Gelman, A., Greenland, S., Imbens, G., Schwab, S., & Goodman, S. N. (2023). A new look at p values for randomized clinical trials. *Nejm evidence*, *3*(1), EVIDoa2300003.

Vovk, V., & Wang, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics*, *49*(3), 1736–1754.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Wang, H., & Ramdas, A. (2025). Anytime-valid t-tests and confidence sequences for Gaussian means with unknown variance. *Sequential Analysis*, *44*(1), 56–110.

Wang, R., & Ramdas, A. (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *84*(3), 822–852.

Wasserman, L., Ramdas, A., & Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, *117*(29), 16880–16890.

Waudby-Smith, I., & Ramdas, A. (2024). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *86*(1), 1–27.

Xu, Z., Solari, A., Fischer, L., de Heide, R., Ramdas, A., & Goeman, J. (2025). Bringing closure to false discovery rate control: A general principle for multiple testing. *arXiv preprint arXiv:2509.02517*.

Zhong, C.-B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, *313*(5792), 1451–1452.

Appendix A
The duality between classical $1 - \alpha$ confidence intervals and $p < \alpha$ tests

A classical confidence interval inverts a p-value test. To set the scene, we consider the t-test setting where any null value φ_0 defines a T-statistic, Eq. (8). The $p < \alpha$ test then rejects the null value φ_0 whenever the observed t-statistic is larger than the threshold $t_{\alpha,\nu}$ corresponding to the $1 - \alpha/2$ quantile of a T-distribution with ν degrees of freedom. Given fixed samples sizes n_1, n_2, if the true population mean difference equals the postulated φ_0 used in the T-statistic, then there is less than α probability to observe outcomes of T with magnitude larger than $t_{\alpha,\nu}$:

$$\text{At each } n_1, n_2 \text{ and all null values } \varphi_0 \in \mathbb{R} : \mathbb{P}_{\varphi_0}\left(\left| \sqrt{n_{\text{eff}}} \frac{\bar{X}_1 - \bar{X}_2 - \varphi_0}{S_p} \right| > t_{\alpha,\nu} \right) \leq \alpha, \tag{A1}$$

A classical $1 - \alpha$ confidence interval inverts Eq. (A1) by negating the event within $\mathbb{P}_{\varphi_0}$. Recall that the probability of all outcomes sum to one, which implies that the probability of the complement of the event A denoted by A^c is given by $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, and yields:

$$\text{At each } n_1, n_2 \text{ and all null values } \varphi_0 \in \mathbb{R} : \mathbb{P}_{\varphi_0}\left(\left| \sqrt{n_{\text{eff}}} \frac{\bar{X}_1 - \bar{X}_2 - \varphi_0}{S_p} \right| \leq t_{\alpha,\nu} \right) > 1 - \alpha. \tag{A2}$$

Rearranging the statement within $\mathbb{P}_{\varphi_0}$ shows that

$$\text{At each } n_1, n_2 \text{ fixed } \text{CI}(1 - \alpha) := \left[\bar{X}_1 - \bar{X}_2 - \tfrac{S_p}{\sqrt{n_{\text{eff}}}} t_{\alpha,\nu}, \bar{X}_1 - \bar{X}_2 + \tfrac{S_p}{\sqrt{n_{\text{eff}}}} t_{\alpha,\nu} \right] \tag{A3}$$

will encapsulate a data-governing φ_0 with at least $1 - \alpha$ probability. The $1 - \alpha$ probability pertains to the interval $\text{CI}(1 - \alpha)$, as that is data-dependent, not the parameter value. This coverage probability drops well below $1 - \alpha$ if n_1, n_2 are not fixed, see Fig. 2.

Appendix B
Markov's inequality

Let X be a non-negative random variable having distribution $\mathbb{P}$, then for any $c > 0$ Markov's inequality states that

$$\mathbb{P}(X \geq c) \leq \mathbb{E}_{\mathbb{P}}[X]/c. \tag{B1}$$

The principle is that if the non-negative random variable X has a low expectation, then large values cannot occur often. If extreme values do occur often, then its expectation cannot be that low. For instance, assume that the population average is $\mathbb{E}_{\mathbb{P}}[X] = 1$, then the event $X \geq 90$ cannot occur with probability 0.8. To show this, we reason to a contradiction. The smallest value of X for which $\mathbb{P}(X \geq 90) = 0.8$ holds occurs when all 80% concentrates at $X = 90$. This minimal case already contributes 72 to the average of X and we are left with a probability of 0.2. Since X is non-negative the smallest value is 0 and giving that 20% yields 0. Summing the two shows that $\mathbb{P}(X \geq 90) = 0.8$ implies that the average of X is at least 72, which contradicts the $\mathbb{E}_{\mathbb{P}}[X] = 1$ assumption.

More formally, we recall that the expectation/population average is a weighted average, i.e. $\mathbb{E}_{\mathbb{P}}[X] = \int x p(x) \mathrm{d}x$, where the integral goes from zero to any $c > 0$ and then to ∞, thus,

$$\mathbb{E}_{\mathbb{P}}[X] = \int_0^c x p(x) \mathrm{d}x + \int_c^\infty x p(x) \mathrm{d}x \geq \int_c^\infty x p(x) \mathrm{d}x \geq c \int_c^\infty p(x) \mathrm{d} = c\mathbb{P}(X \geq c), \tag{B2}$$

where the last inequality is due to all outcomes x on (c, ∞) being no less than c, which implies Markov's inequality Eq. (B1).

Appendix C
Ville's inequality

This proof of Ville's inequality is based on Section 6.1 of Howard et al. (2021), but added for completeness. We assume that the data are governed by a $\mathbb{P} \in \mathcal{M}_0$, and we let E be an arbitrary E-process for $\mathcal{M}_0$. The first-passage time N_α is defined as the smallest time n at which $E_n \geq 1/\alpha$ occurs, and ∞, if it does not occur. That is,

$$N_\alpha := \inf\{n \in \{1, 2, \ldots, n, \ldots, n_{\text{plan}}, \ldots\} : E_n \geq 1/\alpha\}, \tag{C1}$$

where inf is the infimum defined as the largest lower bound, i.e. the minimum. The event $N_\alpha < \infty$ is equivalent to the event V

$$V := \{\text{There exists a finite time point } n \text{ at which } E_n \geq 1/\alpha\}, \tag{C2}$$

which is how the sample size comes into the probability statement. Instead of applying Eq. (5) directly to N_α, we do so for the sequence of stopping times $N_{n_{\text{plan}}}$, where $N_{n_{\text{plan}}}$ is the minimum between the time horizon n_{plan} and the first time E_n crosses $1/\alpha$. In other words, the stopping time $N_{n_{\text{plan}}}$ is defined as

$$N_{n_{\text{plan}}} := \inf\{n \in \{1, 2, \ldots, n_{\text{plan}}\} : E_n \geq 1/\alpha\}. \tag{C3}$$

Note that $N_\alpha = \lim_{n_{\text{plan}} \to \infty} N_{n_{\text{plan}}}$. Since, $N_{n_{\text{plan}}}$ increases in n_{plan} we can pull the limit out of the probability statement, thus, $\mathbb{P}(\lim N_{n_{\text{plan}}} \leq n_{\text{plan}}) = \lim \mathbb{P}(N_{n_{\text{plan}}} \leq n_{\text{plan}})$. Applying Eq. (5) with $N_{n_{\text{plan}}}$ for each n_{plan} with $n_{\text{plan}} \to \infty$ in the limit statement yields

$$\mathbb{P}(N_\alpha < \infty) = \lim \mathbb{P}(N_{n_{\text{plan}}} \leq n_{\text{plan}}) = \lim \mathbb{P}(E_{N_{n_{\text{plan}}}} \geq 1/\alpha) \leq \lim \alpha = \alpha. \tag{C4}$$

Replacing the event $N_\alpha < \infty$ by the event V yields Ville's inequality.

Appendix D
Anytime-valid confidence sequences in the t-test setting

The notable difference between the p-value test, e.g. Eq. (A1) in Appendix A, and a savi test, Eq. (6), is the placement of the sample size within the probability statement, which therefore also needs negating. Note first hat 1 minus Ville's inequality yields

$$1 - \mathbb{P}_{\varphi_0}(V) = \mathbb{P}_{\varphi_0}(V^c) \geq 1 - \alpha, \tag{D1}$$

where V is given in Eq. (C2) and where V^c is its complement given by

$$V^c = \{\text{For all } n : E_n < 1/\alpha\}. \tag{D2}$$

We now invert $E_n < 1/\alpha$. For t-tests, E_n can be rewritten as a function that takes as input the square of the T-statistic centred at φ_0, i.e. Eq. (12), and outputs a non-negative number. We let E_n^{-1} be the pre-image that takes in a non-negative number and outputs a t^2. As such,

$$E_n(T) < 1/\alpha \Leftrightarrow \bar{X}_1 - \bar{X}_2 - \frac{S_p}{\sqrt{n_{\text{eff}}}}\sqrt{E_n^{-1}(1/\alpha)} < \varphi_0 < \bar{X}_1 - \bar{X}_2 + \frac{S_p}{\sqrt{n_{\text{eff}}}}\sqrt{E_n^{-1}(1/\alpha)} \tag{D3}$$

Hence, for any data-governing φ_0

$$\mathbb{P}_{\varphi_0}\Bigg(\text{For all } n : \Big[\bar{X}_1 - \bar{X}_2 \pm \frac{S_p}{\sqrt{n_{\text{eff}}}}\sqrt{E_n^{-1}(1/\alpha)}\Big] \ni \varphi_0\Bigg) \geq 1 - \alpha, \tag{D4}$$

where $A \ni \varphi_0$ means that the (random) set A covers the element φ_0.

Appendix E
Likelihood ratios are E-processes

Let $\mathcal{M}_0$ be a simple null model with distribution $\mathbb{P}$ and $\mathbb{Q}$ be any (other) distribution with densities p and q, respectively. We consider the likelihood ratio $\text{LR}_n := \frac{q(x^{(n)})}{p(x^{(n)})}$, which is non-negative, since both its numerator and denominator are non-negative.

To see that it fulfils the definition of an E-process, it suffices to verify Property (iii) Eq. (4), for which we use the fact that (a) the null model is simple, (b) the law of total probability, and (c) that stopping at time $N = m$ corresponds to some (measurable) set B_m of outcomes $x^{(m)} := (x_1, \ldots, x_m)$. As an example, we consider the z-test with variance known to be one, thus, $z := \sqrt{n}\bar{x}_n$, where $\bar{x}_n$ is the sample mean of n data points. As an example stopping time we take the first time N at which $p < 0.05$. The event B_m corresponding to $N = m$ is then equivalent to data sequences $x^{(m)} = (x_1, \ldots, x_m)$ with sample means $|\bar{x}_j| < \frac{1.96}{\sqrt{j}}$ for $j = 1, \ldots, m - 1$, otherwise m would not be the first time to stop, and $|\bar{x}_m| > \frac{1.96}{\sqrt{m}}$, otherwise we would not have stopped at time m. Regardless of the stopping time N, we can always identify the event $N = m$ to some set B_m of outcomes $x^{(m)}$.

Before applying the law of total probability, we note that

$$\text{LR}_N := \sum_{m=1}^{\infty} \mathbf{1}\{N = m\}\, \text{LR}_m, \tag{E1}$$

where $\mathbf{1}\{N = m\} = 1$ if the stopping time realises m, and zero otherwise. In other words, LR_N equals LR_m whenever $N = m$ as one would expect. The law of total probability allows us to compute the expectation $\mathbb{E}_{\text{LR}_N \sim \mathbb{P}}[\text{LR}_N]$ in two steps: The integral over the outcomes of $X^{(m)}$ conditioned on the event $\{N = m\}$ and a (green) expectation $\mathbb{E}_{\mathbb{P}_0}$ pertaining to the outcomes of N. Swapping the expectation and summation operators yields

$$\mathbb{E}_{\text{LR}_N \sim \mathbb{P}}[\text{LR}_N] = \mathbb{E}_{N \sim \mathbb{P}}\Big[\mathbb{E}_{\text{LR}_m \sim \mathbb{P} \mid \{N=m\}}\Big[\sum_{m=1}^{\infty} \mathbf{1}\{N = m\}\, \text{LR}_m\Big]\Big] \tag{E2}$$

$$= \mathbb{E}_{N \sim \mathbb{P}}\Big[\sum_{m=1}^{\infty} \mathbf{1}\{N = m\}\, \mathbb{E}_{\text{LR}_m \sim \mathbb{P} \mid \{N=m\}}[\text{LR}_m]\Big] \tag{E3}$$

$$= \sum_{m=1}^{\infty} \mathbb{E}_{N \sim \mathbb{P}}\Big[\mathbf{1}\{N = m\}\, \mathbb{E}_{\text{LR}_m \sim \mathbb{P} \mid \{N=m\}}[\text{LR}_m]\Big]. \tag{E4}$$

For Eq. (E3) we took $\mathbf{1}\{N = m\}$ out of the inner integral, because it is a function of the conditioning event $\{N = m\}$. For Eq. (E4) we again use the fact that we can swap sums and expectations of non-negative functions.

As before, the expectation $\mathbb{E}_{\text{LR}_m \sim \mathbb{P} \mid \{N=m\}}[\text{LR}_m]$ defines an integral, but this time with respect to the density $p(x^{(m)} \mid B_m)$. The condition on $N = m$ also restricts the outcomes of length m to the event B_m. Continuing from Eq. (E4), and exploiting the fact that LR_m is a ratio of densities that can be factorised via conditioning, we have that

$$\mathbb{E}_{\mathbb{P}}[\mathrm{LR}_N] = \sum_{m=1}^{\infty} \mathbb{E}_{N\sim\mathbb{P}}\Big[\mathbf{1}\left\{B_m\right\} \mathbb{E}_{\mathrm{LR}_m\sim\mathbb{P}\,|\,B_m}[\mathrm{LR}_m]\Big] \tag{E5}$$

$$= \sum_{m=1}^{\infty} \mathbb{E}_{N\sim\mathbb{P}}\Big[\mathbf{1}\left\{B_m\right\} \int_{B_m} \tfrac{q(x^{(m)}\,|\,B_m)}{p(x^{(m)}\,|\,B_m)} \tfrac{\mathbb{Q}(B_m)}{\mathbb{P}(B_m)} p(x^{(m)}\,|\,B_m)\mathrm{d}x^{(m)}\Big] \tag{E6}$$

$$= \sum_{m=1}^{\infty} \mathbb{E}_{N\sim\mathbb{P}}\Big[\mathbf{1}\left\{B_m\right\} \tfrac{\mathbb{Q}(B_m)}{\mathbb{P}(B_m)} \int_{B_m} q(x^{(m)}\,|\,B_m)\mathrm{d}x^{(m)}\Big], \tag{E7}$$

where again a function of the conditioning event can be taken out of the expectation. Using the fact that B_m corresponds exactly to the outcomes of X^m given B_m we can conclude that the inner integral equals one, that is,

$$\mathbb{E}_{\mathbb{P}}[\mathrm{LR}_N] = \sum_{m=1}^{\infty} \mathbb{E}_{N\sim\mathbb{P}}\Big[\mathbf{1}\left\{B_m\right\} \tfrac{\mathbb{Q}(B_m)}{\mathbb{P}(B_m)}\Big] = \sum_{m=1}^{\infty} \mathbb{E}_{N\sim\mathbb{P}}\Big[\mathbf{1}\left\{N=m\right\} \tfrac{\mathbb{Q}(N=m)}{\mathbb{P}(N=m)}\Big] \tag{E8}$$

$$= \sum_{m=1}^{\infty} \mathbb{Q}(N=m) = 1 \leq 1. \tag{E9}$$

The last equality follows from the fact that the probability of any N taking on any outcome is one, in particular, under the measure $\mathbb{Q}$ corresponding to the numerator of LR_n. Because N was taken arbitrarily and we only have to deal with a single $\mathbb{P}$, we verified that Eq. (4) indeed holds for simple likelihood ratios.

As an aside, the $p < 0.05$ rule corresponds to the fixed threshold on the z-scale, that is, $|z| > 1.96$. This is differs substantially from stopping as soon as the likelihood ratio exceeds$1/\alpha$, as illustrated in Fig. E1. For those interested in the technicalities, the defining Property (iii) of E-processes holds generally for any non-negative supermartingales that starts at 1, and it is known as Doob's optional stopping theorem. In other words, non-negative supermartingales are E-processes. However, not every E-processes is a non-negative supermartingale, see Ruf et al. (2023) for the details.

Appendix F

The `mom` t-test E-process based on the Gaussian non-local moment prior

In Pérez-Ortiz et al. (2024) it was shown that for any savi test defining parameter δ_s the T-likelihood ratio Eq. (7) is an E-process. As such, for any prior $\pi(\delta_s)$ the following mixture also defines an E-process

$$E_{n_{\mathrm{eff}},\nu}(t) := \int \tfrac{T_\nu(t\,|\,\sqrt{n_{\mathrm{eff}}}\delta_s)}{T_\nu(t)} \pi(\delta_s)\mathrm{d}\delta_s. \tag{F1}$$

The computations of Gronau et al. (2020, Appendix A) and Ly, Marsman, and Wagenmakers (2018) show that the use of a symmetric $\pi(\delta_s)$ simplifies the computations to

$$E_{n_{\mathrm{eff}},\nu}(t) = \int e^{-\frac{n_{\mathrm{eff}}\delta_s^2}{2}} \, {}_1\mathrm{F}_1\left(\tfrac{\nu+1}{2}\,;\,\tfrac{1}{2}\,;\,\tfrac{t^2}{\nu+t^2}\tfrac{n_{\mathrm{eff}}\delta_s^2}{2}\right)\pi(\delta_s)\mathrm{d}\delta_s, \tag{F2}$$

where for $|z| < 1$ the confluent hypergeometric function is given by ${}_1\mathrm{F}_1\left(a\,;c\,;z\right) := \sum_{i=0}^{\infty}\frac{(a)_i}{(c)_i}\frac{z^i}{i!}$, where $(a)_i = \Gamma(a+i)/\Gamma(a)$ is known the Pochhammer symbol for a rising factorial, and where $\Gamma(a)$ is the gamma function at a, and $\Gamma(n+1) = n!$ for $n \in \mathbb{N}$.

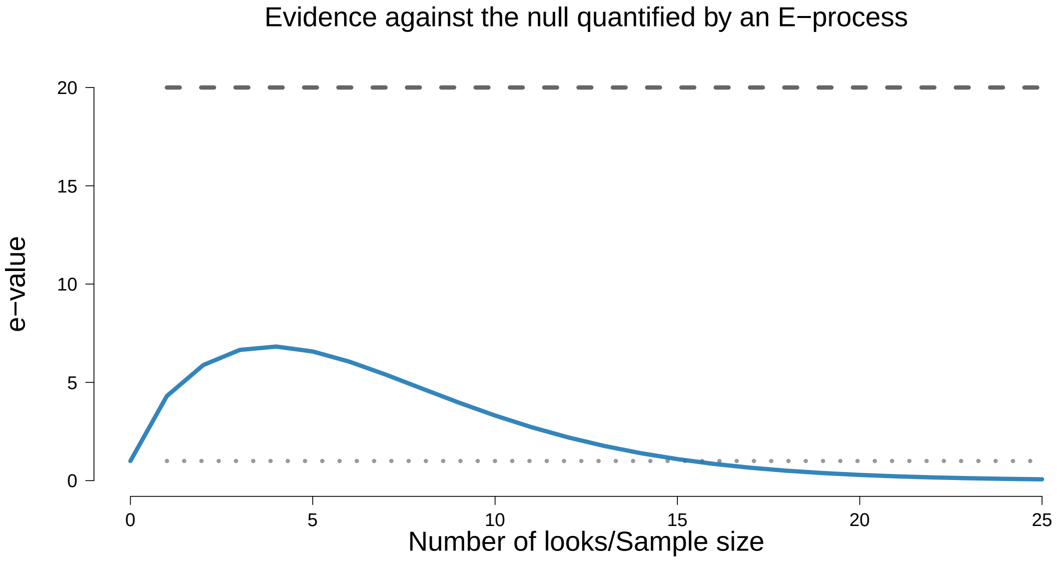


Figure E1

Independent on the sample size n an observed $z = 1.96$ corresponds to $p < \alpha = 0.05$, thus, a null rejection in the classical framework, whereas the evidence for the alternative as quantified by the z-likelihood ratio, thus, e-value would first increase and then decrease as a function of n.

We can express $E_{n_{\text{eff}},\nu}(t)$ analytically if we use a Gaussian kth moment prior density introduced by Johnson and Rossell (2010) for $\pi(\delta_s)$, which is given by

$$\pi(\delta) = \frac{\delta^{2k}\exp(-\frac{\delta^2}{2g})}{p_k(\varnothing)}, \text{ where } p_k(\varnothing) := \int_{\infty}^{\infty} \delta^{2k}\exp(-\tfrac{\delta^2}{2g}) = (2g)^{k+\frac{1}{2}}\Gamma(k+\tfrac{1}{2}), \qquad \text{(F3)}$$

is the normalisation constant of the prior. The computations follow from Gradshteyn and Ryzhik (2007, p. 822) yielding

$$E_n^{\texttt{mom}}(t) = (1 + n_{\text{eff}}g)^{-k-\frac{1}{2}}\,{}_2\mathrm{F}_1\left(\tfrac{\nu+1}{2}, k+\tfrac{1}{2}\,;\,\tfrac{1}{2}\,;\,\tfrac{t^2}{\nu+t^2}\tfrac{n_{\text{eff}}g}{1+n_{\text{eff}}g}\right), \qquad \text{(F4)}$$

where for $|z| < 1$ the Gaussian hypergeometric function is given by ${}_2\mathrm{F}_1(a, b\,;\,c\,;\,z) := \sum_{i=0}^{\infty}\frac{(a)_i(b)_i}{(c)_i}\frac{z^i}{i!}$. An Euler transform and the plugin $k = 1$ then yields Eq. (13).

Appendix G

A computationally convenient (two-sided) two-sample Bayes factor t-test

A Bayes factor is by definition a ratio of marginal likelihoods, where the marginalisation is with respect to a pair of priors, one for each model (e.g. Jeffreys, 1961, Ly et al., 2016a, 2016b).

In the two-sample t-test setting the numerator has a likelihood function that depends on, say, μ_1, μ_2, σ, whereas the denominator has a likelihood function that only depends on two parameters, say, the global mean μ_g and σ. The priors $\pi_1(\mu_1, \mu_2, \sigma)$ and $\pi_0(\mu_g, \sigma)$ used to construct Bayes factor in favour of the alternative over the null for data $x^{(n)}$ is then

$$\mathrm{BF}_{10}(x^{(n)}) := \frac{\int\int\int f(x^{(n)} \,|\, \mu_1, \mu_2, \sigma)\pi_1(\mu_1, \mu_2, \sigma)\mathrm{d}\mu_1\mathrm{d}\mu_2\mathrm{d}\sigma}{\int\int f(x^{(n)} \,|\, \mu_g, \mu_g, \sigma)\pi_0(\mu_g, \sigma)\mathrm{d}\mu_g\mathrm{d}\sigma}. \tag{G1}$$

A computational convenient choice is to exploit conjugacy, which states that normal likelihoods $\mathcal{N}(x^{(n)} \,|\, \mu, \sigma)$ for data $x^{(n)}$ combined with normal priors lead to normal posteriors. For the two-sample t-test, we have under the alternative the likelihood function

$$\begin{aligned} f(x^{(n)} \,|\, \mu_1, \mu_2, \sigma) &= \mathcal{N}(x^{(n_1)} \,|\, \mu_1, \sigma)\mathcal{N}(x^{(n_2)} \,|\, \mu_2, \sigma) && \text{(G2)} \\ &= (2\pi)^{-\frac{n_+}{2}}\sigma^{-n_+}e^{-\frac{\sum_{k=1}^2 \nu_k s_k^2}{2\sigma^2}} \exp\Big(-\tfrac{1}{2\sigma^2}[n_1(\mu_1 - \bar{x}_1)^2 + n_2(\mu_2 - \bar{x}_2)^2]\Big), && \text{(G3)} \end{aligned}$$

where $n_+ := n_1 + n_2$ is the total sample size, and $\bar{x}_k, s_k^2, \nu_k = n_k - 1$ the sample mean, the sample variance and the degrees of freedom for group $k = 1, 2$, respectively. The likelihood under the null is $f(x^{(n)} \,|\, \mu_1, \mu_2, \sigma)$ with the means set to the global mean, i.e. $\mu_1 = \mu_2 = \mu_g$.

For the computationally convenient Bayes factor we use a (conditional) normal prior on μ_1, μ_2, μ_g, that is, $\mu_k \,|\, \sigma \sim \mathcal{N}(a_k, g_k\sigma^2)$ for $k = 1, 2$ in the alternative model and $\mu_g \,|\, \sigma \sim \mathcal{N}(a_0, g_0\sigma^2)$ in the null model. Combined with a conjugate inverse root gamma prior on σ, i.e. $\sigma \sim \mathrm{Gamma}^{-\frac{1}{2}}(\alpha_\sigma, \beta_\sigma)$, we have 8 parameters to play with. The yields

$$\mathrm{BF}_{10\,;\,\eta} := \sqrt{\frac{1+n_+g_0}{(1+n_1g_1)(1+n_2g_2)}} \left(\frac{\frac{n_1n_2}{n_+}(\bar{x}_1-\bar{x}_2)^2 + \frac{n_+}{1+g_0n_+}(\frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_+}-a_0)^2 + 2\beta_\sigma + \sum_{k=1}^2 \nu_k s_k^2}{\frac{n_1(\bar{x}_1-a_1)^2}{1+n_1g_1} + \frac{n_2(\bar{x}_2-a_2)^2}{1+n_2g_2} + 2\beta_\sigma + \sum_{k=1}^2 \nu_k s_k^2} \right)^{\frac{n_+}{2}+\alpha_\sigma}, \tag{G4}$$

where $\eta = (a_1, g_1, a_2, g_2, a_0, g_0, \alpha_\sigma, \beta_\sigma)$ collects the tuning parameters. For the counter example below, we choose $a_1 = 3.98, g_1 = 0.03, a_2 = 4.02, g_2 = 0.05, a_0 = 4, g_0 = 2$ and $\alpha_\sigma = 2$ and $\beta_\sigma = 1/2$. An interpretation for this choice is as follows: If there is a difference in the population mean then it is relatively small with population means at 3.98 and 4.02 and this knowledge is quite concentrated. On the other hand, if the null holds true, then the shared mean is 4 and the prior is relatively spread out. This conjugate Bayes factor is implemented as the function `conjugateBfTStat()` in the `safestats` package.

Appendix H
Not all Bayes factors are E-processes

Frequentist principles such as power and type I error control are the main motivation for this work. It is therefore surprising that ensuring these frequentist principles over time led us to E-processes that were previously derived as Bayes factors, as remarked in Section 1.2.5. The distinction in their definitions is striking: An E-process is characterised by constraints on the expectation under every data-generating $\mathbb{P}$ and all possible stopping time (see Eq. (4)), whereas a Bayes factor is defined as a ratio of marginal likelihoods, which integrates out the parameters in the numerator and the denominator (e.g. Eq. (G1)). Grünwald et al. (2024) provides insights to when the two concepts align. Roughly speaking, optimal E-variables have a Bayes factor representation, because they are derived from projecting a (weighted averaged) alternative model onto the so-called convex hull of the null model. By convex we mean a weighted average of $\mathbb{P} \in \mathcal{M}_0$, thus, a marginal likelihood. In this view, the prior is used as means to solve a minimisation problem and does not represent prior belief.

It is, however, not true that every E-process is a Bayes factor, nor is every Bayes factor an E-process. Firstly, E-processes derived with the universal inference construction are based on plugins, which cannot be written as a ratio of marginal likelihoods (e.g. Pandeva, Bakker, Naesseth, & Forré, 2024, and Ramdas et al., 2023, but also Turner et al., 2024). Secondly, the computationally convenient Bayes factor $\mathrm{BF}_{10\,;\,\eta}$, i.e. Eq. (G4), is an example of a Bayes factor that is not an E-variable. Depending on the values of these 8 tuning parameters, perhaps due to how they were estimated using past data, the procedure that stops as soon as the $\mathrm{BF}_{10\,;\,\eta} \geq 1/\alpha = 20$ can be unreliable; see the dashed brown FPR curve in the left panel of Fig. H1. Hence, $\mathrm{BF}_{10\,;\,\eta}$ cannot be an E-process, because for every E-process Ville's inequality must hold. This reflects a point made earlier by De Heide and Grünwald (2021) that there can be issues with Bayes factors under optional stopping. We introduced the example to show that type I error control due to Ville's inequality does not automatically hold for all Bayes factors in general. The constructed Bayes factor in Fig. H2 is not an E-process because it violates Property (iii) (Eq. (4)) by having conditional expectations larger than one for at least one data generating distribution from the null model, when $\mu_1 = \mu_2 = 4$ and $\sigma = 2$.

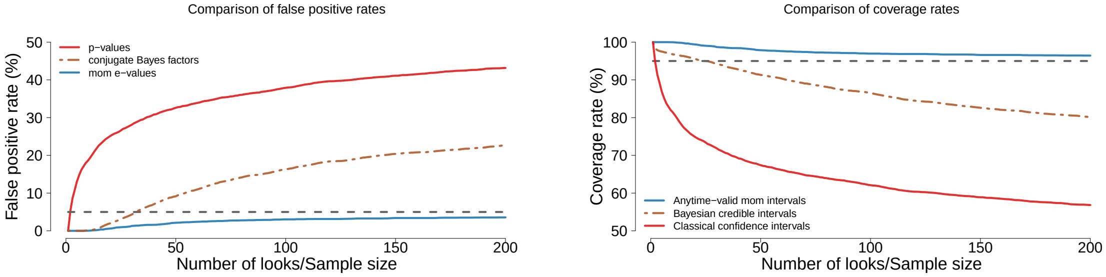


Figure H1

Left panel: The FPR of monitoring the $\mathrm{BF}_{10\,;\,\eta} \geq 1/\alpha$ *test (dashed brown) does not remain below* $\alpha = 0.05$. *It contradicts Ville's inequality and therefore cannot be an* E*-process. Right panel: The coverage rate of the (Bayesian) credible interval (dashed brown) dips below the nominal 95%-level, if it is used sequentially, and actually mimics the behaviour of a frequentist confidence interval, see Fig. H2 below.*

The right panel of Fig. H1 shows that the coverage rate of the associated 95% credible interval will also drop well below the nominal level of 95%. Fig. H2 shows the typical evolution of the chosen computationally convenient Bayes factor BF_{10} for the same data (generated under the null) that were used for Fig. 1. The right panel of Fig. H2 shows three types of intervals in one plot. The (Bayesian) credible interval (yellow) quickly intersects with the classical confidence interval (red), and both unfortunately do not cover the true mean difference, here, $\varphi = 0$ at all times, whereas the anytime-valid confidence interval (blue) does. Hence, as with 95% confidence intervals, we cannot guarantee that the 95% (Bayesian) credible interval covers the true underlying parameter with 95% probability during data collection or at a possibly data-driven stopped time. This is unfortunately a fundamental problem; Some Bayes factors are E-variables, but no credible interval forms a sequence of anytime-valid confidence intervals. The problem cannot be solved by choosing different priors, as typical priors yield credible intervals that (relatively quickly) converge

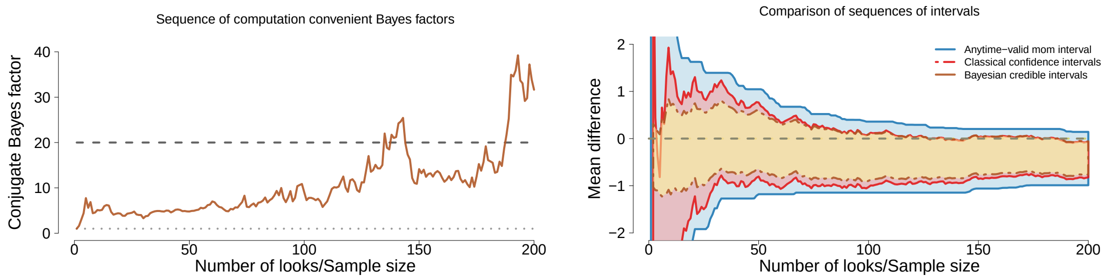


Figure H2

Left panel: The (non-default) computationally convenient Bayes factor (Appendix G) tends to overstate the evidence against the null. Right panel: The (Bayesian) credible interval (yellow) and the classical confidence interval (red) both do not cover the true data generating mean difference of zero at all times, unlike the anytime-valid confidence interval (blue), which is also a wee bit wider.

to, and thus behave as, classical confidence intervals (Ghosal & van der Vaart, 2017).

An anytime-valid confidence interval based on e-values avoids being turned into a classical confidence interval by not updating a prior to a posterior as is the case for credible intervals, but by inverting the savi test. This is, thus, a completely different procedure yielding wider intervals as shown in Fig. H2. We feel that the additional width is a relatively low price to pay for reliability and convenience, as it is guaranteed that the true underlying parameter value is covered with 95% probability regardless of when or even if data collection has stopped. The resulting anytime-valid confidence interval in some cases still has some special, though, non-standard Bayesian interpretation (Pawel, Ly, & Wagenmakers, 2024).

Appendix I

Anytime-valid tests are more flexible compared to sequential Bayes factor tests

In the t-test setting certain Bayes factors, such as the one discussed in Section 1.2.5, are also E-variables. The way they are used in forming a sequential Bayes factor test, however, differs from that of a savi test based on Protocol 1.1. Sequential Bayes factor tests (e.g. Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017, Schnuerch et al., 2022, Pramanik & Johnson, 2022) follow Protocol 6.1. This procedure does not come with an n_{plan} at which the detection of an effect $|\delta| \geq \delta_{\text{min}}$ with at least $1 - \beta$ power is guaranteed. Crucially, it also has an additional evidence boundary for accepting the null. The latter is natural due to a Bayes factor less than 1 being interpreted as evidence for the null.

The current discussion in sequential Bayes factor testing revolves around (1) the selection of the boundaries A and B, thus, *a specific stopping rule*, and (2) which Bayes factor to use. Type I and II errors are approximated with simulations that depends on the specific choices, most importantly, the chosen stopping rule. For E-processes the stopping time is irrelevant, which makes its use easier and more straightforward.

To address the boundary question, Schönbrodt et al. (2017) suggested employing the `eCauchy` E-process/Bayes factor t-test, constructed with a Cauchy prior on δ_s, featuring a prior width of 1, and decision boundaries set to $A = 6$ and $B = 1/6$ for early lines of research. This general recommendation was based on a large-scale simulation study with

```
# Pseudo code: This code does NOT run

n ← 1
bf10 ← 1

while (B < bf10 && bf10 < A) {
  bf10 ← computeBayesFactor(x[1:n], y[1:n], somePriors)

  if (bf10 >= A) {
    "Reject the null and accept the alternative"
    stop()
  } else if (bf10 <= B) {
    "Reject the alternative and accept the null"
    stop()
  } else {
    "Increase sample size and test again
          at the start of the while loop"
    n ← n + 1
  }
}
```

Protocol 6.1: The sequential Bayes factor testing procedure in pseudo R code without a maximum sample size.

data-governing $\delta = 0.5$, for which this particular Bayes factor and boundaries resulted in realised type I and II error rates of 4.7% of 4.6%, respectively .

Inspired by Wald's sequential probability ratio test, Schnuerch et al. (2022) argue for the boundaries $A = \frac{1-\beta}{\alpha}$ and $B = \frac{\beta}{1-\alpha}$ instead. These boundaries were also explored by Pramanik and Johnson (2022) who recommend using the non-local moment prior, that is, the mom E-process with the two bumps at ±0.3 as default. Their large-scale simulation studies show that the resulting sequential Bayes factor test yields average sample sizes close to those of the two-point prior variant at ±0.3. In the large-scale simulation studies the realised type I and type II error were observed to be below the tolerable α and β respectively. This can partially be explained by the employed Bayes factors also being E-processes, at least as far as the type I error is concerned.

The Wald boundaries on their own, however, do not actually guarantee error control. For instance, running Protocol 6.1 with boundaries $A = \frac{1-\beta}{\alpha} = 16$, $B = \frac{\beta}{1-\alpha} = 0.21$ and the computationally convenient Bayes factor of Eq. (G4) leads to 16.8% correct null acceptance, but 24.9% false positive rejections, and no conclusion was reached for the remaining 58.2%. The latter inconclusive category can lead to this procedure requiring more data points than the savi testing procedure.

Stopping rule dependent error control make the procedure less flexible. For instance, during data acquisition a hopeful investigator changes his mind and continues sampling despite observing $\text{BF}_{10} \leq 0.21$ by interpreting $\text{BF}_{10} = 0.21$ as only "moderate evidence" for the null (e.g. Jeffreys, 1961, Appendix B, Lee & Wagenmakers, 2013). Continuing sampling after hitting the lower boundary makes the sequential Bayes factor Protocol 6.1 equivalent to the savi testing Protocol 1.1 that has the potential to run indefinitely due to the absence of an n_{plan}. By Ville's inequality we know that halting as soon as $\text{BF}_{10} \geq 16$ only guarantees a type I error of 6.25% under the null, if the Bayes factor is also an E-process. If it is not, we cannot even say this, as is the case for the computationally convenient Bayes factor. Assuming the Bayes factor is an E-process, a type I error guarantee of level $\alpha = 0.05$ therefore forces the investigator to stop as soon as $\text{BF}_{10} \leq 0.21$ so that the unaccounted probability of 1.25% can be absorbed by the lower boundary. E-processes are robust to adaptation of the stopping rule, as it still defines a stopping time for which Ville's inequality holds. This makes e-value based inference more flexible.

The additional stopping rule for the null also complicates the design of sequential Bayes factor tests. To the best of our knowledge, all sequential Bayes factor tests require extensive simulations to provide an indication of both the realised type I and type II error, and they will all have to be re-performed when the stopping rule or the tuning parameter of the Bayes factor under consideration is changed. In contrast, type I error control for E-processes is mathematically guaranteed by Ville's inequality, and type II error control and estimates of the run times of experiments can be obtained with much cheaper simulations. Specifically, e-value sampling designs provide an indication of how long the experiment should run for in terms of n_{plan}, and monitoring the test until n_{plan} guarantees a $1 - \beta$ probability to detect effects $|\delta| \geq \delta_{\text{min}}$. This n_{plan}, in turn, is found by relatively quick simulations for typical values of α, β and δ_{min}.