

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

Alexander Ly¹, Udo Boehm¹, Peter Grünwald^{1,2}, Aaditya Ramdas³,
and Don van Ravenzwaaij⁴

¹Machine Learning Research Group, Centrum Wiskunde & Informatica

²Mathematical Institute, Leiden University

³Department of Statistics and Data Science, Carnegie Mellon University

⁴Department Psychometrics and Statistics, University of Groningen

Abstract

We show how e -values simplify the design and the conduct of experiments. These e -values yield anytime-valid tests and confidence intervals that preserve type I error guarantee *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 making a false discovery. Early stopping not only preserves resources, but also mitigates risk for participants in clinical settings. Anytime-valid tests always allow for optional continuation, that is, the extension of an experiment regardless of the motivation. For instance, if more funds become available, or 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% chance, cover the true effect size regardless of how, or even if, data collection is stopped. We use the free and open-source software library **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 undertaking: It necessitates reducing the risk of reporting, publication, and hindsight biases by explicitly formulating the theories and hypotheses that we test 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. Incorporating a sampling regimen in a pre-registration document serves to prevent the commonly used classical p -value test from becoming unreliable and invalid, a situation that arises when such a test is conducted during data acquisition. Based on this fact, Simmons et al. (2011, p. 4) argued that reliable science requires authors to explicate and abide by their sampling rule. They also proposed that reviewers should act


as gatekeepers to enforce adherence to the sampling plan. The suggested solution ultimately comes down to removing flexibility by confining researchers to a rigid sampling regimen in service of an old-fashioned statistical tool of inference, which was developed in the 1930s. However, despite all good intentions to sustain a sampling plan, researchers might face challenges such as difficulties in study recruitment, faster-than-expected depletion of funds, or even the impact of a global pandemic. Experiencing these setbacks and subsequently being informed that the statistical inferences are invalid is akin to having salt rubbed into one's wounds.


Rather than forcing researchers to become slaves to rigid classical statistical tools that require strict adherence to a sampling regimen, we propose the use of recently developed, maximally flexible, anytime-valid tests and confidence intervals based on E -processes, which yield realisations referred to as e -values. The main advantage of E -process-based analysis methods is that their validity is independent of any sampling plan; they can be conducted at any moment in time, regardless of planned, current, or future sample sizes. Being free of a sampling plan implies that the realised e -values can help researchers safely make informed decisions about whether to stop or continue an investigation during data acquisition. Crucially, the flexibility in data collection afforded by e -values *does not* permit just any working hypothesis to be falsely presented as significant, in contrast to p -values, as was so convincingly illustrated by Simmons et al. (2011). As such, this e -value-based inference framework relieves authors of the strenuous effort required to adhere to a sampling plan. It also lifts the burden on reviewers to evaluate whether the authors upheld their sampling plan — a reviewer who considers the evidence to be too weak (i.e. the reported e -value to be too small) may even ask authors to gather some additional data and compute an updated e -value. Whenever there truly is no effect, additional data cannot intolerably inflate the chance of a false positive finding.


To clarify our stance, we do not object to the principles of pre-registration and planning. Rather, our concern is with the traditional methods of inference that restrict us from


Draft version 1.3 (14-02-2025). This paper has not been peer reviewed.

Alexander Ly  <https://orcid.org/0000-0003-3925-3833>

Udo Boehm  <https://orcid.org/0000-0002-8677-0721>

Peter Grünwald  <https://orcid.org/0000-0001-9832-9936>

Aaditya Ramdas  <https://orcid.org/0000-0003-0497-311X>

Don van Ravenzwaaij  <https://orcid.org/0000-0002-5030-4091>

Udo Boehm is now at Department Methodology and Statistics, Tilburg University.

The programming code used in the present work is available at <https://osf.io/mdbqe/>. The authors declare no competing interests. This research was partly financed by 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).

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.

Correspondence concerning this article should be addressed to Alexander Ly, Centrum Wiskunde & Informatica, Machine Learning, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands. E-mail: a.ly@cwi.nl

adapting to new circumstances. Our aim is to show that a viable alternative to classical p -value-based methods exists, and making it accessible to practitioners. Notably, our primary contribution consists of insightful yet *non-binding* and maximally flexible sampling designs, which provide explicit guidance on the efficient use of resources for designing and executing experiments. In a nutshell (all jargon will be explained further below), our designs involve a planned sample size 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 add data points at will.

These designs are elaborated on in Section 3. Before doing so, we first illustrate the use of e -values applied to two real-world data sets in Section 1. Section 2 examines the nature of anytime-validity by analysing the behaviour of E -processes under the null (when the effect is truly absent), while Section 3 focuses on their behaviour under the alternative (when the effect is truly present). Section 4 is concerned with the more nuanced differences between e -value based and Bayesian inference, which can be safely skipped upon first reading. Practical guidelines are given in Section 5, which we use to revisit the two real-world examples in Section 6. We conclude with a brief discussion in Section 7.

This tutorial on e -values is intended for practitioners who wish to make reliable inferences with minimal statistical constraints. It prioritises readability by an applied audience over rigorous mathematical presentation, a facet adeptly handled by, for instance, Grünwald et al. (2024), Howard et al. (2021), Ramdas et al. (2023), Ramdas and Wang (2024), and Shafer (2021).

1 Two Real-World Examples Illustrating e -Value Based Inference in Action

Before elaborating on the design of an e -value based experiment, we first illustrate its application by examining two replication attempts from the Many Labs 2 Project (Klein et al., 2018). This project investigated variations across samples and settings in the replicability of 28 classic psychological findings.

1.1 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’s (2009) postulated that age plays a key role in perceived morality. More precisely, they hypothesise that children are viewed (1) as being less responsible for their actions, (2) as having less intention of doing right or wrong, and (3) they are more likely to be perceived as receivers of moral actions compared to adults.

To study this hypothesis, Gray and Wegner (2009) had 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 ($\bar{x}_2 = 3.86, s_2 = 1.64$). The two observed sample means \bar{x}_1 and \bar{x}_2 are assumed to be imperfect

InfoBox 1: Basic terminology, notation and the statistical model underlying the t -test

A hypothesis makes a claim about (unobserved) population *parameters of interest*. For instance, the null hypothesis of no effect $\mathcal{H}_0 : \mu_1 = \mu_2$ postulates that the population mean difference parameter $\varphi := \mu_1 - \mu_2$ equals zero. Note that the null hypothesis specifies one constraint on two populations means, which leaves one parameter free. This is even clearer when we write $\mu_1 := \mu_g + \varphi/2$ and $\mu_2 := \mu_g - \varphi/2$, where μ_g is a so-called grand mean. The parameter μ_g is test irrelevant, thus, a nuisance parameter, because its value does not affect the parameter of interest φ .

Instead of limiting ourselves to the postulate that φ equals zero, we can as easily test null hypotheses where the mean difference parameter is claimed to be -3.142 , or 2.718 , or any other constant φ_0 . The T -statistic is a function of the data that measures the discrepancy between the claimed value φ_0 and the (observed) sample mean difference scaled by the standard error $S_p/\sqrt{n_\delta}$, that is,

$$T := \sqrt{n_\delta} \frac{\bar{X}_1 - \bar{X}_2 - \varphi_0}{S_p}, \quad (1)$$

where $\bar{X}_k := \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ki}$ is the sample mean of population $k = 1, 2$, $n_\delta = \frac{n_1 n_2}{n_1 + n_2}$ the so-called effective sample size with n_k the size of the sample from population $k = 1, 2$, and

$$S_p := \sqrt{\frac{1}{\nu} \left(\sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2 + \sum_{j=1}^{n_2} (X_{2j} - \bar{X}_2)^2 \right)}, \quad (2)$$

is the pooled standard deviation, where $\nu = n_1 + n_2 - 2$ is the degree of freedom.

If the null hypothesis holds true, then the discrepancy as measured by the T -statistic is expected to be small. The null hypothesis does not exclude the possibility of ever seeing large values of T , it only implies that such an event occurs with low chance. What quantifies large values of T occurring with small chance is quantified by a model that links the population parameters to data. Such a model \mathcal{M} specifies a *collection of data-generating distributions* for the problem of interest (e.g. Bickel & Doksum, 2015; Ly et al., 2017). For typical t -test scenarios, the model consists of normal distributions. In context of Section 1.1, the ratings in the adult offending and baby offending condition are assumed to be independently drawn from the normal distributions $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$ respectively, where σ^2 is an unknown population variance, and also a nuisance parameter. The null model denoted by \mathcal{M}_0 consists of the aforementioned normal distributions with mean difference parameter restricted to the null value φ_0 leaving the nuisance parameters μ_g and σ free to vary. \diamond

reflections of two underlying, but unobserved, population means μ_1 and μ_2 , respectively. The null hypothesis postulates that the mean difference is zero, i.e. $\mathcal{H}_0 : \varphi := \mu_1 - \mu_2 = 0$, which implies that the observed mean difference $\bar{x}_1 - \bar{x}_2$ is a result of mere random noise. This null hypothesis was tested and rejected at the typical significance level of $\alpha = 0.05$ due to observing $t(68) = 3.32$ and $p = 0.001 < \alpha$. A more detailed description of the relevant terminology and notation can be found in Box 1.

The resulting null rejection was viewed as support for Gray and Wegner's (2009) (1) perceived responsibility part of the moral typecasting hypothesis. The authors of the original study also found that participants (2) perceived the adult offender to be more intentional compared to the baby, and that (3) the perceived pain of the adult is less than that of the baby. The main focus here, as in Many Labs 2, is on (1) perceived responsibility.

Out of 61 Many Labs 2 replication attempts, a total of 58 (95.08%) led to a signifi-

cant p -value. The pre-registration required each replication attempt to contain at least 80 participants, say, $n_{\text{plan},1} = n_{\text{plan},2} = 40$ in each group. To ensure that the classical p -value tests remain valid (see also the discussion surrounding Fig. 2 below), data collection across all attempts had to stop before the p -value tests were conducted.

By using e -values, we can reverse this process and monitor the test results before stopping an experiment — all while maintaining type I error control, as explained below. The e -value quantifies the evidence against the null hypothesis, ranging from zero (absolutely no evidence against the null) to one (neutral evidence) and to infinity (irrefutable evidence against the null). A null rejection at level α can be concluded as soon as the e -value crosses the threshold $1/\alpha$, e.g. $e \geq 20$ for $\alpha = 0.05$. Monitoring the test implies that we deal with a sequence of e -values, say, $e_1, e_2, \dots, e_n, \dots$. At each “time” n we can check whether the e -value at that moment, denoted by e_n , exceeds the evidence threshold $1/\alpha$. The left panel of Fig. 1 shows in blue the evolution of the e -values for the replication data acquired at Carleton University in Ottawa, Canada. The plot shows the progression of e -values as a

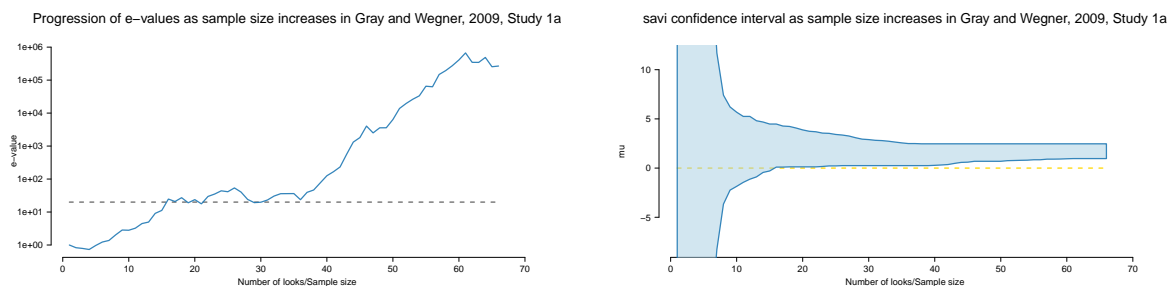


Figure 1

Left panel: The e -value sample path shows that at $n_1 = 16$, and $n_2 = 14$, there is already sufficient evidence to reject the null. Right panel: The 95% anytime-valid confidence intervals encapsulates the data-governing mean difference at all times with 95% chance. The null value of zero mean difference drops out of the intervals the first time the e -value passes the rejection threshold of $1/\alpha = 20$.

function of the sample sizes n_1 and n_2 of the adult offending and baby offending condition, respectively. For simplicity, the ratio between n_1 and n_2 is assumed to be fixed to the ratio of the final sample sizes $n_1 = 66$ and $n_2 = 55$. For instance, the first time the e -value passes the threshold of $1/\alpha = 20$ (indicated by the broken horizontal line) occurs at $n_1 = 16$ and $n_2 = 14$, at which point data collection can already be stopped. Doing so leads to testing 50 and 41 fewer participants from the adult offending and the baby offending condition, respectively. The right panel of Fig. 1 shows the corresponding anytime-valid confidence interval for the mean difference parameter $\varphi = \mu_1 - \mu_2$. The yellow broken horizontal line corresponds to the postulated $\mathcal{H}_0 : \varphi = 0$. The few lines of code needed for this analysis are shown in R Code 1.1.¹ Repeating the e -value test for all replication attempts yields 54 null rejections (88.52%), 4 fewer than the p -value analysis. These conclusions, however,

¹For those interested in coding 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.

```

1 designObj ← designSaviT(nPlan=c(40, 40), alpha=0.05,
2                         testType="twoSample")
3 result ← saviTTest(x, y, designObj=designObj, sequential=TRUE)
4 plot(result, xlim=c(1, 70), xlab="Number of looks/Sample size")
5 plot(result, xlim=c(1, 70), wantConfSeqPlot=TRUE)

```

R Code 1.1: R Code for an e -value t -test. Only a few lines of code suffices to run an e -value t -test. The design object on Line 1 formalises the Many Labs 2 protocol of acquiring at least 80 participants. The e -value test is performed on Line 3 for data vectors x and y . Line 4 plots the evolution of e -values, and Line 5 plots the corresponding confidence interval as a function of the sample size of the first group, see Fig. 1.

can be reached with 2543 and 2474 fewer participants (about 63.1% and 62.3% less) in the offending adult and offending baby condition, respectively, if an experiment is stopped as soon as $e \geq 1/\alpha = 20$ is observed.

1.2 Example 2: The Macbeth Effect — Moral Violations and Desire for Cleansing Zhong and Liljenquist, 2006, Study 2

Exactly the same E -process, that is, the same procedure that takes in data and outputs e -values, yields markedly different results when it is applied to the following example. 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, Zhong and Liljenquist (2006) asked 27 participants to copy 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 ($\bar{x}_2 = 3.75, s_2 = 1.32$). A p -value test was conducted to reject the null hypothesis $\mathcal{H}_0 : \varphi = 0$ that claims that the observed mean difference is due to mere random noise. The null hypothesis was rejected based on $t(25) = 2.64$ and $p = 0.01 < \alpha = 0.05$.

The test applied to the combined data from all replication attempts yielded $t(6954) = 0.096$ and $p = 0.9237$. Only 3 out of the 57 Many Labs 2 replication attempts (5.26%) yielded a significant p -value less than $\alpha = 0.05$.² Recall that if the null hypothesis indeed holds true (e.g. see Earp et al., 2014), then we expect the $p < 0.05$ test to falsely reject in about a 5% proportion of the number of replication attempts. Indeed, the results are fully consistent with the postulate that the Macbeth effect is absent. It is worth noting that these $p < 0.05$ tests were only computed once, at the final sample sizes. If the p -value test was performed as the data accumulate and the experiment stopped as soon as $p < 0.05$ was observed, then we would have ended up with 16 significant p -values out of the 57 replication attempts (28.07%, which is much larger than the tolerable 5%). On the other

²Unfortunately, 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).

hand, monitoring the e -value and rejecting the null as soon as $e_n \geq 1/\alpha = 20$ resulted in 1 null rejection out of 57 attempts (only 1.75% compared to 28.07%). If the null hypothesis holds true, then no matter how far we extend each trial, the chance of ever falsely rejecting the null based on $e \geq 20$ will forever remain below 5%.

Before we proceed, we offer a few remarks:

1. In general, there are various types of E -processes that output e -values. In the examples above, we employed one specific E -process to two sets of replication attempts. A different E -process might have led to faster inference in the first example and fewer false discoveries in the second example, assuming the Macbeth effect does not exist. To avoid e -hacking, one must fix the E -process in advance. R Code 1.1 does so by fixing the e -value type, and the tuning parameters within the type, see Section 3.4 for further details.

2. The specific E -process employed in the two examples is of type `mom` with explicit formulation given in Box 5. As a result of running Lines 1 and 2 of R Code 1.1 the tuning parameter g_{mom} , see Section 3.4 for details, is set to $g_{\text{mom}} = 0.134$. Unless specified otherwise, all simulations performed below, use the same `mom` E -process with $g_{\text{mom}} = 0.134$. The `safestats` package (version 0.8.8) uses the `mom` type E -process for t -tests by default. The other types included in the package are referred to as `eGauss`, `eCauchy`, and `grow`. Motivation for the `mom` and the other E -processes types will be given in Section 5.

3. The underlying tuning parameter of the E -process used in the two examples above is optimised to the minimum sample size of $n_{\text{plan},1} = n_{\text{plan},2} = 40$ within each condition. In Many Labs 1 it was mentioned that this choice was a trade-off between lowering the threshold for laboratories to join the Many Labs project and having a sample size at which estimates would be reasonable. Other inputs could have been used to determine the planned sample sizes. For instance, based on an expected or minimal clinically relevant standardised effect size δ_{min} and a targeted power $1 - \beta$, as elaborated on in Section 3. We revisit these two sets of replication attempts in Section 6.

2 Anytime-Valid Inference and the Definition of E -Processes and e -Values

In the two examples above, we considered the aggressive *data-dependent* so-called first-passage time N at which the e -value passes $1/\alpha$ as a stopping rule. Inference based on e -values does not require us to adhere to this or any other stopping time. We use the term stopping time to describe possibly data-dependent rules for stopping, see Box 2 for a more precise definition and practical examples. Regardless of the stopping time N we choose, or is forced upon us, the chance of $e_N \geq 1/\alpha$ remains forever small if there truly is no effect. This robustness to the stopping time follows quite directly from its definition, which involves some statistical terminology that we briefly review first.

2.1 Type I Error α Control is Relevant for Both Tests and Confidence Intervals

By a (statistical) test or test procedure we refer to a random variable that takes in data, typically via a p -value, Bayes factor, or e -value, and has two potential realisations: Either “reject” or “not reject” the null hypothesis. Ideally, we want the test to only reject the null when it is false, and only refrain from rejecting the null hypothesis when it is true, see Table 1. Regrettably, due to individual differences, no statistical method can entirely

InfoBox 2: Definition of stopping times and examples

A stopping time N is a potentially data-dependent rule for halting data collection, as long as the decision to stop does not depend on future events. Here are some examples:

- a. The **classic stopping rule** to halt data collection at a fixed time, say, $N = 200$, after observing $n_1 = n_2 = 200$ pairs. Halting is not influenced by any observation, particularly those after $N = 200$.
- b. The **first-passage time** N , at which $e_N \geq 1/\alpha$ occurs, only depends on the data up to time N .
- c. The **external stopping time** when we have to halt data collection due to our funds being depleted, or because our measurement instrument, such as an EEG scanner, malfunctions.
- d. **Frustrated stopping time**: The moment the study stops, because the principal investigator quits his job in rage due to his constant struggles with a particularly evil printer.
- e. **Convenient stopping time**: The point in time when a meeting is finally scheduled with a funding agent, governing board and ethics committee during which it is established that the e -value at that time is satisfactory after intermediate assessments.

Stopping times can also be reformulated during data collection:

- A. **Delayed first-passage time**: Assume that the first-passage time N occurs early on. This makes us conservative, leading us to require the e -value to remain above $1/\alpha$ for, say, an additional 6 time points before actually stopping.
- B. **Forced continuation**: We observed $e = 12 < 1/\alpha = 20$, but a reviewer, convinced that there should be significant evidence against the null hypothesis, insists that we test another 20 observations before concluding the experiment.
- C. **Hopeful continuation**: A Bayesian statistician tells us to stop data collection in favour of the null the first instance $e_N \leq 0.21$, and to stop for the alternative as soon as $e_N \geq 16$. Suppose that during data acquisition, $e_n \leq 0.21$ is observed, but that the investigator is hopeful and continues 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. Finally, let M be the (potentially unknown in advance) time at which one must stop due to the exhaustion of resources such as money, time, or energy. It follows that for any stopping time N , the minimum between N and M , i.e. $\min\{N, M\}$ is also a stopping time. \diamond

eliminate the tabulated errors stemming from sampling only a (small) portion of a large population. The general consensus is that a type I error, that is, a false positive rejection of a true null hypothesis, is the worst type of error. A false positive finding arises when random noise is mistaken for a structural effect. Type I errors are costly, as they introduce random noise into a scientific field that typically persists in the literature. Furthermore, type I errors can lead to fruitless research programmes and hurt the credibility of the field (Simmons et al., 2011). Thus, to ensure reliable statistical hypothesis tests, we always first insist that the type I error is controlled.

Type I error control of level α remains a directly relevant quantity if the goal is to report $1 - \alpha$ confidence intervals alongside, or even in place of, a test (Amrhein et al., 2019). This relevance arises because α reflects the chance that a confidence interval will not cover the data-governing parameter, such as the population mean difference φ . Classical $1 - \alpha$

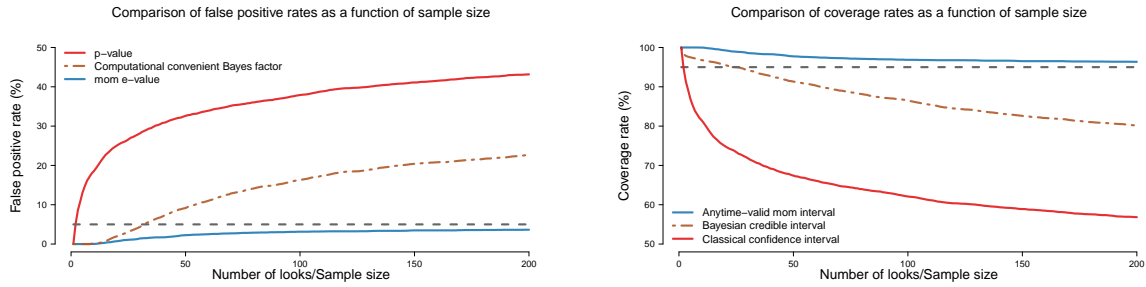
Table 1*Schematic of binary classification errors when testing.*

	Test outcome	
	Reject	Not reject
True \mathcal{H}_0	Type I error	Correct decision
False \mathcal{H}_0	Correct decision	Type II error

confidence intervals are the inversions of the $p < \alpha$ tests; see Section 4.5 Bickel and Doksum, 2015 and Box 3 for further details.

To control the type I error rate for a test or confidence interval, we choose a nominal α , which is typically, perhaps ritually, set to $\alpha = 0.05$. This α serves two purposes. Firstly, it defines the threshold of null rejection, namely, whenever $p < \alpha$. Secondly, α represents the tolerable chance with which the test is allowed to produce a false null rejection. By ‘chance’, we mean the relative frequencies of the potential realisations of the test under repeated uses, *assuming that a hypothesis, such as the null hypothesis, holds true*. For instance, if a $p < \alpha = 0.05$ test were applied to 100,000 experiments with data generated under the null, then we are willing to put up with at most 5,000 incorrect null rejections.

Classical p -value tests require strict adherence to the sampling plan for the False Positive Rate (FPR) to align with the nominal type I error. The FPR defined as the *realised type I error* can be much higher than α , e.g. when the classical $p < \alpha$ test is conducted during data collection (see the red curve in Fig. 2 for more details). The repeated use of

**Figure 2**

Left panel: Both the false positive rates of monitoring the $p < \alpha$ (solid red) and a $\text{BF}_{10} \geq 1/\alpha$ tests (dashed brown) increase well beyond the tolerable $\alpha = 0.05$ -level, where BF_{10} represents a (non-default) computationally convenient Bayes factor (Appendix D). On the other hand, the FPR of the $e \geq 1/\alpha$ test (blue) remains below the tolerable α at all moments in time, see Section 2.2 for more details. Right panel: Both the coverage rates of the classical confidence interval (solid red) and the (Bayesian) credible interval (dashed brown) dip below the nominal 95%-level, if they are used sequentially. The coverage rate of the anytime-valid confidence interval remains above the nominal 95% as promised. Simulation details are provided in Section 2.4.

the standard $p < \alpha$ test is then of level α in name only, hence the adjective nominal, and

InfoBox 3: The duality between classical $1 - \alpha$ confidence intervals and $p < \alpha$ tests

A classical confidence interval inverts a p -value test. For instance, for any null value φ_0 we have a T -statistic, Eq. (1). The $p < \alpha$ test is equivalent to rejecting the null value φ_0 whenever the observed t -statistic is larger than the threshold $t_{\alpha,\nu}$. For instance, when $\alpha = 0.05$ and $n_1 = n_2 = 18$, then the degrees of freedom equal $\nu = 34$ and the null value φ_0 is rejected if $t > t_{\alpha,\nu} = 2.03$ or if $t < -t_{\alpha,\nu}$. For $n_1 = n_2 = 456$ the $p < \alpha$ test corresponds to $|t| > t_{\alpha,\nu} = 1.96$. More generally, if the true population mean difference equals the postulated φ_0 used in the T -statistic, then there is less than α chance to observe outcomes of T with magnitude larger than $t_{\alpha,\nu}$:

$$\text{At each } n_1, n_2 \text{ and all null values } \varphi_0 \in \mathbb{R} : \mathbb{P}_{\varphi_0} \left(\left| \sqrt{n_\delta} \frac{\bar{X}_1 - \bar{X}_2 - \varphi_0}{S_p} \right| > t_{\alpha,\nu} \right) \leq \alpha, \quad (3)$$

where $t_{\alpha,\nu}$ is the $1 - \alpha/2$ quantile of a T -distribution with ν degrees of freedom.

A classical $1 - \alpha$ confidence interval inverts Eq. (3) by negating the event within \mathbb{P}_{φ_0} . This follows from the fact that chances sum to one, which implies that the chance of the complement of the event A denoted by A^c is given by $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$. This leads to:

$$\text{At each } n_1, n_2 \text{ and all null values } \varphi_0 \in \mathbb{R} : \mathbb{P}_{\varphi_0} \left(\left| \sqrt{n_\delta} \frac{\bar{X}_1 - \bar{X}_2 - \varphi_0}{S_p} \right| \leq t_{\alpha,\nu} \right) > 1 - \alpha. \quad (4)$$

Rewriting the statement within \mathbb{P}_{φ_0} shows that

$$\text{At each } n_1, n_2 \text{ fixed, the interval } \text{CI}(1 - \alpha) := \left[\bar{X}_1 - \bar{X}_2 - \frac{S_p}{\sqrt{n_\delta}} t_{\alpha,\nu}, \bar{X}_1 - \bar{X}_2 + \frac{S_p}{\sqrt{n_\delta}} t_{\alpha,\nu} \right] \quad (5)$$

will encapsulate φ_0 with at least $1 - \alpha$ chance. It is worth emphasising that the $1 - \alpha$ chance pertains to the interval, as that is data-dependent, not the parameter value, and this chance drops well below $1 - \alpha$ if n_1, n_2 are not fixed, see Fig. 2. \diamond

has an intolerably inflated chance of a false null rejection. A classical p -value test is valid, if it is conducted once — and only once, as the FPR then (exactly) equals the tolerable α . This is the reason why valid inference based on p -values is limited to the protocol where data collection is first stopped, after which the p -value test is performed. Further data acquisition and analysis leads to an FPR larger than the nominal α . In other words, with a classical p -value test we have a one-shot chance to reliably conduct the test, essentially rendering the collected data single use.

In contrast, the safe anytime-valid inference paradigm allows researchers to reverse the protocol, enabling them to monitor the test results before concluding the experiment, all while maintaining type I error control. The test is also specified by α , and a null rejection is realised whenever the e -value is compellingly large, that is, $e \geq 1/\alpha$ such as $e \geq 20$ for $\alpha = 0.05$. Monitoring the test implies that at each “time” n we can check whether the e -value at that moment, denoted by e_n , exceeds the evidence threshold $1/\alpha$. By construction, as elaborated on in the next section, the FPR for anytime-valid tests will never exceed the tolerable α -level. It is important to note that stopping is *not necessary* for type I error control, it is *allowed*, as we always have the option to continue the study. This feature is not shared by other sequential methods.³

³For example, Wald’s sequential tests require a precise stopping rule to be determined in advance (see

2.2 Inference as an Ongoing Process and the Definition of E -Processes as a Generalisation of Likelihood Ratios

More generally, as the data accrue, the observed e -values form a sequence/sample path of non-negative numbers $e = (e_1, e_2, \dots, e_n, \dots)$, e.g. Fig. 1, that is realised by a so-called E -process $E = (E_1, E_2, \dots, E_n, \dots)$. Following standard conventions, we denote a random variable in upper case and its realisation in lower case. For instance, the random variable E_n refers to the anticipated e -value, that is, before the data are observed at time n . The randomness and variability in E_n arise from data that have not yet been observed. In contrast, after data observation E_n realises a number e_n such as $e_n = 7$, which does not vary. For an E -process to reliably quantify the evidence against the null at any moment in time, it has to fulfil three defining properties:

- (i) It has to quantify neutral evidence at the start of the process, 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 model \mathcal{M}_0 (Box 1), and regardless of the stopping time N (Box 2) we expect E_N to convey at most neutral evidence:

$$\text{For any stopping time } N \text{ and } \mathbb{P} \in \mathcal{M}_0 \text{ we require } \mathbb{E}_{\mathbb{P}}[E_N] \leq 1, \quad (6)$$

where $\mathbb{E}_{\mathbb{P}}$ is the expectation with respect to a data-generating distribution \mathbb{P} from the null model \mathcal{M}_0 . An intuitive analogy to an E -process is a bettor's wealth in a multi-round game of roulette (see Box 4).

2.2.1 Example: Simple Likelihood Ratios

These three defining properties hold naturally for the better-known notion of likelihood ratios when the null model consists of a single data-generating distribution, as is the case for a one-sample z -test. The null model then specifies that the data are normally distributed with known mean and variance such as $X_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu_0, 1)$, where μ_0 is fixed, typically, $\mu_0 = 0$. The typical alternative model consists of any data-generating normal distribution with mean μ some real-valued number that is allowed to vary freely. Let us write $q(x_i) = f(x_i | \mu = \mu_s)$ for the likelihood under the alternative for outcome x_i with μ fixed at μ_s , say, $\mu_s = 1$, and $p_0(x_i) = f(x_i | \mu = \mu_0)$ denotes the likelihood under the null. The likelihood ratio (of the alternative over the null) for observations $x^{(n)} := (x_1, \dots, x_n)$ is then given by the product

$$\text{lr}_n^{\mu_s}(x^{(n)}) = \prod_{i=1}^n \frac{q(x_i)}{p_0(x_i)} = \frac{q(x_1)}{p_0(x_1)} \times \frac{q(x_2)}{p_0(x_2)} \times \dots \times \frac{q(x_n)}{p_0(x_n)}. \quad (7)$$

Section 4 for a brief discussion). Also, Johari et al., 2022 introduce *anytime-valid p -values*, which allow optional continuation; but upon closer inspection their anytime-valid p -values are e -values in disguise (Ramdas et al., 2020).

InfoBox 4: 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. 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, as for a roulette table without the green outcome, the null hypothesis corresponds to $p(\text{RED}) = p(\text{BLACK}) = 1/2$, then the game maker (i.e. the casino/nature) can set $o(x) = 1/p(x) = 2$. At each round n , before seeing x_n , the bettor can decide the strategy with which she spreads the wealth e_{n-1} she has gained so far over all outcomes by putting a fraction $q(x)$ of her money on outcome x ; for example, $q(\text{RED}) = 2/3, q(\text{BLACK}) = 1/3$. She receives e_{n-1} multiplied by $l_n(x_n) := q(x_n) \times o(x_n)$, where x_n is the realised outcome; 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 $e_n = \prod_{i=1}^n l_i(x_i) = l_1(x_1) \times l_2(x_2) \times \dots \times l_n(x_n)$. For example, after observing a sequence **(RED, RED, 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 $E = (E_1, E_2, \dots)$ an E -process. It further is a *good* E -process if additionally, the bettor is expected to get rich fast if the alternative is true. Betting with strategy $q(\text{RED}) = 2/3$ at each round n makes sense if the bettor thinks that the null hypothesis is false and that there is a substantially larger chance of observing **RED** compared to **BLACK**. If this is really 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 guarantees under optional stopping is the *same* mathematical phenomenon as the fact that the chance that you can multiply your initial capital in a real casino by a factor K is bounded by $1/K$ — 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 in the past 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), and Waudby-Smith and Ramdas (2024) and at an introductory level to Ter Schure, 2023, Chapter 1 for further details.

◇

To see that the likelihood ratio statistic, i.e. $\text{LR}_n^{\mu_s} = \text{lr}_n^{\mu_s}(X^{(n)})$ that takes in random data $X^{(n)}$, is an E -process, we verify the properties. For (ii) note that for all outcomes $x^{(n)}$ the likelihoods, thus, also their ratios, are always non-negative. For (i) we use the fact that for $n = 1$ fixed, the expectation involves integrating/summing over all possible realisations of X , and that for each fixed μ_s the probability density function q integrates to one, that is,

$$\mathbb{E}_{\mathbb{P}_0}[\text{LR}_1^{\mu_s}] = \int_{-\infty}^{\infty} \frac{q(x)}{p_0(x)} p_0(x) dx = \int_{-\infty}^{\infty} q(x) dx = 1 \leq 1. \quad (8)$$

The assumption that q and p_0 have the normal form was never used in the derivation above, which is why this argument holds for simple likelihood ratios in general; as long as the null hypothesis consists of a single distribution. If, as in the t -test, this is not the case, most likelihood ratios will not give E -processes, see Section 2.3. The verification of Property (iii) requires some technicalities, but follows the same logic, see Appendix A. These arguments imply that Wald's sequential probability ratio statistic, one of the first statistics that was used in a sequential test, is an E -process.

Moreover, mixtures of E -processes are also E -processes. For instance, instead of choosing a single μ_s in the numerator of the z -likelihood ratio, we can take a weighted average of the likelihood ratio with respect to μ_s . For this we employ a prior distribution $\pi(\mu_s)$ that serves as mixture weights provided that $\int \pi(\mu_s) d\mu_s = 1$. If Property (iii) holds for any μ_s , then it also holds for its mixture by interchanging the order of integration/expectation:

$$\mathbb{E}_{\mathbb{P}_0} \left[\int_{-\infty}^{\infty} \text{LR}_N^{\mu_s} \pi(\mu_s) d\mu_s \right] = \int_{-\infty}^{\infty} \mathbb{E}_{\mathbb{P}_0} [\text{LR}_N^{\mu_s}] \pi(\mu_s) d\mu_s \stackrel{(iii)}{\leq} \int_{-\infty}^{\infty} 1 \pi(\mu_s) d\mu_s = 1, \quad (9)$$

where N is an arbitrary stopping time. This derivation does not explicitly use that $\text{LR}_n^{\mu_s}$ is a likelihood ratio, only Property (iii) for every fixed μ_s and that $\pi(\mu_s)$ integrates to one. Hence, a mixture of E -processes that are not likelihood ratios is also an E -process. Since this derivation holds for any stopping time N it also holds for the deterministic stopping time $N = 1$, thus, Property (i) follows. These arguments imply that E -processes are in general not unique: Any fixed μ_s in the numerator of $\text{LR}_n^{\mu_s}$ yields an E -process parametrised by μ_s , and so does any mixture. Some have higher power than others to detect an effect, see Section 3 below. Regardless of the choice, all E -processes provide anytime-valid type I error control.

2.2.2 Ville's Inequality Implies Anytime-Valid Type I Error Control

The three defining properties of E -processes, in particular, Property (iii) being formulated with respect to *any* stopping time, allows for anytime-valid type I error control via *Ville's inequality*. The essence of Ville's inequality is using the expectation to bound the chance of a rare event. For instance, we expect 100 heads, if we stop flipping a fair coin after 200 flips. This expectation does not rule out the possibility of observing 190 heads in 200 flips, but it does imply that such an extreme event occurs infrequently, that is, only once in many repetitions of 200 coin flips.

Ville's inequality (Ruf et al., 2023) quantifies that for data generated under any distribution \mathbb{P} from the null model \mathcal{M}_0 there is at most a small chance α that an E -process will *ever* yield evidence against the null larger than $1/\alpha$. That is,

$$\text{For all } \mathbb{P} \in \mathcal{M}_0 : \mathbb{P}(\text{There exists a stopping time } N \text{ such that } E_N \geq 1/\alpha) \leq \alpha. \quad (10)$$

This means that when there truly is no effect, there is little chance to nudge the evidence above the threshold, for instance, due to a hopeful reviewer or investigator, see Box 2. The discussion surrounding Fig. 5 below provides some intuition on this fact. Due to its robustness to all stopping times, we call the test that rejects the null whenever the e -value crosses the threshold $1/\alpha$ a safe anytime-valid inference (savi) test of level α .

This does not mean that anytime-valid tests can handle data dredging, which is associated with a strategy for halting data collecting that is not a stopping time. For instance, the cheating strategy of reporting only $n = 40$, while actually collecting 90 observations and removing 50 “outliers” because they led to a low e -value, can be associated with a retroactive stop at $n = 40$. This retroactive decision to stop depends not only on the $n = 40$ data points, but on the full data set up to $n = 90$. Thus, we did not use a valid stopping time and hence the reported e -value is not guaranteed to be lower than $1/\alpha$ with at most α chance: cheating is still prohibited within this framework of inference, and it is not anticipated to be permitted within any reasonable framework of inference.

Similar to how a confidence interval is constructed from a p -value test, see Box 3, we can invert Ville’s inequality, thus, the $e_N \geq 1/\alpha$ test to construct anytime-valid confidence intervals. The notable difference between the p -value test, e.g. Eq. (3), and an anytime-valid test, Eq. (10), is the placement of the sample size within the probability statement, which therefore also needs to be negated. The negation of a “there exists” statement is a “for all” statement and vice versa, resulting in:⁴

$$\text{For all } \mathbb{P} \in \mathcal{M}_0 : \mathbb{P}(\text{For all stopping times } N : E_N < 1/\alpha) \geq 1 - \alpha. \quad (11)$$

This version of Ville’s inequality, thus, states that for data generated by any distribution from the null model, the evidence against the null remains *forever* smaller than $1/\alpha$ with at least $1 - \alpha$ chance. We exploit the fact that the statement holds for all times and convert the evidence to a $1 - \alpha$ confidence interval at a particular time n by gathering all null values φ_0 that have not (yet) led to an e -value larger than $1/\alpha$, e.g. see Box 5. Ville’s inequality Eq. (11) ensures that the thus constructed “running intersection” $1 - \alpha$ confidence interval (Howard et al., 2021) encapsulates the data-governing effect size of interest at all times with at least $1 - \alpha$ chance. These intervals are therefore also attractive to those who suggest to eliminate the concept of statistical significance in favour of confidence intervals (Amrhein et al., 2019).

Ville’s inequality also provides guidance on when data collection can be stopped. Since it holds for any stopping time, it also holds for the earliest time between the aggressive first-passage time N at which $e_N \geq 1/\alpha$ and some planned sample size n_{plan} , or the sample size at which the semester ends, or when the resources are depleted, as was the case for various laboratories in the Many Labs 2 project. Section 3 below on sample size determination with E -processes provides guidance on how n_{plan} can be chosen based on a power analysis. Ville’s inequality thus justifies the e -value testing procedure as described in Protocol 2.1 in pseudo R code.

⁴For example, the existence of just one black swan is enough to disprove the statement “all swans are white”.

```

1 # Pseudo code: This code does NOT run
2 n ← 1
3 eValueAtTime[1] ← currentEValue ← 1
4
5 while (currentEValue < 1/alpha && n <= nPlan) {
6   currentEValue ← saviTTest(x[1:n], y[1:n], designObj=designObj)
7   eValueAtTime[t] ← currentEValue
8
9   if (currentEValue >= 1/alpha) {
10     "Reject the null"
11     stop()
12   } else {
13     "Increase sample size and test again
14       at the start of the while loop"
15     n ← n + 1
16   }
17 }

```

Protocol 2.1: Pseudo R code for the e -value testing procedure with an n_{plan} .

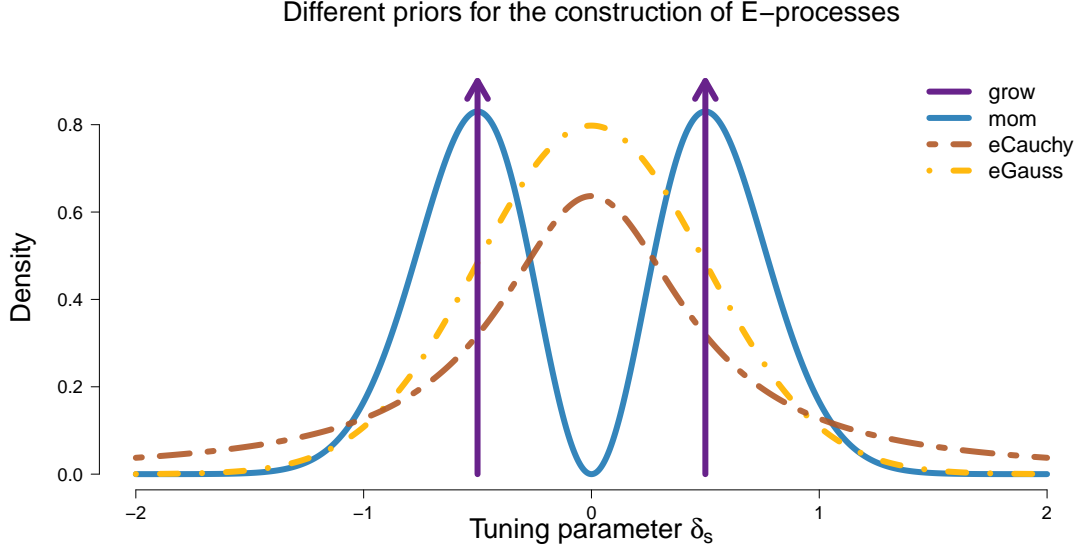
2.3 Four Types of Anytime-Valid t -Tests

The arguments used to show that likelihood ratios are E -processes crucially rely on the null model being simple. Constructing E -processes for composite null models, however, is much more involved. For example, the two-sample t -test used in Section 1 involves a composite null hypothesis. We can take $\mathbb{P} \in \mathcal{M}_0$ to be two identical normal distributions with grand mean $\mu_g = \mu_1 = \mu_2 = 4$ and $\sigma = 2$ or any other value for μ_g and $\sigma > 0$, see Box 1. Some technical effort (Hendriksen et al., 2021, Pérez-Ortiz et al., 2024) is needed to rigorously prove that the following likelihood ratio is an E -process, see also Gronau et al. (2020, Appendix A) for the explicit computations,

$$\text{lr}_n^{\delta_s}(t) := \frac{T_\nu(t|\sqrt{n_\delta}\delta_s)}{T_\nu(t)}, \text{ and where } \text{LR}_n^{\delta_s} := \text{lr}_n^{\delta_s}(T), \quad (12)$$

where $T_\nu(t)$ is the likelihood of the T -distribution with ν degrees of freedom centred at zero, and $T_\nu(t|\sqrt{n_\delta}\delta_s)$ in the numerator above it is the T -likelihood for outcome t at the non-centrality parameter $\sqrt{n_\delta}\delta_s$, where n_δ is the effective sample size (Box 1), and δ_s some savi test defining parameter.

This tuning parameter δ_s , and the parameters for the E -process types given below, can be selected optimally, if we are provided with an expected, or a minimal clinically relevant effect size δ_{\min} . Intuitively, an optimal choice leads to an E -process that *accumulates evidence as fast as possible, thus, requiring the smallest possible average sample size* to reject the null whenever the data-governing parameter equals δ_{\min} (Ter Schure et al., 2024). Below we use the term ‘optimal’ and ‘fastest evidence accumulating’ E -process interchangeably. Extended intuition for the meaning of optimal/fastest evidence accumulating is provided in Section 3.1, and a more formal treatment can be found in Appendix C. For now, we just list the optimal choices for each type of t -test E -process. The different types are arrived at

**Figure 3**

The various types of E -processes for the two-sided anytime-valid t -test 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 (dash-dotted brown) on δ_s . All priors are optimised to a minimal clinically relevant effect size $\delta_{\min} = 0.5$.

by putting different prior distributions π on δ_s , mixing the T -likelihood ratios as in (9) in different ways. Crucially, the type I error guarantee remains valid irrespective of how we mix.

2.3.1 *grow*

One of the main results of Pérez-Ortiz et al. (2024) is that the fastest evidence accumulating *one-sided* anytime-valid t -test (alternative $\delta \geq \delta_{\min}$) is given by δ_s in Eq. (12) set equal to δ_{\min} . It follows that the fastest evidence accumulating *two-sided* anytime-valid t -test, which is relevant for our examples, is given by the mixture π as in (9) that puts half its mass at $\delta_s = -\delta_{\min}$ and at $\delta_s = \delta_{\min}$, see the purple arrows in Fig. 3 for a graphical representation. This choice of prior leads to the *grow* E -process for the t -test, where *grow* stands for growth-rate optimal in the worst case, see (Grünwald et al., 2024) and Appendix C for more details.

2.3.2 *eGauss*

An anytime-valid t -test can also be based on the *eGauss* E -process (Gönen et al., 2005) that uses Gaussian priors $\mathcal{N}(0, g)$ on δ_s , instead of the two-point priors. The prior variance g inherits the role of the tuning parameter. Provided with δ_{\min} the optimal choice (amongst all *eGauss* E -processes) is given by $g = \delta_{\min}^2$ (the dashed yellow curve in Fig. 3).

InfoBox 5: Explicit formula of the **mom** E -process, and the associated anytime-valid confidence sequence.

The **mom** E -process has tuning parameter g_{mom} , which can be adapted to δ_{\min} , see Section 3.1 for details, or Section 3.4 when no δ_{\min} is available. It takes as input the observed t -value, i.e. a realisation of Eq. (1), the sample sizes n_1, n_2 , thus, effective sample size $n_\delta = \frac{n_1 n_2}{n_1 + n_2}$, and the degrees of freedom $\nu = n_1 + n_2 - 2$. Its explicit form (Appendix B) is given by

$$E_{n_\delta, \nu}(t) = (1 + n_\delta g_{\text{mom}})^{-\frac{3}{2}} \left(\frac{1 + \frac{t^2}{\nu}}{1 + \frac{t^2}{\nu(1 + n_\delta g_{\text{mom}})}} \right)^{1 + \frac{\nu+1}{2}} \left(\frac{1 + \frac{1 + n_\delta g_{\text{mom}}(\nu+1)}{\nu(1 + n_\delta g_{\text{mom}})} t^2}{1 + \frac{t^2}{\nu}} \right). \quad (13)$$

The same formula holds for the one-sample case but with t replaced by its one-sample counterpart, $n_\delta = n_1$ and $\nu = n_1 - 1$.

The associate anytime-valid $1 - \alpha$ confidence interval at time N , with sample sizes N_1 and N_2 respectively, is based on Eq. (11) by gathering all φ_0 for which, at all times $n_1 = 1, 2, \dots, N_1$ and $n_2 = 1, 2, \dots, N_2$, we have:

$$E_{n_\delta, \nu} \left(\frac{\sqrt{n_\delta}(\bar{x}_1 - \bar{x}_2 - \varphi_0)}{s_p} \right) \leq 1/\alpha. \quad (14)$$

The inversion is done numerically. In the real-world examples above, the parameter was set to $g_{\text{mom}} = 0.134$, which corresponds to the **mom** confidence interval that, at $n_1 = n_2 = 40$, is the most narrow. \diamond

2.3.3 **mom**

Alternatively, anytime-valid inference can also be based on the **mom** E -process derived from a so-called non-local moment prior on δ_s (Johnson & Rossell, 2010; Pramanik & Johnson, 2022). This 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. The associated E -process can be computed explicitly and it is given in Box 5. The optimal choice (amongst all **mom** E -processes) corresponds to putting the bumps at $-\delta_{\min}$ and δ_{\min} , which mimics the behaviour of the **grow** choice. This is achieved by setting g_{mom} in Box 5 to $g_{\text{mom}} = \delta_{\min}^2/2$.

2.3.4 **eCauchy**

Similarly, we can also opt for an **eCauchy** t -test E -process based on a Cauchy prior $\delta_s \sim \text{Cauchy}(0, \kappa^2)$ (Jeffreys, 1961; Rouder et al., 2009) with scale parameter κ represented by the dash-dotted red curve in Fig. 3. The optimal choice involves setting $\kappa = |\delta_{\min}|$.

All four types of E -processes have in some form been introduced in the literature as Bayes factors (Jeffreys, 1961, Ly et al., 2016a, 2016b) with specifically chosen priors on the nuisance parameters, i.e. μ_g and σ . For instance, not adapting to δ_{\min} by setting $\kappa = |\delta_{\min}|$, but fixing $\kappa = 1/\sqrt{2}$ recovers the default choice in psychology (Rouder et al., 2009; Schönbrodt et al., 2017).

It is not well known that these Bayes factors are also E -processes. Furthermore, it should be noted that (a) not all E -processes are Bayes factors (e.g. Wang and Ramdas, in press), and that (b) not all Bayes factors are E -processes. For (a) we mention the universal inference construction (Wasserman et al., 2020). This method was used to construct

anytime-valid tests where no reliable p -value or Bayesian procedure can be reasonably formulated; see, for instance Pandeva et al. (2024), Ramdas et al. (2023), and the reference therein for further details. There also exists an information projection method (Grünwald et al., 2024) to construct powerful E -processes such as the anytime-valid test for two proportions (Turner et al., 2024), which is also not a Bayes factor. Section 4 below shows an example of a Bayes factor that is not an E -process.

Exploiting the fact that the Bayes factor t -tests listed above are also E -processes allows us to use them in Protocol 2.1 with frequentist type I error guarantees. Frequentist guarantees for a sequential Bayes factor test, on the other hand, uses evidence threshold that require extensive simulations for their justifications, see Section 4.1 for further discussions.

2.4 E -Processes Slowly Drift Towards Zero Under the Null

In reality, we do not know whether the null hypothesis holds true. In simulation studies, when the null hypothesis is known to be true, we expect a good test to maintain a FPR below the tolerable level α . For E -processes, the need for simulating under the null is redundant due to the assurance provided by Ville’s inequality. However, for purely illustrative purposes, we nonetheless generate a data set under the null hypothesis of no effect with $\mu_1 = \mu_2 = 4$ and $\sigma = 2$ until time $n = 200$ at which $n_1 = n_2 = 200$. To compare and to emphasise the point that the $p < \alpha$ test should only be performed once, we also depict the evolution of p -values as the red curve in the top left panel of Fig. 4. The dark grey horizontal dashed line represents the tolerable $\alpha = 0.05$. There are 31 times where the p -value dips below $\alpha = 0.05$, namely, at $n_1 = n_2 = 4, 5, 127, 128, 131, 135, 144, 179, 185$, and between $n_1 = n_2 = 187$ and $n_1 = n_2 = 200$. The bottom left panel of Fig. 4 shows the corresponding confidence interval. The dashed yellow line indicates $\mu_1 - \mu_2 = 0$ representing the true data-governing mean difference, as the data are generated under the null. Note that the true mean difference falls outside the 95% confidence interval whenever the p -value incorrectly rejects the null hypothesis at level $\alpha = 0.05$, e.g. Box 3. Hence, for this particular data set, monitoring and rejecting the null hypothesis as soon as p dips below α yields a false null rejection. Subsequent inference based on the classical interval then also provides incorrect conclusions regarding the magnitude of the mean difference φ .

The top right panel of Fig. 4 shows the evolution of the mom e -values for the same data set, with the parameter set to $g_{\text{mom}} = 0.134$, as in Section 1. The dotted light grey horizontal line represents neutral evidence $e = 1$. The dashed dark grey line at the top represents the evidence threshold of $1/\alpha = 20$, which the e -value sample path correctly remains under. The bottom right panel of Fig. 4 shows the corresponding 95% anytime-valid confidence intervals. At each time n , the interval contains all values of the mean difference parameter φ_0 serving as a null hypothesis $\mathcal{H}_0 : \mu_1 - \mu_2 = \varphi_0$ for which the corresponding e -values have remained below the evidence threshold, here, $1/\alpha = 20$.

Fig. 2, shown further above, illustrates the performance of the two procedures under repeated use. A simulation study was performed based on 5000 data sets generated under the null with $\mu_1 = \mu_2 = 4$ and $\sigma = 2$ as before. The FPR at time $n = n_1 = n_2$ was determined by tallying the number of data sets that, up to that point, led to a false rejection of the null hypothesis. The number of false positives is then divided by the total number of data sets. Similarly, the number of data sets that included the true mean difference of $\mu_1 - \mu_2 = 0$ at all times up to $n_1 = n_2$ was recorded before dividing it by 5000. Fig. 2 depicts in blue

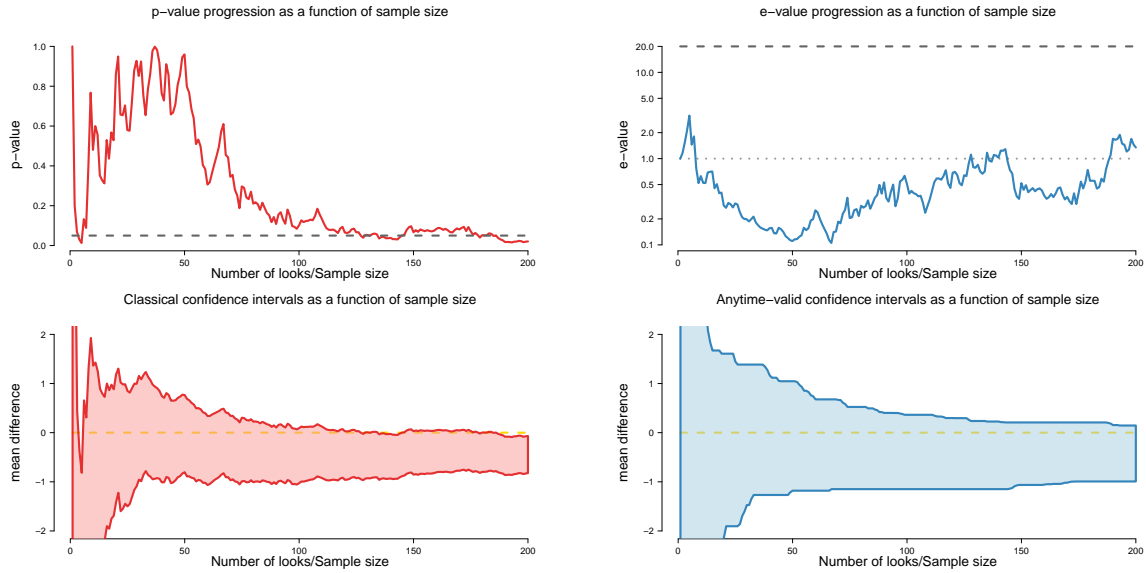


Figure 4

Top left panel: Given a sufficiently large time horizon, here $n_1 = n_2 = 200$, 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 e -values correctly remain under the threshold of $1/\alpha$. Bottom right panel: The anytime-valid confidence intervals cover the true effect size at any moment in time.

the FPR of the anytime-valid test based on the `mom` E -process with $g_{\text{mom}} = 0.134$, which we have used so far. The red curve represents the FPR of the procedure in which the p -value is monitored and the null is rejected as soon as $p < \alpha$ is observed. This red curve will continue to increase to an FPR of 100% resulting in a sure rejection despite the null being true. The left column of Fig. 4 is therefore typical. Moreover, this problem occurs for all classical p -value testing scenarios, not just t -tests. The core issue is that, under the null hypothesis, p -values are uniformly distributed at each sample size, causing them to meander between zero and one and eventually dip below any α threshold.

The dramatic increase in FPR when monitoring the $p < \alpha$ test is caused by the number of looks, not by the test being performed after each pair of observations. For instance, if the $p < \alpha$ test were conducted after 38, 20 and 40 participants, the FPR would be approximately 5%, 8.64%, and 10.80%, respectively, as depicted in Fig. 2. Note that the p -value test already “spent” all the tolerable $\alpha = 0.05$ at the first look, which reiterates the point that a classical p -value test is only reliable if it is conducted once.

In contrast, after 200 looks the FPR of the anytime-valid test is only 3.64%; less than the tolerable $\alpha = 0.05$. Analogously, after 200 looks the coverage rate of the 95% anytime-valid confidence interval was 96.36%. As the number of looks increases, so will the FPR, but only slightly. Ville’s inequality guarantees that increasing the time horizon from $n = n_1 = n_2 = 200$ to 63 million, or even indefinitely, will not cause the FPR of the anytime-valid test to exceed the tolerable 5%. This is due to the distribution of e -values

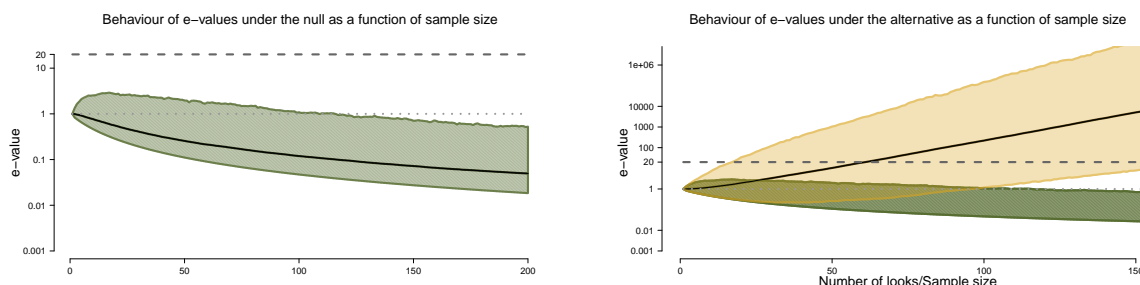


Figure 5

*Left panel: E -processes start at one and under the null do not increase and typically drift towards zero. Right panel: Under the alternative, here with the data-governing δ equal to δ_{\min} , the distribution of *mom* e -values increases. 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 -processes.*

not increasing and typically monotonically drifting to zero under the null. The left panel of Fig. 5 depicts this decreasing drift by illustrating the sampling distribution of the *mom* E -process under the null at each time n . The vertical axis is on a logarithmic scale, which makes e -values of $1/100$ and $1/1000$ as far removed from 1 as e -values of 100 and 1000, respectively. The top and bottom green curves depict the 95% and 5% quantiles of the distribution of e -values, respectively, whereas the black curve represents the mean of the logarithm of the E -process. This general decreasing trend towards zero makes it increasingly harder for the E -process to yield $e_n \geq 1/\alpha$ under the null for larger values of n . As such, additional data will tend to decrease the evidence against the null, if the null holds true.

This also suggests that an inadvertent $e_n \geq 1/\alpha$ under the null will eventually be corrected due to the e -values being pushed towards zero if we continue sampling. This non-increasing drift not only applies to all the aforementioned anytime-valid tests (*mom*, *eGauss*, etc), but to all E -processes in general. Intuitively, we can view an E -process as a non-increasing, and in some cases decreasing, measure of evidence whenever the null model holds true.

3 Designing Sampling Strategies for Experiments Based on E -Processes

A good E -process should not only decrease under any data generating distribution \mathbb{P}_0 from the null model \mathcal{M}_0 , but also accumulate evidence against \mathcal{M}_0 when the null hypothesis is false. For instance, using notation introduced in Box 1, when in a t -test the data-governing standardised population effect size $\delta = (\varphi - \varphi_0)/\sigma$ is non-zero. In such a case, the faster an E -process grows above the threshold $1/\alpha$, the sooner we have the option to reject the null and conserve resources.

For this purpose, we implemented design functions, such as `designSaviT`, in the `safestats` package. These functions select the fastest evidence accumulating, say, *mom* E -process, based on a minimal clinically relevant (standardised) effect size to enable efficient inference. When also provided with a permissible type II error (β) such a design function will then determine the sample size we should plan for to arrive at a correct null rejection

with at least $1 - \beta$ (e.g. 80%) power. For instance, Lines 2 and 3 of R Code 3.1, that

```

1 deltaMin <- 0.5176537
2 designNPlan <- designSaviT(deltaMin=deltaMin, beta=0.2,
3                             alpha=0.05, testType="twoSample", seed=4)
4 designDeltaMin <- designSaviT(nPlan=c(40, 40), beta=0.2, seed=3,
5                               alpha=0.05, testType="twoSample")
6 designBeta <- designSaviT(deltaMin=0.7, nPlan=c(40, 40),
7                            alpha=0.05, testType="twoSample", seed=2)
8 designNarrowestInterval <- designSaviT(nPlan=c(40, 40), alpha=0.05,
9                                         testType="twoSample")

```

R Code 3.1: R Code for designing Savi analyses. The design function `designSaviT()` takes as input α and any two of the three quantities $\beta, \delta_{\min}, n_{\text{plan}}$ to yield an indication of the remaining quantity as output. Lines 1, 2 and 3 specify as input β, δ_{\min} and outputs n_{plan} , Lines 4 and 5 take as input β, n_{plan} and outputs an indication of δ_{\min} , Lines 6 and 7 take as input $\delta_{\min}, n_{\text{plan}}$ and yield β . Lastly, Line 8 takes as input n_{plan} and yields the parameter such that at n_{plan} the confidence interval is the narrowest.

is, the function `designSaviT()` with $\delta_{\min} = 0.5176537$ and $\beta = 0.2$ pre-registers a design object that is going to analyse the data with (i) a `mom` e -value t -test with parameter fixed at $g_{\text{mom}} = 0.134$, the same as the one we used in the examples above,⁵ and (ii) it tells us that we need to plan for 94 participants in each group to observe $e_n \geq 1/\alpha$ with 80% chance/power, if the magnitude of the data governing δ is at least δ_{\min} . Fig. 6 shows the full summary. The next two subsections illustrate the role δ_{\min} and β play in selecting the optimal parameters of the E -process and the derivation of n_{plan} .

3.1 Determining the Fastest Evidence Accumulating E -Process Under $\mathbb{P}_{\delta_{\min}}$

Prior/similar experiments or subject experts might provide us with an indication of an expected or a minimal clinically relevant standardised effect δ_{\min} , which is used to tune the fastest evidence accumulating (`mom`) E -process. To illustrate the sense in which it is optimal, we consider a simulation study in which the null hypothesis is false with normally distributed data (Box 1) and a standardised effect size equal to $\delta = \delta_{\min} = 0.5176537$,⁶ and $\mu_g = 4$ and $\sigma = 2$ as before. The right panel of Fig. 5 shows the 95% and 5% quantiles (golden yellow curves), and the mean of the logarithm of the E -process (black curve) under this alternative at each time $n = n_1 = n_2$. At $n = 200$ the average under the alternative is an e -value of 115,150.4, whereas under the null it is about $1/20.2$. This asymmetric behaviour is typical for a good E -process: Under the null such an E -process (slowly) drifts to zero, whereas it grows rapidly (exponentially fast) under the alternative $\mathbb{P}_{\delta_{\min}}$. Hence, when there is an effect, continuing sampling only makes the evidence against the null stronger.

The selected `mom` E -process being optimal for the given δ_{\min} implies that the slope of the black curve in the right panel of Fig. 5 is the steepest achievable amongst all `mom`

⁵This particular choice for δ_{\min} is derived from the relation $g_{\text{mom}} = \delta_{\min}^2/2$ with $g_{\text{mom}} = 0.134$ we have used so far, see Section 3.4 for further details.

⁶The same δ_{\min} that we used to specify the `mom` design object in R Code 3.1.

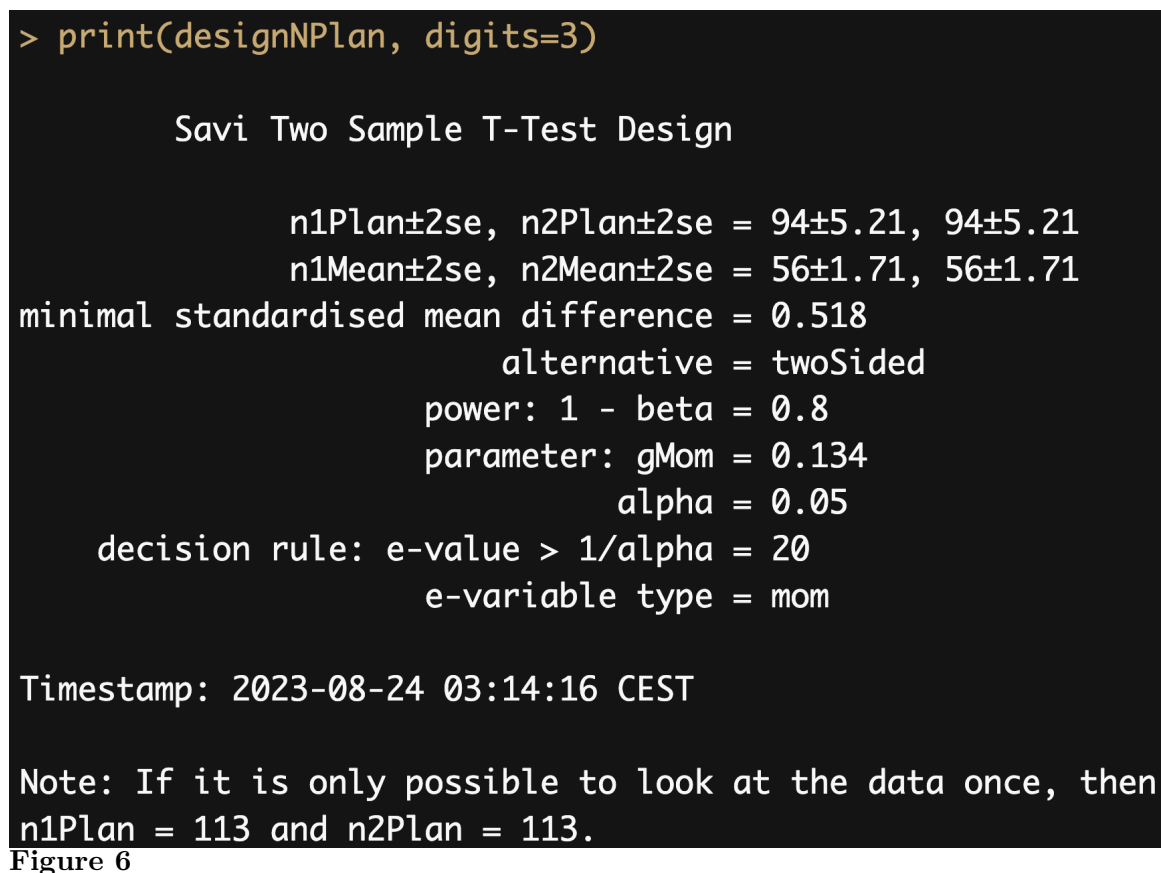


Figure 6

Screenshot of the design object created on Lines 2 and 3 in R Code 3.1.

E -processes under $\mathbb{P}_{\delta_{\min}}$. As such, it is guaranteed to cross the evidence threshold of $1/\alpha$ (broken horizontal grey line) under $\mathbb{P}_{\delta_{\min}}$ the soonest. Provided with β we can also determine when this crossing occurs.

3.2 The Targeted Power $1 - \beta$ /Permissible Type II Error β Defines n_{plan}

The use of the fastest evidence accumulating (mom) E -process under the alternative $\mathbb{P}_{\delta_{\min}}$ together with β determines n_{plan} . Assuming that we aim to stop as soon as we accumulate sufficient evidence to reject the null hypothesis, the natural question arises: how many samples do we need to collect for the optimal mom E -process to reach $1/\alpha$? The answer is the aforementioned n_{plan} of $n_1 = n_2 = 94$, which is derived by horizontally cutting the yellow sampling distribution depicted in the right panel of Fig. 5 at the evidence threshold $1/\alpha$. This is equivalent to studying the distribution of the first times that the E -process passes the evidence threshold $1/\alpha$. To do so (in less than 2.7 seconds on a 2021 iMac M1), Lines 2 and 3 of R Code 3.1 simulate, by default, $m = 1000$ data sets under $\mathbb{P}_{\delta_{\min}}$. Each data set is analysed sequentially, resulting in $m = 1000$ sequences/sample paths of e -values. The first 100 e -value sample paths until they pass the evidence threshold $1/\alpha$ are shown in yellow in Fig. 7. The histogram of the $m = 1000$ first passage times N at which

$e_N \geq 1/\alpha$ occurred is depicted in blue.⁷ The top panel of Fig. 7 summarises the distinct roles of $\alpha, \delta_{\min}, \beta$ simultaneously. Firstly, α defines the test with the evidence threshold $1/\alpha$ depicted as the horizontal black line ($1/\alpha$, e.g. 20). Secondly, δ_{\min} defines the optimal *mom E*-process which is expressed by the steepness of the average overall upward drift of the *e*-value sample paths. Lastly, the blue histogram shows that after $n_1 = n_2 = 94$ observations, a wee few more than 800 out of the $m = 1000$ *e*-value sample paths led to a correct null rejection under $\mathbb{P}_{\delta_{\min}}$. Hence, we can guarantee $1 - \beta$ power (80%, if we set $\beta = 0.2$) by monitoring up to the $1 - \beta$ quantile (94 for $\beta = 0.2$) of the first-passage time distribution, which is therefore recommended as n_{plan} .

To acknowledge that n_{plan} is derived from simulations, the design object also reports twice the bootstrap standard error of 5.21. This uncertainty in the approximation can be decreased by requesting a larger number of sample paths from the design function. The derived n_{plan} serves as an indication of how long the null should be tested for to detect $\delta = \delta_{\min}$, as monitoring $e_n \geq 1/\alpha$ until n_{plan} provides such an effect size ample chance, i.e. $1 - \beta$ power, to reject the null. The output Fig. 6 also shows an average sample size of $n_1 = n_2 = 56$, 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 sample paths that continued until n_{plan} . In comparison, a classical *p*-value test with the same $\alpha, \beta, \delta_{\min}$ always requires $n_1 = n_2 = 60$. Hence, under $\mathbb{P}_{\delta_{\min}}$, the *mom E*-process will require on average four fewer (but in the worst-case, 34 more) participants in both groups compared to the classical *p*-value test. The flexibility of *e*-value based tests comes at the price of a larger sample size to *plan for*, but Fig. 7 shows that in return there is about 57% chance to realise a stopped experiment before $n = 60$, whenever $\delta = \delta_{\min}$.

It is worth emphasising that in practice we do not have to stop the experiment as soon as the evidence crosses $1/\alpha$. If we would like to acquire more evidence against the null, or if the anytime-valid confidence interval is too wide for our liking, we always have the option to continue recruiting new participants without intolerably inflating the FPR, despite this decision to continue being driven by the data.

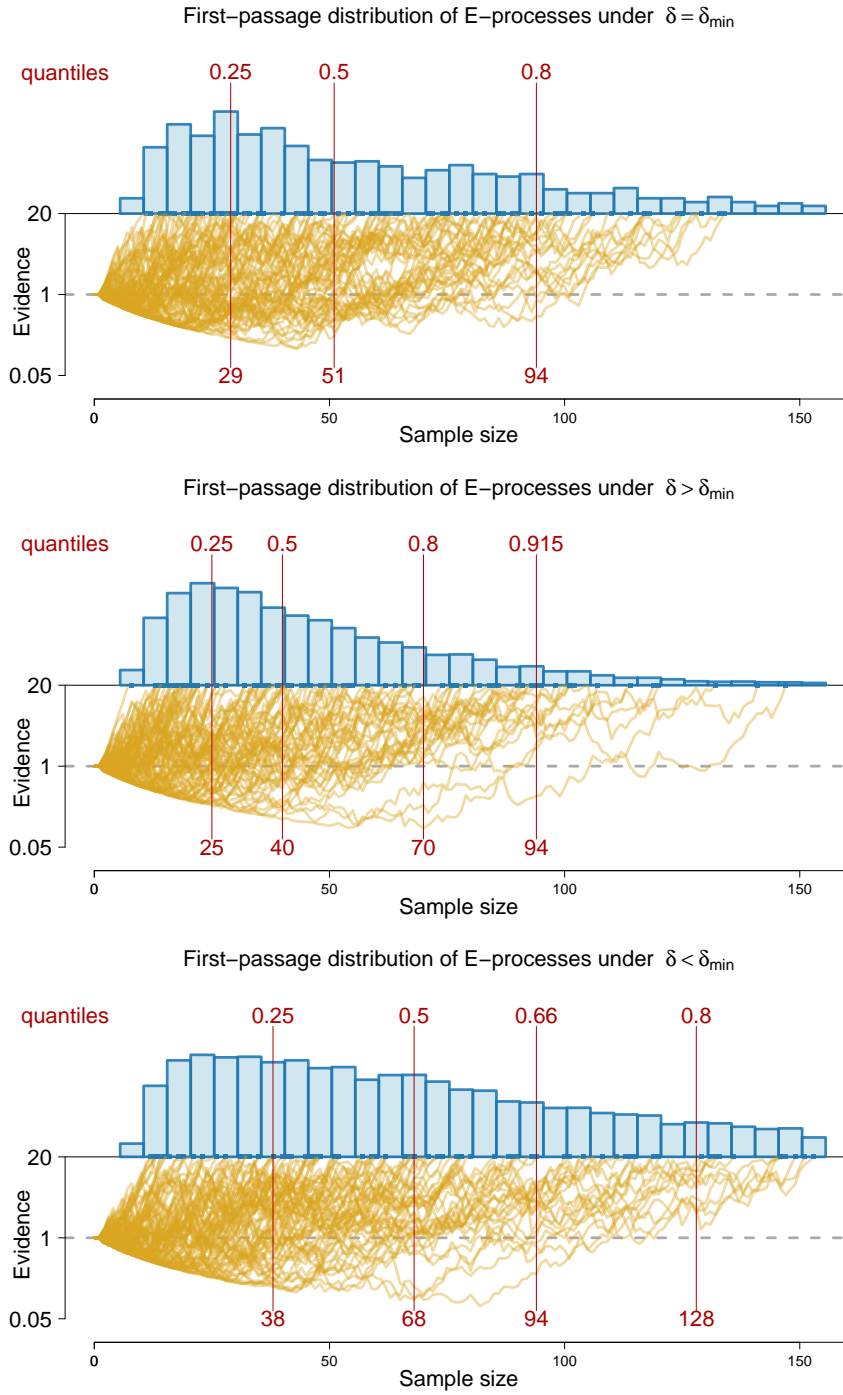
3.3 The Behaviour of n_{plan} Under Other Alternatives

Monitoring the test is even more beneficial when the data-governing effect size is larger than the expected or the minimal clinically relevant δ_{\min} . The middle panel of Fig. 7 shows that under \mathbb{P}_{δ} 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 the experiment is stopped is then $n_1 = n_2 = 46$. This gain in efficiency goes unnoticed for a classical *p*-value test that for type I error control should always be performed at $n_1 = n_2 = 60$.

On the other hand, if the data-governing effect size is smaller than expected or minimal clinically relevant, then more samples are needed to detect this smaller effect with 80% power. The bottom panel of Fig. 7 shows that under \mathbb{P}_{δ} with $\delta = 0.45$ about 66% of the sample paths resulted in a null rejection at $n_1 = n_2 = 94$.

Hence, although the n_{plan} was tuned to the case $\delta = \delta_{\min}$, it also suffices for all data-governing δ larger than δ_{\min} , as one would intuit for a minimal clinically relevant

⁷The first passage time of a sample path that did not (yet) cross the evidence threshold $1/\alpha$ is set to ∞ , as is customary.

**Figure 7**

Regardless of the actual value of the data-governing δ , it can (eventually) be detected with an anytime-valid test. From top to bottom: Distribution of the first-passage distribution when data-governing $\delta = \delta_{\min} = 0.5176$, $\delta = 0.6 > \delta_{\min}$, and $\delta = 0.45 < \delta_{\min}$, respectively.

effect size. Upon reaching n_{plan} without the e -value sample path passing $1/\alpha$, 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 no more than β . Alternatively, we can also continue sampling if the e -value looks promising and we are keen to investigate a smaller effect. The same conclusions can be reached by exploring the anytime-valid confidence intervals. The role of the planned sample size is to guarantee at least $1 - \beta$ power to detect $|\delta| \geq |\delta_{\min}|$ under continuous monitoring. It plays no role in controlling the type I error. If there truly is no effect, we cannot nudge the e -value over the threshold with more than α chance, as the e -value sample path is then expected to move towards zero, see the left panel of Fig. 5. On the other hand, if there is an effect, the (mom) E -process will eventually detect the effect by simply continuing sampling, as the e -value sample path is then expected to increase (see the right panel of Fig. 5). This ability to optionally continue, without breaking type I error control, is, as far as we know, unique to e -value based inference, and forms the basis for flexible sequential learning.

3.4 Alternative Sampling Design Scenarios

There are circumstances where the available budget forms the bottleneck of our investigation, yielding a restriction on the sample sizes, say, at most $n_1 = n_2 = 40$, just enough to be included in the Many Labs 2 project. Before running Protocol 2.1 with $n_{\text{plan}} = n_1 = n_2 = 40$, we might want to get an indication of the effect sizes that we can detect with, say, 80% power. By providing the `designSaviT` function with $\alpha, n_{\text{plan}}, \beta$, e.g. Lines 4 and 5 of R Code 3.1, we see that the smallest effect that we can detect, by continuous monitoring up to $n_{\text{plan}} = n_1 = n_2 = 40$ with 80% chance 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 found if we sample up to $n_1 = n_2 = 40$, but with less chance. Providing `designSaviT` with $\alpha, n_{\text{plan}}, \delta_{\min}$, e.g. Lines 6 and 7 of R Code 3.1, we see that under \mathbb{P}_{δ} with $\delta = 0.7$ we have 64.2 % power to reject the null 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.

Lastly, if there is no prior information regarding δ_{\min} or the permissible type II error rate β , we can run `designSaviT` with only α, n_{plan} , e.g. Line 8 of R Code 3.1. The underlying code then finds the parameter value that *at the specified* n_{plan} has the narrowest confidence interval. For the Many Labs 2 project with $n_{\text{plan}} = n_1 = n_2 = 40$ this yielded $g_{\text{mom}} = 0.134$. 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 and attained by $g_{\text{mom}} = 0.134$, if $s^2 = 1$, whereas it has width 0.85 at $n_1 = n_2 = 100$. Running Line 8 of R Code 3.1 with $n_{\text{plan}} = n_1 = n_2 = 100$ shows that the narrowest mom interval is attained by $g_{\text{mom}} = 0.05193$ leading a to width of 0.82.

In all cases, the design function optimises the requested E -process, which can then be used for subsequent inference as demonstrated in R Code 1.1. The key point is that a sampling strategy for an experiment based on anytime-valid t -tests can be derived with just a few lines of code. Importantly, the planned sample sizes are only an indication, not a

commitment, nor a promise of the number of samples that have 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 specific value of the data-governing δ , the evidence at hand, as quantified by the e -value, will safely guide us in adjusting the experiment during the data collection process.

4 Not All Bayes Factors Are E -Processes

The next two sections may be omitted during the initial reading, as they explore the more nuanced differences between E -processes, Bayes factors, and the tests they form. By definition, a Bayes factor is a ratio of marginal/averaged likelihoods (e.g. Jeffreys, 1961, Ly et al., 2016a, 2016b), whereas the E -processes discussed so far are mixtures of T -likelihood ratios. In other words, the order of marginalising/mixing and taking ratios matters.

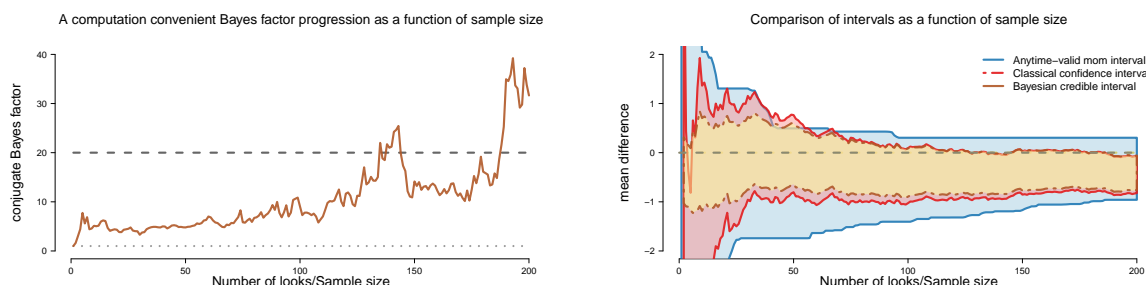
Frequentist principles such as power and type I error control are the main motivation for this work. It is therefore surprising that ensuring these frequentist principles over time led us to E -processes that were previously derived as Bayes factors, as remarked in Section 2.3. The suggestion to employ these Bayes factors (e.g. Jeffreys, 1961, Johnson and Rossell, 2010, Rouder et al., 2009) for inference is not new, but the realisation that these specific Bayes factors are also E -processes is novel. This knowledge allows us to take the best of the two typically competing philosophies of inference.

When E -processes are Bayes factors they have a likelihood interpretation. This allows us to reason from the particular observations to the unobserved general population of interest.⁸ For instance, an e -value of $e = 7$ can then be interpreted as the data being 7 times more likely under the alternative compared to the null, whereas $e = 1/3$ can be interpreted as the data being 3 times more likely under the null compared to the alternative. For E -processes that are not Bayes factors this interpretation holds approximately.

When Bayes factors are E -processes it is guaranteed that the likelihood interpretation is correct with high chance. For instance, under the null there is no more than α chance, say, 1%, to ever observe the alternative being more than $1/\alpha$, say, 100, times more likely compared to the null. This (i.e. the Bayes factor being an E -process) holds whenever the Bayes factor is constructed with so-called Haar priors on the nuisance parameters (Pérez-Ortiz et al., 2024), which is true for the above-mentioned `grow`, `eGauss`, `mom`, and `eCauchy` E -processes.

That said, not all Bayes factors are E -processes. As such, we cannot assume type I error control by simply taking *any* Bayes factor BF_{10} in favour of the alternative over the null and reject the null hypothesis as soon as it crosses the threshold $1/\alpha$. In Appendix D we constructed a (non-default) computationally convenient two-sample Bayes factor t -test with 8 tuning parameters based on priors that do not take the structure of the problem into account. 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 this Bayes factor crosses $1/\alpha = 20$ becomes unreliable; see the dashed brown FPR curve in the left panel of Fig. 2. This reflects a point made earlier by De Heide and Grünwald, 2021 that there can be issues with Bayes factors under optional stopping. The right panel of Fig. 2 shows

⁸As opposed to the notion of chance, which we used to reason from the general population about the relative frequencies of potential realisations.

**Figure 8**

Left panel: The (non-default) computationally convenient Bayes factor (Appendix D) 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.

that the coverage rate of the associated 95% credible interval will also drop well below the nominal level of 95%. Extending this graph further ultimately leads to a coverage rate of 0%. Hence, provided that we sample long enough, we will certainly draw incorrect conclusions if we track the 95% credible interval and reject the null as soon as zero is outside the interval. Fig. 8 shows the typical evolution of the chosen computationally convenient Bayes factor BF_{10} for the same data (under the null) that were used for Fig. 4. The right panel of Fig. 8 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% chance during data collection or at a possibly data-driven stopped time. This problem cannot be solved by choosing different priors, as typical priors yield credible intervals that (relatively quickly) converge 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 anytime-valid test. This is, thus, a completely different procedure yielding wider intervals as shown in Fig. 8. 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% chance 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 et al., 2024).

We introduced the example in this section 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. 8 in this case is not an E -process because it violates Property (iii) (Eq. (6)) by having conditional expectations larger than one for at least one data generating distribution from the null model. Hence, Bayes factors are not necessarily

E -processes, and it was already mentioned that not all E -processes are Bayes factors (e.g. Pandeva et al., 2024, Ramdas et al., 2023 and Turner et al., 2024).

4.1 Anytime-Valid Tests Are More Flexible Compared to Sequential Bayes Factor Tests

Even when a Bayes factor is an E -process, it is used differently in forming a sequential Bayes factor test compared to how it is used in an anytime-valid test. Sequential Bayes factor tests (e.g. Schönbrodt et al., 2017, Schnuerch et al., 2022, Pramanik and Johnson, 2022) follow Protocol 4.1. This procedure does not come with an n_{plan} at which the detec-

```

1 # Pseudo code: This code does NOT run
2
3 n ← 1
4 bf10 ← 1
5
6 while (B < bf10 && bf10 < A) {
7   bf10 ← computeBayesFactor(x[1:n], y[1:n], somePrior)
8
9   if (bf10 >= A) {
10     "Reject the null and accept the alternative"
11     stop()
12   } else if (bf10 <= B) {
13     "Reject the alternative and accept the null"
14     stop()
15   } else {
16     "Increase sample size and test again
17       at the start of the while loop"
18     n ← n + 1
19   }
20 }
```

Protocol 4.1: The sequential Bayes factor testing procedure in pseudo R code without a maximum sample size.

tion of an effect $|\delta| \geq \delta_{\min}$ with at least $1 - \beta$ power is guaranteed. Crucially, it also has an additional evidence boundary for accepting the null. The current discussion in sequential Bayes factor testing revolves around the selection of the boundaries A and B , thus, the specific stopping rule, and which Bayes factor to use. Because sequential Bayes factor tests do not automatically come with explicit type I and type II error guarantees, they have to be estimated in extensive simulations for particularly chosen Bayes factors with specifically chosen tuning and data-governing parameters.

For instance, 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 recommendation was based on the observation that these particular choices resulted in realised type I and II error rates of 4.7% of 4.6%, respectively, in a large-scale simulation study when the data generating δ was set equal to 0.5.

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. Bayes factors that are not E -processes, on the other hand, can yield unfavourable results. For instance, applying the Wald boundaries with $A = \frac{1-\beta}{\alpha} = 16$ and $B = \frac{\beta}{1-\alpha} = 0.21$ to the computationally convenient Bayes factor of Appendix D results in inflated type I error rates. For the same data used in the simulations above, i.e. Fig. 2, the computationally convenient Bayes factor results in 16.8% correct null acceptance, but 24.9% false positive rejections, and no conclusion was reached for the remaining 58.2% due to the Bayes factor sample paths staying between both the lower $B = \frac{\beta}{1-\alpha}$ and upper $A = \frac{1-\beta}{\alpha}$ boundaries until $n = 200$. The latter inconclusive category can lead to this procedure requiring more data points than the savi testing procedure.

Note the difference in order of approach between anytime-valid tests and the sequential Bayes factor tests: For e -value tests, the type I error guarantee comes from Ville’s inequality, which holds for *all* stopping times and, therefore, also *any* stopping time used in simulations. For sequential Bayes factor tests, the protocol comes first. This means that the estimated type I and type II errors depend on the *specific* stopping rule, i.e. halting as soon as $\text{BF}_{10} \geq A$ or $\text{BF}_{10} \leq B$, such as $A = 16$ and $B = 0.21$ for $\alpha = 0.05$ and $\beta = 0.2$. Violating the stopping rule might therefore invalidate the estimated type I and type II error. 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 and Wagenmakers, 2013). Continuing sampling after hitting the lower boundary makes the sequential Bayes factor Protocol 4.1 equivalent to the anytime-valid testing Protocol 2.1 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). Assuming it 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 chance of 1.25% can be absorbed by the lower boundary. E -processes are robust to these types of adaptations, since Ville’s inequality provides type I error protection for all stopping times simultaneously, making anytime-valid tests more flexible in general.

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. Moreover, Schnuerch et al. (2022) notes that there is no guarantee that a sequential Bayes factor test will terminate at, or before reaching a certain sample size. 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 $1 - \beta$ chance to detect effects $|\delta| \geq \delta_{\min}$. This n_{plan} , in turn, is found by relatively quick simulations for typical values of α, β and δ_{\min} .

5 Practical Guidance for Selecting Amongst Various Types of E -Processes

In Section 2.3 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 may guide the choice of the E -process type.

5.1 Default Choice: The mom E -Process

We recommend the mom E -process (Johnson & Rossell, 2010) as a default procedure, because the resulting anytime-valid test appears to provide the best balance between efficiency and robustness to the choice of δ_{\min} . It also yields reasonable anytime-valid confidence intervals (see the blue curves in Fig. 10). Table 2 shows the planned sample sizes for various types of E -processes for $\alpha = 0.05$ and $\beta = 0.2$ in the two-sample t -test scenario with $n = n_1 = n_2$ for expected or minimal clinically relevant $\delta_{\min} = 0.5$ and $\delta_{\min} = 0.2$, which Cohen (1988) interprets as a medium and a small effect size, respectively. The ordering of

Table 2

Planned and expected sample sizes for different E -processes under different data-generating effect sizes δ . The mom E -process provides the best balance between efficiency and robustness to the choice of δ_{\min} . Except for the third and the last column with n_{mean} above them, the table shows the planned sample sizes $n_1 = n_2$ 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 $\delta \geq \delta_{\min}$, but requires an enormous number of samples to detect effects smaller than δ_{\min} with $1 - \beta$ chance. 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 $\delta \geq \delta_{\min}$. 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	$\delta_{\min} = 0.5$					$\delta_{\min} = 0.2$			
	True	n_{plan}	$\delta = \delta_{\min}$	n_{mean}	$\delta = 0.8 > \delta_{\min}$	$\delta = 0.2 < \delta_{\min}$	n_{plan}	$\delta = \delta_{\min}$	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

the types of E -processes in terms of the lowest n_{plan} remains the same for various values of $\alpha, \beta, \delta_{\min}$, see Fig. 9. For a fair comparison between the four types of E -processes, we chose the fastest evidence accumulating E -process for each type as mentioned in Section 2.3. The third and last columns also show the average sample sizes. For instance, under $\delta = \delta_{\min} = 0.5$,

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 in both groups. For the same α, β and δ_{\min} , the classical p -value test should always be performed at $n_1 = n_2 = 64$. Under $\delta = \hat{\delta}_{\min} = 0.2$, we get average sample sizes of $n_{\text{mean}} = n_1 = n_2 = 323$ 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, see Appendix C for more and Grünwald et al. (2024) for full details. However, in case the **grow** t -test was optimised for $\delta_{\min} = 0.5$, but the data-governing effect size is $\delta = 0.2$, then the **grow** E -process requires many more samples to detect the effect with $1 - \beta = 80\%$ power. If the **grow** E -process were tuned to $\hat{\delta}_{\min} = 0.2$ 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. 9.

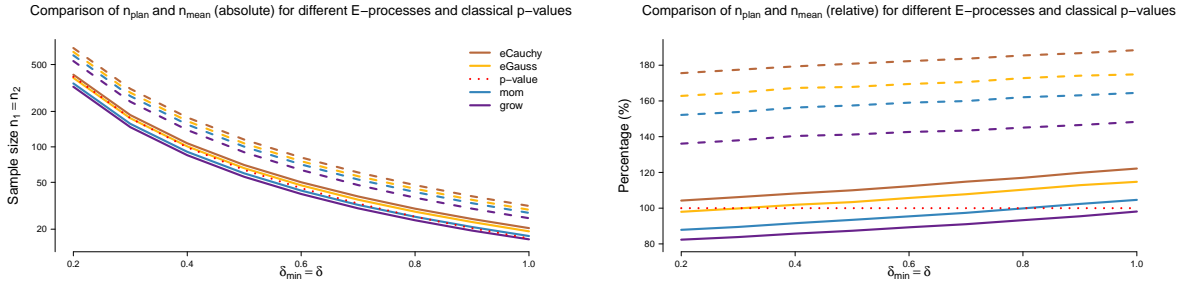


Figure 9

To detect an effect with 80% power, the anytime-valid tests require more samples in the worst-case compared to the classical fixed sample p -value test when the minimal clinical relevant effect size δ_{\min} equals the data-governing δ . On average, however, the **grow** and **mom** E -process tests outperform the classical tests, and the gain is higher for smaller data-governing effect sizes. Left panel: The dotted red line represents the number of samples needed for the classical p -value test to reject the null with 80% under the effect size shown on the x -axis. The four dashed lines at the top represent the worst-case $n_{\text{plan},1} = n_{\text{plan},2}$ of **eCauchy** (brown), **eGauss** (yellow), **mom** (blue) and **grow** (purple). The solid version of these lines represent the average sample sizes. Right panel: The same information as in the left plot is shown, but scaled so the sample sizes of the classical p -value test are set to 100% representing the baseline. Roughly speaking, the worst-case additional data in the 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) still 12% and for **eGauss** (yellow) 6% more data on average. For **mom** (blue) we require 4%, and for **grow** (purple) 10% less data.

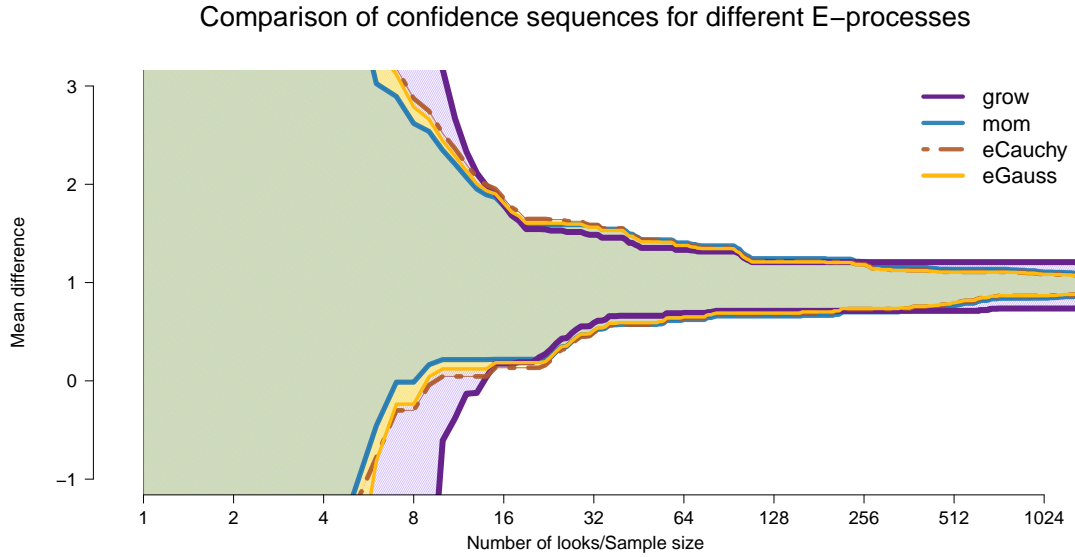


Figure 10

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 red, and **eGauss** in yellow. The data were generated with a true mean difference of $\varphi = 1$.

5.2 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 the \mathcal{M}_0 of interest. Table 2 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. Hence, 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 not be realistic.

5.3 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, and are therefore computationally stable, and easily analysed (Wang & Ramdas, in press). Furthermore, 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. 10 shows the confidence intervals associated with **grow** in purple, **mom** in blue, **eCauchy** in red, 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 $n = 64$.

5.4 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 informative. For the case at hand, overwhelmingly informative data correspond to observing a non-zero sample mean difference without any sampling variability, see Gronau et al. (2020), 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 2 shows that choosing δ_{\min} close to the data generating δ can result in fast detection of $|\delta| \geq \delta_{\min}$, but a relatively harsh penalty in terms of n_{plan} whenever $|\delta| < \delta_{\min}$. 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. The increase in n_{plan} for $|\delta| < \delta_{\min}$ should not pose a problem if δ_{\min} truly represents the minimal clinically relevant effect size. In the ideal situation δ_{\min} is given by the context of the inference problem at hand, perhaps guided by an original finding that we aim to replicate, a meta-analysis, or by conventions in the field such as those posed by Cohen (1988).

6 The Two Real-World Examples Revisited With an e -Value Design

With a specified design object at hand, we revisit the examples discussed at the beginning of this manuscript. The first revised example shows how we can trade off β in favour of further resource conservation. The second revised example shows how, in addition to resource conservation, we get more interpretable results.

6.1 Example 1: Moral Typecasting (Gray & Wegner, 2009, Study 1a) Revisited

A reasonable estimate of the underlying standardised effect size based on the original findings (Gray & Wegner, 2009, Study 1a) is within $(0.769, 0.872)$.⁹ We err on the side of caution by using the lower bound for our power analyses. 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 $\delta_{\min} = 0.769$ and $1 - \beta = 80\%$ power, the one-sided p -value test should be performed after we gathered data from $n_1 = n_2 = 22$ participants in each group.

Lines 1 to 3 of R Code 6.1 show that we should plan for $n_{\text{plan}} = n_1 = n_2 = 37$ participants in each group to detect a data-governing δ of at least $\delta_{\min} = 0.769$ with 80%

⁹The 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.

```

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 6.1: R Code for designing and visualising anytime-valid tests. All the code needed to design, perform and visualise an anytime-valid test. The design object is created on Line 2 and 3. Line 4 yields a plot similar to Fig. 7. The anytime-valid test is performed on Line 6 for data vectors x and y . Line 7 plots the e -value sample path, and Line 8 illustrates the anytime-valid confidence interval as a function of the sample size similar to Fig. 1.

chance.¹⁰ 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 $\delta = \delta_{\min}$.

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.¹¹ By reducing the sample sizes compared to the analysis in Section 1.1, 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 $n_1 = 1315$ of $n_2 = 1324$ participants, thus, on average $n_1 = 21.56$ and $n_2 = 21.71$ in each replication attempt. In other words, 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 null rejections out of a total of 61 trials (83.6%). These conclusions were based on a total of $n_1 = 1031$ and $n_2 = 1062$ participants, which corresponds to $n_1 = 16.90$ and $n_2 = 17.41$ per each replication attempt on average. As a result, the designed e -value test reduced the number of participants needed by 2997 and 2912 in the respective groups, leading to significant resource savings.

This demonstration highlights the benefits of planned analyses in general, and efficiency of e -value based tests in particular. The general conclusions remain the same if we tuned the E -process to the upper bound for δ_{\min} instead. Hence, the precise specification of δ_{\min} is not necessary to gain efficiency and conserve resources.

6.2 Example 2: The Macbeth Effect — Moral Violations and Desire for Cleansing Zhong and Liljenquist (2006, Study 2) Revisited

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,

¹⁰For a two-sided e -value test we require $n_{\text{plan}} = 45$ and the procedure then stops after $n_{\text{mean}} = 28$ in each group on average.

¹¹For 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 the case for the “tanzaniaon” data set, which had only $n_1 = 3$ and $n_2 = 13$ valid responses.

we err on the side of caution by using the lower bound for our design. A classical power analysis shows that for $\delta_{\min} = 0.909$ 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 6.1 with `deltaMin <- (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 detect a data governing δ of at least $\delta_{\min} = 0.909$ with 80% chance. If $\delta = \delta_{\min}$ the procedure is expected to stop after testing $n_{\text{mean}} = 17$ participants in each group.¹² For most replication attempts the one-sided e -value test hit n_{plan} . We get one null rejection 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 δ_{\min} , then we have provided such an effect size ample chance to reject the null. Since this did not occur for most replication attempts, we not only conclude that there was not enough evidence to reject the null, but we can also infer that the postulate $\delta > \delta_{\min}$ is unlikely with high chance. 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 in the analysis of Section 1.2.

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.

7 Summary and Concluding Remarks

Determining the appropriate sample size at which a classical p -value test or confidence interval should be computed is difficult. Especially, before experimentation when no data are present. The fact that classical p -value tests and confidence intervals should be 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 3.1, to construct a design object that derives a *non-binding* planned sample size n_{plan} based on an optimal, say, `mom` type t -test E -process. Analogous code can be used to construct design objects for anytime-valid z -tests (i.e. `designSaviZ()`), anytime-valid tests for two proportions (Turner et al., 2024) (i.e. `designSaviTwoProportions()`), the anytime-valid logrank test (Ter Schure et al., 2024) (i.e. `designSaviLogrank()`), and many more are scheduled to be implemented into the `safestats` package (Ly et al., 2024).

Simulations with the aggressive first-passage time illustrated that type I error control is maintained, despite flexible use of the optimal E -process, whereas simulations under the alternative showed an increase in power compared to classical methods on average. This increase in power translates into smaller expected sample sizes at which we can conclude experiments, which further emphasises the non-binding nature of n_{plan} . The ability to

¹²For 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.

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 a particular set of replication attempts from the Many Labs 2 project the use of the optimal E -process allows the number of participants to be reduced by more than 70% (Section 6.1), whereas a non-optimal E -process already led to a reduction of more than 60% (Section 1.1). These conclusions should not be viewed as criticism on the Many Labs 2 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 yielding narrower anytime-valid confidence intervals, thus, more precise estimates of the effects of interest. Furthermore, it must be noted that anytime-validity does not mean that the E -processes are robust to hypothesising after the results have become available. Reliable science with e -values, thus, still requires us to pre-register the hypotheses (but not the sampling plan) of our confirmatory analyses up front.

Anytime-valid tests overcome the classical trade-off between flexibility in data collection and type I error control forced onto researchers by classical methods. Whereas type I error control cannot be guaranteed when p -values are monitored, e -values can be tracked continuously throughout data collection. The restrictions imposed by classical methods are unhelpful and, in our opinion, can frustrate researchers in their pursuit of uncovering truths about the world. With e -values, our aim is to simplify statistical practice, not to complicate it. Achieving science that is both reproducible and generalisable requires effort, but this effort should not be devoted to futilely maintaining an outdated predetermined sampling plan method, especially since e -values eliminate the need for it.

If monitoring of the test is genuinely impossible and the test is destined for a singular execution, sequential methods may lose their efficacy compared to classical p -value methods, as was seen in Section 3.2. The phenomenon we observed there for the t -test holds for many other models (e.g. the logrank test (Ter Schure et al., 2024) or contingency tables (Turner et al., 2024)) as well: to obtain a desired power, one needs less data than in a classical test *on average*, but one needs more *in the worst-case* (namely, n_{plan}). Nonetheless, we strongly advocate for the use of E -process methods in individual studies as well, given their flexibility. The case for e -value methods is even stronger in reproducibility studies such as Many Labs 2, where studies are repeated many times. Then the *Law of Large Numbers* kicks in and the average, rather than the worst-case, will determine what happens: One can be almost sure that the total amount of data needed over the studies is substantially less than with classical tests. The same holds for meta-analyses, for which e -values are particularly useful (Ter Schure et al., 2022): one can define a meta-analytical e -value that safely combines the evidence as the data from multiple studies accrue, ensuring type I error control even when a study triggers replication attempts, regardless of the sample sizes within each study. But this aspect is beyond the scope of this paper: While a more detailed exploration of e -values in meta-analysis is currently underway, here, we concentrated on E -processes for individual studies, to which they already bring ample advantages.

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.). Chapman; Hall/CRC.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd 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.
- Earp, B. D., Everett, J. A., Madva, E. N., & Hamlin, J. K. (2014). Out, damned spot: Can the “Macbeth Effect” be replicated? *Basic and Applied Social Psychology*, 36(1), 91–98.
- 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.
- 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., & Koolen, W. M. (2025). Supermartingales for one-sided tests: Sufficient monotone likelihood ratios are sufficient. *arXiv preprint arXiv:2502.04208*.
- Grünwald, P. D., De Heide, R., & Koolen, W. (2024). Safe testing [With discussion]. *Journal of the Royal Statistical Society. Series B (Methodological)*, 86(5), 1091–1128.
- 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. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2), 1055–1080.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.
- Johari, R., Koomen, P., Pekelis, L., & Walsh, D. (2022). Always valid inference: Continuous monitoring of *a/b* tests. *Operations Research*, 70(3), 1806–1821.
- 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., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Nosek, B., et al. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443–490.
- Larsson, M., Ramdas, A., & Ruf, J. (2025). The numeraire *e*-variable and reverse information projection [in press]. *Annals of Statistics*.

- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- 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. <https://doi.org/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* [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-evaluating classifier two-sample tests. *Transactions on Machine Learning Research*. <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. <https://doi.org/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.
- Ramdas, A., Ruf, J., Larsson, M., & Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*.
- Ramdas, A., & Wang, R. (2024). Hypothesis testing with e-values. *arXiv preprint arXiv:2410.23614*.
- 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. <https://doi.org/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.
- 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.
- Ter Schure, J. (2023, January). *ALL-IN meta-analysis* [Doctoral dissertation, Leiden University].
- Ter Schure, J., Ly, A., Belin, L., Benn, C. S., Bonten, M. J., Cirillo, J. D., Damen, J. A., Fronteira, I., Hendriks, K. D., Junqueira-Kipnis, A. P., Kipnis, A., Launay, O., Mendez-Reyes, J. E., Moldvay, J., Netea, M. G., Nielsen, S., Upton, C. M., van den Hoogen, G., Weehuizen, J. M., . . . van Werkhoven, C. (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. <https://doi.org/https://doi.org/10.1016/j.jspi.2023.106116>
- Wang, H., & Ramdas, A. (in press). Anytime-valid t-tests and confidence sequences for Gaussian means with unknown variance. *Sequential Analysis*.
- 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.
- Zhong, C.-B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, 313(5792), 1451–1452.

Appendix A

Likelihood Ratios Are E -Processes

To verify Property (iii) for likelihood ratios, we exploit the fact that the null model is simple, the law of total probability, and that a realised stopping time of $N = m$ corresponds to outcomes $x^{(m)}$ belonging to some (measurable) set B_m . Firstly, we can write

$$\text{LR}_N := \sum_{m=1}^{\infty} \mathbf{1}\{N = m\} \text{LR}_m, \quad (\text{A1})$$

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 $\mathbb{E}_{\mathbb{P}_0}[\text{LR}_N]$ 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 , that is,

$$\mathbb{E}_{\mathbb{P}_0}[\text{LR}_N] = \mathbb{E}_{\mathbb{P}_0} \left[\mathbb{E}_{\mathbb{P}_0} \left[\sum_{m=1}^{\infty} \mathbf{1}\{N = m\} \text{LR}_m \mid \{N = m\} \right] \right] \quad (\text{A2})$$

$$= \mathbb{E}_{\mathbb{P}_0} \left[\sum_{m=1}^{\infty} \mathbf{1}\{N = m\} \mathbb{E}_{\mathbb{P}_0}[\text{LR}_m \mid \{N = m\}] \right] \quad (\text{A3})$$

$$= \sum_{m=1}^{\infty} \mathbb{E}_{\mathbb{P}_0} \left[\mathbf{1}\{N = m\} \mathbb{E}_{\mathbb{P}_0}[\text{LR}_m \mid \{N = m\}] \right]. \quad (\text{A4})$$

For Eq. (A3) we took $\mathbf{1}\{N = m\}$ out of the inner integral, because it is a function of the conditioning event $\{N = m\}$. For Eq. (A4) we use the fact that we can swap sums and expectations of non-negative functions.

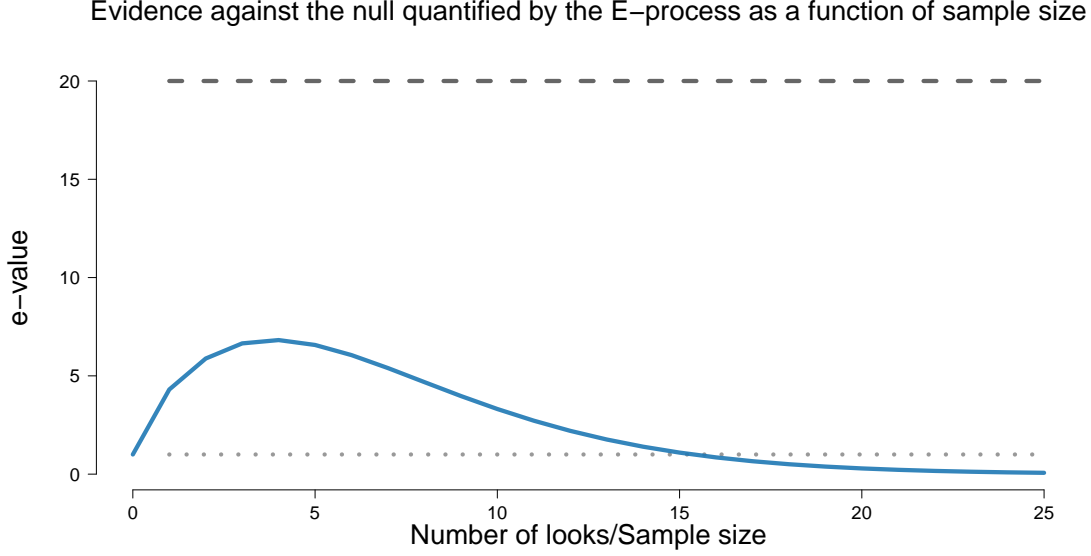
As before, the expectation $\mathbb{E}_{\mathbb{P}_0}[\text{LR}_m \mid \{N = m\}]$ defines an integral. However, by conditioning on $N = m$ the possible outcomes of length m becomes restricted. As an example, we consider the z -test with variance known to be one, thus, $z := \sqrt{n}\bar{x}_n$, where \bar{x}_n is the sample mean of n data points. As an example stopping time we take the first time N at which $p < 0.05$, which corresponds to observing a z -score smaller than -1.96 or larger than 1.96 , or equivalently, $|\bar{x}_n| > \frac{1.96}{\sqrt{n}}$. The event $N = m$ is equivalent to data sequences $x^{(m)} = (x_1, \dots, x_m)$ with sample means $|\bar{x}_j| < \frac{1.96}{\sqrt{j}}$ for $j = 1, \dots, m-1$, 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 . We refer to all outcomes corresponding to this restriction as the event B_m . In general, we can identify $\{N = m\}$ with an event B_m .

Continuing from Eq. (A4), and exploiting the fact that LR_m is a ratio of densities that can be factorised via conditioning, we have that

$$\mathbb{E}_{\mathbb{P}_0}[\text{LR}_N] = \sum_{m=1}^{\infty} \mathbb{E}_{\mathbb{P}_0} \left[\mathbf{1}\{B_m\} \mathbb{E}_{\mathbb{P}}[\text{LR}_m \mid B_m] \right] \quad (\text{A5})$$

$$= \sum_{m=1}^{\infty} \mathbb{E}_{\mathbb{P}_0} \left[\mathbf{1}\{B_m\} \int_{B_m} \frac{q(x^{(m)} \mid B_m)}{p_0(x^{(m)} \mid B_m)} \frac{\mathbb{Q}(B_m)}{\mathbb{P}_0(B_m)} p_0(x^{(m)} \mid B_m) dx^{(m)} \right] \quad (\text{A6})$$

$$= \sum_{m=1}^{\infty} \mathbb{E}_{\mathbb{P}_0} \left[\mathbf{1}\{B_m\} \frac{\mathbb{Q}(B_m)}{\mathbb{P}_0(B_m)} \int_{B_m} q(x^{(m)} \mid B_m) dx^{(m)} \right], \quad (\text{A7})$$

**Figure A1**

Evidence against the null quantified by the E-process as a function of sample size. Independent of the sample size n an observed $z = 1.96$ corresponds to $p < \alpha = 0.05$, thus, a null rejection. On the other hand, the evidence as quantified by the z -likelihood ratio, that is, the e -value, would first increase and then decrease as a function of n .

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,

$$\mathbb{E}_{\mathbb{P}_0}[\text{LR}_N] = \sum_{m=1}^{\infty} \mathbb{E}_{\mathbb{P}_0} \left[\mathbf{1}\{B_m\} \frac{\mathbb{Q}(B_m)}{\mathbb{P}_0(B_m)} \right] = \sum_{m=1}^{\infty} \mathbb{E}_{\mathbb{P}_0} \left[\mathbf{1}\{N = m\} \frac{\mathbb{Q}(N=m)}{\mathbb{P}_0(N=m)} \right] \quad (\text{A8})$$

$$= \sum_{m=1}^{\infty} \mathbb{Q}(N = m) = 1 \leq 1. \quad (\text{A9})$$

The last equality follows from the fact that the chance of N taking on any outcome is one, in particular, under the measure \mathbb{Q} corresponding to the numerator of LR_n .

As an aside, the $p < 0.05$ rule corresponds to the fixed threshold on the z -scale, that is, $|z| > 1.96$. This is completely different than rejecting at a fixed threshold $1/\alpha$ at the likelihood ratio scale, as illustrated in Fig. A1. For those interested in the technicalities, the defining Property (iii) of E -processes also holds for any non-negative supermartingales. For supermartingales, this property 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 B

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. (12) is an E -process. As such, for any prior $\pi(\delta_s)$ the following mixture

also defines an E -process

$$E_{n_\delta, \nu}(t) := \int \frac{T_\nu(t|\sqrt{n_\delta}\delta_s)}{T_\nu(t)} \pi(\delta_s) d\delta_s. \quad (\text{B1})$$

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

$$E_{n_\delta, \nu}(t) = \int e^{-\frac{n_\delta \delta_s^2}{2}} {}_1F_1\left(\frac{\nu+1}{2}; \frac{1}{2}; \frac{t^2}{\nu+t^2} \frac{n_\delta \delta_s^2}{2}\right) \pi(\delta_s) d\delta_s, \quad (\text{B2})$$

where for $|z| < 1$ the confluent hypergeometric function is given by ${}_1F_1(a; c; z) := \sum_{i=0}^{\infty} \frac{(a)_i}{(c)_i} \frac{z^i}{i!}$, 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 $n \in \mathbb{N}$.

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

$$\pi(\delta) = \frac{\delta^{2k} \exp(-\frac{\delta^2}{2g})}{p_k(\emptyset)}, \text{ where } p_k(\emptyset) := \int_{-\infty}^{\infty} \delta^{2k} \exp(-\frac{\delta^2}{2g}) = (2g)^{k+\frac{1}{2}} \Gamma(k+\frac{1}{2}), \quad (\text{B3})$$

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

$$E_{n_\delta, \nu}(t) = (1 + n_\delta g)^{-k-\frac{1}{2}} {}_2F_1\left(\frac{\nu+1}{2}, k+\frac{1}{2}; \frac{1}{2}; \frac{t^2}{\nu+t^2} \frac{n_\delta g}{1+n_\delta g}\right), \quad (\text{B4})$$

where for $|z| < 1$ the Gaussian hypergeometric function is given by ${}_2F_1(a, b; c; z) := \sum_{i=0}^{\infty} \frac{(a)_i (b)_i}{(c)_i} \frac{z^i}{i!}$. An Euler transform and the plugin $k = 1$ then yields the result shown in Box 5.

Appendix C

Growth-Rate Optimal in Worst Case E -Processes

There is a choice between E -processes, because the collection $\mathcal{E}(\mathcal{M}_0)$ of all E -processes for a null model \mathcal{M}_0 is vast. This collection includes *some* specifically constructed Bayes factors, but $\mathcal{E}(\mathcal{M}_0)$ also contains the pathological E -process that always yields the e -value 1, irrespective of the data. This pathological E -process will never commit a type I error, but it will also not reject the null when it is blatantly false. A good E -process should therefore also indicate (large) evidence against the null, whenever it is false. For efficient inference the evidence should grow quickly, as we can then reject the null and conclude the experiment sooner, leading to resource conservation.

To measure the rate of growth, we fix a stopping time N and write $\mathcal{E}_N(\mathcal{M}_0)$ for the collection of E -variables for data samples collected with stopping time N . 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. We prefer $E_N^{s_1}$ over $E_N^{s_2}$, if under the alternative we expect $E_N^{s_1}$ to, say, triple, whereas $E_N^{s_2}$ to only double the evidence against the null at time N . In general, we prefer 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 $\mathbb{E}_{\mathbb{P}_{\delta_1}}[\log(E_N^{s_1})] \geq \mathbb{E}_{\mathbb{P}_{\delta_1}}[\log(E_N^{s_2})]$, where δ_1 is the

data-governing effect size. If data are indeed generated under \mathbb{P}_{δ_1} , then the most preferred E_N from $\mathcal{E}_N(\mathcal{M}_0)$ is the one that achieves the maximum¹³

$$\max_{E_N \in \mathcal{E}_N(\mathcal{M}_0)} \mathbb{E}_{\mathbb{P}_{\delta_1}} [\log(E_N)]. \quad (\text{C1})$$

The E -variable that attains this maximum is referred to as the growth-rate optimal E -variable for \mathbb{P}_{δ_1} .

Now we return to E -processes and we fix an alternative δ_1 . For every stopping time N , any E -process E gives us an E -variable E_N . In general, there do not always exist E -processes E such that, for arbitrary stopping time N , we have that E_N is the growth-rate optimal E -variable relative to the alternative δ_1 . But for certain regular models, this is possible after all: In that case, we have a growth-rate optimal E -process denoted by E^{δ_1} . Further regularity conditions show that E^{δ_1} yields the lowest possible expected stopping time under the alternative \mathbb{P}_{δ_1} (Ter Schure et al., 2024), if we stop at the first n for which $E_n^{\delta_1} \geq 1/\alpha$.

Due to the lack of knowledge regarding \mathbb{P}_{δ_1} we cannot specify the growth-rate optimal E -variable. A possible workaround involves a minimal clinically relevant effect size δ_{\min} and the adoption of a conservative approach. The so-called **grow**, that is, growth-rate optimal in worst case, E -variable then solves the following maximin problem

$$\max_{E_N \in \mathcal{E}_N(\mathcal{M}_0)} \min_{|\delta_1| \geq \delta_{\min}} \mathbb{E}_{\mathbb{P}_{\delta_1}} [\log(E_N)]. \quad (\text{C2})$$

As before, the maximum is over the collection $\mathcal{E}_N(\mathcal{M}_0)$ of all E -variables for data samples collected with stopping time N , and the minimum (the worst-case part) is over all data-generating distributions \mathbb{P}_{δ_1} with $|\delta_1| \geq |\delta_{\min}|$. For models that have the so-called monotone likelihood property, the problem gets easier as the data-governing δ_1 is further away from the null. For these models the inner minimum of Eq. (C2) is attained by $\delta_1 = \delta_{\min}$ and simplifies to the problem Eq. (C1). This is the case for t -tests.

The work of Pérez-Ortiz et al., 2024 implies that for t -tests the optimal two-sided **grow** E -variable depends only on δ_{\min} , but not on n . More specifically, the two-sided **grow** E -process corresponds to the t -likelihood ratio with point priors at $\pm\delta_{\min}$, see Fig. 3.

The optimal **mom** test can be found by solving Eq. (C1) with $\delta_1 = \delta_{\min}$, but with the candidate set of E -variables restricted to only the **mom** E -variables. The optimal solution is asymptotically identified for ν large and by differentiation with respect to g_{mom} , and leads to $g_{\text{mom}} = \delta_{\min}^2/2$ as specified in the main text. For the optimal **eGauss** and **eCauchy** we similarly restrict the search space of candidate E -variables to the respective class.

Appendix D

A Computationally Convenient (Two-Sided) Two-Sample t -Test

To construct a Bayes factor we need to select priors to marginalise out the free parameters. Recall that the alternative model has three free parameters, say, μ_1, μ_2, σ , whereas the null

¹³To ease exposition, we write maximum, but it should in fact be the supremum. Similarly, below we write a minimum, which should actually be an infimum. The differences between the maximum/minimum and supremum/infimum matters substantially, see Grünwald et al. (2024) and Larsson et al. (2025) for further details.

model only has two, say, μ_g, 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

$$\text{BF}_{10}(x^{(n)}) := \frac{\int \int \int f(x^{(n)} | \mu_1, \mu_2, \sigma) \pi_1(\mu_1, \mu_2, \sigma) d\mu_1 d\mu_2 d\sigma}{\int \int f(x^{(n)} | \mu_g, \mu_g, \sigma) \pi_0(\mu_g, \sigma) d\mu_g d\sigma}. \quad (\text{D1})$$

A computational convenient choice is to exploit conjugacy, which states that normal likelihoods/densities $\mathcal{N}(x^{(n)} | \mu, \sigma)$ 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

$$f(x^{(n)} | \mu_1, \mu_2, \sigma) = \mathcal{N}(x^{(n_1)} | \mu_1, \sigma) \mathcal{N}(x^{(n_2)} | \mu_2, \sigma) \quad (\text{D2})$$

$$= (2\pi)^{-\frac{n_+}{2}} \sigma^{-n_+} e^{-\frac{\sum_{k=1}^2 \nu_k s_k^2}{2\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} [n_1(\mu_1 - \bar{x}_1)^2 + n_2(\mu_2 - \bar{x}_2)^2]\right), \quad (\text{D3})$$

where $n_+ = n_1 + n_2$ is the combined 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 $k = 1, 2$. The likelihood under the null is $f(x^{(n)} | \mu_1, \mu_2, \sigma)$ with the means set to the grand mean, i.e. $\mu_1 = \mu_2 = \mu_0$.

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. $\sigma \sim \text{Gamma}^{-\frac{1}{2}}(\alpha_\sigma, \beta_\sigma)$, we have 8 parameters to play with. The resulting Bayes factor is then given by

$$\text{BF}_{10;\eta} := \sqrt{\frac{1+n_+g_0}{(1+n_1g_1)(1+n_2g_2)}} \left(\frac{\frac{n_1n_2}{n_+} (\bar{x}_1 - \bar{x}_2)^2 + \frac{n_+}{1+g_0n_+} (\frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_+} - a_0)^2 + 2\beta_\sigma + \sum_{k=1}^2 \nu_k s_k^2}{\frac{n_1(\bar{x}_1 - a_1)^2}{1+n_1g_1} + \frac{n_2(\bar{x}_2 - a_2)^2}{1+n_2g_2} + 2\beta_\sigma + \sum_{k=1}^2 \nu_k s_k^2} \right)^{\frac{n_+}{2} + \alpha_\sigma}, \quad (\text{D4})$$

where $\eta = (a_1, g_1, a_2, g_2, a_0, g_0, \alpha_\sigma, \beta_\sigma)$ collects the tuning parameters. For the example 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.