ORIGINAL ARTICLE

Discussion Section

Safe Testing

[Read before The Royal Statistical Society at a meeting organized by the Research Section on Wednesday, 24th January, 2024, Professor Robin Evans in the Chair]

Peter Grünwald[1,2] | Rianne de Heide[3] | Wouter Koolen[1,4]

[1]Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands

[2]Leiden University, The Netherlands

[3]Vrije Universiteit Amsterdam, The Netherlands

[4]University of Twente, The Netherlands

Correspondence
Peter Grünwald, Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands
Email: pdg@cwi.nl

Funding information
Does not apply.

Data Availability Statement
This paper only refers to data simulated from computer experiments. All data needed to reproduce the findings in the paper can be generated using the R package `safestats` (**?**) available on CRAN.

We develop the theory of hypothesis testing based on the *e*-value, a notion of evidence that, unlike the *p*-value, allows for effortlessly combining results from several studies in the common scenario where the decision to perform a new study may depend on previous outcomes. Tests based on *e*-values are safe, i.e. they preserve Type-I error guarantees, under such optional continuation. We define growth-rate optimality (GRO) as an analogue of power in an optional continuation context, and we show how to construct GRO *e*-variables for general testing problems with composite null and alternative, emphasizing models with nuisance parameters. GRO *e*-values take the form of Bayes factors with special priors. We illustrate the theory using several classic examples including a one-sample safe *t*-test and the 2×2 contingency table. Sharing Fisherian, Neymanian and Jeffreys-Bayesian interpretations, *e*-values may provide a methodology acceptable to adherents of all three schools.

KEYWORDS
Bayes Factors,E-Values,Hypothesis Testing,Information Projection,Optional Stopping,Test Martingales

1 | INTRODUCTION AND OVERVIEW

We wish to test the veracity of a null hypothesis $\mathcal{H}_0$, often in contrast with some alternative hypothesis $\mathcal{H}_1$, where both $\mathcal{H}_0$ and $\mathcal{H}_1$ represent sets of distributions on some given sample space. Our theory is based on *e*-test statistics. These are simply nonnegative random variables that satisfy the inequality:

$$\text{for all } P \in \mathcal{H}_0\text{: } \mathbf{E}_P[E] \leq 1. \tag{1}$$

We refer to *e*-test statistics as *e*-variables, and to the value they take on a given sample as the *e*-value, emphasizing that they are to be viewed as an alternative to, and in many cases an improvement of, the classical *p*-value. Note that large *e*-values correspond to evidence against the null: for given *e*-variable E and $0 \leq \alpha \leq 1$, we define the threshold test corresponding to E with significance level α, as the test that rejects $\mathcal{H}_0$ iff $E \geq 1/\alpha$. We will see, in a sense to be made more precise, that this test is safe under optional continuation with respect to Type-I error.

Motivation

p-values and standard null hypothesis testing have come under intense scrutiny in recent years (**??**). *e*-variables and safe tests offer several advantages. Most importantly, in contrast to *p*-values, *e*-variables behave excellently under optional continuation, the highly common practice in which the decision to perform additional tests partly depends on the outcome of previous tests. They thus seem particularly promising when used in meta-analysis (**?** provides a first such 'ALL-IN' meta-analysis; see also (**?**)). A second reason is their enhanced interpretability: they have a very concrete (monetary) interpretation as 'evidence against the null' which remains valid even if one dismisses concepts such as 'significance' altogether, as recently advocated by **?**. A third is their flexibility: as we show in this paper, *e*-variables can be based on Bayesian prior knowledge, on earlier data, but also on minimax performance considerations, in all cases preserving frequentist Type I error guarantees.

Overall Contribution and Contents

Although the concept is much older (Section 7), the interest in *e*-values and the related test martingales has exploded over the last four years (**??????**). In this paper, we further develop the theory of *e*-values, by providing general optimality criteria and show how to design *e*-variables that satisfy them. We do this on the basis of four ever more general versions of a single novel theorem, Theorem 1. In its first incarnation, in Section 2, Theorem 1 already tells us that one can design nontrivial, useful *e*-variables for a wide class of testing problems with composite null and alternative. This first instance relies on using a prior W_1 on the alternative $\mathcal{H}_1$. The ensuing *e*-variables, while guaranteeing frequentist Type-I error control, will have a GRO (growth-rate optimality) property under W_1. This GRO *e*-variable will be a Bayes factor with a special prior on the null. More general versions of the theorem allow us to construct *e*-variables when no prior on $\mathcal{H}_1$ is available. These satisfy either a direct worst-case optimality criterion (GROW) or a relative one (REGROW). In our example applications we restrict ourselves to classical testing scenarios such as 1-dimensional exponential families, the 2×2 contingency table, and the t-test. Importantly, the latter two have nuisance parameters and the GRO approach provides a generic methodology for dealing with them. For the t-test setting, GRO *e*-variables turn out to be Bayes factors based on the right Haar prior, as known from objective Bayes analyses (**?**). For the 2×2-setting, GRO *e*-values do not correspond to standard Bayes factors.

We then, in Section 5 and 6, investigate optional continuation, stopping and GRO in more detail, and we assess how competitive the *e*-variables we designed are compared to classical methods in terms of the amount of data needed before a certain desired power or growth rate can be reached. The final three sections put our work in context. We

provide a historical overview of *e*-value related work in Section 7, critically discuss GRO in Section 8, and then, in Section 9, taking a step back, we come to the inescapable conclusion that *e*-variables unify and correct ideas from the three main paradigms of testing: Fisherian, Neyman-Pearsonian and Jeffreysian. But first, in the remainder of this introduction, we explain the three main interpretations of *e*-variables (Section 1.1), we briefly introduce our main theorem (Section 1.2) and, in Section 1.3, we explain the main advantage of *e*-variables over *p*-values in terms of optional continuation. We claim no technical novelty for this part, which mainly restates and reinterprets existing results[1]. We defer to the appendices all longer proofs and details that would distract from the main story.

1.1 | The three main interpretations of *e*-variables

1. First Interpretation: Gambling

The first and foremost interpretation of *e*-variables is in terms of money, or, more precisely, Kelly (?) gambling. Imagine a ticket (contract, gamble, investment) that one can buy for 1\$, and that, after realisation of the data, pays E \$; one may buy several and positive fractional amounts of tickets. (1) says that, if the null hypothesis is true, then one expects not to gain any money by buying such tickets: for any $r \in \mathbb{R}^+$, upon buying r tickets one expects to end up with $r\mathbf{E}[E] \leq r$ \$. Therefore, if the observed value of E is large, say $20 = 1/0.05$, one would have gained a lot of money after all, indicating that something might be wrong about the null.

2. Second Interpretation: Conservative *p*-Value, Type I Error Probability

Recall that a (strict) p-value is a random variable p such that for all $0 \leq \alpha \leq 1$, all $P_0 \in \mathcal{H}_0$,

$$P_0(p \leq \alpha) = \alpha. \tag{2}$$

A conservative p-value is a random variable for which (2) holds with '=' replaced by '$\leq$'. There is a close connection between (small) p- and (large) e-values:

Proposition 1 *For any given e-variable E, define $p_{[e]} := 1/E$. Then $p_{[e]}$ is a conservative p-value. As a consequence, for every e-variable E, any $0 \leq \alpha \leq 1$, the corresponding threshold-based test has Type-I error guarantee α, i.e. for all $P \in \mathcal{H}_0$,*

$$P(E \geq 1/\alpha) \leq \alpha. \tag{3}$$

Proof (of Proposition 1) Markov's inequality gives $P(E \geq 1/\alpha) \leq \alpha \mathbf{E}_P[E] \leq \alpha$.

While reciprocals of e-variables thus give a special type of conservative p-values, reciprocals of standard p-values satisfying (2) are by no means e-variables; if E is an e-variable and p is a standard p-value, and they are calculated on the same data, then we will usually observe $p \ll 1/E$ so with e-values we need more extreme data in order to reject the null (see Section 6 for a nuanced analysis and Section 7 for more on e-p conversions).

Combining 1. and 2.: Optional Continuation

Proposition 2 below shows that multiplying e-variables $E_{(1)}, E_{(2)}, \ldots$ for tests based on respective independent samples $\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}, \ldots$ (with each $\mathbf{Y}_{(j)}$ being the batch of outcomes for the j-th test), gives rise to new e-variables, even if the decision whether or not to perform the test resulting in $E_{(j)}$ was based on the value of earlier test outcomes $E_{(j-1)}, E_{(j-2)}, \ldots$ As a result (Corollary 1), the Type I-Error Guarantee (3) remains valid even under this 'optional

[1] Since the first version of the present paper appeared on arXiv, various subsets of these results have been widely discussed in various recent papers, but we still re-state them here to keep the paper self-contained.

continuation' of testing. Just as importantly, in contrast to p-values, e-variables satisfy an 'optional continuation principle': whether an observed e-value is valid or not does not depend on whether or not you would have performed an additional study and gathered additional evidence in situations that did not occur.

An indication that something like this might be true is immediate from our gambling interpretation: if we start by investing \$1 in $E_{(1)}$ and, after observing $E_{(1)}$, reinvest all our new capital $\$E_{(1)}$ into $E_{(2)}$, then after observing $E_{(2)}$ our new capital will obviously be $\$E_{(1)} \cdot E_{(2)}$, and so on. If, under the null, we do not expect to gain any money for any of the individual gambles $E_{(j)}$, then, intuitively, we should not expect to gain any money under whichever strategy we employ for deciding whether or not to reinvest (just as you would not expect to gain any money in a casino irrespective of your rule for re-investing and/or stopping and going home).

3. Third Interpretation: Bayes Factors

For convenience, from now on we write the models $\mathcal{H}_0$ and $\mathcal{H}_1$ as $\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\}$; $\mathcal{H}_1 = \{P_\theta : \theta \in \Theta_1\}$, where $\Theta_0, \Theta_1 \subset \Theta$, and $\{P_\theta : \theta \in \Theta\}$ represents a general family of distributions or random processes, defined relative to some given sample space and σ-algebra or filtration. $\mathbf{Y} = Y^N = (Y_1, \ldots, Y_N)$, a vector of N outcomes, represents our data. N may be a fixed sample size n but can also be a random stopping time. We assume that, under every P_θ with $\theta \in \Theta$, $\mathbf{Y}$ has a probability density p_θ relative to some fixed underlying measure μ. In the Bayes factor approach to testing, one associates both $\mathcal{H}_j$ with a prior W_j, which is simply a probability distribution on Θ_j, and a Bayes marginal probability distribution P_{W_j}, with density (or mass) function given by

$$p_{w_j}(\mathbf{Y}) := \int_{\Theta_j} p_\theta(\mathbf{Y}) \, \mathrm{d}W_j(\theta). \tag{4}$$

The Bayes factor is then given as:

$$\mathrm{BF} := \frac{p_{w_1}(\mathbf{Y})}{p_{w_0}(\mathbf{Y})}. \tag{5}$$

Whenever $\mathcal{H}_0 = \{P_0\}$ is simple, i.e., a singleton, then the Bayes factor is also a (sharp, i.e. with expectation exactly 1) e-variable, since we must then have that W_0 is degenerate, putting all mass on 0, and $p_{w_0} = p_0$, and then for all $P \in \mathcal{H}_0$, i.e. for P_0, we have, assuming P_0 has strictly positive density,

$$\mathbf{E}_P[\mathrm{BF}] = \int p_0(y) \cdot \frac{p_{w_1}(y)}{p_0(y)} \, \mathrm{d}\mu(y) = 1. \tag{6}$$

For such e-variables that are really simple-$\mathcal{H}_0$-based Bayes factors, Proposition 1 reduces to the well-known universal bound for likelihood ratios (**?**). When $\mathcal{H}_0$ is itself composite, most Bayes factors $\mathrm{BF} = p_{W_1}/p_{W_0}$ will not be e-variables any more, since for BF to be an e-variable we require (6) to hold for all $P_\theta, \theta \in \Theta_0$, whereas in general it only holds for $P = P_{W_0}$. Nevertheless, Theorem 1 (in its first, simplest version in Section 2) implies that, under weak conditions, for every prior W_1 on Θ_1 there always exists a corresponding prior W_0 on Θ_0, for which $\mathrm{BF} = p_{W_1}/p_{W_0}$ is an e-variable after all. More generally, in all our examples e-variables invariably take on a Bayesian form, though sometimes with highly unusual (e.g. degenerate) priors.

1.2 | This Paper: Beyond Simple Nulls, Beyond Available Priors

In this paper, we focus on general, composite $\mathcal{H}_0$. The only assumption on $\mathcal{H}_0$ is the existence of densities as above — we make this assumption because it allows for a completely general characterisation of GRO ('growth-rate optimal') e-variables as in Theorem 1. Still, useful e-variables for nonparametric settings without densities do exist (**??**).

Theorem 1 in its first form in Section 2 tells us how to choose an e-variable that is optimal in the GRO sense if a prior W_1 on $\mathcal{H}_1$ is given. Roughly speaking, GRO means that the e-variable tends to grow fast under $\mathcal{H}_1$ as more data come in, thereby generating substantial evidence against $\mathcal{H}_0$. The generalisations of Section 3–4, extend the GRO idea to e-variables when no such prior is available. Section 3 deals with a basic maximin optimality approach, which is appropriate if there is a single parameter of interest, a minimum relevant effect size, and no nuisance parameter. Section 4 describes relative maximin optimal e-variables, appropriate if there is no minimal effect size and/or nuisance parameters are present. This culminates in the fully general version of Theorem 1 in Section 4.3 which is also applicable to hypotheses with nuisance parameters that satisfy a group invariance, such as in the t-test. To show that the e-variable we propose for the t-test is indeed optimal we need a special case of an additional result, Theorem 4.2 of the paper (**?**), which, for convenience, we restate. But before embarking on these results, we explain the benefits of e-variable based tests in detail.

1.3 | Optional Continuation

We defined e-variables for a single experiment. We now discuss sequential experimentation and how e-values can be combined to accumulate evidence against the null. To this end, let us imagine a sequence of random variables $\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}, \ldots$ representing the outcomes of experiments/studies. We will not (except for illustration purposes later on) make use of any internal structure of the $\mathbf{Y}_{(j)}$, which in particular may come to us as batches of varying lengths.

Definition 1 *Let $\mathcal{H}_0$ be a collection of distributions for a sample space equipped with filtration $(\mathcal{F}_{(m)})_m$. We say that $E_{(m)}$ is a $\mathcal{F}_{(m-1)}$-conditional e-variable (relative to null hypothesis $\mathcal{H}_0$) if it is a nonnegative random variable that is $\mathcal{F}_{(m)}$-measurable and for all $P \in \mathcal{H}_0$: $\mathbf{E}_P[E_{(m)} \mid \mathcal{F}_{(m-1)}] \leq 1$ a.s.. If, for each m, $E_{(m)}$ is an $\mathcal{F}_{(m-1)}$-conditional e-variable, we will call $\{E_{(m)}\}_m$ a conditional e-variable collection relative to $(\mathcal{F}_{(m)})_m$.*

In standard cases, $\mathcal{F}_{(m)}$ represent all that is known to us at time m (that is, after having observed the m-th study). Then it is simply $\sigma(\mathbf{Y}^{(m)})$, with $\mathbf{Y}^{(m)} = (\mathbf{Y}_{(1)}, \ldots, \mathbf{Y}_{(m)})$ the sequence of outcomes of previous studies, and we could then rewrite the expectation elementarily as $\mathbf{E}_P[E_{(m)} \mid \mathbf{Y}^{(m-1)}]$. More generally though, $\mathcal{F}_{(m)}$ is allowed to be a coarser filtration as well: as long as for all m, $E_{(m)}$ is $\mathcal{F}_{(m)}$-measurable, we can safely engage in optional continuation in the sense of Corollary 1 below, as explained in Section 5. On the other hand, $\mathcal{F}_{(m)}$ could also be finer, including nonstochastic side information such as 'there is money to do an additional study' or covariates; we briefly describe such extensions, as well as subtleties that may arise, in Appendix B.1.

Intuitively, $\mathcal{F}_{(m-1)}$-conditional e-values measure the conditional evidence in round m (representing the m-th study) against $\mathcal{H}_0$, and hence their running product measures the total evidence (such a running product would then be a test super-martingale (**?**), i.e. a nonnegative super-martingale with starting value ≤ 1, under every element of the null). We may turn this running product into one quantity by adding a stopping rule. The following result, both parts of which are a direct implication of Doob's optional stopping theorem (**?**) states that, irrespective of the stopping rule/time, we obtain a fair measure of evidence.

Proposition 2 **1.** *Let $\{E_{(m)}\}_m$ be a collection of conditional e-variables relative to filtration $(\mathcal{F}_{(m)})_m$. Then the running product $(E^{(m)})_m$ with $E^{(m)} := \prod_{j=1}^{m} E_{(j)}$ is a test super-martingale w.r.t. each $P \in \mathcal{H}_0$.*

2. *Any process $(E^{(m)})_m$ that is a test super-martingale w.r.t. each $P \in \mathcal{H}_0$ is also an e-process (**?**) w.r.t. each $P \in \mathcal{H}_0$, which by definition means that for any stopping time τ (not necessarily finite), the stopped value $E^{(\tau)}$ is a (standard non-conditional) e-value for $\mathcal{H}_0$.*

Proposition 2 says that, no matter when we stop collecting batches of data, the resulting product is an e-variable and therefore a test based on it preserves Type-I error guarantees by Proposition 1. We note that, after the first version

of the present paper appeared on arxiv, it was found that for some composite $\mathcal{H}_0$ there exist useful e-processes that are not test-martingales (**?**); we cannot build these as products of the conditional e-variables that our main theorem, presented in the next section, generates. On the other hand, as a referee suggested, in our optional continuation context the use of conditional e-variables as basic building blocks may also have advantages over general e-processes (for example, it is easier to switch (**?**) to a different type of e-variable from one study to the next); sorting this out in detail remains a question for future work.

Just-in-Time Conditional *e*-variables: Optional Continuation

As we can see from Proposition 2, the stopped running product $E^{(\tau)}$ of a sequence of conditional *e*-variables only evaluates each member variable $E_{(m)}$ after m rounds, and only on the data $\mathbf{Y}_{(1)}, \ldots, \mathbf{Y}_{(m)}$ that actually happened. It is therefore perfectly fine (for the Type-I error safety guaranteed by combining Propositions 2 and 1) for us to construct $E_{(m)}$ on demand just before round m, as a function of all available information so far (including possibly both stochastically modeled and arbitrary variables), as long as we ensure the conditional safety property in Definition 1. This simple observation gives us tremendous flexibility for testing, much in contrast to traditional *p*-values where the sampling plan needs to be fixed up front. In particular, it allows optional continuation: the practice of deciding after an initial series of experiments whether to output the current accumulated evidence, or perform yet more experiments.

Proposition 2 already indicates that we can safely engage in such optional continuation, assuming that we stop performing studies at some stopping time τ relative to the filtration $(\mathcal{F}_{(m)})_{(m)}$. The following proposition expresses that we have Type-I error safety under optional continuation in an even stronger filtration-independent sense:

Corollary 1 (of Proposition 2): "Ville-Robbins"

$$\textit{For all } P \in \mathcal{H}_0, \textit{ all } 0 < \alpha \leq 1\text{: } P\left(\textit{there exists } m \textit{ such that } E^{(m)} \geq \frac{1}{\alpha}\right) \leq \alpha. \tag{7}$$

Proof Proposition 2 expresses that $E^{(1)}, E^{(2)}, \ldots$ is a super-martingale with starting value ≤ 1. Ville's inequality (**?**) (also known as Ville-Robbins or Ville-Robbins-Wald inequality) then implies (7).

For use later on, we formally define a threshold test based on non-negative process $(E^{(m)})_m$ to be the random function that, when input m and level α, outputs *reject* if $E^{(m)} \geq 1/\alpha$ and outputs *accept* otherwise (in this general definition, $E^{(m)}$ can but does not need to be defined as a product of conditional *e*-variables). We say that a threshold test is safe under optional continuation with respect to Type I error if (7) holds. Thus, no matter when the data collecting and combination process is stopped, the Type-I error probability is preserved. We relate optional continuation to the more common notion of 'optional stopping' and discuss filtration-related subtleties in Section 5.

Whereas Ville-Robbins stresses that we may greedily 'keep combining studies until we can reject or resources run out', it is just as important that our *e*-values keep providing valid Type-I error guarantees if the continuation rule is externally imposed or unknowable.

Example 1 Consider the simple scenario with a single underlying data stream $Y_1, Y_2, \ldots$ with Y_i i.i.d. according to both $\mathcal{H}_0$ and $\mathcal{H}_1$. Assume for simplicity simple $\mathcal{H}_0 = \{P_0\}$ so that Bayes factors provide *e*-variables. For arbitrary prior W on Θ_1, define $e_{n,W}(Y^n) = p_W(Y^n)/p_0(Y^n)$ to be the Bayes factor as in (5) with prior W for Θ_1 applied to data Y^n.

Suppose we perform an initial study on sample $\mathbf{Y}_{(1)} := Y^{N_{(1)}} := (Y_1, \ldots, Y_{N_{(1)}})$ and we equip Θ_1 with prior $W_{(1)}$. We can use as our *e*-variable $E_{(1)}$ the Bayes factor $E_{(1)} := e_{N_{(1)},W_{(1)}}(\mathbf{Y}_{(1)})$. Suppose this leads to a first *e*-value $E_{(1)} = 18$ — promising enough for us to invest our resources into a subsequent study. We decide to gather $N_{(2)}$ data points leading to data $\mathbf{Y}_{(2)} = (Y_{N_{(1)}+1}, \ldots, Y_{N_{(1)}+N_{(2)}})$. For this second data batch, we will use an *e*-variable $E_{(2)} := e_{N_{(2)},W_{(2)}}(\mathbf{Y}_{(2)})$ for a new prior $W_{(2)}$. Crucially, we are allowed to choose both $N_{(2)}$ and $W_{(2)}$ as a function of

past data $\mathbf{Y}^{(1)}$: clearly, because the underlying data stream was assumed i.i.d., $\mathbf{E}_{P_0}[E_{(2)} \mid \mathbf{Y}_{(1)}] \leq 1$ irrespective of our choice (here we use (6)), and this allows us to use Proposition 2. If we want to stick to the Bayesian paradigm, we can choose $W_{(2)} := W_{(1)}(\cdot \mid \mathbf{Y}_{(1)})$, as the Bayes posterior for θ_1 based on data $\mathbf{Y}_{(1)}$ and prior $W_{(1)}$. Bayes' theorem shows that multiplying $E^{(2)} := E_{(1)} \cdot E_{(2)}$ (which gives a new e-variable by Proposition 2), satisfies

$$E^{(2)} = E_{(1)} \cdot E_{(2)} = \frac{p_{W_{(1)}}(\mathbf{Y}_{(1)}) \cdot p_{W_{(1)}(\cdot|\mathbf{Y}_{(1)})}(\mathbf{Y}_{(2)})}{p_0(\mathbf{Y}_{(1)}) \cdot p_0(\mathbf{Y}_{(2)})} = \frac{p_{W_{(1)}}(Y_1, \ldots, Y_{N_{(2)}})}{p_0(Y_1, \ldots, Y_{N_{(2)}})}, \tag{8}$$

which is exactly what one would get by Bayesian updating. This illustrates that, for simple $\mathcal{H}_0$, combining e-variables by multiplication can be done consistently with Bayesian updating.

The Local Perspective

It might also be the case that it is not us who get the additional funding to obtain extra data, but rather some research group at a different location. If the question is, say, whether a medication works, the null hypothesis would still be $\mathcal{H}_0 = \{P_0\}$ but, if it works, its effectiveness might be slightly different due to slight differences in population. In that case, the research group might decide to use a different test statistic $E'_{(2)}$ which is again a Bayes factor, but now with an alternative prior W on θ_1 (for example, the original prior $W_{(1)}$ might be re-used rather than replaced by $W_{(1)}(\cdot \mid \mathbf{Y}_{(1)})$) — one might call this the local perspective. Even though not standard Bayesian, $E_{(1)} \cdot E'_{(2)}$ still gives a valid e-variable, and Type-I error guarantees are preserved — and the same will hold even if the new research group would use an entirely different prior on Θ_1. And, after the second batch of data $\mathbf{Y}_{(2)}$, one might consider obtaining even more samples, each time using a different $W_{(j)}$, that is always allowed to depend on the past in arbitrary ways.

Finally, it is important to note that, when combining studies, we do require all the data batches $\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}, \ldots$ to refer to separate data: obviously it is not allowed to 'borrow' some data from $\mathbf{Y}_{(1)}$ and reuse it as part of $\mathbf{Y}_{(2)}$. E-values based partly on the same data can still be validly combined (e.g. by averaging (**?**)) but not by multiplication as we do here.

2 | THE GRO e-VARIABLE

Mathematical Preliminaries

In this and the coming sections we present our main result, Theorem 1. We first list all required mathematical notations and definitions. We invariably assume that some family $\{P_\theta : \theta \in \Theta\}$ of probability distributions for $\mathcal{Y}$ has been fixed and all P_θ with $\theta \in \Theta$ have densities relative to some underlying measure μ. When we write 'p is a (sub-) probability density', we mean it is a (sub-) probability density relative to μ, i.e. $p \geq 0$ and $\int p(\mathbf{Y})\, d\mu = 1$ for a density and $\int p(\mathbf{Y})\, d\mu \leq 1$ for a sub-density. In the latter case we call the measure P with density p a sub-probability distribution. We use $D(Q\|P)$ to denote the Kullback-Leibler (KL) Divergence between distributions Q and P (**?**). We allow P (but not Q) to be a sub-probability distribution, with $D(Q\|P) = \mathbf{E}_{\mathbf{Y}\sim Q}[\ln q(\mathbf{Y})/p(\mathbf{Y})]$. We say that random variables U^* and U° are essentially equal if, for all $\theta \in \Theta$, $P_\theta(U^* = U^\circ) = 1$. We say that U^* essentially uniquely satisfies property *prop* if all other random variables satisfying property *prop* are essentially equal to U^*. When we write 'P has full support', we mean that its density p satisfies $p(\mathbf{Y}) > 0$ μ-almost everywhere. We assume some suitable σ-algebra including all singleton sets on Θ has been defined, and for $\Theta' \subset \Theta$ we let $\mathcal{W}(\Theta')$ be the set of all probability distributions (i.e., 'proper priors') on Θ' with this σ-algebra. Notably, $\mathcal{W}(\Theta')$ includes, for each $\theta \in \Theta'$, the degenerate distribution W which puts all mass on θ. We say that W essentially uniquely satisfies property *prop* among $\mathcal{W}(\Theta')$ if for all other distributions $W' \in \mathcal{W}(\Theta')$ that satisfy *prop* and all $\theta \in \Theta$, we have $P_\theta(p_W = p'_W) = 1$, where p_W and p'_W are as in (4).

$\mathcal{E}(\Theta_0)$ is defined as the set of all e-variables that can be defined on $\mathbf{Y}$ for Θ_0, i.e. all random variables satisfying (1). We frequently use the fact that if $\Theta_0 = \{0\}$ is a singleton so that $\mathcal{H}_0$ is simple, then the class of e-variables corresponds exactly to the set of likelihood ratios relative to p_0:

$$\mathcal{E}(\{0\}) = \left\{ \frac{q(\mathbf{Y})}{p_0(\mathbf{Y})} \ : q \text{ is a sub-probability density for } \mathbf{Y} \right\}. \tag{9}$$

To see this, note that for every e-variable $E = e(\mathbf{Y})$ we can define $q(\mathbf{Y}) := e(\mathbf{Y}) \cdot p_0(\mathbf{Y})$ and then $\int q d\mu = \int p_0(\mathbf{Y}) e(\mathbf{Y}) d\mu \leq 1$; conversely every sub-density q defines an e-variable by setting $E = q(\mathbf{Y})/p_0(\mathbf{Y})$ which gives $\mathbf{E}_{P_0}[E] \leq 1$.

Our main theorem (proof in Appendix A.1) implies that nontrivial e-variables exist without any further conditions:

Theorem 1 *Suppose Q is a probability distribution with full support and with density q, and assume $\inf_{W_0 \in \mathcal{W}(\Theta_0)} D(Q\|P_{W_0}) < \infty$. Then there exists a (potentially sub-) distribution P_0^* with density p_0^* such that*

$$E^* := \frac{q(\mathbf{Y})}{p_0^*(\mathbf{Y})} \tag{10}$$

is an e-variable. Moreover, E^* satisfies, essentially uniquely,

$$\sup_{E \in \mathcal{E}(\Theta_0)} \mathbf{E}_{\mathbf{Y}\sim Q}[\log E] = \mathbf{E}_{\mathbf{Y}\sim Q}[\log E^*] = \inf_{W_0 \in \mathcal{W}(\Theta_0)} D(Q\|P_{W_0}) = D(Q\|P_0^*). \tag{11}$$

If the minimum is achieved by some W_0^*, i.e. $D(Q\|P_0^*) = D(Q\|P_{W_0^*})$, then $P_0^* = P_{W_0^*}$.

The full support condition is natural and discussed further in Appendix A.3. Following Barron and **?** (see also (**?**)), we call P_0^* the Reverse Information Projection (RIPr) of Q on $\mathcal{P}(\Theta_0) = \{P_{W_0} : W_0 \in \mathcal{W}(\Theta_0)\}$. In all examples in this paper, we have $P_0^* = P_{W_0^*}$: the minimum is achieved and its density integrates to 1 (one can construct special $\mathcal{H}_0$ for which p_0^* integrates to strictly less than 1 (**?**), but we do not know whether this happens for any practically relevant $\mathcal{H}_0$). The following corollary (see Appendix A.1 for details) is useful in applications:

Corollary 2 *E^* is the only e-variable of Bayes factor/likelihood ratio form with q in the numerator. That is, for all $W_0 \in \mathcal{W}(\Theta_0)$: if P_{W_0} is not essentially equal to P_0^* then $q(\mathbf{Y})/p_{W_0}(\mathbf{Y})$ is not an e-variable. In particular this implies: (a) if $P_0^* = P_{W_0^*}$, then W_0^* achieves $\min_{W_0 \in \mathcal{W}(\Theta_0)} D(Q\|P_{W_0})$ essentially uniquely; and (b) if we have found an e-variable of form $q(\mathbf{Y})/p_{W_0}(\mathbf{Y})$ then W_0 must be essentially equal to W_0^*.*

Theorem 1 leaves open the question of how to calculate W_0^*, if it exists. In the examples we encounter below, we can either show that W_0^* is degenerate, putting all its mass on a single distribution $P_{\theta_0^*}$, and θ_0^* can be determined analytically, or, as in the t-test example, we can analytically find it by other means. 'Easy' W_0^* occur in surprisingly many other situations as well (see e.g. (**??**)), but by no means always (**?**). More generally, even if W_0^* is not easy to determine analytically, as long as $\mathcal{Y}$ is finite then, using Carathéodory's theorem we can still show that W_0^* must exist and has finite support. By strict convexity of KL divergence in its second argument it can therefore in principle be found by numerical methods, but more research is needed to see existing methods are fast enough in practice. If $\mathcal{Y}$ is infinite, one can still try to approximate W_0^* numerically but it may be hard to determine the accuracy of such approximations.

2.1 | The GRO criterion when $\mathcal{H}_1$ is simple

We now focus on the case with a given alternative $\mathcal{H}_1 = \{P_\theta : \theta \in \Theta_1\}$, and for now assume $\Theta_1 = \{\theta_1\}$ is a singleton. Applying Theorem 1 above with $Q = P_{\theta_1}$, we call the resulting E^* (or any essentially equal version of it) the θ_1-GRO

e-variable, GRO standing for growth-rate optimal. We define the growth rate achievable with θ_1 as

$$\mathrm{GRO}(\theta_1) := \sup_{E \in \mathcal{E}(\Theta_0)} \mathbf{E}_{Y \sim P_{\theta_1}}[\log E] = D(P_{\theta_1} \| P_0^*), \tag{12}$$

with the equality following from Theorem 1 (we omit Θ_0 in the notations since, in contrast to Θ_1 or θ_1, Θ_0 will always be clear from context). In general, there exist many nontrivial e-variables for a given $\mathcal{H}_0$. The θ_1-GRO e-variable is a special one that is optimal in a natural sense for the given $\mathcal{H}_1$: whereas in the Neyman-Pearson paradigm, one measures the quality of a test at a given significance level α by its power, i.e. the probability of correct decision under θ_1, we will measure it by the expected capital growth rate under θ_1. This is different from power, yet there are close connections to which we return in Section 9.

To explain, we return to the monetary interpretation of e-values. The definition of e-variable ensures that we expect them to stay under 1 (one does not gain money) under any $P \in \mathcal{H}_0$. Analogously, one would like them to be constructed such that they can be expected to grow large as fast as possible (one gets rich, gets evidence against $\mathcal{H}_0$) under $\mathcal{H}_1$. Assuming for now that $\mathcal{H}_1 = \{P_{\theta_1}\}$ is simple, this suggests to define the optimal e-variable E^* as the one that maximises $\mathbf{E}_{P_{\theta_1}}[f(E^*)]$ for some function that is increasing in E^*. At first sight it may seem natural to pick f the identity, but this can lead to adoption of an e-variable E^* such that $P_{\theta_1}(E^* = 0) > 0$. This choice, however, does not go together well with preserving evidence (capital) under optional continuation: if $E^*_{(1)}$ is 0 with positive probability, then it may happen that the evidence $E^{(m)} = \prod_{j=1}^m E_{(j)}$ obtained so far remains 0, no matter how large $E_{(j)}$ for $j \geq 1$ — akin to losing all one's money in the first round at a roulette table. A similar objection applies to any polynomial f, but it does not apply to the logarithm, which is also the asymptotically optimal choice for f if samples are independent: by Kolmogorov's strong law of large numbers, any sequence of e-variables $E_{(1)}, E_{(2)}, \ldots$ based on independent $\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}, \ldots$ with $\sup_{j \in \mathbb{N}} \mathbf{E}_{\mathbf{Y}_{(j)} \sim P_{\theta_1}}[(\log E_{(j)})^2]/(j/\log^2(j+1)) < \infty$ (in particular this holds if the variances are uniformly bounded), will a.s. satisfy $(1/m)\sum_{j=1}^m (\log E_{(j)}) - (1/m)\sum_{j=1}^m \mathbf{E}_{\mathbf{Y}_{(j)} \sim P_{\theta_1}}[\log E_{(j)}] \to 0$. Thus, $E^{(m)} := \prod_{j=1}^m E_{(j)}$ will grow exponentially fast if all $\mathbf{E}_{\mathbf{Y}_{(j)} \sim P_{\theta_1}}[\log E_{(j)}] > 0$, with maximal growth rate attained if the $E_{(j)}$ are chosen to maximize $\mathbf{E}_{\mathbf{Y}_{(j)} \sim P_{\theta_1}}[\log E_{(j)}]$ — a quantity which, for log taken to the base 2, is known as the doubling rate (**??**). This provides a powerful reason for choosing the logarithm; see also the extensive exposition by **?**. We return to GRO's motivation in Section 8.

Example 2 **[2×2 Contingency Tables]** Let $\mathcal{Y}^n = \{0,1\}^n$ and let $\mathcal{X} = \{a, b\}$ represent two categories. We start with a multinomial model $\mathcal{G}$ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, extended to n outcomes by independence. We want to test whether the Y_i are dependent on the X_i. To this end, we condition every distribution in $\mathcal{G}$ on a fixed, given, $X^n = \mathbf{x}$ with $\mathbf{x} = (x_1, \ldots, x_n)$, and we let $\mathcal{H}_1$ be the set of (or a subset of the) conditional distributions on $\mathcal{Z}$ that thus result. We thus assume the design of X^n to be set in advance, but N_1, the number of ones, to be random; alternative choices are possible and would lead to a different analysis. Conditioned on $X^n = \mathbf{x}$, the likelihood of an individual sequence $\mathbf{y} \mid \mathbf{x}$ becomes:

$$p_{\mu_{1|a},\mu_{1|b}}(\mathbf{y} \mid \mathbf{x}) = \mu_{1|a}^{n_{a1}}(1-\mu_{1|a})^{n_{a0}} \cdot \mu_{1|b}^{n_{b1}}(1-\mu_{1|b})^{n_{b0}}, \tag{13}$$

where n_{ji} is the number of times outcome i was observed to fall in category j and $\mu_{1|j}$ is the probability of observing a 1 given category j. These densities define the full model $\{P_{\mu_{1|a},\mu_{1|b}} : (\mu_{1|a}, \mu_{1|b}) \in \Theta\}$ with $\Theta = [0,1]^2$. $\mathcal{H}_0$, the null model, simply has $(X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ independent, with $Y_i, \ldots, Y_n$ i.i.d. $\mathrm{Ber}(\mu)$ distributed, $\mu \in \Theta_0 := [0,1]$, i.e. $p_\mu(\mathbf{y} \mid \mathbf{x}) = p_\mu(\mathbf{y}) = \mu^{n_1}(1-\mu)^{n-n_1}$ with $n_1 = n_{a1} + n_{b1}$. We defer description of the test of the full alternative $\{P_\theta : \theta \in \Theta_1\}$ with $\Theta_1 = (0,1)^2$ against $\mathcal{H}_0$ to Section 4.1. For now, we assume a simple $\mathcal{H}_1 = \{P_{\theta_1}\}$ for a specific $\theta_1 = (\mu_{1|a}, \mu_{1|b})$ with $\mu_{1|a} \neq \mu_{1|b}$. **?** shows that the RIPr for P_{θ_1}, achieving the infimum in (11) is given by $P_0^* = P_{W_0^*}$ where W_0^* is the degenerate prior that puts all its mass on the single point $\mu^\circ = (n_a\mu_{1|a} + n_b\mu_{1|b})/(n_a + n_b)$.

Thus, the θ_1-GRO e-variable has an intuitive form here, being given by

$$E^* = \frac{p_{\mu_{1|a},\mu_{1|b}}(\mathbf{Y} \mid \mathbf{x})}{p_{\mu^\circ}(\mathbf{Y} \mid \mathbf{x})}. \tag{14}$$

The fact that the RIPr is achieved by a point prior is quite specific to contingency tables. We also note that, while the expectation of E^* is bounded by 1 under all $P_\mu \in \mathcal{H}_0$, its actual distribution function varies with P_μ. This is in contrast to the t-test example, in which the GRO E^* turns out to have the same distribution under all distributions in the null.

2.2 | GRO when prior on $\mathcal{H}_1$ is available

We now take a Bayesian stance regarding $\mathcal{H}_1$ and, conditioned on $\mathcal{H}_1$, are prepared to represent our uncertainty by prior distribution W_1 on Θ_1. The marginal distribution of $\mathbf{Y}$ is then $P_{W_1}(\mathbf{Y})$. Applying Theorem 1 with P_{W_1} as Q then leads to the "W_1-GRO e-variable" — it would be optimal in the GRO sense under prior W_1. This e-variable is a Bayes factor, but gives only a quasi-Bayesian notion of evidence since any prior W_1 on $\mathcal{H}_1$ that we wish to adopt forces us to adopt a particular corresponding prior $W_0^* \in \mathcal{H}_0$. One may perhaps consider this a small price to pay for creating a Bayes factor that, by its Type-I error safety under optional continuation, should be acceptable to frequentists as well. Moreover it is often recognised that priors on Θ_0 and Θ_1 should somehow be 'matched' to each other (**?**); we may view the RIPr construction as providing a reasonable (from a frequentist stance) matching.

Example 3 [Gaussian Location with Gaussian prior (z-test)] Now consider $\mathcal{H}_1$ according to which the Y_i are i.i.d. $\sim N(\mu, 1)$ for some $\mu \in \Theta_1 = \mathbb{R}$, so that $p_\mu(\mathbf{Y}) = p_\mu(Y_1, \ldots, Y_n) \propto \exp(-\sum_{i=1}^n (Y_i - \mu)^2/2)$. We let $\mathcal{H}_0 = \{P_0\}$. We perform a Bayes factor test using $E := p_W(\mathbf{Y})/p_0(\mathbf{Y})$ where we take the prior W to have Gaussian density $w(\mu) \propto \exp(-\mu^2/2)$. By (6) we know that E is not just a Bayes factor but also an e-variable. By straightforward calculation:

$$-\frac{1}{2}\log(n+1) + \frac{1}{2}(n+1) \cdot \breve{\mu}_n^2,$$

where $\breve{\mu}_n = (\sum_{i=1} Y_i)/(n+1)$ is the Bayes MAP estimator, which only differs from the ML estimator by $O(1/n^2)$: $\breve{\mu}_n - \hat{\mu}_n = \hat{\mu}_n/(n(n+1))$. If we were to reject Θ_0 when $E \geq 20$ (giving, by Proposition 1 a Type-I error guarantee of 0.05), we would thus reject if

$$|\breve{\mu}_n| \geq \sqrt{\frac{5.99 + \log(n+1)}{n+1}}, \text{ i.e. } |\hat{\mu}_n| \geq \sqrt{(\log n)/n}, \tag{15}$$

where we used $2\log 20 \approx 5.99$. Contrast this with the standard two-sided Neyman-Pearson (NP) test, which would reject (with $\alpha = 0.05$) if $|\hat{\mu}_n| \geq 1.96/\sqrt{n}$, or the one-sided test which would reject if $\hat{\mu}_n \geq 1.645/\sqrt{n}$ or the e-value based tests of the next section: the standard Bayesian test is significantly more conservative, requiring more data to conclude rejection. In Section 6 we investigate this further.

3 | THE GROW e-VARIABLE

We now show how to construct good e-variables if $\mathcal{H}_1$ is composite and no prior on Θ_1 is available. We focus on variations of worst-case (maximin) growth optimality, but this is certainly not the only criterion that might be useful or valuable; see the discussion in Section 8. In the case of simple $\mathcal{H}_1 = \{P_{\theta_1}\}$, we aimed for e-variables that could be expected to grow as fast as possible under P_{θ_1}. Analogously, we would now like them to be constructed such that they can be expected to grow large as fast as possible (one gets rich, gets evidence against $\mathcal{H}_0$) under all $P_1 \in \mathcal{H}_1$.

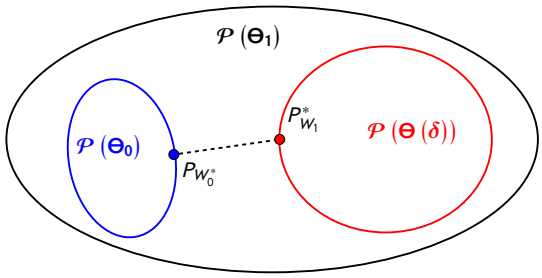

FIGURE 1 Joint Information Projection (JIPr). Θ_0, Θ_1 represent non-overlapping models, $\mathcal{W}(\Theta')$ is the set of all priors over Θ', and $\mathcal{P}(\Theta') = \{P_W : W \in \mathcal{W}(\Theta')\}$. Theorem 1 implies that the GROW e-variable between Θ_0 and Θ_1 is given by $p_{W_1^*}/p_{W_0^*}$, the Bayes factor between the two Bayes marginals that minimise KL divergence $D(P_{W_1} \| P_{W_0})$, assuming the minima are achieved.

We call e-variables satisfying this property GROW: growth-rate optimal in worst-case. We now discuss the simplest, 'raw' form of this criterion — in some settings a modification of this criterion, REGROW, which we discuss in the next section, is more suitable. GROW tells us to pick, among all e-variables relative to $\mathcal{H}_0$, the E^* that achieves the worst-case optimal expected capital growth rate

$$\text{GROW}(\Theta_1) := \sup_{E : E \in \mathcal{E}(\Theta_0)} \inf_{\theta \in \Theta_1} \mathbf{E}_{P_\theta}[\log E]. \tag{16}$$

Theorem 1, First Generalisation *Suppose all $P_\theta, P_{\theta'}$ with $\theta, \theta' \in \Theta_1$ satisfy $D(P_\theta \| P_{\theta'}) < \infty$, and have full support. If $\inf_{W_1 \in \mathcal{W}_1, W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1} \| P_{W_0}) = \inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1^*} \| P_{W_0}) < \infty$, (i.e. the minimum on the left over $\mathcal{W}_1$ is achieved by W_1^*) then then there exists an e-variable*

$$E^* := \frac{p_{W_1^*}(\mathbf{Y})}{p_0^*(\mathbf{Y})}, \tag{17}$$

where p_0^* is the density of P_0^*, a (potentially sub-) distribution satisfying $\inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1^*} \| P_{W_0}) = D(P_{W_1^*} \| P_0^*)$, and E^* achieves (16), satisfying, essentially uniquely: $\inf_{\theta \in \Theta_1} \mathbf{E}_{\mathbf{Y} \sim P_\theta}[\log E^*] = \sup_{E \in \mathcal{E}(\Theta_0)} \inf_{\theta \in \Theta_1} \mathbf{E}_{\mathbf{Y} \sim P_\theta}[\log E] = D(P_{W_1^*} \| P_0^*)$. If further $D(P_{W_1^*} \| P_0^*) = D(P_{W_1^*} \| P_{W_0^*})$ for some $W_0^* \in \mathcal{W}(\Theta_0)$, then $P_0^* = P_{W_0^*}$.

The earlier version of Theorem 1 is the special case we get if we set $\Theta_1 = \{\theta_1\}$ a singleton and $Q := P_{\theta_1}$. This generalized version expresses that the GROW e-variable is once again a Bayes factor — a special one in fact, between the components of the joint information projection $(P_{W_1^*}, P_0^*)$ (?); see Figure 1. As to computing W_1^* in practice, the same remarks apply as we already made (underneath Corollary 2) regarding computing W_0^*.

3.1 | One-parameter models with minimum relevant effect size

Let Θ be a connected subset of $\mathbb{R}$ indexing a 1-parameter parametric model $\{P_\theta : \theta \in \Theta\}$ with θ indicating the size of some effect. If, as is standard practice in e.g. medical statistics, we have a minimum clinically relevant effect size δ^+ and a status quo $\delta^- < \delta^+$ in mind, we want to test

$$\Theta_0 = \{\theta \in \Theta : \theta \leq \delta^-\} \text{ vs. } \Theta_1 = \{\theta \in \Theta : \theta \geq \delta^+\}. \tag{18}$$

In standard cases, often $\delta^- = 0$ and $\Theta_0 := \{0\}$.

Proposition 3 *Suppose there exists a 1-dimensional statistic $T = t(\mathbf{Y})$ such that the family of densities $\{p_\theta : \theta \in \Theta\}$ has a monotone likelihood ratio in T. Then for all $\delta^- < \delta^+$ with $\delta^-, \delta^+ \in \Theta$, the GROW e-variable relative to Θ_1 and Θ_0 as in (18), is given by $E^* = p_{\delta^+}(\mathbf{Y})/p_{\delta^-}(\mathbf{Y})$: it sets W_1^* and W_0^* to be degenerate priors, putting all mass on δ^+ and δ^-, respectively.*

We now illustrate Proposition 3 for 1-dimensional exponential families, but stress that it can be applied to some other families (e.g. location families or the t-test setting in Section 4.3) as well.

Example 4 [GROW for 1-dimensional exponential families] Let $\{P_\theta \mid \theta \in \Theta\}$ represent a 1-parameter exponential family for sample space $\mathcal{Y}$, given in its mean-value parameterisation, where Θ is a connected subset of (and possibly equal to) the full mean-value parameter space. Let $\delta^- < \delta^+$ with δ^-, δ^+ both in Θ. Both $\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\}$ and $\mathcal{H}_1 = \{P_\theta : \theta \in \Theta_1\}$ with Θ_0, Θ_1 as in (18) are extended to outcomes in $\mathbf{Y} = (Y_1, \ldots, Y_n)$ by independence. Let $T = t(\mathbf{Y})$ be the sufficient statistic of the exponential family under consideration, i.e. $\mathbf{E}_{\mathbf{Y}\sim P_\theta}[t(\mathbf{Y})] = \theta$. It is well-known that the monotone likelihood property holds in the statistic T. It thus follows from Proposition 3 above that the GROW e-variable relative to Θ_1 and Θ_0 can be calculated as a likelihood ratio $E^* = p_{\delta^+}(\mathbf{Y})/p_{\delta^-}(\mathbf{Y})$ between two point hypotheses, even though Θ_1 and/or Θ_0 may be composite. Comparison of the ensuing test to the Neyman-Pearson and Bayes factor tests are given in Section 6.

4 | THE REGROW e-VARIABLE: GENERAL COMPOSITE $\mathcal{H}_1$ CASE

Theorem 1, Further Generalisation *Let $f(\theta)$ be a function that is bounded on Θ_1; we abbreviate $f(W) := \mathbf{E}_{\theta\sim W}[f(\theta)]$. Suppose all $P_\theta, P_{\theta'}$ with $\theta, \theta' \in \Theta_1$ satisfy $D(P_\theta \| P_{\theta'}) < \infty$, and have full support. If $\inf_{W_1 \in \mathcal{W}(\Theta_1), W_0 \in \mathcal{W}(\Theta_0)} (D(P_{W_1} \| P_{W_0}) - f(W_1)) = \inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1^*} \| P_{W_0}) - f(W_1^*) < \infty$ then there exists an e-variable E^f given by*

$$E^f := \frac{p_{W_1^*}(\mathbf{Y})}{p_0^*(\mathbf{Y})} \tag{19}$$

where p_0^* is the density of P_0^*, a (potentially sub-) distribution such that $\inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1^*} \| P_{W_0}) = D(P_{W_1^*} \| P_0^*)$, and E^f satisfies, essentially uniquely:

$$\inf_{\theta \in \Theta_1} \left(\mathbf{E}_{\mathbf{Y}\sim P_\theta}[\log E^f] - f(\theta) \right) = \sup_{E \in \mathcal{E}(\Theta_0)} \inf_{\theta \in \Theta_1} \left(\mathbf{E}_{\mathbf{Y}\sim P_\theta}[\log E] - f(\theta) \right) = D(P_{W_1^*} \| P_0^*) - f(W_1^*). \tag{20}$$

If further $D(P_{W_1^*} \| P_0^*) = D(P_{W_1^*} \| P_{W_0^*})$ for some $W_0^* \in \mathcal{W}(\Theta_0)$, then $P_0^* = P_{W_0^*}$.

We call E^f the REGROW (standing for relative growth in the worst-case) e-variable relative to offset f. The previous version of Theorem 1 is the special case with f constant. The offset f will be useful when Θ_0 and Θ_1 are nested and no effect size can be stated in advance (Section 4.1) and/or when nuisance parameters are present (Section 4.2 and 4.3). All these cases can be handled essentially the same way (and we may in fact think of the case of nested models as a situation in which all parameters in Θ_0 are viewed as nuisance): we first consider a modified problem in which Θ_1 is reduced to a singleton. That is, we imagine that some oracle tells us "if $\mathcal{H}_1$ is the case, then the data are sampled from this specific θ_1^*". We then consider the corresponding $\mathrm{GRO}(\theta_1^*) = \mathrm{GROW}(\{\theta_1^*\})$ and view this as the desirable but unobtainable expected growth rate — the one we could have obtained if we had known θ_1^*. We may now aim for the e-variable such that, no matter what θ_1^* turns out to be, our expected growth is close to the optimum we could have obtained had we known θ_1^*. Thus, we want to be worst-case growth optimal relative to $f(\theta_1) := \mathrm{GRO}(\theta_1) = \mathbf{E}_{P_{\theta_1}}[\log E_{\theta_1}^*] = \inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{\theta_1} \| P_{W_0})$ (where we write $E_{\theta_1}^*$ for the GRO e-variable for testing $\{\theta_1\}$ vs. Θ_0 and the second equality follows by (12)). Plugging in this f and taking negatives on both sides, (20) now

becomes:

$$\sup_{\theta_1 \in \Theta_1} \mathbf{E}_{\mathbf{Y}\sim P_{\theta_1}} \left[\log E^*_{\theta_1} - \log E^f \right] = \inf_{E \in \mathcal{E}(\Theta_0)} \sup_{\theta_1 \in \Theta_1} \mathbf{E}_{\mathbf{Y}\sim P_{\theta_1}} [\log E^*_{\theta_1} - \log E] = \mathbf{E}_{\theta_1 \sim W_1^*} \left[\inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{\theta_1} \| P_{W_0}) \right] - D(P_{W_1^*} \| P_0^*), \quad (21)$$

an expression that is always nonnegative, since, by definition of $E^*_{\theta_1}$, for any e-variable E, $\mathbf{E}_{\mathbf{Y}\sim P_{\theta_1}}[\log E^*_{\theta_1}] \geq \mathbf{E}_{\mathbf{Y}\sim P_{\theta_1}}[\log E]$. This shows that E^f can be thought of as a minimax pseudo-regret e-variable, regret being the loss of expected capital growth under $\mathcal{H}_1$ due to not knowing the underlying θ_1 in advance.

4.1 | Composite $\mathcal{H}_1$, no effect size known

Suppose we are interested in detecting whether there is any deviation at all from the null. There is no pre-stated effect size, and $\Theta_0 \subset \Theta_1 = \Theta$ are nested, or more generally, for all $\theta_1 \in \Theta_1$, $\inf_{\theta_0 \in \Theta_0} D(P_{\theta_1} \| P_{\theta_0}) = 0$. In this case, $\mathrm{GROW}(\Theta_1) = 0$ and the GROW e-variable that achieves it is equal to $E^* = 1$, which will never give any evidence against $\mathcal{H}_0$, so clearly, the raw GROW approach is not useful. Instead, in this setting, the REGROW approach is a sensible generalisation of successful existing approaches. We first establish this for simple nulls:

Simple Nulls

If $\Theta_0 = \{0\}$ is simple, we have $\inf_{W_0} D(P_{\theta_1} \| P_{W_0}) = D(P_{\theta_1} \| P_0)$ and $D(P_{W_1^*} \| P_{W_0^*}) = D(P_{W_1^*} \| P_0)$, and terms in (21) involving $-\log p_0(\mathbf{Y})$ cancel. Further using the 1-to-1 mapping (9) between probability densities and e-variables for the case of point 0's to rewrite (21) and using $E^f = p_{W_1^*}(\mathbf{Y})/p_0(\mathbf{Y})$, (21) simplifies to:

$$\sup_{\theta \in \Theta_1} \mathbf{E}_{\mathbf{Y}\sim P_\theta} \left[-\log \frac{p_{W_1^*}(\mathbf{Y})}{p_\theta(\mathbf{Y})} \right] = \inf_q \sup_{\theta \in \Theta_1} \mathbf{E}_{\mathbf{Y}\sim P_\theta} \left[-\log \frac{q(\mathbf{Y})}{p_\theta(\mathbf{Y})} \right] = \sup_{W_1 \in \mathcal{W}(\Theta_1)} \mathbf{E}_{\theta \sim W_1} \left[D(P_\theta \| P_{W_1}) \right], \quad (22)$$

where the infimum is over all sub-probability densities q over $\mathbf{Y}$. (22) is just the redundancy-capacity theorem (**?**) of information theory, and it has a data-compression interpretation. In a nutshell, for any e-variable of the form $p_{W_1}(\mathbf{Y})/p_0(\mathbf{Y})$, the log evidence $\log p_{W_1}(\mathbf{Y})/p_0(\mathbf{Y})$ is thought of as a difference between the code length needed to code the data using two lossless codes, one with lengths $-\log p_{W_1}$, associated with $\mathcal{H}_1$, and one with lengths $-\log p_0$, associated with $\mathcal{H}_0$. (22) expresses that when choosing $W_1 = W_1^*$, one associates $\mathcal{H}_1$ with the code that minimises worst-case redundancy (the additional expected number of bits needed compared to an encoder that knows θ_1^*). This is in accordance with the MDL (Minimum Description Length) Principle, in which code length difference between the same two codes is used to measure evidence (**??**).

Example 5 [Exponential Families with a point null: Jeffreys' Prior on Θ_1] To make this more concrete, let $\{P_\theta : \theta \in \Theta\}$ represent a d-dimensional exponential family given in either the mean or the canonical parameterisation. We restrict the parameter set to Θ_1 that is a compact subset of the interior of Θ and let Θ_0 be a singleton subset in the interior of Θ_1. By standard properties of exponential families, the finite KL condition of Theorem 1 applies and the problem reduces to finding the prior W_1^* on Θ_1 that satisfies (22). (**?**) showed that, for large n, this prior converges in an L_1-sense to Jeffreys' prior ('least favourable under entropy loss') , which is the main reason for adopting it in MDL model selection. They also showed that (22) and hence (21) is of size $(d/2)\log n + O(1)$. Thus, for point nulls and suitably truncated parameter spaces, this approach is consistent with the MDL Principle and with objective Bayes approaches based on Jeffreys prior.

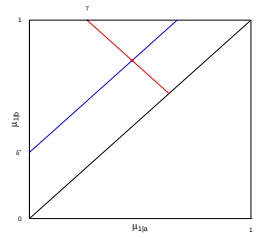

FIGURE 2 The 2×2 model. The diagonal represents the null, the decreasing line the set of parameters with nuisance parameter $\gamma = 1/4$ and the blue increasing line is Θ_{1,δ^+} for $\delta^+ = 1/3$.

Example 6 **[2×2 Tables, Continued]** If Θ_0 is not a singleton then the simplification of (21) to (22) is not possible, and numerical simulation can be used to determine (21) and the priors appearing therein. Consider for example the 2×2 model, but now with unrestricted $\Theta_1 = (0,1)^2$. This does satisfy the regularity conditions needed for Theorem 1 to be applicable (see Appendix A.3), but it has Θ_1 2-dimensional and Θ_0 1-dimensional. We saw in the previous example that for 1-vs. 0-dimensional exponential family models, (21) would take on value $(1/2)\log n + O(1)$, which suggests that it is the same here, for 2- vs. 1-dimensional. This is confirmed by numerical simulations (**?**).

4.2 | Composite $\mathcal{H}_1$, nuisance parameters present

We now consider the common situation of models that can be parameterised by $\Theta = \{(\delta, \gamma) : \delta \in \Delta, \gamma \in \Gamma\}$ where δ is a single parameter of interest (for simplicity taken to be a scalar) and γ represents a nuisance parameter (scalar or vector). As in Section 3.1, we want to test whether $\delta \geq \delta^+$ or $\delta \leq \delta^-$ for some $\delta^- < \delta^+$. We thus let

$$\Theta_0 = \{(\delta, \gamma) : \delta \leq \delta^-, \gamma \in \Gamma\}, \quad \text{vs. } \Theta_1 = \{(\delta, \gamma) : \delta \geq \delta^+, \gamma \in \Gamma\}. \tag{23}$$

We will first consider the simplified problem in which we test $\Theta_{0,\delta^-} := \{(\delta^-, \gamma) : \gamma \in \Gamma\}$ vs. $\Theta_{1,\delta^+} := \{(\delta^+, \gamma) : \gamma \in \Gamma\}$, and later extend to (23). This simplified problem can be handled via a REGROW *e*-variable just like in the previous subsection, with now $\theta = (\delta^+, \gamma)$: we take $f((\delta^+, \gamma)) = \mathrm{GRO}((\delta^+, \gamma))$ and apply Theorem 1 as in (21). This gives an *e*-variable $E^*_{\delta^+} := p_{W_1^*}(\mathbf{Y})/p_0^*(\mathbf{Y})$ with W_1^* a prior on $\{(\delta^+, \gamma) : \gamma \in \Gamma\}$. Using this REGROW rather than GROW approach reflects a particular interpretation of what it means for a parameter γ to be nuisance: we have no idea of what the true γ might be and are therefore prepared to incur the same expected loss of growth for not knowing γ, irrespective of what γ is. Solving this problem for all $\delta^+ \geq \delta^-$ gives us a collection of *e*-variables $\mathcal{E}_{\geq \delta^-} := \{E^*_\delta : \delta \geq \delta^-\}$. Now suppose there exists another *e*-variable E^* such that

$$\sup_{E \in \mathcal{E}_{\geq \delta^-}} \inf_{\theta \in \Theta_1} \mathbf{E}_{P_\theta}[\log E] = \inf_{\theta \in \Theta_1} \mathbf{E}_{P_\theta}[\log E^*] \tag{24}$$

That is, we pick the worst-case optimal *e*-variable among $\mathcal{E}_{\geq \delta^-}$, thereby applying GROW on a meta-level as it were, after restricting ourselves to *e*-variables that are themselves REGROW for fixed δ and unknown γ. This E^* is then our choice for solving the original problem (23).

Example 7 **[2×2 Tables, Continued]** We can reparameterise $\{P_{\mu_{1|a},\mu_{1|b}} : (\mu_{1|a}, \mu_{1|b}) \in [0,1]^2\}$ as $\Theta = \{(\delta, \gamma) : \delta \in \Delta, \gamma \in [0,1]\}$ using $\gamma = (n_a \mu_{1|a} + n_b \mu_{1|b})/(n_a + n_b)$ as a nuisance parameter: the marginal probability of observing a 1. For

δ we can take, for example, $\delta = \mu_{1|b} - \mu_{1|a}$ to be our notion of effect size, the substantive difference, with $\Delta = [-1, 1]$. Another popular choice, like substantive difference considered by **??** is $\delta = \log((\mu_{1|b}/(1-\mu_{1|b})) \cdot (1-\mu_{1|a})/\mu_{1|a})$, i.e. the log-odds ratio, but for simplicity we stick to the substantive difference here. We take a Θ_1 and Θ_0 relative to some effect size δ^+ and δ^- as in (23) above, where for simplicity we will take $\delta^- = 0$ and $\Delta = [0, 1]$ and also $n_a = n_b$ so that $\gamma = (\mu_{1|a} + \mu_{1|b})/2$. The situation is depicted in Figure 2, where we took, as an example, Θ_1 and Θ_0 defined relative to $\delta^+ = 1/3$ and $\delta^- = 0$.

Now, assume first that the true value of γ were given in advance. We would then be dealing with a one-parameter exponential family model represented by a straight decreasing line in Figure 2; the Figure illustrates this for $\gamma = 2/3$. We would then be in the situation of Section 3.1, Example 4, and find, for any δ^+, that $\mathrm{GRO}((\delta^+, \gamma)) = \inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{\delta^+,\gamma} \| P_{W_0}) = D(P_{\delta^+,\gamma} \| P_{0,\gamma})$ where the latter equality was already stated as (14) in Example 2.

Now we look at unknown γ. As suggested above, we first set $\delta := \delta^+$ and test $\Theta_{1,\delta}$ vs. $\Theta_{1,0}$ (the increasing lines in Figure 2), taking the REGROW e-variable relative to $f((\delta, \gamma)) = \mathrm{GRO}((\delta, \gamma)) = D(P_{\delta,\gamma} \| P_{0,\gamma})$. Then the minimum W_1^* on $\Theta_{1,\delta}$ and W_0^* (with W_1^* putting mass 1 on δ and spreading its mass over γ, and $P_0^* = P_{W_0^*}$) as in (21) are achieved and have finite support, and the finite KL condition of the theorem applies (Appendix A.3). This solves the problem for testing $\Theta_{1,\delta}$ vs. $\Theta_{1,0}$ for $\delta = \delta^+$; by varying δ we get a collection of e-variables $\mathcal{E}_{\geq 0}$ containing an E_δ^* for each fixed $\delta \geq 0$. We then pick the E^* among $\mathcal{E}_{\geq 0}$ satisfying (24), which turns out equal to $E_{\delta^+}^*$: it has a point mass on δ^+ again.

Discussion

In our examples, we have used (and will keep using in the next section) the REGROW approach to first eliminate nuisance parameters, if they are present, followed by a GROW approach for the parameters of interest.(**??**) find that this gives intuitive e-variables that perform well in practical applications, not just directly in the GRO sense but also in terms of secondary measures such as a power analysis (Section 6) or as the basis of anytime-valid confidence intervals (**?**). Still, it may not always be the best way to go. For example, a REGROW approach for the parameter of interest when a minimum effect size is given may sometimes be sensible as well. Let us consider this a bit further for (for simplicity) the case with a minimum effect size δ^+ but without nuisance parameters, as in Example 4. REGROW would amount here to using $E_{W_1}^* = p_{W_1}(\mathbf{Y})/p_{\delta^-}(\mathbf{Y})$ with some prior W_1 spread out on the set $\Theta_1 = \{\delta : \delta \geq \delta^+\}$. Then we would get $\mathbf{E}_{P_\delta}[\log E_{W_1}^*] < \mathrm{GROW}(\Theta_1)$ for δ close to δ^+; $\mathbf{E}_{P_\delta}[\log E_{W_1}^*] \gg \mathrm{GROW}(\Theta_1)$ for $\delta \gg \delta^+$ so we would win if we are 'lucky' and it turns out that $\delta \gg \delta^+$. However, in practice we often deal with small sample sizes, and δ's that may very well be very close to δ^+. Then (as our experiments done for the papers above indicate) the difference in '$<$' above is non-negligible, and the GROW approach seems safer, since for the GROW e-variable we automatically have $\mathbf{E}_{P_\delta}[\log E_{W_1}^*] \geq \mathrm{GROW}(\Theta_1)$ for all $\delta \in \Theta_1$.

4.3 | Theorem 1 in Full: Application to Bayesian and Sequential t-test

If we try to apply Theorem 1 as above to the t-test, a prototypical nuisance setting, we run into the issue that the minimum KL is not achieved. This problem can be solved by extending the theorem further, allowing for densities on a coarsening of $\mathbf{Y}$. This is any random variable $\mathbf{V}$ that can be written as a function of $\mathbf{Y}$, i.e. $\mathbf{V} = g(\mathbf{Y})$ for some function g; we retrieve the previous version of Theorem 1 if we take g the identity and $\mathbf{V} = \mathbf{Y}$. We now present Theorem 1 in full generality, allowing for such coarsening and additionally for considering the best e-variable on a modified $\mathcal{H}_1$, consisting of any convex set of Bayes marginal distributions with priors on Θ_1. This is needed for accommodating the t-test. It also allows us to incorporate robust Bayesian settings (**??**), but we will not further pursue those here. In the theorem we use the following notation: for (sub-) distribution P for $\mathbf{Y}$, $P^{[\mathbf{V}]}$ denotes the marginal (sub-) distribution of P for $\mathbf{V}$, and p' denotes its density.

Theorem 1, Full Generality *Let $f(\theta)$ be a function that is bounded on Θ_1. Suppose all $P_\theta, P_{\theta'}$ with $\theta, \theta' \in \Theta_1$ satisfy $D(P_\theta \| P_{\theta'}) < \infty$, and have full support. Let $\mathcal{W}_1 \subseteq \mathcal{W}(\Theta_1)$ be convex. If for some coarsening $\mathbf{V}$ of $\mathbf{Y}$ we have:*

$$\inf_{W_1 \in \mathcal{W}_1} \inf_{W_0 \in \mathcal{W}(\Theta_0)} \left(D(P_{W_1} \| P_{W_0}) - f(W_1) \right) =$$

$$\min_{W_1 \in \mathcal{W}_1} \inf_{W_0 \in \mathcal{W}(\Theta_0)} \left(D(P^{[\mathbf{V}]}_{W_1} \| P^{[\mathbf{V}]}_{W_0}) - f(W_1) \right) = \inf_{W_0 \in \mathcal{W}_0(\Theta_0)} D(P^{[\mathbf{V}]}_{W_1^*} \| P^{[\mathbf{V}]}_{W_0}) - f(W_1^*) < \infty, \quad (25)$$

then there exists an e-variable

$$E^f := \frac{p'_{W_1^*}(\mathbf{V})}{p_0^{*'}(\mathbf{V})}, \quad (26)$$

$p_0^{*'}$ being the density of $P_0^{*[\mathbf{V}]}$, a (potentially sub-) distribution for $\mathbf{V}$ that satisfies $\inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1^*} \| P_{W_0}) = D(P^{[\mathbf{V}]}_{W_1^*} \| P_0^{*[\mathbf{V}]})$. E^f satisfies, essentially uniquely:

$$\inf_{W \in \mathcal{W}_1} \left(\mathbf{E}_{\mathbf{Y} \sim P_W}[\log E^f] - f(W) \right) = \sup_{E \in \mathcal{E}(\Theta_0)} \inf_{W \in \mathcal{W}_1} \left(\mathbf{E}_{\mathbf{Y} \sim P_W}[\log E] - f(W) \right) = D(P^{[\mathbf{V}]}_{W_1^*} \| P_0^{*[\mathbf{V}]}) - f(W_1^*). \quad (27)$$

If further $D(P^{[\mathbf{V}]}_{W_1^*} \| P_0^{*[\mathbf{V}]}) = D(P^{[\mathbf{V}]}_{W_1^*} \| P^{[\mathbf{V}]}_{W_0^*})$ for some $W_0^* \in \mathcal{W}(\Theta_0)$, then $P_0^{*[\mathbf{V}]} = P^{[\mathbf{V}]}_{W_0^*}$.

The previous version of Theorem 1 is the special case obtained by setting $\mathcal{W}_1 = \mathcal{W}(\Theta_1)$, $\mathbf{Y} = \mathbf{V}$ and using linearity of expectation. We call E^f as in (27) the REGROW *e*-variable relative to offset f and set of priors $\mathcal{W}_1$. If f is constant (no offset), we call it $\mathcal{W}_1$-GROW, noting that it gives worst-case optimal growth rate under all priors in $\mathcal{W}_1$.

The t-test Setting

We return to the setting with a nuisance parameter with notation as in Section 4.2. **?** proposed a Bayesian version of the t-test; see also (**?**). We start with the models $\mathcal{H}_0$ and $\mathcal{H}_1$ for data $\mathbf{Y} = (Y_1, \ldots, Y_n)$ given as $\mathcal{H}_0 = \{P_{0,\sigma}(\mathbf{Y}) \mid \sigma \in \Gamma\}$; $\mathcal{H}_1 = \{P_{\delta,\sigma}(\mathbf{Y}) \mid (\delta, \sigma) \in \Theta_1\}$, where $\Delta = \mathbb{R}, \Gamma = \mathbb{R}^+, \Theta_1 := \Delta \times \Gamma$ and $\Theta_0 = \{(0, \sigma) : \sigma \in \Gamma\}$, and $P_{\delta,\sigma}$ has density (with $\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i$)

$$p_{\delta,\sigma}(y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{n}{2} \left[\left(\frac{\overline{y}}{\sigma} - \delta \right)^2 + \left(\frac{\frac{1}{n}\sum_{i=1}^n (y_i - \overline{y})^2}{\sigma^2} \right) \right] \right),$$

Jeffreys proposed to equip $\mathcal{H}_1$ with a Cauchy prior $W^c[\delta]$ on the effect size δ, and both $\mathcal{H}_1$ and $\mathcal{H}_0$ with the scale-invariant prior measure with density $w^H(\sigma) \propto 1/\sigma$ on the variance. The same formula with the same prior on σ but other priors on δ was suggested by **?** with a non-Bayesian, martingale interpretation. Below we will see that, even though $w^H(\sigma)$ is improper (whereas the priors appearing in Theorem 1 are invariably proper), the resulting Bayes factor E^* is an *e*-variable. We then present Theorem 2 which shows that, for priors $W[\delta]$ with more than 2 moments, E^* in fact even is $\mathcal{W}_1$-GROW with $\mathcal{W}_1$ the set of all product priors on $\delta \times \sigma$ with marginal $W[\delta]$ on δ, i.e. it has a worst-case optimal growth rate property relative to all distributions in $\mathcal{H}_1$ compatible with $W[\delta]$. Thus, a form of GROW-optimality holds for most priors $W[\delta]$ one might want to use, including standard choices (such as a standard normal) or the point prior we will suggest further below — but we do not know if it holds for the moment-less Cauchy proposed by Jeffreys.

Almost Bayesian Case: prior on δ available

For any proper prior distribution $W[\delta]$ on δ and any proper prior distribution $W[\sigma]$ on σ, we define $p_{W[\delta],W[\sigma]}(y) = \int_{\delta\in\Delta}\int_{\sigma\in\Gamma} p_{\delta,\sigma}(y)\,\mathrm{d}W[\delta]\,\mathrm{d}W[\sigma]$, as the Bayes marginal density under the product prior $W[\delta]\times W[\sigma]$.

For convenience later on we set the sample space to be $\mathcal{Y}^n = (\mathbb{R}\setminus\{0\})\times\mathbb{R}^{n-1}$, assuming beforehand that the first outcome will not be 0. Now we define $\mathbf{V} := (V_1,\ldots,V_n)$ with $V_i = Y_i/|Y_1|$. We have that $\mathbf{Y}$ determines $\mathbf{V}$, and $(\mathbf{V}, Y_1)$ determines $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$. The distributions in $\mathcal{H}_0\cup\mathcal{H}_1$ can thus alternatively be thought of as distributions on the pair $(\mathbf{V}, Y_1)$. $\mathbf{V}$ is "$\mathbf{Y}$ with the scale divided out": as is well-known (**??**) and easily shown (Appendix B.3), under all $P\in\mathcal{H}_0$, i.e. all $P_{0,\sigma}$ with $\sigma>0$, $\mathbf{V}$ has the same distribution $P_0[\mathbf{V}]$ with density p_0'. In the same way, one shows that under all $P_{W[\delta],\sigma}$ with $\sigma>0$, $\mathbf{V}$ has the same pdf $p'_{W[\delta]}$ (which therefore does not depend on the prior on σ). We now get that, with

$$E^* := \frac{p'_{W[\delta]}(\mathbf{V})}{p_0'(\mathbf{V})}, \tag{28}$$

we must have $\mathbf{E}_{\mathbf{V}\sim P}[E^*] = 1$ for all $P\in\mathcal{H}_0$, hence it is an e variable. Remarkably, this 'scale-free' e-variable coincides with the Bayes factor one gets if one uses, for σ, the prior $w^H(\sigma) = 1/\sigma$ suggested by Jeffreys, and treats σ and δ as independent. That is (**?**, page 273) (a full derivation is in Appendix B.3), we have

$$\frac{\int_\sigma p_{W[\delta],\sigma}(\mathbf{Y})\,w^H(\sigma)\,\mathrm{d}\sigma}{\int_\sigma p_{0,\sigma}(\mathbf{Y})\,w^H(\sigma)\,\mathrm{d}\sigma} = \frac{p'_{W[\delta]}(\mathbf{V})}{p_0'(\mathbf{V})} = E^*. \tag{29}$$

Despite its improperness, w^H induces a valid e-variable when used in the Bayes factor. The equivalence of this Bayes factor to E^* simply means that it manages to ignore the 'nuisance' part of the model and models the likelihood of the scale-free $\mathbf{V}$ instead. The reason this is possible is that w^H coincides with the right-Haar prior for this problem (**??**), about which we will say more below. Amazingly, it turns out that the e-variable (29) has a GROW property (among all e-variables for data $\mathbf{Y}$, not just the coarsened $\mathbf{V}$) under the weak condition that the prior $W[\delta]$ has a $(2+\epsilon)$th moment. This follows from a special case of Theorem 4.2. of (**?**) (for the case that $W[\delta]$ puts all its mass on a single δ) and Corollary 8.3. (for general $W[\delta]$. For convenience we re-state this special case here. Let, for priors $W[\delta], W[\sigma]$, $P^{[\mathbf{V}]}_{W[\delta],W[\sigma]}$ be the marginal distribution with density $p'_{W[\delta],W[\sigma]}$. We have:

Theorem 2 [Special case of Theorem 4.2./Corollary 8.3. of ?] Let $W[\delta]$ be a distribution on δ such that $\mathbf{E}_{\delta\sim W[\delta]}[|\delta|^{2+\epsilon}] < \infty$ for some $\epsilon>0$ (in particular this includes all degenerate priors with mass 1 on a single δ). Let $\mathcal{W}[\Gamma]$ be the set of all probability distributions $W[\sigma]$ on the variance σ. Let $\mathcal{W}_1$ be the set of all product distributions on $\delta\times\sigma$ such that, for each $W'\in\mathcal{W}_1$, δ and σ are independent and its marginal on δ, i.e. $W'[\delta]$, coincides with $W[\delta]$. We have:

$$\inf_{W\in\mathcal{W}_1}\inf_{W[\sigma]\in\mathcal{W}[\Gamma]} D(P_W\|P_{0,W[\sigma]}) = \inf_{W[\sigma],W'[\sigma]\in\mathcal{W}(\Gamma)} D(P_{W[\delta],W[\sigma]}\|P_{0,W'[\sigma]}) = D(P^{[\mathbf{V}]}_{W[\delta]}\|P^{[\mathbf{V}]}_0). \tag{30}$$

The theorem allows us to use Theorem 1 as above with constant $f(\delta,\sigma)=0$ (note that $\mathcal{W}_1$ is convex) to conclude that E^* as in (28) is equal to E^f as in (27): the Bayes factor based on the right Haar prior, is not just an e-variable, but is even GROW relative to the set of all priors on $\delta\times\sigma$ that are compatible with $W[\delta]$.

REGROW-GROW safe t-test with minimum effect sizes

Suppose we want to test Θ_1 vs. Θ_0 as in (23) with fixed effect sizes δ^+ and δ^- and with σ^2 in the role of γ. We proceed exactly as we did underneath (23): we first consider the test $\{(\delta^+,\sigma^2):\sigma^2>0\}$ vs. $\{(\delta^-,\sigma^2):\sigma^2>0\}$ for the fixed given δ^+ using the REGROW criterion with $f((\delta,\sigma)) = \mathrm{GRO}(\delta,\sigma)$. We have (**?**, Section 4.3) that $f(\delta^+,\sigma) =$

$(n/2)\log(1+\delta^{+2})$ is constant on σ. Therefore we can use Theorem 1 in its most general form above in combination with Theorem 2 (applied with point prior $W[\delta]$ on δ^+) to conclude that (both the GROW and) the REGROW e-variable are given by $E^*_{\delta^+} := p'_{\delta^+}(\mathbf{V})/p'_{\delta^-}(\mathbf{V})$. Since Proposition 3 is applicable to sets of distributions defined on $\mathbf{V}$ rather than $\mathbf{Y}$ (details in Appendix B.3), we find that, with $\mathcal{E}_{\geq\delta^+} := \{E^*_\delta : \delta \geq \delta^-\}$ that $\sup_{E\in\mathcal{E}_{\geq\delta^+}} \inf_{\sigma>0,\delta\geq\delta^+} \mathbf{E}_{\mathbf{Y}\sim P_{\delta,\sigma}}[\log E] = \inf_{\sigma>0,\delta\geq\delta^+} \mathbf{E}_{\mathbf{Y}\sim P_{\delta,\sigma}}[\log E^*_{\delta^+}]$ so $E^*_{\delta^+}$ may be thought of as first applying REGROW, to get rid of the nuisance parameter, and then applying GROW — just like in the 2×2 Example 7.

Extension to General Group Invariant Bayes Factors

In a series of papers (**???**), Berger and collaborators developed a theory of Bayes factors for $\mathcal{H}_0 = \{P_{0,\gamma} : \gamma \in \Gamma\}$ and $\mathcal{H}_1 = \{P_{\delta,\gamma} : \delta \in \Delta, \gamma \in \Gamma\}$ with a nuisance parameter (vector) γ that appears in both models and that satisfies a group invariance; the Bayesian t-test is the special case with $\gamma = \sigma, \Gamma = \mathbb{R}^+$ and with the scalar multiplication group and δ an 'effect size'. Other examples include regression based on mixtures of g-priors (**?**), testing a Weibull vs. the log-normal and many more (**?**). The reasoning of the first part of this section straightforwardly generalises to all such cases: the Bayes factor based on using the right Haar measure on γ in both models gives rise to an e-variable. Theorem 4.2. of **?** shows that, if the underlying group satisfies a condition called amenability (which holds, e.g., for scaling as in the t-test, but also for e.g. rotations and affine transformations as in parametric linear regression models), then the resulting Bayes factor is GROW relative to a suitably defined set $\mathcal{W}_1$. Theorem 2 above is the very special case of their result when instantiated with γ instantiated to the variance in the t-test (scaling). Although its proof is quite different, the general result may be viewed as the 'e-variant' of the classical Hunt-Stein theorem (**?**, Section 8.5), with 'power' in that theorem replaced by 'GROW'. (**?**, Proposition 4.4) then implies that in all such cases, this GROW Bayes factor is in fact also REGROW. Remarkably therefore, with parameters representing group transformations, unlike e.g. for the 2×2 case, GROW and REGROW e-variables generally coincide.

5 | (RE)GRO(W), OPTIONAL CONTINUATION AND STOPPING

We now address two related questions:

1. We focused on Type-I error safety under optional continuation (OC). Can we also get safety under optional stopping (OS), and what is the difference?
2. The GRO-criteria were chosen to optimise expected capital (logarithmic) growth 'locally', within a study. How well do GRO-criteria go together with OC over several studies?

To make the questions concrete, we consider a specific set-up with data stream $Y_{j,1}, Y_{j,2}, \ldots$ corresponding to the j-th study to be performed. In the first study we observe batch of outcomes $\mathbf{Y}_{(1)} = (Y_{1,1}, \ldots, Y_{1,N_{(1)}})$; in the second study (if it is performed at all) $\mathbf{Y}_{(2)} = (Y_{2,1}, \ldots, Y_{2,N_{(2)}})$; and so on. For further simplicity we will assume that all $Y_{j,i}$ are i.i.d. The set-up slightly differs from Example 3 in which the second study's data was part of the same stream as the first; at the expense of additional notation, everything that follows can be formalized in that setting as well.

The conditional e-variables determining our test martingale are now determined by a sequence of stopping times $N_{(1)}, N_{(2)}, \ldots$. The first stopping time $N_{(1)}$ is defined as a stopping time on the first sequence $Y_{1,1}, Y_{1,2}, \ldots$ relative to filtration $(\sigma(Y_1^n))_n$ and defines a stopped σ-algebra $\mathcal{F}_{(1)} := \sigma(Y_1^{N_{(1)}}) = \sigma(\mathbf{Y}_{(1)})$. $N_{(2)}$ is defined on the second sequence $Y_{2,1}, Y_{2,2}, \ldots$, but it is also allowed to depend on previous data $\mathbf{Y}_{(1)}$, i.e it is a stopping time relative to filtration $(\sigma(\mathbf{Y}_{(1)}, Y_2^n))_n$, and defines a stopped σ-algebra $\mathcal{F}_{(2)} := \sigma(\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)})$. In general, $N_{(m)}$ is defined relative to filtration $\sigma(\mathbf{Y}_{(1)}, \ldots, \mathbf{Y}_{(m-1)}, Y_m^n)_n$, and defines a stopped σ-algebra $\mathcal{F}_{(m)} := \sigma(\mathbf{Y}_{(1)}, \ldots, \mathbf{Y}_{(m)})$. We let $\mathcal{N}_{(m)}$ be the collection of all stopping times for the m-th study, i.e. relative to $\sigma(\mathbf{Y}_{(1)}, \ldots, \mathbf{Y}_{(m-1)}, Y_{m,1}, \ldots, Y_{m,n})_n$. The sequence of stopped

σ-algebras thus defines a new filtration $(\mathcal{F}_{(m)})_m$, which we call the filtration at the study level, since $\mathcal{F}_{(m)}$ denotes all information available after m studies or trials have been completed — it is the filtration referred to in Proposition 2 (for the t-test we need to extend this set-up a little, see Example 8).

OC vs. OS

In the optional continuation setting, we assume that, after having observed and analyzed $m-1$ studies, we either stop, or continue to the next study. In the latter case we need to specify a stopping time $N_{(m)} \in \mathcal{N}_{(m)}$ and a conditional e-variable $E_{(m)} \in \mathcal{E}_{(m)}$ where $\mathcal{E}_{(m)}$ is the set of all $\mathcal{F}_{(m-1)}$-conditional e-variables. For example, in the setting of Example 1, with a simple null $\mathcal{H}_0 = \{P_0\}$ and composite alternative $\mathcal{H}_1 = \{P_\theta : \theta \in \Theta_1\}$ and $N_{(m)} \in \mathcal{N}_{(m)}$ an arbitrary stopping time for the m-th study, $e_{N_{(m)},W}(Y_{m,1},\ldots,Y_{m,N_{(m)}}) := p_W(Y_{m,1},\ldots,Y_{m,N_{(m)}})/p_0(Y_{m,1},\ldots,Y_{m,N_{(m)}})$ is an e-variable for every 'prior' distribution W on Θ_1 (here we generalize the notation of Example 1 to allow for data-dependent stopping times). In this setting the analyst can freely choose (or somebody else can impose) any $N_{(m)}$ and any W and use the corresponding e-variable $e_{N_{(m)},W}$; all such e-variables are contained in the set $\mathcal{E}_{(m)}$. As explained in Example 1, this includes the choice to set $W_{(m)} := W_{(1)}$ (re-use the original prior) and the choice to set $W_{(m)} := W_{(1)}(\cdot \mid \mathbf{Y}^{(m-1)})$, i.e. use the Bayes posterior based on the previous studies.

We can now interpret our various GRO criteria as each providing a prescription to choose specific elements of $\mathcal{E}_{N_{(m)}}$, for all stopping times $N_{(m)}$ that are constant given the outcomes of previous studies $\mathbf{Y}^{(m-1)}$, i.e. that are $\mathcal{F}_{(m-1)}$-measurable. To see this, note that each GRO criterion in combination with $\mathcal{H}_0$ and $\mathcal{H}_1$ defines, for each n, an e-variable $S^{[n]}$ for a single sequence $Y_1,\ldots,Y_n$. In all cases, this can be written as $S^{[n]} = s^{[n]}(Y^n)$ for some function $s^{[n]}$; the function $s^{[n]}$ is what is really specified; we call $s^{[1]}, s^{[2]},\ldots$ an e-specification. If $N_{(m)} := n_m$, we can thus simply set $E_{(m)} := s^{[n_m]}(Y_{m,1},\ldots,Y_{m,n_m})$. We call this the plug-in method for constructing $E_{(m)}$ from specification $s^{[1]}, s^{[2]},\ldots$ (such specification may correspond to one of our GRO criteria, but in general it could be arrived at in different ways as well) . The running product of $E_{(m)}$ thus constructed provides, via Proposition 2, a test martingale at the study level.

In contrast, optional stopping scenarios usually concern only a single data-level process $Y_1, Y_2,\ldots$, without any super-structure in terms of subsequent studies. The scenario is completely determined by a sequence of conditional e-variables $S_1, S_2,\ldots$, i.e., applying Definition 1 at the data level, such that for all $n \in \mathbb{N}$, for all $P \in \mathcal{H}_0$: $\mathbf{E}_P[S_n \mid \mathcal{F}_{n-1}] \leq 1$ a.s., with $\mathcal{F}_n = \sigma(Y^n)$. Their running product $S^{[1]}, S^{[2]},\ldots$ (with $S^{[n]} = \prod_{i=1}^n S_i$) then forms a test martingale. Recall from Corollary 1 in Section 1.3 that for any nonnegative random process $E^{(1)}, E^{(2)},\ldots$ at the study level (adapted to $(\mathcal{F}_{(m)})_m$) we say that the corresponding threshold test is safe under OC (with respect to Type-I error) if the Ville-Robbins inequality (7) holds. Extending this definition in the natural way, we say, for a data-level process of nonnegative random variables $S^{[1]}, S^{[2]},\ldots$, i.e. with $S^{[j]}$ adapted to $\mathcal{F}_j$, that the corresponding threshold test is safe under OS (with respect to Type-I error) if again the Ville-Robbins inequality holds (with $E^{(n)} := S^{[n]}$).

Sequentially Decomposable e-Specifications

In many (not all) cases, the GRO-specification $s^{[1]}, s^{[2]},\ldots$ forms itself a test martingale relative to some filtration $(\mathcal{G}_n)_n$: there exist a sequence of functions $s_1, s_2,\ldots$, with s_i a function on $\mathcal{Y}^i$, such that, for all n, $s^{[n]}(Y^n) = \prod_{i=1}^n S_i$ with $S_i := s_i(Y^i)$ and $\{S_i\}_i$ is a conditional e-variable collection relative to filtration $(\mathcal{G}_n)_n$. We will say that such an e-variable specification is sequentially decomposable, or seqdec for short, relative to filtration $(\mathcal{G}_n)_n$; in all our examples except the t-test (Example 8) we can take $\mathcal{G}_n = \sigma(Y^n)$. Seqdec specifications have a direct link with the OS setting, resulting in three remarkable properties: first, assuming still that data are i.i.d., any study-level test martingale process we can construct via the plug-in method (see above) based on a seqdec specification also defines a sequence of conditional e-variables and hence a test martingale at the corresponding concatenated data level $Y'_1, Y'_2,\ldots$ where Y'_i is arrived at by relabeling $Y_{1,1}, Y_{1,2},\ldots,Y_{1,N_{(1)}}, Y_{2,1}, Y_{2,2},\ldots,Y_{2,N_{(2)}}, Y_{3,1},\ldots$ in order (so that e.g. $Y'_{N_{(1)}+N_{(2)}} = Y_{2,N_{(2)}}$).

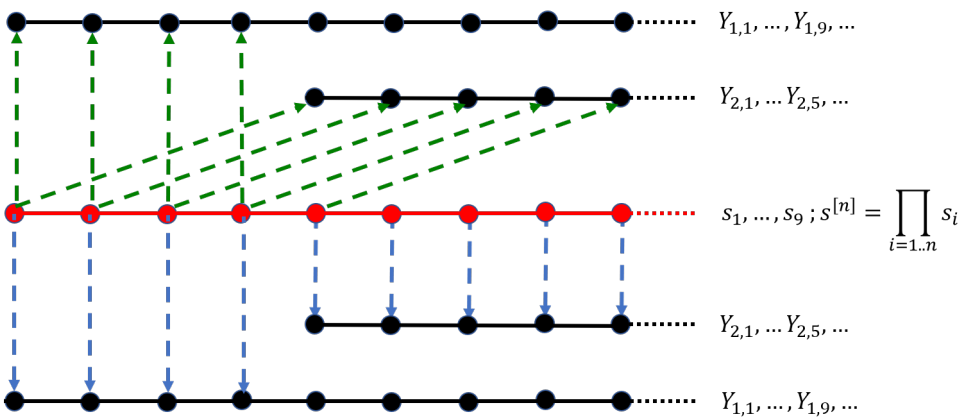

FIGURE 3 Two ways of applying seqdec e-variable specifications to subsequent studies. We observe $N_{(1)} = 4$ data points in the first study (represented by the top and bottom-most line), and $N_{(2)} = 5$ data points in the second (represented by the 2nd and 4th line above). Since the specification is seqdec, it provides a sequence of functions s_i on Y^i, represented by the dots on the red-line. The plug-in application uses $s^{[4]}$ for the first batch and $s^{[5]}$ applied to the second data batch, as depicted in the top two lines. The sequential application uses $s^{[4]}$ for the first and $\prod_{i=1}^{5} s_{N_{(1)}+i}$ for the second, as depicted on the bottom two lines. If the specification is not seqdec, then $s^{[n]}$ does not decompose into a product of s_i and in general only the plug-in application can be used.

This defines a concatenated data-level filtration $(\mathcal{G}'_t)_t$ with $\mathcal{G}'_t = \sigma(Y'_1, \ldots, Y'_t)$. The corresponding sequence of conditional *e*-variables is then given by $S'_1, S'_2, \ldots$ where, for $T_{(0)} := 0$, $m \in \mathbb{N}$, $1 \leq i < N_{(m)}$, $T_{(m)} := \sum_{j=1}^{m} N_{(j)}$, we set $S'_{T_{(m)}+i} := s_i(Y_{m,1}, \ldots, Y_{m,i})$. This means that besides engaging in optional continuation, we can also safely do optional stopping at this concatenated-data level, since the Ville-Robbins inequality holds at this level by Corollary 1.

Second, we can use seqdec specifications to extend the plug-in method (which required constant stopping times $N_{(m)}$) to prescribe conditional E-variables for stopping times $N_{(m)}$ that are not constant given $\mathbf{Y}^{(m-1)}$: for arbitrary $N_{(m)} \in \mathcal{N}_{(m)}$, we set $E_{(m)} := \prod_{i=1}^{N_{(m)}} s_i(Y_{m,1}, \ldots, Y_{m,i})$. By construction, this reduces to the plug-in method whenever $N_{(m)}$ is constant given $\mathbf{Y}^{(m-1)}$, so it is a proper extension, and it follows as a direct corollary of the fact that $E_{(m)}$ can be rewritten as $\prod_{i=1}^{N_{(m)}} S'_{T_{(m-1)}+i}$, i.e. a product of factors in a test martingale, that any $E_{(m)}$ constructed in this manner for any seqdec specification is a $\mathcal{F}_{(m-1)}$-conditional *e*-variable.

Third, seqdec specifications allow for alternative ways to create study-level processes from *e*-specifications beyond the plug-in method used thus far. For example, we can set $E_{(m)} := \prod_{i=1}^{N_{(m)}} s_{T_{(m-1)}+i}(\mathbf{Y}^{(m-1)}, Y_{m,1}, \ldots, Y_{m,i})$ for arbitrary stopping times $N_{(1)}, \ldots, N_{(m)}$. Once again, this also defines a martingale at the concatenated data-level — it is simply the martingale that arises if we view the m studies as one single, long sequence of $T_{(m)}$ data points. We call this the sequential application of the *e*-variable specification — see Figure 3.

Example 8 All GRO-type specifications based on a simple $\mathcal{H}_0 = \{P_0\}$ are likelihood ratios $s^{[i]} = q(Y^i)/p_0(Y^i)$, and hence will be seqdec and can thus be combined with optional stopping. In Example 1, choosing $W_{(m)} := W_{(1)} \mid \mathbf{Y}^{(m-1)}$ to be the Bayesian posterior corresponds to the sequential application of the $W_{(1)}$-GRO specification of Section 2.2; using $W_{(m)} := W_{(1)}$ corresponds to the plug-in application of the $W_{(1)}$-GRO application. This illustrates that we may think of our GRO criteria not as prescribing a single choice $E_{(m)} \in \mathcal{E}_{N_{(m)}}$, but rather as suggesting to choose $E_{(m)}$ from a preferred subset $\mathcal{E}'_{N_{(m)}}$ of $\mathcal{E}_{N_{(m)}}$; the end-user may then pick any e-variable in $\mathcal{E}'_{N_{(m)}}$. For example, in the case of Example 1 with simple $\mathcal{H}_0$, we may further specify a set of distributions $\mathcal{W}_1 \mid \mathbf{Y}^{(m-1)}$ on Θ_1 that we deem

'reasonable' given previous outcomes $\mathbf{Y}^{(m-1)}$, which may include the full Bayesian posterior, the originally used prior, combinations of these, tempered posteriors and so on; and we may then suggest the set $\mathcal{E}'_{N_{(m)}}$ of all e-variables for the m-th study based on a prior in $\mathcal{W}_1 \mid \mathbf{Y}^{(m-1)}$.

For the (RE)GRO(W)-specifications with composite null, one immediately verifies that those of Proposition 3 and Example 4 also are seqdec, since they equal the likelihood ratio between the same two distributions P_{δ^+} and P_{δ^-} irrespective of n. The t-test GROW/REGROW e-variables for arbitrary prior $W[\delta]$ on δ as in (28) are seqdec as well, but to formalize this statement we have to extend the setting. In general, the seqdec definition makes sense for every filtration $(\mathcal{G}_n)_n$ with $\mathcal{G}_n = \sigma(V^n)$ where $V_n = v_n(Y^n)$ for some sequence of functions $v_1, v_2, \ldots$ defined on $\mathcal{Y}^1, \mathcal{Y}^2, \ldots$ respectively. The previous definition is the special case with $v_n(Y^n) = Y_n$. In the t-test example we can take V_n as in Section 4.3 such that $v_n(Y_1, \ldots, Y_n) = Y_n/|Y_1|$. Let us illustrate how, with this coarser filtration, we can still apply the plug-in method for non-constant stopping times $N_{(m)}$. For this, we also have to coarsen the filtrations relative to which the $N_{(m)}$ are defined : the set of allowed stopping times $\mathcal{N}_{(m)}$ is now restricted to lie in $(\mathcal{F}_{(m-1)} \cup \sigma(V_{m,1}, \ldots, V_{m,n}))_n$ with $V_{j,n} = v_{j,n}(Y_{j,1}, \ldots, Y_{j,n})$ for some collection of functions $(v_{j,n})_n$ (recall that before they were members of $(\mathcal{F}_{(m-1)} \cup \sigma(Y_{m,1}, \ldots, Y_{m,n}))_n$) . For the t-test example we set $v_{j,n}(Y_{j,1}, \ldots, Y_{j,n}) = Y_{j,n}/|Y_{j,1}|$. The study-level filtrations $\mathcal{F}_{(1)} := \sigma(Y_{(1)}), \mathcal{F}_{(2)} := \sigma(Y_{(1)}, Y_{(2)}), \ldots$ remain unchanged and do not hide any information in the $Y_{(j)}$. In practice the restriction of $\mathcal{N}_{(m)}$ will not be of much concern since 'most' stopping times are still allowed, including the most aggressive stopping rule: stop the m-th study at the smallest n' such that $\prod_{i=1}^{n'} s_i(V_{m,i}) \geq 1/\alpha_{(m)}$, where $\alpha_{(m-1)}$ is some threshold that is allowed to depend on $\mathbf{Y}^{(m-1)}$. We can also allow for the sequential (rather than plug-in) application of the t-test e-variable specification so that effectively we view all studies as subsequent outcomes of a single study, by restricting the filtrations in a slightly different way; we omit the details. We can even let the choice between a sequential or plug-in choice for $E_{(m)}$ depend on past data, but this requires further generalizations of the $\mathcal{F}_m$ and $(v_{j,n})_n$ definitions that we shall not pursue here.

Summarizing, the practical setting we have in mind when we speak about OS and OC respectively is quite different: OC concerns study-level martingales constructed by deciding, on the fly , after the $m-1$-st study, whether to continue to the m-th study and if so, what new $\mathcal{F}_{(m-1)}$-conditional e-variable to take from the set $\mathcal{E}_{(m)}$ of possible e-variables of use. OS is about data-level martingales with only a stop/continue choice. Nevertheless, the formal definitions of (Type-I error) 'safety under OS' and 'safety under OC' only differ in that 'study-level' is replaced by 'data-level'. We may say that combining e-variables $E_{(m)}$ by multiplication is always Type-I error safe under OC. If the e-variable prescription used to construct $E_{(m)}$ has the seqdec property, then the stopping times $N_{(m)}$ used in each study do not need to be specified before the study starts and can even be externally imposed, so that we have Type-I error safety not just under OC but also under OS within each individual study.

There is one final subtlety to consider: in the OS setting, with a single stream of data $Y_1, Y_2, \ldots$ and conditional e-variables $S_1, S_2, \ldots$ and test martingale $S^{[1]}, S^{[2]}, \ldots$, the Ville-Robbins inequality (7) implies that our Type-I error bound α is guaranteed no matter when we stop — in particular, the actual stopping time does not have to be taken relative to the filtration $(\mathcal{G}_n)_n$ — we may even peek into the future to decide whether to stop now. This suggests that our care in specifying the correct filtrations for the t-test was unnecessary — it seems we can use any stopping rule we like! But this becomes incorrect once we move from OS at the data-level to OC at the study-level: if, in the t-test setting with the plug-in construction of the e-variable $E_{(m)}$ for the m-th study, we were to set the $N_{(m)}$ so that they are not stopping times relative to $(\mathcal{F}_{(m-1)} \cup \sigma(V_{m,1}, \ldots, V_{m,n}))_n$ but only relative to the more refined $(\mathcal{F}_{(m-1)} \cup \sigma(Y_{m,1}, \ldots, Y_{m,n}))_n$, we could end up creating fake conditional e-variables at the study-level, i.e. so that $\mathbf{E}_{P_0}[E_{(m)} \mid \mathcal{F}_{(m-1)}] > 1$ for all m ((**?**, Appendix B) constructs such a random variable for the t-test). And then the Ville-Robbins inequality may not hold any more at the study-level, and we loose the Type-I error guarantee under

optional continuation.

Local vs. Global GRO

Now consider two data streams, $Y_{1,1}, \ldots, Y_{1,n_1}$ and $Y_{2,1}, \ldots, Y_{2,n_2}$ of fixed lengths n_1 and n_2. We may alternatively model these two streams as a single stream $Y'_1, \ldots, Y'_{n_1+n_2}$ of length $n_1 + n_2$. If we use an e-variable specification with the seqdec property to generate a study-level test martingale in the sequential way (i.e. not the plug-in way) described above, we will get that, $E^{(2)} = E_{(1)} \cdot E_{(2)}$ constructed for the first two data streams (with $E_{(1)} = \prod_{i=1}^{n_1} s_i(Y^i)$ and $E_{(2)} = \prod_{i=1}^{n_2} s_{n_1+i}(Y'^{n_1+i})$) coincides with $E'_{(1)} = \prod_{i=1}^{n_1+n_2} s^{[i]}(Y'_i)$ constructed for the single alternative stream. We may thus say that the sequential application of seqdec e-specifications is always coherent: applying the specification sequentially-'locally' (separately for both studies) or sequentially-globally (for the concatenated data viewed as one study) gives the same result. For example, the Bayesian W-GRO specification of Section 2.2, the GROW specification of Proposition 3 and Example 4 and the GROW and REGROW specifications in the t-test example are all seqdec and hence all have this coherence property when applied sequentially. Sometimes the sequential application of an E-prescription is not feasible or desirable; for example, not all details of previous data may be known. We may then prefer the plug-in application of the E-prescription. Unfortunately, the seqdec property is not sufficient to get coherence for the plug-in method: clearly, if we use the same prior $W_{(1)}$ as prior in the Bayesian Example 3 for the first and the second study, this leads to different $E^{(2)}$ and $E'_{(1)}$, the latter being equivalent to using the posterior of the first study as prior in the second. A sufficient condition for a plug-in application to satisfy coherence after all is that it satisfies both the seqdec property, and further that s_i, with $S_i = s_i(Y^i)$ as in the definition of seqdec, can be rewritten as $s_i(Y^i) = s'(Y_i)$ for a single function s', for all i. Then in fact the plug-in and the sequential application of the e-prescription will coincide, and coherence is guaranteed. This happens in the subset of our examples in which $s^{[i]}(Y^i)$ takes the form $q(Y^i)/p(Y^i)$ for the same p and q, for all i, as happens e.g. in Example 4.

An Open Question concerning GRO

In practice we may very well be in a situation in which OS at the data-level is desirable (see the next section for why it would be), so we want to use a seqdec specification, yet the GRO criterion we are interested in does not give one — Example 9 illustrates this for the 2×2 case. We may then try the following approach, which for simplicity we only describe for the REGROW criterion: let E_n^f be the REGROW e-variable (19) achieving (20) with $f(\theta) = \mathrm{GRO}(\theta)$ for samples of size n. We try to find a sequence of e-variables $E_1, E_2, \ldots$ such that E_i is a $\sigma(Y^{i-1})$-conditional e-variable for Y_i and the product e-variable $E^{[n]} := \prod_{i=1}^n E_i$ achieves (20) to within some fixed ϵ for all n larger than some minimal n_0, i.e.

$$\inf_{\theta \in \Theta_1} \left(\mathbf{E}_{\mathbf{Y} \sim P_\theta}[\log E^{[n]}] - \mathrm{GRO}(\theta) \right) \geq \inf_{\theta \in \Theta_1} \left(\mathbf{E}_{\mathbf{Y} \sim P_\theta}[\log E_n^f] - \mathrm{GRO}(\theta) \right) - \epsilon. \tag{31}$$

By construction, the sequence $E^{[1]}, E^{[2]}, \ldots$ is seqdec and allows for optional stopping, and if we can find $E_1, E_2, \ldots$ such that ϵ is small for all n larger than or equal to the n_0 corresponding to the smallest sample we'd ever be interested in analyzing, we can say that the full process (and not just an instance at a fixed n) is 'almost' REGROW in the desired sense.

Example 9 **?** successfully use this idea for the 2×2 model. We illustrate this confining ourselves for simplicity to a stream of paired data, i.e. $X_1 = a, X_2 = b, X_3 = a, X_4 = b, \ldots$. First, we note that directly applying the idea above will not work. To see this, consider the simple alternative $\Theta_1 = \{(\mu_{1|a}, \mu_{1|b})\}$. According to the composite null, the Y_i are i.i.d. Bernoulli with parameter $\mu \in [0,1]$, but the $\mathrm{GROW}(\Theta_1) = \mathrm{GRO}((\mu_{1|a}, \mu_{1|b}))$- e-variable for such a $\mathcal{H}_0$ and single data point Y_i is the trivial $E \equiv 1$. Thus we would get all E_i equal to 1, and zero growth. However, if we analyze

the data in batches of size 2, so set $Y_1' = (Y_1, Y_2), Y_2' = (Y_3, Y_4)$ and take as $E_i' = \mathrm{e}(Y_i')$ the (nontrivial) $(\mu_{1|a}, \mu_{1|b})$-GRO-*e*-variable for $Y_i' = (Y_{2i-1}, Y_{2i})$ then $E'^{[n]} = \prod_{i=1}^n E_i'$ is the $(\mu_{1|a}, \mu_{1|b})$-GRO-*e*-variable for all n: we have the seqdec property and coherence, both for the sequential and for the plug-in method of applying the $(\mu_{1|a}, \mu_{1|b})$-GRO-prescription. Now in practice we want to consider a composite alternative — say we consider the full alternative $\Theta_1 = [0,1]^2$. Then the REGROW prescription will not be seqdec - the prior W_1^* in (19) depends on the sample size n. However it turns out that for a particular choice of prior W (**?** find it to be the beta-prior with parameters $\alpha = \beta = 0.18$) we have the following: if, for all i, we take E_i' the $W \mid Y'^{i-1}$-GRO *e*-variable, $W \mid Y'^{i-1}$ being the posterior based on Y'^{i-1}, then we numerically find that $E'^{[n]} = \prod_{i=1}^n E_i'$ is, for all but the smallest n, very close to the REGROW *e*-variable E_n^f for that n, i.e. it achieves (31) for small ϵ.

The example raises an important question: under what conditions (on model, minimal batch sizes and the like) can we create a seqdec specification that behaves optimally for our desired GRO criterion (as in Section 2.2 with W-GRO, and in Example 4, with the GROW criterion) or almost optimally (as in the 2×2 example above with batch size 2, with the REGROW criterion)?

6 | COMPETITIVENESS: GRO AND POWER

What sample size should we minimally plan for in a study so that we may expect a useful result? The answer depends on whether one looks at *e*-values purely as measures of evidence, without an accept/reject decision attached, or whether one considers such decisions after all. In the latter case, we can ask, more generally, how competitive e-value based tests are, in terms of required sample size, compared to the standard fixed-sample size Neyman-Pearson approach. We consider the cases without and with accept/reject decisions in turn.

e-Values as Evidence

e-values may be viewed simply as a measure of evidence, extending the evidential interpretation of likelihood ratios (**?**). They are then certainly competitive in every sense: for simple $\mathcal{H}_0$ they coincide with likelihood ratios and Bayes factors, and will give thus as much evidence as these notions do; for composite $\mathcal{H}_0$, GRO(W) *e*-variables are designed to give as much expected log-evidence against $\mathcal{H}_0$ as possible without violating the optional continuation requirement — in practice in some cases giving a bit more, and in some cases a bit less evidence against the null than standard Bayes factors (see **?** for a practical example).

Now suppose we have a minimal effect size δ in mind, and we plan a study in which obtaining data is expensive. What sample size should we plan for? One option is to pick a certain target growth L (essentially the logarithm of **?**'s notion of "implied target") and determine the sample size at which we expect to gain L. To illustrate, consider the 1-dimensional exponential family case of Example 4 with $\Theta_0 = \{0\}$. We know that, for a sample of size n , under all $\theta_1 \in \Theta_1$ with $\Theta_1 = \{\theta_1 : \theta_1 \geq \delta\}$, we have $\mathrm{GROW}(\Theta_1) = nD(P_\delta \| P_0)$ where $D(P_\delta \| P_0)$ is the KL divergence for 1 outcome. We then calculate n_{GROW} as the smallest n such that $nD(P_\delta \| P_0) \geq L$, i.e. $n_{\mathrm{GROW}} = \lceil L / D(P_\delta \| P_0) \rceil$. In the Gaussian location model, $D(P_\delta \| P_0) = \delta^2/2$, so $n = \lceil 2L/\delta^2 \rceil$. We return to the question of choosing L below.

e-Values for Decisions

We can also use *e*-values in the traditional setting, in which a study ends with an accept/reject decision — with the proviso that any decision is provisional, since there always is an option to continue and combine the results with a new study. For better or worse, this is the paradigm that researchers often have to work in, and within this paradigm they will inevitably be interested in the power for the experiment ahead. They will then plan for a certain sample

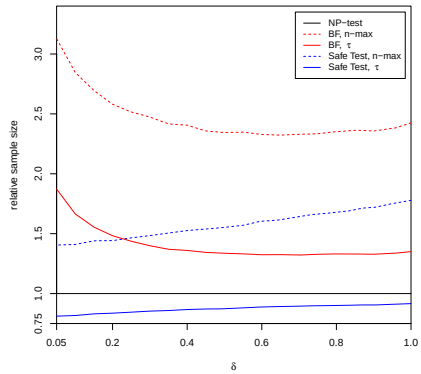

FIGURE 4 Factor of additional data needed to obtain a power of $1-\beta = 80\%$ compared to a Neyman-Pearson z-test, as a function of effect size (mean) for the Bayesian e-variable as in Example 3 with normal prior and the GROW e-variable with minimum relevant effect size δ as in Example 4 under optional stopping, both in expectation ($\tau_{0.8}$) and in worst-case ($n_{\max}(0.8)$) (for very small and large δ, the normal prior we used for the Bayesian e-variable may not be the prior of choice, but the results are representative for other priors one might use at $\delta \approx 0.5$).

size n to achieve such power, with a minimal relevant effect size δ in mind. As long as the e-variables themselves are chosen according to a GRO criterion, such a use of power as a 'secondary' criterion used merely to determine sample size is consistent with the GRO approach. In order for e-variables to be embraced by practitioners, we would hope that the sample sizes required to achieve a certain desired power with GRO-e-variables would be competitive with the standard approach based on Neyman-Pearson tests. We now study whether this is the case. For simplicity we only consider the Gaussian location model of Example 3, where $\mathcal{H}_0$ is the standard normal $N(0,1)$ and $\mathcal{H}_1$ the set $\{P_\mu : \mu \in \Theta\}$ of normals with variance 1. All results readily generalise to 1-dimensional exponential families.

Power: planning for a Fixed n

For comparison, recall that a standard one-sided NP test at level α would reject if $\hat{\mu} \geq z_\alpha/\sqrt{n}$ with z_α the $(1-\alpha)$-quantile of the standard normal with $z_{0.05} = 1.645$, $z_{0.01} = 2.33$. By standard calculation (see Appendix B.4), under an alternative with mean $\geq \delta$, the sample size needed with this test to get power at least $1-\beta$ satisfies $n_{np} = C_{np}/\delta^2$ with $C_{np} = (z_\alpha + z_\beta)^2$; for $\alpha = 0.05, \beta = 0.2$ we get $C_{np} \approx 6.180$. For the same $\Theta_1 = \{\mu : \mu \geq \delta\}$, we can also calculate the sample size needed to get power $1-\beta$ using the GROW e-variable of Example 4. If we use a fixed sample size n, we reject if $\log p_\delta(Y^n)/p_0(Y^n) \geq -\log \alpha$. By a simple calculation, for $\alpha > 1/2$, the smallest n at which we have power at least $1-\beta$, is given by setting

$$n_{\text{grow-fixed}} = 2 \cdot \frac{-\log \alpha}{\delta^2} \cdot \left(1 + \frac{z_\beta}{z_\alpha}\right)^2 = c_\alpha n_{np} \text{ with } c_\alpha = \frac{2 \cdot (-\log \alpha)}{z_\alpha^2}. \tag{32}$$

We have $c_{0.05} \approx 2.2$; $c_{0.01} \approx 1.7$ and c_α very slowly converges to 1 in the limit $\alpha \downarrow 0$: up to a constant factor of about two we need the same amount of data as in a classical approach, and the width of the induced confidence interval is of the same order. We can therefore choose a GROW E^* that is qualitatively more similar to a standard NP test than a standard Bayes factor approach. Using instead a standard Bayesian prior W_1 on Θ_1 with the W_1-GRO e-variable has the advantage of not needing to specify any δ in advance, but the number of samples to get power $1-\beta$ is larger by a logarithmic factor (Appendix B.4).

The evidential target growth and the maximal power approach are not contradictory: for any particular choice of α and β, there is a choice of L such that the planned-for sample sizes become the same function of δ (but of course the L resulting from $\alpha = 0.05, \beta = 1 - 0.8$ will be just as arbitrary as these choices were in the first place).

Power with Optional Stopping — a Tragedy of the Commons?

No matter the considerable advantages of being safe under optional continuation, the factor of about 2 of extra data needed to get a desired power might scare away practitioners from adopting *e*-variables. The situation changes completely once one adopts optional stopping. As we saw above, many testing problems allow us to use *e*-variables that remain safe under optional stopping — and we can use the most aggressive stopping rule that stops as soon as either $E_n \geq -\log \alpha$ (and we reject) or a pre-set maximum $n_{\max}$ is achieved (and we reject if $E_{n_{\max}} \geq -\log \alpha$ and otherwise accept). A simple but quite accurate approximation of the resulting stopping time τ_1 for i.i.d. data in the GROW setting of Example 4, when setting $n_{\max}$ to ∞ is given by using Wald's equality in a manner first set out by **?**; **?** give details. It gives for data $Y_1, Y_2, \ldots \sim P_\delta$, that $\mathbf{E}_{P_\delta}[\tau_1] \approx (-\log \alpha)/D(P_\delta \| P_0)$ with $D(P_\delta \| P_0)$ the KL divergence for a single outcome. For the Gaussian location family $D(P_\delta \| P_0) = \delta^2/2$, and we get $\mathbf{E}_{P_\delta}[\tau_1] \approx 2(-\log \alpha)/\delta^2$. Comparing to (32), this gives that with $n_{\max} = \infty$ (so that the power of our test is 1), the expected stopping time will be in fact already smaller than the fixed stopping time we get with the Neyman-Pearson approach at power $1 - \beta$ set to 0.8. In practice we will choose $n_{\max}(\beta)$ to be the smallest number so that the overall procedure has power $1 - \beta$, resulting in a stopping time $\tau_\beta = \min\{\tau_1, n_{\max}(\beta)\}$. The expected stopping time is then really even smaller. Figure 4 demonstrates this for the Gaussian location family. As the figure illustrates for the case $\beta = 0.8$, we have $n_{\max}(\beta) = C_{\beta,\delta}/\delta^2$ and $\mathbf{E}[\tau_\beta] = C'_{\beta,\delta}/\delta^2$ for $C_{\beta,\delta}$ and $C'_{\beta,\delta}$ that remain within constant bounds as δ varies. We can in fact heuristically derive analytic expressions (integrals) for the limits $C_{\beta,0}$ and $C'_{\beta,0}$ for $\delta \downarrow 0$ by rescaling the log-likelihood process $(\log p_\delta(X^n)/p_0(X^n))_n$ to become compatible with a Brownian motion with drift, see Appendix B.4. These give, for $\beta = 0.2$, $C_{\beta,0} = 8.5936$ and $C'_{\beta,0} = 4.971$ in accordance with Figure 4. We performed experiments with the publicly (on CRAN) available R package (**?**) safestats that implements the e-value based z-test, logrank test (**?**), t-test (**?**) and the 2×2 contingency table test (**?**). Experiments within the latter three settings confirm the picture that arises from Figure 4: with *e*-variables based on optional stopping one needs on average less data to achieve a certain desired power, but one needs to prepare for more data in the worst-case. We tentatively conclude that if current standard null hypothesis tests were replaced by *e*-value-based tests, and the standard practice to determine study sizes were replaced by the one above, and the percentage of studies in which the alternative is true is not too small, the world would need on average about the same or even a bit less data than it does now, to reach substantially more robust conclusions and better meta-analyses. Yet — at least as long as scientists insist on power requirements — each individual study would have to plan for substantially more data, giving researchers an incentive not to adapt these new methods. We see this Tragedy of the Commons as one of the biggest obstacles for uptake of *e*-variables in practical settings.

7 | EARLIER AND RELATED WORK

***e*-Variables, Test Martingales, Information Projections, General Novelty**

As seen in Section 1.3, *e*-variables are the building blocks of test (super-) martingales, which go back to **?**. E-variables themselves have probably been originally introduced by Levin (of *P* vs *NP* fame) (**?**) (see also (**?**)) under the name test of randomness, but Levin's abstract context is quite different from ours. Independently discovered by **?** (under the name PBR (prediction-based ratio)) they were later analyzed by **?** (calling them, with hindsight confusingly, Bayes factors); **??** (e-variables/values) and **?** (bets/betting scores) — we originally called them S-values ourselves. The literature seems to converge to *e*-variables and -values. Here the *e* may either stand for evidence or for expectation.

Test martingales themselves have been thoroughly investigated by **??**. They themselves underlie AV (anytime-valid) p-values (**?**), AV tests (which we call 'tests that are safe for optional stopping') and AV confidence sequences. The latter were recently developed in great generality by A. Ramdas and collaborators; see e.g. (**??**). Both AV tests and confidence sequences have first been developed by H. Robbins and his students (**???**). Like we do for e-variables, Ramdas et al. (and also e.g. **?**) stress the promise of AV notions for a safer kind of statistics that is significantly more robust than standard tests and confidence intervals.

Just like regular tests can be turned into confidence intervals by varying the null and 'inverting' the resulting tests, AV confidence sequences can be created by starting with a collection of test martingales, one for each null, and then varying the null (it is sometimes claimed that problematic aspects of null hypothesis testing are mostly due to the very idea of a 'null hypothesis' or a significance level (**??**). Without wanting to take sides in this issue, we note that standard confidence intervals are just as unsafe under optional continuation as standard Neyman-Pearson hypothesis tests). The work on AV tests and confidence sequences is therefore very similar in spirit to ours, with our work stressing analysis at the level of batches of data rather than individual data points, and with the AV work bringing out the difference to Bayesian approaches more explicitly (AV $1-\alpha$-confidence intervals are typically wider than Bayesian $1-\alpha$-credible sets). In fact we do not claim any real novelty for the 'safe' or 'AV' setting per se: the real novelty of this paper is in the four versions of Theorem 1. As far as we know, these results are new, with the exception of a special case of the simplest version of Theorem 1 (Section 2): the case of discrete outcome spaces, simple $\mathcal{H}_1$ and convex $\mathcal{H}_0$ was already formulated and proved by **?**. Theorem 1 heavily builds on properties of standard- and reverse and joint information projections about which there is a rich literature, key references being **????**. Both the standard and the reverse information projection are special cases of F-divergence projections, investigated in great detail in both the information-theoretic community (a pioneering paper being (**?**)) and the mathematical finance community (**??**); the robust optimization problems in the latter two papers, when instantiated to logarithmic utility, bear some resemblance to our GROW criterion. While the details and the motivation are quite different, it would be of interest to study the connections further.

Relation to Sequential Testing

Sequential testing (**??**), pioneered by **?** and Barnard, is mathematically very similar to, but conceptually quite different from, testing based on test martingales and (therefore) e-variables. Sequential tests are made for streams of data $Y_i, Y_2, \ldots$ as in Example 1 and Section 5 and are based on random processes $(S_i)_{i\in\mathbb{N}}$ such that, for each i, S_i is a conditional e-variable given $Y_1, \ldots, Y_{i-1}$ under $\mathcal{H}_0$, and $1/S_i$ is a conditional e-variable given $Y_1, \ldots, Y_{i-1}$ under $\mathcal{H}_1$. Of course, this two-sided e-variable property only holds in quite special cases — roughly under the same conditions as Proposition 3 (monotone likelihood ratio), i.e. in our Example 4 and for the t-test with point prior on δ^+, δ^-. In such a setting, the sequential test based on $S_1, S_2, \ldots$ with prespecified parameters α, β proceeds by calculating $S_1, S_2, \ldots$ and stopping at τ^*, the smallest τ at which either $S_\tau \geq (1-\beta)/\alpha$ ('accept') or $S_\tau \leq (1-\alpha)/\beta$ ('reject'). Wald showed that this test has Type I error probability bounded by α and Type II error bounded by β. The reason one can stop at a smaller threshold ($(1-\beta)/\alpha$ rather than $1/\alpha$) is that one has to stop at τ^*, Thus, the method does not allow for optional stopping in our sense: conceptually, sequential tests were designed for special, pre-specified stopping times. Still, much work in sequential testing can be re-cycled to obtain test martingales and e-values — but not always vice versa since e-variables are often not 'two-sided'.

Related Work on Relating p-values and e-variables

? and **?** give a general formula for calibrators f (see also **?** and **??** for early work in this direction). These are decreasing functions $f : [0,1] \to [0,\infty]$ so that for any p-value p, $E := f(p)$ is an e-variable. The choice of any such calibrator is

essentially arbitrary, but, following **?**, let us consider one that is especially simple: $f(p) = 1/\sqrt{p} - 1$. For example, for any calibrator f suggested for practice, rejection under the e-variable based test with significance level $\alpha = 0.05$, so that $E \geq 20$, would then correspond to reject only if $p \leq f^{-1}(20) = 1/441 = 0.0023$, requiring a substantial amount of additional data for rejection under a given alternative. Note that the e-variables we developed for given models in previous sections are more sensitive than such generic calibrators though. For example, consider the normal location family of the previous section. With the calibrator above, we would reject if $\hat{\mu} \geq z_{0.0023}/\sqrt{n} \approx 2.8/\sqrt{n}$. The amount of data to plan for to obtain power 80% would then be $\approx (2.8/1.65)^2 n_{np} \approx 3.0 n_{np}$, whereas for the e-value based on the normal likelihood ratio we would need $\approx 2.2 n_{np}$, and even significantly less under optional stopping.

8 | GRO: DISCUSSION AND OPEN PROBLEMS

In this paper we provided several motivations for our various GRO criteria as we introduced them, in Section 2–4. Here we reflect on the strength of our arguments in sequence-of-studies settings when taking the running products of the per-study GRO e-variables. Let us first consider a simple alternative Q as in Section 2.1, so that GROW and REGROW criteria coincide with GRO. In an optional continuation (OC) context, the justification of GRO is strongest if all study outcomes $\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}, \ldots$ are independent and the variances of $\log E_{(m)}$ under Q for the GRO e-variables $E_{(m)}$ are not too large. As pointed out by a referee, if these variances are large, there could be prolonged periods of 'drawdowns' — in the gambling interpretation, independence ensures, by the law of large numbers, that asymptotically the GRO e-variables maximize our capital; but if there is high variance, there may be several studies in a row during which the product of all e-values so far remains low, a fact well-known among economists. Moreover, i.i.d. data and the seqdec property are required for another central justification of GRO in the optional stopping setting, namely Breiman's insight that the expected stopping time before one can reject is minimized by maximizing expectation of the log capital (**?** explain this in detail). This issue may perhaps in some cases be resolved by instead of adopting the 'global' GRO e-variable (among all e-variables for the null), taking an alternative e-variable that is GRO among a subset of e-variables with sufficiently low variance under Q. Such low variance e-variables would presumably look quite different from likelihood ratios and Bayes factors though, since for given Θ_0, the only e-variable that can be obtained by projecting on $\mathcal{W}(\Theta_0)$ is the unconstrained GRO e-variable: e-variables that are optimal in a non-GRO sense cannot be obtained by projection, even if KL is replaced by another distance or divergence D': Corollary 2 implies that, with $W_0^* = \arg\min_{W) \in \mathcal{W}(\Theta_0)} D'(Q; P_W)$, we have that $E' = q(\mathbf{Y})/p_{W_0^*}(\mathbf{Y})$ does not give an E-variable unless W_0^* also minimizes the KL divergence, i.e. if E' is an E-variable at all, it must be GRO.

In case $\mathcal{H}_1$ is composite, then even if the $\mathbf{Y}_{(j)}$ are independent and the variances are low, there may be alternatives to GROW and REGROW that are sometimes preferable. For example, if we are in an OS setting and we use our tests as the basis for always-valid confidence intervals (see the previous section), we may want to aim for the conditional e-variables that lead to the confidence intervals that shrink to 0 at the fastest possible rate, which for regular 1-dimensional parameters of interest is usually $O((\log\log n)/n)$ (**?**). These are not obtained by REGROW e-variables (which, by extending the reasoning of Example 5, in parametric problems achieve a width of $O((\log n)/n)$). The narrower $O((\log\log n)/n)$ can be achieved by the switching strategies of **?** or the stitching method used to design test martingales by **?**. A more precise understanding of whether such methods can also be re-understood as optimizing a variation of GRO, and more generally what meta-GRO criteria are reasonable at all, and in what situations, is needed. This includes the question of when any variation of GRO automatically provides seqdec, or close-to-seqdec specifications, allowing us to engage in OS (Example 9). Answering these questions is a major avenue for future research: it ultimately determines how widely applicable GRO criteria really are.

9 | COULD FISHER, JEFFREYS AND NEYMAN HAVE AGREED ON A CURRENCY FOR TESTING?

The three main approaches towards null hypothesis testing are Jeffreys' Bayes factors, Fisher's p-value-based testing and the Neyman-Pearson method. In the paper Could Fisher, Jeffreys and Neyman Have Agreed on Testing?, **?** noted that, while these methodologies seem superficially highly contradictory, there exist methods that have a place within all three. The developments in this paper lead to the conclusion that e-variable based testing — although it differs in some technical respects from Berger's proposals — is very much in the same spirit:

Concerning the Neyman-Pearson approach: e-variables lead to tests with Type-I error guarantees at any fixed significance level α, which is also the first requirement of a Neyman-Pearson test — requiring safety under optional continuation or optional stopping simply enforces the requirement to hold over a non-pre-specified sequence of studies, which is a natural requirement in scientific applications. Since there is then no single study any more, the concept of 'power' looses its centrality (and may be upgraded to requiring power one), and growth-rate optimality is a natural quantitative refinement. The fact that a high growth-rate corresponds to a high value of $\mathbf{E}_{P_1}[\log E]$ under $\mathcal{H}_1$ whereas a high power corresponds to a high probability that $P_1(\log E \geq -\log \alpha)$ also shows that GRO and power remain intimately connected in e-variable theory as well.

Concerning the Fisherian approach: here, p-values are interpreted as indicating amounts of evidence against the null, and their definition does not need to refer to any specific alternative $\mathcal{H}_1$. Exactly the same holds for e-values: the basic interpretation 'a large e-value provides evidence against $\mathcal{H}_0$' holds no matter how the e-variable is defined, as long as it satisfies (1). If they are defined relative to $\mathcal{H}_1$ that is close to the actual process generating the data they will grow fast and provide a lot of evidence, but the basic interpretation holds regardless. In contrast to evidence based on standard p-values however, (a) e-based evidence has a concrete additional interpretation in terms of money (the higher E, the more money one has gained in a game that is not favourable under the null); (b) it remains valid under optional continuation, and (c) unlike the p-value, it is compatible with (provides the same evidence as) likelihood ratios do in simple-vs.-simple testing — the one case where the use of likelihood ratio as evidence is standard.

Concerning the Bayesian approach: despite their monetary interpretation, all e-variables that we encountered can also be written as Bayes factors, and Theorem 1 strongly suggests that this is a very general phenomenon. Subjective prior knowledge can be accounted for using the W_1-GRO e-variable (Section 2.2), whereas maximin optimal GROW and REGROW e-variables sometimes correspond to 'objective' Bayes approaches based on Jeffreys' and/or right-Haar priors. Still, there seem to be two fundamental differences: first, in a standard Bayesian analysis, one would require error guarantees and safety under OC under the prior instead of under all $P \in \mathcal{H}_0$, and second, one would insist on using full, standard likelihoods — whereas e-variables may also be based on partial (**?**) or Dawid's (**?**) prequential (**?**) likelihoods rather than full likelihoods — which then however may be combined with priors (on $\mathcal{H}_1$) after all. Even though we emphasise Type-I error safety throughout, because of this generic freedom in using priors on $\mathcal{H}_1$ the link to Bayesian methods remains close.

The Dream

With the massive criticisms of p-values in recent years, there seems to be growing consensus that, in the context of hypothesis testing, p-values should either not be used at all, or at least, with utter care (**??**). Yet otherwise, the disputes among adherents of the three schools continue. For example, some highly accomplished statisticians reject the idea of testing without a clear alternative outright; others say that such goodness-of-fit tests are an essential part of data analysis. Some insist that significance testing (with binary decisions) should be abolished altogether (**?**), others (perhaps slightly cynically) acknowledge that significance may be silly in principle, yet maintain that journals and conferences

will always require a significance-style 'bar' in practice and that therefore such bars should be made as meaningful as possible. Finally, within the Bayesian community, the Bayes factor is sometimes presented as a panacea for most testing ills, while others warn against its use, protesting, for example, against claims that Bayes factors can 'handle optional stopping' (**?**). *Wouldn't it be nice if all these accomplished but disagreeing people could continue to go their way, yet would have a common language or 'currency' to express amounts of evidence, and would be able to combine their results in a meaningful way?* This is what *e*-variables can provide: consider three tests with the same null hypothesis $\mathcal{H}_0$, based on samples $\mathbf{Y}_{(1)}$, $\mathbf{Y}_{(2)}$ and $\mathbf{Y}_{(3)}$ respectively. The results of a GROW *e*-variable test aimed to optimise power on sample $\mathbf{Y}_{(1)}$ for $\delta \geq \delta^+$, an *e*-variable test for sample $\mathbf{Y}_{(2)}$ based on a Bayesian prior W_1 on $\mathcal{H}_1$ and a Fisherian *e*-variable test in which the alternative $\mathcal{H}_1$ is not explicitly formulated, can all be multiplied — and the result will be meaningful, both in terms of monetary gain and in terms of error probability.

Acknowledgments

Many thanks to A. Barron, J. Berger, R. Frongillo, P. Harremoës, A. Hendriksen, A. Henzi, A. Ly, R. Meester, M. Perez, A. Ramdas, J. ter Schure, G. Shafer, R. Turner, V. Vovk, B. Wagonner, J. Ziegel who all made helpful remarks.

Web-based supporting materials for *Safe Testing* by Grünwald, De Heide and Koolen

A | THEOREM 1, COROLLARIES AND CONDITIONS

Here we prove Theorem 1 and its Corollary 2. We also, in Appendix A.3, discuss the required regularity conditions for Theorem 1 and we prove that they are applicable in all our examples.

A.1 | Proof of Theorem 1, Simplest Version, and Corollary 2

The proof of the first version of Theorem 1 relies on quite technical results from **?**, but if the minimum in (11) is achieved by some prior W_0^*, and under the further condition that we can exchange differentiation and expectation, then the partial and crucial result that (10) is an e-variable has a very simple proof, which we provide first as a 'warm-up': evaluate the derivative $f(\alpha) = (d/\alpha)D(Q\|(1-\alpha)P_{W_0^*} + \alpha P_\theta)$ at $\alpha = 0$ for arbitrary $\theta \in \Theta_0$ and note that it is ≥ 0 iff $\mathbf{E}_{\mathbf{Y}\sim\mathbf{Q}}[p_\theta(\mathbf{Y})/p_{W_0^*}(\mathbf{Y})] = \mathbf{E}_{\mathbf{Y}\sim\mathbf{P}_\theta}[q(\mathbf{Y})/p_{W_0^*}(\mathbf{Y})] \leq 1$. Differentiating again gives that $f(\alpha)$ is convex, and the result follows from convexity of $\mathcal{P}_0 := \{P_W : W \in \mathcal{W}(\Theta_0)\}$.

We proceed to give the complete and fully general proof. Note that $\mathcal{P}_0$ is convex, and (by assumption of the theorem) every distribution in $\mathcal{P}_0$ as well as Q has a density relative to μ and $\inf_{P\in\mathcal{P}_0} D(Q\|P) < \infty$. These three givens allow us to use a range of results about the reverse information projection (RIPr) established in the Ph.D. thesis (**?**) (additional proofs of (extensions of) all of Li's results we need below can be found in the refereed paper **?**).

First, the existence and uniqueness of a measure P_0^* (not necessarily a probability measure) with density p_0^* that satisfies $D(Q\|P_0^*) = \inf_{P\in\mathcal{P}_0} D(Q\|P)$ (i.e. it is the RIPr), and furthermore has the property

$$\text{for all } p \text{ that are densities of some } P \in \mathcal{P}_0\text{: } \mathbf{E}_{\mathbf{Y}\sim Q}\left[\frac{p(\mathbf{Y})}{p_0^*(\mathbf{Y})}\right] \leq 1, \tag{33}$$

follows directly from (**?**, Theorem 4.3). But by writing out the integral in the expectation explicitly we immediately see that we can rewrite (33) as:

$$\text{for all } P \in \mathcal{P}_0\text{: } \mathbf{E}_{\mathbf{Y}\sim P}\left[\frac{q(\mathbf{Y})}{p_0^*(\mathbf{Y})}\right] \leq 1.$$

Li's Theorem 4.3 still allows for the possibility that $\int p_0^*(y)\,\mathrm{d}\mu(y) > 1$. To see that in fact this is impossible, i.e. p_0^* defines a (sub-) probability density, use Lemma 4.5 of **?**. This shows that $E^* = q(\mathbf{Y})/p_0^*(\mathbf{Y})$ is an e-variable, and (using that P_0^* is the RIPr) the second and third equality of (11). The final line of the result ('if ... then $P_0^* = P_{W_0^*}$') follows directly from Lemma 4.1 of **?**.

It remains to show the first equality of (11) and essential uniqueness of E^*. For the former, it is sufficient to show that for all e-variables, i.e. all $E \in \mathcal{E}(\Theta_0)$,

$$\mathbf{E}_{\mathbf{Y}\sim Q}[\log E] \leq \mathbf{E}_Q\left[\log E^*\right]. \tag{34}$$

To show this, fix any e-variable $E = e(\mathbf{Y})$ in $\mathcal{E}(\Theta_0)$. Now further fix $\epsilon > 0$ and fix a $W_{(\epsilon)} \in \mathcal{W}(\Theta_0)$ with $D(Q\|P_{W_{(\epsilon)}}) \leq \inf_{W_0\in\mathcal{W}(\Theta_0)} D(Q\|P_{W_0}) + \epsilon$. We must have, with $q'(y) := e(y)p_{W_{(\epsilon)}}(y)$, that $\int q'(y)\,\mathrm{d}\mu = \mathbf{E}_{\mathbf{Y}\sim P_{W_{(\epsilon)}}}[E] \leq 1$, so q' is

a sub-probability density, and by properness of the log scoring rule,

$$\mathbf{E}_Q[\log E] = \mathbf{E}_Q\left[\log \frac{q'(\mathbf{Y})}{p_{W_{(\epsilon)}}(\mathbf{Y})}\right] \leq \mathbf{E}_Q\left[\log \frac{q(\mathbf{Y})}{p_{W_{(\epsilon)}}(\mathbf{Y})}\right] = D(Q\|P_{W_{(\epsilon)}}) \leq \inf_{W_0 \in \mathcal{W}(\Theta_0)} D(Q\|P_{W_0}) + \epsilon.$$

Since we can take ϵ to be arbitrarily close to 0, (34) follows.

To show essential uniqueness of E^*, let E be any e-variable with $\mathbf{E}_Q[\log E] = \mathbf{E}_Q[\log E^*]$. By linearity of expectation, $E' = (1/2)E^* + (1/2)E$ is then also an e-variable, and by Jensen's inequality applied to the logarithm we must have $\mathbf{E}_Q[\log E'] > \mathbf{E}_Q[\log E^*]$ unless $Q(E = E^*) = 1$. Since we have already shown that for any e-variable E', $\mathbf{E}_Q[\log E'] \leq \mathbf{E}_Q[\log E^*]$, it follows that $Q(E \neq E^*) = 0$. But then, by our assumption that Q has full support, i.e. $q(\mathbf{Y}) > 0$ hold μ-almost everywhere, we must have that $P_\theta(E \neq E^*) = 0$ for all $\theta \in \Theta$, so E^* is essentially unique.

Proof of Corollary 2

Let W_0 be as in the corollary statement. By definition of E^* as in Theorem 1, simplest version, and then using strict convexity of the KL divergence in its second argument (**?**) and the fact that $D(Q\|P)$ is minimised, over $P \in \{P_W : W \in \mathcal{W}(\Theta_0)\}$, we have:

$$\mathbf{E}_Q\left[\log E^*\right] = D(Q\|P_0^*) < D(Q\|P_{W_0}) = \mathbf{E}_Q\left[\log \frac{q}{p_{W_0}}\right]$$

so that, if q/p_{W_0} were an e-variable, we would have a contradiction with the first equality in (11).

A.2 | Proof of full version of Theorem 1

The proof consists of two sub-parts, Part (a) relying on the simple version of the theorem presented in Section 2 and proven above (henceforth called 'the simple theorem'), and Part (b) relying on a nonstandard minimax/saddle-point theorem from **?** (GD from now on), itself relying heavily on an earlier result from **?**.

Part (a). We first show that E^f as in (26) is an e-variable. This follows by the simple theorem, with $\mathbf{V}$ in the role of $\mathbf{Y}$, Q in the theorem statement substituted by $P^{[\mathbf{V}]}_{W_1^*}$ and P_W for $W \in \mathcal{W}(\Theta_0)$ replaced by $P^{[\mathbf{V}]}_W$ and using that (25) implies that $\inf_{W_0 \in \mathcal{W}_0(\Theta_0)} D(P^{[\mathbf{V}]}_{W_1^*}\|P^{[\mathbf{V}]}_{W_0}) < \infty$. Next, we show that <u>if</u> (27) holds for E^f as in (26), then all e-variables E for which it holds must be essentially equal to E^f. To see this, suppose that E is another e-variable satisfying (27). Then we have

$$\inf_{W \in \mathcal{W}_1}\left(\mathbf{E}_{\mathbf{Y}\sim P_W}[\log E] - f(W)\right) = D(P^{[\mathbf{V}]}_{W_1^*}\|P_0^{*[\mathbf{V}]}) - f(W_1^*) = \mathbf{E}_{\mathbf{Y}\sim P_{W_1^*}}\left[\log E^f\right] - f(W_1^*)$$

as follows by writing out the definition of $D(\cdot\|\cdot)$. On the other hand, using the definition of inf, we must have

$$\inf_{W \in \mathcal{W}_1}\left(\mathbf{E}_{\mathbf{Y}\sim P_W}[\log E] - f(W)\right) \leq \mathbf{E}_{\mathbf{Y}\sim P_{W_1^*}}[\log E] - f(W_1^*).$$

The only way these two displays can be reconciled is if $\mathbf{E}_{\mathbf{Y}\sim P_{W_1^*}}[\log E] \geq \mathbf{E}_{\mathbf{Y}\sim P_{W_1^*}}\left[\log E^f\right]$. But since E^f is the RIPr of $P^{[\mathbf{V}]}_{W_1^*}$, we can use the simple theorem again, applied with $P_{W_1^*}$ in the role of Q and $\mathcal{H}_1 = \{P_{W_1^*}\}$, to conclude that E^f must be essentially equal to E. The final line of the fully general Theorem 1 follows again by reduction to the analogous statement of the simple theorem.

Finally, we will use the simple theorem to show the following one-sided version of (27):

$$\inf_{W \in \mathcal{W}_1} \left(\mathbf{E}_{\mathbf{Y} \sim P_W} [\log E^f] - f(W) \right) \leq \sup_{E \in \mathcal{E}(\Theta_0)} \inf_{W \in \mathcal{W}_1} \left(\mathbf{E}_{\mathbf{Y} \sim P_W} [\log E] - f(W) \right) \leq D(P^{[\mathbf{V}]}_{W_1^*} \| P^{*[\mathbf{V}]}_0) - f(W_1^*). \tag{35}$$

The first inequality is trivial since $E^f \in \mathcal{E}(\Theta_0)$. The second follows if we can show that

$$\sup_{E \in \mathcal{E}(\Theta_0)} \inf_{W \in \mathcal{W}_1} \mathbf{E}_{P_W} [\log E - f(W)] \leq \inf_{W \in \mathcal{W}_1} \inf_{W_0 \in \mathcal{W}_0} \left(D(P_W \| P_{W_0}) - f(W_1) \right) \tag{36}$$

and recognizing that by assumptions of the theorem, the right-hand side of (36) coincides with the right-hand side of (35). To prove (36), note that by the simple version of the theorem we already have for each fixed $W_1 \in \mathcal{W}_1$ that

$$\inf_{W_0 \in \mathcal{W}_0} D(P_{W_1} \| P_{W_0}) = \sup_{E \in \mathcal{E}(\Theta_0)} \mathbf{E}_{P_{W_1}} [\log E]$$

and this directly implies the inequality by adding $-f(W_1)$ to both sides and using a standard "inf sup $\geq$ sup inf" argument (the trivial side of the minimax theorem). The equality follows by assumption of the Theorem.

Part (b). Taking stock, we see that the only thing that is left to prove is (35) with the reversed inequalities. For this, it suffices to show that

$$D(P^{[\mathbf{V}]}_{W_1^*} \| P^{*[\mathbf{V}]}_0) - f(W_1^*) \leq \inf_{W \in \mathcal{W}_1} \mathbf{E}_{P_W} [\log E^f - f(W)]. \tag{37}$$

Since all distributions occurring in (37) are marginals on $\mathbf{V}$, and E^f can be written as a function of $\mathbf{V}$, we will from now on simply refer to the marginal densities on $\mathbf{V}$ corresponding to P_W as p_W (rather than p'_W as in the main text), and we will omit the superscripts $[\mathbf{V}]$ from P; thus we take as our basic outcome now $\mathbf{V}$ rather than $\mathbf{Y}$.

We will show the stronger statement that (37) holds with equality, by using a minimax/saddle point result that holds for general functions $L : \Theta_1 \times \mathcal{W}_1 \to \mathbb{R} \cup \{\infty\}$ such that $L(W_\mathrm{n}, W_\mathrm{u}) := \mathbf{E}_{\theta \sim W_\mathrm{n}}[L(\theta, W_\mathrm{u})]$ is well-defined for all $W_\mathrm{n} \in \mathcal{W}_1$ (the condition 'well-defined' is necessary since the expectation is over a function that may neither be bounded from below nor from above; see for example Section 3.1 of GD for the (standard) definitions). These L are interpreted as loss functions, with $\theta_1 \in \Theta_1$ denoting a state of nature and $\mathcal{W}_1$ an arbitrary convex set of distributions on Θ_1, each $W \in \mathcal{W}_1$ being interpreted as an action. Following GD, we can associate a decision-theoretic entropy $H(W_\mathrm{n}) := \inf_{W_\mathrm{u} \in \mathcal{W}} L_0(W_\mathrm{n}, W_\mathrm{u}) = L(W_\mathrm{n}, W_\mathrm{n})$ with any such L. The following result holds for all Θ_1, $\mathcal{W}_1$ and L as defined above but we will apply it to the instantiation of Θ_1 and $\mathcal{W}_1$ in Theorem 1.

GD's Theorem 6.3

Assume that (a) L is a proper scoring rule, i.e. for all $W_\mathrm{n} \in \mathcal{W}_1$, $H(W_\mathrm{n}) = L(W_\mathrm{n}, W_\mathrm{n})$. Suppose that (b) $W_1^* \in \mathcal{W}_1$ is 'maximum entropy' i.e. $\sup_{W_\mathrm{n} \in \mathcal{W}_1} H(W_\mathrm{n}) = H(W_1^*) < \infty$ and (c) the lower semi-continuity condition below holds. Then (W_1^*, W_1^*) is a saddle-point relative to L, i.e.

$$H(W_1^*) = L(W_1^*, W_1^*) = \sup_{W \in \mathcal{W}_1} L(W, W_1^*). \tag{38}$$

Lower Semicontinuity Condition (GD's Condition 6.1)

Let (W_n) be a sequence of distributions in $\mathcal{W}_1$ such that $H(W_n)$ is bounded below and such that (W_n) converges weakly to some distribution W° on Θ_1. Then $L(W^\circ, W_{\mathrm{u}})$ is well-defined for all $W_{\mathrm{u}} \in \mathcal{W}$ and for all $W' \in \mathcal{W}$, $L(W', W^\circ) \leq \liminf_{n\to\infty} L(W', W_n)$.

We now define the specific loss function to which we will apply the above theorem. $\mathcal{W}_1$ and Θ_1 are defined as in the statement of Theorem 1. (25) implies that $D(P_{W_1^*} \| P_0^*) < \infty$ for some P_0^* with density p_0^*. Similarly, P_W must have some density p_W under all $W \in \mathcal{W}_1$. We can therefore define, using these densities,

$$L(\theta, W_{\mathrm{u}}) = \mathbf{E}_{\mathbf{V}\sim P_\theta}\left[-\log \frac{p_{W_{\mathrm{u}}}(\mathbf{V})}{p_0^*(\mathbf{V})} - f(\theta)\right]$$
$$L(W_{\mathrm{n}}, W_{\mathrm{u}}) = \mathbf{E}_{\theta\sim W_{\mathrm{n}}}\mathbf{E}_{\mathbf{V}\sim P_\theta}\left[-\log \frac{p_{W_{\mathrm{u}}}(\mathbf{V})}{p_0^*(\mathbf{V})} - f(\theta)\right].$$

Since P_θ has full support for all $\theta \in \Theta_1$, $P_{W_1^*}$ has full support and so $p_0^*(\mathbf{V}) > 0$ a.s. under $P_{W_1^*}$ and hence under all P_θ with $\theta \in \Theta_1$. Similarly $p_W(\mathbf{V}) > 0$ a.s. under all P_θ with $\theta \in \Theta_1$. Thus, the quantity inside the expectation is almost surely well-defined. To see that the expectations are themselves well-defined (using standard definitions, see again Section 3.1 of GD), note that we can write

$$L(W_{\mathrm{n}}, W_{\mathrm{u}}) = +\mathbf{E}_{W_{\mathrm{n}}}\mathbf{E}_{P_\theta}\left[-\log \frac{p_{W_{\mathrm{u}}}}{p_{W_{\mathrm{n}}}}\right]_+ + \mathbf{E}_{W_{\mathrm{n}}}\mathbf{E}_{P_\theta}\left[-\log \frac{p_{W_{\mathrm{n}}}}{p_0^*}\right]_+ - \mathbf{E}_{W_{\mathrm{n}}}\mathbf{E}_{P_\theta}\left[f(\theta)\right]_+$$
$$- \mathbf{E}_{W_{\mathrm{n}}}\mathbf{E}_{P_\theta}\left[-\log \frac{p_{W_{\mathrm{u}}}}{p_{W_{\mathrm{n}}}}\right]_- - \mathbf{E}_{W_{\mathrm{n}}}\mathbf{E}_{P_\theta}\left[-\log \frac{p_{W_{\mathrm{n}}}}{p_0^*}\right]_- + \mathbf{E}_{W_{\mathrm{n}}}\mathbf{E}_{P_\theta}\left[f(\theta)\right]_-$$

with $[x]_+ := \max\{x, 0\}$ and $[x]_- = \max\{-x, 0\}$. The expectation would be undefined iff there is both a term equal to ∞ and a term equal to $-\infty$ on the right. We will show that this is not the case. We assume $f(\theta)$ bounded, and, under our finite KL condition, $D(P_{W_{\mathrm{n}}} \| P_{W_{\mathrm{u}}}) = \mathbf{E}_{W_{\mathrm{n}}}\mathbf{E}_{P_\theta}\left[-\log \frac{p_{W_{\mathrm{u}}}}{p_{W_{\mathrm{n}}}}\right]_+ - \mathbf{E}_{W_{\mathrm{n}}}\mathbf{E}_{P_\theta}\left[-\log \frac{p_{W_{\mathrm{u}}}}{p_{W_{\mathrm{n}}}}\right]_- < \infty$, so we only need to worry about the terms involving p_0^*. But these are also the positive and negative parts of a minus KL divergence, so the full expectation is well-defined as a number in $\mathbb{R} \cup \{-\infty\}$. The expectations are therefore welldefined and we can write

$$L(W_{\mathrm{n}}, W_{\mathrm{u}}) = D(P_{W_{\mathrm{n}}} \| P_{W_{\mathrm{u}}}) - D(P_{W_{\mathrm{n}}} \| P_0^*) - \mathbf{E}_{\theta\sim W_{\mathrm{n}}}[f(\theta)] \tag{39}$$

and analogously for $L(\theta, W_{\mathrm{u}})$.

Applying GD's theorem to L

We apply GD's theorem to the loss function L above with W_1^* as in the statement of the theorem. From (39) we see that $L(W_1^*, W)$ is minimised, over $\mathcal{W}_1$, by $W = W_1^*$ and then finite, so that GD's requirements (a) and (b) hold for loss function L. We can now reason as follows. If the lower semicontinuity condition (c) also holds, then the theorem applies and (38) implies, taking minus on both sides,

$$-L(W_1^*, W_1^*) = \inf_{W\in\mathcal{W}_1} -L(W, W_1^*).$$

which, rewriting the left-hand side using (39) and the right-hand side using definition of L, is in turn seen to be equivalent to (37), and the desired result follows.

It thus only remains to show that the lower semicontinuity condition holds. Using (39) we can write $H(W_\mathrm{n}) = -D(P_{W_\mathrm{n}} \| P_0^*) + f(W_\mathrm{n})$ for all $W_\mathrm{n} \in \mathcal{W}_1$. Take a sequence $(W_n)_n$ and W° as in the condition. Then $(P_{W_n})_n$ converges weakly to P_{W° (this is easy to see but see the proof of Lemma 9.2. of GD for an explicit proof). Since also f is bounded, $f(W_n)$ converges to $f(W^\circ)$. Also, for some $K \in \mathbb{R}$, for all n, we have $H(W_n) \geq K$ so for some $K' \in \mathbb{R}$, by boundedness of f, we have, for all n, $D(P_{W_n} \| P_0^*) \leq K' < \infty$. By Posner's (**?**) theorem, $D(P_W \| P_0^*)$ is lower semi-continuous in its first argument. Posner only proves the result for P_0^* a probability measure; but it still holds even if P_0^* is a strict sub-probability measure, since then $p'(\mathbf{Y}) = p_0^*(\mathbf{Y}) / \int p_0^*(\mathbf{Y}) d\mu = 1$ represents a distribution P' and the result follows by applying Posner's result to P' and noting that $D(P_W \| P_0^*) = D(P_W \| P') + C$ for some constant C not depending on W.

The lower-semicontinuity in the first argument implies $D(P_{W^\circ} \| P_0^*) \leq \liminf_{n\to\infty} D(P_{W_n} \| P_0^*) \leq K' < \infty$. Following an argument exactly parallel to the proof of well-definedness for $L(W_\mathrm{n}, W_\mathrm{u})$ given above, it now follows that $L(W^\circ, W_\mathrm{u})$ is well-defined for all $W_\mathrm{u} \in \mathcal{W}_1$ as required. Next, using (39), we see that it is sufficient to show that for all $W \in \mathcal{W}_1$:

$$D(P_W \| P_{W^\circ}) \leq \lim \inf_{n\to\infty} D(P_W \| P_{W_n}).$$

But this again follows directly from Posner's theorem, which also says that KL divergence is lower semi-continuous in its second argument.

A.3 | Remarks on and Checking of Conditions for Theorem 1

The Full Support and Finite KL Condition

Requiring full support in the simplest version of the theorem ensures that E^* is a.s. well-defined: without it, there may be an outcome $\mathbf{y}$ such that for some $\theta \in \Theta_0$, $P_\theta(\mathbf{y}) > 0$ whereas $P_{W_0}(\mathbf{y}) = Q(\mathbf{y}) = 0$. Then E^* is undefined with positive probability under this θ. The finite KL condition $D(P_\theta \| P_{\theta'}) < \infty$ imposed in the generalised versions of the theorem is just slightly stronger than the full support condition. It is required to make sure that all expectations in the proof are well-defined.

For standard parametric models in standard parameterisations (e.g. all multivariate exponential families in their mean-value parameterisation), both conditions will hold automatically as long as one excludes points at the boundary of the parameter space, if those exist. For example, in the 2×2 setting without a pre-specified effect size we restrict Θ_1 to $(0,1)^2$, requiring the Bernoulli probabilities $\mu_{1|a}$ and $\mu_{1|b}$ to be non-degenerate. But, since the condition only refers to Θ_1, it is o.k. to set $\Theta_0 = [0,1]$ to include the boundary points in the null.

Additional Condition: Existence of W_1^*

The requirement for composite $\mathcal{H}_1$ that a W_1^* exists achieving the minimum in (27) is strong in general, but it holds in all our examples with composite $\mathcal{H}_1$: Example 4 (W_1^* is shown to be a point prior in the example), Example 5 (since there we restrict Θ_1 to be compact) and Example 6 and 7 (here verifying the condition requires some work, see below). It also holds in the t-test setting underneath Theorem 2 with effect sizes δ^+ and δ^- (W_1^* reduces to a point prior on δ^+). By allowing e-variables to be functions of $\mathbf{V}$ that are themselves functions of $\mathbf{Y}$ (i.e. $\sigma(\mathbf{Y})$-measurable) as in the latter example, we make the condition considerably weaker.

Applicability of Theorem 1 and existence of minimizing W_1^* and W_0^* in Example 6 and 7

We have $\mathcal{H}_1 = \{P_{\mu_{1|a},\mu_{1|b}} : (\mu_{1|a}, \mu_{1|b}) \in \Theta_1\}$ and $\mathcal{H}_0 = \{P_\mu : \mu \in \Theta_0\}$, $\Theta_0 = [0,1]$ with definitions as in Example 2. In Example 6, we take $\Theta_1 = (0,1)^2$ and we can take $\Theta_0 = [0,1]$ or $\Theta_0 = (0,1)$ (the same minima will be achieved in both

cases). In Example 7 we take $\Theta_1 = \{(\mu_{1|a}, \mu_{1|a} + \delta) : 0 < \mu_{1,a} < 1 - \delta\}$ and $\Theta_0 = (0,1)$.

We only give the proof for Example 6; the proof for Example 7 is entirely analogous.

The requirement for applying Theorem 1 that $P_\theta, P_{\theta'}$ with $\theta, \theta' \in \Theta_1$ satisfy $D(P_\theta \| P_{\theta'}) < \infty$, and have full support trivially holds by our exclusion of the boundary points in Θ_1. The only remaining condition for applying Theorem 1 is the existence of a KL minimizing prior W_1^*. We will show the stronger result that there exists a pair of minimizing priors (W_1^*, W_0^*) with $W_1^* \in \mathcal{W}(\Theta_1)$ and $W_0^* \in \mathcal{W}(\Theta_0)$ such that

$$\inf_{W_1 \in \mathcal{W}(\Theta_1), W_0 \in \mathcal{W}(\Theta_0)} (D(P_{W_1} \| P_{W_0}) - f(W_1)) = D(P_{W_1^*} \| P_{W_0^*}) - f(W_1^*) < \infty, \tag{40}$$

with $f(W) = \mathbf{E}_{(\mu_{1|a}, \mu_{1|b}) \sim W}[f(\mu_{1|a}, \mu_{1|b})]$ and $f(\mu_{1|a}, \mu_{1|b}) = D(P_{\mu_{1|a}, \mu_{1|b}} \| P_{\mu^\circ})$ with μ° as in (14). We do this by first, in Part (a), showing that there exists such a pair with $W_1^* \in [0,1]^2$, i.e. with Θ_1 extended to include its boundary points. We then, in Part (b), show that the resulting W_1^* puts no mass on these boundary points, so that it also achieves the minimum on $\mathcal{W}(\Theta_1)$.

Part (a) The sets $\mathcal{W}([0,1]^2)$ and $\mathcal{W}(\Theta_0)$ are convex and compact in the weak topology; by Posner's (**?**) theorem, $D(P_{W_1} \| P_{W_0})$ is lower-semicontinuous in its second argument in the weak topology on $\{P_{W_0} : W_0 \in \mathcal{W}(\Theta_0)\}$ and hence on $\mathcal{W}(\Theta_0)$ itself (see Section 9 of GD) and $f(W)$ is linear and bounded on $\mathcal{W}(\Theta_0)$; this shows that for each W_1, the corresponding minimizing W_0^* is achieved; since $D(P_{W_1} \| P_{W_0^*}) \leq D(P_{W_1} \| P_{1/2}) < \infty$ (with $P_{1/2} \in \mathcal{H}_0$ representing Bernoulli$(1/2)$) and f is bounded, the finiteness in (40) is guaranteed as well. To see that the minimum W_1 is achieved as well, note that, again by Posner's theorem, $D(P_{W_1} \| P_{W_0})$ is also lower-semicontinuous in its first argument in the weak topology on $\{P_{W_1} : W_1 \in \mathcal{W}([0,1]^2)\}$. The same argument as before now gives that the minimum W_1^* is achieved.

Part (b) It now suffices to show that $P_{W_1^*}$ has full support, for this implies that W_1^* assigns mass 1 on $\Theta_1 = (0,1)^2$ and then W_0^* must assign mass 1 on $(0,1)$ (otherwise the KL divergence in (40) would be infinite, and we already established it is not). To show full support of $P_{W_1^*}$, note that by symmetry considerations, we must have that, with our choice of f,

$$M := D(P_{W_1^*} \| P_{W_0^*}) - f(W_1^*) = D(P_{W_1^\circ} \| P_{W_0^\circ}) - f(W_1^\circ) \tag{41}$$

for a prior W_1° such that, for all $\mathbf{y} \in \{0,1\}^n$, $p_{W_1^\circ}(\mathbf{y} \mid \mathbf{x}) = p_{W_1^*}(\mathbf{y}' \mid \mathbf{x})$ with $\mathbf{y}'$ is the modification of $\mathbf{y}$ with all 0s and 1s interchanged, and similarly for W_0°. Now if $P_{W_1^*}$ would not have full support, we have $P_{W_1^*}(Y_1 = y_1, Y_2 = y_2 \mid X_1 = a, X_2 = b) = 0$ for some $(y_1, y_2) \in \{0,1\}^2$. Then $P_{W_1^\circ}(Y_1 = \bar{y}_1, Y_2 = \bar{y}_2 \mid X_1 = a, X_2 = b) = 0$ for $\bar{y}_j = 1 - y_j$. But $\inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1^*} \| P_{W_0}) - f(W_1^*)$ as a function of W_1^* is easily checked to be strictly convex on $\mathcal{W}(\Theta_1)$, so by (41) we must have that, for $W' = (1/2)W_1^* + (1/2)W_1^\circ$, it holds that $\inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{W'} \| P_{W_0}) + f(W') < M$. But this contradicts that M is the minimum. This shows that $P_{W_1^*}$ has full support.

Finiteness of Support in Example 6 and 7

We claimed that the supports of the priors W_1^* and W_0^* in Example 7 (restricted Θ_1) are finite. In fact they are finite also in Example 6 (unrestricted Θ_1). We verify this for W_1^*, the case for W_0^* is analogous. Note that for given sample size n, the probability distribution P_W is completely determined by the probabilities assigned to the sufficient statistics $N_{1|a}, N_{1|b}$. This means that for each prior $W \in \mathcal{W}(\Theta_1)$, the Bayes marginal P_W can be identified with a vector of $M_n := (n_a + 1) \cdot (n_b + 1)$ real-valued components. Every such P_W can also be written as a mixture of P_θ's for $\theta = (\mu_{a|1}, \mu_{b|1}) \in \Theta_1$, a convex set. By Carathéodory's theorem we need at most M_n mixture components to describe an arbitrary P_W as a mixture of the P_θ's; this proves the claim.

B | ADDITIONAL CLARIFICATIONS AND PROOFS

B.1 | Section 1.3

We discuss two extensions of the filtration to be used in Definition 1. For concreteness and simplicity we consider the sequential set-up of Section 5 in which, for each study m, there is an underlying data stream $Y_{m,1}, Y_{m.2}, \ldots$.

Filtration: Conditional Distributions

In the 2×2 setting the Θ_1 represent conditional distributions of $Y \mid X$. While neither Section 1.3 nor Section 5 formally allowed for that setting, the extension is straightforward. We simply assume the underlying streams are of the form $(X_{m,1}, Y_{m,1}), (X_{m,2}, Y_{m,2}), \ldots$ (with, in the 2×2 example, $X_{m,i} \in \{a, b\}$). The distributions in $\mathcal{H}_1$ are now extended to define a random process of independent outcomes with the same conditional distribution for a single such stream $(X_1, Y_1), (X_2, Y_2), \ldots$, i.e. for all $\theta_1 \in \Theta_1$, we set $p_{\theta_1}(y^n \mid x^n) = P_{\theta_1}(Y^n = y^n \mid X^n = x^n) := \prod_{i=1}^n p_{\theta_1}(y_i \mid x_i)$. We then only need to extend Definition 1 of conditional e-variables to deal with this extension. This is achieved by setting $\mathcal{F}_{(m-1)}$ in the definition to $\sigma(\mathbf{Y}^{(m-1)}, \mathbf{X}^{(m)})$.

Filtration: Side Information

Now we consider how the set-up can be extended to deal with side-information that may be used e.g. after j studies to decide whether to start a new, $j+1$st study at all, and if so, what the sample size of that study will be. For this we again need to extend $\mathcal{H}_0$ and $\mathcal{H}_1$ so that its elements define a conditional random process, with at the time that the j-th study has just been observed, also the additional variables $\mathbf{R}_{(1)}, \ldots, \mathbf{R}_{(j)}$ observed. Even if there are underlying streams of data $Y_{j,1}, Y_{j,2}, \ldots$ so that the $\mathbf{Y}_{(j)}$ have an internal structure, the $\mathbf{R}_{(j)}$ are not required to have such a structure. To make all desired probabilities well-defined and at the same time make sure that the side-information is really external, we impose a conditional independence on the underlying stream: under each $P \in \mathcal{H}_0$, for each $m, i \in \mathbb{N}$, the conditional distribution of $Y_{m,i}$ given Y_m^{i-1} and $\mathbf{R}^{(m-1)}$ is defined to be the same as the distribution under P of $Y_{m,i}$ given Y_m^{i-1}, which is already well-defined once $\mathcal{H}_0$ is specified. With this definition, we can set $\mathcal{F}_{(j)} = \sigma(\mathbf{Y}^{(j)}, \mathbf{R}^{(j)})$ (or $\mathcal{F}_{(j)} = \sigma(\mathbf{Y}^{(j)}, \mathbf{R}^{(j)}, \mathbf{X}^{(j+1)})$ if the $\mathcal{H}_j$ already contain conditional distributions, or $\mathcal{F}_{(j)} = \sigma(\mathbf{U}^{(j)}, \mathbf{R}^{(j)}, \mathbf{X}^{(j+1)})$ with $\mathbf{U}^{(j)}$ a coarsening of $\mathbf{Y}^{(j)}$ if required, such as in the t-test setting). Focusing on the simplest case with $\mathcal{F}_{(j)} = \sigma(\mathbf{Y}^{(j)}, \mathbf{R}^{(j)})$, the construction ensures that $\mathbf{E}_P[E_{(j)} \mid \mathbf{Y}^{(j-1)}, \mathbf{R}^{(j-1)}] \leq 1$ (and hence we have safety under OC) in both cases discussed in Section 5: we can either use the plug-in application of any e-variable specification, or, if the specification is seqdec, we can also use the sequential application. Note that with this construction, the $\mathbf{R}^{(j)}$ are allowed to depend on $\mathbf{Y}_{(1)}, \ldots, \mathbf{Y}_{(j)}$ in unspecified ways. This is unproblematic because (other than in the case with conditional distributions and $\mathbf{X}^{(j+1)}$) the $E_{(j+1)}$ cannot depend on $\mathbf{R}^{(j+1)}$. For example, based on what she sees in the $\mathbf{Y}_{(j)}$, your boss may decide to announce 'we have money to do an additional study with 100 patients', which can be encoded as a particular outcome of $\mathbf{R}_{(j)}$. This may then be used to decide to continue (i.e. set τ in Proposition 2 to be larger than j) and set $N_{(j+1)}$ to be 100.

B.2 | Section 3

Proof of Proposition 3

The monotone likelihood ratio property implies stochastic dominance (**?**), i.e. with $P[T]$ denoting the distribution of the statistic T, we must have $\mathbf{E}_{P_\delta[T]}[f(T)] \geq \mathbf{E}_{P_{\delta'}}[f(T)]$ for $\delta \geq \delta'$ and every increasing function f. This implies

that

$$D(P_{\delta^+}\|P_{\delta^-}) = \mathbf{E}_{\mathbf{Y}\sim P_{\delta^+}}\left[\log \frac{p_{\delta^+}(\mathbf{Y})}{p_{\delta^-}(\mathbf{Y})}\right] = \inf_{\delta\geq\delta^+} \mathbf{E}_{\mathbf{Y}\sim P_{\delta}}\left[\log \frac{p_{\delta^+}(\mathbf{Y})}{p_{\delta^-}(\mathbf{Y})}\right]. \tag{42}$$

We also have, by the same stochastic dominance result, for $\delta \leq \delta^-$,

$$\mathbf{E}_{\mathbf{Y}\sim P_{\delta}}\left[\frac{p_{\delta^+}(\mathbf{Y})}{p_{\delta^-}(\mathbf{Y})}\right] \leq \mathbf{E}_{\mathbf{Y}\sim P_{\delta^-}}\left[\frac{p_{\delta^+}(\mathbf{Y})}{p_{\delta^-}(\mathbf{Y})}\right] = 1,$$

so that $E^* = p_{\delta^+}(\mathbf{Y})/p_{\delta^-}(\mathbf{Y})$ is an e-variable, which directly leads to the first inequality in the chain of (in)equalities (43) below. The first equality follows by (42), the second because, since E^* is of form p_{δ^+}/p_{W_0}, with $W_0 \in \mathcal{E}(\Theta_0)$ (namely, W_0 puts all mass on δ^-), it must, by Corollary 2, be the GRO-e-variable for testing between $\Theta_1' = \{\delta^+\}$ and Θ_0. The final two inequalities are immediate:

$$\begin{aligned}&\sup_{E\in\mathcal{E}(\Theta_0)} \inf_{\delta\geq\delta^+} \mathbf{E}_{\mathbf{Y}\sim P_{\delta}}\left[\log E\right] \geq \inf_{\delta\geq\delta^+} \mathbf{E}_{\mathbf{Y}\sim P_{\delta}}\left[\log \frac{p_{\delta^+}(\mathbf{Y})}{p_{\delta^-}(\mathbf{Y})}\right] = \mathbf{E}_{\mathbf{Y}\sim P_{\delta^+}}\left[\log \frac{p_{\delta^+}(\mathbf{Y})}{p_{\delta^-}(\mathbf{Y})}\right] = \\ &\sup_{E\in\mathcal{E}(\Theta_0)} \mathbf{E}_{\mathbf{Y}\sim P_{\delta^+}}\left[\log E\right] \geq \inf_{\delta\geq\delta^+} \sup_{E\in\mathcal{E}(\Theta_0)} \mathbf{E}_{\mathbf{Y}\sim P_{\delta}}\left[\log E\right] \geq \sup_{E\in\mathcal{E}(\Theta_0)} \inf_{\delta\geq\delta^+} \mathbf{E}_{\mathbf{Y}\sim P_{\delta}}\left[\log E\right].\end{aligned} \tag{43}$$

This chain of inequalities implying that all its parts are equal, the result follows.

B.3 | Section 4.3

Proof of (29)

(29) follows from (**?**, page 273) or as special case of Theorem 2.1. of **?**, but the first proof leaves out details and the second is very abstract, so for convenience we give a direct proof. For simplicity restrict to the case with W putting all its mass on a particular δ. Fix arbitrary $\sigma > 0$ and $n \geq 2$ and note that $V_1 \in \{-1, 1\}$ and $P_\delta(V_1 = 1) = P_{\delta,\sigma}(Y_1 > 0)$ (note that $p'_{W[\delta]}(V_1)$ is a probability mass function, whereas $p'_{W[\delta]}(V_i \mid V^{i-1})$ is defined as density relative to Lebesgue measure for $i > 1$). We must then have:

$$\begin{aligned}&P_\delta(V_1 = 1)\cdot p'_\delta(v_2,\ldots,v_n \mid V_1 = 1) = P_\delta(V_1 = 1)\cdot \int_0^\infty p_{\delta/|y_1|,\sigma/|y_1|}(v^n \mid Y_1 = y_1) p_{\delta,\sigma}(y_1 \mid Y_1 > 0) dy_1 \\ &= \int_0^\infty \prod_{i=2}^n \left(\frac{1}{\sqrt{2\pi}}\cdot\frac{|y_1|}{\sigma}\cdot e^{-\frac{1}{2}\left(\frac{v_i}{\sigma/|y_1|}-\delta\right)^2}\right)\cdot\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{y_1}{\sigma}-\delta\right)^2}\right) dy_1 \\ &= \int_0^\infty \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi}}\cdot\frac{|y_1|}{\sigma}\cdot e^{-\frac{1}{2}\left(\frac{v_i}{\sigma/|y_1|}-\delta\right)^2}\right)\cdot\frac{1}{|y_1|} dy_1 \\ &= \int_0^\infty \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\tau}\cdot e^{-\frac{1}{2}\left(\frac{v_i}{\tau}-\delta\right)^2}\right)\cdot\frac{\tau}{\sigma}\left|\frac{dy_1}{d\tau}\right| d\tau = \int_0^\infty \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\tau}\cdot e^{-\frac{1}{2}\left(\frac{v_i}{\tau}-\delta\right)^2}\right)\cdot\frac{1}{\tau} d\tau\end{aligned}$$

Here in the first equality we used that, given $Y_1 = y_1$, the V_i are independent Gaussian, with variance $\sigma/|y_1|$ and mean $\delta\sigma/|y_1|$ hence effect size $\delta/|y_1|$ and hence density $p_{\delta/|y_1|,\sigma/|y_1|}$. The second equality replaces the conditional density of Y_1 by the marginal (so that $P_\delta(V_1 = 1)$ cancels) and exploits that $y_1 = v_1|y_1|$, the third is a standard change-of-variable from $\sigma/|y_1|$ to τ using the Jacobian transformation and the fourth is immediate.

This shows the desired result if $V_1 = 1$. The case for $V_1 = -1$ is analogous.

Applicability of Proposition 3 to the t-test Setting

(Section 4.3 underneath Theorem 2) A simple calculation gives that $p'_{\delta^+}(\mathbf{V})/p'_{\delta^-}(\mathbf{V})$ can be re-expressed as a density ratio of the t-statistic $T = t(\mathbf{Y})$, i.e. $p'_{\delta^+}(\mathbf{V})/p'_{\delta^-}(\mathbf{V}) = p''_{\delta^+}(t(\mathbf{Y}))/p'_{\delta^-}(t(\mathbf{Y}))$, where p''_δ is the density of a noncentral t-distribution with $\nu := n-1$ degrees of freedom and noncentrality parameter $\mu = \sqrt{n}\delta$. But these densities are well-known to form a monotone likelihood ratio family in the T statistic, so that we can apply Proposition 3 to $\{p''_\delta : \delta \in \Delta\}$.

B.4 | Section 6

Determining Sample Size for a Desired Power

Consider a 1-sided test for the normal location family with variance 1 which rejects if $\hat{\mu} \geq f(n)/\sqrt{n}$ where $\hat{\mu}$ is the MLE at sample size n and f is some increasing function of n. We want to find the smallest n at which we achieve power $1-\beta$ under mean δ, i.e. such that

$$P_\delta(\sqrt{n}(\hat{\mu}-\delta) \geq f(n) - \delta\sqrt{n}) \geq 1-\beta,$$

where under P_δ, the $Y_1, \ldots, Y_n$ are i.i.d. $N(\delta, 1)$. This is the smallest n at which $f(n) - \delta\sqrt{n} \geq -z_\beta$, i.e. $\sqrt{n} \geq (z_\beta + f(n))/\delta$. The standard result for n_{np} now follows by taking $f(n) = z_\alpha$. For the Bayesian test, to very good approximation, $f(n) = \sqrt{6 + \log n}$ (Example 3). Since, n_{Bayes}, the smallest n for the Bayesian test must be larger than n_{np}, it satisfies $n_{Bayes}/n_{np} \geq (z_\beta + \sqrt{6+\log n_{Bayes}})^2/(z_\beta + z_\alpha)^2 \geq (z_\beta + \sqrt{6+\log n_{np}})^2/(z_\beta+z_\alpha)^2$, giving a logarithmic factor as claimed.

Brownian Motion

Fix $a > 0$ and, for a standard Brownian motion $X_t = B_t + ct$ with drift c, define $S = \min\{t > 0 : X_t \geq a\}$. The distribution of S is well-known (see e.g. (**?**)) and has density given by

$$f_{a,c}(s) = \frac{a \exp\left(-\frac{(a-cs)^2}{2s}\right)}{\sqrt{2\pi s^3}}. \tag{44}$$

Fix δ and let $n_t = t/\delta^2$ and $\mathcal{T} = \{\delta^2, 2\delta^2, 3\delta^2, \ldots\}$. Consider, for $t \in \mathcal{T}$, the discrete time process

$$W_t := \log \frac{p_\delta(Y^{n_t})}{p_0(Y^{n_t})}$$

where p_δ is the density of Y^n under $N(\delta, 1)$. Writing out the definition and re-arranging, we find that if $Y_1, Y_2, \ldots$ are i.i.d. $\sim N(\delta, 1)$, then for all $\mathcal{T}' \subset \mathcal{T}$ the conditional distribution of W_t given $\{W_t : t \in \mathcal{T}'\}$ (in particular, this includes the marginal distribution of W_t if we take $\mathcal{T}'$ empty) agrees with the conditional distribution of $X_t = B_t + (1/2)t$ and we can thus approximate the distribution of the first time when W_t exceeds $-\log\alpha$ by the distribution with density (44) with $c = 1/2$ and $a = -\log\alpha$ — the distribution of the first hitting time of B_t will be shifted slightly to the left, because when stopping W_t we are only checking the process at intervals of size δ^2. Intuitively, as $\delta \downarrow 0$, we expect the distribution functions to converge. To make this concrete, let q_β be the β-quantile of the distribution given by $f_{-\log\alpha, 1/2}(s)$. We want to calculate $n_{\max}(\beta)$, the smallest n such that $P_\delta(\tau_1 \leq n) > 1-\beta$, i.e. we want to find the smallest n such that, with $t^* = \delta^2 n$, we have $P_\delta(\delta^2\tau_1 > t^*) < \beta$. The correspondence between W_t and X_t suggests that in the limit for $\delta \downarrow 0$, t^* converges to q_β, giving that $n_{\max}(\beta) \sim C_{\beta,0}/\delta^2$ with $C_{\beta,0} = q_\beta$ and $\mathbf{E}_{P_\delta}[\tau_\beta] \sim C'_{\beta,0}/\delta^2$ with $C'_{\beta,0} = \int_0^{q(\beta)} t f(t)\,dt + \beta q_\beta$.