Synthese (2026) 208:130
https://doi.org/10.1007/s11229-026-05779-4

ORIGINAL RESEARCH

E-values as statistical evidence: a comparison to Bayes factors, likelihoods, and p-values

Ben Chugg[1] · **Aaditya Ramdas**[1] · **Peter Grünwald**[2,3]

Received: 25 March 2026 / Accepted: 11 August 2026

Abstract

A recurring debate in the philosophy of statistics concerns what, exactly, should count as a measure of evidence for or against a given hypothesis. P-values, likelihood ratios, and Bayes factors all have their defenders. In this paper we add two additional candidates to this list: the e-value and its sequential analogue, the e-process. E-values enjoy several desirable properties as measures of evidence: they combine naturally across studies, handle composite hypotheses, provide long-run error rates, and admit a useful interpretation as the wealth accrued by a bettor in a game against the null distribution. E-processes additionally handle optional stopping and optional continuation. This work examines the extent to which e-values and e-processes satisfy the evidential desiderata of different statistical traditions, concluding that they combine attractive features of p-values, likelihood ratios, and Bayes factors, and merit serious consideration as interpretable and intuitive measures of statistical evidence.

Keywords E-values · E-processes · Statistical evidence · Bayes factors · Likelihood ratios · P-values · Philosophy of statistics

1 Introduction

It is sometimes asserted that the p-value, first introduced by Karl Pearson (Pearson, 1900) but popularized and developed by Ronald Fisher (Fisher, 1925, 1935), is a measure of evidence against the null hypothesis. There is—to put it mildly—signifi-

✉ Ben Chugg
benchugg@cmu.edu

[1] Carnegie Mellon University, Pittsburgh, PA, USA

[2] Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

[3] Leiden University, Leiden, The Netherlands

Published online: 26 August 2026

cant disagreement over this claim (Berger & Sellke, 1987; Greenland, 2019; Hubbard & Lindsay, 2008; Lakens, 2022; Muff et al., 2022; Wagenmakers, 2007; Wright, 1992).

P-values lack several properties that many would expect of a measure of evidence, including consistency, sample size invariance, and the ability to handle optional stopping. These drawbacks have caused many statisticians to renounce p-values as appropriate measures of statistical evidence (though, of course, they remain useful for other purposes). For example, Principle 6 in the American Statistical Association's statement on p-values reads "by itself, the p-value does not provide a good measure of evidence regarding a model or hypothesis" (Wasserstein & Lazar, 2016).

Critics of p-values turn to other proposed notions of evidence, often the likelihood ratio and the Bayes factor. No single notion has satisfied everyone, however, and debates about the best object for quantifying statistical evidence continue to rage (e.g., Forster, 2006; Lele, 2004; Mayo & Spanos, 2011; Royall, 1997; Taper & Ponciano, 2016).

In this article we do not attempt to resolve these disputes. Instead, we hope to simply add two new candidates to the discussion about statistical evidence: *e-values/e-variables* and *e-processes*. These objects (together, *e-statistics*) are simple yet rich objects in mathematical statistics that have garnered significant attention over the past five years. They have proven indispensable to recent progress across several areas, such as multiple testing, mean estimation, and changepoint detection. We refer to Grünwald et al. (2024) and Ramdas and Wang (2025) for more on such applications.

1.1 Contributions and outline

We investigate the extent to which e-statistics satisfy various evidential criteria commonly used in the philosophy of statistics. To do so, Sect. 3 collects 21 desiderata that a measure of statistical evidence should satisfy according to several schools of thought, including Bayesian confirmation theory and various flavors of frequentism. Section 5 then evaluates e-statistics against these criteria and compares them with p-values, likelihood ratios, and Bayes factors.

While e-statistics do not satisfy all criteria for evidence that we identify throughout the literature, we find that they perform admirably well across a variety of measures. This stems partly from the fact that they can be viewed as generalized likelihood ratios, thereby inheriting many of the likelihood ratio's evidential properties. But they also satisfy criteria that the likelihood ratio fails, such as providing meaningful evidence against a single hypothesis without necessarily requiring an explicit alternative hypothesis.[1] That said, the flexibility of e-values and e-processes also has drawbacks. We discuss these in Sect. 6, where we also discuss their relationship to Birnbaum's theorem and the likelihood principle (Birnbaum, 1962).

Beyond relating e-statistics to the literature on statistical evidence, a secondary contribution of the paper is to clarify and sharpen several of the desiderata themselves. In particular, we distinguish between *static* and *dynamic* evidential criteria. Static criteria are evaluated on a fixed batch of data, whereas dynamic criteria con-

[1] Though whether this is valuable, or even sensible, depends partly on one's view of statistical evidence.

cern sequential and counterfactual settings in which additional data may be collected or further studies run. We further distinguish between different kinds of counterfactual scenarios and different forms of optional continuation, providing examples of each. We hope this taxonomy clarifies a set of issues that are often conflated in the literature.

We begin with an overview of e-statistics.

2 E-Statistics and the betting game

An e-variable E for a statistical hypothesis (i.e., collection of distributions) $\mathcal{H}$ is a nonnegative random variable whose expectation is at most 1 under all $P \in \mathcal{H}$:

$$\mathbb{E}_{\mathcal{H}}[E] := \sup_{P \in \mathcal{H}} \mathbb{E}_P[E] \leq 1. \tag{1}$$

We sometimes differentiate between the random variable E, which we call the e-variable, and its realization which we call the e-value. In practice, there is a random quantity X which represents our data (for example X may be a vector of 100 independent observations in an experiment, or a summary of these data such as a z-score), and E can be written as a (measurable) function of X, i.e. E is a *statistic*.

An e-process is the sequential analogue of the e-value. Here, we model our data as a sequence $X_1, X_2, \ldots$ (which we abbreviate henceforth to $(X_t)_{t \geq 1}$) where each X_i may be either a simple data point or a vector. Now consider a sequence of random variables $(E_t)_{t \geq 1} = E_1, E_2, \ldots$ where each E_i can be written as a function of the first i outcomes, $X_1, \ldots, X_i$. We call such a sequence an e-process for $\mathcal{H}$ if all E_i are nonnegative and $\mathbb{E}_{\mathcal{H}}[E_\tau] \leq 1$ for every stopping time τ. Informally, a stopping time is a random (that is, data-dependent) time determined by a stopping *rule* that for sequences $X_1, \ldots X_t$ of arbitrary length either says 'stop' or 'continue'. Crucially, whether we stop at time t must be decided on the basis of $X_1, \ldots, X_t$ alone, and not on data yet to arrive. The realized stopping time τ is then the first t at which the rule says 'stop', and is random because it is a function of the random data. Examples are 'stop at $t = 5$', 'stop as soon as you have seen two outcomes larger than 1', or 'stop at the smallest t such that the p-value based on t outcomes is smaller than 0.05.'[2]

It is worth flagging immediately that a sequence of e-variables is not in general an e-process. That is, validity at each fixed time does not imply validity at data-dependent times, since one could always stop at whichever E_t happens to be largest. What *is* guaranteed is that a running product of e-variables, each of which is valid conditionally on the past, forms an e-process. We make this precise in Example 1 and the discussion following it, and return to the gap between the two notions in Sect. 5.2.

[2] Formally, what we covered here is just a special case of the more general, measure-theoretic definition of e-processes and stopping times. There we have a set of filtered probability spaces $\{(\Omega, (\mathcal{F}_t)_{t \in \mathbb{N}}, P)\}_{P \in \mathcal{H}}$ on which (E_t) is an adapted process (i.e., E_t is $\mathcal{F}_t$-measurable for each t). Recall that a stopping time τ is defined relative to such a filtration as a random variable such that $\{\tau = t\}$ is $\mathcal{F}_t$-measurable for each t.

E-statistics have long appeared implicitly in the statistics and probability literature, especially in relation to nonnegative supermartingales (which are a special case of e-processes). It is only recently however (roughly 2020; see Ramdas & Wang, 2025, Section 1.8) that they have begun to be studied as objects in their own right.

E-statistics have often been referred to as measures of evidence against $\mathcal{H}$ (Ramdas et al., 2023; Shafer, 2021): the larger the value the more evidence. This view is buoyed by their interpretations as betting scores in a game between the statistician and nature (Shafer & Vovk, 2005, 2019). This game proceeds as follows.

2.1 The betting game

Fix some distribution $P \in \mathcal{H}$. At time $t = 1, 2, \ldots$, the statistician designs a nonnegative random variable B_t, which is required to be a function of data $X_1, \ldots, X_t$ arriving at or before time t, satisfying[3]

$$\mathbb{E}_P[B_t | X_1, X_2, \ldots, X_{t-1}] \leq 1. \tag{2}$$

We imagine the statistician buying a share of B_t for \$1 and his payoff is $\$B_t$ once its value is revealed. The inequality in (2) expresses that, under P, the statistician does not expect to make any money, regardless of how cleverly B_t is designed. We assume that the statistician can buy an arbitrary (nonnegative) number of shares. If the statistician starts with \$1 in round 1, then after the first outcome X_1 is revealed he will have $\$B_1$. If he reinvests all this money in round 2, then after X_2 is revealed he will have $B_1 B_2$. Thus, if the statistician starts with \$1 and re-invests his money at each point in time, his total capital at time t is given by[4] $K_t := \prod_{i \leq t} B_i$.

The intuition behind the betting game is that large values of K_t may be considered evidence against P. To see this, note that if data were truly drawn according to P, then the statistician does not expect to gain any money in this game, regardless of whatever stopping time τ is employed. More formally, for any stopping time τ:

$$\mathbb{E}_P[K_\tau] \leq 1, \tag{3}$$

which follows from (2) and the optional stopping theorem.[5] Next, note that it is unlikely that K_t can become large. How unlikely, exactly? By Markov's inequality,

[3] More formally and generally we can write $\mathbb{E}_P[B_t | \mathcal{F}_{t-1}] \leq 1$ where $\mathcal{F}_{t-1}$ is all the information available until and including time $t-1$, which is typically (though not always) the σ-algebra generated by data $X_1, \ldots, X_{t-1}$. For example, the process B_t in Example 6 depends on a coarser filtration. Strictly speaking X_i in (2) should be be replaced by B_i. The subtle difference, elaborated on by (Pérez-Ortiz et al., 2024), is beyond the scope of this paper.

[4] In the e-process literature, K_t is called a *test supermartingale*.

[5] The optional stopping theorem makes precise the idea that a fair game cannot be made favorable by a clever choice of when to quit. Condition (2) says that $\mathbb{E}_P[K_t \mid X_1, \ldots, X_{t-1}] = K_{t-1}\mathbb{E}_P[B_t \mid X_1, \ldots, X_{t-1}] \leq K_{t-1}$, i.e. that $(K_t)_{t \geq 0}$ is a nonnegative *supermartingale*, so in expectation the capital never grows. The theorem states that for such a process, $\mathbb{E}_P[K_\tau] \leq \mathbb{E}_P[K_0]$ for every stopping time τ. Since the statistician starts with one dollar, $K_0 = 1$, this gives (3). The nonnegativity of K_t is what allows τ to be unbounded.

(3) immediately gives that $P(K_\tau \geq 1/\alpha) \leq \alpha$. Further, by Ville's inequality (Ville, 1939), a time-uniform extension of Markov's inequality,

$$P\left(\text{there exists } t \geq 1 : K_t \geq \frac{1}{\alpha}\right) \leq \alpha. \tag{4}$$

That is, the probability of the process $(K_t)_{t\geq 1}$ ever exceeding $1/\alpha$, *no matter how long we continue the game*, is at most α.

We can relate e-variables and e-processes to the betting game as follows. When $\mathcal{H} = \{P\}$ is a singleton (which we henceforth call a *point* or *simple* hypothesis), an e-variable may be viewed as the payoff from a single-round game, $E = K_1$. The data X_1 in this game need not be a single observation—it may itself be a batch of data presented all at once. An e-process, meanwhile, corresponds to an evolving capital process $(K_t)_{t\geq 1}$.

For composite $\mathcal{H}$, for each $P \in \mathcal{H}$ let $(K_t^P)_{t\geq 1}$ be the capital process of a game played against the singleton distribution P. Now every e-variable corresponds to the *minimum* payoff in a single-round game, $E \leq \inf_{P\in\mathcal{H}} K_1^P$. Analogously, it can be shown that every e-process satisfies $E_t \leq \inf_{P\in\mathcal{H}} K_t^P$ (and *admissible* e-processes satisfy this with equality; see Ramdas et al., 2022). In other words, an e-process for $\mathcal{H}$ reports the minimum wealth across many simultaneous betting games, one for each $P \in \mathcal{H}$, all played using the same data $X_1, X_2, \ldots$.

To relate e-statistics to scientific practice, consider the following scenario.

Example 1 Suppose a research group analyzes an initial batch of data X_1—for example, from a clinical trial—and reports an e-variable $S_1 = S_1(X_1)$. Next, perhaps because the initial findings appear promising, the same or another group analyzes a new, independent batch of data X_2, reporting an e-variable S_2. The process may then continue, with further groups analyzing independent batches $X_3, X_4, \ldots$ and reporting e-variables $S_3, S_4, \ldots$. To combine the accumulating evidence—for instance, in a meta-analysis (Ter Schure & Grünwald, 2022)—the researchers multiply the successive e-variables.

We may think of the e-variable S_t as the payoff B_t at time t. Since the design of the e-variable $S_t = B_t$ at time t may depend on the earlier data $X_1, \ldots, X_{t-1}$, viewing the batches as arriving sequentially turns the defining e-variable condition $\mathbb{E}_P[S_t] \leq 1$ into the conditional requirement $\mathbb{E}_P[S_t \mid X_1, \ldots, X_{t-1}] \leq 1$. It follows that, for any stopping time τ, $E_\tau := \prod_{i\leq\tau} S_i$ is an e-variable, and that the running product $E_t := \prod_{i\leq t} S_i$ defines an e-process.

Example 1 constructs an e-process by multiplying e-variables. Conversely, when $\mathcal{H}$ is simple, every admissible e-process $(E_t)_{t\geq 0}$ for $\mathcal{H}$ can be decomposed as a product of 'past-conditional' e-variables, i.e. of the form B_t satisfying (2). Explicitly distinguishing between these two construction-directions (creating e-processes out of e-variables and vice-versa) is not always important, yet it will become highly useful in Sect. 5.2.

Next let us provide an example of a fundamental e-statistic.

Example 2 (Likelihood ratio) Consider a simple hypothesis $\mathcal{H} = \{P\}$ where P has density p. Then the likelihood ratio q/p for any distribution Q with density q is an e-variable for $\mathcal{H}$ (see below for assumptions implicit in this notation). To see this, note that

$$\mathbb{E}_P\left[\frac{q(X)}{p(X)}\right] = \int \frac{q(x)}{p(x)} p(x)\mathrm{d}x = \int q(x)\mathrm{d}x = 1. \tag{5}$$

By similar reasoning, for a sequence of random variables $(X_t)_{t\geq 1}$, the process $(L_t)_{t\geq 1}$ where $L_t = q(X_1, \ldots, X_t)/p(X_1, \ldots, X_t)$ defines an e-process for the hypothesis that the data are distributed according to P. If further they are independent and identically distributed (iid) under P and Q, we can write $L_t = \prod_{i\leq t} q(X_i)/p(X_i)$.

By recovering the likelihood ratio for point hypotheses, Example 2 highlights that one useful way to understand e-statistics is as extensions of likelihood ratios to composite settings. In particular, as opposed to some other popular extensions to the composite setting such as the generalized likelihood ratio (Sect. 3), they choose to preserve the property $\mathbb{E}_{\mathcal{H}}[E] \leq 1$, which we can view as "fairness" in the betting game: evidence is difficult to exaggerate when playing against the null. We will return to the connection between simple-versus-simple likelihood ratios and general e-statistics later on.

We provide additional examples in Sect. 4, but to orient unfamiliar readers, let us provide a second e-statistic here. This example can be viewed as an instance of the former.

Example 3 (Gaussian e-variable) For any $\lambda \in \mathbb{R}$, the object $E_\lambda(X) = \exp(\lambda X - \lambda^2\sigma^2/2)$ is an e-variable for $\mathcal{H} = \{N(0, \sigma^2)\}$, i.e., a Gaussian with mean 0 and variance σ^2. Note that E_λ is the likelihood ratio between $N(\sigma^2\lambda, \sigma^2)$ and $N(0, \sigma^2)$. Further, the process $(E_t)_{t\geq 1}$ where $E_t = \exp(\lambda \sum_{i\leq t} X_i - t\lambda^2\sigma^2/2) = \prod_{i\leq t} \exp(\lambda X_i - \lambda^2\sigma^2/2)$ is an e-process for the hypothesis that the data $(X_t)_{t\geq 1}$ are iid from $N(0, \sigma^2)$.

Remark on notation and definitions: Although the theory of e-statistics can be developed in much greater generality (Larsson et al., 2025), throughout this paper we assume that all distributions admit densities or probability mass functions. We use capital letters such as P and Q to denote distributions, and corresponding lowercase letters such as p and q to denote their densities/mass functions. Whenever we refer to likelihood ratio q/p with $P \in \mathcal{H}$, we assume it is almost surely well-defined under both $\mathcal{H}$ and Q.

2.2 Evidence for or against?

2.2.1 The meaning of small E

The betting game provides a notion of evidence *against* a hypothesis $\mathcal{H}$, as opposed to a notion of evidence *for* the hypothesis. In particular, if E is an e-variable for $\mathcal{H}$, then in general small values of E need not provide evidence in favor of $\mathcal{H}$. This is

because one can play the betting game conservatively. As an extreme example, the bet $B_t \equiv 1$ is an e-variable but will never make any money regardless of the discrepancy between $\mathcal{H}$ and the true data generation process.

That said, e-statistics are often used to compare two hypotheses. In hypothesis testing, for instance, we compare a null $\mathcal{H}_0$ to an alternative $\mathcal{H}_1$ and search for e-statistics under $\mathcal{H}_0$ which grow quickly under $\mathcal{H}_1$. For example, for singletons $\mathcal{H}_0 = \{P_0\}$ and $\mathcal{H}_1 = \{P_1\}$, we can consider the likelihood ratio (Example 2). This suggests that there are indeed cases when a small e-statistic E for $\mathcal{H}_0$ should be viewed as evidence in favor of $\mathcal{H}_0$.

A necessary requirement for such a scenario is that the reciprocal E^{-1} of E be an e-statistic when we reverse the roles of $\mathcal{H}_0$ and $\mathcal{H}_1$, as is the case for the likelihood ratio. It is not, however, sufficient. Example 4 below exhibits an E whose reciprocal is a perfectly good e-statistic for the reversed problem, and yet whose alternative is not one that anybody should endorse. When the condition holds *and* the alternative is a bona fide hypothesis about the world, evidence against $\mathcal{H}_0$ can indeed be viewed as evidence for $\mathcal{H}_1$ and vice versa.

2.2.2 The meaning of large E

Often, however, $\mathcal{H}_1$ is implicit or chosen pragmatically based on the problem at hand. This is the case in recent work on bounded mean testing (Waudby-Smith & Ramdas, 2024) (Example 7 below) or in various conformal e-statistics (Vovk, 2025), for instance. In such cases, the perspective of a large e-value providing evidence against $\mathcal{H}_0$ remains clear whereas the interpretation as evidence for a specific $\mathcal{H}_1$ less so. The following example shows that this may happen even when $\mathcal{H}_0 = \{P\}$ is simple and the e-statistic E is a likelihood ratio. In general, while an alternative can be inferred from the bets or, more indirectly, from the e-statistic's definition, such an alternative might be best viewed as instrumentally useful, and not representing a bona fide hypothesis that represents a true state of the world.

Example 4 Ryabko and Monarev (2005) show that bit strings produced by standard random number generators can be substantially compressed by lossless data compression algorithms such as `zip`, which is a clear indication that the bits are not so random after all. Here, the null hypothesis states that data are 'random' (independent fair coin flips), and one measures 'amount of evidence against $\mathcal{H}_0$ provided by data $x^\tau = x_1, \ldots, x_\tau$' as

$$\tau - \mathrm{CL}_{\mathrm{zip}}(x^\tau),$$

where $\mathrm{CL}_{\mathrm{zip}}(x^\tau)$ is the number of bits needed to code x^τ using (say) `zip`. Kraft's inequality (Cover & Thomas, 1991) says that for arbitrary prefix codes for encoding binary strings of length τ into other binary strings of non-fixed length, the code length $\mathrm{CL}(x^\tau)$ as measured in bits satisfies $\sum_{x^\tau \in \mathcal{X}^{(\tau)}} 2^{-\mathrm{CL}(x^\tau)} \leq 1$, where $\mathcal{X}^{(\tau)}$ is the set of binary strings determined by stopping time τ (if $\tau = n$ for fixed n, this is simply $\{0,1\}^n$). Thus, if we set $q(x^\tau) := 2^{-\mathrm{CL}_{\mathrm{zip}}(x^\tau)}$ we get $\sum_{x^\tau \in \mathcal{X}^{(\tau)}} q(x^\tau) \leq 1$: q represents a *sub-probability distribution*. At the same time, for the null we have $\mathcal{H}_0 = \{P\}$,

where P represents a sequence of fair coin flips, so it has mass function p with for each $x^\tau \in \mathcal{X}^{(\tau)}$, $p(x^\tau) = 2^{-\tau}$. Defining $E_\tau = q(X_1, \ldots, X_\tau)/p(X_1, \ldots, X_\tau)$ we thus find

$$\mathbb{E}_P[E_\tau] = \sum_{x^\tau \in \mathcal{X}^{(\tau)}} \frac{q(x^\tau)}{p(x^\tau)} p(x^\tau) \leq 1 \text{ and } \log E_\tau = \tau - \mathrm{CL}_{\mathrm{zip}}(X_1, \ldots, X_\tau).$$

Thus, the Ryabko-Monarev bit-difference is the logarithm of an e-value. But note that there is no explicitly defined alternative. Being able to significantly compress a string by `zip` intuitively provides strong evidence that the null hypothesis is false, and this is formalized by e-processes. Nevertheless, even though we measure evidence by the log likelihood ratio between $\mathcal{H}_0 = \{P\}$ and Q, this evidence against $\{P\}$ should clearly *not* be construed as evidence in favor of the sub-distribution Q, which is a complicated object that helps to *detect* some structure in $x_1, \ldots, x_\tau$ but is not best thought of as either generating or predicting that structure.

It turns out that this example generalizes: any e-statistic has a code length-difference interpretation, which connects it to notions of evidence that have implicitly been suggested by the information theory community, often in the context of *Minimum Description Length* model comparison (Grünwald, 2007; Grünwald & Roos, 2020). Since it is less well-known, we do not spell out the desiderata coming from this tradition here but refer instead to (Grünwald & Roos, 2020) for those interested.

To summarize: e-statistics are best regarded as measures of evidence against a designated hypothesis, with "evidence for an alternative" emerging only in distinctive scenarios.

2.3 Optimality of e-statistics

Example 4 notwithstanding, in some settings there is a clearly defined alternative hypothesis $\mathcal{H}_1$. In such cases, it is natural to call an e-variable for testing $\mathcal{H}_0$ *optimal* if it tends to become large under $\mathcal{H}_1$. In particular, if the data are in fact generated under $\mathcal{H}_1$, we would like evidence against $\mathcal{H}_0$ to accumulate quickly. This idea of "growing quickly under the alternative" can be formalized in several ways. The most natural, and by far the most studied, is *log-optimality*; the e-variable achieving it has been called *GRO* (growth-rate optimal) (Grünwald et al., 2024; Lardy et al., 2024) or the *numeraire* (Larsson et al., 2025) (the latter paper also describes alternative notions of optimality, as does Koning, 2024). In fact, the likelihood ratio from Example 2 is the log-optimal e-variable for testing P against Q. More broadly, this highlights that a given hypothesis may admit many different e-variables and e-processes. We return to this point in Sect. 6. For now, let us turn to a discussion of the various statistical traditions and what properties they believe a measure of evidence should possess.

3 An overview of evidential desiderata

Unsurprisingly given the fierce debates at the foundations of statistics, there isn't a single checklist of evidential criteria adopted by all statisticians and philosophers of statistics alike. Instead, different statistical traditions argue for different criteria. Following Forster (2006), we group these traditions into three broad groups:

(i) likelihood-based accounts, which treat evidence as relative support between hypotheses and elevate likelihood ratios as fundamental;
(ii) Bayesian accounts, which quantify evidence by changes in model odds, typically via Bayes factors and their approximations; and
(iii) error-probability accounts, which connect evidential strength to the ability of procedures to detect discrepancies (e.g., through type-I/II error rates or confidence intervals).

The gray box below summarizes the main criteria proposed by these three constellations of statistical philosophies, augmented with some of our own. But let us first provide some more background on these three traditions. We begin with the latter.

3.1 The error-probability tradition

The error-probability tradition traces back to Fisher, Neyman, and Pearson. There is no unified view within this tradition of how statistics should be practiced. In fact, Fisher famously had significant disagreements with Neyman and Pearson, who proposed a fully decision-theoretic theory of statistics in which 'evidence' really plays no role—there are only decisions, risks and error probabilities. Nevertheless, Fisher, Neyman and Pearson were all frequentists and informal references to 'evidence' within the Neyman-Pearson tradition abound.[6] And, despite these disagreements, contemporary frequentist practice largely draws on a hybrid error-probability tradition combining ideas from Fisher, Neyman, and Pearson. We can thus still identify core assumptions on the desired properties within this tradition of a measure of evidence. In particular, a measure of evidence should bound or diagnose the probability of being misled.

In other words, the evidential strength of a given object or procedure is inseparable from its (frequentist) control of various error rates (type I/II error, risk control, false discovery control, etc.) Evidence is stronger when the method would rarely give such supportive results if the underlying claim were false. This is precisely the guarantee offered by a p-value, which is a statistic $T = T(X)$ such that $P(T \leq \alpha) \leq \alpha$ for all

[6] In fact, even Neyman (1976) himself wrote 'my own preferred substitute for 'do not reject H' is 'no evidence against H is found'.

$\alpha \in (0, 1)$ and all $P \in \mathcal{H}$. (Here we would say that T is a p-value for $\mathcal{H}$.)[7] Thus, on the evidential view of a p-value, small values of T are evidence against $\mathcal{H}$.[8]

Extensions of p-values such as s-values (Greenland, 2019), second-generation p-values (Blume et al., 2019), and replication values (Killeen, 2005) are all rooted in the error-probability tradition. Confidence curves/distributions are also sometimes proposed as evidential summaries on this view (Xie & Singh, 2013). The error-statistics framework of Mayo and Spanos (Mayo, 1996; Mayo & Spanos, 2011) (which downplays the differences between Fisher and Neyman and focuses on commonalities instead) and Allan Birnbaum's "confidence concept of evidence" (Birnbaum, 1977) likewise sit squarely in the error-probability family.

3.2 Bayesian confirmation theory

The Bayesian statistical tradition is, naturally, rooted in the Bayesian view of probability, which treats model parameters as random variables instead of unknown constants. An agent places an initial prior distribution π over parameters θ (often referred to as the agents' *beliefs*) and, upon observing data X, updates $\pi(\theta)$ via Bayes' rule to a posterior $\pi(\theta \mid X)$. Instead of controlling error, evidence on the Bayesian view is linked to *confirmation*: data provide evidence for a hypothesis or model to the extent that they increase its posterior support relative to its prior support.

The most common tool for quantifying evidence on the Bayesian view is the Bayes factor (Jeffreys, 1939; Kass & Raftery, 1995), which is typically used to compare two distinct hypotheses. In particular, for $\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\}$ and $\mathcal{H}_1 = \{P_\theta : \theta \in \Theta_1\}$ equipped with priors π_0 and π_1, we can write the change in posterior odds after observing X as the prior odds multiplied by the Bayes factor:

$$\frac{\mathbb{P}(\mathcal{H}_1 \mid X)}{\mathbb{P}(\mathcal{H}_0 \mid X)} = \frac{\mathbb{P}(\mathcal{H}_1)}{\mathbb{P}(\mathcal{H}_0)} \cdot \mathrm{BF}(X), \text{ where } \mathrm{BF}(X) = \frac{\mathbb{P}(X|\mathcal{H}_1)}{\mathbb{P}(X|\mathcal{H}_0)}, \tag{6}$$

where $\mathbb{P}(X|\mathcal{H}_j) = \int_{\Theta_j} P_\theta(X)\pi_j(\theta)\mathrm{d}\theta$ is the marginal likelihood under hypothesis j. Note that the posterior odds are defined by both prior distributions over hypotheses,

[7] Readers more familiar with the textbook formulation—"the probability of obtaining test results at least as extreme as the result actually observed, assuming the null hypothesis is correct"—may find it helpful to see how the two are related. Suppose evidence against $\mathcal{H}$ is summarized by some test statistic $S(X)$, with larger values counting as more extreme, and define the tail probability $T(x) := \sup_{P \in \mathcal{H}} P\big(S(X) \geq S(x)\big)$. This T is precisely the textbook p-value, and a short calculation shows that it satisfies $P(T \leq \alpha) \leq \alpha$ for every α and every $P \in \mathcal{H}$. The definition above simply isolates that property—which is what all the error-probability guarantees actually rest on—and drops the requirement that T arise from any particular S.

[8] We say "on the evidential view" advisedly. Many within the error-probability tradition would deny that the p-value is a measure of evidence at all—for the strict Neyman-Pearsonian it is an input to a decision rule and nothing more—and would therefore regard the question we pose in this paper as not quite the right one to ask. However, our aim is not to adjudicate whether each tradition ought to be in the business of measuring evidence, but to take the criteria each has articulated and ask how e-statistics fare against them. Readers who reject the evidential reading of p-values may simply read the corresponding rows of Table 1 as reporting what the p-value would deliver were it forced into evidential service.

denoted as $\mathbb{P}(\mathcal{H}_j)$, as well as prior distributions over parameters Θ_j, denoted by π_j. For the purposes of the Bayes factor, however, only π_0 and π_1 are necessary to define.

Bayes factors can be difficult to compute and approximations such as BIC are therefore sometimes offered (Wagenmakers, 2007). However, for the purposes of an evidential standard, exact Bayes factors are the most popular among Bayesian confirmation theorists, and will thus be our main focus here. For more detailed discussion of various measures of Bayesian confirmation, see Fitelson (1999), Eells and Fitelson (2000).

3.3 The likelihood tradition

Finally, there is the likelihood camp, which emerged as a rival to both Bayesianism, which it saw as too subjective, and error-probability accounts, which it saw as overly concerned with decision-making instead of evidence (Edwards, 1972; Royall, 1997; Hacking, 2016). That said, likelihoodists and Bayesians share a number of commitments, emphasizing criteria such as combination within and across studies, continuity, consistency, and sharing the view that evidence is fundamentally comparative between two hypotheses (Lele, 2004; Taper & Lele, 2010; Taper & Ponciano, 2016). But likelihoodists side with the frequentists on one decisive point: they want an objective, prior-free measure of evidence, one fixed by the data and the models under consideration rather than by the analyst's degrees of belief. But this is the only sense in which we call them frequentists here since they do not, in general, accept the further frequentist commitment that evidential strength must be underwritten by long-run error rates.

Unlike the Bayes factor, however, which marginalizes over the prior, the likelihood ratio is not immediately well-defined for composite hypotheses. In that case, one must appeal to some generalization, such as the generalized likelihood ratio (GLR)

$$\mathrm{GLR}(\Theta_0, \Theta_1) := \frac{\sup_{\theta_1 \in \Theta_1} p_{\theta_1}(X)}{\sup_{\theta_0 \in \Theta_0} p_{\theta_0}(X)}, \tag{7}$$

or others (cf. Bickel, 2011, 2012). Indeed, the Bayes factor is one possible extension to the composite setting, though likelihoodists tend to prefer using objective priors (Jeffreys, 1939) in this case. E-statistics, as we will continue to discuss, are another possible generalization. As a consequence, it is sometimes natural to refer to e-statistics as "generalized likelihood ratios," and we will do so throughout this paper. However, it is important to keep in mind that they are distinct from (7).

Aside from these two, the evidential properties of extensions of the likelihood ratio to composite settings are not always clear, and often do not preserve the comfortable logic of the likelihood ratio: that the observed data are a times as likely under θ_1 as under θ_0 if $p_{\theta_1}(X) = a p_{\theta_0}(X)$. In Sect. 5 we focus on the GLR as the natural

composite extension, but we trust that readers can extend the logic to a different generalization if desired.[9]

With this background established, we turn to the evidential desiderata endorsed by these three traditions. Of course, it should not be assumed that every entry carries equal weight, even among those traditions which endorse it. The list is thus not a scorecard, and a measure which satisfies more entries is not automatically better. Still, the list gives a useful comparison point.

As noted in the introduction, we divide these into two categories: static and dynamic. The static criteria are relatively standard in the literature and require little explanation. The dynamic criteria, by contrast, are more nuanced—and in some cases, entirely novel. We further divide the dynamic criteria into two sub-categories: inter-experiment and intra-experiment. The former concerns the relationship between multiple experiments, while the latter focuses on the internal dynamics of a single experiment. Because these dynamic criteria are subtle, Sect. 5.2 devotes significant space to clarifying their meaning, comparing them with related criteria, and illustrating them through specific examples.

4 Evidential desiderata

4.1 Static criteria

Scalar measure. Evidence for or against a hypothesis should be quantified by a single real number. This is a criterion shared by all traditions.

Continuous measure. Evidence should be graded rather than dichotomous. That is, there is no fixed threshold at which something abruptly becomes evidence. Rather, evidence comes in degrees. This is again endorsed by all three traditions.[10] See, e.g., Royall (1997), Lele (2004).

Irrelevance of the analyst. The evidence should depend on the data only and not on whoever is running the experiment. That is, evidence should be objective. This is endorsed by the likelihoodists and error-probabilists, and explicitly rejected by the Bayesians. See Royall (1997), Bickel (2012), Taper and Ponciano (2016).

Coherence. A measure of evidence should not assign higher evidence to any implication of a hypothesis than to the hypothesis itself (Gabriel, 1969). Formally: If $\mathcal{H} \subset \mathcal{H}'$ then you should have at least as much support for $\mathcal{H}'$ as you do for $\mathcal{H}$ (Schervish, 1996). Or, equivalently, for $\mathcal{H} \subset \mathcal{H}'$, the evidence against $\mathcal{H}'$ should not exceed that against $\mathcal{H}$ (Bickel, 2024). This desideratum arose from multiple testing, and was only later applied to criticize p-values as valid measures of evidence (Schervish, 1996).[11]

[9] We note that in many important instances (i.e. choices of $\mathcal{H}$), *conditional likelihoods* (Royall, 1997) and *partial likelihoods* such as those underlying the Cox (1972) regression model do turn out to be e-statistics (Hao & Grünwald, 2024).

[10] Note this is distinct from *decision-making* in the error-probability tradition, which thresholds p-values in order to either reject or sustain the null hypothesis. But as a measure of evidence, the p-value is meant to be continuous.

[11] This is distinct from the notion of coherence often discussed in a Bayesian context, which is the requirement that a set of probabilities be immune from a Dutch book (Berger, 1983).

Consistency. A measure of evidence is consistent if, roughly speaking, it favors the true hypothesis as the sample size increases. This has been formalized in various ways (cf. Berger et al., 2003; Grendár, 2012). Here we will say that, when comparing $\mathcal{H}_0$ and $\mathcal{H}_1$, a measure of evidence M_n on n observations is consistent if M_n tends to ∞ in probability under $\mathcal{H}_1$, and tends to 0 in probability under $\mathcal{H}_0$. Consistency is typically promoted by likelihoodists (Royall, 1997) and Bayesians (Chib & Kuffner, 2016).

Sample size invariance. A given numerical value of the evidence should correspond to the same strength of evidence regardless of sample size or other aspects of the sampling design (i.e., "same score, same evidence"). For instance, an evidence value of x obtained with $n = 20$ should be comparable to the same value x obtained with $n = 200$. This is supported by Bayesians and likelihoodists; see Royall (1986), Goodman and Royall (1988), Wagenmakers (2007).

Scale invariance. One-to-one transformations of the sample space (e.g., unit changes) should not change the evidence. For instance, measuring distance in meters versus feet should not change the interpretation. This is supported by likelihoodists and Bayesians (Taper & Lele, 2011).

Reparameterization invariance. One-to-one transformations of the model parameters should not change the evidence. This is supported by likelihoodists (Hartigan, 1967; Berger, 1983; Lele, 2004; Taper & Lele, 2011) and (some) Bayesians (Jeffreys, 1946; Ghosh, 2011).

Single hypothesis validity. One should be able to give evidence for, or at least against, a single hypothesis, without comparing it to another hypothesis. This is explicitly endorsed by the error-probability tradition, and explicitly rejected by the likelihood tradition (Edwards, 1972; Royall, 1997).

Long-run error rates. Thresholds on the evidential scale should correspond to explicit bounds on the long-run frequency with which they are exceeded when the hypothesis is true. This is the main desideratum of the error-probability tradition. See, e.g., Neyman and Pearson (1928), Mayo (1996, 2018),Mayo and Spanos (2011).

Composite hypotheses. A measure of evidence should handle composite hypotheses. This is implicitly endorsed by Bayesians, who place priors over composite hypotheses, error-probabilists, and by those likelihoodists who seek to generalize the standard likelihood ratio. It has also motivated a significant amount of recent work in the e-statistics literature (Ramdas et al., 2023).

Nonparametric hypotheses. A measure of evidence should handle nonparametric hypotheses, where likelihood ratios may be very hard to define (due to technical reasons, like not having reference measures to define densities). As above, this is implicitly endorsed by many proponents of most traditions, but some of them handle it significantly more satisfactorily than others.

The likelihood principle. If two data sets induce proportional likelihood functions for the parameters of interest, then they should carry the same evidence about those parameters. This principle is foundational for the likelihood tradition (Edwards, 1972; Berger & Wolpert, 1988; Royall, 1997; Forster & Sober, 2004) and is also adopted by Bayesians who interpret likelihood function as marginal likelihood.

Composability under dependence. It should be possible to combine evidence from multiple sources without requiring strong assumptions on their dependence structure.

Such a desideratum arises more from practice than from philosophy. In multiple testing, for instance, it is common to want to combine p-values which can rarely be considered independent (Efron, 2012; Vovk et al., 2022). But variants of composability are supported by both the likelihoodist and Bayesian traditions. See Royall (1997), Bickel (2012), Morey et al. (2016).

4.2 Inter-experiment dynamic criteria

Accumulation I: Fixed design. When data arise from a fixed number of independent experiments, the total evidence should accumulate from the evidence contributed by the component parts. This is supported by both the likelihoodist and Bayesian traditions. See Royall (1997), Bickel (2012), Morey et al. (2016).
Accumulation II: Flexible design. When data arise from a sequence of independent experiments, and the decision to perform the next experiment may depend on the result of the previous one, the total evidence should accumulate from the evidence contributed by the component parts. This generalizes Accumulation I by allowing the number of experiments to be data-dependent. The criterion is discussed frequently in the literature on e-values, where it serves as part of their motivation (Grünwald et al., 2024).
Dynamic consistency. When assessing the evidence in an overall sample, we should obtain the same value regardless of whether we analyze the sample as a whole or partition it into sub-samples and combine the results. Variations of this desideratum have been studied extensively in the economics and decision-theory literature (Epstein & Schneider, 2003; Grünwald & Halpern, 2011). Despite the surface similarity, this is a strictly stronger demand than Accumulation I. Accumulation asks only that the pieces *can* be combined into a valid measure of the total evidence. Dynamic consistency asks in addition that the combined value *agree* with what one would have obtained by analyzing the pooled sample directly. A measure can therefore satisfy the former and fail the latter, as products of e-variables do (Sect. 5.2).
Inter-experiment counterfactuals. A measure of evidence should not depend on data that were not observed. More concretely, it should not depend on decisions about whether or not to continue collecting a new batch of data in counterfactual scenarios in which the data were different from what they actually were. Where Accumulation II concerns a further experiment that *is* run, and asks that its evidence combine properly with what came before, the present criterion concerns a further experiment that is *not* run, but would have been had the first batch come out differently.

4.3 Intra-experiment dynamic criteria

Intra-experiment optional stopping. The reason an experiment was stopped should not affect its evidential value. For example, stopping early because the initial results appear promising, or continuing longer than originally planned, should not change the evidence. This is closely related to, but distinct from, Accumulation II: here the issue is whether to stop or continue a single experiment, rather than whether to run an additional one. Handling optional stopping satisfactorily is a major motivation in modern sequential analysis (Ramdas et al., 2023).

Intra-experiment counterfactuals. The measure of evidence should depend only on the realized event that the observer learns has occurred, and not on unrealized contingencies in the experimental protocol. Thus, if two descriptions of the experiment yield the same realized event Y, they should yield the same evidence. This criterion, together with the previous one, has motivated criticism of p-values (Forster & Sober, 2004; Wagenmakers, 2007). The general issue of counterfactuals is often brought up by likelihoodists.

5 E-statistics, P-values, likelihood ratios, Bayes factors

Before investigating to what extent e-statistics satisfy the desiderata listed above, let us provide more detail on how they relate to p-values, likelihood ratios, and Bayes factors. We begin with p-values.

E-values can be converted to p-values and vice-versa. Indeed, for an e-value E, $1/E$ is a p-value by Markov's inequality: $P(1/E < \alpha) = P(E > 1/\alpha) \leq \alpha$. The inverse of a p-value is not an e-value in general. E-values tend to be more conservative than p-values, i.e. more extreme data is needed to reach a threshold $1/\alpha$. p-values can be converted to e-values via *calibrators* (Vovk & Wang, 2021): Non-negative functions such that $\mathbb{E}_{\mathcal{H}} f(P) \leq 1$. Examples of calibrators include $f(p) = \kappa p^{1-\kappa}$ for any $\kappa \in (0, 1)$, and $f(p) = \int_0^1 \kappa p^{\kappa-1} d\kappa$. Unfortunately, the e-values obtained via calibrators are usually not very powerful. In particular, they grow far more slowly under the alternative than an e-value designed for the problem at hand, and good (for example, log-optimal) e-values must typically be designed directly.

Next, let us introduce the following general method for constructing e-variables and e-processes when comparing two composite hypotheses.

Example 5 (Universal inference) Consider independent and identically distributed (iid) observations $X_1, X_2, \ldots, X_t$, null $\mathcal{H}_0$, and alternative $\mathcal{H}_1$, both of which may be composite. Wasserman et al. (2020) introduce the universal inference estimator

$$U_t := \prod_{i=1}^{t} \frac{\hat{q}_i(X_i)}{\hat{p}_t(X_i)}, \tag{8}$$

where $\hat{q}_i$ is some density in the alternative $\mathcal{H}_1$ which may be based on the first $i-1$ samples, and $\hat{p}_t$ is the maximum likelihood estimate among all distributions in the null, based on all t observations.[12] Alternatively, one can also equip $\mathcal{H}_1$ with a prior π (which must be independent of $X_1, \ldots, X_t$) and consider

[12] Here, for ease of comparison to other e-processes, we consider the case without a holdout dataset. Universal inference is more commonly known for the case where we do have a holdout dataset, we might compute $\hat{q}_i$ on these data to maximize U_t, in which case $\hat{q}_1 = \cdots = \hat{q}_t$. This is called "split universal inference."

$$V_t := \int \prod_{i \le t} \frac{q(X_i)}{\hat{p}_t(X_i)} \mathrm{d}\pi(q), \tag{9}$$

which we call the "method of mixtures." To see that U_t is indeed an e-variable for $\mathcal{H}_0$, note that $\prod_{i \le t} \hat{p}_t(X_i) \ge \prod_{i \le t} p(X_i)$ for all densities p in the null by definition, so for all $P \in \mathcal{H}_0$, $\mathbb{E}_P[U_t] \le \mathbb{E}_P[\prod_{i \le t} \frac{\hat{q}_i(X_i)}{p(X_i)}] = \int \left(\prod_{i \le t} \frac{\hat{q}_i(x_i)}{p(x_i)}\right) p(x_1) \cdots p(x_n) \mathrm{d}x_1 \cdots \mathrm{d}x_n = \prod_{i \le t} \int \hat{q}_i(x_i) \mathrm{d}x_i = 1$

, using that $\hat{q}_i$ is a probability density. Similar arithmetic applies for V_t. Moreover, $(U_t)_{t \ge 1}$ and $(V_t)_{t \ge 1}$ are e-processes.

When both hypotheses are simple, the universal inference e-variable collapses to the likelihood ratio. While it serves as an illuminating example, we do not advocate its use in all situations. In case of a simple alternative and a composite null, the numeraire e-variable (Larsson et al., 2025) is inherently preferable to the universal inference e-variable (even if we have a different notion of optimality in mind than log-optimality): for all outcomes, the numeraire is at least as large as the universal inference e-variable, whereas for some outcomes the inequality is strict. The reason this ordering is decisive is that both objects are e-variables for the same null, and so carry exactly the same type-I error guarantee. Universal inference is therefore *inadmissible* in this setting, in the usual decision-theoretic sense of being weakly dominated. On the other hand, the numeraire e-variable defines, in general, an e-variable only and not an e-process.[13]

The method of mixtures—pioneered by Herbert Robbins in the context of obtaining confidence sequences (Robbins, 1970)—and the method of predictable plug-ins, respectively illustrated by V_t and U_t above, are common strategies in the world of e-statistics and illustrate how e-statistics can serve as a bridge between Bayesians and frequentists. Indeed, while e-statistics are fundamentally frequentist objects, the method of mixtures shows how they can accommodate priors while retaining their frequentist properties. Meanwhile, the method of predictable plug-ins can be seen as learning a distribution over the alternative over time, which also has a very Bayesian spirit: the plug-in $\hat{q}_i$ is fit to the data $X_1, \ldots, X_{i-1}$ and is used to predict X_i, so that the sequence $\hat{q}_1, \hat{q}_2, \ldots$ progressively concentrates on whichever part of $\mathcal{H}_1$ best explains the data. This is analogous to the role played by a Bayesian posterior, but arrived at by estimation rather than by conditioning on a prior.

With Bayesianism in mind, let us give an example of a Bayes factor that is also an e-variable.

Example 6 (t-test Bayes factor) Consider iid observations $X_1, \ldots, X_t \sim N(\mu, \sigma^2)$, where $\sigma > 0$ is unknown, and suppose we wish to test the null $\mathcal{H}_0 := \{N(0, \sigma^2) : \sigma > 0\}$ against alternatives with nonzero standardized effect size $\delta := \mu/\sigma$. Let W be a proper prior on δ, and equip the nuisance scale σ with the (improper) right-Haar/Jeffreys prior $w_H(\sigma) \propto 1/\sigma$. The resulting Bayes factor is (Rouder et al., 2009):

[13] By this we mean that if we construct the numeraire e-variable $E_1, E_2, \ldots$ with E_i the numeraire variable for sample $(X_1, \ldots, X_i)$, the resulting process $(E_t)_{t \ge 1}$ is not an e-process Ramdas et al. (2023).

$$B_t := \frac{\int_{\mathbb{R}} \int_0^\infty \prod_{i=1}^t \frac{1}{\sigma} \rho\left(\frac{X_i - \sigma\delta}{\sigma}\right) \frac{\mathrm{d}\sigma}{\sigma} W(\mathrm{d}\delta)}{\int_0^\infty \prod_{i=1}^t \frac{1}{\sigma} \rho\left(\frac{X_i}{\sigma}\right) \frac{\mathrm{d}\sigma}{\sigma}}, \tag{10}$$

where ρ denotes the standard normal density. Although the prior on σ is improper, B_t^{RH} is an e-variable for $\mathcal{H}_0$ (in fact, (B_t) is an e-process for $\mathcal{H}_0$) (Grünwald et al., 2024; Wang & Ramdas, 2025). Indeed, under a mild $2+\epsilon$-moment condition, this e-variable is log-optimal (Pérez-Ortiz et al., 2024).

To say more about the relationship of e-statistics to Bayes factors, if $\Theta_0 = \{\theta_0\}$ is simple, then the Bayes factor is an e-variable; the logic is the same as in (5). In general, however, the Bayes factor is not an e-variable: it only has expectation 1 under the prior-averaged null, not uniformly over every $\theta_0 \in \Theta_0$. That said, Grünwald et al. (2024) show that, for sufficiently regular parametric null hypotheses with simple alternatives, log-optimal e-variables are Bayes Factors under a specific prior on Θ_0.

If the alternative hypothesis is composite, then one can obtain e-variables by equipping the alternative with an arbitrary prior determined by the analyst. There is then, again under mild regularity conditions, a unique prior on Θ_0, depending on the prior on Θ_1, such that the resulting Bayes factor is an e-variable. In this way, optimal e-variables can be seen as a subset of Bayes factors, though only the prior on the alternative, not on the null, is independent of the analyst.

As for the relationship between likelihood ratios and Bayes factors, the latter reduces to the former in point versus point settings as the priors are determined (they are atoms). For composite hypotheses, the Bayes factor can be seen as an extension of the likelihood ratio if one allows for distributions over the parameter space.

Next let us turn to a powerful e-variable for bounded random variables.

Example 7 (Bounded mean testing) For iid random variables $X_1, \ldots, X_t$ lying in [0,1] with mean $\mu \in (0,1)$, consider the object $M_t := \prod_{i=1}^t [1 + \lambda_i (X_i - \mu)]$, where λ_i lies in $[-1/(1-\mu), 1/\mu]$ and can be based on $X_1, \ldots, X_{i-1}$ but not X_j for $j \geq i$ (that is, it is *predictable*). Then M_t is nonnegative and satisfies $\mathbb{E}_{\mathcal{H}}[M_t] = 1$, where $\mathcal{H}$ is the set of all distributions on [0,1] with mean μ (whether continuous, discrete, or mixed). In fact, the process $(M_t)_{t \geq 0}$ is an e-process for testing the null hypothesis that the data are iid from P for any $P \in \mathcal{H}$. Such e-processes were leveraged recently by Waudby-Smith and Ramdas (2024) to obtain state-of-the-art confidence intervals/sequences.

The predictable sequence $(\lambda_t)_{t \geq 1}$ is typically chosen via the method of predictable plug-ins (see Waudby-Smith & Ramdas, 2024, Appendix B for an overview), but the method of mixtures has also been studied in this context (Stark, 2020).

Note that unlike in universal inference, the example above defines an e-statistic for all distributions with mean μ without an explicit alternative in mind—similar to what we have seen in Example 4. It is thus a good example of using a pragmatic, or instrumental, alternative to design a powerful e-variable for the null, which in this case is all distributions on $[0,1]$ with mean μ. To elaborate, while no alternative is stated, one is nonetheless implicit. Indeed, every e-variable defines an implicit alternative. What varies from case to case then is not whether an alternative exists but whether it is one that should be considered as actually having generated the data, or simply some-

thing mathematically useful (e.g., Example 4). Example 7 also shows that in practice, e-variables may superficially look very different from likelihood ratios. Nevertheless, by Ramdas et al. (2020, Proposition 4), it turns out that for all $P \in \mathcal{H}$, there exists some distribution Q such that $M_t = q(X_1, \ldots, X_t)/p(X_1, \ldots, X_t)$, i.e., the likelihood ratio between Q and P. Here the q in the numerator varies along with the p in the denominator. Such a "likelihood-like" interpretation holds for general e-statistics. We discuss this more in Sect. 6.

It's worth emphasizing that all of the e-statistics discussed here obey the basic logic described in Sect. 2 which lends them a unified evidential interpretation. This is unlike the situation with likelihood ratios, whose proposed generalizations do not sit comfortably together.

6 Comparing evidence measures

We now spell out in more detail how p-values, Bayes factors, likelihood ratios, and e-statistics fare as measures of evidence. Table 1 summarizes the discussion.

Regarding Bayes factors, we will draw a distinction between the objective and subjective traditions. In the objective tradition, priors are chosen by formal rules that are determined by the problem only and independent of the analyst (Jeffreys, 1939; Kass & Wasserman, 1996; Berger, 2006), whereas in the subjective tradition priors can be any distribution. References to "objective/subjective Bayes factors" should be assumed to mean Bayes factors in the objective/subjective tradition.

We will also discuss both the simple-versus-simple likelihood ratio, which we will refer to as simply the likelihood ratio, and the generalized likelihood ratio defined in (7), which we will refer to as GLR.

For many of the desiderata listed in Sect. 3, the distinction between e-variables and e-processes is immaterial. In these cases, we will refer simply to e-statistics without differentiating the two. The distinction becomes important when discussing counterfactuals and optional continuation; there we will be sure to explicitly distinguish them.

6.1 Static criteria

Continuous measure. P-values, likelihood ratios, and Bayes factors are all graded measures of evidence (Schervish, 1996; Royall, 1997). Smaller p-values are interpreted as more evidence against the null; larger likelihood ratios and Bayes factors are interpreted as evidence in favor of Θ_1 over Θ_0. E-statistics are likewise graded: larger values are more evidence against the null. We stress that the criterion concerns the evidential scale rather than the topology of the map from data to evidence. For an event A with $P(A) > 0$, the indicator e-variable $E(X) = \mathbf{1}_A/P(A)$ is itself not a continuous function of X, yet the evidence it reports is 'continuous' in the following sense: for any other e-value E' relative to any other null hypothesis $\mathcal{H}_0'$, it holds that, if we observe $E'(X) = E(X) + \epsilon$ for $\epsilon > 0$, then we are justified in saying that E' provides more evidence against $\mathcal{H}_0$ than E does against $\mathcal{H}_0$.

Table 1 Comparison of how different evidence functions fare with respect to various evidential desiderata

Desideratum	Trad.	p-values	LRs		Bayes factors		e-statistics	
			simp.	GLR	subj.	obj.	all	some
Scalar measure	B,E,L	✓	✓	✓	✓	✓	✓	✓
Continuous measure	B,E,L	✓	✓	✓	✓	✓	✓	✓
Irrelevance of analyst	E,L	≈	✓	✓	-	✓	≈	✓
Coherence	O	-	-	✓	-	-	-	✓
Consistency	B,L	-	✓	≈	✓	✓	-	✓
Sample size invariance	B,L	-	✓	✓	✓	≈	✓	✓
Scale invariance	B,L	-	✓	✓	✓	✓	≈	✓
Reparam. invariance	B,L	-	✓	✓	≈	≈	≈	✓
Single hypothesis validity	E	✓	-	-	-	-	✓	✓
Error rates	E	✓	✓	≈	-	-	✓	✓
Composite hypotheses	O	✓	-	✓	✓	✓	✓	✓
Nonparametric hypotheses	O	✓	-	✓	≈	≈	✓	✓
Likelihood principle (Strict)	L	-	✓	✓	✓	-	-	≈
Likelihood principle (Loose)	L	-	✓	✓	✓	≈	✓	✓
Composability	O	✓	-	-	-	-	✓	✓
Accumulation I (Fixed)	B,E,L	$\checkmark^+$	$\checkmark^+$	-	✓	≈	$\checkmark^+$	$\checkmark^+$
Accumulation II (Flexible)	O	-	$\checkmark^+$	-	✓	-	$\checkmark^+$	$\checkmark^+$
Dynamic Consistency	B, O	-	$\checkmark^+$	-	✓	-	-	$\checkmark^+$
Inter-experiment counterfactuals	O	-	$\checkmark^+$	-	✓	-	$\checkmark^+$	$\checkmark^+$
Intra-experiment counterfactuals	B,L	-	$\checkmark^+$	$-^*$	✓	-	-	≈
Intra-experiment optional stopping	B,L	-	$\checkmark^+$	$-^*$	✓	-	-	$\checkmark^+$

Here ✓ indicates the desideratum is largely satisfied, "–" that it is either not satisfied or cannot be meaningfully defined, "≈" that it is partially satisfied. In the final six lines, $\checkmark^+$ means that the evidence is well-defined, has a clear interpretation and provides valid error rates. ✓ means that it is well-defined and has a clear interpretation, but in general cannot be used to infer valid error rates. '$-^*$' means that the evidence remains well-defined—and, in the case of intra-experiment counterfactuals, numerically unchanged—so the desideratum is formally satisfied, yet it lacks any clear justification and does not lead to valid error rates, so it is hard to imagine that one would ever want to use it. The "trad." column indicates which tradition the desideratum comes from: L=likelihood tradition, B=Bayesian tradition, E=error-probability tradition, O=other. Since there are typically multiple e-statistics for a given hypothesis, we break their analysis into two categories: 'all' meaning all e-statistics which satisfy the criteria, or 'some' meaning that there exists at least one e-statistic which does (though typically there are many). See the text for more discussion in each case

Irrelevance of the analyst. The desideratum says that two analysts, analyzing the same data and the same $\mathcal{H}_0$ (and, if present, $\mathcal{H}_1$), should come to the same conclusions. This desideratum may be violated if (i) they employ different prior distributions within $\mathcal{H}_0$ (and potentially $\mathcal{H}_1$), whenever these are composite; (ii) they employ different test statistics; (iii) they employ different sampling plans (i.e., the rule that specifies how data will be collected) but happen to obtain the same statistic. Here we are concerned with whether the candidate notions satisfy the desideratum in the sense of not violating either (i) or (ii). (iii) is discussed further below in Sect. 5.2.

It is easy to check that likelihood ratios and GLRs then satisfy the desideratum. Bayes factors in the subjective tradition do not satisfy it by design, whereas Bayes factors in the objective tradition do, as the prior is given by the structure of the problem. Intriguingly, while the "subjectivity" of Bayes factors is a common criticism of

Bayesians made by frequentists, it's questionable whether p-values satisfy this criterion. Indeed, there are often multiple p-values for any given problem. Which one is calculated and reported is, of course, dependent on the analyst. We have thus marked p-values as only partially satisfying this requirement.

Similar comments apply to e-statistics. There are multiple ways to play the betting game, and different analysts might thus use different e-statistics and come to different conclusions. Further, while they are perfectly well-defined in the frequentist setting, e-statistics *can* be defined in the Bayesian setting and, as discussed above, optimal e-variables can often be recovered as Bayes factors with (sometimes peculiar) priors. Like p-values then, we mark e-statistics as only partially satisfying this criterion.

Coherence. Both p-values and Bayes factors can be incoherent (Lavine & Schervish, 1999; Fossaluza et al., 2017). Moreover, since the notion of coherence requires evidence for or against *composite* hypotheses, the ordinary simple-versus-simple likelihood ratio is too narrow to address this desideratum. The GLR, however, is coherent (Bickel, 2012; Fossaluza et al., 2017).

Not all e-statistics are coherent. Bickel (2024) observed that the numeraire e-variable (Larsson et al., 2025) is not coherent, whereas the universal inference e-variable (and e-process) is. The upshot is thus not that e-statistics are coherent to some intermediate degree, but that coherence is satisfied by some, but not all, e-statistics. See Ramdas and Wang (2025, Chap. 6) for further discussion on coherent e-variables.

Consistency. Not all e-statistics are consistent, but many are. In fact, it is typically easy to construct e-variables and e-processes which satisfy such a property. To give but a short list: those designed for testing means of bounded random variables (Waudby-Smith & Ramdas, 2024), those designed for testing exponential families (Grünwald et al., 2024, Hao & Grünwald, 2024), two-sample testing (Shekhar & Ramdas, 2023) and others (Waudby-Smith et al., 2025). Still, there are obviously e-statistics that are inconsistent—an example being the constant process $E_\tau \equiv 1$ that ignores all data. Meanwhile, recent work shows that universal inference can be conservative asymptotically, suggesting that universal inference e-values may not satisfy asymptotic consistency (Takatsu, 2025).

The p-value is not consistent, while the likelihood ratio is (Grendár, 2012). The GLR is consistent under regularity and separability assumptions (Bickel, 2012), but not in general. Bayes factors are consistent in most practical settings (so we marked the desideratum as satisfied), though there are some exceptions depending on the prior and the hypothesis class (Casella et al., 2009; Moreno et al., 2010).

Sample size invariance. Both likelihood ratios and subjective Bayes factors have sample size invariant interpretations. The likelihood ratio, for instance, is interpreted as: *if* $p_{\theta_1}(X)/p_{\theta_0}(X) = c$, *then the data are* c *times as likely under* p_{θ_1} *as they are under* p_{θ_0}, a statement which does not mention the sample size. Or, in the words of Goodman and Royall (1988), "the likelihood ratio has the same meaning in trials of different designs and sizes." Similarly for the GLR. Meanwhile, e-statistics can be seen as monetary gains in a betting game, an interpretation which also doesn't depend on the sample size. p-values, on the other hand, are not sample size invariant (Goodman & Royall, 1988). For objective Bayes, the answer depends on the hypotheses under consideration. For some hypotheses (such as in linear regression), the prior advocated in some objective Bayes methods depends on the covariates (De Heide &

Grünwald, 2021) and therefore, implicitly, also on the sample size. Thus, sample size invariance fails.

Scale invariance. Both likelihood ratios and Bayes factors are invariant to one-to-one transformations of the data because they are ratios of probabilities. To elaborate, under any such transformation both densities are multiplied by the same quantity: If $Y = g(X)$ then the new density for Y is $p_{\theta_j}(y) = p_{\theta_j}(x)|\det J_{g^{-1}}(y)|$ where $J_{g^{-1}}$ is the Jacobian. Thus, the change to both numerator and denominator cancel out. (Note the Jacobian only appears in continuous-valued data; for discrete data the invariance is immediate.) The GLR is scale invariant for the same reason.

Scale invariance for e-statistics is more subtle. Some e-statistics automatically satisfy this property, such as universal inference (again using that it is the ratio of probabilities) and those based on self-normalized processes (Wang & Ramdas, 2025) (since self-normalized statistics are scale free).

However, not all e-statistics are immediately scale invariant. Consider $E(X) = \mathbf{1}\{X \leq 1\}/P(X \leq 1)$ for any distribution P with $P(X \leq 1) > 0$. This is clearly an e-variable. If $Y = aX$, then $E(Y) \neq E(X)$, so E is not scale invariant. This is because a threshold has been hard-coded in the example. One should arguably parameterize this threshold and instead consider the family $E_s(X) = \mathbf{1}\{X \leq s\}/P(X \leq s)$. In this case, $E_s(Y) = E_{s/a}(X)$.

Allowing the parameter space to change alongside the data is natural. For instance, following Example 3, $E_\lambda(X) = \exp(\lambda X - \lambda^2/2)$ is an e-variable for $\mathcal{H} = \{N(0,1)\}$. Suppose we let $Y = aX$, so under the null, $Y \sim N(0, a^2)$. Then E is scale invariant in the sense that $E_\lambda(X) = E_{\lambda/a}(Y)$, where on the right E denotes the family of Example 3 with null variance a^2. Notice that $E_\lambda(X)$ is the likelihood ratio of $N(\lambda, 1)$ and $N(0, 1)$. That is, the parameter λ is encoding the alternative distribution. Hence, when we refer to the likelihood ratio being scale invariant, we are implicitly allowing the parameter to change.

Reparameterization invariance. Using the same logic as above, likelihood ratios are invariant to transformations of parameters (cf. McCullagh & Cox, 1986). So too are Bayes factors if one allows the prior to change along with the reparameterization. That is, if we begin with a prior π over parameters θ, reparameterize $\theta \mapsto \phi(\theta)$, and consider the pushforward measure $f(\phi) := \pi(\theta)|\det \mathrm{d}\theta/\mathrm{d}\phi|$ as the new prior, then the Bayes factor is invariant.

However, considering the pushforward is not always done in practice. For instance, choosing a uniform prior in one parameterization will not necessarily yield a uniform prior in another. Thus, if one is committed to a specific class of priors, the Bayes factor can depend on the particular parameterization. Some Bayesians find this troubling, which led to invariant priors (Jeffreys, 1946). Neither objective Bayes factors nor subjective Bayes factors always use invariant priors, however.

A similar nuance applies to e-statistics as it did in the case of scale invariance. As an object satisfying $\sup_{P \in \mathcal{H}} \mathbb{E}_P[E] \leq 1$, a change in the parameter space simply relabels the null hypothesis, and does not affect the guarantee afforded by the e-statistic. In other words, e-statistics are invariant to transformations of the parameter in the sense that they remain e-statistics. However, if one is committed to a specific "generating procedure" for e-variables (e.g., universal inference with prior $\pi(\theta)$, or constructing E_λ using a specific λ), then its value might change after a transforma-

tion of the parameters. Therefore, we mark general e-statistics as only partially satisfying this requirement.

Finally, not all p-values are reparameterization invariant, leading to the search for some which are (Evans & Jang, 2010).

Single hypothesis validity. P-values and e-statistics are defined only in terms of a single hypothesis $\mathcal{H}$, whereas both likelihood ratios and Bayes factors require two. See, e.g., Example 4 and 7. That is, the latter provide relative evidence between two hypotheses and the former provide a notion of absolute evidence against a single hypothesis.

Long-run error rates. This is perhaps the defining feature of evidence from the error-probability tradition. It holds that if one uses a measure of evidence to make decisions (which is itself already a controversial endeavor, and marks one of the divides between the Fisherian and Neyman-Pearson schools of thought), then the frequentist error on those decisions should be controlled. More precisely, if we repeat the decision process many times, we should have some guarantee on how often the decision will be wrong.

As discussed in Sect. 2, Markov's inequality (and Ville's inequality for e-processes) gives us control over type-I error rates for e-statistics. The likelihood ratio, being an e-value, also gives error control, but the GLR on its own does not. There, error rates must be obtained by appealing to the sampling plan, underlying distribution, or other machinery. Since in practice, GLRs are almost always used within a context in which such error rates are analyzed as function of the GLR for a given sampling distribution, we still marked them as approximately satisfying this criterion. Subjective Bayes factors also do not provide type-I error control (Grünwald et al., 2024); for objective Bayes factors, there are some cases in which type-I error control is provided, such as the t-test Bayes factor of Example 6, but usually it is not (note that we refer to exact, nonasymptotic error rates here).

It's worth mentioning so-called "post-hoc validity," a new development in the theory of e-statistics. When measures of evidence are used to make decisions, e-statistics continue to provide meaningful error guarantees, in the form of control on expected risk, under data-dependent significance levels (Grünwald, 2023, 2024; Chugg et al., 2026). This is not allowed by traditional p-values. The property matters because significance levels are, in practice, often not fixed before the data are seen. An analyst who sets out with $\alpha = 0.05$ but, on seeing a striking result, reports it at $\alpha = 0.01$ has broken the contract under which the p-value's risk guarantee was issued. Post-hoc validity says that the guarantee attached to an e-value survives such after-the-fact choices of level. Further, the only subset of p-values which do provide such a guarantee turn out to be precisely the inverse of e-variables, both in asymptotic and non-asymptotic settings (Wang & Ramdas, 2022; Koning, 2023; Chugg et al., 2026).

Composite and nonparametric hypotheses. The only evidential measure that, by definition, cannot handle composite hypotheses is the simple-versus-simple likelihood ratio. All other measures handle composite hypotheses and, in principle, even handle highly complex nonparametric hypotheses.

Some evidential measures, however, handle such complex classes more naturally than others. For instance, there exist simple e-statistics for the class of all bounded distributions with mean μ (Example 7), which is a large nonparametric family. By

contrast, to define a Bayes factor for this same hypothesis one must place a prior on an infinite-dimensional space of distributions (or otherwise restrict attention to a parametric or semiparametric sub-model). This can certainly be done, but it is considerably less direct and substantially more prior-dependent (and can lead to less evidence against hypotheses that are evidently wrong in light of the data; see Li et al., 2023). For this reason, we regard this criterion as only partially satisfied by both subjective and objective Bayes factors.

The likelihood principle. The likelihood principle is something like the founding charter, or the constitution, of the likelihoodists. It is useful to distinguish between two versions: the *strict* and *loose* likelihood principle. The strict version holds that if two data sets induce proportional likelihood functions for the parameters of interest (i.e., for all parameters in $\mathcal{H}_0$ and $\mathcal{H}_1$), then the evidence should be the same for both data sets. The loose version holds that evidence can always be written as, or at least upper bounded by, a likelihood ratio with some element of $\mathcal{H}_0$ in the denominator.

Let us deal first with the strict likelihood principle. This is, of course, satisfied by the likelihood ratio. For the subjective Bayes factor, it is satisfied if one identifies likelihood with marginal likelihood over the prior. Here we assume, as is usually done, that the priors are chosen *independently* of the sampling plan, i.e. the analyst's prior assessment of the parameters determining the actual distribution is independent of the sampling plan employed. This is in contrast to objective Bayesian approaches, in which priors are chosen by formal means that often depend on aspects of the sampling plan such as the stopping time (Berger & Wolpert, 1988). Indeed, Berger and Wolpert (1988), De Heide and Grünwald (2021) give concrete examples where Jeffreys' prior depends on the sampling plan, hence the value of the objective Bayes factor is not merely determined by the likelihoods and as such it violates the strict likelihood principle. P-values violate it for the same reason, whereas the GLR satisfies it.

As for e-processes, the strict likelihood principle is satisfied by the likelihood ratio e-process and the universal inference e-process, when the alternative is specified by a prior as in Example 5, Eq. (9). There is a subtle catch which prevents us from saying that the strict version is always satisfied by e-statistics of the form (9), however. We describe this issue further below under intra-experiment counterfactuals.

The loose version of the likelihood principle is clearly satisfied by any evidential measure satisfying the strict version. Moreover, it is once again violated by p-values. As to e-statistics, as shown in Ramdas and Wang (2025, Theorem 14.5) (a special case of which we mentioned in Sect. 4), any e-variable E for $\mathcal{H}$ can be shown to be dominated by likelihood ratios, in the sense that for any $P \in \mathcal{H}$, there exists some Q such that $E(X) \leq q/p$; see Sect. 6 for a simple proof. For objective Bayes, this is not the case, but since objective Bayesian evidence can always be written as a likelihood ratio with a (potentially improper) prior over Θ_0 in the denominator, we still mark it as partially satisfied.

Composability under dependence. The convex combination of any set of (arbitrarily dependent) e-statistics remains an e-statistic: If $E^1, \ldots, E^K$ are e-variables then $\lambda_0 + \sum_{1 \leq k \leq K} \lambda_k E^k$ is an e-variable where $\sum_{0 \leq i \leq K} \lambda_i \leq 1$, $\lambda_i \geq 0$. Similarly for e-processes. In fact, the only admissible way of merging e-variables is with

a combination such that $\sum_{0\leq i\leq K}\lambda_i = 1$ (Wang, 2025). There also exist rules for combining arbitrarily dependent p-values, including two times the average and the median of the p-values. See Vovk et al. (2022) for a nice overview. Marginal likelihood-ratios or Bayes-factors from separate sources cannot in general be merged when their dependence structure is unknown.

6.2 Dynamic criteria

To assess the dynamic criteria introduced in Sect. 3, it will be convenient to introduce some unified notation that we can use throughout the section. We will assume that we want to measure the evidence of some data X^* which decomposes as $X^* = (X_{(1)}, X_{(2)}, \ldots, X_{(\tau)})$. For the inter-experiment dynamic criteria (the first four criteria below), $X_{(i)}$ is itself the data produced by some study. For the intra-experiment criteria, $X_{(i)}$ is more fruitfully considered to be a single datum and X^* are the data of one study. In both scenarios, the total number of studies run or observations collected, τ, may or may not be known in advance (depending on the desideratum).

Accumulation I: Fixed-Design. If the $X_{(j)}$ are independent and τ is known in advance or determined independently of the data, then the likelihood ratio decomposes as $\mathrm{LR} = \prod_{j\leq\tau} p_1(X_{(j)})/p_0(X_{(j)})$, meaning that we can combine evidence by multiplication.

As is well known, the same multiplicative accumulation holds for Bayes factors when the priors π_0 and π_1 are fixed in advance, independently of the data-collection process (what we call the subjective Bayesian setting). In that case, independent studies can be combined sequentially by updating each prior to the corresponding posterior after each study. The overall Bayes factor is then the product of the stage-wise Bayes factors.

The situation is less clear for objective Bayes methods. There, the prior is often not a genuinely fixed ingredient, but is instead chosen by a rule that may depend on features of the experimental design, such as the sample space, stopping plan, or covariate structure. If later studies are added, the prior deemed appropriate for the combined experiment may differ from the prior used for the first study considered in isolation. As a result, sequential accumulation becomes ambiguous: the product of the stagewise Bayes factors need not coincide with the Bayes factor that would have been computed had the full experiment been specified from the start (De Heide & Grünwald, 2021). Since this happens only for some, and not nearly all, choices of $\mathcal{H}$, we marked the requirement as only partially satisfied.

The GLR, meanwhile, is not additive on any scale, as the parameter achieving (or approximating) the supremum may change for different samples. Independent p-values can be combined using Fisher's method, which adds not the p-values themselves but their negative logarithms, and the product of two conditionally independent e-statistics remains an e-statistic, a property which follows easily from the definition of the expected value. See Ramdas and Wang (2025, Chap. 8) for more discussion.

Accumulation II: Flexible-Design. Now suppose that the decision to perform an additional study depends on the results of previous studies. If the data $X_{(j)}$ is

independent of $X_{(1)}, \ldots, X_{(j-1)}$ and $X_{(i)}$ is associated with e-variable E_i, then the product $\prod_{i \leq \tau} E_i$ is an e-variable. This was a major motivation in the original development of e-statistics (Grünwald et al., 2024). This property is also satisfied by e-processes that, in the inter-experimental setting, are constructed on the fly from a sequence of e-variables.

GLRs fail to satisfy Accumulation II for the same reason they failed to satisfy Accumulation I. Objective Bayes factors fail this criterion more dramatically than they did Accumulation I: the prior for the initial experiment may not only be chosen in different ways, its definition will in general depend on the number of studies that will eventually be done. Hence the objective Bayes factor may in fact be undefined. Subjective Bayes factors, on the other hand, continue to satisfy this criterion (Rouder, 2014; De Heide & Grünwald, 2021).

P-values fail for this and all subsequent desiderata because they require that X^* is fully specified in advance of running the study. Consider the following example.

Example 8 Suppose that a randomized clinical trial to test a new medical treatment is performed on 50 patients represented by 50-dimensional data vector $X_{(1)}$. The result turns out to be *promising but not conclusive*: the researchers observed a p-value of $p_1 = 0.1$ while they had a significance level of 0.05 in mind. But their boss is optimistic at the news and agrees to supply the resources to test another 30 patients, resulting in data $X_{(2)}$.

Is this good news? Not if one insists on measuring evidence in the second trial by another p-value, say p_2. For some realizations of $X_{(1)}$, we might decide not to gather $X_{(2)}$. As a result, standard combination methods for p-values like Fisher's cannot be employed any more, since they invariably require that, no matter what is observed in each sample, we always combine both. Similarly, joining the two data sets and recalculating the p-value leads to a wrong answer as well. Indeed, take a method that stops for some values of $X_{(1)}$ and in that case outputs $p^* = p_1$, a sharp, nonconservative p-value for $X_{(1)}$, yet continues for other values of $X_{(1)}$ and in that case, after observing $X_{(2)}$, outputs $p^* = p'$ with p' a number strictly smaller than 1. It is easily shown that for any such method, the resulting p^* is not a valid p-value anymore. In fact, the only known way to obtain a sequence of *anytime-valid* p-values in such settings is to convert each p_i, $i = 1, 2$, into an e-value E_i by calibration (Sect. 4), and then report not p_i itself but $p'_i := 1/E_i$. The quantity $p'_{12} := 1/(E_1 E_2)$ can then be interpreted as a p-value for the combined data. In practice, however, this approach is typically inefficient (Grünwald et al., 2024, Section 7), and one can do better by working directly with e-values.

Dynamic Consistency. Dynamic consistency asks that we obtain the same evidence when we treat X^* as a batch versus when we treat samples $X_{(j)}$ separately, sequentially, and then combine them. For p-values, likelihood ratios, and Bayes factors, satisfying dynamic consistency is equivalent to satisfying accumulation under a flexible design (here we again assume that τ can be data-dependent).

For e-statistics we must distinguish between several cases. If we design an e-process for the stream of data $X_{(1)}, X_{(2)}, \ldots$, then dynamic consistency is satisfied (Ramdas et al., 2023). If, however, we model the individual $X_{(j)}$ by separate

e-variables and combine them via multiplication, then this desideratum is not satisfied in general (Grünwald et al., 2024, Section 6). In particular, for some e-variables the product $\prod_{j=1}^{\tau} E_j$ of the e-values E_j for study $X_{(j)}$ does not coincide with the single e-value E^* for X^* that one would have obtained if one had designed a log-optimal e-variable for X^* directly (though we emphasize that, as stated earlier, the product $\prod_{j=1}^{\tau} E_j$ is a valid e-variable, which of course yields valid error rates as discussed in Sect. 2).

Inter-experiment counterfactuals. A common criticism of p-values is that, paradoxically, they depend on data that are not observed, or, more precisely, on what decisions an experimenter would have taken in counterfactual situations in which the data were different (Barnard, 1947; Wagenmakers, 2007). There are multiple kinds of counterfactuals to consider, which we delineate as inter- and intra-experiment counterfactuals[14]. The former, discussed here, concerns counterfactuals with respect to the number of observations, which might also be thought of as optional continuation in counterfactual situations.

Example 9 (Example 8, continued) In a variation of the previous example, suppose the researchers told their boss merely that the p-value was small enough for the result to be "promising but not conclusive," and not its actual value. The boss, once again, responds optimistically and requests that they continue the trial on a second batch of 30 patients. The researchers, who know their statistics, are now worried about invalidating the results. But then news reaches the boss that the p-value is 0.1. Dismayed, he decides to stop the trial.

Should the researchers be relieved? *The answer is an emphatic no*: a calculation similar to the one in the previous example shows that, if the originally reported p-value was sharp, then the mere fact that the sample would have been 80 patients (thus different from the originally planned 50) in a counterfactual situation (i.e. if the first 50 data points had been different than they actually were) makes the p-value invalid. That is, counterfactuals can ruin the validity of the p-value even if the sample plan was not in fact changed for the data which were actually observed.

The dependence of the p-value on decisions that would have been made in counterfactual situations is disturbing, since in reality, it is often quite unknowable (or even meaningless to speak about) what exactly would have happened had the data been different from what they actually were. That p-values cannot handle counterfactuals is an immediate consequence of the fact that they cannot deal with accumulation. Similarly for the objective Bayes factor and the GLR.

For e-variables and e-processes in inter-experimental settings (where the e-processes are constructed by multiplying the e-variables), this desideratum is satisfied because the constructed e-variables are valid at arbitrary stopping times. That is, regardless of the data witnessed thus far, stopping an e-process will always result in

[14] We have not found this distinction anywhere else in the literature, but it is important to make since the criteria behave quite differently under each sub-division!

a valid e-value. Similarly for subjective Bayes factors, which also remain valid at stopping times.

Note that this implies that p-values which can be written as the inverse of e-values also satisfy this property. In general, however, such p-values are conservative and are not the ones used in practice.

Intra-experiment optional stopping. Now we suppose that X^* represents data from a single study in which we observe τ data points, $X^* = (X_1, \ldots, X_\tau)$. As usual, τ is not fixed: it is a stopping time whose definition the analyst may not know or may not want to fix in advance. Consider the following simple example in the spirit of Lindley and Phillips (1976).

Example 10 Suppose the data are binary and we observe $X^* = (0, 1, 0, 1, 1)$, but do not know whether the sample size was fixed to be 5 in advance or if the sampling plan was "stop as soon as you see two 1s in a row." Or, then again, maybe each data point is expensive and the p-value for a sample size of 5 was already so small that your boss decided data collection should be stopped.

As we stated in Sect. 3, this criterion is very similar to Accumulation II, but making the distinction between optional continuation 'inside' a single study versus 'between' studies sheds light on several subtleties for e-statistics. As we saw in Sect. 2, e-processes are defined such that they allow optional stopping. Likelihood ratios, being e-processes, also satisfy this criterion. On the other hand, for e-variables (when directly defined on the batch X^*, as they often are in practice), optional stopping is an undefined operation.

As for the other notions of evidence, it is well-known that p-values do not allow for data-dependent stopping times. In fact, picking the sample size as a function of the data and reporting the resulting p-value has come to be known as "p-hacking," and is often cited among the most common "questionable research practices" (Simmons et al., 2011; John et al., 2012). Bayes factors, meanwhile, allow for optional stopping as long as the priors are chosen a priori, independently of the stopping time (though note that they do not in general provide type-I error rates, in contrast to likelihood ratios). We refer to De Heide and Grünwald (2021) and Rouder (2014) for more discussion on Bayes factors and optional stopping.

Like the Bayes factor with subjective prior, the GLR remains well-defined under optional stopping, but like the Bayes factor, it fails to provide valid error rates, as the supremum can exaggerate evidence for the alternative (Ramdas et al., 2023). Therefore, while the Bayes factor still has a clear interpretation and can be justified in case the priors reflect trustworthy prior knowledge, using the GLR here has no clear justification at all. In practice, the GLR is usually employed within the error-probability tradition, and its main goal is to provide error rates, which it cannot provide here.

Intra-experiment counterfactuals. Researchers within the likelihood tradition have observed that the p-values' dependence on decisions made in counterfactual situations stretches beyond only the size of the dataset, as illustrated by the following celebrated example:

Example 11 (Pratt's voltmeter, cf. Edwards, 1972; Savage et al., 1962) Suppose we observe $X_1, \ldots, X_n$ where n is fixed and the X_i represent voltages of electron tubes, measured with an accurate voltmeter. A statistician examines the X_i assuming they are normally distributed with fixed variance and some mean μ. He aims to use a p -value to measure the evidence against the null hypothesis that $\mu = 7.0$. Later he visits the engineer's laboratory, and notices that the voltmeter reads only as far as 10: the population appears to be *censored*. Even though none of the X_i were 10, this makes the standard p-value invalid; it necessitates a new calculation based on a p-value that takes into account the (potential, counterfactual) censoring. However, the engineer says she also has a super-high-range-meter, equally accurate, which she would have used if any of the measurements had turned out ≥ 10. This is a relief to the statistician, because it means the original p-value is correct after all. But the next day the engineer telephones and says, *"I just discovered my high-range voltmeter was not working the day I did the experiment"*. The statistician then informs her that a new analysis will be required after all! The engineer is astounded. She says, "But the experiment turned out just the same as if the high-range meter had been working. *I learned exactly what I would have learned if the high-range meter had been available*. Next you'll be asking about my oscilloscope!"

We may think of both the voltmeter example and Example 10 as an instance of the following generic phenomenon.

The random variable X^*, taking values in some set $\mathcal{X}^*$, defines a partition of the sample space Ω, say $\{\mathcal{E}_x : x \in \mathcal{X}^*\}$, where $\mathcal{E}_x$ is the set of exactly those $\omega \in \Omega$ with $X^*(\omega) = x$. Observing $X^* = x$ is equivalent to observing that the event $\mathcal{E}_x$ occurred. Now, it is usually tacitly assumed that the definition of X^* is known to the observer. In practice, however, an analyst simply observes an event $\mathcal{E}_x$, for some $x \in \mathcal{X}^*$. The analyst then knows that $\mathcal{E}_x$ happened but often does not know how X^* was even defined. This is the case in Example 10, where we observed $X^* = (X_1, \ldots, X_\tau) = (0, 1, 0, 1, 1)$ but did not know whether τ was defined as the constant 5 or as 'stop when you see two 1s'. It is also the case in the voltmeter example, where we observed $(x_1, \ldots, x_n)$ with all $x_i < 10$ but do not know whether $X^* = (X_1, \ldots, X_n)$ or $X^* = (X_1 \wedge 10, \ldots, X_n \wedge 10)$.

Now, above we defined likelihood ratios as functions of the random variable X^* , but we may equivalently define them as functions of events, i.e. measurable subsets of Ω. To avoid difficult mathematics, let us illustrate this in the case of discrete data. For any event $\mathcal{E}$, any random variable X, and any value x such that $X = x$ corresponds to event $\mathcal{E}_x$, we have $p_1(x)/p_0(x) = P_1(\mathcal{E})/P_0(\mathcal{E})$ as long as $\mathcal{E} = \mathcal{E}_x$. Thus, we can re-define likelihood ratios as functions of events and will get the same results as before, irrespective of the definition of X^*. As a consequence, the likelihood ratio is immune to decisions in counterfactual situations in a very general sense. The same holds for the subjective Bayes factor. For the objective Bayes factor and isolated e-variables, the desideratum fails to be met for the same reasons as for inter-experiment counterfactuals. The GLR is a different case. Being a functional of the likelihood alone, its value is unchanged under such counterfactuals. In the voltmeter example, if every observation falls below the censoring point, the likelihood—and hence the GLR—is exactly what it would have been had the voltmeter been

unrestricted. What does change is its sampling distribution, so thresholds calibrated under one protocol need not retain their error guarantees under the other. Table 1 thus records '$-^*$' for the GLR.

The situation is trickier for e-processes. Since likelihood ratios define e-processes, we might say this desideratum holds for at least some e-processes. On the other hand, one invariably defines e-processes on filtrations (essentially sequences of random variables), not on arbitrary events, and at the time of writing it is not clear whether they can be extended to events beyond the simplest case in which $\mathcal{H}_0$ and $\mathcal{H}_1$ are simple (this is also the reason we only marked the strict likelihood principle to be only approximately satisfied by e-processes—likelihood ratios are defined on arbitrary events, and e-processes on the more restricted notion of filtrations). Further work is required to answer this question.

Overall, we find that while no single e-statistic satisfies all evidential desiderata in Sect. 3, each desideratum is (at least partially) satisfied by at least one e-statistic. Moreover, when we restrict attention to criteria satisfied by *all* e-variables, they satisfy the largest set of desiderata among those measures which can handle composite hypotheses (which we deem to be a very important property!). Indeed, as we've emphasized throughout and will explore more shortly, we may think of e-statistics as providing a novel extension to likelihood ratios, arguably more sophisticated than the generalized likelihood ratio, that behaves particularly well on most common desiderata.

With that said, let us not pretend that e-statistics are the perfect measure of evidence. We end by considering some of their drawbacks.

7 Limitations and counterarguments

One could of course object to e-statistics as a measure of evidence by appealing to any of the criteria listed above that they do not wholly satisfy. Instead, in this section we consider two broader critiques.

7.1 Non-uniqueness

The first is the fact that there can be many e-variables for a given problem. Even if one uses e-statistics, the amount of evidence against a hypothesis is not determined by the structure of the problem alone, but also by how the analyst chooses to bet in the betting game.

We are sympathetic to this concern, but let us point out that the situation is not so different from other measures of evidence. For instance, there are typically multiple p-values that one may use for a given problem, and the Bayes factor depends on the choice of prior which may vary from analyst to analyst. Further, while the likelihood

ratio for simple hypotheses is uniquely determined, when moving to composite settings the analyst has their choice of possible generalizations.[15]

Nevertheless, the situation is arguably somewhat clearer for p-values than for e-values. In particular, within the error-statistics tradition, there is broad consensus that the quality of a p-value, when used for testing, is determined by the maximal power it can achieve. For sufficiently simple $\mathcal{H}_0$ and $\mathcal{H}_1$—though only in such cases—one can define a uniformly most powerful p-value. In the e-value literature, by contrast, the preferred optimality criterion under a simple alternative is often taken to be log-optimality. For composite alternatives, however, there are multiple ways of generalizing log-optimality, and these lead to quite different e-statistics. More broadly, there is less consensus in the e-statistics literature that log-optimality is the right criterion than there is in the error-statistics tradition that power is the right one (Koning, 2024).

7.2 Birnbaum's theorem and the likelihood principle

Birnbaum (1962) showed that two conditions entail the likelihood principle: the sufficiency principle and the conditionality principle. The result has come to be known as Birnbaum's theorem.

The sufficiency principle states that if there exists a sufficient statistic for the statistical model, then it captures everything about evidence. That is, if two experiments give the same sufficient statistic, then they provide the same evidence for the parameter. The conditionality principle states that evidence should only depend on which experiment was actually run. For instance, if an analyst decides between two experiments by flipping a coin, the evidence should depend only on that which was performed.

Birnbaum's theorem is controversial because the conditionality and sufficiency principles are relatively weak, striking many frequentists as perfectly plausible. The conclusion, however, contradicts the frequentist's favorite inferential methods. The theorem has thus fallen under various forms of scrutiny (Durbin, 1970; Kalbfleisch, 1975; Evans et al., 1986; Evans, 2013; Mayo, 2014; Gandenberger, 2015).

To make the controversy more concrete, consider again Example 10. The likelihood ratio is entirely determined from these data, thus the likelihoodist is satisfied that they have enough evidence about the situation to make a pronouncement. But the frequentist who wants to employ a p-value is not. For them, the sampling plan matters: the p-value differs in all cases.

How do e-statistics fit into this story? As we've emphasized throughout the article, e-statistics are intimately related to likelihood ratios. For one, the simple-versus-simple likelihood ratio is an e-variable, and when testing two point hypotheses, it is the unique log-optimal e-variable (see Shafer (2021) for a simple proof). Second, we can upper bound any e-statistic E for $\mathcal{H}$ via a likelihood ratio as follows. Given $P \in \mathcal{H}$

[15] One might also tie this non-uniqueness to a familiar epistemological point, which is that data are always interpreted through our theories, and there is thus never only one way to interpret them. In the jargon, observation is always *theory-laden* (Popper, 1963).

, define a distribution Q via the density $q = p \cdot E/\mathbb{E}_P[E]$ if $\mathbb{E}_P[E] > 0$, and $Q = P$ otherwise. Then, since $\mathbb{E}_P[E] \leq 1$,

$$E \leq \frac{q}{p}. \tag{11}$$

We have already seen an instance of this in Example 4, where the data-compression e-variable was re-expressed as a likelihood ratio. When $\mathbb{E}_P[E] = 1$ (as it is for $E = E_\tau$ for all bounded stopping times τ in the e-process in Example 7, for example), (11) becomes an equality. See Ramdas and Wang (2025, Theorem 14.5) for a slightly more general statement.

Thus, while e-statistics commit to a distinct core axiom ($\mathbb{E}_P[E] \leq 1$), they often recover the likelihood ratio, or can be expressed or bounded in terms of likelihood ratios. E-statistics therefore occupy an intermediate position in the debate around Birnbaum's theorem: they are closely connected to likelihood ratios, yet are not defined by a commitment to the likelihood ratio as the canonical measure of evidence. That stance may leave ardent likelihoodists unsatisfied, but others may regard it as a reasonable price to pay for the broader advantages of e-statistics.

8 Summary and future work

We have positioned e-statistics in the broader literature on statistical evidence. They can be viewed as generalizations of the likelihood ratio, and thus inherit many of the likelihood ratio's appealing evidential properties. E-statistics are fundamentally frequentist objects (though they are allowed to depend on priors via the method of mixtures (Example 5) and recover Bayes factors in particular settings), thus avoiding some of the criticisms of Bayesian confirmation theory. Their game-theoretic meaning as the wealth accumulated by a statistician betting against nature lends them an intuitive interpretation-one that, like p-values, is well-defined even when considering a single hypothesis. Overall, e-statistics blend several attractive features of p-values, likelihood ratios, and Bayes factors, and we hope to have convinced the reader that they are worthy of consideration as compelling measures of statistical evidence.

With all that said, many evidential aspects of e-statistics remain underexplored. First, Koning (2024), building on work of Ramdas and Manole (2026), introduced a novel notion of evidence against $\mathcal{H}$: the maximum probability with which a certain randomized test, while controlling type-I error, rejects the null. We did not include this notion in our comparison because it differs substantially from existing desiderata, though it may well deserve further study. Second, e-processes as currently defined do not fully satisfy the intra-experiment counterfactual desideratum. Is it possible to generalize their definition so that some do?

Acknowledgements We thank the anonymous referees for helpful comments and Jürgen Landes for his patience.

Author contributions Ben Chugg: ideation, writing. Peter Grünwald: ideation, writing. Aaditya Ramdas: ideation, writing.

Funding Open Access funding provided by Carnegie Mellon University. BC and AR acknowledge support from NSF grants IIS-2229881 and DMS-2310718. PG was co-funded by the European Union (ERC Advanced Grant FLEX 101142168).

Data availability Not applicable.

Code availability Not applicable.

Materials availability Not applicable.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication All authors consent to the publication of this manuscript.

Competing interests The authors declare that they have no competing interests.

References

Barnard, G. A. (1947). The meaning of a significance level. *Biometrika*, *34*(1/2), 179–182. https://doi.org/10.1093/biomet/34.1-2.179

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*(3), 385–402. https://doi.org/10.1214/06-BA115

Berger, J. O. (1983). *In defense of the likelihood principle: Axiomatics and coherency*. Purdue University. Department of Statistics.

Berger, J. O., Ghosh, J. K., & Mukhopadhyay, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *Journal of Statistical Planning and Inference*, *112*(1–2), 241–258. https://doi.org/10.1016/S0378-3758(02)00336-1

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, *82*(397), 112–122. https://doi.org/10.1080/01621459.1987.10478397

Berger, J. O., & Wolpert, R. L. (1988). The likelihood principle. In *Institute of mathematical statistics: Lecture notes–monograph series*. IMS.

Bickel, D. R. (2011). A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons. *Canadian Journal of Statistics*, *39*(4), 610–631. https://doi.org/10.1002/cjs.10109

Bickel, D. R. (2012). The strength of statistical evidence for composite hypotheses: Inference to the best explanation. *Statistica Sinica*, 1147–1198.

Bickel, D. R. (2024). David R. Bickel's contribution to the discussion of 'Safe testing' by Grünwald, De Heide, and Koolen. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *86*(5), 1133–1134. https://doi.org/10.1093/jrsssb/qkae089

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, *57*(298), 269–306. https://doi.org/10.1080/01621459.1962.10480660

Birnbaum, A. (1977). The Neyman-Pearson theory as decision theory, and as inference theory; with a criticism of the Lindley-Savage argument for Bayesian theory. *Synthese*, *36*(1), 19–49. https://doi.org/10.1007/BF00485690

Blume, J. D., Greevy, R. A., Welty, V. F., Smith, J. R., & Dupont, W. D. (2019). An introduction to second-generation p-values. *The American Statistician*, *73*(1), 157–167. https://doi.org/10.1080/00031305.2018.1537893

Casella, G., Girón, F. J., Martínez, M. L., & Moreno, E. (2009). Consistency of Bayesian procedures for variable selection. *The Annals of Statistics*, *37*(3), 1207. https://doi.org/10.1214/08-AOS606

Chib, S., & Kuffner, T. A. (2016). Bayes factor consistency. arXiv preprint arXiv:1607.00292.

Chugg, B., Gauthier, E., Jordan, M. I., Ramdas, A., & Waudby-Smith, I. (2026). Post-hoc large-sample statistical inference. arXiv preprint arXiv:2603.08002.

Chugg, B., Lardy, T., Ramdas, A., & Grünwald, P. (2026). On admissibility in post-hoc hypothesis testing. *International Journal of Approximate Reasoning*, *191*, 109634. https://doi.org/10.1016/j.ijar.2026.109634

Cover, T., & Thomas, J. (1991). *Elements of information theory*. Wiley-Interscience.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society Series B*, *34*(2), 187–220.

De Heide, R., & Grünwald, P. D. (2021). Why optional stopping can be a problem for Bayesians. *Psychonomic Bulletin & Review*, *28*(3), 795–812. https://doi.org/10.3758/s13423-020-01803-x

Durbin, J. (1970). On Birnbaum's theorem on the relation between sufficiency, conditionality and likelihood. *Journal of the American Statistical Association*, *65*(329), 395–398. https://doi.org/10.1080/01621459.1970.10481088

Edwards, A. (1972). Likelihood. An account of the statistical concept of likelihood and its application to scientific inference. *British Journal for the Philosophy of Science*, *23*(2), 490–490. https://doi.org/10.1088/0031-9112/23/8/030

Eells, E., & Fitelson, B. (2000). Measuring confirmation and evidence. *The Journal of Philosophy*, *97*(12), 663–672. https://doi.org/10.2307/2678462

Efron, B. (2012). *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction* (Vol. 1). Cambridge University Press.

Epstein, L. G., & Schneider, M. (2003). Recursive multiple priors. *The Journal of Economic Theory*, *113*(1), 1–31. https://doi.org/10.1016/S0022-0531(03)00097-8

Evans, M. (2013). What does the proof of Birnbaum's theorem prove? *Electronic Journal of Statistics*, *7*, 2645–2655. https://doi.org/10.1214/13-EJS857

Evans, M., & Jang, G. H. (2010). Invariant p-values for model checking. *The Annals of Statistics*, *38*(1), 512–525. https://doi.org/10.1214/09-AOS727

Evans, M. J., Fraser, D. A., & Monette, G. (1986). On principles and arguments to likelihood. *Canadian Journal of Statistics*, *14*(3), 181–194. https://doi.org/10.2307/3314794

Fisher, R. A. (1925). *Statistical methods for research Workers*. Oliver & Boyd (Edinburgh).

Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd Ltd.

Fitelson, B. (1999). The plurality of Bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science*, *66*(S3), S362–S378. https://doi.org/10.1086/392738

Forster, M., & Sober, E. (2004). Why likelihood? In *The nature of scientific evidence: Statistical, philosophical, and empirical considerations* (pp. 153–190).

Forster, M. R. (2006). Counterexamples to a likelihood theory of evidence. *Minds and Machines*, *16*(3), 319–338. https://doi.org/10.1007/s11023-006-9038-y

Fossaluza, V., Izbicki, R., da Silva, G. M., & Esteves, L. G. (2017). Coherent hypothesis testing. *The American Statistician*, *71*(3), 242–248. https://doi.org/10.1080/00031305.2016.1237893

Gabriel, K. R. (1969). Simultaneous test procedures–some theory of multiple comparisons. *Annals of Mathematical Statistics*, *40*(1), 224–250. https://doi.org/10.1214/aoms/1177697819

Gandenberger, G. (2015). A new proof of the likelihood principle. *British Journal for the Philosophy of Science*, *66*(3), 475–503. https://doi.org/10.1093/bjps/axt039

Ghosh, M. (2011). Objective priors: An introduction for frequentists. *Statistical Science*, *26*(2), 187–202. https://doi.org/10.1214/10-STS338

Goodman, S. N., & Royall, R. (1988). Evidence and scientific research. *American Journal of Public Health*, *78*(12), 1568–1574. https://doi.org/10.2105/AJPH.78.12.1568

Greenland, S. (2019). Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. *The American Statistician*, *73*(sup1), 106–114. https://doi.org/10.1080/00031305.2018.1529625

Grendár, M. (2012). Is the -value a good measure of evidence? Asymptotic consistency criteria. *Statistics & Probability Letters*, *82*(6), 1116–1119. https://doi.org/10.1016/j.spl.2012.02.018

Grünwald, P. (2007). *The minimum description length principle*. MIT Press.

Grünwald, P., de Heide, R., & Koolen, W. (2024). Safe testing. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *86*(5), 1091–1128. https://doi.org/10.1093/jrsssb/qkae011

Grünwald, P., & Halpern, J. (2011). Making decisions using sets of probabilities: Updating, time consistency, and calibration. *Journal of Artificial Intelligence Research (JAIR)*, *42*, 393–426.

Grünwald, P., Lardy, T., Hao, Y., Bar-Lev, S. K., & de Jong, M. (2024). Optimal e-values for exponential families: The simple case. arXiv preprint arXiv:2404.19465.

Grünwald, P., & Roos, T. (2020). Minimum description length revisited. *International Journal of Mathematics for Industry*, *11*(1). https://doi.org/10.1142/S2661335219300018

Grünwald, P. D. (2023). The e-posterior. *Philosophical Transactions of the Royal Society A*, *381*(2247), 20220146. https://doi.org/10.1098/rsta.2022.0146

Grünwald, P. D. (2024). Beyond neyman–Pearson: E-values enable hypothesis testing with a data-driven alpha. *Proceedings of the National Academy of Sciences*, *121*(39), e2302098121. https://doi.org/10.1073/pnas.2302098121

Hacking, I. (2016). *Logic of statistical inference*. Cambridge University Press.

Hao, Y., & Grünwald, P. (2024). E-values for exponential families: The general case. arXiv preprint arXiv:2409.11134.

Hartigan, J. (1967). The likelihood and invariance principles. *Journal of the Royal Statistical Society: Series B (Methodological)*, *29*(3), 533–539. https://doi.org/10.1111/j.2517-6161.1967.tb00715.x

Hubbard, R., & Lindsay, R. M. (2008). Why p values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, *18*(1), 69–88. https://doi.org/10.1177/0959354307086923

Jeffreys, H. (1939). *Theory of probability*. The Clarendon Press.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *186*(1007), 453–461. https://doi.org/10.1098/rspa.1946.0056

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kalbfleisch, J. D. (1975). Sufficiency and conditionality. *Biometrika*, *62*(2), 251–259. https://doi.org/10.1093/biomet/62.2.251

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, *91*(435), 1343–1370. https://doi.org/10.1080/01621459.1996.10477003

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, *16*(5), 345–353. https://doi.org/10.1111/j.0956-7976.2005.01538.x

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

Koning, N. W. (2023). Post-hoc α hypothesis testing and the post-hoc p-value. arXiv preprint arXiv:2312.08040.

Lakens, D. (2022). Why P values are not measures of evidence. *Trends in Ecology & Evolution*, *37*(4), 289–290. https://doi.org/10.1016/j.tree.2021.12.006

Lardy, T., Grünwald, P., & Harremoës, P. (2024). Reverse information projections and optimal e-statistics. *IEEE Transactions on Information Theory.*, *70*(11), 7616–7631. https://doi.org/10.1109/TIT.2024.3444458

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

Lavine, M., & Schervish, M. J. (1999). Bayes factors: What they are and what they are not. *The American Statistician*, *53*(2), 119–122. https://doi.org/10.1080/00031305.1999.10474443

Lele, S. R. (2004). Evidence functions and the optimality of the law of likelihood. *The Nature of Scientific Evidence*, 191–216.

Li, J., Li, Y., & Dai, X. (2023). Li, li, and dai's contribution to the discussion of "estimating means of bounded random variables by betting" by waudby-smith and aaditya ramdas. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *86*(1), 41–43. https://doi.org/10.1093/jrsssb/qkad111

Lindley, D. V., & Phillips, L. D. (1976). Inference for a bernoulli process (a Bayesian view). *The American Statistician*, *30*(3), 112–119. https://doi.org/10.1080/00031305.1976.10479154

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press.

Mayo, D. G. (2014). On the Birnbaum argument for the strong likelihood principle. *Statistical Science*, *29*(2), 227–239. https://doi.org/10.1214/13-STS457

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press.

Mayo, D. G., & Spanos, A. (2011). Error statistics. In *Philosophy of statistics* (pp. 153–198). Elsevier.

McCullagh, P., & Cox, D. (1986). Invariants and likelihood ratio statistics. *The Annals of Statistics*, *14*(4), 1419–1430. https://doi.org/10.1214/aos/1176350167

Moreno, E., Girón, F. J., & Casella, G. (2010). Consistency of objective Bayes factors as the model dimension grows. *The Annals of Statistics*, *38*(4), 1937. https://doi.org/10.1214/09-AOS754

Morey, R. D., Romeijn, J. W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*(6–18), 36.

Muff, S., Nilsen, E. B., O'Hara, R. B., & Nater, C. R. (2022). Rewriting results sections in the language of evidence. *Trends in Ecology & Evolution*, *37*(3), 203–210. https://doi.org/10.1016/j.tree.2021.10.009

Neyman, J. (1976). Tests of statistical hypotheses and their use in studies of natural phenomena. *Communications in Statistics: Theory and Methods*, *5*(8), 737–751. https://doi.org/10.1080/03610927608827392

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, *20*(1/2), 175–240.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh & Dublin Philosophical Magazine & Journal of Science*, *50*(302), 157–175. https://doi.org/10.1080/14786440009463897

Pérez-Ortiz, M. F., Lardy, T., de Heide, R., & Grünwald, P. D. (2024). E-statistics, group invariance and anytime-valid testing. *The Annals of Statistics*, *52*(4), 1410–1432. https://doi.org/10.1214/24-AOS2394

Popper, K. R. (1963). *Conjectures and refutations: The growth of scientific knowledge*. Routledge.

Ramdas, A., Grünwald, P., Vovk, V., & Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, *38*(4), 576–601. https://doi.org/10.1214/23-STS894

Ramdas, A., & Manole, T. (2026). Randomized and exchangeable improvements of markov's, chebyshev's and chernoff's inequalities. *Statistical Science*, *41*(1), 121–142. https://doi.org/10.1214/24-STS952

Ramdas, A., Ruf, J., Larsson, M., & Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. arXiv preprint arXiv:2009.03167.

Ramdas, A., Ruf, J., Larsson, M., & Koolen, W. M. (2022). Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, *141*, 83–109. https://doi.org/10.1016/j.ijar.2021.06.017

Ramdas, A., & Wang, R. (2025). Hypothesis testing with e-values. *Foundations and Trends® in Statistics*, *1*(1–2), 1–390. https://doi.org/10.1561/3600000002

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *Annals of Mathematical Statistics*, *41*(5), 1397–1409. https://doi.org/10.1214/aoms/1177696786

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Royall, R. (1997). *Statistical evidence: A likelihood paradigm* (Vol. 71). CRC Press.

Royall, R. M. (1986). The effect of sample size on the meaning of significance tests. *The American Statistician*, *40*(4), 313–315. https://doi.org/10.1080/00031305.1986.10475424

Ryabko, B. Y., & Monarev, V. (2005). Using information theory approach to randomness testing. *Journal of Statistical Planning and Inference*, *133*(1), 95–110. https://doi.org/10.1016/j.jspi.2004.02.010

Savage, L. J., Barnard, G., Cornfield, J., Bross, I., Good, I., Lindley, D., Clunies-Ross, C., Pratt, J. W., Levene, H., & Goldman, T., et al (1962). On the foundations of statistical inference: Discussion. *Journal of the American Statistical Association*, *57*(298), 307–326.

Schervish, M. J. (1996). P values: What they are and what they are not. *The American Statistician*, *50*(3), 203–206. https://doi.org/10.1080/00031305.1996.10474380
Shafer, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *184*(2), 407–431. https://doi.org/10.1111/rssa.12647
Shafer, G., & Vovk, V. (2005). *Probability and finance: It's only a game!* (Vol. 491). John Wiley & Sons.
Shafer, G., & Vovk, V. (2019). *Game-theoretic foundations for probability and finance*. John Wiley & Sons.
Shekhar, S., & Ramdas, A. (2023). Nonparametric two-sample testing by betting. *IEEE Transactions on Information Theory*, *70*(2), 1178–1203. https://doi.org/10.1109/TIT.2023.3305867
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632
Stark, P. B. (2020). Sets of half-average nulls generate risk-limiting audits: SHANGRLA. In *International conference on financial cryptography and data security* (pp. 319–336). Springer.
Takatsu, K. (2025). On the precise asymptotics of universal inference. arXiv preprint arXiv:2503.14717.
Taper, M. L., & Lele, S. R. (2010). *The nature of scientific evidence: Statistical, philosophical, and empirical considerations*. University of Chicago Press.
Taper, M. L., & Lele, S. R. (2011). Evidence, evidence functions, and error probabilities. In *Philosophy of statistics* (pp. 513–532). Elsevier.
Taper, M. L., & Ponciano, J. M. (2016). Evidential statistics as a statistical modern synthesis to support 21st century science. *Population Ecology*, *58*(1), 9–29. https://doi.org/10.1007/s10144-015-0533-y
Ter Schure, J., & Grünwald, P. (2022). ALL-IN meta-analysis: Breathing life into living systematic reviews. *F1000research*, *11*(549), 549. https://doi.org/10.12688/f1000research.74223.1
Ville, J. (1939). Étude critique de la notion de collectif. *Bulletin of the American Mathematical Society*, *45*(11), 824.
Vovk, V. (2025). Conformal e-prediction. *Pattern Recognition*, *166*, 111674. https://doi.org/10.1016/j.patcog.2025.111674
Vovk, V., Wang, B., & Wang, R. (2022). Admissible ways of merging p-values under arbitrary dependence. *The Annals of Statistics*, *50*(1), 351–375. https://doi.org/10.1214/21-AOS2109
Vovk, V., & Wang, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics*, *49*(3), 1736–1754. https://doi.org/10.1214/20-AOS2020
Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105
Wang, H., & Ramdas, A. (2025). Anytime-valid t-tests and confidence sequences for Gaussian means with unknown variance. *Sequential Analysis*, *44*(1), 56–110. https://doi.org/10.1080/07474946.2024.2428245
Wang, R. (2025). The only admissible way of merging arbitrary e-values. *Biometrika*, *112*(2), asaf020. https://doi.org/10.1093/biomet/asaf020
Wang, R., & Ramdas, A. (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *84*(3), 822–852. https://doi.org/10.1111/rssb.12489
Wasserman, L., Ramdas, A., & Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, *117*(29), 16880–16890. https://doi.org/10.1073/pnas.1922664117
Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108
Waudby-Smith, I., & Ramdas, A. (2024). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *86*(1), 1–27. https://doi.org/10.1093/jrsssb/qkad009
Waudby-Smith, I., Sandoval, R., & Jordan, M. I. (2025). Universal log-optimality for general classes of e-processes and sequential hypothesis tests. arXiv preprint arXiv:2504.02818.
Wright, S. P. (1992). Adjusted p-values for simultaneous inference. *Biometrics*, *48*(4), 1005–1013. https://doi.org/10.2307/2532694
Xie, M. G., & Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, *81*(1), 3–39. https://doi.org/10.1111/insr.12000