E-value, a better alternative for testing?

Mingzhe Liu

April 2026

1 Introduction

Since its introduction by *Sir Ronald Aylmer Fisher*, the p-value has become one of the central tools in statistical inference, particularly in hypothesis testing and confidence estimation. In modern statistical practice, software such as `R` or `Julia` routinely reports p-values for tasks such as regression model assessment and parameter significance testing.

Despite their widespread use, p-values have several well-documented limitations as measures of statistical evidence. First, commonly used thresholds such as 0.05 are largely conventional and lack a principled justification. Second, a p-value greater than 0.05 does *not* constitute evidence in favor of the null hypothesis; it merely indicates insufficient evidence against it.[1] These issues have led to increasing criticism of p-values as tools for quantifying evidence.

To further illustrate their limitations, consider the classical normal location model:

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0,$$

where observations $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ with known variance σ^2. The usual test statistic is

$$T_n = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sigma}.$$

Under H_0, $T_n \sim \mathcal{N}(0, 1)$ for all n, and hence the corresponding p-value is uniformly distributed on $[0, 1]$. Under the alternative hypothesis H_1, however, T_n diverges to $\pm\infty$ as $n \to \infty$, implying that the p-value converges to 0 in probability. Thus, p-values are *consistent* for detecting deviations from the null.

Nevertheless, this consistency does not resolve the fundamental interpretational issues: p-values do not provide a direct, quantitative measure of evidence in favor of either hypothesis, nor do they behave well under sequential analysis. These shortcomings motivate the development of alternative frameworks, such as *e-values*, which quantify evidence in terms of expectations rather than tail probabilities [3, 2].

The classical foundation of hypothesis testing is often traced to *Cournot's principle*, which informally states that events with very small probability under the null hypothesis are not expected to occur.[3] In the traditional p-value

framework, this idea is implemented by rejecting the null when the p-value is sufficiently small, indicating that the observed data would be unlikely if the null were true. E-values adopt a closely related philosophical standpoint, but quantify evidence in a different way. Instead of focusing on tail probabilities, e-values are based on nonnegative random variables with expectation at most one under the null hypothesis, and large observed values suggest that the null hypothesis is unlikely.

Formally, an *e-variable* is a nonnegative random variable $E : \Omega \to [0, \infty]$ such that

$$\mathbb{E}_P[E] \leq 1 \quad \text{for all } P \in H_0.$$

Any realized value of E is called an *e-value*. This definition immediately yields a testing rule via Markov's inequality: for any $\alpha \in (0, 1)$,

$$P\big(E \geq 1/\alpha\big) \leq \alpha,$$

for all $P \in H_0$. Thus, rejecting the null hypothesis when $E \geq 1/\alpha$ controls the Type-I error at level α.

An appealing interpretation of e-values is given through a gambling or betting perspective. One may think of E as the payoff of a gamble that costs one unit: under the null hypothesis, the expected payoff is at most one, so no strategy can systematically yield profit. However, observing a large value of E corresponds to an unexpectedly large gain, providing evidence against the null hypothesis. This interpretation, emphasized by [2], gives e-values a direct and intuitive meaning as accumulated evidence against H_0.

E-variables also arise naturally in classical statistical constructions. In the case of testing simple hypotheses $H_0 = \{P_0\}$ versus $H_1 = \{P_1\}$, where both distributions admit densities p_0 and p_1 with respect to a common dominating measure μ, the likelihood ratio

$$E(x) = \frac{p_1(x)}{p_0(x)}$$

is an e-variable. Indeed, under P_0, we have $\mathbb{E}_{P_0}[E] = 1$, so E satisfies the defining property of an e-variable. From a Bayesian perspective, this likelihood ratio coincides with the Bayes factor comparing H_1 to H_0, showing that Bayes factors are a special case of e-values for simple null hypotheses [2]. Later in this article, we will see that even in composite settings, e-variables can often be constructed in forms closely related to Bayes factors, sometimes involving carefully chosen priors and mixtures.

In this report, we explore both the theoretical and practical aspects of e-values as tools for statistical inference. We first study conditions under which specific constructions of e-variables yield consistent tests, providing reliable evidence both against and in favor of the null hypothesis. We then examine several

practical advantages of e-values compared to p-values, including their ability to be combined through simple averaging and their validity under optional continuation and sequential analysis. At the same time, we also discuss important limitations of the e-value framework, such as the lack of universally accepted significance thresholds and the challenges involved in constructing powerful e-variables in complex or composite testing problems. Together, these perspectives provide a balanced understanding of the strengths and limitations of e-values in modern statistical practice.

2 Consistent e-Variables

Suppose we observe a sequence of independent e-variables $E_1, \ldots, E_n$ with respect to a null hypothesis H_0, all generated from the same data-generating process. A fundamental property of e-values is that they can be combined multiplicatively: the product

$$E_n^* := \prod_{i=1}^{n} E_i$$

is itself an e-variable with respect to H_0. More generally, the running product

$$X_m := \prod_{i=1}^{m} E_i, \quad m \geq 1,$$

forms a nonnegative test supermartingale under any probability measure $P \in H_0$.

This observation leads to the following result, which describes the asymptotic behavior of such products under the null hypothesis.

Proposition 2.1. *Let $E_1, E_2, \ldots$ be i.i.d. non-degenerate e-variables with respect to a (possibly composite) null hypothesis H_0. Then, for any $P \in H_0$, the running product*

$$X_m = \prod_{i=1}^{m} E_i$$

converges to 0 almost surely.

Proof. For any $P \in H_0$, the process $(X_m)_{m \geq 1}$ is a nonnegative supermartingale with $\mathbb{E}_P[X_m] \leq 1$ for all m. By the martingale convergence theorem, $X_m \to X_\infty$ almost surely for some finite random variable X_∞.

We claim that $X_\infty = 0$ almost surely. Since Ei is non-degenerated, for sufficiently small $\lambda > 0$, we have

$$P(|E_n - 1| > \lambda) > 0$$

, and since X_n converges a.s., $|X_{n+1} - X_n| = |E_{n+1} - 1||X_n|$ converges to 0 in probability, thus $P(|X_n| > \epsilon)$ converges to 0 as n goes large, for all sufficiently small ϵ, this shows that X_n converges to 0 almost surely. □

This proposition shows that, under the null hypothesis, the cumulative evidence measured by the product of i.i.d. e-values vanishes almost surely. In other words, repeated application of valid e-tests does not spuriously accumulate evidence against H_0.

The converse behavior—namely, that X_m grows large under alternative distributions—requires additional structure and is more delicate to establish. In the remainder of this section, we investigate conditions under which e-variables can be constructed to yield consistent tests, meaning that the accumulated evidence diverges under H_1 while vanishing under H_0.

2.1 Simple Alternative: The GRO e-Variables

In this and the following sections, we introduce a general framework and notation for hypothesis testing with e-variables. Let $\{P_\theta : \theta \in \Theta\}$ be a statistical model on a measurable space $(\Omega, \mathcal{F})$, and suppose that all P_θ are dominated by a common σ-finite measure μ, with corresponding densities $p_\theta = \frac{dP_\theta}{d\mu}$. The null and alternative hypotheses are specified as subsets $H_0, H_1 \subseteq \Theta$.

We assume that Θ, H_0, and H_1 are equipped with a σ-algebra $\mathcal{A}$ such that all singletons are measurable. For any subset $\Theta_a \subseteq \Theta$, we denote by $\mathcal{W}(\Theta_a)$ the set of all probability measures (priors) on $(\Theta_a, \mathcal{A})$.

Given a prior $W \in \mathcal{W}(\Theta_a)$, we define the corresponding mixture distribution

$$p_W(x) := \int_{\Theta_a} p_\theta(x)\, dW(\theta),$$

which will play a central role in constructing e-variables.

Equivalence and Essential Uniqueness. Two e-variables E and E' are said to be *equivalent* (or essentially equal) under $P \in H_0$ if $P(E = E') = 1$. More generally, we say that a construction of e-variables is *essentially unique* if any two valid constructions agree almost surely under all $P \in H_0$. This notion reflects the fact that e-variables are only defined up to null sets under the null hypothesis.

Kullback–Leibler Divergence. For two probability measures P and Q dominated by μ, the Kullback–Leibler (KL) divergence is defined as

$$D(P \,\|\, Q) := \int \log\left(\frac{dP}{dQ}\right) dP,$$

provided the Radon–Nikodym derivative exists. In the parametric setting, this becomes

$$D(P_\theta \,\|\, P_{\theta'}) = \int p_\theta(x) \log\left(\frac{p_\theta(x)}{p_{\theta'}(x)}\right) d\mu(x).$$

We introduce the existence of the GRO(*Growth Rate Optimal*) e-variable.

Theorem 2.2 (GRO e-variable). Suppose $H_0, H_1 = \{\theta_a\} \subseteq \Theta$ and $\inf_{W_0 \in \mathcal{W}(H_0)}(D(P_{\theta_a} \,\|\, P_{W_0})) < \infty$. Then there exists a prior $W_0 \in \mathcal{W}(H_0)$ such

that the likelihood ratio

$$E^*(x) := \frac{p_{\theta_a}(x)}{p_{W_0}(x)}$$

is an e-variable with respect to H_0, i.e.,

$$\mathbb{E}_{P_\theta}[E^*] \leq 1 \quad \text{for all } \theta \in H_0.$$

Moreover, among all e-variables of this form, E is *growth-rate optimal*: it maximizes the expected logarithmic growth rate

$$\mathbb{E}_{P_{\theta_a}}[\log E^*(x)] = \sup_{E \in \mathcal{E}(H_0)} (\mathbb{E}_{P_{\theta_a}}[\log E(x)]) = \inf_{W_0 \in \mathcal{W}(H_0)} (D(P_{\theta_a} \,\|\, P_{W_0}))$$

where $\mathcal{E}(H_0)$ denotes all e-variables with respect to H_0.

Proof: Consult [2],Appendix A.1

Notice that if the alternative hypothesis is simple and sufficiently separated from H_0 in the sense that

$$\inf_{W_0 \in \mathcal{W}(H_0)} D(P_{\theta_a} \,\|\, P_{W_0}) > 0,$$

then the optimal e-variable E^* satisfies

$$\mathbb{E}_{P_{\theta_a}}[\log E^*(X)] > 0.$$

Now suppose we observe i.i.d. samples and construct independent copies $E_1^*, E_2^*, \ldots$ of E^*. By the strong law of large numbers,

$$\frac{1}{n} \sum_{i=1}^{n} \log E_i^* \xrightarrow{a.s.} \mathbb{E}_{P_{\theta_a}}[\log E^*(X)] > 0 \quad \text{under } P_{\theta_a}.$$

apply continuous mapping theorem we get

$$X_n := \prod_{i=1}^{n} E_i^* \xrightarrow{a.s.} \infty,$$

showing that the accumulated evidence diverges under the alternative hypothesis. Hence, the test based on the running product is consistent.

The growth-rate optimal (GRO) property means that the e-variable E^* maximizes the expected logarithmic growth of evidence against H_0 under the alternative. In particular, $\mathbb{E}_{P_{\theta_a}}[\log E(X)]$ can be interpreted as the average rate at which evidence accumulates per observation. From a gambling perspective, this corresponds to maximizing the long-run exponential growth rate of capital when repeatedly betting against H_0.

As emphasized in [2], this notion is analogous to power in classical hypothesis testing, but adapted to sequential and optional continuation settings. Rather than maximizing rejection probability at a fixed sample size, GRO e-variables maximize the speed at which evidence accumulates over time. This makes them particularly well-suited for settings where data arrive sequentially and decisions may depend on past observations.

2.2 Beyond Simple Alternative: GROW and reGROW

We see that if the alternative is simple (a singleton set), then we can find an e-variable that achieves the optimal growth rate. We now extend this idea to composite alternatives H_1 in a truly frequentist fashion.

For composite H_1, we define the *GROW* (growth-rate optimal in the worst case) of the null hypothesis H_0 as

$$\sup_{E \in \mathcal{E}(H_0)} \inf_{\theta_1 \in H_1} \mathbb{E}_{\theta_1}[\log E],$$

where $\mathcal{E}(H_0)$ denotes the class of all e-variables valid for H_0. The motivation is that if we can ensure positive logarithm growth in the worst case over all $\theta_1 \in H_1$, then, *a fortiori*, we obtain positive log-growth for every individual alternative. This extends the simple-alternative result over the entire composite alternative hypothesis. In this sense, GROW serves as a natural analogue of power in the composite setting, capturing the worst-case evidence accumulation rate against H_0.

Theorem 2.3 GROW e-Variable: Let $\Theta_0, \Theta_1 \subseteq \Theta$ and $H_0 = \{P_\theta : \theta \in \Theta_0\}$, $H_1 = \{P_\theta : \theta \in \Theta_1\}$ be null and alternative hypotheses, and all the densities p_θ has full support. Further suppose that $\inf_{W_1 \in \mathcal{W}_\infty, W_0 \in \mathcal{W}} D(P_{W_1} \,\|\, P_{W_0}) = \inf_{W_0 \in \mathcal{W}} D\big(P_{W_1^*} \,\|\, P_{W_0}\big) < \infty$ for some prior W_1^* on Θ_1, and for any $\theta_1, \theta_2 \in \Theta_1$, $D(P_{\theta_1} \,\|\, P_{\theta_2}) < \infty$.

Then there exists an e-variable of the form

$$E(x) = \frac{p_{W_1^*}(x)}{p^*(x)},$$

where $p_{W_1^*}$ is the density of $P_{W_1^*}$ and p^* is the density of P^* which achieves $\inf_{W_0 \in \mathcal{W}} D\big(P_{W_1^*} \,\|\, P_{W_0}\big)$, such that E is GROW-optimal, i.e.,

$$\inf_{\theta_1 \in \Theta_1} \mathbb{E}_{\theta_1}[\log E] = \sup_{E' \in \mathcal{E}(H_0)} \inf_{\theta_1 \in \Theta_1} \mathbb{E}_{\theta_1}[\log E'] = D(P_{W_1^*} \,\|\, P^*).$$

Proof: Consult [2],Appendix A.1

If we can ensure that Θ_0 and Θ_1 are *uniformly distinguishable*, i.e.,

$$\inf_{W_1 \in \mathcal{W}(\Theta_1),\, W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1} \,\|\, P_{W_0}) > 0,$$

then, similarly to the simple alternative case, the optimal e-process

$$X_n = \prod_{m=1}^{n} E_m^*$$

exhibits strictly positive worst-case log-growth. In particular, this implies that $X_n \to +\infty$ almost surely under any $P_\theta \in H_1$, so that evidence against H_0 accumulates indefinitely.

However, this uniform distinguishability condition is rather strong and fails in many classical testing problems. A canonical example is the scale-normal model with known mean and unknown variance.

Example (Scale Normal: known mean, unknown variance). Consider the Gaussian scale model with known mean 0 and unknown variance:

$$P_\sigma = \mathcal{N}(0, \sigma^2), \quad \sigma > 0.$$

Let $H_0 = \{P_{\sigma_0}\}$ be a simple null and $H_1 = \{P_\sigma : \sigma \neq \sigma_0\}$ be the composite alternative.

A direct calculation shows that the Kullback–Leibler divergence between two such distributions is

$$D(P_{\sigma_1} \,\|\, P_{\sigma_0}) = \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_0^2} - 1 - \log\frac{\sigma_1^2}{\sigma_0^2}\right).$$

Now take $\sigma_1 \to \sigma_0$. Writing $r := \sigma_1^2/\sigma_0^2$, we have $r \to 1$, and by Taylor's expansion we have

$$D(P_{\sigma_1} \,\|\, P_{\sigma_0}) \;\to\; 0 \quad \text{as } \sigma_1 \to \sigma_0.$$

Thus, for any $\varepsilon > 0$, we can find σ_1 arbitrarily close to σ_0 such that

$$D(P_{\sigma_1} \,\|\, P_{\sigma_0}) < \varepsilon.$$

It follows that

$$\inf_{\theta_1 \in H_1} D(P_{\theta_1} \,\|\, P_{\theta_0}) = 0,$$

and hence H_0 and H_1 are not uniformly distinguishable.

If the parameter space exhibits a *minimum effect*, the uniform distinguishability issue can be avoided. For instance, suppose $\Theta_0 = (0, \lambda_-]$ and $\Theta_1 = [\lambda_+, \infty)$ with $\lambda_- < \lambda_+$, so that the null and alternative are separated by a fixed gap. In this case, the KL divergence between any $P_{\theta_1} \in H_1$ and $P_{\theta_0} \in H_0$ is bounded away from zero, implying uniform distinguishability. [2]

However, such minimum effect assumptions are often unrealistic in practice. A natural attempt to address this limitation is to introduce an offset in the growth criterion, leading to the notion of *reGROW* e-variables.

Theorem 2.4 reGROW e-Variable: Let f be a bounded function on Θ_1, we abbreviate $\mathbb{E}_W(f(\theta))$ as $f(W)$. If for any $\theta_1, \theta_2 \in \Theta_1$, $D(P_{\theta_1} \,\|\, P_{\theta_2}) < \infty$ and $p_{\theta_0}, p_{\theta_1}$ have full support. Furthermore, $\inf_{W_1 \in \mathcal{W}_\infty, W_0 \in \mathcal{W},} D(P_{W_1} \,\|\, P_{W_0}) - f(W_1)$ $= \inf_{W_0 \in \mathcal{W},} D\big(P_{W_1^*} \,\|\, P_{W_0}\big) - f(W_1^*) < \infty$ for some prior W_1^* on Θ_1, then there exist an e-variable E^f given by:

$$E^f(x) = \frac{p_{W_1^*}(x)}{p^*(x)},$$

for p^* is the density of P^* which achieves $\inf_{W_0 \in \mathcal{W},} D\big(P_{W_1^*} \,\|\, P_{W_0}\big) - f(W_1^*)$, such that E^* is optimal in the relative growth-rate sense, i.e.,

$$\inf_{\theta \in \Theta_1} \big(\mathbb{E}_\theta[\log E^f] - f(\theta)\big) = \sup_{E \in \mathcal{E}(\Theta_0)} \inf_{\theta \in \Theta_1} \big(\mathbb{E}_\theta[\log E] - f(\theta)\big) = \inf_{W_1 \in \mathcal{W}_\infty, W_0 \in \mathcal{W},} D(P_{W_1} \,\|\, P_{W_0}) - f(W_1)$$

over all e-variables E. *Proof*: [2],Appendix A

In summary, the GRO/GROW/reGROW e-variables provide a principled framework for constructing tests with strong asymptotic guarantees. In particular, when the corresponding growth rate is positive (or positive in the relative sense for reGROW), the associated e-process exhibits consistency: the product process diverges to infinity almost surely under the alternative, ensuring reliable long-run detection of deviations from H_0.

However, these e-variables are typically not straightforward to compute in practice. Their construction involves solving nontrivial optimization problems over distributions, such as finding the optimal prior W_1^* (and corresponding W_0^*), which may not admit closed-form solutions and can be computationally demanding.

Finally, the existence results for these optimal e-variables rely on certain technical assumptions, including measurability of the parameter space and the existence of appropriate σ-algebra structures, as well as regularity conditions ensuring that the relevant KL projections are well-defined. These assumptions, while mild in many parametric settings, may require additional care in more general or nonparametric models.

2.3 Bayesian Perspective for GRO

Consider first the simple versus simple setting, where $H_0 = \{\theta_0\}$ and $H_1 = \{\theta_1\}$. In this case, the GRO e-variable reduces to the likelihood ratio,

$$E = \frac{p_{\theta_1}}{p_{\theta_0}},$$

which coincides with the Bayes factor for testing H_0 against H_1 under point-mass priors. Thus, in the simplest setting, GRO e-variables admit a direct Bayesian interpretation.

Moreover, without appealing to large-sample properties, we observe a symmetry: the reciprocal $1/E = \frac{p_{\theta_0}}{p_{\theta_1}}$ is itself an e-variable for testing H_1 versus H_0. This highlights that e-values quantify evidence in a directional manner—large values of E provide evidence against H_0, while small values (equivalently, large values of $1/E$) provide evidence against H_1.

If we observe independent data and construct corresponding e-values $E_1, \ldots, E_n$ from the same testing procedure, then their product

$$X_n = \prod_{i=1}^{n} E_i$$

is again an e-variable. This multiplicative structure naturally aggregates evidence across samples: a large value of X_n indicates increasing evidence against H_0, whereas a very small value of X_n indicates increasing evidence against H_1.

From a Bayesian testing perspective, suppose we place a proper prior W_1 on the alternative parameter space Θ_1. This induces a marginal distribution which can be viewed as a single "aggregated" alternative. In this case, we may

bypass the explicit construction of GROW or reGROW e-variables by reducing the composite problem to a simple-versus-composite one, and applying a GRO-type construction to test H_0 against the mixture model P_{W_1}. The resulting e-variable then takes the familiar Bayes factor form.

However, this approach yields only a *quasi-Bayesian* interpretation. While W_1 can be chosen freely as a prior on Θ_1, the validity of the resulting e-variable requires a corresponding choice of W_0^* on Θ_0, typically determined through an optimization (such as a KL projection) rather than subjective prior specification. Thus, unlike standard Bayesian testing—where priors on both hypotheses are chosen a priori—the e-variable framework retains a fundamentally frequentist constraint on the null, and the resulting construction should be viewed as a hybrid between Bayesian modeling and frequentist validity.[2]

Importantly, despite this quasi-Bayesian interpretation, the resulting e-variables retain their key advantage: they remain valid under optional stopping and allow for safe, sequential accumulation of evidence against the null.

3 Compare P-value and E-value

In this final section, we compare p-values and e-values from a practical perspective. While both serve as measures of evidence against a null hypothesis, they differ substantially in their interpretation, behavior under repeated testing, and ease of use. As we will see, e-values offer several appealing properties—such as straightforward combination across studies and validity under optional stopping—that address well-known limitations of p-values. However, these advantages come at a cost: e-values are often less intuitive to interpret, lack widely accepted reporting standards, and rely on somewhat arbitrary thresholds when used for decision-making. This trade-off highlights that, despite their theoretical appeal, e-values are not yet a drop-in replacement for p-values in routine statistical practice.

3.1 E-merging

A key practical advantage of e-values is their simplicity in aggregation. In particular, e-values can be merged by taking their arithmetic mean, and the result remains a valid e-value even without any independence assumptions. That is, given e-values $E_1, \ldots, E_K$, the quantity $\frac{1}{K}\sum_{k=1}^{K} E_k$ is again an e-value. This stands in sharp contrast to p-values, for which merging is considerably more delicate and typically requires specific assumptions or specialized procedures (see, e.g.,[4]). From a practical standpoint, this means that if multiple e-tests are conducted—say, one yields strong evidence (a large e-value) while another is inconclusive—we can simply average them to obtain a combined measure of evidence that remains valid. This robustness and flexibility make e-values particularly attractive in settings involving multiple analyses or data sources.

A further advantage is that independence is not required to justify sequential multiplication of e-values. In particular, to ensure that the running product

$$X_n = \prod_{m=1}^{n} E_m$$

remains a valid e-variable (indeed, a test supermartingale), it suffices that each E_m is a *conditional* e-variable with respect to the past, i.e.,

$$\mathbb{E}_P[E_m \mid \mathcal{F}_{m-1}] \leq 1 \quad \text{for all } P \in H_0.$$

This notion, introduced in [**grunwald2023safe**], allows each E_m to depend on previous data in an arbitrary (possibly adaptive) manner, as long as the above conditional expectation constraint is satisfied. As a result, the product process $(X_n)_n$ forms a test supermartingale and retains its validity under optional stopping. This flexibility is in stark contrast with classical p-value methods, which typically require independence or carefully pre-specified sampling plans.

3.2 Optional Continuation

A central advantage of e-values is their validity under optional continuation. In particular, any e-process $(E_n)_n$—that is, a nonnegative process satisfying the conditional e-variable property—forms a test supermartingale under H_0. As a consequence, for any (possibly data-dependent) stopping time τ, the stopped value E_τ remains a valid e-variable. This implies that we may freely decide, based on the observed data, whether to continue collecting evidence or to stop, without invalidating Type-I error guarantees. In contrast to classical p-value methods, which typically require fixed sample sizes or carefully pre-specified stopping rules, e-values allow fully adaptive and sequential testing procedures. This property, formalized via the optional stopping theorem and Ville's inequality, provides substantial flexibility in practice and is one of the key motivations for e-value based testing [2].

3.3 Drawback of E-values

Despite these advantages, e-values also come with notable drawbacks. First, their use in decision-making still relies on thresholds (e.g. rejecting when $E \geq 1/\alpha$), and these cutoffs are ultimately just as arbitrary as the significance levels used for p-values. While interpretations such as Jeffreys' scale provide some guidance, there is no universally accepted standard in practice.

Second, e-values are generally more difficult to interpret than p-values. Whereas p-values have a familiar (albeit often misunderstood) probabilistic meaning tied to tail probabilities, e-values quantify evidence through expectations or betting interpretations, which can be less intuitive for practitioners. As a result, communicating and calibrating the strength of evidence using e-values remains a challenge, particularly outside of specialized settings.

In summary, e-values offer a compelling alternative to p-values, particularly in settings involving sequential testing, multiple analyses, or adaptive data collection. Their ability to combine evidence transparently and remain valid under optional stopping addresses several well-known limitations of classical p-value based methods. However, these advantages come with practical challenges, including less intuitive interpretation and the lack of standardized thresholds for decision-making. As such, e-values should be viewed not as a wholesale replacement for p-values, but as a complementary tool—especially valuable in modern, data-adaptive settings where flexibility and robustness are essential.

4 Conclusion

In this paper, we explored the theory and applications of e-values as an alternative framework for statistical hypothesis testing. Starting from their definition and connection to likelihood ratios, we examined their key properties, including their interpretation as measures of evidence, their behavior under sequential testing, and their flexibility in combining information across experiments. We also studied optimal constructions such as GRO, GROW, and reGROW e-variables, highlighting both their theoretical guarantees and practical limitations.

From a practical standpoint, e-values offer several important advantages over p-values, including straightforward merging, validity under optional stopping, and a natural interpretation in terms of evidence accumulation. At the same time, they come with challenges, such as computational complexity, lack of standardized thresholds, and less intuitive interpretation.

Overall, despite these drawbacks, e-values provide a more coherent and principled measure of statistical evidence than p-values, particularly in modern settings involving sequential and adaptive data analysis. They should therefore be viewed as a promising direction for developing more robust and flexible statistical methods.

References

[1] Michael Evans. *Theory of Statistical Inference: Lecture III.7.* `https://utstat.utoronto.ca/mikevans/sta422/lectureIII.7.pdf`. STA422/STA2162, University of Toronto lecture notes. 2026.

[2] Peter Grünwald, Rianne de Heide, and Wouter Koolen. "Safe Testing". In: *arXiv preprint arXiv:1906.07801* (2023).

[3] Vladimir Vovk and Ruodu Wang. "Confidence and discoveries with e-values". In: *arXiv preprint arXiv:1912.13292* (2022).

[4] Vladimir Vovk and Ruodu Wang. "E-values: Calibration, combination, and applications". In: *arXiv preprint arXiv:1912.06116* (2020).

5 AI Contribution Statement

Portions of this report were developed with the assistance of an AI language model (ChatGPT, OpenAI). The AI was used to help improve clarity of exposition, refine mathematical writing, and suggest structural organization of the material. It also assisted in restating known results, drafting explanations, and polishing LaTeX code.

All mathematical content, interpretations, and final edits were verified and approved by the author. Any errors or omissions remain the responsibility of the author.