Biometrika (2024), **0**, 0, *pp.* 1–22
https://doi.org/10.1093/biomet/asae049
Advance Access publication 30 September 2024

Testing the mean and variance by *e*-processes

By YIXUAN FAN

Centre for Applied Mathematics, Tianjin University, 92 Weijin Road, Tianjin 300072, China
yxfanjlu@gmail.com

ZHANYI JIAO

Department of Mathematics, Illinois State University,
300 S. School Street, Normal, Illinois 61761, U.S.A.
zjiao1@ilstu.edu

and RUODU WANG

Department of Statistics and Actuarial Science, University of Waterloo,
200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada
wang@uwaterloo.ca

Summary

We address the problem of testing the conditional mean and conditional variance for nonstationary data. We build *e*-values and *p*-values for four types of nonparametric composite hypothesis with specified mean and variance as well as other conditions on the shape of the data-generating distribution. These shape conditions include symmetry, unimodality and their combination. Using the obtained *e*-values and *p*-values, we construct tests via *e*-processes, also known as testing by betting, as well as some tests based on combining *p*-values for comparison. Although we mainly focus on one-sided tests, the two-sided test for the mean is also studied. Simulation and empirical studies are conducted under a few settings, and they illustrate features of the methods based on *e*-processes.

Some key words: *E*-process; *E*-value; *P*-value; Symmetry; Unimodality.

1. Introduction

Testing the mean and variance in various settings is a classic problem in statistics. In parametric inference concerning testing the mean, well-known tests like the Student's *t*-test and *z*-test, as well as tests related to variance such as the chi-squared test and the *F*-test, are commonly employed; see, e.g., Lehmann et al. (1986). Parametric tests always come with assumptions about the forms of the population distribution from which samples are derived. Deviating from these assumptions can lead to significantly flawed results. For situations where these assumptions might be compromised, nonparametric methods provide a great alternative. Certainly, nonparametric methods may also make strong assumptions on the underlying population, such as finite or bounded moments, but not on the specific parametric forms. Comprehensive and well-established methods of nonparametric techniques for testing means and variances can be found in Conover (1999) and Hollander et al. (2013). Different from the classic settings, we consider the problem of testing composite hypotheses in which data are not stationary.

Suppose that a tester has sequentially arriving, possibly dependent, data points $X_1, X_2, \ldots$, each from an unknown distribution, possibly different. The tester is interested in testing whether

$$\mathbb{E}(X_i \mid \mathcal{F}_{i-1}) \leqslant \mu_i \quad \text{and} \quad \text{var}(X_i \mid \mathcal{F}_{i-1}) \leqslant \sigma_i^2 \quad \text{for each } i, \tag{1}$$

where $\mathcal{F}_{i-1}$ is the σ-algebra generated by $X_1, \ldots, X_{i-1}$, and μ_i and σ_i are $\mathcal{F}_{i-1}$ measurable. All conditional expectations are in the almost-sure sense. If independence is further assumed then this problem reduces to the classic problem of testing the mean and variance. Testing the conditional mean and conditional variance is common in some contexts such as forecasting (e.g., Henzi & Ziegel, 2022) and financial risk assessment (e.g., Fissler & Ziegel, 2016).

Problem (1) can be interpreted in two different ways, omitting 'conditional' here:

(A) testing both the mean and the variance,
(B) testing the mean under knowledge of an upper bound on the variance.

Interpretation (A) is relevant when the tester is interested in whether a time series has switched away from a given regime with specified mean and variance bounds. We mainly use interpretation (A), while keeping in mind that interpretation (B) is useful when comparing with the literature. Of course, one could also interpret (1) as testing the variance under knowledge of an upper bound on the mean.

Clearly, problem (1) is a composition of many complicated, nonparametric, composite hypotheses on each observation. The key challenge in this setting is that the data points are not independent and identically distributed, and hence we cannot make inference of the distributions themselves.

This problem can be addressed with the following general methodology, called e-testing or testing by betting, a successful example being Waudby-Smith & Ramdas (2024). We first consider a simpler problem: constructing an e-value from one random variable from each data point with the corresponding hypothesis on its mean and variance, which corresponds to $n = 1$. For a general background on e-values in hypothesis testing, see Vovk & Wang (2021), the review by Ramdas et al. (2023) and Grünwald et al. (2024). After obtaining these e-values, we combine them, usually by forming an e-process, to construct a test for the overall hypothesis. Alternatively, we can construct p-values instead of e-values, but the power of such a strategy is usually quite weak, as seen from our experiments.

We formally describe the hypotheses and define e-variables, e-processes and p-variables. As mentioned above, we first address the case of one data point, i.e., $n = 1$. We consider four types of composite hypotheses on the mean, variance and the shape of the distribution: symmetry, unimodality and their combination. Our main results are ways that are optimal, in a natural sense, to constructions of p-values and e-values in this setting. Although our main methodology is based on e-processes, we also present results for p-values, which may be useful in multiple testing, not treated in this paper; for instance, p-values are the inputs of the standard procedure of Benjamini & Hochberg (1995). Considering a nonparametric composite hypothesis with a given mean and variance as the baseline case, assuming symmetry approximately improves the baseline p-variable by a multiplicative factor of 1/2, unimodality by a factor of 4/9 and both by a factor of 2/9. Similarly, the corresponding baseline e-variable is improved by multiplicative factors of 2, 1 and 2, respectively, in these scenarios; recall that smaller p-values are more useful, whereas larger e-values are more useful.

We propose several methods to test using multiple data points, thus addressing the main task of the tester. The main proposals are e-process-based tests, which follow the idea of testing by betting of Wasserman et al. (2020), Shafer (2021) and Waudby-Smith & Ramdas (2024). Although we mainly focus on one-sided hypotheses, our methodology can be easily adapted to test the two-sided hypothesis on the mean, that is,

$$\mathbb{E}(X_i \mid \mathcal{F}_{i-1}) \in [\mu_i^L, \mu_i^U] \quad \text{and} \quad \operatorname{var}(X_i \mid \mathcal{F}_{i-1}) \leqslant \sigma_i^2 \quad \text{for each } i,$$

where $[\mu_i^L, \mu_i^U]$ is an interval or a singleton for each i; this is discussed in § 4.3.

The closest methodological work related to this paper is that of Waudby-Smith & Ramdas (2024), where the authors tested in a nonparametric setting the conditional mean of sequential data, which are assumed to be bounded within a prespecified range, and thus a generally smaller class of distributions. Our problem and methodology are different from those of Waudby-Smith & Ramdas (2024) in the sense that we assume a bounded variance instead of a bounded range. Since a bounded range implies bounded variance, the assumption needed to apply our methodology is weaker than in the setting of Waudby-Smith & Ramdas (2024), following interpretation (B) of the main testing problem. Moreover, we are able to utilize the additional information on the distributional shape to obtain better e-values than without such information. A great advantage of the tests of Waudby-Smith & Ramdas (2024) is that their power adapts to the unknown true variance of the distribution if data come from an independent and identically distributed population. Our method based on the growth rate of empirical e-values has a similar feature, which uses a betting strategy similar to that of Waudby-Smith & Ramdas (2024). Another closely related methodology is that of Wang et al. (2024), who tested statistical functions other than the mean. Once e-variables are constructed, we build e-processes in a similar way to Wang et al. (2024). The methods of Howard et al. (2020, 2021) and Wang & Ramdas (2023) based on exponential test supermartingales, exponential processes that form supermartingales with initial value one, which are e-processes, can also be applied to test (1). These methods differ from ours as our e-process is obtained by combining individual e-variables.

We provide simulation studies for the proposed methods and compare them with the method of Waudby-Smith & Ramdas (2024) when the model has both bounded support and bounded variance and with methods based on the exponential test supermartingale of Howard et al. (2021) and Wang & Ramdas (2023). Empirical studies using financial asset return data during the 2007–8 financial crisis further demonstrate the effectiveness of the e-process-based methods. All proofs in the paper are provided in the Supplementary Material.

2. General setting

2.1. *Hypotheses to test*

We first describe our main testing problem. Let n be a positive integer or ∞, and denote by $[n] = \{1, \ldots, n\}$. Throughout, fix a sample space. Suppose that data points $(X_i)_{i\in[n]}$ arrive sequentially, each possibly from a different distribution, and not necessarily independent. A hypothesis is a collection H of probability measures that govern $(X_i)_{i\in[n]}$. Denote by $\mathcal{F}_i$

the σ-field generated by $X_1, \ldots, X_i$ for $i \in [n]$ with $\mathcal{F}_0$ being the trivial σ-field. The main hypotheses of interest are variations, by adding shape information, of the hypothesis

$$H = \{Q\colon \mathbb{E}^Q(X_i \mid \mathcal{F}_{i-1}) \leqslant \mu_i \text{ and } \mathrm{var}^Q(X_i \mid \mathcal{F}_{i-1}) \leqslant \sigma_i^2 \text{ for } i \in [n]\}, \tag{2}$$

where μ_i and σ_i are $\mathcal{F}_{i-1}$ measurable for each $i \in [n]$; that is, they can be data-dependent on past observations. A simple case is

$$H = \{Q\colon \mathbb{E}^Q(X_i \mid \mathcal{F}_{i-1}) \leqslant \mu \text{ and } \mathrm{var}^Q(X_i \mid \mathcal{F}_{i-1}) \leqslant \sigma^2 \text{ for } i \in [n]\}, \tag{3}$$

where μ and σ are two constants; that is, we would like to test whether data exhibit conditional mean and conditional variance in $(-\infty, \mu] \times [0, \sigma^2]$. Although (3) looks simpler, it is indeed equivalent to (2) by noting that μ_i and σ_i are $\mathcal{F}_{i-1}$ measurable and can be absorbed into X_i by considering $(X_i - \mu_i)/\sigma_i$ instead of X_i. Therefore, we focus on formulation (3) for the rest of the paper. If data are independent, but not necessarily identically distributed, then the problem is to test the unconditional mean and variance. We sometimes omit Q in $\mathbb{E}^Q$ and var^Q when it is clear.

We further consider hypotheses with additional shape information, by assuming that some, or all of the distributions of $X_1, \ldots, X_n$ are unimodal, symmetric or both. Below, all terms like 'increasing' and 'decreasing' are in the nonstrict sense. A distribution on $\mathbb{R}$ is unimodal if there exists $x \in \mathbb{R}$ such that the distribution has an increasing density on $(-\infty, x)$ and a decreasing density on (x, ∞); it may have a point mass at x. A distribution on $\mathbb{R}$ with mean μ is symmetric if, for all $x \in \mathbb{R}$, it assigns equal probabilities to $(-\infty, \mu - x]$ and $[\mu + x, \infty)$. If a distribution with mean μ is both unimodal and symmetric then its mode must be either μ or an interval centred at μ.

Remark 1. The main problem in Waudby-Smith & Ramdas (2024) is to test the conditional mean m with data taking values in $[0, 1]$. Any random variable with mean at most m and range $[0, 1]$ has variance at most $1/4$ if $m \geqslant 1/2$, or $m(1 - m)$ if $m < 1/2$, both of which are attained by a Bernoulli random variable. Therefore, our hypothesis with $\mu = m$ and $\sigma^2 = 1/4$ or $\sigma^2 = m(1 - m)$ has less restrictive assumptions than their setting, except they formulated two-sided hypotheses (see Remark 2 below), and, in particular, our setting can handle unbounded data.

Remark 2. Our hypotheses are formulated as one sided on both μ and σ^2. Certainly, all validity results remain true for the two-sided hypotheses. Testing $\mathbb{E}^Q(X_i) \geqslant \mu$ is symmetric to testing $\mathbb{E}^Q(X_i) \leqslant \mu$, but such symmetry does not hold for testing the variance. Building e-processes to test the two-sided hypothesis on the mean is discussed in §4.3 below.

2.2. *P-variables and e-variables*

We formally define p-variables and e-variables, following Vovk & Wang (2021). A p-variable P for a hypothesis H is a random variable that satisfies $Q(P \leqslant \alpha) \leqslant \alpha$ for all $\alpha \in (0, 1)$ and all $Q \in H$. In other words, a p-variable is stochastically larger than $\mathrm{Un}[0, 1]$, often truncated at 1. An e-variable E for a hypothesis H is a $[0, \infty]$-valued random variable satisfying $\mathbb{E}^Q(E) \leqslant 1$ for all $Q \in H$. Often e-variables are obtained from stopping an e-process $(E_t)_{t \geqslant 0}$, which is a nonnegative stochastic process adapted to a prespecified filtration, $(\mathcal{F}_i)_{i \in [n]}$ in our problem, such that $\mathbb{E}^Q(E_\tau) \leqslant 1$ for any stopping time τ and any $Q \in H$.

Some p-variables and e-variables are useless, like $P = 1$ or $E = 1$. A p-variable P for H is precise if $\sup_{Q \in H} Q(P \leqslant \alpha) = \alpha$ for each $\alpha \in (0, 1)$, and an e-variable E for H is precise if $\sup_{Q \in H} \mathbb{E}^Q(E) = 1$. In other words, a p-variable or an e-variable being precise means that it is not wasteful in a natural sense. For instance, if $\sup_{Q \in H} \mathbb{E}^Q(E) < 1$ then we can multiply E by a constant larger than 1. Some imprecise e-variables may also be useful, such as those built on the Hoeffding inequality; see Hoeffding (1963), Howard et al. (2021) and Waudby-Smith & Ramdas (2024).

A p-variable P is semiprecise for H if $\sup_{Q \in H} Q(P \leqslant \alpha) = \alpha$ for each $\alpha \in (0, 1/2]$. Semiprecise p-variables require the sharp probability bound $\sup_{Q \in H} Q(P \leqslant \alpha) = \alpha$ only for the case $\alpha \leqslant 1/2$, which is relevant for testing purposes. We will see that, for some hypotheses, precise p-variables do not exist unless we rely on external randomization, but semiprecise ones do exist.

Realizations of p-variables and e-variables are referred to as p-values and e-values. As is customary in the literature, we sometimes, but never in mathematical statements, use the two terms *e-value* and *e-variable* interchangeably.

3. Best p- and e-variables for one data point

3.1. *Setting*

We begin by considering the simple setting where one data point X is available, from which we build a p-variable or e-variable for the hypothesis. Although it may be unconventional to test based on one observation, there are several situations where this construction becomes useful.

(1) *Testing by betting.* To construct an e-process, one needs to sequentially obtain one e-value from each observation, or a batch of observations. This is the main setting in the current paper.

(2) *Testing multiple hypotheses.* One observation is obtained for each hypothesis, and p-values or e-values for each of them are computed and fed into a multiple testing procedure such as that of Benjamini & Hochberg (1995); this setting is particularly relevant for the procedure of Wang & Ramdas (2022) based on e-values, which yields false discovery rate control under arbitrary dependence. Even if, for some hypotheses, there is only one data point, a p-value or e-value, even moderate, say $e = 0.8$ or $e = 1.2$, from this hypothesis may be useful for the overall testing problem; see Ignatiadis et al. (2024), where e-values are used as weights, so $e = 0.8$ or $e = 1.2$ matters.

(3) *Testing a global null.* One may first obtain a p-value or e-value for each experiment and then combine them to test the global null, as in meta-analysis; see Vovk & Wang (2020, 2021) and the references therein.

The e-values are relevant for all three contexts, and p-values are relevant for the second and third contexts.

We focus on p-variables, which are decreasing functions of X, and e-variables, which are increasing functions of X. Thus, a larger value of X indicates stronger evidence against the null; this is intuitive because we are testing the mean less than or equal to μ in (3). This assumption on p-variables and e-variables will be made throughout the rest of the paper.

Remark 3. In the contexts of multiple testing and sequential e-values, the dependence among several e-values or p-values obtained is preserved from the dependence among the data points, if the monotonicity assumption above holds. This will be helpful when applying statistical methods based on dependence assumptions; see Benjamini & Yekutieli (2001) for the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995) with positive dependence and Chi et al. (2024) for the Benjamini–Hochberg procedure with negative dependence. Both concepts of dependence are preserved under monotone transforms.

3.2. *Two technical lemmas*

The following lemma establishes that the infimum of p-variables based on the same data point X is still a p-variable. This result relies on our assumption that p-variables are decreasing functions of X.

LEMMA 1. *For a given observation X and hypothesis H, the infimum of p-variables, which are assumed to be decreasing functions of X, is a p-variable. As a consequence, there exists the smallest p-variable.*

Although the smallest p-variable for H exists, it may not be precise. Indeed, in Theorems 2 and 4 below we will see that there may not exist any precise p-variable for some hypotheses.

The following lemma allows us to convert conditions on distribution functions into conditions on the corresponding quantile functions. For a probability measure Q, define

$$T_Y^Q(\alpha) = \inf\{x \in \mathbb{R} \colon Q(Y \leqslant x) \geqslant \alpha\} \quad \text{for } \alpha \in (0,1);$$

that is, T_Y^Q is the left-quantile function of Y under Q.

LEMMA 2. *For a random variable P and a hypothesis H,*

(i) *P is a p-variable if and only if $\inf_{Q\in H} T_P^Q(\alpha) \geqslant \alpha$ for all $\alpha \in (0,1)$,*
(ii) *P is a precise p-variable if and only if $\inf_{Q\in H} T_P^Q(\alpha) = \alpha$ for all $\alpha \in (0,1)$,*
(iii) *P is a semiprecise p-variable if and only if $\inf_{Q\in H} T_P^Q(\alpha) = \alpha$ for all $\alpha \in (0,1/2)$ and $\inf_{Q\in H} T_P^Q(\alpha) \geqslant \alpha$ for $\alpha \in [1/2,1)$.*

The proof of Lemma 2 is essentially identical to that of Lemma 1 of Vovk & Wang (2020), which gives the equivalence between probability statements and quantile statements for merging functions of p-values. Our construction for precise and semiprecise p-variables will be based on computing $\alpha \mapsto \sup_{Q\in H} T_X^Q(1-\alpha)$ and its inverse function.

3.3. *Main results*

Recall that we have only one observation, denoted X. We consider the following four classes of nonparametric composite hypotheses, where $\mu \in \mathbb{R}$ and $\sigma > 0$:

$$\begin{aligned} H(\mu,\sigma) &= \{Q \colon \mathbb{E}^Q(X) \leqslant \mu \text{ and } \mathrm{var}^Q(X) \leqslant \sigma^2\}, \\ H_{\mathrm{S}}(\mu,\sigma) &= \{Q \in H(\mu,\sigma) \colon X \text{ is symmetrically distributed}\}, \\ H_{\mathrm{U}}(\mu,\sigma) &= \{Q \in H(\mu,\sigma) \colon X \text{ is unimodally distributed}\}, \\ H_{\mathrm{US}}(\mu,\sigma) &= H_{\mathrm{U}}(\mu,\sigma) \cap H_{\mathrm{S}}(\mu,\sigma). \end{aligned}$$

For our main results on the best p-variables and e-variables, it will be clear from our proofs that the condition $\mathrm{var}^Q(X) \leqslant \sigma^2$ in each hypothesis can be replaced by $\mathrm{var}^Q(X) = \sigma^2$, and the condition $\mathbb{E}^Q(X) \leqslant \mu$ in each hypothesis can be replaced by $\mathbb{E}^Q(X) = \mu$. All results remain true with any combinations of the above alternatives. Possible improvement for the two-sided test is discussed in § 4.3 below.

The above four sets of distributions are studied in a very different context by Li et al. (2018) to compute worst-case risk measures under model uncertainty in finance. Some of our techniques for constructing p-variables use results from Li et al. (2018) and Bernard et al. (2020) for finding bounds on the quantile, which is called the value at risk in finance.

In what follows, for $x \in \mathbb{R}$, we write $x_+ = \max(x, 0)$, $x_- = \max(-x, 0)$, $x_+^2 = (x_+)^2$ and $x_-^2 = (x_-)^2$. We first consider the simplest case of testing $H(\mu, \sigma)$.

THEOREM 1. *A precise p-variable for $H(\mu, \sigma)$ is $P = \{1 + (X - \mu)_+^2/\sigma^2\}^{-1}$, and a precise e-variable for $H(\mu, \sigma)$ is $E = (X - \mu)_+^2/\sigma^2$.*

Theorem 1 can be seen as a consequence of Cantelli's inequality. It may be interesting to compare P and $1/E$ obtained from Theorem 1. Any e-variable can be converted to a p-variable via the so-called calibrator $e \mapsto \min(1/e, 1)$; see, e.g., Vovk & Wang (2021); this is an immediate consequence of Markov's inequality. As $1/E$ is a p-variable for an e-variable E, we have $P \leqslant 1/E$. In Theorem 1, we obtain $1/P = 1 + E > E$, as expected.

In the subsequent analysis, we compare p-variables and e-variables for other hypotheses with those in Theorem 1. For a concise presentation, we always write

$$P_0 = \{1 + (X - \mu)_+^2/\sigma^2\}^{-1} \quad \text{and} \quad E_0 = (X - \mu)_+^2/\sigma^2,$$

which are the p-variable and e-variable in Theorem 1, and note the connection $P_0 = (1 + E_0)^{-1}$.

We next consider hypothesis $H_S(\mu, \sigma)$ of symmetric distributions.

THEOREM 2. *A semiprecise p-variable for $H_S(\mu, \sigma)$ is $P = \min\{(2E_0)^{-1}, P_0\}$, and a precise e-variable for $H_S(\mu, \sigma)$ is $E = 2E_0$. Precise p-variables do not exist for $H_S(\mu, \sigma)$.*

From Theorem 2, the e-variable for $H_S(\mu, \sigma^2)$, which we denote by E_S, is improved by a factor of 2 from E_0 for $H(\mu, \sigma^2)$ due to the additional assumption of symmetry. On the other hand, the p-variable in Theorem 2, denoted P_S, is improved from P_0 by taking an extra minimum with $1/E_S$. In the most relevant case that $P_0 \leqslant 1/2$, or, equivalently, $E_0 \geqslant 1$, indicating some evidence against the null, we have $P_S = 1/E_S$.

Next, we will see that hypothesis $H_U(\mu, \sigma)$ of unimodal distributions admits the same precise e-variable, but a quite improved p-variable, compared to P_0 and E_0. This class includes, for instance, the commonly used gamma, beta and log-normal distributions.

THEOREM 3. *A precise p-variable for $H_U(\mu, \sigma)$ is*

$$P = \max\left(\frac{4}{9}P_0, \frac{4P_0 - 1}{3}\right),$$

and a precise e-variable for $H_U(\mu, \sigma)$ is $E = E_0$.

We denote the p-variable in Theorem 3 by P_U and the e-variable by E_U. If P_0 is smaller than $3/8$, corresponding to $(X-\mu)/\sigma > (5/3)^{1/2}$, then $P_U = 4P_0/9$; that is, the unimodality assumption reduces the p-variable by a multiplicative factor of $4/9$ compared to $H(\mu, \sigma)$. On the other hand, the e-variable E_U does not get improved at all compared to E_0.

The proof of Theorem 3, in particular on the factor of 4/9 for the p-variable, is based on Theorem 1 of Bernard et al. (2020), which gives

$$\sup_{Q\in H_{\mathrm{U}}(0,1)} T_X^Q(1-\alpha) = \max\left\{\left(\frac{4-9\alpha}{9\alpha}\right)^{1/2}, \left(\frac{3-3\alpha}{1+3\alpha}\right)^{1/2}\right\} \quad \text{for } \alpha \in (0,1),$$

and applying Lemma 2 by inverting of the above curve as a function of α.

Finally, we consider hypothesis $H_{\mathrm{US}}(\mu,\sigma)$ of unimodal-symmetric distributions. This class includes, for instance, the popular normal, t- and Laplace distributions. To construct a semiprecise p-variable for this hypothesis, we use the following lemma of quantile bounds within $H_{\mathrm{US}}(\mu,\sigma)$, which may be of independent interest. In what follows, $\mathbb{1}$ is the indicator function; that is, $\mathbb{1}_A(x) = 1$ if $x \in A$ and $\mathbb{1}_A(x) = 0$ otherwise.

LEMMA 3. *For $\alpha \in (0,1)$, it holds that*

$$\sup_{Q\in H_{\mathrm{US}}(0,1)} T_X^Q(1-\alpha) = \left(\frac{2}{9\alpha}\right)^{1/2} \mathbb{1}_{(0,1/6]}(\alpha) + 3^{1/2}(1-2\alpha)\mathbb{1}_{(1/6,1/2]}(\alpha).$$

The general formula for $H_{\mathrm{US}}(\mu,\sigma)$ can be easily obtained from Lemma 3 via

$$\sup_{Q\in H_{\mathrm{US}}(\mu,\sigma)} T_X^Q(1-\alpha) = \mu + \sigma \sup_{Q\in H_{\mathrm{US}}(0,1)} T_X^Q(1-\alpha).$$

THEOREM 4. *A semiprecise p-variable for $H_{\mathrm{US}}(\mu,\sigma)$ is*

$$P = \frac{2}{9E_0}\mathbb{1}_{[4/3,\infty)}(E_0) + \frac{3-(3E_0)^{1/2}}{6}\mathbb{1}_{(0,4/3)}(E_0) + \mathbb{1}_{\{0\}}(E_0),$$

and a precise e-variable for $H_{\mathrm{US}}(\mu,\sigma)$ is $E = 2E_0$. Precise p-variables do not exist for $H_{\mathrm{US}}(\mu,\sigma)$.

The proof of Theorem 4 relies on Lemma 3, which is a new technical result. The value 2/9 appeared earlier in Table 1 of Li et al. (2018) for $\alpha \leqslant 1/6$, a result weaker than Lemma 3.

We denote the p-variable obtained from Theorem 4 by P_{US} and the e-variable by E_{US}. One may check that P_{US} is smaller than both P_{U} and P_{S} unless $X \leqslant \mu$, in which case they are equal to 1. For $(X-\mu)/\sigma \geqslant (5/3)^{1/2}$, or, equivalently, $P_0 \leqslant 3/8$, we have the simple relations

$$P_{\mathrm{S}} = \frac{P_0}{2(1-P_0)}, \qquad P_{\mathrm{U}} = \frac{4}{9}P_0 \quad \text{and} \quad P_{\mathrm{US}} = \frac{2P_0}{9(1-P_0)},$$

implying the order $P_0 > P_{\mathrm{S}} > P_{\mathrm{U}} > P_{\mathrm{US}}$ unless $P_0 = 0$. For instance, if we observe $(X-\mu)/\sigma = 3$ then the p-values are

$$P_0 = \tfrac{1}{10} = 0.1, \qquad P_{\mathrm{S}} = \tfrac{1}{18} \approx 0.056, \qquad P_{\mathrm{U}} = \tfrac{2}{45} \approx 0.044 \quad \text{and} \quad P_{\mathrm{US}} = \tfrac{2}{81} \approx 0.025.$$

Table 1. *Formulas for p-variables and e-variables*

Hypothesis	P-variable	E-variable
$H(0,1)$	$(1+X_+^2)^{-1}$	X_+^2
$H_{\mathrm{S}}(0,1)$	$\frac{1}{2}X^{-2}$ if $X \geqslant 1$	$2X_+^2$
	$(1+X_+^2)^{-1}$ if $X<1$	
$H_{\mathrm{U}}(0,1)$	$\frac{4}{9}(1+X^2)^{-1}$ if $X \geqslant (5/3)^{1/2}$	X_+^2
	$\frac{4}{3}(1+X_+^2)^{-1}-\frac{1}{3}$ if $X<(5/3)^{1/2}$	
$H_{\mathrm{US}}(0,1)$	$\frac{2}{9}X^{-2}$ if $X \geqslant (4/3)^{1/2}$	$2X_+^2$
	$\frac{1}{2}-\frac{3^{1/2}}{6}X$ if $0<X<(4/3)^{1/2}$	
	1 if $X \leqslant 0$	

On the other hand, the corresponding e-values are

$$E_0 = 9, \qquad E_{\mathrm{S}} = 18, \qquad E_{\mathrm{U}} = 9 \quad \text{and} \quad E_{\mathrm{US}} = 18.$$

For a comparison, if we are testing the simple parametric hypothesis $N(0,1)$ against $N(3,1)$ with one observation $X=3$, then the corresponding Neyman–Pearson p-value is 0.001 35 and the corresponding likelihood ratio e-value is 90.02. This is not surprising as, generally, p-values and e-values built for composite hypotheses are more conservative than those for simple hypotheses based on the same data.

We summarize our construction formulas for p-variables and e-variables in Table 1 by breaking them down using ranges of X. To obtain the formulas for a general (μ,σ) other than $(0,1)$, it suffices to replace X in Table 1 by $(X-\mu)/\sigma$.

We conclude the section by making a few technical remarks on the obtained results.

First, all results hold if the conditions $\mathbb{E}^Q(X) \leqslant \mu$ and $\mathrm{var}^Q(X) \leqslant \sigma^2$ in each hypothesis is replaced by $\mathbb{E}^Q(X)=\mu$ and $\mathrm{var}^Q(X)=\sigma^2$, respectively. Such modifications narrow the hypotheses and hence all validity statements hold. The precision statements can be checked with similar arguments to our proofs, and we omit them. Therefore, knowing $\mathrm{var}^Q(X)=\sigma^2$ on top of $\mathrm{var}^Q(X) \leqslant \sigma^2$, or $\mathbb{E}^Q(X)=\mu$ on top of $\mathbb{E}^Q(X) \leqslant \mu$, does not lead to more powerful one-sided p-variables or e-variables.

Second, admissibility of the proposed p-variables and e-variables needs future research. For e-variables, admissibility is not difficult to establish, but the picture is different for p-variables. By Lemma 1, there always exists a smallest p-variable. It remains unclear whether the p-variables we obtained in Theorems 1–4 are the smallest ones for the four hypotheses, respectively.

Third, for any hypothesis H, we can define a function $g\colon \alpha \mapsto \sup_{Q\in H} T_X^Q(1-\alpha)$. If g is strictly decreasing on $(0,1)$, as in the case of $H(\mu,\sigma)$ and $H_{\mathrm{U}}(\mu,\sigma)$, then choosing $f=g^{-1}$ yields a precise p-variable $f(X)$. For H being $H_{\mathrm{S}}(\mu,\sigma)$ and $H_{\mathrm{US}}(\mu,\sigma)$, g is flat on $[1/2,1)$, making it impossible to find a decreasing f such that $\inf_{Q\in H} T_{f(X)}^Q(\alpha)=\alpha$ for all $\alpha \in (0,1)$.

4. Testing the null hypotheses

4.1. *Constructing e-processes*

We next build tests based on e-values and p-values. In this subsection we describe the main methodology based on e-processes for the one-sided testing problem.

Let $\mu \in \mathbb{R}$ and $\sigma > 0$. We consider the following hypotheses by keeping the same notation as in § 3:

$$\begin{aligned}
H(\mu,\sigma) &= \{Q\colon \mathbb{E}^Q(X_i \mid \mathcal{F}_{i-1}) \leqslant \mu \text{ and } \mathrm{var}^Q(X_i \mid \mathcal{F}_{i-1}) \leqslant \sigma^2 \text{ for } i \in [n]\},\\
H_{\mathrm{S}}(\mu,\sigma) &= \{Q \in H(\mu,\sigma)\colon X_i \mid \mathcal{F}_{i-1} \text{ is symmetrically distributed for } i \in [n]\},\\
H_{\mathrm{U}}(\mu,\sigma) &= \{Q \in H(\mu,\sigma)\colon X_i \mid \mathcal{F}_{i-1} \text{ is unimodally distributed for } i \in [n]\},\\
H_{\mathrm{US}}(\mu,\sigma) &= H_{\mathrm{U}}(\mu,\sigma) \cap H_{\mathrm{S}}(\mu,\sigma).
\end{aligned}$$

Recall that, without loss of generality, we consider μ and σ^2 as constants. We can also test the hypotheses where some data are symmetric or unimodal and some are not, because we build e-values from each of them separately. For simplicity, we only list the above four representative cases. Using a similar formulation, the hypothesis of Waudby-Smith & Ramdas (2024) is

$$H_{\mathrm{WSR}}(\mu) = \{Q \in H(\mu,1)\colon X_i \mid \mathcal{F}_{i-1} \text{ is supported in } [0,1] \text{ almost surely for } i \in [n]\}.$$

In the above formulation, the choice of $\sigma = 1$ is simply to remove the variance constraint; see Remark 1.

There are several simple ways to use the results in § 3 to construct an e-variable or p-variable for the above hypotheses; some of these methods are more useful than others. In general, we can compute an e-variable E_i or p-variable P_i based on X_i for $i \in [n]$ using Theorems 1–4, and then combine them.

Our main proposal is to use e-processes. An e-process $M = (M_t)_{t\in[n]}$ can be constructed using

$$M_t = \prod_{i=1}^{t}(1 - \lambda_i + \lambda_i E_i), \tag{4}$$

where λ_i is $\mathcal{F}_{i-1}$ measurable and takes values in $[0,1)$. This idea is the main methodology behind game-theoretic statistics; see Shafer & Vovk (2019), Shafer (2021) and Waudby-Smith & Ramdas (2024, Proposition 3). It has been used by Waudby-Smith & Ramdas (2024) for testing the mean and by Wang et al. (2024) for testing risk measures. To find good choices of $\lambda = (\lambda_i)_{i\in[n]}$ is a nontrivial task. We propose to specify λ in two different ways.

(a) *E-mixture method.* We first take several $\lambda_i = \lambda \in [0,1)$, which is a constant for each $i \in [n]$, and then average the resulting e-processes from (4) over these choices to get an e-process. An uninformative choice of the values of λ may be some points in $[0, 0.2]$. We avoid choosing λ close to 1 because our e-value may take value 0 with substantial probability, leading to a small value of $\mathbb{E}^Q\{\log(1-\lambda+\lambda E)\}$. This quantity measures the growth rate of an e-process; see Grünwald et al. (2024) and Waudby-Smith & Ramdas (2024). In our simulation and empirical studies, we average over $\lambda = 0.01 \times \{1,\ldots,20\}$.

(b) *E-GREE method.* In the GREE (growth-rate for empirical e-statistics) method of Wang et al. (2024) for λ_i, $i \in [n]$, in (4), λ_i is determined by solving the optimization problem

$$\lambda_i = \left\{\operatorname*{arg\,max}_{\lambda\in[0,1)} \frac{1}{i-1}\sum_{j=1}^{i-1}\log(1-\lambda+\lambda E_j)\right\} \wedge \frac{1}{2}. \tag{5}$$

To simplify the maximization in (5), a fast and approximate solution can be obtained using a Taylor expansion, as in Waudby-Smith & Ramdas (2024). This leads to the simple formula

$$\lambda_i = \left\{ \frac{\sum_{j=1}^{i-1}(E_j - 1)}{\sum_{j=1}^{i-1}(E_j - 1)^2} \right\}_+ \wedge \frac{1}{2}. \tag{6}$$

We use (6) for all e-GREE related calculations for the following results. Our unreported simulation suggests that using (5) and (6) yield very similar results.

When the hypothesis to test is $H_{\mathrm{WRS}}(\mu)$, the e-GREE method reduces to the method of Waudby-Smith & Ramdas (2024); see § 5.2 below. An optimization procedure related to (5) is studied by Kumon et al. (2011).

For either the e-GREE or the e-mixture method, we fix $\alpha \in (0, 1)$ and reject the null hypothesis if the e-process M goes beyond $1/\alpha$, that is, when $M_t \geqslant 1/\alpha$ for the first time. The Type-I error control is guaranteed by Ville's inequality (Ville, 1939) as $\mathbb{P}(\sup_{t\in[n]} M_t \geqslant 1/\alpha) \leqslant \alpha$, because any e-process is almost surely upper bounded by nonnegative supermartingales with initial value one; see Ramdas et al. (2022).

The result below clarifies consistency of the e-GREE method in the most idealistic setting.

PROPOSITION 1. *Suppose that data are independent and identically distributed and generated from an alternative probability Q. The e-GREE method has asymptotic power approaching 1 as $n \to \infty$, that is, $Q(\sup_{t\in[n]} M_t \geqslant 1/\alpha) \to 1$ for any $\alpha \in (0, 1)$ if and only if* $\mathbb{E}^Q(E_1) > 1$.

Although Proposition 1 requires an independent and identically distributed assumption, this assumption is not needed for consistency in practical situations; a simulation example is given in § 5.1 below.

4.2. *Some other methods*

Below we list some other methods, where we assume that n is finite. They do not generally work well, as shown by the simulation studies, but nevertheless we list them as they follow from our results in § 3, and they are presented only for a comparison.

(c) *P-Fisher method.* Construct a p-variable P using the Fisher combination

$$P = 1 - \chi_{2n}\{-2(\log P_1 + \cdots + \log P_n)\},$$

where χ_{2n} is the cumulative distribution function of a chi-square distribution with $2n$ degrees of freedom.

(d) *P-Simes method.* Construct a p-variable P using the Simes combination (see Simes, 1986),

$$P = \min_{i\in[n]} \frac{n}{i} P_{(i)},$$

where $P_{(i)}$ is ith order statistic of $P_1, \ldots, P_n$ from the smallest to the largest.

Although in general the p-Fisher and p-Simes methods require independence among p-variables, they are valid in our setting since our p-variables are conditionally valid, and they can be combined as if they are independent and identically distributed; a proof of this is presented in the Supplementary Material.

The next two methods use all data directly, and require independence among $X_1, \ldots, X_n$. A most natural statistic is the sample mean $T = \sum_{i=1}^n X_i/n$. Under $H(\mu, \sigma)$, T has at most mean μ and variance at most σ^2/n. Moreover, symmetry of T follows from symmetry of $X_1, \ldots, X_n$. Nevertheless, T is not necessarily unimodal even if $X_1, \ldots, X_n$ are unimodal, and hence unimodality of T cannot be used. The following e-variables and p-variables are constructed by directly applying Theorems 1–4.

(e) *E-batch method.* An e-variable for $H(\mu, \sigma)$ or $H_{\mathrm{U}}(\mu, \sigma)$ is

$$E_0 = n(T - \mu)_+^2/\sigma^2;$$

an e-variable for $H_{\mathrm{S}}(\mu, \sigma)$ or $H_{\mathrm{US}}(\mu, \sigma)$ is

$$E_{\mathrm{S}} = 2n(T - \mu)_+^2/\sigma^2.$$

(f) *P-batch method.* A p-variable for $H(\mu, \sigma)$ or $H_{\mathrm{U}}(\mu, \sigma)$ is

$$P_0 = (1 + E_0)^{-1};$$

a p-variable for $H_{\mathrm{S}}(\mu, \sigma)$ or $H_{\mathrm{US}}(\mu, \sigma)$ is

$$P_{\mathrm{S}} = \min\{(2E_0)^{-1}, P_0\}.$$

All methods described in this section have Type-I error control under the null hypothesis and with finite sample without requiring that the data are identically distributed. Methods (e) and (f) additionally require independence.

4.3. *Two-sided e-values testing the mean given variance*

We briefly discuss the two-sided mean testing problem, where the main hypothesis $H(\mu^L, \mu^U, \sigma)$ to test is

$$\{Q \colon \mathbb{E}^Q(X_i \mid \mathcal{F}_{i-1}) \in [\mu^L, \mu^U] \text{ and } \mathrm{var}^Q(X_i \mid \mathcal{F}_{i-1}) \leqslant \sigma^2 \text{ for } i \in [n]\},$$

where $\mu^L \leqslant \mu^U$ are constants. The case $\mu^L = \mu^U$ corresponds to testing whether the mean is equal to a precise value.

Our methodology can be easily adapted to test this hypothesis. The e-variable E given by

$$E = \frac{(X - \mu^U)_+^2 + (X - \mu^L)_-^2}{\sigma^2} \tag{7}$$

is a precise e-variable for $H(\mu^L, \mu^U, \sigma)$ formulated on a single observation X. To see this, it suffices to note that, for $Q \in H(\mu^L, \mu^U, \sigma)$,

$$\begin{aligned} \mathbb{E}^Q(E) &= \mathbb{E}^Q\left\{\frac{(X - \mu^U)_+^2 + (X - \mu^L)_-^2}{\sigma^2}\right\} \\ &\leqslant \mathbb{E}^Q\left[\frac{\{X - \mathbb{E}^Q(X)\}_+^2 + \{X - \mathbb{E}^Q(X)\}_-^2}{\sigma^2}\right] = \frac{\mathrm{var}^Q(X)}{\sigma^2} \leqslant 1. \end{aligned}$$

The statement on its precision can be verified similarly to Theorem 1.

If $\mu^L = \mu^U = \mu$ then the e-variable in (7) is

$$E = (X - \mu)^2/\sigma^2.$$

This e-variable satisfies the property that $\mathbb{E}^Q(E) > 1$ if $\mathbb{E}^Q(X) \neq \mu$ and $\mathrm{var}^Q(X) = \sigma^2$; this condition is useful to establish consistency in Proposition 1.

Following the same procedure in § 4.1 using (7), we obtain e-processes for the two-sided problem $H(\mu^L, \mu^U, \sigma)$. Because of a smaller null hypothesis, this e-process is generally more powerful than that in § 4.1 testing the one-sided mean.

There are special, adversarial scenarios where such two-sided tests may not be powerful. For instance, if data are independent with $\mathbb{E}(X_i) < \mu$ and $\mathbb{E}(X_j) > \mu$ appearing in an alternating sequence; this forms a dataset that looks like independent and identically distributed data with mean μ, and is thus very difficult to detect. The same challenge exists for other methods based on e-processes, such as that of Waudby-Smith & Ramdas (2024).

Remark 4. Under the additional information of symmetry, the e-variable in (7) can be used, but it cannot be multiplied by 2 as in Theorem 2. In this case, an alternative way to take advantage of symmetry is to build two e-processes in § 4.1: one to test $\mathbb{E}(X_i \mid \mathcal{F}_{i-1}) \leqslant \mu^U$ and another to test $\mathbb{E}(-X_i \mid \mathcal{F}_{i-1}) \leqslant -\mu^L$. Taking the average of these two e-processes yields a valid e-process for the null hypothesis. As long as one of the two e-processes has good power for the true data-generating procedure, the average e-process has good power.

4.4. *Power of the e-values with fixed mean and growing variance*

In this section, we analyse the power of the e-variables. For a given e-variable E, its e-power, using the terminology of Vovk & Wang (2024), for an alternative probability Q is defined as $\mathbb{E}^Q(\log E)$; see Shafer (2021) and Grünwald et al. (2024) for using this quantity as a notion of power. Certainly, the power depends on the specific alternative Q. We are particularly interested in how the e-power changes as the variance in the alternative hypothesis grows.

For this purpose, we consider a simplistic, yet representative setting, where a class of simple alternatives $(Q_\sigma)_{\sigma>1}$ is indexed by $\sigma > 1$, such that our data point X under Q_σ is distributed as σZ, where Z has a fixed distribution with mean 0 and variance 1 satisfying the null hypothesis, which can be one of $H(0, 1)$, $H_{\mathrm{S}}(0, 1)$, $H_{\mathrm{U}}(0, 1)$ and $H_{\mathrm{US}}(0, 1)$. In this setting, the mean of the data is always 0, and only its variance grows under the alternative. We denote by Q_0 a null probability. Below, we show that the e-power of each an e-variable grows at a rate of $\log \sigma$ as the alternative variance σ^2 grows, regardless of the distribution of Z.

Let E be the e-variable computed based on X as in § 3. Because of the construction of the e-process M in (4), the e-power of relevance is defined as

$$\Pi^{Q_\sigma} = \sup_{\lambda \in [0,1]} \mathbb{E}^{Q_\sigma}\{\log(1 - \lambda + \lambda E)\} = \sup_{\lambda \in [0,1]} \mathbb{E}^{Q_0}\{\log(1 - \lambda + \lambda \sigma^2 E)\},$$

that is, the best-achievable e-power in each multiplicative term in the e-process M.

PROPOSITION 2. *Suppose that* $p := Q_0(E \geqslant 1) > 0$. *For* $\sigma > 1$,

$$(2p \log \sigma - \log 2)_+ \leqslant \Pi^{Q_\sigma} \leqslant 2 \log \sigma. \tag{8}$$

Moreover, $0 \leqslant \Pi^{Q_\sigma} - \Pi^{Q_\delta} \leqslant 2(\log \sigma - \log \delta)$ *for* $\sigma > \delta > 1$.

Proposition 2 suggests that the growth rate of the e-process M is roughly a constant times $\log\sigma$ when the alternative variance σ^2 is larger than 1. An additional negative term $-\log 2$ in (8) is not surprising, because our conditions do not guarantee $\Pi^{Q_\sigma} > 0$ for σ very close to 1. Below, we give an example to illustrate the sharpness of the bounds in (8).

Example 1. Suppose that $Q_0(E=0)=Q_0(E=2)=1/2$. We can compute

$$\Pi^{Q_\sigma} = \sup_{\lambda\in[0,1]} \frac{1}{2}[\log(1-\lambda)+\log\{1+\lambda(2\sigma^2-1)\}] = \frac{1}{2}\log\frac{\sigma^4}{2\sigma^2-1}.$$

It is clear that Π^{Q_σ} is approximately equivalent to $\log\sigma$ for large σ, corresponding to the left-hand side of (8) with $p=1/2$.

5. Simulation studies

5.1. *A comparison of different e-combining methods*

In this section, we conduct simulation studies for the nonparametric hypotheses in § 4. We set $\mu=0$ and $\sigma=1$ without loss of generality.

We first concentrate on the null hypothesis $H(0,1)$, as the other four cases are similar. For all the methods stated in § 4, we do not make the assumption that the data are identically distributed. Thus, we generate a sample of n independent data points, although independence is not needed for methods (a)–(d), alternating from two different distributions: $X_1, X_3, \ldots$ follow a normal distribution, and $X_2, X_4, \ldots$ follow a Laplace distribution, with the same mean ν and the same variance η^2. The assumption that the two distributions have the same mean and variance is not necessary when evaluating the power of the methods. We assume this only for simplicity. We denote this data-generating process as $\mathrm{NL}(\nu,\eta^2)$ with the null parameters being $(\nu,\eta^2)=(0,1)$. We consider two alternatives: (i) data generated from $\mathrm{NL}(0,\eta^2)$, where $\eta>1$; (ii) data generated from $\mathrm{NL}(\nu,1)$, where $\nu>0$. In our setting, the tester does not know the alternating data-generating mechanism. For each alternative model, we compute the rejection rate over 1000 runs using the thresholds of $E \geqslant 1/\alpha$ and $P\leqslant\alpha$, with $\alpha=0.05$, for e-values and p-values, respectively.

For the e-mixture method, we experiment by averaging λ in the interval $[0.01, 0.20]$ with step size 0.01. The e-GREE method is similar to the e-mixture method, except that λ_i is dynamically updated with different $i\in[n]$ using (5).

Figure 1 shows the rejection rates for all methods with data generated from $\mathrm{NL}(0,\eta^2)$ for $\eta\in[1,4]$, and from $\mathrm{NL}(\nu,1)$ for $\nu\in[0,1]$. For the alternative model $\mathrm{NL}(0,\eta^2)$, we see that the e-mixture and e-GREE methods outperform the other methods, with the e-mixture method being the most powerful. For $\eta<1.5$, the rejection rates of all methods are very low, making it challenging to distinguish their efficiency. As $\eta>1.5$, both the e-mixture method and the e-GREE method exhibit significantly higher rejection rates compared to other methods, demonstrating their effectiveness in testing $H(0,1)$. The other four methods have almost no power. For the alternative model $\mathrm{NL}(\nu,1)$, we observe that the e-batch method and the p-batch method show significant high rejection rates, since they are quite sensitive to the sample mean. Recall that these methods rely on independence, so the central limit theorem kicks in.

Among all methods, only the e-process-based methods satisfy anytime validity, that is, a decision can be made at any stopping time when data arrive sequentially. This situation

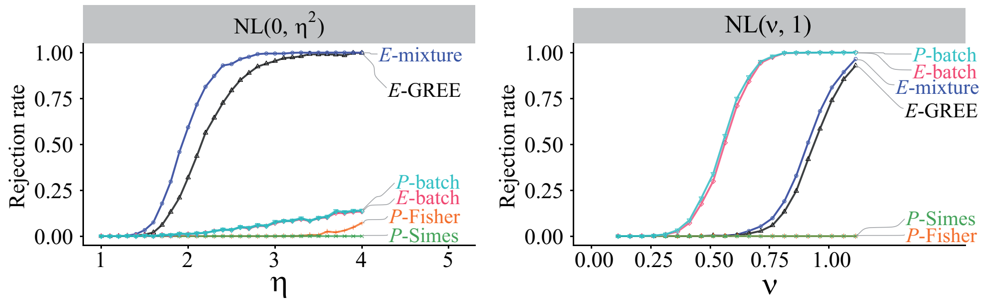

Fig. 1. Rejection rates for all methods for testing $H(0,1)$ with sample size $n = 100$ over 1000 runs using the threshold 20.

Table 2. *Rejection rates for testing* H*,* H_{S}*,* H_{U} *and* H_{US} *with* $n = 100$ *data generated from model* $\mathrm{NL}(0.5, 2)$

	E-mixture	E-GREE	P-Fisher	P-Simes	E-batch	P-batch
H	0.419	0.315	0.000	0	0.639	0.664
H_{S}	0.998	0.882	0.000	0	0.900	0.900
H_{U}	0.419	0.315	0.006	0	0.639	0.664
H_{US}	0.998	0.882	0.763	0	0.900	0.900

is common in financial applications, where realized losses accumulate over time; see the empirical study in § 6 below.

The testing procedures for H_{S}, H_{U} and H_{US} are the same as for testing H. We generate 100 data points from $\mathrm{NL}(0.5, 2)$ and calculated the rejection rates for testing H_{S}, H_{U} and H_{US} with null hypotheses $\mu = 0$ and $\sigma = 1$. Table 2 displays the rejection rates for all hypotheses. It is clear that the extra information of symmetry improves the power.

5.2. *A comparison with the GRAPA method*

Recall that our model can also be interpreted as testing the mean under knowledge of an upper bound on the variance. This allows us to compare our testing approach with the growth rate adaptive to the particular alternative (GRAPA) method proposed by Waudby-Smith & Ramdas (2024). The GRAPA method is similar to the e-GREE method discussed in § 4, but it requires the random variable to be bounded. The e-process $(M_t)_{t\in[n]}$ for the GRAPA method is constructed as

$$M_t = \prod_{i=1}^{t} \{1 + \lambda_i (X_i - \mu)\}, \tag{9}$$

where μ is the conditional mean being tested and λ_i is $\mathcal{F}_{i-1}$ measurable and takes value in $(-1/\{1-\mu\}, 1/\mu)$. It is clear that $1 + \lambda_i(X_i - \mu)$ is an e-variable for each $i \in [n]$. Thus, maximizing the growth of (9) is similar to (5), where λ_i is determined by solving the optimization problem

$$\lambda_i = \underset{\lambda\in[-c/(1-\mu), c/\mu]}{\arg\max} \frac{1}{i-1} \sum_{j=1}^{i-1} \log\{1 + \lambda(X_i - \mu)\}, \tag{10}$$

where $c \in (0, 1]$ is fixed. For faster computation in the context of confidence sequences, Waudby-Smith & Ramdas (2024) also offered an alternative way to obtain λ_i, which they called the approximate GRAPA method, and λ_i is determined by

$$\lambda_i = -\frac{c}{1-\mu} \vee \frac{\hat{\mu}_{i-1} - \mu}{\hat{\sigma}_{i-1}^2 + (\hat{\mu}_{i-1} - \mu)^2} \wedge \frac{c}{\mu}, \tag{11}$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ are the empirical mean and variance of observations $X_1, \ldots, X_i$. From (11), it is clear that the GRAPA method is able to use the sample variance information adaptively. In particular, our e-GREE method in (6) is adaptive to the empirical variance of the e-values. In the simulation results, we use (10) and choose $c = 1/2$.

We compare the following five methods for testing the mean under various conditions.

(a) GRAPA: the GRAPA method with a bounded support $[0, 1]$.
(b) E-GREE: the e-GREE method with the variance upper bound σ^2.
(c) E-mixture: the e-mixture method with the variance upper bound σ^2.
(d) E-GREE-2s: the two-sided e-GREE method with the variance upper bound σ^2.
(e) E-mixture-2s: the two-sided e-mixture method with the variance upper bound σ^2.

GRAPA is designed as a two-sided test, although it can easily be adjusted by restricting λ_i in (9) to be nonnegative.

Remark 5. We could also implement the e-GREE and e-mixture methods without an upper bounded variance, but using the bounded support, as described in Remark 1. Although these methods are valid, they have poor power in our setting, because their assumption is strictly weaker than both bounded variance and bounded support. We omit these results.

We set $\mu = 0.35$ and apply both one-sided and two-sided tests on the same dataset. We generate a sample consisting of n independent data points from a beta distribution, denoted $\mathrm{Be}(\nu, \sigma^2)$, where ν and σ^2 represent the mean and variance of the beta distribution. None of the methods requires that the data follow identical distributions; we use a single distribution just for simplicity. Here, we use ν and σ^2 instead of the standard beta parameters α and β for the sake of convenience. Parameters α and β can be easily recovered based on the given mean ν and variance σ^2: $\alpha = \nu(\nu - \nu^2 - \sigma^2)/\sigma^2$ and $\beta = (\nu^2 + \sigma^2 - \nu)(\nu - 1)/\sigma^2$. Since the beta distribution has a bounded support $[0, 1]$, we can make meaningful comparisons between the GRAPA method and the e-GREE and e-mixture methods.

We first compare the rejection rates, using a threshold of 20 over 1000 runs, for all methods mentioned above under different ν with fixed σ^2. We consider $\nu \geqslant 0.35$ and $\sigma = 0.05$, $\sigma = 0.1$ and $\sigma = 0.3$. We use 20 data points for each run.

Figure 2 shows the performance of the three methods. First, the e-GREE method is always better than the e-mixture method. Second, the two-sided versions of both the e-GREE and e-mixture methods show a slight improvement over their respective one-sided methods, as expected. Third, in the case in which $\sigma = 0.05$ and $\sigma = 0.1$, the e-GREE method outperforms the GRAPA method; in the case in which $\sigma = 0.3$, the GRAPA method demonstrates superior performance compared to the other methods. This is intuitive, because the variance information is less useful for larger σ; recall that, for any distribution supported in $[0, 1]$ with mean $\mu \leqslant 0.35$, the maximum possible variance is 0.2275, and $\sigma \approx 0.477$.

Figure 3 shows the average logarithmic e-processes for n up to 50 by using $\nu = \mu + \sigma$ for each alternative model. The relative rankings of these methods are consistent with their rejection rates, with e-GREE performing the best when σ is relatively small.

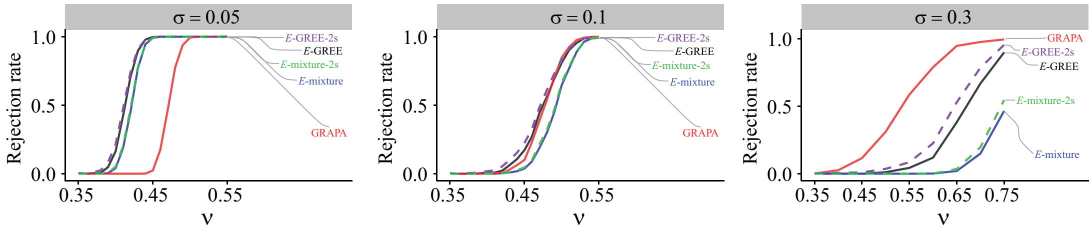

Fig. 2. Rejection rates for the GRAPA, e-GREE, e-mixture and the two-sided e-GREE-2s and e-mixture-2s methods over 1000 runs using the threshold 20 and $\mu = 0.35$. Data are generated from $\mathrm{Be}(\nu, \sigma^2)$ with sample size $n = 20$, where $\nu \geqslant 0.35$ and $\sigma \in \{0.05, 0.1, 0.3\}$.

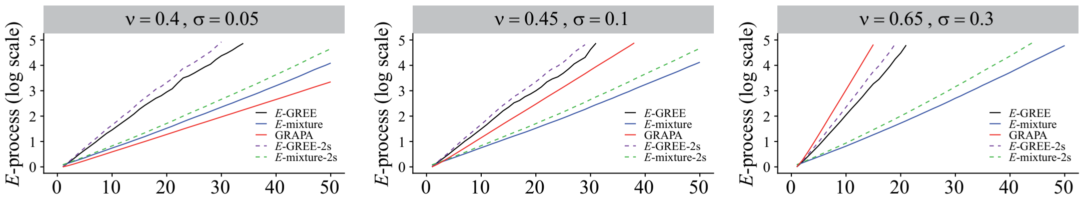

Fig. 3. Average logarithmic e-processes for the GRAPA, e-GREE, e-mixture and the two-sided e-GREE-2s and e-mixture-2s methods with varying sample size and $\mu = 0.35$. Data are generated from $\mathrm{Be}(\nu, \sigma^2)$, where $\sigma \in \{0.05, 0.1, 0.3\}$ and $\nu = \mu + \sigma$.

From the simulation results, our general recommendation is to use the e-GREE method to construct the e-process when the variance to be tested is relatively small, and to use the GRAPA method when the variance to be tested is relatively large compared to the bounded support.

5.3. *A comparison with exponential test supermartingale methods*

Next, we compare our methods with the exponential test supermartingale methods that directly construct e-processes, rather than using a betting strategy to combine sequential e-variables.

Wang & Ramdas (2023) extended the idea of Catoni (2012) to construct a nonnegative test supermartingale called the Catoni supermartingale to test the mean and variance in sequential settings. The test supermartingale is constructed as

$$M_t^{\mathrm{C}} = \prod_{i=1}^{t} \exp\left[\phi\{\lambda_i(X_i - \mu)\} - \frac{\lambda_i^2 \sigma^2}{2}\right],$$

where ϕ is the influence function and $(\lambda_i)_{i \in [n]}$ is any predictable process. Following the recommendation of Wang & Ramdas (2023), we choose the influence function

$$\phi(x) = \begin{cases} \log(1 + x + x^2/2) & \text{if } x \geqslant 0, \\ -\log(1 - x + x^2/2) & \text{if } x < 0, \end{cases}$$

and $(\lambda_i)_{i \in [n]}$ as

$$\lambda_i = \left\{\frac{2\log(1/\alpha)}{i(\sigma^2 + \eta_i^2)}\right\}^{1/2}, \quad \text{where} \quad \eta_i = \left\{\frac{2\sigma^2 \log(1/\alpha)}{i - 2\log(1/\alpha)}\right\}^{1/2}.$$

A different approach by Howard et al. (2021) is to use a framework for nonparametric confidence sequences based on the concept of exponential supermartingales. They introduced the concept of a 'sub-ψ process' in Howard et al. (2021, Definition 1). Informally, a sub-ψ process is a pair of $\mathcal{F}_t$-adapted processes (S_t, V_t) such that S_t is the zero-mean deviation of the sample sum from its estimand at time t and V_t and ψ make the process

$$M_t^{\psi} = \exp\{\lambda S_t - \psi(\lambda) V_t\},$$

dominated by a supermartingale for each λ in an interval $[0, \lambda_{\max})$. This framework allows for testing the mean and variance under a wide variety of assumptions, including bounded supports, self-normalized bounds and symmetric conditions. We refer the reader to Howard et al. (2021, Appendix J, Table 3) for a collection of commonly used ψ functions and variance processes for $S_t = \sum_{i=1}^{t}(X_i - \mu)$ under various assumptions. We choose two special cases for comparison with our methods: the self-normalized bound test supermartingale, denoted $M_t^{\psi,\mathrm{SN}}$, and the symmetric condition test supermartingale, denoted $M_t^{\psi,\mathrm{sym}}$. For $\lambda \in [0, \infty)$, these test supermartingales are constructed as

$$M_t^{\psi,\mathrm{SN}} = \prod_{i=1}^{t} \exp\left\{\lambda(X_i - \mu) - \frac{\lambda^2 (X_i - \mu)^2 + 2\sigma^2}{6}\right\}, \quad (12)$$

which also appears in Wang & Ramdas (2023, § 5), and

$$M_t^{\psi,\mathrm{sym}} = \prod_{i=1}^{t} \exp\left\{\lambda(X_i - \mu) - \frac{\lambda^2 (X_i - \mu)^2}{2}\right\}. \quad (13)$$

We follow a simple method of choosing λ suggested by Howard et al. (2021, §3.2), that is, to use the mixture supermartingale $\int \exp\{\lambda S_t - \psi(\lambda) V_t\}\, \mathrm{d}\Phi(\lambda)$ by assuming that $\lambda \sim \Phi = N(0, 1)$. Now, we compare the following methods.

(f) WR23-Catoni: the Catoni method with the variance upper bound σ^2.
(g) HRMS21-SN: the self-normalized method with the variance upper bound σ^2.
(h) HRMS21-sym: the sub-ψ method with symmetry, but without variance information.
(i) E-GREE-sym: the e-GREE method with the variance upper bound σ^2 and symmetry.
(j) E-mixture-sym: the e-mixture method with the variance upper bound σ^2 and symmetry.

We compare the above five methods, along with methods (a) and (b), the e-GREE and e-mixture methods, that do not utilize symmetric information, in testing $H(0, 1)$. Following the same data-generating process as described in § 5.2, we generate n independent data points alternating between the normal and Laplace distributions, denoted by $\mathrm{NL}(\nu, \eta^2)$. Figure 4 shows rejection rates for the above methods with data generated from three cases: $\mathrm{NL}(\nu, 1^2)$ for $\nu \in [0, 1]$, $\mathrm{NL}(\nu, (1+\nu)^2)$ for $\nu \in [0, 1]$ and $\mathrm{NL}(\nu/5, (1+\nu)^2)$ for $\nu \in [0, 2]$.

For $\mathrm{NL}(\nu, 1^2)$, the Catoni method outperforms other methods, while methods utilizing symmetric information generally perform well. For $\mathrm{NL}(\nu, (1+\nu)^2)$, where both the mean and variance of the data-generating process change, the power of the methods from Howard et al. (2021) reduces. In contrast, the power of our e-value-based methods increases, as our construction of e-values is sensitive to the changes to variance. In the last case, $\mathrm{NL}(\nu/5, (1+\nu)^2)$, the impact of changes in the mean is small and the variance effect is large;

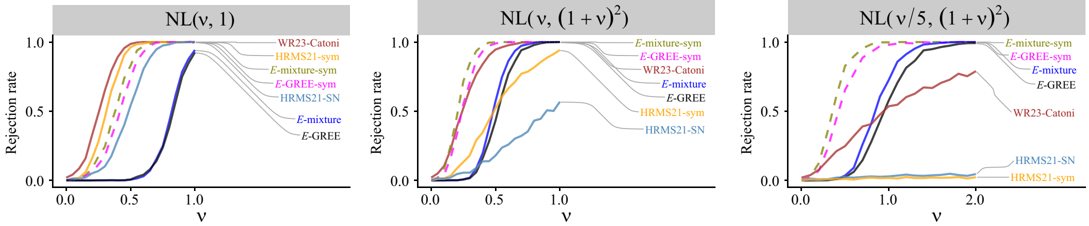

Fig. 4. Rejection rates for methods (a), (b) and (f)–(j) for testing $H(0, 1)$ with sample size $n = 100$ over 1000 runs using the threshold 20.

e-value-based methods generally outperform others. Although method (h) benefits from not requiring information about the variance or even the existence of variance, it demonstrates minimal power when testing the mean with varying variance, due to its penalization term $-(X_i - \mu)^2$ in the exponential form of (12) and (13). In summary, our methods are comparatively more powerful when the alternative variance defers from the null.

6. Empirical study with financial data

We now conduct an empirical study to test the hypothesis $H(\mu, \sigma)$ on the daily losses of financial assets. We aim to calculate the number of trading days required to detect evidence for rejecting the null hypothesis $H(\hat{\mu}, \hat{\sigma})$ during the 2007–8 financial crisis period. Here, $\hat{\mu}$ and $\hat{\sigma}$ represent the sample mean and sample variance estimated from historical data prior to the testing period. That is, we are testing whether the historical estimations before the testing period are still valid. If the null hypothesis can be rejected at a reasonable threshold level rather swiftly, this will serve as evidence of the effectiveness of *e*-process methods and could help investors switch strategies in a timely manner.

We choose 20 stocks from 10 different sectors of the S&P 500 list with large market capitalization in each sector. Moreover, we include two companies with the largest market capitalization from the to-be real estate sector. Real estate became the 11th sector of the S&P 500 in 2016. We first calculate the daily losses for each of the selected stocks from 1 January 2001 to 31 December 2010. The daily losses are expressed as percentages and calculated as $L_t = -(S_{t+1} - S_t)/S_t$, where S_t is the close price at day t. A positive value represents a loss and a negative value represents a gain. We could also use the log-loss data instead of the linear loss data, but the difference between the two is minor. We use the loss data from 1 January 2001 to 31 December 2006 to estimate the mean and variance for the null hypothesis. We compute the *e*-values using both the *e*-mixture method and the *e*-GREE method based on the construction of (4) as the daily loss from 1 January 2007 fed into the *e*-process.

Following a methodology similar to the simulation study in § 5, we report evidence against the null hypothesis when the *e*-process exceeds thresholds of 2, 5, 10 and 20. In accordance with Jeffrey's rule of thumb about *e*-values (see Jeffreys, 1998 and Vovk & Wang, 2021), if the *e*-value falls within the interval of $(10^{1/2}, 10)$, evidence against the null hypothesis is considered substantial; if the *e*-value falls within the interval of $(10, 10^{3/2})$, evidence against the null hypothesis is regarded as strong. Thus, *e*-values exceeding 5 or 10 provide substantial evidence to reject the null hypothesis, while a threshold of 20 offers strong evidence against the null hypothesis. Although a threshold of 2 may not be substantial enough to reject the null hypothesis, it can still serve as an early warning that the stock's performance may be different from its historical path.

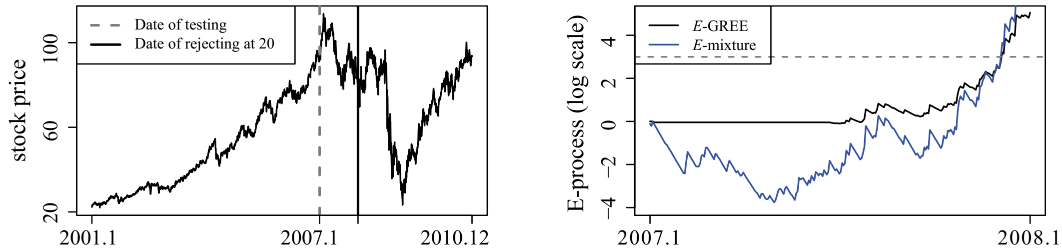

Fig. 5. Sample path and logarithmic e-process using the e-GREE and e-mixture methods' testing of $H(\hat{\mu}, \hat{\sigma})$ for Simon Property (SPG) stock from January 2007 to January 2008, where $\hat{\mu} = -0.001\,028$ and $\hat{\sigma} = 0.012\,123$ are the sample mean and variance estimated from historical data for stock SPG from 1 January 2001 to 31 December 2006.

To illustrate the e-process detection procedure, we first focus on a single stock as an example. Figure 5 reports the stock price for Simon Property (SPG) throughout the detection period and its corresponding e-process initiated on 1 January 2007. Observing from the e-process figure, it is evident that both the e-mixture method and the e-GREE method effectively reject the null hypothesis at thresholds of 2, 5, 10 and 20 before the financial crisis ends. Notably, the e-GREE method generally takes fewer trading days compared to the e-mixture method to achieve this rejection across various threshold levels. Also, the null hypothesis is rejected using the e-GREE method prior to another significant decline in the stock price during February 2009 to June 2009, thus preventing potential larger losses and underscoring the effectiveness of e-process methods.

Compared to the e-batch and other p-variable-based methods stated in § 4, the e-process-based methods exhibit a unique advantage in sequential settings, particularly in financial applications where actual losses accumulate sequentially over time. In such scenarios, the e-process permits the early termination without a specified sampling period, potentially preventing further losses at an earlier stage.

Table 3 displays the number of trading days required to reject the null hypothesis at various threshold levels for the selected 20 stocks from 10 different sectors and the two stocks in real estate. The table shows that stocks in sectors significantly impacted by the 2007–8 subprime crisis, such as financials, consumer discretionary and energy, could generally be detected using e-process-based methods. In particular, the representative companies in real estate are rejected the earliest; see the last rows of Table 3. In contrast, for stocks in sectors less affected by the subprime crisis, such as technology, health care and consumer staples, we are unable to reject the null hypothesis. This is intuitive, given that their prices and returns remain relatively stable or even increase during the financial crisis.

7. DISCUSSION

As shown in the simulation studies, in comparison with the GRAPA method of Waudby-Smith & Ramdas (2024) and with the exponential supermartingale methods of Howard et al. (2020, 2021) and Wang & Ramdas (2023), our proposed methods have superior performance in some settings and have inferior performance in other settings. A full picture of the comparative advantages requires future work.

Our constructions of p-values and e-values are potentially useful for multiple testing, which is not addressed in this paper. The literature on using e-values in multiple testing is growing recently. For instance, e-values are used for false discovery control in knockoffs; see Ren & Barber (2024) for derandomization, Ahn et al. (2023) for Bayesian linear models and

Table 3. *The number of trading days taken to detect evidence against $H(\hat{\mu}, \hat{\sigma})$ using the e-GREE method and the e-mixture method for different stocks from 1 January 2007 to 31 December 2010; an endash means that no detection was observed up to 31 December 2010*

		E-GREE				*E*-mixture			
	Threshold	2	5	10	20	2	5	1	20
Financials	Bank of America	378	385	385	393	393	394	395	403
	Morgan Stanley	429	439	445	447	447	447	447	447
Utilities	The Southern	–	–	–	–	–	–	–	-
	Duke Energy	–	–	–	–	–	–	–	–
Communication services	Verizon Comms.	–	–	–	–	–	–	–	–
	AT&T	–	–	–	–	–	–	–	–
Consumer staples	Walmart	–	–	–	–	–	–	–	–
	PepsiCo	–	–	–	–	–	–	–	–
Consumer discretionary	Ford Motor	476	491	498	565	546	594	594	594
	Las Vegas Sands	442	445	447	450	451	454	457	457
Energy	Texas Pacific Land	158	244	261	269	242	261	261	263
	Pioneer	496	622	–	–	–	–	–	–
Material	Southern Copper	476	496	537	–	539	–	–	–
	Air Products	476	516	537	–	–	–	–	–
Health care	Johnson & Johnson	–	–	–	–	–	–	–	–
	Pfizer	–	–	–	–	–	–	–	–
Technology	Int. Business Machines	–	–	–	–	–	–	–	–
	Microsoft	–	–	–	–	–	–	–	–
Industrials	General Electric	537	546	578	–	–	–	–	–
	United Parcel Service	476	524	542	632	542	604	–	–
Real estate	Simon Property	165	224	242	254	223	239	250	253
	Prologis	264	271	271	296	270	271	271	275

Gablenz & Sabatti (2024) for resolution-adaptive variable selection. Finally, the obtained e-variables may also be useful to build e-confidence regions, as in Vovk & Wang (2023), and the e-posterior, as in Grünwald (2023) for (μ, σ^2), although we mainly consider a nonparametric setting.

Acknowledgement

We thank the editor, an associate editor and two referees for constructive comments. We also thank Aaditya Ramdas, Qiuqi Wang and Ian Waudby-Smith for helpful discussions. Wang was partly supported by the Natural Sciences and Engineering Research Council of Canada.

Supplementary material

The Supplementary Material contains proofs of all the results.

References

Ahn, T., Lin, L. & Mei, S. (2023). Near-optimal multiple testing in Bayesian linear models with finite-sample FDR control. *arXiv:* 2211.02778v3.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289–300.

BENJAMINI, Y. & YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–88.

BERNARD, C., KAZZI, R. & VANDUFFEL, S. (2020). Range value-at-risk bounds for unimodal distributions under partial information. *Insur.: Math. Econ.* **94**, 9–24.

CATONI, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. H. Poincaré Prob. Statist.* **48**, 1148–85.

CHI, Z., RAMDAS, A. & WANG, R. (2024). Multiple testing under negative dependence. To appear in *Bernoulli*.

CONOVER, W. J. (1999). *Practical Nonparametric Statistics*. New York: John Wiley and Sons.

FISSLER, T. & ZIEGEL, J. F. (2016). Higher order elicitability and Osband's principle. *Ann. Statist.* **44**, 1680–707.

GABLENZ, P. & SABATTI, C. (2024). Catch me if you can: signal localization with knockoff *e*-values. *J. R. Statist. Soc.* B doi: 10.1093/jrsssb/qkae042.

GRÜNWALD, P. (2023). The e-posterior. *Phil. Trans. R. Soc.* A **381**, 20220146.

GRÜNWALD, P., DE HEIDE, R. & KOOLEN, W. M. (2024). Safe testing. *J. R. Statist. Soc. B*, doi: 10.1093/jrsssb/qkae011.

HENZI, A. & ZIEGEL, J. F. (2022). Valid sequential inference on probability forecast performance. *Biometrika* **109**, 647–63.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Am. Statist. Assoc.* **58**, 13–30.

HOLLANDER, M., WOLFE, D. A. & CHICKEN, E. (2013). *Nonparametric Statistical Methods*. New York: John Wiley and Sons.

HOWARD, S. R., RAMDAS, A., MCAULIFFE, J. & SEKHON, J. (2020). Time-uniform Chernoff bounds via nonnegative supermartingales. *Prob. Surv.* **17**, 257–317.

HOWARD, S. R., RAMDAS, A., MCAULIFFE, J. & SEKHON, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *Ann. Statist.* **49**, 1055–80.

IGNATIADIS, N., WANG, R. & RAMDAS, A. (2024). *E*-values as unnormalized weights in multiple testing. *Biometrika* **111**, 417–39.

JEFFREYS, H. (1998). *The Theory of Probability*, 3rd ed. New York: Oxford University Press.

KUMON, M., TAKEMURA, A. & TAKEUCHI, K. (2011). Sequential optimizing strategy in multidimensional bounded forecasting games. *Stoch. Proces. Appl.* **121**, 155–83.

LEHMANN, E. L., ROMANO, J. P. & CASELLA, G. (1986). *Testing Statistical Hypotheses*. New York: Springer.

LI, L., SHAO, H., WANG, R. & YANG, J. (2018). Worst-case range value-at-risk with partial information. *SIAM J. Finan. Math.* **9**, 190–218.

RAMDAS, A., GRÜNWALD, P., VOVK, V. & SHAFER, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statist. Sci.* **38**, 576–601.

RAMDAS, A., RUF, J., LARSSON, M. & KOOLEN, W. (2022). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv:* 2009.03167v3.

REN, Z. & BARBER, R. F. (2024). Derandomized knockoffs: leveraging *e*-values for false discovery rate control. *J. R. Statist. Soc.* B **86**, 122–54.

SHAFER, G. (2021). The language of betting as a strategy for statistical and scientific communication. *J. R. Statist. Soc.* A **184**, 407–31.

SHAFER, G. & VOVK, V. (2019). *Game-Theoretic Foundations for Probability and Finance*. New York: John Wiley.

SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–4.

VILLE, J. (1939). *Étude Critique de la Notion de Collectif* (thèses de l'entre-deux-guerres **218**). Paris: Gauthier-Villares.

VOVK, V. & WANG, R. (2020). Combining *p*-values via averaging. *Biometrika* **107**, 791–808.

VOVK, V. & WANG, R. (2021). E-values: calibration, combination, and applications. *Ann. Statist.* **49**, 1736–54.

VOVK, V. & WANG, R. (2023). Confidence and discoveries with *e*-values. *Statist. Sci.* **38**, 329–54.

VOVK, V. & WANG, R. (2024). Nonparametric e-tests of symmetry. *New Engl. J. Statist. Data Sci.* **2**, 261–70.

WANG, H. & RAMDAS, A. (2023). Catoni-style confidence sequences for heavy-tailed mean estimation. *Stoch. Proces. Appl.* **163**, 168–202.

WANG, Q., WANG, R. & ZIEGEL, J. (2024). E-backtesting. *arXiv:* 2209.00991v4.

WANG, R. & RAMDAS, A. (2022). False discovery rate control with *e*-values. *J. R. Statist. Soc.* B **84**, 822–52.

WASSERMAN, L., RAMDAS, A. & BALAKRISHNAN, S. (2020). Universal inference. *Proc. Nat. Acad. Sci.* **117**, 16880–90.

WAUDBY-SMITH, I. & RAMDAS, A. (2024). Estimating means of bounded random variables by betting. *J. R. Statist. Soc.* B **86**, 1–27.

[*Received on* 24 *November* 2023. *Editorial decision on* 23 *September* 2024]