

# Testing mean and variance by e-processes

Yixuan Fan\*      Zhanyi Jiao†      Ruodu Wang‡

September 24, 2024

## Abstract

We address the problem of testing conditional mean and conditional variance for non-stationary data. We build e-values and p-values for four types of non-parametric composite hypotheses 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.

**Keywords:** P-values, e-values, e-processes, symmetry, unimodality

## 1 Introduction

Testing mean and variance in various settings is a classic problem in statistics. In parametric inference concerning testing the mean, well-known tests like 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. [20]. 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, non-parametric methods provide a great alternative. Certainly, non-parametric 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 non-parametric techniques for testing means and variances can be found in

---

\*Center for Applied Mathematics, Tianjin University, Tianjin, China E-mail: [yxfan-jlu@gmail.com](mailto:yxfan-jlu@gmail.com)

†Department of Mathematics, Illinois State University, Normal, Illinois, USA. E-mail: [zjiao1@ilstu.edu](mailto:zjiao1@ilstu.edu)

‡Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada. E-mail: [wang@uwaterloo.ca](mailto:wang@uwaterloo.ca).

e.g., Conover [7] and Hollander et al. [14]. Different from the classic settings, we will consider the problem of testing composite hypotheses in which data are not stationary.

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

$$\mathbb{E}[X_i | \mathcal{F}_{i-1}] \leq \mu_i \quad \text{and} \quad \text{var}(X_i | \mathcal{F}_{i-1}) \leq \sigma_i^2 \quad \text{for each } i, \quad (1)$$

where  $\mathcal{F}_{i-1}$  is the  $\sigma$ -algebra generated by  $X_1, \dots, X_{i-1}$ , and  $\mu_i$  and  $\sigma_i$  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 mean and variance. Testing conditional mean and conditional variance is common in some contexts such as forecasting (e.g., Henzi and Ziegel [12]) and financial risk assessment (e.g., Fissler and Ziegel [8]).

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 the knowledge of an upper bound on the variance.

The 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 the knowledge of an upper bound on the mean.

Clearly, problem (1) is a composition of many complicated, non-parametric, composite hypotheses on each observation. The key challenge in this setting is that the data points are not iid, 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 and Ramdas [37]. 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 and Wang [30], Grünwald et al. [10], and the review by Ramdas et al. [22]. 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.

Section 2 formally describes the hypotheses and defines e-variables, e-processes, and p-variables. As mentioned above, we will first address the case of one data point, i.e.,  $n = 1$ , presented in Section 3. We consider four types of composite hypotheses on 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 present results also 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 and Hochberg [2]. Considering a non-parametric 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 in Section 4 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 in Shafer [26], Wasserman et al. [36] and Waudby-Smith and Ramdas [37]. 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,

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

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

The closest methodological work related to this paper is Waudby-Smith and Ramdas [37], where the authors test in a non-parametric setting the conditional mean of sequential data, which are assumed to be bounded within a pre-specified range, thus a generally smaller class of distributions. Our problem and methodology are different from Waudby-Smith and Ramdas [37] 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 and Ramdas [37], 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 and Ramdas [37] is that their power adapts to the unknown true variance of the distribution if data come from an iid 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 and Ramdas [37]. Another closely related methodology is Wang et al. [33], where, other statistical functions are tested other than the mean. Once e-variables are constructed, we will build e-processes in a similar way to Wang et al. [33]. The methods of Howard et al. [15, 16] and Wang and Ramdas [35] 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.

Section 5 provides simulation studies for the proposed methods and compare them with the method of Waudby-Smith and Ramdas [37] when the model

has both bounded support and bounded variance and with methods based on exponential test supermartingale of Howard et al. [16] and Wang and Ramdas [35]. Section 6 contains empirical studies using financial asset return data during the 2007–2008 financial crisis, further demonstrating the effectiveness of the e-process based methods. Section 7 concludes the paper. 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  $\infty$ , and denote by  $[n] = \{1, \dots, n\}$ . 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  $\sigma$ -field generated by  $X_1, \dots, X_i$  for  $i \in [n]$  with  $\mathcal{F}_0$  being the trivial  $\sigma$ -field. The main hypotheses of interest are variations (by adding shape information) of the following hypothesis

$$H = \{Q : \mathbb{E}^Q[X_i | \mathcal{F}_{i-1}] \leq \mu_i \text{ and } \text{var}^Q(X_i | \mathcal{F}_{i-1}) \leq \sigma_i^2 \text{ for } i \in [n]\}, \quad (2)$$

where  $\mu_i$  and  $\sigma_i$  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

$$H = \{Q : \mathbb{E}^Q[X_i | \mathcal{F}_{i-1}] \leq \mu \text{ and } \text{var}^Q(X_i | \mathcal{F}_{i-1}) \leq \sigma^2 \text{ for } i \in [n]\}, \quad (3)$$

where  $\mu$  and  $\sigma$  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  $\mu_i$  and  $\sigma_i$  are  $\mathcal{F}_{i-1}$ -measurable, they can be absorbed into  $X_i$  by considering  $(X_i - \mu_i)/\sigma_i$  instead of  $X_i$ . Therefore, we will focus on the 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  $\text{var}^Q$  when it is clear.

We will further consider hypotheses with additional shape information, by assuming that some, or all of the distributions of  $X_1, \dots, X_n$  are unimodal, symmetric, or both. Below, all terms like “increasing” and “decreasing” are in the non-strict 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  $(x, \infty)$ ; it may have a point-mass at  $x$ . A distribution on  $\mathbb{R}$  with mean  $\mu$  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  $\mu$  is both unimodal and symmetric, then its mode must be either  $\mu$  or an interval centered at  $\mu$ .

*Remark 1.* The main question in Waudby-Smith and Ramdas [37] is to test the conditional mean  $m$  with data taking values in  $[0, 1]$ . We note that any

random variable with mean at most  $m$  and range  $[0, 1]$  has variance at most  $1/4$  (if  $m \geq 1/2$ ) or  $m(1 - m)$  (if  $m < 1/2$ ), 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  $\mu$  and  $\sigma^2$ . Certainly, all validity results remain true for the two-sided hypotheses. Testing  $\mathbb{E}^Q[X_i] \geq \mu$  is symmetric to testing  $\mathbb{E}^Q[X_i] \leq \mu$ , 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 Section 4.3.

## 2.2 P-variables and e-variables

We formally define p-variables and e-variables, following Vovk and Wang [30]. A *p-variable*  $P$  for a hypothesis  $H$  is a random variable that satisfies  $Q(P \leq \alpha) \leq \alpha$  for all  $\alpha \in (0, 1)$  and all  $Q \in H$ . In other words, a p-variable is stochastically larger than  $U[0, 1]$ , often truncated at 1. An *e-variable*  $E$  for a hypothesis  $H$  is a  $[0, \infty]$ -valued random variable satisfying  $\mathbb{E}^Q[E] \leq 1$  for all  $Q \in H$ . E-variables are often obtained from stopping an *e-process*  $(E_t)_{t \geq 0}$ , which is a non-negative stochastic process adapted to a pre-specified filtration,  $(\mathcal{F}_i)_{i \in [n]}$  in our problem, such that  $\mathbb{E}^Q[E_\tau] \leq 1$  for any stopping time  $\tau$  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  $\sup_{Q \in H} Q(P \leq \alpha) = \alpha$  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 [13], Howard et al. [16] and Waudby-Smith and Ramdas [37].

A p-variable  $P$  is *semi-precise* for  $H$  if  $\sup_{Q \in H} Q(P \leq \alpha) = \alpha$  for each  $\alpha \in (0, 1/2]$ . Semi-precise p-variables require the sharp probability bound  $\sup_{Q \in H} Q(P \leq \alpha) = \alpha$  only for the case  $\alpha \leq 1/2$  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 semi-precise 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 and Hochberg [2]; this setting is particularly relevant for the procedure of Wang and Ramdas [34] 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. [17] 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 and Wang [29, 30] and the references therein.

E-values are relevant for all of the three contexts, and p-values are relevant for the second and the third contexts.

We will 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 or equal to  $\mu$  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 and Yekutieli [3] for the BH (Benjamini and Hochberg [2]) procedure with positive dependence and Chi et al. [6] for BH 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 a 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  $Q$ , denote by

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

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  $\inf_{Q \in H} T_P^Q(\alpha) \geq \alpha$  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 semi-precise p-variable if and only if  $\inf_{Q \in H} T_P^Q(\alpha) = \alpha$  for all  $\alpha \in (0, 1/2)$  and  $\inf_{Q \in H} T_P^Q(\alpha) \geq \alpha$  for  $\alpha \in [1/2, 1)$ .*

The proof of Lemma 2 is essentially identical to that of Lemma 1 of Vovk and Wang [29], which gives the equivalence between probability statements and quantile statements for merging functions of p-values. Our construction for precise and semi-precise 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 by  $X$ . We consider the following four classes of non-parametric composite hypotheses, where  $\mu \in \mathbb{R}$  and  $\sigma > 0$ .

$$\begin{aligned} H(\mu, \sigma) &= \{Q : \mathbb{E}^Q[X] \leq \mu \text{ and } \text{var}^Q(X) \leq \sigma^2\}; \\ H_S(\mu, \sigma) &= \{Q \in H(\mu, \sigma) : X \text{ is symmetrically distributed}\}; \\ H_U(\mu, \sigma) &= \{Q \in H(\mu, \sigma) : X \text{ is unimodally distributed}\}; \\ H_{US}(\mu, \sigma) &= H_U(\mu, \sigma) \cap H_S(\mu, \sigma). \end{aligned}$$

For our main results on the “best” p-variables and e-variables, it will be clear from our proofs that the condition  $\text{var}^Q(X) \leq \sigma^2$  in each hypothesis can be replaced by  $\text{var}^Q(X) = \sigma^2$ , and the condition  $\mathbb{E}^Q[X] \leq \mu$  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 Section 4.3.

The above four sets of distributions are studied in a very different context by Li et al. [21] 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. [21] and Bernard et al. [4] for finding bounds on 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 consequence of Cantelli's inequality. It may be interesting to compare  $P$  and  $1/E$  obtained from Theorem 1. Note that any e-variable can be converted into a p-variable via the so-called calibrator  $e \mapsto \min\{1/e, 1\}$ ; see e.g., Vovk and Wang [30]; this is an immediate consequence of Markov's inequality. As  $1/E$  is a p-variable for an e-variable  $E$ , we have  $P \leq 1/E$ . In Theorem 1, we obtain  $1/P = 1 + E > E$ , as expected.

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

$$P_0 = (1 + (X - \mu)_+^2/\sigma^2)^{-1} \quad \text{and} \quad E_0 = (X - \mu)_+^2/\sigma^2, \quad (4)$$

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

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

**Theorem 2.** *A semi-precise 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 two 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 by  $P_S$ , is improved from  $P_0$  by taking an extra minimum with  $1/E_S$ . In the most relevant case that  $P_0 \leq 1/2$ , or equivalently,  $E_0 \geq 1$ , indicating some evidence against the null, we have  $P_S = 1/E_S$ .

Next, we will see that the 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. [4], which gives

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

and applying Lemma 2 by inverting of the above curve as a function of  $\alpha$ .

Finally, we consider the hypothesis  $H_{US}(\mu, \sigma)$  of unimodal-symmetric distributions. This class includes, for instance, the popular normal, t-, and Laplace distributions. To construct a semi-precise p-variable for this hypothesis, we will use the following lemma of quantile bounds within  $H_{US}(\mu, \sigma)$ , 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*

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

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

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

**Theorem 4.** *A semi-precise p-variable for  $H_{US}(\mu, \sigma)$  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  $H_{US}(\mu, \sigma)$  is  $E = 2E_0$ . Precise p-variables do not exist for  $H_{US}(\mu, \sigma)$ .

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. [21] for  $\alpha \leq 1/6$ , 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  $X \leq \mu$ , in which case they are equal to 1. For  $(X - \mu)/\sigma \geq (5/3)^{1/2}$ , or equivalently,  $P_0 \leq 3/8$ , we have the following simple relation:

$$P_S = \frac{P_0}{2(1 - P_0)}, \quad P_U = \frac{4}{9}P_0, \quad \text{and} \quad P_{US} = \frac{2P_0}{9(1 - P_0)},$$

Table 1: Formulas for p-variables and e-variables

Hypothesis	p-variable		e-variable
$H(0, 1)$	$(1 + X_+^2)^{-1}$		$X_+^2$
$H_S(0, 1)$	$\frac{1}{2}X^{-2}$	if $X \geq 1$	$2X_+^2$
	$(1 + X_+^2)^{-1}$	if $X < 1$	
$H_U(0, 1)$	$\frac{4}{9}(1 + X^2)^{-1}$	if $X \geq (5/3)^{1/2}$	$X_+^2$
	$\frac{4}{3}(1 + X_+^2)^{-1} - \frac{1}{3}$	if $X < (5/3)^{1/2}$	
$H_{US}(0, 1)$	$\frac{2}{9}X^{-2}$	if $X \geq (4/3)^{1/2}$	$2X_+^2$
	$\frac{1}{2} - \frac{3^{1/2}}{6}X$	if $0 < X < (4/3)^{1/2}$	
	1	if $X \leq 0$	

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

$$P_0 = \frac{1}{10} = 0.1, \quad P_S = \frac{1}{18} \approx 0.056, \quad P_U = \frac{2}{45} \approx 0.044, \quad \text{and} \quad P_{US} = \frac{2}{81} \approx 0.025.$$

On the other hand, the corresponding e-values are

$$E_0 = 9, \quad E_S = 18, \quad E_U = 9, \quad \text{and} \quad E_{US} = 18.$$

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 0.00135 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  $(\mu, \sigma)$  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 holds true if the conditions  $\mathbb{E}^Q[X] \leq \mu$  and  $\text{var}^Q(X) \leq \sigma^2$  in each hypothesis is replaced by  $\mathbb{E}^Q[X] = \mu$  and  $\text{var}^Q(X) = \sigma^2$ , 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  $\text{var}^Q(X) = \sigma^2$  on top of  $\text{var}^Q(X) \leq \sigma^2$ , or  $\mathbb{E}^Q[X] = \mu$  on top of  $\mathbb{E}^Q[X] \leq \mu$ , 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  $g : \alpha \mapsto \sup_{Q \in H} T_X^Q(1 - \alpha)$ . If  $g$  is strictly decreasing on  $(0, 1)$ , as in the case of  $H(\mu, \sigma)$  and  $H_U(\mu, \sigma)$ , then choosing  $f = g^{-1}$  yields a precise p-variable  $f(X)$ . For  $H$  being  $H_S(\mu, \sigma)$  and  $H_{US}(\mu, \sigma)$ ,  $g$  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 Section 3. Section 4.1 describes the main methodology based on e-processes for the one-sided testing problem; Section 4.2 describes a few other methods using our results in Section 3; and Section 4.3 discusses the two-sided testing problem on the mean with given variance.

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

$$\begin{aligned} H(\mu, \sigma) &= \{Q : \mathbb{E}^Q[X_i | \mathcal{F}_{i-1}] \leq \mu \text{ and } \text{var}^Q(X_i | \mathcal{F}_{i-1}) \leq \sigma^2 \text{ for } i \in [n]\}; \\ H_S(\mu, \sigma) &= \{Q \in H(\mu, \sigma) : X_i | \mathcal{F}_{i-1} \text{ is symmetrically distributed for } i \in [n]\}; \\ H_U(\mu, \sigma) &= \{Q \in H(\mu, \sigma) : X_i | \mathcal{F}_{i-1} \text{ is unimodally distributed for } i \in [n]\}; \\ H_{US}(\mu, \sigma) &= H_U(\mu, \sigma) \cap H_S(\mu, \sigma). \end{aligned}$$

Recall that it is without loss of generality to consider  $\mu$  and  $\sigma^2$  as constants. We can also test the hypotheses where some data are symmetric or unimodal and some are not, because we will build e-values from each of them separately. For simplicity, we only list the above four representative cases. Using a similar formulation, the hypothesis in Waudby-Smith and Ramdas [37] is

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

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 results in Section 3 to construct an e-variable or p-variable for the above hypotheses; some of these methods are more useful than the 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

$$M_t = \prod_{i=1}^t (1 - \lambda_i + \lambda_i E_i), \quad (5)$$

where  $\lambda_i$  is  $\mathcal{F}_{i-1}$ -measurable and takes values in  $[0, 1)$ . This idea is the main methodology behind game-theoretic statistics; see Shafer [26], Shafer and Vovk

[27], and Waudby-Smith and Ramdas [37, Proposition 3]. It has been used by Waudby-Smith and Ramdas [37] for testing the mean and Wang et al. [33] for testing risk measures. To find good choices of  $\lambda = (\lambda_i)_{i \in [n]}$  is a non-trivial task. We propose to specify  $\lambda$  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 (5) over these choices to get an e-process. An uninformative choice of the values of  $\lambda$  may be some points in  $[0, 0.2]$ . We avoid choosing  $\lambda$  close to 1 because our e-value may take the value 0 with substantial probability, leading 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. [10] and Waudby-Smith and Ramdas [37]. In our simulation and empirical studies, we average over  $\lambda = 0.01 \times \{1, \dots, 20\}$ .
- (b) **E-GREE method:** In the GREE (growth-rate for empirical e-statistics) method of Wang et al. [33] for  $\lambda_i$ ,  $i \in [n]$  in (5),  $\lambda_i$  is determined by solving the following optimization problem:

$$\lambda_i = \left( \arg \max_{\lambda \in [0, 1)} \frac{1}{i-1} \sum_{j=1}^{i-1} \log(1 - \lambda + \lambda E_j) \right) \wedge \frac{1}{2}. \quad (6)$$

To simplify the maximization in (6), a fast and approximate solution can be obtained using Taylor expansion as in Waudby-Smith and Ramdas [37]. This leads to the following simple formula

$$\lambda_i = \left( \frac{\sum_{j=1}^{i-1} (E_j - 1)}{\sum_{j=1}^{i-1} (E_j - 1)^2} \right)_+ \wedge \frac{1}{2}. \quad (7)$$

We will use (7) for all e-GREE related calculations for the following results. Our unreported simulation suggests that using (6) and using (7) yield very similar results.

When the hypothesis to test is  $H_{\text{WRS}}(\mu)$ , the e-GREE method reduces to the method of Waudby-Smith and Ramdas [37]; see Section 5.2. An optimization procedure related to (6) is studied by Kumon et al. [19].

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  $M_t \geq 1/\alpha$  for the first time. The Type-I error control is guaranteed by Ville's inequality (Ville [28]) as  $\mathbb{P}(\sup_{t \in [n]} M_t \geq 1/\alpha) \leq \alpha$ , because any e-process is almost surely upper bounded by nonnegative supermartingales with initial value one; see Ramdas et al. [23].

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

**Proposition 1.** *Suppose that data are iid and generated from an alternative probability  $Q$ . The e-GREE method has asymptotic power approaching 1 as*

$n \rightarrow \infty$ , that is,  $Q(\sup_{t \in [n]} M_t \geq 1/\alpha) \rightarrow 1$  for any  $\alpha \in (0, 1)$  if and only if  $\mathbb{E}^Q[E_1] > 1$ .

Although Proposition 1 requires an iid assumption, this assumption is not needed for consistency in practical situations; a simulation example is in Section 5.1.

## 4.2 Some other methods

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

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

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

where  $\chi_{2n}$  is the cdf 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 [25],

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

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

Although in general p-Fisher and p-Simes 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 iid; a proof of this is presented in the Supplementary Material.

Then next two methods use all data directly, and requires independence among  $X_1, \dots, X_n$ . 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  $\mu$  and variance at most  $\sigma^2/n$ . Moreover, symmetry of  $T$  follows from symmetry of  $X_1, \dots, X_n$ . Nevertheless,  $T$  is not necessarily unimodal even if  $X_1, \dots, X_n$  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  $H_U(\mu, \sigma)$  is

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

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

$$E_S = 2n(T - \mu)_+^2 / \sigma^2.$$

(f) **P-batch method:** A p-variable for  $H(\mu, \sigma)$  or  $H_U(\mu, \sigma)$  is

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

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

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

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

### 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

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

where  $\mu^L \leq \mu^U$  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. First, we note that the e-variable  $E$  given by

$$E = \frac{(X - \mu^U)_+^2 + (X - \mu^L)_-^2}{\sigma^2}, \quad (8)$$

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  $Q \in H(\mu^L, \mu^U, \sigma)$ ,

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

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

If  $\mu^L = \mu^U = \mu$ , then the e-variable in (8) 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  $\text{var}^Q(X) = \sigma^2$ ; this condition is useful to establish consistency in Proposition 1.

Following the same procedure in Section 4.1 using (8), we obtain e-processes for the two-sided problem  $H(\mu^L, \mu^U, \sigma)$ . Due to a smaller null hypothesis, this e-process is generally more powerful than the one in Section 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 iid data with mean  $\mu$ , thus very difficult to detect. The same challenge exists for other methods based on e-processes, such as that of Waudby-Smith and Ramdas [37].

*Remark 4.* Under the additional information of symmetry, the e-variable in (8) can be used, but it cannot be multiplied by two as in Theorem 2. In this case, an alternative way to take advantage of symmetry is to build two e-processes in Section 4.1: one to test  $\mathbb{E}[X_i|\mathcal{F}_{i-1}] \leq \mu^U$  and another one to test  $\mathbb{E}[-X_i|\mathcal{F}_{i-1}] \leq -\mu^L$ . 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 analyze the power of the e-variables. For a given e-variable  $E$ , its e-power, using the terminology of Vovk and Wang [32], for an alternative probability  $Q$  is defined as  $\mathbb{E}^Q[\log E]$ ; see Shafer [26] and Grünwald et al. [10] 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  $Q_\sigma$  is distributed as  $\sigma Z$ , where  $Z$  has a fixed distribution with mean 0 and variance 1 satisfying the null hypothesis, which can be one of  $H(0, 1)$ ,  $H_S(0, 1)$ ,  $H_U(0, 1)$  and  $H_{US}(0, 1)$ . Note that 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 will show that the e-power of each e-variable grows at a rate of  $\log \sigma$  as the alternative variance  $\sigma^2$  grows, regardless of the distribution of  $Z$ .

Let  $E$  be the e-variable computed based on  $X$  as in Section 3. Due to the construction of the e-process  $M$  in (5), 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 of in each multiplicative term in the e-process  $M$ .

**Proposition 2.** *Suppose  $p := Q_0(E \geq 1) > 0$ . For  $\sigma > 1$ ,*

$$(2p \log \sigma - \log 2)_+ \leq \Pi^{Q_\sigma} \leq 2 \log \sigma. \tag{9}$$

*Moreover,  $0 \leq \Pi^{Q_\sigma} - \Pi^{Q_\delta} \leq 2(\log \sigma - \log \delta)$  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  $\sigma^2$  is larger than 1. An additional negative term  $-\log 2$  in (9) is not surprising, because our conditions do not guarantee  $\Pi^{Q_\sigma} > 0$  for  $\sigma$  very close to 1. Below, we give an example to illustrate the sharpness of bounds in (9).

**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  $\Pi^{Q_\sigma}$  is approximately equivalent to  $\log \sigma$  for large  $\sigma$ , corresponding to the left side of (9) 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 non-parametric hypotheses in Section 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 Section 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:  $X_1, X_3, \dots$ , follow a normal distribution, and  $X_2, X_4, \dots$ , follow a Laplace distribution, with the same mean  $\nu$  and the same variance  $\eta^2$ .<sup>1</sup> We denote this data generating process as  $NL(\nu, \eta^2)$  with the null parameters being  $(\nu, \eta^2) = (0, 1)$ . We consider two alternatives: (1) Data generated from  $NL(0, \eta^2)$  where  $\eta > 1$ ; (2) Data generated from  $NL(\nu, 1)$  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  $E \geq 1/\alpha$  and  $P \leq \alpha$ , with  $\alpha = 0.05$ , for e-values and p-values, respectively.

For the e-mixture method, we experiment by averaging  $\lambda$  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  $\lambda_i$  is dynamically updated with different  $i \in [n]$  using the formula (6).

Figure 1 shows the rejection rates for all methods with data generated from  $NL(0, \eta^2)$  for  $\eta \in [1, 4]$ , and from  $NL(\nu, 1)$  for  $\nu \in [0, 1]$ . For alternative model  $NL(0, \eta^2)$ , we see that the e-mixture and the 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 alternative model  $NL(\nu, 1)$ , we observe that 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.

<sup>1</sup>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.



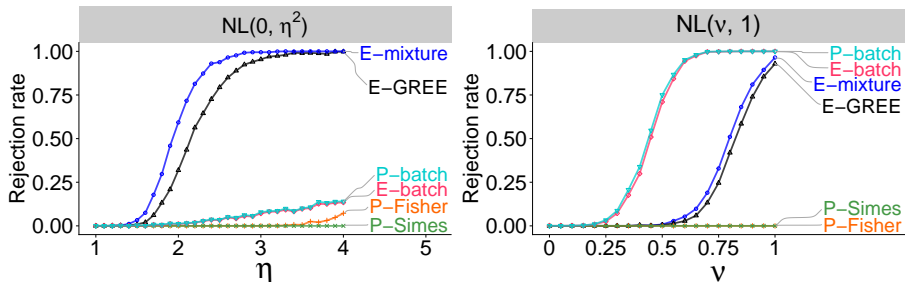


Figure 1: Rejection rates for all methods for testing  $H(0, 1)$  with sample size  $n = 100$  over 1000 runs using the threshold 20.

Among all methods, only the e-process based methods satisfy anytime-validity, that is, decision can be made at any stopping time when data arrive sequentially. This situation is common in financial applications, where realized losses accumulate over time; see the empirical study in Section 6.

The testing procedures for  $H_S$ ,  $H_U$  and  $H_{US}$  are the same as for testing  $H$ . We generate 100 data points from  $NL(0.5, 2)$  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.

Table 2: Rejection rates of testing  $H$ ,  $H_S$ ,  $H_U$  and  $H_{US}$  with  $n = 100$  data generated from the model  $NL(0.5, 2)$ .

	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

## 5.2 A comparison with the GRAPA method

Recall that our model can also be interpreted as testing the mean under the knowledge of an upper bound on the variance. This allows us to compare our testing approach with the GRAPA (Growth Rate Adaptive to the Particular Alternative) method proposed by Waudby-Smith and Ramdas [37]. GRAPA is similar to the e-GREE method discussed in Section 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 follows:

$$M_t = \prod_{i=1}^t (1 + \lambda_i (X_i - \mu)), \quad (10)$$

where  $\mu$  is the conditional mean being tested and  $\lambda_i$  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 (10) is similar to (6), where  $\lambda_i$  is determined by solving the following optimization problem:

$$\lambda_i = \arg \max_{\lambda \in [-c/(1-\mu), c/\mu]} \frac{1}{i-1} \sum_{j=1}^{i-1} \log(1 + \lambda(X_j - \mu)), \quad (11)$$

where  $c \in (0, 1]$  is fixed. For faster computation in the context of confidence sequences, Waudby-Smith and Ramdas [37] also offered an alternative way to obtain  $\lambda_i$ , which they called approximate GRAPA method, and  $\lambda_i$  is determined by

$$\lambda_i = -\frac{c}{1-\mu} \vee \frac{\hat{\mu}_{i-1} - \mu}{\hat{\sigma}_{i-1}^2 + (\hat{\mu}_{i-1} - \mu)^2} \wedge \frac{c}{\mu}, \quad (12)$$

where  $\hat{\mu}_i$  and  $\hat{\sigma}_i^2$  are empirical mean and variance of the observations  $X_1 \cdots, X_i$ . From (12), it is clear that the GRAPA method is able to use the sample variance information adaptively. In particular, our e-GREE method in (7) is adaptive to the empirical variance of the e-values. In the simulation results, we use (11) and choose  $c = 1/2$ .

We compare 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  $\sigma^2$ .
- (c) E-mixture: The e-mixture method with the variance upper bound  $\sigma^2$ .
- (d) E-GREE-2s: The two-sided e-GREE method with the variance upper bound  $\sigma^2$ .
- (e) E-mixture-2s: The two-sided e-mixture method with the variance upper bound  $\sigma^2$ .

We note that GRAPA is designed as a two-sided test, although it can easily be adjusted by restricting  $\lambda_i$  in (10) to be non-negative.

*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 their 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 by  $\text{Beta}(\nu, \sigma^2)$ , where  $\nu$  and  $\sigma^2$  represent the mean and variance of the beta distribution.<sup>2</sup> Here, we use  $\nu$  and  $\sigma^2$  instead of the

<sup>2</sup>None of the methods requires that the data follow identical distributions; we use a single distribution just for simplicity.

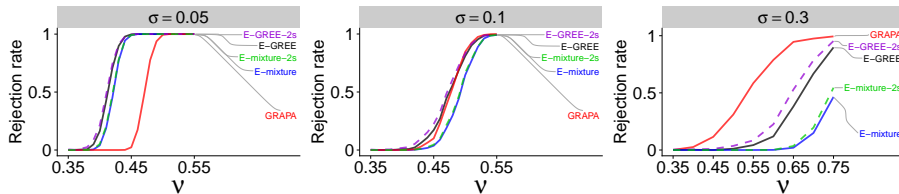


Figure 2: Rejection rates for the GRAPA, the e-GREE, the e-mixture and the two-sided e-GREE-2s and the e-mixture-2s methods over 1000 runs using the threshold 20 and  $\mu = 0.35$ . Data are generated from  $\text{Beta}(\nu, \sigma^2)$  with sample size  $n = 20$ , where  $\nu \geq 0.35$  and  $\sigma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

standard beta parameters  $\alpha$  and  $\beta$  for the sake of convenience. Note that the parameters  $\alpha$  and  $\beta$  can be easily recovered based on given mean  $\nu$  and variance  $\sigma^2$ :  $\alpha = \nu(\nu - \nu^2 - \sigma^2)/\sigma^2$  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  $\nu$  with fixed  $\sigma^2$ . We consider  $\nu \geq 0.35$  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 case  $\sigma = 0.05$  and  $\sigma = 0.1$ , the e-GREE method outperforms the GRAPA method; in case  $\sigma = 0.3$ , the GRAPA method demonstrates superior performance compared to the other methods. This is intuitive, because the variance information is less useful for larger  $\sigma$ ; recall that for any distribution supported in  $[0, 1]$  with mean  $\mu \leq 0.35$ , 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  $\sigma$  is relatively small.

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

### 5.3 A comparison with exponential test supermartingale

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.

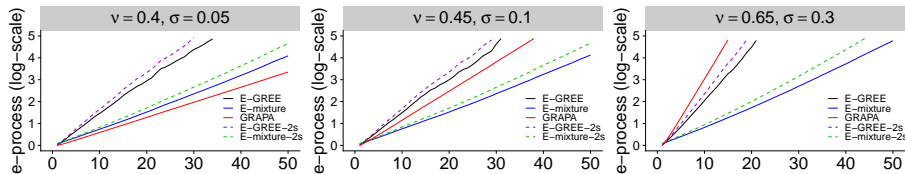


Figure 3: Average logarithmic e-processes for the GRAPA, the e-GREE, the e-mixture and the two-sided e-GREE-2s and the e-mixture-2s methods with varying sample size and  $\mu = 0.35$ . Data are generated from Beta( $\nu, \sigma^2$ ) where  $\sigma \in \{0.05, 0.1, 0.3\}$  and  $\nu = \mu + \sigma$ .

Wang and Ramdas [35] extends the idea from Catoni [5] to construct a non-negative test supermartingale called the Catoni supermartingale to test mean and variance in sequential settings. The test supermartingale is constructed as follows:

$$M_t^C = \prod_{i=1}^t \exp \left( \phi(\lambda_i(X_i - \mu)) - \frac{\lambda_i^2 \sigma^2}{2} \right), \quad (13)$$

where  $\phi$  is the influence function and  $(\lambda_i)_{i \in [n]}$  is any predictable process. Following the recommendation of Wang and Ramdas [35], we choose the influence function

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

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

$$\lambda_i = \left( \frac{2 \log(1/\alpha)}{i(\sigma^2 + \eta_i^2)} \right)^{1/2} \quad \text{where} \quad \eta_i = \left( \frac{2\sigma^2 \log(1/\alpha)}{i - 2 \log(1/\alpha)} \right)^{1/2}. \quad (14)$$

A different approach by Howard et al. [16] is to use a framework for non-parametric confidence sequences based on the concept of exponential supermartingales. They introduce the concept of a “sub- $\psi$  process” in Howard et al. [16, Definition 1]. Informally, a sub- $\psi$  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  $\psi$  make the following process

$$M_t^\psi = \exp\{\lambda S_t - \psi(\lambda)V_t\} \quad (15)$$

dominated by a supermartingale for each  $\lambda$  in an interval  $[0, \lambda_{\max})$ . This framework allows for testing mean and variance under a wide variety of assumptions, including bounded supports, self-normalized bounds, and symmetric conditions. We refer to Howard et al. [16, Appendix J, Table 3] for a collection of commonly used  $\psi$  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 bounds test supermartingale, denoted by  $M_t^{\psi, \text{SN}}$ , and the

symmetric condition test supermartingale, denoted by  $M_t^{\psi, \text{sym}}$ . For  $\lambda \in [0, \infty)$ , these test supermartingales are constructed as follows:

$$M_t^{\psi, \text{SN}} = \prod_{i=1}^t \exp \left( \lambda(X_i - \mu) - \frac{\lambda^2(X_i - \mu)^2 + 2\sigma^2}{6} \right), \quad (16)$$

which also appears in Wang and Ramdas [35, Section 5], and

$$M_t^{\psi, \text{sym}} = \prod_{i=1}^t \exp \left( \lambda(X_i - \mu) - \frac{\lambda^2(X_i - \mu)^2}{2} \right). \quad (17)$$

We follow a simple method of choosing  $\lambda$  suggested by Howard et al. [16, Section 3.2], that is, to use the mixture supermartingale  $\int \exp(\lambda S_t - \psi(\lambda) V_t) d\Phi(\lambda)$  by assuming  $\lambda \sim \Phi = N(0, 1)$ . Now, we further compare the following methods:

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

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

For  $NL(\nu, 1^2)$ , the Catoni method outperforms other methods, while methods utilizing symmetric information generally perform well. For  $NL(\nu, (1 + \nu)^2)$ , where both the mean and variance of the data generating process change, the power of methods from Howard et al. [16] 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,  $NL(\nu/5, (1 + \nu)^2)$ , the impact of changes in mean is small and the variance effect is large, e-value based methods generally outperform others. Although method (h) benefits from not requiring information about variance or even the existence of variance, it demonstrates minimal power when testing mean with varying variance, due to its penalization term  $-(X_i - \mu)^2$  in the exponential form of (16) and (17). In summary, our methods are comparatively more powerful when the alternative variance defers from the null.

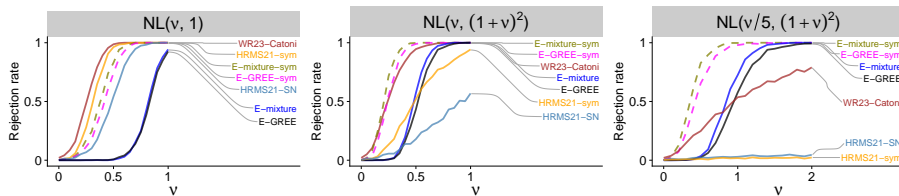


Figure 4: Rejection rates for methods (a), (b) and (f)-(j) for testing  $H(0, 1)$  with sample size  $n = 100$  over 1000 runs using the threshold 20.

## 6 Empirical study with financial data

In this section, we 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 2007–2008 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 thresholds 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 the large market capitalization in each sector. Moreover, we include two companies with the largest market capitalization from the to-be Real Estate sector.<sup>3</sup> We first calculate the daily losses for each of the selected stocks from January 1, 2001 to December 31, 2010. The daily losses are expressed by percentage and calculate by  $L_t = -(S_{t+1} - S_t)/S_t$ , where  $S_t$  is the close price at day  $t$ . Note that the positive value represents a loss and 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 January 1, 2001 to December 31, 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 (5) as the daily loss from January 1, 2007 fed into the e-process.

Following a methodology similar to the simulation study in Section 5, we report the evidence against the null hypothesis when the e-process exceeds thresholds of 2, 5, 10, and 20.<sup>4</sup> 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. It is important to note that, although a threshold of 2 may not be substantial enough to reject the null hypothesis, it

<sup>3</sup>Real Estate becomes the 11th sector of S&P500 in 2016.

<sup>4</sup>In accordance with Jeffrey’s rule of thumb about e-values (see Jeffrey [18] and Vovk and Wang [30]), if the e-value falls within the interval of  $(10^{1/2}, 10)$ , the evidence against the null hypothesis is considered substantial; If the e-value falls within the interval of  $(10, 10^{3/2})$ , the evidence against the null hypothesis is regarded as strong.

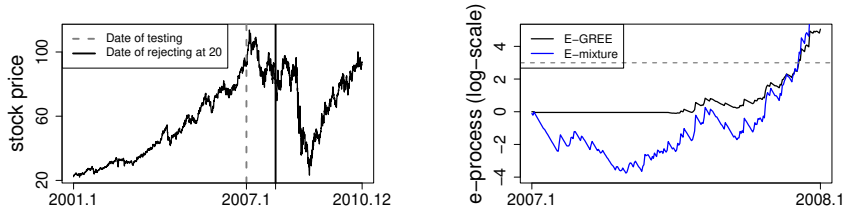


Figure 5: Sample path and logarithmic e-process using the e-GREE and the e-mixture methods testing of  $H(\hat{\mu}, \hat{\sigma})$  for Simon Property (SPG) stock from January 2007 to January 2008, where  $\hat{\mu} = -0.001028$  and  $\hat{\sigma} = 0.012123$  are the sample mean and variance estimated from historical data for stock SPG from January 1, 2001 to December 31, 2006.

can still serve as an early warning that the stock’s performance may be different from its historical path.

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 January 1, 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 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 e-batch and other p-variable based methods stated in Section 4, 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 2007–2008 subprime crisis, such as Financials and Consumer Discretionary, 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

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 January 1, 2007 to December 31, 2010; “-” means no detection is observed till December 31, 2010.

	Threshold	E-GREE				E-mixture			
		2	5	10	20	2	5	10	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

increase during the financial crisis.

## 7 Discussion

This paper proposes an e-process based approach for testing mean and variance from non-stationary data. We consider four classes of non-parametric composite hypotheses with specified mean and variance bound along with additional constraints of distribution, such as symmetry, unimodality, or a combination thereof. For this purpose, our main technical results give the best p-variables and e-variables in the simple setting where one summary data point is observed. The explicit formulas are summarized in Table 1. Using the obtained e-variables, we construct an e-process using either the e-mixture method or the e-GREE method. Simulation studies and empirical analysis are conducted to show the performance of the proposed methods in comparison with GRAPA of Waudby-Smith and Ramdas [37] and with the exponential supermartingale methods of Howard et al. [15, 16] and Wang and Ramdas [35].

As mentioned in Section 3, 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 and Barber [24] for derandomization, Ahn et al. [1] for Bayesian linear models, and Gablenz and Sabatti [9] for resolution-adaptive variable selection. Finally, the obtained e-variables may also be useful to build e-confidence regions (see Vovk and Wang [31]) and e-posterior as (see Grünwald [11]) for  $(\mu, \sigma^2)$ , although we mainly consider a non-parametric setting.

## 8 Proofs of all results

We collect all proofs in the paper in this section.

*Proof of Lemma 1.* Let  $\mathcal{P}$  be any collection of p-variables for  $H$ . For  $Q \in H$ , using the fact that the elements of  $\mathcal{P}$  are comonotonic, we have

$$Q(\inf\{P \in \mathcal{P}\} > \alpha) = Q\left(\bigcap_{P \in \mathcal{P}} \{P > \alpha\}\right) = \inf_{P \in \mathcal{P}} Q(P > \alpha) \geq 1 - \alpha.$$

This implies

$$Q(\inf\{P \in \mathcal{P}\} \leq \alpha) \leq \alpha.$$

Hence, the infimum of all p-variables for  $H$  is still a p-variable, which is the smallest one.  $\square$

For all theorems below, we will prove precision statements for the formulation of  $\mathbb{E}^Q[X] = \mu$  instead of  $\mathbb{E}^Q[X] \leq \mu$ , making these statements stronger. For the validity statements, it is easy to verify that those p-variables and e-variables are valid under both formulations.

*Proof of Theorem 1.* Since the problem is invariant under location shift and scaling, it suffices to consider the normalized case of  $(\mu, \sigma) = (0, 1)$ .

It is clear that  $P$  is decreasing in  $X$  and  $E$  is increasing in  $X$ .

For  $Q \in H(0, 1)$ , Cantelli's inequality implies  $Q(X > x) \leq 1/(1 + x^2)$  for  $x > 0$ , which implies, for each  $\alpha \in (0, 1)$ ,

$$Q(P \leq \alpha) = Q(1 + X_+^2 \geq 1/\alpha) = Q\left(X \geq \sqrt{(1 - \alpha)/\alpha}\right) \leq \frac{1}{1 + 1/\alpha - 1} = \alpha.$$

The inequality above is an equality if  $Q$  is chosen such that

$$Q\left(X = \sqrt{(1 - \alpha)/\alpha}\right) = \alpha = 1 - Q\left(X = -\sqrt{\alpha/(1 - \alpha)}\right), \quad (18)$$

and we can easily verify that  $\mathbb{E}^Q[X] = 0$  and  $\text{var}^Q(X) = 1$ . This implies that  $\sup_{Q \in H(0, 1)} Q(P \leq \alpha) = \alpha$  for each  $\alpha \in (0, 1)$ , and therefore  $P = 1/(1 + X_+^2)$  is a precise p-variable for  $H(0, 1)$ .

For  $Q \in H(0, 1)$ , we have  $\mathbb{E}^Q[X_+^2] \leq \mathbb{E}^Q[X^2] \leq 1$ . To show that  $E$  is precise, let  $Q$  be given by (18), which satisfies  $\mathbb{E}^Q[X_+^2] = \alpha$ . By taking  $\alpha \uparrow 1$  we know  $\sup_{Q \in H(0, 1)} \mathbb{E}^Q[E] = 1$ , and therefore  $E = X_+^2$  is a precise e-variable for  $H(0, 1)$ .  $\square$

*Proof of Theorem 2.* We first show the statement on the e-variable. Set  $(\mu, \sigma) = (0, 1)$  as in the proof of Theorem 1. For  $Q \in H_S(0, 1)$ , we have  $2\mathbb{E}^Q[X_+^2] = \mathbb{E}^Q[X^2] \leq 1$ , with equal sign holding if  $\text{var}^Q(X) = 1$ . Therefore,  $E = 2X_+^2$  is a precise e-variable for  $H_S(0, 1)$ .

Since  $E = 2X_+^2$  is an e-variable, by Markov's inequality,  $1/E = (2X_+)^{-2}$  is a p-variable for  $H_S(0, 1)$ . In Theorem 1 we have seen that  $P_0$  is a p-variable for  $H(0, 1)$ , and hence also a p-variable for  $H_S(0, 1) \subseteq H(0, 1)$ . Using Lemma 1, the minimum of  $P_0$  and  $(2E_0)^{-1}$  is a p-variable for  $H_S(0, 1)$ .

Next, we show that  $P$  is semi-precise. For  $\alpha \in (0, 1/2]$ , let  $Q$  be chosen such that

$$Q\left(X = (2\alpha)^{-1/2}\right) = \alpha = Q\left(X = -(2\alpha)^{-1/2}\right) \text{ and } Q(X = 0) = 1 - 2\alpha.$$

We can verify that  $\mathbb{E}^Q[X] = 0$ ,  $\text{var}^Q(X) = 1$ , and  $X$  is symmetrically distributed. It follows that  $Q(P \leq \alpha) = Q(X = (2\alpha)^{-1/2}) = \alpha$ . This implies that  $\sup_{Q \in H_S(0,1)} Q(P \leq \alpha) = \alpha$  for  $\alpha \in (0, 1/2]$ . Therefore,  $P$  is a semi-precise p-variable for  $H_S(0, 1)$ .

Finally, we show that there do not exist precise p-variables for  $H_S(0, 1)$ . Suppose that  $P = f(X)$  is a precise p-variable, where  $f$  is a decreasing function. Note that  $Q(X \leq 0) \geq 1/2$  for all  $Q \in H_S(0, 1)$ . It follows that  $Q(P \geq f(0)) \geq 1/2$  and  $Q(P < f(0)) \leq 1/2$ . If  $f(0) > 1/2$ , then for  $\alpha \in [1/2, f(0)]$ ,  $Q(P \leq \alpha) \leq 1/2 < \alpha$ , implying that  $P$  is not precise. If  $f(0) \leq 1/2$ , then, by taking  $Q$  as the point-mass at 0, we have  $Q(P \leq 1/2) = 1$ , implying that  $P$  is not a p-variable. Either way we have a contradiction, and hence does not exist a precise p-variable.  $\square$

*Proof of Theorem 3.* Set  $(\mu, \sigma) = (0, 1)$  as in the proof of Theorem 1. By Theorem 1 of Bernard et al. [4],

$$\sup_{Q \in H_U(0,1)} T_X^Q(1 - \alpha) = \max \left\{ \sqrt{\frac{4}{9\alpha} - 1}, \sqrt{\frac{3 - 3\alpha}{1 + 3\alpha}} \right\} \text{ for } \alpha \in (0, 1). \quad (19)$$

Note that  $P$  is a decreasing function of  $X$ , and we denote this by  $P = f(X)$  where

$$f(x) = \max \left\{ \frac{4}{9}(1 + x_+^2)^{-1}, \frac{4}{3}(1 + x_+^2)^{-1} - \frac{1}{3} \right\}.$$

For  $\alpha \in (0, 1/6]$ , we have

$$\sup_{Q \in H_U(0,1)} T_X^Q(1 - \alpha) = \sqrt{\frac{4}{9\alpha} - 1},$$

and hence

$$\inf_{Q \in H_U(0,1)} T_P^Q(\alpha) = f \left( \sup_{Q \in H_U(0,1)} T_X^Q(1 - \alpha) \right) = \frac{4}{9} \left( 1 + \frac{4}{9\alpha} - 1 \right)^{-1} = \alpha.$$

For  $\alpha \in (1/6, 1)$ , it is

$$\sup_{Q \in H_U(0,1)} T_X^Q(1-\alpha) = \sqrt{\frac{3-3\alpha}{1+3\alpha}},$$

and hence

$$\inf_{Q \in H_U(0,1)} T_P^Q(\alpha) = f\left(\sup_{Q \in H_U(0,1)} T_X^Q(1-\alpha)\right) = \frac{4}{3} \left(1 + \frac{3-3\alpha}{1+3\alpha}\right)^{-1} - \frac{1}{3} = \alpha.$$

Using Lemma 2, we obtain that  $P$  is a precise p-variable for  $H_U(0,1)$ .

As  $E$  is an e-variable for  $H(0,1)$ , it is also an e-variable for  $H_U(0,1)$ . To show that it is precise, fix any  $a \in (0,1)$ , and let  $p > 0$  and  $b > 0$  satisfy

$$a^2 = \frac{3-3p}{3p+2-p^2} \quad \text{and} \quad b = \frac{1+p}{1-p}a.$$

Note that such  $p$  exists for any  $a \in (0,1)$  since the range of  $(3-3p)/(3p+2-p^2)$  covers  $(0,1)$ . Choose  $Q$  such that the distribution of  $X$  has a point-mass at  $-a$  with probability  $p$  and a uniform density on  $[-a,b]$ . We can compute

$$\mathbb{E}^Q[X] = -ap + \frac{b-a}{2}(1-p) = -ap + ap = 0,$$

and

$$\mathbb{E}^Q[X^2] = a^2p + \frac{a^2}{3}(1-p) + \frac{b^2}{3}(1-p) = \frac{a^2(3p+2-p^2)}{3(1-p)} = 1.$$

Therefore  $Q \in H_U(0,1)$ . We also have

$$\mathbb{E}^Q[E] = \mathbb{E}^Q[X_+^2] = 1 - a^2p - \frac{a^2}{3}(1-p) \geq 1 - a^2.$$

Since  $a \in (0,1)$  is arbitrary, we get  $\sup_{Q \in H_U(0,1)} \mathbb{E}^Q[E] = 1$ , and hence  $E$  is a precise e-variable.  $\square$

*Proof of Lemma 3.* For  $\alpha \geq 1/2$ , since  $Q \in H_{US}(0,1)$  is symmetric about 0, we have  $T_X^Q(1-\alpha) \leq 0$ , with  $T_X^Q(1-\alpha) = 0$  if  $Q$  is the point-mass at 0. We assume  $\alpha < 1/2$  below.

Take any  $Q \in H_{US}(0,1)$ , and we will find another distribution  $R$  with smaller variance and the same  $\alpha$ -quantile (we omit “left” because the quantile is unique for  $Q$  and  $R$ ). Note that  $Q$  has a decreasing density on  $(0, \infty)$  and possibly a point-mass at 0. Denote by  $x_0 = T_X^Q(1-\alpha)$  and  $g$  the density function of  $Q$  on  $(0, \infty)$ . Consider a different distribution  $R$  symmetric with respect to 0 which has uniform density equal to  $g(x_0)$  on  $(0, b)$  for some  $b > x_0$  and a point-mass at 0, such that  $R([x_0, b]) = \alpha = Q([x_0, \infty)) = R([x_0, \infty))$ . Denote by  $h$  the density function of  $R$  on  $(0, \infty)$ , and note that  $h(x) = 0$  for  $x > b$ . Since  $Q$  has

a decreasing density  $g$  on  $(0, \infty)$ ,  $g \geq h$  on  $(0, x_0)$  and  $g \leq h$  on  $(x_0, b)$ . The above conditions imply

$$\int_0^{x_0} x^2 g(x) dx \geq \int_0^{x_0} x^2 h(x) dx \quad \text{and} \quad \int_{x_0}^{\infty} x^2 g(x) dx \geq \int_{x_0}^{\infty} x^2 h(x) dx, \quad (20)$$

where the second inequality is due to  $R([x_0, \infty)) = Q([x_0, \infty))$ . Note that both inequalities in (20) are equalities if and only if  $g = h$ , and equivalently,  $Q = R$ . It follows that  $\mathbb{E}^Q[X^2] \geq \mathbb{E}^R[X^2]$ , and hence  $R \in H_{\text{US}}(0, 1)$ . Note that the condition  $Q([x_0, \infty)) = \alpha = R([x_0, \infty))$  guarantees  $T_X^Q(1-\alpha) = x_0 = T_X^R(1-\alpha)$ ; that is  $R$  has the same  $\alpha$ -quantile as  $G$ .

The above argument shows that it suffices for us to consider distributions  $Q$  which can be represented by a mixture of point-mass at 0 and a uniform distribution on  $[-b, b]$ . We also assume that  $Q$  has variance 1; if the variance is less than 1, then a rescaled distribution from  $Q$  has variance 1 and a larger  $\alpha$ -quantile. Let  $p = Q((0, \infty)) \in (0, 1/2]$ . We can compute  $\mathbb{E}^Q[X^2] = 2pb^2/3 = 1$ , and hence  $b = 3^{1/2}(2p)^{-1/2}$ . This gives

$$T_X^Q(1-\alpha) = b(1-\alpha/p) = \sqrt{\frac{3}{p}} \left(1 - \frac{\alpha}{p}\right).$$

Maximizing the above term over  $p \in (0, 1/2]$  gives  $p = 3\alpha$  if  $\alpha \leq 1/6$  and  $p = 1/2$  if  $\alpha \in (1/6, 1/2]$ , showing the desired supremum formula in the lemma.  $\square$

*Proof of Theorem 4.* Set  $(\mu, \sigma) = (0, 1)$  as in the proof of Theorem 1. By Theorem 3,  $E = 2E_0$  is an e-variable for  $H_{\text{US}}(0, 1)$ . It is precise because  $\mathbb{E}^Q[2X_+^2] = 1$  for any  $Q \in H_{\text{US}}(0, 1)$  with  $\text{var}^Q(X) = 1$ .

The fact that precise p-variables do not exist for  $H_{\text{US}}(0, 1)$  follows from the same argument as in the proof of the corresponding statement in Theorem 2.

It remains to show that  $P$  is a semi-precise p-variable for  $H_{\text{US}}(0, 1)$ . Write  $P = f(X)$  where

$$f(x) = \frac{2}{9x^2} \mathbf{1}_{[4/3, \infty)}(x_+^2) + \frac{3 - \sqrt{3}x}{6} \mathbf{1}_{(0, 4/3)}(x_+^2) + \mathbf{1}_{(-\infty, 0]}(x).$$

Using Lemma 3, for  $\alpha \in (0, 1/6]$ , we have

$$\sup_{Q \in H_{\text{US}}(0, 1)} T_X^Q(1-\alpha) = \sqrt{\frac{2}{9\alpha}} \geq \sqrt{\frac{4}{3}}$$

and

$$\inf_{Q \in H_{\text{US}}(0, 1)} T_P^Q(\alpha) = f \left( \sup_{Q \in H_{\text{US}}(0, 1)} T_X^Q(1-\alpha) \right) = \frac{2}{9} \times \frac{9\alpha}{2} = \alpha.$$

Similarly, for  $\alpha \in (1/6, 1/2)$ , we have

$$\sup_{Q \in H_{\text{US}}(0, 1)} T_X^Q(1-\alpha) = \sqrt{3}(1-2\alpha) \in (0, \sqrt{4/3})$$

and

$$\inf_{Q \in H_{\text{US}}(0,1)} T_P^Q(\alpha) = f \left( \sup_{Q \in H_{\text{US}}(0,1)} T_X^Q(1 - \alpha) \right) = \frac{3 - 3(1 - 2\alpha)}{6} = \alpha.$$

Finally, for  $\alpha \in [1/2, 1)$ , we have  $\inf_{Q \in H_{\text{US}}(0,1)} T_P^Q(\alpha) = 1$  since  $\mathbb{P}(X \leq 0) \geq 1/2$ . Using Lemma 2, the above three cases together imply that  $P$  is a semi-precise p-variable for  $H_{\text{US}}(0, 1)$ .  $\square$

*Proof of a statement in Section 4.2.* Here we show that p-Simes and p-Fisher can be applied to conditionally valid p-values. Assume  $\mathbb{P}(P_t \leq \alpha | \mathcal{F}_{t-1}) \leq \alpha$  for each  $t = 1, \dots, n$  and  $\alpha \in (0, 1)$  under  $H_0$ . This implies that there exists  $\tilde{P}_t \leq P_t$  such that  $\mathbb{P}(\tilde{P}_t \leq \alpha | \mathcal{F}_{t-1}) = \alpha$  for all  $\alpha \in (0, 1)$ . Hence,  $\tilde{P}_1, \dots, \tilde{P}_n$  are iid. Applying the combination methods to  $\tilde{P}_1, \dots, \tilde{P}_n$  yields a valid Type-I error control. Since  $\tilde{P}_t \leq P_t$  for each  $t$  and the two combination methods are monotone, we also have a valid Type-I error control when combining  $P_1, \dots, P_n$ .  $\square$

*Proof of Proposition 1.* The assumption that data are iid implies that  $E_1, E_2, \dots$  are iid. The “only if” statement is trivial since  $\mathbb{E}^Q[E_1] \leq 1$  implies that  $(M_t)_{t \geq 1}$  is an e-process for  $Q$ , and hence  $Q(\sup_{t \in [n]} M_t \geq 1/\alpha) \leq \alpha$  for all  $n \in \mathbb{N}$ . Next we show the “if” statement. For this, we use Theorem 3 of Wang et al. [33], which states that, under the iid assumption,

$$\frac{1}{t} \left( \log M_T(\boldsymbol{\lambda}^{\text{GREE}}) - \log M_t(\boldsymbol{\lambda}^{\text{GRO}}) \right) \xrightarrow{L^1(Q)} 0 \quad \text{as } t \rightarrow \infty,$$

where  $M_t(\boldsymbol{\lambda}^{\text{GREE}})$  is given by (5) with each  $\lambda_i$  computed from the e-GREE method, and  $M_t(\boldsymbol{\lambda}^{\text{GRO}})$  is given by (5) with each  $\lambda_i$  given by its theoretically growth-rate optimal value

$$\lambda^* = \arg \max_{\lambda \in (0,1)} \mathbb{E}^Q[\log(1 - \lambda + \lambda E_1)],$$

and this gives

$$\frac{1}{t} \log M_t(\boldsymbol{\lambda}^{\text{GRO}}) = \max_{\lambda \in (0,1]} \mathbb{E}^Q[\log(1 - \lambda + \lambda E_1)].$$

Therefore, we have

$$\frac{1}{t} \log M_t \xrightarrow{Q} \max_{\lambda \in (0,1]} \mathbb{E}^Q[\log(1 - \lambda + \lambda E_1)] \quad \text{as } t \rightarrow \infty.$$

It remains to verify  $\max_{\lambda \in (0,1]} \mathbb{E}^Q[\log(1 - \lambda + \lambda E_1)] > 1$ . Note that  $\mathbb{E}[E_1] > 1$  implies  $\mathbb{E}[E_1 \wedge K] > 1$  for some  $K \geq 1$ . We denote by  $Y = E_1 \wedge K$ . Since  $\mathbb{E}[(Y - 1)_+] - \mathbb{E}[(Y - 1)_-] = \mathbb{E}[Y - 1] > 0$ , there exists some  $\epsilon \in (0, 1)$  such that

$$\frac{1}{1 + \epsilon} \mathbb{E}[(Y - 1)_+] - \frac{1}{1 - \epsilon} \mathbb{E}[(Y - 1)_-] > 0.$$

Note that  $\log(1+x) \geq x/(1+\epsilon)$  for  $x \in [0, \epsilon)$  and  $\log(1+x) \geq x/(1-\epsilon)$  for  $x \in (-\epsilon, 0)$ , that is,

$$\log(1+x) \geq \frac{x_+}{1+\epsilon} - \frac{x_-}{1-\epsilon} \quad \text{for } x \in (-\epsilon, \epsilon).$$

Hence, for  $\lambda \in (0, \epsilon/K)$ , implying  $\lambda(Y-1) \in (-\epsilon, \epsilon)$ , we have

$$\begin{aligned} \mathbb{E}[\log(1-\lambda+\lambda E_1)] &\geq \mathbb{E}[\log(1+\lambda(Y-1))] \\ &\geq \frac{1}{1+\epsilon} \mathbb{E}[\lambda(Y-1)_+] - \frac{1}{1-\epsilon} \mathbb{E}[\lambda(Y-1)_-] > 0, \end{aligned}$$

thus showing the desired inequality.  $\square$

*Proof of Proposition 2.* First, it is clear that  $\Pi^{Q_\sigma} \geq 0$  by choosing  $\lambda = 0$  in the supremum. Second, by Jensen's inequality, for  $\sigma > 1$ ,

$$\mathbb{E}^{Q_0}[\log(1-\lambda+\lambda\sigma^2 E)] \leq \log(1-\lambda+\lambda\sigma^2 \mathbb{E}^{Q_0}[E]) \leq \log(1-\lambda+\lambda\sigma^2) = 2 \log \sigma.$$

We next show  $\Pi^{Q_\sigma} \geq 2p \log \sigma - \log 2$ . Note that

$$\mathbb{E}^{Q_0}[\log(1-\lambda+\lambda\sigma^2 E)] \geq (1-p) \log(1-\lambda) + p \log(1-\lambda+\lambda\sigma^2).$$

Maximizing the above term over  $\lambda \in [0, 1]$ , the maximizer is  $\lambda^* = (p\sigma^2 - 1)/(\sigma^2 - 1)$ . The corresponding maximum value satisfies

$$\begin{aligned} (1-p) \log \frac{(1-p)\sigma^2}{\sigma^2 - 1} + p \log(p\sigma^2) &\geq (1-p) \log(1-p) + p \log p + p \log \sigma^2 \\ &\geq -\log 2 + p \log \sigma^2, \end{aligned}$$

where we used the fact that  $x \log x + (1-x) \log(1-x)$  on  $[0, 1]$  is maximized at  $x = 1/2$ . This shows  $\Pi^{Q_\sigma} \geq 2p \log \sigma - \log 2$ , completing the proof of (9).

Finally, we prove the last statement  $0 \leq \Pi^{Q_\sigma} - \Pi^{Q_\delta} \leq 2(\log \sigma - \log \delta)$  for  $\sigma > \delta > 1$ . For any  $\lambda \in [0, 1]$ , let  $\lambda' = \lambda\delta^2/\sigma^2 \in [0, 1]$ . We have

$$\Pi^{Q_\sigma} \geq \log(1-\lambda'+\lambda'\sigma^2 E) \geq \log(1-\lambda+\lambda\delta^2 E).$$

Taking a supremum over  $\lambda \in [0, 1]$  yields  $\Pi^{Q_\sigma} \geq \Pi^{Q_\delta}$ . To show the other inequality,

$$\begin{aligned} \Pi^{Q_\sigma} &\leq \sup_{\lambda \in [0, 1]} \log \left( \frac{\sigma^2}{\delta^2} (1-\lambda) + \lambda\sigma^2 E \right) \\ &= \log \frac{\sigma^2}{\delta^2} + \sup_{\lambda \in [0, 1]} \log(1-\lambda+\lambda\delta^2 E) = 2 \log \frac{\sigma}{\delta} + \Pi^{Q_\delta}. \end{aligned}$$

This gives  $\Pi^{Q_\sigma} - \Pi^{Q_\delta} \leq 2(\log \sigma - \log \delta)$  and completes the proof.  $\square$

## Acknowledgements

We thank the Editor, an Associate Editor, and two anonymous 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.

## References

- [1] Ahn, T., Lin, L. and Mei, S. (2023). Near-optimal multiple testing in Bayesian linear models with finite-sample FDR control. *arXiv:2211.02778*.
- [2] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**(1), 289–300.
- [3] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**(4), 1165–1188.
- [4] Bernard, C., Kazzi, R. and Vanduffel, S. (2020). Range value-at-risk bounds for unimodal distributions under partial information. *Insurance: Mathematics and Economics*, **94**, 9–24.
- [5] Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l’IHP Probabilités et statistiques*, Vol. 48, No. 4, pp. 1148–1185.
- [6] Chi, Z., Ramdas, A. and Wang, R. (2024). Multiple testing under negative dependence. *Bernoulli*, forthcoming.
- [7] Conover, W. J. (1999). *Practical Nonparametric Statistics*, John Wiley & Sons.
- [8] Fissler, T. and Ziegel, J. F. (2016). Higher order elicibility and Osband’s principle. *The Annals of Statistics*, **44**(4), 1680–1707.
- [9] Gablenz, P. and Sabatti, C. (2024). Catch me if you can: Signal localization with knockoff e-values. *Journal of the Royal Statistical Society Series B*, forthcoming.
- [10] Grünwald, P., de Heide, R. and Koolen, W. M. (2024). Safe testing. *Journal of the Royal Statistical Society Series B*, forthcoming.
- [11] Grünwald, P. (2023). The e-posterior. *Philosophical Transactions of the Royal Society A*, **381**(2247), 20220146.
- [12] Henzi, A. and Ziegel, J. F. (2022). Valid sequential inference on probability forecast performance. *Biometrika*, **109**(3), 647–663.

- [13] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, **58**(301), 13–30.
- [14] Hollander, M., Wolfe, D. A. and Chicken, E. (2013). *Nonparametric Statistical Methods*, John Wiley & Sons.
- [15] Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2020). Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, **17**, 257–317.
- [16] Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, **49**(2), 1055–1080.
- [17] Ignatiadis, N., Wang, R. and Ramdas, A. (2024). E-values as unnormalized weights in multiple testing. *Biometrika*, **111**(2), 417–439.
- [18] Jeffreys, H. (1998). *The Theory of Probability*, 3rd Ed. Oxford University Press.
- [19] Kumon, M., Takemura, A. and Takeuchi, K. (2011). Sequential optimizing strategy in multidimensional bounded forecasting games. *Stochastic Processes and their Applications*, **121**(1), 155–183.
- [20] Lehmann, E. L., Romano, J. P. and Casella, G. (1986). *Testing Statistical Hypotheses*, New York: Springer.
- [21] Li, L., Shao, H., Wang, R. and Yang, J. (2018). Worst-case Range Value-at-Risk with partial information. *SIAM Journal on Financial Mathematics*, **9**(1), 190–218.
- [22] Ramdas, A., Grünwald, P., Vovk, V. and Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, **38**(4), 576–601.
- [23] RAMDAS, A., RUF, J., LARSSON, M. AND KOOLEN, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv*: 2009.03167.
- [24] Ren, Z. and Barber, R. F. (2024). Derandomized knockoffs: Leveraging e-values for false discovery rate control. *Journal of the Royal Statistical Society Series B*, **86**(1), 122–154.
- [25] Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**(1), 751–754.
- [26] Shafer, G. (2021). The language of betting as a strategy for statistical and scientific communication. *Journal of the Royal Statistical Society, Series A*, **184**(2), 407–431.



- [27] Shafer, G. and Vovk, V. (2019). *Game-Theoretic Foundations for Probability and Finance*. Wiley, New York, 2019.
- [28] Ville, J. (1939). Étude critique de la notion de collectif. *Thèses de l'entre-deux-guerres*, 218.
- [29] Vovk, V. and Wang, R. (2020). Combining p-values via averaging. *Biometrika*, **107**(4), 791–808.
- [30] Vovk, V. and Wang, R. (2021). E-values: Calibration, combination, and applications. *Annals of Statistics*, **49**(3), 1736–1754.
- [31] Vovk, V. and Wang, R. (2023). Confidence and discoveries with e-values. *Statistical Science*, **38**(2), 329–354.
- [32] Vovk, V. and Wang, R. (2024). Nonparametric e-tests of symmetry. *New England Journal of Statistics in Data Science*, **2**(2), 261–270.
- [33] Wang, Q., Wang, R. and Ziegel, J. (2022). E-backtesting. *arXiv:2209.00991*.
- [34] Wang, R. and Ramdas, A. (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B*, **84**(3), 822–852.
- [35] Wang, H. and Ramdas, A. (2023). Catoni-style confidence sequences for heavy-tailed mean estimation. *Stochastic Processes and their Applications*, **163**, 168–202.
- [36] Wasserman, L., Ramdas, A. and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, **117**(29), 16880–16890.
- [37] Waudby-Smith, I. and Ramdas, A. (2024). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B*, **86**(1), 1–27.