The Annals of Statistics
2021, Vol. 49, No. 3, 1736–1754
https://doi.org/10.1214/20-AOS2020

E-VALUES: CALIBRATION, COMBINATION AND APPLICATIONS

BY VLADIMIR VOVK[1] AND RUODU WANG[2]

[1]*Department of Computer Science, Royal Holloway, University of London, v.vovk@rhul.ac.uk*
[2]*Department of Statistics and Actuarial Science, University of Waterloo, wang@uwaterloo.ca*

Multiple testing of a single hypothesis and testing multiple hypotheses are usually done in terms of p-values. In this paper, we replace p-values with their natural competitor, e-values, which are closely related to betting, Bayes factors and likelihood ratios. We demonstrate that e-values are often mathematically more tractable; in particular, in multiple testing of a single hypothesis, e-values can be merged simply by averaging them. This allows us to develop efficient procedures using e-values for testing multiple hypotheses.

1. Introduction. The problem of multiple testing of a single hypothesis (also known as testing a global null) is usually formalized as that of combining a set of p-values. The notion of p-values, however, has a strong competitor, which we refer to as e-values in this paper. E-values can be traced back to various old ideas, but they have started being widely discussed in their pure form only recently: see, for example, Shafer (2019), who uses the term "betting score" in the sense very similar to our "e-value", Shafer and Vovk (2019, Section 11.5), who use "Skeptic's capital", and Grünwald, de Heide and Koolen (2020). The power and intuitive appeal of e-values stem from their interpretation as results of bets against the null hypothesis (Shafer (2019, Section 1)).

Formally, an *e-variable* is a nonnegative extended random variable whose expected value under the null hypothesis is at most 1, and an *e-value* is a value taken by an e-variable. Whereas p-values are defined in terms of probabilities, e-values are defined in terms of expectations. As we regard an e-variable E as a bet against the null hypothesis, its realized value $e := E(\omega)$ shows how successful our bet is (it is successful if it multiplies the money it risks by a large factor). Under the null hypothesis, it can be larger than a constant $c > 1$ with probability at most $1/c$ (by Markov's inequality). If we are very successful (i.e., e is very large), we have reasons to doubt that the null hypothesis is true, and e can be interpreted as the amount of evidence we have found against it. In textbook statistics, e-variables typically appear under the guise of likelihood ratios and Bayes factors.

The main focus of this paper is on combining e-values and multiple hypothesis testing using e-values. The picture that arises for these two fields is remarkably different from, and much simpler than, its counterpart for p-values. To clarify connections between e-values and p-values, we discuss how to transform p-values into e-values, or *calibrate* them, and how to move in the opposite direction.

We start the main part of the paper by defining the notion of e-values in Section 2 and reviewing known results about connections between e-values and p-values; we will discuss how the former can be turned into the latter and vice versa (with very different domination structures for the two directions). In Section 3, we show that the problem of merging e-values is more or less trivial: a convex mixture of e-values is an e-value, and symmetric merging functions are essentially dominated by the arithmetic mean. For example, when several analyses are conducted on a common (e.g., public) dataset each reporting an e-value,

Received December 2019; revised September 2020.
MSC2020 subject classifications. Primary 62G10, 62F03; secondary 62C07, 62C15.
Key words and phrases. Hypothesis testing, multiple hypothesis testing, global null, Bayes factor, test martingale, admissible decisions.

it is natural to summarize them as a single e-value equal to their weighted average (the same cannot be said for p-values). In Section 4 we assume, additionally, that the e-variables being merged are independent and show that the domination structure is much richer; for example, now the product of e-values is an e-value. The assumption of independence can be replaced by the weaker assumption of being *sequential*, and we discuss connections with the popular topic of using martingales in statistical hypothesis testing: see, for example, Duan et al. (2019) and Shafer and Vovk (2019). In Section 5, we apply these results to multiple hypothesis testing. In the next section, Section 6, we briefly review known results on merging p-values (e.g., the two classes of merging methods in Rüger (1978) and Vovk and Wang (2020)) and draw parallels with merging e-values; in the last subsection we discuss the case where p-values are independent. Section 7 is devoted to experimental results; one finding in this section is that, for multiple testing of a single hypothesis in independent experiments, a simple method based on e-values outperforms standard methods based on p-values. Section 8 concludes. The Supplementary Material to this paper (Vovk and Wang (2021)) consists of a series of appendices.

2. Definition of e-values and connections with p-values. For a probability space $(\Omega, \mathcal{A}, Q)$, an *e-variable* is an extended random variable $E : \Omega \to [0, \infty]$ satisfying $\mathbb{E}^Q[E] \le 1$; we refer to it as "extended" since its values are allowed to be ∞, and we let $\mathbb{E}^Q[X]$ (or $\mathbb{E}[X]$ when Q is clear from context) stand for $\int X \, \mathrm{d}Q$ for any extended random variable X. The values taken by e-variables will be referred to as *e-values*, and we denote the set of e-variables by $\mathcal{E}_Q$. It is important to allow E to take value ∞; in the context of testing Q, observing $E = \infty$ for an *a priori* chosen e-variable E means that we are entitled to reject Q as null hypothesis.

Until Section 5, we will concentrate on the case of a simple null hypothesis Q. A composite null hypothesis $\mathcal{Q}$ is a set of probability measures on $(\Omega, \mathcal{A})$ and an e-variable for $\mathcal{Q}$ is defined as an extended nonnegative random variable that integrates to at most 1 under any probability measure in $\mathcal{Q}$. The main results that we state for simple null hypotheses remain true for composite null hypotheses; see Appendix D.

Our emphasis in this paper is on e-values, but we start from discussing their connections with the familiar notion of p-values. A *p-variable* is a random variable $P : \Omega \to [0, 1]$ satisfying

$$\forall \epsilon \in (0, 1) : \quad Q(P \le \epsilon) \le \epsilon.$$

The set of all p-variables is denoted by $\mathcal{P}_Q$.

A calibrator is a function transforming p-values to e-values. Formally, a decreasing function $f : [0, 1] \to [0, \infty]$ is a *calibrator* (or, more fully, *p-to-e calibrator*) if, for any probability space $(\Omega, \mathcal{A}, Q)$ and any p-variable $P \in \mathcal{P}_Q$, $f(P) \in \mathcal{E}_Q$. A calibrator f is said to *dominate* a calibrator g if $f \ge g$, and the domination is *strict* if $f \ne g$. A calibrator is *admissible* if it is not strictly dominated by any other calibrator.

The following proposition says that a calibrator is a nonnegative decreasing function integrating to at most 1 over the uniform probability measure.

PROPOSITION 2.1. *A decreasing function $f : [0, 1] \to [0, \infty]$ is a calibrator if and only if $\int_0^1 f \le 1$. It is admissible if and only if f is upper semicontinuous, $f(0) = \infty$, and $\int_0^1 f = 1$.*

Of course, in the context of this proposition, being upper semicontinuous is equivalent to being left-continuous.

PROOF. Proofs of similar statements are given in, for example, Vovk (1993), Theorem 7, Shafer et al. (2011), Theorem 3 and Shafer and Vovk (2019), Proposition 11.7, but we will give an independent short proof using our definitions. The first "only if" statement is obvious. To show the first "if" statement, suppose that $\int_0^1 f \le 1$, P is a p-variable, and P' is uniformly distributed on $[0,1]$. Since $Q(P < x) \le Q(P' < x)$ for all $x \ge 0$ and f is decreasing, we have

$$Q\big(f(P) > y\big) \le Q\big(f(P') > y\big)$$

for all $y \ge 0$, which implies

$$\mathbb{E}\big[f(P)\big] \le \mathbb{E}\big[f(P')\big] = \int_0^1 f(p)\,\mathrm{d}p \le 1.$$

The second statement in Proposition 2.1 is obvious. □

The following is a simple family of calibrators. Since $\int_0^1 \kappa p^{\kappa-1}\,\mathrm{d}p = 1$, the functions

$$f_\kappa(p) := \kappa p^{\kappa-1} \tag{1}$$

are calibrators, where $\kappa \in (0,1)$. To solve the problem of choosing the parameter κ, sometimes the maximum

$$\mathrm{VS}(p) := \max_{\kappa\in[0,1]} f_\kappa(p) = \begin{cases} -\exp(-1)/(p\ln p) & \text{if } p \le \exp(-1), \\ 1 & \text{otherwise} \end{cases}$$

is used (see, e.g., Benjamin and Berger (2019), Recommendations 2 and 3); we will refer to it as the *VS bound* (abbreviating "Vovk–Sellke bound", as used in, e.g., the JASP package). It is important to remember that $\mathrm{VS}(p)$ is not a valid e-value, but just an overoptimistic upper bound on what is achievable with the class (1). Another way to get rid of κ is to integrate over it, which gives

$$F(p) := \int_0^1 \kappa p^{\kappa-1}\,\mathrm{d}\kappa = \frac{1-p+p\ln p}{p(-\ln p)^2}. \tag{2}$$

(See Appendix B in the Supplementary Material for more general results and references. We are grateful to Aaditya Ramdas for pointing out the calibrator (2).) An advantage of this method is that it produces a bona fide e-value, unlike the VS bound. As $p \to 0$, $F(p) \sim p^{-1}(-\ln p)^{-2}$, so that $F(p)$ is closer to the ideal (but unachievable) $1/p$ (cf. Remark 2.3 below) than any of (1).

In the opposite direction, an e-to-p calibrator is a function transforming e-values to p-values. Formally, a decreasing function $f : [0,\infty] \to [0,1]$ is an *e-to-p calibrator* if, for any probability space $(\Omega, \mathcal{A}, Q)$ and any e-variable $E \in \mathcal{E}_Q$, $f(E) \in \mathcal{P}_Q$. The following proposition, which is the analogue of Proposition 2.1 for e-to-p calibrators, says that there is, essentially, only one e-to-p calibrator, $f(t) := \min(1, 1/t)$.

PROPOSITION 2.2. *The function $f : [0,\infty] \to [0,1]$ defined by $f(t) := \min(1,1/t)$ is an e-to-p calibrator. It dominates every other e-to-p calibrator. In particular, it is the only admissible e-to-p calibrator.*

PROOF. The fact that $f(t) := \min(1,1/t)$ is an e-to-p calibrator follows from Markov's inequality: if $E \in \mathcal{E}_Q$ and $\epsilon \in (0,1)$,

$$Q\big(f(E) \le \epsilon\big) = Q(E \ge 1/\epsilon) \le \frac{\mathbb{E}^Q[E]}{1/\epsilon} \le \epsilon.$$

On the other hand, suppose that f is another e-to-p calibrator. It suffices to check that f is dominated by $\min(1,1/t)$. Suppose $f(t) < \min(1,1/t)$ for some $t \in [0,\infty]$. Consider two cases:

- If $f(t) < \min(1, 1/t) = 1/t$ for some $t > 1$, fix such t and consider an e-variable E that is t with probability $1/t$ and 0 otherwise. Then $f(E)$ is $f(t) < 1/t$ with probability $1/t$, whereas it would have satisfied $P(f(E) \le f(t)) \le f(t) < 1/t$ had it been a p-variable.
- If $f(t) < \min(1, 1/t) = 1$ for some $t \in [0, 1]$, fix such t and consider an e-variable E that is 1 a.s. Then $f(E)$ is $f(t) < 1$ a.s., and so it is not a p-variable. □

Proposition 2.1 implies that the domination structure of calibrators is very rich, whereas Proposition 2.2 implies that the domination structure of e-to-p calibrators is trivial.

REMARK 2.3. A possible interpretation of this section's results is that e-variables and p-variables are connected via a rough relation $1/e \sim p$. In one direction, the statement is precise: the reciprocal (truncated to 1 if needed) of an e-variable is a p-variable by Proposition 2.2. On the other hand, using a calibrator (1) with a small $\kappa > 0$ and ignoring positive constant factors (as customary in the algorithmic theory of randomness, discussed in Section A.2), we can see that the reciprocal of a p-variable is approximately an e-variable. In fact, $f(p) \le 1/p$ for all p when f is a calibrator; this follows from Proposition 2.1. However, $f(p) = 1/p$ is only possible in the extreme case $f = 1_{[0,p]}/p$.

3. Merging e-values. An important advantage of e-values over p-values is that they are easy to combine. This is the topic of this section, in which we consider the general case, without any assumptions on the joint distribution of the input e-variables. The case of independent e-variables is considered in the next section.

Let $K \ge 2$ be a positive integer (fixed throughout the paper apart from Section 7). An *e-merging function* of K e-values is an increasing Borel function $F : [0, \infty]^K \to [0, \infty]$ such that, for any probability space $(\Omega, \mathcal{A}, Q)$ and random variables $E_1, \dots, E_K$ on it,

$$E_1, \dots, E_K \in \mathcal{E}_Q \quad \Longrightarrow \quad F(E_1, \dots, E_K) \in \mathcal{E}_Q \tag{3}$$

(in other words, F transforms e-values into an e-value). In this paper we will also refer to increasing Borel functions $F : [0, \infty)^K \to [0, \infty)$ satisfying (3) for all probability spaces and all e-variables $E_1, \dots, E_K$ taking values in $[0, \infty)$ as e-merging functions; such functions are canonically extended to e-merging functions $F : [0, \infty]^K \to [0, \infty]$ by setting them to ∞ on $[0, \infty]^K \setminus [0, \infty)^K$ (see Proposition C.1 in the Supplementary Material).

An e-merging function F *dominates* an e-merging function G if $F \ge G$ (i.e., $F(\mathbf{e}) \ge G(\mathbf{e})$ for all $\mathbf{e} \in [0, \infty)^K$). The domination is *strict* (and we say that F *strictly dominates* G) if $F \ge G$ and $F(\mathbf{e}) > G(\mathbf{e})$ for some $\mathbf{e} \in [0, \infty)^K$. We say that an e-merging function F is *admissible* if it is not strictly dominated by any e-merging function; in other words, admissibility means being maximal in the partial order of domination.

A fundamental fact about admissibility is proved in Appendix E (Proposition E.5): any e-merging function is dominated by an admissible e-merging function.

Merging e-values via averaging. In this paper, we are mostly interested in symmetric merging functions (i.e., those invariant w.r. to permutations of their arguments). The main message of this section is that the most useful (and the only useful, in a natural sense) symmetric e-merging function is the *arithmetic mean*

$$M_K(e_1, \dots, e_K) := \frac{e_1 + \dots + e_K}{K}, \quad e_1, \dots, e_K \in [0, \infty). \tag{4}$$

In Theorem 3.2 below we will see that M_K is admissible (this is also a consequence of Proposition 4.1). But first we state formally the vague claim that M_K is the only useful symmetric e-merging function.

An e-merging function F *essentially dominates* an e-merging function G if, for all $\mathbf{e} \in [0,\infty)^K$,

$$G(\mathbf{e}) > 1 \quad \Longrightarrow \quad F(\mathbf{e}) \geq G(\mathbf{e}).$$

This weakens the notion of domination in a natural way: now we require that F is not worse than G only in cases where G is not useless; we are not trying to compare degrees of uselessness. The following proposition can be interpreted as saying that M_K is at least as good as any other symmetric e-merging function.

PROPOSITION 3.1. *The arithmetic mean M_K essentially dominates any symmetric e-merging function.*

In particular, if F is an e-merging function that is symmetric and positively homogeneous (i.e., $F(\lambda\mathbf{e}) = \lambda F(\mathbf{e})$ for all $\lambda > 0$), then F is dominated by M_K. This includes the e-merging functions discussed later in Section 6.

PROOF OF PROPOSITION 3.1. Let F be a symmetric e-merging function. Suppose for the purpose of contradiction that there exists $(e_1,\dots,e_K) \in [0,\infty)^K$ such that

$$b := F(e_1,\dots,e_K) > \max\left(\frac{e_1+\cdots+e_K}{K}, 1\right) =: a. \tag{5}$$

Let Π_K be the set of all permutations of $\{1,\dots,K\}$, π be randomly and uniformly drawn from Π_K, and $(D_1,\dots,D_K) := (e_{\pi(1)},\dots,e_{\pi(K)})$. Further, let $(D'_1,\dots,D'_K) := (D_1,\dots,D_K)1_A$, where A is an event independent of π and satisfying $P(A) = 1/a$ (the existence of such random π and A is guaranteed for any atomless probability space by Lemma D.1 in the Supplementary Material).

For each k, since D_k takes the values $e_1,\dots,e_K$ with equal probability, we have $\mathbb{E}[D_k] = (e_1+\cdots+e_K)/K$, which implies $\mathbb{E}[D'_k] = (e_1+\cdots+e_K)/(Ka) \leq 1$. Together with the fact that D'_k is nonnegative, we know $D'_k \in \mathcal{E}_Q$. Moreover, by symmetry,

$$\mathbb{E}\big[F\big(D'_1,\dots,D'_K\big)\big] = Q(A)F(e_1,\dots,e_K) + \big(1-Q(A)\big)F(0,\dots,0) \geq b/a > 1,$$

a contradiction. Therefore, we conclude that there is no $(e_1,\dots,e_K)$ such that (5) holds. □

It is clear that the arithmetic mean M_K does not dominate every symmetric e-merging function; for example, the convex mixtures

$$\lambda + (1-\lambda)M_K, \quad \lambda \in [0,1], \tag{6}$$

of the trivial e-merging function 1 and M_K are pairwise noncomparable (with respect to the relation of domination). In the theorem below, we show that each of these mixtures is admissible and that the class (6) is, in the terminology of statistical decision theory (Wald (1950), Section 1.3), a complete class of symmetric e-merging functions: every symmetric e-merging function is dominated by one of (6). In other words, (6) is the minimal complete class of symmetric e-merging functions.

THEOREM 3.2. *Suppose that F is a symmetric e-merging function. Then F is dominated by the function $\lambda + (1-\lambda)M_K$ for some $\lambda \in [0,1]$. In particular, F is admissible if and only if $F = \lambda + (1-\lambda)M_K$, where $\lambda = F(\mathbf{0}) \in [0,1]$.*

The proof of Theorem 3.2 is put in Appendix E as it requires several other technical results in the Supplementary Material. Finally, we note that, for $\lambda \neq 1$, the functions in the class (6) carry the same statistical information.

4. Merging independent e-values. In this section, we consider merging functions for independent e-values. An *ie-merging function* of K e-values is an increasing Borel function $F : [0,\infty)^K \to [0,\infty)$ such that $F(E_1,\dots,E_K) \in \mathcal{E}_Q$ for all independent $E_1,\dots,E_K \in \mathcal{E}_Q$ in any probability space $(\Omega, \mathcal{A}, Q)$. As for e-merging functions, this definition is essentially equivalent to the definition involving $[0,\infty]$ rather than $[0,\infty)$ (by Proposition C.1 in the Supplementary Material, which is still applicable in the context of merging independent e-values). The definitions of domination, strict domination and admissibility are obtained from the definitions of the previous section by replacing "e-merging" with "ie-merging".

Let $i\mathcal{E}_Q^K \subseteq \mathcal{E}_Q^K$ be the set of (component-wise) independent random vectors in $\mathcal{E}_Q^K$, and $\mathbf{1} := (1,\dots,1)$ be the all-1 vector in $\mathbb{R}^K$. The following proposition has already been used in Section 3 (in particular, it implies that the arithmetic mean M_K is an admissible e-merging function).

PROPOSITION 4.1. *For an increasing Borel function $F : [0,\infty)^K \to [0,\infty)$, if $\mathbb{E}[F(\mathbf{E})] = 1$ for all $\mathbf{E} \in \mathcal{E}_Q^K$ with $\mathbb{E}[\mathbf{E}] = \mathbf{1}$ (resp., for all $\mathbf{E} \in i\mathcal{E}_Q^K$ with $\mathbb{E}[\mathbf{E}] = \mathbf{1}$), then F is an admissible e-merging function (resp., an admissible ie-merging function).*

PROOF. It is obvious that F is an e-merging function (resp., ie-merging function). Next, we show that F is admissible. Suppose for the purpose of contradiction that there exists an ie-merging function G such that $G \ge F$ and $G(e_1,\dots,e_K) > F(e_1,\dots,e_K)$ for some $(e_1,\dots,e_K) \in [0,\infty)^K$. Take $(E_1,\dots,E_K) \in i\mathcal{E}_Q^K$ with $\mathbb{E}[(E_1,\dots,E_K)] = \mathbf{1}$ such that $Q((E_1,\dots,E_K) = (e_1,\dots,e_K)) > 0$. Such a random vector is easy to construct by considering any distribution with a positive mass on each of $e_1,\dots,e_K$. Then we have

$$Q\big(G(E_1,\dots,E_K) > F(E_1,\dots,E_K)\big) > 0,$$

which implies

$$\mathbb{E}\big[G(E_1,\dots,E_K)\big] > \mathbb{E}\big[G(E_1,\dots,E_K)\big] = 1,$$

contradicting the assumption that G is an ie-merging function. Therefore, no ie-merging function strictly dominates F. Noting that an e-merging function is also an ie-merging function, admissibility of F is guaranteed under both settings. □

If $E_1,\dots,E_K$ are independent e-variables, their product $E_1 \dots E_K$ will also be an e-variable. This is the analogue of Fisher's (1932) method for p-values (according to the rough relation $e \sim 1/p$ mentioned in Remark 2.3; Fisher's method is discussed at the end of Section 6). The ie-merging function

$$(e_1,\dots,e_K) \mapsto e_1 \dots e_K \tag{7}$$

is admissible by Proposition 4.1. It will be referred to as the *product* (or *multiplication*) ie-merging function. The betting interpretation of (7) is obvious: it is the result of K successive bets using the e-variables $E_1,\dots,E_K$ (starting with initial capital 1 and betting the full current capital $E_1 \dots E_{k-1}$ on each E_k).

More generally, we can see that the U-statistics

$$U_n(e_1,\dots,e_K) := \frac{1}{\binom{K}{n}} \sum_{\{k_1,\dots,k_n\} \subseteq \{1,\dots,K\}} e_{k_1} \dots e_{k_n}, \quad n \in \{0,1,\dots,K\}, \tag{8}$$

and their convex mixtures are ie-merging functions. Notice that this class includes product (for $n = K$), arithmetic average M_K (for $n = 1$), and constant 1 (for $n = 0$). Proposition 4.1 implies that the U-statistics (8) and their convex mixtures are admissible ie-merging functions.

The betting interpretation of a U-statistic (8) or a convex mixture of U-statistics is implied by the betting interpretation of each component $e_{k_1} \dots e_{k_n}$. Assuming that $k_1, \dots, k_n$ are sorted in the increasing order, $e_{k_1} \dots e_{k_n}$ is the result of n successive bets using the e-variables $E_{k_1}, \dots, E_{k_n}$; and a convex mixture of bets corresponds to investing the appropriate fractions of the initial capital into those bets.

Let us now establish a very weak counterpart of Proposition 3.1 for independent e-values (on the positive side it will not require the assumption of symmetry). An ie-merging function F *weakly dominates* an ie-merging function G if, for all $e_1, \dots, e_K$,

$$(e_1, \dots, e_K) \in [1, \infty)^K \quad \Longrightarrow \quad F(e_1, \dots, e_K) \geq G(e_1, \dots, e_K).$$

In other words, we require that F is not worse than G if all input e-values are useful (and this requirement is weak because, especially for a large K, we are also interested in the case where some of the input e-values are useless).

PROPOSITION 4.2. *The product $(e_1, \dots, e_K) \mapsto e_1 \dots e_K$ weakly dominates any ie-merging function.*

PROOF. Indeed, suppose that there exists $(e_1, \dots, e_K) \in [1, \infty)^K$ such that

$$F(e_1, \dots, e_K) > e_1 \dots e_K.$$

Let $E_1, \dots, E_K$ be independent random variables such that each E_k for $k \in \{1, \dots, K\}$ takes values in the two-element set $\{0, e_k\}$ and $E_k = e_k$ with probability $1/e_k$. Then each E_k is an e-variable but

$$\begin{aligned} \mathbb{E}\big[F(E_1, \dots, E_K)\big] &\geq F(e_1, \dots, e_K) Q(E_1 = e_1, \dots, E_K = e_K) \\ &> e_1 \dots e_K (1/e_1) \dots (1/e_K) = 1, \end{aligned}$$

which contradicts F being an ie-merging function. □

REMARK 4.3. A natural question is whether the convex mixtures of (8) form a complete class. They do not: Proposition 4.1 implies that

$$f(e_1, e_2) := \frac{1}{2}\left(\frac{e_1}{1 + e_1} + \frac{e_2}{1 + e_2}\right)(1 + e_1 e_2)$$

is an admissible ie-merging function, and it is easy to check that it is different from any convex mixture of (8).

Testing with martingales. The assumption of the independence of e-variables $E_1, \dots, E_K$ is not necessary for the product $E_1 \dots E_K$ to be an e-variable. Below, we say that the e-variables $E_1, \dots, E_K$ are *sequential* if $\mathbb{E}[E_k \mid E_1, \dots, E_{k-1}] \leq 1$ almost surely for all $k \in \{1, \dots, K\}$. Equivalently, the sequence of the partial products $(E_1 \dots E_k)_{k=0,1,\dots,K}$ is a supermartingale in the filtration generated by $E_1, \dots, E_K$ (or a *test supermartingale*, in the terminology of Grünwald, de Heide and Koolen (2020), Howard et al. (2020), Shafer et al. (2011), meaning a nonnegative supermartingale with initial value 1). A possible interpretation of this test supermartingale is that the e-values $e_1, e_2, \dots$ are obtained by laboratories $1, 2, \dots$ in this order, and laboratory k makes sure that its result e_k is a valid e-value given the previous results $e_1, \dots, e_{k-1}$. The test supermartingale is a *test martingale* if $\mathbb{E}[E_k \mid E_1, \dots, E_{k-1}] = 1$ almost surely for all k (intuitively, it is not wasteful).

It is straightforward to check that all convex mixtures of (8) (including the product function) produce a valid e-value from sequential e-values. On the other hand, independent e-variables are sequential, and hence merging functions for sequential e-values form a subset

of ie-merging functions. In this class of merging functions, the convex mixtures of (8) are admissible, as they are admissible in the larger class of ie-merging functions (by Proposition 4.1). For the same reason (and by Proposition 4.2), the product function in (7) weakly dominates every other merging function for sequential e-variables. This gives a (weak) theoretical justification for us to use the product function as a canonical merging method in Sections 5 and 7 for e-values as long as they are sequential. Finally, we note that it suffices for $E_1, \dots, E_K$ to be sequential in any order for these merging methods (such as Algorithm 2 in Section 5) to be valid.

5. Application to testing multiple hypotheses. As in Vovk and Wang (2020), we will apply results for multiple testing of a single hypothesis (combining e-values in the context of Sections 3 and 4) to testing multiple hypotheses. As we explain in Appendix A (Section A.3), our algorithms just spell out the application of the closure principle (Goeman and Solari (2011), Marcus, Peritz and Gabriel (1976)), but our exposition in this section will be self-contained.

Let $(\Omega, \mathcal{A})$ be our sample space (formally, a measurable space) and $\mathfrak{P}(\Omega)$ be the family of all probability measures on it. Remember that E is an *e-variable* w.r. to a composite null hypothesis $H \subseteq \mathfrak{P}(\Omega)$ if $\mathbb{E}^Q[E_k] \le 1$ for any $Q \in H_k$.

In multiple hypothesis testing, we are given a set of composite null hypotheses H_k, $k = 1, \dots, K$. Suppose that, for each k, we are also given an e-variable E_k w.r. to H_k. Our multiple testing procedure is presented as Algorithm 1. The procedure adjusts the e-values $e_1, \dots, e_K$, perhaps obtained in K experiments (not necessarily independent), to new e-values $e_1^*, \dots, e_K^*$; the adjustment is downward in that $e_k^* \le e_k$ for all k. Applying the procedure to the e-values $e_1, \dots, e_K$ produced by the e-variables $E_1, \dots, E_K$, we obtain extended random variables $E_1^*, \dots, E_K^*$ taking values $e_1^*, \dots, e_K^*$. The output $E_1^*, \dots, E_K^*$ of Algorithm 1 satisfies a property of validity which we will refer to as *family-wise validity* (FWV); in Section A.3 we will explain its analogy with the standard family-wise error rate (FWER).

A *conditional e-variable* is a family of extended nonnegative random variables E_Q, $Q \in \mathfrak{P}(\Omega)$, that satisfies

$$\forall Q \in \mathfrak{P}(\Omega): \quad \mathbb{E}^Q[E_Q] \le 1$$

(i.e., each E_Q is in $\mathcal{E}_Q$). We regard it as a system of bets against each potential data-generating distribution Q.

Algorithm 1 Adjusting e-values for multiple hypothesis testing

Require: A sequence of e-values $e_1, \dots, e_K$.

Find a permutation $\pi : \{1, \dots, K\} \to \{1, \dots, K\}$ such that $e_{\pi(1)} \le \dots \le e_{\pi(K)}$.
Set $e_{(k)} := e_{\pi(k)}$, $k \in \{1, \dots, K\}$ (these are the *order statistics*).
$S_0 := 0$
for $i = 1, \dots, K$ **do**
    $S_i := S_{i-1} + e_{(i)}$
for $k = 1, \dots, K$ **do**
    $e^*_{\pi(k)} := e_{\pi(k)}$
    for $i = 1, \dots, k-1$ **do**
        $e := \frac{e_{\pi(k)} + S_i}{i+1}$
        if $e < e^*_{\pi(k)}$ **then**
            $e^*_{\pi(k)} := e$

Algorithm 2 Adjusting sequential e-values for multiple hypothesis testing

Require: A sequence of e-values $e_1, \dots, e_K$.

Let a be the product of all $e_k < 1$, $k = 1, \dots, K$ (and $a := 1$ if there are no such k).
for $k = 1, \dots, K$ **do**
    $e_k^* := a e_k$

Extended random variables $E_1^*, \dots, E_K^*$ taking values in $[0, \infty]$ are *family-wise valid* (*FWV*) for testing $H_1, \dots, H_K$ if there exists a conditional e-variable $(E_Q)_{Q \in \mathfrak{P}(\Omega)}$ such that

$$\forall k \in \{1, \dots, K\} \, \forall Q \in H_k : \quad E_Q \geq E_k^* \tag{9}$$

(where $E_Q \geq E_k^*$ means, as usual, that $E_Q(\omega) \geq E_k^*(\omega)$ for all $\omega \in \Omega$). We can say that such $(E_Q)_{Q \in \mathfrak{P}(\Omega)}$ *witnesses* the FWV property of $E_1^*, \dots, E_K^*$.

The interpretation of family-wise validity is based on our interpretation of e-values. Suppose we observe an outcome $\omega \in \Omega$. If $E_Q(\omega)$ is very large, we may reject Q as the data-generating distribution. Therefore, if $E_k^*(\omega)$ is very large, we may reject the whole of H_k (i.e., each $Q \in H_k$). In betting terms, we have made at least $\$E_k^*(\omega)$ risking at most \$1 when gambling against any $Q \in H_k$.

Notice that we can rewrite (9) as

$$\forall Q \in \mathfrak{P}(\Omega) : \quad \mathbb{E}^Q \Big[\max_{k : Q \in H_k} E_k^* \Big] \leq 1.$$

In other words, we require joint validity of the e-variables E_k^*.

We first state the validity of Algorithm 1 (as well as Algorithm 2), and our justification follows.

THEOREM 5.1. *Algorithms 1 and 2 are family-wise valid.*

Let us check that the output $E_1^*, \dots, E_K^*$ of Algorithm 1 is FWV. For $I \subseteq \{1, \dots, K\}$, the composite hypothesis H_I is defined by

$$H_I := \Big(\bigcap_{k \in I} H_k \Big) \cap \Big(\bigcap_{k \in \{1, \dots, K\} \setminus I} H_k^{\mathsf{c}} \Big), \tag{10}$$

where H_k^{c} is the complement of H_k. The conditional e-variable witnessing that $E_1^*, \dots, E_K^*$ are FWV is the arithmetic mean

$$E_Q := \frac{1}{|I_Q|} \sum_{k \in I_Q} E_k, \tag{11}$$

where $I_Q := \{k \mid Q \in H_k\}$ and E_Q is defined arbitrarily (say, as 1) when $I_Q = \varnothing$. The optimal adjusted e-variables E_k' can be defined as

$$E_k' := \min_{Q \in H_k} E_Q \geq \min_{I \subseteq \{1, \dots, K\} : k \in I} \frac{1}{|I|} \sum_{i \in I} E_i, \tag{12}$$

but for computational efficiency we use the conservative definition

$$E_k^* := \min_{I \subseteq \{1, \dots, K\} : k \in I} \frac{1}{|I|} \sum_{i \in I} E_i. \tag{13}$$

REMARK 5.2. The inequality "$\geq$" in (12) holds as the equality "$=$" if all the intersections (10) are nonempty. If some of these intersections are empty, we can have a strict inequality. Algorithm 1 implements the definition (13). Therefore, it is valid regardless of whether some of the intersections (10) are empty; however, if they are, it may be possible to improve the adjusted e-values. According to Holm's (1979) terminology, we allow "free combinations". Shaffer (1986) pioneered methods that take account of the logical relations between the base hypotheses H_k.

To obtain Algorithm 1, we rewrite the definitions (13) as

$$E^*_{\pi(k)} = \min_{i\in\{0,\dots,k-1\}} \frac{E_{\pi(k)} + E_{(1)} + \cdots + E_{(i)}}{i+1}$$
$$= \min_{i\in\{1,\dots,k-1\}} \frac{E_{\pi(k)} + E_{(1)} + \cdots + E_{(i)}}{i+1}$$

for $k \in \{1,\dots,K\}$, where π is the ordering permutation and $E_{(j)} = E_{\pi(j)}$ is the jth order statistic among $E_1,\dots,E_K$, as in Algorithm 1. In lines 3–5 of Algorithm 1, we precompute the sums

$$S_i := e_{(1)} + \cdots + e_{(i)}, \quad i = 1,\dots,K,$$

in lines 8–9 we compute

$$e_{k,i} := \frac{e_{\pi(k)} + e_{(1)} + \cdots + e_{(i)}}{i+1}$$

for $i = 1,\dots,k-1$, and as result of executing lines 6–11 we will have

$$e^*_{\pi(k)} = \min_{i\in\{1,\dots,k-1\}} e_{k,i} = \min_{i\in\{1,\dots,k-1\}} \frac{e_{\pi(k)} + e_{(1)} + \cdots + e_{(i)}}{i+1},$$

which shows that Algorithm 1 is an implementation of (13).

The computational complexity of Algorithm 1 is $O(K^2)$.

In the case of sequential e-variables, we have Algorithm 2. This algorithm assumes that, under any $Q \in \mathfrak{P}(\Omega)$, the base e-variables E_k, $k \in I_Q$, are sequential (remember that I_Q is defined by (11) and that independence implies being sequential). The conditional e-variable witnessing that the output of Algorithm 2 is FWV is the one given by the product ie-merging function,

$$E_Q := \prod_{k\in I_Q} E_k,$$

where the adjusted e-variables are defined by

$$E^*_k := \min_{I\subseteq\{1,\dots,K\}:k\in I} \prod_{i\in I} E_i. \tag{14}$$

A remark similar to Remark 5.2 can also be made about Algorithm 2. The computational complexity of Algorithm 2 is $O(K)$ (unusually, the algorithm does not require sorting the base e-values).

6. Merging p-values and comparisons. Merging p-values is a much more difficult topic than merging e-values, but it is very well explored. First, we review merging p-values without any assumptions, and then we move on to merging independent p-values.

A *p-merging function* of K p-values is an increasing Borel function $F : [0,1]^K \to [0,1]$ such that $F(P_1,\dots,P_K) \in \mathcal{P}_Q$ whenever $P_1,\dots,P_K \in \mathcal{P}_Q$.

For merging p-values without the assumption of independence, we will concentrate on two natural families of p-merging functions. The older family is the one introduced by Rüger (1978), and the newer one was introduced in our paper Vovk and Wang (2020). Rüger's family is parameterized by $k \in \{1, \dots, K\}$, and its kth element is the function (shown by Rüger (1978) to be a p-merging function)

$$(p_1, \dots, p_K) \mapsto \frac{K}{k} p_{(k)} \wedge 1, \tag{15}$$

where $p_{(k)} := p_{\pi(k)}$ and π is a permutation of $\{1, \dots, K\}$ ordering the p-values in the ascending order: $p_{\pi(1)} \le \dots \le p_{\pi(K)}$. The other family (Vovk and Wang (2020)), which we will refer to as the *M-family*, is parameterized by $r \in [-\infty, \infty]$, and its element with index r has the form $a_{r,K} M_{r,K} \wedge 1$, where

$$M_{r,K}(p_1, \dots, p_K) := \left(\frac{p_1^r + \dots + p_K^r}{K} \right)^{1/r} \tag{16}$$

and $a_{r,K} \ge 1$ is a suitable constant. We also define $M_{r,K}$ for $r \in \{0, \infty, -\infty\}$ as the limiting cases of (16), which correspond to the geometric average, the maximum, and the minimum, respectively.

The initial and final elements of both families coincide: the initial element is the Bonferroni p-merging function

$$(p_1, \dots, p_K) \mapsto K \min(p_1, \dots, p_K) \wedge 1, \tag{17}$$

and the final element is the maximum p-merging function

$$(p_1, \dots, p_K) \mapsto \max(p_1, \dots, p_K).$$

Similarly to the case of e-merging functions, we say that a p-merging function F *dominates* a p-merging function G if $F \le G$. The domination is *strict* if, in addition, $F(\mathbf{p}) < G(\mathbf{p})$ for at least one $\mathbf{p} \in [0, 1]^K$. We say that a p-merging function F is *admissible* if it is not strictly dominated by any p-merging function G.

The domination structure of p-merging functions is much richer than that of e-merging functions. The maximum p-merging function is clearly inadmissible (e.g., $(p_1, \dots, p_K) \mapsto \max(p_1, \dots, p_K)$ is strictly dominated by $(p_1, \dots, p_K) \mapsto p_1$) while the Bonferroni p-merging function (17) is admissible, as the following proposition (proved in Appendix H in the Supplementary Material) shows.

PROPOSITION 6.1. *The Bonferroni p-merging function is admissible.*

The general domination structure of p-merging functions appears to be very complicated, and is the subject of future planned work.

Connections to e-merging functions. The domination structure of the class of e-merging functions is very simple, according to Theorem 3.2. It makes it very easy to understand what the e-merging analogues of Rüger's family and the M-family are; when stating the analogues we will use the rough relation $1/e \sim p$ between e-values and p-values (see Remark 2.3). Let us say that an e-merging function F is *precise* if cF is not an e-merging function for any $c > 1$.

For a sequence $e_1, \dots, e_K$, let $e_{[k]} := e_{\pi(k)}$ be the order statistics numbered from the largest to the smallest; here π is a permutation of $\{1, \dots, K\}$ ordering e_k in the descending order: $e_{\pi(1)} \ge \dots \ge e_{\pi(K)}$. Let us check that the Rüger-type function $(e_1, \dots, e_K) \mapsto (k/K) e_{[k]}$ is a

precise e-merging function. It is an e-merging function since it is dominated by the arithmetic mean: indeed, the condition of domination

$$\frac{k}{K}e_{[k]} \le \frac{e_1 + \cdots + e_K}{K}, \tag{18}$$

can be rewritten as

$$ke_{[k]} \le e_1 + \cdots + e_K$$

and so is obvious. As sometimes we have a strict inequality, the e-merging function is inadmissible (remember that we assume $K \ge 2$). The e-merging function is precise because (18) holds as equality when the k largest e_i, $i \in \{1, \dots, K\}$, are all equal and greater than 1 and all the other e_i are 0.

In the case of the M-family, let us check that the function

$$F := \left(K^{1/r-1} \wedge 1\right) M_{r,K} \tag{19}$$

is a precise e-merging function, for any $r \in [-\infty, \infty]$. For $r \le 1$, $M_{r,K}$ is increasing in r (Hardy, Littlewood and Pólya (1952), Theorem 16), and so $F = M_{r,K}$ is dominated by the arithmetic mean M_K; therefore, it is an e-merging function. For $r > 1$, we can rewrite the function $F = K^{1/r-1} M_{r,K}$ as

$$F(e_1, \dots, e_K) = K^{1/r-1} M_{r,K}(e_1, \dots, e_K) = K^{-1}\left(e_1^r + \cdots + e_K^r\right)^{1/r},$$

and we know that the last expression is a decreasing function of r (Hardy, Littlewood and Pólya (1952), Theorem 19); therefore, F is also dominated by M_K and so is a merging function. The e-merging function F is precise (for any r) since

$$\begin{aligned} r \le 1 \quad &\Longrightarrow \quad F(e, \dots, e) = M_K(e, \dots, e) = e, \\ r > 1 \quad &\Longrightarrow \quad F(0, \dots, 0, e) = M_K(0, \dots, 0, e) = e/K, \end{aligned}$$

and so by Proposition 3.1 (applied to a sufficiently large e) cF is not an e-merging function for any $c > 1$. But F is admissible if and only if $r = 1$ as shown by Theorem 3.2.

REMARK 6.2. The rough relation $1/e \sim p$ also sheds light on the coefficient, $K^{1/r-1} \wedge 1 = K^{1/r-1}$ for $r > 1$, given in (19) in front of $M_{r,K}$. The coefficient $K^{1/r-1}$, $r > 1$, in front of $M_{r,K}$ for averaging e-values corresponds to a coefficient of $K^{1+1/r}$, $r < -1$, in front of $M_{r,K}$ for averaging p-values. And indeed, by Proposition 5 of Vovk and Wang (2020), the asymptotically precise coefficient in front of $M_{r,K}$, $r < -1$, for averaging p-values is $\frac{r}{r+1} K^{1+1/r}$. The extra factor $\frac{r}{r+1}$ appears because the reciprocal of a p-variable is only approximately, but not exactly, an e-variable.

REMARK 6.3. Our formulas for merging e-values are explicit and much simpler than the formulas for merging p-values given in Vovk and Wang (2020), where the coefficient $a_{r,K}$ is often not analytically available. Merging e-values does not involve asymptotic approximations via the theory of robust risk aggregation (e.g., Embrechts, Wang and Wang (2015)), as used in that paper. This suggests that in some important respects e-values are easier objects to deal with than p-values.

Merging independent p-values. In this section, we will discuss ways of combining p-values $p_1, \dots, p_K$ under the assumption that the p-values are independent.

One of the oldest and most popular methods for combining p-values is Fisher's (1932), Section 21.1, which we already mentioned in Section 4. Fisher's method is based on the product statistic $p_1 \dots p_K$ (with its low values significant) and uses the fact that $-2\ln(p_1 \dots p_K)$ has the χ^2 distribution with $2K$ degrees of freedom when p_k are all independent and distributed uniformly on the interval $[0, 1]$; the p-values are the tails of the χ^2 distribution.

Simes (1986) proves a remarkable result for Rüger's family (15) under the assumption that the p-values are independent: the minimum

$$(p_1, \dots, p_K) \mapsto \min_{k \in \{1,\dots,K\}} \frac{K}{k} p_{(k)} \tag{20}$$

of Rüger's family over all k turns out to be a p-merging function. The counterpart of Simes's result still holds for e-merging functions; moreover, now the input e-values do not have to be independent. Namely,

$$(e_1, \dots, e_K) \mapsto \max_{k \in \{1,\dots,K\}} \frac{k}{K} e_{[k]}$$

is an e-merging function. This follows immediately from (18), the left-hand side of which can be replaced by its maximum over k. And it also follows from (18) that there is no sense in using this counterpart; it is better to use the arithmetic mean.

7. Experimental results. In this section, we will explore the performance of various methods of combining e-values and p-values and multiple hypothesis testing, both standard and introduced in this paper.

In order to be able to judge how significant results of testing using e-values are, Jeffreys's (1961), Appendix B, rule of thumb may be useful:

- If the resulting e-value e is below 1, the null hypothesis is supported.
- If $e \in (1, \sqrt{10}) \approx (1, 3.16)$, the evidence against the null hypothesis is not worth more than a bare mention.
- If $e \in (\sqrt{10}, 10) \approx (3.16, 10)$, the evidence against the null hypothesis is substantial.
- If $e \in (10, 10^{3/2}) \approx (10, 31.6)$, the evidence against the null hypothesis is strong.
- If $e \in (10^{3/2}, 100) \approx (31.6, 100)$, the evidence against the null hypothesis is very strong.
- If $e > 100$, the evidence against the null hypothesis is decisive.

Our discussions in this section assume that our main interest is in e-values, and p-values are just a possible tool for obtaining good e-values (which is, e.g., the case for Bayesian statisticians in their attitude towards Bayes factors and p-values; cf. Section A.1 and Appendix B). Our conclusions would have been different had our goal been to obtain good p-values.

Combining independent e-values and p-values. First, we explore combining independent e-values and independent p-values; see Figure 1. The observations are generated from the Gaussian model $N(\mu, 1)$ with standard deviation 1 and unknown mean μ. The null hypothesis is $\mu = 0$ and the alternative hypothesis is $\mu = \delta$; for Figures 1 and 2 we set $\delta := -0.1$. The observations are IID. Therefore, one observation does not carry much information about which hypothesis is true, but repeated observations quickly reveal the truth (with a high probability).

For Figures 1 and 2, all data are generated from the alternative distribution (an example with some of the data coming from the null distribution will be given in the Supplementary

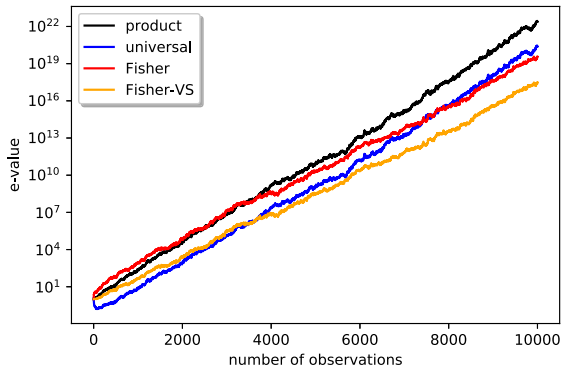

FIG. 1. *Combining p-values using Fisher's method vs combining e-values by multiplication* (*details in text*).

Material, Appendix I). For each observation, the e-value used for testing is the likelihood ratio

$$E(x) := e^{-(x-\delta)^2/2}/e^{-x^2/2} = e^{x\delta-\delta^2/2} \tag{21}$$

of the alternative probability density to the null probability density, where x is the observation. It is clear that (21) is indeed an e-variable under the null hypothesis: its expected value is 1. As the p-value we take

$$P(x) := N(x), \tag{22}$$

where N is the standard Gaussian distribution function; in other words, the p-value is found using the most powerful test, namely the likelihood ratio test given by the Neyman–Pearson lemma.

In Figure 1, we give the results for the product e-merging function (7) and Fisher's method described in the last subsection of Section 6. (The other methods that we consider are vastly less efficient, and we show them in the following figure, Figure 2.) Three of the values plotted in Figure 1 against each $K = 1, \dots, 10{,}000$ are:

- the product e-value $E(x_1) \dots E(x_K)$; it is shown as the black line;
- the reciprocal $1/p$ of Fisher's p-value p obtained by merging the first K p-values $P(x_1), \dots, P(x_K)$; it is shown as the red line;
- the VS bound applied to Fisher's p-value; it is shown as the orange line.

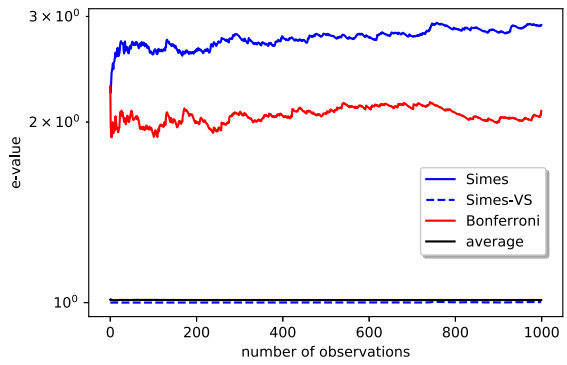

FIG. 2. *Combining p-values using Simes's and Bonferroni's methods and combining e-values using averaging* (*details in text*).

The plot depends very much on the seed for the random number generator, and so we report the median of all values over 100 seeds.

The line for the product method is below that for Fisher's over the first 2000 observations but then it catches up. If our goal is to have an overall e-value summarizing the results of testing based on the first K observations (as we always assume in this section), the comparison is unfair, since Fisher's p-values need to be calibrated. A fairer (albeit still unfair) comparison is with the VS bound, and the curve for the product method can be seen to be above the curve for the VS bound. *A fortiori*, the curve for the product method would be above the curve for any of the calibrators in the family (1).

It is important to emphasize that the natures of plots for e-values and p-values are very different. For the red and orange lines in Figure 1, the values shown for different K relate to different batches of data and cannot be regarded as a trajectory of a natural stochastic process. In contrast, the values shown by the black line for different K are updated sequentially, the value at K being equal to the value at $K-1$ multiplied by $E(x_K)$, and form a trajectory of a test martingale. Moreover, for the black line we do not need the full force of the assumption of independence of the p-values. As we discuss at the end of Section 4, it is sufficient to assume that $E(x_K)$ is a valid e-value given $x_1, \dots, x_{K-1}$; the black line in Figure 1 is then still a trajectory of a test supermartingale.

What we said in the previous paragraph can be regarded as an advantage of using e-values. On the negative side, computing good (or even optimal in some sense) e-values often requires more detailed knowledge. For example, whereas computing the e-value (21) requires the knowledge of the alternative hypothesis, for computing the p-value (22) it is sufficient to know that the alternative hypothesis corresponds to $\mu < 0$. Getting μ very wrong will hurt the performance of methods based on e-values. To get rid of the dependence on μ, we can, for example, integrate the product e-value over $\delta \sim N(0,1)$ (taking the standard deviation of 1 is somewhat wasteful in this situation, but we take the most standard probability measure). This gives the "universal" test martingale (see, e.g., Howard et al. (2020))

$$\begin{aligned} S_K &:= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-\delta^2/2) \prod_{k=1}^{K} \exp(x_k \delta - \delta^2/2) \, \mathrm{d}\delta \\ &= \frac{1}{\sqrt{K+1}} \exp\left(\frac{1}{2(K+1)} \left(\sum_{k=1}^{K} x_k \right)^2 \right). \end{aligned} \tag{23}$$

This test supermartingale is shown in blue in Figure 1. It is below the black line but at the end of the period it catches up even with the line for Fisher's method (and beyond that period it overtakes Fisher's method more and more convincingly).

Arithmetic average (4) and Simes's method (20) have very little power in the situation of Figure 1: see Figure 2, which plots the e-values produced by the averaging method, the reciprocals $1/p$ of Simes's p-values p, the VS bound for Simes's p-values, and the reciprocals of the Bonferroni p-values over 1000 observations, all averaged (in the sense of median) over 1000 seeds. They are very far from attaining statistical significance (a p-value of 5% or less) or collecting substantial evidence against the null hypothesis (an e-value of $\sqrt{10}$ or more according to Jeffreys).

Multiple hypothesis testing. Next, we discuss multiple hypothesis testing. Figure 3 shows plots of adjusted e-values and adjusted p-values resulting from various methods for small numbers of hypotheses, including Algorithms 1 and 2. The observations are again generated from the statistical model $N(\mu, 1)$.

We are testing 20 null hypotheses. All of them are $\mu = 0$, and their alternatives are $\mu = -4$. Each null hypothesis is tested given an observation drawn either from the null or from the

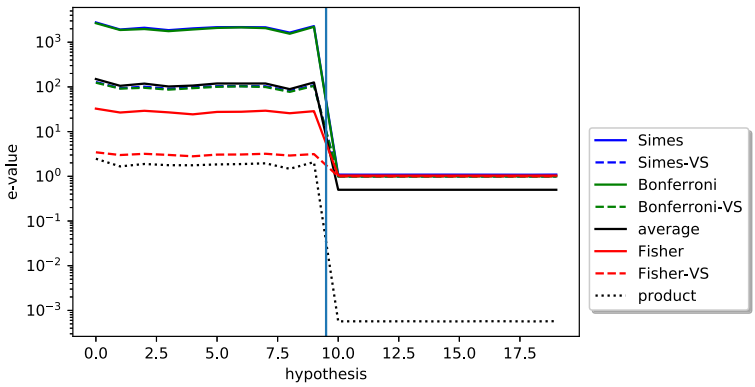

FIG. 3. *Multiple hypothesis testing for* 20 *hypotheses using p-values and e-values, with some graphs indistinguishable* (*details in text*).

alternative. The first 10 null hypotheses are false, and in fact the corresponding observations are drawn from the alternative distribution. The remaining 10 null hypotheses are true, and the corresponding observations are drawn from them rather than the alternatives. The vertical blue line at the centre of Figure 3 separates the false null hypotheses from the true ones: null hypotheses 0 to 9 are false and 10 to 19 are true. We can see that at least some of the methods can detect that the first 10 null hypotheses are false.

Since some of the lines are difficult to tell apart, we will describe the plot in words. The top two horizontal lines to the left of the vertical blue line are indistinguishable but are those labeled as Simes and Bonferroni in the legend; they correspond to e-values around 2×10^3. The following cluster of horizontal lines to the left of the vertical blue line (with e-values around 10^2) are those labeled as average, Simes-VS, and Bonferroni-VS, with average slightly higher. To the right of the vertical blue line, the upper horizontal lines (with e-values 10^0) include all methods except for average and product; the last two are visible.

Most of the methods (all except for Bonferroni and Algorithm 1) require the observations to be independent. The base p-values are (22), and the base e-values are the likelihood ratios

$$E(x) := \frac{1}{2} e^{x\delta - \delta^2/2} + \frac{1}{2} \tag{24}$$

(cf. (21)) of the "true" probability density to the null probability density, where the former assumes that the null or alternative distribution for each observation is decided by coin tossing. Therefore, the knowledge encoded in the "true" distribution is that half of the observations are generated from the alternative distribution, but it is not known that these observations are in the first half. We set $\delta := -4$ in (24), keeping in mind that accurate prior knowledge is essential for the efficiency of methods based on e-values.

A standard way of producing multiple testing procedures is applying the closure principle described in Appendix A and already implicitly applied in Section 5 to methods of merging e-values. In Figure 3, we report the results for the closures of five methods, three of them producing p-values (Simes's, Bonferroni's and Fisher's) and two producing e-values (average and product); see Section 5 for self-contained descriptions of the last two methods (Algorithms 1 and 2). For the methods producing p-values we show the reciprocals $1/p$ of the resulting p-values p (as solid lines) and the corresponding VS bounds (as dashed lines). For the closure of Simes's method, we follow the Appendix of Wright (1992), the closure of Bonferroni's method is described in Holm (1979) (albeit not in terms of adjusted p-values), and for the closure of Fisher's method we use Dobriban's (2020) FACT (FAst Closed Testing) procedure. To make the plot more regular, all values are averaged (in the sense of median) over 1000 seeds of the Numpy random number generator.

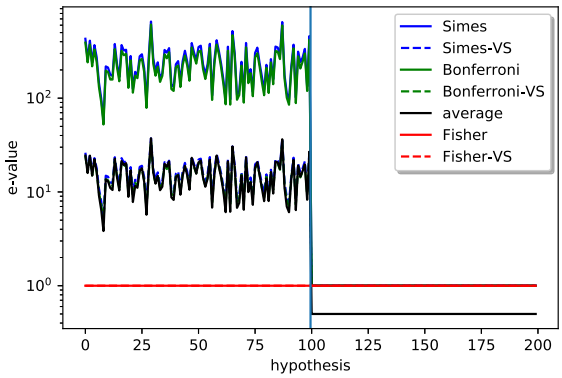

FIG. 4. *The analogue of Figure 3 without the product method, with* 200 *observations, and with some graphs indistinguishable* (*details in text*).

According to Figure 3, the performance of Simes's and Bonferroni's methods is very similar, despite Bonferroni's method not depending on the assumption of independence of the p-values. The e-merging method of averaging (i.e., Algorithm 1) produces better e-values than those obtained by calibrating the closures of Simes's and Bonferroni's methods; remember that the line corresponding to Algorithm 1 should be compared with the VS versions (blue and green dashed, which almost coincide) of the lines corresponding to the closures of Simes's and Bonferroni's methods, and even that comparison is unfair and works in favour of those two methods (since the VS bound is not a valid calibrator). The other algorithms perform poorly.

Figure 4 is an analogue of Figure 3 that does not show results for merging by multiplication (for large numbers of hypotheses its results are so poor that, when shown, differences between the other methods become difficult to see). To get more regular and comparable graphs, we use averaging (in the sense of median) over 100 seeds.

Since some of the graphs coincide, or almost coincide, we will again describe the plot in words (referring to graphs that are straight or almost straight as lines). To the left of the vertical blue line (separating the false null hypotheses 0–99 from the true null hypotheses 100–199), we have three groups of graphs: the top graphs (with e-values around 2×10^2) are those labeled as Simes and Bonferroni in the legend, the middle graphs (with e-values around 10^1) are those labeled as average, Simes-VS and Bonferroni-VS, and the bottom lines (with e-values around 10^0) are those labeled as Fisher and Fisher-VS. To the right of the vertical blue line, we have two groups of lines: the upper lines (with e-values 10^0) include all methods except for average, which is visible.

Now the graph for the averaging method (Algorithm 1) is very close to the graphs for the VS versions of the closures of Simes's and Bonferroni's methods, which is a very good result (in terms of the quality of e-values that we achieve): the VS bound is a bound on what can be achieved whereas the averaging method produces a bona fide e-value.

A key advantage of the averaging and Bonferroni's methods over Simes's and Fisher's is that they are valid regardless of whether the base e-values or p-values are independent.

8. Conclusion. This paper systematically explores the notion of an e-value, which can be regarded as a betting counterpart of p-values that is much more closely related to Bayes factors and likelihood ratios. We argue that e-values often are more mathematically convenient than p-values and lead to simpler results. In particular, they are easier to combine: the average of e-values is an e-value, and the product of independent e-values is an e-value. We apply e-values in two areas, multiple testing of a single hypothesis and testing multiple

hypotheses, and obtain promising experimental results. One of our experimental findings is that, for testing multiple hypotheses, the performance of the most natural method based on e-values almost attains the Vovk–Sellke bound for the closure of Simes's method, despite that bound being overoptimistic and not producing bona fide e-values.

Acknowledgments. The authors thank Aaditya Ramdas, Alexander Schied and Glenn Shafer for helpful suggestions. Thoughtful comments by the Associate Editor and four reviewers have led to numerous improvements in presentation and substance.

Funding. The first author was supported by Amazon, Astra Zeneca and Stena Line.

The second author was supported by NSERC grants RGPIN-2018-03823 and RGPAS-2018-522590.

SUPPLEMENTARY MATERIAL

E-values: Online supplement (DOI: 10.1214/20-AOS2020SUPP; .pdf). The online supplement discusses connections with literature and provides further theoretical results and simulation studies.

REFERENCES

BENJAMIN, D. J. and BERGER, J. O. (2019). Three recommendations for improving the use of p-values. *Amer. Statist.* **73** 186–191. MR3925724 https://doi.org/10.1080/00031305.2018.1543135

DOBRIBAN, E. (2020). Fast closed testing for exchangeable local tests. *Biometrika* **107** 761–768. MR4138990 https://doi.org/10.1093/biomet/asz082

DUAN, B., RAMDAS, A., BALAKRISHNAN, S. and WASSERMAN, L. (2019). Interactive martingale tests for the global null. Technical report. Available at arXiv:1909.07339 [stat.ME].

EMBRECHTS, P., WANG, B. and WANG, R. (2015). Aggregation-robustness and model uncertainty of regulatory risk measures. *Finance Stoch.* **19** 763–790. MR3413935 https://doi.org/10.1007/s00780-015-0273-z

FISHER, R. A. (1932). *Statistical Methods for Research Workers*, 4th ed. Oliver and Boyd, Edinburgh. Section 21.1 on combining independent p-values first appears in this edition and is present in all subsequent editions.

GOEMAN, J. J. and SOLARI, A. (2011). Multiple testing for exploratory research. *Statist. Sci.* **26** 584–597. MR2951390 https://doi.org/10.1214/11-STS356

GRÜNWALD, P., DE HEIDE, R. and KOOLEN, W. M. (2020). Safe testing. Technical report. Available at arXiv:1906.07801 [math.ST].

HARDY, G. H., LITTLEWOOD, J. E. and PÓLYA, G. (1952). *Inequalities*, 2nd ed. Cambridge Univ. Press, Cambridge. MR0046395

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6** 65–70. MR0538597

HOWARD, S. R., RAMDAS, A., MCAULIFFE, J. and SEKHON, J. (2020). Time-uniform, nonparametric, nonasymptotic confidence sequences. Technical report. Available at arXiv:1810.08240 [math.ST]. *Ann. Statist.* To appear.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Clarendon Press, Oxford. MR0187257

MARCUS, R., PERITZ, E. and GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655–660. MR0468056 https://doi.org/10.1093/biomet/63.3.655

RÜGER, B. (1978). Das maximale Signifikanzniveau des Tests: "Lehne H_0 ab, wenn k unter n gegebenen Tests zur Ablehnung führen". *Metrika* **25** 171–178. MR0526476 https://doi.org/10.1007/BF02204362

SHAFER, G. (2019). The language of betting as a strategy for statistical and scientific communication. Technical report. Available at arXiv:1903.06991 [math.ST]. *J. R. Stat. Soc., A*. To appear.

SHAFER, G. and VOVK, V. (2019). *Game-Theoretic Foundations for Probability and Finance*. Wiley, Hoboken, NJ.

SHAFER, G., SHEN, A., VERESHCHAGIN, N. and VOVK, V. (2011). Test martingales, Bayes factors and p-values. *Statist. Sci.* **26** 84–101. MR2849911 https://doi.org/10.1214/10-STS347

SHAFFER, J. P. (1986). Modified sequentially rejective multiple test procedures. *J. Amer. Statist. Assoc.* **81** 826–831.

SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751–754. MR0897872 https://doi.org/10.1093/biomet/73.3.751

VOVK, V. G. (1993). A logic of probability, with application to the foundations of statistics. *J. Roy. Statist. Soc. Ser. B* **55** 317–351. With discussion and a reply by the author. MR1224399

VOVK, V. and WANG, R. (2020). Combining p-values via averaging. Technical report. Available at arXiv:1212.4966 [math.ST]. *Biometrika*. To appear. https://doi.org/10.1093/biomet/asaa027.

VOVK, V. and WANG, R. (2021). Supplement to "E-values: Calibration, combination, and applications." https://doi.org/10.1214/20-AOS2020SUPP

WALD, A. (1950). *Statistical Decision Functions*. Wiley, New York, NY. MR0036976

WRIGHT, S. P. (1992). Adjusted p-values for simultaneous inference. *Biometrics* **48** 1005–1013.