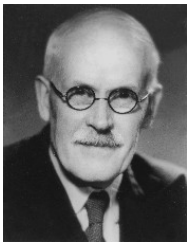
Testing with p*-values: Between p-values and e-values

Ruodu Wang[1]



Users of these tests speak of the
5 per cent. point [p-value of 5%]
in much the same way as I should
speak of the $K = 10^{-1/2}$ point
[e-value of $10^{1/2}$], and of the 1
per cent. point [p-value of 1%]
as I should speak of the
$K = 10^{-1}$ point [e-value of 10].

Project "Hypothesis testing with e-values"

Working Paper #10

November 21, 2020

Project web site:
http://alrw.net/e

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

Abstract

We introduce the notion of p*-values (p*-variables), which generalizes p-values (p-variables) in several senses. The new notion has four natural interpretations: probabilistic, operational, Bayesian, and frequentist. The simplest interpretation of a p*-value is the average of several p-values. We show that there are four equivalent definitions of p*-values. The randomized p*-test is proposed, which is a randomized version of the simple p-test. Admissible calibrators of p*-values to and from p-values and e-values are obtained with nice mathematical forms, revealing the role of p*-values as a bridge between p-values and e-values. The notion of p*-values becomes useful in many situations even if one is only interested in p-values and e-values. In particular, tests based on p*-values can be applied to improve several classic methods for p-values and e-values.

Keywords: randomized test; arbitrary dependence; average of p-values; posterior predictive p-values; calibration

Contents

1 Introduction

Hypothesis testing is usually conducted with the classic notion of p-values. E-values have been recently introduced to the statistical community by Vovk and Wang [21], and they have several advantages in contrast to p-values, especially via their connections to Bayes factors and test martingales (Shafer et al. [15]), betting scores (Shafer [14]), universal inference (Wasserman et al. [23]), anytime-valid tests (Grünwald et al. [7]), and conformal tests (Vovk [18]); see also our Section 8.

In this paper, we introduce the abstract notion of *p*-variables*, with *p*-values* as their realizations, defined via a simple inequality in stochastic order, in a way similar to p-variables and p-values. By definition, p*-variables are generalized p-variables, and they are motivated by four natural interpretations: probabilistic, operational, Bayesian, and frequentist; we will explain them in Section 3. The simplest interpretation of a p*-value is the average of several p-values; it is a posterior predictive p-value of Meng [11] in the Bayesian context. Moreover, p*-values are naturally connected to e-values as we will explain later.

In discussions where the probabilistic specification as random variables is not emphasized, we will loosely use the term "p/p*/e-values" for both p/p*/e-variables and their realizations, similarly to [20, 21], and this should be clear from the context.

There are four equivalent definitions of p*-variables: by stochastic order (Definition 2.2), by conditional probability (Theorem 3.1), by averaging p-variables (Theorem 3.2), and by randomized tests (Theorem 4.6); each of them represents a natural path to a generalization of p-values, and these paths lead to the same mathematical object of p*-values.

The randomized p*-test for p*-values is introduced in Section 4, which is a randomized version of the traditional p-test with several attractive properties; the randomization is necessary because p*-values are weaker than p-values. We explore in Sections 5 and 6 the connections among p-values, p*-values, and e-values by establishing results for admissible calibrators. Figure 1 summarizes these calibrators where the ones between p-values and e-values are obtained in [21]. Notably, for an e-value e, $(2e)^{-1}$ is a calibrated p*-value, which has an extra factor of $1/2$ compared to the standard calibrated p-value e^{-1}. A composition of the e-to-p* calibration $p^* = (2e)^{-1} \wedge 1$ and the p*-to-p calibration $p = (2p^*) \wedge 1$ leads to the unique admissible e-to-p calibration $p = e^{-1} \wedge 1$, thus showing that p*-values serve as a bridge between e-values and p-values.

We do not suggest directly using p*-values in classic statistical settings where precise (uniform on $[0, 1]$) p-values are available. Nevertheless, applying the randomized p*-test in situations where precise p-values are unavailable leads to many improvements on classic methods for p-values and e-values. We discuss and numerically illustrate such applications in detail in Sections 7 and 8. As a consequence, the notion of p*-values is useful even when one is only interested in p-values or e-values. As merging methods are useful in multiple hypothesis testing for both p-values and e-values, we study merging functions for p*-values in Section 9, which turn out to be much nicer mathematically than those for

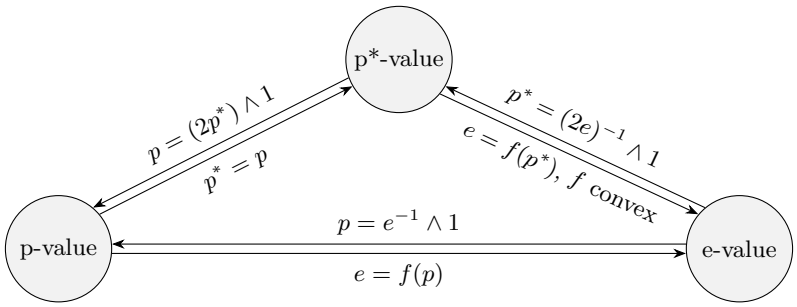


Figure 1: Calibration among p-values, p*-values and e-values, where $f : [0, 1] \to [0, \infty]$ is left-continuous and decreasing with $f(0) = \infty$ and $\int_0^1 f(t)\,\mathrm{d}t = 1$.

p-values.

The paper is written such that p-values, p*-values and e-values are treated as abstract measure-theoretical objects, following the setting of [20, 21]. Our null hypothesis is a generic and unspecified one, and it can be simple or composite; nevertheless, for the discussions of our results, it would be harmless to keep a simple hypothesis in mind as a primary example.

2 P-values, p*-values, and e-values

Following the setting of [21], we directly work with a fixed atomless probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where our (global) null hypothesis is set to be the singleton $\{\mathbb{P}\}$. As explained in Appendix D of [21], no generality is lost as all mathematical results (of the kind in this paper) are valid also for general composite hypotheses. We assume that $(\Omega, \mathcal{A}, \mathbb{P})$ is rich enough so that we can find a uniform random variable of independent of a given random vector as we wish. We first define stochastic orders, which will be used to formulate the main objects in the paper.

Definition 2.1. Let X and Y be two random variables.

1. X is first-order stochastically smaller than Y, written as $X \le_1 Y$, if $\mathbb{E}[f(X)] \le \mathbb{E}[f(Y)]$ for all increasing real functions f such that the expectations exist.

2. X is second-order stochastically smaller than Y, written as $X \le_2 Y$, if $\mathbb{E}[f(X)] \le \mathbb{E}[f(Y)]$ for all increasing concave real functions f such that the expectations exist.

Using stochastic orders, we define p-variables and e-variables (for the null hypothesis) and our new concept, called p*-variables. Here, we allow both p-variables and p*-variables to take values above one, although such values are uninteresting, and one may safely truncated them at 1.

Definition 2.2. Let U be a uniform random variable on $[0,1]$.

1. A random variable P is a *p-variable* if $U \leq_1 P$.
2. A random variable P is a *p*-variable* if $U \leq_2 P$.
3. A random variable E is an *e-variable* if $0 \leq_2 E \leq_2 1$.

Since $\leq_1$ is stronger than $\leq_2$, a p-variable is also a p*-variable, but not vice versa. Due to the close proximity between p-variables and p*-variables, we often use P for both of them; this should not create any confusion. We refer to *p-values* as realizations of p-variables, *p*-values* as those of p*-variables, and *e-values* as those of e-variables. By definition, both a p-variable and a p*-variable have a mean at least $1/2$. We allow an e-variable E to take the value ∞ (but with probability 0 under the null), which corresponds to a p-variable taking the value 0 (also with probability 0).

Recall the classic definitions of p-variables and e-variables (see [21]):

1. P is a *p-variable* if $\mathbb{P}(P \leq \alpha) \leq \alpha$ for all $\alpha \in (0,1)$;
2. E is an *e-variable* if $\mathbb{E}[E] \leq 1$ and $E \geq 0$.

The above formulations are equivalent to our Definition 2.2 because

$$U \leq_1 P \iff \mathbb{P}(P \leq \alpha) \leq \mathbb{P}(U \leq \alpha) \text{ for all } \alpha \in (0,1);$$

$$0 \leq_2 E \iff 0 \leq E; \qquad E \leq_2 1 \iff \mathbb{E}[E] \leq 1.$$

We choose to express our definition via stochastic orders to make an analogy among the three concepts of p-values, p*-values, and e-values, and stochastic orders will be a main technical tool for results in this paper.

Our main focus is the notion of p*-values, which will be motivated from four interpretations in Section 3: probabilistic, frequentist, Bayesian, and operational; each interpretation leads to a different way to construct p*-variables.

Remark 2.3. There are many equivalent conditions for the stochastic order $U \leq_2 P$. One of the most convenient conditions, which will be used repeatedly in this paper, is

$$U \leq_2 P \iff \int_0^v G_P(u)\,\mathrm{d}u \geq \frac{v^2}{2} \text{ for all } v \in (0,1), \tag{1}$$

where G_P is the left-quantile function of P; see e.g., Theorem 3.A.5 of [16].

3 Four interpretations of p*-values

3.1 Probabilistic interpretation

To explain the motivation for p*-values, we first recall the usual practice to obtain p-values. Let T be a test statistic which is a function of the observed

data, represented by a vector X. The p-variable P is usually computed from the conditional probability

$$P = F(T) = \mathbb{P}(T' \leq T|T) = \mathbb{P}(T' \leq T|X), \tag{2}$$

where F is the distribution of T, and T' is a copy of T independent of X; here and below, a copy of T is a random variable identically distributed as T. Indeed, any p-variable as a function of the data X can be obtained from (2) for some T; this will be made rigorous in Theorem 3.1 below.

Next, we show that p*-variables have a similar representation to (2). The only difference is that T here may not be a function of X; thus, T can include some unobservable random factors in the statistical experiment. We call this interpretation a probabilistic one as we will interpret both p-variables and p*-variables as conditional probabilities.

Theorem 3.1. *For $\sigma(X)$-measurable random variables $P : \Omega \to [0, \infty)$,*

(i) P is a p-variable if and only if there exists a $\sigma(X)$-measurable function T such that $P \geq \mathbb{P}(T' \leq T|X)$ where T' is a copy of T independent of X;

(ii) P is a p-variable if and only if there exists a random variable T such that $P \geq \mathbb{P}(T' \leq T|X)$ where T' is a copy of T independent of (T, X).*

Proof. For the "only-if" statement in (i), since P is a p-variable, we know that its distribution F satisfies $F(t) \leq t$ for $t \in (0, 1)$. Therefore, by setting $T = P$,

$$P \geq F(P) = F(T) = \mathbb{P}(T' \leq T|T) = \mathbb{P}(T' \leq T|X),$$

where the last equality holds since T' is independent of X. To check the "if" direction of (i), we have $\mathbb{P}(T' \leq T|X) = \mathbb{P}(T' \leq T|T) = F(T)$ where F is the distribution of T. Note that $F(T)$ is stochastically larger or equal to than a uniform random variable on $[0, 1]$, and hence $\mathbb{P}(F(T) \leq t) \leq t$.

Next, we show (ii). First, suppose that $P \geq \mathbb{P}(T' \leq T|X)$. Let U be a uniform random variable on $[0, 1]$. By Jensen's inequality, we have $\mathbb{E}[F(T)|X] \geq_2 F(T)$. Hence, $P \geq_2 \mathbb{P}(T' \leq T|X) = \mathbb{E}[F(T)|X] \geq_2 F(T) \geq_2 U$, and thus P is a p*-variable.

For the converse direction, suppose that P is a p*-variable. By Strassen's Theorem in the form of [16, Theorem 3.A.4], there exists a uniform random variable U' on $[0, 1]$ and a random variable P' identically distributed as P such that $P' \geq \mathbb{E}[U'|P']$. Let $G(\cdot|p)$ be the quantile function of a regular conditional distribution of U' given $P' = p \in [0, 1]$. Further, let V be a uniform random variable on $[0, 1]$ independent of (P, X), and $U := G(V|P)$. It is clear that (U, P) has the same law as (U', P'). Therefore, $P \geq \mathbb{E}[U|P]$. Moreover, $\mathbb{E}[U|X] = \mathbb{E}[G(V|P)|X] = \mathbb{E}[G(V|P)|P]$ since V is independent of X. Hence, $P \geq \mathbb{E}[U|P] = \mathbb{E}[U|X]$. □

3.2 Operational interpretation

Our second interpretation of p*-variables is operational: we will see that a p*-variable is precisely the arithmetic average of some p-values. This characterization relies on a recent technical result of Mao et al. [10] on the sum of standard uniform random variables.

Theorem 3.2. *A random variable is a p*-variable if and only if it is the convex combination of some p-variables. Moreover, any p*-variable can be expressed as the arithmetic average of three p-variables.*

Proof. We first show that a convex combination of p-variables is a p*-variable. Let U be a uniform random variable on $[0,1]$, $P_1, \dots, P_K$ be K p-variables, $(\lambda_1, \dots, \lambda_K)$ be an element of the standard K-simplex, and f be an increasing concave function. By monotonicity and concavity of f, we have

$$\mathbb{E}\left[f\left(\sum_{k=1}^{K} \lambda_k P_k\right)\right] \geq \mathbb{E}\left[\sum_{k=1}^{K} \lambda_k f(P_k)\right] \geq \mathbb{E}\left[\sum_{k=1}^{K} \lambda_k f(U)\right] = \mathbb{E}[f(U)].$$

Therefore, $\sum_{k=1}^{K} \lambda_k P_k \geq_2 U$ and thus $\sum_{k=1}^{K} \lambda_k P_k$ is a p*-variable.

Next, we show the second statement that any p*-variable can be written as the average of three p-variables, which also justifies the "only if" direction of the first statement.

Let P be a p*-variable satisfying $\mathbb{E}[P] = 1/2$. Note that $P \geq_2 U$ and $\mathbb{E}[P] = \mathbb{E}[U]$ together implies $P \geq_{\mathrm{cv}} U$ (see e.g., [16, Theorem 4.A.35]), where $\leq_{\mathrm{cv}}$ is the concave order, meaning that $\mathbb{E}[f(P)] \geq \mathbb{E}[f(U)]$ for all concave f. Theorem 5 of [10] says that any $P \geq_{\mathrm{cv}} U$, there exist three standard uniform random variables P_1, P_2, P_3 such that $3P = P_1 + P_2 + P_3$ (this statement is highly non-trivial). This implies that P can be written as the arithmetic average of three p-variables P_1, P_2, P_3.

Finally, assume that the p*-variable P satisfies $\mathbb{E}[P] > 1/2$. In this case, using Strassen's Theorem in the form of [16, Theorems 4.A.5 and 4.A.6], there exists a random variable Z such that $U \leq_{\mathrm{cv}} Z \leq P$. As we explained above, there exist p-variables P_1, P_2, P_3 such that $3Z = P_1 + P_2 + P_3$. For $i = 1, 2, 3$, let $\tilde{P}_i := P_i + (P - Z) \geq_1 P_i$. Note that $\tilde{P}_1, \tilde{P}_2, \tilde{P}_3$ are p-variables and $3P = \tilde{P}_1 + \tilde{P}_2 + \tilde{P}_3$. Hence, P can be written as the arithmetic average of three p-variables. □

Remark 3.3. As implied by Theorem 5 of [10], a p*-variable can always be written as the arithmetic average of n p-variables for any $n \geq 3$, but the statement is not true for $n = 2$ ([10, Proposition 1]).

As a consequence of Theorem 3.2, the set of p*-variables is convex, in contrast to that of p-variables; indeed, the set of p*-variables is the convex hull of the set of p-variables. In the next proposition we summarize some closure properties of these two sets.

Proposition 3.4. *The set of p*-variables is closed under convex combinations. Both sets of p-variables and p*-variables are closed under distribution mixtures and convergence in distribution.*

Proof. By Theorem 3.2, the set of p*-variables is the convex hull of the set of p-variables, and thus convex.

To show that the set of p*-variables is closed under distribution mixtures, it suffices to note that the stochastic orders $\leq_1$ and $\leq_2$ (indeed, any order induced by inequalities via integrals) is closed under distribution mixture.

Closure under convergence for $\leq_1$ is justified by Theorem 1.A.3 of [16], and closure under convergence for $\leq_2$ is justified by Theorem 1.5.9 of [12]. □

3.3 Bayesian interpretation

In the Bayesian context, the posterior predictive p-value of Meng [11] is a p*-value. Let X be the data vector in Section 3.1. The null hypothesis H_0 is given by $\{\psi \in \Psi_0\}$ where Ψ_0 is a subset of the parameter space Ψ on which a prior distribution is specified. The posterior predictive p-value is defined as the realization of the random variable

$$P_B := \mathbb{P}(D(X', \psi) \geq D(X, \psi)|X),$$

where D is a function (taking a similar role as test statistics), X' and X are iid conditional on ψ, and the probability is computed under the joint posterior distribution of (X', ψ). Note that P_B can be rewritten as

$$P_B = \int \mathbb{P}(D(X', y) \geq D(X, y)|X, y)\, \mathrm{d}\Pi(y|X)$$

where Π is the posterior distribution of ψ given the data X. One can check that P_B is a p*-variable by using Jensen's inequality; see Theorem 1 of [11] where $D(X, \psi)$ is assumed to be continuously distributed conditional on ψ.

In this formulation, p*-variables are obtained by integrating p-variables over the posterior distribution of some unobservable parameter. Since p*-variables are treated as measure-theoretic objects in this paper, we omit a detailed discussion of the Bayesian interpretation; nevertheless, it is reassuring that p*-values has a natural appearance in the Bayesian context as put forward by Meng [11]. One of our later results is related to an observation of [11] that two times a p*-variable is a p-variable (see Theorem 5.1).

3.4 Frequentist interpretation

In this section we illustrate another simple way to construct p*-variables. As in (2), the classic formulation of p-variables relies on the probability of T' being more extreme than T via $P := \mathbb{P}(T' \leq T|T)$. If T is continuously distributed, then the P has a standard uniform distribution. Instead, if T has a discrete distribution, then P is strictly first-order stochastically larger than a uniform

random variable on $[0,1]$. The extreme case of T being a constant leads to P being the constant 1.

One may choose either $\leq$ or $<$ to represent exceedance. Thus, we can alternatively define

$$P' := \mathbb{P}(T' < T|T).$$

However, such a definition of P' does not satisfy the requirement of a p-variable anymore in case of a discrete distribution. Hence, the definition of p-variables creates a bias in the sense of penalizing discrete distributions. To eliminate this bias, it is tempting to use

$$\tilde{P} := \frac{P + P'}{2} = \frac{1}{2}\mathbb{P}(T' \leq T|T) + \frac{1}{2}\mathbb{P}(T' < T|T) \tag{3}$$

which is again not a p-variable in general, as it takes the constant value $1/2$ in the case of a purely atomic T. Nevertheless, we can show that $\tilde{P}$ is a p*-variable.

Proposition 3.5. *The variable $\tilde{P}$ defined by (3) is a p*-variable, where T is any random variable, and T' is an independent copy of T.*

Proof. Let U be a uniform random variable on $[0,1]$ that $G_T(U) = T$ a.s., where G_T is the left-quantile function of T. Let U' be an iid copy of U. We have

$$\mathbb{P}(U' \leq U|T) = \frac{1}{2}\mathbb{P}(T' \leq T|T) + \frac{1}{2}\mathbb{P}(T' < T|T) = \tilde{P}.$$

Hence, by Theorem 3.1 (replacing X therein by our T), $\tilde{P}$ is a p*-variable. □

We note that in Theorem 3.1 (ii), $\mathbb{P}(T' \leq T|X)$ can be safely replaced by by $\mathbb{P}(T' \leq T|X)/2 + \mathbb{P}(T' < T|X)/2$. The "only-if" direction follows from the proof of Theorem 3.1 by noting that U constructed therein has a continuous distribution. The "if" direction follows from

$$\begin{aligned}\frac{1}{2}(\mathbb{P}(T' \leq T|X) + \mathbb{P}(T' < T|X)) &= \frac{1}{2}(\mathbb{E}[F(T)|X] + \mathbb{E}[F(T-)|X]) \\ &\geq_2 \frac{1}{2}(F(T) + F(T-)) \geq_2 U,\end{aligned}$$

where the second-last inequality is Jensen's, and the last inequality is implied by Proposition 3.5. Therefore, in contrast to p-variables, one can use the unbiased version (3) of the probability of exceedance to define p*-variables, thus a midway tie-breaking in the classic formulation of p-values for discrete test statistics without any correction. Discrete test statistics appear in many applications, especially when data represent frequencies or counts; see e.g., [3] and the references therein.

4 Testing with p*-values

Recall that the defining property of a p-variable P is that the standard p-test

$$\text{rejecting the null hypothesis} \iff P \leq \alpha \tag{4}$$

has size (i.e., probability of type-I error) at most α for each $\alpha \in (0,1)$. Since p*-values are a weaker version of p-values, one cannot guarantee that the test (4) for a p*-variable P has size at most α. Below, we explain how to test with p*-values, which turns out to be valid for a randomized version of (4).

The following density condition (D) for a $(0,1)$-valued random variable V will be useful; here and throughout, monotonicity is in the non-strict sense.

(D) V has a decreasing density function on $(0,1)$.

The canonical choice of V satisfying (D) is a uniform random variable on $[0, 2\alpha]$ for $\alpha \in (0, 1/2]$, which we will explain later. For a $(0,1)$-valued random variable V with mean α and a p*-variable P independent of V, we consider the test

$$\text{rejecting the null hypothesis} \iff P \le V. \tag{5}$$

The following theorem justifies the validity of the test (5) with the necessary and sufficient condition (D).

Theorem 4.1. *Suppose that V is a $(0,1)$-valued random variable with mean α.*

(i) The test (5) has size at most α for all p-variables P independent of V if and only if V satisfies (D).*

(ii) The test (5) has size at most α for all p-variables P independent of V, and the size is precisely α if P is uniformly distributed on $[0,1]$.

The proof of Theorem 4.1 relies on the following lemma.

Lemma 4.2. *For any non-negative random variable V with a decreasing density function on $(0,\infty)$ (with possibly a probability mass at 0) and any p*-variable P independent of V, we have $\mathbb{P}(P \le V) \le \mathbb{E}[V]$.*

Proof. Let F_V be the distribution function of V, which is an increasing concave function on $[0,\infty)$ because of the decreasing density. Since P is a p*-variable, we have $\mathbb{E}[F_V(P)] \ge \int_0^1 F_V(u)\,\mathrm{d}u$. Therefore,

$$\mathbb{P}(P \le V) = \mathbb{E}[\mathbb{P}(P \le V|P)] = \mathbb{E}[1 - F_V(P)] \le \int_0^1 (1 - F_V(u))\,\mathrm{d}u = \mathbb{E}[V].$$

Hence, the statement in the lemma holds. □

Proof of Theorem 4.1. We first note that (ii) is straightforward: for a uniform random variable U on $[0,1]$ independent of V, then $\mathbb{P}(U \le V) = \mathbb{E}[V]$. If $P \ge_1 U$, then $\mathbb{P}(P \le V) \le \mathbb{P}(U \le V) \le \mathbb{E}[V]$.

The "if" statement of point (i) directly follows from Lemma 4.2, noting that condition (D) is stronger than the condition in Lemma 4.2. Below, we show the "only if" statement of point (i).

Let F_V be the distribution function of V and U be a uniform random variable on $[0,1]$. Suppose that F_V is not concave on $(0,1)$. It follows that there exists

$x, y \in (0,1)$ such that $F_V(x) + F_V(y) > 2F_V((x+y)/2)$. By the right-continuity of F_V, there exists $\epsilon \in (0, |x-y|)$ such that

$$F_V(x) + F_V(y) > 2F_V\left(\frac{x+y+\epsilon}{2}\right). \tag{6}$$

Let $A[x, x+\epsilon]$ and $B = [y, y+\epsilon]$, which are disjoint intervals. Define a random variable P by

$$P = U\mathbb{1}_{\{U \notin A \cup B\}} + \frac{x+y+\epsilon}{2}\mathbb{1}_{\{U \in A \cup B\}}. \tag{7}$$

We check that P defined by (7) is a p*-variable. For any concave function g, Jensen's inequality gives

$$\begin{aligned}
\mathbb{E}[g(P)] &= \int_{[0,1]\setminus(A\cup B)} g(u)\,\mathrm{d}u + 2\epsilon g\left(\frac{x+y+\epsilon}{2}\right) \\
&\geq \int_{[0,1]\setminus(A\cup B)} g(u)\,\mathrm{d}u + \int_{A\cup B} g(u)\,\mathrm{d}u = \mathbb{E}[g(U)].
\end{aligned}$$

Hence, P is a p*-variable. It follows from (6) and (7) that

$$\begin{aligned}
\mathbb{E}\left[F_V\left(P\right)\right] &= \int_{[0,1]\setminus(A\cup B)} F_V(u)\,\mathrm{d}u + \int_{A\cup B} F_V\left(\frac{x+y+\epsilon}{2}\right)\mathrm{d}u \\
&< \int_{[0,1]\setminus(A\cup B)} F_V(u)\,\mathrm{d}u + \int_{A\cup B} \frac{F_V(x)+F_V(y)}{2}\,\mathrm{d}u \\
&= \int_{[0,1]\setminus(A\cup B)} F_V(u)\,\mathrm{d}u + \epsilon(F_V(x)+F_V(y)) \\
&= \int_{[0,1]\setminus(A\cup B)} F_V(u)\,\mathrm{d}u + \int_A F_V(x)\,\mathrm{d}u + \int_B F_V(y)\,\mathrm{d}u \\
&\leq \int_0^1 F_V(u)\,\mathrm{d}u = 1 - \mathbb{E}[V] = 1-\alpha.
\end{aligned}$$

Therefore,

$$\mathbb{P}\left(P \leq V\right) = 1 - \mathbb{E}\left[F_V\left(P\right)\right] > 1 - (1-\alpha) = \alpha.$$

Since this contracts the validity requirement, we know that V has to have a concave distribution function, and hence a decreasing density on $(0,1)$. □

Lemma 4.2 gives $\mathbb{P}(P \leq V) \leq \mathbb{E}[V]$ for V possibly taking values larger than 1 and possibly having a probability mass at 0. We are not interested designing a random threshold with positive probability to be 0 or larger than 1, but this result will become helpful in Section 8. Since condition (D) implies $\mathbb{E}[V] \leq 1/2$, we will assume $\alpha \in (0, 1/2]$, which is certainly harmless for practice.

With the help of Theorem 4.1, we formally define α-random thresholds and the randomized p*-test.

Definition 4.3. For a significance level $\alpha \in (0, 1/2]$, an *α-random threshold* V is a $(0, 1)$-valued random variable independent of the test statistics (a p*-variable P in this section) with mean α satisfying (D). For an α-random threshold V and a p*-variable P, the *randomized p*-test* is given by (5), i.e., rejecting the null hypothesis $\iff P \le V$.

Theorem 4.1 implies that the randomized p*-test always has size at most α, just like the classic p-test (4). Since the randomized p*-test (5) has size equal to α if P is uniformly distributed on $[0, 1]$, the size α of the randomized p*-test cannot be improved in general.

The drawback of the randomized p*-test is also obvious: an extra layer of randomization is needed, and hence, like any other randomized methods, different scientists may arrive at different statistical conclusions for the same data set generating the p*-value. Unfortunately, because of assumption (D), which is necessary for validity by Theorem 4.1, we cannot reduce the α-random threshold V to a deterministic α. This undesirable feature is the price one has to pay when a p-variable is weakened to a p*-variable.

If one needs to test with a deterministic threshold, then $\alpha/2$ needs to be used instead of α. In other words, the test

$$\text{rejecting the null hypothesis } \iff P \le \alpha/2 \tag{8}$$

has size α for all p*-variable P. The validity of (8) was noted by Meng [11], and it is a direct consequence of Theorem 5.1 below. Unlike the random threshold $\mathrm{U}[0, 2\alpha]$ which gives a size precisely α in realistic situations, the deterministic threshold $\alpha/2$ is often overly conservative in practice (see discussions in [11, Section 5]), but it cannot be improved in general when testing with the average of p-variables ([20, Proposition 3]); recall that the average of p-variables is a p*-variable.

We will see an important application of the randomized p*-test in Section 7, leading to new tests on the weighted average of p-values, which can be made deterministic if one of the p-values is independent of the others. Moreover, the randomized p*-test can be used to improve the power of tests with e-values and martingales in Section 8.

As mentioned above, the extra randomness introduced by the random threshold V is often considered undesirable. One may wish to choose V such that the randomness is minimized. The next result shows that $V \sim \mathrm{U}[0, 2\alpha]$ is the optimal choice if the randomness is measured by variance or convex order.

Proposition 4.4. *For any α-random threshold V, we have* $\mathrm{var}(V) \ge \alpha^2/3$, *and this smallest variance is attained by $V^* \sim \mathrm{U}[0, 2\alpha]$. Moreover, $\mathbb{E}[f(V)] \ge \mathbb{E}[f(V^*)]$ for any convex function f (hence $V \le_2 V^*$ holds).*

Proof. We directly show $\mathbb{E}[f(V)] \ge \mathbb{E}[f(V^*)]$ for all convex functions f, which implies the statement on variance as a special case since the mean of V is fixed as α. Note that V has a concave distribution function F_V on $[0, 1]$, and V^* has a linear distribution function F_{V^*} on $[0, 2\alpha]$. Moreover, they have the same

mean. Hence, there exists $t \in [0, 2\alpha]$ such that $F_V(x) \geq F_{V^*}(x)$ for $x \leq t$ and $F_V(x) \leq F_{V^*}(x)$ for $x \geq t$. This condition is sufficient for $\mathbb{E}[f(V)] \geq \mathbb{E}[f(V^*)]$ by Theorem 3.A.44 of [16]. □

Combining Theorem 4.1 and Proposition 4.4, the canonical choice of the threshold in the randomized p*-test has a uniform distribution on $[0, 2\alpha]$.

We note that it is also possible to use some V with mean less than α and variance less than $\alpha^2/3$. This reflects a tradeoff between power and variance. Such a random threshold does not necessarily have a decreasing density. For instance, the point-mass at $\alpha/2$ is a valid choice; the next proposition gives some other choices.

Proposition 4.5. *Let V be an α-random threshold and V' is a random variable satisfying $V' \leq_1 V$. We have $\mathbb{P}(P \leq V') \leq \alpha$ for arbitrary p*-variable P independent of V'.*

Proof. Let F be the distribution function of $\bar{P}$. For any increasing function f, we have $\mathbb{E}[f(V')] \leq \mathbb{E}[f(V)]$, which follows from $V' \leq_1 V$. Hence, we have

$$\mathbb{P}(P \leq V') = \mathbb{E}[F(V')] \leq \mathbb{E}[F(V)] = \mathbb{P}(P \leq V) \leq \alpha,$$

where the last inequality follows from Theorem 4.1. □

Proposition 4.5 can be applied to a special situation where a p-variable P and an independent p*-variable P^* are available for the same null hypothesis. Note that in this case $2\alpha(1 - P) \leq_1 V \sim \mathrm{U}[0, 2\alpha]$. Hence, by Proposition 4.5, the test

$$\text{rejecting the null hypothesis} \iff \frac{P^*}{2(1-P)} \leq \alpha. \tag{9}$$

has size at most α. Alternatively, using the fact that $\mathbb{P}(P \leq 2\alpha(1 - P^*)) \leq \alpha$ implied by Theorem 4.1 (ii), we can design a test

$$\text{rejecting the null hypothesis} \iff \frac{P}{2(1-P^*)} \leq \alpha. \tag{10}$$

The tests (9) and (10) both have a deterministic threshold of α. This observation will be useful in Section 7.

Before ending this section, we give a further result showing that p*-variables admit another equivalent definition: p*-variables can be defined by the randomized p*-test (5) just like p-variables are defined by the deterministic p-test (4). For this result, we only need the size requirement to hold for uniformly distributed α-random thresholds.

Theorem 4.6. *Let $V_\alpha \sim \mathrm{U}[0, 2\alpha]$ and a random variable P be independent of V_α. Then $\mathbb{P}(P \leq V_\alpha) \leq \alpha$ for all $\alpha \in (0, 1/2]$ if and only if P is a p*-variable.*

Proof. The "if" statement is implied by Theorem 4.1. To show the "only if" statement, denote by F_P the distribution function of P and F_U be the distribution function of a uniform random variable U on $[0,1]$. We have

$$\alpha \geq \mathbb{P}(P \leq V_\alpha) = \int_0^{2\alpha} \frac{F_P(u)}{2\alpha} \,\mathrm{d}u.$$

Therefore, for $v \in (0,1]$, we have

$$\int_0^v F_P(u) \,\mathrm{d}u \leq \frac{v^2}{2} = \int_0^v u \,\mathrm{d}u = \int_0^v F_U(u) \,\mathrm{d}u.$$

By Theorem 4.A.2 of [16], the above inequality implies $U \leq_2 P$. Hence, P is a p*-variable. □

Theorem 4.6 implies that p*-variables are precisely test statistics which can pass the randomized p*-test with the specified probability, thus a further equivalent definition of p*-variables. As a consequence, the randomized p*-test cannot be applied to objects more general than the class of p*-variables.

5 Calibration between p-values and p*-values

In this section, we discuss calibration between p-values and p*-values. A *p-to-p* calibrator* is an increasing function $f : [0,1] \to [0,\infty)$ that transforms p-variables to p*-variables, and a *p*-to-p calibrator* is an increasing function $g : [0,1] \to [0,\infty)$ which transforms in the reverse direction. Clearly, the values of p-values larger than 1 are irrelevant, and hence we restrict the domain of all calibrators in this section to be $[0,1]$; in other words, input p-variables and p*-variables larger than 1 will be treated as 1. A calibrator is said to be *admissible* if it is not strictly dominated by another calibrator of the same kind (for calibration to p-values and p*-values, f dominates g means $f \leq g$, and for calibration to e-values in Section 6 it is the opposite inequality).

Theorem 5.1. *(i) A p-variable is a p*-variable, and the sum of two p*-variables is a p-variable.*

(ii) The p-to-p calibrator $u \mapsto (2u) \wedge 1$ dominates all other p*-to-p calibrators.*

(iii) An increasing function f on $[0,1]$ is an admissible p-to-p calibrator if and only if f is left-continuous, $f(0) = 0$, $\int_0^v f(u) \,\mathrm{d}u \geq v^2/2$ for all $v \in (0,1)$, and $\int_0^1 f(u) \,\mathrm{d}u = 1/2$.*

Proof. Let U be a uniform random variable on $[0,1]$.

(i) The first statement is trivial by definition. For the second statement, let P_1, P_2 be two p*-variables. For any $\epsilon \in (0,1)$,

$$\begin{aligned}\mathbb{P}(P_1+P_2 \le \epsilon) &= \mathbb{E}[\mathbb{1}_{\{P_1+P_2\le\epsilon\}}] \\ &\le \mathbb{E}\left[\frac{1}{\epsilon}(2\epsilon - P_1 - P_2)_+\right] \\ &\le \mathbb{E}\left[\frac{1}{\epsilon}(\epsilon - P_1)_+\right] + \mathbb{E}\left[\frac{1}{\epsilon}(\epsilon - P_2)_+\right] \\ &\le 2\mathbb{E}\left[\frac{1}{\epsilon}(\epsilon - U)_+\right] = \epsilon,\end{aligned}$$

where the last inequality is because $U \le_2 P_1, P_2$ and $u \mapsto (\epsilon - u)_+$ is convex and decreasing. Therefore, $P_1 + P_2$ is a p-variable.

(ii) The validity of the calibrator $u \mapsto (2u) \wedge 1$ is implied by (i), and below we show that it dominates all others. For any function g on $[0,\infty)$, suppose that $g(u) < 2u$ for some $u \in (0, 1/2]$. Consider the random variable V defined by $V = U1_{\{U>2u\}} + u1_{\{U\le 2u\}}$. Clearly, V is a p*-variable. Note that

$$\mathbb{P}(g(V) \le g(u)) \ge \mathbb{P}(U \le 2u) = 2u > g(u),$$

implying that g is not a p*-to-p calibrator. Hence, any p*-to-p calibrator g satisfies $g(u) \ge 2u$ for all $u \in (0, 1/2]$, thus showing that $u \mapsto (2u) \wedge 1$ dominates all p*-to-p calibrators.

(iii) By (1), we know that $f(U) \ge_2 U$, and thus f is a valid p-to-p* calibrator. To show its admissibility, it suffices to notice that f is left-continuous (lower semi-continuous) function on $[0,1]$, and if $g \le f$ and $g \ne f$, then $\int_0^1 g(u)\,\mathrm{d}u < 1/2$, implying that $g(U)$ cannot be a p*-variable. □

Theorem 5.1 (ii) states that two times a p*-value is a p-value, and this is the best calibrator that works for all p*-values. This observation justifies the deterministic threshold $\alpha/2$ in the test (8) for p*-values, as mentioned in Section 4. Although Theorem 5.1 (iii) implies that there are many admissible p-to-p* calibrators, it seems that there is no obvious reason to use anything other than the identity in (i) when calibrating from p-values to p*-values. Finally, we note that the conditions in Theorem 5.1 (iii) imply that the range of f is contained in $[0,1]$, an obvious requirement for an admissible p-to-p* calibrator.

6 Calibration between p*-values and e-values

Next, we discuss calibration between e-values and p*-values. A *p*-to-e calibrator* is a decreasing function $f : [0,1] \to [0,\infty]$ that transforms p-variables to p*-variables, and an *e-to-p* calibrator* $g : [0,\infty] \to [0,1]$ is a decreasing function which transforms in the reverse direction. We include $e = \infty$ in the calibrators, which corresponds to $p = 0$.

First, since a p-variable is a p*-variable, any p*-to-e calibrator is also a p-to-e calibrator. Hence, the set of p*-to-e calibrators is contained in the set of p-to-e calibrators. By Proposition 2.1 of [21], any admissible p-to-e calibrator $f : [0,1] \to [0,\infty]$ is a decreasing function such that $f(0) = \infty$, f is left-continuous, and $\int_0^1 f(t)\,\mathrm{d}t = 1$. Below we show that some of these admissible p-to-e calibrators are also p*-to-e calibrators.

Theorem 6.1. *(i) A convex p-to-e calibrator is a p*-to-e calibrator.*

(ii) A convex admissible p-to-e calibrator is an admissible p-to-e calibrator.*

(iii) An admissible p-to-e calibrator is a p-to-e calibrator if and only if it is convex.*

Proof. (i) Let f be a convex p-to-e calibrator. Note that $-f$ is increasing and concave. For any $[0,1]$-valued p*-variable P, by definition, we have $\mathbb{E}[-f(P)] \geq \mathbb{E}[-f(U)]$. Hence,

$$\mathbb{E}[f(P)] = -\mathbb{E}[-f(P)] \leq -\mathbb{E}[-f(U)] = \mathbb{E}[f(U)] \leq 1.$$

Since a $[0,\infty)$-valued p*-variable is first-order stochastically larger than some $[0,1]$-valued p*-variable (e.g., [16, Theorem 4.A.6]), we know $\mathbb{E}[f(P)] \leq 1$ for all p*-variables P. Thus, f is a p*-to-e calibrator.

(ii) From (i), a convex admissible p-to-e calibrator f is a p*-to-e calibrator. Since the class of p-to-e calibrators is larger than the class of p*-to-e calibrators, f is not strictly dominated by any p*-to-e calibrator.

(iii) We only need to show the "only if" direction, since the "if" direction is implied by (i). Suppose that a non-convex function f is an admissible p-to-e calibrator. Since f is not convex, there exist two points $t, s \in [0,1]$ such that

$$f(t) + f(s) < 2f\left(\frac{t+s}{2}\right).$$

Left-continuity of f implies that there exists $\epsilon \in (0, |t-s|)$ such that

$$f(t-\epsilon) + f(s-\epsilon) < 2f\left(\frac{t+s}{2}\right).$$

Note that

$$\int_{t-\epsilon}^{t} \left(f\left(\frac{t+s}{2}\right) - f(u)\right) \mathrm{d}u \geq \epsilon \left(f\left(\frac{t+s}{2}\right) - f(t-\epsilon)\right),$$

and the inequality also holds if the positions of s and t are flipped. Hence, by letting $A = [t-\epsilon, t] \cup [s-\epsilon, t]$, we have

$$\begin{aligned} &\int_A \left(f\left(\frac{t+s}{2}\right) - f(u)\right) \mathrm{d}u \\ &\geq \epsilon \left(2f\left(\frac{t+s}{2}\right) - f(t-\epsilon) - f(s-\epsilon)\right) > 0. \end{aligned} \tag{11}$$

Let U be a uniform random variable on $[0,1]$ and P be given by

$$P = U\mathbb{1}_{\{U \notin A\}} + \frac{t+s}{2}\mathbb{1}_{\{U \in A\}}.$$

For any increasing concave function g and $x \in [t-\epsilon, t]$ and $y \in [s-\epsilon, s]$, we have

$$2g\left(\frac{t+s}{2}\right) \geq g(t) + g(s) \geq g(x) + g(y).$$

Therefore, $\mathbb{E}[g(P)] \geq \mathbb{E}[g(U)]$, and hence $U \leq_2 P$. Thus, P is a p*-variable. Moreover, using (11), we have

$$\mathbb{E}[f(P)] = \int_0^1 f(u)\,\mathrm{d}u + \int_A \left(f\left(\frac{t+s}{2}\right) - f(u)\right)\,\mathrm{d}u > \int_0^1 f(u)\,\mathrm{d}u = 1.$$

Hence, f is not a p*-to-e calibrator. Thus, f has to be convex if it is both an admissible p-to-e calibrator and a p*-to-e calibrator. □

All practical examples of p-to-e calibrators are convex and admissible; see [21, Section 2 and Appendix B] for a few classes (which are all convex). By Theorem 6.1, all of these calibrators are admissible p*-to-e calibrators. A popular class of p-to-e calibrators is given by, for $\kappa \in (0,1)$,

$$p \mapsto \kappa p^{\kappa-1}, \quad p \in [0,1]. \tag{12}$$

Another simple choice, proposed by Shafer [14], is

$$p \mapsto p^{-1/2} - 1, \quad p \in [0,1]. \tag{13}$$

Clearly, the p-to-e calibrators in (12) and (13) are convex and thus they are p*-to-e calibrators.

Next, we discuss the other direction, namely e-to-p* calibrators. As shown by [21], there is a unique admissible e-to-p calibrator, which is given by $e \mapsto e^{-1} \wedge 1$. Since the set of p*-values is larger than that of p-values, the above e-to-p calibrator is also an e-to-p* calibrator. The interesting questions are whether there is any e-to-p* calibrator stronger than $e \mapsto e^{-1} \wedge 1$, and whether an admissible e-to-p* calibrator is also unique.

The constant map $e \mapsto 1/2$ is an e-to-p* calibrator since $1/2$ is a constant p*-variable. If there exists an e-to-p* calibrator f which dominates all other e-to-p* calibrators, then it is necessary that $f(e) \leq 1/2$ for all $e \geq 0$; however this would imply $f = 1/2$ since any p*-variable has mean at least $1/2$. Since $e \mapsto 1/2$ does not dominate $e \mapsto e^{-1} \wedge 1$, we conclude that there is no e-to-p* calibrator which dominates all others, in contrast to the case of e-to-p calibrators.

Nevertheless, we have an optimality result with a slight and natural modification. We say that an e-to-p* calibrator f *essentially dominates* another e-to-p* calibrator f' if $f(e) \leq f'(e)$ whenever $f'(e) < 1/2$. That is, we only require dominance when the calibrated p*-value is useful (relatively small); this consideration is similar to the essential domination of e-merging functions in [21]. It turns out that the e-to-p calibrator $e \mapsto e^{-1} \wedge 1$ can be improved by a factor of $1/2$, which essentially dominates all other e-to-p* calibrators.

Theorem 6.2. *The e-to-p* calibrator $e \mapsto (2e)^{-1} \wedge 1$ essentially dominates all other e-to-p* calibrators.*

Proof. First, we show that $f : e \mapsto (2e)^{-1} \wedge 1$ is an e-to-p* calibrator. Clearly, it suffices to show that $1/(2E)$ is a p*-variable for any e-variable E with mean 1, since any e-variable with mean less than 1 is dominated by an e-variable with mean 1. Let δ_x be the point-mass at x.

Assume that E has a two-point distribution (including the point-mass δ_1 as a special case). With $\mathbb{E}[E] = 1$, the distribution F_E of E can be characterized with two parameters $p \in (0,1)$ and $a \in (0, 1/p]$ via

$$F_E = p\delta_{1+(1-p)a} + (1-p)\delta_{1-pa}.$$

The distribution F_P of $P := 1/(2E)$ (we allow P to take the value ∞ in case $a = 1/p$) is given by

$$F_P = p\delta_{1/(2+2(1-p)a)} + (1-p)\delta_{1/(2-2pa)}.$$

Let G_P be the left-quantile function of P on $(0,1)$. We have

$$G_P(t) = \frac{1}{2+2(1-p)a}\mathbb{1}_{\{t\in(0,p]\}} + \frac{1}{2+2pa}\mathbb{1}_{\{t\in(p,1)\}}.$$

Define two functions g and h on $[0,1]$ by $g(v) := \int_0^v G_P(u)\,\mathrm{d}u$ and $h(v) := v^2/2$. For $v \in (0,p]$, we have, using $a \le 1/p$,

$$g(v) = \int_0^v G_P(u)\,\mathrm{d}u = \frac{v}{2+2(1-p)a} \ge \frac{v}{2+2(1-p)/p} = \frac{vp}{2} \ge \frac{v^2}{2} = h(v).$$

Moreover, Jensen's inequality gives

$$g(1) = \int_0^1 G_P(u)\,\mathrm{d}u = \mathbb{E}[P] = \mathbb{E}\left[\frac{1}{2E}\right] \ge \frac{1}{2\mathbb{E}[E]} = \frac{1}{2} = h(1).$$

Since g is linear on $[p,1]$, and v is convex, $g(p) \le h(p)$ and $g(1) \le h(1)$ imply $g(v) \le h(v)$ for all $v \in [p,1]$. Therefore, we conclude that $g \le h$ on $[0,1]$, namely

$$\int_0^v G_P(u)\,\mathrm{d}u \le \frac{v^2}{2} \quad \text{for all } v \in [0,1].$$

Using (1), we have that P is a p*-variable.

For a general e-variable E with mean 1, its distribution can be rewritten as a mixture of two-point distributions with mean 1 (see e.g., the construction in Lemma 2.1 of [22]). Since the set of p*-variables is closed under distribution mixtures (Proposition 3.4), we know that $f(E)$ is a p*-variable. Hence, f is an e-to-p* calibrator.

To show that f essentially dominates all other e-to-p* calibrators, we take any e-to-p* calibrator f'. Using Theorem 5.1, the function $e \mapsto (2f'(e)) \wedge 1$

is an e-to-p calibrator. Using Proposition 2.2 of [21], any e-to-p calibrator is dominated by $e \mapsto e^{-1} \wedge 1$, and hence

$$(2f'(e)) \wedge 1 \geq e^{-1} \wedge 1 \quad \text{for } e \in [0, \infty),$$

which in term gives $f'(e) \geq (2e)^{-1}$ for $e \geq 1$. Since f' is decreasing, we know that $f'(e) < 1/2$ implies $e > 1$. For any $e \geq 0$ with $f'(e) < 1/2$, we have $f'(e) \geq f(e)$, and thus f essentially dominates f'. □

The result in Theorem 6.2 shows that the unique admissible e-to-p calibrator $e \mapsto e^{-1} \wedge 1$ can actually be achieved by a two-step calibration: first use $p^* = (2e)^{-1} \wedge 1$ to get a p*-value, and then use $p = (2p^*) \wedge 1$ to get a p-value.

On the other hand, all p-to-e calibrators f in [21] are convex, and they can be seen as a composition of the calibrations $p^* = p$ and $e = f(p^*)$. Therefore, p*-values serve as an intermediate step in both directions of calibration between p-values and e-values, although one of the directions is less interesting since the p-to-p* calibrator is an identity.

Figure 1 in the Introduction illustrates our recommended calibrators among p-values, p*-values and e-values based on Theorems 5.1, 6.1 and 6.2, and they are all admissible.

Example 6.3. Suppose that U is uniformly distributed on $[0, 1]$. Using the calibrator (12), for $\kappa \in (0, 1)$, $E := \kappa U^{\kappa-1}$ is an e-variable. By Theorem 6.2, we know that $P := (2E)^{-1}$ is a p*-variable. Below we check this directly. The quantile function G_P of P satisfies

$$G_P(u) = \frac{u^{1-\kappa}}{2\kappa}, \quad u \in (0, 1).$$

Using $\kappa(2 - \kappa) \leq 1$ for all $\kappa \in (0, 1)$, we have

$$\int_0^v G_P(u)\,\mathrm{d}u = \frac{v^{2-\kappa}}{2\kappa(2-\kappa)} \geq \frac{v^{2-\kappa}}{2} \geq \frac{v^2}{2}, \quad v \in (0, 1).$$

Hence, P is a p*-variable by verifying (1). Moreover, for $\kappa \in (0, 1/2]$, P is even a p-variable, since $G_P(u) \geq u$ for $u \in (0, 1)$.

In the next result, we show that a p*-variable obtained from the calibrator in Theorem 6.2 is a p-variable under a further condition (D'):

(D') $E \leq_1 E'$ for some e-variable E' which has a decreasing density on $(0, \infty)$.

In particular, condition (D') is satisfied if E itself has a decreasing density on $(0, \infty)$. Examples of e-variables satisfying (D') are those obtained from applying a non-constant convex p-to-e calibrator f with $f(1) = 0$ to any p-variable, e.g., the p-to-e calibrator (13) but not (12); this is because convexity of the calibrator yields a decreasing density when applied to a uniform p-variable.

Proposition 6.4. *For any e-variable E, $P := (2E)^{-1} \wedge 1$ is a p*-variable, and if E satisfies (D'), then P is a p-variable.*

Proof. The first statement is implied by Theorem 6.2. For the second statement, we note that 1/2 is a p*-variable. For $\alpha \in (0,1)$, applying (5) with $P = 1/2$ and $V = \alpha E$, we obtained a p*-test with a random threshold with mean at most α. Using Proposition 4.5, this test has size at most α, that is, $\mathbb{P}(2E \geq 1/\alpha) \leq \alpha$. Hence, P is a p-variable. □

Remark 6.5. In a spirit similar to Proposition 6.4, smoothing techniques leading to an extra factor of 2 in the Markov inequality have been studied by Huber [8].

7 Testing with averages of p-values

In this section we illustrate applications of the randomized p*-test to tests with averages of arbitrarily dependent p-values.

Let $P_1, \dots, P_K$ be K p-variables for a global null hypothesis which are generally not independent. Vovk and Wang [20] proposed testing using generalized means of the p-values, so that the type-I error is controlled at a level α under arbitrage dependence. We focus on the weighted (arithmetic) average $\bar{P} := \sum_{k=1}^{K} w_k P_k$ for some weights $w_1, \dots, w_K \geq 0$ with $\sum_{k=1}^{K} w_k = 1$. In case $w_1 = \cdots = w_K = 1/K$, we speak of the arithmetic average.

The method of [20] on arithmetic average is given by

$$\text{rejecting the null hypothesis} \iff \bar{P} \leq \alpha/2. \tag{14}$$

We will call (14) the *arithmetic averaging test*. The extra factor of 1/2 is needed to compensate for arbitrary dependence among p-values. Since $\bar{P}$ is a p*-variable by Theorem 3.2, the test (14) is a special case of (8). This method is quite conservative, and it often has relatively low power compared to the Bonferroni correction and other similar methods unless p-values are very highly correlated, as illustrated by the numerical experiments in [20].

To enhance the power of the test (14), we apply the randomized p*-test in Section 4 to design the *randomized averaging test* by

$$\text{rejecting the null hypothesis} \iff \bar{P} \leq V. \tag{15}$$

where V is an α-random threshold independent of $(P_1, \dots, P_K)$. Comparing the fixed-threshold test (14) and the randomized averaging test (15) with $V \sim \mathrm{U}[0, 2\alpha]$, there is a 3/4 probability that the randomized averaging test has a better power, with the price of randomization.

Next, we consider a special situation, where a p-variable among $P_1, \dots, P_K$ is independent of the others under the null hypothesis. In this case, we can apply (9), and the resulting test is no longer randomized, as it is determined by the observed p-values.

Without loss of generality, assume that P_1 is independent of $(P_2, \dots, P_K)$. Let $\bar{P}_{(-1)} := \sum_{k=2}^{K} w_k P_k$ be a weighted average of $(P_2, \dots, P_K)$. Using $\bar{P}_{(-1)}$ as the p*-variable, the test (9) becomes

$$\text{rejecting the null hypothesis} \iff \bar{P}_{(-1)} + 2\alpha P_1 \leq 2\alpha. \tag{16}$$

Following directly from the validity of (9), for any p-variables $P_1, \dots, P_K$ with P_1 independent of $(P_2, \dots, P_K)$, the test (16) has size at most α.

Comparing (16) with (14), we rewrite (16) as

$$\text{rejecting the null hypothesis } \iff \bar{P} := \sum_{k=1}^{n} w'_k P_k \le \frac{2\alpha}{1+2\alpha},$$

where

$$w'_1 = \frac{2\alpha}{1+2\alpha} \quad \text{and} \quad w'_k = \frac{w_k}{1+2\alpha},\ k = 2, \dots, K.$$

Note that $\bar{P}$ is a weighted average of $(P_1, \dots, P_K)$. Since α is small, the rejection threshold is increased by almost three times, compared to the test (14) applied to $\sum_{k=1}^{n} w'_k P_k$ using the threshold $\alpha/2$. For this reason, we will call (16) the *enhanced averaging test*.

In particular, if $2\alpha(K-1) = 1$, and $\bar{P}_{(-1)}$ is the arithmetic average of $(P_2, \dots, P_K)$, then $\bar{P}$ is the arithmetic average of $(P_1, \dots, P_K)$. For instance, if $K = 51$ and $\alpha = 0.01$, then the rejection condition for test (16) is $\bar{P} \le 1/51$, and the rejection condition for (14) is $\bar{P} \le 1/200$.

Simulation experiments

We compare by simulation the performance of a few tests via merging p-values. For the purpose of illustration, we conduct correlated z-tests for the mean of normal samples with variance 1. More pecisely, the null hypothesis H_0 is $\mathrm{N}(0,1)$ and the alternative is $\mathrm{N}(\delta,1)$ for some $\delta > 0$. The p-variables $P_1, \dots, P_K$ are specified as $P_k = 1 - \Phi(X_k)$ from the Neyman-Pearson lemma, where Φ is the standard normal distribution function, and $X_1, \dots, X_K$ are generated from $\mathrm{N}(\delta, 1)$ with pair-wise correlation ρ. As illustrated by the numerical studies in [20], the arithmetic average test performs poorly unless p-values are strongly correlated. Therefore, we consider the cases where p-values are highly correlated, e.g., parallel experiments with shared data or scientific objects. We set $\rho = 0.9$ in our simulation studies; this choice is harmless as we are interested in the relative performance of the averaging methods in this section, instead of their performance against other methods (such as method of Simes [17]) that are known work well for lightly correlated or independent p-values.

The significance level α is set to be 0.01. For a comparison, we consider the following tests:

(a) the arithmetic averaging test (14): reject H_0 if $\bar{P} \le \alpha/2$;

(b) the randomized averaging test (15): reject H_0 if $\bar{P} \le V$ where $V \sim \mathrm{U}[0, 2\alpha]$ independent of $\bar{P}$;

(c) the Bonferroni method: reject H_0 if $\min(P_1, \dots, P_K) \le \alpha/K$;

(d) the Simes method: reject H_0 if $\min_k(KP_{(k)}/k) \le \alpha$ where $P_{(k)}$ is the k-th smallest p-value;

(e) the harmonic averaging test of [20]: reject H_0 if $(\sum_{k=1}^{K} P_k^{-1})^{-1} \le \alpha/c_K$ where $c_K > 1$ is a constant in [20, Proposition 6].

The validity (size no larger than α) of the Simes method is guaranteed under some dependence conditions on the p-values; see [13, 1]. Moreover, as shown recently by Vovk et al. [19, Theorem 6], the Simes method dominates any symmetric and deterministic p-merging method valid for arbitrary dependence (such as the (a), (c) and (e); the Simes method itself is not valid for arbitrary dependence).

In the second setting, we assume that one of the p-variables (P_1 without loss of generality) is independent of the rest, and the rest p-variables have a pair-wise correlation of $\rho = 0.9$. For this setting, we further include

(f) the enhanced averaging test (16): reject H_0 if $\bar{P}_{(-1)} + 2\alpha P_1 \le 2\alpha$.

The power (i.e., the probability of rejection) of each test is computed from the average of 10,000 replications for varying signal strength δ and for $K \in \{20, 100, 500\}$. Results are reported in Figure 2.

In the first setting of correlated p-values, the randomized averaging test (b) improves the performance of (a) uniformly, at the price of randomization. The Bonferroni method (d) and the harmonic averaging test (e) perform poorly and are both penalized significantly as K increases. None of these methods visibly outperforms the Simes method, although in some situations the test (b) performs comparably to the Simes method.

In the second setting where an independent p-value exists, the enhanced arithmetic test (f) performs quite well; it outperforms the Simes method for most parameter values especially for small signal strength δ. This illustrates the significant improvement via incorporating an independent p-value.

We remark that the averaging methods (a), (b) and (f) should not be used in situations in which correlation among p-values is known to be not very strong. This is because the arithmetic mean does not benefit from an increasing number of independent p-values of similar strength, unlike the methods of Bonferroni and Simes.

8 Testing with e-values and martingales

In this section we discuss applications of the randomized p*-test to tests with e-values and martingales. Note that e-values and test martingales are usually used for purposes more than rejecting a null hypothesis whiling controlling type-I error; in particular they offer anytime validity and different interpretations of statistical evidence. We compare the power of several methods here for a better understanding of their performance, while keeping in mind that single-run detection power (which is maximized by p-values if they are available) is not the only purpose of e-values.

Suppose that E is an e-variable, usually obtained from likelihood ratios or

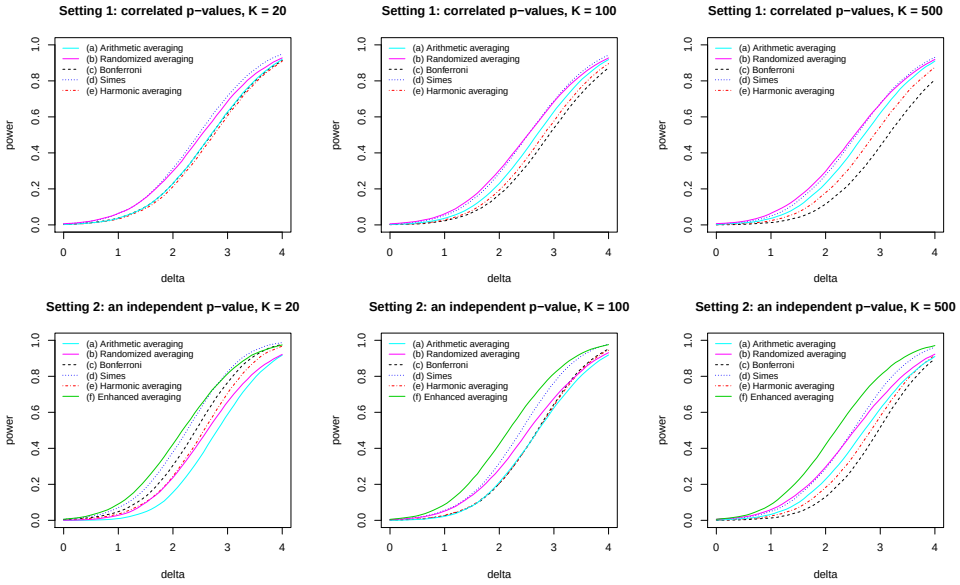


Figure 2: Tests based on combining p-values

stopped test supermartingales (e.g., [15], [14]). A traditional e-test is

$$\text{rejecting the null hypothesis} \iff E \geq \frac{1}{\alpha}. \tag{17}$$

Using the fact that $(2E)^{-1}$ is a p*-variable in Theorem 6.2, we can design the randomized test

$$\text{rejecting the null hypothesis} \iff 2E \geq \frac{1}{V}, \tag{18}$$

where $V \sim \mathrm{U}[0, 2\alpha]$ is independent of E. Just like the comparison between (14) and (15), the test (18) has $3/4$ chance of being more power than the traditional choice of testing E against $1/\alpha$ in (17).

Next, suppose that one has two independent e-variables E_1 and E_2 for a null hypothesis. As shown by [21], it is optimal in a weak sense to use the combined e-variable E_1E_2 for testing the null. Assume further that one of E_1 and E_2 satisfies condition (D').

Using (18) and Proposition 4.5, we get $\mathbb{P}((2E_1)^{-1} \leq \alpha E_2) \leq \alpha$ (note that the positions of E_1 and E_2 are symmetric here). Hence, the test

$$\text{rejecting the null hypothesis} \iff 2E_1E_2 \geq \frac{1}{\alpha}, \tag{19}$$

has size at most α. The threshold of the test (19) is half the one obtained by directly applying (17) to the e-variable E_1E_2. Thus, the test statistic is boosted

by a factor of 2 via condition (D') on either E_1 or E_2. No assumption is needed for the other e-variable. In particular, by setting $E_2 = 1$, we get a p-variable $(2E_1)^{-1}$ if E_1 satisfies (D'), as we see in Proposition 6.4.

E-values calibrated from p-values are useful in the context of testing randomness online (see [18]) and designing test martingales (see [4]). More specifically, for a possibly infinite sequence of independent p-variables $(P_t)_{t\in\mathbb{N}}$ and a sequence of p-to-e calibrators $(f_t)_{t\in\mathbb{N}}$, the stochastic process

$$X_t = \prod_{k=1}^{t} f_k(P_k) \quad t = 0, 1, \dots$$

is a supermartingale (with respect to the filtration of $(P_t)_{t\in\mathbb{N}}$) with initial value $X_0 = 0$ (it is a martingale if P_t, $t \in \mathbb{N}$ are standard uniform and f_t, $t \in \mathbb{N}$ are admissible). As a supermartingale, $(X_t)_{t\in\mathbb{N}}$ satisfies the anytime validity, i.e., X_τ is an e-variable for any stopping time τ; moreover, Ville's inequality gives

$$\mathbb{P}\left(\sup_{t\in\mathbb{N}} X_t \geq \frac{1}{\alpha}\right) \leq \alpha \text{ for any } \alpha > 0. \tag{20}$$

Anytime validity is crucial in the design of online testing where evidence arrives sequentially in time, and scientific discovery is reported at a stopping time considered with sufficient evidence.

Notably, the most popular choice of p-to-e calibrators are those in (12) and (13) (see e.g., [18]), which are convex. Theorem 6.1 implies that if the inputs are not p-values but p*-values, we can still obtain test martingales using convex calibrators such as (12) and (13), without calibrating these p*-values to p-values. This observation becomes useful when each observed P_t is only a p*-variable (e.g., an average of several p-values from parallel experiments using shared data).

Moreover, for a fixed $t \in \mathbb{N}$, if there is a convex f_s for some $s \in \{1, \dots, t\}$ with $f_s(1) = 0$, and P_s is a p-variable (the others can be p*-variables with any p*-to-e calibrators), then (D') is satisfied by $f_s(P_s)$, and we have $\mathbb{P}(X_t \geq 1/\alpha) \leq \alpha/2$ by using the test (19); see our numerical experiments below.

Simulation experiments

In the simulation results below, we generate test martingales following [21]. Similarly to Section 7, the null hypothesis H_0 is $\mathrm{N}(0, 1)$ and the alternative is $\mathrm{N}(\delta, 1)$ for some $\delta > 0$. We generate iid $X_1, \dots, X_n$ from $\mathrm{N}(\delta, 1)$. Define the e-variables from the likelihood ratios of the alternative to the null density,

$$E_t := \frac{\exp(-(X_t - \delta)^2/2)}{\exp(-X_t^2/2)} = \exp(\delta X_t - \delta^2/2), \quad t = 1, \dots, n. \tag{21}$$

The test martingale $S = (S_t)_{t=1,\dots,n}$ is defined as $S_t = \prod_{s=1}^{t} E_s$. Such a martingale S is *growth optimal* in the sense of [14], as it maximizes the expected log growth among all test martingales built on the data $(X_1, \dots, X_n)$; indeed, S is Kelly's strategy under the betting interpretation. Here, we constructed

the martingale S assuming that we know δ; otherwise we can use universal test martingales (e.g., [2]) by taking a mixture of S over δ under some probability measure.

Note that each E_t is log-normally distributed and it does not satisfy (D'). Hence, (19) cannot be applied to S_n. Nevertheless, we can replace E_1 by another e-variable E_1' which satisfies (D'). We choose E_1' by applying the p-to-e calibrator (13) to the p-variable $P_1 = 1 - \Phi(X_1)$, namely, $E_1' = (P_1)^{-1/2} - 1$.

Replacing E_1 by E_1', we obtain the new test martingale $S' = (S_t')_{t=1,\dots,n}$ by $S_t' = E_1' \prod_{s=2}^{t} E_s$. The test martingale S' is not growth optimal, but as E_1' satisifies (D'), we can test via the rejection condition $2S_n' \geq 1/\alpha$, thus boosting the terminal value by a factor of 2. Let $V \sim \mathrm{U}[0, 2\alpha]$ be independent of the test statistics. We compare five different tests, all with size at most α:

(a) applying (17) to S_n: reject H_0 if $S_n \geq 1/\alpha$ (benchmark case);

(b) applying (18) to S_n: reject H_0 if $2S_n \geq 1/V$;

(c) applying (19) to S_n': reject H_0 if $2S_n' \geq 1/\alpha$;

(d) applying a combination of (18) and (19) to S_n': reject H_0 if $2S_n' \geq 1/V$;

(e) applying (20) to the maximum of S: reject H_0 if $\max_{1 \leq t \leq n} S_t \geq 1/V$.

Since test (a) is strictly dominated by test (e), we do not need to use (a) in practice; nevertheless we treat it as a benchmark for comparison on tests based on e-values as it is built on the fundamental connection between e-values and p-values: the e-to-p calibrator $e \mapsto e^{-1} \wedge 1$.

The significance level α is set to be 0.01. The power of the five tests is computed from the average of 10,000 replications for varying signal strength δ and for $n \in \{2, 10, 100\}$. Results are reported in Figure 3. For most values of δ and n, either the deterministic test (c) for S' or the maximum test (e) has the best performance. The deterministic test (c) performs very well in the cases $n = 2$ and $n = 10$, especially for weak signals; this may be explained by the factor of 2 being substantial when the signal is weak. If n is large and the signal is not too weak, the effect of using the maximum of S in (e) is dominating; this is not surprising. Although the randomized test (b) usually improves the performance from the benchmark case (a), the advantages seem to be quite limited, especially in view of the extra randomization, often undesirable.

9 Merging p*-values

Merging p-values and e-values is extensively studied in the literature; see [9], [20] and [21] and the references therein. Using a p-merging or e-merging function gives rise to generalized Bonferroni-Holm procedures (see [20] for p-values and [21] for e-values), or general false discovery control procedures similar to those in [5] and [6]. We discuss merging p*-variables, which is surprisingly nice, specially noting that p-variables and p*-variables are very similar as seen from Theorem

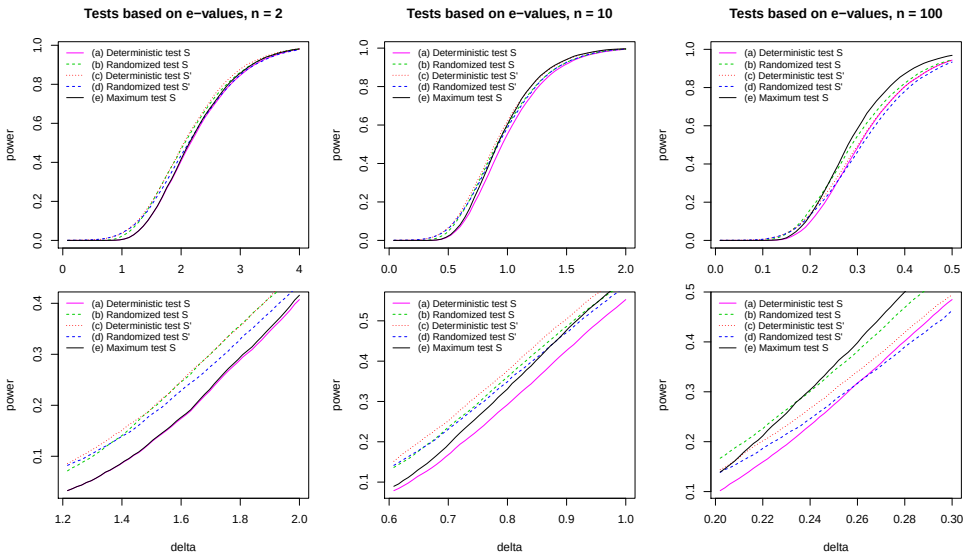


Figure 3: Tests based on e-values; the second row is zoomed in from the first row

5.1 (i), but merging p-values is generally very complicated as illustrated by [20, 19].

A *p*-merging function* in dimension K is an increasing Borel function M on $[0,\infty]^K$ such that $M(P_1,\dots,P_K)$ is a p*-variable for all p*-variables $P_1,\dots,P_K$; p-merging and e-merging functions are defined analogously; see also [21]. A p*-merging function M is *admissible* if it is not strictly dominated by another p*-merging function.

Proposition 9.1. *The arithmetic average M_K is an admissible p*-merging function in any dimension K.*

Proof. The validity of M_K as a p*-merging function is implied by Theorem 3.2. To show its admissibility, suppose that there exists a p*-merging function M that strictly dominates M_K. Let $P_1,\dots,P_K$ be iid uniform random variables on $[0,1]$. The strict domination implies $M \le M_K$ and $\mathbb{P}(M(P_1,\dots,P_K) < M_K(P_1,\dots,P_K)) > 0$. We have

$$\mathbb{E}[M(P_1,\dots,P_K)] < \mathbb{E}[M_K(P_1,\dots,P_K)] = \frac{1}{2}.$$

This means that $M(P_1,\dots,P_K)$ is not a p*-variable, a contradiction. □

Proposition 9.1 illustrates that p*-values are very easy to combine using an arithmetic average; recall that M_K is not a valid p-merging function since the average of p-values is not necessarily a p-value. On the other hand, M_K is an

admissible e-merging function which essentially dominates all other symmetric admissible e-merging functions ([20, Proposition 3.1]).

Another benchmark merging function is the Bonferroni merging function

$$M_B : (p_1, \dots, p_K) \mapsto \left(K \bigwedge_{k=1}^{K} p_k \right) \wedge 1.$$

The next result shows that M_B is an admissible p*-merging function. The Bonferroni merging function M_B is known to be an admissible p-merging function ([21, Proposition 6.1]), whereas its transformed form (via $e = 1/p$) is an e-merging function but not admissible; see [21, Section 6] for these claims.

Proposition 9.2. *The Bonferroni merging function M_B is an admissible p*-merging function in any dimension K.*

Proof. Let $P_1, \dots, P_K$ be p*-variables, and P be a random variable such that the distribution of P is the equally weighted mixture of those of $P_1, \dots, P_K$. Note that P a p*-variable by Proposition 3.4. Let $P_{(1)} = \bigwedge_{k=1}^{K} P_k$. Using the Bonferroni inequality, we have, for any $\epsilon \in (0, 1)$,

$$\mathbb{P}(P_{(1)} \le \epsilon) \le \sum_{k=1}^{K} \mathbb{P}(P_k \le \epsilon) = K\mathbb{P}(P \le \epsilon). \tag{22}$$

Let G_1 be the left-quantile function of $P_{(1)}$ and G_2 be that of P. By (22), we have $G_1(Kt) \ge G_2(t)$ for all $t \in (0, 1/K)$. Hence, for each $y \in (0, 1/K)$, using the equivalent condition (1), we have

$$\int_0^y KG_1(t)\,\mathrm{d}t \ge \int_0^y KG_2(t/K)\,\mathrm{d}t = K^2 \int_0^{y/K} G_2(t)\,\mathrm{d}t \ge K^2 \frac{y^2}{2K^2} = \frac{y^2}{2}.$$

This implies, via the equivalent condition (1) again, that $KP_{(1)}$ is a p*-variable.

Next we show the admissibility of M_B for $K \ge 2$, since the case $K = 1$ is trivial. Suppose that there is a p*-merging function M which strictly dominates M_B. Since M is increasing, there exists $p \in (0, 1/K]$ such that $q := M(p, \dots, p) < M_B(p, \dots, p) = Kp$. First, assume $2Kp \le 1$. Define identically distributed random variables $P_1, \dots, P_K$ by

$$P_k = p\mathbb{1}_{A_k} + \mathbb{1}_{A_k^c}, \quad k = 1, \dots, K,$$

where $A_1, \dots, A_K$ are disjoint events with $\mathbb{P}(A_k) = 2p$ for each k. It is easy to check that $P_1, \dots, P_K$ are p*-variables, and

$$\mathbb{P}(M(P_1, \dots, P_K) = q) = \mathbb{P}\left(\bigcup_{k=1}^{K} A_k \right) = \sum_{k=1}^{K} \mathbb{P}(A_k) = 2Kp.$$

Thus, $M(P_1, \dots, P_K)$ takes the value $q < Kp$ with probability $2Kp$, and it takes the value 1 otherwise. Let G be the left-quantile function of $M(P_1, \dots, P_K)$. The above calculation leads to

$$\int_0^{2Kp} G(t)\,\mathrm{d}t = 2qKp < \frac{(2Kp)^2}{2},$$

showing that $M(P_1, \dots, P_K)$ is not a p*-variable by (1), a contradiction.

Next, assume $2Kp > 1$. In this case, let $r = p - 1/(2K)$, and define identically distributed random variables $P_1, \dots, P_K$ by

$$P_k = r\mathbb{1}_{B_k} + p\mathbb{1}_{A_k} + \mathbb{1}_{(A_k \cup B_k)^c}, \quad k = 1, \dots, K,$$

where $A_1, \dots, A_K, B_1, \dots, B_K$ are disjoint events with $\mathbb{P}(A_k) = 1/K - 2r$ and $\mathbb{P}(B_k) = 2r$, $k = 1, \dots, K$. Note that the union of $A_1, \dots, A_K, B_1, \dots, B_K$ has probability 1. It is easy to verify that $P_1, \dots, P_K$ are p*-variables. Moreover, we have $q' := M(r, \dots, r) \le Kr$ since M dominates M_B. Hence, $M(P_1, \dots, P_K)$ takes the value $q' \le q$ with probability $2Kr$, and it takes the value q otherwise. Let G be the left-quantile function of $M(P_1, \dots, P_K)$. Using $q' \le Kr$ and $q < Kp = Kr + 1/2$, we obtain

$$\begin{aligned}\int_0^1 G(t)\,\mathrm{d}t &= \int_0^{2Kr} q'\,\mathrm{d}t + \int_{2Kr}^1 q\,\mathrm{d}t \\ &\le 2(Kr)^2 + q(1 - 2Kr) \\ &< 2(Kr)^2 + \left(Kr + \frac{1}{2}\right)(1 - 2Kr) = \frac{1}{2},\end{aligned}$$

showing that $M(P_1, \dots, P_K)$ is not a p*-variable by (1), a contradiction. As M cannot strictly dominate M_B, we know that M_B is admissible. □

10 Conclusion

In this paper we introduced p*-values (and p*-variables) as an abstract measure-theoretic object. The notion of p*-values is a manifold generalization of p-values, and it enjoys many attractive theoretical properties in contrast to p-values. Calibration between p*-values and e-values is particularly useful, as p*-values serve as an intermediate step in both the standard e-to-p and p-to-e calibrations. We discuss testing with p*-values and introduced the randomized p*-test. Although we may not test directly with p*-values, we illustrate that they are useful in the design of tests with averages of p-values and with e-values. As a consequence of results in this paper, the concept of p*-values serves as least as a useful technical tool which enhances the extensive and growing applications of p-values and e-values.

Acknowledgements

The author thanks Ilmun Kim, Aaditya Ramdas and Vladimir Vovk for helpful comments on an earlier version of the paper, and acknowledges financial support from and the Natural Sciences and Engineering Research Council of Canada (RGPIN-2018-03823, RGPAS-2018-522590).

References

[1] 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.

[2] Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2020). Time-uniform, nonparametric, nonasymptotic confidence sequences. *Annals of Statistics*, forthcoming.

[3] Döhler, S., Durand, G. and Roquain, E. (2018). New FDR bounds for discrete and heterogeneous tests. *Electronic Journal of Statistics*, **12**(1), 1867–1900.

[4] Duan, B., Ramdas, A., Balakrishnan, S. and Wasserman, L. (2019). Interactive martingale tests for the global null. *arXiv*: 1909.07339.

[5] Genovese, R. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Annals of Statistics*, **32**, 1035–1061.

[6] Goeman, J. J. and Solari, A. (2011). Multiple testing for exploratory research. *Statistical Science*, **26**(4), 584–597.

[7] Grünwald, P., de Heide, R. and Koolen, W. M. (2020). Safe testing. *arXiv*: 1906.07801v2.

[8] Huber, M. (2019). Halving the bounds for the Markov, Chebyshev, and Chernoff Inequalities using smoothing. *The American Mathematical Monthly*, **126**(10), 915–927.

[9] Liu, Y. and Xie, J. (2020), Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, **115**(529), 393–402.

[10] Mao, T., Wang, B. and Wang, R. (2019). Sums of uniform random variables. *Journal of Applied Probability*, **56**(3), 918–936.

[11] Meng, X. L. (1994). Posterior predictive p-values. *Annals of Statistics*, **22**(3), 1142–1160.

[12] Müller, A. and Stoyan, D. (2002). *Comparison Methods for Stochastic Models and Risks*. Wiley, England.

[13] Sarkar, S. K. (1998). Some probability inequalities for ordered MTP2 random variables: a proof of the Simes conjecture. *Annals of Statistics*, **26**(2), 494–504.

[14] Shafer, G. (2020). The language of betting as a strategy for statistical and scientific communication. *Journal of the Royal Statistical Society, Series A*, forthcoming.

[15] Shafer, G., Shen, A., Vereshchagin, N. and Vovk, V. (2011). Test martingales, Bayes factors, and p-values. *Statistical Science*, **26**, 84–101.

[16] Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic Orders*. Springer Series in Statistics.

[17] Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.

[18] Vovk, V. (2020). Testing randomness online. *Statistical Science*, forthcoming.

[19] Vovk, V., Wang, B. and Wang, R. (2020). Admissible ways of merging p-values under arbitrary dependence. *arXiv*: 2007.14208

[20] Vovk, V. and Wang, R. (2020). Combining p-values via averaging. *Biometrika*, forthcoming.

[21] Vovk, V. and Wang, R. (2020). E-values: Calibration, combination, and applications. *Annals of Statistics*, forthcoming.

[22] Wang, B. and Wang, R. (2015). Extreme negative dependence and risk aggregation. *Journal of Multivariate Analysis*. **136**, 12–25.

[23] Wasserman, L., Ramdas, A. and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, **117**(29), 16880–16890.