Submitted to Statistical Science

Confidence and discoveries with *e*-values

Vladimir Vovk and Ruodu Wang

Abstract. We discuss systematically two versions of confidence regions: those based on *p*-values and those based on *e*-values, a recent alternative to *p*-values. Both versions can be applied to multiple hypothesis testing, and in this paper we are interested in procedures that control the number of false discoveries under arbitrary dependence between the base *p*- or *e*-values. We introduce a procedure that is based on *e*-values and show that it is efficient both computationally and statistically using simulated and real-world datasets. Comparison with the corresponding standard procedure based on *p*-values is not straightforward, but there are indications that the new one performs significantly better in some situations.

MSC2020 subject classifications: Primary 62G10; secondary 62C15.
Key words and phrases: Hypothesis testing, multiple hypothesis testing, Bayes factor, discovery vector, discovery matrix.

1. INTRODUCTION

Starting from the introduction of confidence regions in the work of Jerzy Neyman (1937), confidence estimation and hypothesis testing have been regarded as dual tasks. We start our discussion from hypothesis testing and then extend it to confidence estimation.

The usual approaches to hypothesis testing and confidence estimation are based on *p*-values, but our emphasis will be on alternative approaches based on *e*-values, as discussed in, e.g., Shafer (2021) (who uses "betting score" for our "*e*-value"), Shafer and Vovk (2019, Section 11.5) (who use "Skeptic's capital"), Grünwald, de Heide and Koolen (2020), and Vovk and Wang (2021) (who proposed the term "*e*-values").

E-values can be defined as values taken by *e*-variables, and an *e-variable* is a random variable taking values in $[0,\infty]$ whose expectation is at most 1 under the null hypothesis. In many areas of statistics *e*-variables appear naturally as likelihood ratios: if Q is a simple null hypothesis and Q' is an alternative probability measure, the Radon–Nikodym derivative $\mathrm{d}Q'/\mathrm{d}Q$ is an *e*-variable. In Bayesian statistics, Q or Q' or both may be defined as marginal probability measures for Bayesian models, in which case likelihood ratios are known as Bayes factors. The fundamental monograph treating Bayes factors is Jeffreys's (1961); see, e.g., Ly, Verhagen and Wagenmakers (2016) for a recent appreciation. The notions of *e*-values and Bayes factors coincide for simple null hypotheses but diverge for composite ones (for *e*-variables, the expectation should be at most 1 under any probability measure in the null hypothesis).

The existing statistical methods are often divided into Bayesian and classical (we will say more about the latter in Section 3). While *p*-values are the standard classical tool of hypothesis testing, Bayes factors are the standard Bayesian tool (Benjamini et al., 2021). One way of looking at *e*-values is as a way of modelling Bayes factors inside classical statistics inasmuch as they do not require prior distributions for their definition. This hints at the difficulty of comparisons between results based on *p*-values and those based on *e*-values; it is a manifestation of the oft-acknowledged chasm between classical and Bayesian statistics.

Roughly, the Bayesian interpretation of an *e*-variable $\mathrm{d}Q'/\mathrm{d}Q$ is that, when deciding between Q' and Q as possible explanations for the data and observing a very large *e*-value, the optimal decision is to reject Q unless the prior probability of Q is high or a mistaken rejection of Q is much more costly than a mistaken rejection of Q' (see Bernardo and Smith (2000, Proposition 6.1) for a precise decision-theoretic statement).

Another important and popular source of *e*-values, especially in the context of sequential observations, is *e-processes*, which are stochastic processes $(E_t)_{t\ge0}$ such that E_τ is an *e*-variable for any stopping time τ (with respect to a pre-specified filtration); see, e.g., Shafer and Vovk (2019), Grünwald, de Heide and Koolen (2020), Vovk and Wang (2021), and Wang and Ramdas (2022). The use of *e*-processes ensures validity under optional

Department of Computer Science, Royal Holloway, University of London, Egham, Surrey, UK, (e-mail: v.vovk@rhul.ac.uk); Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada, (e-mail: wang@uwaterloo.ca)

stopping and allows sequential update of statistical evidence. These advantages are discussed extensively in the existing literature and are not the focus of this paper.

If E is an *e*-variable and $\alpha > 1$, Markov's inequality implies that, under the null hypothesis, $E \geq \alpha$ with probability at most $1/\alpha$. Therefore, observing a large value of E provides evidence against the null hypothesis in classical statistics as well. In typical uses of *e*-values, however, we are not given a threshold α in advance (or ever), and simply regard an *e*-value as the strength of evidence against the null.

The defining property of a *p*-value is that it is α or less with a probability of at most α. This definition involves a quantifier over thresholds α and sometimes is considered misleading in situations where no threshold α is fixed in advance. There have been proposals to turn ("calibrate") *p*-values into Bayes factors (Sellke, Bayarri and Berger, 2001) to help intuition, and Jeffreys (1961, Appendix B) proposes an informal correspondence between *p*-values and Bayes factors; both can be used for establishing connections between *p*-values and *e*-values. Ways of turning *p*-values into *e*-values and vice versa have been systematically discussed in Vovk and Wang (2021, Section 2) and are the topic of Section 3. They provide ways of comparing results based on *e*-values and *p*-values, albeit crude ones.

An area of statistics where we can see both *e*-values and *p*-values in action is controlling the number of false discoveries in multiple hypothesis testing. A known procedure of controlling the number of false discoveries (Genovese and Wasserman, 2004; Goeman and Solari, 2011a; Goeman et al., 2019), which we call the *GWGS procedure*, uses *p*-values, but it can be easily adapted to *e*-values. In this paper we demonstrate the performance of both versions.

Both versions of the GWGS procedure control the number of false discoveries in a stronger sense than the well-known procedure of Benjamini and Hochberg (1995) controlling the false discovery rate (FDR). Whereas FDR is the expected value of the false discovery proportion, the GWGS procedure provides upper confidence bounds on the number of false discoveries. Procedures that control FDR using *e*-values are studied by Wang and Ramdas (2022).

The GWGS procedure involves, at least implicitly, merging several *p*-values into a single *p*-value. Merging *p*-values is difficult: see, e.g., Vovk and Wang (2020a); Vovk, Wang and Wang (2022). The situation with *e*-values is radically different: arithmetic averaging is essentially the only symmetric method of merging (Vovk and Wang, 2021, Proposition 3.1). This contrast shows in the observation that the *e*-version of the GWGS procedure produces seemingly better results than the *p*-version; we cannot be more categorical since comparison between *p*-values and *e*-values is not straightforward.

We start the main part of the paper by discussing testing in Sections 2 and 3, defining confidence regions in Section 4, and repackaging them as necessity measures in Section 5. In Section 6 we introduce the *e*-version of the GWGS procedure, postponing the *p*-version to an appendix. A special case that is easy to visualize is introduced under the name of discovery *e*-matrices. In Sections 7 and 8 we demonstrate the advantages of the *e*-version in simulation and empirical studies, respectively. In Section 9 we give its computationally efficient implementation. Section 10 concludes.

The main content of the paper is complemented by five appendixes, A–E. Appendix A contains some further information on a toy example in the main paper. Appendix B explores other procedures of controlling false discoveries with *e*-values, including the one based on a Bonferroni-type procedure of merging *e*-values. If the goal is family-wise validity, such procedures (including the one in Holm (1979)) usually work very well, but if the goal is to control the number of false discoveries, they work much worse than arithmetic averaging. In this appendix we also discuss a Simes-type procedure based on *e*-values. Appendix C makes connections with Goeman and Solari's (2011a) work explicit. As Hemerik, Solari and Goeman (2019, Supplementary material) explain, the method of Goeman and Solari (2011a) is equivalent to a method in Genovese and Wasserman (2004). Appendix D points out the importance of generalized Bayes factors. Finally, Appendix E summarizes results of further biomedical studies related to the dataset that we use in Section 8.

2. THREE APPROACHES TO HYPOTHESIS TESTING

The basic principle of hypothesis testing is sometimes referred to as Cournot's principle (Shafer, 2007). Augustin Cournot's bridge between probability theory and the world is that if a given event has a small probability, we do not expect it to happen. It is shown at the top of Figure 1 and has entered (without its name) countless statistics textbooks: the simplest approach to hypothesis testing consists in selecting *a priori* a critical region A of a small probability under the null hypothesis and rejecting the null hypothesis when A happens. Cournot's principle is the basis of the classical approach to statistics; it was known to and used by James Bernoulli (1713), and Cournot's (1843) contribution was to say that this is the *only* bridge.

We are mostly interested in two generalizations of Cournot's principle. To give formal definitions, we fix a measurable space $(\Omega, \mathcal{A})$. This is our *sample space*; to complete it to a probability space we need a probability measure $Q \in \mathfrak{P}(\Omega)$, where $\mathfrak{P}(\Omega)$ is the set of all probability measures on $(\Omega, \mathcal{A})$.

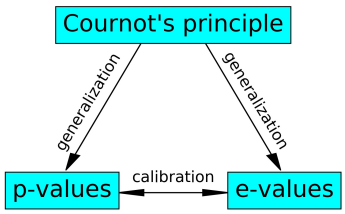

FIG 1. *Cournot's principle and its two generalizations*

A *statistical model* is a family $(Q_\theta \mid \theta \in \Theta)$ of probability measures on $(\Omega, \mathcal{A})$. We do not require measurability, in any sense, of Q_θ in θ; in particular, the *parameter space* Θ is just a set (not a measurable space). We are mostly interested in the case where $\Theta = \mathfrak{P}(\Omega)$ and $Q_\theta = \theta$ for all $\theta \in \Theta$, but in the first few sections our exposition will be general, which may make our definitions more familiar to some of the readers. For a given parameter $\theta \in \Theta$, we have the notion of expectation $\mathbb{E}^\theta(E) := \int E \, \mathrm{d}Q_\theta \in [0, \infty]$ for each extended random variable E taking nonnegative values (we call it "extended" since it may take value ∞) and the notion of probability $\mathbb{P}^\theta(A) := \mathbb{E}^\theta(1_A) = Q_\theta(A)$ for each event $A \in \mathcal{A}$.

A *simple statistical hypothesis* is an element θ of Θ. A *statistical hypothesis* (or *composite statistical hypothesis*, or simply *hypothesis*) is a set $H \subseteq \Theta$ of parameters. We embed the simple statistical hypotheses into the composite statistical hypotheses by identifying $\theta \in \Theta$ with the corresponding singleton $\{\theta\} \subseteq \Theta$. We will say "null hypothesis" to emphasize that we are interested in whether the hypothesis should be rejected in view of the data $\omega \in \Omega$.

We already mentioned that the most basic way of testing a simple hypothesis $\theta \in \Theta$ is to choose a critical region $A \in \mathcal{A}$ with probability $\mathbb{P}^\theta(A) \le \alpha$, α (the *size*) being a small positive number, and to reject the hypothesis θ at level α after observing an outcome $\omega \in A$. A disadvantage of this way of testing is that it is binary; either we completely reject the null hypothesis or we find no evidence whatsoever against it. We will discuss two ways to graduate the notion of a critical region: the classical one using *p*-values and a more recent one using *e*-values.

A *p-variable* for testing a simple hypothesis θ is a nonnegative random variable P such that, for any $\alpha \in (0,1)$, $\mathbb{P}^\theta(P \le \alpha) \le \alpha$. For each threshold α we have a critical region $\{P \le \alpha\}$, and a *p*-variable provides a nested family of critical regions. An *e-variable* for testing a simple hypothesis $\theta \in \Theta$ is a nonnegative extended random variable E such that $\mathbb{E}^\theta(E) \le 1$.

Suppose we are testing a simple null hypothesis θ (it might correspond to a default parameter value). In *p*-testing, we choose a *p*-variable P in advance and reject the null hypothesis θ when the observed value $P(\omega)$ of P (the *p-value*) is small, and in *e*-testing, we choose an *e*-variable E in advance and reject the null hypothesis θ when the observed value $E(\omega)$ of E (the *e-value*) is large. In both cases, we get a measure of the amount of evidence found against the null hypothesis.

We can embed basic testing into both *p*-testing and *e*-testing: namely, to each critical region A corresponds the *p*-variable

$$P(\omega) := \begin{cases} \alpha & \text{if } \omega \in A \\ 1 & \text{if not} \end{cases} \tag{1}$$

and *e*-variable

$$E(\omega) := \begin{cases} 1/\alpha & \text{if } \omega \in A \\ 0 & \text{if not,} \end{cases} \tag{2}$$

where α is the size of the critical region A. These two random variables carry the same information as A. This justifies the two arrows marked "generalization" in Figure 1.

The special case of basic testing corresponds to concentrating on only one threshold, denoted α in the case of *p*-values, (1), and $1/\alpha$ in the case of *e*-values, (2). It is instructive to see how we could extend the basic *p*-variable (1) and the basic *e*-variable (2). It is easy to extend (1); e.g., we can take another critical region $A' \supset A$ of size $\alpha' > \alpha$ and define a p-variable,

$$P(\omega) := \begin{cases} \alpha & \text{if } \omega \in A \\ \alpha' & \text{if } \omega \in A' \setminus A \\ 1 & \text{if } \omega \in \Omega \setminus A', \end{cases}$$

that strongly dominates (1). As it were, for each α we have a separate budget of α that can be spent on a critical region. On the other hand, there is no way to improve the *e*-variable (2) in a non-trivial way (make it larger on a set of positive probability). Now we have a single budget of 1, which has been fully spent in (2).

The definitions of critical regions, *p*-variables, and *e*-variables extend to the case of composite hypotheses as follows. A *critical region* of size α for a composite hypothesis H is an event $A \in \mathcal{A}$ satisfying $\mathbb{P}^\theta(A) \le \alpha$ for all $\theta \in H$. A *p-variable* for testing a composite hypothesis H is a nonnegative random variable P such that, for any $\alpha \in (0,1)$, $\mathbb{P}^\theta(P \le \alpha) \le \alpha$ for all $\theta \in H$. And an *e-variable* for testing a composite hypothesis H is an extended nonnegative E satisfying $\mathbb{E}^\theta(E) \le 1$ for all $\theta \in H$.

While *p*-variables (referred to as valid *p*-values in Casella and Berger (2002, Definition 8.3.26)) are standard, *e*-variables (Shafer, 2021; Vovk and Wang, 2021) have not been used widely.

Observing a small *p*-value or a large *e*-value provide evidence against H. It is convenient to have conventional thresholds for *p*-values and *e*-values. For *p*-values, the

standard thresholds are 1% and 5%, and they go back to Fisher. If $p \le 0.05$, we say that the evidence against the null hypothesis is significant, and if $p \le 0.01$, we say that the evidence is highly significant. For *e*-values, we will use Jeffreys's (1961, Appendix B) rule of thumb:

- If the *e*-value is below 1, the null hypothesis is supported. In our plots (such as in Figure 3 below) in the experimental sections, 7 and 8, such *e*-values will be shown in dark green.
- If the *e*-value is in the interval $(1, \sqrt{10}) \approx (1, 3.16)$, the evidence against the null hypothesis is not worth more than a bare mention. Such *e*-values will be shown in green.
- If the *e*-value is in $(\sqrt{10}, 10) \approx (3.16, 10)$, the evidence against the null hypothesis is substantial. Shown in yellow.
- If it is in $(10, 10^{3/2}) \approx (10, 31.6)$, the evidence against the null hypothesis is strong. Shown in red.
- If it is in $(10^{3/2}, 100) \approx (31.6, 100)$, the evidence against the null hypothesis is very strong. Shown in dark red.
- If the *e*-value exceeds 100, the evidence is decisive. Shown in black.

Fisher's and Neyman–Pearson's views of testing

A common view is that, in our terminology, Fisher preferred *p*-testing, whereas Neyman preferred basic testing (with the null hypothesis complemented by an alternative hypothesis). The full story is, however, more complex: see, e.g., Lehmann (2011, Section 4.4).

Fisher's interpretation of hypothesis testing was in terms of a disjunction (Fisher, 1973, Section III.1). If A is a critical region of a small size α and we observe an outcome in A, then *either the null hypothesis is wrong or "a rare chance has occurred"*. To avoid any frequentist connotations, we may express it in the equivalent form *the null hypothesis is wrong unless the outcome is strange* ("unless" being one of the ways to express the idea of disjunction (Kleene, 1967, Section 14)). A similar interpretation is applicable to *p*-values and *e*-values: e.g., if we observe a large *e*-value, then the null hypothesis is wrong unless the outcome is strange.

In Neyman and Pearson's approach to hypothesis testing, a big role is played by alternative hypotheses. In *e*-testing, the notion of an alternative hypothesis plays a less independent role: choosing an *e*-variable E can often be interpreted as choosing an alternative hypothesis in such a way that E is the likelihood ratio of the alternative hypothesis to the null (Shafer, 2021, 2.2).

A toy example

Let us see how these definitions work in a simple example. We would like to test the null hypothesis $x \sim N(0,1)$ given an observation $x \in \Omega := \mathbb{R}$. Suppose we believe that $|x|$ reflects the amount of evidence against the null hypothesis. Therefore, we will be interested in *p*-variables $P(x)$ and *e*-variables $E(x)$ that depend on x only via $|x|$ and are monotonic functions (increasing for E and decreasing for P) of $|x|$. There is a unique *p*-variable P (uniformly distributed on $[0,1]$) satisfying this property, namely $P(x) := 2\Phi(-|x|)$, where Φ is the standard Gaussian distribution function. On the other hand, there is a huge variety of *e*-variables satisfying this property. A natural class of such *e*-variables is

$$E(x) := \frac{|x|^d}{\pi^{-1/2} 2^{d/2} \Gamma\left(\frac{d+1}{2}\right)}, \quad d > 0, \tag{3}$$

where the denominator is just the normalizing constant ensuring $\int E \, \mathrm{d}N(0,1) = 1$ (the dth absolute moment of the standard Gaussian distribution, which is well known and easily found by direct integration).

Figure 2 gives the *p*-values as the black solid line in both panels and gives the *e*-values for $d \in \{2, 10, 50\}$ in the left panel. On Fisher's scale, *p*-values are significant when their decimal logarithms drop below $\log_{10} 0.05$ and highly significant when they drop below -2; these levels are shown as thin black lines. Jeffreys's levels 0.5, 1, 1.5, and 2 for *e*-values on the $\log_{10}$ scale are shown as thin orange lines. The *e*-variables are not comparable, in the sense that none of them dominates any other everywhere.

Each of the *e*-variables in the left panel of Figure 2 defines an alternative to the null hypothesis, as discussed above, so that the *e*-variable becomes the likelihood ratio of the alternative to the null hypothesis. The alternative hypothesis corresponding to (3) has the density proportional to $|x|^d \exp(-x^2/2)$. Since x ranges over $\mathbb{R}$, it is not exactly the χ density with $d+1$ degrees of freedom, which we will denote χ_{d+1}, but it is a slight variation: after generating x from χ_{d+1}, we change its sign (i.e., multiply it by -1) with probability $1/2$. We will call this alternative distribution the *signed* χ_{d+1} *distribution*; for $d = 2$ this is the *signed Maxwell–Boltzmann distribution*, which is abbreviated to "M.-B." in the legend in the left panel of Figure 2.

The signed χ alternatives, corresponding to $E(x) \propto |x|^d$, appear to be the simplest unconstrained choice, but a more standard approach is to look for alternatives inside a parametric family of distributions. Let us embed $N(0,1)$ into the statistical model $N(\mu, 1)$, $\mu \in \mathbb{R}$ (there are other natural embeddings, and in Appendix A we will also discuss the embedding into the statistical model $N(0, \sigma^2)$, $\sigma > 0$). The right panel of Figure 2 shows three more *e*-variables, which are based on the likelihood ratios

$$E^{(\delta)}(x) := \frac{\mathrm{d}N(\delta,1)}{\mathrm{d}N(0,1)}(x) = \frac{\exp(-(x-\delta)^2/2)}{\exp(-x^2/2)} = \exp(\delta x - \delta^2/2). \tag{4}$$

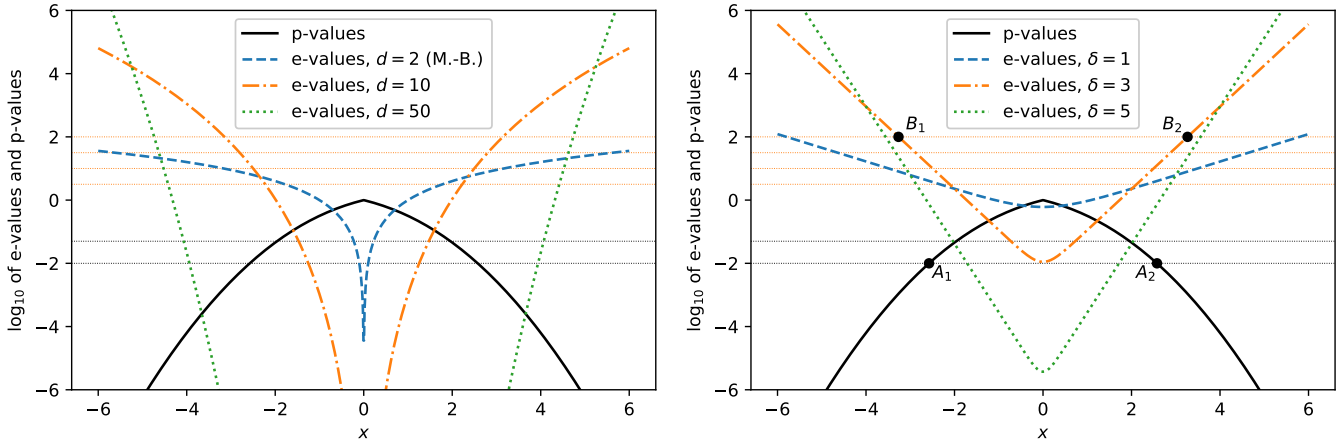

FIG 2. *The p-values (black solid lines) and e-values on the decimal log scale for testing the null hypothesis* $N(0,1)$. *Left panel: for the signed* χ *alternatives. Right panel: for the alternatives in the family* $N(\mu,1)$.

To obtain an *e*-variable that depends on x only via $|x|$, let us average $E^{(\delta)}$ and $E^{(-\delta)}$:

$$\bar{E}^{(\delta)}(x) := \frac{E^{(\delta)}(x) + E^{(-\delta)}(x)}{2}. \tag{5}$$

The right panel of Figure 2 shows $\bar{E}^{(1)}$, $\bar{E}^{(3)}$, and $\bar{E}^{(5)}$.

Four points are indicated in the right panel to illustrate the interpretation of our plots. The points A_1 and A_2 are at the intersection of the graph of the *p*-variable with the horizontal line at level -2, and the points B_1 and B_2 are at the intersection of the graph of the *e*-variable with $\delta := 3$ with the horizontal line at level 2. The x-coordinates of the points A_1 and A_2 are approximately ± 2.58, and the x-coordinates of the points B_1 and B_2 are approximately ± 3.27. The observations x with $|x|$ exceeding (approximately) 3.27 provide decisive evidence against the null hypothesis $N(0,1)$, and the other observations do not, according to Jeffreys's scale. Similarly, the observations x with $|x|$ exceeding (approximately) 2.58 provide highly significant evidence against the null hypothesis $N(0,1)$, while the other observations do not, according to Fisher's scale. Using similar interpretation for the left panel, we can see, e.g., that the *e*-variable for $d = 2$ makes a wider range of observations x provide substantial evidence against the null hypothesis than the *e*-variable for $d = 10$ does, whereas for strong evidence we have the opposite situation.

Later in the paper we will use methods related to both panels of Figure 2. In our simulation studies in Section 7, we will generate observations from $N(\mu,1)$ and use again the likelihood ratios (4) that we used in the right panel. In our empirical studies in Section 8, where we have no idea of the true distribution of the data, we use the *e*-variables proportional to $|x|^d$, as in the left panel.

3. IS IT POSSIBLE TO COMPARE *E*-VALUES AND *P*-VALUES?

Starting from the next section we will describe various methods based on *p*-values and *e*-values. In Section 1 we already alluded to difficulties of comparing such results. There is no overarching testing framework (at least at this time) containing both *p*-testing and *e*-testing that could be used for comparing such results. The best we can do rigorously is to convert, albeit imperfectly, *p*-values to *e*-values and vice versa.

Sometimes the user of statistical procedures has a clear preference for *p*-values or *e*-values. These are some possible categories of users (this is not an exhaustive list, of course):

1. Some users will find the frequentist interpretation of *p*-variables P appealing: for any threshold α, the long-run frequency of observing $P \le \alpha$ in a sequence of independent identical trials is at most α. This is typical of the frequentist school of classical statistics.
2. Other users will prefer a direct application of Cournot's principle: for a small α and pre-specified P, we do not expect to observe $P \le \alpha$ under the null hypothesis. This school is referred to as Bernoullian statistics by Glenn Shafer (Shafer, 2022), following Francis Edgeworth, Richard von Mises, Arthur Dempster, and Ian Hacking.
3. Another category, representing Bayesian statistics, will like the Bayesian interpretation of *e*-variables referred to in Section 1.
4. Our final category will accept the betting interpretation of an *e*-variable E (see, e.g., Shafer (2021)): the *e*-value $E(\omega)$ is the pay-off of a lottery that is fair under the null hypothesis, and a large pay-off casts doubt on the null hypothesis. The idea of

betting is often regarded as an important ingredient of Bayesian statistics (see, e.g, de Finetti (2017, 3.3.5)), but it is used there in a very different way, in the form of no Dutch book requirement.

Communication may be easier between the users in the first two categories (classical statistics), or between the users in the last two categories. But otherwise, we need methods of conversion between p- and e-values. Therefore, as a first step we discuss rigorous ways of turning p-values into e-values (known as calibrating p-values) and vice versa. For further details, see Vovk and Wang (2021, Section 2).

A decreasing function $f : [0,1] \to [0,\infty]$ is a *calibrator* if, for any p-variable P, $f(P)$ is an e-variable. In other words, a calibrator transforms p-values to e-values. A very natural family of calibrators is

$$f_\kappa(p) := \kappa p^{\kappa-1}, \tag{6}$$

where $\kappa \in (0,1]$. The maximum possible e-value

$$\mathrm{VS}(p) := \max_{\kappa\in(0,1]} f_\kappa(p) = \tag{7}$$

$$\begin{cases} -\exp(-1)/(p\ln p) & \text{if } p \le \exp(-1) \\ 1 & \text{otherwise} \end{cases}, \quad p \in (0,1],$$

attainable by this family will be referred to as the *VS bound* (abbreviating "Vovk–Sellke bound" (Sellke, Bayarri and Berger, 2001; Shafer and Vovk, 2019, Section 11.5)), but due to the maximum operation, $\mathrm{VS}(P)$ need not be an e-variable even if P is a p-variable.

In the opposite direction, a decreasing function $f : [0,\infty] \to [0,1]$ is an *e-to-p calibrator* if, for any e-variable E, $f(E)$ is a p-variable. It is a function transforming e-values to p-values. As explained and formalized in Vovk and Wang (2021, Proposition 2.2),

$$t \in [0,\infty] \mapsto \min(1, 1/t) \tag{8}$$

is the only reasonable e-to-p calibrator.

In general, calibrating p-values and e-values are crude processes. A strong e-value of 20 barely attains statistical significance when transformed into a p-value (namely, 5%) using (8). The VS bound for the borderline significant p-value of 5% is approximately 2.456, and so "is not worth more than a bare mention", according to Jeffreys. The low "round-trip efficiency" in the domain of p-values can be illustrated by

$$1/\mathrm{VS}(0.005) \approx 0.072. \tag{9}$$

The round trip turns the highly significant p-value of 0.5% into the non-significant p-value of 7.2%. And this is despite the VS bound being achievable as e-value only in hindsight.

In view of the low round-trip efficiency, it is natural to expect that users of statistical procedures who insist on using p-values will be best served by methods producing directly p-values. A method producing e-values will have to be vastly superior to result in better, or even equally good, p-values after conversion. Interestingly, we will see such an example in Section 7 (see the discussion of Figure 4). Symmetrically, a method producing p-values will have to be vastly superior to a method producing directly e-values in order to result in better or equally good e-values after conversion. The caveat here is that the result of comparison still depends on using the bound (7) (using the e-to-p calibrator (8) is uncontroversial).

The user who is uncertain whether to use p-values or e-values usually needs a more accurate comparison than that provided by the crude procedures of calibration and e-to-p calibration. We do not have objective ways of doing that. One subjective way to compare results using e-values to those using p-values is to appeal to Jeffreys's (1961, Appendix B) authority: "Users of these tests speak of the 5 per cent. point in much the same way as I should speak of the $K = 10^{-1/2}$ point, and of the 1 per cent. point as I should speak of the $K = 10^{-1}$ point." In our terminology, people doing p-testing speak of a p-value of 5% (resp. 1%) in much the same way as Jeffreys should speak of an e-value of $10^{1/2}$ (resp. 10). The approximate equivalences are

$$p\text{-value of } 5\% \sim e\text{-value of } 10^{1/2} \approx 3.16 \tag{10}$$

$$p\text{-value of } 1\% \sim e\text{-value of } 10. \tag{11}$$

Another subjective way is to use Good's (Good, 1958, Appendix IV) rule of thumb. According to Good, the e-value corresponding to a p-value of p should lie in the range

$$\left(\frac{1}{30p}, \frac{3}{10p}\right) \tag{12}$$

when $0.001 < p < 0.2$ (which Good felt were the values of p that are usually of most practical interest). In Good's picture the p-value of p is obtained using the standard recipe from the Bayes factor as test statistic (this condition is always satisfied in this paper). If we take the geometric mean $1/(10p)$ of the end-points of the interval (12), we will obtain

$$p\text{-value of } 5\% \sim e\text{-value of } 2 \tag{13}$$

$$p\text{-value of } 1\% \sim e\text{-value of } 10 \tag{14}$$

in place of (10)–(11). While (10)–(11) and (13)–(14) are close, Good acknowledges the significant uncertainty surrounding the correspondence.

A slightly more objective way of comparing methods based on p-values and e-values is to consider their mathematical simplicity. A great advantage of e-values is that they are very easy to combine; as we mentioned in Section 1, arithmetic averaging is essentially the only symmetric method of combination (Vovk and Wang, 2021,

Proposition 3.1). This leads to simple and intuitive algorithms (and in Vovk, Wang and Wang (2022) merging *e*-values is even used as a technical tool for designing admissible ways of merging *p*-values).

4. CONFIDENCE REGIONS

The notion of a confidence region was introduced by Neyman (1934; 1937) only in its basic version. (See Lehmann (2011, Section 6.4) for Neyman's predecessors; the word "confidence" is a translation of the Polish "ufność" (Neyman, 1941), and Neyman's adjectival use of it was at first made fun of by his English listeners (Neyman, 1934, comments by Bowley and Fisher).) The *p*-version is usually implicit, and the *e*-version has not been used in mainstream statistics. However, the *e*-version has been used for a long time, in some form, in the algorithmic theory of randomness (Levin, 1976; Gács, 2005; Vovk and V'yugin, 1993), and in this paper we will use the terminology close to that of the algorithmic theory of randomness.

Let us fix a statistical model $(Q_\theta \mid \theta \in \Theta)$. A *basic test* of size α is a family of critical regions $(A_\theta \mid \theta \in \Theta)$ of size α. Therefore, for each simple statistical hypothesis θ, a basic test fixes a critical region A_θ for testing θ: $\mathbb{P}^\theta(A_\theta) \le \alpha$.

The interpretation of a basic test that is symmetric between the parameter space Θ and sample space Ω is that $\omega \in A_\theta$ means poor agreement between θ and ω. This binary relation of poor agreement and its complementary relation of good agreement have two sides:

- on the testing side, we start from θ and divide the ωs into those that conform to θ ($\omega \notin A_\theta$) and those that do not ($\omega \in A_\theta$);
- on the estimation side, we start from ω and divide the θs into those that agree with ω ($\omega \notin A_\theta$) and those that do not ($\omega \in A_\theta$).

In particular, on the estimation side we have the notion of a confidence estimator as introduced by Neyman: the confidence estimator corresponding to a basic test $(A_\theta \mid \theta \in \Theta)$ is

$$\Gamma(\omega) := \{\theta \in \Theta \mid \omega \notin A_\theta\}. \tag{15}$$

In the context of a basic test $(A_\theta \mid \theta \in \Theta)$ of a small size α we may say that an outcome $\omega \in \Omega$ is *strange* for a parameter value $\theta \in \Theta$ if $\omega \in A_\theta$. According to Cournot's principle, we do not expect the outcome ω to be strange for the true θ. Our interpretation of the confidence region (15) is that $\Gamma(\omega)$ covers the true θ unless ω is strange.

Graduated notions of a confidence estimator are discussed surprisingly rarely in statistics textbooks, especially in full generality (e.g., the popular textbook Cox and Hinkley (1974) is one of the few places where they are discussed, but only in the context of a linearly ordered parameter space Θ). A *p-test* is a family of *p*-variables $(P_\theta \mid \theta \in \Theta)$, and the corresponding *p-confidence regions* are defined as

$$\Gamma_{p,\alpha}(\omega) := \{\theta \in \Theta \mid P_\theta(\omega) > \alpha\}, \quad \alpha \in (0,1). \tag{16}$$

We regard $P_\theta(\omega)$ as a measure of agreement between θ and ω, with small values indicating poor agreement, and define $\Gamma_{p,\alpha}(\omega)$ to be the set of θ that agree with ω at level α. The definition of a p-confidence estimator is only a slight variation on the definition of a basic estimator: namely, (16) can be obtained from (15) by setting $A_\theta := \{\omega \mid P_\theta(\omega) \le \alpha\}$ for each α. Notice that the *p*-confidence regions $\Gamma_{p,\alpha}(\omega)$ are *nested*: $\alpha_1 < \alpha_2$ implies $\Gamma_{p,\alpha_2}(\omega) \subseteq \Gamma_{p,\alpha_1}(\omega)$; this property is sometimes discussed or at least mentioned in statistics textbooks (e.g., in Cox and Hinkley (1974, Section 7.2), Casella and Berger (2002, 9.5.2), and Stuart, Ord and Arnold (1999, Section 19.15)).

Similarly, an *e-test* is a family of *e*-variables $(E_\theta \mid \theta \in \Theta)$. We also regard $E_\theta(\omega)$ as a measure of agreement between θ and ω, but now large values indicate poor agreement. Analogously to (16), we define the *e-confidence regions* as

$$\Gamma_{e,\alpha}(\omega) := \{\theta \in \Theta \mid E_\theta(\omega) < \alpha\}, \quad \alpha \in (0,\infty). \tag{17}$$

The definitions (16) and (17) of *p*-confidence regions and *e*-confidence regions generalize the basic definition (15), which corresponds to using the *p*-test and the *e*-test defined by (1) and (2), respectively, with added subscripts θ.

The notions of *p*-test and *e*-test provide graduated notions of strangeness. Let $\alpha > 0$; we will sometimes refer to it as the *significance level* (the interesting values are $\alpha < 1$ for *p*-testing and $\alpha > 1$ for *e*-testing). In the context of a *p*-test $(P_\theta \mid \theta \in \Theta)$, we say that $\omega \in \Omega$ is *α-strange* for $\theta \in \Theta$ if $P_\theta(\omega) \le \alpha$ (i.e., if we reject θ at level α after observing ω). And in the context of an *e*-test $(E_\theta \mid \theta \in \Theta)$, we say that $\omega \in \Omega$ is *α-strange* for $\theta \in \Theta$ if $E_\theta(\omega) \ge \alpha$. If there is any risk of confusion, we will use the fuller expressions "(p,α)-strange" and "(e,α)-strange".

The interpretation of the confidence region (16) in terms of a Fisher-type disjunction is that $\Gamma_{p,\alpha}(\omega)$ covers the true θ unless ω is (p,α)-strange. Similarly, we interpret (17) by saying that $\Gamma_{e,\alpha}(\omega)$ covers the true θ unless ω is (e,α)-strange.

Starting from Section 7, we will visualize *e*-confidence regions for a range of thresholds, including 10. Inspired by the terminology of Jeffreys (1961, Appendix B), already discussed in Section 2, we will refer to an *e*-confidence region at level 1 as *weak*, at level $10^{1/2}$ as *substantial*, at level 10 as *strong*, at level $10^{3/2}$ as *very strong*, and at level 100 as *extremely strong*.

Simultaneous confidence regions

Sometimes we are interested not in θ but in some derivative parameter (as in Schervish's textbook (1995, 5.2.1)). For example, if $\Omega = \mathbb{R}$ and $\Theta = \mathfrak{P}(\Omega)$, we might be interested in the median of $\theta \in \Theta$. Let $g : \Theta \to \Theta_g$ be the function mapping the original parameter θ to a new parameter, $g(\theta)$.

The confidence regions for the derived parameter $g(\theta)$ become:

$$\Gamma^g(\omega) := \{g(\theta) \mid \theta \in \Theta \ \& \ \omega \notin A_\theta\} \tag{18}$$

in place of (15),

$$\Gamma^g_{p,\alpha}(\omega) := \{g(\theta) \mid \theta \in \Theta \ \& \ P_\theta(\omega) > \alpha\}, \quad \alpha \in (0,1),$$

in place of (16), and

$$\Gamma^g_{e,\alpha}(\omega) := \{g(\theta) \mid \theta \in \Theta \ \& \ E_\theta(\omega) < \alpha\}, \quad \alpha \in (0,\infty), \tag{19}$$

in place of (17).

It is important that we can have a family of functions g, and the confidence estimator (18) will be valid simultaneously for all of them, provided the same basic test $(A_\theta \mid \theta \in \Theta)$ is used for all g. The same is true for *p*-confidence estimators and *e*-confidence estimators; what is important is that the notion of strangeness should not depend on g. For example, for any family of functions $g : \Theta \to \Theta_g$, the confidence region $\Gamma^g_{e,\alpha}(\omega)$ in (19) contains $g(\theta)$ for all g simultaneously unless the outcome is α-strange for the true parameter θ.

Confidence regions in the toy example

Here we continue our discussion of the toy example started in the previous section. Now our statistical model $(Q_\theta \mid \theta \in \Theta)$ is $\Theta := \mathbb{R}$ and $Q_\theta := N(\theta, 1)$ for all θ. For a fixed δ, such as $\delta := 3$, let us generalize (5) to

$$\bar{E}^{(\delta)}_\theta(x) := \frac{E^{(\delta)}_\theta(x) + E^{(-\delta)}_\theta(x)}{2}, \tag{20}$$

where, generalizing (4),

$$E^{(\delta)}_\theta(x) := \frac{\mathrm{d}N(\theta+\delta,1)}{\mathrm{d}N(\theta,1)}(x) \tag{21}$$

$$= \frac{\exp(-(x-\theta-\delta)^2/2)}{\exp(-(x-\theta)^2/2)} = \exp(\delta(x-\theta) - \delta^2/2).$$

This gives us an *e*-test.

Remember that the x-coordinate of the point B_2 in the right panel of Figure 2 is approximately 3.27, and let us fix $\delta := 3$. Therefore, the extremely strong *e*-confidence regions (*e*-confidence intervals in this case) are

$$\Gamma_{e,100}(x) \approx [x - 3.27, x + 3.27], \quad x \in \mathbb{R}, \tag{22}$$

where "$\approx$" refers to 3.27 being an approximate value. For Jeffreys's other thresholds the *e*-confidence intervals are

$$\Gamma_{e,10^{3/2}}(x) \approx [x - 2.88, x + 2.88], \tag{23}$$

$$\Gamma_{e,10}(x) \approx [x - 2.50, x + 2.50] \tag{24}$$

$$\Gamma_{e,10^{1/2}}(x) \approx [x - 2.11, x + 2.11], \tag{25}$$

and the *p*-confidence intervals are, as usual, $\Gamma_{p,0.01}(x) \approx [x - 2.58, x + 2.58]$.

The *e*-confidence intervals (22)–(25) will change if the alternative hypotheses $\theta \pm 3$ are replaced by other ones, such as $\theta \pm 1$ or $\theta \pm 5$. It can be considered an advantage of *p*-confidence intervals, and *p*-values in general, that for an important (albeit small) set of popular statistical models there is no dependence on the choice of the alternative hypothesis. This is closely related to the existence of uniformly most powerful statistical tests (Lehmann and Romano, 2022, Chapter 3).

5. NECESSITY AND POSSIBILITY MEASURES

The notions of a test discussed in the previous sections allow us to associate measures of confidence with subsets of the parameter space in view of an outcome. These are just a different way to package confidence regions.

For a *p*-test $(P_\theta \mid \theta \in \Theta)$, the *p-necessity measure* of a set $B \subseteq \Theta$ in view of an outcome $\omega \in \Omega$ is defined as

$$\Box_p(B \mid \omega) := \sup_{\theta \notin B} P_\theta(\omega). \tag{26}$$

Now the Fisher-type disjunction for the true θ is: $\theta \in B$ *unless* ω *is* $\Box_p(B \mid \omega)$*-strange for* θ. Therefore, we expect $\theta \in B$ for a small $\Box_p(B \mid \omega)$. Of course, this disjunction remains true if we replace "$\Box_p(B \mid \omega)$-strange" by "c-strange" for any $c \geq \Box_p(B \mid \omega)$, but in statements of this kind we usually choose the c that makes them as strong as possible.

Similarly, for an *e*-test $(E_\theta \mid \theta \in \Theta)$, the *e-necessity measure* of $B \subseteq \Theta$ given $\omega \in \Omega$ is

$$\Box_e(B \mid \omega) := \inf_{\theta \notin B} E_\theta(\omega), \tag{27}$$

with the analogous interpretation: $\theta \in B$ *unless* ω *is* $\Box_e(B \mid \omega)$*-strange for* θ.

If we are interested in a derivative parameter $g(\theta)$, where $g : \Theta \to \Theta_g$, the *p-necessity measure* and *e-necessity measure* of $B \subseteq \Theta_g$ in view of $\omega \in \Omega$ are now defined as

$$\Box^g_p(B \mid \omega) := \sup_{\theta \in \Theta : g(\theta) \notin B} P_\theta(\omega) = \Box_p(g^{-1}(B) \mid \omega), \tag{28}$$

$$\Box^g_e(B \mid \omega) := \inf_{\theta \in \Theta : g(\theta) \notin B} E_\theta(\omega) = \Box_e(g^{-1}(B) \mid \omega), \tag{29}$$

respectively, with the same interpretations as before.

Analogously to (26)–(29) we can define the *p-possibility measure* and *e-possibility measure* by

$$\Diamond_p(B \mid \omega) := \sup_{\theta \in B} P_\theta(\omega) = \Box_p(B^c \mid \omega),$$

$$\Diamond_e(B \mid \omega) := \inf_{\theta \in B} E_\theta(\omega) = \Box_e(B^c \mid \omega),$$

$$\Diamond_p^g(B \mid \omega) := \Diamond_p(g^{-1}(B) \mid \omega) = \Box_p^g(B^c \mid \omega),$$

$$\Diamond_e^g(B \mid \omega) := \Diamond_e(g^{-1}(B) \mid \omega) = \Box_e^g(B^c \mid \omega),$$

where $B^c := \Theta \setminus B$ is the complement of B. For example, a large value of $\Diamond_e(B \mid \omega)$ means that $\theta \in B$ is hardly possible for the true θ in view of the outcome ω.

REMARK 5.1. Our notation is borrowed from modal logic, which has two basic modalities, $\Box$ (necessity) and $\Diamond$ (possibility), analogous to the quantifiers $\forall$ and $\exists$, respectively. The notions of necessity and possibility measures discussed in this section are closely related to the necessity and possibility measures of possibility theory (Dubois and Prade, 1988) (which they include in a wider class of what they call confidence measures), and also somewhat related to the belief and plausibility functions of the Dempster–Shafer theory (Shafer, 1976), and to confidence and credibility in conformal prediction (Vovk, Gammerman and Shafer, 2005, (3.66)). However, unlike their counterparts in those theories, our notions just re-express the idea of confidence regions without adding new information.

Necessity measures in the toy example

In the toy example considered at the end of the previous section (with the same *p*-test and *e*-tests), we can write the *p*-necessity measure of a set $B \subseteq \mathbb{R}$ of parameter values in view of an observation $x \in \mathbb{R}$ as

$$\Box_p(B \mid x) = 2\Phi\left(-\inf_{\theta \notin B} |x - \theta|\right).$$

According to the definition, $\Box_p(B \mid x)$ is determined by the parameter value outside B (assuming the inf is attained) that makes the observed x least strange.

The main application of necessity measures in this paper (described in the following section) will be "one-sided", in that the corresponding confidence regions will provide only a lower bound (on the quantity called the number of true discoveries; equivalently, they provide an upper bound on the number of false discoveries). If instead of (20) we use the *e*-test (21) with $\delta > 0$, we will have prediction regions in the form of rays pointing left, and the necessity measure will be

$$\Box_e(B \mid x) = \exp\left(\delta\left(x - \inf B^c\right) - \delta^2/2\right).$$

6. CONTROLLING THE NUMBER OF FALSE DISCOVERIES

Starting from this section we specialize our setting. Our sample space $(\Omega, \mathcal{A})$ is still arbitrary, but now we take $\Theta := \mathfrak{P}(\Omega)$ as our parameter space and $Q_\theta := \theta$ for all $\theta \in \Theta$ as our statistical model; remember that $\mathfrak{P}(\Omega)$ is the set of all probability measures on $(\Omega, \mathcal{A})$. Since our statistical model contains all probability measures on Ω, there is no real loss of generality.

Suppose that we are given K *e*-variables $E_1, \dots, E_K$ for testing hypotheses $H_1, \dots, H_K$, which are our *base hypotheses*; we would like to reject some of them (in fact, as many of them as possible under a validity constraint). The realized values of $E_1, \dots, E_K$ are denoted by $e_1, \dots, e_K$, so that $e_k := E_k(\omega)$ for the realized outcome ω.

If we do not know anything about the nature of the hypotheses $H_1, \dots, H_K$, it makes sense to reject a number of them with the largest e_k. But in general, we can consider an arbitrary non-empty *rejection set* $R \subseteq \{1, \dots, K\}$; this is the set of base hypotheses, represented by their indices, that the researcher chooses to reject. Goeman and Solari (2011a) argue convincingly that in some practically relevant cases R will not necessarily correspond to the largest e_k; e.g., R may include hypotheses connected by a common theme, such as all relevant genes related to the gastrointestinal tract (Goeman and Solari, 2011a, 4.1).

In this section we will find functions D providing a measure of confidence in the number of true discoveries (to be formally defined momentarily) in the following sense: a rejection set R contains at least j true discoveries unless the outcome ω is $D^R(j)$-strange. This statement is uniform in R and j, in the sense of the strangeness of outcomes being measured by a fixed *e*-test. Therefore, a large $D^R(j)$ means high confidence in the number of true discoveries being at least j.

For each $\theta \in \mathfrak{P}(\Omega)$, we define

$$I_\theta := \{k \in \{1, \dots, K\} \mid \theta \in H_k\}$$

to be the set of indices of hypotheses containing θ. If the researcher rejects H_k, we refer to this decision as a *discovery*. We say that the discovery is *true* if $\theta \notin H_k$, and it is *false* if $\theta \in H_k$, where θ is the true (unknown) probability measure governing the data generation. For a rejection set R, the number of true discoveries is

$$g_R(\theta) := |R \setminus I_\theta| = |\{k \in R \mid \theta \notin H_k\}|, \tag{30}$$

and the number of false discoveries is

$$|R \cap I_\theta| = |\{k \in R \mid \theta \in H_k\}|.$$

The sum of these two numbers is $|R|$, the total number of discoveries, and so controlling the number of false discoveries is the same thing as controlling the number of

true discoveries. Our functions $D^R(j)$ will provide measures of confidence in lower bounds j on the number of true discoveries (equivalently, upper bounds $|R| - j$ on the number of false discoveries). Researchers are sometimes interested in the proportion of true or false discoveries $|R \setminus I_\theta| / |R|$ or $|R \cap I_\theta| / |R|$, respectively. We can also control those with the bounds $j/|R|$ or $(|R| - j)/|R|$ (lower for true and upper for false discoveries), respectively.

REMARK 6.1. The researcher may be interested in parameters $g(\theta)$ that differ from (30) more substantially. For example, $g(\theta)$ may be the weighted number of true discoveries in R (e.g., some genes can be more important than other genes). Or, for a partition of R into groups (one of which can be, e.g., the genes related to the gastrointestinal tract), $g(\theta)$ may depend on the number of groups containing true discoveries. In this paper we restrict ourselves to the simplest case.

For e-confidence bounds, we need an e-test $(E_\theta)_{\theta \in \mathfrak{P}(\Omega)}$. For each $k \in I_\theta$, E_k is an e-variable for testing θ. We will obtain E_θ by merging $(E_k)_{k \in I_\theta}$. This can be achieved by using e-merging functions studied in Vovk and Wang (2021). An *e-merging function* is a Borel function $F : \cup_{n=0}^\infty [0,\infty]^n \to [0,\infty]$ that is increasing in each of its arguments and maps any finite sequence of e-variables to an e-variable: if $E_1, \dots, E_n$ are e-variables, $F(E_1, \dots, E_n)$ is required to be an e-variable as well. We always set $F := 1$ if the input sequence is empty. An example (of paramount importance, as discussed earlier) is the arithmetic mean

$$(e_1, \dots, e_n) \mapsto \frac{1}{n} \sum_{i=1}^n e_i. \tag{31}$$

An e-merging function is *symmetric* if it does not depend on the order of its arguments, like the arithmetic mean.

Let F be a symmetric e-merging function; we define for each $\theta \in \Theta$ the e-variable

$$E_\theta := F(E_k : k \in I_\theta). \tag{32}$$

Our main object of interest is $\square_e^{g_R}(\{j, j+1, \dots\} \mid \omega)$ for this e-test. In agreement with Section 5, $\square_e^{g_R}(\{j, j+1, \dots\} \mid \omega)$ gives a confidence bound on the number of true discoveries: the rejection set R contains at least j true discoveries unless the outcome ω is $\square_e^{g_R}(\{j, j+1, \dots\} \mid \omega)$-strange. Therefore, we can count on there being at least j true discoveries in R for a large observed $\square_e^{g_R}(\{j, j+1, \dots\} \mid \omega)$.

Let us now replace $\square_e^{g_R}(\{j, j+1, \dots\} \mid \omega)$ by a more explicit and easily computable expression. Set, for a rejection set R,

$$\begin{aligned} \square_e^{g_R}(\{j, j+1, \dots\} \mid \omega) &= \min_{\theta \in \mathfrak{P}(\Omega) : g_R(\theta) < j} E_\theta \\ &= \min_{\theta \in \mathfrak{P}(\Omega) : |R \setminus I_\theta| < j} F(E_k : k \in I_\theta) \\ &\geq \min_{I \subseteq \{1,\dots,K\} : |R \setminus I| < j} F(E_k : k \in I) =: D^R_{e,F}(j), \end{aligned} \tag{33}$$

where, as usual, $\min \emptyset := \infty$, and the equality $=:$ in the last line of (33) signifies $D^R_{e,F}(j)$ being defined, with the subscripts (e, F) dropped if clear from the context.

Intuitively, in (33) we go over all I for which there are fewer than j true discoveries and evaluate their strangeness; if all of them are strange, we are entitled to reject there being fewer than j true discoveries. The interpretation in terms of the Fisher-type disjunction of the value $D^R(j)$ defined in (33) is: the rejection set R contains at least j true discoveries unless the outcome ω is $D^R(j)$-strange. The values $D^R(j)$ are informative for $j = 1, \dots, |R|$, and we will sometimes refer to $D^R(j)$ (as function of j) as *discovery e-vector*.

Let us say (following Holm (1979, Section 1)) that $H_1, \dots, H_K$ satisfy the *free combinations condition for* R if the sets I_θ, $\theta \in \Theta$, include all subsets of R:

$$\forall S \subseteq R \ \exists \theta \in \Theta : I_\theta = S. \tag{34}$$

The "$\geq$" in (33) becomes "$=$" under the free combinations condition, but this condition is not required for the validity of our methods. In particular, the Fisher-type disjunction still holds for $D^R(j)$ when the condition is not satisfied.

Technically, the choice of F in (32) may depend on θ, but using the same F leads to convenient properties of monotonicity for $D^R(j)$ described in the next proposition. These properties will help us to visualize D in our plots in the experimental sections.

PROPOSITION 6.2. *For any $R, R' \subseteq \{1, \dots, K\}$, $j, j' \in \{0, 1, \dots\}$, and e-merging function F:*

(1) $D^R_{e,F}(j) \leq D^{R'}_{e,F}(j)$ if $R \subseteq R'$;
(2) $D^R_{e,F}(j) \leq D^R_{e,F}(j')$ if $j' \leq j$;
(3) $D^{R'}_{e,F}(j + |R' \setminus R|) \leq D^R_{e,F}(j)$.

PROOF. For item (1), we can rewrite the inequality $D^R_{e,F}(j) \leq D^{R'}_{e,F}(j)$ as

$$\min_{I : |R \setminus I| < j} F(E_i : i \in I) \leq \min_{I : |R' \setminus I| < j} F(E_i : i \in I),$$

and it suffices to notice that any I satisfying $|R' \setminus I| < j$ satisfies $|R \setminus I| < j$.

Item (2) can be rewritten as

$$\min_{I : |R \setminus I| < j} F(E_i : i \in I) \leq \min_{I : |R \setminus I| < j'} F(E_i : i \in I),$$

which follows from $|R \setminus I| < j'$ implying $|R \setminus I| < j$.

Item (3) can be rewritten as

$$\min_{I : |R' \setminus I| < j + |R' \setminus R|} F(E_i : i \in I) \leq \min_{I : |R \setminus I| < j} F(E_i : i \in I),$$

Algorithm 1 Discovery *e*-vector for a given rejection set

Input: A symmetric *e*-merging function F, the rejected hypotheses $R \subseteq \{1,\dots,K\}$, and an increasing sequence of *e*-values $e_1 \le \dots \le e_K$.

for $j = 1,\dots,|R|$ **do**
  let R_j be R without its $j-1$ largest elements
  $D^R(j) := F_{\mathbf{e}}(R_j)$
  for $i = 1,\dots,K$ **do**
    $e := F_{\mathbf{e}}(R_j \cup \{1,\dots,i\})$
    if $e < D^R(j)$ **then**
      $D^R(j) := e$

Algorithm 2 Discovery *e*-matrix D

Input: A symmetric *e*-merging function F and an increasing sequence of *e*-values $e_1 \le \dots \le e_K$.

for $r = 1,\dots,K$ **do**
  for $j = 1,\dots,r$ **do**
    $S_{r,j} := \{K-r+1,\dots,K-j+1\}$
    $D_{r,j} := F_{\mathbf{e}}(S_{r,j})$
    for $i = 1,\dots,K-r$ **do**
      $e := F_{\mathbf{e}}(S_{r,j} \cup \{1,\dots,i\})$
      if $e < D_{r,j}$ **then**
        $D_{r,j} := e$

which follows from

$$|R \setminus I| < j \Longrightarrow \left|R' \setminus I\right| < j + \left|R' \setminus R\right|,$$

which in turn follows from the obvious

$$\left|R' \setminus I\right| \le |R \setminus I| + \left|R' \setminus R\right|. \qquad \square$$

An algorithm for computing the discovery vector D^R is given as Algorithm 1; it is polynomial-time if the underlying *e*-merging function F, assumed symmetric, is polynomial-time. It uses the notation

$$F_{\mathbf{e}}(I) := F(e_i : i \in I), \quad I \subseteq \{1,\dots,K\}, \ I \ne \emptyset, \tag{35}$$

where $\mathbf{e} := (e_1,\dots,e_K)$. Without loss of generality we assume that the *e*-values are sorted in the ascending order,

$$e_1 \le \dots \le e_K. \tag{36}$$

Our definitions so far are essentially translations of Goeman and Solari's (2011a, Section 2) definitions into the language of *e*-values. We will explain the connection in detail in Appendix C. As the procedure for *p*-values was first proposed in Genovese and Wasserman (2004, Theorem 6.3), we refer to it as the *GWGS procedure*. In Appendix C we will also comment on the recent result by Goeman, Hemerik and Solari (2021) about the GWGS procedure being the only admissible one for controlling true discoveries (under a property of validity based on *p*-values).

A special and important choice of F is the arithmetic average (31). Using this *e*-merging function in (33), the *arithmetic-mean discovery e-vector* is defined as

$$\mathrm{AV}^R(j) := \min_{I \subseteq \{1,\dots,K\}: |R \setminus I| < j} \frac{1}{|I|} \sum_{i \in I} E_i,$$
$$R \subseteq \{1,\dots,K\}, \ \ j \in \{1,\dots,|R|\}.$$

As we said earlier, arithmetic averaging is the only useful symmetric *e*-merging function (Vovk and Wang, 2021, Proposition 3.1). It is computed by Algorithm 1 with

$$F_{\mathbf{e}}(I) := \frac{1}{|I|} \sum_{i \in I} e_i, \quad I \subseteq \{1,\dots,K\}, \ \ I \ne \emptyset.$$

Discovery *e*-matrices

Let us return to the case of a general symmetric *e*-merging function F, although we are mainly interested in the arithmetic mean. Next we will discuss a less flexible method in which we consider a family of rejection sets R that are chosen in an optimal way, in some sense. For each $r \in \{1,\dots,K\}$, the set

$$R_r := \{K-r+1,\dots,K\}$$

is the optimal rejection set of size r (assuming (36)), meaning that $D^{R_r}_{e,F} \ge D^R_{e,F}$ for any other set $R \subseteq \{1,\dots,K\}$ of size r. In the terminology of statistical decision theory (Wald, 1950, Section 1.3), R_r is a complete class of rejection sets.

Let us call $D_{r,j} := D^{R_r}_{e,F}(j)$ the *discovery e-matrix*. In terms of the Fisher-type disjunction, there are at least j true discoveries among the r hypotheses with the largest *e*-values unless the outcome ω is $D_{r,j}$-strange. An algorithm for computing the discovery *e*-matrix D is given as Algorithm 2.

We are particularly interested in the *arithmetic-mean discovery matrix* AM, i.e., the discovery *e*-matrix

$$\mathrm{AM}_{r,j}(e_1,\dots,e_K) := \min_{I: |R_r \setminus I| < j} \frac{1}{|I|} \sum_{i \in I} e_i.$$

In Appendix B we illustrate discovery *e*-matrices with some other choices of the *e*-merging function F, which, according to Vovk and Wang (2021, Proposition 3.1), are essentially dominated by the arithmetic mean.

REMARK 6.3. For simplicity, in this remark we will only discuss the arithmetic-mean discovery matrix AM under the free combinations condition (34) with $R := \{1,\dots,K\}$ (although our observations are applicable more widely). An alternative, often more convenient, parameterization of the arithmetic-mean discovery matrix AM is

$$\mathrm{AM}^-_{r,j}(e_1,\dots,e_K) := \mathrm{AM}_{r,j+1}(e_1,\dots,e_K),$$

where j now ranges starting from 0. We can then write

$$\mathrm{AM}^-_{r,j}(e_1,\dots,e_K) = \Diamond_e^{g_{R_r}}(\{j\} \mid \omega)$$

$$= \min_{\theta\in\mathfrak{P}(\Omega):g_{R_r}(\theta)=j} E_\theta = \min_{I:|R_r\setminus I|=j} \frac{1}{|I|}\sum_{i\in I} e_i.$$

In words, $\mathrm{AM}^-_{r,j}$ is the e-possibility of there being exactly j true discoveries among the top r e-values.

REMARK 6.4. It is true that arithmetic averaging is the only useful symmetric e-merging function when no assumptions are made about the dependence structure of the base e-values. On the other hand, if the base e-values are supposed to be independent, it is clear that the product of e-variables is always an e-variable. Therefore, when defining the e-test (32), we have plenty of alternatives to the arithmetic mean in the role of F; F can be the product, or a combination of the arithmetic mean and the product. The product does not work well for problems of the type considered in this paper, since small base e-values then have disproportionate effect. However, the combination

$$E_\theta := \sum_{\{i,j\}\subseteq I_\theta: i\neq j} E_iE_j \Big/ \binom{|I_\theta|}{2}$$

(with $\{i,j\}$ ranging over the 2-element subsets of I_θ) of the product and arithmetic averaging gives excellent results, much better than what we can get without the assumption of independence. See Vovk and Wang (2020b) for details.

7. SIMULATION STUDIES

In our simulation studies we will visualize the arithmetic-mean discovery matrix in some simple cases and compare Algorithm 2 with a method based on p-values. Our setting will be similar to that of Vovk and Wang (2021, Section 5), where family-wise validity is studied.

The observations are generated from the Gaussian model $N(\mu,1)$. The null hypotheses are $N(0,1)$ and the alternatives are $N(\delta,1)$, where we take $\delta := -3$ throughout the section. We generate $K/2$ observations from $N(\delta,1)$ (the alternative distribution) and then $K/2$ observations from $N(0,1)$ (the null distribution), where K (the overall number of hypotheses) is an even number.

In this paper, we colour-code the entries of discovery e-matrices according to Jeffreys's rule of thumb discussed in Section 2. The full colour map is shown on the right of Figure 3 with the thresholds between different colours given in terms of the decimal logarithm of e-values. The most interesting parts of our plots of discovery e-matrices are those in yellow and red; green and dark green parts carry little or no evidence and so are useless for us, and dark red and black parts carry so much evidence that they are rare in a wide range of practical applications (cf. Section 8). In all our discussions below we will ignore the boundaries between the green and dark green parts.

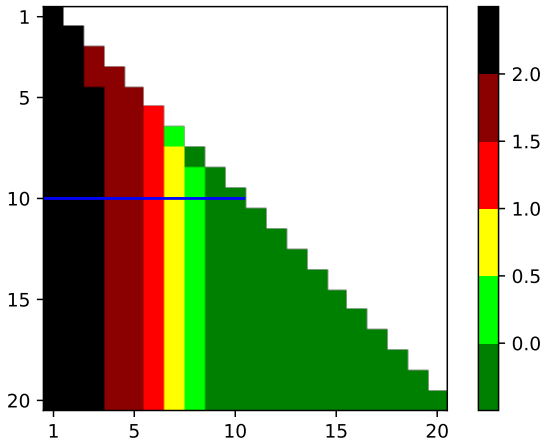

FIG 3. *The arithmetic-mean discovery matrix for 10 false and 10 true null hypotheses, as described in text. The colour map on the right gives Jeffreys's thresholds, the boundaries between different colours in most of our plots, on the decimal log scale. Row 10 is highlighted in blue.*

Figure 3 shows the arithmetic-mean discovery matrix that Algorithm 2 gives for $K = 20$: we generate 10 observations from $N(\delta,1)$ and then 10 from $N(0,1)$. The base e-values are the likelihood ratios

$$E(x) := \frac{\mathrm{d}N(\delta,1)}{\mathrm{d}N(0,1)}(x) = \exp(\delta x - \delta^2/2) \tag{37}$$

(cf. (4)) of the alternative to the null density, where $x \sim N(\mu,1)$ is the corresponding observation. For example, row 10 (highlighted in blue) of the matrix in Figure 3 shows that there is decisive evidence that the number of true discoveries among the 10 hypotheses with the largest e-values is at least 3. Similarly, there is very strong evidence that the number of true discoveries is at least 5, there is strong evidence that the number of true discoveries is at least 6, etc.

REMARK 7.1. When discussing confidence regions, it is more convenient to use the discovery e-matrix AM^- as discussed in Remark 6.3, with the same colour code; it looks as Figure 3 except that 1 is replaced by 0, 5 is replaced by 4, etc. The non-black part of the figure is then formed by the extremely strong confidence regions in each row. Row 10 now shows that the extremely strong confidence region for the number of true discoveries among the 10 hypotheses with the largest e-values is $\{3,\dots,10\}$. The red/yellow/green part (not including dark red) is formed by the very strong confidence regions, so that the very strong confidence region for the number of true discoveries among the 10 hypotheses with the largest e-values is $\{5,\dots,10\}$. Similarly, the yellow/green and green parts are formed by strong and substantial confidence regions, respectively.

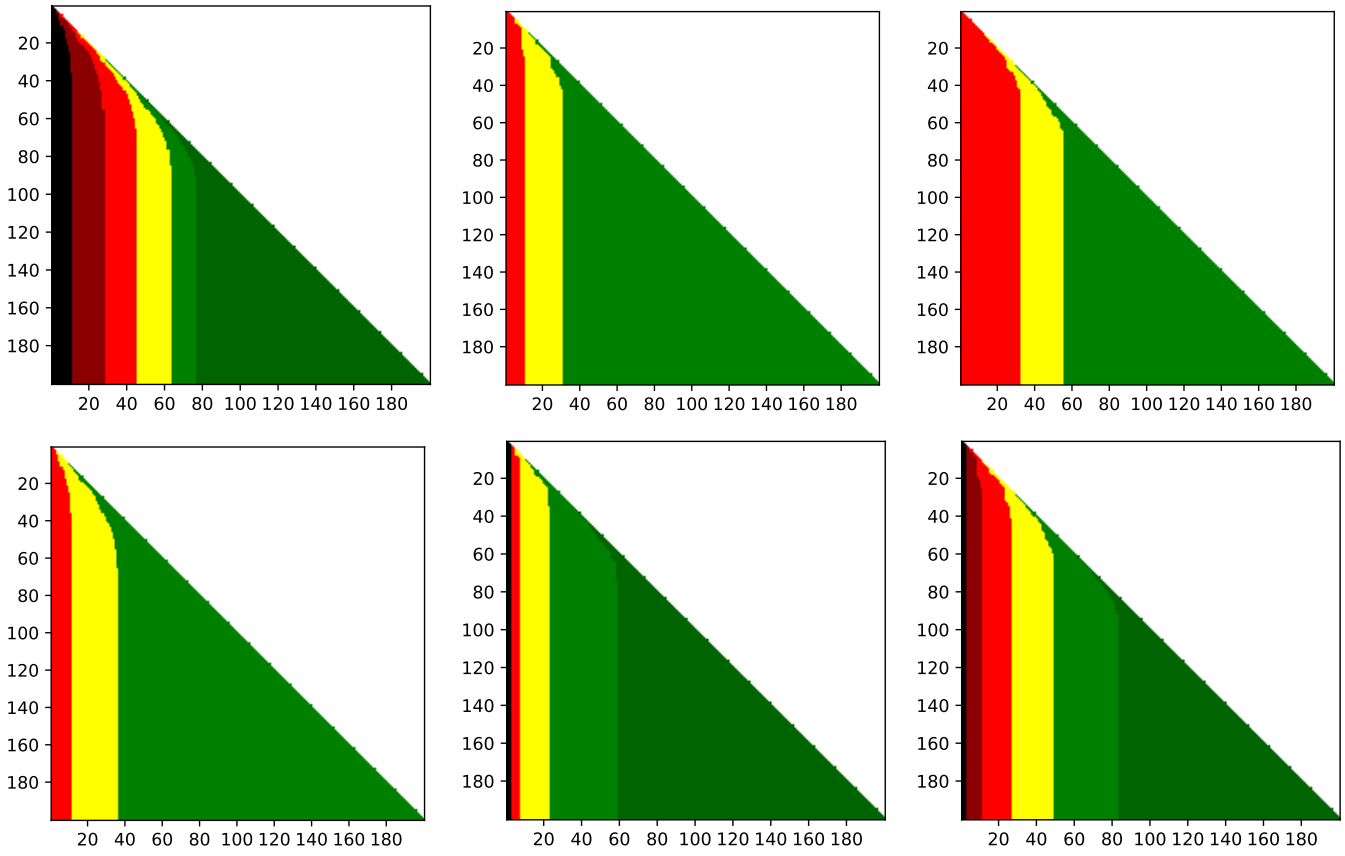

FIG 4. *Upper left panel: the arithmetic-mean discovery matrix for 100 false and 100 true null hypotheses. Upper middle panel: the GWGS discovery p-matrix in the same situation for Fisher's thresholds* 1% *and* 5% *(with values below* 1% *shown in red and between* 1% *and* 5% *in yellow) under arbitrary dependence. Upper right panel: as the upper middle panel but assuming independence.*
Lower left panel: the e-to-p calibrated arithmetic-mean discovery matrix in the upper left panel using Fisher's thresholds. Lower middle panel: the VS-transformed GWGS discovery p-matrix in the upper middle panel of Figure 4 (under arbitrary dependence) using Jeffreys's thresholds. Lower right panel: the VS-transformed GWGS discovery p-matrix in the upper right panel (under independence) using Jeffreys's thresholds.

The upper left panel of Figure 4 is the counterpart of Figure 3 for a larger number of hypotheses, $K = 200$: we generate 100 observations from $N(\delta, 1)$ and then 100 from $N(0, 1)$. We will refer to this set of observations as the *simulation data*.

In practice, a discovery *e*-matrix, such as that shown in the upper left panel of Figure 4, can be used in different ways, for example:

- The researcher may have budget for a limited number of follow-up studies of the hypotheses. For example, if in the situation of that panel her budget is 50 hypotheses, she just concentrates on row 50 (studying the 50 hypotheses H_k with the largest *e*-values). For the first 11 entries in this row the *e*-value exceeds 100, and so she has decisive evidence that there are at least 11 true discoveries among those 50 hypotheses. Similarly,
  - she has very strong evidence that there are at least 27 true discoveries,
  - she has strong evidence that there are at least 40 true discoveries,
  - she has substantial evidence that there are at least 46 true discoveries.

  For the relevant *e*-values, see the bold entries in the row $r = 50$ of Table 1. In terms of confidence regions, we can say, e.g., that our method gives the substantial *e*-confidence region $\{46, 47, \dots\}$, so that 46 may be called the substantial lower *e*-confidence bound on the number of true discoveries among the 50 hypotheses.
- The researcher might have some idea of what proportion of false discoveries she is willing to tolerate (in the spirit of choosing the false discovery rate *a priori* (Benjamini and Hochberg, 1995)). For example, if she is willing to tolerate 10% of false discoveries and willing to use Jeffreys's standard (*e*-value greater than 10) of strong evidence, she should concentrate on row 31 (i.e., study the 31 hypotheses with the largest *e*-values), which is the

TABLE 1
The values $D_{r,j}$ of the discovery matrix shown in Figure 4 for several rows r and columns j.

r \ j	11	12	27	28	29	40	41	42	43	46	47
31	95.6	87.5	14.1	**10.8**	**7.63**						
32	96.5	88.5	16.0	**12.7**	**9.63**						
50	**103**	**96.0**	**33.2**	**30.6**	28.2	**11.2**	**9.94**	8.73	7.52	**4.04**	**3.11**
51	103	96.0	33.4	30.9	28.5	11.6	**10.4**	**9.21**	8.01	4.61	3.69
52	103	96.0	33.6	31.1	28.7	12.0	**10.8**	**9.61**	8.43	5.10	4.19

lowest row with at most 10% of entries below 10. See the bold entries in the rows $r \in \{31, 32\}$ of Table 1 (we have strong evidence that there are at most $3/31 \approx 9.7\%$ of false discoveries in row 31 and at most $4/32 = 12.5\%$ of false discoveries in row 32).

- Alternatively, the researcher might have some idea of how many false discoveries she is willing to tolerate (in the spirit of k-FWER (Romano and Wolf, 2007)). If she is willing to tolerate at most 10 false discoveries and still willing to use Jeffreys's standard of strong evidence, she should concentrate on row 51, which is the lowest row with at most 10 entries (in fact, exactly 10 entries) below 10. See the bold entries in the rows $r \in \{51, 52\}$ of Table 1 (we have strong evidence that there are at most 10 false discoveries in row 51 and at most 11 false discoveries in row 52).

Of course, the researcher may know her hypotheses and relations between them very well, and after looking at the discovery e-matrix she may come up with her own rejection set R, as discussed in Section 6. In this case she should also use Algorithm 1.

In accordance with Proposition 6.2, all discovery e-matrices in our figures satisfy three properties of monotonicity. For example, the entries $\mathrm{AM}_{r,j}$ are decreasing in j and increasing in r. They are also monotonic along the main diagonal and the lines parallel to it: for any $c = 0, 1, \dots$, $\mathrm{AM}_{r,r-c}$ is decreasing in r.

Comparisons

This paper concentrates on multiple hypothesis testing using e-values, but in scientific practice p-values are more popular, despite recent criticism. In this subsection we will report results of our simulation studies in terms of p-values and compare them to our results, as best we can in view of the difficulties discussed in Section 3.

For comparison with methods based on p-values, we use the GWGS procedure applied to standard procedures for combining p-values and to the same nested rejection sets (initial subsets of $\{1, \dots, 200\}$ assuming the p-values are given in ascending order). These procedures admit computationally efficient shortcuts (Goeman et al., 2019, Theorem 1) and are implemented in the R package `hommel` (Goeman, Meijer and Krebs, 2019). As the base p-values we take $P(x) := \Phi(x)$, where Φ is, as before, the standard Gaussian distribution function; these are the p-values found using the most powerful test given by the Neyman–Pearson lemma. The GWGS procedure can be interpreted as producing an analogue of a discovery e-matrix, which we call a *discovery p-matrix*, with e-values replaced by p-values. For details, see Appendix C. In particular, a version of the notion of a discovery p-vector was introduced in Goeman and Solari (2011b, Section 3).

The package `hommel` has an option (`simes`) that controls the choice of the procedure for combining p-values, and the resulting discovery p-matrix is valid either under arbitrary dependence, in which case Hommel's (1986) procedure is used for combining p-values, or under certain assumptions on the dependence structure for the input p-values, in which case Simes's (1986) procedure is used. In particular, Simes's procedure is valid under the assumption of independence; it is also valid under relaxations of independence such as positive dependence (Sarkar, 2011), but not under arbitrary dependence. For brevity we will talk about p-values that are either arbitrarily dependent or independent, but it should be remembered that the assumption of independence may be relaxed. We never make such assumptions about base e-values (but cf. Remark 6.4).

The upper middle panel of Figure 4 shows the discovery p-matrix found using `hommel` applied to the simulation data under arbitrary dependence. The upper right panel of Figure 4 is analogous but assumes independent base p-values. Both panels use Fisher's thresholds 1% and 5%; the values below 1% are shown in red, between 1% and 5% in yellow, and above 5% in green (so that red means "highly significant" and yellow means "significant but not highly significant"). According to Jeffreys as quoted in Section 3 (p. 6), the red and yellow areas are somewhat comparable between p-values and e-values, but we can draw some conclusions even without such cross-comparisons.

Remember that our method does not require any assumptions about the dependence structure of the e-values. It is true that our simulated data are independent, but this information is typically unavailable, and the performance of methods that do not depend on independence or similar

TABLE 2
The Benjamini–Hochberg and Benjamini–Yekutieli procedures applied to the simulation data for FDR (false discovery rate) 5% *and* 1%.

assumption	5%	1%
independence	87	61
arbitrary dependence	55	28

assumptions is still interesting. Comparing the upper left and upper middle panels of Figure 4, we can see that our method produces better confidence bounds if we are willing to use Jeffreys's informal correspondence between *e*-values and *p*-values. The upper left panel is even better, in this sense, than the upper right panel, which makes an assumption on the dependence structure of the base *p*-values.

The three lower panels of Figure 4 are the transformed versions of the corresponding upper panels. In the lower left panel, we transform the arithmetic-mean discovery matrix (upper left panel) by applying the canonical *e*-to-*p* calibrator $e \mapsto 1/e$. In the other two lower panels, we transform the corresponding upper panels by applying the VS transformation (7). Therefore, the lower left panel contains valid *p*-values, whereas the other two lower panels contain upper bounds on *e*-values.

It is interesting that even after the crude step of *e*-to-*p* calibration (remember the woeful round-trip efficiency illustrated by (9)), the lower left panel of Figure 4 still looks slightly better than the upper middle panel. In this comparison there is no uncertainty in the choice of the *e*-to-*p* calibrator, since (8) is the only reasonable one (namely, it dominates any other *e*-to-*p* calibrator). And even the optimistic VS transformation (the lower middle panel) looks much worse than the arithmetic-mean discovery matrix in the upper left panel. We can see, even without using Jeffreys's informal correspondence, that the method based on *e*-values produces better results in this case, despite the crude calibration steps.

Not surprisingly, assuming independence makes direct treatment of *p*-values more efficient: compare the upper right panel and the lower left panel. What is more surprising is that, even assuming independence and using the optimistic VS transformation, the lower right panel still look worse than the upper left panel.

Table 2 gives the numbers of null hypotheses rejected by the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) and its version for arbitrary dependence (Benjamini and Yekutieli, 2001). (The relationship between the Benjamini–Yekutieli procedure and the `hommel` package without the `simes` option is the same as that between the Benjamini–Hochberg procedure and the `hommel` package with the `simes` option; this is made explicit in Goeman et al. (2019, Section 5).) Here the results are more difficult to interpret, since the kind of guarantees provided by those procedures is so different from the guarantees provided by the GWGS methods. Roughly, we get comparable results between the row "arbitrary dependence" in Table 2 and the confidence bounds for the number of true discoveries in, say, row 50 of the discovery matrix in the upper left of Figure 4: as discussed earlier (cf. Table 1), the strong and substantial confidence bounds are 40 and 46, respectively. (Remember that, following Jeffreys, "strong" refers to the threshold of 10 for *e*-values and regarded as roughly corresponding to "highly significant", and "substantial" refers to the threshold of $10^{1/2}$ for *e*-values and regarded as roughly corresponding to "significant".)

8. EMPIRICAL STUDIES

In this section we will demonstrate how the methods of this paper can, in principle, be used in practice. It is important that we will make no assumptions of independence.

We will use the classical dataset first described in Hedenfalk et al. (2001) and then carefully studied in Storey and Tibshirani (2003). Essentially, we will adapt Storey and Tibshirani's analysis to using *e*-values in place of *p*-values (see the end of the section for a discussion of differences). In our experiments we use the version of the dataset made available as part of the R package `qvalue` (Storey et al., 2019).

The main content of the dataset is the expression levels of 3226 genes in 15 samples of tissues. Seven samples are coming from carriers of mutations in the BRCA1 gene, and the remaining eight from carriers of mutations in the BRCA2 gene. We will say that each sample is labelled with its BRCA status: seven are labelled BRCA1, and eight are labelled BRCA2. The core of the dataset is the 3226×15 matrix of gene expressions in the samples; all entries are positive numbers. Following Storey and Tibshirani (2003, Appendix, Remark C), we remove all rows containing at least one entry exceeding 20, which leaves us with a 3170×15 data matrix. Each row of the data matrix corresponds to a gene and each column to a sample, and the entry in row k and column j is the expression level of gene k in sample j. For each gene we are interested in the scientific hypothesis that the gene expression does not depend on the BRCA status of the sample.

Storey and Tibshirani's version of the dataset also contains some further information, such as the *p*-value for each gene. For further information about this dataset, which we will refer to as the *BRCA dataset*, see, e.g., Storey, Dai and Leek (2007, Section 5) and Guindani, Müller and Zhang (2009, 5.2).

For this dataset methods ensuring family-wise validity do not work well. For example, the ten smallest *p*-values

in Storey and Tibshirani's list multiplied by the number of genes 3170 are

$$0,\ 0.040,\ 0.050,\ 0.060,\ 0.060,\ 0.100,\ 0.140,\ 0.160,\ 0.240,\ 0.240 \tag{38}$$

and so the Bonferroni correction leads to only three statistically significant p-values. Moreover, one of the p-values is exactly zero, and so cannot be a valid p-value (for details, see p. 17). Hedenfalk et al. conclude that 9–11 genes are differentially expressed. Storey and Tibshirani's informal analysis suggests that many more, at least 33%, of the examined genes are differentially expressed. However, their informal analysis assumes what they call "weak independence": they rely on the law of large numbers when inspecting histograms of p-values, assuming that the probabilities of the p-values lying in various ranges will manifest themselves as empirical frequencies seen in the histograms. Their formal analysis is asymptotic and also assumes weak independence: see their Appendix, Remark D.

We formalize the scientific theory of interest as the following statistical hypothesis about each gene k: given the multiset of expression levels in row k of the data matrix, each ordering of the row has the same probability. In our current context, a *nonconformity measure* is a measurable function of two multisets; we will use it by applying, for a given gene, to the multiset (of size 7) of the expression levels for the samples labelled BRCA1 and the multiset (of size 8) of the expression levels for the samples labelled BRCA2; the resulting value will be called the *nonconformity score*. For computing base e-values, we use the formula

$$e_k := \frac{T_k}{\frac{1}{B+1}\left(\sum_{b=1}^{B} T_k^{0b} + T_k\right)}, \quad k = 1, \dots, 3170, \tag{39}$$

where T_k is the nonconformity score computed from the kth row of the data matrix with the true labels (BRCA1 or BRCA2) for each sample, T_k^{0b} is the nonconformity score computed from the same row with randomly permuted labels, and B is the number of permutations. In our experiments, the case $0/0$ of the right-hand side of (39) never occurs. We will call (39) the *Monte Carlo e-value*. We are justified in calling it an e-value since, under the null hypothesis, the expected value of the right-hand side of (39) is 1 if we set $0/0 := 1$. (Moreover, the conditional expectation of the right-hand side of (39) is 1 given the multiset of nonconformity scores $\{T_k^{01}, \dots, T_k^{0B}, T_k\}$, the expectation being over all choices of the position of the true nonconformity score in the multiset.)

Our nonconformity measure will be defined in terms of the t-statistic. Let x_{kj} be the base two logarithm of the value in row k and column j of the data matrix (although the base does not matter in our empirical studies). The two-sample t-statistic for the kth gene is

$$t_k := \frac{\bar{x}_{k2} - \bar{x}_{k1}}{\sqrt{s_{k1}^2/n_1 + s_{k2}^2/n_2}}, \tag{40}$$

where $n_1 = 7$ is the number of BRCA1 columns, $n_2 = 8$ is the number of BRCA2 columns, and

$$\bar{x}_{k1} := \frac{1}{n_1} \sum_{j \in \text{BRCA1}} x_{kj},$$

$$s_{k1}^2 := \frac{1}{n_1 - 1} \sum_{j \in \text{BRCA1}} (x_{kj} - \bar{x}_{k1})^2$$

are the sample mean and variance for the BRCA1 entries, with the analogous expressions for BRCA2. The variances of the two groups (BRCA1 and BRCA2) are not assumed to be equal (following Storey and Tibshirani (2003)), but using equal-variance two-sample t-statistics would lead to similar results.

We define the nonconformity score as $T_k := f(t_k)$ for some function f of the t-statistic t_k (see (40)). A natural nonconformity score is $|t_k|$, but we generalize it to $|t_k|^d$ for some $d > 0$. This choice of f is motivated by the Bayesian two-sample t-test widely discussed in recent literature starting from Gönen et al. (2005) and briefly reviewed in Gönen et al. (2019, Section 3). A standard expression for the Bayes factor produced by such a test via the t-statistic t is

$$f(t) := c\left(1 + at^2\right)^{d/2} \tag{41}$$

for positive constants a, c, and d involving the number of degrees of freedom and effective sample size; see, e.g., Wang and Liu (2016, (14)) and Rouder et al. (2009, (1)); the form (41) goes back to Jeffreys (Ly, Verhagen and Wagenmakers, 2016, (12)). However, different constants are used in different papers. We set, without loss of generality, $c := 1$, since c cancels out when using (39). We further simplify (41) by ignoring the "$1+$"; this makes a and any constant factors in the definition of the t-statistic t (there is a non-trivial factor under the assumption of equal variances for the two groups) irrelevant, as they also cancel out when applying (39). Of course, this step does not affect the validity of our methods.

The left panel of Figure 5 gives a key part of the arithmetic-mean discovery matrix for the BRCA dataset with $f(t) := |t|^d$ for $d := 10$, with $B := 10000$, and with base Monte Carlo e-values (39). We can see that there is strong evidence that the number of differentially expressed genes is at least as large as Hedenfalk at al.'s number. If we settle for substantial evidence, the number is much larger. Arguably, it is not as large as in Storey and Tibshirani's study, but we are not using any exchangeability or independence assumptions.

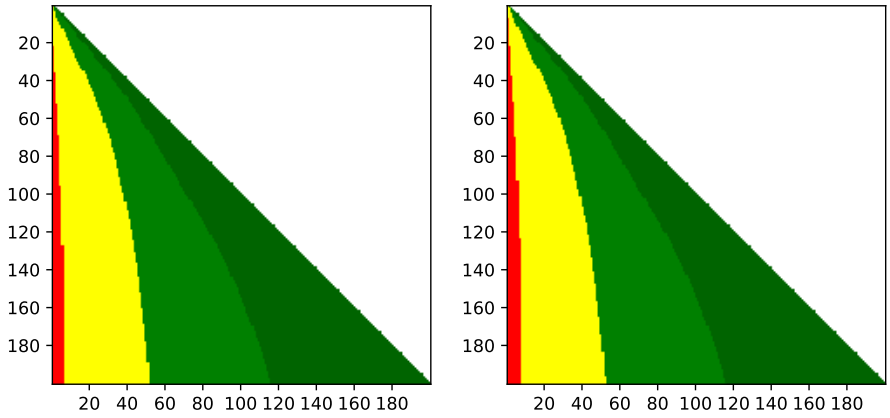

FIG 5. *Left panel: the top-left* 200×200 *corner of the arithmetic-mean discovery matrix for the* BRCA *dataset for* $B := 10000$*, using Jeffreys's thresholds. Right panel: its version (based on* (44)*) that is only approximately valid.*

TABLE 3

Summaries of the last row of the arithmetic-mean discovery matrix for different values of d. Column "strong" contains the number of entries that are greater than 10 (all of them are below $10^{3/2}$*, and so they provide strong evidence, i.e., red in our pictures). Column "at least substantial" contains the number of entries that are greater than* $10^{1/2}$ *(providing at least substantial evidence).*

d	strong	at least substantial
4	0	62
6	0	82
8	4	70
10	7	56
12	8	46
20	9	29
50	8	17
100	7	14

Dependence on the initial state of the random numbers generator (always set to 1 in our experiments) is fairly significant but does not affect our conclusions. Dependence on the value of d is also significant; the values below 10 tend to lead to higher numbers of true discoveries for Jeffrey's standard of substantial evidence, and the values above 10 to higher numbers of true discoveries for strong evidence (up to a limit; see Table 3). The literature on the Bayesian two-sample t-test quoted above seems to suggest that d should have the same order of magnitude as the number of degrees of freedom.

Comparisons

We start by comparing our methodology with that of Storey and Tibshirani (2003), which was our main source of data and ideas in this section. The main differences are:

- Storey and Tibshirani use p-values whereas we use e-values.
- Storey and Tibshirani implicitly assume that the genes are exchangeable under the null hypothesis.
- Moreover, Storey and Tibshirani assume that the p-values are weakly independent.

Strictly speaking, Storey and Tibshirani's method does not produce valid p-values, even under their null hypothesis implicitly involving gene exchangeability. This can be seen from their formula for computing the p-values,

$$p_k := \frac{\sum_{b=1}^{B} \left| \left\{ j : \left| t_j^{0b} \right| \geq |t_k| , j = 1, \dots, 3170 \right\} \right|}{3170 \cdot B} \tag{42}$$

(the last displayed equation in their Appendix, Remark C), where t_k is the t-statistic for gene k and t_j^{0b} is the t-statistic for gene j with the labels BRCA1 and BRCA2 randomly permuted (for the bth random permutation, $b = 1, \dots, B$ and $B := 100$). The numerator of (42) can well be zero (and it is in one case: see (38)).

To turn the expression (42) into a valid p-value (under the null hypothesis of gene exchangeability and label uninformativeness), it suffices to add 1 to the numerator and denominator of (42); cf. Lehmann and Romano (2022, (17.7)), Hemerik and Goeman (2018), and the method of conformal prediction (Vovk, Gammerman and Shafer, 2005). Namely,

$$\frac{\sum_{b=1}^{B} \left| \left\{ j : \left| t_j^{0b} \right| \geq |t_k| , j = 1, \dots, 3170 \right\} \right| + 1}{3170 \cdot B + 1} \tag{43}$$

is a valid p-value. The intuition behind the expression (43) is that, to see how well t_k conforms to the multiset of size $3170 \cdot B$ consisting of t_j^{0b}, we add t_k to the multiset before computing the rank p-value. Since we are comparing the t-statistic for gene k with t-statistics for other genes in (42) and (43), we are implicitly assuming gene exchangeability.

Under gene exchangeability, for computing base e-values, we can use the formula

$$e_k := \frac{T_k}{\frac{1}{3170 \cdot B + 1} \left(\sum_{b=1}^{B} \sum_{j=1}^{3170} T_j^{0b} + T_k \right)},$$

TABLE 4
The Benjamini–Hochberg and Benjamini–Yekutieli procedures applied to the BRCA *dataset (the three entries of "1" are unreliable as they are based on a zero p-value).*

assumption	5%	1%
independence	88	1
arbitrary dependence	1	1

in analogy with (43). This gives an e-variable under the assumption that the labels are uninformative and the genes are exchangeable.

To avoid the assumption of gene exchangeability, we use the expression (39) thus avoiding comparing the statistic pertaining to gene k to statistics pertaining to other genes. Our value of B, $B = 10000$, is much larger than Storey and Tibshirani's $B = 100$.

We can also introduce a simplified version of (39):

$$e_k := \frac{T_k}{\frac{1}{B}\sum_{b=1}^{B} T_k^{0b}}, \tag{44}$$

in analogy with (42). This version may be more intuitive, but it is only approximately valid for large B and ceases to be valid for small B. The difference shows, e.g., in the fact that e_k defined via (39) is bounded above by $B + 1$ whereas e_k defined via (44) is potentially unbounded; such a difference can be significant if B is small. When $B = 10000$, there is not much difference between using (39) and using (44): see the right panel of Figure 5, which uses (44).

A useful role of the version (44) may be to check whether the value of B in (39) is sufficiently large. In the case of Figure 5, the approximation is good, which suggests that B is sufficiently large. However, in the case of Figure 6, where $B = 100$ (as in Storey and Tibshirani (2003)), the right-hand panel, which uses (44), looks far too good to be valid. On the other hand, the left-hand panel, which uses (39), is valid but extremely conservative.

Results given by `hommel` are either poor (when independence is assumed) or extremely poor (under arbitrary dependence). The former are given in Figure 7 and the latter are given in Appendix C (Figure 11).

The Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) rejects 88 null hypotheses at FDR $q := 0.05$ and 1 null hypothesis at FDR $q := 0.01$ for Storey and Tibshirani's list of p-values (of course, we will always reject at least 1 null hypothesis because of the zero p-value on their list). However, this procedure assumes independence. Under arbitrary dependence, we can control FDR by replacing q by $q/\sum_{k=1}^{K} k^{-1}$ (Benjamini and Yekutieli, 2001, Theorem 1.3). This leads to rejecting 1 null hypothesis even at FDR 0.05, which is both poor and unwarranted. These results are summarized in Table 4.

Algorithm 3 Arithmetic-mean discovery matrix AM

Input: An increasing sequence of e-values $e_1 \le \dots \le e_K$.

1: $s_1 := e_1$
2: **for** $k = 2, \dots, K$ **do**
3: $\quad s_k := s_{k-1} + e_k$
4: **for** $r = 1, \dots, K$ **do**
5: $\quad \sigma_{r,r} := e_{K-r+1}$
6: $\quad$ **for** $j = r-1, \dots, 1$ **do**
7: $\quad\quad \sigma_{r,j} := \sigma_{r,j+1} + e_{K-j+1}$
8: **for** $r = 1, \dots, K$ **do**
9: $\quad$ **for** $j = 1, \dots, r$ **do**
10: $\quad\quad \mathrm{AM}_{r,j} := \sigma_{r,j}/(r-j+1)$
11: $\quad\quad$ **for** $i = 1, \dots, K-r$ **do**
12: $\quad\quad\quad e := (\sigma_{r,j} + s_i)/(r-j+1+i)$
13: $\quad\quad\quad$ **if** $e < \mathrm{AM}_{r,j}$ **then**
14: $\quad\quad\quad\quad \mathrm{AM}_{r,j} := e$

9. EFFICIENT IMPLEMENTATION OF ALGORITHM 2 FOR THE ARITHMETIC MEAN

Algorithm 2 is a generic algorithm that works for any symmetric e-merging function F. In general, computing one row of the discovery e-matrix takes time $O(K^3)$ if we assume that the base e-merging function F can be computed in time linear in the number of arguments. This assumption is correct for the arithmetic mean and, provided the arguments are sorted, the Simes e-merging function (see Appendix B). The overall computational complexity for the full discovery e-matrix is very high, $O(K^4)$.

A more efficient implementation of Algorithm 2 for the arithmetic mean is given as Algorithm 3, which uses arrays s_k (the sum of the first k base e-values) and $\sigma_{r,j}$ (the sum of the base e-values with indices in $S_{r,j}$ in the notation of Algorithm 2). There is a preprocessing stage (lines 1–3) taking time $O(K)$ and another preprocessing stage (lines 4–7) taking time $O(K^2)$; the loop in lines 6–7 is executed in the decreasing order of j, and in particular it is not executed when $r = 1$ (this also applies to two similar loops in Algorithm 4). After that computing each row of the arithmetic mean discovery matrix takes time $O(K^2)$. The overall time is $O(K^3)$.

An even more efficient implementation of Algorithm 2 is given as Algorithm 4. This algorithm computes one row of the arithmetic mean discovery matrix in time $O(K)$, which gives the overall time $O(K^2)$. Both $O(K)$ and $O(K^2)$ are clearly optimal in this context. The ability to compute efficiently individual rows is useful when the discovery matrix is big; e.g., it can be too big to fit in computer memory.

We are using essentially the same array s as in Algorithm 3 (now we extend it by adding $s_0 := 0$), and the array σ in Algorithm 4 is one row of the array σ in Algorithm 3;

$$s_k = e_1 + \dots + e_k, \qquad k = 0, \dots, K-r,$$
$$\sigma_j := e_{K-r+1} + \dots + e_{K-j+1}, \qquad j = 1, \dots, r.$$

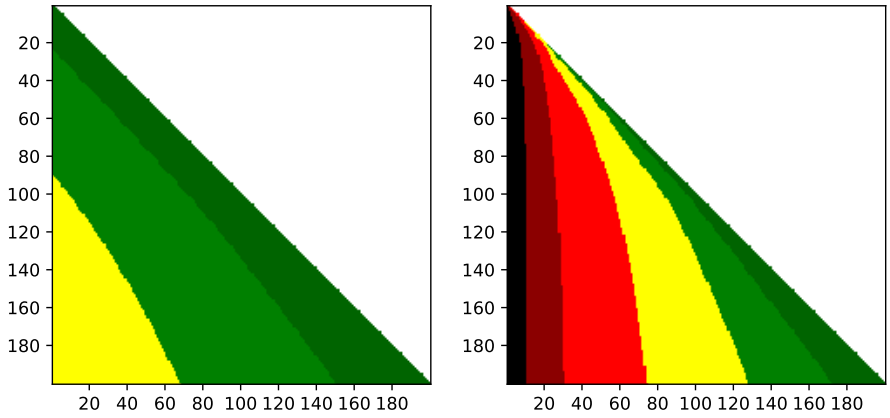

FIG 6. *Left panel: the top-left* 200×200 *corner of the arithmetic-mean discovery matrix for the* BRCA *dataset for* $B := 100$. *Right panel: its simplified version whose lack of validity is visible.*

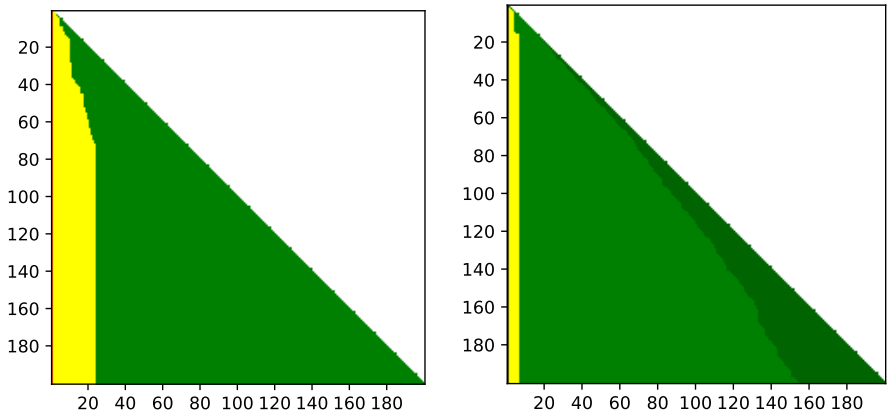

FIG 7. *Left panel: the top-left corner of the GWGS discovery p-matrix for the* BRCA *dataset for Fisher's thresholds 1% and 5%, assuming independence. Right panel: analogous picture for Jeffreys's thresholds applied to the VS bounds for the entries of this matrix.*

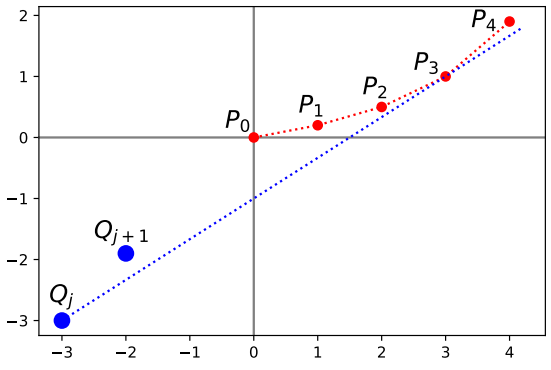

FIG 8. *Geometry behind Algorithm 4.*

Of course, there is no need to recompute the array s for each row r of the discovery matrix.

The geometry behind Algorithm 4 is shown in Figure 8. The coordinates of each of the points P_k, $k = 0, \dots, K - r$, are (k, s_k), and the coordinates of the point Q_j, where $j \in \{1, \dots, r\}$, are $-(r - j + 1, \sigma_j)$. Since the sequence $e_1, \dots, e_{K-r}$ is increasing, connecting the points $P_0, P_1, \dots$ in this order (see the red line in Figure 8) gives us the graph of a convex function.

The command **break** in line 12 means leaving the loop, as in Python or R; in this context, it is equivalent to "go to line 15". The variable k in line 13 is the index of the "current vertex" P_k; we start from the rightmost P_k in line 7 and then keep moving left. Figure 8 illustrates the execution of Algorithm 4 when $k = 3$, so that the current vertex is P_3.

For each $j = 1, \dots, r$, the iteration of the loop in lines 9–15 of Algorithm 4 computes the slope of the straight line (shown in blue) passing through Q_j and touching the red line from below. The validity of the algorithm follows from the point Q_{j+1} lying at or above the blue line, for each $j = 1, \dots, r - 1$. Let us check the last statement. If $k < K - r$, the slope of the blue line is at most $e_{k+1} \le$

Algorithm 4 One row of the arithmetic-mean discovery matrix in time $O(K)$

Input: Increasing sequence of e-values $e_1 \le \dots \le e_K$ and row number $r \in \{1, \dots, K\}$ of the discovery matrix.

$s_0 := 0$
for $k = 1, \dots, K - r$ **do**
  $s_k := s_{k-1} + e_k$
$\sigma_r := e_{K-r+1}$
for $j = r - 1, \dots, 1$ **do**
  $\sigma_j := \sigma_{j+1} + e_{K-j+1}$
$k := K - r$
for $j = 1, \dots, r$ **do**
  $\texttt{slope} := \frac{s_k + \sigma_j}{k+r-j+1}$
  for $i = k - 1, \dots, 0$ **do**
    $\texttt{new\_slope} := \frac{s_i + \sigma_j}{i+r-j+1}$
    if $\texttt{new\_slope} > \texttt{slope}$ **then break**
    $k := i$
    $\texttt{slope} := \texttt{new\_slope}$
  $\mathrm{AM}_{r,j} := \texttt{slope}$

e_{K-r}. On the other hand, the slope of the line going from Q_j to Q_{j+1} is $e_{K-j+1} \ge e_{K-r}$. It remains to consider the case $k = K - r$. In this case, it suffices to notice that the slope e_{K-j+1} of the line going from Q_j to Q_{j+1} is greater than or equal to the average of $e_1, \dots, e_{K-j+1}$, which is the slope of the line going from Q_j to P_{K-r}.

If we are only interested in the positions where discovery vectors or matrices exceed a given threshold, we can also use algorithms described in Tian et al. (2021).

10. CONCLUSION

The main technical tool of this paper, *e*-values, has important advantages over *p*-values. The advantage that we have found most useful here is the easiness of merging *e*-values: the arithmetic average of *e*-values is an *e*-value, and this is the only useful symmetric method of merging *e*-values. Other advantages were mentioned in Section 1, such as the open nature of *e*-values allowing their sequential updating.

We have described methods for multiple hypothesis testing using *e*-values and demonstrated their use in simulation and empirical studies. We believe that these methods, being simpler and more powerful, are preferred to methods using *p*-values unless the final result must be stated in terms of *p*-values. Besides, our methods do not depend on the base *e*-values being independent, and under arbitrary dependence, they are sometimes competitive with results based on *p*-values even when the final result is to be stated in terms of *p*-values.

One of the obvious directions of further research is to extend our methods to non-symmetric problems of multiple hypothesis testing (cf. Genovese, Roeder and Wasserman (2006)), in which different *e*-values may be assigned different weights. Our procedure for multiple hypothesis testing is generic and does not have to rely on unweighted arithmetic averaging.

Acknowledgments

We are grateful to Peter Westfall for his advice about the literature on Bayesian two-sample *t*-tests. We thank Glenn Shafer, Aaditya Ramdas, and participants in the course "Game-theoretic statistics" (January–April 2021) for helpful comments. The presentation was greatly improved as result of the comments by two referees, an Associate Editor, and the Editor (Sonia Petrone). For most of our simulation and empirical studies in Sections 7–8 we used Python. We also used the R package `hommel` (Goeman, Meijer and Krebs, 2019) and a dataset available in the R package `qvalue` (Storey et al., 2019).

V. Vovk's research has been partially supported by Amazon, Astra Zeneca, and Stena Line. R. Wang is supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2018-03823, RGPAS-2018-522590).

REFERENCES

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57** 289–300.

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29** 1165–1188.

BENJAMINI, Y., VEAUX, R. D. D., EFRON, B., EVANS, S., GLICKMAN, M., GRAUBARD, B. I., HE, X., MENG, X.-L., REID, N., STIGLER, S. M., VARDEMAN, S. B., WIKLE, C. K., WRIGHT, T., YOUNG, L. J. and KAFADAR, K. (2021). The ASA president's task force statement on statistical significance and replicability. *Annals of Applied Statistics* **15** 1084–1085.

BERNARDO, J. M. and SMITH, A. F. M. (2000). *Bayesian Theory*. Wiley, Chichester.

BERNOULLI, J. (1713). *Ars Conjectandi*. Thurnisius, Basel.

CASELLA, G. and BERGER, R. L. (2002). *Statistical Inference*, second ed. Duxbury, Pacific Grove, CA.

COURNOT, A.-A. (1843). *Exposition de la théorie des chances et des probabilités*. Hachette, Paris.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

DE FINETTI, B. (2017). *Theory of Probability*. Wiley, Chichester.

DUBOIS, D. and PRADE, H. (1988). *Possibility Theory*. Plenum Press, New York.

FISHER, R. A. (1973). *Statistical Methods and Scientific Inference*, Third ed. Hafner, New York.

GÁCS, P. (2005). Uniform test of algorithmic randomness over a general space. *Theoretical Computer Science* **341** 91–137.

GENOVESE, C. R., ROEDER, K. and WASSERMAN, L. (2006). False discovery control with *p*-value weighting. *Biometrika* **93** 509–524.

GENOVESE, C. R. and WASSERMAN, L. (2004). A stochastic process approach to false discovery control. *Annals of Statistics* **32** 1035–1061.

GOEMAN, J. J., HEMERIK, J. and SOLARI, A. (2021). Only closed testing procedures are admissible for controlling false discovery proportions. *Annals of Statistics* **49** 1218–1238.

GOEMAN, J. J., MEIJER, R. and KREBS, T. (2019). `hommel`: Methods for closed testing with Simes inequality, in particular Hommel's method R package version 1.5, available on CRAN.

GOEMAN, J. J., ROSENBLATT, J. D. and NICHOLS, T. E. (2019). The harmonic mean p-value: Strong versus weak control, and the assumption of independence. *Proceedings of the National Academy of Sciences* **116** 23382–23383.

GOEMAN, J. J. and SOLARI, A. (2011a). Multiple testing for exploratory research. *Statistical Science* **26** 584–597. Correction: **28** 464.

GOEMAN, J. J. and SOLARI, A. (2011b). Multiple testing for exploratory research: Rejoinder. *Statistical Science* **26** 608–612.

GOEMAN, J. J., MEIJER, R. J., KREBS, T. J. P. and SOLARI, A. (2019). Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika* **106** 841–856.

GÖNEN, M., JOHNSON, W. O., LU, Y. and WESTFALL, P. H. (2005). The Bayesian two-sample *t* test. *American Statistician* **59** 252–257.

GÖNEN, M., JOHNSON, W. O., LU, Y. and WESTFALL, P. H. (2019). Comparing objective and subjective Bayes factors for the two-sample comparison: the classification theorem in action. *American Statistician* **73** 22–31.

GOOD, I. J. (1958). Significance tests in parallel and in series. *Journal of the American Statistical Association* **53** 799–813.

GRÜNWALD, P., DE HEIDE, R. and KOOLEN, W. M. (2020). Safe testing Technical Report No. arXiv:1906.07801 [math.ST], arXiv.org e-Print archive.

GRÜNWALD, P. and VAN OMMEN, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis* **12** 1069–1103.

GUINDANI, M., MÜLLER, P. and ZHANG, S. (2009). A Bayesian discovery procedure. *Journal of the Royal Statistical Society B* **71** 905–925.

HEDENFALK, I., DUGGAN, D., CHEN, Y., RADMACHER, M., BITTNER, M., SIMON, R., MELTZER, P., GUSTERSON, B., ESTELLER, M., KALLIONIEMI, O.-P., WILFOND, B., BORG, A. and TRENT, J. (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine* **344** 539–548.

HELD, L. (2019). On the Bayesian interpretation of the harmonic mean *p*-value. *Proceedings of the National Academy of Sciences* **116** 5855–5856.

HEMERIK, J. and GOEMAN, J. J. (2018). Exact testing with random permutations. *Test* **27** 811–825.

HEMERIK, J., SOLARI, A. and GOEMAN, J. J. (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika* **106** 635–649.

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6** 65–70.

HOMMEL, G. (1986). Multiple test procedures for arbitrary dependence structures. *Metrika* **33** 321–336.

JEFFREYS, H. (1961). *Theory of Probability*, Third ed. Oxford University Press, Oxford.

KLEENE, S. C. (1967). *Mathematical Logic*. Wiley, New York.

LEHMANN, E. L. (2011). *Fisher, Neyman, and the Creation of Classical Statistics*. Springer, New York.

LEHMANN, E. L. and ROMANO, J. P. (2022). *Testing Statistical Hypotheses*, Fourth ed. Springer, Cham.

LEVIN, L. A. (1976). Uniform tests of randomness. *Soviet Mathematics Doklady* **17** 337–340.

LY, A., VERHAGEN, J. and WAGENMAKERS, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology* **72** 19–32.

NEYMAN, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection (with discussion). *Journal of the Royal Statistical Society* **97** 558–625.

NEYMAN, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London A* **236** 333–380.

NEYMAN, J. (1941). Fiducial argument and the theory of confidence intervals. *Biometrika* **32** 128–150.

ROMANO, J. P. and WOLF, M. (2007). Control of generalized error rates in multiple testing. *Annals of Statistics* **35** 1378–1408.

ROUDER, J. N., SPECKMAN, P. L., SUN, D. and MOREY, R. D. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review* **16** 225–237.

SARKAR, S. K. (2011). Simes' test in multiple testing. In *International Encyclopedia of Statistical Science* (M. Lovric, ed.) 1325–1327. Springer, Berlin.

SCHERVISH, M. J. (1995). *Theory of Statistics*. Springer, New York.

SELLKE, T., BAYARRI, M. J. and BERGER, J. (2001). Calibration of P-values for testing precise null hypotheses. *American Statistician* **55** 62–71.

SHAFER, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ.

SHAFER, G. (2007). From Cournot's principle to market efficiency. In *Augustin Cournot: Modelling Economics* (J.-P. Touffut, ed.) 55–95. Edward Elgar, Cheltenham.

SHAFER, G. (2021). The language of betting as a strategy for statistical and scientific communication (with discussion). *Journal of the Royal Statistical Society A* **184** 407–478.

SHAFER, G. (2022). Bayesian, fiducial, frequentist. In *Handbook on Bayesian, Fiducial and Frequentist (BFF) Inferences* (J. Berger, X.-L. Meng, N. Reid and M. Xie, eds.) Chapman and Hall (to appear).

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

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

STOREY, J. D., DAI, J. Y. and LEEK, J. T. (2007). The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics* **8** 414–432.

STOREY, J. D. and TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the USA* **100** 9440–9445.

STOREY, J. D., BASS, A. J., DABNEY, A. and ROBINSON, D. (2019). `qvalue`: Q-value estimation for false discovery rate control R package version 2.18.0, available on Bioconductor.

STUART, A., ORD, K. J. and ARNOLD, S. (1999). *Kendall's Advanced Theory of Statistics* **2a: Classical inference and the linear model**, Sixth ed. Arnold, London.

TIAN, J., CHEN, X., KATSEVICH, E., GOEMAN, J. and RAMDAS, A. (2021). Large-scale simultaneous inference under dependence Technical Report No. arXiv:2102.11253 [math.ST], arXiv.org e-Print archive. To appear in the *Scandinavian Journal of Statistics*.

VOVK, V., GAMMERMAN, A. and SHAFER, G. (2005). *Algorithmic Learning in a Random World*. Springer, New York.

VOVK, V. and V'YUGIN, V. V. (1993). On the empirical validity of the Bayesian method. *Journal of the Royal Statistical Society B* **55** 253–266.

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

VOVK, V. and WANG, R. (2020b). True and false discoveries with independent e-values Technical Report No. arXiv:2003.00593 [stat.ME], arXiv.org e-Print archive.

VOVK, V. and WANG, R. (2021). E-values: Calibration, combination, and applications. *Annals of Statistics* **49** 1736–1754.

VOVK, V., WANG, B. and WANG, R. (2022). Admissible ways of merging p-values under arbitrary dependence. *Annals of Statistics* **50** 351–375.

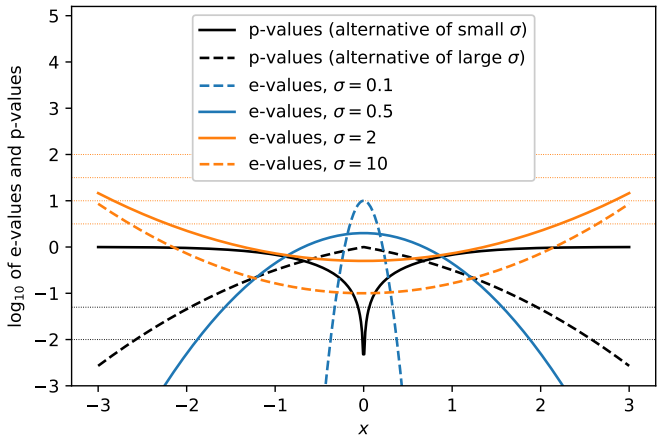

FIG 9. *Some p-values and e-values for testing $N(0,1)$ embedded into the family $N(0,\sigma^2)$. This complements the right panel of Figure 2.*

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

WANG, M. and LIU, G. (2016). A simple two-sample Bayesian t-test for hypothesis testing. *American Statistician* **70** 195–201.

WANG, R. and RAMDAS, A. (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society B* **84** 822–852.

WIGGINS, G. A. R., WALKER, L. C. and PEARSON, J. F. (2020). Genome-wide gene expression analyses of BRCA1- and BRCA2-associated breast and ovarian tumours. *Cancers* **12** 3015.

WILSON, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences* **116** 1195–1200.

APPENDIX A: MORE *E*-VARIABLES FOR TESTING $N(0,1)$

Embedding the null hypothesis $N(0,1)$ into the family $N(\mu,1)$, as in the right panel of Figure 2, is not the only option, and Figure 9 gives results for the family $N(0,\sigma^2)$, $\sigma>0$ being the standard deviation. The figure shows the likelihood ratios $\mathrm{d}N(0,\sigma^2)/\mathrm{d}N(0,1)$ as *e*-variables for a range of σ, and it shows $P_<(x):=\chi^2(x^2)$ and $P_>(x):=1-\chi^2(x^2)$ as *p*-variables, where χ^2 is the distribution function of χ^2 with 1 degree of freedom (with $P_<$ based on $\sigma<1$ as the alternative hypothesis, and $P_>$ based on $\sigma>1$).

To get a non-trivial two-sided confidence interval for σ, we need to merge the two *p*-variables (by, say, using the Bonferroni merging function $B(P_<,P_>):=2\min(P_<,P_>)$) or merge two of the *e*-variables, one for $\sigma>1$ and the other for $\sigma<1$ (which can be done more efficiently, simply by averaging them).

APPENDIX B: USING OTHER *E*-MERGING FUNCTIONS

In this appendix, we briefly explore discovery *e*-matrices using *e*-merging functions F other than the arithmetic mean. The *Bonferroni e-merging function* is the following lower bound for (31):

$$B(e_1,\dots,e_n):=\frac{1}{n}\max_{i\in\{1,\dots,n\}}e_i. \tag{45}$$

A better lower bound for (31) is given by the *Simes e-merging function*

$$S(e_1,\dots,e_n):=\max_{i\in\{1,\dots,n\}}\frac{ie_{[i]}}{n}, \tag{46}$$

where $e_{[i]}$ is the ith largest *e*-value among e_i, $i\in\{1,\dots,n\}$ (Vovk and Wang, 2021, end of Section 6): $e_{[1]},\dots,e_{[n]}$ is the permutation of $e_1,\dots,e_n$ satisfying $e_{[1]}\ge\dots\ge e_{[n]}$.

Our discussion of *e*-values and *p*-values in Section 3 suggests that the function $t\mapsto 1/t$ transforms *e*-values into *p*-values (cf. (8)) and transforms *p*-values into approximate *e*-values (cf. (6) for a small $\kappa\in(0,1)$); of course, the word "approximate" is used here in a crude sense (in the spirit of the algorithmic theory of randomness). Under this correspondence, the Bonferroni *e*-merging function (45) turns into the Bonferroni merging function for *p*-values, and the Simes *e*-merging function (46) turns into the Simes merging function for *p*-values. The dominating arithmetic-mean *e*-merging function (31) corresponds to using the harmonic mean for merging *p*-values, and indeed the harmonic mean has been discussed recently in this role (Wilson, 2019), sometimes with a similar justification based on the VS bound (Held, 2019). However, the harmonic mean is not a valid function for merging *p*-values (Goeman, Rosenblatt and Nichols, 2019) unless multiplied by, say, $2.5\ln K$ for $K\ge 3$ (Vovk and Wang, 2020a).

With $F_{\mathbf{e}}(I)$ of (35) specialized to the Bonferroni lower bound

$$F_{\mathbf{e}}(I)=B_{\mathbf{e}}(I):=\frac{1}{|I|}\max_{i\in I}e_i,\quad I\subseteq\{1,\dots,K\},\ I\ne\emptyset,$$

Algorithms 1 and 2 have the same interpretation as before (although the results are not as good since they are based on more conservative *e*-values). However, they simplify, especially Algorithm 2, whose Bonferroni implementation is given as Algorithm 5. In line 1 of Algorithm 5 we initialize the adjusted Bonferroni *e*-value, in line 3 we compute the raw Bonferroni *e*-value, and in line 4 we adjust it. The algorithm produces a matrix BM with constant

Algorithm 5 Bonferroni discovery *e*-matrix BM

Input: An increasing sequence of *e*-values $e_1\le\dots\le e_K$.

1: $a:=\infty$
2: **for** $j=1,\dots,K$ **do**
3: $\quad B:=e_{K-j+1}/(K-j+1)$
4: $\quad$ **if** $a>B$ **then** $a:=B$
5: $\quad$ **for** $r=j,\dots,K$ **do**
6: $\quad\quad \mathrm{BM}_{r,j}:=a$

columns and takes time $O(K)$ per column; this time is spent simply by writing one value repeatedly. The resulting computational complexity $O(K^2)$ is clearly the optimal one.

Whereas Bonferroni-type procedures often perform well when the goal is family-wise validity (see, e.g., Vovk and Wang (2021, Figures 3 and 4)), their performance tends to deteriorate for less demanding notions of validity. (In terms of *p*-values, this phenomenon is discussed in, e.g., Goeman and Solari (2011b, Section 1).) Comparing the left panel of Figure 10 with the upper left panel of Figure 4 we can see that the *e*-Bonferroni method is much worse than arithmetic averaging when the goal is to control the number of false discoveries.

The poor performance of the *e*-Bonferroni method is clear already from the upper left panel of Figure 4: the areas of different colours are far from been vertical at the top, where they curve left. It is clear that every discovery *e*-matrix that is dominated by this one and has vertical boundaries between different colours (such as *e*-Bonferroni) is going to be much worse.

The right panel of Figure 10 shows the Simes *e*-matrix, based on (46), in the situation of Figure 4. It is intermediate between AM and Bonferroni and, remarkably, it looks better than the GWGS discovery *p*-matrix transformed by applying the VS bound to its elements (the lower right panel of Figure 4).

REMARK B.1. We can quantify the quality of the lower bounds (45) and (46) of the arithmetic mean F by the inequalities

$$B(e_1,\dots,e_n) \le S(e_1,\dots,e_n) \le F(e_1,\dots,e_n),$$

$$1 \le \frac{F(e_1,\dots,e_n)}{S(e_1,\dots,e_n)} \le \sum_{k=1}^{n} \frac{1}{k} \le \ln n + 1,$$

$$1 \le \frac{F(e_1,\dots,e_n)}{B(e_1,\dots,e_n)} \le n, \quad 1 \le \frac{S(e_1,\dots,e_n)}{B(e_1,\dots,e_n)} \le n,$$

all of which are tight (achievable as equality for any n), apart from $\le \ln n + 1$ (which is tight only for $n = 1$). Since $n := |I| \le K$ in (33), this gives bounds for the ratios of the corresponding elements of the discovery *e*-matrices built on top of B, S, and F.

APPENDIX C: COMPARISON WITH THE GWGS PROCEDURE

First we discuss the GWGS multiple testing procedure in the form described in Goeman and Solari (2011a, Section 2) and in terms of our definitions. Let $F : \cup_{n=1}^{\infty}[0,1]^n \to [0,1]$ be a *p-merging function*, i.e., a monotonic function transforming *p*-variables into a *p*-variable: whenever $P_1,\dots,P_n$ are *p*-variables for some $n \in \{1,2,\dots\}$, $F(P_1,\dots,P_n)$ is a *p*-variable. Suppose that F is symmetric. With such an F we can associate the following analogue of (33) in terms of *p*-values:

$$D^R_{p,F}(j) := \max_{I:|R\setminus I|<j} F(p_i, i \in I) \ge \square_p^{g_R}(\{j, j+1, \dots\} \mid \omega), \tag{47}$$

where the *p*-test P is defined by $P_\theta := F(P_k : k \in I_\theta)$ (analogously to (32)); we leave the dependence on $p_1,\dots,p_K$ implicit, following Goeman and Solari and similarly to the case of *e*-values.

Goeman and Solari prefer the inverse to the function (47), which they denote $f_\alpha(R)$, suppressing the dependence on $p_1,\dots,p_K$; we consider it as function of $\alpha \in [0,1]$, which is interpreted as significance level. We will see that this function satisfies

$$f_\alpha(R) \ge j \iff D^R_{p,F}(j) \le \alpha \tag{48}$$

(and this equivalence can serve as definition of f). Therefore, it gives us the lower *p*-confidence bound on the number of true discoveries at significance level α.

For the reader familiar with Goeman and Solari (2011a), we will check that their definition indeed satisfies (48). They first define their bound

$$t_\alpha(R) := \max\{|I| \mid I \subseteq R, I \notin \mathcal{X}\}$$

on the number of false discoveries, where

$$\mathcal{X} := \{I \mid \forall J \supseteq I : J \in \mathcal{U}\}$$

are the subsets of $\{1,\dots,K\}$ rejected by the closed testing procedure, and

$$\mathcal{U} := \{I \mid F(p_i, i \in I) \le \alpha\}$$

are the subsets of $\{1,\dots,K\}$ rejected by F; in general, I and J will run over the subsets of $\{1,\dots,K\}$. Then they define their bound on the number of true discoveries as

$$f_\alpha(R) := |R| - t_\alpha(R). \tag{49}$$

The equivalence (48) can be checked as follows:

$$\begin{aligned}
&f_\alpha(R) \ge j \iff t_\alpha(R) \le |R| - j \\
&\iff \max\{|I| \mid I \subseteq R, I \notin \mathcal{X}\} \le |R| - j \\
&\iff (\forall I \subseteq R : I \notin \mathcal{X} \Rightarrow |I| \le |R| - j) \\
&\iff (\forall I \subseteq R : |I| > |R| - j \Rightarrow I \in \mathcal{X}) \\
&\iff (\forall I \subseteq R : |I| > |R| - j \Rightarrow (\forall J \supseteq I : J \in \mathcal{U})) \\
&\iff (\forall J : |J \cap R| > |R| - j \Rightarrow J \in \mathcal{U}) \\
&\iff (\forall J : |R \setminus J| < j \Rightarrow J \in \mathcal{U}) \\
&\iff \max_{J:|R\setminus J|<j} F(p_i, i \in J) \le \alpha \iff D^R_{p,F}(j) \le \alpha.
\end{aligned}$$

In Section 7 we mentioned that Goeman and Solari (2011b, Section 3) introduced a version of the notion of

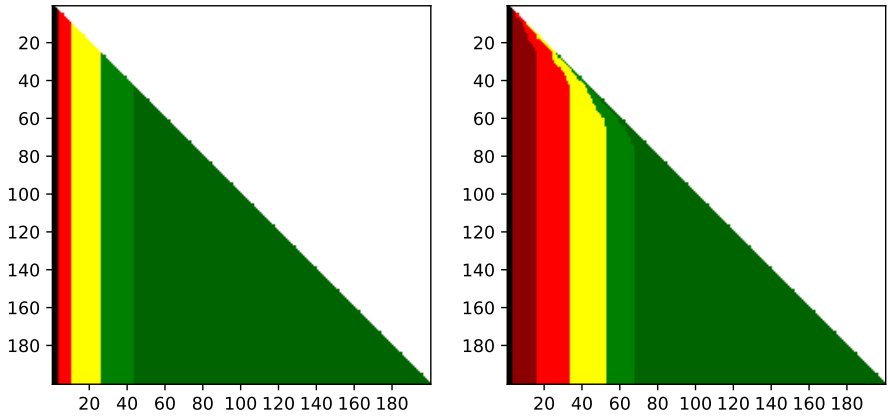

FIG 10. *Left panel: the Bonferroni discovery e-matrix (given by Algorithm 5) in the situation of Figure 4, using Jeffreys's thresholds. Right panel: the corresponding Simes discovery e-matrix.*

a discovery p-vector. Namely they introduced the confidence distribution whose quantile function is $\alpha \mapsto t_\alpha(R)$. We can interpret (48) as $D^R_{p,F}$ being the distribution function whose quantile function is $\alpha \mapsto f_\alpha(R)$. Since f and t are so closely connected (see (49)), the discovery p-vector is closely connected to Goeman and Solari's confidence distribution function.

A property of completeness for the GWGS procedure

Goeman, Hemerik and Solari (2021) have shown that the GWGS procedure is the only admissible one for controlling true discoveries. Doesn't this mean that the e-version of this procedure, Algorithm 1, is inadmissible?

Similarly to (33), the interpretation of the property of validity (47) in terms of a Fisher-type disjunction is: the rejection set R contains at least j true discoveries unless the outcome ω is $D^R_{p,F}(j)$-strange. We can indeed obtain a property of validity of the same kind (i.e., in terms of p-values) for the procedure of Algorithm 1. Our interpretation of (33) was that the rejection set R contains at least j true discoveries unless ω has an e-value of $D^R_{e,F}(j)$ or more. Applying the canonical e-to-p calibrator (8), we can see that R contains at least j true discoveries unless ω has a p-value of $1/D^R_{e,F}(j)$ or less. We have the same property of validity, but with $1/D^R_{e,F}(j)$ in place of $D^R_{p,F}(j)$. By Goeman et al.'s result, the procedure with $1/D^R_{e,F}(j)$ is either a GWGS procedure or inadmissible. Since the operation of e-to-p calibration is so crude, there is no doubt that this procedure is inadmissible in non-degenerate cases. The source its inadmissibility is the inefficiency of converting e-values into p-values, and it can be shown that Algorithm 1 itself is admissible when F is arithmetic averaging.

More results for the `hommel` package

In conclusion, we give one more figure demonstrating the work of the `hommel` package. Figure 11 shows results (very poor) for the `BRCA` dataset without assuming independence. In particular, the right panel is much worse than the left panel of Figure 5, which also does not assume independence.

APPENDIX D: GENERALIZED BAYES AND BOOSTING A WEAK SIGNAL

When defining the base e-values for use in our simulation studies we just used the likelihood ratio $E(x)$ defined by (37). This is the simplest version of a Bayes factor. It usually works very well, but in some cases can be improved. Later in this appendix we will see an example where a weak signal needs to be boosted, but we start from developing tools that will allow us to do so.

Let us choose a constant $\eta > 0$ (the *learning rate*) and refer to

$$E_\eta(x) := \frac{1}{c}E(x)^\eta = \frac{1}{c}\exp(\eta\delta x - \eta\delta^2/2) \tag{50}$$

as the *generalized Bayes factor* (see, e.g., Grünwald and van Ommen (2017, 2.4) and references therein). Here $c > 0$ is the normalizing constant ensuring $\int E_\eta \,\mathrm{d}N(0,1) = 1$; a simple calculation gives

$$c = \exp\left(\eta(\eta-1)\delta^2/2\right).$$

Plugging this into (50) we obtain

$$E_\eta(x) = \exp(\eta\delta x - \eta^2\delta^2/2).$$

This gives a useful interpretation of the generalized Bayes factor: it is still the likelihood ratio, but we replace the true alternative $N(\delta, 1)$ by a false one, $N(\eta\delta, 1)$. For $\eta > 1$ we are boosting the difference between the null and alternative hypotheses.

One situation in which the likelihood ratio (37) does not work well is where we have a large number of false null hypotheses, but the true data-generating distributions

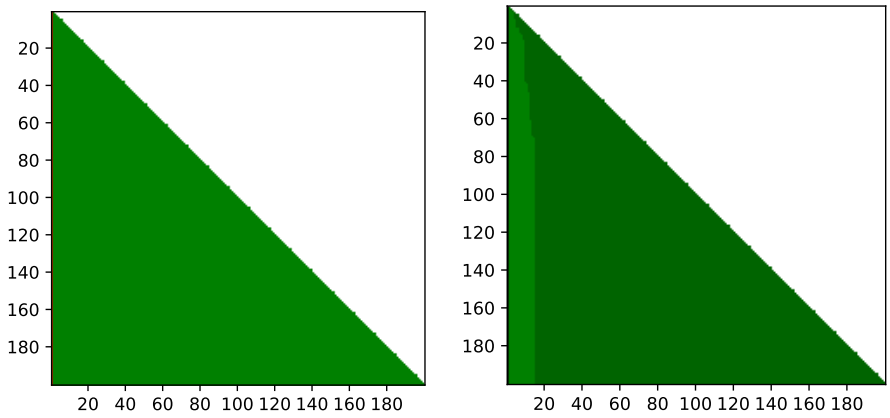

FIG 11. *Left panel: the top-left corner of the GWGS discovery p-matrix for the* BRCA *dataset for Fisher's thresholds 1% and 5%, under arbitrary dependence. Right panel: analogous picture with each entry replaced by the corresponding VS bound and using Jeffreys's thresholds.*

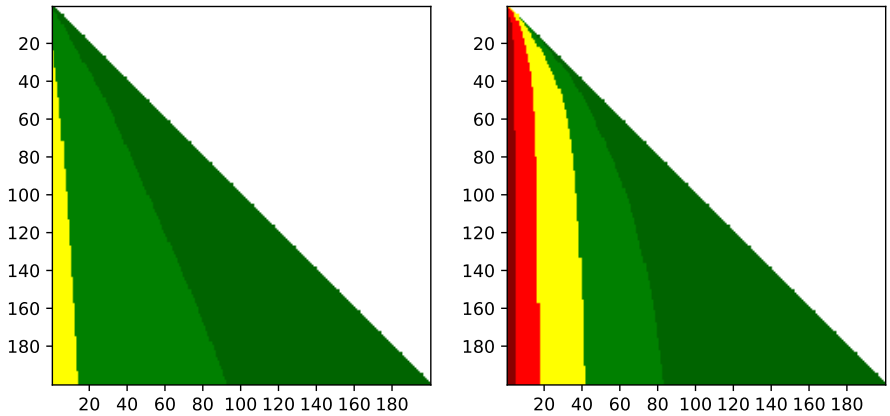

FIG 12. *Left panel: the top-left* 200×200 *corner of the arithmetic-mean discovery matrix for the simulation data with 10,000 observations,* 10% *of false hypotheses, and weak signal, using Bayes factors as base e-values, as described in text. Right panel: using generalized Bayes factors with learning rate* $\eta = 2$*. Both panels use Jeffreys's thresholds.*

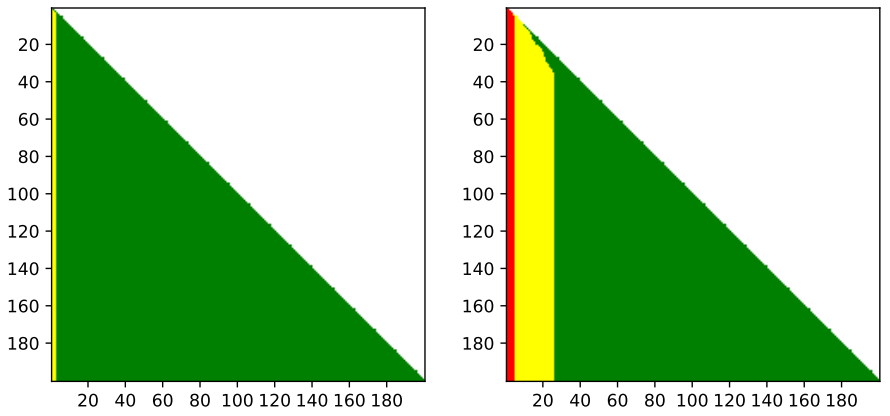

FIG 13. *Left panel: the top-left* 200×200 *corner of the GWGS discovery p-matrix for the simulation data with 10,000 observations,* 10% *of false hypotheses, and weak signal, under general dependence. Right panel: assuming independence. The colour code is based on Fisher's thresholds.*

TABLE 5
The Benjamini–Hochberg and Benjamini–Yekutieli procedures applied to the simulation data with 10,000 observations, 10% of false hypotheses, and weak signal for FDR 5% and 1%.

assumption	5%	1%
independence	84	18
arbitrary dependence	10	0

are fairly close to the null hypotheses (as it were, we have a weak signal). Figures 12–13 illustrate the case of 10,000 null hypotheses $N(0,1)$ of which 1000 are false, the true alternatives being $N(-2,1)$ (which makes the signal much weaker than in Section 7). In the left panel of Figure 12 we use the Bayes factor (37), whereas in its right panel we use the generalized Bayes factor (50) for $\eta = 2$. Using the generalized Bayes factor greatly improves the discovery *e*-matrix. The results for the GWGS procedure are given in Figure 13; they look poor, particularly so for arbitrary dependence.

Table 5 gives the numbers of rejections for the Benjamini–Hochberg and Benjamini–Yekutieli procedures. In view of Figure 12 (right panel), the results for arbitrary dependence are poor.

APPENDIX E: EMPIRICAL STUDY: GROUND TRUTH

In Section 8 we discussed a pioneering biomedical study whose results were published a long time ago (Hedenfalk et al., 2001). To evaluate the performance of various statistical techniques and their assumptions, it is natural to analyze the developments in this area of biomedicine since 2001.

The main goal of Hedenfalk et al. (2001) was to test the hypothesis that different genes are expressed by hereditary malignant breast tumours that are due to mutations in the BRCA1 and BRCA2 genes and to identify differentially expressed genes. A recent review (Wiggins, Walker and Pearson, 2020, Sections 2 and 5) compares results of nine studies, starting from Hedenfalk et al. (2001), pursuing this goal and mostly using different biological samples (therefore, not including Storey and Tibshirani (2003)). The overlap between the lists of differentially expressed genes produced by different studies is poor. In particular, only one gene has been identified as associated with BRCA1 by more than two studies. This gene, TOB1, was among the genes identified in Hedenfalk et al. (2001, Figure 2A). In Storey and Tibshirani's list of *p*-values used in Figure 7 the TOB1 gene has rank 77; in our list of *e*-values used in Figure 5 TOB1 has a slightly better rank of 51.

One reason (Wiggins, Walker and Pearson, 2020, Section 2) for the poor overlap between different studies is the genuine difficulty of the problem of differentiating mutations in the two BRCA genes while controlling for potential confounders, first of all the estrogen- and progesterone-receptor status and the subtype, which are known to affect gene expression greatly. For the dataset used in this paper, differentiation between mutations in the two genes is facilitated, e.g., by all BRCA1 samples being negative for both estrogen and progesterone receptors and majority of the BRCA2 samples being positive for both (Hedenfalk et al., 2001, Table 1). Other studies reported in Wiggins, Walker and Pearson (2020) tried to control for these confounders.

We can draw only limited conclusions from these follow-up studies. There is often a big difference between the statistical null hypothesis and the scientific hypothesis of interest. Whereas there are genuine significant differences between the BRCA1 and BRCA2 samples in the dataset, the differences are not necessarily due to their different BRCA status. A possible lesson is that in our assumptions we should err on the side of caution avoiding assuming independence or weak independence, which lead us to expect very large numbers of discoveries.