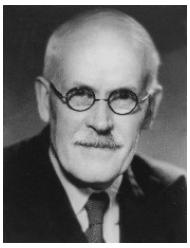
True and false discoveries with independent e-values

Vladimir Vovk Ruodu Wang



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

Project "Hypothesis testing with e-values"

Working Paper #4

First posted March 1, 2020. Last revised March 30, 2020.

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

Abstract

In this paper we use e-values (a non-Bayesian version of Bayes factors) in the context of multiple hypothesis testing assuming that the base tests produce independent e-values. Our simulation and empirical studies and theoretical considerations suggest that, under this assumption, our new algorithms are superior to the known algorithms using independent p-values and to our recent algorithms using e-values that are not necessarily independent.

Contents

1 Introduction **1**

2 Discovery matrices for independent e-values **1**

3 A simulation study **3**

4 An empirical study **5**

5 An attempt of a theoretical explanation **8**

6 Conclusion **9**

References **9**

A Towards an efficient algorithm for U_2 and other U_n **11**

1 Introduction

Our recent paper [9] gives a generic procedure for turning e-merging functions into discovery matrices and applies it to arithmetic mean. Using arithmetic mean is very natural in the case of arbitrary dependence between the base e-values, at least in the symmetric case, since arithmetic mean essentially dominates any e-merging function [9, Theorem 5.1]. But in this paper we will show that in the case of independent e-values we can greatly improve on arithmetic mean.

2 Discovery matrices for independent e-values

To make our exposition self-contained, we start from basic definitions (see our previous papers [7, 8, 9] exploring e-values for further information).

An *e-variable* on a probability space $(\Omega, \mathcal{A}, P)$ is a nonnegative extended random variable $E : \Omega \to [0, \infty]$ such that $\int E \, \mathrm{d}P \le 1$. A measurable function $F : [0, \infty]^K \to [0, \infty]$ for an integer $K \ge 1$ is an *ie-merging function* if, for any probability space and any independent e-variables $E_1, \dots, E_K$ on it, the extended random variable $F(E_1, \dots, E_K)$ is an e-variable. We will only consider ie-merging functions that are increasing in each argument and are symmetric (do not depend on the order of their arguments).

Important examples of ie-merging functions [8] are

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

We will refer to them as the *U-statistics* (they are the standard U-statistics with product as kernel). The statistics U_1 play a special role since they belong to the narrower class of *e-merging functions*, meaning that $U_1(E_1, \dots, E_k)$ is an e-variable whenever $E_1, \dots, E_K$ are e-variables (not necessarily independent).

Multiple hypothesis testing using U_1 was explored in [8, 9], and in this paper we will mainly concentrate on U_2. It will be convenient to generalize (1) to the case $n > K \ge 1$; namely, we set

$$U_n(e_1, \dots, e_K) := U_K(e_1, \dots, e_K), \quad n > K$$

(we are mostly interested in the case $n = 2$ and $K = 1$).

Let us fix the underlying *sample space* $(\Omega, \mathcal{A})$, which is simply a measurable space. Let $\mathfrak{P}(\Omega)$ be the set of all probability measure on the sample space. A *simple hypothesis* is $Q \in \mathfrak{P}(\Omega)$ and a *(composite) hypothesis* is $H \subseteq \mathfrak{P}(\Omega)$. An *e-variable* w.r. to a hypothesis H is an extended random variable $E : \Omega \to [0, \infty]$ such that $\int E \, \mathrm{d}Q \le 1$ for all $Q \in H$. It is clear that any ie-merging function transforms independent e-variables w.r. to H (i.e., independent e-variables w.r. to any $Q \in H$) to an e-variable w.r. to H.

An *e-value* is a value taken by an e-variable. Let us fix $K \ge 2$, K hypotheses $H_1, \dots, H_K$, and *independent e-variables* $E_1, \dots, E_K$ *for testing* $H_1, \dots, H_K$, in the following sense:

Algorithm 1 Discovery matrix (lower triangular)

Require: ie-merging functions F_k, $k \in \{1, \dots, K\}$.
Require: an increasing sequence of e-values $e_1 \le \dots \le e_K$.

```
for r = 1, ..., K do
    for j = 1, ..., r do
        S_{r,j} := {K − r + 1, ..., K − j + 1}
        DM^F_{r,j} := F_{r−j+1}(e_k : k ∈ S_{r,j})
        for i = 1, ..., K − r do
            e := F_{r−j+1+i}(e_k : k ∈ S_{r,j} ∪ {1, ..., i})
            if e < DM^F_{r,j} then
                DM^F_{r,j} := e
```

- for each $k \in \{1, \dots, K\}$, E_k is an e-variable w.r. to H_k;
- for any $Q \in \mathfrak{P}(\Omega)$, the e-variables E_k for which $Q \in H_k$ are independent under Q.

An *e-test* is a family $(E_Q)_{Q \in \mathfrak{P}(\Omega)}$ of nonnegative extended random variables such that $\int E_Q \, \mathrm{d}Q \le 1$ for all Q.

Let us say that a measurable function $D : \{1, \dots, K\}^2 \times [0, \infty]^K \to [0, \infty]$ is an *i-discovery matrix* if there exists an e-test $(E_Q)_{Q \in \mathfrak{P}(\Omega)}$ such that

$$\forall r \in \{1, \dots, K\} \, \forall R \subseteq \{1, \dots, K\} \, \forall j \in \{1, \dots, r\} \, \forall Q \in \mathfrak{P}(\Omega) \, \forall \omega \in \Omega : \\ \left(|R| = r \text{ and } \min_{k \in R} E_k(\omega) \ge \max_{k \notin R} E_k(\omega) \right) \\ \implies \left(|\{k \in R \mid Q \notin H_k\}| \ge j \text{ or } E_Q(\omega) \ge D_{r,j}(E_1(\omega), \dots, E_K(\omega)) \right). \quad (2)$$

To emphasize that we interpret D as a matrix, we write its values as $D_{r,j}(e_1, \dots, e_K)$. We are not interested in $D_{r,j}$ for $j > r$ and regard D as a lower triangular matrix. The intuition behind (2) is that if $D_{r,j}(E_1(\omega), \dots, E_K(\omega))$ is large and we reject r hypotheses H_k with K largest E_k, we can count on at least j true discoveries.

Algorithm 1 is one way of constructing a discovery matrix based on a family of ie-merging functions F_k, $k \in \{1, \dots, K\}$. It uses the notation $f(e_k : k \in I)$, where $I \subseteq \{1, \dots, K\}$ and f is a symmetric function of $|I|$ arguments, to mean the value of f on the sequence of e_k, $k \in I$, arranged in any order. The algorithm implements the formula

$$\mathrm{DM}^F_{r,j} := \min_{I \subseteq \{1, \dots, K\} : |R \setminus I| < j} F_{|I|}(E_i : i \in I),$$

where R is a set of indices of hypotheses with r largest e-values, and is an obvious modification of Algorithm 2 in [9]; now we apply it to arbitrary ie-merging functions (such as U_2) rather than just to arithmetic mean (i.e., U_1). As in [9], the e-values are assumed to be ordered, without loss of generality.

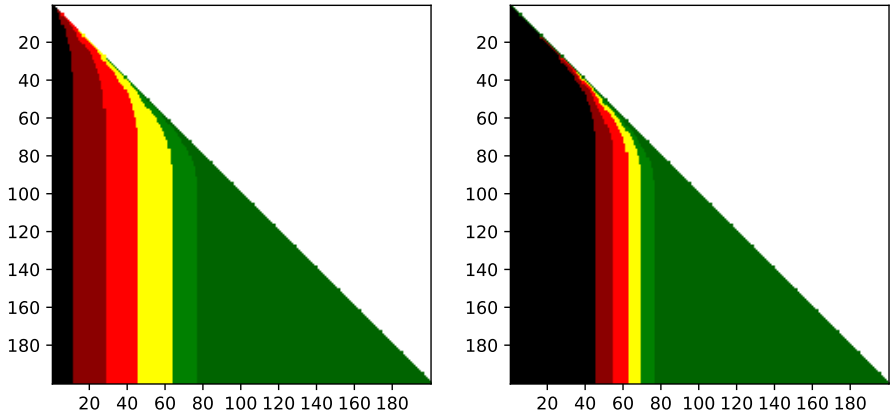


Figure 1: Left panel: the discovery matrix for the U_1 statistic (i.e., arithmetic mean) for 100 false and 100 true null hypotheses. Right panel: the U_2 analogue.

The validity of Algorithm 1 can be demonstrated by the argument in the proof of Theorem 2.1 in [9]. It is clear that, in the case of U_2, the assumption of independence of $E_1, \dots, E_n$ can be relaxed to the assumption that all covariances $\operatorname{cov}(E_i, E_j)$, $i \neq j$, are nonpositive.

The discovery matrix constructed in Algorithm 1 does not depend on the sample space $(\Omega, \mathcal{A})$, hypotheses $H_1, \dots, H_K \subseteq \mathfrak{P}(\Omega)$, or e-variables $E_1, \dots, E_K$, and in this sense is universal (in the terminology of [9, Section 5]).

3 A simulation study

In this section we run Algorithm 1 applied to U_1 and U_2. Slightly generalizing the explanation in [9, Appendix B in Working Paper 27], we can see that the U_n discovery matrix can be computed in time $O(K^{3+n})$. For $n = 1$, the time can be improved from $O(K^4)$ to $O(K^2)$ [9, Appendix B in Working Paper 27]. For $n = 2$, we can improve the time $O(K^5)$ to $O(K^3)$, as we show in Appendix A, and this is sufficient to cope not only with the case $K = 200$ that we usually use in our simulation studies but also with K of a few thousand (as used in our empirical studies, where we compute only part of the discovery matrix).

We generate the base e-values as in Section 3 of [9]: the null hypothesis is $N(0, 1)$, $K = 200$, the first 100 observations x are generated from $N(-3, 1)$, the last 100 from $N(0, 1)$, all independently, and the base e-variables are the likelihood ratios

$$E(x) := \frac{\exp(-(x+3)^2/2)}{\exp(-x^2/2)} = \exp(-3x - 9/2). \tag{3}$$

The results are shown in Figure 1 (whose left panel is identical to the left

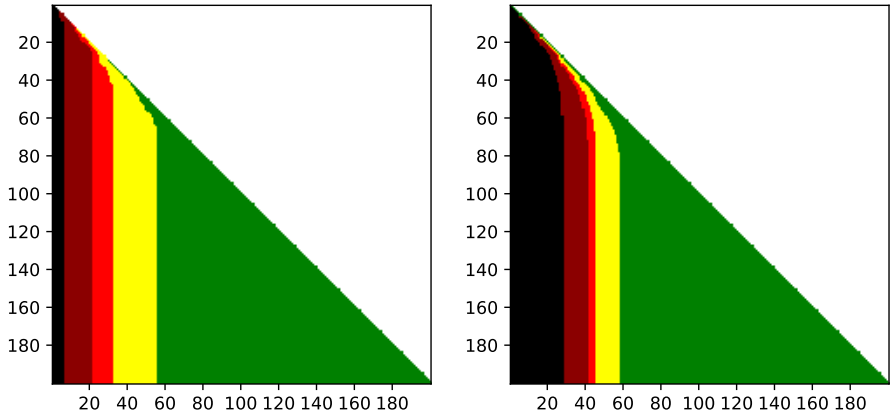


Figure 2: Left panel: the discovery p-matrix for the GWGS procedure. Right panel: the U_2 discovery matrix e-to-p calibrated.

panel of Figure 2 in [9]); they are much better for U_2. Each panel shows the lower triangular matrix $\mathrm{DM}^F_{r,j}$, the left for $F = U_1$ and the right for $F = U_2$. The colour scheme used in this figure is inspired by Jeffreys's [5, Appendix B] (as in [9]):

- The entries with $\mathrm{DM}^F_{r,j}$ below 1 are shown in dark green; there is no evidence that there are at least j true discoveries among r hypotheses with the largest e-values.
- The entries $\mathrm{DM}^F_{r,j} \in (1, \sqrt{10}) \approx (1, 3.16)$ are shown in green. For them the evidence is poor.
- The entries $\mathrm{DM}_{r,j} \in (\sqrt{10}, 10) \approx (3.16, 10)$ are shown in yellow. The evidence is substantial.
- The entries $\mathrm{DM}^F_{r,j} \in (10, 10^{3/2}) \approx (10, 31.6)$ are shown in red. The evidence is strong.
- The entries $\mathrm{DM}^F_{r,j} \in (10^{3/2}, 100) \approx (31.6, 100)$ are shown in dark red. The evidence is very strong.
- Finally, the entries $\mathrm{DM}^F_{r,j} > 100$ are shown in black, and for them the evidence is decisive.

It is interesting that after the crude e-to-p calibration $e \mapsto 1/e$ our method produces p-values that look even better than the p-values produced by the GWGS procedure (in the terminology of [9]) designed specifically for p-values: see Figure 2.

In Figure 2 we use what we called Fisher's scale in [9], but now we extend it by two further thresholds, one of which is 0.5%, as advocated by [1]. Our colour scheme is:

- P-values above 5% are shown in green; they are not significant.
- P-values between 1% and 5% are shown in yellow; they are significant but not highly significant.
- P-values between 0.5% and 1% are shown in red; they are highly significant (but fail to attain the more stringent criterion of significance advocated in [1]).
- P-values between 0.1% and 0.5% are shown in dark red.
- P-values below 0.1% are shown in black; they can be regarded as providing decisive evidence against the null hypothesis (to use Jeffreys's expression).

4 An empirical study

In this section we will use the `prostate` dataset first described in [6]; it is also analyzed by Efron in [2, Chapter 2] and then by Efron and Hastie [3, Chapter 15]. Efron makes the assumption of independence of the test statistics (while analyzing its shortfalls in [2, Section 2.5]), along with several other substantial assumptions, such as the Gaussian distribution of the genes' activities. In this paper, however, we will avoid making any other assumptions apart from independence.

The `prostate` dataset represents a 6033×102 matrix whose rows correspond to 6033 genes and columns correspond to 102 men. The first 50 men are healthy controls and the remaining 52 are patients with prostate cancer. Each entry $x_{k,j}$ represents the activity of the jth gene in the kth man. We proceed as in [9].

The base e-values are computed as

$$e_k := \frac{T_k}{\frac{1}{B+1}\left(\sum_{b=1}^{B} T_k^{(b)} + T_k\right)}, \quad k = 1, \dots, 6033, \tag{4}$$

where T_k is the nonconformity score computed (see below for details) from the kth row of the data matrix, $T_k^{(b)}$ is the nonconformity score computed from the same row with randomly permuted labels, and B is the number of permutations. (Ideally, the random permutations should be drawn without replacement, but we can, and we do, safely ignore this requirement.)

We define the nonconformity score as $T_k := |t_k|^d$, where $d > 0$ is a parameter of the algorithm and

$$t_k := \frac{\bar{x}_{k,1} - \bar{x}_{k,0}}{s_k}$$

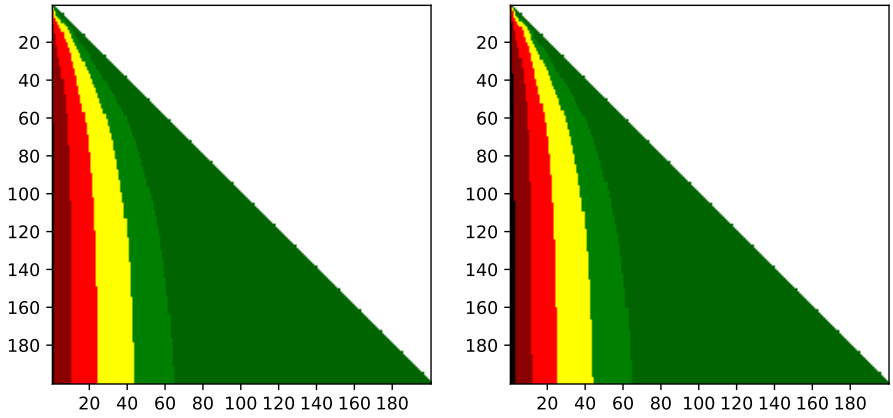


Figure 3: Left panel: the top-left 200×200 corner of the U_2 discovery matrix for the `prostate` dataset for $B := 10000$, using Jeffreys's thresholds. Right panel: its version (based on (5)) that is only approximately valid.

is the two-sample t-statistic for the kth gene, where $\bar{x}_{k,1}$ is the average entry in the kth row for patients, $\bar{x}_{k,0}$ is the average entry in the kth row for healthy controls, and

$$s_k^2 := \sum_j \left(x_{k,j} - \bar{x}_{k,y_j}\right)^2$$

is the sample variance of row k (ignoring a constant factor, which cancels out) and y_j is the label (1 for patients and 0 for healthy controls).

The left panel of Figure 3 gives the top-left 200×200 corner of the discovery matrix for $d := 10$ and $B := 10000$. The result is much better than for the U_1 discovery matrix: see the left panel of Figure 4. In the right panels of Figures 3 and 4 we give analogous plots but with the proper e-values (4) replaced by their commonly used simplified versions

$$e_k := \frac{T_k}{\frac{1}{B}\sum_{b=1}^{B} T_k^{(b)}}, \quad k = 1, \dots, 6033. \tag{5}$$

The closeness of the left and right panels suggests that the number of iterations $B = 10000$ is sufficiently large.

Finally, Figure 5 is the analogue of Figures 3 and 4 for p-values and the GWGS method, as described in [9]. In this case, there is no dependence on d as the p-values only depend on the ranks of T_k. The valid p-values used in the left panel of Figure 5 give a very poor result, and the comparison with the right-hand panel shows that the number of iterations $B = 10000$ is far too small when using p-values.

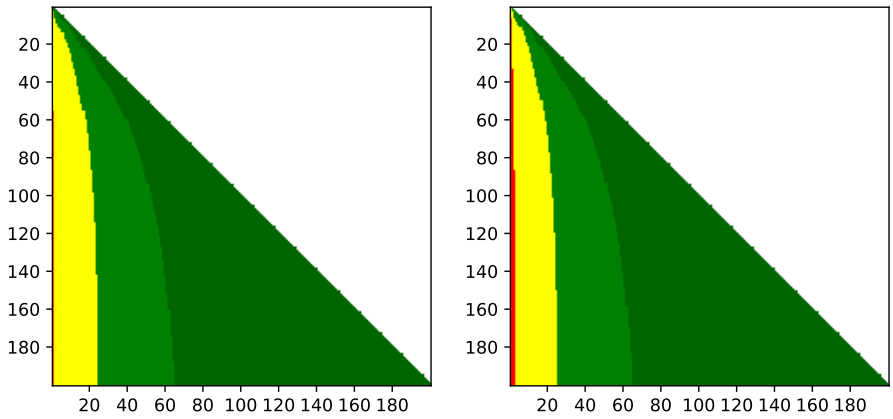


Figure 4: The analogue of Figure 3 for the U_1 discovery matrix.

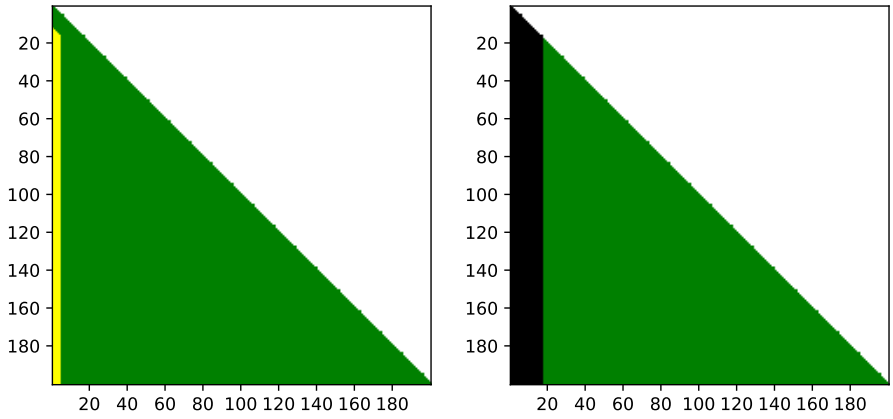


Figure 5: The analogue of Figure 3 for the discovery matrix using p-values and the GWGS method.

5 An attempt of a theoretical explanation

We start from an alternative representation of U_2, which will shed some light on the expected performance of our algorithm.

Let $\mathbf{e} := (e_1, \dots, e_K) \in [0, \infty)^K$, $M_1 = U_1(\mathbf{e})$ be the arithmetic mean of $e_1, \dots, e_K$, M_2 be the quadratic mean of $e_1, \dots, e_K$, and

$$\operatorname{var}(\mathbf{e}) := \frac{1}{K} \sum_{k=1}^{K} (e_k - M_1)^2 = M_2^2 - M_1^2$$

be the sample variance of $\mathbf{e}$.

Lemma 5.1. *For any* $\mathbf{e}$*,*

$$U_2(\mathbf{e}) = M_1^2 - \frac{1}{K-1} \operatorname{var}(\mathbf{e}). \tag{6}$$

Proof. By definition,

$$\begin{aligned} U_2(\mathbf{e}) &= \frac{1}{K(K-1)/2} \sum_{i<j} e_i e_j = \frac{1}{K(K-1)} \left(\left(\sum_i e_i \right)^2 - \sum_i e_i^2 \right) \\ &= \frac{K}{K-1} M_1^2 - \frac{1}{K-1} M_2^2 = M_1^2 - \frac{1}{K-1} \operatorname{var}(\mathbf{e}). \end{aligned}$$

□

Corollary 5.2. *For any* $\mathbf{e}$*,*

$$\operatorname{var}(\mathbf{e}) \le (K-1) M_1^2.$$

For some $\mathbf{e} \neq 0$ *the equality holds as equality.*

Proof. The first statement follows from $U_2(\mathbf{e}) \ge 0$, and an example for the second one is $\mathbf{e} := (K, 0, \dots, 0)$. □

According to Corollary 5.2,

$$\operatorname{rvar}(\mathbf{e}) := \frac{\operatorname{var}(\mathbf{e})}{(K-1) M_1^2},$$

which we will call the *relative (sample) variance* of $\mathbf{e}$, is a dimensionless quantity in the interval $[0, 1]$. When $\mathbf{e} = 0$, we set $\operatorname{rvar}(\mathbf{e}) := 0$. The relative variance is zero if and only if all e_i coincide, and it is 1 if and only if all e_i but one are zero.

Using the notion of relative variance, we can rewrite (6) as

$$U_2(\mathbf{e}) = M_1^2 (1 - \operatorname{rvar}(\mathbf{e})). \tag{7}$$

We can see the method of this paper based on U_2 has a potential for improving on the method of [9], but the best it can achieve is squaring the entries of the discovery matrix. An entry is squared if the multiset of e-values on which the

infimum in the algorithm of [9] is attained consists of a single value. Otherwise we suffer as the e-values become more diverse.

Replacing the equality sign "=" in (7) by ">" and solving the resulting inequality, we can see that U_2 improves on $U_1 = M_1$ if and only if

$$\operatorname{rvar}(\mathbf{e}) < 1 - \frac{1}{M_1}.$$

For example, if the result of applying U_1 is borderline strong evidence, $U_1 = 10$, it is improved by U_2 if and only if $\operatorname{rvar}(\mathbf{e}) < 0.9$. For comparison, the relative variance of the whole set of the likelihood ratios (3) is approximately 0.24.

6 Conclusion

This paper has given examples of ie-merging functions, namely U_n, which can be successfully applied to multiple hypothesis testing. An interesting question is whether ie-merging functions different from U_n and their convex combinations can also be useful for this purpose. Such functions definitely exist; e.g., by [8, Proposition 3.1],

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

is an admissible ie-merging function.

Acknowledgments

We are grateful to Yuri Gurevich for useful discussions. In our simulation studies we used Python and R, including the package `hommel` [4].

V. Vovk's research has been partially supported by 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

[1] Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, Eric-Jan Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael

McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman, and Valen E. Johnson. Redefine statistical significance: We propose to change the default p-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries (Comment). *Nature Human Behaviour*, 2:6–10, 2018.

[2] Bradley Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, volume 1 of *IMS Monographs*. Cambridge University Press, Cambridge, 2010.

[3] Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*, volume 5 of *IMS Monographs*. Cambridge University Press, Cambridge, 2016.

[4] Jelle J. Goeman, Rosa Meijer, and Thijmen Krebs. `hommel`*: Methods for closed testing with Simes inequality, in particular Hommel's method*, 2019. R package version 1.5, available on CRAN.

[5] Harold Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, third edition, 1961.

[6] Dinesh Singh, Phillip G. Febbo, Kenneth Ross, Donald G. Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A. Renshaw, Anthony V. D'Amico, Jerome P. Richie, Eric S. Lander, Massimo Loda, Philip W. Kantoff, Todd R. Golub, and William R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209, 2002.

[7] Vladimir Vovk. Non-algorithmic theory of randomness. Technical Report arXiv:1910.00585 [math.ST], arXiv.org e-Print archive, October 2019. For the latest version, see alrw.net/e, Working Paper 1. The conference version is to appear in: *Fields of Logic and Computation III: Essays Dedicated to Yuri Gurevich on the Occasion of His 80th Birthday*, ed. by Andreas Blass, Patrick Cégilski, Nachum Dershowitz, Manfred Droste, and Bernd Finkbeiner. Springer, 2020.

[8] Vladimir Vovk and Ruodu Wang. Combining e-values and p-values. Technical Report arXiv:1912.06116 [math.ST], arXiv.org e-Print archive, December 2019. For the latest version, see alrw.net/e, Working Paper 2.

[9] Vladimir Vovk and Ruodu Wang. True and false discoveries with e-values. Technical Report arXiv:1912.13292 [math.ST], arXiv.org e-Print archive, December 2019. For the latest version, see alrw.net/e, Working Paper 3.

Algorithm 2 One row of the discovery matrix $\mathrm{DM} = \mathrm{DM}^{U_2}$ in time $O(K^2)$

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

1: $s_0 := 0$
2: **for** $k = 1, \dots, K$ **do**
3: $\quad s_k := s_{k-1} + e_k$
4: **for** $j = r, \dots, 1$ **do**
5: $\quad V := V_2(\{K - r + 1, \dots, K - j + 1\})$
6: $\quad e := \frac{2V}{(r-j+1)(r-j)}$ (unless $j = r$)
7: $\quad$ **for** $k = 1, \dots, K - r$ **do**
8: $\quad\quad V' := V_2(\{1, \dots, k\} \cup \{K - r + 1, \dots, K - j + 1\})$
9: $\quad\quad e' := \frac{2V'}{(k+r-j+1)(k+r-j)}$
10: $\quad\quad$ **if** $e' < e$ **then** $e := e'$
11: $\quad \mathrm{DM}_{r,j} := e$

A Towards an efficient algorithm for U_2 and other U_n

In this appendix we will see that each row of the discovery matrix based on U_2 can be computed in time $O(K^2)$, and so the computation of the full discovery matrix takes time $O(K^3)$. This is not as good as the $O(K^2)$ algorithm for U_1 given as Algorithm 5 in [9], and the existence of such an algorithm for U_2 remains an open problem (it can be shown that the direct translation of Algorithm 5 in [9] to U_2 loses its validity).

Algorithm 2 starts (in lines 1–3) from defining an array s of partial sums, which not used in it explicitly, but we will explain how it enables an efficient update of the variables V and V'. In the algorithm we use the notation

$$V_2(I) := \sum_{i,i' \in I : i < i'} e_i e_{i'}, \quad \emptyset \subset I \subseteq \{1, \dots, K\};$$

this expression is the key component of the U-statistic computed from e_i, $i \in I$. When $|I| = 1$, we will treat the value of $V_2(I)$ as undefined.

The first value of V in line 5, $V_2(\{K - r + 1\})$, is undefined, and we then set

$$e := e_{K-r+1}$$

instead of the formula given in line 6. The next value of V,

$$V := V_2(\{K - r + 1, K - r + 2\} = e_{K-r+1} e_{K-r+2},$$

is computed from scratch in time $O(1)$, and the following values are computed in time $O(1)$ using the previous value:

$$V = V + e_{K-j+1} \left(s_{K-j} - s_{K-r}\right).$$

The very first value of V' computed in line 8 (for $j = r$ and $k = 1$) can be found in time $O(1)$ from scratch,

$$V' = V_2(\{1, K - r + 1\}) = e_1 e_{K-r+1}.$$

After that the first value of V' in each execution of the loop starting in line 7 (i.e., V' for $j < r$ and $k = 1$) can be found in time $O(1)$ from the current value of V using

$$V' := V + e_1 \left(s_{K-j+1} - s_{K-r}\right).$$

For the following iterations of this loop the value of V' can be updated in time $O(1)$ using

$$V' := V' + e_k s_{k-1} + e_k \left(s_{K-j+1} - s_{K-r}\right).$$

Modifications for U_n, $n > 2$

It is easy (but tiresome) to modify Algorithm 2 so that it computes a row of DM^{U^n} for a fixed $n > 2$ in time $O(K^2)$. Let us consider, for simplicity, the case $n = 3$.

Both entries of V_2 in Algorithm 2 should be changed to V_3, where

$$V_3(I) := \sum_{i,i',i'' \in I : i < i' < i''} e_i e_{i'} e_{i''},$$

line 6 should be changed to

$$e := \frac{6V}{(r - j + 1)(r - j)(r - j - 1)}$$

(unless $j = r$ or $j = r - 1$), and line 9 should be changed to

$$e' := \frac{6V'}{(k + r - j + 1)(k + r - j)(k + r - j - 1)}$$

(unless $j = r$ and $k = 1$). The lines 1–3 for computing the array

$$s_k := V_1(\{1, \ldots, k\}),$$

where

$$V_1(I) := \sum_{i \in I} e_i,$$

should be complemented by computing the array

$$t_k := V_2(\{1, \ldots, k\}), \quad k = 0, \ldots, K.$$

The array t can be computed in time $O(K)$ starting from $t_0 := 0$ and setting, for $k = 1, \ldots, K$,

$$t_k := t_{k-1} + s_{k-1} e_k.$$

In line 5, we can compute V in time $O(1)$ given its previous value using the identity (true unless j is very close to r)

$$\begin{aligned} &V_3(\{K-r+1,\dots,K-j+1\}) \\ &\qquad = V_3(\{K-r+1,\dots,K-j\}) + V_2(\{K-r+1,\dots,K-j\})e_{K-j+1} \end{aligned}$$

and the identity

$$V_2(\{a+1,\dots,b\}) = t_b - t_a - s_a(s_b - s_a). \tag{8}$$

Finally, in line 8 we can compute V' in time $O(1)$ given its previous value using the identity (true unless j is very close to r or k is very small)

$$\begin{aligned} &V_3(\{1,\dots,k\} \cup \{K-r+1,\dots,K-j+1\}) \\ &\qquad = V_3(\{1,\dots,k-1\} \cup \{K-r+1,\dots,K-j+1\}) \\ &\quad + t_{k-1}e_k + s_{k-1}e_k(s_{K-j+1} - s_{K-r}) + e_k V_2(\{K-r+1,\dots,K-j+1\}) \end{aligned}$$

and the identity (8). The simple but numerous special cases (signalled by "unless") should be considered separately.