Applications of e-values to multiple hypothesis testing (joint work with Ruodu Wang)

Vladimir Vovk

Centre for Reliable Machine Learning
Department of Computer Science
Royal Holloway, University of London

Department of Statistics
London School of Economics & Political Science
14 November 2022

My plan

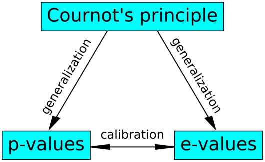
- Cournot's principle and its 2 natural developments: p-values (standard) and e-values.
- Two versions of confidence regions: based on p-values and based on e-values.
- Applying both versions to multiple hypothesis testing: controlling the number of true discoveries
  - under arbitrary dependence between the base p- or e-values,
  - under independence (or sequential dependence).

Plan

1. Testing and confidence regions
2. Controlling true discoveries (in general)
3. Controlling true discoveries (under independence)

Cournot's principle and its variants

Augustin Cournot's bridge between probability theory and the world: if a given event has a small probability, we do not expect it to happen.



Cournot's principle is the basis of the classical approach to statistics (testing statistical hypotheses and confidence regions).

Testing a probability measure Q

- The most basic way: choose a critical region A with probability $Q(A) \le \alpha$, α (the size) being a small positive number; reject Q 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. Two ways to graduate the notion of a critical region: using p-values and using e-values.
- A p-variable for testing Q is a nonnegative random variable P such that, for any $\alpha \in (0,1)$, $Q(P \le \alpha) \le \alpha$.
- An e-variable for testing Q is a nonnegative extended random variable E such that $\mathbb{E}_Q(E) \le 1$. (Example: likelihood ratio dQ'/dQ; Bayesian flavour.)

Embedding

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

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

and e-variable

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

where α is the size of the critical region A. These two random variables carry the same information as A.

An advantage of e-values

- e-Values (=values taken by e-variables) can be merged simply by averaging them ("multiple testing of a single hypothesis").
- Averaging dominates (in a natural sense) any other symmetric way of merging e-values (V. & Ruodu Wang, 2021).
- This will show in testing multiple hypotheses: procedures for controlling the numbers of false (or true) discoveries based on e-values look more efficient.

Conventional thresholds for p-values

- Observing a small p-value or a large e-value provide evidence against Q.
- For p-values, the standard thresholds are 1% and 5%, and they go back to Fisher.
- If $p \leq 0.05$, the evidence against the null hypothesis is significant.
- If $p \leq 0.01$, the evidence is highly significant.

Conventional thresholds for e-values

For e-values, this is Jeffreys's (1961 book, Appendix B) proposal (e-variables are likelihood ratios, i.e., Bayes factors for simple statistical hypotheses):

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

Jeffreys's correspondence

- “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).

Different versions

- Confidence regions were introduced by Neyman (1934) only in their basic version.
- The p-version is usually implicit, and the e-version may have been introduced only by Glenn Shafer in his 2021 RSS discussion paper.
- Suppose we only know that the true probability measure $Q \in \mathcal{Q}$ for some $\mathcal{Q} \subseteq \mathfrak{P}(\Omega)$ ($\mathcal{Q}$ is our statistical model on the sample space Ω).

Basic tests

- A basic test of size α is a family of critical regions $(A_Q \mid Q \in \mathcal{Q})$ of size α.
- A symmetric interpretation of a basic test is that $\omega \in A_Q$ means poor agreement between Q 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 Q and divide the ωs into those that conform to Q ($\omega \notin A_Q$) and those that do not ($\omega \in A_Q$); the latter are strange;
  - on the estimation side, we start from ω and divide the Qs into those that agree with ω ($\omega \notin A_Q$) and those that do not ($\omega \in A_Q$).

Parameters

- We are often interested in a parameter θ, which is a function of Q: $\theta := \Theta(Q)$ for some function Θ on $\mathcal{Q}$ (e.g., $\Theta : \mathcal{Q} \to \mathbb{R}^d$).
- Suppose we want a confidence region for θ.
- (In our applications, Θ is often chosen post hoc; Cournot's principle only requires that the test be chosen in advance.)

Basic confidence regions

- On the estimation side we have the notion of a confidence estimator as introduced by Neyman:

$$\Gamma(\omega) := \{\Theta(Q) \mid Q \in \mathcal{Q}, \omega \notin A_Q\}.$$

- Our interpretation of the confidence region $\Gamma(\omega)$ is that $\Gamma(\omega)$ covers the true $\theta = \Theta(Q)$ unless ω is strange.

p-Tests and confidence regions

- A p-test is a family of p-variables $(P_Q \mid Q \in \mathcal{Q})$, and the corresponding p-confidence regions are defined as

$$\Gamma(\omega) := \{\Theta(Q) \mid Q \in \mathcal{Q}, P_Q(\omega) > \alpha\}, \quad \alpha \in (0,1).$$

- We regard $P_Q(\omega)$ as a measure of agreement between Q and ω, with small values indicating poor agreement, and define $\Gamma(\omega)$ to be the set of $\Theta(Q)$ for Q that agree with ω at level α.

e-Tests and confidence regions

- Similarly, an e-test is a family of e-variables $(E_Q \mid Q \in \mathcal{Q})$.
- We also regard $E_Q(\omega)$ as a measure of agreement between Q and ω, but now large values indicate poor agreement.
- We define the e-confidence regions as

$$\Gamma(\omega) := \{\Theta(Q) \mid Q \in \mathcal{Q}, E_Q(\omega) < \alpha\}, \quad \alpha \in (0, \infty).$$

Plan

1. Testing and confidence regions
2. Controlling true discoveries (in general)
3. Controlling true discoveries (under independence)

Setting (for e-values, for concreteness)

- Let us specialize our setting. Now we take $\mathcal{Q} := \mathfrak{P}(\Omega)$.
- Suppose that we are given K e-variables $E_1, \dots, E_K$ for testing composite hypotheses $H_1, \dots, H_K$ (our base hypotheses); we would like to reject some of them.
- Being an e-variable for H means being an e-variable for any $Q \in H$. [This is where e-variables diverge from Bayes factors.]
- 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.

Rejection sets

- 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.
- For example, R may include hypotheses connected by a common theme (such as all relevant genes related to the gastrointestinal tract in a medical application).

True and false discoveries (1)

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

  $$I_Q := \{k \in \{1, \dots, K\} \mid Q \in H_k\}$$

  to be the set of indices of hypotheses containing Q.
- If the researcher rejects H_k, this is a discovery.
- The discovery is true if $Q \notin H_k$ and false if $Q \in H_k$, where Q is the true (unknown) probability measure governing the data generation.

True and false discoveries (2)

- For a rejection set R, the number of true discoveries is

  $$|R \setminus I_Q| = |\{k \in R \mid Q \notin H_k\}|,$$

  and the number of false discoveries is

  $$|R \cap I_Q| = |\{k \in R \mid Q \in H_k\}|.$$

- 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.

True and false discoveries (3)

- Researchers are sometimes interested in the proportion of true or false discoveries $|R \setminus I_Q| / |R|$ or $|R \cap I_Q| / |R|$, respectively.
- The researcher may be interested in other parameters θ (e.g., θ may be the weighted number of true discoveries in R: e.g., some genes can be more important than other genes). These are processed in the same way.

Merging e-values

- For e-confidence regions, we need an e-test $(E_Q)_{Q\in\mathfrak{P}(\Omega)}$.
- For each $k \in I_Q$, E_k is an e-variable for testing Q. We will obtain E_Q by merging $(E_k)_{k\in I_Q}$.
- 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.)

Symmetric merging functions

- An e-merging function is symmetric if it does not depend on the order of its arguments. An example (essentially dominating any symmetric merging function) is

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

- Let F be a symmetric e-merging function. The e-test

$$E_Q := F(E_k : k \in I_Q)$$

uniquely determines e-confidence regions.

Confidence regions for the number of true discoveries

- We will use the arithmetic-mean e-test

$$E_Q := \frac{1}{|I_Q|} \sum_{k \in I_Q} E_k.$$

- Once we have the e-test and the parameter $|R \setminus I_Q|$ (number of true discoveries), we have the e-confidence region for each significance level α, as defined earlier.
- This definition is essentially the translation of Genovese and Wasserman's (2004) and Goeman and Solari's (2011) into the language of e-values.

Optimal rejection sets

- Let us now consider a family of rejection sets R that are chosen in an optimal way. 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 the e-values are sorted in the ascending order), meaning that R_r leads to smaller (in the sense of $\subseteq$) confidence regions than any other rejection set $R \subseteq \{1, \dots, K\}$ of size r.
- In the terminology of statistical decision theory, R_r is a complete class of rejection sets.

Discovery e-matrices

- The confidence regions for R_r can be visualized as a discovery e-matrix (pictures will follow momentarily).
- It can be computed very efficiently. It takes time $O(K)$ to compute one row of the arithmetic-mean discovery e-matrix (exact under free combinations, perhaps conservative in general).

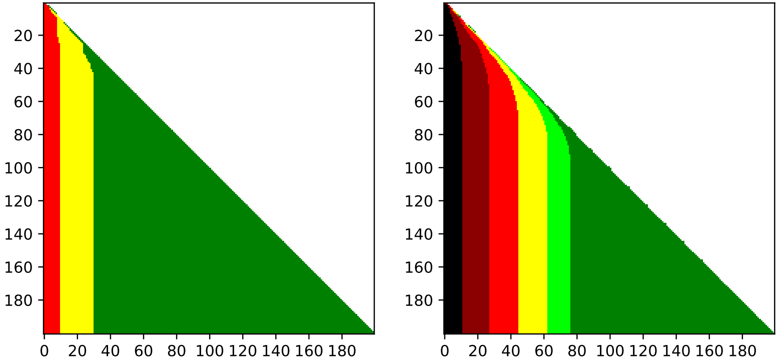
Simulation study

- Let us compute the arithmetic-mean discovery matrix for $K = 200$: we generate 100 observations from $N(-3, 1)$ and then 100 from $N(0, 1)$ (independently, but this is not known).
- The base e-values are the likelihood ratios

$$E(x) := \frac{\mathrm{d}N(-3,1)}{\mathrm{d}N(0,1)}(x)$$

  of the alternative to the null $N(0, 1)$, where $x \sim N(\mu, 1)$ is the corresponding observation.
- The base p-values are computed from E as the test statistic (Neyman–Pearson).

Discovery matrices $D_{r,j}$ (based on p-values, `hommel`, vs e-values)



Rows: r; columns: j, the number of true discoveries.

Interpretation

- The interesting colour codes are from black (decisive) to yellow (substantial) on Jeffreys's scale and red (highly significant) to yellow (significant) on Fisher's scale.
- The black colour means that those cells cannot be the numbers of true discoveries at level 100; we have decisive evidence that the number of true discoveries in covered by another colour.
- Dark red: those cells cannot be the numbers of true discoveries at level $10^{3/2}$; we have very strong evidence that the number of true discoveries is light red, yellow, or green.
- Et cetera.
- Comparison is informal, but for the e-values the picture looks better.

Hommel p-merging function and its admissible modification

- The p-merging function used in the previous picture is (Hommel, 1983)

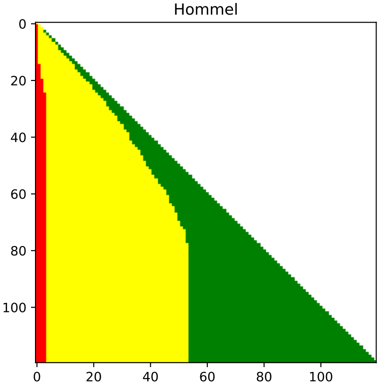
  $$(p_1, \ldots, p_K) \mapsto \ell_K \bigwedge_{k=1}^{K} \frac{K}{k} p_{(k)}$$

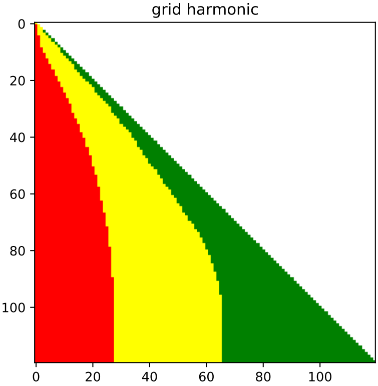
  (truncated at 1), where $\ell_K := \sum_{k=1}^{K} k^{-1}$ (not needed under independence (Simes, 1986)).
- It is not admissible (V., Wang, Wang, 2022) and dominated by the "grid harmonic p-merging function".

Another toy example

- Next slide: the upper left corners of size 120 × 120 of the discovery p-matrices for p-variables $P_1, \dots, P_{1000}$ with the first 100 observations coming from the alternative distribution $N(-4, 1)$ and the remaining 900 from the null distribution $N(0, 1)$.
- The correlation is 0.9 for all pairs of observations, except for the last one (-0.9 with the rest, to violate MTP_2).
- Improvement is not as impressive as when moving to e-values (unless high correlation), but more tangible (direct comparability).
- In fact, I will show the median over 10 simulations (to reduce noise).

Discovery p-matrix with Hommel and grid-harmonic merging





Plan

1. Testing and confidence regions
2. Controlling true discoveries (in general)
3. Controlling true discoveries (under independence)

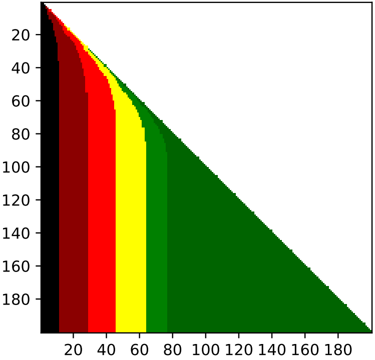
Merging e-values under independence

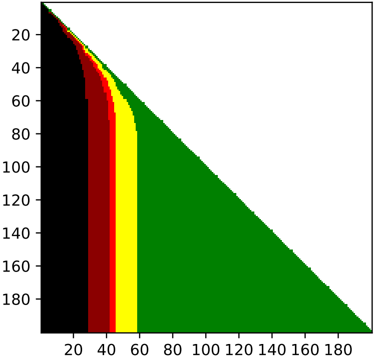
- Under independence, it's obvious that the product of e-variables is again an e-variable ($\mathbb{E}_Q(E_1 E_2) = \mathbb{E}_Q(E_1)\mathbb{E}_Q(E_2) \le 1$).
- Taking the product $e_1 \dots e_K$ is too radical! (Destroyed by a single small e-value.)
- Instead we use the U-statistic

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

  for a small n (such as 2). (Or their convex mixture.)
- This class includes product (for $n = K$), arithmetic average (for $n = 1$), and constant 1 (for $n = 0$).
- The U-statistics and their convex mixtures are admissible ie-merging functions.

Not using ($n = 1$) vs using ($n = 2$) independence for e-values

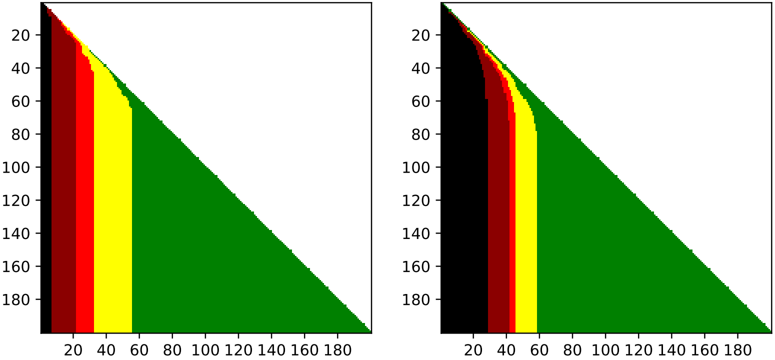




Another picture

- The setting: testing 200 hypotheses, as before.
- Now we extend Fisher's scale: yellow is significant (5%), red is highly significant (1%), dark red (0.5%), and black (0.1%).
- The e-values can be transformed into p-values ($p := 1 \vee \frac{1}{e}$ by Markov's inequality; this is the best way) and vice versa (lots of ways that are not comparable). Atrocious round-trip efficiency.
- Now the comparison will be less informal.

p-Values: Simes vs transformed U_2



References (1)

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

- Vladimir Vovk and Ruodu Wang.
  e-Values: calibration, combination, and applications.
  Annals of Statistics **49**, 1736–1754, 2021.

- Vladimir Vovk, Bin Wang, and Ruodu Wang.
  Admissible ways of merging p-values under arbitrary dependence.
  Annals of Statistics **50**, 351–375, 2022.

References (2)

- Vladimir Vovk and Ruodu Wang.
  Confidence and discoveries with e-values.
  To appear in *Statistical Science*, arXiv 2022.

- Vladimir Vovk and Ruodu Wang.
  True and false discoveries with independent e-values.
  arXiv 2020.

- Jelle J. Goeman, Rosa Meijer, and Thijmen Krebs.
  `hommel`: Methods for closed testing. . . .
  R package, available on CRAN (2019).

Thank you for your attention!