Lecture 2: Multiple hypothesis testing with e-values

Vladimir Vovk

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

Mahalanobis Memorial Lectures 2020–21
15 March, 2021

Main points of this lecture

- P-values have a promising alternative, e-values.
- E-values are useful measures of evidence in their own right, akin to Bayes factors.
- They are also a useful technical tool, even if we are only interested in p-values.

Plan

One-step test martingales

- In lecture 1, we discussed test martingales $X_0 = 1, X_1, X_2, \dots$.
- Let's see what happens if we are interested in time horizon 1: X_1 only.
- It is an e-variable: $\mathbb{E}(X_1) = 1$.
- E-variables are far from trivial objects and provide a useful alternative to p-values.

E-values vs p-values

- Slightly more general definition: an e-variable is a nonnegative extended random variable whose expected value under the null hypothesis is at most 1.
- An e-value is a value taken by an e-variable.
- A p-variable is a random variable $P : \Omega \to [0, 1]$ satisfying

$$\forall \epsilon \in (0, 1) : \mathbb{P}(P \leq \epsilon) \leq \epsilon.$$

  (The underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with expectation $\mathbb{E}$ is fixed; let us assume that $\mathbb{P}$ is atomless.)
- A p-value is a value taken by a p-variable.

Statistical interpretation of e-values

- The interpretation of e-values: observing a large e-value (for an e-variable that is chosen in advance) is evidence against the null hypothesis $\mathbb{P}$.
- This makes sense in view of Markov's inequality: $\mathbb{P}(E \geq 1/\epsilon) \leq \epsilon$ for any $\epsilon > 0$.
- More generally, let us say that 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 (i.e., f transforms e-values to p-values).

Admissible e-to-p calibrators

- An e-to-p calibrator f is said to dominate an e-to-p calibrator g if $f \leq g$, and the domination is strict if $f \neq g$.
- An e-to-p calibrator is admissible if it is not strictly dominated by any other e-to-p calibrator.

Proposition

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

Admissible p-to-e calibrators

- A calibrator is a function transforming p-values to e-values. Formally, a decreasing function $f : [0,1] \to [0,\infty]$ is a calibrator (or, more fully, p-to-e calibrator) if, for any p-variable P, $f(P)$ is an e-variable.
- A calibrator f is said to dominate a calibrator g if $f \geq g$, and the domination is strict if $f \neq g$.
- A calibrator is admissible if it is not strictly dominated by any other calibrator.

Proposition

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

Examples of calibrators

- A popular class of calibrators (used, e.g., by Sellke et al.) is

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

where $\kappa > 0$.

- Roughly, if we take $\kappa \approx 0$ and ignore constant factors (as in the algorithmic theory of randomness), $e \sim 1/p$.
- Another class of calibrators:

$$H_\kappa(p) := \begin{cases} \infty & \text{if } p = 0 \\ \kappa(1+\kappa)^\kappa p^{-1}(-\ln p)^{-1-\kappa} & \text{if } p \in (0, \exp(-1-\kappa)] \\ 0 & \text{if } p \in (\exp(-1-\kappa), 1] \end{cases}$$

where $\kappa > 0$. Advantage: as $p \to 0$, $H_\kappa(p)$ are closer than $f_\kappa(p)$ to the ideal (but impossible) $1/p$.

Connections with Bayes factors

- An attractive specific calibrator has been proposed by Glenn Shafer (2020):

$$S(p) := p^{-1/2} - 1.$$

- For simple null hypotheses, e-variables are almost indistinguishable from likelihood ratios and, therefore, Bayes factors.
- E is an e-variable if and only if $E = dQ/d\,\mathbb{P}$.
- Shafer's calibrator agrees well with the borderline values proposed for Bayes factors.

Which e-values are significant?

Jeffreys says about users of p-values:

> *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].*

For $p = 5\%$, Shafer give 3.47 instead of Jeffreys's 3.16, and for $p = 1\%$, Shafer gives 9 instead of Jeffreys's 10.

Similarly, Shafer's calibrator agrees very well with Good's rule of thumb.

Combining sequential e-values

- In science, important null hypotheses are often tested repeatedly (e.g., there can be a “gold rush” of follow-up studies after an initial discovery).
- If the results of consecutive studies are p-values, how do we combine them?
- On the other hand, e-values produced by various laboratories sequentially can be combined by multiplying them.
- At each point in time the product is the overall amount of evidence found against the null hypothesis.

Judith ter Schure and Peter Grünwald (2019).
Accumulation Bias in meta-analysis: the need to consider time in error control.
arXiv:1905.13494 [stat.ME]

Combining e-values in general

- The meta-analysis example works for e-values $E_1, E_2, \dots$ that are sequential (e.g., independent): $\mathbb{E}(E_n \mid E_1, \dots, E_{n-1}) = 1$, in which case $E_1 \dots E_n$, $n = 0, 1, \dots$, is a test martingale.
- But even if not, we can always combine them by averaging:

$$E := \frac{E_1 + \cdots + E_K}{K}$$

is again an e-value. (Or weighted averaging.)

Averaging is best

- An e-merging function is an increasing Borel function $F : [0, \infty]^K \to [0, \infty]$ such that $F(E_1, \dots, E_K)$ is an e-variable whenever $E_1, \dots, E_K$ are e-variables.
- An e-merging function F essentially dominates an e-merging function G if, for all $\mathbf{e} \in [0, \infty]^K$,

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

Proposition

The arithmetic mean essentially dominates any symmetric e-merging function.

E-variables are not Bayes factors in general

- A composite null hypothesis is a set H of probability measures on the sample space Ω.
- $E : \Omega \to [0, \infty]$ is an e-variable w.r. to H if $\int E \, \mathrm{d}Q \le 1$ for any $Q \in H$.
- In my older papers I used "Bayes factors" to mean "e-variables", which baffled Bayesian statisticians.
- Bayes factors are only required to satisfy $\int E \, \mathrm{d}P \le 1$, where $P := \int Q \, \mu(\mathrm{d}Q)$ for some prior distribution μ.

Plan

Terminology (1)

- Suppose we are given $K \geq 2$ e-values $e_1, \dots, e_K$ for testing composite hypotheses $H_1, \dots, H_K$ (our base hypotheses). We would like to reject some of them.
- 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 H_k with the largest e_k.
- But in general, we can consider an arbitrary non-empty rejection set

$$R \subseteq \{1, \dots, K\}$$

(the indices of the base hypotheses that the researcher chooses to reject).

Terminology (2)

- R may include hypotheses connected by a common theme, such as being related to the gastrointestinal tract.
- If the researcher rejects H_k, we refer to this decision as a discovery.
- If Q is the true probability measure (unknown to the researcher), then the discovery is true if $Q \notin H_k$ and false if $Q \in H_k$.

Discovery vectors (1)

- Let E_k be an e-variable for testing H_k, $k = 1, \dots, K$.
- The arithmetic-mean discovery vector is defined as

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

  (notice that $I = \emptyset$ is excluded for any j).
- The arithmetic-mean discovery vector controls the number of true discoveries (and is optimal) in a natural sense.
- If $\mathrm{AV}_R(j)$ is large, we have a Fisher-type disjunction: either there are at least j true discoveries (rejections of false null hypotheses) or a rare chance has occurred (namely, the observed e-value is at least $\mathrm{AV}_R(j)$).

Discovery vectors (2)

- In Jeffreys's terminology (1961 book, Appendix B), $\mathrm{AV}_R(j) \geq 10$ provides strong evidence for there being at least j true discoveries.
- Notice that, intuitively, controlling true discoveries and controlling false discoveries are the same thing, since the total number of discoveries $|R|$ is known.
- The procedure I am describing is an e-value counterpart of the "GWGS procedure" for p-values (Genovese, Wasserman, Goeman, Solari).

Discovery matrices

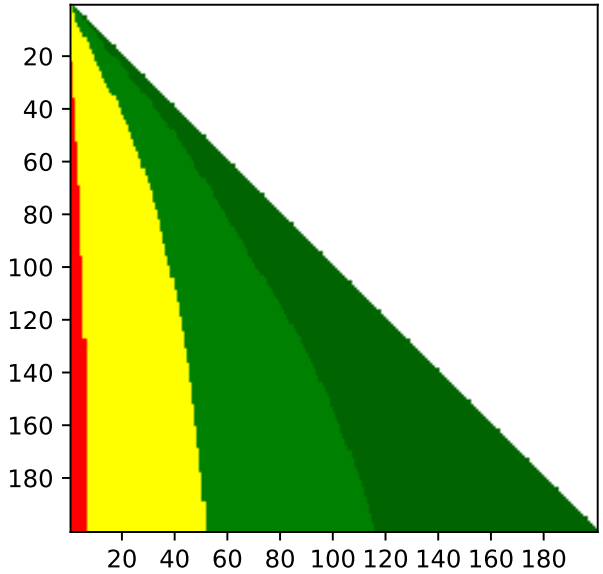
- Let R_i be the indices of the i largest e-values.
- The discovery matrix is defined as

$$\mathrm{AM}_{i,j} := \mathrm{AV}_{R_i}(j), \quad i \in \{1, \dots, K\}, \ j \in \{1, \dots, i\}.$$

- A large value of $\mathrm{AM}_{i,j}$ is evidence for the statement "there are at least j true discoveries among the i hypotheses (with the largest e-values) that we choose to reject".

Results for a medical dataset (3170 genes)

Using the arithmetic average we obtain this discovery matrix (in Jeffreys's colour code, with red meaning "strong evidence").



For a comparable figure for p-values, only one p-value in each row is significant.

Jeffreys' scale

- If $\mathrm{AM}_{i,j} < 1$, the null hypothesis is supported. Dark green.
- If $\mathrm{AM}_{i,j} \in (1, \sqrt{10}) \approx (1, 3.16)$, the evidence against the null hypothesis is not worth more than a bare mention. Green.
- If $\mathrm{AM}_{i,j} \in (\sqrt{10}, 10) \approx (3.16, 10)$, the evidence against the null hypothesis is substantial. Yellow.
- If $\mathrm{AM}_{i,j} \in (10, 10^{3/2}) \approx (10, 31.6)$, the evidence against the null hypothesis is strong. Red.

His scale also includes "very strong" ($\mathrm{AM}_{i,j} \in (10^{3/2}, 100)$) and decisive ($\mathrm{AM}_{i,j} > 100$).

Problem with data splitting

- In 1975 David Cox discovered that splitting data at random not only allows flexible testing of statistical hypotheses but also achieves high efficiency.
- A serious objection to the method is that different people analyzing the same data may get very different answers (thus violating "inferential reproducibility").
- Using e-values instead of p-values remedies the situation.

David R. Cox (1975).
A note on data-splitting for the evaluation of significance levels.
Biometrika, 62:441–444.

Cox's procedure (1)

- We are given m independent random samples of size r from normal populations with means $\mu_1, \dots, \mu_m$ and known common variance σ_0^2.
- The null hypothesis is that all means are zero, and the alternative is that just one of the means is positive, $\mu > 0$.
- We apply the method of data splitting by dividing each sample into two portions of sizes pr and $(1-p)r$.
- We then take the population for which the first-portion sample mean is largest.
- Finally we apply the standard one-sided normal test to the mean of the corresponding second portion, ignoring the second-portion samples of the other $m-1$ populations.

Cox's procedure (2)

- Cox also defines an exact procedure that tests the means collectively for significance using the largest mean as test statistic.
- The efficiency of the simple data splitting procedure is surprisingly close to that of the exact procedure.
- Using e-values instead of p-values allows us to repeat data splitting many times and average the results, thus achieving inferential reproducibility.

Vladimir Vovk (2020).
A note on data splitting with e-values: online appendix to my comment on Glenn Shafer's "Testing by betting".
arXiv:2008.11474 [stat.ME]

Plan

Definitions

- Even if we are only interested in p-values, e-values are still a powerful technical tool.
- In a natural sense, the duality theorem of optimal transport established a duality between p-values and e-values.
- Let me state corollaries of this duality for homogenous p-merging functions.
- A p-merging function of K p-values is an increasing Borel function $F : [0, \infty)^K \to [0, \infty)$ such that $F(P_1, \dots, P_K)$ is a p-variable whenever $P_1, \dots, P_K$ are p-variables.
- A p-merging function F is symmetric if it is invariant under any permutation of its arguments, and it is homogenous if $F(\lambda \mathbf{p}) = \lambda F(\mathbf{p})$ for all $\mathbf{p} \in [0, \infty)^K$ and $\lambda > 0$.

Rejection regions

- A p-merging function can be characterized by its rejection regions.
- The rejection region of a p-merging function F at level $\epsilon > 0$ is

$$R_\epsilon(F) := \left\{ \mathbf{p} \in [0,\infty)^K : F(\mathbf{p}) \le \epsilon \right\}.$$

- If F is homogenous, then $R_\epsilon(F)$, $\epsilon \in (0,1)$, takes the form $R_\epsilon(F) = \epsilon A$ for some $A \subseteq [0,\infty)^K$.

Theorem ($\approx$ duality)

Theorem

For any admissible homogenous p-merging function F, there exist $(\lambda_1, \dots, \lambda_K) \in \Delta_K$ and admissible calibrators $f_1, \dots, f_K$ such that

$$R_\epsilon(F) = \epsilon \left\{ \mathbf{p} \in [0, \infty)^K : \sum_{k=1}^K \lambda_k f_k(p_k) \geq 1 \right\}$$

for each $\epsilon \in (0, 1)$.

Conversely, for any $(\lambda_1, \dots, \lambda_K) \in \Delta_K$ and calibrators $f_1, \dots, f_K$, this equation determines a homogenous p-merging function.

Δ_K is the standard simplex.

Simplified theorem

If the homogenous p-merging function F is symmetric, then $f_1, \dots, f_K$, as well as $\lambda_1, \dots, \lambda_K$, can be chosen identical.

Theorem

For any F that is admissible within the family of homogenous symmetric p-merging functions, there exists an admissible calibrator f such that

$$R_\epsilon(F) = \epsilon \left\{ \mathbf{p} \in [0, \infty)^K : \frac{1}{K} \sum_{k=1}^{K} f(p_k) \geq 1 \right\} \quad \textit{for each } \epsilon \in (0, 1).$$

Conversely, for any calibrator f, this equation determines a homogenous symmetric p-merging function.

In this case, we say that f induces F.

More definitions

- A p-merging function F dominates a p-merging function G if $F \leq G$.
- The domination is strict if, in addition, $F \neq G$.
- A p-merging function is admissible if it is not strictly dominated by any p-merging function.
- Analogously, we can define admissibility within smaller classes of p-merging functions, such as the class of symmetric p-merging functions.

Hommel function is not admissible

- An important p-merging function is Hommel's (1983), given by

$$H_K := \ell_K \bigwedge_{k=1}^{K} \frac{K}{k} p_{(k)},$$

where

$$\ell_K := \sum_{k=1}^{K} \frac{1}{k}.$$

- Hommel's function is not admissible but can be improved to admissible using our theorem.

Making Hommel admissible

- Our admissible modification $H_K^* \le H_K$ of the Hommel function is induced by the grid harmonic calibrator

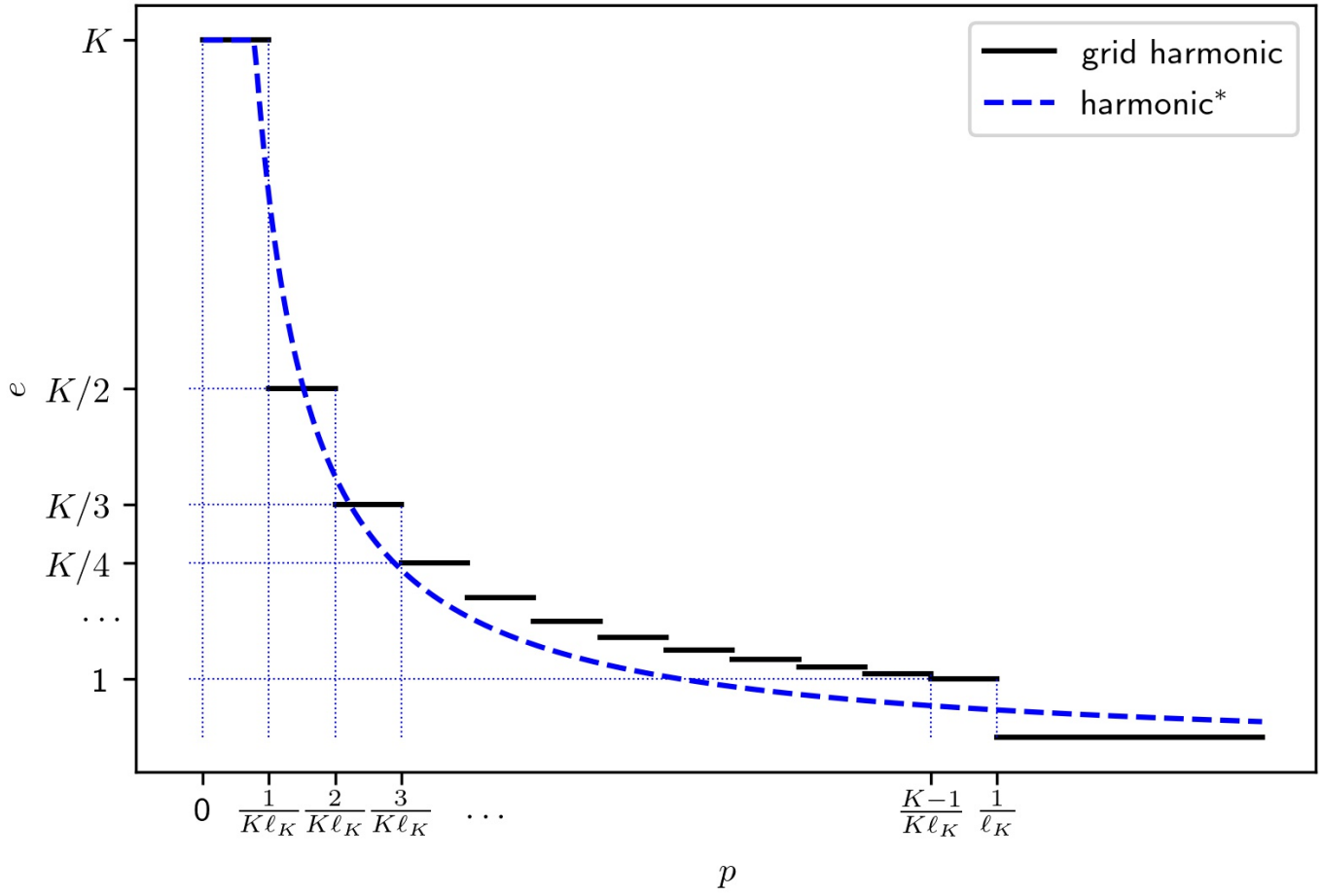
$$f : x \mapsto \frac{K 1_{\{\ell_K x \le 1\}}}{\lceil K \ell_K x \rceil},$$

- H_K^*: the grid harmonic p-merging function.

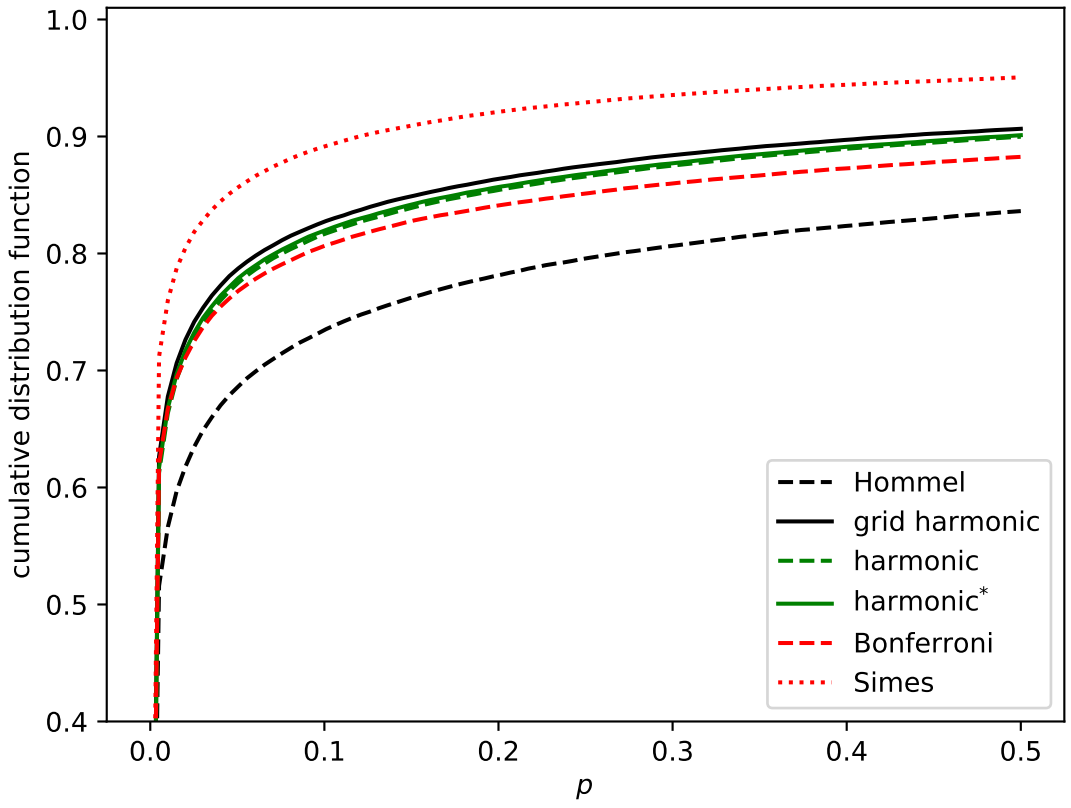
Theorem

The p-merging function H_K is dominated (strictly if $K \ge 4$) by the grid harmonic p-merging function H_K^*. Moreover, H_K^* is always admissible among symmetric p-merging functions, and it is admissible if K is not a prime number.

What the grid harmonic calibrator looks like

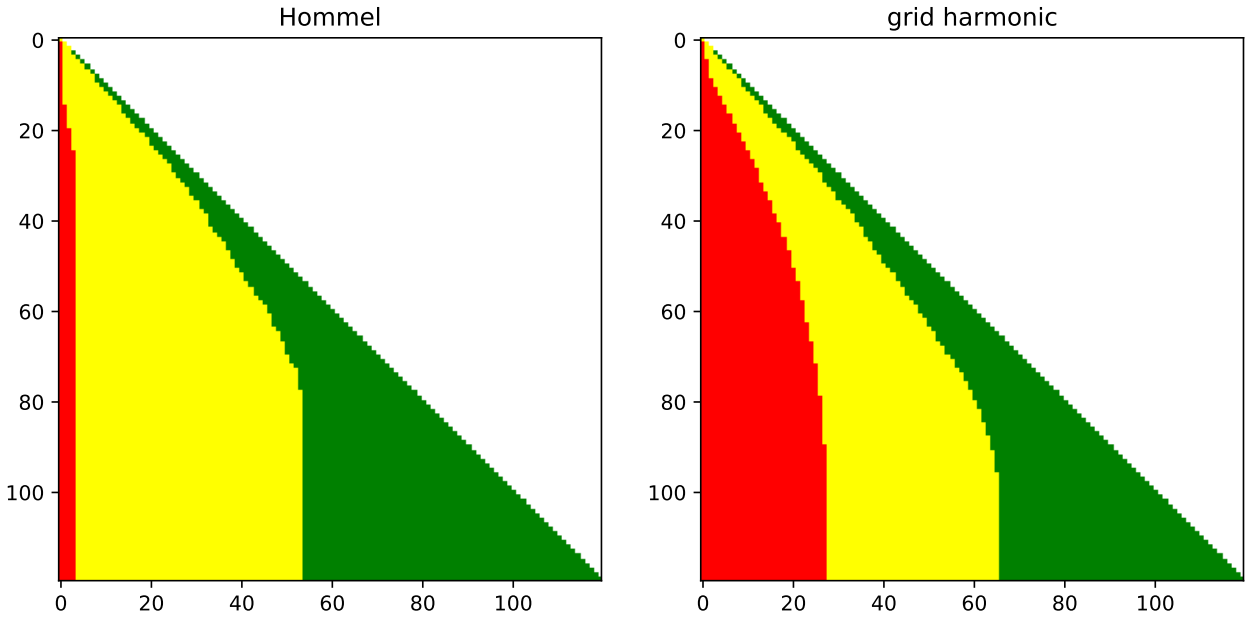


Cdf's for various p-merging functions



$K = 10^6$ observations, 10^3 of them with alternative distribution ($N(5,1)$ instead of $N(0,1)$, with correlation 0.9), z-tests

Results for the GWGS procedure



red is highly significant ($< 1\%$), yellow merely significant ($\in (1\%, 5\%)$); $K = 1000$ observations, 100 of them with alternative distribution ($N(4, 1)$ instead of $N(0, 1)$, with correlation 0.9), z-tests

Advantages of p-values (1)

Both p-values and e-values have important advantages, and I think they should complement (rather than compete with) each other.

Advantages of p-values:

- P-values can be more robust to our assumptions (perhaps implicit). For some natural classes of alternative hypotheses, the Neyman–Pearson optimal p-value will not depend on the choice of the alternative hypothesis in the class (there are numerous examples in statistics textbooks). This is not true for the likelihood ratio itself (which is the optimal e-value in a natural sense).

Advantages of p-values (2)

- There are many known efficient ways of computing p-values for testing nonparametric hypotheses that are already widely used in science.
- In many cases, we know the distribution of p-values under the null hypothesis: it is uniform on the interval $[0, 1]$. If the null hypothesis is composite, we can test it by testing the simple hypothesis of uniformity for the p-values. A recent application of this idea is the use of conformal martingales for detecting deviations from the IID model (lecture 4).

Advantages of e-values

- For many people, betting scores are more intuitive than p-values. Betting intuition has been acclaimed as the right approach to uncertainty even in popular culture (Duke, 2018).
- Betting can be opportunistic. Outcomes of experiments performed sequentially by different research groups can be combined seamlessly into a nonnegative martingale.
- Mathematically, averaging e-values still produces a valid e-value, which is far from being true for p-values. This is useful in, e.g., multiple hypothesis testing and statistical testing with data splitting.
- E-values appear naturally as a technical tool when applying the duality theorem in deriving admissible functions for combining p-values.

Bibliography

- Vladimir Vovk and Ruodu Wang (2020).
  E-values: Calibration, combination, and applications.
  Annals of Statistics, to appear.
- Vladimir Vovk and Ruodu Wang (2020).
  True and false discoveries with e-values.
  arXiv:1912.13292 [math.ST]
- Vladimir Vovk, Bin Wang, and Ruodu Wang (2020).
  Admissible ways of merging p-values under arbitrary dependence.
  arXiv:2007.14208 [math.ST]

Thank you for your attention!