A tiny review on e-values and e-processes

Ruodu Wang*

June 2023

E-values ("e" for "expectation") are an alternative to p-values ("p" for "probability"). Since 2019, e-values have been used for statistical testing by Shafer (2021), Vovk and Wang (2021), Grünwald et al. (2023) and Howard et al. (2021). These authors used various names for the concept, but the literature has converged on the terminology "e-value" proposed by Vovk and Wang (2021). Tests with e-values are usually based on martingale techniques, and the notion of "e-processes" generalizes the notion of likelihood ratios to composite hypotheses.

The use of martingales in statistical testing can be traced back to Wald (1945) and has been an important part of sequential analysis since the work by Darling and Robbins (1967), Lai (1976) and Siegmund (1978). The recent work, which is intimately connected to the game-theoretic probability and statistics of Shafer and Vovk (2001, 2019), emphasizes optional stopping or continuation of experiments. For a review on e-values and game-theoretic statistics, see Ramdas et al. (2022).

Definitions

Fix a measurable space $(\Omega, \mathcal{F})$ which is our sample space. A *hypothesis* is a collection H of probability measures on the sample space. A hypothesis is *simple* if it contains only one probability measure, and for simplicity we use the probability measure Q to represent the simple hypothesis $\{Q\}$.

An *e-variable* E for a hypothesis H (or, an e-variable testing H) is a $[0,\infty]$-valued random variable satisfying $\mathbb{E}^Q[E] \le 1$ for all $Q \in H$. In contrast, a *p-variable* for a hypothesis H is a $[0,1]$-valued random variable P satisfying $Q(P \le \alpha) \le \alpha$ for all $\alpha \in (0,1)$. An *e-process* $M = (M_t)_{t=0,1,\dots,T}$, where T can be finite or infinite, is a nonnegative stochastic process adapted to a pre-specified filtration such that $\mathbb{E}^Q[M_\tau] \le 1$ for any stopping time τ and any $Q \in H$; in other words, M_τ is an e-variable for H. This filtration is often chosen as the one generated by sequentially observed data points.

An e-variable is allowed to take the value ∞; observing $E = \infty$ for an e-variable E means that we are entitled to reject the null hypothesis; this corresponds to observing 0 for a p-variable. Realizations of p-variables and e-variables

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

are referred to as p-values and e-values. Many authors use "e/p-variables" and "e/p-values" interchangeably when the distinction is not essential for their study. A large observed e-value suggests evidence against the null hypothesis similarly to a small observed p-values.

Basic examples and properties

For the simplest example, suppose that we are testing a simple hypothesis Q_0 versus a simple hypothesis Q_1, where Q_1 is absolutely continuous with respect to Q_0. For this setting, a natural e-variable is the likelihood ratio $E = \mathrm{d}Q_1/\mathrm{d}Q_0(X)$ where X is the observed data. It is straightforward to verify that $E \geq 0$ and it satisfies $\mathbb{E}^{Q_0}[E] = 1$. If we observe iid data $X_1, X_2, \ldots$ sequentially, then the likelihood ratio process M given by

$$M_0 = 1 \quad \text{and} \quad M_t = \prod_{k=1}^{t} \frac{\mathrm{d}Q_1}{\mathrm{d}Q_0}(X_k) \text{ for } t = 1, 2, \ldots$$

is an e-process adapted to the filtration generated by the data. Moreover, we can easily see that M is a martingale. Indeed, when testing simple hypotheses, it is optimal in a natural sense to use a martingale to construct e-processes. For composite hypotheses, the situation is much more complicated, as non-trivial composite martingales may not exist while non-trivial e-processes may exist (Ramdas et al. (2020)).

Let E be an e-variable for H. An important property of e-variables is the inequality $Q(E \geq 1/\alpha) \leq \alpha$ for any $\alpha \in (0,1)$ and $Q \in H$, due to Markov's inequality. Moreover, for any non-negative supermartingale M under Q with $M(0) = 1$, Ville (1939)'s inequality gives

$$Q\left(\sup_{t=0,1,\ldots,T} M_t \geq \frac{1}{\alpha}\right) \leq \alpha, \quad \alpha \in (0,1);$$

here T may be finite or infinite. Moreover, any e-process for H is dominated by a class of supermartingales M^Q with initial value 1 for $Q \in H$, all with respect to the same filtration (Ramdas et al. (2020)). This insight implies that tests formulated by rejecting the null hypothesis if an e-process goes above $1/\alpha$ are *anytime-valid*; that is, its type-I error is controlled at α regardless of the stopping rule.

Calibration

P-values and e-values can be converted between each other. A *p-to-e calibrator* (we often omit "p-to-e") is a decreasing (in the non-strict sense) function $f : [0,1] \to [0,\infty]$ such that $f(P)$ is an e-variable any p-variable P testing the same hypothesis. An *e-to-p calibrator* is a decreasing function $g : [0,\infty] \to [0,1]$ such that $g(E)$ is a p-variable for any e-variable E testing the same hypothesis. A calibrator f is said to *dominate* a calibrator g if $f \geq g$ (p-to-e) or $f \leq g$ (e-to-p), and the domination is *strict* if $f \neq g$. A calibrator is *admissible* if it is not strictly dominated by any other calibrator.

Calibrators, under various names, are studied by Shafer et al. (2011), Shafer (2021) and Vovk and Wang (2021). A decreasing function $f : [0,1] \to [0,\infty]$ is an admissible p-to-e calibrator if and only if f is upper semicontinuous, $f(0) = \infty$, and $\int_0^1 f = 1$. Simple examples of p-to-e calibrators are $f(p) = \kappa p^{\kappa-1}$ for some $\kappa \in (0,1)$ and $f(p) = p^{-1/2} - 1$ (Shafer's). On the other hand, the only admissible e-to-p calibrator is given by $f : [0,\infty] \to [0,1]$, $f(e) = \min(1/e, 1)$; this is again due to Markov's inequality. Hence, for any e-variable E, $1/E$ truncated at 1 is a p-variable. If further E has a decreasing density on $(0,\infty)$, then $1/(2E)$ is a p-variable (Wang (2023)). Converting a p-value to an e-value using a p-to-e calibrator and then back to p-value using an e-to-p calibrator generally loses quite a lot of evidence. For instance starting with $p = 0.01$, a conversion with the p-to-e calibrator $p \mapsto p^{-1/2} - 1$ gives $e = 9$, and another conversion with the e-to-p calibrator $e \mapsto \min(1/e, 1)$ yields $p' = 1/9$.

A compromise between the above two calibration directions is the Vovk-Sellke (VS) bound $f(p) = \max(-(\exp(1)p \log p)^{-1}, 1)$. This function is not a p-to-e calibrator, but it is the supremum of the class of the p-to-e calibrators $f(p) = \kappa p^{\kappa-1}$ for $\kappa \in (0,1)$.

Recommended thresholds

In testing scientific hypotheses, thresholds for p-values are often chosen as 0.01 or 0.05 which correspond to type-I errors controlled at these levels. The e-to-p calibrator $e \mapsto \min(1/e, 1)$ implies that thresholds of 100 and 20 for e-values also have the above type-I error control. However, it is not recommended in practice to directly use these thresholds as the conversion $e \mapsto \min(1/e, 1)$ is typically wasteful.

In order to judge how significant results of testing using e-values are, the type-I error, based on which p-values are defined, may not be the desirable metric. Although there is no universally agreed thresholds to use for e-values, the rule of thumb of (Jeffreys, 1961, Appendix B), originally designed for likelihood ratios, may be useful as e-values are generalizations of likelihood ratios. We summarize this rule of thumb in Table 1 below. Our rough recommendation in line with Jeffreys (1961) is to use $e > 4$ in place of $p < 0.05$ and $e > 10$ in place of $p < 0.01$, but one should keep in mind that these choices are quite arbitrary since p-values and e-values are not one-to-one corresponding to each other. Kelter (2021, Table 1) summarizes some variations of Jeffreys's rule.

Growth optimality and e-power

The most simple and well-accepted criterion to quantify the power of an e-variable E is through its growth rate under an alternative probability measure Q_1, defined as $\mathbb{E}^{Q_1}[\log E]$. This ideas goes back to Kelly (1956), and it is studied by Shafer (2021), Grünwald et al. (2023) and Waudby-Smith and Ramdas (2023)

e-value	evidence	Shafer's p-value
$0 \leq e < 1$	null hypothesis is supported	$0.25 < p \leq 1$
$1 < e < 3.16$	no more than a bare mention	$0.0577 < p < 0.25$
$3.16 < e < 10$	substantial	$8.3\times10^{-3} < p < 0.0577$
$10 < e < 31.6$	strong	$9.4\times10^{-4} < p < 8.3\times10^{-3}$
$31.6 < e < 100$	very strong	$9.8\times10^{-5} < p < 9.4\times10^{-4}$
$100 < e$	decisive	$0 \leq p < 9.8\times10^{-5}$

Table 1: Applying Jeffreys (1961)'s rule of thumb for likelihood ratios to e-values. For comparison, we also reported Shafer's p-value, which corresponds to the range of p via $e = p^{-1/2} - 1$. The boundary values can be put in either of the two adjacent categories.

in detail. The quantity $\mathbb{E}^{Q_1}[\log E]$ is called the *e-power* of E by Vovk and Wang (2022).

The intuition behind e-power is built on the fact that e-variables for sequential data are often multiplicative. That is, very often one relies on the e-process M given by $M_t = \prod_{k=1}^{t} E_k$ where $E_1, E_2, \dots$ are *sequential e-variables*, meaning that $\mathbb{E}[E_k \mid E_{k-1}, \dots, E_1] \leq 1$ for each k. If $E_1, E_2, \dots$ are iid, then the asymptotic growth rate of the e-process M, $\lim_{t\to\infty}(\log M_t)/t$, is the e-power of E_1 by the Law of Large Numbers.

For the test of a simple null Q_0 versus a simple alternative Q_1 which absolutely continuous with respect to Q_0, the growth rate is maximized by the likelihodo ratio $E = \mathrm{d}Q_1/\,\mathrm{d}Q_0$. Grünwald et al. (2023) developed a theory on finding the optimal e-variable maximizing the e-power for a given composite null hypothesis and a composite alternative hypothesis.

Multiple testing with e-values

One advantage of e-values is that they can be combined in a straightforward manner; this is different from the situation of p-values where many complicated methods exist (e.g., Vovk and Wang (2020)).

Let E_k be an e-variable for a hypothesis H_k, for $k \in [K] := \{1, \dots, K\}$. Let $H = \bigcap_{k\in[K]} H_k$ which represents the global null (it does not hurt to think about the situation where $H_1 = \cdots = H_K$). The arithmetic average $(\sum_{k=1}^{K} E_k)/K$ is an e-variable for H, regardless of how $E_1, \dots, E_K$ are dependent. If we know that $E_1, \dots, E_K$ are independent, then the product $\prod_{k=1}^{K} E_k$ is also an e-variable for H. These two choices are admissible ways of merging e-variables, and each of them is optimal in a different sense (Vovk and Wang (2021)).

In the context of testing multiple hypotheses, a popular metric is the false discovery rate (FDR), which is the expected proportion of false rejections among all rejections. The celebrated BH procedure of Benjamini and Hochberg (1995) controls the FDR for p-values which are independent or positively dependent in the sense of Benjamini and Yekutieli (2001). Wang and Ramdas (2022) developed the so called e-BH procedure which uses e-values instead of p-values. They showed that the e-BH procedure controls FDR under arbitrary dependence structures.

Constructing e-processes based on betting strategies

There is a standard way of constructing e-processes from data based on betting strategies, studied by Shafer and Vovk (2019) and Shafer (2021). An e-process $(M_t)_{t\in\mathbb{T}}$, where $\mathbb{T} = \{1,\ldots,T\}$ with T possibly finite or infinite, is often constructed by combining several sequential e-variables $E = (E_t)_{t\in\mathbb{T}}$ from the data via a method of martingale:

$$M_t = \prod_{s=1}^{t} \left(1 - \lambda_s(E_s - 1)\right),\ t \in \mathbb{T},$$

where $\lambda = (\lambda_t)_{t\in\mathbb{T}}$ is called a betting strategy. A betting strategy λ takes values in $[0,1]^{\mathbb{T}}$ and is predictable, in the sense that λ_t is determined by $X_1,\ldots,X_{t-1}$ for each t, where $X_1,\ldots,X_t$ are the data points available to time t (from which E_t is computed). One can easily verify that M defined in this way is a valid e-process for any choice of the betting strategy λ. Nevertheless, optimally choosing a betting strategy λ that maximizes the e-power is nontrivial since we typically do not know the true data-generating probability. Some methods, such as those computing λ from the empirical distribution of the data, are studied by Waudby-Smith and Ramdas (2023) and Wang et al. (2022).

E-confidence regions

Like p-values, e-values can be used to construct e-confidence regions. These confidence regions are historically studied as confidence sequences (Robbins (1970)); see Howard et al. (2021) for a more recent study. Suppose that $\theta \in \Theta$ is a parameter of interest, which corresponds to a probability measure Q_θ. The usual confidence region at level α formulated via a class of p-variables P_θ testing Q_θ for $\theta \in \Theta$, is defined by

$$\{\theta \in \Theta \mid P_\theta > \alpha\}, \quad \alpha \in (0,1).$$

Analogously, an *e-confidence region* at level α is defined by

$$\{\theta \in \Theta \mid E_\theta < 1/\alpha\}, \quad \alpha \in (0,\infty),$$

where E_θ is an e-variable testing Q_θ for each $\theta \in \Theta$ (Shafer (2021) and Vovk and Wang (2023)). Since $1/E_\theta$ is a p-variable for Q_θ, an e-confidence region is a confidence region in the classic sense, but it offers something stronger. For instance, when the procedure of Benjamini and Yekutieli (2005) is applied to e-confidence regions, the false coverage rate can be controlled under arbitrary dependence (Xu et al. (2022)); this is not the case for the classic confidence regions based on p-values.

Remark 1. The e-values in this article should not be confused with other concepts bearing the name of "e-value". For instance, the term "e-value" of VanderWeele and Ding (2017) is a different object which measures causality.

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 Series B*, **57**(1), 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**(4), 1165–1188.

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, **100**(469), 71–81.

Darling, D. A. and Robbins, H. (1967). Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences USA*, **58**, 66–68.

Grünwald, P., de Heide, R. and Koolen, W. M. (2023). Safe testing. *Journal of the Royal Statistical Society, Series B*, forthcoming.

Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2020). Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, **17**, 257–317.

Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *Annals of Statistics*, **49**(2), 1055–1080.

Kelter, R. (2021). Bayesian and frequentist testing for differences between two groups with parametric and nonparametric two-sample tests. *Wiley Interdisciplinary Reviews: Computational Statistics*, **13**(6), e1523.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, Oxford, third edition.

Kelly, J. L. (1956). A new interpretation of information rate. *Bell System Technical Journal*, **35**(4), 917–926.

Lai, T. L. (1976). On confidence sequences. *Annals of Statistics*, **4**, 265–280.

Ramdas, A., Grünwald, P., Vovk, V. and Shafer, G. (2022). Game-theoretic statistics and safe anytime-valid inference. *arXiv*: 2210.01948.

Ramdas, A., Ruf, J., Larsson, M. and Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv*: 2009.03167.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *Annals of Mathematical Statistics*, **41**(5), 1397–1409.

Shafer, G. (2021). The language of betting as a strategy for statistical and scientific communication. *Journal of the Royal Statistical Society, Series A*, **184**(2), 407–431.

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

Shafer, G. and Vovk, V. (2001). *Probability and Finance: It's Only a Game*. Wiley, New York, 2001.

Shafer, G. and Vovk, V. (2019). *Game-Theoretic Foundations for Probability and Finance*. Wiley, New York, 2019.

Siegmund, D. (1978). Estimation following sequential tests. *Biometrika*, **65**, 341–349.

VanderWeele, T. J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine*, **167**(4), 268–274.

Ville, J. (1939). Étude critique de la notion de collectif. *Thèses de l'entre-deux-guerres*, 218.

Vovk, V. and Wang, R. (2020). Combining p-values via averaging. *Biometrika*, **107**(4), 791–808.

Vovk, V. and Wang, R. (2021). E-values: Calibration, combination, and applications. *Annals of Statistics*, **49**(3), 1736–1754.

Vovk, V. and Wang, R. (2022). Efficiency of nonparametric e-tests. *arXiv*: 2208.08925.

Vovk, V. and Wang, R. (2023). Confidence and discoveries with e-values. *Statistical Science*, **38**(2), 329–354.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, **16**(2), 117–186.

Wang, Q., Wang, R. and Ziegel, J. (2022). E-backtesting. *arXiv*: 2209.00991.

Wang, R. (2023). Testing with p*-values: Between p-values, mid p-values, and e-values. *Bernoulli*, forthcoming.

Wang, R. and Ramdas, A. (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B*, **84**(3), 822–852.

Waudby-Smith, I. and Ramdas, A. (2023). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B*, forthcoming.

Xu, Z., Wang, R. and Ramdas, A. (2022). Post-selection inference for e-value based confidence intervals. *arXiv*: 2203.12572.