Comment on Glenn Shafer’s “Testing by betting” [15]

Vladimir Vovk

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 #8


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

This note is my comment on Glenn Shafer’s discussion paper “Testing by betting” [15], together with an online appendix comparing p-values and betting scores.

Contents

Main comment 1
References 1
Appendix A Cournot’s principle, p-values, and e-values 3
Main comment

Glenn Shafer’s paper is a powerful appeal for a wider use of betting ideas and intuitions in statistics. He admits that p-values will never be completely replaced by betting scores, and I discuss it further in Appendix A (one of the two online appendices that I have prepared to meet the word limit). Both p-values and betting scores generalize Cournot’s principle [13], but they do it in their different ways, and both ways are interesting and valuable.

Other authors have referred to betting scores as Bayes factors [16] and e-values [23, 7]. For simple null hypotheses, betting scores and Bayes factors indeed essentially coincide [7, Section 1, interpretation 3], but for composite null hypotheses they are different notions, and using “Bayes factor” to mean “betting score” is utterly confusing to Bayesians [11]. However, the Bayesian connection still allows us to apply Jeffreys’s [9, Appendix B] rule of thumb to betting scores; namely, a p-value of 5% is roughly equivalent to a betting score of $10^{1/2}$, and a p-value of 1% to a betting score of 10. This agrees beautifully with Shafer’s rule (6), which gives, to two decimal places:

- for $p = 5\%$, 3.47 instead of Jeffreys’s 3.16 (slight overshoot);
- for $p = 1\%$, 9 instead of Jeffreys’s 10 (slight undershoot).

The term “e-values” emphasizes the fundamental role of expectation in the definition of betting scores (somewhat similar to the role of probability in the definition of p-values). It appears that the natural habitat for “betting scores” is game-theoretic while for “e-values” it is measure-theoretic [14]; therefore, I will say “e-values” in the online appendices (Appendix A and [19]), which are based on measure-theoretic probability.

In the second online appendix [19] I give a new example showing that betting scores are not just about communication; they may allow us to solve real statistical and scientific problems (more examples will be given by my co-author Ruodu Wang). David Cox [4] 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” [6, 8]). Using e-values instead of p-values remedies the situation.

Acknowledgments

Thanks to Ruodu Wang for useful discussions and for sharing with me his much more extensive list of advantages of e-values. This research has been partially supported by Amazon, Astra Zeneca, and Stena Line.

References


Appendix A  Cournot’s principle, p-values, and e-values

This is an online appendix to the main comment. It is based, to a large degree, on Glenn Shafer’s ideas about the philosophy of statistics. After a brief discussion of p-values and e-values as different extensions of Cournot’s principle, I list some of their advantages and disadvantages.

A.1 Three ways of testing

Both p-values and e-values are developments of Cournot’s principle [13], which is referred to simply as the standard way of testing in Shafer’s [15, Section 2.1]. If a given event has a small probability, we do not expect it to happen; this is Cournot’s bridge between probability theory and the world. (This bridge was discussed already by James Bernoulli [2]; Cournot’s [3] contribution was to say that this is the only bridge.) See Figure 1.

Cournot’s principle requires an a priori choice of a rejection region $E$. Its disadvantage is that it is binary: either the null hypothesis is completely rejected or we find no evidence whatsoever against it. A $p$-variable is a nonnegative random variable $p$ such that, for any $\alpha \in (0, 1)$, $P(p \leq \alpha) \leq \alpha$; one way to define $p$-variables is via Shafer’s (3). An $e$-variable is a nonnegative random
variable $e$ such that $E_P(e) \leq 1$; one way to define $e$-variables is via Shafer’s first displayed equation in Section 2. In p-testing, we choose a p-variable $p$ in advance and reject the null hypothesis $P$ when the observed value of $p$ (the p-value) is small, and in e-testing, we choose an e-variable $e$ in advance and reject the null hypothesis $P$ when the observed value of $e$ (the e-value) is large. In both cases, binary testing becomes graduated: now we have a measure of the amount of evidence found against the null hypothesis.

We can embed Cournot’s principle into both p-testing,

$$p(y) := \begin{cases} \alpha & \text{if } y \in E \\ 1 & \text{if not} \end{cases}$$

and e-testing (as Shafer [15, Section 2.1, (1)] explains),

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

where $\alpha := P(E)$.

There are numerous ways to transform p-values to e-values (to calibrate them) and essentially one way ($e \mapsto 1/e$) to transform e-values to p-values, as discussed in detail in [22]. The idea of calibrating p-values originated in Bayesian statistics ([1, Section 4.2], [18, Section 9], [12]), and there is a wide range of admissible calibrators. Transforming e-values into p-values is referred to as e-to-p calibration in [22], where $e \mapsto 1/e$ is shown to dominate any e-to-p calibrator [22, Proposition 2.2].

Moving between the p-domain and e-domain is, however, very inefficient. Borrowing the idea of “round-trip efficiency” from energy storage, let us start from the highly statistically significant ($\leq 1\%$) p-value 0.05%, transform it to an e-value using Shafer’s [15, (6)] calibrator

$$S(0.005) = \frac{1}{\sqrt{0.005}} \approx 13.14,$$
and then transform it back to a p-value using the only admissible e-to-p calibrator: $1/13.14 \approx 0.076$. The resulting p-value of 7.6% is not even statistically significant ($>5\%$).

A.2 Some comparisons

Both p-values and e-values have important advantages, and I think they should complement (rather than compete with) each other. Let me list a few advantages of each that come first to mind. Advantages of p-values:

- P-values can be more robust to our assumptions (perhaps implicit). Suppose, for example, that our null hypothesis is simple. When we have a clear alternative hypothesis (always assumed simple) in mind, the likelihood ratio has a natural property of optimality as e-variable (Shafer [15, Section 2.2]), and the p-variable corresponding to the likelihood ratio as test statistic is also optimal (Neyman–Pearson lemma [10, Section 3.2, Theorem 1]). For some natural classes of alternative hypotheses, the resulting p-value will not depend on the choice of the alternative hypothesis in the class (see, e.g., [10, Chapter 3] for numerous examples; a simple example can be found in [19, Section 4]). This is not true for e-values.

- 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 [20].

Advantages of e-values (starting from advantages mentioned by Shafer [15, Section 1]):

- As Shafer [15] powerfully argues, betting scores are more intuitive than p-values. Betting intuition has been acclaimed as the right approach to uncertainty even in popular culture [5].

- Betting can be opportunistic, in Shafer’s words [15, Sections 1 and 2.2]. Outcomes of experiments performed sequentially by different research groups can be combined seamlessly into a nonnegative martingale [17] (see also [7, Section 1]).

- Mathematically, averaging e-values still produces a valid e-value, which is far from being true for p-values [24]. This is useful in, e.g., multiple hypothesis testing [22] and statistical testing with data splitting [19].

- E-values appear naturally as a technical tool when applying the duality theorem in deriving admissible functions for combining p-values [21].