e-values

Perspectives from 'Testing by Betting'

Marius Puke*

CSH-Forum
University of Hohenheim
July 06, 2023

*Institute of Economics & Computional Science Hub, University of Hohenheim. marius.puke@uni-hohenheim.de

Motivation

Shafer (2021, *JRSS*)

- Problem: *p*-values are to complicated
  - applied scientists are apt to answer questions about the meaning of *p*-values incorrectly

    McShane and Gal (2017, *JASA*)
  - **Common p-value alternatives such as** ...

    ... confidence intervals instead of tests ...

    ... practical instead of statistical significance ...

    ... Bayesian interpretation of one-sided p-values ...

    ... Bayes factors ...

    ... **will not work for communicating statistical evidence!**

    Gelman and Carlin (2017, *JASA*)

Frame statistical tests and conclusions based on bets!

e-values: intuition

Shafer (2021, *JRSS*)

Hypothesis: $\mathbb{P}$ describes random variable Y.

Question: How do we use $Y = y$ to test $\mathbb{P}$?

How do we measure the strength of evidence against $\mathbb{P}$?

P-values

- Use a test statistic and chose $\alpha \in (0, 1)$.
- P-value is the smallest α for which the test rejects.
- The smaller the p-value, the more evidence against $\mathbb{P}$.

E-values

- Construct a bet on Y that can pay many different amounts.
- Such a bet is a function $E(Y)$.
- Chose E so that $\mathbb{E}_{\mathbb{P}}(E) \leq 1$.
- Pay \$1 and get back \$$E(Y)$.
- The larger \$$E(Y)$ the more evidence against $\mathbb{P}$.

What are *e*-values?

Vovk and Wang (2021, *AoS*)

An e-variable is a random variable $E \geq 0$ under hypothesis $\mathbb{P}$ if

$$\mathbb{E}_{\mathbb{P}}(E) \leq 1.$$

An e-variable, E, realizes in an e-value which ...

... is just defined under the expected value.

... has a betting interpretation i.e. the factor by which we multiply the wager.

... implies an alternative hypothesis, $\mathbb{Q}$, under which $\mathbb{E}_{\mathbb{Q}}(E) > 1$.

... yields a p-variable $\frac{1}{E}$ by Markov's ineq. as under the null $\Pr(\frac{1}{E} \leq \alpha) \leq \alpha$.

... allows for combinations and sequential testing.

★ Instead classical tests require asymptotic distributions of the test statistics.

Simple Example

Shafer (2021, *JRSS*)

- We observe $y_1 = 30, y_2 = 0.5, y_3 = -3$, and aim to perform the simple test

$$\mathcal{H}_0 : Y \sim \underbrace{N(0, 10)}_{\mathbb{P}} \qquad \text{vs} \qquad \mathcal{H}_1 : Y \sim \underbrace{N(1, 10)}_{\mathbb{Q}}.$$

- In that simple test a valid e-variable is given by the likelihood ratio

$$E(y_i) = \frac{f_{\mathbb{Q}}(y_i)}{f_{\mathbb{P}}(y_i)} = \exp\left(\frac{2y_i - 1}{200}\right).$$

- The e-values are $E(y_1) = 1.34, \quad E(y_2) = 1, \quad E(y_3) = 0.96.$
- Advantage of *e*-values:
  - combinations: $E_y = 1.34 \cdot 1 \cdot 0.96 = 1.3$ is an e-value (so is the mean)
  - valid under optional stopping

- Wald test pendants can easily be constructed based on R package `safestats`:

$\mathcal{H}_0 : X \sim N(0, \sigma^2) \quad \text{vs} \quad \mathcal{H}_1 : X \sim N(\mu, \sigma^2)$ for some $\mu \neq 0$ and unknown σ^2.

In case you are interested . . .

- **Readings** if you want to familiarize yourself with the topic:
  - Shafer and Vovk (2019): insightful book on basics of testing by betting
  - Shafer (2021, *JRSS*): seminal work on how to bet against the null
  - Grünwald et al. (2020, *JRSS* forthcoming), Vovk and Wang (2021, *AoS*)
  - Ramdas et al. (2023, *arXiv*): summary on recent developments
- See also workshop on SAVI and Game-theoretic Statistics. ▸ Workshop Material
- **Videos**:
  - There are some nice talks of Glen Shafer on testing by betting
    ▸ Video 1 ▸ Video 2 ▸ Video 3
  - Aaditya Ramdas
    ▸ Video 1 ▸ Video 2
  - Johanna Ziegel
    ▸ Video
- Short lecture series by Glenn Shafer. ▸ Lectrue Material

References

Gelman, A. and Carlin, J. (2017). Some natural solutions to the p-value communication problem—and why they won't work. *Journal of the American Statistical Association*, 112(519):899–901.

Grünwald, P., de Heide, R., and Koolen, W. (2020). Safe testing. *Preprint*. arXiv: 1906.07801.

Henzi, A., Puke, M., Dimitriadis, T., and Ziegel, J. (2023). A safe hosmer-lemeshow test. *Preprint*. arXiv: 2203.00426.

Henzi, A. and Ziegel, J. F. (2022). Valid sequential inference on probability forecast performance. *Biometrika*, 109(3):647–663.

McShane, B. B. and Gal, D. (2017). Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association*, 112(519):885–895.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. arXiv: 2210.01948.

Shafer, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184:407–431.

Shafer, G. and Vovk, V. (2019). *Game-Theoretic Foundations for Probability and Finance*. John Wiley & Sons, Ltd.

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

Applications: forecast evaluation

- Many experts repeatedly make probability predictions, P, about an binary event Y (credit default, rain, low birth weight, mortality, ...).

  Q1 How to assess the reliability of a forecaster?

  Q2 How to compare two different rival forecasters?

- Q1 Henzi et al. (2023, *arXiv*)
  - $\mathcal{H}_0$: F1 reliable
  - Test calibration / GoF
  - automatic binning

- Q2 Henzi and Ziegel (2022, *Biometrika*)
  - $\mathcal{H}_0$: F1 dominates F2
  - Test forecast dominance
  - optional stopping

- For Q1, $i = 1, \ldots, n$, an e-variable is given by

$$E_n^{\text{id}} = \prod_{i=1}^{n} E_{q_i}(P_i, Y_i) \quad \text{where} \quad E_{qi}(P_i, Y_i) = \frac{q_i^{Y_i}(1-q_i)^{1-Y_i}}{P_i^{Y_i}(1-P_i)^{1-Y_i}}.$$