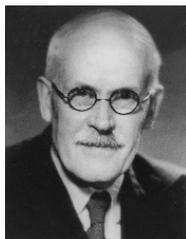# A note on data splitting with e-values

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].

Abstract

This note reanalyzes Cox's idealized example of testing with data splitting using e-values (Shafer's [8] betting scores). Cox's exciting finding was that the method of data splitting, while allowing flexible data analysis, achieves quite high efficiencies, of about 80%. The most serious objection to the method was that it involves splitting data at random, and so different people analyzing the same data may get very different answers. Using e-values instead of p-values remedies this disadvantage.

Contents

1 Introduction

Data splitting is a simple method for performing hypothesis testing using procedures chosen in the light of the data. The data is randomly split into disjoint parts, some of which used for choosing a test and another for performing it.

The method was analysed computationally and theoretically by Cox [1] in 1975 and further discussed in his 1977 review [2, Section 3.2], where he describes the method as well known and refers to an *American Statistician* paper [7] with a wide-ranging discussion of "snooping", "fishing", and "hunting" in data analysis.

As summarized in [2, Section 3.2], Cox [1] had analyzed the method of data splitting theoretically in a simple idealized situation showing that it achieves quite high efficiencies, of about 80%. However, Cox [1] also points out an obvious disadvantage of the method:

> Any method involving randomization in analysis is such that different investigators analyzing the same data by nominally the same method get different answers.

This criticism is reiterated in [2], including the discussion.

Using e-values [10] (also know as betting scores [8]) makes the problem of getting different answers by different investigators much less serious. The average of e-values is always an e-value, and so we can average e-values resulting from several random data splits obtaining a less random valid e-value (an almost deterministic one if the number of splits is very large). This note studies the dependence of the resulting p-values in Cox's idealized situation on the random split of the data and then performs similar (but simpler) analysis using e-values instead of p-values.

For a recent review of testing by data splitting, see [3, Section 1]. That paper also discusses methods based on combining p-values resulting from different data splits, both using the median (Section 2.1) and the arithmetic mean (Section 2.2). Both methods of combination, however, result in an extra factor of 2 (which has served as an inspiration for this note). Combination of e-values by averaging is discussed, in a similar context, in [11, Section 4], which refers to averaging a large number of e-values as derandomization.

Section 3 discusses the dependence on the random data split of the resulting p-values in Cox's [1] setting (discussed in Section 2). We will see that in many cases it is significant. Section 4 replaces p-values by e-values. To compare e-values and p-values, we use Shafer's [8] calibrator and Jeffreys's rule of thumb [4, Appendix B, p. 435]: a p-value of 0.05 corresponds to an e-value of $10^{1/2}$, and a p-value of 0.01 corresponds to an e-value of 10.

This note is not self-contained in that it does not define e-values; see [8] or [10] for definitions.

| | | $\alpha = 0.1$ | | | | $\alpha = 0.01$ | | | |
m	Δ	p = 0.2	p = 0.4	p = 0.6	exact	p = 0.2	p = 0.4	p = 0.6	exact
2	1	0.22	0.21	0.18	0.26	0.047	0.041	0.032	0.058
	2	0.31	0.49	0.43	0.64	0.22	0.18	0.12	0.28
	4	0.89	0.93	0.88	0.99	0.80	0.75	0.57	0.92
10	1	0.065	0.071	0.070	0.092	0.014	0.014	0.012	0.018
	2	0.21	0.26	0.26	0.37	0.091	0.094	0.076	0.14
	4	0.60	0.79	0.82	0.95	0.54	0.64	0.53	0.82
	6	0.85	0.97	0.99	1.000	0.85	0.96	0.93	0.998

Table 1: Cox's results for the effective power of data splitting for different p and the effective power of the exact test

2   Cox's ideal situation

We are given m independent random samples of size r from normal populations with means $\mu_1, \ldots, \mu_m$ 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 defines the *effective* level α power of this data-splitting procedure to be the probability that the correct population is chosen and that the second-portion sample is significant at least at level α. It is given by the formula

$$\Phi\left(-k_\alpha^* + \Delta\sqrt{1-p}\right) \int_{-\infty}^{\infty} \Phi(v)^{m-1}\phi\left(v - \Delta\sqrt{p}\right) dv \tag{1}$$

where $\Delta := \mu\sqrt{r}/\sigma_0$, and ϕ and Φ are, respectively, the standard normal density function and distribution function, and k_α^* is defined by $\Phi(-k_\alpha^*) = \alpha$.

Cox also defines an *exact procedure* that tests the means collectively for significance using the largest mean as test statistic. (Cox often surrounds "exact" by quotes, but let us omit those.) If the largest mean is significant at level α_m in the usual test for a single mean, its level of significance after allowing for selection is

$$\alpha := 1 - (1 - \alpha_m)^m. \tag{2}$$

The effective level α power (where "effective" is used in the same sense of choosing the correct population) is

$$\int_{k_{\alpha_m}^*}^{\infty} \Phi(v)^{m-1}\phi\left(v - \Delta\right) dv, \tag{3}$$

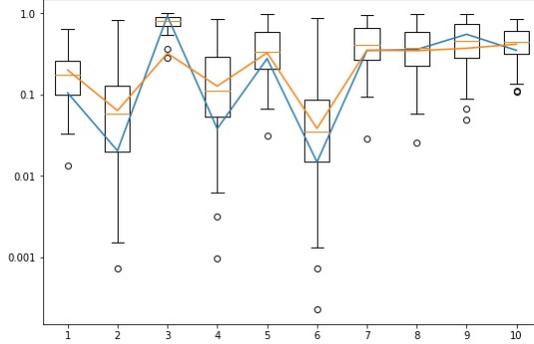where α_m is defined by (2).

Figure 1: Box plots for 10 randomly generated datasets with $m = 2$ and $\Delta = 1$, with the exact p-values added in blue (details in text)

Notice that (1) and (3) are only lower bounds for the power as it is usually defined. Cox's numeric results about the effective power of these tests are given in Table 1. In our experiments we will always use $p := 0.4$ (the middle value).

3   Experiments 1: variability of p-values

Let us assume, without loss of generality, that $\sigma_0 = 1$. The results of our computational experiments are very much affected by the choice of the random seed for the random number generator (but our conclusions will not be affected, of course).

The *data-split p-value* is computed as

$$k_\alpha^*, \tag{4}$$

where $\alpha := b/\sigma = \sqrt{(1-p)r}b$, b is the mean of the second portion of the sample with the largest mean of the first portion, and

$$\sigma := 1/\sqrt{(1-p)r} \tag{5}$$

is the standard deviation of the second portion of each sample.

Similarly, the *exact p-value* is computed as

$$1 - (1 - k_\alpha^*)^m, \tag{6}$$

where $\alpha := \sqrt{r}c$ and c is the largest mean of all m samples.

Results for one of the low-power cases (according to Table 1) are shown as Figure 1. The 10 boxplots correspond to 10 datasets (each consisting of $m = 2$ samples) randomly generated from the normal distribution $N(\mu, 1)$ making $\Delta = 1$. The value of r (sample size) is set to 100 (see Remark 1 below); therefore, $\mu = \Delta/\sqrt{r} = 0.1$. The exact p-values for the same 10 datasets are shown in blue,
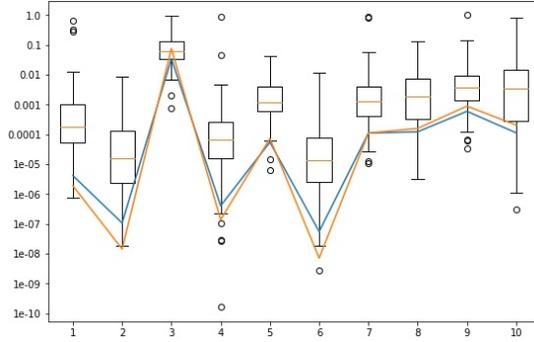
Figure 2: The analogue of Figure 1 for $m = 2$ and $\Delta = 4$
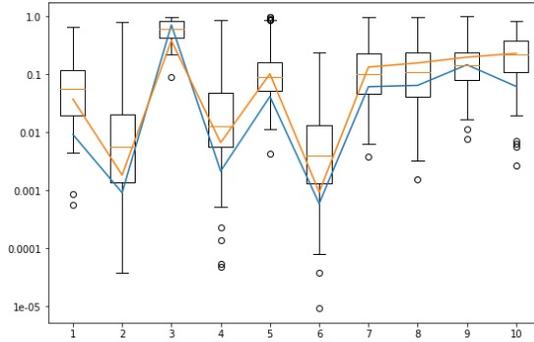

Figure 3: The analogue of Figure 1 for $m = 2$ and $\Delta = 2$

and for visibility they are connected with blue lines. (The orange lines should be ignored in this section; they will be explained in Section 4.) Each dataset is split randomly 100 times, and the corresponding 100 p-values (4) are summarized as boxplot (whose box is bounded by the quartiles and contains the median; for the rules governing the whiskers see the `matplotlib` documentation).

Remark 1. The role of r is not essential in Cox's calculations as long as pr (the size of the first portion of each sample) is integer (in the context of Table 1, as long as r is divisible by 5). In particular, r does not enter (1) or (3). In our experiments, however, we can't set $r := 5$ since this would lead to only $\binom{5}{2} = 10$ possible data splits for each sample, which would show in the boxplots, especially for $m = 2$.

A high-power case is shown as Figure 2. Most of the results are highly statistically significant (the p-values are below 1%), and so, from the point of view of the conventional levels 1% and 5%, it is not as informative as the mid-power case shown as Figure 3. In that figure, the statistical significance of
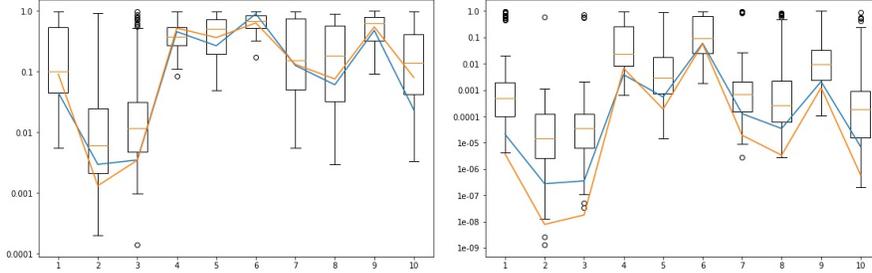
Figure 4: The analogues of Figure 1 for $m = 10$ (both panels) and $\Delta = 2$ (left panel) and $\Delta = 4$ (right panel)

data-split p-values strongly depends on the random data split.

Figure 4 reports results for $m = 10$ and $\Delta \in \{2, 4\}$ (the two mid-power cases). The effect of the randomness in the data split on the statistical significance of the resulting p-values is again substantial.

4 Experiments 2: using e-values

There is a natural way to define data-split e-values. As Shafer [8] discusses in Section 2.2, a natural choice is the likelihood ratio between an alternative and the null hypotheses. The true alternative (namely, its parameter μ) is not known, but in Cox's data-splitting scheme we can use the largest mean a of the first portion as an estimate of μ, which gives us the likelihood ratio

$$e := \frac{(2\pi)^{-1/2} \exp\left(-\frac{(b-a)^2}{2\sigma^2}\right)}{(2\pi)^{-1/2} \exp\left(-\frac{b^2}{2\sigma^2}\right)} = \exp\left(\frac{ab}{\sigma^2} - \frac{a^2}{2\sigma^2}\right), \tag{7}$$

where b is the mean of the second portion of the sample with the largest mean a of the first portion, and σ is the standard deviation (5) of the mean of the second portion.

It is not immediately clear, however, how to define an e-value analogue for Cox's exact p-values. Not knowing the alternative hypothesis hurts more in the case of e-values; while the Neyman–Pearson p-values do not depend on the parameter μ, we cannot set the e-value to the likelihood ratio since it does depend on μ. A natural way out is to choose a "prior distribution" over μ and mix the likelihood ratios corresponding to different μ with respect to that distribution, but this appears more ad hoc and more complicated than what we did in the case of data splitting.

Figure 5 presents the e-values (7) for $m = 2$ and $\Delta = 2$; therefore, it contains some common information with Figure 3, which is, however, presented differently. We adapt the method of [3, Section 1] to e-values (as already mentioned in Section 1). Instead of the boxplots of p-values in Figure 3, we can now average
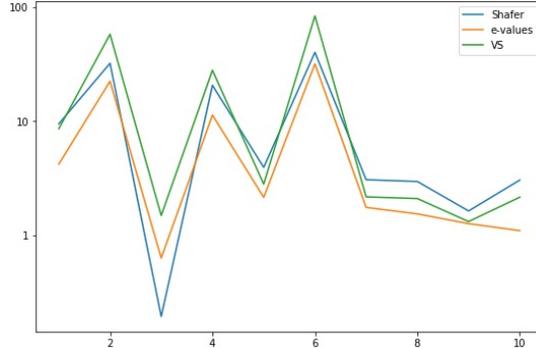
5

Figure 5: The average e-values for $m = 2$ and $\Delta = 2$ in orange complemented by calibrated p-values in blue and green

the 100 e-values (7) for $m = 2$ and $\Delta = 2$, obtained from the same datasets and the same data splits as in Figure 3 (the remaining randomness in the average e-values is tiny and boxplots would not be informative for them). The average e-values are shown in orange in Figure 5. Using Jeffreys's [4, Appendix B] rule of thumb already mentioned in Section 1, we can regard e-values above $10^{1/2}$ as statistically significant and e-values above 10 as highly statistically significant. We can see that two or three resulting e-values are highly statistically significant in this sense.

Remark 2. We could eliminate *all* randomness in the average e-values by averaging over all splits of the samples into first and second portions, as discussed in [11, Section 4]. This would be feasible for $m = 2$ and $r = 5$ (with 10 possible splits for each sample) but not in our case $r = 100$ (with about 10^{65} possible splits for each sample).

As Shafer says in section "Comparing scales" [8, Section 3], there is no one way to compare p-values, such as (6), and e-values, such as (7). As discussed in detail in [10], there are various ways of *calibrating* p-values, i.e., making them into e-values (although there is essentially one way, $e \mapsto 1/e$, of making e-values into p-values).

Let us first apply Shafer's *calibrator* (i.e., function transforming p-values into e-values)

$$S(p) := \frac{1}{\sqrt{p}} - 1 \tag{8}$$

[8, (6)] to Cox's exact p-values. Table 2 (an extension of Shafer's [8] Table 2 "Making a p-value into a betting score") gives the e-values produced by Shafer's calibrator in the third column and Jeffreys's [4, Appendix B] estimates, which are remarkably close to Shafer's.

The exact p-values shown in blue in Figure 3 are shown in Figure 5, also in blue, after being transformed into e-values by Shafer's calibrator (8). They

p-value p	$1/p$	$S(p)$	Jeffreys	VS(p)
1	1	0		1
0.1	10	2.2		1.6
0.05	20	3.5	3.2	2.5
0.01	100	9.0	10	8.0
0.005	200	13.1		13.9
0.001	1,000	30.6		53.3
0.000001	1,000,000	999		26628

Table 2: Calibrating p-values

are fairly close to the orange lines. Notice that in one respect the comparison between the blue and orange lines is unfair to the average e-values: the input p-values are exact while the average e-values are based on data splitting. (On the other hand, our treatment of p-values and e-values is not symmetric in that we compare them in the e-domain, which makes sense in discussing a paper promoting betting scores as language for statistical and scientific communication.)

The orange lines in the previous figures, Figures 1–4, show the average e-values in the p-domain using the inverse transformation S^{-1} to Shafer's calibrator

$$S^{-1}(e) = (e+1)^{-2}.$$

Of course, these are not valid p-values; they are the p-values we would need to obtain our average e-values by using Shafer's calibrator. We can see that for all those figures we obtain similar e-values directly by using data splitting and by applying Shafer's calibrator to exact p-values.

A particularly natural class of calibrators is

$$p \mapsto \epsilon p^{\epsilon-1}, \tag{9}$$

where $\epsilon \in (0,1)$ is the parameter. For a small ϵ, it comes close to the ideal (but not attainable) calibrator $p \mapsto 1/p$ (if we ignore constant factors, as customary in the algorithmic theory of randomness, an area where many of these ideas originated). To get rid of the parameter ϵ, we may consider the upper bound

$$\mathrm{VS}(p) := \sup_\epsilon \epsilon p^{\epsilon-1} = \begin{cases} -\exp(-1)/(p \ln p) & \text{if } p \leq \exp(-1) \\ 1 & \text{otherwise.} \end{cases}$$

It was proposed in [9, Section 9] and, independently, [6]. It should be remembered that it is not a valid e-value and just shows what is attainable with the calibrators in the class (9).

Figure 5 also shows the VS transformations of the exact p-values (in green). As expected, they are slightly above the bona fide e-values shown in orange and blue. The VS values are also given in Table 2 (last column). We can see that the VS transformation is also not so different from Shafer's calibrator and Jeffreys's intuition unless the input p-value is very large or extremely small.
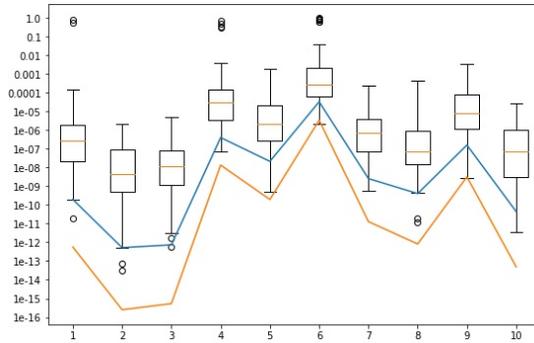
7

Figure 6: The analogue of Figure 1 for $m = 10$ and $\Delta = 6$
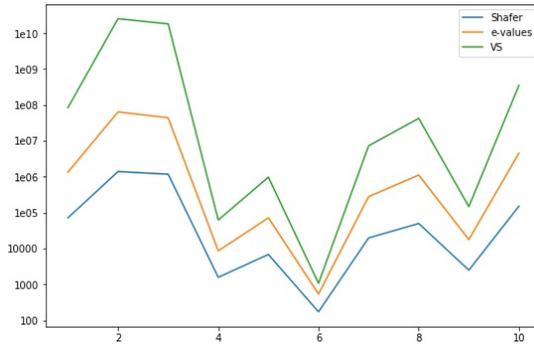

Figure 7: Blue and orange lines of Figure 6 in the e-domain complemented by the VS plot

We can see already in Figure 4 that for a strong signal the performance of e-values tends to improve as compared with Shafer-calibrated p-values. This is illustrated further by Figure 6, where $\Delta = 6$. The e-values produced by Shafer's calibrator are significantly worse than the e-values obtained by data splitting (which has a whiff of superefficiency).

Figure 7 shows the blue and orange lines of Figure 6 in the e-domain, complemented by the VS bound. We can see that using some calibrators in the class (9) allows us to beat the e-values obtained by data splitting. (The problem, however, is that we do not know how to choose the right ϵ in advance.)

5 Conclusion

Despite the high efficiency of the method of data splitting applied to p-values (at least in Cox's idealized situation), the adoption of the method may have been

hindered by the significant amount of randomness at the stage of data analysis, with different analysts potentially arriving at very different conclusions while using the same approach. A byproduct of adopting the language of betting and e-values as a means of statistical and scientific communication [8] is that this drawback disappears.

Acknowledgments

References

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

[2] David R. Cox. The role of significance tests (with discussion). *Scandinavian Journal of Statistics*, 4:49–70, 1977.

[3] Cyrus J. DiCiccio, Thomas J. DiCiccio, and Joseph P. Romano. Exact tests via multiple data splitting. *Statistics and Probability Letters*, 166:108865, 2020.

[4] Harold Jeffreys. *Theory of Probability.* Oxford University Press, Oxford, third edition, 1961.

[5] Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses.* Springer, New York, third edition, 2005.

[6] Thomas Sellke, M. J. Bayarri, and James Berger. Calibration of p-values for testing precise null hypotheses. *American Statistician*, 55:62–71, 2001.

[7] Hanan C. Selvin and Alan Stuart. Data-dredging procedures in survey analysis. *American Statistician*, 20(3):20–23, 1966.

[8] Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. To be read before the Royal Statistical Society on 9 September, 2020, and to appear as discussion paper in the *Journal of the Royal Statistical Society A*, 2020.

[9] Vladimir Vovk. A logic of probability, with application to the foundations of statistics (with discussion). *Journal of the Royal Statistical Society B*, 55:317–351, 1993.

[10] Vladimir Vovk and Ruodu Wang. Combining e-values and p-values. Technical Report arXiv:1912.06116 [math.ST], arXiv.org e-Print archive, May 2020.
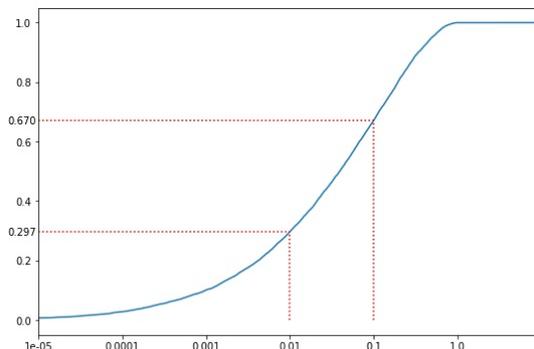
Figure 8: The distribution function of exact p-values for $m = 2$ and $\Delta = 2$

[11] Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences of the USA*, 117:16880–16890, 2020.

A   The distribution of p-values under the alternative hypothesis

It is customary in statistical hypothesis testing to talk about the size and power of statistical tests in the toy situation of a simple null hypothesis and a simple alternative (see, e.g., [5, Section 3.1]). This was essentially the language of Cox's paper [1] and Section 2, except that power was replaced by its lower bound, effective power. The experiments of Section 3, however, concerned the behaviour of the p-values under the alternative hypothesis. This appendix will spell out the connections and explore the distribution of p-values experimentally.

The distribution function F of the p-values under the alternative hypothesis is exactly the power as function of the test size. Let us check this, assuming that the p-value is distributed uniformly on $[0, 1]$ under the null hypothesis. For any significance level $\alpha \in (0, 1)$, the critical region of size α consists of all observations producing a p-value $p \leq \alpha$. Its probability $F(\alpha)$ is, by definition, the power at significance level α.

Figure 8 gives the distribution function of exact p-values for $m = 2$ and $\Delta = 2$, which are the values in Table 1 that have the effective powers farthest from the end-points of the interval $[0, 1]$. The powers corresponding to the sizes $\alpha \in \{0.1, 0.01\}$ are greater than the effective powers reported in Table 1 ($0.670 > 0.64$ and $0.297 > 0.28$), but the difference is modest. To plot the distribution function in Figure 8, I have generated from the alternative distribution for the data 10,000 exact p-values (valid under the null distribution).