

P-values

July 15, 2019

We consider a non-parametric test with an unknown distribution of values¹. We don't want to assume that the distribution is symmetric around its mean, nor that it is centered on 0. We also want a two-tailed test.

We compute the distribution of the value under H_0 , and then use it to compute p-values.

There are two methodological questions associated with this analysis (even though they are intricate):

- How to write the p-value properly?
- How to compare them to the Type I error rate? (possibly with a FDR-correction for multiple testing so we can't directly look at quantiles)

P-value definition

There are different ways to define the p-value with a non-parametric method and an unknown distribution.

Based on [1]:

$$p_1 = \frac{\#(|\eta^*| \geq |\eta_{emp}|) + 1}{\#(\eta^*) + 1}$$

where η^* are the values of the index under H_0 and η_{emp} is the observed value.

Based on [2]:

$$p_2 = 2 \min (Pr(\eta^* \leq \eta_{emp}), Pr(\eta^* \geq \eta_{emp}))$$

where we can assume that $Pr(\eta^* \leq \eta_{emp}) = \frac{\#(\eta^* \leq \eta_{emp}) + 1}{\#(\eta^*) + 1}$ (and conversely).

Test

We assume that η_{H0} follows a Beta(2,5) distribution which we shift to the left by 0.4 (eq. 1)².

$$\eta_{H0} \sim \text{Beta}(2, 5) - 0.4 \tag{1}$$

We build the distribution of η_{H0} with 10 000 draws, corresponding to Fig 1. We first compared fixed potential values of the observed index to see if H_0 is kept or rejected. In Fig. 1, the thick black line indicates the position of 0 while the dotted lines correspond to the 5th and 95th percentile of the distribution. Two possible observed values are also shown in the plot: the blue line corresponds to a case where H_0 should not be rejected (and is not with either pval1 or pval2) while the red line corresponds to a case where H_0 should be rejected as it is above the 95th percentile. However, pval1 > 0.05 (keep H_0) while pval2 < 0.05 (reject H_0). For now, it seems that p_2 is more accurate (?).

Then, we drew 1000 possible observed values from the H_0 distribution and compute the two p-values for each of them. We first note that both p-values are uniformly distributed between 0 and 1 and that the FDR is kept at 5% for both p-values (Fig. 2). However, pval1 is shifted towards 0 compared to the mode of the distribution, which means that there can be a bias in the estimate of the pvalue. This is not the case with the 2nd computation of the pvalue.

¹It all came from questions on the distribution of η by Gross et al., 2013, but this applies to more complex distributions

²Before, it was 0.6 but it was deemed too unrealistic.

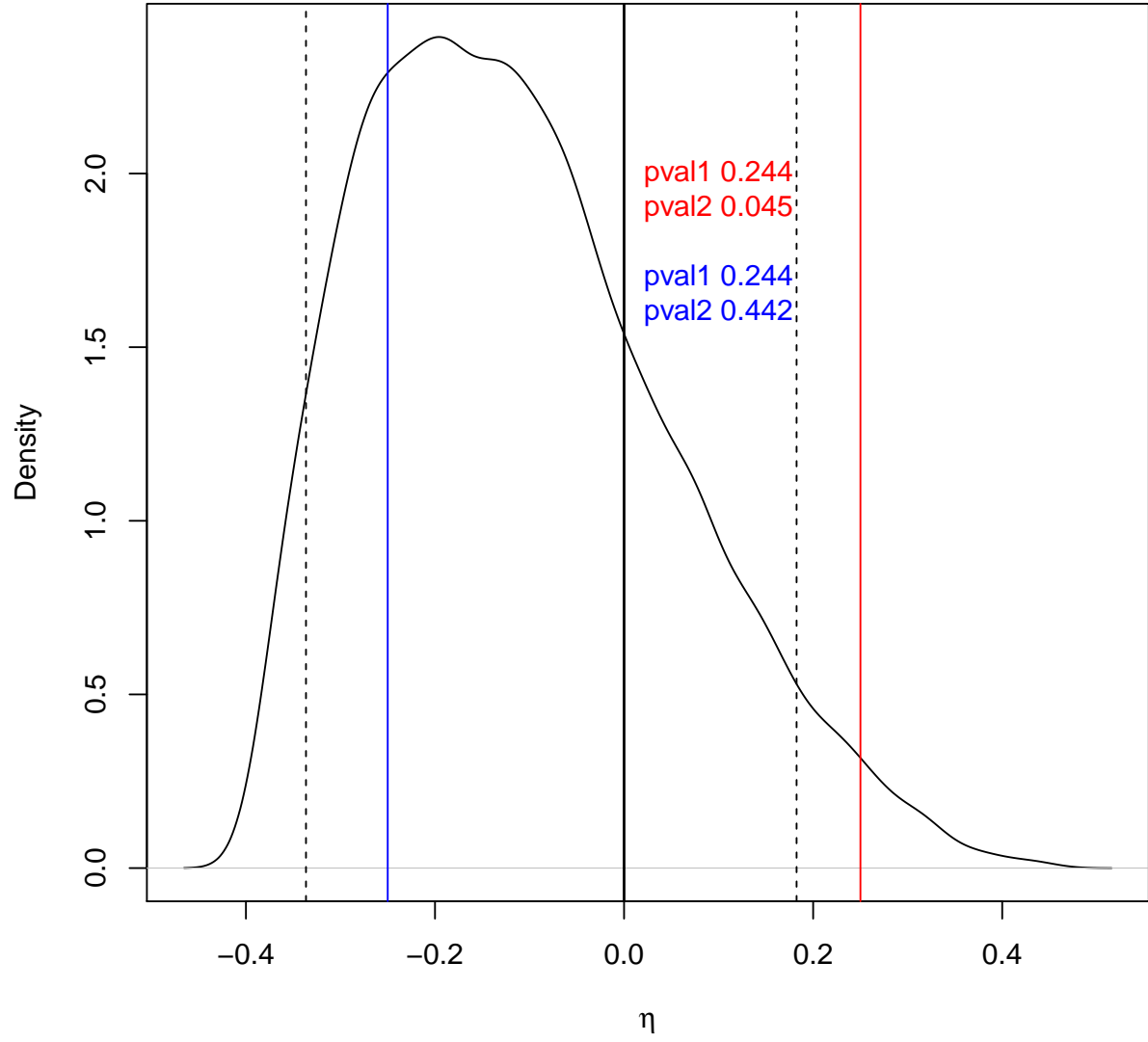


Figure 1: Distribution of the studied index under H_0 (eq. 1) with possible observed values in blue and red (see text for more details)

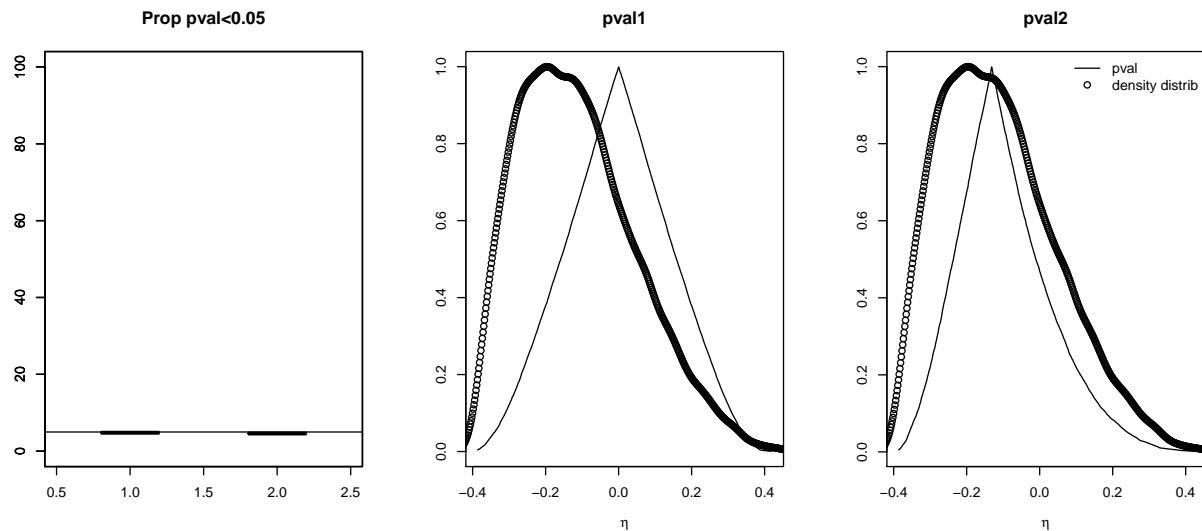


Figure 2: Comparison between the two different p-values and a 5% threshold: both p-values lead to a 5% FDR. In the center and right panels, the distribution of the p-values according to the possibly observed index under H_0 is shown with lines while dots indicate the real distribution of this index. There is a mode shift and possibly a bias with $pval1$.

References

Gross, K., B. J. Cardinale, J. W. Fox, A. Gonzalez, M. Loreau, H. Wayne Polley, P. B. Reich, and J. van Ruijven, 2013. Species richness and the temporal stability of biomass production: a new analysis of recent biodiversity experiments. *The American Naturalist* **183**:1–12.