

Simulating null distributions

Sayani Gupta

19/08/2020

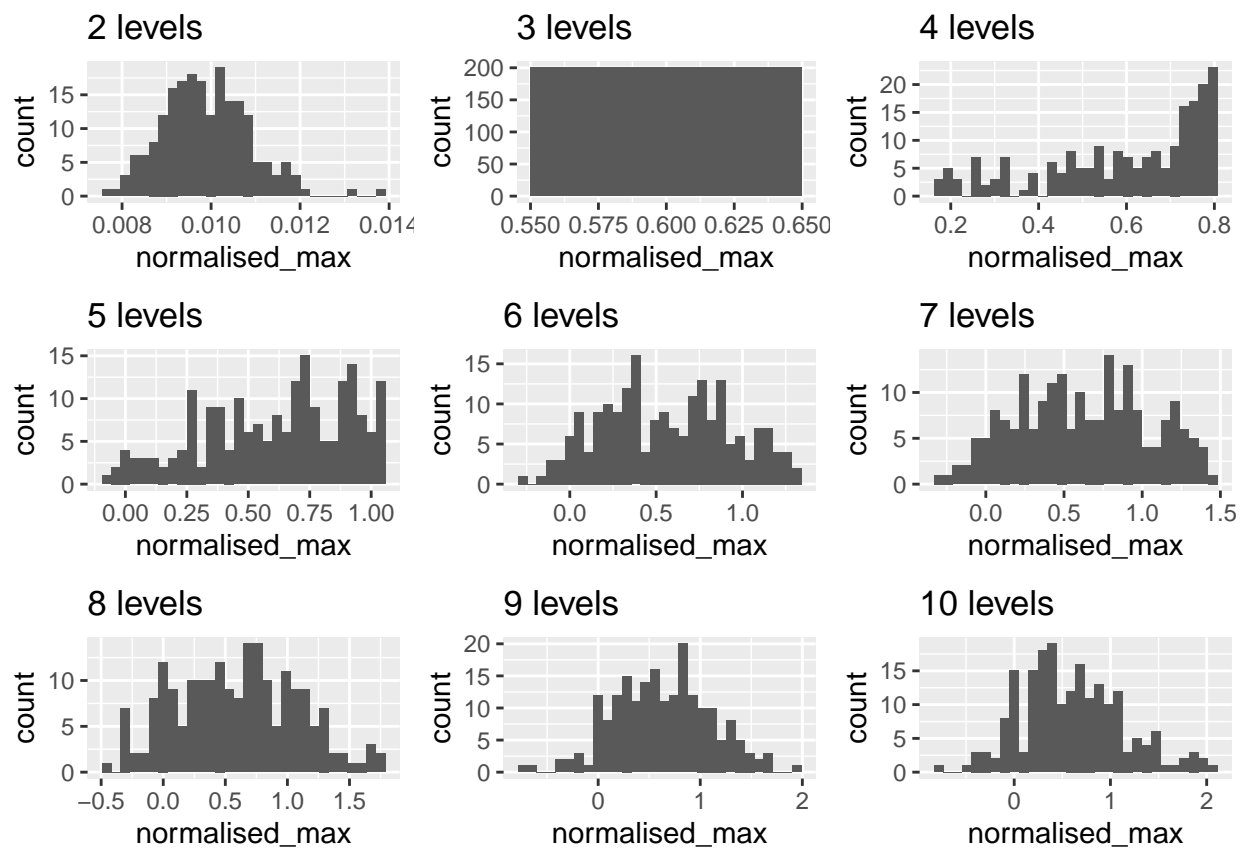
If we suppose that MMPD is useful to normalize for the number of levels, then distribution of MMPD should be same for different levels, provided the response variable for all the levels comes from the same distribution. Following scenarios are considered to test that by breaking down the problem and checking - A) Does normalizing works for different x-axis levels? B) if yes, then does further normalization using median works for different facet categories?

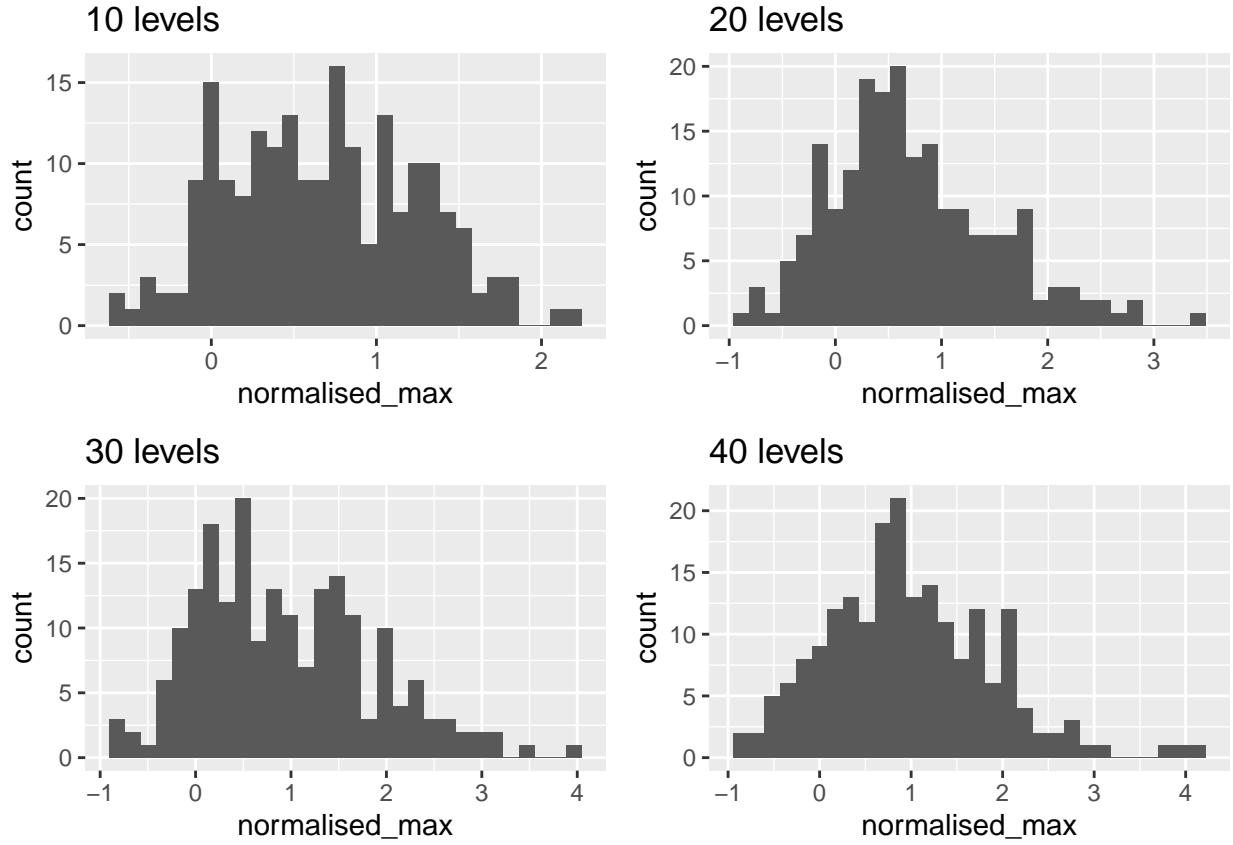
A) To test if normalization works for different x-axis levels:

For x-axis, first maximum of pairwise JS distances are computed and it is normalized using Fisher–Tippett–Gnedenko theorem. Different x-axis levels viz, 10, 20, 30, 40 are considered and data is simulated 200 times for each of these cases to learn about the distribution of these normalized distances by plotting their histograms.

Null distribution of normalized maximum distances at this stage should be standardized Gumbel distribution. This is because we have assumed that distribution of JS distances are Normal and the maximum of such distances would follow a Gumbel distribution. When we normalize these maximum distances, ideally the distribution of those normalized distances should be standard Gumbel distribution (which is positively skewed).

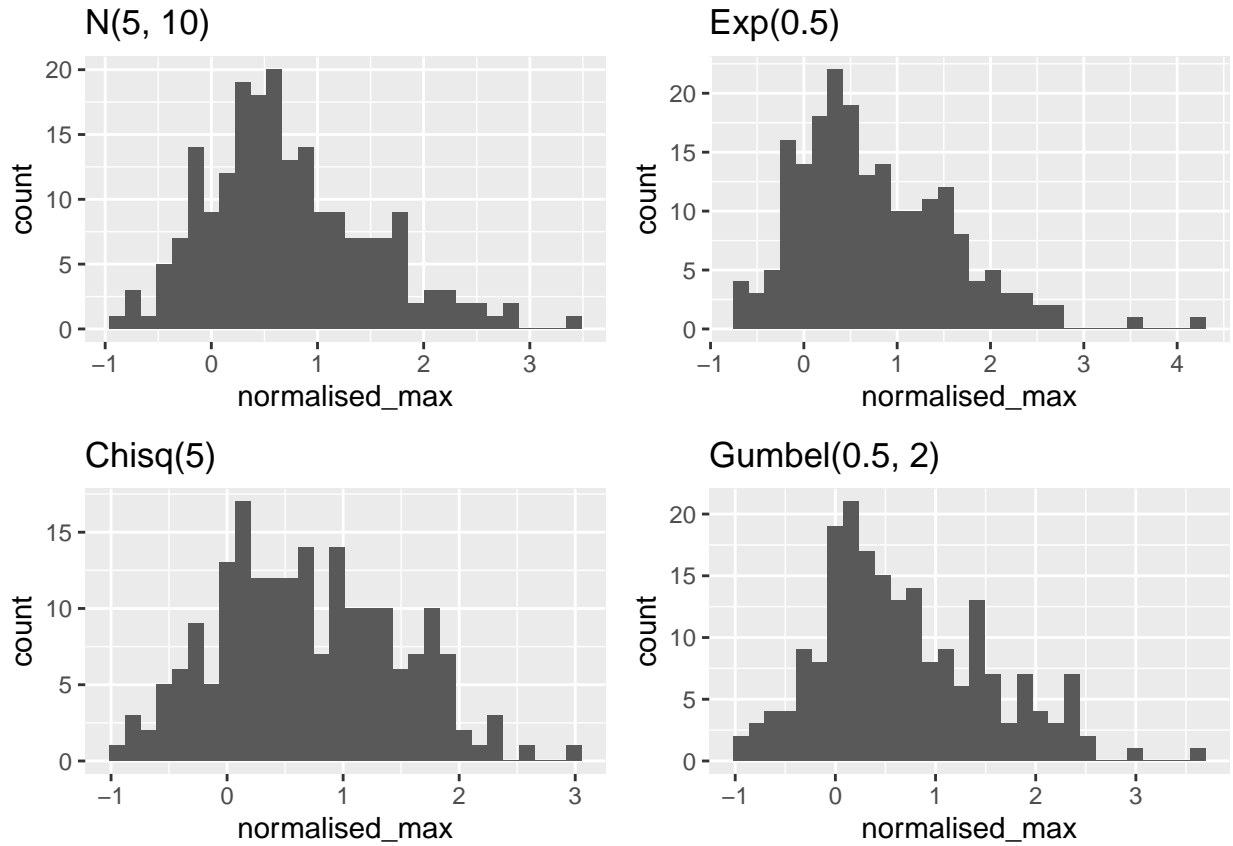
All levels drawn from $N(5, 10)$ distribution





From the above graphs, we could see that irrespective of the levels, the distribution of the normalized distances look like a positively skewed distribution when the underlying observations are drawn from a $N(5, 10)$ distribution. Let us check if the same holds true if the underlying distributions are from $\text{Exp}(0.5)$, $\text{Chi-squared}(5)$, $\text{Gumbel}(0.5, 2)$. Here, we fix the number of levels to 20.

Different distributions with 20 levels

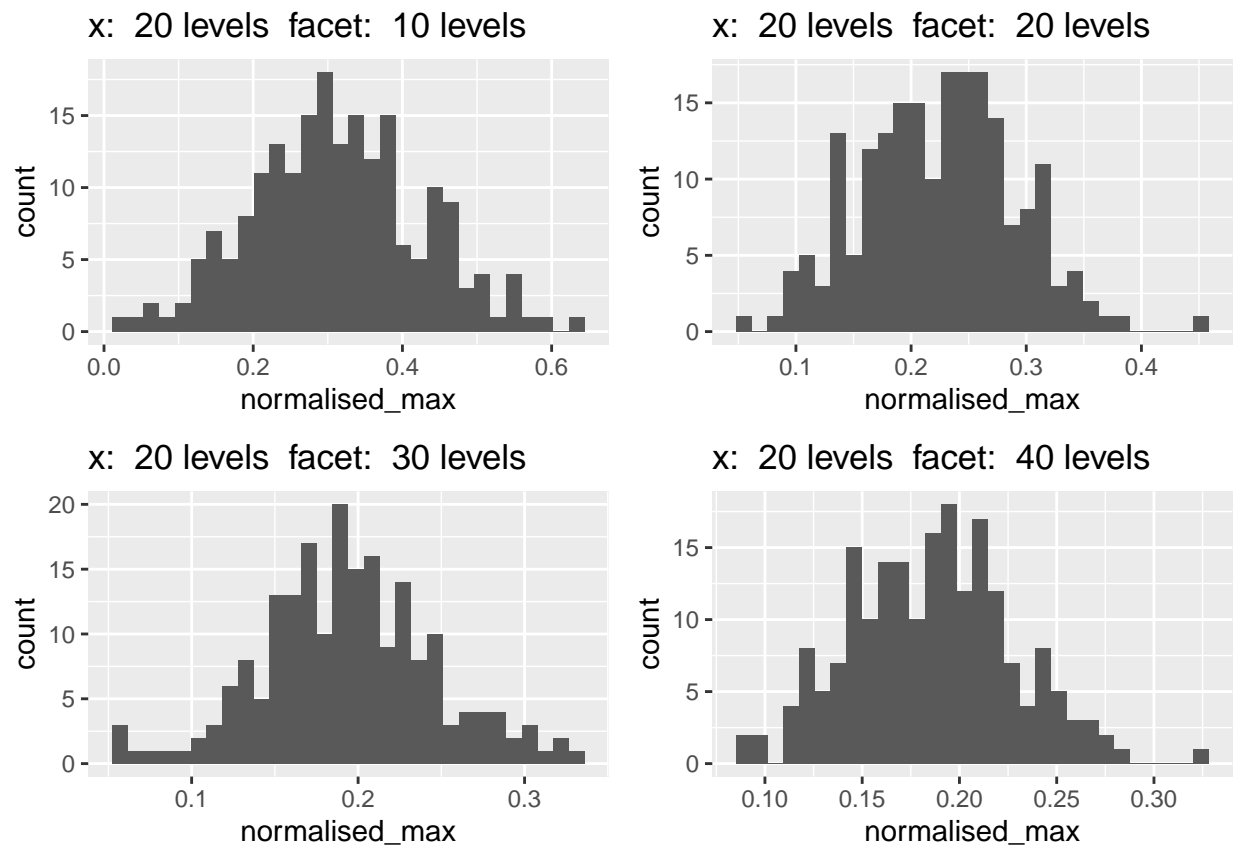


B) To test if further normalization using median works for different facet categories?

If the normalization along x-axis is working, we can vary the levels across facets keeping levels of x-axis constant to see if normalization along facets work or not.

For facets, first all normalized maximum distances are obtained and then their median is taken and further divided by $\log(\text{number of facet levels})$. Different facet levels viz, 10, 20, 30, 40 are considered for a fixed x-axis level and data is simulated 200 times for each of these cases to learn about the distributions.

All levels drawn from $\text{Exp}(2/3)$ distribution



It looks like the normalization is also working at this stage.