Choosing and rating harmonies

Contents

1 Idea

Even after excluding clashes, the list of harmonies left could be large and overwhelming for human consumption. Hence, there is a need to rank the harmonies basis how well they capture the variation in the measured variable and additionally reduce the number of harmonies for further exploration/visualization. Assuming a numeric response variable, our graphics are displays of distributions compared across combinations of categorical variables, one placed at x-axis and the other on the facet. Gestalt theory suggests that when items are placed in close proximity, people assume that they are in the same group because they are close to one another and apart from other groups. Hence, displays that capture more variation within different categories in the same group would be important to bring out different patterns of the data. Here, we have two main objectives:

- To choose harmonies for which distributions of categories are significantly different

- To rank the selected harmonies from highest to lowest variation in the distribution of their categories. The idea here is to rate a harmony pair higher if this variation between different levels of the x-axis variable is higher on an average across all levels of facet variables.

2 Computing distances

One of the potential ways to evaluate this variation is by computing the pairwise distances between the distributions of the measured variable. We do this through Jensen-Shannon distance which is based on Kullback-Leibler divergence.

The Jensen-Shanon distance between two probability distribution p_1 and p_2 is given by

$$d = [D(p_1, r) + D(p_2, r)]/2 \quad where \quad r = (p_1 + p_2)/2$$

where,

$$D(p_1, p_2) = \int_{-\infty}^{\infty} p_1(x) log \frac{p_1(x)}{p_2(x)}\, dx$$

is the Kullback-Leibler divergence between p_1 and p_2. Probability distributions are estimated through quantiles instead of kernel density so that there is minimal dependency on selecting kernel or bandwidth.

We call this measure of variation as Median Maximum Pairwise Distances (MMPD).

2.1 Normalize distances

The harmony pairs could be arranged from highest to lowest average maximum pairwise distances across different levels of the harmonies. But maximum is not robust to the number of levels and is higher for harmonies with higher levels. Thus these maximum pairwise distances need to be normalized for different harmonies in a way that eliminates the effect of different levels. The Fisher–Tippett–Gnedenko theorem in the field of Extreme Value Theory states that the maximum of a sample of iid random variables after proper re- normalization can converge in distribution to only one of Weibull, Gumbel or Freschet distribution, independent of the underlying data or process.

More formally, $d_1, d_2 \ldots, d_n$ be a sequence of independent and identically-distributed pairwise distances and $M_n = \max\{d_1, \ldots, d_n\}$. Then Fisher–Tippett–Gnedenko theorem (Haan and Ferreira 2007) suggests that if a sequence of pairs of real numbers (a_n, b_n) exists such that each $a_n > 0$ and $\lim_{m\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F(x)$, where F is a non-degenerate distribution function, then the limit distribution F belongs to either the Gumbel, Fréchet or Weibull family. The normalizing constants (a_n, b_n) vary depending on the underlying distribution of the pairwise distances. Hence to normalize appropriately, it is important to assume a distribution of these distances.

2.2 Distribution of distances

Theoretical: JS distances are distributed as chi-squared with m df where we discretize the continuous distribution with m discrete values. Taking sample percentiles to approximate the integral would mean taking $m = 99$. With large m, chi-squared is asymptotically normal by the CLT. Thus, by CLT, ${\chi^2}_m \tilde{} N(m, 2m)$, which would depend on the number of discretization used to approximate the continuous distribution. Then $b_n = 1 - 1/n$ quantile of the normal distribution and $a_n = 1/[n * \phi(b_n)]$ where ϕ is the normal density function. n is the number of pairwise comparisons being made.

Empirical: Distribution of JS distances is assumed to be normal but the mean and variance are estimated from the sample, rather than deducing it from the number of discretization used to approximate the continuous distribution.

3 Choose thresholds for harmonies through permutation test

Assumption: random permutation without considering ordering (global)

1. Given the data; $\{v_t : t = 0, 1, 2, \ldots, T-1\}$, the MMPD is computed and is represented by $MMPD_{obs}$.
2. From the original sequence a random permutation is obtained: $\{v_t^* : t = 0, 1, 2, \ldots, T-1\}$.
3. MMPD is computed for all random permutation of the data and is represented by $MMPD_{sample}$.
4. Steps (2) and (3) are repeated a large number of times M (e.g. 1000).
5. For each permutation, one $MMPD_{sample}$ value is obtained.
6. 95^{th} percentile of this $MMPD_{sample}$ distribution is computed and stored in $MMPD_{threshold}$.
7. If $MMPD_{obs} > MMPD_{threshold}$, harmony pairs are accepted. Only one threshold for all harmony pairs.

Pros: Considering thresholds global for all harmony pairs would imply less computation time.

Cons: Only one threshold for all harmony pairs means we are assuming distribution of all harmonies pairs are similar, which might not be the case.But nevertheless, it is a good benchmark.

4 Does normalization work? - Minimal examples

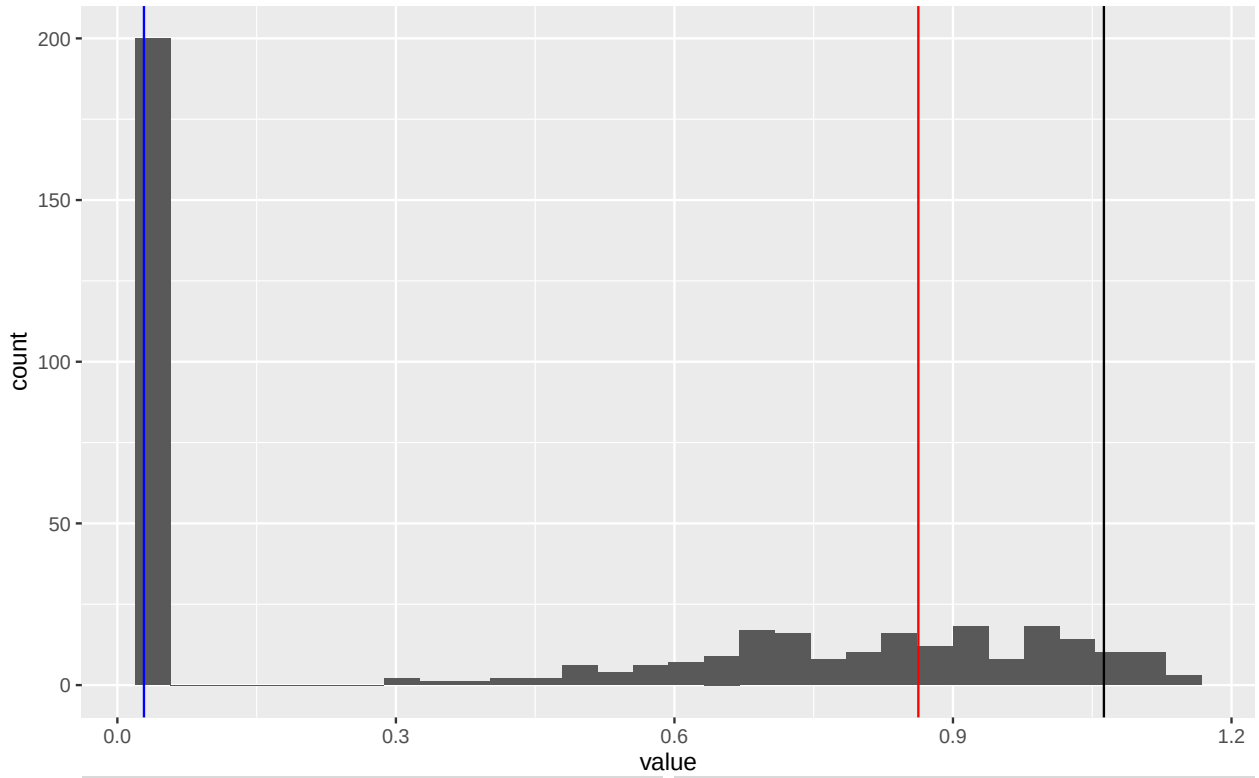
Consider two cyclic granularities A and B with 2 and 3categories. Thus, the harmony table consisting of all possible harmony pairs (assuming all pairs are harmonies), would look like the following:

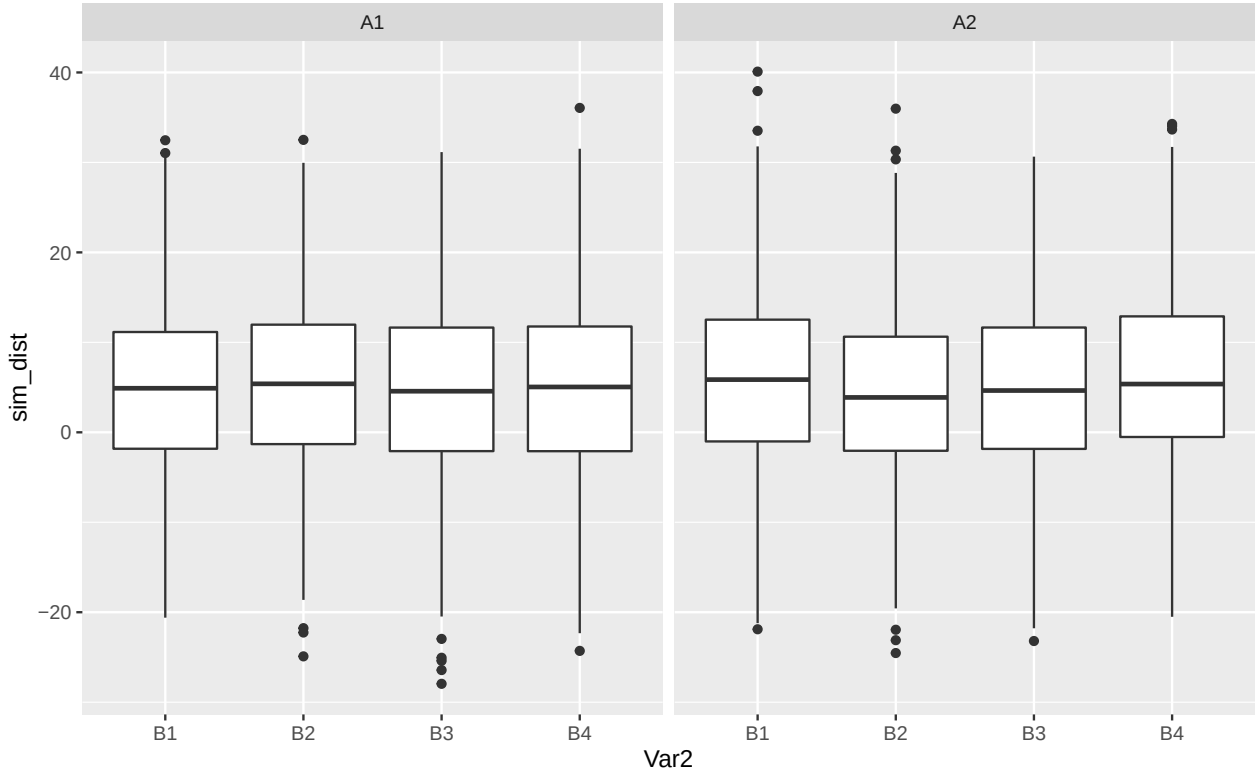
facet_variable	x_variable	facet_levels	x_levels
A	B	2	3
B	A	3	2

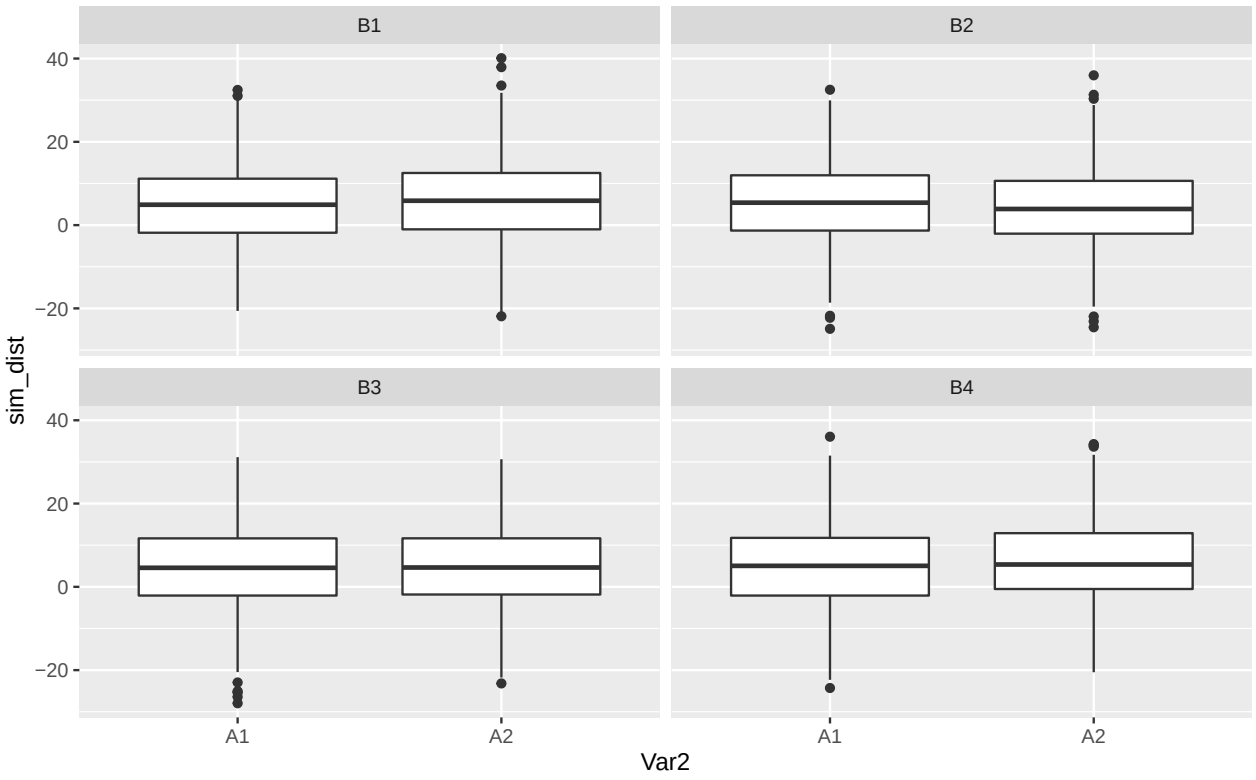
The output table has the value of MMPD (normalized median maximum pairwise distances), max_pd (un-normalized maximum pairwise distances), r(Rank of max_pd), gt_MMPD(global threshold of MMPD indicator) and gt_maxpd(global threshold of max_pd indicator).

4.1 Scenario 1: Simulated same normal distributions for all combinations of categories for all harmony pairs.

Failure to reject the null hypothesis when there is in fact no significant effect.

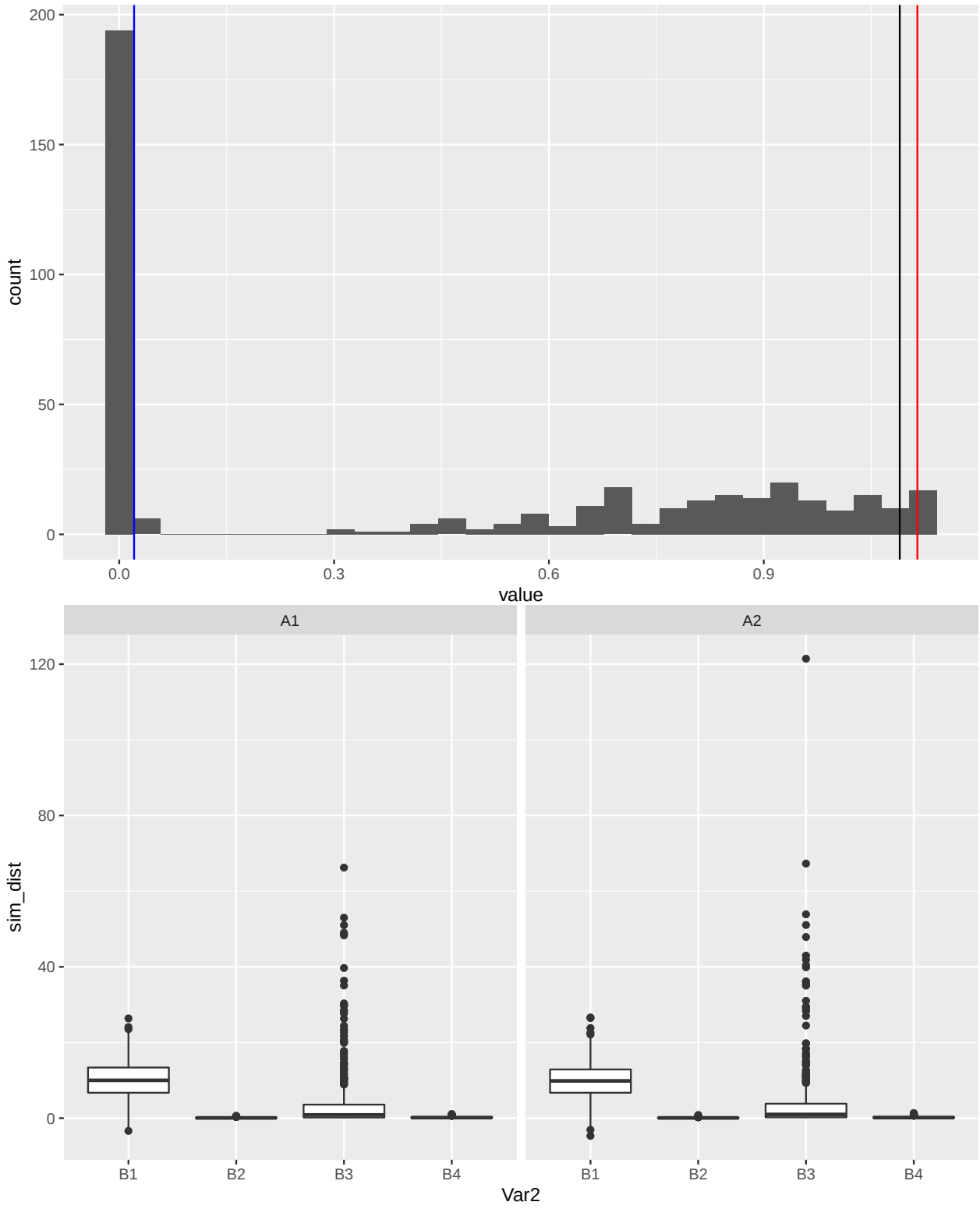

count
200
150
100
50
0
0.0
0.3
0.6
0.9
1.2
value

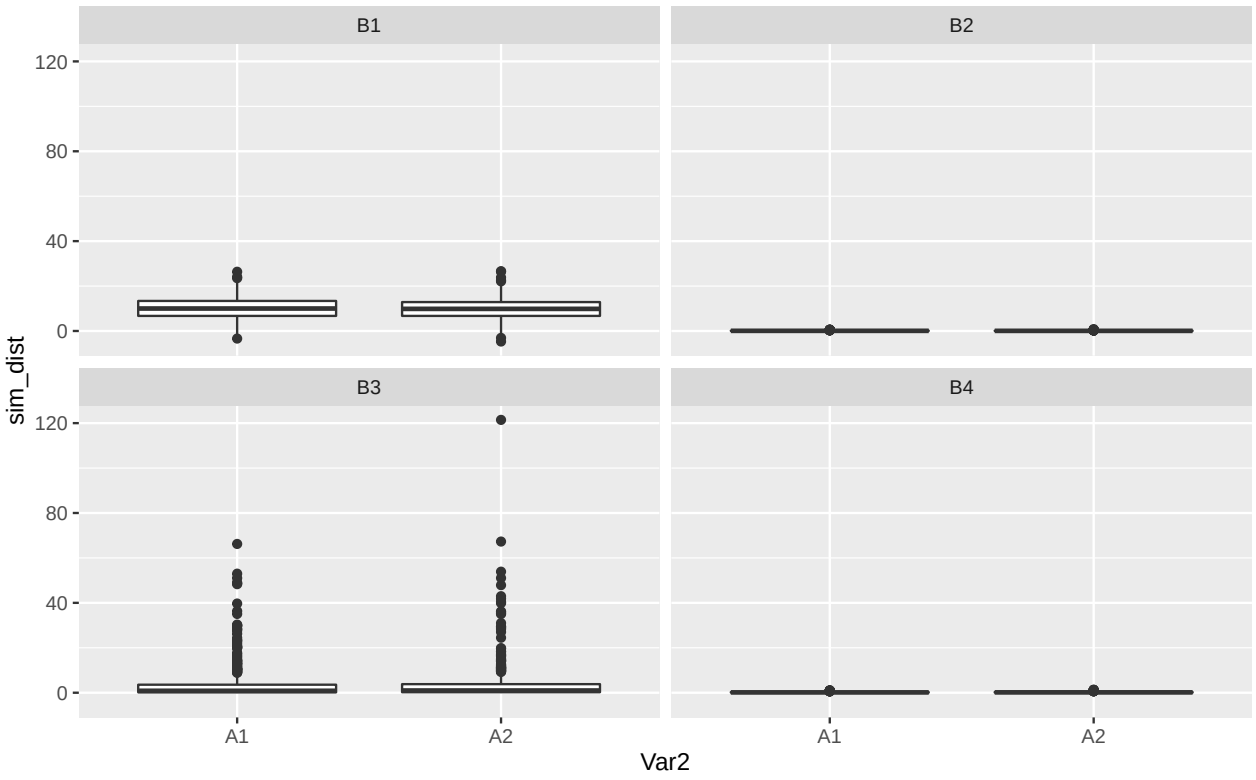

A1
A2
sim_dist
40
20
0
−20
B1
B2
B3
B4
B1
B2
B3
B4
Var2



Conclusion: The test does not select harmony pairs when distributions of categories are same

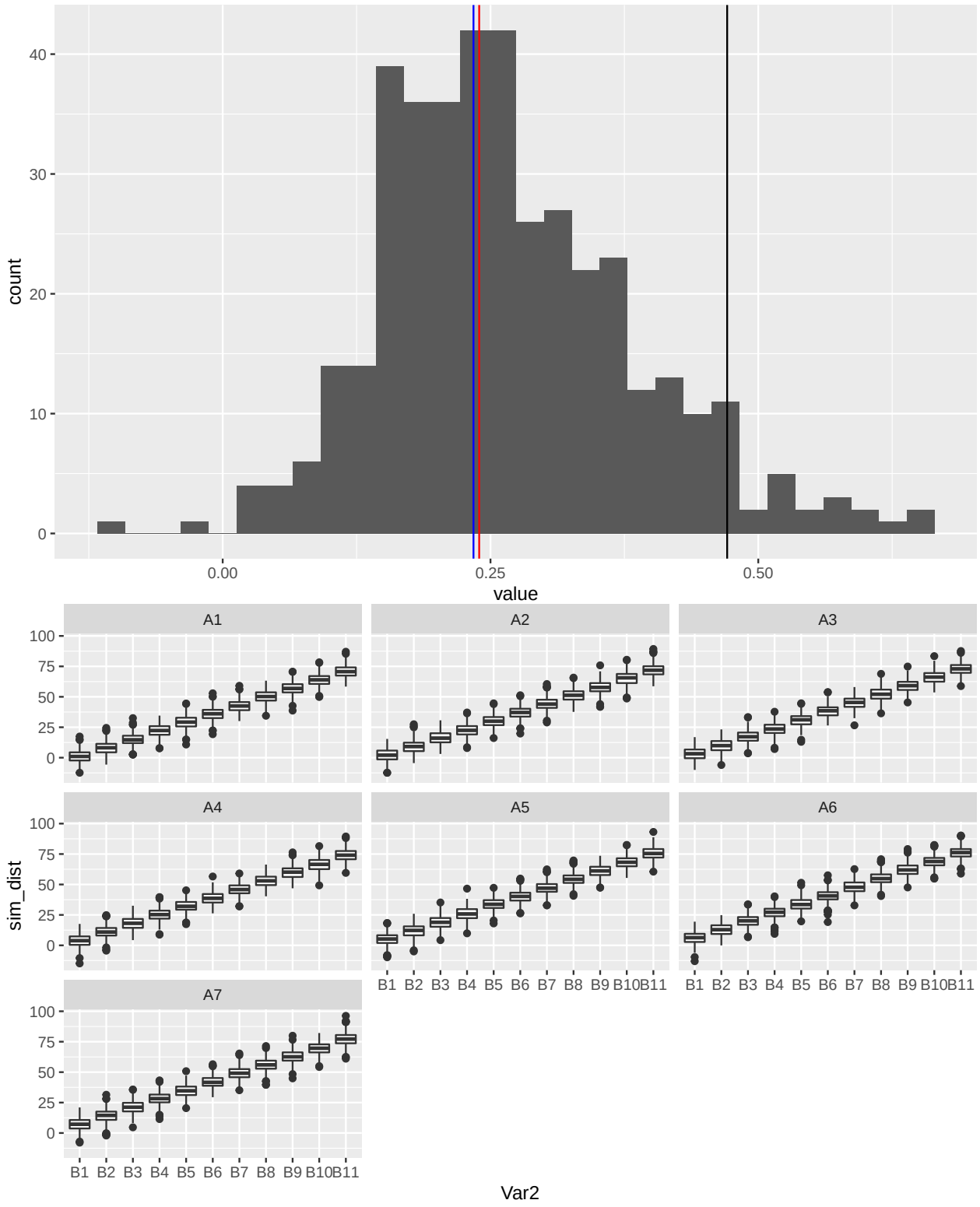
4.2 Scenario 2: Simulated different distributions for all combinations of categories for harmony pairs for few levels.

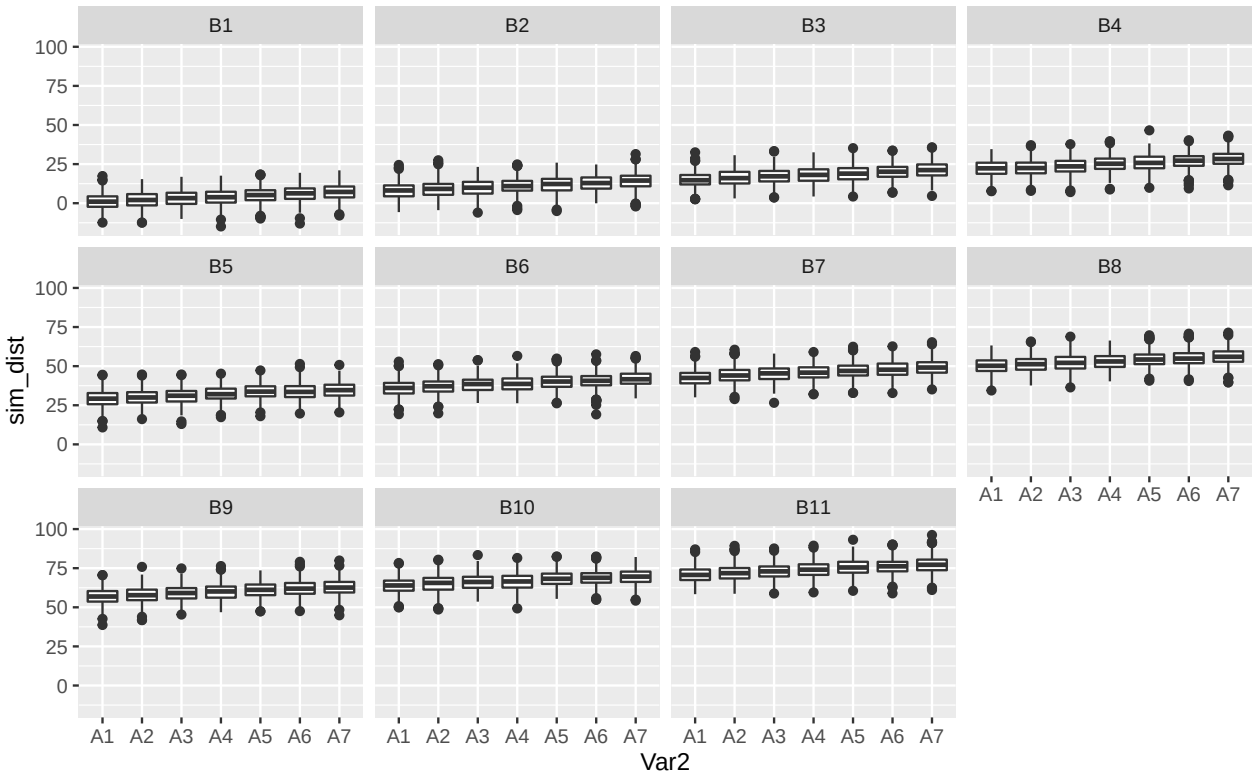




Conclusion: The test select the harmony pair for which distribution of x-axis categories are significantly different

4.3 Scenario 3: Simulated different distributions for all combinations of categories for all harmony pairs with many levels.

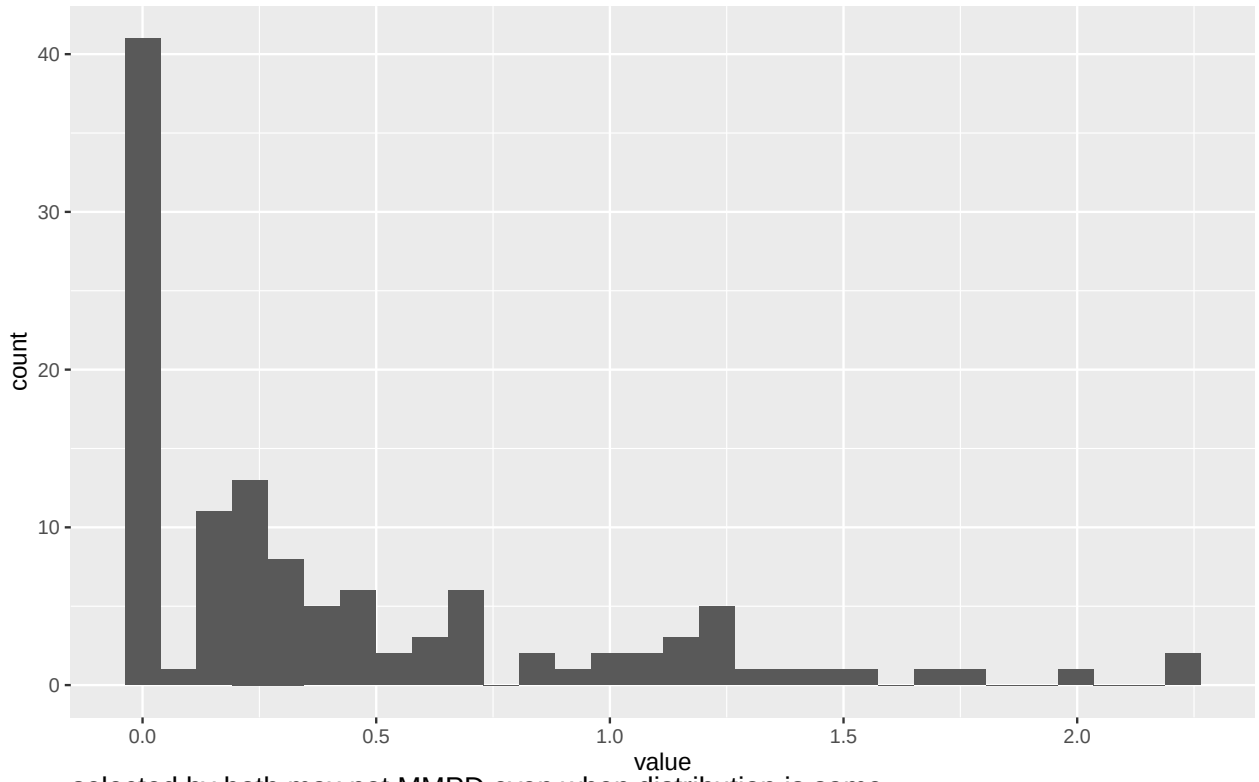




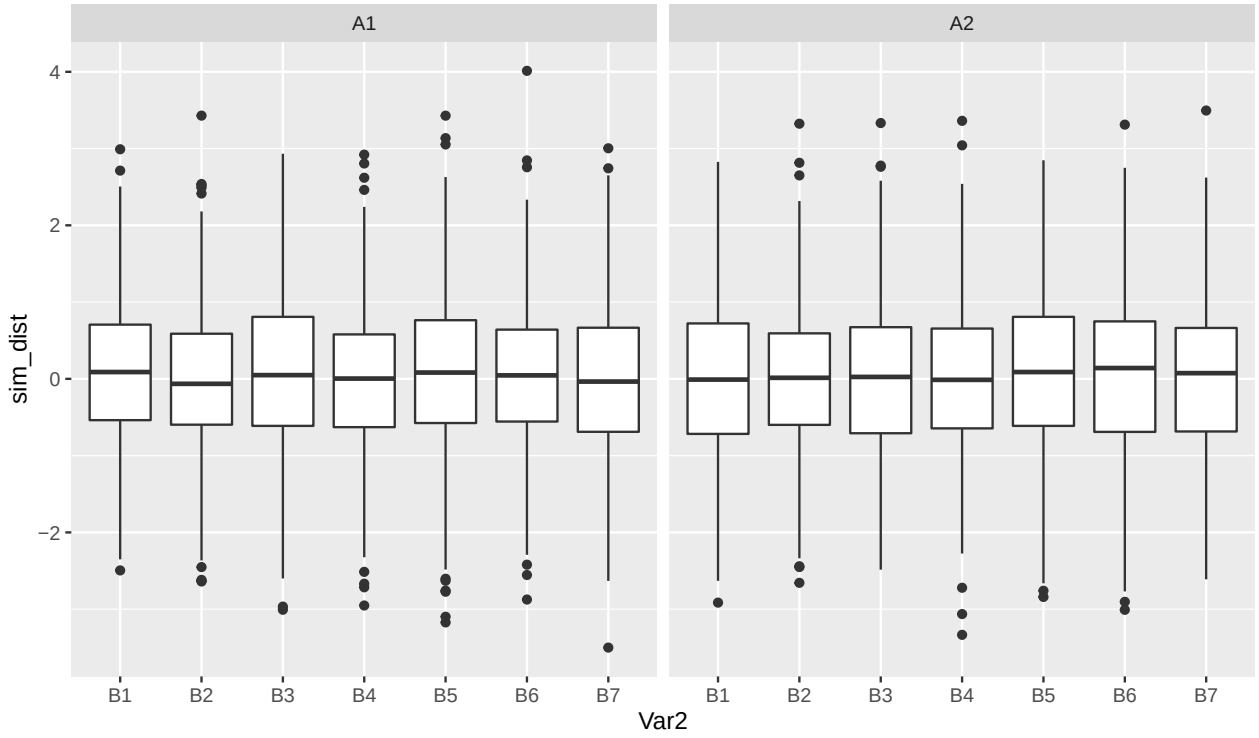
Conclusion: The test with MMPD rejects the harmony pair even for which distribution of x-axis categories are significantly different. But the test with only maximum selects the pair with varying x-axis categories.

4.4 Scenario 4: Cumulative 3 levels with 2, 7 and 11 and testing level and power

```
#> # A tibble: 120 x 1
#>     value
#>     <dbl>
#>  1 1.27
#>  2 0.871
#>  3 0.333
#>  4 0.251
#>  5 0.0224
#>  6 0.0165
#>  7 1.01
#>  8 0.938
#>  9 0.412
#> 10 0.213
#> # ... with 110 more rows
```



selected by both max not MMPD even when distribution is same



5 Levels and Power of the test statistic

Conclusion: With 3 levels the test incorrectly chooses 1 harmony pair with similar distribution. The harmony pair which is displayed.

Conclusion: The test with MMPD selects just one pair, as opposed to the test with maximum. This needs to be checked against what we expect from the test. The harmony pairs which are selected (either through MMPD or maximum) are displayed.

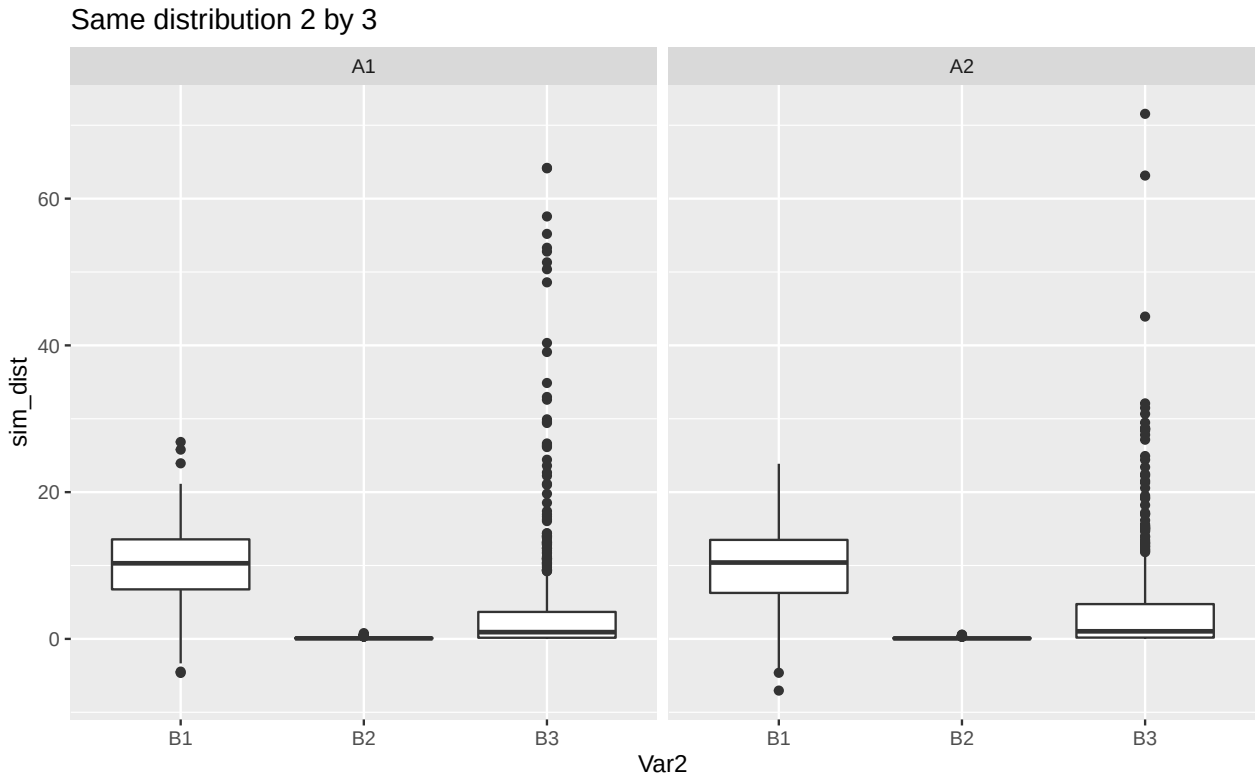
6 histograms of MMPD when levels are different

When all distributions of response variables are generated from normal, how do the histogram looks

7 Permutation test

To estimate the sampling distribution of the test statistic we need many samples generated under the null hypothesis. If the null hypothesis is true, changing the exposure would have no effect on the outcome. By randomly shuffling the exposures we can make up as many data sets as we like. If the null hypothesis is true the shuffled data sets should look like the real data, otherwise they should look different from the real data. The ranking of the real test statistic among the shuffled test statistics gives a p-value.

```
#> # A tibble: 6 x 4
#> # Groups:   Var1, Var2 [6]
#>   Var1  Var2             dist sim_dist
#>   <fct> <fct>          <dist> <list>
#> 1 A1    B1          N(10, 25) <dbl [500]>
#> 2 A2    B1          N(10, 25) <dbl [500]>
#> 3 A1    B2            Exp(10) <dbl [500]>
#> 4 A2    B2            Exp(10) <dbl [500]>
#> 5 A1    B3    Weibull(0.5, 2) <dbl [500]>
#> 6 A2    B3    Weibull(0.5, 2) <dbl [500]>
```



Haan, Laurens de, and Ana Ferreira. 2007. *Extreme Value Theory: An Introduction*. Springer Science & Business Media.