Median Maximum Pairwise Distance

Contents

1 Introduction

Take an example of a data set which are observed at fine temporal scales, like that of NYC bike usage available at https://www.citibikenyc.com/system-data. We use the `nyc_bikes` data set from the R package `tsibbledata` which takes a sample of 10 bikes for the year 2018. The `start_time` and the `stop_time` are recorded to a fineness of seconds. We can look at pair of cyclic granularities (hour_day, wknd_wday) or (week_month, day_week) to see how these periodicities interact. But there could be other pairs that are important too. How to understand which pairs are sufficient to explore given the data set without losing much information about the data.

When we need to understand the interplay of different periodicities in a high frequency temporal datasets, we have many choices to consider. In (Wang, Cook, and Hyndman 2020a) and (Wang, Cook, and Hyndman 2020b), periodicities are explored across hour of the day and day of the week or months. But calendar effects are not restricted to conventional day-of-week or month-of-year deconstructions. The paper (gravitas) describes how we can start with all possible combinations of cyclic time granularities and narrow down our search to harmonies and remove clashes without losing any perspectives about the data. 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.

To be able to dice time in all possible ways in order to have multiple perspectives about peridicities in the data, there are too many things that we should look at ideally. Exploratory analysis is detective work and it is difficult to traverse through all possible ways without any systematic approach. It is immensely useful to be able to make transition from all possible ways to only ways that could potentially be important given a situation.

Talk about Scagnostics: Tukey

2 Computing distance measure

2.1 Distance between distributions

The most common divergence measure between distributions is the Kullback-Leibler (KL) divergence introduced by Solomon Kullback and Richard Leibler in 1951. The KL divergence, denoted $D(p(x), q(x))$ is a non-symmetric measure of the difference between two probability distributions $p(x)$ and $q(x)$ and is interpreted as the amount of information lost when $q(x)$ is used to approximate $p(x)$. Although the KL divergence measures the "distance" between two distributions, it is not a distance measure since it is not symmetric. The Jensen-Shannon divergence based on the Kullback-Leibler divergence is symmetric and it always has a finite value. The square root of the Jensen-Shannon divergence is a metric, often referred to as Jensen-Shannon distance. Other common measures of distance are Hellinger distance, total variation distance and Fisher information metric.

In the context of this paper, the pairwise distances between the distributions of the measured variable are computed through Jensen-Shannon distance which is based on Kullback-Leibler divergence and is defined by,

$$JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M)$$

where $M = \frac{P+Q}{2}$ and $D(P||Q) := \int_{-\infty}^{\infty} p(x) f(\frac{p(x)}{q(x)})$ is the KL divergence between distributions $p(x)$ and $q(x)$. Probability distributions are estimated through quantiles instead of kernel density so that there is minimal dependency on selecting kernel or bandwidth.

2.2 Normalize distances

Maximum pairwise distances are 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.

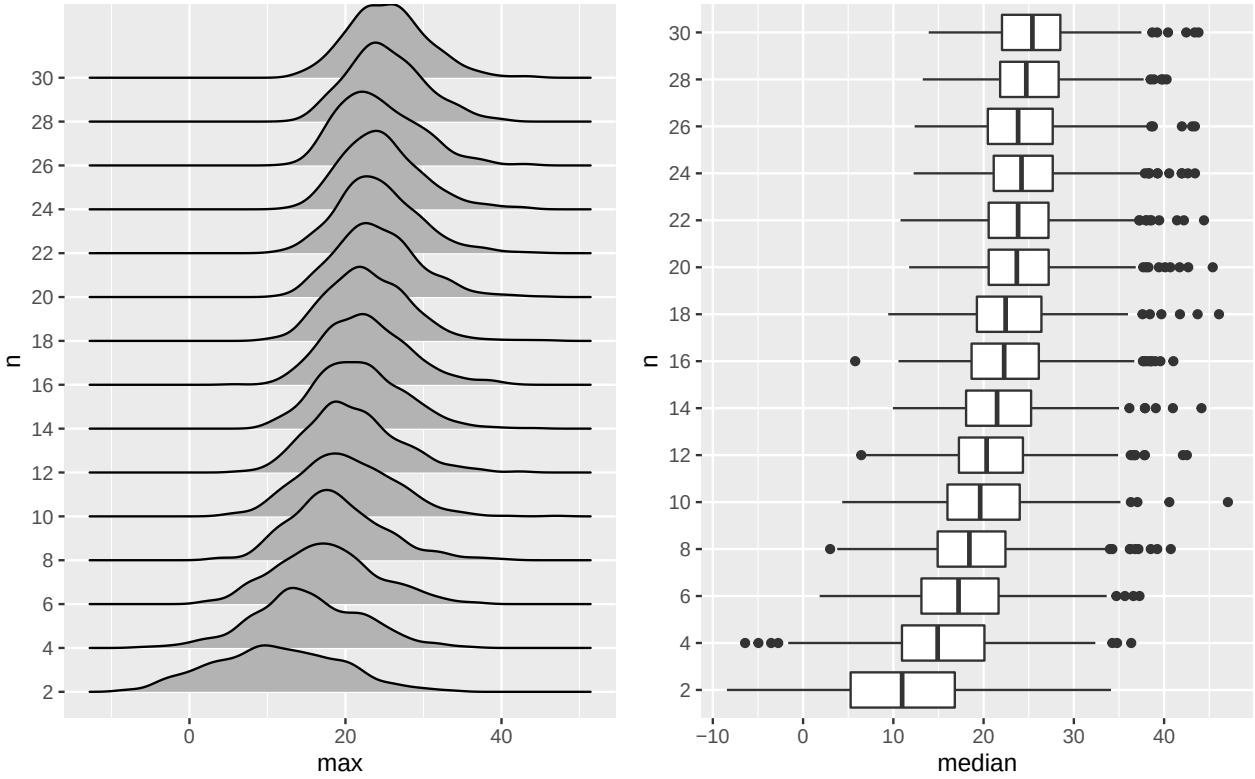
More formally, $d_1, d_2 \ldots, d_n$ be a sequence of independent and identically-distributed pairwise distances and $M_n = \max\{d_1, \ldots, d_n\}$. Then Fisher–Tippett–Gnedenko theorem (Haan and Ferreira 2007) suggests that if a sequence of pairs of real numbers (a_n, b_n) exists such that each $a_n > 0$ 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.

2.2.1 why normalize

Suppose that X_1, X_2 , ... , X_n are i.i.d. random variables with expected values $E(X_i) = \mu < \infty$ and variance $Var(X_i) = \sigma^2 < \infty$. Let $Y = max(X_1, X_2, \ldots, X_n)$.

Let $F_X(x)$ be the common distribution of the variables X_i and let $F_Y(y)$ be the corresponding distribution of Y. $F_Y(y)$ could be obtained from $F_X(x)$ simply by using: $F_Y(y) = P[(X_1 \leq y) \cap (X_2 \leq y) \cap ... \cap (X_n \leq y)] = F_X(y)^n$. For large n, the distribution of Y approaches a standard shape, which does not depend on F_X. But what about the case when n is not large enough? The distribution of maximum in that case will indeed depend on n and the underlying distribution of X. If $F_X(x)$ is the CDF of X, then $F_Y(y) = F_x(y)^n$. Suppose Φ nd ϕ are the cdf and pdf of a standard normal distribution, then $f_Y(y) = n\Phi(y)^{n-1}\phi(y)$, which depends on n. Hence, we are trying to normalise for n. Also, it depends on the underlying distribution of X, which we have assumed as normal in our case. As n grows, we can see the right tail growing, which implies that the probability that we will get a higher maximum is more. Now, for large n, we used EVT to normalise for n, that is, we brought them to the same scale without distorting the range of the distribution. But in our case, we will mostly have small n. It is important to ensure that they have the same mean and variation, for being able to compare the maximum value across n. We observe from the following graphs that our normalisation works after $n = 6$, after which the difference in mean and standard deviation flattens out a lot.

2.3 Mean and standard deviation of the distribution of maximum



2.4 Distribution of distances

2.4.1 Theoretical evidence

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.

2.4.2 Empirical evidence

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. We look at different scenarios, where observations are collected from Normal, Exponential, Chi-squared and Gumbel distribution and found the distribution of JS distances are similar, irrespective of which distribution they are drawn from.

2.4.2.1 Initial distribution of observed variables shown in plot title

2.5 Median Maximum Pairwise distance

2.5.1 Definition

2.5.2 Algorithm for computation for all harmony pairs

The algorithm employed for computing MMPD is summarized as follows:

- **Input:** Data corresponding to all harmony pairs, i.e., data sets of the form (C_i, C_j, v) $\forall i, j \in N_C$
- **Output:** MMPD (Median Maximum Pairwise Distances) measuring the average variation across different levels of C_i and C_j $\forall i, j \in N_C$

1. Fix harmony pair (C_i, C_j).
2. Fix k. Then there are L groups corresponding to level A_k of C_i.
3. Compute $m = \binom{L}{2}$ pairwise distances between distributions of L unordered levels and $m = L - 1$ pairwise distances for L ordered categories.
4. Identify maximum within the m computed distances.
5. Compute normalized maximum distance (NM) using appropriate norming constants.
6. Use Steps 1-5 to compute normalized maximum distance for $\forall k \in \{1, 2, \ldots, K\}$.
7. Compute MMPD = median $(NM_1, NM_2, \ldots, NM_K)/\log(K)$.
8. Repeat Steps 1 to 7 for all harmony pairs.

2.5.3 Bounds

This is not correct becuase MMPD should be median of standardized Gumbel distribution. So no bound?

By Lin (1991),

$$0 \leq JSD(P||Q) \leq ln(2)$$

.

Thus,

$$0 \leq MMPD \leq \frac{ln(2)}{ln(k)}$$

. Now, by asumption $k \geq 2$ and hence,

$$\frac{ln(2)}{ln(k)} : \begin{cases} 1 & if\ k = 2 \\ < 1 & if\ k \geq 2 \end{cases}$$

Thus,

$$0 \leq MMPD \leq\ 1$$

3 The statistical test

3.1 Definition

3.1.1 Algorithm for computation for all harmony pairs

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.

3.2 Null distribution

3.2.1 Normalised maximum distances follow standard Gumbel distribution

3.2.2 Limiting distribution of median of normalised maximum distances is normal

Let a continuous population be given with cdf F(x) (cumulative distribution function) and median ξ (assumed to exist uniquely). For a sample of size $2n+1$, let $\tilde{x}$ denote the sample median. The distribution of $\tilde{x}$,under certain conditions, to be asymptotically normal with mean ξ and variance $\sigma_n^2 = \frac{1}{4}[f(\xi)]^2(2n+1)$, where $f(x) = F'(x)$ is the pdf (probability density function).

3.2.3 Confidence interval of test statistic

3.3 Simulation design

Behavior of the statistic - control simulation

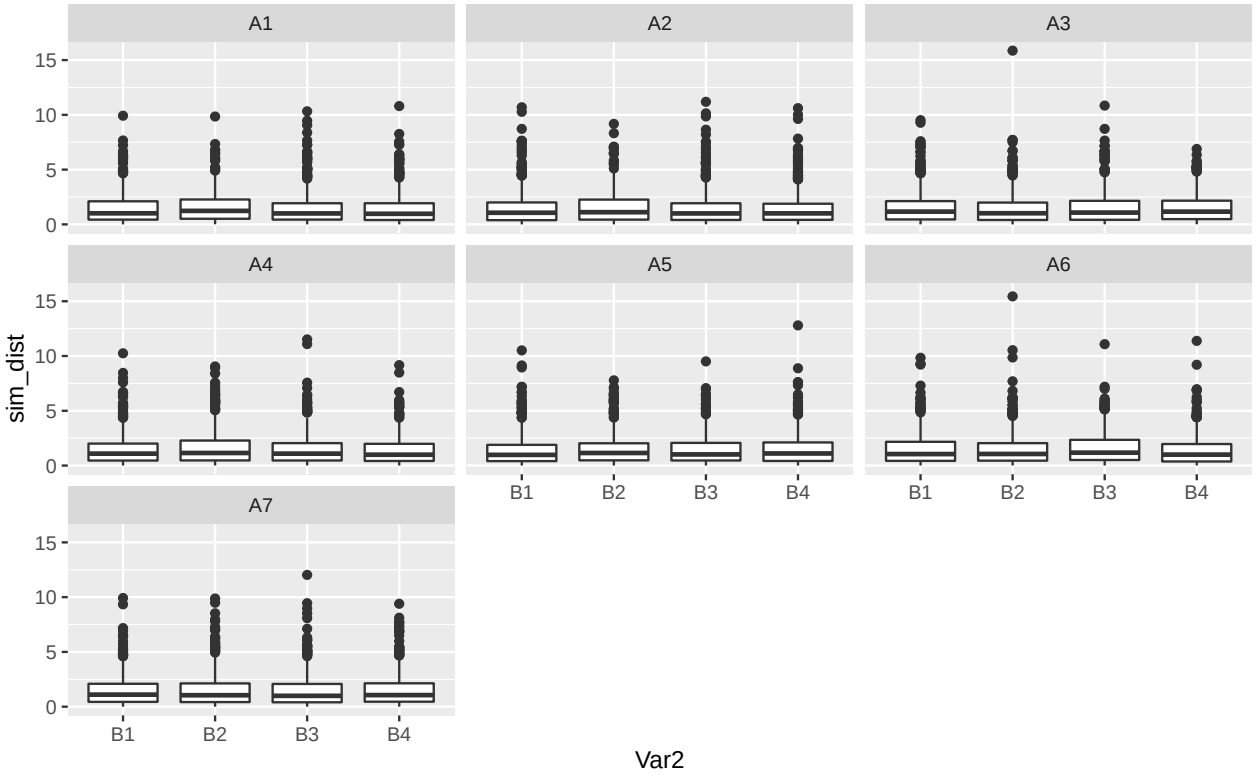
- To check if different distributions impact (simulate with different distributions but same for all levels)
- To check if x-levels are normalised (simulate with different distributions) (simulate with different x-levels)
- To check if facet-levels are normalised (simulate with different distributions) (simulate with different facet-levels for fixed x-levels)

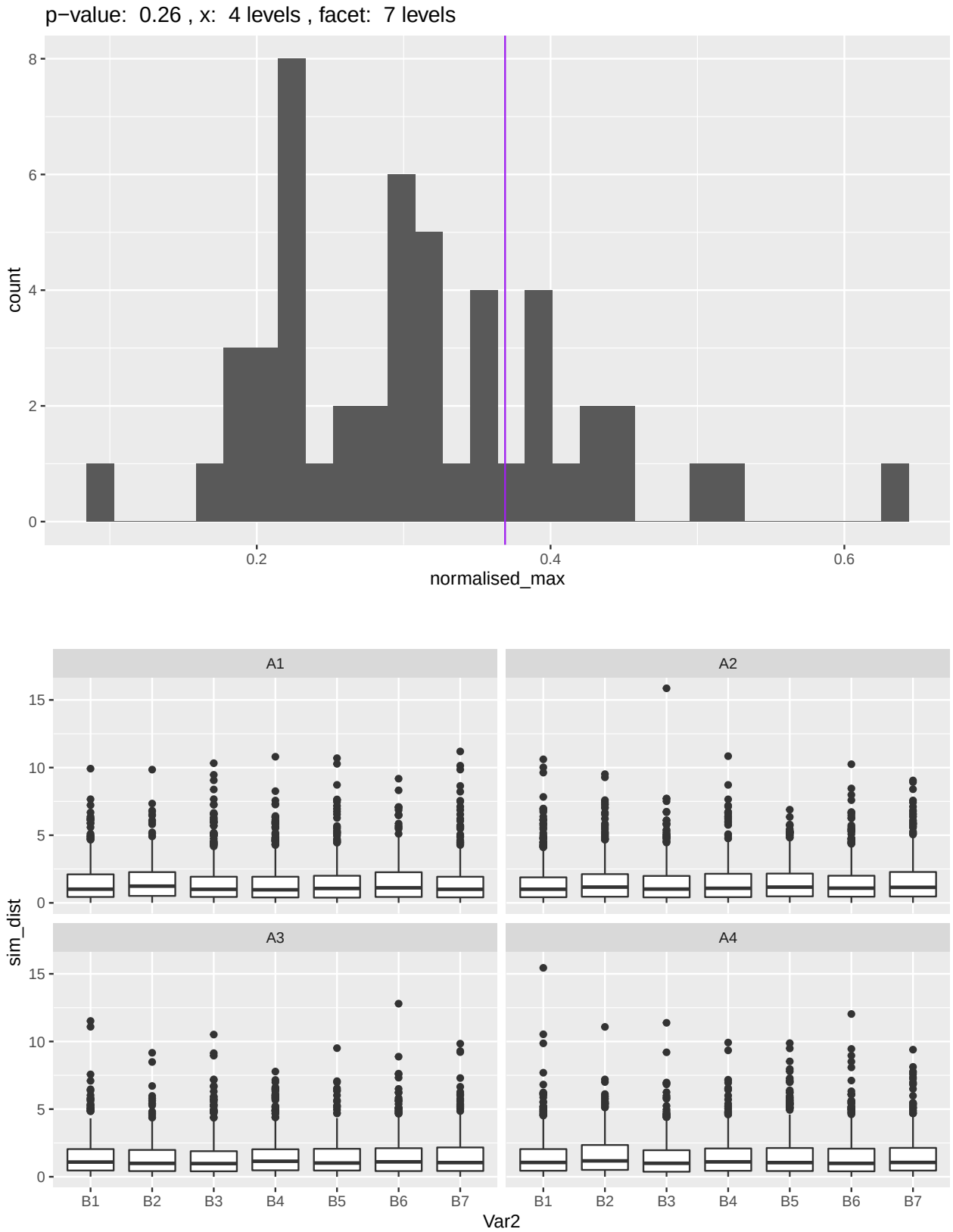
3.4 Size and power

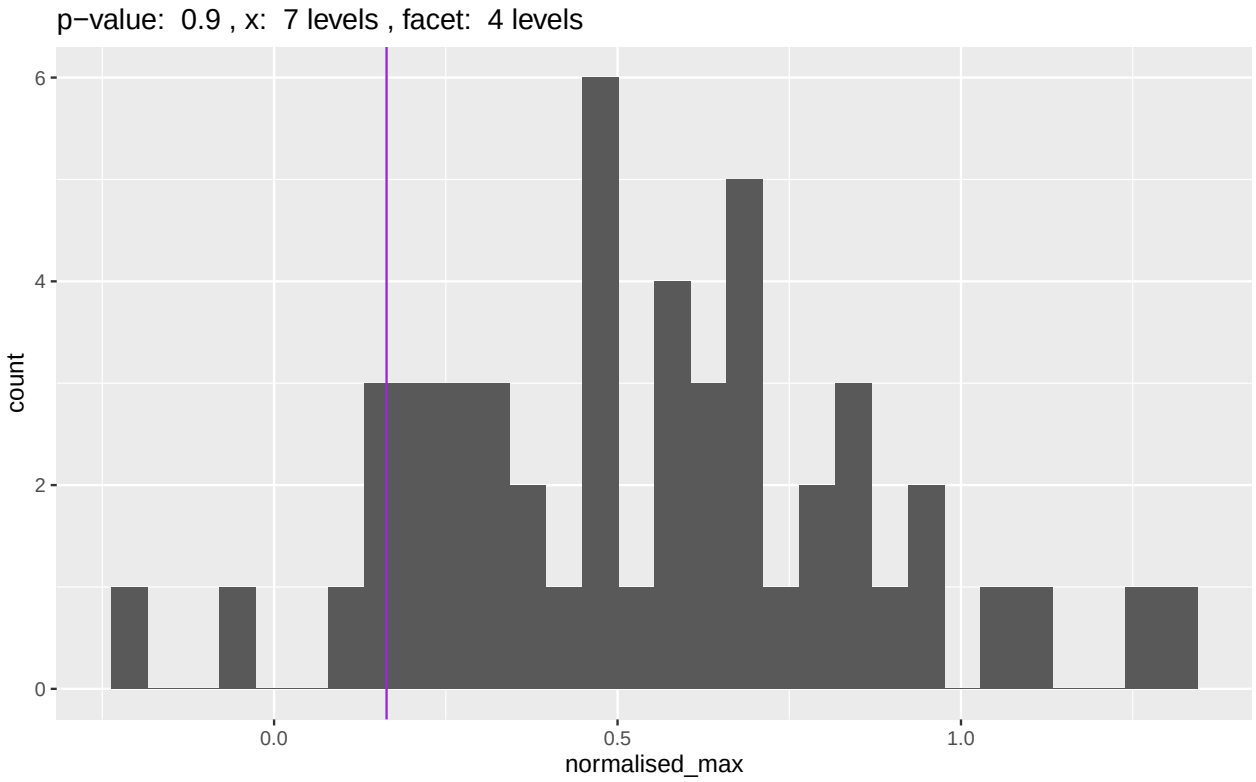
To estimate the sampling distribution of the test statistic we need many samples generated under the null hypothesis. If the null hypothesis is true, changing the exposure would have no effect on the outcome. By randomly shuffling the exposures we can make up as many data sets as we like. If the null 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.

3.4.1 Size: Simulated same distribution for all combinations of categories for all harmony pairs.

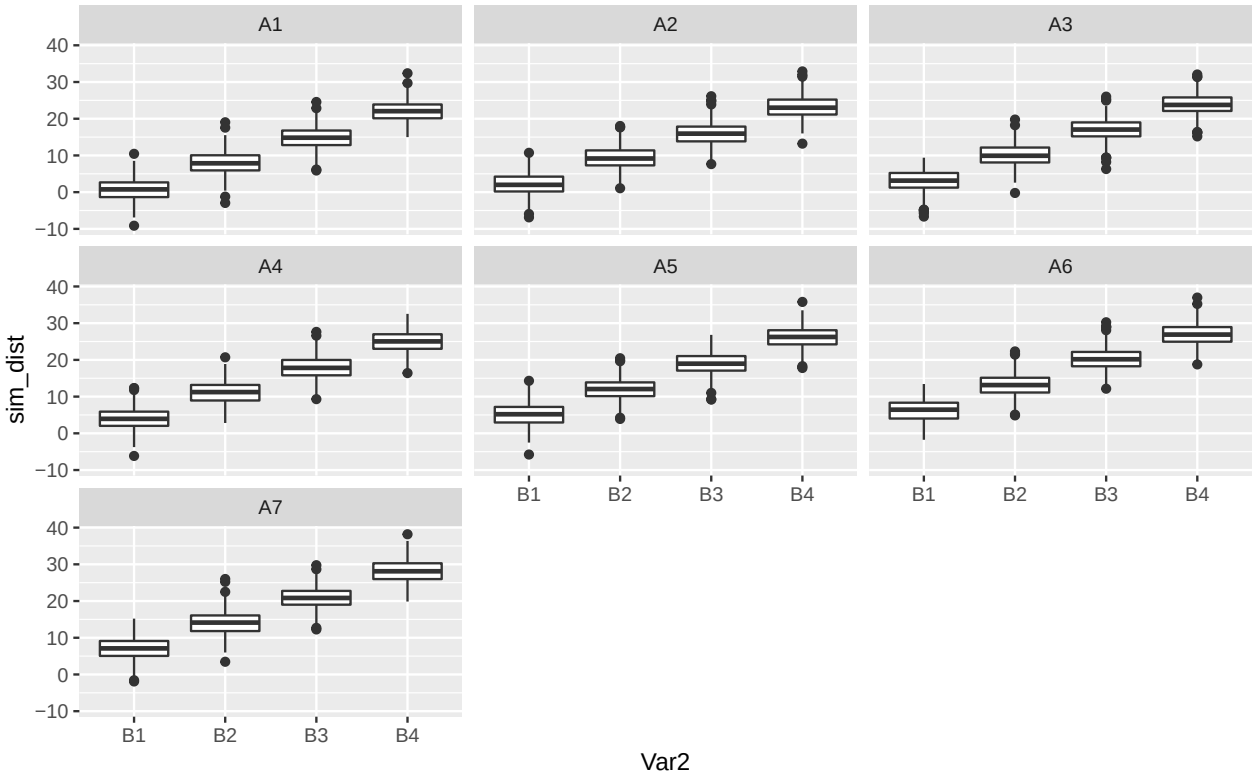
Failure to reject the null hypothesis when there is in fact no significant effect.

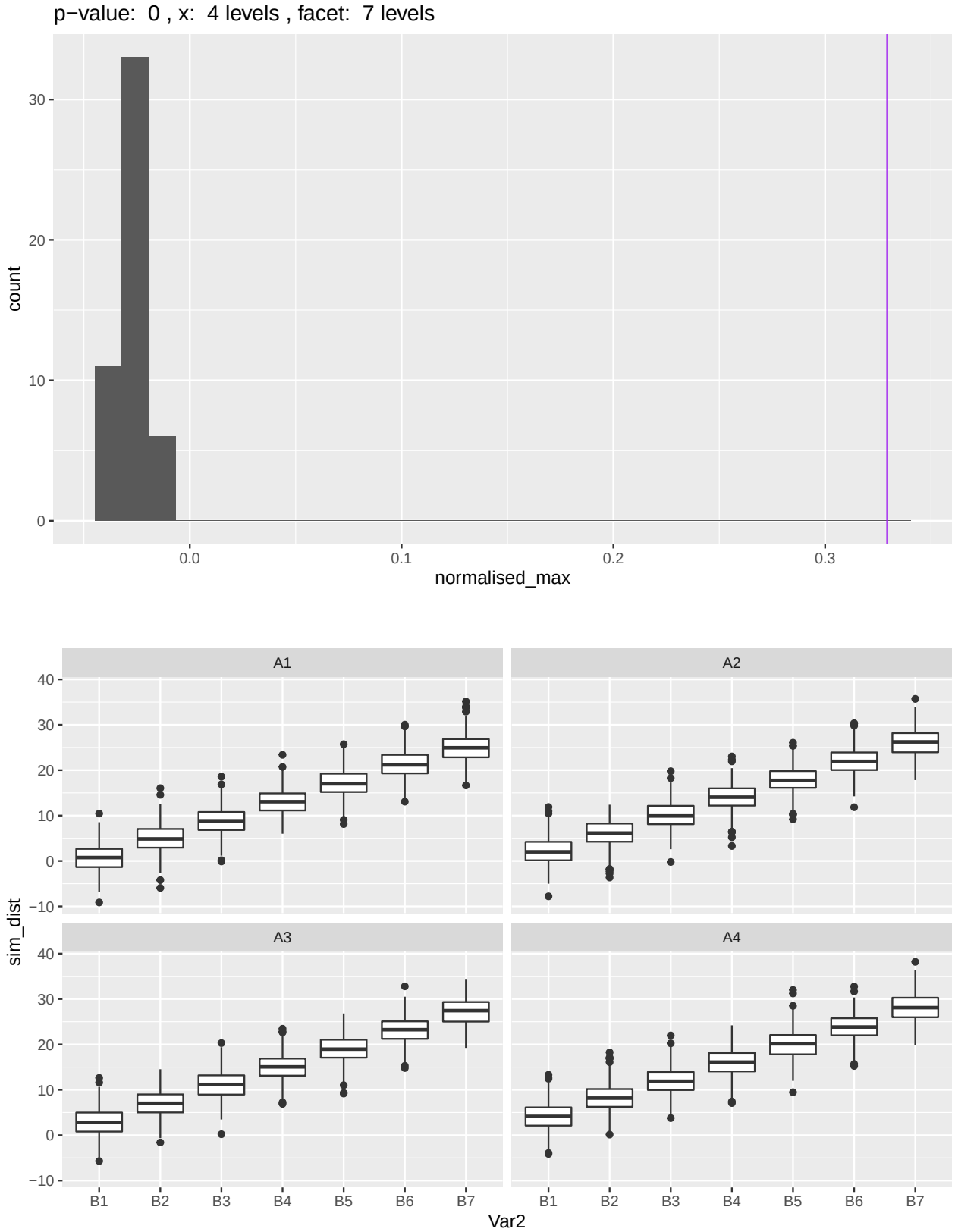


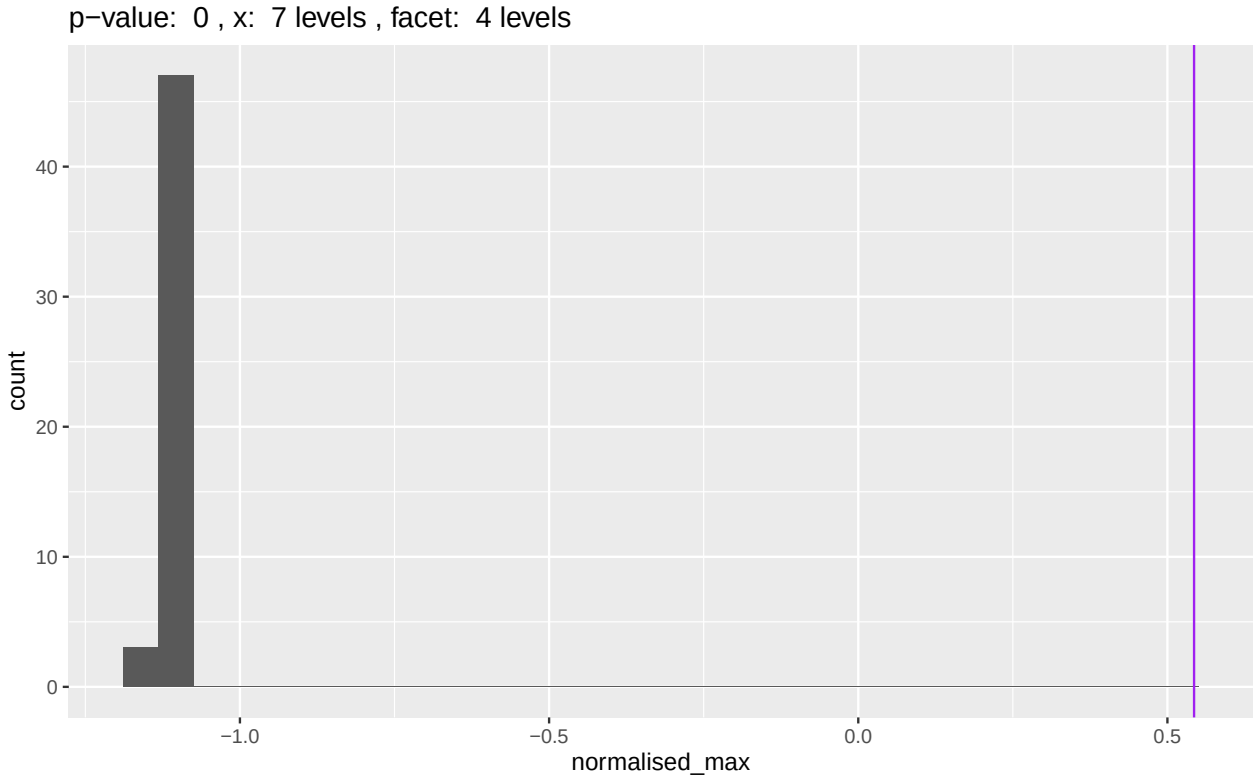

p−value: 0.26 , x: 4 levels , facet: 7 levels
count
normalised_max
A1
A2
A3
A4
sim_dist
Var2
B1
B2
B3
B4
B5
B6
B7



3.4.2 Power: Simulated same distribution for all combinations of categories for all harmony pairs.




p−value: 0 , x: 4 levels , facet: 7 levels
count
normalised_max
A1
A2
A3
A4
sim_dist
Var2
B1
B2
B3
B4
B5
B6
B7



Conclusion: The test rejects the null hypothesis if distributions are different.

3.4.3 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

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

Conclusion: The test indicates that both harmony pairs do not have significant variation.

3.4.5 Scenario 4: Simulated different distributions for all combinations of categories for all harmony pairs with many levels - very different distribution across x-axis

Conclusion: The test indicates that only the first harmony pair has significant variation.

3.4.6 Scenario 5: Simulated different distributions for all combinations of categories for all harmony pairs with many levels - very different distribution across facets

–>

–>

4 Application

Summary and discussion

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

Lin, J. 1991. "Divergence Measures Based on the Shannon Entropy." *IEEE Trans. Inf. Theory* 37 (1): 145–51.

Wang, Earo, Dianne Cook, and Rob J Hyndman. 2020a. "A New Tidy Data Structure to Support Exploration and Modeling of Temporal Data." *Journal of Computational and Graphical Statistics*. https://doi.org/10.1080/10618600.2019.1695624.

———. 2020b. "Calendar-Based Graphics for Visualizing People's Daily Schedules." *Journal of Computational and Graphical Statistics*. https://doi.org/10.1080/10618600.2020.1715226.