Implementing significance testing for the Overlapping Index using Permutation Test

Ambra Perugini, Giulia Calignano, Massimo Nucci, Livio Finos, Massimiliano Pastore

September 27, 2024

1 Overlapping Index

The overlapping index (η) is an intuitive way to define the area intesected by two or more probability density functions (Pastore & Calcagnì, 2019). In a simple way, two distributions are similar when their distribution functions overlap, and as η diminishes, the two distributions differ. The index η of two empirical distributions varies from zero – when the distributions are completely disjoint – and one – when they are completely overlapped (Pastore et al., 2018). The simple interpretation of the overlapping index (η) makes its use particularly suitable for many applications (Moravec, 1988; Viola & Wells III, 1997; Inman & Bradley Jr, 1989; Milanovic & Yitzhaki, 2002).

Assuming two probability density functions $f_A(x)$ and $f_B(x)$, the overlapping index $\eta : \mathbb{R}^n \times \mathbb{R}^n \to [0, 1]$ is formally defined in the following way:

$$\eta(A, B) = \int_{\mathbb{R}^n} min[f_A(x), f_B(x)]dx \tag{1}$$

where, in the discrete case, the integer can be replaced by summation. As previously mentioned, $\eta(A, B)$ is normalized to one and when the distributions of A and B do not have points in common, meaning that $f_A(x)$ and $f_B(x)$ are disjoint, $\eta(A, B) = 0$. This index provides an intuitive way to quantify the agreement between A and B based on their density functions (Inman & Bradley Jr, 1989).

In theory the two distributions are defined in the following way: $y_1 \sim \text{Normal}(10, 2)$ $y_2 \sim \text{Unif}(0, 20)$
The true $\eta = 0.43$.

To quickly illustrate a visual representation of the overlapping area in two given distributions we present the following example: a sample of 30 observations generated from a normal distribution with mean of 10 and standard deviation on 2 and a sample of 30 generated from a random uniform with the minimum value of 0 and the maximum of 20.

The figure 1 shows how two distributions with almost same mean could still be very different from each other with the overlapping area being $\hat{\eta} = 0.46$.

In this case, a t-test would not be able to detect such difference, as it does not take into account the different variance in the two groups. Even when using a Welch test, which does not assume equal variance, the test does is less informative ($t = 0.881$, $p = 0.384$) compared to the overlapping index.

1.1 Permutation approach

Now we will introduce another approach which does not rely on the assumptions of linear models: the permutation approach. This is a non-parametric statistical method that can be used to determine statistical significance and it is most useful when the assumptions of parametric tests are not met (Pesarin & Salmaso, 2010). What the test does is to rearrange the data in many different ways and recalculates the test statistic each time. If we are thinking about a simple mean comparison (a t-test), the data in the two groups are mixed over and over and the t-value is calculated each time. If the two groups come from the same population, mixing the labels should give similar results to the ones observed. Else, if the two groups come from different populations, mixing tags should lead to very different results. From the empirical density of the permuted values it is possible to calculate the p-value as the probability to obtain an equal or more extreme value compared to the observed one.

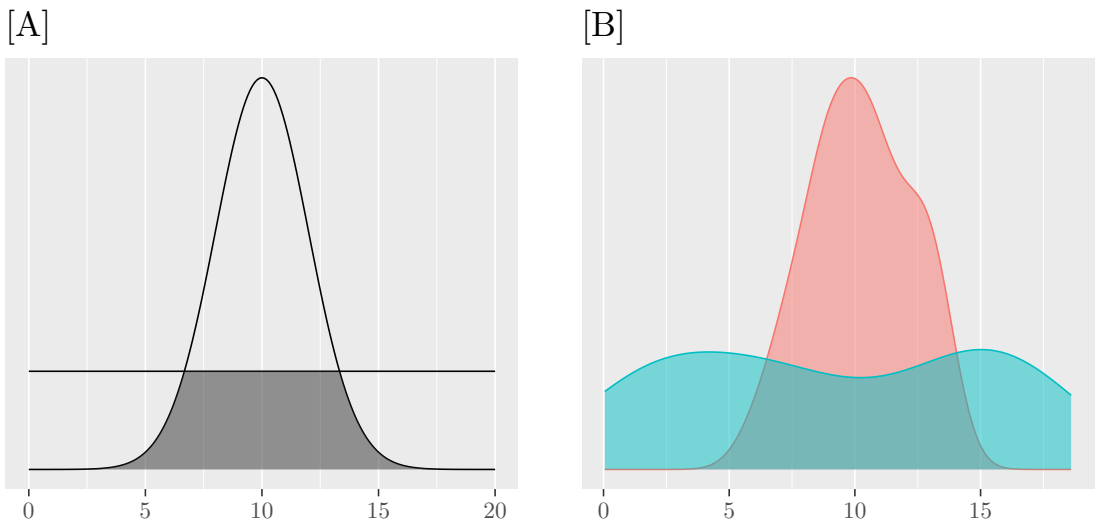

Figure 1: Comparison of a normal distribution and a uniform distribution with same mean.

1.2 Application of permutation test to the overlapping index

If we are reasoning from the perspective of Null Hypothesis Significance Testing (NHST), we should define the null hypothesis as follows: $H_0 : \eta = 1$, meaning there is no difference between the distributions of data in the population. For this reason, it is more intuitive to work with the complement of η, which is $1 - \eta = \zeta$ which is the area of non-overlap, therefore, defining the null hypothesis as $H_0 : \zeta = 0$. When testing the difference between the two distributions, we will no longer be working with η, but with the complement ζ.

Even though the overlapping index has a simple interpretation, one could argue that it does not provide information on the significance of the parameter η, therefore, we decided to implement permutation testing to offer to the ones interested a value of significance. In particular, we implemented permutations test, to give a tool that tests differences in distributions in cases where other tests' assumptions would be violated.

The algorithm estimates the value of ζ on the observed data ($\hat{\zeta}$). Then, through permutation, the values of the two groups are randomly re-assigned to the groups for B times, estimating again the new value of $\hat{\zeta}_b$. The times in which the estimate of $\hat{\zeta}_b$ on permuted data is higher than the one observed on real data is estimated ($\hat{\zeta}_b > \hat{\zeta}$) and then the found value is divided by B, returning the p-value. This approach is equivalent to the traditional parametric tests.

A typical example of data not respecting previously said assumptions is reaction times and for this purpose we present a real case of a dataset available online (citation of the OSF repository) on reaction times of word reading of high and low frequency words in English and we implement on the overlapping function the permutation test.

In the figure 2[A] are represented the densities of reaction times of word reading of high and low frequency words in English; the obtained value of $\hat{\zeta}$ is 0.56. In figure 2[B] is represented the distribution of the values of $\hat{\zeta}$ obtained with 2000 permutations; let us calculate the *p-value*:

```
sum( zperm > obsz ) / length( zperm )

[1] 0
```

The difference is statistically significant and the t test:

```
> with( xList, t.test( x1, x2 ) )

Welch Two Sample t-test

data:  x1 and x2
t = -3, df = 46, p-value = 0.002
```

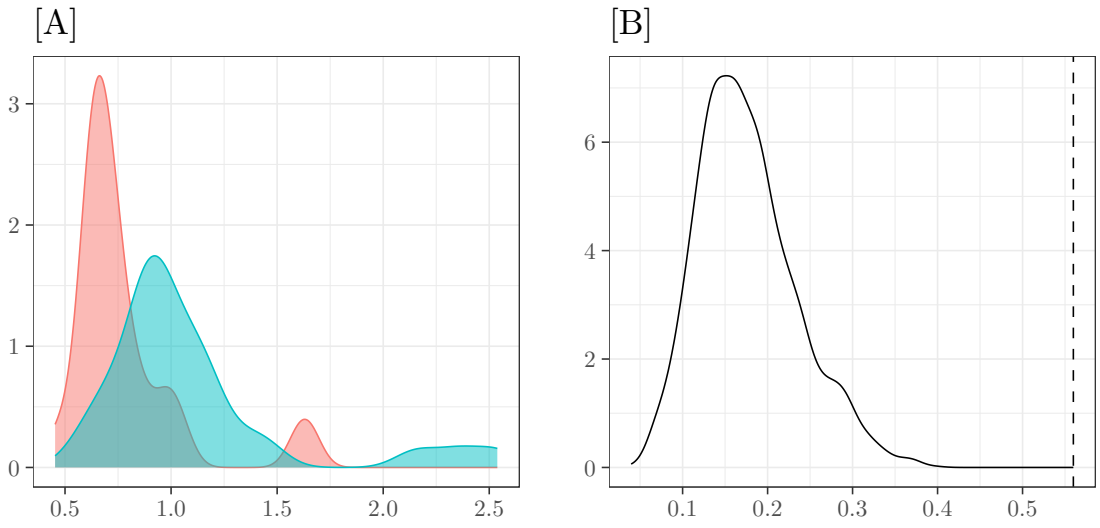

Figure 2: $\hat{\zeta} = 0.56$. [A] Distribution of reaction times of word reading of high and low frequency words in English; [B] Distribution of $\hat{\zeta}$ obtained with 2000 permutations of the data.

2 Simulation study

To evaluate the performance of the permutation test applied to the overlapping index, we performed a simulation study. The aim is to generate data for a set of scenarios distinguishing mean, variance and shape of the populations and compare the ζ perm test to other commonly used tests in terms of type I error control and power.

2.1 Data generation

In the simulation, two density distributions will be compared for many different scenarios. The first distribution will always be a normal standard distribution with $\mu = 0$ and $\sigma = 1$. To simulate data for the second distribution we use the Skew-Normal distribution (Azzalini, 1985), which is defined in the following way: given $\xi \in \mathbb{R}$, $\omega \in \mathbb{R}^+$ and $\alpha \in \mathbb{R}$, then for $y \in \mathbb{R}$ we have

$$\mathcal{SN}(y|\xi,\omega,\alpha) = \frac{1}{\omega\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{y-\xi}{\omega}\right)^2\right]\left[1 + \text{erf}\left(\alpha\left(\frac{y-\xi}{\omega\sqrt{2}}\right)\right)\right] \tag{2}$$

in which

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt$$

is the *error function*. When $\xi = 0$, $\omega = 1$ and $\alpha = 0$ the distribution is a standard normal distribution.

The parameter α determines the symmetry, ξ is the mean value and ω determines the variance. Therefore, this distribution is suitable to generate data modelling both the distance between means (the effect size), symmetry and variance.

Mean and variance of the Skew-Normal are respectively:

$$\begin{aligned} \mu &= \xi + \omega\delta\sqrt{2/\pi} \\ \sigma^2 &= \omega^2[1 - (2\delta^2)/\pi] \end{aligned} \tag{3}$$

in which $\delta = \alpha/\sqrt{1+\alpha^2}$. Based on the equations (3) we can determine the values to assign to the parameters ξ e ω in function of μ and σ with the equations:

$$\begin{aligned} \xi &= \mu - \omega\delta\sqrt{2/\pi} \\ \omega &= \sqrt{\sigma^2/[1 - (2\delta^2)/\pi]} \end{aligned} \tag{4}$$

The Skwe-Normal distribution is optimal for our purpose as it allows to have control over parameters of skewness and kurtosis, as shown in figure 3.

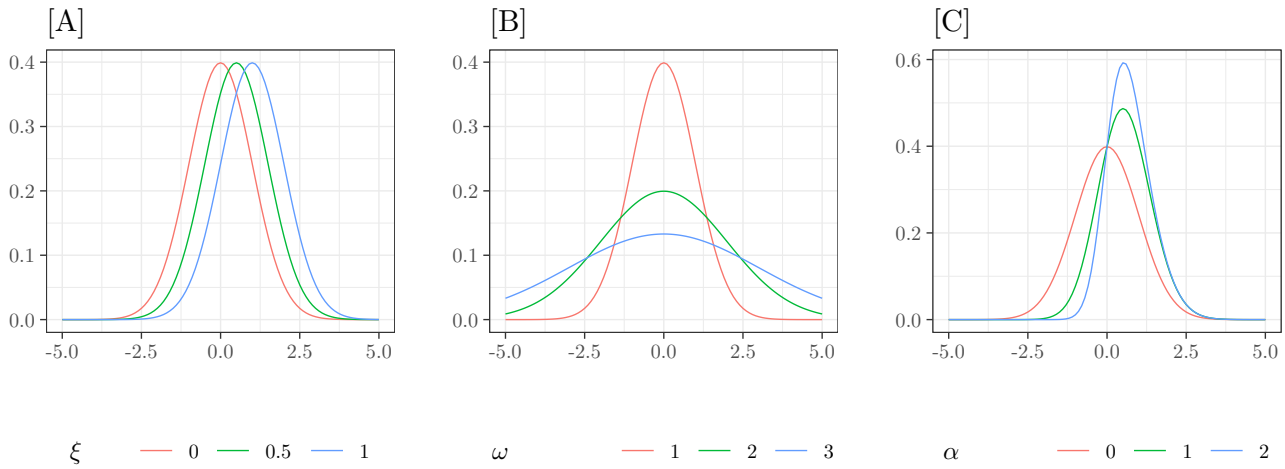

Figure 3: Skew-Normal distribution (ξ,ω,α); [A] the parameter ξ controls the mean, [B] the parameter ω the variance and [C] the parameter α the simmetry.

2.2 Simulation design

In the simulation we confront two samples extracted from a Skew-Normal, the first one is generated from $\mathcal{SN}(0, 1, 0)$, which is the Standard-Normal distribution, and the second one from $\mathcal{SN}(\xi, \omega, \alpha)$ where parameters are chosen each time based on the experimental design as follows:

- $n = (10, 20, 50, 100, 500)$; sample size, equal in the two samples;
- $\delta = (0, 0.2, 0.5, 0.8)$; mean of the second sample, which corresponds also to the difference between the two groups, the first one has always $\mu = 0$;
- $\sigma = (1, 2, 3)$; standard deviation of the second sample;
- $\alpha = (0, 2, 10)$; degree of asymmetry (skewness) of the second sample.
- N simulation: 1000 for each combination of parameters

For each of the $5 \times 4 \times 3 \times 3 = 180$ conditions we generated 1000 sets of data on which we performed the analysis.
In figure 4 are graphically represented the 36 scenarios of data generation, the black curves are the first sample, always a $\mathcal{SN}(0, 1, 0)$, and the red curves are relative to the second sample $\mathcal{SN}(\xi, \omega, \alpha)$.
For each combination $n \times \delta \times \sigma \times \alpha$, on the generated data were performed the following tests:

- t test for independent samples, assuming equal variance
- Welch test for independent samples
- Wilcoxon test for independent samples
- Permutation test on the complement of the overlapping index, $\zeta = 1 - \eta$, which therefore becomes an index of difference between groups
- F test of omogeneity of variances
- Kolmogorov-Smirnov test for confronting two distributions

2.3 Definition of Null Hypothesis

Each test relies on different assumptions and tests a specific null hypothesis.

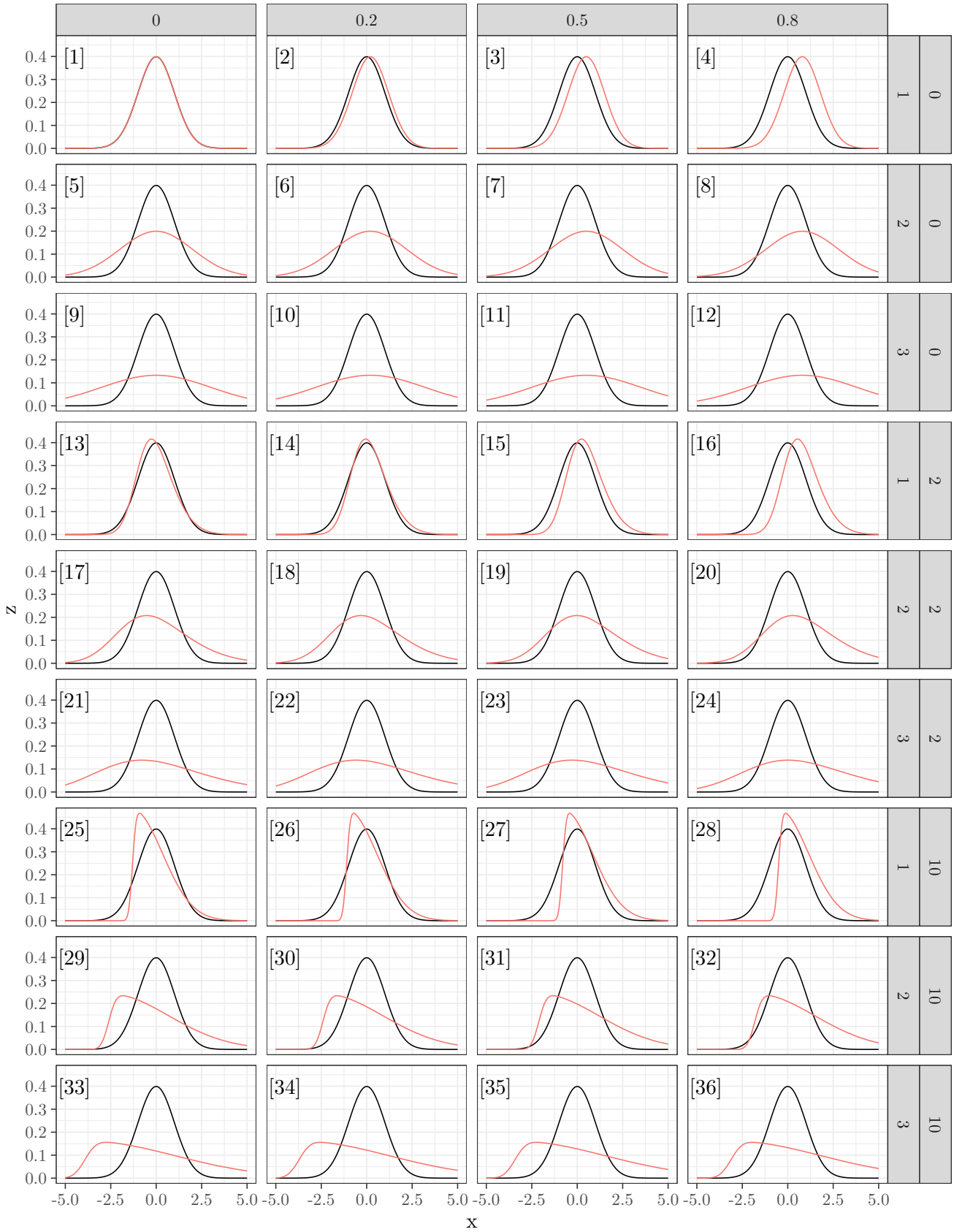

Figure 4: Generative data distributions in function of δ (column panels) and σ (row panels). The black curves are the first sample, $\mathcal{SN}(0, 1, 0)$, the red ones represent second sample.

2.3.1 t test

This is the classic case of a test for independent samples assuming equal variances:

$$H_0 : \mu_1 - \mu_2 = 0 \text{ con } \sigma_1 = \sigma_2$$

Therefore, in the scenarios from which the samples come from populations with same mean $\mathcal{SN}(0, \sigma, \alpha)$ – figure 4, panels in the first left column – type I error control is estimated, meanwhile, power is estimated for the other scenarios.

2.3.2 Welch test

This is the t test modified when homogeneity of variances is not respected:

$$H_0 : \mu_1 - \mu_2 = 0 \text{ con } \sigma_1 \neq \sigma_2$$

Control of type I error is estimated for the same scenarios as for the t test, as well as for the power.

2.3.3 Wicoxon-Mann-Whitney test

This is the test on ranks which assumes

$$H_0 : P(X_1 > X_2) = P(X_2 > X_1) = 0.5$$

in which X_1 and X_2 are the random variables representing the observations extracted from the two populations. In this case, the only scenario in which H_0 is true is in panel [1].

2.3.4 ζ permutation test

Since $\zeta = 1 - \eta$, in which η is the area of overlapping of the empirical distributions, the null hypothesis of the test is

$$H_0 : \zeta = 0$$

which implies that the data comes from the same population, or from populations with same shape (mean, variance and skewness). Therefore, the only condition in which H_0 is true is the first panel.

2.3.5 F test

This is the test of homogeneity of variances

$$H_0 : \sigma_1^2 = \sigma_2^2$$

the condition is true in all scenarios where $(\delta, 1)$, panels [1:4, 13:16, 25:28]. In those scenarios we estimate type I error, in all the others we calculate power.

2.3.6 Kolmogorov-Smirnov test

This test compares the cumulative distributions

$$H_0 : F(X_1) = F(X_2)$$

therefore, the null hypothesis should be true in panel [1], as it is for the ζ permutation test.

Taking into account those null hypothesis and assumptions, we will compute type I error by counting how many times the test is significant when the null is true, and the power by counting how many times it will be significant when H_0 is not true. Then we will consider separately the cases in which assumptions are respected and when they are not.

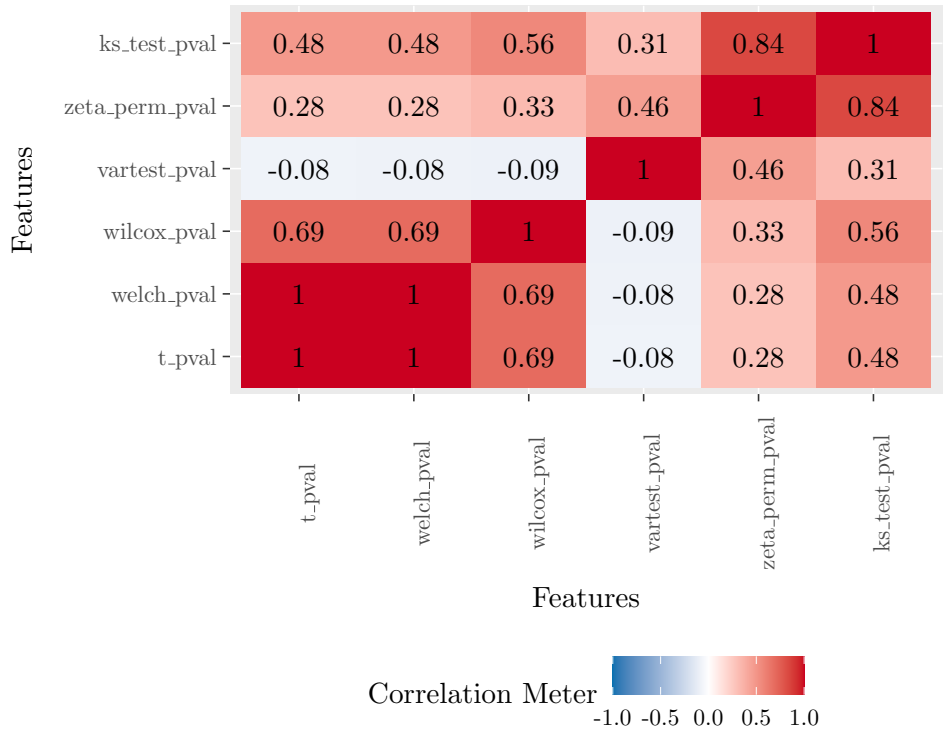

Figure 5: Correlation matrix among p-values ($N = 180000$).

3 Results

Figure 5 represents the correlation matrix between the indexes in all experimental conditions, calculated on 180000 indexes obtained from the simulation. two subgroups are clearly visible: the first group with tests on mean and ranks, and the second one on tests about the shape, the F test is not correlated with the others.

3.1 Global type I error and power

In figure 6, is represented type I error in panel A and power in panel B, for all scenarios it is evaluated when H_0 is either true or false, as different tests have different null hypothesis. Panel A shows how they all control well enough for type I error, except for the F test. The ζ perm test outperforms all other tests in terms of power, already from small sample sizes, once more, the F test is the exception, as it is a test on variance.

3.2 Assumptions and type I error and power

In the top row of figure 7 are represented type I error and power for cases in which assumptions are respected. The patter is similar to the scenario in 6 where there was no distinction for the assumptions, confirming the good control of type I error of the ζ perm test and greater power of the test in comparison to the others. As not all tests that we performed imply assumptions, we only computed type I error and power for those tests that can have the assumptions violated (t test, F test, Welch test). What emerges is a bad control of type I error of the F test.

4 Discussion

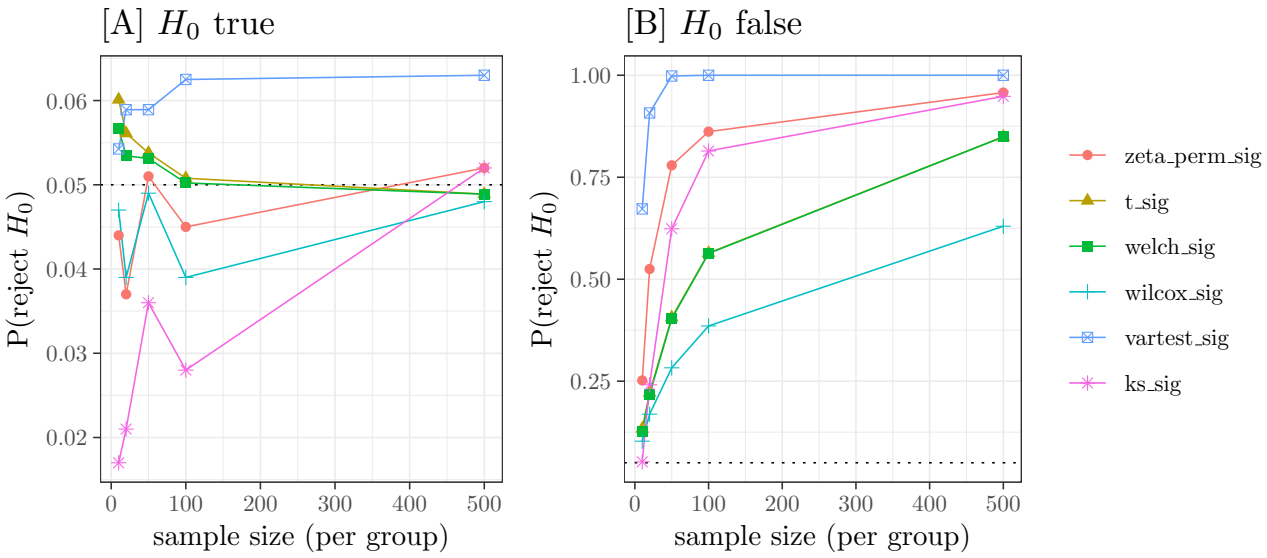

Figure 6: Control of type I error [A] and power [B] in various tests taking into account for each of them in which scenario H_0 is true or false.

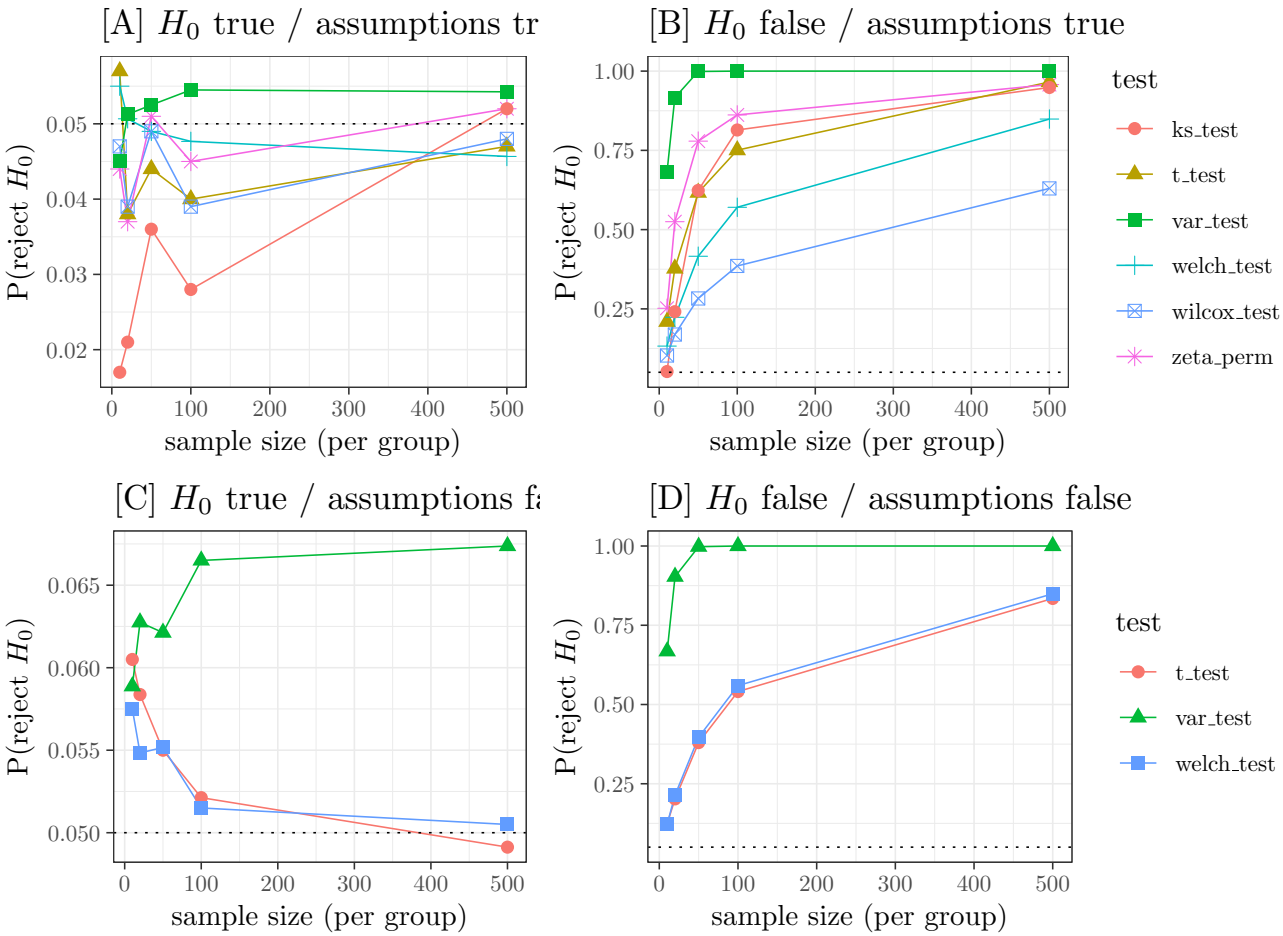

Figure 7: Control of type I error and power for scenarios in which assumptions are respected (top panels) and when they are not (bottom panels).

5 Legenda

η is the area of overlap

ζ is the area of non overlap, therefore $1 - \eta$

μ is the parameter of the mean of the normal standard

σ is the standard deviation of the normal standard

α determins the simmetry of the skew-normal

ξ is the mean value of the skew-normal

ω determines the variance of the skew-normal

δ is the difference between the two means

References

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, 171–178.

Inman, H. F., & Bradley Jr, E. L. (1989). The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics-theory and Methods*, *18*(10), 3851–3874.

Milanovic, B., & Yitzhaki, S. (2002). Decomposing world income distribution: Does the world have a middle class? *Review of income and wealth*, *48*(2), 155–178.

Moravec, H. P. (1988). Sensor fusion in certainty grids for mobile robots. *AI magazine*, *9*(2), 61–61.

Pastore, M., & Calcagnì, A. (2019). Measuring distribution similarities between samples: a distribution-free overlapping index. *Frontiers in psychology*, *10*, 1089.

Pastore, M., et al. (2018). Overlapping: a r package for estimating overlapping in empirical distributions. *Journal of Open Source Software*, *3*(32), 1023.

Pesarin, F., & Salmaso, L. (2010). The permutation testing approach: a review. *Statistica*, *70*(4), 481–509.

Viola, P., & Wells III, W. M. (1997). Alignment by maximization of mutual information. *International journal of computer vision*, *24*(2), 137–154.