How do my distributions differ? Significance testing for the Overlapping Index using Permutation Test

Giulia Calignano [1], Ambra Perugini [1], Massimo Nucci [2], Livio Finos [3], Massimiliano Pastore [1]

Abstract

Psychological research frequently relies on statistical tests targeting single distributional parameters, typically means, despite empirical data often differing in variance, skewness, or overall shape. We introduce the ζ-Overlapping test, a permutation-based inferential procedure built on the Overlapping Index, an effect size quantifying similarity between empirical distributions. The proposed approach evaluates global distributional differences without relying on parametric assumptions. Through simulations manipulating mean, variance, skewness, and sample size, we compare the ζ-Overlapping test with commonly used procedures (t, Welch, Wilcoxon–Mann–Whitney, Kolmogorov–Smirnov, and variance tests). Results show accurate Type I error control and substantially higher power than parameter-specific tests across a wide range of non-normal scenarios, with strong performance even from small sample sizes. An applied example using reaction-time data demonstrates how distributional overlap detects differences missed by mean-based analyses. Rather than replacing traditional tests, the method provides a theoretically aligned global assessment that encourages distribution-aware inference and integration of visualization and descriptive analysis into statistical workflows. The ζ-Overlapping framework supports ongoing methodological shifts in psychological science toward robust, assumption-light, and interpretable statistical reasoning.

Keywords: simulation, type I error, data visualization, reaction time, Nonparametric inference

1 Statistical testing choices in Psychology

Cognitive psychology frequently relies on comparisons between experimental conditions to infer psychological effects. Standard analyses typically focus on single summary statistics, such as mean differences, and depend on assumptions that are rarely satisfied by behavioral data. Yet cognitive measures often differ in dispersion, skewness, or tail behavior rather than central tendency alone. Researchers also often need to demonstrate that groups are comparable before interpreting experimental effects, for instance, to argue that observed differences are attributable to a manipulation rather than pre-existing sample characteristics. In practice, this comparability is usually inferred from non-significant differences in group means, sometimes accompanied by separate tests of variance, implicitly treating equality of means as evidence of overall similarity. However, groups with similar averages may still differ substantially in distributional structure, leading researchers to rely on multiple parameter-specific tests that provide only a partial assessment of similarity. In this work, we propose a distribution-based statistical test that evaluates both similarity and difference by measuring the overlap between empirical distributions. The test derives form the Overlapping Index (Gini & Livada, 1943; Pastore, Loro, Mingione, & Calcagni', 2024), which is the area intersected by two or more probability density functions, allowing to easily quantify the similarity or difference among samples (Inman & Bradley Jr, 1989). The implementation of the test through permutations permits to test global differences without relying on rigid assumptions.

The need for a distribution-level approach is especially clear in research areas where psychological data naturally depart from symmetry and homogeneity assumptions. Reaction time paradigms- such as attentional cueing e.g., Posner cueing task; (?, ?, ?), lexical decision tasks (?, ?, ?), social cueing tasks (?, ?, ?), and eye-tracking or pupillometric studies (?, ?, ?), typically yield right-skewed distributions containing outliers (?, ?, ?). In these contexts, mean-based tests may indicate no difference even when the distributions differ substantially in spread or shape. Nevertheless, even significant mean effects would fail to capture the full structure of the data (?, ?, ?; Pastore & Calcagnì, 2019).

Comparable patterns emerge in other areas of psychological science. Developmental studies often reveal changes in variability and behavioral strategies rather than average performance (?, ?, ?, ?). Similarly, in clinical contexts, comparisons of anxiety scores between clinical and control groups, or between patients and caregivers (?, ?), may involve similar mean scores alongside pronounced differences in extreme responses (?, ?). Across these domains, the inferential question shifts from whether averages differ to whether distributions themselves differ, motivating methods that directly quantify distributional similarity e.g., overlapping indices;(Pastore & Calcagnì, 2019; Inman & Bradley Jr, 1989).

To date, awareness on the importance of the generative process of the data is rising and there is a growing attention against blindly using statistical tools and analytical methods without a deep understanding of their assumptions and implications. Scheel, Tiokhin, Isager, and Lakens (2021) caution against the routine and uncritical application of hypothesis tests detached from theoretical grounding advocating for theory-aligned analysis. In this spirit, the present proposal not only aims to guide the researcher in an aware and thoughtful choice of statistical tests, but also advocates for the implementation of descriptive statistics and data visualization in the routine workflow of data analysis, encouraging a more explicit focus on the structure of empirical distributions. As originally noted by Fisher (1925), the choice of a statistical test should always be guided by the context and purpose of its application.

This contribution introduces a novel approach to statistical testing, particularly for comparing two groups/conditions. Specifically, the paper proposes a new test, called the ζ-Overlapping test, which applies the Permutation Test (Pesarin, 2001) alongside the Overlapping Index (η, Pastore & Calcagnì, 2019) to compare empirical distributions. The Overlapping index, and its proposed test, aims to promote reflection on the structure and shape of the data instead of a blind use of a new inferential tool. The application of the test should not be used as a mere alternative test with fewer assumptions but as a thoughtful choice in tailored cases (i.e. preliminary checks between groups/conditions, cases of violation of assumptions).

The ζ-Overlapping test has four main advantages. First, the core strength of the ζ-Overlapping test is that it derives from an effect size and as such is highly intuitive. By quantifying the degree of overlap two density distributions, it allows for an easy comprehension of the test. The focus of the test on the entire distribution of the data, combined with the solicitation to implement in workflow descriptive statistics and data visualization contributes to increasing awareness on the generative process of the data, as already mentioned, essential for drawing meaningful conclusions in psychological research.

The second advantage is its ability to simultaneously account for and compare the mean, variance, and shape with a single test. The ζ-Overlapping test is a global distributional test, as such one can lower the inflation of false positive rates that can arise when multiple parameter-specific tests are applied sequentially, a scenario commonly encountered in preliminary or baseline checks. This particular use of the test aligns with Scheel et al. (2021) recommendations by avoiding fishing expeditions for significant effects.

The third advantage relies in the fact that the proposed test can also be naturally extended to paired samples, making it suitable for repeated-measures designs commonly adopted in experimental psychology. In these settings, differences between conditions are evaluated within the same individuals, allowing individual variability to be explicitly preserved. The permutation procedure is implemented

by shuffling condition labels within participants rather than across groups, thus maintaining the dependency structure of the data. This approach incorporates individual differences directly into the inferential process while providing a distribution-level comparison between conditions. An example based on real data illustrating this application is presented in the article, and all code required to reproduce the analyses is openly available.

Lastly, the ζ-Overlapping test offers an alternative tool to address the assumptions issue. Given that parametric tests commonly used for statistical inference rely on strong assumptions, such as normality and homoscedasticity, assumptions that are unlikely to be met in many areas of psychological research (i.e. reaction times, accuracy, proportions), alternative methods like this one become optimal. In cases of small sample sizes, when tests are more sensitive to assumptions violations, an alternative choice that is more adherent to the empirical distribution of the data is preferable.

The remainder of this article is structured as follows. First, we introduce the concept of the Overlapping Index, providing foundational definitions and highlighting its importance. Next, we define the Permutation approach and explore its application to the Overlapping Index. With a practical application on a real case of RTs, evidence is provided for its relevance in statistical analysis. Subsequently, we present a Simulation study to illustrate the practical implications and the performance of the Overlapping index utilizing permutations. By simulating different scenarios, we compare statistical tests used in the psychological sciences with this novel approach, evaluating their control of Type I error and power. Finally, we discuss the results and provide the readers with an easy to implement workflow using the ζ-Overlapping test, offering insight into the strengths and limitations of the Permutation-based Overlapping Index and its potential applications in psychological sciences.

2 Methods

2.1 Overlapping Index

The Overlapping Index (η) is an intuitive way to define the area intesected by two or more empirical 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 η index varies from zero – when the distributions are completely disjoint – and one – when they are completely overlapped (Pastore, 2018). The simple interpretation of the overlapping index (η) makes its use particularly suitable for many applications (e.g. Jensen & Sanner, 2021; Garofalo, Giovagnoli, Orsoni, Starita, & Benassi, 2022; Schuetze & Yan, 2023; Karrobi et al., 2023; Sirbiladze, Midodashvili, & Manjafarashvili, 2024; Habibi, Achour, Bounaceur, Benaradj, & Aulagnier, 2024; Ricote et al., 2024; Rossi et al., 2024; Einbeck, Coolen-Maturi, Uwimpuhwe, & Singh, 2024; Wachendörfer & Oeberst, 2024; Rohrbach, 2024; Hawkins et al., 2024; Upadhayay, Granger, & Collins, 2024; Conversano, Frigau, & Contu, 2024; Pietrabissa et al., 2024; Nougaret et al., 2024; Greene, Guitard, Forsberg, Cowan, & Naveh-Benjamin, 2024).

More formally, 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 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).

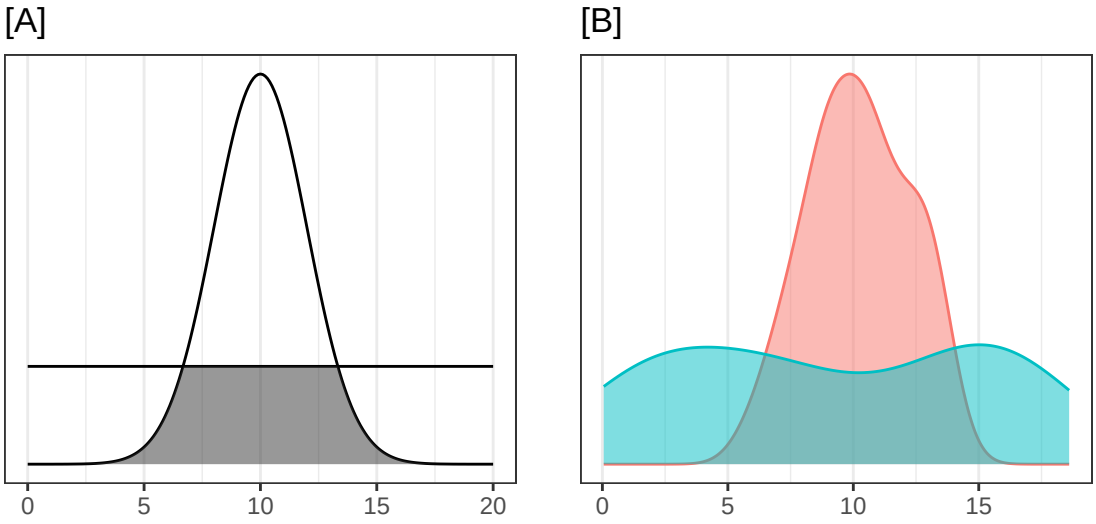

Figure 1: Comparison of a normal distribution and a uniform distribution with same mean, 10, and different variances, 4 and 33.3, respectively.

To quickly illustrate an example of the overlapping area, in figure 1 are represented two different empirical densities. In panel [A], are depicted two density distributions, a Normal(10,2) and a Uniform(0,20); note that the two distributions have the same mean (10), but different variance, 4 and 33.3 respectively. In the panel [B] are represented the empirical densities of two random samples of 30 observations drawn from the two populations specified as in panel [A]; the estimated overlapping area being $\eta = 0.46$.

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, the t-test focuses on mean differences, therefore correctly does not rejects the null hypothesis, even though the degree of similarity of the two densities is only 46%, in other words, the difference is about 54%. Moreover, we remind that the t-test in this case is far from ideal as the two distributions have different variances.

Of note, although the ζ-Overlapping test is nonparametric in that it does not assume any specific parametric form for the underlying distributions and relies on permutation for inference, it does require estimating empirical density functions using kernel density estimation (KDE). Like all nonparametric estimators, KDE introduces smoothing parameters (e.g., bandwidth) that can affect the precision of the overlap estimate. While standard bandwidth rules (e.g., Silverman's rule) work well in many scenarios, special care is required in cases of bounded data, where standard kernels may yield biased estimates near the boundaries. In such cases, boundary-corrected KDE methods or data transformations (e.g., logit for proportion data) can be used to improve accuracy. We note that KDE is widely accepted in psychological research (e.g., for modeling reaction time distributions). Nonetheless, researchers should be mindful of KDE characteristics when applying the method in bounded or sparse data contexts.

2.2 Permutation approach

Permutation tests, also known as randomization tests, are a class of nonparametric statistical significance tests. The concept dates back to the work of R.A. Fisher in the 1930s, in particular his book *The Design of Experiments* (Fisher, 1935). The theoretical foundations were further developed by E.J.G. Pitman in his seminal papers of 1937 (Pitman, 1937) and 1938 (Pitman, 1938). The basic principle of permutation testing is based on the idea of rearranging observed data to generate a null distribution.

This approach assumes that if the null hypothesis is true, then all possible arrangements of the data are equally likely, i.e., each permuted sample has the same probability as the observed one. By resampling the data, we can obtain the distribution of the test statistic under the null hypothesis without making any assumptions about the underlying data generating process. This is particularly valuable when dealing with small sample sizes or when the assumptions of parametric tests are not met. The observed test statistic is then compared to this empirically derived null distribution to determine the probability of obtaining such a result by chance alone (Pesarin, 2001).

The permutation approach allows the adoption of any test statistic chosen by the user, including statistics designed for paired samples. In the case of paired data, observations are not exchanged between participants; instead, condition labels are permuted within each participant. This means that the relationship between paired observations is preserved while testing whether the observed difference between conditions could arise by chance. As a result, the procedure naturally accounts for individual differences while maintaining the logic of permutation-based inference. For example, if we are thinking about comparing the means of the two manipulation checks within a sample, we could choose a t-test statistic for repeated measures; the data in the two conditions are permuted and the t value is calculated each time. If the two conditions come from the same population, the t-statistic computed on the observed data should be close to 0; the t-statistic computed on randomly permuted data will also give values close to zero. Therefore, the randomly generated test statistic and the observed one have the same – nonparametric – distribution. Otherwise, if the two groups come from populations with different means, the t-statistic computed on the observed data will be far from zero, while the t-statistic computed on the permuted data will be around zero.

The p-value is the probability of obtaining an equal or more extreme t-statistic compared to the observed one:

$$p = \frac{(\#_{b=1}^{B} |t_b| \geq |t|) + 1}{B + 1} \tag{2}$$

where B is the number of random permutations, t is the t-statistic computed on the observed data, t_b are those computed on the permuted data.

The test will have power – i.e., the probability of getting a $p \leq \alpha$ when the two conditions or two samples are really different – very close to the parametric t-statistic, and it will retain control over false positives even when the assumptions of normality are not met.

It is important to note that the choice of which t-statistic to use is a user choice; different test statistics (e.g., difference of mean ranks, Kolmogorov-Smirnov, etc.) will produce tests with different power. For example, if the two samples differ only in their variability and not in their mean, the permutation test based on the t-statistic will have little or no power to detect that the two samples come from different populations. In this direction, the present paper proposes to use the Overlapping Index as a test statistic that results to be powerful under a wide range of differences in distributions.

Remark 1 The choice to add a +1 in the numerator and denominator is a choice supported by many authors (Phipson & Smyth, 2010; Hemerik & Goeman, 2018) and ensures that the probability of false positives is less than or equal to α.

Remark 2 As one can understand, the p-value may change depending on the permutations that are drawn. By increasing the number of permutations B, the results will change less and less. Since the number of possible permutations is finite, it is preferable, if possible, to explore the set of them (i.e., to compute the statistics on all possible permutations of the data). This set of all possible rearrangements of the data is, in fact, the orbit of the sample that allows us to compute the exact p-value - i.e., the exact probability of observing a test statistic that is as extreme or more extreme than that observed

in the data. In this case, $B = \binom{n}{n_1} = \frac{n!}{n_1!(n-n_1)!}$ and the p-value formula reduces to

$$p = \frac{(\#_{b=1}^{B}|t_b| \geq |t|)}{B}$$

since the test statistic computed on the observed data is certainly in this set.

2.3 Application of permutation test to the Overlapping Index

Even though the Overlapping Index has a simple interpretation, it does not provide information about the statistical significance of η. Therefore, we implemented permutation testing to provide a measure of statistical significance. In particular, we implemented permutation testing, to give a tool that tests differences in distributions without assumptions, offering a valid alternative in cases in which traditional assumptions are not met.

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 that there is complete overlap between the theoretical densities in the two populations from which we sample the data. 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 $H_0 : \zeta = 0$, once again meaning that there is no difference between the densities of the two populations. Obviously, this does not change the results, but only the way in which they are interpreted. When testing the difference between the two distributions, we will no longer be working with η, but with the complement ζ.

The algorithm estimates the value of ζ on the observed data ($\hat{\zeta}$). Then, through permutation, the observed 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 or equal than the one observed on real data is estimated ($\hat{\zeta}_b \geq \hat{\zeta}$) and then the found value is divided by B, returning the p-value.

2.4 Illustrative example with real data

To show a realistic application and possible workflow we present a real case of a dataset available online (Oksuz & Rebuschat, 2024)[1] on reaction times of word reading of high and low frequency words in English. For the applied example we selected two conditions of word reading of high and low frequency words in English of 29 participants.

2.4.1 Step 1: descriptive statistics and data visualization

The first step is to plot the two conditions/groups (panel [A] figure 2). This can easily be done with the `overlapping` R package (Pastore et al., 2024). First the user has to upload the package in the R environment, then a list object is created with the two vectors of values of the two groups/conditions, subsequently the function `overlap` is used allowing to easily plot densities by setting `plot = T`. Here a quick illustration of code:

```
library( overlapping )
yList <- list( y1 = y1, y2 = y2 )
overlap( yList, plot = TRUE )
```

Simultaneously, calculate descriptive statistics (see table 1). The two conditions have means of 0.86 and 0.77, variance of 0.07 for both conditions, and skewness of 1.47 and 2.2, for Low and High frequency

[1] Link to the dataset: https://osf.io/muwjz/files/ubtyq

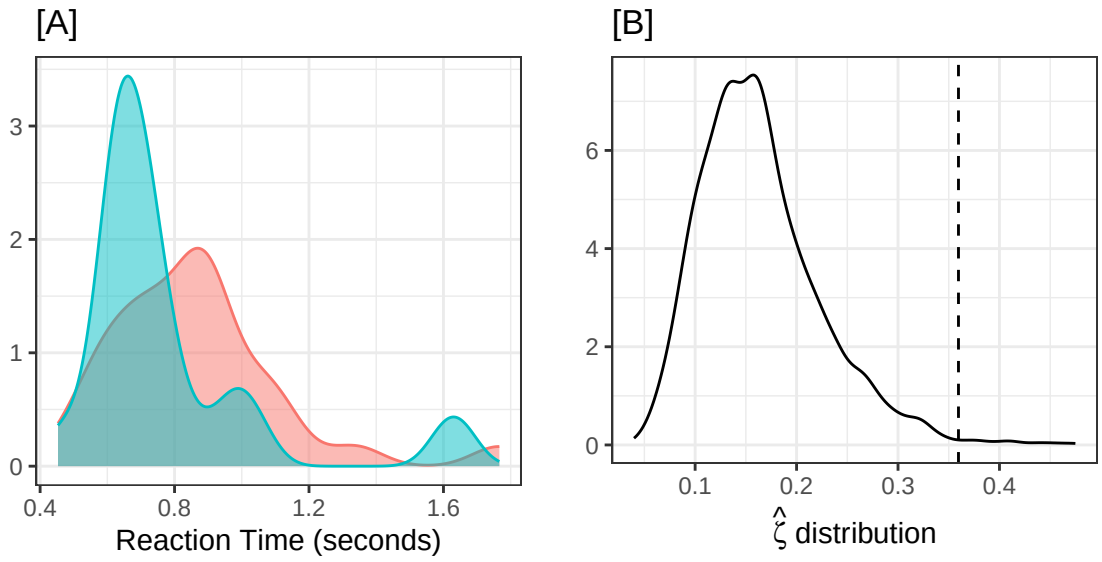

Figure 2: [A] Distribution of reaction times of word reading of high (blue) and low (red) frequency words in English. The overlapping area is $\hat{\eta} = 0.64$, corresponding to a non-overlapping area of $\hat{\zeta} = 1 - \hat{\eta} = 0.36$; [B] Distribution of $\hat{\zeta}$ obtained with 3000 permutations of the data.

	Low frequency	High frequency
Mean	0.86	0.77
Median	0.87	0.69
Variance	0.07	0.07
Skewness	1.47	2.20
Kurtosis	6.25	7.49

Table 1: Descriptive statistics for low and high frequency words

respectively. The overlapping area between the two distributions is $\eta = 0.64$, and consequently $\zeta = 1 - 0.64 = 0.36$. Already from this information, one can have insight on the characteristics of the data.

2.4.2 Step 2: Significance testing

After visualizing the data and carefully evaluating descriptive statistics, one can move to hypothesis testing. In the case that a global test is appropriate to answer the research question, or it is a case of small sample size and assumptions violation, the performance of a statistical test on the Overlapping Index is appropriate. To perform the ζ-Overlapping test, the procedure is the same as to use the `overlap` function, with the only difference that the function `perm.test` is used instead, as shown below:

```
perm.test( yList, paired = TRUE )
```

The function will give the following output:

```
$Zobs
[1] 0.36
```

```
$pval
[1] 0.008
```

`Zobs` is the observed ζ and `pval` is the p-value obtained through permutations.

In figure 2[B] is represented the distribution of the values of $\hat{\zeta}$ obtained with 3000 permutations; the *p-value* is calculated as follows:

$$p = \frac{(\#\hat{\zeta}_b \geq \hat{\zeta}) + 1}{B + 1} = \frac{24}{3001}$$

Using the permutations to obtain a p-value on ζ, gives a $p < .01$. Based on this test, we can conclude that there is a statistically significant difference between the two distributions, with an area of non-overlap equal to 0.36. Instead, performing a t-test gives a non significant result: $t(28) = 1.46$, $p = .16$.

This result suggests that the overlapping method has detected differences that the previous t-tests did not identify, highlighting the potential sensitivity of this approach.

In our example, we believe that no additional testing is needed, as the graphical representation, the descriptive statistics and the ζ-Overlapping test give a clear insight of the difference between the two conditions, with a $\zeta = 0.36$ and a clear delay in reaction time of the Low frequency condition.

Similarly, if two groups were not to differ significantly, especially with a small sample, the researcher should have focused on the the effect size and reason if the area of non-overlap could be considered reasonably a trivial difference (based on the field of study and the specific effect) or if more data was needed to reach a meaningful conclusion. One can not simply reject H_0 and should carefully evaluate the effect size, as absence of significance does not mean absence of an effect. In this case, reasoning on the smallest effect size of interest (SESOI) and running a sensitivity analysis (ideally a priori power analysis should be run before the data collection) can be beneficial for the researcher.

This approach highlights the importance of visualizing data and stresses out the invaluable insight offered by descriptive statistics (Wilkinson, 1999; Tay, Parrigon, Huang, & LeBreton, 2016; Pastore, Lionetti, & Altoè, 2017). However, in case of clear differences raising in Step 1 and confirmed in Step 2, one can decide to move forward with Step 3.

2.4.3 Step 3: Tailored parameter specific testing

Optional for deeper inquiry, this phase encourages researchers to openly weigh theoretical hypotheses against data characteristics, selecting tools (such as independent t-tests/Welch for mean differences, Levene's/F for variances, or skewness/kurtosis tests via Kolmogorov-Smirnov) in a critically informed way. Data quality and structure guide these choices, with clear labeling of confirmatory vs. exploratory elements to address type I error concerns transparently (cf. p-hacking simulations (Stefan & Schönbrodt, 2023)). Such an open, layered approach unpacks ζ differences effectively, balancing rigor with adaptability.

We also believe that Step 2 can be substituted directly by Step 3 if the researcher has a clear parameter specific hypothesis and descriptive statistics support such choice by not showing clear assumptions violation.

2.5 The 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.

2.6 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

$$\mathcal{SN}(y|\xi,\omega,\alpha) = \frac{1}{\omega\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y-\xi}{\omega}\right)^2\right]\left[1 + \text{erf}\left(\alpha\left(\frac{y-\xi}{\omega\sqrt{2}}\right)\right)\right] \tag{3}$$

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.

ξ is the location parameter, ω is the scale parameter and α is related to the skewness of the distribution. Therefore, this distribution is suitable to generate data modelling both the distance between means (the effect size), symmetry and variance.

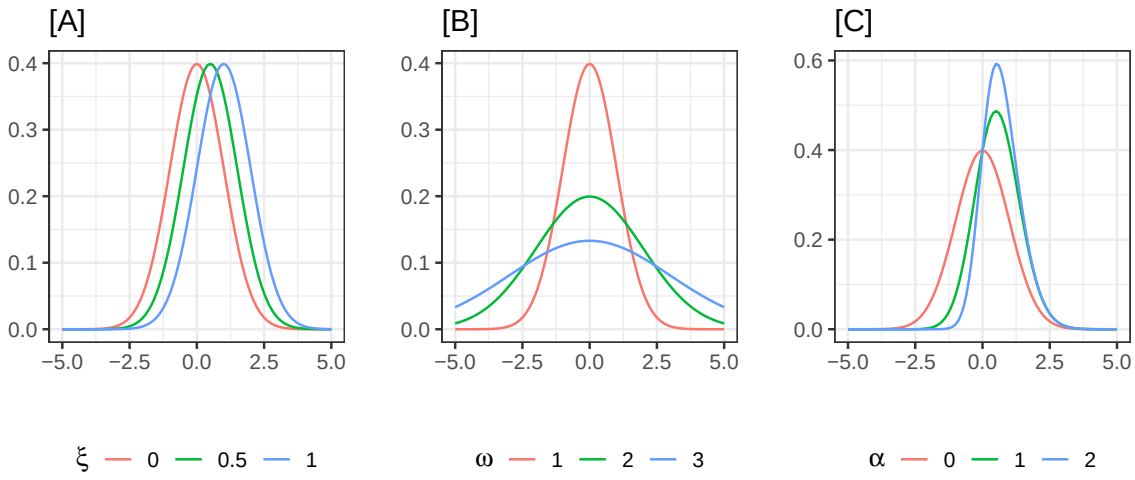

Figure 3: Examples of Skew-Normal distributions (ξ,ω,α); [A] three densities with same scale and shape but different location parameter values ($\xi = 0, 0.5, 1$), [B] three densities with same location and shape but different scale parameter values ($\omega = 1, 2, 3$) and [C] three densities with same location and scale but different shape parameter values ($\alpha = 0, 1, 2$).

Mean and variance of the Skew-Normal are respectively:

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

in which $\gamma = \alpha/\sqrt{1+\alpha^2}$. Based on the equations (4) we can determine the values to assign to the parameters ξ e ω in function of μ and σ with the equations:

$$\begin{aligned} \xi &= \mu - \omega\gamma\sqrt{2/\pi} \\ \omega &= \sqrt{\sigma^2/[1 - (2\gamma^2)/\pi]} \end{aligned} \tag{5}$$

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

2.7 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 $\mathcal{SN}(\xi, \omega, \alpha)$. Consequently, the first sample derives always from a population with mean 0 and variance 1. To define the various scenarios, we manipulate the parameters of the second population in orther to obtain specific differences in means (δ), standard deviations (σ) and skewness (α). Four factors were sistematically varied in a complete four-factors design as follows:

- $\delta = (0, 0.2, 0.5, 0.8)$; mean of the second population, 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 population;
- $\alpha = (0, 2, 10)$; degree of asymmetry (skewness) of the second population;
- $n = (10, 20, 50, 100, 500)$; sample size, equal in the two samples.

For each of the $4 \times 3 \times 3 \times 5 = 180$ conditions we generated 3000 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 population, always a $\mathcal{SN}(0, 1, 0)$, and the red curves are relative to the second population $\mathcal{SN}(\xi, \omega, \alpha)$.

For each combination $\delta \times \sigma \times \alpha \times n$, 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 homogeneity of variances;
- Kolmogorov-Smirnov test for comparing two distributions.

The whole procedure generated a total of 540000 datasets as well as 3240000 of statistical tests and corresponding p-values.

2.8 Definition of Statistical tests

We introduce the chosen statistical tests summarizing the specific hypothesis and assumptions for each one.

2.8.1 t test

This is the classic case of a test for independent samples assuming equal variances and the normality of the two distributions:

$$H_0 : \mu_1 - \mu_2 = 0 \text{ with } \sigma_1^2 = \sigma_2^2$$

Therefore, in the scenarios from which the samples come from populations with same mean – figure 4, panels in the first left column, [1, 5, 9, 13, 17, 21, 25, 29, 33] – type I error control is estimated, meanwhile, power is estimated for the remaining scenarios. Note that assumption of homogeneity of variance for this test are met only in the scenarios in the first row.

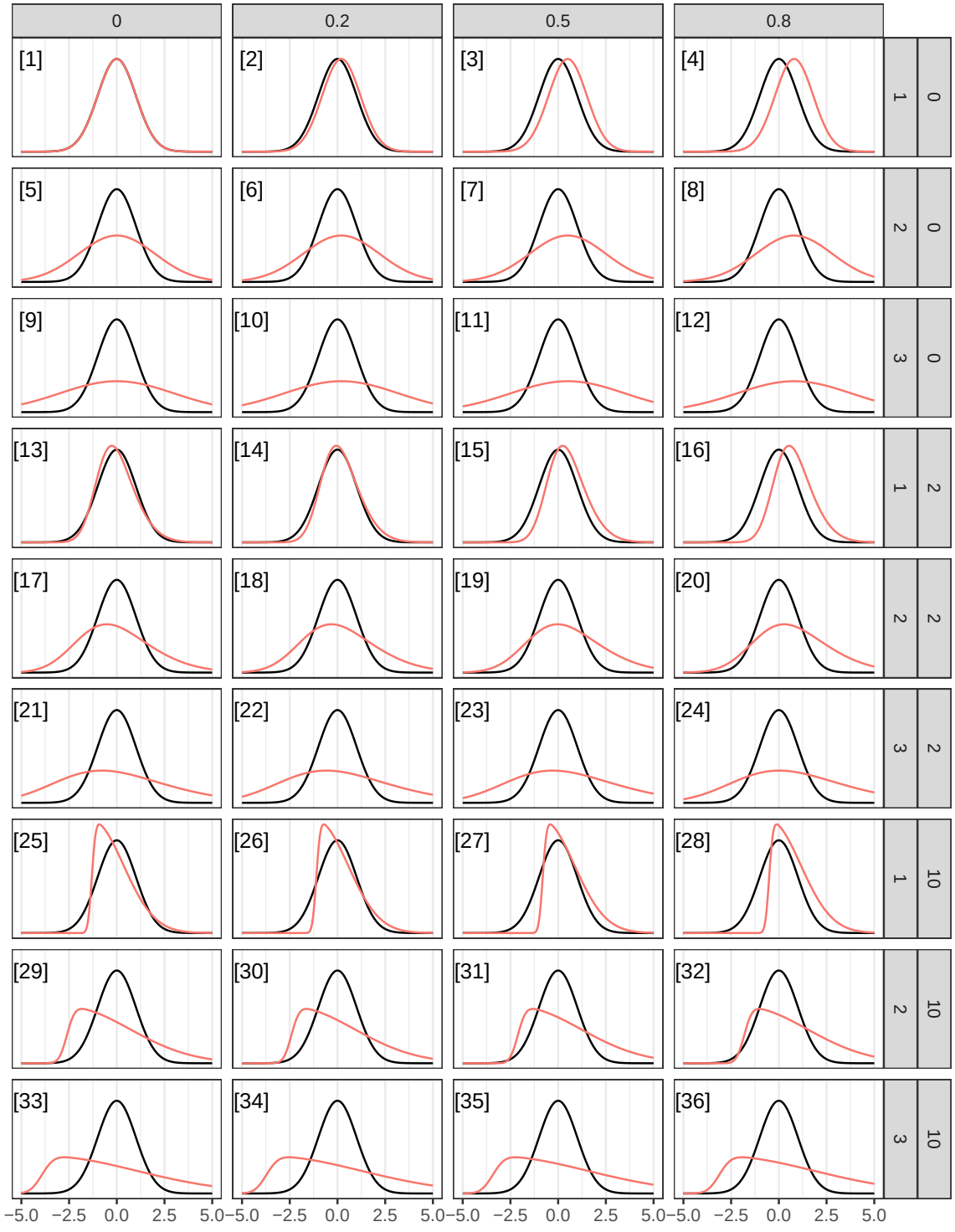

Figure 4: Generative data distributions in function of δ (column panels), σ and α (row panels). The black curves are the first population, $\mathcal{SN}(0, 1, 0)$, the red ones represent second population, $\mathcal{SN}(\xi, \omega, \alpha)$.

2.8.2 Welch (W) test

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

$$H_0 : \mu_1 - \mu_2 = 0 \text{ with } \sigma_1^2 \neq \sigma_2^2$$

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

2.8.3 Wilcoxon-Mann-Whitney (WMW) test

This is the test on ranks which evaluates the following hypothesis without assumptions on distributions:

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

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]. Given that this is a distribution free test, assumptions are not required.

2.8.4 Kolmogorov-Smirnov (KS) test

This test compares the cumulative distributions

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

without assumptions on distributions. The null hypothesis is true in panel [1], as it is for the ζ permutation test.

2.8.5 F test

This is the test of homogeneity of variances

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

assuming the normality. The condition is true in all scenarios where $\sigma = 1$, panels [1-4, 13-16, 25-28]. In those scenarios we estimate type I error, in all the others we calculate power.

2.8.6 ζ-Overlapping ($\zeta_{\mathbf{ov}}$) 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) but without specific assumptions. Therefore, the only condition in which H_0 is true is the first panel, [1]. Also in this case, the test does not require particular assumptions.

3 Results

First of all, we analysed correlations between the p-values of the considered tests in order to assess how much they are associated independently from the experimental condition. Next, we considered separately for each test in which scenarios H_0 is true, as reported in the previous section. Consequently, we computed type I error by counting how many times the test is significant in those scenarios, and the power by counting how many times it will be significant in all other scenarios. In this way, we evaluated type I error and power based on the experimental conditions.

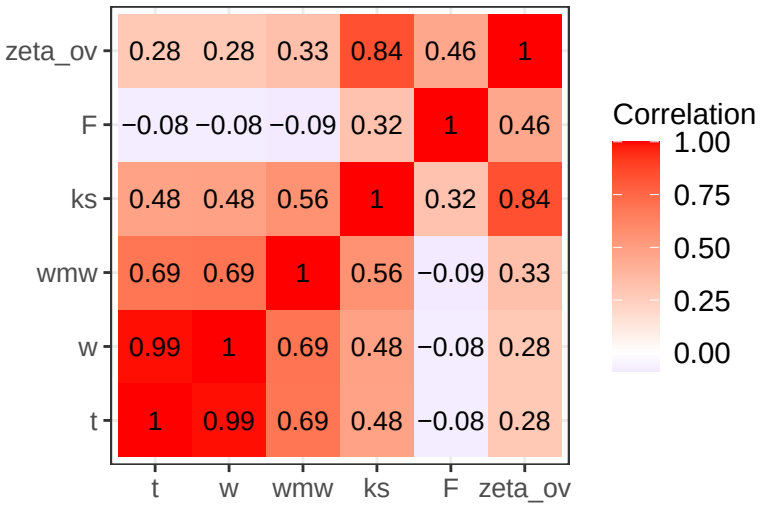

Figure 5: Correlation matrix among p-values ($N = 540000$) in chosen tests. Note: $\zeta_{ov} = \zeta$ overlapping test, F = variance test, ks = Kolmogorov-Smirnov test, wmw = Wicoxon-Mann-Whitney test, w = Welch test, t = Student's t test.

3.1 Correlations among tests

Figure 5 represents the correlation matrix between the p-values for the different tests in all experimental conditions. The classical tests show an order in the way they correlate. More specifically, t and W tests show a perfect correlation, WMW is highly correlated with the aforesaid tests, and the KS shows a lower but still medium-large correlation. F presents no correlation with t, W and WMW tests, and medium correlation with the ζ_{ov} and KS tests.

The ζ_{ov} test is highly correlated with the KS test, has a lower correlation with tests on means (t and W) and ranks (WMW), and a medium correlation with the F test.

The lower correlations observed among the test statistics are expected, given that each test is designed to detect different aspects of the data - whether central tendency, variance, or overall distributional shape. This pattern reinforces the idea that tests are not interchangeable but rather complementary in what they capture. The ζ-Overlapping test, by design, integrates sensitivity to multiple features, which explains its lower correlations with more narrowly focused tests. Therefore, while the correlation matrix does not offer direct interpretive insight into test performance, it supports the broader claim that distinct tests yield distinct inferential perspectives.

3.2 Type I error and power

Figure 6, shows the type I error in panel [A] and power in panel [B], estimated considering as true null hypothesis the situations reported in the section *Definition of Statistical tests*.

Remark It is important to note that the six tests compared in this simulation evaluate different null hypotheses and are therefore sensitive to different aspects of distributional differences. Specifically: the ζ-Overlapping test evaluates $H_0 : \zeta = 0$ (complete distributional overlap); the F-test evaluates $H_0 : \sigma_1^2 = \sigma_2^2$ (homoscedasticity); the Kolmogorov-Smirnov test evaluates $H_0 : F(X_1) = F(X_2)$ (identical cumulative distribution functions); the Wilcoxon-Mann-Whitney test evaluates $H_0 : P(X_1 > X_2) = 0.5$ (stochastic equality); the Welch test evaluates $H_0 : \mu_1 = \mu_2$ allowing unequal variances; and the Student's t-test evaluates $H_0 : \mu_1 = \mu_2$ assuming equal variances. This aspect is taken into account when defining for which scenarios to compute power and when type I error, ensuring a fair

comparison between tests.

In relation to type I error, almost all tests show a good performance, whereas the KS test is too conservative for small samples and the F-test is always above the nominal level of 0.05 (even with 500 observations per group).

Concerning power, the ζ-Overlapping test is the second outperforming other tests with good control of Type I error, already with small sample sizes, with the only exception of the F-test which has higher power but bad control of Type I error. From the graphical representation it is visible how two subgroups can be identified: one including the tests on means and ranks, not reaching adequate power even with large samples, and the second one formed by the ζ_{ov} and KS tests, reaching good power already from 100 observations, with the ζ_{ov} outperforming the KS test reaching good power already from 50 observations.

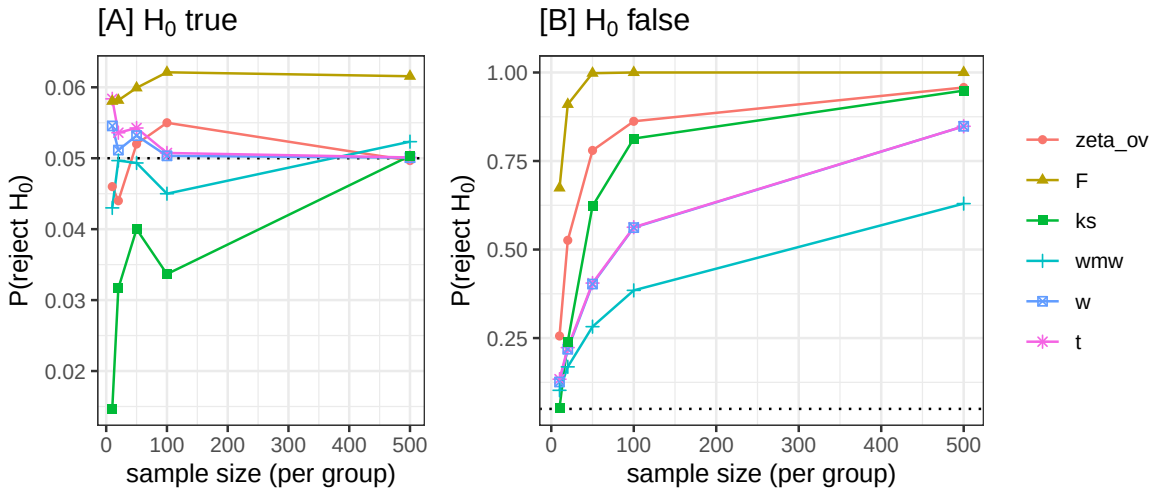

Figure 6: Control of type I error [A] and power [B] in the various tests. Note: ζ_{ov} = ζ-Overlapping test, F = variance test, ks = Kolmogorov-Smirnov test, wmw = Wicoxon-Mann-Whitney test, w = Welch test, t = Student's t test.

4 Discussion

The analysis of p-value correlations among the tests provides insight into their relationships. High correlation between the t-test and W test, for instance, reflects their similar objectives and shared focus on mean differences. However, lower correlations between parametric and non-parametric methods, such as the WMW and KS tests, indicate that these tests capture distinct aspects of the data, such as ranks or distribution shapes rather than means. The permutation-based tests show intermediate correlations with both parametric and non-parametric methods, which suggests that they may align with either types of tests depending on the underlying data structure, highlighting the versatility of the ζ_{ov} test.

Moreover, the present analysis evaluated the performance of various statistical significance tests across simulated scenarios comparing each test on the appropriate null hypothesis. The tests include both parametric (such as the t-test, Welch test, and F-test for variance) and non-parametric methods (such as the Wilcoxon Signed-Rank, Kolmogorov-Smirnov, and permutation-based approaches including the ζ_{ov} test). Through these scenarios, we assess each test's Type I error control, and power.

In figure 6, panel [A], the only test which is over-conservative in terms of Type I error is the KS test, yet reaching the nominal .05 level with bigger samples (over 400 observations). Instead, the F-test

shows a bad control of Type I error, always exceeding the nominal level even with big sample size. All other tests, already from small samples, are in the range of $.045 - .055$, converging closer to .05 as sample size grows. From this analysis we can then conclude that most of the tests sufficiently control the Type I error, but caution is needed for the F-test even with big samples and with the KS test with small samples.

Concerning power, figure 6 panel [B], the F-test is the most sensitive even with small variance differences, at the cost of not controlling Type I error. Second for power is the ζ_{OV} test outperforms all other tests. Competing with good power is the KS test, reaching a power of 60% with 50 observations and aligning with the ζ_{OV} at 500 observations. W and t tests performances are identical, in line with their almost perfect correlation of p-values, never exceeding 60% of power and performing consistently worse than the ζ_{OV} and KS tests, especially with small samples. Lastly, the WMW test performs similarly to t and W tests with slightly less power throughout the increase of sample size. These results highlight the outstanding performance of the ζ_{OV} test already from small samples, showing an advantage in choosing this test in research settings regardless of data distribution and assumptions.

While the ζ-Overlapping test demonstrates consistently high power across the simulation scenarios, it is not designed to replace traditional tests focused on specific parameters like the mean. Rather, it offers a global option when researchers are not interested in a single parameter, or when data violate the assumptions of parametric tests. In this way, it complements rather than competes with existing approaches, and supports the growing shift in psychological science toward robust, distribution-aware inference.

Moreover, it is important to note that this performance reflects its broader sensitivity to differences in distributional shape, not just central tendency therefore it may detect effects that more narrowly focused tests (e.g., t-test or F-test) are not designed to identify. This feature is highlighted in the example presented in the Practical Application section, showing how the ζ-Overlapping test is responsive to a wider set of deviations from the null.

The main limitation of the study is that it simulates only scenarios in which the first population is a Standard-Normal distribution ($\mathcal{SN}(0, 1, 0)$) and it does not consider the presence of outliers, which would give more insight into the performance of the ζ-Overlapping test. Another concern could be that the test does not inform on specific parametric differences, as its design focuses on distributional overlap rather than mean differences, which means it does not directly inform on mean, variance or skewness differences specifically. But we argue that this is rather a chore caracteristic of the test and not a limitation. The ζ_{OV} test offers a great starting point to evaluate whether two distributions differ in the first place with high power already from small samples. Moreover, the `overlapping` R package to easily compute the index also offers the possibility to plot densities and the area of overlap, therefore making it extremely intuitive to visualize how the two distributions practically differ.

While the ζ-test offers a broader diagnostic scope, it is not intended to replace targeted hypothesis tests when those are clearly aligned with the research goal and respect the nature of the data. When the research hypothesis specifically targets a single parameter, such as a mean difference, the use of more focused tests (e.g., Welch's t-test) may offer higher statistical power under valid assumptions. A complementary simulation study comparing power across such focused and omnibus approaches would be a valuable avenue for future work.

5 Conclusion

Many researchers focus on differences in means and may not initially consider the full distribution of their data. One of the strengths of the ζ-Overlapping test is precisely that it encourages a more holistic view, prompting researchers to explore whether groups differ not just in central tendency but in dispersion or shape as well. More precisely, it is easy to interpret as an effect size, with high values of ζ signaling differences between the two empirical distributions and low values indicating similarity. It is

robust to distributional assumptions, as it calculates p-values through permutations rather than relying on parametric assumptions like normality or equal variance, making it particularly useful in scenarios where other tests may be sub-optimal due to assumption violations, also providing a conservative Type I error rate when H_0 is true and robust power when H_0 is false. In practice, the ζ-Overlapping test can serve as a global test that prompts a broader examination of the data's characteristics.

By exploring alternative scenarios, the study offers a practical indication to operate a shift in the philosophical approach to data analysis and significance testing. In fact, the Overlapping Index forces the functional interpretation of the results to move beyond significance testing alone (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016; Gelman, 2018; Pastore & Calcagnì, 2019). In psychological research, considering the distribution of data rather than relying solely on significance testing offers a deeper, more nuanced understanding of results. Traditional significance testing does not provide information about the nature or magnitude of that effect (see Cohen, 1994; Wagenmakers, 2007; Ziliak & McCloskey, 2008; Wasserstein & Lazar, 2016). By visualizing and considering the entire distribution of data, researchers can observe the spread, central tendency, and shape of the data, which often reveal valuable insight about variability and individual differences within the sample. As in example presented in figure 2, reporting a mean difference without an understanding of the data distribution could lead to misrepresentation of the consistency or generalizability of the observed effect. Therefore, incorporating distributional analyses allows psychologists to present a fuller picture of their findings, improving both interpretability and transparency in their research conclusions.

While classical concerns regarding normality and homoscedasticity tend to diminish with increasing sample sizes, the ζ-Overlapping test offers unique advantages that persist even in large-sample scenarios. Specifically, it provides a formal and assumption-free way to test whether full empirical distributions differ beyond just location parameters allowing researchers to assess global differences with one test. The permutation-based p-value offers a rigorous statistical complement to data visualization: while plotting distributions is essential, ζ offers an objective inferential check that enhances transparency and reproducibility, particularly when interpretation may be ambiguous.

Importantly, not all issues resolve with large sample sizes. Psychological measures such as reaction times and event-related potentials (ERPs) components typically exhibit non-normal, right-skewed distributions (Luck & Gaspelin, 2017; Blanca, Arnau, López-Montiel, Bono, & Bendayan, 2013). Reaction-time modeling relies heavily on ex-Gaussian distributions (e.g., (Lacouture & Cousineau, 2008) because skew violates parametric test assumptions even with $n > 50$. Meta-analyses show that only ~ 5.5 % of behavioral datasets approximate normality (skew ≈ 0; kurtosis ≈ 0), even in moderate sample sizes (n 10–30). These issues are compounded in physiological measures like ERPs, which often follow heavy-tailed distributions due to individual differences and noise. In these contexts, the Overlapping Index η and the ζ test offers a robust, nonparametric alternative: it quantifies full distributional differences and provides p-values via permutation, without requiring normality or equal variances. Moreover, beyond traditional significance testing, the ζ-Overlapping framework can be readily adapted for equivalence testing and minimum-effect testing (Lakens et al., 2018), in line with current recommendations (Riesthuis, 2024) Murphy & Myors, 1999). Because η (and its complement ζ) directly quantify the similarity or difference between empirical distributions, researchers can define meaningful thresholds (e.g., $\eta \geq 0.90$ or $\zeta \leq 0.10$) that reflect negligible differences for practical purposes. A permutation-based test of equivalence can then assess whether the observed ζ falls within the predefined bounds, supporting equivalence. Conversely, a minimum-effect test can assess whether ζ significantly exceeds a lower threshold, indicating a difference of at least a meaningful size. Most importantly, thresholds should always be case specific and reasoned upon, rather than conventional benchmarks. These extensions preserve the nonparametric and assumption-free nature of the ζ-test while allowing for more nuanced and informative inferential conclusions.

Moreover, the present study further underscores the necessity of reasoning on the most suitable statistical tools contingent on the specific characteristics of the data and the assumptions inherent in the analytical techniques employed. Such a switch in the philosophical approach to data analysis

in psychological sciences (Vasishth & Gelman, 2021) may improve the robustness and validity of psychological research findings, allowing for more aware interpretations and generalizations. We stress this by making open available data and material so that such an approach might be useful for a wide range of psychologists interested in increasing the understandability of their results.

Ultimately, statistics in psychology should reflect both theoretical knowledge and an appreciation for the distributional nuances of psychological variables. Rather than a rigid application of conventional methods, statistical analysis should be a deliberate choice that aligns with the nature of the data and the research question. The approach of the ζ-Overlapping test embodies this principle, capturing the depth and complexity of psychological effects in a way that is both methodologically rigorous and sensitive to the real-world structure of psychological phenomena.

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
δ is the difference between the two means
ξ is the location parameter of the skew-normal
ω is the scale parameter of the skew-normal
α is the shape parameter of the skew-normal

Funding

No Funding supported this project.

Conflicts of interest

The authors declare no conflict of interests.

Ethics approval

Not applicable.

Consent to practice

Not applicable.

Consent for pubblication

Not applicable.

Availability of data and materials

Data and materials to reproduce the present work are openly available at OSF

Code availability

The code to reproduce the present work is openly available at OSF

Authors' contributions

All authors contributed substantially to the conceptualization, code development and refinement of this manuscript. X. XXX and X. XXX performed the primary writing task with substantial input from the other authors.

References

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

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology*, *9*(2), 78–84.

Cohen, J. (1994). The earth is round (p¡. 05). *American psychologist*, *49*(12), 997.

Conversano, C., Frigau, L., & Contu, G. (2024). Overlapping coefficient in network-based semi-supervised clustering. *Computational Statistics*, 1–24.

Einbeck, J., Coolen-Maturi, T., Uwimpuhwe, G., & Singh, A. (2024). A comparison of threshold-free measures for assessing the effectiveness of educational interventions. *The Journal of Experimental Education*, 1–18.

Fisher, R. A. (1925). Theory of statistical estimation. In *Mathematical proceedings of the cambridge philosophical society* (Vol. 22, pp. 700–725).

Fisher, R. A. (1935). *The design of experiments*. Macmillan.

Garofalo, S., Giovagnoli, S., Orsoni, M., Starita, F., & Benassi, M. (2022). Interaction effect: Are you doing the right thing? *PLoS One*, *17*(7), e0271668.

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin*, *44*(1), 16–23.

Gini, C., & Livada, G. (1943). Nuovi contributi alla teoria della transvariazione. *Roma: Atti della VI Riunione della Società Italiana di Statistica*.

Greene, N. R., Guitard, D., Forsberg, A., Cowan, N., & Naveh-Benjamin, M. (2024). Working memory limitations constrain visual episodic long-term memory at both specific and gist levels of representation. *Memory & Cognition*, 1–25.

Habibi, I., Achour, H., Bounaceur, F., Benaradj, A., & Aulagnier, S. (2024). Predicting the future distribution of the barbary ground squirrel (atlantoxerus getulus) under climate change using niche overlap analysis and species distribution modeling. *Environmental Monitoring and Assessment*, *196*(11), 1–18.

Hawkins, S. J., Gärtner, Y., Offner, T., Weiss, L., Maiello, G., Hassenklöver, T., & Manzini, I. (2024). The olfactory network of larval Xenopus laevis regenerates accurately after olfactory nerve transection. *European Journal of Neuroscience*.

Hemerik, J., & Goeman, J. (2018). Exact testing with random permutations. *Test*, *27*(4), 811–825.

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.

Jensen, T. M., & Sanner, C. (2021). A scoping review of research on well-being across diverse family structures: Rethinking approaches for understanding contemporary families. *Journal of Family Theory & Review*, *13*(4), 463–495.

Karrobi, K., Tank, A., Fuzail, M. A., Kalidoss, M., Tilbury, K., Zaman, M., ... Roblyer, D. (2023). Fluorescence Lifetime Imaging Microscopy (FLIM) reveals spatial-metabolic changes in 3D breast cancer spheroids. *Scientific Reports*, *13*(1), 3624.

Lacouture, Y., & Cousineau, D. (2008). How to use matlab to fit the ex-gaussian and other probability functions to a distribution of response times. *Tutorials in quantitative methods for psychology*, *4*(1), 35–45.

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any erp experiment (and why you shouldn't). *Psychophysiology*, *54*(1), 146–157.

Nougaret, S., Ferrucci, L., Ceccarelli, F., Sacchetti, S., Benozzo, D., Fascianelli, V., ... Genovesio, A. (2024). Neurons in the monkey frontopolar cortex encode learning stage and goal during a fast learning task. *Plos Biology*, *22*(2), e3002500.

Oksuz, D. C., & Rebuschat, P. (2024, Mar). *Collocational processing in typologically different languages, english and turkish.* OSF. Retrieved from `osf.io/muwjz`

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

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

Pastore, M., Lionetti, F., & Altoè, G. (2017). When one shape does not fit all: A commentary essay on the use of graphs in psychological research. *Frontiers in psychology*, *8*, 1666.

Pastore, M., Loro, P. A. D., Mingione, M., & Calcagni', A. (2024). overlapping: Estimation of overlapping in empirical distributions [Computer software manual]. (R package version 2.2)

Pesarin, F. (2001). *Multivariate permutation tests: with applications in biostatistics.* Wiley.

Phipson, B., & Smyth, G. K. (2010). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, *9*(1).

Pietrabissa, G., Semonella, M., Marchesi, G., Mannarini, S., Castelnuovo, G., Andersson, G., & Rossi, A. A. (2024). Validation of the Italian version of the web screening questionnaire for common mental disorders. *Journal of Clinical Medicine*, *13*(4), 1170.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations. ii. the correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society*, *4*(2), 225–232.

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any populations: Iii. the analysis of variance test. *Biometrika*, *29*(3/4), 322–335.

Ricote, N., Weinberger, C., Ramírez-Otarola, N., Bustamante, S., Málaga, M. L., Barceló, G., ... Maldonado, K. (2024). The interplay of resource availability and parent foraging strategies on juvenile sparrow individual specialization. *Journal of Avian Biology*, e03391.

Riesthuis, P. (2024). Simulation-based power analyses for the smallest effect size of interest: A confidence-interval approach for minimum-effect and equivalence testing. *Advances in Methods and Practices in Psychological Science*, *7*(2), 25152459241240722.

Rohrbach, T. (2024). Are women politicians kind and competent? disentangling stereotype incongruity in candidate evaluations. *Political Behavior*, 1–24.

Rossi, A. A., Panzeri, A., Fernandez, I., Invernizzi, R., Taccini, F., & Mannarini, S. (2024). The impact of trauma core dimensions on anxiety and depression: a latent regression model through the Post-Traumatic Symptom Questionnaire (PTSQ). *Scientific Reports*, *14*(1), 23036.

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755.

Schuetze, B. A., & Yan, V. X. (2023). Psychology faculty overestimate the magnitude of cohen'sd effect sizes by half a standard deviation. *Collabra: Psychology*, *9*(1), 74020.

Sirbiladze, G., Midodashvili, B., & Manjafarashvili, T. (2024). Divergence and similarity characteristics for two fuzzy measures based on associated probabilities. *Axioms*, *13*(11), 776.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712.

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science*, *10*(2).

Tay, L., Parrigon, S., Huang, Q., & LeBreton, J. M. (2016). Graphical descriptives: A way to improve data transparency and methodological rigor in psychology. *Perspectives on Psychological Science*, *11*(5), 692–701.

Upadhayay, H. R., Granger, S. J., & Collins, A. L. (2024). Comparison of sediment biomarker signatures generated using time-integrated and discrete suspended sediment samples. *Environmental Science and Pollution Research*, *31*(15), 22431–22440.

Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*, *59*(5), 1311–1342.

Wachendörfer, M. M., & Oeberst, A. (2024). Differences between true and false memories using the criteria-based content analysis. *Applied Cognitive Psychology*, *38*(5), e4246.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, *14*(5), 779–804.

Wasserstein, R. L., & Lazar, N. A. (2016). *The asa statement on p-values: context, process, and purpose* (Vol. 70) (No. 2). Taylor & Francis.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American psychologist*, *54*(8), 594.

Ziliak, S. T., & McCloskey, D. N. (2008). The cult of statistical significance. *Ann Arbor, MI: University of Michigan Press*, *326*.