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

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

November 3, 2024

## Abstract

The present contribution aims to compare both commonly and less commonly used statistical methods in psychological sciences to evaluate their utility in tailored cases. Specifically, the paper proposes applying the Permutation test alongside the Overlapping index to estimate effects of interest in psychological science. Starting from real and openly available data, we simulated different scenarios focusing on residual distribution characteristics. The present contribution provides practical tools for considering, and deciding which statistical methods are useful and sufficient considering the features of data distribution. Subsequently, we present a Simulation study to illustrate the practical implications and reliability of each approach, particularly valuable in scenarios commonly encountered in quantitative psychology, where navigating data characteristics and adhering to or deviating from test assumptions is crucial. The findings underscore the necessity of choosing statistical methods that are resilient to the complexities inherent in psychological data, where assumption violations are often inevitable.

## 1 Statistical testing choices in Psychology

Methodological choices in cognitive and behavioral sciences aim to combine data richness with data collection feasibility, and at the same time they aim to land on valid interpretation based on reliable and robust statistical methods. Classic examples, like reaction times, demonstrate how specific measures have achieved such an acceptable trade-off, and for example, this is true even by comparing the framework of in lab *vs* online data collection (Semmelmann & Weigelt, 2017). Nevertheless, even in the fortunate case of reaction times which have widespread and solid epistemic rationale of use (Grosjean, Rosenbaum, & Elsinger, 2001; Proctor & Schneider, 2018; Silverman, 2010) significance testing often relies on the rigid application of a few statistical methods that have gained popularity among the scientific community and are perpetrated *perinde ac cadaver* by formal guidelines (Cumming, Fidler, Kalinowski, & Lai, 2012), even if their limits and risks have always been noted in the field of psychology and beyond (Boneau, 1960).

In fact, there is a growing caution against blindly using statistical tools and analytical methods without a deep understanding of their assumptions and implications (Scheel, Tiokhin, Isager, & Lakens, 2021). In other words, it is increasingly apparent that relying solely on significance testing as a trustworthy measure is improbable without considering the assumptions inherent to specific statistical methods, such as the t-test, across various scenarios in psychology. In fact, considering the particular circumstances of application has consistently been crucial advice when deciding on significance testing methods (Fisher, 1925).

The present contribution aims to compare both commonly and less commonly used statistical methods in psychological sciences to evaluate their utility in specific and tailored cases. Through an illustrative example, we re-analyze reaction times coming from a real and available dataset of a reading task with high- and low-frequency words. Specifically, the present work proposes applying the Permutation test (Pesarin & Salmaso, 2010) alongside the Overlapping index (Pastore & Calcagnì, 2019) to estimate effects of interest in psychological sciences. Of note, the proposed approach is practically declined by using a word reading study, however the very same logic can extend to other measures in psychology sciences. Importantly, by simulating different scenarios focusing on residual distribution characteristics, the paper provides practical tools for considering, and deciding which statistical methods are useful and sufficient considering the features of data distribution. Understanding and applying significance testing properly is crucial to deriving meaningful conclusions from psychological research. Accordingly, we offer practical and reproducible tools to manage the assumptions underlying these analytical approaches, and increase awareness in significance testing in psychology.

In particular, when the assumptions of linear regres-

sion, such as normality and homoscedasticity, are not met, alternative methods become optimal. The use of indices calculated on empirical distributions is particularly beneficial when these assumptions are violated (Pastore, 2015). Specifically, when using a t-test to compare two groups or two experimental conditions using a given variable, it functions as a straightforward version of linear regression. This statistical process necessitates assumptions about the residuals, such as their independence and normal distribution, to be met. In cases such as reaction times, these assumptions might be violated if they are not properly addressed. Sometimes, two populations might present the same mean for a given variable, yet their distributions largely differ in other parameters, leading to genuinely distinct groups (see figure 1).

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, showcasing its relevance in statistical analysis. Subsequently, we present a Simulation study to illustrate the practical implications and reliability of the overlapping index utilizing permutations.

In the following section, we compare several statistical tests: the t-test for independent samples assuming equal variance, the Welch test for independent samples, the Wilcoxon test for independent samples, the Permutation test on the complement of the Overlapping index ($\zeta = 1 \ \ \eta$), which serves as a measure of intergroup differences, the F test for homogeneity of variances, and the Kolmogorov-Smirnov test for comparing two distributions.

The rationale behind these steps involves first introducing the concept of the Overlapping Index ($\eta$), which is crucial because it provides an intuitive measure of similarity between distributions by quantifying the overlapping area of their probability density functions, a common question in quantitative psychology. The Permutation approach is then defined and applied to the Overlapping Index, demonstrating how non-parametric methods can offer insights without relying on typical parametric assumptions. Specifically, the Permutation test involves shuffling data points to generate a sampling distribution, allowing the calculation of a p-value and highlighting its utility in assessing the statistical significance of the Overlapping Index. Next, a Simulation study uses a real dataset to simulate various scenarios that might meet or violate the assumptions of different statistical tests, modeling a range of conditions reflective of real-world complexities in psychological research. This simulation facilitates the evaluation of the statistical power (probability of correctly rejecting a false null hypothesis) and the type I error rate (likelihood of incorrectly rejecting a true null hypothesis) of each approach. To this end, several statistical tests are compared: the t-test for independent samples, assuming equal variances; the Welch test, which does not assume equal variances; the Wilcoxon test, suitable for ordinal data or when normality is not assumed; the Permutation Test on the Overlapping Index Complement ($\zeta = 1 \ \ \eta$), providing a non-parametric approach to evaluate intergroup differences; the F test for examining the homogeneity of variances; and the Kolmogorov-Smirnov test for comparing two distributions regardless of their underlying forms. These results enable researchers to visualize and comprehend the reliability and utility of each approach, particularly valuable in scenarios commonly encountered in quantitative psychology, where navigating data characteristics and adhering to or deviating from test assumptions is crucial.

Finally, we discuss the results, offering insights into the strengths and limitations of the Permutation-based Overlapping index and its potential applications in psychological sciences.

# 2 Overlapping Index

Cognitive and experimental researchers regularly strive to uncover evidence that supports their hypotheses by examining statistical differences or similarities among groups or conditions. Frequently, this involves measuring the difference between two distribution within the same dependent variable, basically relying on their mean values using metrics like the t statistic, Cohen's $d$, or $U$ statistics. The goal in each scenario is to estimate the magnitude of these differences to identify them as significant effects. However, a complementary perspective can be gained through the overlapping index ($\eta$), which intuitively quantifies the common area between two or more probability density functions. This measure serves as an additional tool for comparing distributions, where greater overlap indicates similarity, and a decrease in $\eta$ signals divergence (Pastore & Calcagnì, 2019).

The index $\eta$ 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 ($\eta$) 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 $\eta : \mathbb{R}^n \times \mathbb{R}^n \rightarrow [0, 1]$ 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

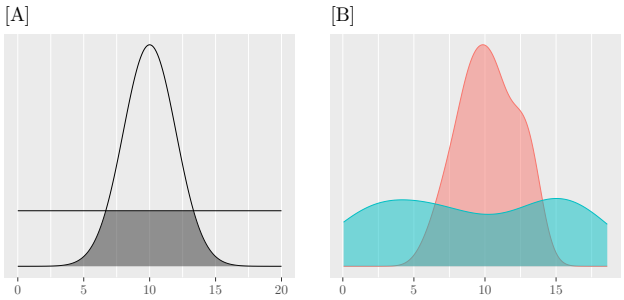


Figure 1: Comparison of a normal distribution and a uniform distribution with same mean.

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.

## 2.1 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.

## 2.2 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 $\eta$, 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 $\eta$, but with the complement $\zeta$.

Even though the overlapping index has a simple interpretation, one could argue that it does not provide information on the significance of the parameter $\eta$, 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 $\zeta$ 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:

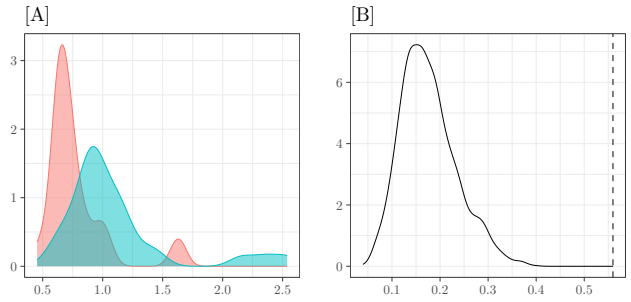


Figure 2: $\hat{\zeta} = 0.56$. [A] Distribution of reaction times of word reading of high and low frequency words in English; [B] Distribution of $\hat{\zeta}$ obtained with 2000 permutations of the data.

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

 Welch Two Sample t-test

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

# 3 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 $\zeta$ perm test to other commonly used tests in terms of type I error control and power.

## 3.1 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

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

in which

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

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

The parameter $\alpha$ determines the symmetry, $\xi$ is the mean value and $\omega$ determines the variance. Therefore, this distribution is suitable to generate data modelling both the distance between means (the effect size), symmetry and variance.

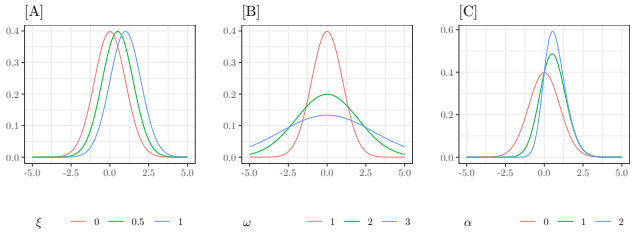


Figure 3: Skew-Normal distribution ($\xi$,$\omega$,$\alpha$); [A] the parameter $\xi$ controls the mean, [B] the parameter $\omega$ the variance and [C] the parameter $\alpha$ the simmetry.

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 $\xi$ e $\omega$ in function of $\mu$ and $\sigma$ 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.

## 3.2 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:

- $n = ()$; sample size, equal in the two samples;
- $\delta = ()$; mean of the second sample, which corresponds also to the difference between the two groups, the first one has always $\mu = 0$;
- $\sigma = ()$; standard deviation of the second sample;
- $\alpha = (0,1,2)$; 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,

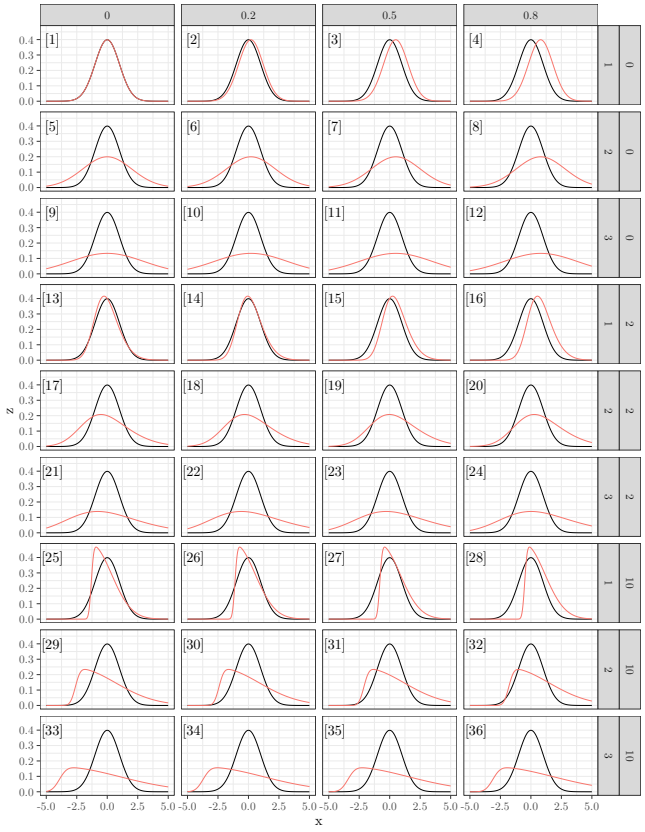


Figure 4: Generative data distributions in function of $\delta$ (column panels) and $\sigma$ (row panels). The black curves are the first sample, $\mathcal{SN}(0, 1, 0)$, the red ones represent second sample.

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

## 3.3 Definition of Null Hypothesis

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

### 3.3.1 $t$ 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.

### 3.3.2 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.

### 3.3.3 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].

### 3.3.4 $\zeta$ permutation test

Since $\zeta = 1 - \eta$, in which $\eta$ 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.

### 3.3.5 $F$ 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.

### 3.3.6 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 $\zeta$ 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.

# 4 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.

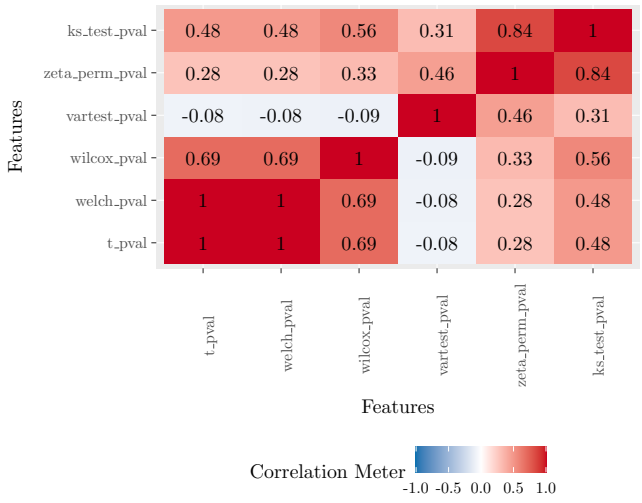


Figure 5: Correlation matrix among $p$-values ($N = 180000$).

## 4.1 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 $\zeta$ 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.

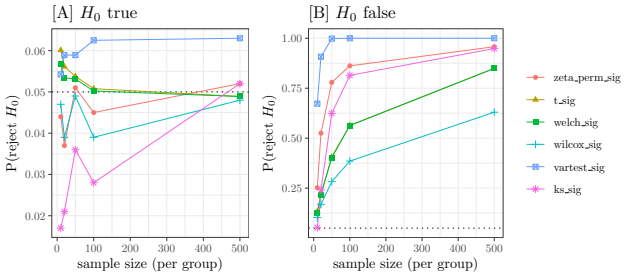


Figure 6: Control of type I error [A] and power [B] in various tests taking into account for each of them in which scenario $H_0$ is true or false.

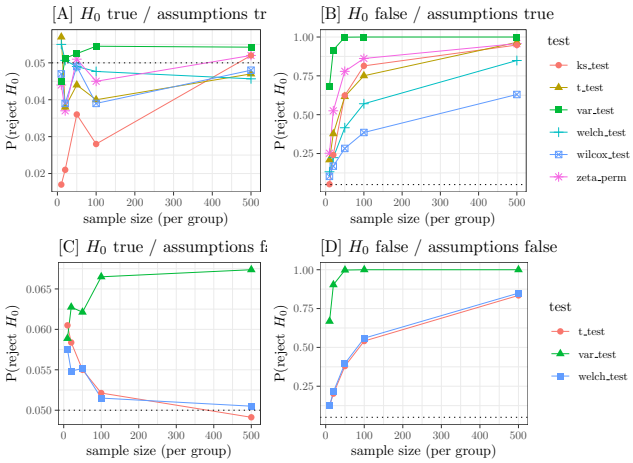


Figure 7: Control of type I error and power for scenarios in which assumptions are respected (top panels) and when they are not (bottom panels).

## 4.2 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 $\zeta$ 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.

# Comparison of Statistical Tests Across Simulated Scenarios

In this analysis, we compared the performance of several statistical significance tests across different simulated scenarios starting from a real data collection, where assumptions were either met or violated, and the null

hypothesis ($H_0$) was either true or false. The simulation study starting from the real dataset emerged to be an informative approach that allowed to evaluate the significance testing performance across statistical test knowing the *ground truth* beyond the data generation. The tests evaluated include both parametric (e.g., T-test, Welch test, F-test for variance) and non-parametric tests (e.g., Wilcoxon Signed-Rank, Kolmogorov-Smirnov, and permutation-based tests, including $\zeta$ permutation). Each scenario provides insight into the robustness, power, and Type I error control of these methods under varied conditions.

## Scenario A: $H_0$ True, Assumptions Met

In Scenario A, where $H_0$ is true and the assumptions are satisfied, all tests ideally should maintain Type I error close to the nominal level of 0.05. Here, we observe that:

- The **T-test** and **Welch test** control Type I error well, as expected for parametric tests under ideal conditions. The Welch test shows slightly more variability in Type I error, likely due to its adjustment for unequal variances, though this remains within an acceptable range.
- The **Wilcoxon Signed-Rank test** and the **Kolmogorov-Smirnov (KS) test** appear slightly conservative, producing Type I error rates below the nominal level. This conservative behavior is typical of non-parametric tests that are more robust to non-normality, though they may reduce sensitivity when assumptions are fully met.
- The $\zeta$ **permutation test** is also conservative, which could indicate its suitability for controlling Type I error in scenarios where overlap between distributions is the focus.
- The **permutation-based T-test** and **F-test** yield Type I error rates close to the nominal level, benefiting from the flexibility of permutation-based p-value calculation even when assumptions are met.

In summary, under ideal conditions, most tests perform as expected, with parametric tests aligning closely with the nominal Type I error and non-parametric and permutation methods showing slight conservatism.

## Scenario B: $H_0$ False, Assumptions Met

When $H_0$ is false and assumptions are met (Scenario B), power is the primary metric of interest. Here, we find:

- The **Welch test** demonstrates high power, surpassing the T-test as sample size increases, due to its flexibility with unequal variances. This makes it a robust choice when variances may differ even if assumptions of normality are met.
- The **permutation-based T-test** and $\zeta$ **permutation test** also exhibit high power, highlighting their effectiveness in detecting true effects without relying on strict distributional assumptions.
- Non-parametric tests, such as the **Wilcoxon** and **KS** tests, show moderate power, though they are generally less sensitive to mean differences than parametric alternatives. Their focus on distribution shapes or ranks limits their power when mean differences are the primary effect.

In this scenario, the Welch test and permutation-based methods emerge as highly effective for detecting differences, especially when variances may differ, while non-parametric tests are somewhat limited in sensitivity.

## Scenario C: $H_0$ True, Assumptions Violated

When assumptions are violated but $H_0$ remains true (Scenario C), Type I error control becomes crucial. This scenario reveals the robustness of each test under non-ideal conditions:

- The **Welch test** maintains Type I error control effectively, showcasing its robustness to heteroscedasticity and other violations. This highlights its utility in practical scenarios where equal variance cannot be guaranteed.
- The **permutation-based tests** (both for means and variances) also perform well, maintaining Type I error near the nominal level, thanks to their non-parametric approach to p-value calculation.
- The **T-test** and **Variance (F) test** exhibit increased sensitivity to assumption violations, particularly the F-test, which shows inflated Type I error rates under heteroscedasticity and non-normality. This sensitivity reduces their reliability in practical applications where assumptions are not met.
- Non-parametric tests, like the **Wilcoxon** and **KS tests**, handle assumption violations effectively, producing conservative Type I error rates. Their robustness to non-normality makes them a safer choice when parametric assumptions are doubtful.

Under assumption violations with a true null hypothesis, the Welch and permutation-based tests stand out as reliable choices, while the F-test is notably sensitive to violations.

## Scenario D: $H_0$ False, Assumptions Violated

In the final scenario, where both $H_0$ is false and assumptions are violated (Scenario D), the adaptability of each test is evaluated under the most challenging conditions:

- The **Welch test** maintains high power, adapting well to heteroscedasticity and other assumption violations. This underscores its suitability for real-world data where variances may be unequal and normality cannot be assumed.

- The **permutation-based T-test** and $\zeta$ **permutation test** also demonstrate strong performance, showing that permutation-based approaches can be powerful alternatives when assumptions do not hold.

- Non-parametric tests like **Wilcoxon** and **KS** show moderate power but generally lag behind parametric tests in detecting mean differences. They remain robust to assumption violations but are less efficient in detecting differences in means, particularly with skewed distributions.

- The **Variance (F) test** performs poorly in this scenario, with both reduced power and increased error rates, underscoring its sensitivity to assumption violations. Its reliance on equal variance assumptions makes it unsuitable in situations where homoscedasticity cannot be assured.

In this scenario, the Welch test and permutation methods again emerge as the most adaptable, providing good power even when assumptions are substantially violated.

### Correlation Analysis of p-values

An analysis of p-value correlations among tests provides additional insights. The high correlation between the **T-test** and **Welch test** reflects their similar objectives, particularly in testing mean differences. However, the lower correlations between parametric and non-parametric tests, such as the **Wilcoxon** and **KS** tests, indicate that these tests capture different aspects of the data (e.g., ranks or distribution shapes rather than means). The **permutation-based tests** exhibit intermediate correlations with both parametric and non-parametric methods, indicating that their results may align with both types depending on the underlying data structure.

## 5 Discussion

The present analysis evaluated the performance of various statistical significance tests across simulated scenarios that altered whether the null hypothesis ($H_0$) is true or false and whether the assumptions required by each test are met. 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 $\zeta$ permutation test). Through these scenarios, we assess each test's robustness, Type I error control, and power under ideal and non-ideal conditions.

In **Scenario A** (true $H_0$, assumptions met), the T-test, Welch test, and permutation-based tests maintain Type I error close to the nominal level of 0.05, with the Welch test showing minor variability due to its robustness to variance differences. Non-parametric tests (Wilcoxon, KS, $\zeta$-permutation) are conservative, slightly undershooting the nominal error—a typical trade-off for robustness.

In **Scenario B** (false $H_0$, assumptions met), power is key. The Welch test excels as sample sizes grow, especially with unequal variances, while permutation-based tests (T-test, $\zeta$) show strong power, making them effective in detecting true effects without strict distributional requirements. Non-parametric tests display moderate power, better suited for cases where shape differences are of interest rather than mean differences.

**Scenario C** (true $H_0$, assumptions violated) tests robustness. The Welch and permutation-based tests maintain Type I control under assumption violations, whereas parametric T and F-tests struggle, particularly with heteroscedasticity. Non-parametric tests (Wilcoxon, KS) remain conservative and reliable, showing resilience to non-normality.

Finally, in **Scenario D** (false $H_0$, assumptions violated), the Welch test and permutation-based tests (T-test, $\zeta$) stand out with high power and robustness, ideal for real-world data with heteroscedasticity or non-normality. Non-parametric tests retain robustness but lack power, while the F-test proves unreliable under these challenging conditions.

Further analysis of p-value correlations among the tests provides additional insight into their relationships. High correlation between the T-test and Welch 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 Wilcoxon 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 type of test depending on the underlying data structure.

### Advantages and Limitations of the $\zeta$ Permutation Test

The $\zeta$ **permutation test**, designed to measure the degree of overlap between distributions, has specific advantages and limitations. Its main strength lies in its robustness to distributional assumptions, as it calculates p-values through permutations rather than relying on parametric assumptions like normality or equal variance. This makes it particularly useful in scenarios where other

tests may fail due to assumption violations, providing a conservative Type I error rate when $H_0$ is true and robust power when $H_0$ is false.

However, the $\zeta$ permutation test's conservative nature may limit its sensitivity in detecting small mean differences, especially when distributions overlap substantially. Its design focuses on distributional overlap rather than mean differences, which means it may lack power relative to parametric tests that specifically target mean shifts. In scenarios where the primary effect is a shift in central tendency rather than overlap, the $\zeta$ permutation test may not be the optimal choice.

The $\zeta$ permutation test is indeed a valuable tool for non-parametric inference, particularly when distributional assumptions do not meet those required by common statistical test e.g. t-test. These are particularly relevant points given that in psychological sciences studies often involve small sample sizes, and relying on small changes in location parameters like the mean can be risky. For example, small samples are highly susceptible to the influence of extreme values, which can skew the mean and lead to misleading conclusions about effect sizes. On the contrary, as demonstrated in simulations, the $\zeta$ permutation test is less prone to being dramatically impacted by extreme values, as it directly measures the distributional overlap between groups rather than focusing solely on mean differences. This characteristic makes the $\zeta$ permutation test particularly valuable in small-sample contexts, like in psychological science where robustness to outliers is critical for obtaining reliable insights into group differences.

# 6 Conclusion

By exploring alternative scenarios, the study offers 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 (Pastore et al., 2018; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016; Gelman, 2018). 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 doesn't provide information about the nature or magnitude of that effect. 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 insights about variability and individual differences within the sample. As presented in Figure 1, reporting a mean difference without an understanding of the data's variability could misrepresent 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.

Moreover, the present study further underscores the necessity of selecting 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 improves 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.

The findings underscore the necessity of choosing statistical methods that are resilient to the complexities inherent in psychological data, where assumption violations are often inevitable. Tests like the Welch and $\zeta$ permutation tests exemplify robust alternatives that accommodate data with unequal variances or non-normal distributions, offering reliable results even when classic parametric conditions are unmet. In this way, these tools extend the flexibility of significance testing, enabling a nuanced understanding of effects in psychology.

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. Approaches such as the overlapping index and permutation-based methods embody 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.

# 7 Legenda

$\eta$ is the area of overlap
$\zeta$ is the area of non overlap, therefore $1 - \eta$
$\mu$ is the parameter of the mean of the normal standard
$\sigma$ is the standard deviation of the normal standard
$\alpha$ determins the simmetry of the skew-normal
$\xi$ is the mean value of the skew-normal
$\omega$ determines the variance of the skew-normal
$\delta$ is the difference between the two means

# 8 Ethical considerations

Ethical approval was not required

# 9 Conflicting interest

The authors declare no conflict of interests.

## 10 Funding statement

No Funding supported this project.

## 11 Data availability statement

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

## References

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

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological bulletin*, *57*(1), 49.

Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the american psychological association publication manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology*, *64*(3), 138–146.

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

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.

Grosjean, M., Rosenbaum, D. A., & Elsinger, C. (2001). Timing and reaction time. *Journal of Experimental Psychology: General*, *130*(2), 256.

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. (2015). Analisi dei dati in psicologia. *Il mulino*.

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.

Proctor, R. W., & Schneider, D. W. (2018). Hick's law for choice reaction time: A review. *Quarterly Journal of Experimental Psychology*, *71*(6), 1281–1299.

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.

Semmelmann, K., & Weigelt, S. (2017). Online psychophysics: Reaction time effects in cognitive experiments. *Behavior Research Methods*, *49*, 1241–1260.

Silverman, I. W. (2010). Simple reaction time: It is not what it used to be. *The American journal of psychology*, *123*(1), 39–50.

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

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

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