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

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

[1] Department of Developmental and Social Psychology, University of Padua, Italy

[2] Department of General Psychology, University of Padua, Italy

[3] Department of Statistical Sciences, University of Padua, Italy

January 28, 2026

Abstract

The present study introduces the application of the permutation test to the Overlapping Index, an effect size measure for comparing density distributions of groups or conditions, to estimate effects of interest in psychological science. Starting with common scenarios in psychological science research, the paper highlights the importance of relying on statistical methods that are resilient to the complexities inherent in psychological data, where assumption violations are often inevitable. A Simulation study is presented to illustrate the practical implications and reliability of the proposed test compared to commonly used alternatives. The findings demonstrate the good control of Type I error of the ζ-Overlapping test and how this approach outperforms in terms of power all other tests considered in the simulation, already with small samples. The paper offers practical guidance and demonstrates the advantages of this method, emphasizing its potential to enhance transparency and rigor in psychological data analysis by shifting focus from traditional significance testing to comprehensive distributional evaluations.

1 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-Overlapping permutation test to other commonly used tests in terms of type I error control and power.

1.1 Data generation

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

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

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.

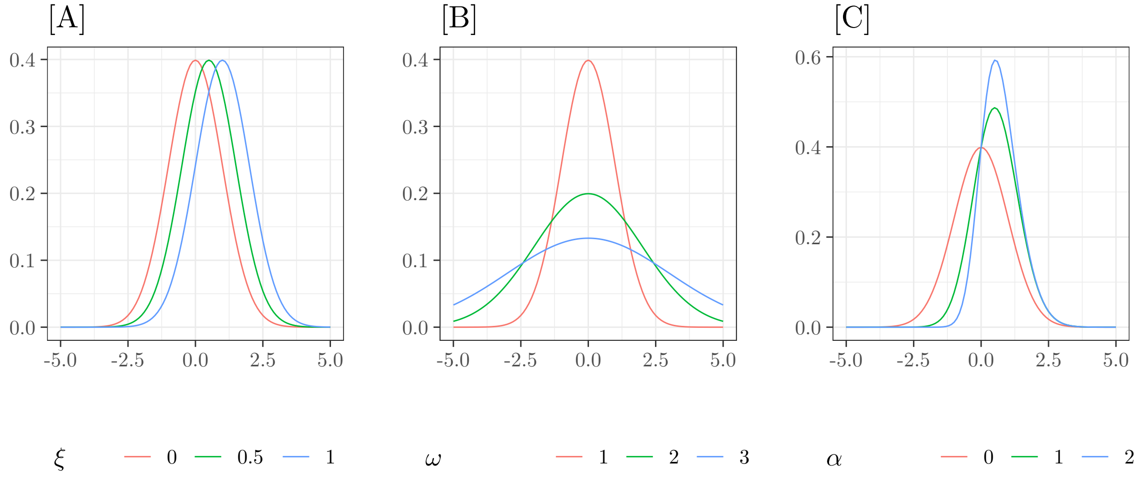

Figure S1: Examples of Skew-Normal distributions (ξ,ω,α); [A] three densities with same variance and shape but different location parameter values ($\xi = 0, 0.5, 1$), [B] three densities with same mean and shape but different scale parameter values ($\omega = 1, 2, 3$) and [C] three densities with same mean and variance but different shape parameter values ($\alpha = 0, 1, 2$).

Mean and variance of the Skew-Normal are respectively:

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

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

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

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 S1.

1.2 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 $\mathcal{SN}(\xi,\omega,\alpha)$. 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 ina 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 S2 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 4 (mean differences) × 3 (standard deviation differences) × 3 (shape differences) × 5 (sample sizes) × 3000 (replications) = 540,000 datasets as well as 3,240,000 of statistical tests and corresponding p-values.

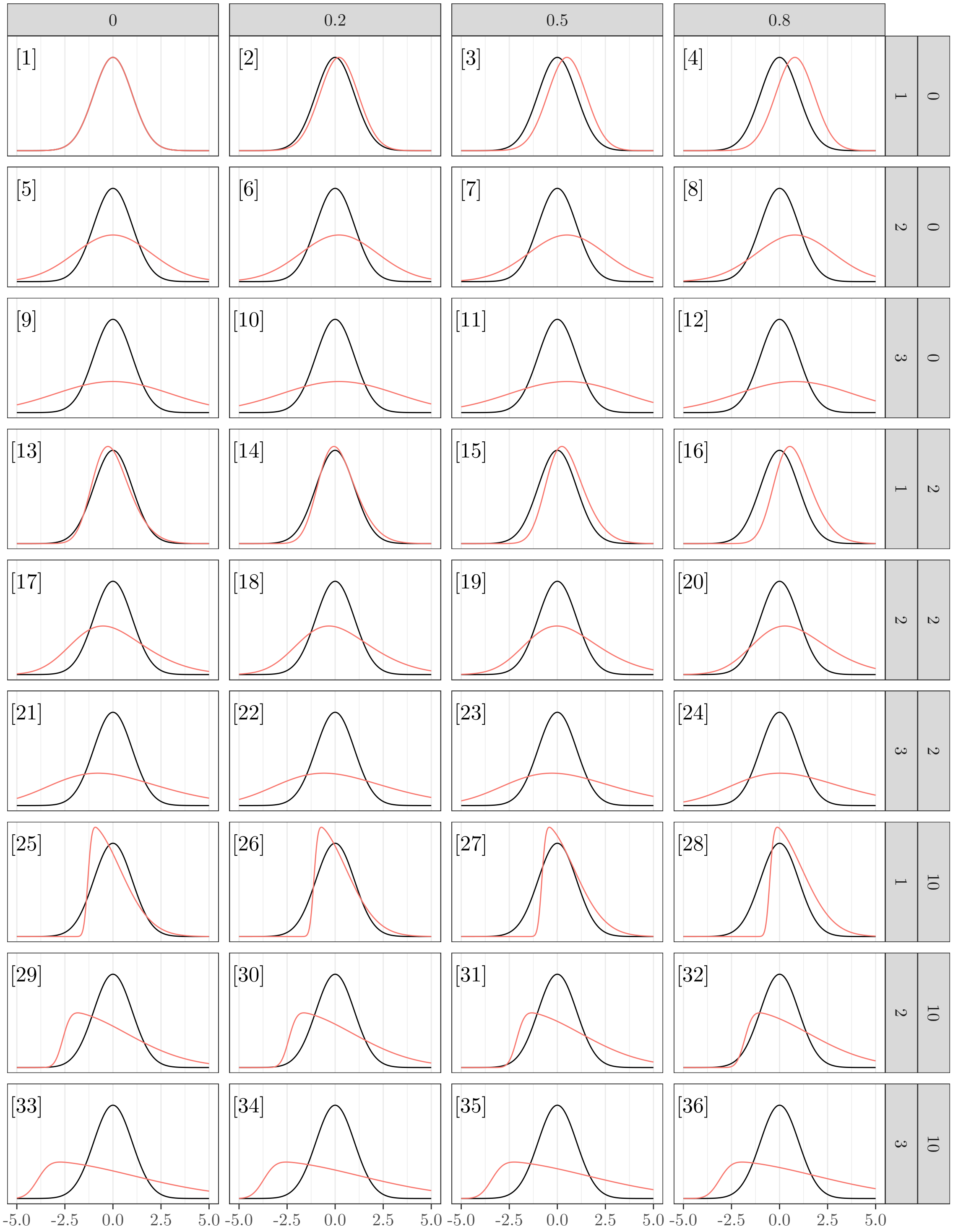

Figure S2: Generative data distributions in function of δ (column panels), σ and α (row panels). The black curves are the first population, $\mathcal{SN}(0, 1, 0)$, the red ones represent second population, $\mathcal{SN}(\xi, \omega, \alpha)$.

1.3 Results

The final data set had 540,000 × 23 variables. We included the following informations:

1	mx1	sample 1 mean
2	sx1	sample 1 standard deviation
3	mx2	sample 2 mean
4	sx2	sample 2 standard deviation
5	eta1	type I overlapping index
6	eta2	type II overlapping index
7	n	sample size
8	delta	true mean difference (i.e. mean of second population)
9	alpha	true skewness of second population
10	omega	true scale parameter of second population
11	true_overlap	true overlapping between the two populations
12	t_pval	t-test p-value
13	welch_pval	Welch-test p-value
14	wilcox_pval	Wilcoxon-Mann-Whitney-test p-value
15	vartest_pval	F-test p-value
16	zeta_perm_pval	ζ_{OV}-test p-value
17	mean_perm_pval	t-test via permutation p-value
18	F_perm_pval	F-test via permutation p-value
19	x1_norm_pval	Shapiro-test in sample 1 p-value
20	x2_norm_pval	Shapiro-test in sample 2 p-value
21	ks_test_pval	Kolmogorov-Smirnov-test p-value
22	mu	true mean of second population
23	sigma	true standard deviation of second population

In the paper, we considered only the following variables: delta, sigma, alpha, and n (representing the experimental conditions) and t_pval, welch_pval, wilcox_pval, zeta_perm_pval, vartest_pval, and ks_test_pval (representing the statistical tests under consideration).

1.3.1 Simulation check

Table S1 reports the means of means and standard deviations of the 540,000 simulated samples. The first sample was extracted from a $Normal(0, 1)$, consequently the mean (mx1) and standard deviation (sx1) are always close to 0 and 1, respectively. The parameters μ and σ represent the mean and standard deviation of the second population from which the second sample was extracted. The mean of means (mx2) and standard deviation (sx2) are close to the expected values μ and σ.

μ	mx1	mx2	σ	sx1	sx2
0	-0.00	0.00	1	0.99	0.99
0.2	0.00	0.20	2	0.99	1.98
0.5	0.00	0.50	3	0.99	2.97
0.8	-0.00	0.80			

Table S1: Means of means and standard deviations of the 540000 simulated samples. μ and σ are the true mean and standard deviation of the second population (the first population has $\mu = 0$ and $\sigma = 1$), mx1 and mx2 are the means of means in the first and second sample, respectively, sx1 and sx2 are the means of standard deviations.

1.3.2 Type I error and power

Figure S3 replicates the figure presented in the paper. To obtain this figure, we used the following algorithm:

1. For each test, we created a dummy variable indicating with `TRUE` a statistically significant result ($p \leq .05$) and with `FALSE` a non-significant result ($p > .05$).
2. We created a dummy variable indicating the experimental conditions for which the Null hypothesis is `TRUE` (i.e., conditions where there is no real difference between the populations being compared: $\mu_1 = \mu_2 = \delta = 0$, $\sigma_1 = \sigma_2 = \sigma = 1$, $\alpha_1 = \alpha_2 = \alpha = 0$). Note that this condition is illustrated in panel [1] of Figure S1.
3. We aggregated results by calculating the proportion of significant results for each condition to obtain type I error and power curves for each test.

In Figure S3, panel [A] shows the type I error curves as a function of sample size (n), while panel [B] shows the power curves as a function of sample size (n).

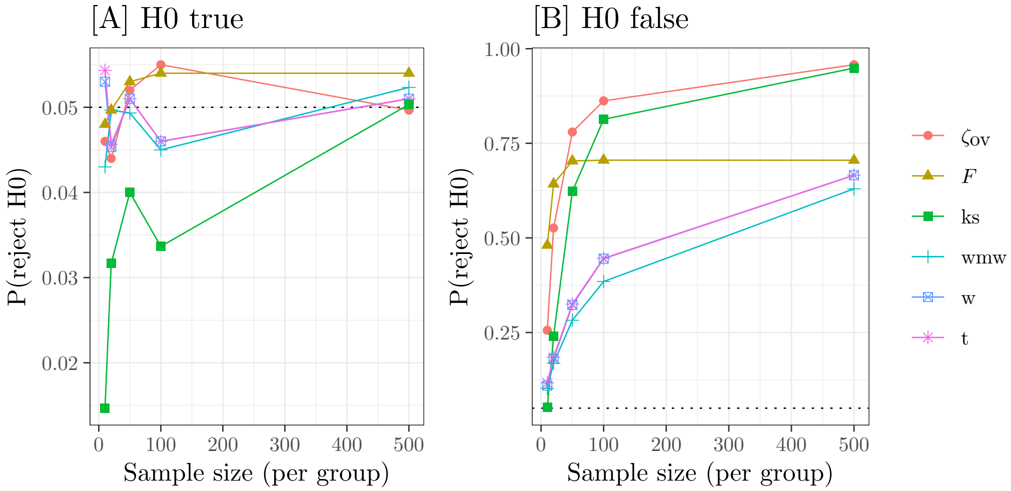

Figure S3: Control of type I error [A] and power [B] in the various tests. Note: ζ_{OV} = ζ-Overlapping test, F = variance test, ks = Kolmogorov-Smirnov test, wmw = Wilcoxon-Mann-Whitney test, w = Welch test, t = Student's t test.

2 Overlapping vs overall test

Usually, researchers are mainly interested in the difference in means, but it's important to remember that, for comparing means, variances should be homogeneous and data should be normally distributed. This implies that two additional tests need to be performed beforehand. The ζ_{OV} test is essentially an overall test that considers the means, variances, and shapes of the two groups simultaneously. This guarantees good control of Type I error and a sufficient level of power (see Fig. S3) without particular assumptions.

Now, let's consider three separate tests: the t-test (for comparing means), the F-test (for comparing variances), and the KS-test (for comparing distributions). We'll also consider an overall test that is statistically significant if at least one of these three tests results in significance at the $\alpha = .05$ level. If the aim is to compare means, ideally, the F-test and the KS-test should not be statistically significant. This does not means that you can support the Null hypothesis, but simply that based on your data you don't have evidence to reject it (Wilkinson & the Task Force on Statistical Inference, 1999).

	FALSE	TRUE
t	0.66	0.34
F	0.37	0.63
KS	0.48	0.52
overall	0.21	0.79

Table S2: Proportion of statistically significant (TRUE) and non-significant (FALSE) results for the three separate tests and the overall test ($N = 540,000$).

Table S2 reports the proportions of statistically significant (TRUE) and non-significant (FALSE) results for the three separate tests and the overall test. You should note that the proportion of significant results for the overall test is the highest (0.79), meaning that in approximately 79% of cases, at least one of the three separate tests was significant.

2.1 Familywise Error Rate

Now, we generate three dummy columns indicating the Null hypothesis status (TRUE or FALSE) for the three separate tests, and one column for the overall test for which the Null is TRUE when all three Nulls are TRUE; i.e. $\delta = 0$, $\sigma = 1$ and $\alpha = 0$.

Table S3 reports the Overall test results. By rows, we can read the proportion of tests that are statistically significant (sig.) or not significant (non-sig.); by columns, we can read the proportion of cases in which the overall Null hypothesis is TRUE or FALSE.

	H_0 status	
Test result	FALSE	TRUE
non-sig.	0.19	0.89
sig.	0.81	0.11

Table S3: Overall test results. Table rows report the proportion of tests that are statistically significant (sig.) or not significant (non-sig.); table columns report the proportion of cases in which the overall Null hypothesis is TRUE or FALSE.

From this table, it is clear that the overall test does not control for the Familywise Type I error: the proportion of false alarms (i.e., significant results when the Null is TRUE) is 0.11. This result is independent of sample size; in the five experimental chosen n values (10, 20, 50, 100, and 500), the proportion of false alarms lies in the interval $[0.101, 0.118]$.

The ζ_{OV}-test, instead, controls well the Type I error for each different sample size; the proportion of false alarms with respect sample size ranges from 0.044 to 0.055, as you can see in Fig. S3[A].

2.2 Adjusting p-values

Based on these considerations, we need to adjust the p-values when performing the three tests separately, for example by using the Bonferroni correction (see Bonferroni, 1936). In this case (three tests), the adjustment can be performed with the formula: $p_{\text{adj}} = \min(3p, 1)$, where p indicates the original p-value.

Table S4 reports the same information as Table S3 after Bonferroni adjustment has been applied. By rows, we have the proportions of tests that are statistically significant (sig.) or not significant (non-sig.); by columns, we have the proportions of cases in which the overall Null hypothesis is TRUE or FALSE. Now, Type I error is controlled at the .05 alpha level, but at the same time, the power, which is the proportion of significant results when the Null is FALSE, has decreased from 0.81 to 0.74.

However, this does not take into account the fact that if we are interested in the difference between the means, the other two tests (F and KS) should be non-significant. Therefore, we need to define an indicator that takes into account the outcome of the t-test conditional on the outcome of the other

	H_0 status	
Test result	FALSE	TRUE
non-sig.	0.26	0.96
sig.	0.74	0.04

Table S4: Overall Bonferroni-adjusted test results. Table rows report the proportion of tests that are statistically significant (sig.) or not significant (non-sig.); table columns report the proportion of cases in which the overall Null hypothesis is TRUE or FALSE.

two tests. In other words, power relative to the t-test is the probability of correctly rejecting the Null hypothesis of no difference between the means when sigma and shape are the same in the two samples.

In Figure S4 are represented the probability of rejecting t-test Null hypothesis after Bonferroni adjustment as a function of sample size (per group). 0 indicates the scenario in which t-test assumptions are met and there are no mean difference (i.e. H_0 is TRUE); all these probabilities are under the .05 alpha level, meaning that the Type I error is under control. 1 indicates the scenario in which t-test assumptions are met and there is a difference in means (H_0 is FALSE); this curve represents the actual power level and is generally lower than the power of the ζ_{OV}-test (represented by the dashed line). 2 indicates the scenario in which t-test assumptions are not met and there is a difference in means (H_0 is FALSE); it represents the case in which we detect a difference between means but without respecting the correct conditions for performing the t-test. Finalli, the dashed line (ζ_{OV}) represents the ζ-Overlapping test power curve; we remember that this test evaluate simultaneously means, variances, and shapes, and do not have any assumptions.

In Figure S4, the probability of rejecting the t-test Null hypothesis after Bonferroni adjustment is represented as a function of sample size (per group). 0 indicates the scenario in which t-test assumptions are met and there is no mean difference (i.e., H_0 is TRUE); all these probabilities are under the .05 alpha level, meaning that the Type I error is under control. 1 indicates the scenario in which t-test assumptions are met and there is a difference in means (H_0 is FALSE); this curve represents the actual

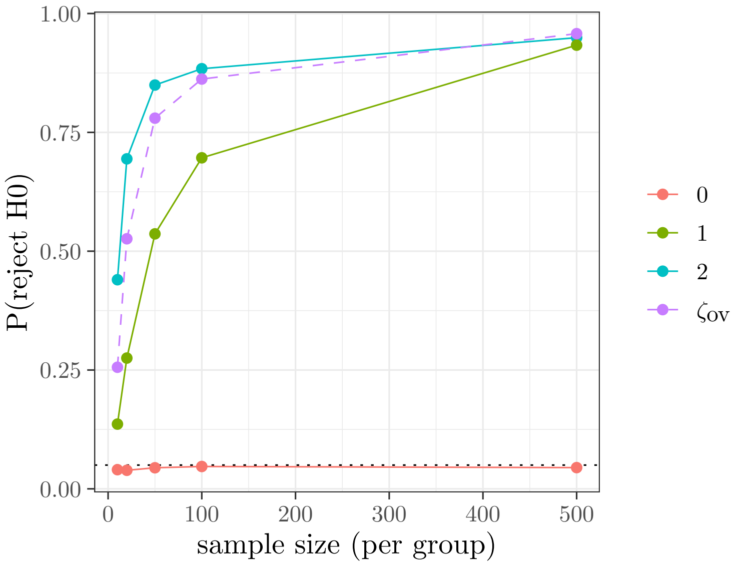

Figure S4: Probability of rejecting t-test Null hypothesis after Bonferroni adjustment as a function of sample size (per group). 0 = scenario in which t-test assumptions are respected and there are no mean difference (H_0 is TRUE); 1 = scenario in which t-test assumptions are respected and there is a difference in means (H_0 is FALSE); 2 = scenario in which t-test assumptions are not respected and there is a difference in means (H_0 is FALSE); ζ_{OV} = ζ overlapping test power curve.

power level and is generally lower than the power of the ζ_{OV}-test (represented by the dashed line). 2 indicates the scenario in which t-test assumptions are not met and there is a difference in means (H_0 is FALSE); it represents the case in which we detect a difference between means but without respecting the correct conditions for performing the t-test. Finally, the dashed line (ζ_{OV}) represents the ζ-Overlapping test power curve; we recall that this test evaluates simultaneously means, variances, and shapes, and does not have any assumptions.

2.3 Assumptions and type I error and power

In the top row of figure S5 are represented type I error and power for cases in which assumptions are respected. The patter is similar to the scenario in **??** where there was no distinction for the assumptions, confirming the good control of type I error of the ζ perm test and greater power of the test in comparison to the others. As not all tests that we performed imply assumptions, we only computed type I error and power for those tests that can have the assumptions violated (t test, F test, Welch test). What emerges is a bad control of type I error of the F test.

```
Error in eval(predvars, data, env):  object ’zeta_perm_H0_true’ not found
Error in eval(predvars, data, env):  object ’t_pval_sig’ not found
Error:  object ’PW2’ not found
Error:  object ’PW2’ not found
Error:  object ’PW2’ not found
Error in eval(predvars, data, env):  object ’welch_pval_sig’ not found
Error:  object ’PW2’ not found
Error:  object ’PW2’ not found
Error:  object ’PW2’ not found
Error in eval(predvars, data, env):  object ’wilcox_pval_sig’ not found
Error:  object ’PW2’ not found
Error:  object ’PW2’ not found
Error:  object ’PW2’ not found
Error in eval(predvars, data, env):  object ’vartest_pval_sig’ not found
Error:  object ’PW2’ not found
Error:  object ’PW2’ not found
Error:  object ’PW2’ not found
Error in eval(predvars, data, env):  object ’ks_test_pval_sig’ not found
Error:  object ’PW2’ not found
Error:  object ’PW2’ not found
Error:  object ’PW2’ not found
```

```
‘geom_line()‘:  Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
```

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 ζ 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 ζ **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

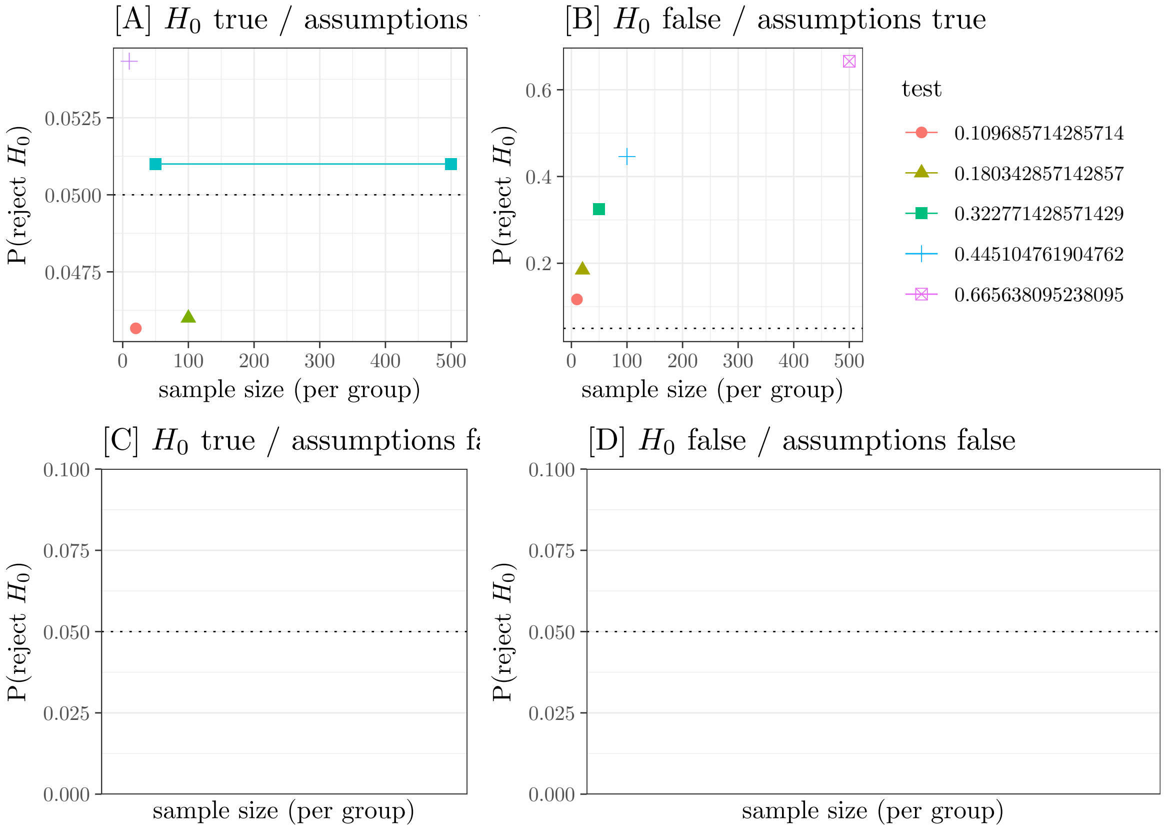

Figure S5: Control of type I error and power for scenarios in which assumptions are respected (top panels) and when they are not (bottom panels).

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 ζ **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 ζ **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.

References

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

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, *8*, 3–62.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Used R packages

- `brms`. Bürkner P (2017). "brms: An R Package for Bayesian Multilevel Models Using Stan." _Journal of Statistical Software_, *80*(1), 1-28. doi:10.18637/jss.v080.i01 https://doi.org/10.18637/jss.v080.i01. Bürkner P (2018). "Advanced Bayesian Multilevel Modeling with the R Package brms." _The R Journal_, *10*(1), 395-411. doi:10.32614/RJ-2018-017 https://doi.org/10.32614/RJ-2018-017. Bürkner P (2021). "Bayesian Item Response Modeling in R with brms and Stan." _Journal of Statistical Software_, *100*(5), 1-54. doi:10.18637/jss.v100.i05 https://doi.org/10.18637/jss.v100.i05.
- `cowplot`. Wilke C (2024). _cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'_. R package version 1.1.3, https://CRAN.R-project.org/package=cowplot.
- `ggplot2`. Wickham H (2016). _ggplot2: Elegant Graphics for Data Analysis_. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.
- `knitr`. Xie Y (2025). _knitr: A General-Purpose Package for Dynamic Report Generation in R_. R package version 1.50, https://yihui.org/knitr/. Xie Y (2015). _Dynamic Documents with R and knitr_, 2nd edition. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 978-1498716963, https://yihui.org/knitr/. Xie Y (2014). "knitr: A Comprehensive Tool for Reproducible Research in R." In Stodden V, Leisch F, Peng RD (eds.), _Implementing Reproducible Computational Research_. Chapman and Hall/CRC. ISBN 978-1466561595.

- `R`. R Core Team (2024). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

- `Rcpp`. Eddelbuettel D, Francois R, Allaire J, Ushey K, Kou Q, Russell N, Ucar I, Bates D, Chambers J (2025). _Rcpp: Seamless R and C++ Integration_. R package version 1.0.14, https://CRAN.R-project.org/package=Rcpp. Eddelbuettel D, François R (2011). ”Rcpp: Seamless R and C++ Integration.” _Journal of Statistical Software_, *40*(8), 1-18. doi:10.18637/jss.v040.i08 https://doi.org/10.18637/jss.v040.i08. Eddelbuettel D (2013). _Seamless R and C++ Integration with Rcpp_. Springer, New York. doi:10.1007/978-1-4614-6868-4 https://doi.org/10.1007/978-1-4614-6868-4, ISBN 978-1-4614-6867-7. Eddelbuettel D, Balamuta J (2018). ”Extending R with C++: A Brief Introduction to Rcpp.” _The American Statistician_, *72*(1), 28-36. doi:10.1080/00031305.2017.137599 https://doi.org/10.1080/00031305.2017.1375990.

- `report`. Makowski D, Lüdecke D, Patil I, Thériault R, Ben-Shachar M, Wiernik B (2023). ”Automated Results Reporting as a Practical Tool to Improve Reproducibility and Methodological Best Practices Adoption.” _CRAN_. https://easystats.github.io/report/.

- `scales`. Wickham H, Pedersen T, Seidel D (2025). _scales: Scale Functions for Visualization_. R package version 1.4.0, https://CRAN.R-project.org/package=scales.

- `sn`. Azzalini AA (2023). _The R package `sn`: The skew-normal and related distributions such as the skew-*t* and the SUN (version 2.1.1)._. Home page: http://azzalini.stat.unipd.it/SN/, https://cran.r-project.org/package=sn.

- `xtable`. Dahl D, Scott D, Roosen C, Magnusson A, Swinton J (2019). _xtable: Export Tables to LaTeX or HTML_. R package version 1.8-4, https://CRAN.R-project.org/package=xtable.

Session Info

```
R version 4.4.0 (2024-04-24)
Platform: x86_64-apple-darwin20
Running under: macOS Sonoma 14.3

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.4-x86_64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.4-x86_64/Resources/lib/libRlapack.dylib;  LAPACK v

attached base packages:
[1] stats4    stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
[1] report_0.5.8  cowplot_1.1.3 xtable_1.8-4  brms_2.21.0   Rcpp_1.0.14
[6] scales_1.4.0  ggplot2_4.0.1 sn_2.1.1      knitr_1.50

loaded via a namespace (and not attached):
 [1] gtable_0.3.6         tensorA_0.36.2.1     xfun_0.52
 [4] QuickJSR_1.1.3       insight_1.4.2        inline_0.3.19
 [7] lattice_0.22-6       numDeriv_2016.8-1.1  vctrs_0.6.5
[10] tools_4.4.0          generics_0.1.4       datawizard_1.3.0
[13] curl_6.4.0           parallel_4.4.0       tibble_3.3.0
[16] highr_0.11           pkgconfig_2.0.3      Matrix_1.7-0
[19] checkmate_2.3.1      RColorBrewer_1.1-3   S7_0.2.1
```

```
[22] distributional_0.4.0 RcppParallel_5.1.7   filehash_2.4-5
[25] lifecycle_1.0.4      compiler_4.4.0       farver_2.1.2
[28] stringr_1.5.1        Brobdingnag_1.2-9    tinytex_0.57
[31] mnormt_2.1.1         codetools_0.2-20     bayesplot_1.11.1
[34] pillar_1.10.2        StanHeaders_2.32.7   bridgesampling_1.1-2
[37] abind_1.4-5          nlme_3.1-164         posterior_1.5.0
[40] rstan_2.32.6         tidyselect_1.2.1     digest_0.6.37
[43] mvtnorm_1.2-5        stringi_1.8.7        dplyr_1.1.4
[46] labeling_0.4.3       grid_4.4.0           colorspace_2.1-0
[49] cli_3.6.5            magrittr_2.0.4       loo_2.7.0
[52] pkgbuild_1.4.8       withr_3.0.2          backports_1.5.0
[55] estimability_1.5.1   matrixStats_1.3.0    emmeans_1.10.2
[58] gridExtra_2.3        coda_0.19-4.1        evaluate_1.0.5
[61] V8_4.4.2             rstantools_2.4.0     rlang_1.1.6
[64] glue_1.8.0           rstudioapi_0.17.1    tikzDevice_0.12.6
[67] jsonlite_2.0.0       R6_2.6.1
```