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

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.

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. The detailed description of the simulation study can be found in the main draft.

1.1 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.

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) $\times$ 3 (standard deviation differences) $\times$ 3 (shape differences) $\times$ 5 (sample sizes) $\times$ 3000 (replications) = 540,000 datasets as well as 3,240,000 of statistical tests and corresponding p-values.

1.2 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).

1.2.1 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.

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. **??**) 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 (?, ?).

	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. **??**[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 ?, ?). 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.

	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.

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 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 S1 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

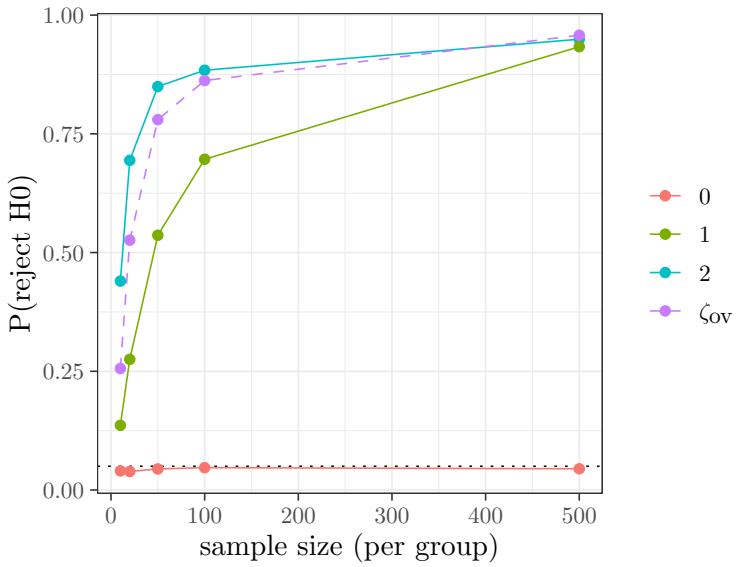

Figure S1: Probability of rejecting t-test Null hypothesis after Bonferroni adjustment as a function of sample size (per group). 0 = scenario in which t-test assumptions are respected and there are no mean difference (H_0 is TRUE); 1 = scenario in which t-test assumptions are respected and there is a 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); $\zeta_{\text{ov}} = \zeta$ overlapping test power curve.

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 S1, 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 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.

References

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.

- `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 14.8.4

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  sn_2.1.1      knitr_1.50

loaded via a namespace (and not attached):
 [1] bridgesampling_1.1-2 tensorA_0.36.2.1     generics_0.1.4
 [4] stringi_1.8.7        lattice_0.22-6       digest_0.6.37
 [7] magrittr_2.0.4       estimability_1.5.1   evaluate_1.0.5
[10] grid_4.4.0           RColorBrewer_1.1-3   mvtnorm_1.2-5
[13] jsonlite_2.0.0       filehash_2.4-5       Matrix_1.7-0
[16] pkgbuild_1.4.8       backports_1.5.0      gridExtra_2.3
[19] Brobdingnag_1.2-9    QuickJSR_1.1.3       codetools_0.2-20
[22] numDeriv_2016.8-1.1  abind_1.4-5          mnormt_2.1.1
[25] cli_3.6.5            rlang_1.1.7          StanHeaders_2.32.7
[28] datawizard_1.3.0     inline_0.3.19        tools_4.4.0
[31] rstan_2.32.6         parallel_4.4.0       rstantools_2.4.0
[34] checkmate_2.3.1      coda_0.19-4.1        dplyr_1.1.4
[37] colorspace_2.1-0     ggplot2_4.0.2        tikzDevice_0.12.6
[40] curl_6.4.0           vctrs_0.7.1          posterior_1.5.0
[43] R6_2.6.1             emmeans_1.10.2       matrixStats_1.3.0
[46] lifecycle_1.0.5      stringr_1.5.1        V8_4.4.2
```

```
[49] insight_1.4.2        pkgconfig_2.0.3      RcppParallel_5.1.7
[52] pillar_1.10.2        gtable_0.3.6         loo_2.7.0
[55] glue_1.8.0           xfun_0.52            tibble_3.3.0
[58] tidyselect_1.2.1     highr_0.11           rstudioapi_0.17.1
[61] farver_2.1.2         bayesplot_1.11.1     nlme_3.1-164
[64] compiler_4.4.0       S7_0.2.1             distributional_0.4.0
```