The Journal of Open Source Software

robnptests – An R package for robust two-sample location and variability tests

Sermad Abbas[1], **Barbara Brune**[2], **and Roland Fried**[1]

1 TU Dortmund University, Faculty of Statistics, 44221 Dortmund, Germany **2** Technical University of Vienna, Institute of Statistics and Mathematical Methods in Economics, 1040 Vienna, Austria

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

The package `robnptests` is a compilation of two-sample tests selected by two criteria: The tests are (i) robust against outliers and (ii) (approximately) distribution free. Regarding the latter aspect, we implemented tests that keep an intended significance level and provide a reasonably high efficiency, both under a variety of continuous distributions. Robustness is achieved by using test statistics that are based on robust location and scale measures.

In what follows, we give a brief description of the package's contents. More details can be found in the introductory vignette of the package, which can be called by `vignette("robnptests")`, and in the cited papers.

Data situation

We consider two samples of independent and identically distributed (i.i.d.) random variables $X_1, ..., X_m$ and $Y_1, ..., Y_n$, respectively. The underlying distributions are assumed to be continuous with cumulative distribution functions F_X and F_Y.

The tests can be used for either of the following scenarios:

- Two-sample location problem: Assuming that both distributions are equal except that F_Y may be a shifted version of F_X, i.e. $F_X(x) = F_Y(x + \Delta)$ for all $x \in \mathbb{R}$ and some $\Delta \in \mathbb{R}$, the tests can be used to detect such a shift.
- Two-sample scale problem: In case of a difference only in scale, i.e. $F_X(x) = F_Y(x/\theta)$ for some $\theta > 0$, a transformation of the observations enables to identify differing scale parameters.

Statement of need

A popular test for the location setting is the two-sample t-test. It is considered to be robust against deviations from the normality assumption because it keeps the specified significance level asymptotically due to the central limit theorem in case of finite variances. However, non-normality can result in a loss of power (Wilcox, 2003). In addition, the t-test is vulnerable to outliers (Fried & Dehling, 2011). Distribution-free tests, like the two-sample Wilcoxon rank-sum test, can be nearly as powerful as the t-test under normality and may have higher power under non-normality. Still, they also can be vulnerable to

The Journal of Open Source Software

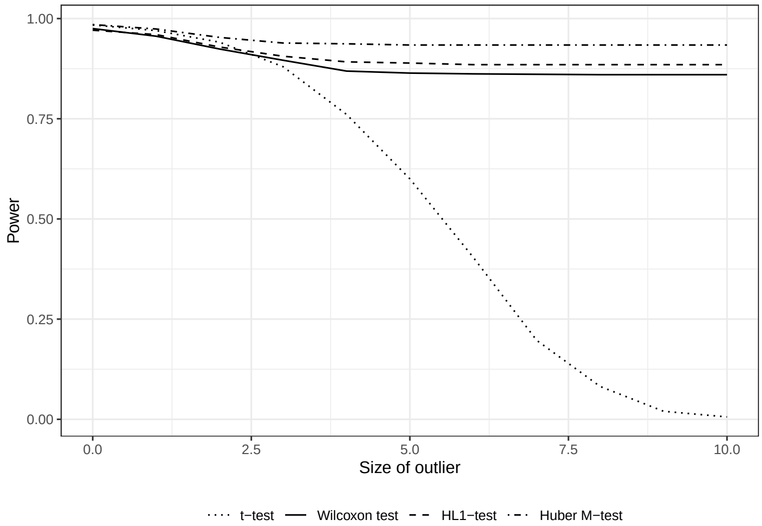

Figure 1: Power of the two-sample t-test, the Wilcoxon rank-sum test, and two robust tests - one based on the one-sample Hodges-Lehmann estimator and one based on Huber's M-estimator - on two samples of size $m = n = 10$ from two normal distributions with unit variance, a location difference of $\Delta = 2$, and an additive single outlier of increasing size.

outliers, particularly for small samples (Fried & Gather, 2007). The two-sample tests in `robnptests` combine (approximate) distribution independence and robustness against outliers. Thus, they are well suited for outlier-corrupted samples from unknown data-generating distributions. At the same time, such tests can be nearly as powerful as popular procedures like the aforementioned t-test or the Wilcoxon test on uncontaminated samples for somewhat larger computational costs.

Figure 1 compares the power of the t-test, the Wilcoxon test and two robust tests. The HL1-test is based on the one-sample Hodges-Lehmann estimator (Hodges & Lehmann, 1963) and the Huber M-test uses Huber's M-estimator (Huber, 1964). We consider a fixed location difference between the samples and a single outlier of increasing size. The power of the t-test decreases to zero, while the loss in power of the Wilcoxon test and both robust tests is small. The robust tests provide a somewhat higher power than the Wilcoxon test. The differences between the Wilcoxon test and robust tests like those in the example can become larger when more outliers are involved (Fried & Dehling, 2011).

Common parametric and non-parametric tests for scale differences have similar problems as described above for the location tests. In addition, some non-parametric tests for the scale problem do not cope well with asymmetry. A possible solution while retaining the robustness is to apply the robust location tests to transformed observations as proposed by Fried (2012). Such tests can yield good results in terms of power and size under both asymmetry and outlier corruption. However, the tests may be less efficient under symmetry than classical procedures like the Mood test.

Implemented two-sample tests

Each test statistic consists of a robust estimator for the location difference between the two populations that are compared. This difference is divided by a robust estimator for the within-sample variability.

To obtain a distribution-free test decision, the p-value can be computed by using the permutation principle, the randomization principle, or a normal approximation. With the permutation principle, the tests hold the desired significance level exactly at the cost of large computing times even for quite small samples such as $m = n = 10$. The time can be reduced by using a randomization distribution and, even more, by taking advantage of the asymptotic normality of the location-difference estimators. The latter approach, however, is only advisable for large sample sizes $m, n > 30$.

The tests based on the following location estimators are described in Fried & Dehling (2011):

- The *difference of the sample medians* achieves high robustness. However, this estimator is not very efficient under the normal distribution and distributions close to it.
- To improve the efficiency, one can use the difference between the *one-sample Hodges-Lehmann estimators* (Hodges & Lehmann, 1963) at the cost of losing some robustness due to the lower breakdown point.
- Similarly, the *two-sample Hodges-Lehmann estimator* leads to a robust test with a higher power under normality than the tests based on the sample median and similar robustness.

For scaling, we use different estimators based on medians and pairwise differences, see Fried & Dehling (2011) for a detailed description.

In addition, we implemented tests based on M-estimators. This approach to robust location estimation allows for flexibility in how outliers are treated through the specification of the tuning constants of the corresponding ρ-function. We focus on Huber's ρ-function, the bisquare function and the Hampel ρ-function. The estimator for the within-sample variance is a pooled estimator derived from the asymptotic normality of the M-estimators (Maronna et al., 2006, p. 36ff). Moreover, the package contains Yuen's t-test which uses the difference of *trimmed means* to estimate the location difference and a scale estimator based on the pooled winsorized variances (Yuen & Dixon, 1973).

In case of data with many ties (e.g. caused by discrete sampling), the ties may carry over to the permutation distribution. This can happen in real-world applications when the measurements are rounded or stem from discrete distributions and may lead to a loss in power or conservative tests. Additionally, the robust scale estimators may become zero, so that the test statistic cannot be calculated. Both issues can be addressed by adding random noise from a uniform distribution with a small variance to each observation ("wobbling", see Fried & Gather (2007)).

A more detailed overview of the implemented tests and corresponding test statistics can be found in the vignette `vignette("robnptests")`.

Applications

Besides conventional two-sample problems, the tests can be used for the online detection of structural breaks in outlier-contaminated time series. Abbas et al. (2016) describe how intensity changes in image sequences generated by a virus sensor are automatically detected by applying the tests to the individual pixel time series of the sequence. Moreover, the test statistics can be used as control statistics for robust, (approximately) distribution-free control charts for time series with a time-varying signal (Abbas & Fried, 2017, 2020). In Abbas et al. (2019), the tests are applied to detect unusual sequences in time series of crack widths in concrete beams by searching for sudden scale changes.

Other packages with robust two-sample tests

The CRAN Task View for robust statistical methods currently lists the three packages `WRS2` (Mair & Wilcox, 2019), `walrus` (Love & Mair, 2018), and `robeth` (Marazzi, 2020) that explicitly deal with robust hypothesis tests for the two-sample problem. `WRS2` contains a large collection of robust procedures which are presented in the book *Introduction to Robust Estimation and Hypothesis Testing* by (Wilcox, 2012). `walrus` provides a different user interface for the functions in `WRS2`. The package `robeth` (Marazzi, 2020) contains some robust tests for linear hypotheses.

The functions in `WRS2` concentrate on the heteroscedastic setting, whereas our focus lies on the homoscedastic case. The reason is that especially for small samples, estimating the within-sample variance separately for both samples, as is the case under heteroscedasticity, may lead to unreliable estimates. Moreover, choosing equal sample sizes $m = n$, can protect against a deteriorating test performance of our implemented tests under heteroscedasticity in terms of size and power (Staudte & Sheather, 1990, p. 179).

The R package `nptest` (Helwig, 2021) contains nonparametric versions of the two-sample t-test, which is realized by using the permutation and randomization principles described above on the t-statistic. This principle has also been studied in Abbas & Fried (2017), and, while achieving distribution independence, the test statistic lacks robustness against outliers.

References

Abbas, S., & Fried, R. (2017). Control charts for the mean based on robust two-sample tests. *Journal of Statistical Computation and Simulation*, *87*(1), 138–155. https://doi.org/10.1080/00949655.2016.1194839

Abbas, S., & Fried, R. (2020). Robust control charts for the mean of a locally linear time series. *Journal of Statistical Computation and Simulation*, *90*(15), 2741–2765. https://doi.org/10.1080/00949655.2020.1788562

Abbas, S., Fried, R., & Gather, U. (2016). Detection of local intensity changes in grayscale images with robust methods for time-series analysis. In S. Michaelis, N. Piatkowski, & M. Stolpe (Eds.), *Solving large scale learning tasks: Challenges and algorithms: Essays dedicated to katharina morik on the occasion of her 60th birthday* (pp. 251–271). Springer. https://doi.org/10.1007/978-3-319-41706-6

Abbas, S., Fried, R., Heinrich, J., Horn, M., Jakubzik, M., Kohlenbach, J., Maurer, R., Michels, A., & Müller, C. H. (2019). Detection of anomalous sequences in crack data of a bridge monitoring. In N. Bauer, K. Ickstadt, K. Lübke, G. Szepannek, H. Trautmann, & M. Vichi (Eds.), *Applications in statistical computing: From music data analysis to industrial quality improvement* (pp. 251–269). Springer. https://doi.org/10.1007/978-3-030-25147-5_16

Fried, R. (2012). On the online estimation of piecewise constant volatilities. *Computational Statistics & Data Analysis*, *56*(11), 3080–3090. https://doi.org/10.1016/j.csda.2011.02.012

Fried, R., & Dehling, H. (2011). Robust nonparametric tests for the two sample location problem. *Statistical Methods & Applications*, *20*(4), 409–422. https://doi.org/10.1007/s10260-011-0164-1

Fried, R., & Gather, U. (2007). On rank tests for shift detection in time series. *Computational Statistics & Data Analysis*, *52*, 221–233. https://doi.org/10.1016/j.csda.2006.12.017

Helwig, N. E. (2021). *Nptest: Nonparametric bootstrap and permutation tests*. https://CRAN.R-project.org/package=nptest

Hodges, J. L., & Lehmann, E. L. (1963). Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, *34*(2), 598–611. https://doi.org/10.1214/aoms/

1177704172

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, *35*(1), 73–101. https://doi.org/10.1214/aoms/1177703732

Love, J., & Mair, P. (2018). *Walrus: Robust statistical methods*. https://CRAN.R-project.org/package=walrus

Mair, P., & Wilcox, R. (2019). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01246-w

Marazzi, A. (2020). *Robeth: R functions for robust statistics*. https://CRAN.R-project.org/package=robeth

Maronna, R. A., Martin, D. R., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. Wiley, Chichester. https://doi.org/10.1002/0470010940

Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. Wiley. https://doi.org/10.1002/9781118165485

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. Academic Press. https://doi.org/10.1016/B978-0-12-751541-0.X5021-4

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (Fourth). Elsevier. https://doi.org/10.1016/C2010-0-67044-1

Yuen, K. K., & Dixon, W. T. (1973). The approximate behaviour and performance of the two-sample trimmed t. *Biometrika*, *60*(2), 369–374. https://doi.org/10.2307/2334550