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

26 September 2021

Summary

The package `robnptests` is a compilation of two-sample tests, which is based on 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.

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 are designed for either one of the following cases:

1. Under the assumption that the scale parameters of both distributions are equal, i.e. $F_X(x) = F_Y(x+\Delta)$, $\Delta \in \mathbb{R}$, the tests can be used to detect a location difference between the samples.
2. If the location parameters are equal with unequal scale parameters, i.e. $F_X(x) = F_Y(x/\theta)$, $\theta > 0$, the tests are designed to identify a scale difference between the samples.

In practice, observations may be rounded to a few decimal places. This may cause a loss in efficiency or a violation of the significance level. Therefore, we implemented the idea of `wobbling` (Fried and Gather 2007), which is the addition of random noise to dampen the effect of the discretization due to rounding.

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. It be called by `vignette("robnptests-vignette")`.

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 due to the central limit theorem and holds the specified significance level. However, there can be much more powerful tests under non-normal distributions (Wilcox 2003). In addition, particularly for small samples, the test is prone to outliers (Fried and 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 surpass it under non-normal distributions. Still, they can also be vulnerable to outliers (Fried and Gather 2007). Our package contains two-sample tests which are (nearly) distribution free and robust against outliers, which makes them better suited for outlier-corrupted samples with an unknown data-generating distribution. 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.

Figure 1 compares the power of the t-test, the Wilcoxon test and two robust tests: one based on the one-sample Hodges-Lehmann estimator (HL1 test) (Hodges and Lehmann 1963), and one based on Huber's M-estimator (Huber 1964) for 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 power drop of the Wilcoxon test and both robust tests is bounded. The robust tests provide a higher power. The differences between the Wilcoxon test and our robust tests may become more apparent when more outliers are involved (Fried and Dehling 2011).

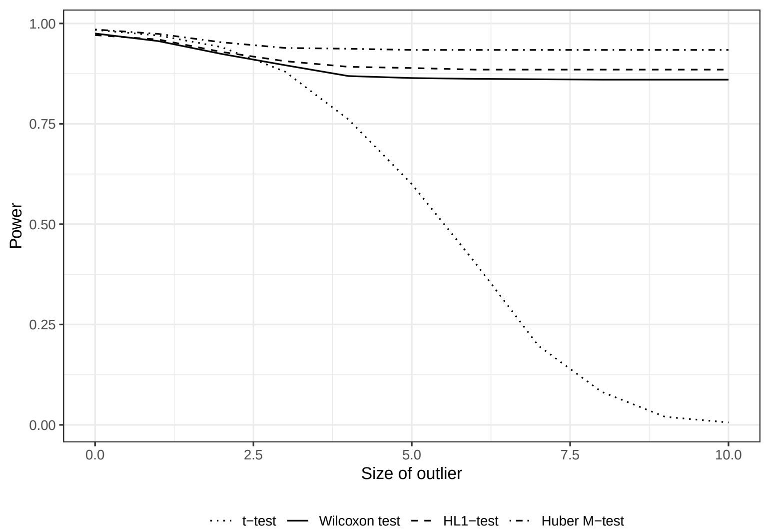

Figure 1: Power of the two-sample t-test, the Wilcoxon rank sum test and a robust test based on the one-sample Hodges-Lehmann estimator on two samples of size $m = n = 10$ from two normal distributions with a location difference of $\Delta = 2$ and a single outlier of increasing size.

When comparing variances, the tests can be more efficient under asymmetry or outliers than commonly used tests like the F-test, the Mood test, or the Ansari-Bradley tests. However, the efficiency may be lower under symmetric distributions.

Applications

The package contains tests that were used in a variety of applications to identify structural breaks in outlier-contaminated time series. In Abbas, Fried, and Gather (2016), intensity changes in image sequences generated by a virus sensor were automatically detected by applying the tests to the individual pixel time series of the sequence. Moreover, several of the implemented tests serve as a basis for (approximately) distribution-free, robust control charts for time series with a time-varying signal (Abbas and Fried 2017, 2020). Some of the tests were also applied to detect unusual sequences in time series of crack-width measurements obtained by monitoring concrete bridges (Abbas et al. 2019).

Other packages with robust two-sample tests

The CRAN Task View currently lists three packages that explicitly deal with robust hypothesis tests for the two-sample problem. `WRS2` (Mair and Wilcox 2019) contains a large collection of robust procedures which are presented in the book *Introduction to Robust Estimation and Hypothesis Testing* by (Wilcox 2012). A different user interface for the functions can be found in the package `walrus` (Love and Mair 2018). 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 and Sheather 1990, 179).

Implemented two-sample tests

Each test statistic consists of a robust estimator for the location difference between the two populations that should be compared, which is divided by a robust estimator for the within-sample variance.

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 $m, n > 30$.

One group of tests is taken from Fried and Dehling (2011). The tests make use of the following three location estimators:

- The *difference of the sample medians* helps to achieve high robustness. However, this estimator is not very efficient under the normal distribution or distributions that do not deviate too much from it.
- To improve the efficiency one can use the difference of the *one-sample Hodges-Lehmann estimators* (Hodges and Lehmann 1963) at the cost of some robustness.
- 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.

In addition to the above, we implemented tests based on the following location estimators:

- The class of *M-estimators* provides an approach to robust location estimation that allows for flexibility in how outliers are treated by specification of the parameters of the ρ-function. We focus on Huber's ρ-function, the bisquare function and the Hampel ρ-function in a similar manner as Aboukalam (1992).
- Yuen's t-test uses the difference of *trimmed means* to estimate the location difference (Yuen and Dixon 1973).

The within-sample variance is estimated by pooled scale estimators that are also robust and chosen so that they fit to the selected estimators for the location difference. A shortcoming of some of these scale estimators is that those may become zero due to observations with identical values. This happens in real-world applications when the measurements may be rounded or generated by discrete distributions. To cope with this, we add random noise from a uniform distribution with a small variance to each observation. This procedure is called `wobbling`. The goal is not to reproduce the original, unrounded values, but to enable the computation of the test statistic without distorting the observations too much.

A more detailed overview of the implemented tests may be found in the vignette.

References

Abbas, Sermad, and Roland Fried. 2017. "Control Charts for the Mean Based on Robust Two-Sample Tests." *Journal of Statistical Computation and Simulation* 87 (1): 138–55. https://doi.org/10.1080/00949655.2016.1194839.

———. 2020. "Robust Control Charts for the Mean of a Locally Linear Time Series." *Journal of Statistical Computation and Simulation* 90 (15): 2741–65. https://doi.org/10.1080/00949655.2020.1788562.

Abbas, Sermad, Roland Fried, and Ursula Gather. 2016. "Detection of Local Intensity Changes in Grayscale Images with Robust Methods for Time-Series Analysis." In *Solving Large Scale Learning Tasks: Challenges and Algorithms: Essays Dedicated to Katharina Morik on the Occasion of Her 60th Birthday*, edited by Stefan Michaelis, Nico Piatkowski, and Marco Stolpe, 251–71. Cham: Springer.

Abbas, Sermad, Roland Fried, Jens Heinrich, Melanie Horn, Mirko Jakubzik, Johanna Kohlenbach, Reinhard Maurer, Anne Michels, and Christine H. Müller. 2019. "Detection of Anomalous Sequences in Crack Data of a Bridge Monitoring." In *Applications in Statistical Computing: From Music Data Analysis to Industrial Quality Improvement*, edited by Nadja Bauer, Katja Ickstadt, Karsten Lübke, Gero Szepannek, Heike Trautmann, and Maurizio Vichi, 251–69. "Cham": Springer. https://doi.org/10.1007/978-3-030-25147-5_16.

Aboukalam, M. A. F. 1992. “Some Robust Two-Sample Test Statistics Based on M-Estimators of Location.” *Communications in Statistics - Simulation and Computation* 21 (1): 133–48.

Fried, R., and H. Dehling. 2011. “Robust Nonparametric Tests for the Two Sample Location Problem.” *Statistical Methods & Applications* 20 (4): 409–22. https://doi.org/10.1007/s10260-011-0164-1.

Fried, Roland, and Ursula Gather. 2007. “On Rank Tests for Shift Detection in Time Series.” *Computational Statistics & Data Analysis* 52: 221–33. https://doi.org/10.1016/j.csda.2006.12.017.

Hodges, Joseph L., and Erich L. Lehmann. 1963. “Estimates of Location Based on Rank Tests.” *The Annals of Mathematical Statistics* 34 (2): 598–611.

Huber, Peter 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, Jonathon, and Patrick Mair. 2018. *Walrus: Robust Statistical Methods*. https://CRAN.R-project.org/package=walrus.

Mair, Patrick, and Rand Wilcox. 2019. “Robust Statistical Methods in R Using the WRS2 Package.” *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01246-w.

Marazzi, Alfio. 2020. *Robeth: R Functions for Robust Statistics*. https://CRAN.R-project.org/package=robeth.

Staudte, Robert G., and Simon J. Sheather. 1990. *Robust Estimation and Testing*. New York: Wiley.

Wilcox, Rand R. 2003. *Applying Contemporary Statistical Techniques*. Amsterdam: Academic Press.

———. 2012. *Introduction to Robust Estimation and Hypothesis Testing*. Fourth. Amsterdam: Elsevier.

Yuen, Karen K., and W. T. Dixon. 1973. “The Approximate Behaviour and Performance of the Two-Sample Trimmed T.” *Biometrika* 60 (2): 369–74.