The Journal of Open Source Software

Purify: An R package for resampling and stratified data

Jeremy VanderDoes [1*¶] **and Yuling Max Chen** [1*]

1 Department of Statistics, University of Waterloo, Waterloo, ON, Canada ¶ Corresponding author * These authors contributed equally.

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Susan Holmes

Reviewers:

- @pachadotdev

Submitted: 12 May 2025
Published: unpublished

Summary

`Purify` is an R package designed for resampling and testing of data, aimed at researchers and practitioners who analyze complex datasets, often with unbalanced strata, multiple predictors, or a dependent output variable. This package enables users to perform robust statistical analyses by test statistics and resampling of variables with and without relation to the output and other stratification variables. The resampling analysis can be conducted on statistical summaries (e.g. mean square error), coefficient estimates of models, forecasts, or model statistics. `Purify` offers versatile resampling settings, including block, sliding window, and stratified resampling with and without replacement. These methods are also extended to cross-validation and confidence intervals. Each method can be tailored to specific data structures and research questions. With its intuitive interface and customizable options, `Purify` streamlines the process of hypothesis testing and estimation of statistical significance.

Statement of need

Unbalanced data are widely observed, yet methods for analysis can be unwieldy, overly complex, or missing implementations. `Purify` fills this gap by providing and organizing many statistical tests robust to various assumptions and an extensive collection of resampling methodology. These tests include normality, anova, and two-sample tests along with statistics to quantify stratification effects.

Resampling is a fundamental technique in estimating the distribution of statistics, testing hypotheses, and deriving confidence intervals, especially when analytical solutions are impractical. When assumptions for statistical tests are under question, resampling provides another tool to assess their effectiveness. Many R packages provide basic resampling methods but lack specialized support for complex structures. Such structures may contain dependence and variables which should not be resampled. Analysis of these structures can be unwieldy to investigate in other packages. `Purify` offers a flexible framework to enable data scientists and researchers to perform and compare targeted, customizable resampling schemes that account for data structure. The methods are compatible for user-defined functions and outside models, making `Purify` ideal for rigorous hypothesis testing and model evaluation.

The package's support for stratified and segmented sampling further allows users to address scenarios with grouped or ordered data (even under dependence), providing a critical resource for modern applied statistical research. By incorporating sophisticated resampling techniques, `Purify` enhances the robustness and reliability of statistical inferences drawn from complex and potentially unbalanced datasets.

Permutation tests can be naturally computationally intensive and speed is an important consideration throughout `Purify`. Users can use it in a variety of problems. Additional functions

provide use of resampling in the context of cross-validation and forecasting. Supplemental functions provide visualizations and summary functions to illuminate the methods and results. Detailed documentation makes `Purify` accessible to users of varying statistical understanding.

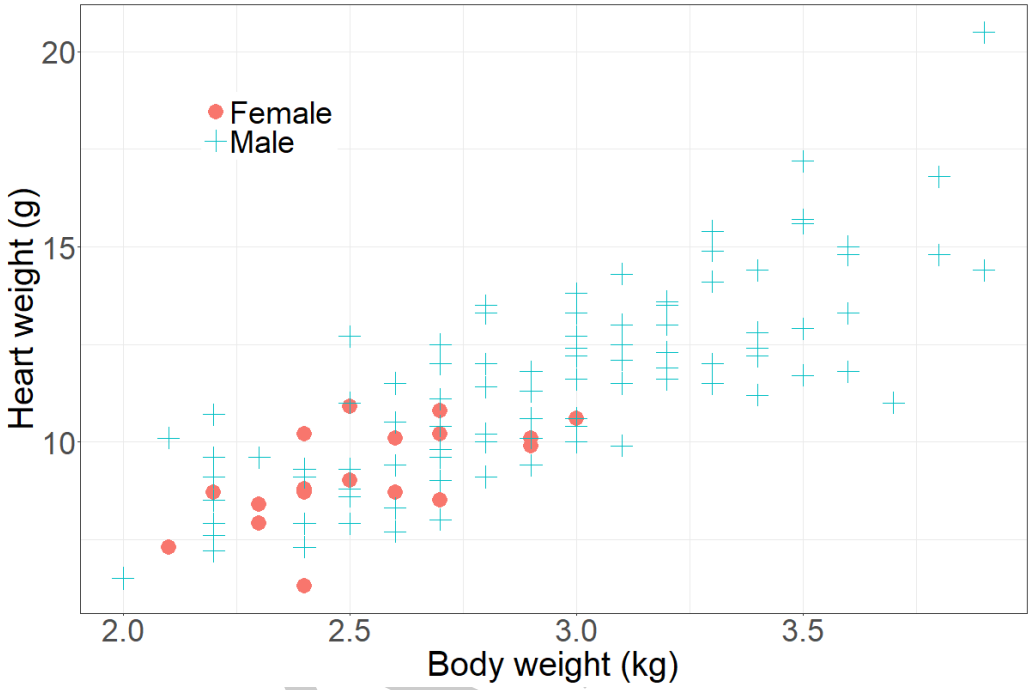

Figure 1: Subcats. Body and heart weights of cats with respect to their sex.

Package Functionality

Assessing whether the data is homogeneous in its variance, exhibits normality, or has significant differences between strata often requires extensive testing. `Purify` offers functions such as `normality_tests()`, `variance_tests()`, and `group_tests()` to investigate several test statistics at once. Resampled versions of statistics are provided to determine estimates and confidence intervals with fewer assumptions, e.g. see `resample_variance()`.

A primary function in `Purify` is `resample()`. This function offers clear input parameters to simplify the process of selecting or evaluating the correct methodology, even for user-defined functions. The multistep selection allows for combinations of dependent data, unbalanced data, and resampling to be performed with and without replacement. The flexibility enables users to adapt `Purify` to diverse data contexts and hypothesis-testing requirements. See also `cross_validation()` and `confidence_intervals()`.

Supporting functions offer additional insight into data. Functions such as `summarize_resample()` provide information to the user on resampling, and other functions highlight the probability of the observed data. The functions `plot_strata_bar()` and `plot_strat_box()` visualize group sizes in stratified samples.

Example

`Purify` provides in-depth articles on the package website. For example,

- The *purify* article describes the core features of 'purify' and includes simulations and real data examples to demonstrate the functions.
- The *cats* article details investigation on real data.

We consider a subset of the cats (*subcats*) data set below. We use sex and body weight to estimate heart weight; see Figure 1. Similar to many real-world examples, the data is highly imbalanced. Nonetheless, sex and body weight are both useful in understanding heart weight. In particular, female cats have lower body weights and have a lower heart weight even for the same body weight when compared to male cats.

Table 1: Subcats models. Models of cats using body weight and sex to predict heart weight.

	Intercept 95% Conf Int	Sex (M) 95% Conf Int	Body weight 95% Conf Int	MSE
Single model	-1.486 (-3.236, 0.264)	0.617 (-0.139, 1.372)	4.208 (3.573, 4.843)	2.258
Resampled model	-1.427 (-3.603, 0.610)	0.620 (0.051, 1.167)	4.186 (3.425, 5.005)	2.192

Let the linear model be defined as heart weight predicted by an intercept, body weight, and sex. Estimates for the coefficients and the confidence intervals for each parameter are given in Table 1. When applying the linear model directly on the data, only body weights appear to significantly impact heart weight. For resampled data, where samples are taken to create more evenly sized groups based on sex, both sex and body weight are determined to be significant. The cost for this simple example is that the confidence interval on body weight is larger. While additional simulations or modifying the resampling scheme may mitigate such losses, it is important to consider such effects. Often the prediction errors, e.g. mean square error (MSE), is more important and in this case, the resampled model also performs better. See articles for information on other functions and additional analysis on this and other data.

Implementation

`Purify` is implemented in R, following standard stylization and using vectorized operations for efficient computation. The package's modular design and clear documentation make it easy to adapt to various research needs, allowing users to integrate their own statistical functions or modify resampling parameters to meet specific analytic requirements. `Purify` has been used in Tetui et al. (2022), Bui et al. (2024), and Alexander et al. (2024).

Acknowledgements

Development of the `Purify` package was inspired by foundational methods in statistical resampling and permutation testing along with the rich literature on stratified data. Special thanks to the open-source R community for support and resources.

Contributions to `Purify` are welcome and notable recognition is given to all who raise awareness of deficiencies in the package via the GitHub repository.

References

Alexander, K. L., Hall, K., & Chen, Y. M. (2024). Librarian involvement on knowledge synthesis articles and its relationship to article citation count and journal impact factor. *The Journal of the Canadian Health Libraries Association*, *45*(3), 137–146.

Bui, T., Chen, M., Hagar, L., Ramsay, K., Shi, Y., Tompkins, G., VanderDoes, J., & Zhu, F. (2024). *Topics in statistical consulting*. https://scsru.github.io/Modules/

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press. ISBN: 9780521573917

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman; Hall. ISBN: 0412042312

Kuhn, & Max. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Tetui, M., Grindrod, K., Waite, N., VanderDoes, J., & Taddio, A. (2022). Integrating the CARD (comfort ask relax distract) system in a mass vaccination clinic to improve the experience of individuals during COVID-19 vaccination: A pre-post implementation study. *Human Vaccines & Immunotherapeutics*, *18*, 2089500–2089500. http://dx.doi.org/10.1080/21645515.2022.2089500

Thompson, S. K. (2012). *Sampling* (3rd ed.). Wiley. ISBN: 9780470402313