stats

MDPI

Communication

eduSTAT—Automated Workflows for the Analysis of Small- to Medium-Sized Datasets

Rudolf Golubich

Independent Researcher, Satzgasse 16, 7100 Neusiedl am See, Austria; rudolf.golubich@statistikmacher.de

Abstract

This communication provides a citable methodological reference for eduSTAT (v1), an automated, rule-based workflow for the statistical analysis of small- to medium-sized datasets ($N \approx 30$–3000). The web application is initially available in German and will be offered in English once it is established in German-speaking regions. It is developed with the aim of supporting early training in the scientific method and reducing the risk of spurious or inappropriate statistical analyses. The paper establishes the foundation for subsequent meta-analyses based on citation tracking of studies that apply eduSTAT, enabling iterative, data-driven improvement of the software.

Keywords: educational statistics; automated analysis; statistics software; robustness; confounding; power analysis; reproducibility

1. Introduction

An inadequate or incorrect introduction of students to the scientific method, particularly statistical analysis, contributes to systematic analytical errors, including misuse of parametric tests, unrecognized confounding, p-hacking [1] and publication bias [2]. These issues are widely regarded as contributors to the interdisciplinary replication crisis [3,4].

Modern statistical software, in combination with AI, lowers technical barriers but often obscures methodological assumptions, increasing the risk of inappropriate application, particularly by early-stage researchers [5].

Addressing this gap, eduSTAT offers a rule-based system that automates the selection and evaluation of statistical tests. Results are generated in a Word template suitable for scientific reporting. Reports include citations for relevant methods and reference this paper.

This communication serves as the primary methodological reference for eduSTAT, supporting pilot deployment and enabling subsequent meta-analyses to assess the software's impact through citation tracking.

2. Methodological Framework

eduSTAT ingests tabular data from common formats (e.g., CSV, XLSX), extracts features, and classifies them as nominal, ordinal, or cardinal via heuristic inference based on data type and value structure. Because automated scale inference is inherently error-prone, users can manually override all inferred properties. Feature construction is supported through aggregation into composed features or binary recoding into nominal groups.

Statistical analysis is initiated by selecting two features. Test selection is inferred from their measurement scales, pairing structure, and normality diagnostics using the decision tree given in Figure 1.

Check for updates

Academic Editor: Wei Zhu

Received: 31 December 2025
Revised: 29 January 2026
Accepted: 3 February 2026
Published: 4 February 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

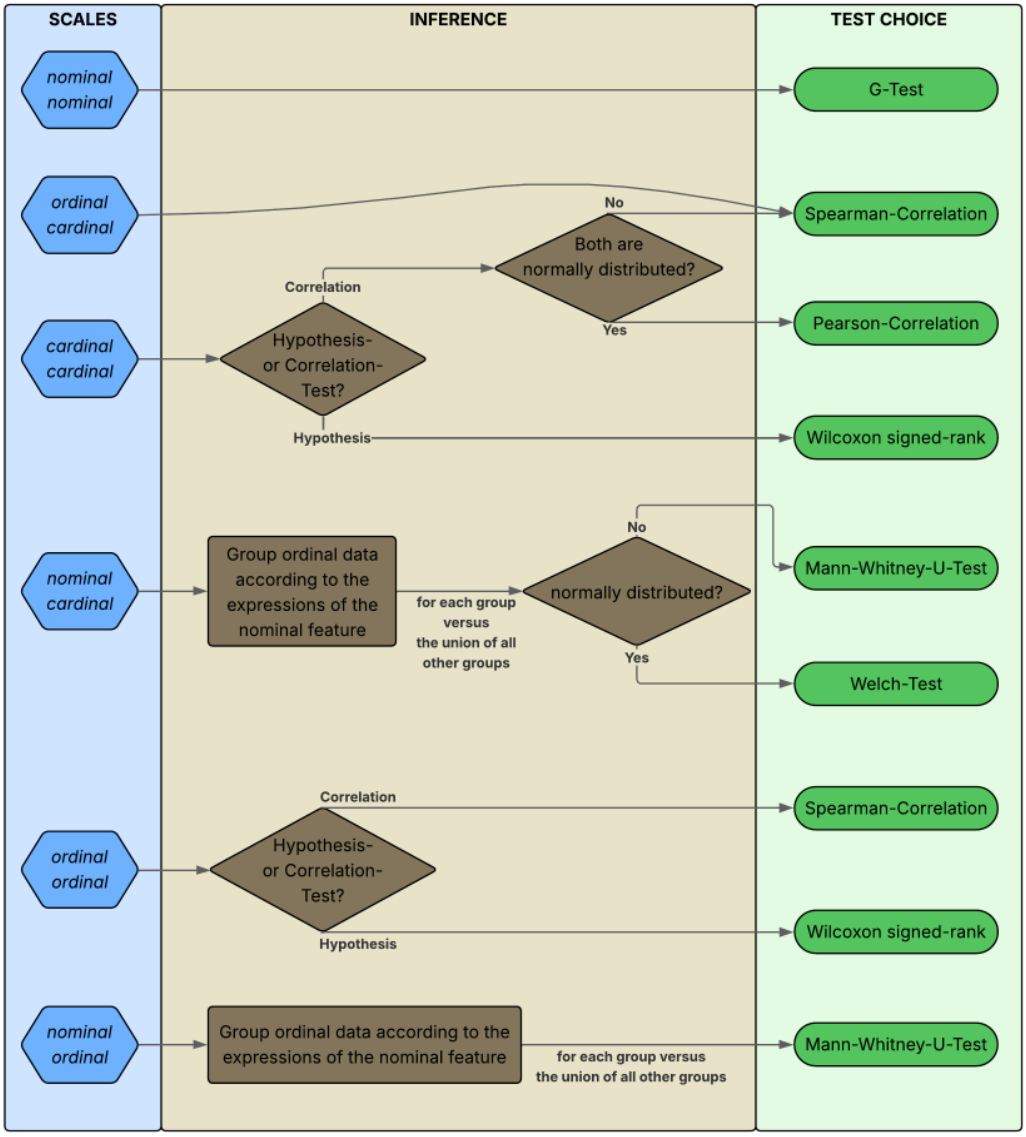

Figure 1. A decision tree based on the scales of two features and normality tests is used to infer the respective hypothesis test or correlation test to be evaluated.

Normality is assessed with the Anderson–Darling test, which is sensitive to tail deviations and outliers to favor robust, non-parametric methods. Parametric tests are used if normality holds; otherwise, rank-based alternatives are applied. Reliance on a single normality test may lead to conservative decisions, particularly for large samples or mild deviations from normality. If two selected features share the same measurement scale and a common term in their names, a hypothesis test is applied; otherwise, a correlation analysis is used. This heuristic is limited by feature-name semantics and may not capture all substantively meaningful pairings.

For each test, eduSTAT provides numerical results (e.g., *p*-value, test statistics and effect size) and a natural-language interpretation together with visualizations. It reports estimates of the sample size required for a test power of 0.8 or the detection threshold, based on the following methodologies:

- Anderson–Darling: sample size estimate for the Anderson–Darling test [6] is based on empirically determined power curves from [7]. The minimum sample size required to achieve a target power of 0.8 is reported.

- Mann–Whitney U Test: for the Mann–Whitney U test [8] the required sample size for a power of 0.8 is estimated using the rank-based effect size, accounting for group variances and relative group sizes based on the methods of Ref. [9].
- Wilcoxon Signed-Rank Test: for the Wilcoxon signed-rank test [10,11], the required sample size for a power of 0.8 is estimated by adapting the standard t-test formula [12], replacing the parametric mean with the Hodges–Lehmann estimator [11] and the standard deviation with a robust measure of dispersion (MAD): $\left\lceil \left(\frac{(1.96+0.84)\cdot 1.4826\cdot \mathrm{MAD}}{|\mathrm{HL}|} \right)^2 \right\rceil$, with MAD being median absolute deviation and HL the Hodges–Lehmann estimator.
- Welch t-Test: the sample size requirement for the Welch's t-test [13] is evaluated using standard deviation and mean, explicitly accounting for unequal variances and sample sizes based on the formula for the t-test [12]: $\left(1+\frac{1}{r}\right)\frac{\left(s_\ell^2+r\,s_r^2\right)(1.96+0.84)^2}{(\langle x_l\rangle-\langle x_r\rangle)^2}$ with x_L and x_r referring to the data of the respective groups and r to the ratio of the group sizes. The detection threshold is determined by scaling the variances for significance.
- G-Test: for the G-test [14], the sample size required for a power of 0.8 is estimated via the Kullback–Leibler divergence as $\left\lceil \frac{\chi^2_{1-\alpha,\,\mathrm{df}}+\chi^2_{0.8,\,\mathrm{df}}}{2\sum_k p_k(\log p_k-\log q_k)} \right\rceil$ with p_k referring to the observed distribution, q_k referring to the expected distribution based on the data synthesized for the power analysis and df referring to the degrees of freedom. By scaling observed contingency counts and iteratively testing hypothetical sample sizes, the likelihood ratio is recalculated as the sample size increases until the critical threshold for significance is reached, also providing a dataset-specific estimate of the detection threshold.
- Pearson and Spearman Correlation: the detection threshold for the Pearson correlation [15] and Spearman correlation [16] are determined by transforming the correlation coefficient into Fisher-z values [17,18], enabling a normal approximation of the test statistic. The detection threshold is determined numerically as the smallest sample size for which a confidence interval based on Fisher's transformation excludes zero.

The implementation of the statistical tests and associated algorithms relies on *math.js* (*math.js* https://mathjs.org/ (accessed on 25 January 2026) under Apache-License 2.0) to improve the numerical precision of floating-point operations and to reduce rounding errors in intermediate computations. At the same time, the implementation is constrained by the JavaScript execution environment, which is not optimized for numerical algorithms of high computational complexity. Consequently, eduSTAT adopts a deliberate trade-off between numerical precision and computational efficiency: it prioritizes stable and sufficiently accurate results for interactive, browser-based analysis over maximally optimized high-performance computation.

Validation results for the sample size estimates, p-values, test-statistics and power analysis of the test implementations are publicly available online at https://edustat.at/Validierungen/?ref=mdpi (accessed on 25 January 2026).

Multiple comparison corrections are presently not implemented in eduSTAT automatisms, as the necessity for such adjustments cannot be definitively inferred from the dataset's structure without additional contextual information.

A confounder analysis is performed to assess the potential impact of additional variables on statistical results based on a simplified ANCOVA to reduce requirements and simplify interpretation:

When comparing groups in a hypothesis test, the homogeneity of the groups with respect to each confounding variable is examined. For cardinal or ordinal confounders, a linear regression on the outcome variable is used to model how differences in group composition translate into differences in the outcome. The system implements a heuristic for selecting between OLS, TLS, and robust linear regression depending on data characteristics.

For nominal confounders, the mean outcome is calculated for each level of the confounding variable to capture compositional effects. Based on a correlation analysis between the confounder and the outcome variable, the effect of the confounder on the target measure is then estimated. This effect is subtracted from the original result, and the confidence interval is re-evaluated. If the adjusted confidence interval crosses zero, the result is considered potentially biased.

For correlation analyses, partial correlation controlling for the confounding variable is applied. If the resulting value of the partial correlation falls outside the confidence interval of the original correlation, the result is considered potentially biased. By distinguishing whether the partial correlation lies above or below the original confidence interval, eduSTAT can provide guidance for the construction of structural equation models.

This approach to confounder analysis is susceptible to biases resulting from multiple comparisons. Since an automated inference of multiple testing cannot be performed without ambiguity in the absence of additional context, no multiple comparison corrections are currently implemented. Consequently, the results of the confounder analysis should be interpreted with caution and primarily serve as an indicator. Furthermore, mutual correlations or covariances between potential confounders are not taken into account. The current implementation in eduSTAT does not aim to provide a quantitative adjustment for confounding, but rather to establish a qualitative basis for potential optimizations of the study design. It is intended to serve as a diagnostic indicator for the robustness of the primary study results.

All results are reported in a structured format that can be exported as a Microsoft Word document, encompassing methods, results, conclusions, and references, systematically linking statistical analyses, visualizations, and citations. The Word export is implemented using *docx.js* (docx.js under the MIT license: https://docx.js.org/ accessed on 25 January 2026) and *FileSaver.js* (FileSaver.js under the MIT license: https://github.com/eligrey/FileSaver.js accessed on 25 January 2026), while visualization generation is based on *plotly.js* (plotly.js under the MIT license: https://plotly.com/javascript/ accessed on 25 January 2026).

The report structure is intended to be complemented by standardized template texts to support transparent reporting. These templates will be introduced in future development stages and iteratively refined based on insights gained from subsequent meta-analyses derived from citation tracking of studies applying eduSTAT.

Scope and Methodological Streamlining

This communication establishes a citable methodological documentation of eduSTAT, providing the transparency required for its deployment in pilot studies and subsequent empirical evaluation. Future work will focus on domain-specific validation of the automated decision logic and confounder diagnostics, iterative refinement of reporting templates guided by citation-based meta-analyses, and methodological extensions such as user-driven test selection, additional variance and normality diagnostics, and Monte Carlo-based power simulations. Together, these steps aim to strengthen both the scientific robustness and pedagogical utility of eduSTAT in applied research and teaching contexts. Once eduSTAT is established in the German-speaking region, the web application will also be offered in English.

Funding: This research and the development of eduSTAT where partly funded by the Austrian Research Promotion Agency (FFG).

Data Availability Statement: All data and results produced during validation of the software components are available at https://edustat.at/Validierungen/?ref=mdpi (accessed on 25 January 2026).

Acknowledgments: The development of eduSTAT was supported by funding from the Austrian Research Promotion Agency (FFG). Their contribution is gratefully acknowledged. I would like to thank Oscar Seiler and Mike Andreas Bähr for supporting the development of eduSTAT through alpha testing and Mike Andreas Bähr also reviewed an preliminary version of this publication for clarity.

Conflicts of Interest: The author received funding from the Austrian Research Promotion Agency (FFG) for the development of the software eduSTAT and is also the primary developer and prospective distributor of this software product. These roles are disclosed in the interest of transparency.

References

1. Head, M.L.; Holman, L.; Lanfear, R.; Kahn, A.T.; Jennions, M.D. The Extent and Consequences of P-Hacking in Science. *PLoS Biol.* **2015**, *13*, e1002106. [CrossRef] [PubMed]
2. Sterling, T.D. Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa. *J. Am. Stat. Assoc.* **1959**, *54*, 30–34. [CrossRef]
3. Pashler, H.; Wagenmakers, E.J. Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspect. Psychol. Sci.* **2012**, *7*, 528–530. [CrossRef]
4. Fidler, F.; Wilcox, J. Reproducibility of Scientific Results. In *The Stanford Encyclopedia of Philosophy*, Winter 2018 ed.; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2018.
5. Khlaif, Z.N.; Mousa, A.; Hattab, M.K.; Itmazi, J.; Hassan, A.A.; Sanmugam, M.; Ayyoub, A. The Potential and Concerns of Using AI in Scientific Research: ChatGPT Performance Evaluation. *JMIR Med. Educ.* **2023**, *9*, e47049. [CrossRef] [PubMed]
6. Anderson, T.W.; Darling, D.A. Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *Ann. Math. Stat.* **1952**, *23*, 193–212. [CrossRef]
7. Mohd Razali, N.; Yap, B. Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. *J. Stat. Model. Anal.* **2011**, *2*, 21–33.
8. Mann, H.B.; Whitney, D.R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.* **1947**, *18*, 50–60. [CrossRef]
9. Happ, M.; Bathke, A.C.; Brunner, E. Optimal sample size planning for the Wilcoxon-Mann–Whitney test. *Stat. Med.* **2019**, *38*, 363–375. [CrossRef] [PubMed]
10. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]
11. Hodges, J.L., Jr.; Lehmann, E.L. Estimates of Location Based on Rank Tests. *Ann. Math. Stat.* **1963**, *34*, 598–611. [CrossRef]
12. Student. The Probable Error of a Mean. *Biometrika* **1908**, *6*, 1–25. [CrossRef]
13. Welch, B.L. The generalization of 'Student's' problem when several different population varlances are involved. *Biometrika* **1947**, *34*, 28–35. [CrossRef]
14. Williams, D.A. Improved likelihood ratio tests for complete contingency tables. *Biometrika* **1976**, *63*, 33–37. [CrossRef]
15. Pearson, K. VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia. *Philos. Trans. R. Soc. Lond. Ser. A Contain. Pap. A Math. Phys. Character* **1896**, *187*, 253–318. [CrossRef]
16. Spearman, C. The Proof and Measurement of Association Between Two Things. *Am. J. Psychol.* **1904**, *15*, 88–103. [CrossRef]
17. Fisher, R.A. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika* **1915**, *10*, 507–521. [CrossRef]
18. Fisher, R.A. On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. *Metron* **1921**, *1*, 3–32.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.