Robust tests should be the default, not the backup

Michael Höfler[1]

[1] Clinical Psychology and Behavioural Neuroscience, Institute of Clinical Psychology and Psychotherapy, Technische Universität Dresden, Dresden, Germany. Chemnitzer Straße 46, 01187 Dresden, Germany. Mail: michael.hoefler@tu-dresden.de, Phone: +49 351 46936921, ORCID 0000-0001-7646-8265

Abstract

The assumptions of standard tests such as the *t*-test, ANOVA and ordinary least squares regression are frequently violated. This can impact the desired error rates in statistical hypothesis testing. Robust tests like the Mann–Whitney U test and robust linear regression do not rely on assumptions such as normality and equal variances. Using them can counteract non-replicated findings that are just due to data anomalies, such as extreme values and outliers, which occur differently across studies. Employing them from the outset bypasses the pitfalls of deciding on the usability of a standard test with data. In this opinion piece, I summarize the epistemic benefits of robust alternatives. Restricting to a robust test instead of conducting it in addition to the standard test avoids generating multiple results, thus counteracting fishing for the desired, which can occur subtly. From a practical standpoint, running a single test simplifies analysis, and many robust methods are readily available in R. However, it is important to understand what a robust method does and what it is actually robust against. I also address common defenses of standard tests, discuss why they remain widespread, and suggest how these arguments should be countered.

Introduction

Standard methods such as *t*-test, ANOVA and ordinary least squares regression assume normal distribution, equal variances and the absence of extreme values and outliers. However, deviations from these assumptions have been found to be the rule rather than the exception in psychological science (Micceri, 1989; Wilcox, 2017). They may have consequences on the decisions of statistical hypothesis testing through inflated error rates—increase the risk of false positives (rejecting a true null hypothesis, Type I (α) error) and false negatives (retaining a false null hypothesis, Type II (β) error) respectively. This especially applies to instances—common in real data—where several issues occur concurrently; for instance, generally skewed distributions combined with additional extreme values (Cressie & Whitford, 1986; Field & Wilcox, 2018; Micceri, 1989; Wilcox et al., 2013; Wilcox, 2017; Tukey, 1960). The literature on *when exactly* standard tests are robust against violated assumptions is vast and sometimes contradictory (Wilcox, 1998; Avella-Medina & Ronchetti, 2015). Therefore, it is appealing to circumvent these issues by using robust methods like the Mann-Whitney U test and robust linear regression that rely much less on these assumptions.

This opinion piece concerns confirmatory research, for which adherence to scientific rigor and the desired error rates is imperative. I review the drawbacks of using data to assess a test's usability and the employment of a robust test only as a backup to the standard test in reaction to anomalies in data. I summarize the epistemic benefits of using robust tests in general, and from the outset. The paper concludes with some practical considerations and a discussion of common defenses of standard tests.

Main part

Deciding on the applicability of a standard test with the given data

In research practice, statistical tests are commonly used to determine if a standard test can be used to examine a substantive hypothesis. If such a test, for example, the Shapiro-Wilk test for normality, finds that deviations are statistically significant (commonly $p < .05$), a more robust test is usually employed to test the hypothesis. However, statistical tests for model assumptions provide a poor decision rule. In small samples, the statistical power to detect deviations is low. This often leads to the standard test being chosen, despite the departures in the sample having a consequential extent (Field & Wilcox, 2018). In large samples, these tests can be overly sensitive, flagging negligible departures as important (Lumley et al., 2002). Yet conceptually, statistical tests are disputable here because they infer to populations, whereas the usability of a test actually depends on the distribution in the sample at hand (Altman, 1991; Lix et al., 1996). Instead of relying on statistical tests, one may apply *graphical methods* to decide upon the usability of a standard test. With visual data inspection such as density or residual plots, it is less likely that the sample size will lead to a different assessment of the accuracy of the model assumptions. However, visual methods require much experience and are highly subjective in their application (Razali & Wah, 2011), which opens the door for fishing experiments. Robust alternatives bypass the need for any data-based decisions to the extent that they per se consider the assumptions that otherwise need to be checked in data.

Robust tests as back-up

If the data suggest violated assumptions, it is common to also run a robust test. Like others (Erceg-Hurn & Mirosevich, 2008; Wilcox, 2017), I argue for carrying out robust tests from the start, instead of as a backup option. Generating multiple results creates the danger of p-hacking (commonly fishing for $p < \alpha$, e.g., when testing for the existence of an effect). Preregistration can and must specify how conflicting results are handled—commonly, by preferring the robust alternative over the standard test when one yields $p < \alpha$ and the other $p \geq \alpha$ (Wagenmakers et al., 2012). In these cases, though, running only the robust test from the outset would have led to the same conclusion. The standard test is redundant. Another prevalent practice may yet subtly undermine scientific rigor: Researchers openly report both results and take them together as 'unclear' or 'partial evidence.' However, because they still need to make a binary decision (e.g., pursue further research or try an intervention assuming the tested effect exists), they may actually behave as if there were evidence. In this case, the nominal α has been subtly exceeded (Gelman & Loken, 2014).

Robust testing from the outset is epistemically well-founded

Fundamentally, empirical science should subject hypotheses to *risky testing*—so that if a hypothesis were false, the test would likely produce contradictory evidence. That is, tests should be *severe*, with large and adhered to falsification rates of $1 - \alpha$ (false positives) and $1 - \beta$ (false negatives) (Mayo, 2018). However, anomalies in data—such as extreme values or outliers that disproportionately influence results—can compromise the intended error rates (in addition to other model–reality mismatch; Gigerenzer, 2004), making a false hypothesis appear corroborated or a true hypothesis appear uncorroborated (Wilcox, 2017). (Note that outliers, by definition, stem from a different population and should therefore conceptually be omitted from the analysis, while extreme values should remain in, but not dominate the results.)

Empirical testing has to be robust across random perturbations to increase reliability and protect inference from flaws in the analytical model. A test should not be passed (or unpassed) just because of faulty assumptions embedded within it (Popper, 1959). Moreover, anomalies such as extreme values and outliers are likely to occur inconsistently across studies. When analyzed, for instance, with the two independent samples *t*-test, one study might corroborate an effect while another might not. As Popper (1959, p. 66) stated, 'non-reproducible single occurrences are of no significance to science.' Such lack of replication distorts scientific communication and leads to unnecessary and misleading debates about substantive reasons for differing results, where the variation is just due to unmet assumptions in the statistical method. Robust statistical tests can mitigate these problems by reducing the influence of outliers, non-normality, and heteroscedasticity (Erceg-Hurn & Mirosevich, 2008; Field & Wilcox, 2017; Rousselet & Wilcox, 2020; Wilcox, 2017).

Scientific communication must be clear about the scope of the hypothesis being tested. *T*-test, ANOVA, and ordinary least squares regression are routinely used to test *population-average effects* (though this is rarely made explicit), via mean group differences or, in the regression context, the average outcome change per unit of a predictor. Estimates of these effects carry broader interpretability only when they closely reflect the true effect in *many individuals*. If anomalies such as extreme values or outliers dominate them, they are only applicable to a few individuals. This results in a tacit and unjustified narrowing of the inference scope (Altman & Krzywinski, 2016; Huber & Ronchetti, 2009). For example, robust linear regression, as an alternative to ordinary least squares (OLS) regression, exactly compensates for this. After applying an outlier criterion to identify and omit individuals—who then have no impact on the results—it weights the remaining individuals so that each contributes approximately the equal amount to the parameter estimates, in the same way as in ordinary least squares regression under its assumptions (normally distributed residuals with equal variance; Huber & Ronchetti, 2009; Wilcox, 2017).

The final argument is an epistemic advantage of *not reacting* to unexpected data features. In the Popperian tradition, the substantive hypothesis should make a *prediction*, for example about an average effect, which may turn out to be right or wrong (Popper, 1959; Mayo, 2018). Together with a decision rule (e.g. the one-tailed p-value in the chosen statistical test must be smaller than α), it then *predetermines* which observations support the hypothesis and which do not. This requires fully specifying an analytical model, so that once data are collected, the test yields either $p < \alpha$ or $p \geq \alpha$ (Lakens & DeBruine, 2021). When $p < \alpha$ as predicted, the test retains evidential value, simply because the prediction succeeded—even if the analytical model is imperfect and can be improved post-hoc (Box, 1976; Huber & Ronchetti, 2009; Uygun Tunç et al., 2023).

Which alternative test is appropriate?

Although unlike other papers (Kim & Lee, 2023; Mair & Wilcox 2020; Wilcox, 2017, and Wilcox & Rousselet, 2018) this article is not a review of robust alternatives, some basic guidance can be given in condensed form. The Mann-Whitney U test as an alternative to the two independent samples t-test has been argued to be largely robust against non-normality, unequal variances, extreme values and outliers (outliers are still included in the analysis, they count as the largest values) (Zimmerman, 1994). However, it does not examine the same type of hypothesis as the standard test. Whereas the *t*-test is based on the difference in the mean between two populations, the U-test compares rank sums between groups. Although this often makes no practical difference, exceptions have been found—for example when distributions differ in spread or shape but have equal means, the U-test may signal a difference although the means are equal in the population (Fay & Proschan, 2010; Bürkner et al., 2017).

Unlike the U-test, the *exact t*-test also compares means. Its robustness comes from computing p-values via all possible data permutations rather than relying on distributional assumptions (Winkler et al., 2014). However, because extreme values and outliers reappear in many permutations, the exact *t*-test is only partially robust to them. Full robustness is achieved by *trimmed* and *Winsorized versions* of the *t*-test, which down-weight or remove extreme values and outliers according to pre-defined criteria and are implemented in the R package WRS2 (Mair & Wilcox, 2020). Importantly, one can do conceptually the same with robust linear regression. Since any linear regression model can test mean differences via dummy-coding of the factor (Rohrer & Arel-Bundock, 2025), robust linear regression constitutes a fully robust substitute for the *t*-test and ANOVA. The method handles extreme values and outliers also by down-weighting or excluding observations, and is implemented in the R package *robustbase* (Maechler et al., 2021). The *robustlmm* package extends robust linear regression to more complex multilevel data, enabling the fitting of robust mixed-effects models (Koller, 2016).

Why standard test are standard—and why the associated arguments have shortcomings

The widespread practice of sticking to standard tests and employing robust methods, at best, as a backup is mainly defended by appeals to tradition and common usage. I address three specific arguments.

1. Measurements are expected to produce normally distributed data.

This assumption is frequently invoked to justify the use of classical statistical methods, even though normally distributed data are rare (Wilcox & Field, 2018; Micceri, 1989; Blanca et al., 2013). At least approximate normality should arise, so the common argument goes, when measurements result from the additive effects of many small, independent influences, as in Gauss's original derivation. However, this justification has been criticized as superficial when the data-generating process is poorly understood (Erceg-Hurn & Mirosevich, 2008). Sometimes, it is clearly implausible—for instance, in clinical psychology where the distributions of key constructs (e.g., symptoms of mental disorders) naturally exhibit heavy tails and skewness (Micceri, 1989). Nevertheless, the normality assumption often leads researchers to interpret extreme values as legitimate instances of a normal distribution, rather than safeguarding statistical inference against them. When normality is at least in doubt, it is clearly preferable to prioritize robust methods that protect against data anomalies.

2. Standard statistical tests possess optimality properties under the assumptions they make

For example, the two-sample *t*-test is the most powerful test for detecting differences in means between independent groups when its assumptions hold. Yet, the U-test requires only about 5% more participants to achieve comparable power under these same conditions (Blair & Higgins, 1980). Similar efficiency applies to robust linear regression compared to ordinary least squares regression (Wilcox, 2017). Thus, the error from unwarrantedly choosing these alternatives seems small.

3. Standard methods are generally robust against their assumptions.

This argument has been criticized for relying on outdated simulations that consider only isolated assumption violations—such as non-normality or unequal variances—rather than the complex combinations of violations found in real data, where the claimed robustness of standard methods often breaks down (Field & Wilcox, 2017; Wilcox, 2017). Traditional textbook presentations rely on simplified narratives: the assumptions of standard methods are 'usually met,' their violations have minimal impact, the *t*-test is valid whenever both groups have sample sizes of at least 30 (Lumley et al., 2002; Boneau, 1960). These simplifications reduce the cognitive burden of statistical analysis and discourage deeper engagement with the consequences of assumption derivations. Most researchers, especially senior researchers, have been trained under these conventions, and questioning them may call into doubt the reliability of earlier analyses and long-standing disciplinary practices. However, pointing out the prospects of using a robust test from the outset may be helpful: The effort to test assumptions in data, compute multiple tests, and integrate results can be avoided. If scientific reputation shifts from publication count to replication success (Nosek et al., 2022), robust tests are preferable.

Discussion

Time has come to overcome standard statistical tests whenever their assumptions are unlikely to be fulfilled. This would reduce unnecessary variation in results both within and between studies. *Within studies*, the flexibility in dealing with the variation must be rigorously disclosed to ensure rigorous testing, which can otherwise be subtly undermined. *Between studies*, the common usage of robust methods shall reduce the number of non-replicated findings, facilitating scientific communication in a field that is already burdened by otherwise occurring variance between study results (Nosek et al., 2022). To make robust testing more commonplace, teaching should connect it to the replication crisis and appeal to the advantages of more reliable and sustainable scientific results. At the same time, the toolbox of robust alternatives should be extended. Ideally, until it covers the vast range of statistical methods used in psychological science. The robustness of each and every method needs to be clearly understood. Ultimately, the methods should be implemented, summarized and explained in a way that serves the goals of easy access and use.

Acknowledgements

None

Funding

No funding was received to assist with the preparation of this manuscript.

Conflict of interest disclosure

The author declares that he complies with the PCI rule of having no financial conflicts of interest in relation to the content of the article, and has no non-financial conflicts of interest.

References

Albers, C. J., Boon, P., & Kallenberg, W. C. M. (2000). Why tests of normality are inappropriate prior to testing hypotheses. *Psychometrika, 65*(4), 401–415.

Altman, D. G. (1991). *Practical statistics for medical research*. Chapman & Hall.

Altman, N., & Krzywinski, M. (2016). Analyzing outliers: Influential or nuisance? *Nature Methods*, 13(4), 281–282. https://doi.org/10.1038/nmeth.3812

Avella-Medina, M., & Ronchetti, E. (2015). Robust statistics: a selective overview and new directions. *WIREs Computational Statistics, 7*(6), 372–393. https://doi.org/10.1002/wics.1363

Blair, R. C., & Higgins, J. J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's *t* statistic under various nonnormal distributions. *Journal of Educational Statistics*, 5(4), 309–335. https://doi.org/10.3102/10769986005004309

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and Kurtosis in Real Data Samples. *Methodology*, *9*(2), 78–84. https://doi.org/10.1027/1614-2241/a000057

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin, 57*(1), 49–64. https://doi.org/10.1037/h0041412

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association, 71*(356), 791–799.

Bürkner, P. C., Doebler, P., & Holling, H. (2017). Optimal design of the Wilcoxon-Mann-Whitney-test. *Biometrical Journal. Biometrische Zeitschrift*, *59*(1), 25–40. https://doi.org/10.1002/bimj.201600022

Cressie, N. A., & Whitford, H. J. (1986). How to use the two-sample *t*-test. *Biometrical Journal, 28*(2), 131–148. https://doi.org/10.1002/bimj.4710280202

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *The American psychologist*, *63*(7), 591–601. https://doi.org/10.1037/0003-066X.63.7.591

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, *4*, 1–39. https://doi.org/10.1214/09-SS051

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour Research and Therapy*, 98, 19–38. https://doi.org/10.1016/j.brat.2017.05.013

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460–465. https://doi.org/10.1511/2014.111.460

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*, 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42(3), 237–288. https://doi.org/10.3102/00346543042003237

Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics* (2nd ed.). Wiley.

Kim, J., & Li, J. C.-H. (2023). Which robust regression technique is appropriate under violated assumptions? A simulation study. Methodology, 19(4), 323–347. https://doi.org/10.5964/meth.8285

Koller, M. (2016). *robustlmm: An R package for robust estimation of linear mixed-effects models*. *Journal of Statistical Software, 75*(6), 1–24. https://doi.org/10.18637/jss.v075.i06

Lakens, D., & DeBruine, L. M. (2021). Improving transparency, falsifiability, and rigor by making hypothesis tests machine-readable. *Advances in Methods and Practices in Psychological Science, 4*(2). https://doi.org/10.1177/2515245920970949

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66*(4), 579–619. https://doi.org/10.2307/1170654

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual review of public health*, *23*, 151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L. T., & Anna di Palma, M. (2021). *robustbase: Basic Robust Statistics*. https://cran.r-project.org/web/packages/robustbase/index.html

Mair, P., & Wilcox, R. R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods, 52*(2), 464–488. https://doi.org/10.3758/s13428-019-01246-w

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press. https://doi.org/10.1017/9781107286184

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156–166. https://doi.org/10.1037/0033-2909.105.1.156

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology, 73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. *Journal of Statistical Modeling and Analytics, 2*(1), 21–33.

Rohrer, J. M., & Arel-Bundock, V. (2025, August 25). Models as Prediction Machines: How to Convert Confusing Coefficients into Clear Quantities. https://doi.org/10.31234/osf.io/g4s2a_v1

Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing* (4th ed.). Academic Press.

Wilcox, R., Carlson, M., Azen, S., & Clark, F. (2013). Avoid lost discoveries, because of violations of standard assumptions, by using modern robust statistical methods. *Journal of Clinical Epidemiology, 66*(3), 319–329. https://doi.org/10.1016/j.jclinepi.2012.09.003

Wilcox, R. R., & Rousselet, G. A. (2018). A guide to robust statistical methods in neuroscience. *Current Protocols in Neuroscience, 82*(1), 8.42.1–8.42.30. https://doi.org/10.1002/cpns.41

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. NeuroImage, 92, 381–397. https://doi.org/10.1016/j.neuroimage.2014.01.060

Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin, S. Ghurye, W. Hoeffding, W. Madow, & H. Mann (Eds.), *Contributions to probability and statistics* (pp. 448–485). Stanford University Press.

Uygun Tunç, D., Tunç, M. N., & Lakens, D. (2023). The epistemic and pragmatic function of dichotomous claims based on statistical hypothesis tests. *Theory & Psychology*, *33*(3), 403-423. https://doi.org/10.1177/09593543231160112 (Original work published 2023)

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Zimmerman, D. W. (1994). A note on the influence of outliers on parametric and nonparametric tests. *The Journal of General Psychology*, 121(4), 391–401. https://doi.org/10.1080/00221309.1994.9921214

Appendix

First author name (last/family, first/given):

Höfler, Michael

Preprint DOI or URL:

https://osf.io/preprints/psyarxiv/6v3cz_v1

Section 1: Data

Does your manuscript contain reports of any data?

☐ Yes (continue with next question)

x☐ No (skip to Section 2):

Are appropriately anonymised raw data available within a trusted digital repository?

☐ Yes, available at this link:

☐ No, justification:

Are third-party data cited in the manuscript, with a DOI? (e.g., for preexisting data, data deposited in a repository; see Data citation – A guide to best practice)

☐ Yes, the DOI is as follows:

☐ No, justification:

Is there a data dictionary and/or readme file included with the data to make it interpretable?

☐ Yes, available at this link:

☐ No, justification:

Do you indicate in the manuscript how the sample size was determined?

☐ Yes.

☐ No, justification:

Do you report all data exclusions (e.g., outliers, careless responders)?

☐ Yes.

☐ No, justification:

Do you report all inclusion/exclusion criteria and when they were established?

☐ Yes.

☐ No, justification:

Are all measures, questions, and/or conditions used in the study described in the manuscript or available in the supplemental material?

☐ Yes.

☐ No, justification:

Section 2: Analysis Scripts/Code/Codebooks

Does your manuscript contain any analysis of quantitative or qualitative data?

☐ Yes (continue with next question)

x☐ No (skip to Section 3):

Are third-party analysis scripts/code (e.g., R, Stata), codebooks, or other relevant documentation available within a trusted digital repository?

☐ Yes, available at this link:

☐ No, justification:

Are the analysis scripts/code (e.g., R, Stata), codebooks, or other relevant documentation cited in the manuscript, with a DOI?

☐ Yes, the DOI is as follows:

☐ No, justification:

Section 3: Study Materials

Does your manuscript contain any research materials (e.g., stimuli, programming code, questionnaires, interview protocols)?

☐ Yes (continue with next question)

x☐ No (skip to Section 4):

Are all study materials and descriptions of study procedures available within a trusted digital repository?

☐ Yes, available at this link:

☐ No, justification:

Are all third-party study materials, descriptions of study procedures, or other relevant documents cited in the manuscript, with a DOI?

☐ Yes, the DOI is as follows:

☐ No, justification:

Section 4: Preregistration

Were any aspects of your manuscripts preregistered?

☐ Yes (continue with next question)

x☐ No (do not complete the rest of the form):

Does the manuscript contain an accessible link to the preregistration?

☐ Yes, available at this link:

☐ No, justification:

Do you clearly indicate in the manuscript which parts were preregistered and which parts were not?

☐ Yes.

☐ No, justification:

Are all preregistered analyses reported in the text or linked in the supplemental material?

☐ Yes.

☐ No, justification:

Are all deviations from the preregistration plan clearly disclosed in the manuscript (either in text or in a table)?

☐ Yes.

☐ No, justification: