The Role of Simulation Studies in Evaluating Statistical Test Performance Under Model Assumption Violations

Leo Stewart, Lillian Adams, Logan Rivera

1 Introduction

Statistical hypothesis testing represents a cornerstone of empirical research across scientific disciplines, providing formal mechanisms for drawing inferences from data. The theoretical foundations of these tests typically rely on specific mathematical assumptions about the underlying data generating processes. Classical statistical education emphasizes these ideal conditions, often presenting tests as having well-defined properties when assumptions such as normality, independence, homogeneity of variance, and linearity are satisfied. However, in practical research applications, these assumptions are frequently violated to varying degrees, raising critical questions about test robustness and the validity of resulting inferences.

The challenge of assumption violations has been recognized since the early development of inferential statistics, with seminal work by Box and others examining the consequences of specific violations. Traditional approaches to this problem have often taken binary perspectives, classifying tests as either 'robust' or 'non-robust' to particular violations. This oversimplification fails to capture the nuanced reality that test performance typically degrades gradually rather than catastrophically, and that the severity of performance deterioration depends on multiple factors including sample size, effect size, and the specific nature and degree of assumption violation.

This research introduces a novel simulation-based framework that moves beyond traditional robustness assessments by quantifying test performance across continuous spectra of assumption violations. Our approach recognizes that real-world data rarely conform perfectly to theoretical assumptions, but

rather exhibits varying degrees of deviation. By systematically exploring these deviations through carefully designed simulation studies, we provide a more realistic and practical understanding of how statistical tests perform under conditions that researchers actually encounter.

The primary contribution of this work lies in developing a comprehensive methodology for evaluating test performance across multidimensional violation spaces. Rather than examining violations in isolation, as is common in existing literature, our framework simultaneously varies multiple assumption violations to better represent the complex interdependencies that occur in actual research data. This approach reveals that the combined effects of multiple minor violations can sometimes produce more severe performance degradation than single major violations, a finding with important implications for statistical practice.

Furthermore, we introduce the concept of 'violation tolerance thresholds' - specific points at which test performance degrades beyond acceptable limits for practical application. These thresholds provide concrete guidance for researchers making decisions about test selection and interpretation in the presence of assumption violations. Our work also develops novel visualization techniques for representing test performance across complex violation landscapes, making the results accessible to applied researchers without advanced statistical training.

Through extensive simulation studies focusing on commonly used parametric tests, we demonstrate that conventional wisdom about test robustness often requires revision. Some tests show unexpected resilience to certain types of violations while exhibiting surprising sensitivity to others. These findings challenge simplified recommendations found in many statistical textbooks and provide a more nuanced, evidence-based foundation for statistical test selection in applied research contexts.

2 Methodology

Our simulation framework employs a systematic approach to evaluating statistical test performance under assumption violations. The methodology consists of several interconnected components designed to comprehensively assess how tests behave when their underlying assumptions are not fully met.

The foundation of our approach lies in the careful specification of violation parameters that systematically deviate from ideal conditions. For

normality violations, we employ the generalized lambda distribution to generate data with controlled levels of skewness and kurtosis, ranging from minor deviations to severe departures from normality. This approach allows us to independently manipulate different aspects of distribution shape, providing insights into how specific distributional characteristics affect test performance. For violations of homogeneity of variance, we systematically vary variance ratios between groups while maintaining controlled sample size relationships. Independence violations are introduced through carefully specified autocorrelation structures in time series contexts and through cluster-based dependencies in hierarchical data scenarios.

A key innovation in our methodology is the development of multidimensional violation spaces that simultaneously vary multiple assumption violations. Traditional simulation studies typically examine violations in isolation, but real-world data often exhibits correlated violations across multiple assumptions. Our framework systematically explores these complex violation landscapes through factorial designs that combine different types and degrees of violations. This approach enables us to identify interaction effects between different types of violations and to understand how combined minor violations might produce more substantial performance degradation than isolated major violations.

The simulation design incorporates a comprehensive set of performance metrics beyond the conventional Type I error rate and power. While these traditional metrics remain important, we additionally examine effect size estimation accuracy, confidence interval coverage rates, and the behavior of p-value distributions under various violation conditions. We also introduce novel metrics that quantify the rate of performance degradation as violations increase, providing insights into how quickly tests become unreliable as assumptions are violated.

Our simulation studies examine several commonly used statistical tests across different research contexts. For comparing group means, we evaluate independent samples t-tests, paired t-tests, and one-way ANOVA under various violation conditions. For correlation and regression analyses, we examine Pearson correlation tests and linear regression models under violations of normality, homoscedasticity, and independence. Each test is evaluated across a range of sample sizes typically encountered in applied research, from small samples common in pilot studies to larger samples typical in well-powered investigations.

The simulation procedure follows a rigorous Monte Carlo approach with

extensive replication to ensure stable estimates of performance metrics. For each combination of violation parameters and sample size, we conduct 10,000 replications to obtain precise estimates of Type I error rates and power. This extensive replication is particularly important for accurately characterizing performance in the tails of the violation distribution where tests may exhibit unstable behavior.

Data generation employs computationally efficient algorithms that ensure precise control over violation parameters while maintaining realistic data structures. We implement careful randomization procedures and use high-quality pseudorandom number generators to minimize simulation artifacts. The entire simulation framework is implemented in a modular architecture that facilitates the addition of new tests and violation types, ensuring the methodology remains extensible for future research.

Analysis of simulation results employs both traditional statistical summaries and novel visualization techniques. We develop specialized plots that represent test performance across multidimensional violation spaces, making complex relationships accessible to researchers with varying statistical backgrounds. These visualizations help identify patterns in test performance that might be obscured in traditional tabular presentations of results.

3 Results

The simulation results reveal complex and often counterintuitive patterns in how statistical tests perform under assumption violations. Our findings challenge several conventional understandings of test robustness and provide new insights into the practical implications of assumption violations in applied research.

For the independent samples t-test, we observed that the test's sensitivity to normality violations depends critically on sample size and the specific nature of the distributional departure. Under skewness violations, the t-test maintained reasonable Type I error control with sample sizes as small as 30 per group, contradicting common recommendations requiring larger samples. However, under kurtosis violations, particularly with heavy-tailed distributions, the test showed substantial inflation of Type I error rates even with moderate sample sizes. The combination of skewness and kurtosis violations produced interactive effects that were not predictable from examining either violation in isolation.

The one-way ANOVA demonstrated surprising resilience to heterogeneity of variance when group sizes were equal, supporting the classical Box's finding. However, when group sizes were unequal in the direction of larger variances accompanying smaller samples, the test showed severe Type I error inflation even with moderate variance heterogeneity. This pattern was particularly pronounced with smaller sample sizes, suggesting that researchers should be especially cautious about variance heterogeneity in studies with unequal group sizes and limited participants per group.

Linear regression analyses revealed complex relationships between assumption violations and test performance. Violations of normality in the error distribution had minimal impact on Type I error rates for regression coefficients, supporting the often-cited robustness of regression to normality violations. However, these same violations substantially affected the accuracy of confidence intervals and power calculations. Heteroscedasticity produced more serious consequences, leading to both inflated Type I error rates and inaccurate standard error estimates. The severity of these effects depended on the pattern of heteroscedasticity, with increasing variance patterns producing different consequences than decreasing variance patterns.

Our examination of violation tolerance thresholds revealed that commonly used rules of thumb for acceptable violation levels often provide inadequate guidance. For example, the frequently cited guideline that F-max ratios below 4:1 indicate acceptable variance heterogeneity proved unreliable across different sample size conditions and test types. Instead, our results suggest that tolerance thresholds should be context-dependent, considering sample size, effect size, and the specific research question.

The multidimensional violation analyses produced particularly insightful results. We found that combinations of minor violations across multiple assumptions sometimes produced more severe performance degradation than single major violations. For instance, the combination of mild non-normality and slight heteroscedasticity, neither of which would be considered problematic individually, could produce substantial Type I error inflation in certain conditions. This finding highlights the importance of comprehensive assumption checking rather than focusing on individual assumptions in isolation.

Visualization of results across the violation spaces revealed complex performance landscapes with regions of stable performance, gradual degradation, and sudden performance collapse. These visual patterns provide intuitive guidance for researchers assessing the potential impact of assumption violations in their specific research contexts. The visualization techniques also

helped identify interaction effects between different types of violations that would be difficult to detect through traditional analytical approaches.

Performance degradation patterns varied non-linearly across violation severity continua. For many tests, performance remained relatively stable through initial violation levels, then deteriorated rapidly after crossing specific threshold points. The location of these threshold points depended on multiple factors including sample size, effect size, and the presence of other violations. This non-linear pattern suggests that researchers cannot safely assume that small increases in violation severity will produce proportionally small changes in test performance.

4 Conclusion

This research demonstrates the critical importance of simulation studies for understanding how statistical tests perform under realistic conditions where model assumptions are violated. Our findings challenge several conventional understandings of test robustness and provide a more nuanced, evidence-based foundation for statistical practice.

The primary contribution of this work lies in developing a comprehensive simulation framework that moves beyond traditional binary robustness classifications to provide continuous characterizations of test performance across spectra of assumption violations. This approach recognizes that real-world data rarely conforms perfectly to theoretical assumptions, and that researchers need guidance about how tests perform under the imperfect conditions they actually encounter. Our introduction of violation tolerance thresholds and multidimensional violation spaces provides practical tools for researchers making decisions about test selection and interpretation.

Our results reveal that test performance under assumption violations follows complex patterns that are often non-linear and interactive. The common practice of examining violations in isolation provides an incomplete picture of how tests behave in applied research contexts. The finding that combinations of minor violations can sometimes produce more severe consequences than individual major violations has important implications for statistical practice, suggesting that researchers should conduct comprehensive assumption checking rather than focusing on individual assumptions.

The simulation methodology developed in this research provides a template for future investigations of statistical test performance. The modular

framework can be extended to examine additional tests, violation types, and performance metrics. Future research could apply this approach to more complex statistical models, including multivariate techniques, mixed effects models, and structural equation models.

Several practical recommendations emerge from our findings. First, researchers should move beyond simple dichotomous decisions about assumption violations and instead consider the severity and combination of violations present in their data. Second, sample size considerations should include not only power calculations but also assessments of robustness to potential assumption violations. Third, graphical methods for assessing assumptions should be supplemented with quantitative measures that can be related to known performance characteristics.

This research also has implications for statistical education. Traditional emphasis on ideal conditions and binary robustness classifications should be supplemented with more realistic discussions of how tests perform under the imperfect conditions typical of applied research. Our visualization techniques provide accessible ways to communicate these complex relationships to students with varying statistical backgrounds.

In conclusion, simulation studies provide indispensable tools for understanding statistical test performance in real-world research contexts. By systematically exploring how tests behave across continua of assumption violations, we can develop more realistic guidance for statistical practice and move beyond oversimplified rules of thumb. The framework developed in this research represents a step toward evidence-based statistical decision making that acknowledges the complex reality of research data while maintaining the rigorous foundations of statistical inference.

References

Box, G. E. P. (1953). Non-normality and tests on variances. Biometrika, 40(3/4), 318-335.

Cohen, J. (1994). The earth is round (p ¡ .05). American Psychologist, 49(12), 997-1003.

Field, A. P. (2013). Discovering statistics using IBM SPSS Statistics: And sex and drugs and rock 'n' roll (4th ed.). Sage Publications.

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.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. Journal of Educational Statistics, 17(4), 315-339.

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.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105(1), 156-166.

Rasmussen, J. L. (1989). Parametric and non-parametric tests of homogeneity for independent groups. Psychological Bulletin, 105(3), 430-437.

Sawilowsky, S. S., Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the t test to departures from population normality. Psychological Bulletin, 111(2), 352-360.

Wilcox, R. R. (2012). Introduction to robust estimation and hypothesis testing (3rd ed.). Academic Press.