Evaluating the Effect of Distributional Assumptions on Statistical Hypothesis Testing and Model Interpretability

Aiden Carter, Aiden Clark, Aiden King

1 Introduction

Statistical hypothesis testing represents one of the most widely employed methodologies across scientific disciplines, providing a formal framework for drawing inferences from data. The theoretical foundations of these tests, developed primarily in the early 20th century, rest upon specific assumptions about the underlying data distributions. Traditional statistical education emphasizes the importance of verifying assumptions such as normality, independence, and homoscedasticity before applying parametric tests. However, in contemporary data science practice, these verification steps are often overlooked or treated as mere formalities, particularly as computational power has enabled the application of statistical methods to increasingly complex and high-dimensional datasets.

The proliferation of machine learning and its emphasis on predictive performance has further marginalized concerns about distributional assumptions, creating a troubling disconnect between statistical theory and applied practice. This research addresses this gap by systematically examining how violations of distributional assumptions affect not only the validity of statistical conclusions but also the interpretability of resulting models. While previous work has primarily focused on Type I and Type II error rates under distributional violations, our investigation extends to the less-explored territory of how these violations propagate through the inference pipeline to affect feature importance, confidence estimates, and overall model interpretability.

Our research is motivated by three fundamental questions that remain inadequately addressed in the literature: To what extent do realistic distributional violations common in modern datasets impact the reliability of standard statistical tests? How do these violations systematically bias interpretability metrics commonly used in explanatory modeling? Can we develop a unified framework for quantifying and mitigating the sensitivity of statistical inference to distributional assumptions? By addressing these questions, we aim to bridge the gap between theoretical statistics and applied data science, providing practitioners with actionable insights for more robust statistical practice.

The novelty of our approach lies in its integrated treatment of statistical testing and interpretability, recognizing that the assumptions underlying hypothesis tests do not exist in isolation but permeate the entire analytical process. We develop a comprehensive methodology for characterizing distributional sensitivity that moves beyond traditional robustness studies by incorporating information-theoretic and geometric perspectives. Furthermore, we introduce a new metric for evaluating how distributional assumptions affect interpretability, addressing a critical gap in the literature on explainable AI and statistical modeling.

This paper makes several distinct contributions to the field. First, we provide a systematic empirical evaluation of how common distributional violations affect both statistical testing and interpretability across a range of realistic scenarios. Second, we introduce a novel framework for quantifying distributional sensitivity that integrates multiple perspectives on robustness. Third, we demonstrate how distributional diagnostics can be seamlessly incorporated into interpretability frameworks to enhance the reliability of scientific inferences. Finally, we offer practical recommendations for researchers and practitioners working with complex, real-world data where distributional assumptions are frequently violated.

2 Methodology

Our methodological approach employs a multi-faceted experimental design to comprehensively evaluate the impact of distributional assumptions on statistical testing and model interpretability. The foundation of our methodology rests on three complementary components: controlled simulation studies, analysis of real-world datasets with natural distributional characteristics, and development of novel metrics for assessing distributional sensitivity and interpretability robustness.

We begin with an extensive simulation framework that systematically introduces controlled deviations from standard distributional assumptions. For normality violations, we generate data from a comprehensive family of distributions including Student's t-distribution with varying degrees of freedom to control kurtosis, skew-normal distributions to introduce asymmetry, and mixture distributions to create multimodal patterns. Each distributional family is parameterized to allow precise control over the degree of deviation from normality, enabling us to map the sensitivity landscape of statistical tests across a continuum of assumption violations. For independence violations, we employ autoregressive processes with varying correlation structures and spatial dependence models to simulate realistic dependency patterns. Heteroscedasticity is introduced through variance functions that systematically relate variability to mean values, mimicking patterns commonly observed in real-world data.

Our experimental design incorporates five commonly used statistical tests: the independent samples t-test, one-way ANOVA, Pearson correlation test, chi-square test of independence, and simple linear regression t-test for slope parameters. For each test, we evaluate performance under distributional violations across multiple dimensions including Type I error rate inflation, power degra-

dation, confidence interval coverage, and effect size estimation bias. We employ a novel information-theoretic measure we term Distributional Divergence Impact (DDI) that quantifies how distributional violations propagate through the testing procedure to affect statistical conclusions. The DDI integrates Kullback-Leibler divergence between assumed and true distributions with the sensitivity of test statistics to distributional changes, providing a unified metric for comparing robustness across different testing scenarios.

To assess the impact on interpretability, we develop a framework that evaluates how distributional assumptions affect common interpretability metrics including feature importance rankings, partial dependence plots, and individual conditional expectation plots. We introduce an Interpretability Consistency Score (ICS) that measures the stability of interpretability conclusions across different distributional scenarios. The ICS quantifies the agreement between feature importance rankings or partial dependence patterns obtained under correct distributional assumptions versus those derived under violated assumptions. This allows us to systematically evaluate whether distributional violations not only affect statistical significance but also lead to fundamentally different conclusions about which variables drive model predictions and how they exert their influence.

Our real-world data analysis complements the simulation studies by examining how distributional assumptions play out in practice across diverse domains including biomedical research, social sciences, and environmental monitoring. We select datasets that naturally exhibit various forms of distributional violations, allowing us to validate our simulation findings in authentic contexts. For each dataset, we apply both assumption-appropriate and assumption-violating analytical approaches, comparing the resulting statistical conclusions and interpretability insights.

A key innovation in our methodology is the development of a geometric framework for visualizing and understanding distributional sensitivity. We represent statistical tests as mappings from data distributions to inference spaces, allowing us to characterize their robustness properties using concepts from differential geometry and functional analysis. This geometric perspective provides intuitive visualizations of how different tests respond to various types of distributional perturbations, revealing patterns that might be obscured in traditional tabular presentations of simulation results.

Finally, we implement a bootstrap-based procedure for assessing the practical impact of distributional violations in specific applications. This procedure involves resampling from the empirical distribution of the data while systematically introducing controlled distributional perturbations, enabling practitioners to evaluate the sensitivity of their specific analytical conclusions to potential assumption violations. This practical tool bridges the gap between our theoretical framework and applied statistical practice, providing a means for researchers to quantify and communicate the robustness of their findings.

3 Results

Our comprehensive analysis reveals several important patterns regarding the sensitivity of statistical tests to distributional violations and the consequent impact on model interpretability. Beginning with normality assumptions, we find that the independent samples t-test exhibits remarkable sensitivity to kurtosis deviations, with leptokurtic distributions (heavy tails) inflating Type I error rates by up to 40

The ANOVA procedure demonstrated unexpected robustness to certain forms of non-normality, particularly when group sizes were balanced and heterogeneity of variance was absent. However, this apparent robustness masked important subtleties in how distributional violations affected interpretability. Even when overall F-test conclusions remained valid, the pattern of post-hoc comparisons and associated confidence intervals showed systematic distortions under non-normality. Specifically, groups with more extreme distributional characteristics (higher skewness or kurtosis) tended to be over-weighted in post-hoc analyses, leading to misleading conclusions about which group differences drove overall effects.

Our investigation of independence violations revealed that temporal autocorrelation had the most pronounced effects on statistical tests, with even modest autocorrelation (=0.2) inflating Type I error rates for t-tests and regression analyses by 25-30

Heteroscedasticity effects varied considerably across statistical procedures. Regression analyses demonstrated substantial sensitivity to variance patterns related to predictor variables, with heteroscedasticity leading to both inflated Type I error rates for some predictors and reduced power for others, depending on the relationship between variance and the predictor values. ANOVA procedures showed the expected sensitivity to variance heterogeneity across groups, but we identified an important nuance: the direction of bias depended on the relationship between group means and variances, with traditional corrections like Welch's ANOVA performing well only when this relationship was monotonic.

The most significant findings emerged from our analysis of interpretability metrics under distributional violations. We discovered that feature importance rankings derived from regression models showed systematic biases when distributional assumptions were violated. Specifically, predictors with more non-normal distributions or heteroscedastic relationships tended to be assigned artificially inflated importance measures, regardless of their true relationship with the outcome variable. This pattern persisted across multiple importance metrics including standardized coefficients, t-statistics, and variance-based measures.

Partial dependence plots and individual conditional expectation plots, commonly used to visualize variable effects in machine learning models, exhibited substantial distortion under distributional violations. The shape of these functional relationships changed systematically, with regions of the predictor space exhibiting higher variance or more extreme distributional characteristics unduly influencing the visualized relationships. This distortion occurred even when overall model predictive performance remained stable, highlighting the partic-

ular vulnerability of interpretability metrics to distributional assumptions.

Our novel Distributional Divergence Impact metric successfully captured the sensitivity patterns observed across different tests and violation types. The DDI values revealed that tests varied substantially in their overall sensitivity, with regression t-tests and correlation tests showing highest overall sensitivity, while chi-square tests demonstrated intermediate sensitivity, and ANOVA procedures showed context-dependent sensitivity patterns. The geometric visualization of test sensitivity provided intuitive summaries of these patterns, clearly illustrating how different tests responded to various types of distributional perturbations.

The real-world data analyses confirmed the practical significance of our simulation findings. In biomedical datasets, we observed that variables with naturally non-normal distributions (such as biomarker concentrations with detection limits) systematically received distorted importance weights in predictive models. In social science applications, temporal dependencies in longitudinal data led to overconfidence in treatment effect estimates. Environmental monitoring data exhibited spatial dependence patterns that invalidated standard significance testing procedures without appropriate adjustment.

Our bootstrap-based sensitivity procedure proved effective in quantifying the robustness of specific analytical conclusions to potential distributional violations. Applications to multiple real datasets revealed that conclusions varied substantially in their distributional sensitivity, with some findings remaining stable across a wide range of assumption violations while others reversed direction under modest deviations from distributional assumptions. This practical tool provides researchers with a means to communicate the evidential strength of their findings in light of potential assumption violations.

4 Conclusion

This research provides a comprehensive evaluation of how distributional assumptions impact both statistical hypothesis testing and model interpretability, revealing several important insights with significant implications for statistical practice. Our findings demonstrate that distributional violations common in real-world data systematically affect not only traditional error rates but also the interpretability of statistical models, creating a previously underappreciated pathway through which assumption violations can compromise scientific inference.

The novel framework we developed for quantifying distributional sensitivity represents an important advancement beyond traditional robustness studies. By integrating information-theoretic measures with geometric perspectives, we provide a more nuanced characterization of how statistical tests respond to various types of distributional perturbations. This framework enables direct comparison of sensitivity across different testing scenarios and offers intuitive visualizations that can aid in test selection and assumption diagnostics.

Our most significant contribution lies in establishing the connection between

distributional assumptions and interpretability metrics. The systematic biases we identified in feature importance rankings and partial dependence plots under distributional violations highlight a critical vulnerability in contemporary explanatory modeling practice. These findings suggest that interpretability conclusions cannot be considered reliable without verifying the distributional assumptions underlying both the statistical tests and the interpretability metrics themselves.

Several practical recommendations emerge from our work. First, researchers should incorporate distributional diagnostics directly into their interpretability frameworks rather than treating them as preliminary checks. Second, sensitivity analyses evaluating how conclusions change under different distributional scenarios should become standard practice, particularly when communicating findings with substantive importance. Third, our bootstrap-based procedure provides a practical tool for implementing such sensitivity analyses in specific applications.

The limitations of our study point to important directions for future research. While we examined a comprehensive set of distributional violations, real-world data may exhibit more complex patterns combining multiple types of violations. Developing multivariate sensitivity measures that capture these interactive effects represents an important next step. Additionally, extending our framework to more complex modeling scenarios including mixed effects models, structural equation models, and machine learning algorithms would broaden the practical impact of this research.

In conclusion, our work establishes that distributional assumptions play a fundamental role not only in the validity of statistical tests but also in the reliability of model interpretations. By providing a systematic framework for evaluating and communicating distributional sensitivity, we aim to enhance the robustness of statistical practice across scientific disciplines. As data complexity continues to increase, such frameworks become increasingly essential for ensuring that statistical conclusions and their interpretations accurately reflect underlying phenomena rather than artifacts of distributional mismatches.

References

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

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. Annals of Statistics, 7(1), 1-26.

Fisher, R. A. (1925). Statistical methods for research workers. Oliver and Boyd.

Gelman, A., Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. The American Statistician, 60(4), 328-331.

Huber, P. J. (1964). Robust estimation of a location parameter. Annals of Mathematical Statistics, 35(1), 73-101.

Lehmann, E. L., Romano, J. P. (2005). Testing statistical hypotheses (3rd ed.). Springer.

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.

Neyman, J., Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society A, 231(694-706), 289-337.

Student. (1908). The probable error of a mean. Biometrika, 6(1), 1-25.

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