EVALUATION OF THE UTILITY OF NORMALITY TESTS IN STATISTICAL PRACTICE

by

Benedict Kongyir

Ph.D. Student

A Dissertation Proposal

To be Presented to Committee Members

Dr. Pratyaydipta Rudra, *Advisor*

Dr. Joshua Habiger, *Member*

Dr. Liang Ye, *Member*

Dr. Liz Mccullagh, *Outside Member*

Department of Statistics

Oklahoma State University

Abstract

Normality testing is widely used for assessing whether a given dataset conforms to the normal distribution. While it has been discussed in the literature that pre-testing normality can be problematic, there has not been research quantifying the effects of normality pre-testing on downstream inference. For instance, the normality tests can lack utility due to their sensitivity to the sample size. There is also a concern of selective inference if a choice about downstream tests is made based on the results of the normality tests. This research aims to study and quantify these effects by developing a comprehensive simulation-based framework enabling informed decisions regarding normality testing and downstream test procedures. The framework is implemented in R providing a user interface with R Shiny. It allows users to assess the normality of their data using several existing methods, as well as information on the impact of normality test procedures on their chosen downstream procedure. In addition, our research also explores robust machine-learning models for classifying the probability distribution generating the data into 'normal' and 'non-normal'. We demonstrate that our proposed method generalizes across various sample sizes and distributions and delivers a favorable performance compared to the traditional normality tests.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

The validity of most statistical inferences in scientific research is dependent upon the assumptions of the models used. Consider a scenario in clinical trials where a statistically significant results for a new drug is published. If the underlying statistical procedure, such as a t-test which validity is reliant on the normality assumption, is applied to non-normal data, the actual Type I error rate may be inflated. For instance, a nominal $\alpha = 0.05$ could inflate to 10%, meaning one in ten positive conclusions maybe incorrect. This illustrates the critical, yet often overlooked, role of assumption checking in statistical analysis.

Violations of distributional assumptions, such as normality, constitute a pervasive concern in statistical practice because they can be difficult to detect yet substantially impact inferences. Standard tests for normality such as Shapiro-Wilk test (Shapiro and Wilk, 1965), and Kolmogorov-Smirnov test (An, 1933) are often less powerful to detect non-normality in small samples, yet overly sensitive to trivial, practically non-significant deviations in large samples (Thode, 2002). This can lead researchers to either retain parametric tests when they are inappropriate or unnecessarily switch to less powerful nonparametric tests, potentially compromising the integrity of their conclusions.

The assumption of normality is foundational to many classical statistical procedures, including t-test, ANOVA, and linear regression. The theoretical justification for these methods often depend on the premise that the data, or the errors of the model, are normally distributed. When this assumption is significantly violated, the theoretical properties of resulting p-values and confidence intervals may not hold, potentially invalidating statistical inferences (Box, 1976). Such erroneous inferences can have tangible implications. For example, they may impact the evaluation of medical treatments, the efficacy of business interventions, and public policy (Ioannidis, 2005).

As we shall demonstrate in our simulation study in chapter 2, violations of normality contributes to inflation of type I error rates and or loss in statistical power for many common downstream procedure, thereby undermining validity of conclusions in statistical research (Rochon et al., 2012). To address these concerns, statistical practice has evolved to include normality pre-testing a preliminary assessment procedure that evaluates whether observed data are consistent with a normal distribution before implementing parametric methods. However, the utility of such pre-testing remains heavily contested (Schucany and Tony Ng, 2006). In small samples, where violations of normality assumptions are most consequential for parametric procedures, normality tests typically lack sufficient power to detect meaningful departures from normality. This creates a particularly challenging scenario; precisely when researchers most need to identify non-normality to avoid inappropriate use of parametric methods, the tests are least capable of doing so. Conversely, in large samples where the Central Limit Theorem provides asymptotic protection to many parametric procedures, normality tests become hypersensitive, flagging trivial deviations that may hold little practical relevance for robust parametric methods (Thode, 2002; Kwak and Kim, 2017).

This sample size paradox reveals a fundamental disconnect between the operating characteristics of normality tests and the actual robustness properties of the parametric procedures they are meant to protect. As sample sizes increase, parametric methods become increasingly robust to moderate deviations from normality due to the Central Limit Theorem, yet this is precisely when normality tests have the highest power to detect even minor departures (Kwak and Kim, 2017).

On one hand, pre-testing for normality can be beneficial. Correctly identifying non-normality may prevent the inappropriate use of parametric methods that depend on normality, thereby reducing the risk of invalid inference. Conversely, if the assumption of normality holds, parametric procedures often provide higher power than corresponding nonparametric tests. Thus, a correct pre-test outcome could protect against unnecessary loss of efficiency and retain the advantages of parametric inference (Thomas Lumley, 2002; Lehmann and Romano, 2005).

On the other hand, using normality tests as a preliminary decision rule can introduces complications. Pre-testing for normality adds an additional stage to the analysis process. It can also reduce the downstream test power. An incorrect rejection of normality may lead to the unnecessary use of less powerful nonparametric procedures, even in contexts where the intended parametric method is robust to the existing deviation. These concerns are well documented in the literature. For example Schucany and Ng (2006); Rochon and Kieser (2011) have demonstrated that pretesting can distort the distribution of test statistics, alter the correlation structure among sample moments, and ultimately lead to unreliable inferential outcomes, including inflated Type I error rates and loss of power.

Beyond theoretical concerns, the performance of common normality tests presents practical challenges. Tests such as Shapiro–Wilk, Anderson–Darling, and Kolmogorov–Smirnov differ in sensitivity depending on the nature of deviation from normality, but all suffer from context-dependent limitations making it unclear what test to use in a given situation.

For decades, the task of assessing normality has been entrusted to a suite of traditional statistical tests, such as the Shapiro-Wilk test (Shapiro and Wilk, 1965), Anderson-Darling test (Anderson and Darling, 1954), etc. These tests often target different aspects of distribution characteristics, such as skewness, kurtosis, or tail behavior,meaning none is uniformly most powerful. This makes the choice of test non-trivial and fragments the diagnostic process.

Graphical assessment tools such as histograms, box-and-whisker plots, and the quantile-quantile plots are also commonly used to make decisions about normality of data. Despite good at detecting outliers and deviations at the tails, they are not very reliable due to subjectivity in their interpretation.

The fundamental challenge extends beyond the limitations of individual tests or plots. The very practice of using data-driven methods to select analytical procedures creates a selection bias that distorts the sampling distribution of subsequent test statistics, ultimately compromising the validity of inferential conclusions (Benjamini, 2020). This creates a fundamental problem for anyone analyzing data. There is no way to know in advance whether checking for normality will help or hurt their results. Standard textbooks and software defaults offer no practical solution to this problem. This work aims to address this methodological gap. We develop a framework to quantify the impact of normality pre-testing on subsequent inference, finally providing actionable advice on when to use it.

1.2 Research Objectives

Based on the critical review of existing literature, this dissertation seeks to address fundamental limitations in normality assessment through three primary research objectives:

1. **Developing a Framework for Evaluation of Normality Test Utility**
   Develop a framework based on systematic simulation studies to quantify the practical value and limitations of existing normality tests across diverse scenarios, including varying sample sizes, distributional characteristics, and downstream analytical contexts. Develop a metric for quantifying the practical utility of pre-testing for normality in varying scenarios that can help researchers in selecting appropriate normality assessment strategies based on their specific data characteristics, analytical goals, and tolerance for inference errors.

2. **Develop a Robust Machine Learning Approach to Normality Testing**
   Design, train, and validate advanced machine learning models for normality assessment that demonstrate superior accuracy, robustness to varying distributional scenarios, and generalizability across different sample sizes and data characteristics.

3. **Address Selective Inference in Normality Pre-testing**
   Develop and validate methodological frameworks to mitigate selection bias introduced by normality pre-testing, adapting and extending selective inference strategies to preserve the validity of subsequent statistical conclusions.

1.3 Significance of the Research

This research addresses a fundamental paradox in statistical practice: while many introductory statistics course teaches normality testing as an essential preliminary step, mounting evidence suggests this practice may sometimes do more harm than good. The significance of this work lies in its comprehensive approach to resolving this contradiction through creating a user decision framework. Below we further elaborate different aspects of our contribution.

1.3.1 Empirical Validation of Normality Test Utility

The systematic evaluation of normality test utility addresses a critical gap between statistical theory and practice. For decades, researchers have often followed textbook recommendations to "check for normality" before conducting parametric tests, without understanding the actual consequences of this practice. As Rochon et al. (2012) demonstrated, this pre-testing approach can distort Type I error rates and lead to invalid conclusions. This research provides the first comprehensive simulation-based framework to quantify when normality testing actually helps versus when it harms downstream inference. By systematically varying sample sizes, distributional characteristics, and analytical contexts, this work moves beyond the oversimplified "pass/fail" mentality that dominates current practice. It answers the crucial question that practitioners actually face: "Given my specific research context, will testing for normality improve or degrade my conclusions?"

1.3.2 Paradigm Shift in Distributional Assessment

The development of machine learning approaches represents a paradigm shift in how we conceptualize distributional assessment. Traditional normality tests suffer from what we might call the "single-alternative problem"—each test is optimized to detect specific types of non-normality but performs poorly against others. The Shapiro-Wilk test excels against asymmetric distributions but may miss heavy-tailed alternatives, while the Anderson-Darling test detects tail deviations but can overlook symmetry issues (Razali and Wah, 2011). This research breaks this limitation by developing ensemble machine learning classifiers that simultaneously consider multiple distributional characteristics. Drawing inspiration from early neural network approaches by Wilson and Engel (1990) and more recent work by Simić (2021), this research advances the field by creating models that learn the complex, multi-faceted "signature" of normality rather than relying on any single statistical property. This represents a fundamental shift from hypothesis testing to pattern recognition in distributional assessment.

1.3.3 Addressing the Silent Killer of Replicability

The issues arising with selective inference is a well known research area. Benjamini (2020) called selective inference the "silent killer" of replicability. The two-stage process of testing for normality then choosing an analysis based on the result creates a form of selection bias that traditional statistical theory doesn't account for. When researchers use the same data to check assumptions and conduct inference, they engage in what statisticians call "double-dipping"—a practice that invalidates nominal error rates. Our research explores the effect of such selective inference in the context of normality pre-testing and proposes solutions using existing frameworks, particularly the conditional inference approaches developed by Lee et al. (2016) and the data carving methods of Fithian et al. (2014), to specifically address the unique dependency structures created by normality pre-testing. By developing correction procedures tailored to this context, the research provides a methodological bridge between modern selective inference theory and everyday statistical practice.

1.3.4 Broader Impact and Scientific Contribution

The broader impact of this research extends beyond normality testing to challenge how we think about statistical assumptions more generally. Many statistical procedures rely on assumptions that are routinely checked using the same data used for inference. This research provides a template for how to rigorously evaluate the costs and benefits of such pre-testing strategies across different statistical contexts.

Perhaps most importantly, this work recognizes that statistical methods exist to serve scientific inquiry, not the other way around. By moving beyond the rigid dichotomy of "normal" versus "non-normal" and toward a more nuanced understanding of distributional characteristics and their practical implications, this research empowers scientists to make better-informed analytical decisions. It replaces blind ritual with evidence-based practice, contributing to what might be called a "post-normal" approach to statistical inference—one that acknowledges the complexity of real-world data while providing practical tools for valid scientific discovery.

In an era where statistical rigor and reproducibility are paramount concerns across scientific disciplines (Ioannidis, 2005), this research provides both a critical examination of current practices and a constructive path forward. It demonstrates that sometimes the most fundamental statistical practices deserve the closest scrutiny, and that improving them requires not just new methods, but new ways of thinking about the relationship between data, assumptions, and inference.

1.4 Organization of the Dissertation

This dissertation is structured to systematically address the research objectives through a logical progression from foundational concepts to methodological development and practical application. The organization reflects the iterative nature of the research process, moving from critical assessment of existing methods to the development and validation of novel approaches, culminating in practical implementation and future directions.

1.4.1 Chapter Overview

Chapter 1: Introduction establishes the research context by outlining the fundamental importance of normality assumptions in statistical inference and identifying the key limitations of current normality testing practices. It presents the research problem with a practical example, and articulates the four primary research objectives that guide this investigation. The chapter concludes by highlighting the broader significance of this work for statistical practice and scientific reproducibility.

Chapter 2: Evaluating Normality Test Utility and Developing a Decision Framework presents a comprehensive methodology for assessing the practical value of normality tests and integrates these findings into a cohesive decision-support system. It begins with a historical survey of traditional normality tests, tracing their development from early goodness-of-fit measures to modern specialized tests. The review then synthesizes comparative studies that benchmark test performance across different scenarios. This chapter details the simulation design, including data generation procedures, performance metrics, and analytical scenarios. It describes the comprehensive evaluation protocol that examines test behavior across varying sample sizes, distributional characteristics, and downstream analytical contexts. The results from this chapter provide the empirical foundation for understanding when normality testing provides genuine utility versus when it introduces more problems than it solves. Building on these empirical results, the chapter develops a user-friendly framework that guides researchers in selecting appropriate normality assessment strategies based on their specific context, including interactive tools for cost-benefit analysis and protocol recommendations. The chapter concludes with validation through case studies and real-data applications across multiple scientific domains.

Chapter 3: Machine Learning Framework for Normality Assessment introduces the development of robust machine learning classifiers for normality testing. The chapter details the feature engineering process, model selection rationale, and training methodology. It presents the validation framework that ensures model generalizability across different distributional scenarios and sample sizes. Special attention is given to addressing the unique challenges of applying machine learning to statistical testing, including interpretability concerns and robustness requirements.

Chapter 4: Addressing Selective Inference in Normality Pre-testing tackles the critical methodological challenge of selection bias. This chapter adapts and extends selective inference frameworks specifically for the context of normality assessment. It presents both theoretical derivations and practical implementations of correction procedures, including conditional inference methods and data carving approaches. The chapter evaluates the effectiveness of these methods in preserving Type I error rates and maintaining statistical power across various pre-testing scenarios.

Chapter 5: Conclusion and Future Directions combines the major findings and contributions of the research. It revisits the four research objectives and summarizes how each was addressed through the methodological developments presented in the dissertation. The chapter discusses the broader implications for statistical practice, methodological limitations, and promising avenues for future research. It concludes with specific recommendations for researchers, educators, and statistical software developers.

Chapter 2

A Framework for Assessing the Utility of Normality Tests

2.1 Introduction

This chapter is dedicated to developing a rigorous framework for evaluating the utility of normality pre-testing for downstream statistical inference. Normality remains a critical assumption in many parametric procedures including t-tests, ANOVA, and linear regression. However, the practice of pre-testing for normality involves a critical trade-off. While it can guide test selection, it can also distort Type I error and power. It can even introduce selection bias into the downstream procedure leading misleading results. We evaluate this trade-off via large-scale simulation for common statistical procedures, measuring its impact through expected power loss, power gain, and Type I error inflation.

2.2 Literature Review

The assumption of normality is fundamental to many statistical methods, and the tools for assessing it have evolved considerably. While a large body of research compares these tests under different conditions, their practical utility remains uncertain. A primary concern is their sensitivity to sample size, which affects both power and Type I error control. Furthermore, using these tests to choose a subsequent analysis can introduce selection bias, invalidating the final inference (Benjamini, 2020; Rochon and Kieser, 2011).

This section reviews the development and performance of normality tests, from early methods to modern approaches. It examines their comparative effectiveness and critically assesses their practical value, focusing on the often overlooked consequences of using them in a two-stage testing procedure.

2.2.1 The Evolution of Normality Tests

Methods for evaluating normality typically fall into two categories: formal statistical tests and informal graphical assessments such as quantile-quantile (Q-Q) plots and histograms. Graphical approaches allow for flexible, intuitive examination of distributional features, yet their interpretation can be subjective and not suited for formal inference (Yap and Sim, 2011). Early contributions to graphical approach included those of pointwise confidence bands (Wilk and Gnanadesikan, 1968; Filliben, 1975; Rosenkrantz, 2000). Building on those foundations, Aldor-Noiman et al. (2013) proposed simultaneous confidence bands to provide an objective decision guide to Q-Q plot interpretation, obtaining performance comparable to traditional tests like Shapiro-Wilk, and Kolmogorov-Smirnov tests. However, the authors noted limitations, including the absence of a closed-form expression for the confidence bands and their non-standard asymptotic behavior compared to established tests.

The Kolmogorov-Smirnov (KS) test (kolmogorov, 1933) , developed by Andrey Kolmogorov in 1933, was one of the first formal goodness-of-fit tests. It compares empirical and theoretical cumulative distribution functions. While the KS test is straightforward, it can perform poorly and not very recommended especially when distribution parameters such as the mean and variance are to be estimated from the data (Ghasemi and Zahediasl, 2012).

To address some of the limitations of the KS test, Lilliefors (1967) proposed a modification specifically for normality testing with unknown parameters. The Lilliefors test is generally preferred over the KS test when the mean and variance must be estimated from the sample (Lilliefors, 1967). Around the same time, Anderson and Darling (1954) introduced Anderson-Darling(AD) test, which extends the KS approach by placing more weight to tail deviations, making it more sensitive to heavy tailed distributions like t and Cauchy. This makes the AD test particularly sensitive to heavy-tailed distributions like the Student's t and Cauchy. It is however less powerful in detecting deviations from symmetric, light-tailed distributions such as the uniform (Thode, 2002).

A major advancement came with the introduction of the Shapiro-Wilk (SW) test (Shapiro and Wilk, 1965). This regression-based test assesses normality by measuring the correlation between ranked sample values and their expected counterparts from a normal distribution. It performs especially well with moderate to large samples and is consistently ranked among the most powerful tests for detecting different types of non-normal data, particularly skewed distributions (Razali and Wah, 2011). To improve computational efficiency, Francia (1972) proposed a simplified version, the

Shapiro-Francia (SF) test, which, while slightly less powerful in small samples, offers a practical balance between speed and reliability.

The 1970s also saw the development of moment-based tests. d'Agostino (1971) introduced an omnibus test that transforms sample skewness and kurtosis into approximately standard normal variables, to detect departures from normality. Similarly, the Jarque-Bera (JB) test (Jarque and Bera, 1980) became popular for assessing normality through sample skewness and kurtosis. However, the JB test has been criticized for its low power, particularly in small samples or for slight deviations (Islam, 2019; Bonett and Seier, 2002). A robust version of the JB test (RJB) has been proposed to improve performance in the presence of outliers (Bayoud, 2021).

Further refinements have led to specialized tests that combine elements from multiple approaches. The Bonett-Seier test uses robust estimators of skewness and kurtosis and has demonstrated high power against slightly skewed and leptokurtic alternatives (Bonett and Seier, 2002; Islam, 2019). More recently, Sadhanala et al. (2019) proposed a Higher-Order Kolmogorov-Smirnov (HOKS) test that incorporates smoothness constraints to enhance tail sensitivity. In a different vein, Meng and Jiang (2023) developed a Cauchy Combination Omnibus Test (CCOT) that integrates the p-values of the AD, SW, and JB tests via a weighted Cauchy transformation, aiming to harness the distinct strengths of each individual test.

2.2.2 Comparative Performance of Normality Tests

A significant part of the literature has compared the performance of these diverse normality tests, typically through Monte Carlo simulation studies. A consistent finding across this literature is that no single test is uniformly most powerful; their relative performance depends critically on the nature of the departure from normality, the sample size, and the specific alternative distribution (Razali and Wah, 2011; Yap and Sim, 2011).

Early work by d'Agostino (1971) established that different tests excel at detecting different types of non-normality. This was confirmed by Razali and Wah (2011), who found the Shapiro-Wilk test generally most powerful but noted each test has distinct strengths. Yap and Sim (2011) further demonstrated that while the Shapiro-Wilk test performs best overall, the Anderson-Darling test is particularly sensitive to heavy-tailed distributions.

Islam (2019) provided a detailed ranking of tests using a framework that measured deviations from theoretical Neyman-Pearson power limits. They found the Bonett and Seier (2002) test to be most effective for slightly skewed distributions, while the Anderson-Darling and Shapiro-Wilk tests performed better with moderate skewness. The Jarque-Bera test and its robust variant consistently exhibited poor power across all sample sizes. Similarly, Bayoud (2021) proposed a new EDFbased test and found it to be competitive, particularly for bounded or asymmetric alternatives, but concluded that no single test dominates all others in all situations.

The consensus is clear: the choice of a normality test should be informed by prior knowledge of the likely deviations. However, in practice, the true underlining distribution is usually unknown, creating a significant dilemma for the user. Common recommendations include combining graphical methods with formal tests, though this does not fully resolve the fundamental issues of power and sensitivity inherent in the tests themselves.

2.2.3 The Utility of Normality Testing

The primary purpose of normality testing is not to prove normality, but to ensure the validity of subsequent parametric inference (Pearson, 1930). However, these tests face significant practical limitations that question their routine application.

The effectiveness of these tests varies dramatically with sample size. They often fail to detect meaningful violations in small samples while flagging trivial deviations in large datasets (Rochon et al., 2012; Schucany and Tony Ng, 2006). This creates a paradox where tests perform worst when most needed. Furthermore, many common procedures like the t-test are robust to moderate non-normality, particularly in equal samples (Schucany and Tony Ng, 2006), reducing the necessity for strict testing. Geary (1947) famously noted that "normality is a myth," emphasizing that perfect normality rarely exists in practice. This perspective suggests that researchers should focus on the degree and practical significance of non-normality rather than pursuing statistical significance in normality tests. Alternative strategies such as choosing between parametric and non-parametric procedures based on substantive differences in their results rather than preliminary normality testing are possible options though not perfect (Zimmerman, 2011). Others have advocated for robust statistical methods that maintain efficiency across various distributional scenarios (Wilcox, 2012). However, these alternatives present their own challenges, including potential efficiency loss when data are truly normal and ambiguous decision criteria.

2.3 Methodology

This section outlines the process of quantifying the utility of normality test through methodological development of assessment criteria and obtaining estimates of them using Monte-Carlo simulation.

2.3.1 Criteria quantifying the utility of normality pre-testing

We begin with some simple definitions and propositions to lay the foundation for the framework of quantifying the utility of pre-testing for normality. We present two approaches: 1) for any fixed sample size, and 2) across a given set of sample sizes.

Definition 2.3.1 (Normality Pre-test)**.** *Let $\mathcal{D} = \{X_1, \ldots, X_n\}$ denote a random sample from an unknown distribution F with finite mean and variance. Let N_T denote any normality test procedure. Define the* normality pre-test *as a function $\phi : \mathbb{R}^n \to \{0, 1\}$ where $\phi(\mathbf{X}) = 1$ indicates rejection of the null hypothesis $H_0 : F \in \mathcal{N}$ at significance level α_{pre}, with $\mathcal{N}$ denoting the family of normal distributions.*

Definition 2.3.2 (Adaptive Test Procedure)**.** *Let **Test 1** be a downstream test procedure whose validity relies on the normality assumption (e.g., a two-sample t-test) and **Test 2** be a downstream test procedure that is valid without distributional assumptions (e.g., a permutation test). The adaptive test procedure based upon a given normality pre-test N_T is defined as:*

$$T_{adaptive} = \begin{cases} T_2 & \textit{if } N_T \textit{ is significant at } \alpha_{pre} \\ T_1 & \textit{otherwise} \end{cases}$$

where T_1 and T_2 represent Test 1 and Test 2 respectively, applied to the same dataset $\mathcal{D}$.

Definition 2.3.3 (Conditional Type I Error)**.** *Let $A = \phi(\mathbf{X}) = 1$ denote the event that the normality test rejects H_0 of normality. Then, the conditional Type I error rates under the null hypothesis $H_0 : \theta \in \Theta_0$ are:*

$$\alpha_1(F|A^c) = \mathbb{P}_F(\delta_1 = 1|\phi(\mathbf{X}) = 0, H_0) \quad \textit{and} \quad \alpha_2(F|A) = \mathbb{P}_F(\delta_2 = 1|\phi(\mathbf{X}) = 1, H_0)$$

where

$$\delta_j = \begin{cases} 1 & \textit{if } T_j \textit{ rejects } H_0 \\ 0 & \textit{otherwise} \end{cases}$$

for $j \in \{1, 2\}$

Proposition 2.3.4 (Unconditional Type I Error Rate of Adaptive Procedure)**.** *The overall Type I error rate of the adaptive procedure is:*

$$\alpha_{adapt}(F) = \mathbb{P}_F(A^c)\alpha_1(F|A^c) + \mathbb{P}_F(A)\alpha_2(F|A) \tag{2.1}$$

where $A = \{\phi(\mathbf{X}) = 1\}$.

Definition 2.3.5 (Expected Type I Error Inflation)**.** *The expected inflation in Type I error for using the adaptive procedure:*

$$\Delta_{\mathit{size}}(F) \equiv \alpha_{\mathit{adapt}}(F) - \alpha \tag{2.2}$$

where α is the chosen nominal level, commonly taken to be 0.05.

Definition 2.3.6 (Power Function of Adaptive Procedure)**.** *Let F_1 be a distribution under the alternative and $\vartheta \in H_1$. Define the power of test j as $\pi_j(F_1, \vartheta)$. Then, the power function of the adaptive procedure is:*

$$\pi_{\mathit{adapt}}(F, \vartheta) = \mathbb{P}_{F_1}(A^c)\pi_1(F, \vartheta \mid A^c) + \mathbb{P}_{F_1}(A)\pi_2(F, \vartheta \mid A) \tag{2.3}$$

Proposition 2.3.7 (Expected Power Loss)**.** *The expected power loss relative to always using T_1 is:*

$$\Delta_\pi(F_1, \vartheta) = \pi_{\mathit{adapt}}(F_1, \vartheta) - \pi_1(F_1, \vartheta) = \mathbb{P}_{F_1}(A)\left[\pi_2(F_1, \vartheta, |, A) - \pi_1(F_1, \vartheta, |, A)\right] \tag{2.4}$$

Proof.

$$\begin{aligned}
\Delta_\pi(F_1, \vartheta) &= \pi_{\text{adapt}}(F_1, \vartheta) - \pi_1(F_1, \vartheta) \\
&= \cancel{\mathbb{P}_{F_1}(A^c)\pi_1(F, \vartheta \mid A^c)} + \mathbb{P}_{F_1}(A)\pi_2(F, \vartheta \mid A) - \cancel{\mathbb{P}_{F_1}(A^c)\pi_1(F, \vartheta \mid A^c)} - \mathbb{P}_{F_1}(A)\pi_1(F, \vartheta \mid A) \\
&= \mathbb{P}_{F_1}(A)\pi_2(F, \vartheta \mid A) - \mathbb{P}_{F_1}(A)\pi_1(F, \vartheta \mid A)
\end{aligned}$$

□

The expected power loss/Type-I error inflation provides a measure that can be used to measure the utility of N_T for a given sample size (n). To quantify the overall loss of power or the overall inflation of the Type I error rate across different sample sizes, we calculate the area under the curve for each procedure.

Definition 2.3.8 (Integrated Type I Error and Power)**.** *For sample sizes $n_1 < n_2 < \cdots < n_k$, the integrated type I error(IER) and integrated power(IPR) across sample sizes is defined as:*

$$\mathit{IER} = \frac{1}{n_k - n_1}\int_{n_1}^{n_k} \alpha(F_0, n)dn \quad \mathit{and} \quad \mathit{IPR} = \frac{1}{n_k - n_1}\int_{n_1}^{n_k} \pi_{(}F_1, n)dn \tag{2.5}$$

Where $\alpha(F_0, n)$ and $\pi(F_1, n)$ denote type I error and power functions respectively. $\frac{1}{n_k - n_1}$ is normalization factor to scale results to within 1.

Lemma 2.3.9 (Numerical Approximation via Trapezoidal Rule)**.** *Let $p(n)$ be a continuous performance metric over sample size $n \in [n_1, n_k]$. Then the area under the curve (AUC) can be approximated using the trapezoid rule as:*

$$\widehat{AUC}(p) = \frac{1}{n_k - n_1} \int_{n_1}^{n_k} p(n), dn \approx \frac{1}{n_k - n_1} \sum_{i=2}^{k} \frac{p(n_i) + p(n_{i-1})}{2} (n_i - n_{i-1}) \tag{2.6}$$

Thus, IER, and IPR can be approximated by Equation 2.6 above.

The two metrics, expected Type I error inflation and expected power loss, assess the two fundamental risks of the adaptive procedure. The expected Type I error inflation measures the risk of increases false positives. Power loss measures the cost in efficiency–how much power is sacrificed by making the wrong test choice. A good normality test minimizes both, but they often involve a trade-off. A test that is too conservative controls false positives but may cause large power losses, while a sensitive test may maintain good power but inflate Type I error. The utility of a pre-test is therefore a balance between these two competing inferential costs.

2.3.2 Monte Carlo Simulation Design

We use Monte Carlo simulation to estimate the utility of a normality pre-test for a specific use case. This requires the user to specify some elements including the sample size n, effect size, a plausible non-normal distribution F_1 they wish to guard against, etc. The procedure then quantifies the trade-offs of using an adaptive test that checks normality before choosing between a parametric test (T_1) and a robust test (T_2).

Since the adaptive test itself can be chosen in more than one possible ways based on the choice of the normality test (N_T) and the level of the normlaity test (α_{pre}, the first two steps of the framework explores these parameters to find the 'best-case-scenario' choice of the adaptive procedure.

For a given n and F_1, the simulation follows these steps:

1. **Evaluate Normality Tests:** We first identify the most effective normality test for the given F_1. Multiple normality tests (e.g., Shapiro-Wilk, Anderson-Darling) are compared by simulating data from both the normal distribution and F_1. The test with the highest Area Under the Receiver Operating Characteristic Curve (AUROC) is selected as the pre-test N_T with true positive rate (TPR) and false positive rate (FPR) defined as:

$$\begin{aligned} \text{TPR} &= P(\text{Reject normality} \mid \text{Non-normal sample}), \\ \text{FPR} &= P(\text{Reject normality} \mid \text{Normal sample}). \end{aligned}$$

2. **Optimize the Pre-test Significance Level:**

   (a) Estimate Key Metrics: Using the selected N_T, we run a large number of simulations M for both normal data and data from F_1. For each simulated dataset, we run T_1, T_2, and the adaptive test. This allows us to estimate the core metrics:

   - **Expected Power Loss/Gain:** The change in power from using the adaptive test versus always using T_1.
   - **Expected Type I Error Inflation:** The deviation of the adaptive test's Type I error rate from the nominal level α.

   (b) Choosing Pre-test Significance Level: We repeat Step 2 for a range of pre-test significance levels (α_{pre}). This identifies the α_{pre} that best balances the trade-off between power loss and Type I error inflation (see demonstration in Section 2.6 for further details). Note that the choice of α_{pre} does not need to be standard values such as 0.05 since the normality testing is only useful to classify the data generation mechanism into 'normal' and 'non-normal'. It is merely a threshold for this classifier. Users are offered an opportunity to decide a tolerance(tol) levels to allow a given amount of expected Type I error inflation to add a bid more gains or decrease in losses in power.

3. **Evaluate Downstream Test Performance:**

   (a) Performance Across Effect Sizes: For the optimal α_{pre} identified in Step 2, we evaluate the power of T_1, T_2, and the adaptive test across a range of effect sizes at the fixed sample size n. This reveals how the choice of testing strategy affects power across different effect sizes.

   (b) Downstream ROC Analysis: We construct ROC curves for the downstream testing procedures by varying the downstream significance level α of T_1, T_2, and the adaptive test. This analysis identifies the optimal α level that maximizes power while controlling Type I error for each testing strategy.

4. **Summarize Performance Across Sample Sizes:** Finally, we repeat the entire process for a range of sample sizes. We calculate the Area Under the Power Curve (AUPC) and the Area Under the Type I Error Curve (AUTIEC) to provide a single-number summary of the adaptive test's overall efficiency and validity across the considered sample sizes.

This workflow provides a complete empirical assessment of whether a normality pre-test is beneficial for a given inferential scenario and, if so, how to implement it optimally.

2.4 Examples of application areas

We applied these criteria to evaluate one-sample, two-sample, ANOVA, and regression tests, implemented in parametric, nonparametric, and adaptive hybrid forms based on normality pre-test outcomes.

2.4.1 One-Sample Location Test Procedures

For one sample location test, consider a random sample $X_1, \ldots, X_n \sim F$, where F denotes the unknown distribution. The primary inferential target is a location parameter $\theta(F)$, which may be the mean or median depending on the test procedure. The goal is to test a hypothesis of the form:

$$H_0 : \theta(F) = \theta_0 \qquad \text{vs.} \qquad H_1 : \theta(F) \neq \theta_0.$$

However, it is crucial to ensure that the inferential target (θ) matches the procedure and data-generating mechanism. For example the t-test targets the mean, the sign test targets the median, and the Wilcoxon signed rank test targets a symmetric location parameter. In the presence of skewness or contamination, care is required to interpret comparisons correctly, as mean and median may diverge, misaligned test-statistics may yield misleading results (Lehmann and Romano, 2005; Hollander et al., 2013) and the comparison may not be meaningful. A rigorous evaluation of one-sample location testing thus involves explicit definition of null/alternative hypotheses, calibration, and attention to underlying assumptions. We will briefly discuss the various test options available for test 1 and test 2 in the study.

Test 1: Parametric test. In our situation, an appropriate choice for the test 1 is the one-sample t-test which is valid under normality assumption. Assume F is normal with mean $\mu < \infty$ and variance σ^2. The standard t-test targets $\theta(F) = \mu$, using the statistic:

$$T = \frac{\overline{X} - \theta_0}{s/\sqrt{n}},$$

where $\overline{X}$ is the sample mean and s is the sample standard deviation. Under the null and normality, $T \sim t_{n-1}$. This procedure is optimal and exact under normality, however, under significant deviations such as skewness, and heavy tails, finite-sample, type I error rate and power may be distorted (Lehmann and Romano, 2005).

Test 2: Nonparametric/Distribution-Free Procedures. For test 2, possible choices such as the sign test, bootstrap, permutation, etc. are available, each valid depending on some further assumptions. We briefly discuss some of these tests.

- *Sign test*: Tests $H_0 : \text{med}(F) = \theta_0$ under continuity and is valid for all continuous F. The statistic, $S = \sum_{i=1}^{n} \mathbf{1}\{X_i > \theta_0\}$ follows a binomial distribution (Hollander et al., 2013).
- *Wilcoxon signed-rank test*: This test evaluates whether the data are symmetric around a hypothesized center θ_0. It considers both the sign and magnitude of the differences $D_i = X_i - \theta_0$ by ranking their absolute values and summing the ranks from the positive differences. This procedure requires the assumption of a symmetric probability distribution generating the data (Wilcoxon, 1945; Hollander et al., 2013).
- *Approximate One-Sample Permutation t-test*: Exchanges signs of centered data under the null hypothesis, recalculating t-statistics to build an exact finite-sample reference distribution under the assumption of symmetry (Hollander et al., 2013). See Algorithm in Appendix 3.
- *One-Sample Bootstrap Location Test*: This method evaluates $H_0 : \theta = \theta_0$ without assuming a specific parametric form. The sample is first centered at the null value θ_0 to satisfy the hypothesis. The sampling distribution of a location statistic T (mean, median, or trimmed mean) is approximated through bootstrap resampling from this centered sample, yielding a reference distribution for assessing the observed statistic(Efron and Tibshirani, 1994).See Algorithm in Appendix 1.
- *Box-Cox Transformed t-Test:* Applies a Box-Cox transformation to the data to approximate normality, followed by a standard t-test. This can be particularly effective when data are positively skewed or exhibit heteroscedasticity (Box and Cox, 1964). See discussion on Box-Cox transformation in Section 2.5 and Appendix 6

Adaptive Choice. For any given choices of test 1 and test 2, one may use an adaptive test based on a chosen normality test following Definition 2.3.2. Note that special care must be taken to ensure that test 1, test 2 and the adaptive test are all testing the same hypothesis. See Section 2.6 for further details.

2.4.2 Two-Sample Location Test Procedures

Consider two independent random samples $X_1, \ldots, X_{n_X} \overset{iid}{\sim} F$ and $Y_1, \ldots, Y_{n_Y} \overset{iid}{\sim} G$. The goal is to test whether the respective location parameters $\theta_X = \theta(F)$ and $\theta_Y = \theta(G)$ are equal:

$$H_0 : \theta_X = \theta_Y \qquad \text{vs.} \qquad H_1 : \theta_X \neq \theta_Y.$$

Similarly, different test procedures target different parameters and that must be handled carefully when making comparisons among tests.

Test 1: Parametric (Two-Sample t-Test, Welch Variant). Assume F and G are normal and potentially have unequal variances. The Welch two-sample t-test targets $\theta(F) = \mu_F$, $\theta(G) = \mu_G$, using the statistic (Welch, 1947):

$$T = \frac{\overline{X} - \overline{Y} - (\mu_F + \mu_G)}{\sqrt{s_X^2/n_X + s_Y^2/n_Y}} \sim t_{df}, \quad \text{where} \quad df = \frac{(s_x^2/n_X + s_Y^2/n_Y)^2}{\frac{(s_X^2/n_X)^2}{n_X-1} + \frac{(s_Y^2/n_Y)^2}{n_Y-1}} \quad \text{under } H_0$$

where $\overline{X}$ and $\overline{Y}$ are sample means, s_X^2 and s_Y^2 sample variances. This test is exact and most powerful unbiased under normality, but sensitive to deviations such as skewness or heavy tails, where type I error rates and power may be affected (Lehmann and Romano, 2005).

Test 2: Nonparametric/Distribution-Free Procedures. The following are candidate choices for test 2.

- *Wilcoxon rank-sum (Mann–Whitney U) test*: The Mann–Whitney U test, tests the null hypothesis of equal distributions and is sensitive to shift alternatives. It relies on ranks of pooled data, assessing whether one sample tends to produce larger values than the other. This test is valid for independent samples from continuous distributions and is distribution-free under the null (Mann and Whitney, 1947; Hollander et al., 2013).

- *Permutation Test*: Under the null hypothesis of equal distributions, all observed values are exchangeable between groups. The observed difference in means or other statistics is compared to the permutation distribution obtained by repeatedly reallocating labels between groups (Efron and Tibshirani, 1994).

- *Two-Sample Bootstrap Test for Location*: For testing $H_0 : \theta_1 = \theta_2$, the two samples are recentered by removing their respective sample means (or other location estimators) to align them under the null hypothesis. Bootstrap resamples are then repeatedly drawn from the combined, centered data to estimate the null distribution of the difference in location statistics, providing an empirical basis for inference (Efron and Tibshirani, 1994). See Algorithm in Appendix 2.

Adaptive Choice. Will be chosen following Definition 2.3.2.

2.4.3 One-Way ANOVA Test Procedures

Consider k independent groups with observations $X_{ij} \sim F_i$, $j = 1, \ldots, n_i$, $i = 1, \ldots, k$. The objective is to test the equality of group means:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{versus} \quad H_1 : \mu_i \neq \mu_j \text{ for some } i \neq j.$$

Test 1: Parametric One-Way ANOVA. The standard parametric One-Way ANOVA model assumes (Lehmann and Romano, 2005):

$$F = \frac{\frac{1}{k-1} \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})^2}{\frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2} \sim F_{k-1,N-k} \quad \text{under } H_0$$

Test 2: Nonparametric and Permutation-Based Alternatives. Possible candidate choices for test 2 are the Kruskal–Wallis test and the Permutation ANOVA test.

- *Kruskal–Wallis Test*: This nonparametric procedure extends the Mann–Whitney U/Wilcoxon concept to k independent groups. It assesses whether the groups originate from the same distribution, with the alternative that at least one tends to produce systematically larger or smaller values. The test statistic has an asymptotic χ^2 distribution with $k - 1$ degrees of freedom under the null hypothesis (Kruskal and Wallis, 1952; Hollander et al., 2013).
- *Permutation ANOVA*: This procedure uses the standard ANOVA F-statistic but determines statistical significance through label shuffling rather than theoretical F-distributions. By randomly reassigning group labels many times, it builds an empirical null distribution that requires only exchangeability under the null hypothesis, meaning the data are exchangeable across groups if no true differences exist. This approach provides valid inference without relying on normality assumptions (Good, 2013).

Adaptive Choice. Will be chosen following Definition 2.3.2.

2.4.4 Linear Regression Test Procedures

In simple linear regression with n independent observations, the model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where ε_i are random errors. The main test evaluates

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0.$$

One may also consider a multiple linear regression model with more than one predictors.

Test 1: Parametric t-Test for Regression Slope. This standard test uses ordinary least squares (OLS) assuming normally distributed errors, $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$

The statistic

$$t = \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)} \sim t_{n-2}$$

under the null (Seber and Lee, 2003).

Test 2: Robust Alternatives. In cases where normality or homoscedasticity are questionable:

- *Permutation Tests*: Methods like Freedman-Lane create an empirical null by shuffling responses or residuals, requiring only exchangeability (Freedman and Freedman, 1983).
- *Bootstrap Tests*: Resampling with replacement approximates the sampling distribution of regression coefficients, enabling inference without parametric assumptions (Efron and Tibshirani, 1994).

Adaptive Choice. Will be chosen following Definition 2.3.2.

2.4.5 Review of Asymptotic Relative Efficiency (ARE) Framework

The choice between parametric and nonparametric methods is a fundamental trade-off between power and robustness. Under regularity conditions, such as normality, parametric tests are believed to be most powerful, however when model assumptions are violated they can be less efficient compared to their robust nonparametric counterparts.

The central question becomes, how do we quantify this trade-off? The theoretical answer, provided by the works of Pitman, Chernoff, Bahadur, and Hoeffding (Serfling, 2009; Lehmann, 1999) is generally the Asymptotic Relative Efficiency (ARE). The ARE provides an elegant, large-sample framework for comparing two statistical tests but are limited to large sample. It is not clear if the same can be applied in finite sample case.

To put this into perspective we refer to Lehmann (1999) and Serfling (2009) to explore these existing theoretical frameworks and highlight their limitations in our application case.

Pitman's Framework. Pitman introduced the concept of *efficacy* to measure the sensitivity of a test under local alternatives. His framework assumes contiguous alternatives of the form $\theta_n = \theta_0 + \delta/\sqrt{n}$ and defines efficacy as the derivative of the test statistic's mean with respect to δ at $\delta = 0$ (Lehmann, 1999). This leads directly to the ARE formula:

$$e_{2,1} = \left(\frac{c_2}{c_1}\right)^2,$$

where c_i is the efficacy of test i.

Definition 2.4.1 (Efficacy)**.** *Efficacy, c measures the rate at which a test statistic's mean shifts under a local alternative $(\theta_n = \theta_0 + \delta/\sqrt{n})$(Lehmann, 1999). A test with higher efficacy can detect smaller deviations with the same sample size. That is higher efficacy is associated to higher power.*

Definition 2.4.2 (Asymptotic Relative Efficiency (ARE))**.** *The ARE of test 1 with respect to test 2, $e_{2,1} = (c_2/c_1)^2$ is the square of the efficacy ratio which is equivalent to the limiting ratio of sample sizes N_1/N_2 required to achieve identical power $e_{2,1} = \lim_{k\to\infty} N_1^k/N_2^k$ where N_i^k are sample sizes Lehmann (1999)*

Chernoff's Index. Chernoff developed a theory based on large deviations and exponential decay rates of error probabilities. His index quantifies the rate at which the probability of Type II error decays exponentially under fixed alternatives. This is particularly useful in sequential testing and provides a more refined efficiency measure than Pitman's efficacy in certain contexts (Serfling, 2009).

Bahadur Efficiency. Bahadur proposed an alternative efficiency measure based on the exact slopes of the test statistic's tail probabilities under the alternative hypothesis. His approach evaluates the rate at which the p-value converges to zero and is especially relevant for nonparametric tests where tail behavior is critical (Serfling, 2009).

Hoeffding's Contributions. Hoeffding laid the groundwork for rank-based and U-statistics, which are central to nonparametric inference. He derived asymptotic distributions and efficiencies for tests like Wilcoxon and Sign, showing their robustness and efficiency under broad conditions. His work supports the use of ARE to compare parametric and nonparametric tests (Serfling, 2009).

For example, suppose $X_1, \ldots, X_n \sim F$, the relative efficiencies of the Sign, Wilcoxon signed rank, and t-tests to each other are:

$$e_{S,t}(F) = 4\sigma^2 f^2(0), \quad e_{W,t}(F) = 12\sigma^2 \left(\int f^2(z)\right)^2, \quad \text{and} \quad e_{S,W}(F) = f^2(0)/3 \left(\int f^2(z)\right)^2$$

respectively (Lehmann, 1999). The above formulas provide a basis for comparing the three tests for different distributions F. For $F \sim N$ we have

$$e_{S,t}(\Phi) = \frac{2}{\pi} \approx 0.637, \quad e_{W,t}(\Phi) = \frac{3}{\pi} \approx 0.955.$$

This theoretical framework provides a beautiful and powerful ranking of tests. Under normality, the hierarchy is that t-test $>$ Wilcoxon $>$ Sign. However, when normality is violated, the relative efficiencies can change drastically.

The ARE formulas depend critically on the true, unknown F. The above assumed that $F = \Phi$. If F is slightly contaminated the actual finite-sample relative efficiency can be very different. One key limitation is that these theoretical frameworks are usually true under some regularity conditions including that we have large sample size, making them not suited for finite sample case which is what is often desired in practice.

While the classical Asymptotic Relative Efficiency provides essential theoretical framework for comparing statistical tests, its limitations in the finite-sample case render it an incomplete guide for practical data analysis.

Our dissertation develops a comprehensive finite-sample comparison framework that integrates theoretical asymptotics with empirical simulation and robust metrics, providing a more reliable and actionable tool for method selection.

2.5 Data Transformation

Data transformation is a common practice in statistical applications often used to stabilize data to meet certain required conditions, and normality is one of those important conditions. A handful of data transformation methods exist and one of the most popular one is the Box-Cox transformation method Box and Cox (1964).

2.5.1 Box-Cox Transformation

The Box-Cox transformation (Box and Cox, 1964) is a widely used method for addressing non-normality and heteroscedasticity in data by applying a power transformation to strictly positive values. It is defined as:

$$g_\lambda(x) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log x, & \lambda = 0, \end{cases} \tag{2.7}$$

where $x > 0$ and the parameter λ is estimated using a profile likelihood approach. The transformation is monotone increasing, preserving the order and the median, i.e., $\mathrm{med}\{g_\lambda(X)\} = g_\lambda(\mathrm{med}(X))$, though it does not generally preserve means, $\mathbb{E}[g_\lambda(X)] \neq g_\lambda(\mathbb{E}[X])$.

One limitation of the Box-Cox formulation is its restriction to strictly positive data. To overcome this, alternative transformations allowing for zero and negative values have been proposed.

2.5.2 Yeo-Johnson Transformation

The Yeo-Johnson transformation (Yeo and Johnson, 2000), extends the Box-Cox approach to accommodate zero and negative values without shifting or truncation. It is defined as a piecewise power transformation:

$$g_\lambda(x) = \begin{cases} \dfrac{(x+1)^\lambda - 1}{\lambda}, & x \geq 0, \lambda \neq 0, \\ \log(x+1), & x \geq 0, \lambda = 0, \\ -\dfrac{(-x+1)^{2-\lambda} - 1}{2-\lambda}, & x < 0, \lambda \neq 2, \\ -\log(-x+1), & x < 0, \lambda = 2. \end{cases}$$

This family retains many desirable properties of the Box-Cox family and is particularly useful when data contain negative or zero values (Yeo and Johnson, 2000; Weisberg, 2001).

2.5.3 Applications and Considerations

Both transformations are aimed at achieving approximate normality, variance stabilization, and improving linear model fitting. The selection of λ typically involves maximization of the profile likelihood or other criteria balancing normality and homoscedasticity (Box and Cox, 1964). While the Box-Cox transformation is well established for strictly positive data, the Yeo-Johnson approach broadens applicability, allowing use in a wider range of contexts without data modification.

These transformations play crucial roles in regression analysis, ANOVA, and other parametric frameworks where model assumptions are sensitive to distributional forms.

2.5.4 Potential Challenges in Transformation Applications

Power transformations such as Box-Cox aim to address non-normality and heteroscedasticity but introduce several interpretive and practical challenges.

Transformed parameters often lack direct interpretability. Although the transformation preserves order statistics and medians, the mean on the transformed scale is not equivalent to the mean on the original scale. Back-transformation of means typically results in bias unless explicit correction methods are employed (Carroll and Ruppert, 1984). Therefore, if one wishes to test parameters such as the mean, the only option is to test it in the transformed scale which may not be comparable to methods that do it in the original scale.

For tests based on two or more samples, assuming a common transformation parameter λ across groups can be restrictive. When distributions differ greatly, a single λ may fail to properly normalize one or both groups, either hiding true differences or creating misleading effects. Additionally, the transformation can change variance-covariance patterns, compromising the validity of later parametric tests.

In linear models, transformations that improve normality worsen variance heterogeneity and vice versa (Box and Cox, 1964). A Box-Cox transformation of the response variable is a common practice, but it modifies the scale of the response and changes the nature of the predictor-response relationship, complicating coefficient interpretation. Unlike simple log-transformations, coefficients from Box-Cox transformed models often lack intuitive meaning on the original scale (Bickel and Doksum, 1981). Additionally, inference can be sensitive to the chosen λ, raising concerns about overfitting and data-driven selection.

A Box-Cox transformation or any commonly used ad-hoc transformation (such as log-transformation) may also create complex issues in regard to the selective inference. This will be further explored in Chapter 4.

Overall, while power transformations are valuable tools for improving model assumptions, careful consideration of their limitations and clear reporting of interpretational nuances are critical for valid inference.

2.6 Demonstration of our framework

We start with the usual comparison of some common normality test methods for different distributions across a given set of sample sizes. Though this has been explored in the literature, we revisit it as a base for our analysis to highlight the inconsistent performance of the existing normality test methods in varied scenarios as shown in Figure 2.1 below. Figure 2.1, Shapiro-Wilk, and Shapiro-Francia are generally most powerful.

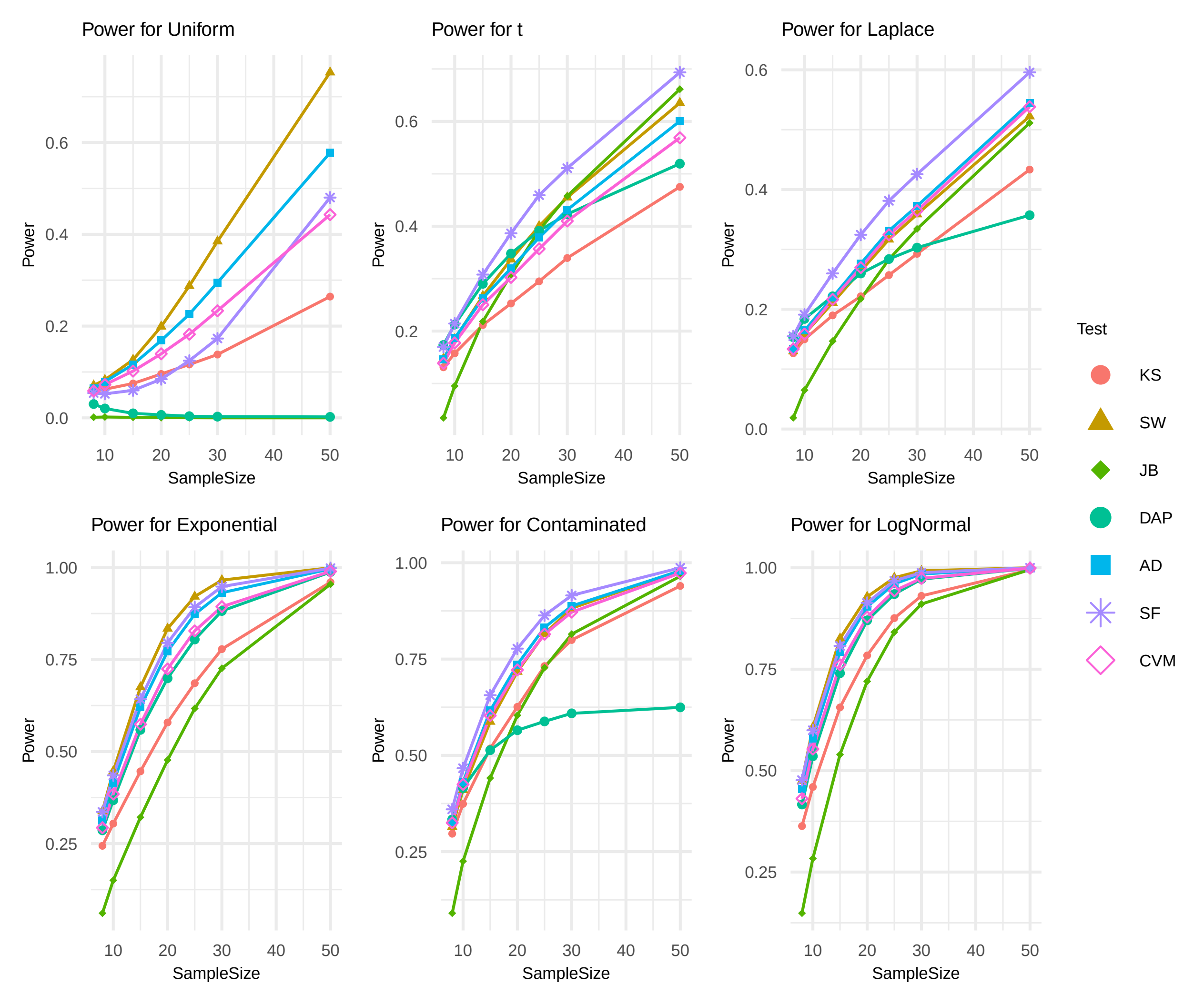


Figure 2.1: *Showing power comparison of different normality test methods for different distributions(Uniform(0,1),* t_3*, Laplace(0,4), Exp(1), Contaminated*$\sim 0.75N(0,1) + 0.25N(0,25)$*, and Lognormal(0,1)) for varied sample sizes.*

2.6.1 One-Sample Location Test - The Setup

To illustrate our framework, we examine examples in the one-sample location test setting, selecting test 1 as the one-sample t-test, and varying test 2 based on its requirements. Depending on the chosen test 2, the target parameter may be the mean or median to ensure comparability (note that under normality, t-test is a test of mean as well as median). For our illustrations, we generate data from an alternative, non-normal distribution and also from a standard normal distribution. In situations where symmetry is a requirement, we choose the alternative distribution appropriately to meet that. In this illustrations, the alternative distribution is the exponential distribution, except for the one-sample permutation case where we use Laplace.

For one-sample test, the actual sample data are required to follow the normal distribution. Following the procedures outlined in Section 2.3.2 above, we constructed the Receiver Operating characteristic (ROC) curves for different normality test methods for a given sample size, $n = 10$ as shown in Figure 2.2. In this particular scenario, all test methods are close in performance with the Shapiro-Wilk(SW) test having a slight edge over the rest. Thus, we chose the SW test in the adaptive test procedure. In this and all subsequent sections all simulation procedure with the exception of where multiple sample sizes like area under curves, which we did for N=10000 iterations were done for N=1000000.

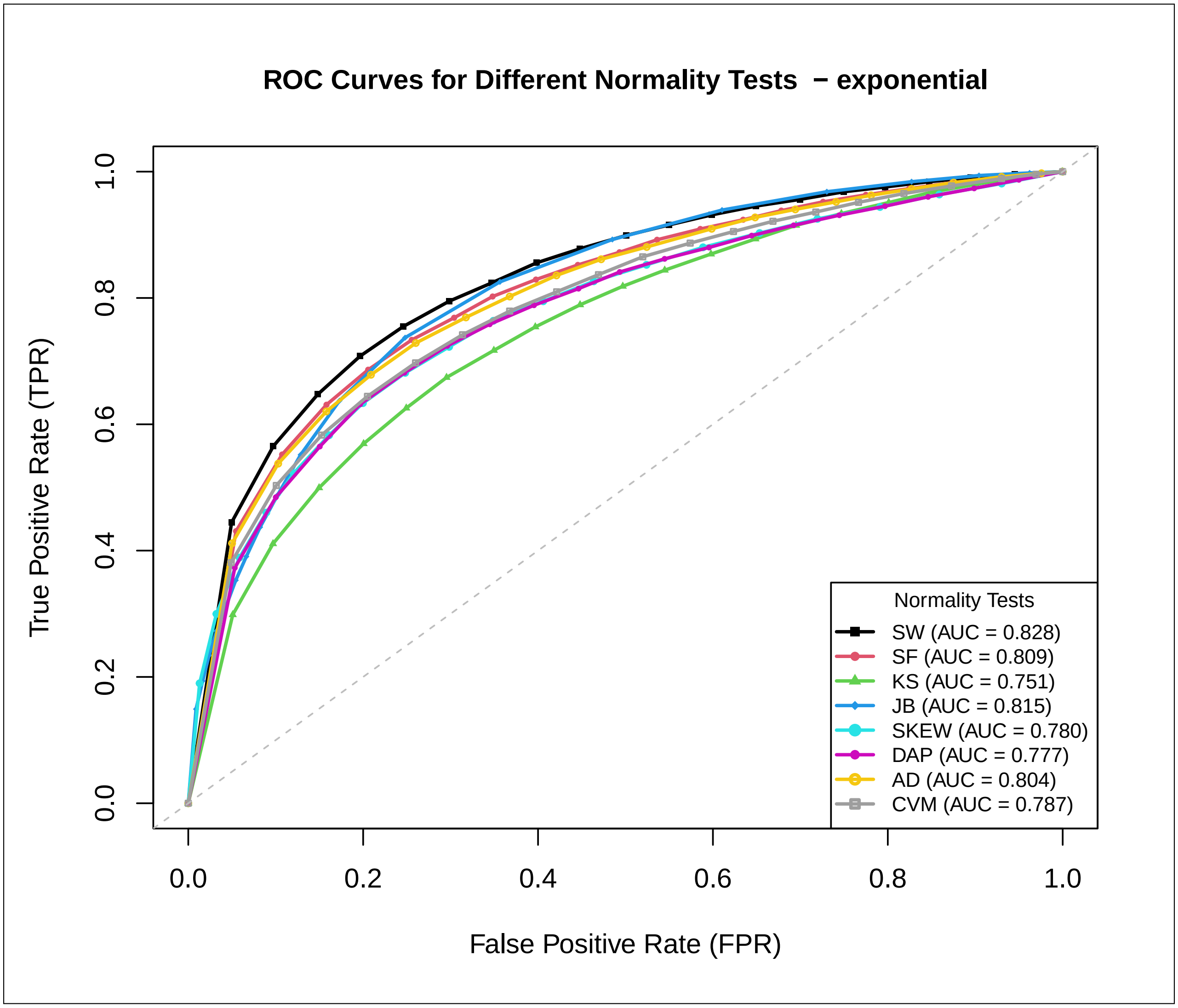


Figure 2.2: *Showing ROC curves for different normality test methods for sample size 10 drawn from normal and exponential distributions.*

2.6.2 Demonstration 1: One-Sample t-Test versus Sign Test

To fairly compare the one-sample t-test and the sign test, we consider testing the same location parameter – median. We test $H_0 : \text{med}(F) = \theta_0$, where both tests are valid if normality holds. We follow the steps outlined in Section 2.3.2 in the demonstration below.

Guided by the results in Figure 2.2, we selected the Shapiro-Wilk test (N_T) for the normality assessment. We quantify pre-testing utility through three metrics across α_{pre} levels: expected power loss (adaptive vs. test 1 under normality), expected power gain (adaptive vs. test 1 under non-normality), and expected Type I error inflation (adaptive error rate minus nominal $\alpha = 0.05$). The above definitions are assumed in the rest of the sections in this work.

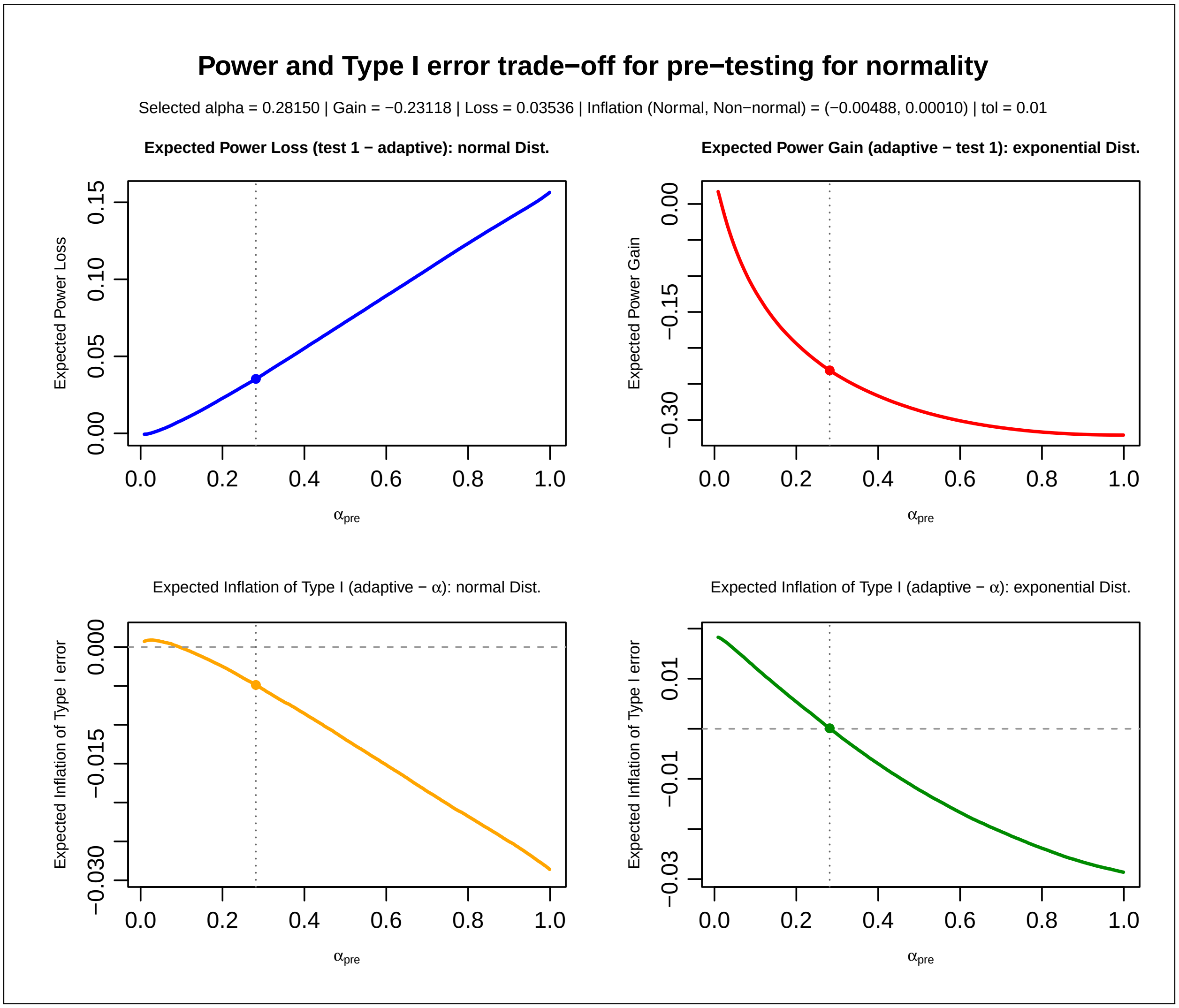


Figure 2.3: *Showing expected power loss and the expected power gain for pre-testing in row 1. Row 2 displays expected inflation of Type I error rates for samples from normal and exponential distribution for sample of sizes 10. The optimal α_{pre}, denoted α_{pre} is about 0.2815 with a tolerance of 0.01.*

The adaptive procedure generally have lower power compared to the standard t-test, since the Sign test is less powerful. However, unlike the parametric test, it achieves near control of Type I error at smaller pre-test alpha levels.

To maintain valid Type I error control (within 0.01 tolerance), the optimal pre-test alpha is approximately 0.2185. Next, for a given sample size $n = 10$, and for any given distribution $\mathcal{D}$, we create a plot of power versus effect size as shown in Figure 2.4. The adaptive test achieves higher power than the Sign test alone while preserving error control, demonstrating the utility of pre-testing in this context (Figure 2.4).

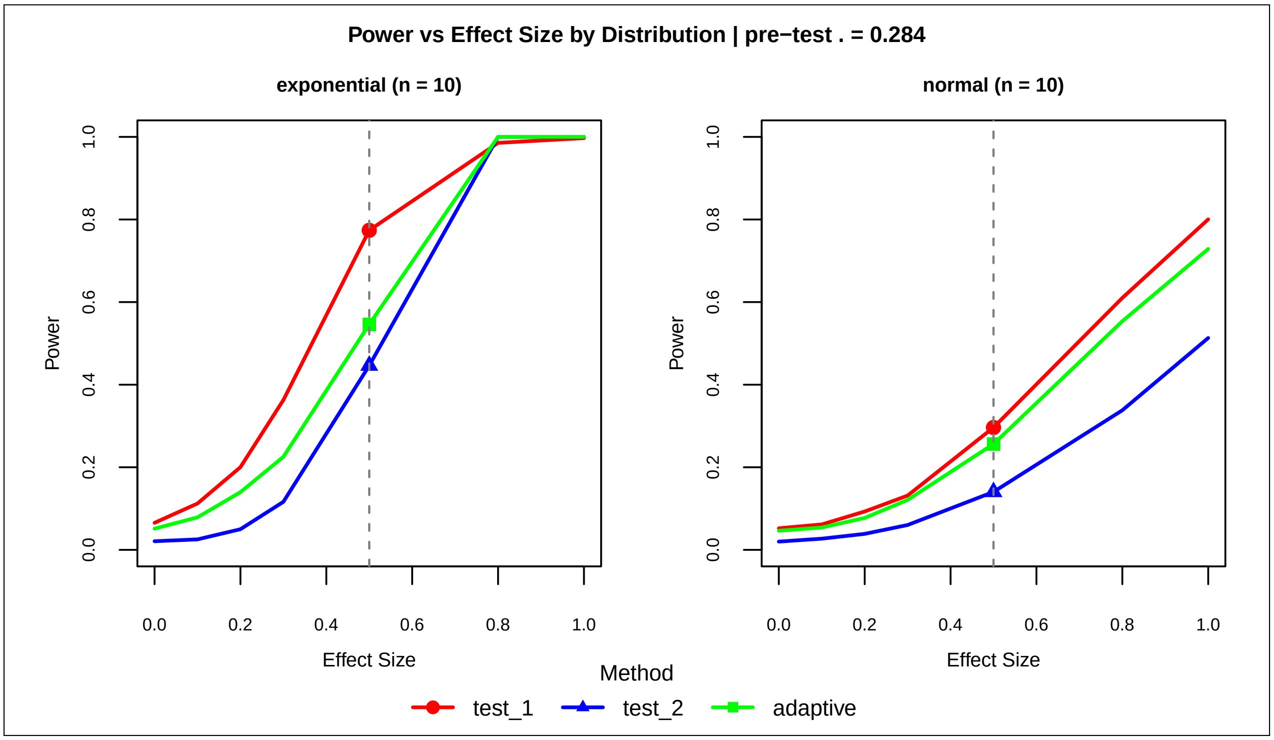


Figure 2.4: *Showing power versus effect size for samples of size 10 drawn from normal and exponential distributions for test 1(t-test), test 2(sign test), and the adaptive test.*

To evaluate the trade-off between power and Type I error for the adaptive procedure, we construct ROC-like curves (Figure 2.5). Using the optimal pre-test threshold α_{pre} identified previously and with the researcher-specified effect size, guided by Figure 2.4 above, we construct ROC like curves for the full range of Type I error values and also zoom-in only to Type I error rates up to 0.1, for better view.

The Sign test, while less powerful than the t-test, provides exact Type I error control. Also note that the adaptive procedure using α_{pre} maintains Type I error at the nominal level while achieving power above the Sign test suggesting some benefits in using the adaptive test.

The jagged ROC curves are a direct result of the Sign Test's discrete Binomial null distribution. Because the test statistic can only take integer values, the actual Type I error rate and power remain constant over ranges of the nominal alpha level, changing only when alpha crosses a threshold that alters the critical values. This step-function behavior is a fundamental property of exact tests based on discrete statistics (Hollander et al., 2013).

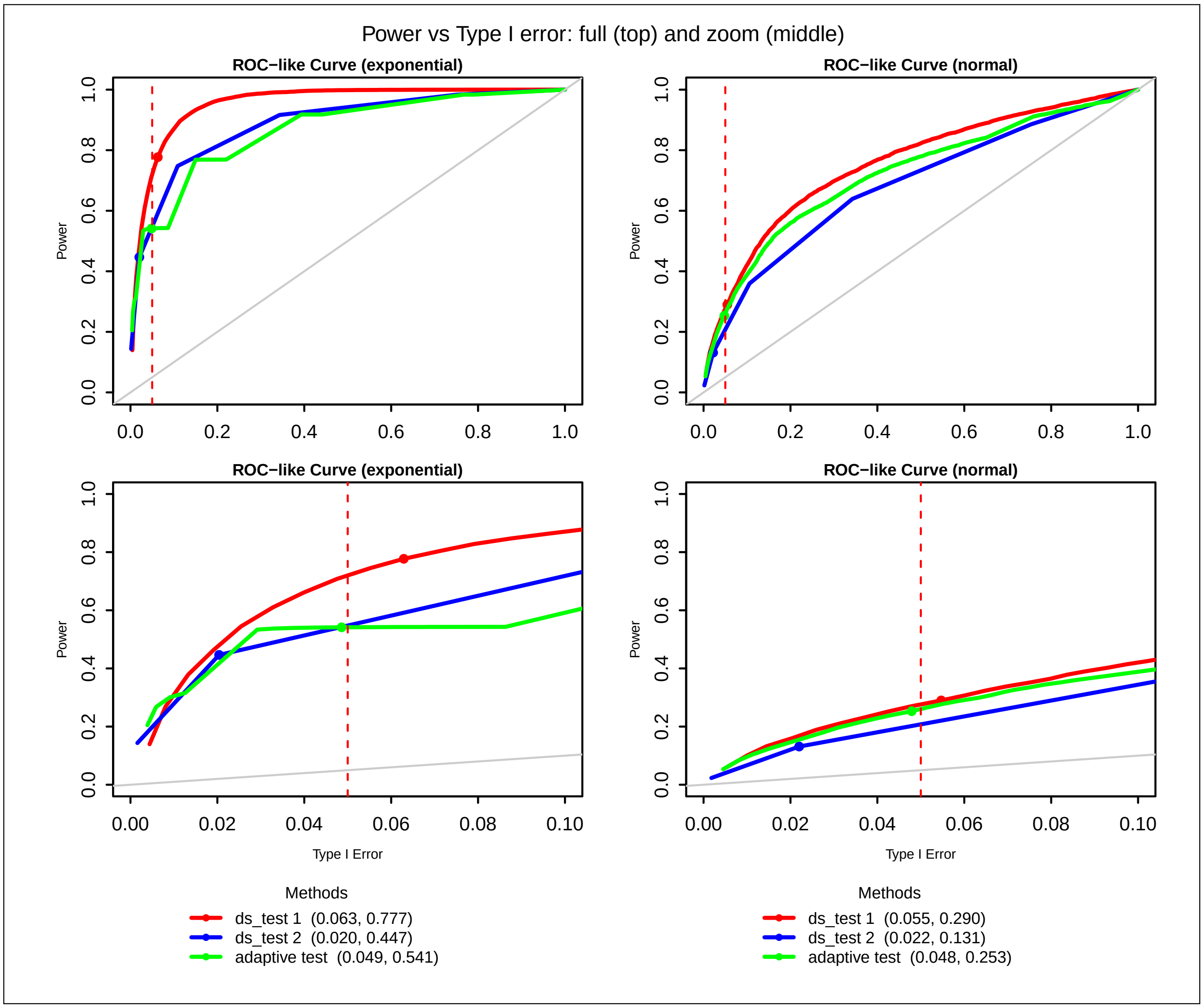


Figure 2.5: *Showing the power vs Type I error plot for t-test, Sign test and the adaptive test based on Shapiro-Wilk test for normality in an ROC-Like curve for a given sample size* $n = 10$. *The optimal alpha for normality pre-testing,* α_{pre} *is* 0.2185 *as obtained from 2.3 above. ds_test 1 and ds_test 2 mean test 1 and test 2 respectively.*

With the optimal alpha values for normality pre-test, researcher's chosen effect size, and downstream significance level chosen to maintain Type I error rates, we evaluate the performance of the three test methods, t-test, Sign test, and adaptive procedure by comparing their power and Type I error rates across different sample sizes using a vector of optimal alpha values, α_{pre} for each sample size . The area under each power curve serves as a summary measure of overall performance (Figure 2.6). In practice, we should obtain different optimum alpha pre-test values for each sample size. This will require applying the techniques in Figure 2.3 for each sample size. As a demonstration, this is done for just one case, the one-sample t-test vs Sign test and skipped the rest due to space and computation time in this write up.

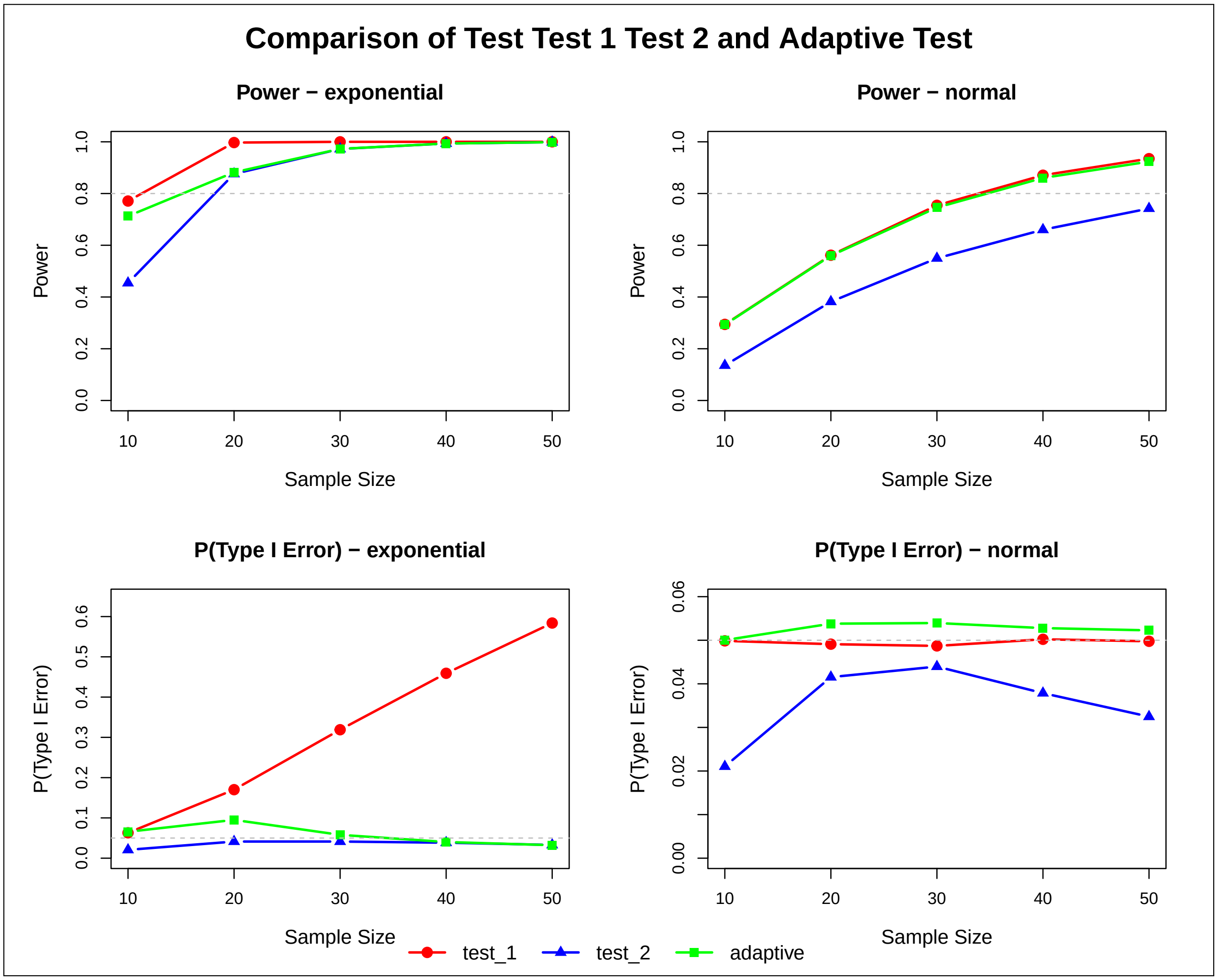


Figure 2.6: *Power and Type I error curves for t-test, Sign test and the adaptive test by pre-testing for samples drawn from exponential and normal distributions.*

Figure 2.6 displays power and Type I error rates for three testing procedures applied to exponential and normal distributions. The one-sample t-test (Test 1) exhibits substantial Type I error inflation under exponential sampling, with inflation increasing with sample size. This occurs because the test targets the population mean while the exponential distribution's median and mean diverge, violating the t-test's underlying assumptions. Both the Sign test and adaptive procedure maintain better error control across sample sizes.

Distribution	Method	AUC-Power	AUC-TypeI
Exponential			
	Parametric (t-test)	0.9704	0.3191
	Nonparametric (Sign test)	0.8933	0.0381
	Adaptive	0.9247	0.0625
Normal			
	Parametric (t-test)	0.7019	0.0507
	Nonparametric (Sign test)	0.5100	0.0393
	Adaptive	0.6928	0.0512

Table 2.1: *AUC for power and Type I error for one sample t, Sign, and adaptive tests by method and distribution.*

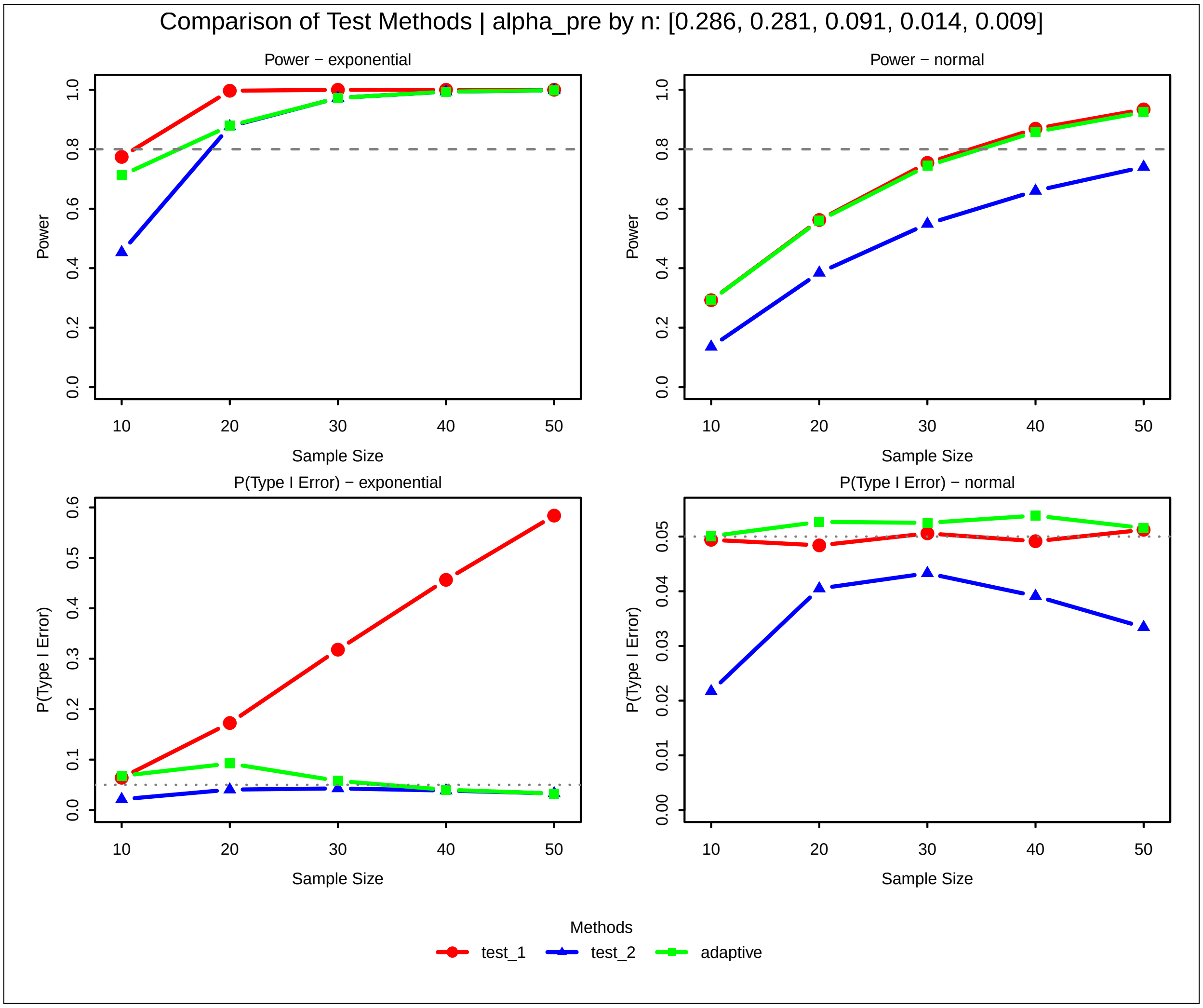


Figure 2.7: *Power and probability of Type I error curves for t-test, Sign test and the adaptive test by pre-testing for samples drawn from exponential and normal distributions. The adaptive test is calculated using the vector of optimal* α_{pre}.

Distribution	**Method**	**AUC-Power**	**AUC-TypeI**
Exponential			
	Parametric (t-test)	0.9710	$0.317\,70$
	Nonparametric (Sign test)	0.8922	$0.037\,28$
	Adaptive	0.9251	$0.060\,09$
Normal			
	Parametric (t-test)	0.6997	$0.049\,61$
	Nonparametric (Sign test)	0.5079	$0.037\,61$
	Adaptive	0.6930	$0.052\,46$

Table 2.2: *AUC for power and Type I error for one sample t, Sign, and adaptive tests by method and distribution. Adaptive procedure is obtained using a vector of* α_{pre} *for each sample size.*

Observations and Recommendations: The analysis reveals nuanced trade-offs between testing procedures. The adaptive approach demonstrates practical value in this specific context (exponential distribution, sample sizes 10-50), offering improved power over the Sign test while maintaining reasonable Type I error control. However, this advantage appears sensitive to the alternative distribution and may not generalize to other non-normal scenarios.

Also, using individual, specific optimal alpha values for each sample size did not show much difference from using a single optimal alpha value as depicted by Figures 2.6 and 2.7. This we believe is due to the fact that as sample size increases, the classical normality test becomes overly sensitive, and thus will always rejects normality irrespective of the value of α_{pre}.

2.6.3 Demonstration 2: One-Sample t-Test vs Bootstrap Test

We now compare the t-test against a bootstrap test targeting the mean. The bootstrap procedure resamples the t-statistic to assess significance. Using the Shapiro-Wilk test (N_T) at various α_{pre} levels, we evaluate the trade-offs of normality pre-testing as shown in Figure 2.8 below.

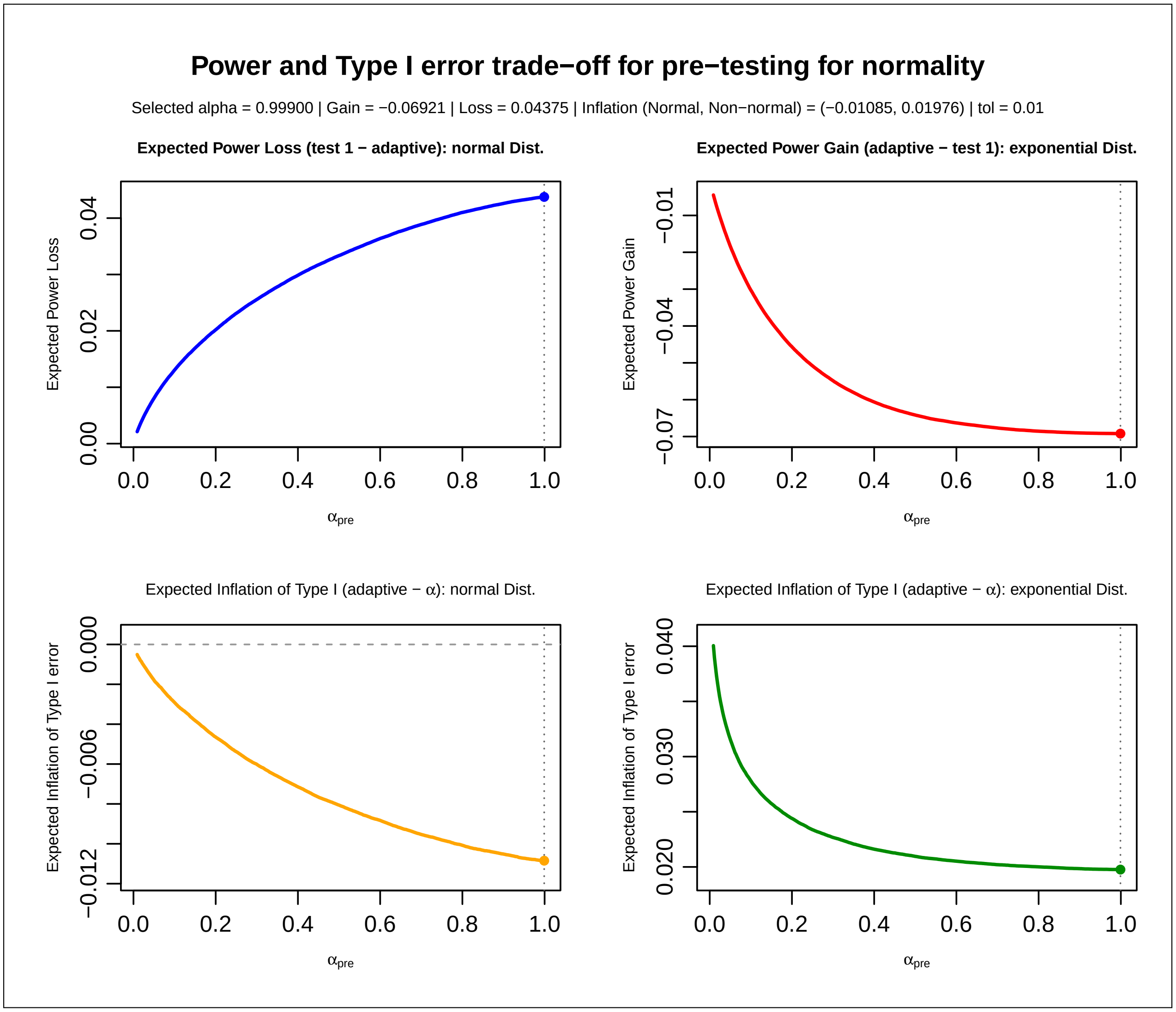


Figure 2.8: *Showing expected gains and losses for pretesting for normality for the t-test, bootstrap, and the adaptive test for samples sample of sizes 10 from normal and exponential distribution. The optimal α_{pre} is 0.999*

In this particular situation, there seems to be no benefits for using the adaptive procedure because we fail to control Type I error rate for non-normal samples case and lose power in both normal and non-normal sample cases as shown in Figure 2.8 above.

Next, for a given sample size $n = 10$, and for any given distribution $\mathcal{D}$, we create a plot of power versus effect size as shown in Figure 2.9. Again, the power of the adaptive procedure is similar to that is test 2 (bootstrap test) due to the extremely optimum alpha value (0.999) selected.

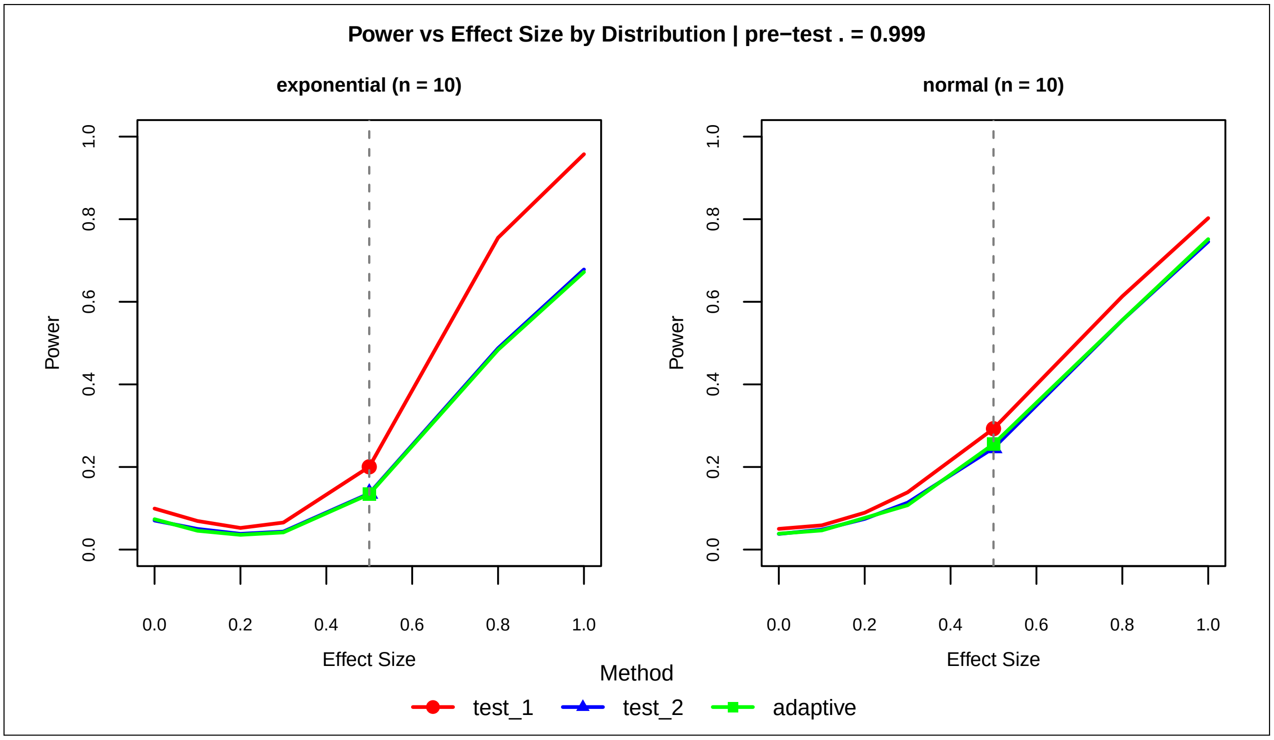


Figure 2.9: *Showing power versus effect size for samples of size 10 drawn from normal and exponential distributions for the t, bootstrap, and adaptive tests.*

To evaluate the trade-off between power and Type I error for the adaptive procedure, we construct ROC-like curves (Figure 2.10). Using the optimal pre-test threshold α_{pre} identified previously and with the researcher-specified effect size, guided by Figure 2.9 above, we construct ROC like curves for the full range Type I error values and also zoom-in only to Type I error rates up to 0.1, for better view.

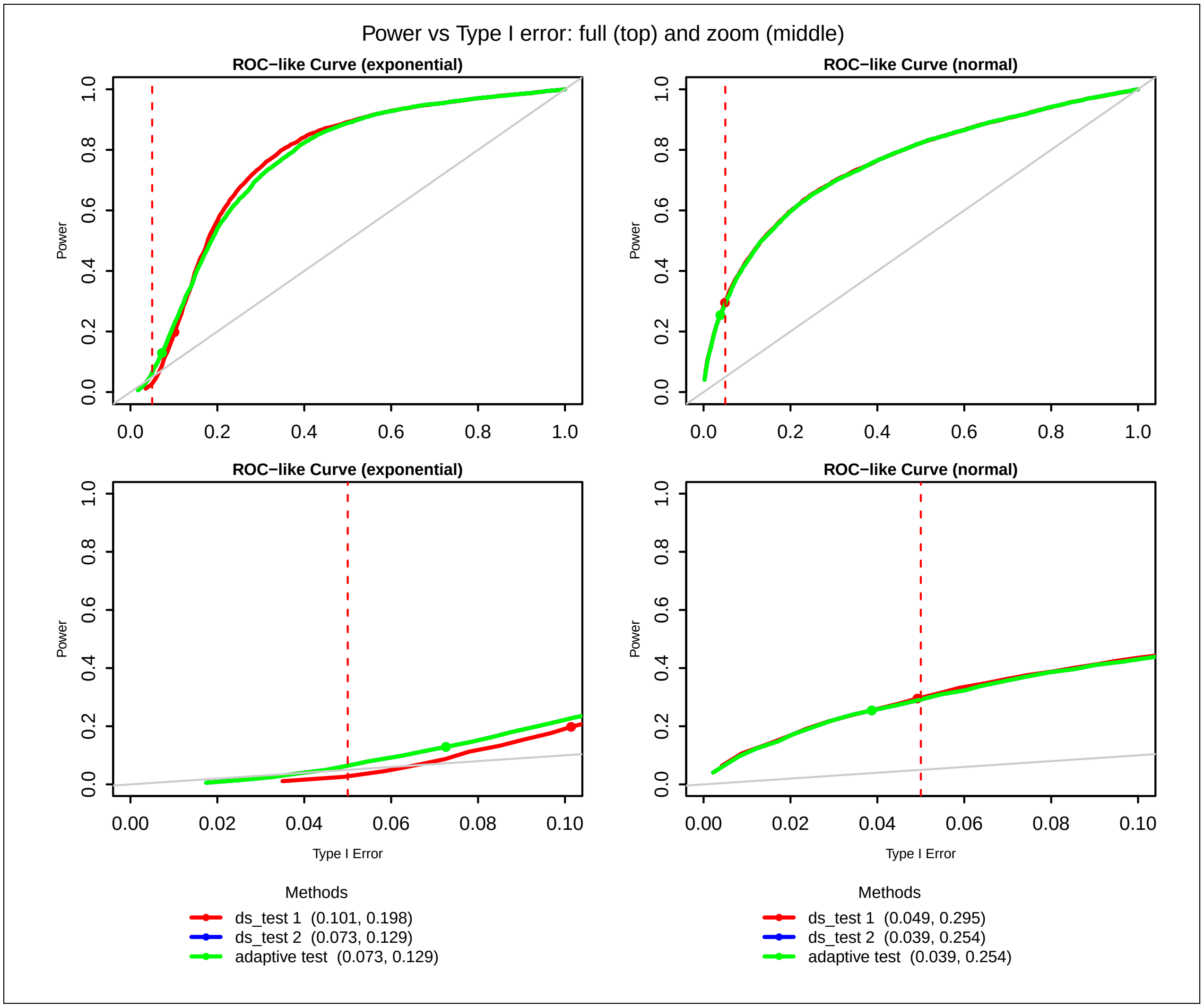


Figure 2.10: *Showing the power vs Type I error plot for t-test, bootstrap test and the adaptive test based on the select* N_T, *in an ROC-Like curve for a given sample size* $n = 10$. *ds_test 1 and ds_test 2 mean test 1 and test 2 respectively.*

We next evaluate the performance of the *t-test*, *bootstrap test* and the *adaptive procedure* by comparing their power and Type I error rates across different sample sizes using a vector of optimal alpha values, α_{pre} for each sample size . The area under each power curve serves as a summary measure of overall performance (Table 2.4).

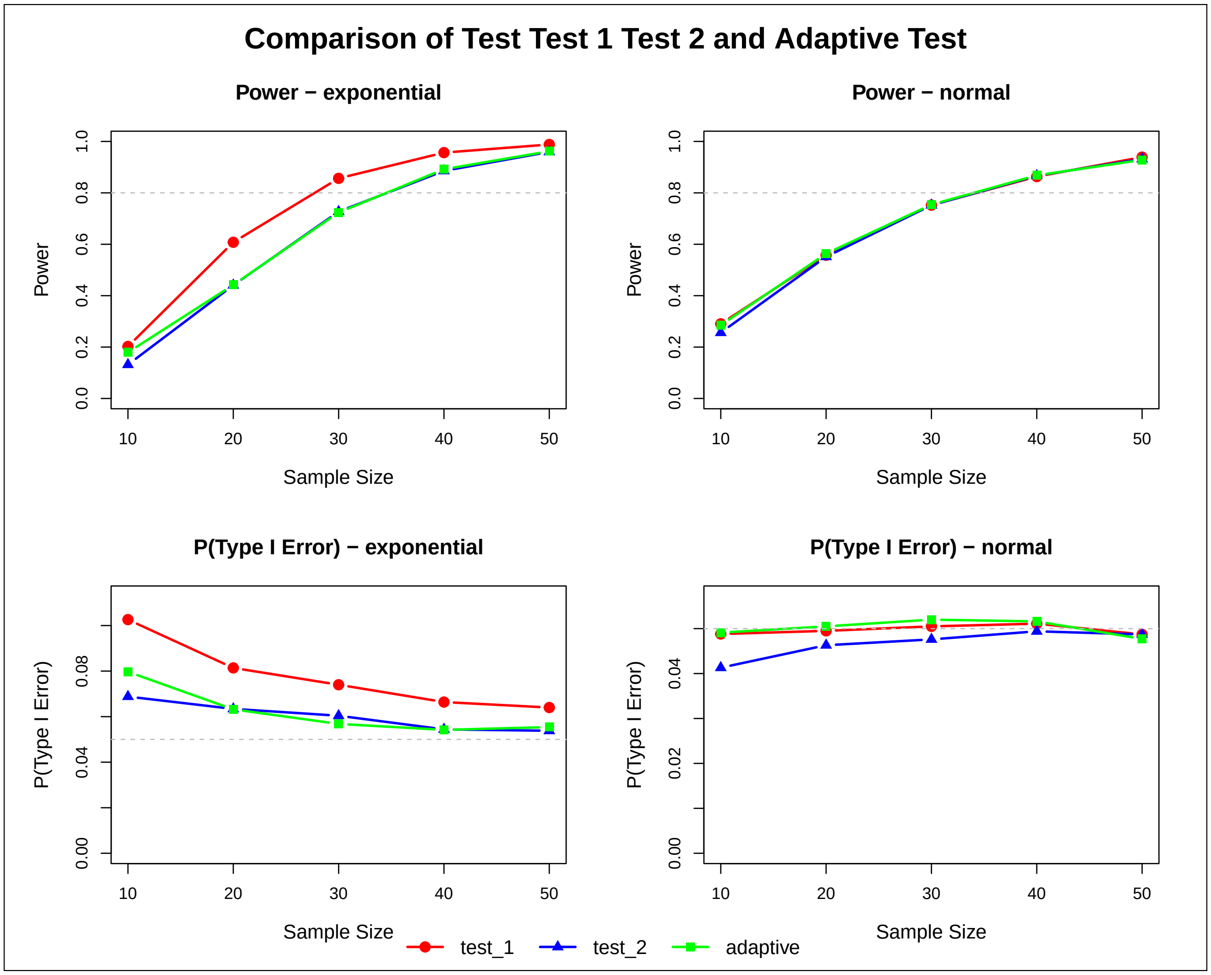


Figure 2.11: *Showing the area under the power curves and area under the Type I error curves for t-test, bootstrap test and the adaptive test by pre-testing for samples drawn from exponential and normal distributions.*

Distribution	Method	AUC-Power	AUC-TypeI
Exponential			
	Parametric (t-test)	0.753 95	0.075 71
	Nonparametric (bootstrap)	0.646 24	0.058 70
	Adaptive	0.655 55	0.060 47
Normal			
	Parametric (t-test)	0.6998	0.050 79
	Nonparametric (bootstrap)	0.6892	0.046 24
	Adaptive	0.6991	0.048 43

Table 2.3: *AUC for power and Type I error by method and distribution for t, bootstrap, and adaptive tests.*

Conclusion and Recommendations Based on our analysis, the adaptive procedure provides no practical advantage, as it fails to improve power while compromising Type I error control compared to using either test individually. We therefore do not recommend normality pre-testing for this application. When distributional uncertainty exists, the bootstrap method offers a preferable alternative. It maintains robustness without introducing the additional complexity and risk associated with pre-testing strategies. These conclusions are specific to the distributional settings and sample sizes examined in our study.

2.6.4 Demonstration 3: One-sample t-test vs permutation test:

We compare the one-sample t-test (Test 1) against a permutation test (Test 2) that uses the t-statistic for mean comparison. This requires symmetric alternative distribution, we chose Laplace distribution.

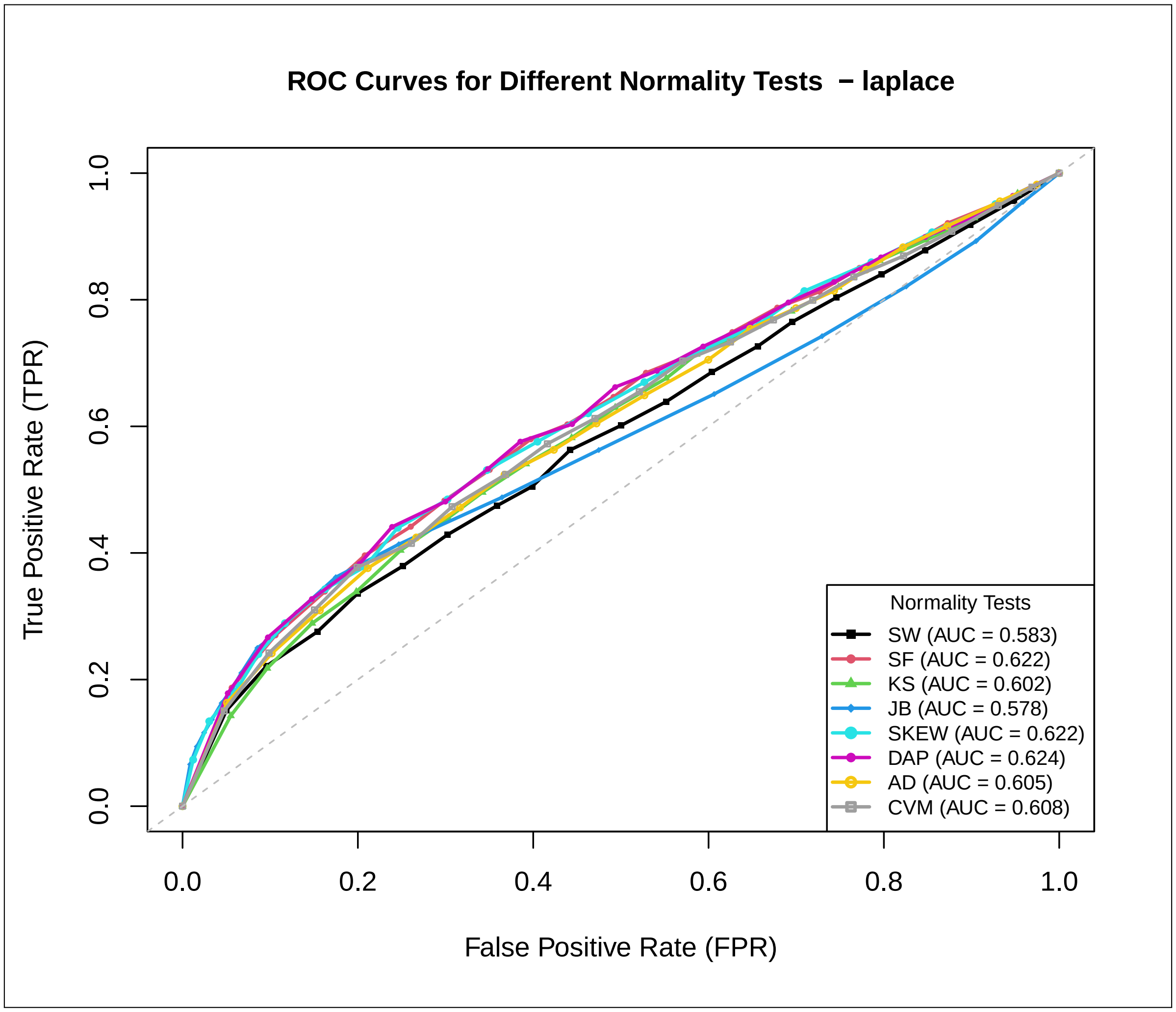


Figure 2.12: *Showing ROC curves for different normality test methods for sample size 10 drawn from normal and Laplace distributions.*

The weakness of the existing normality test methods in detecting departures from symmetric distributions is glaring here when we chose Laplace as our alternative distribution. We also see a sharp switch in Shapiro-Wilk test to nearly the worst performing test, compared to the others. This highlights the unpredictable nature in their performance for different alternative distributions.

we then assess the expected gains and losses for pre-testing for normality at different normality alpha value, α_{pre} with the chosen normality test method, N_T as Shapiro-Francia using the expected power loss, expected power gain, and the expected type I error inflation as shown in Figure 2.13 below.

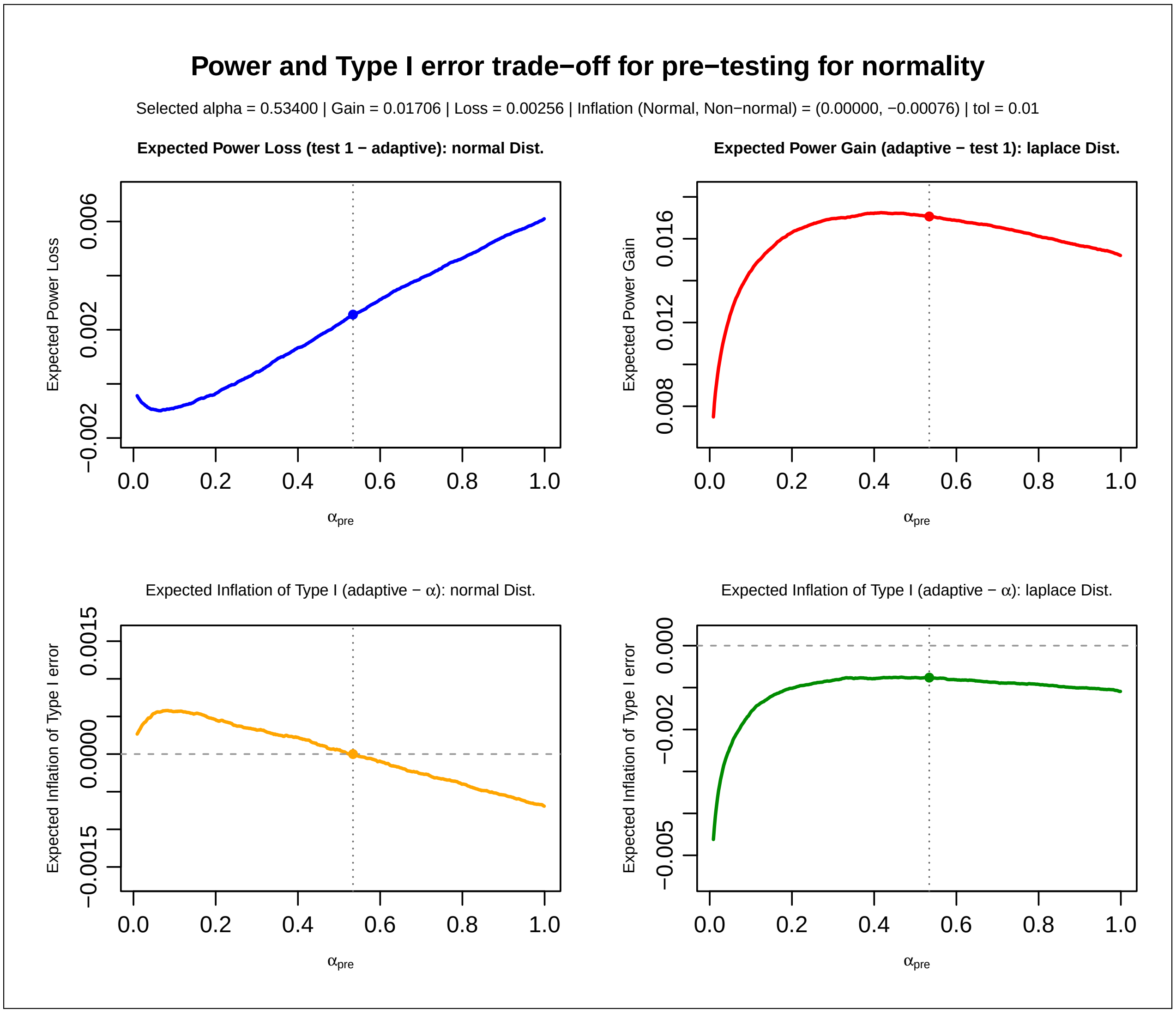


Figure 2.13: *Showing expected power loss for pretesting, left plot and the expected power gain for pre-testing, right plot in row 1. Row 2 displays expected inflation of Type I error rates for samples from normal and Laplace distribution for sample of sizes 10.*

Using Figure 2.13 above, to maintain valid Type I error control (within 0.01 tolerance), the optimal pre-test alpha is approximately 0.534. Next, for a given sample size $n = 10$, and for any given distribution $\mathcal{D}$, we create a plot of power versus effect size as shown in Figure 2.14. There are both gains and loss in power at the chosen optimum alpha pre-test where Type I error rates are controlled in both normal and non-normal samples however, the gains far exceeds the loses, signifying some good benefits in pre-testing for normality.

Next, for a given sample size n, and for any given distribution $\mathcal{D}$, we create a plot of power versus effect size. All test procedures produced similar power as shown in Figure 2.14 below.

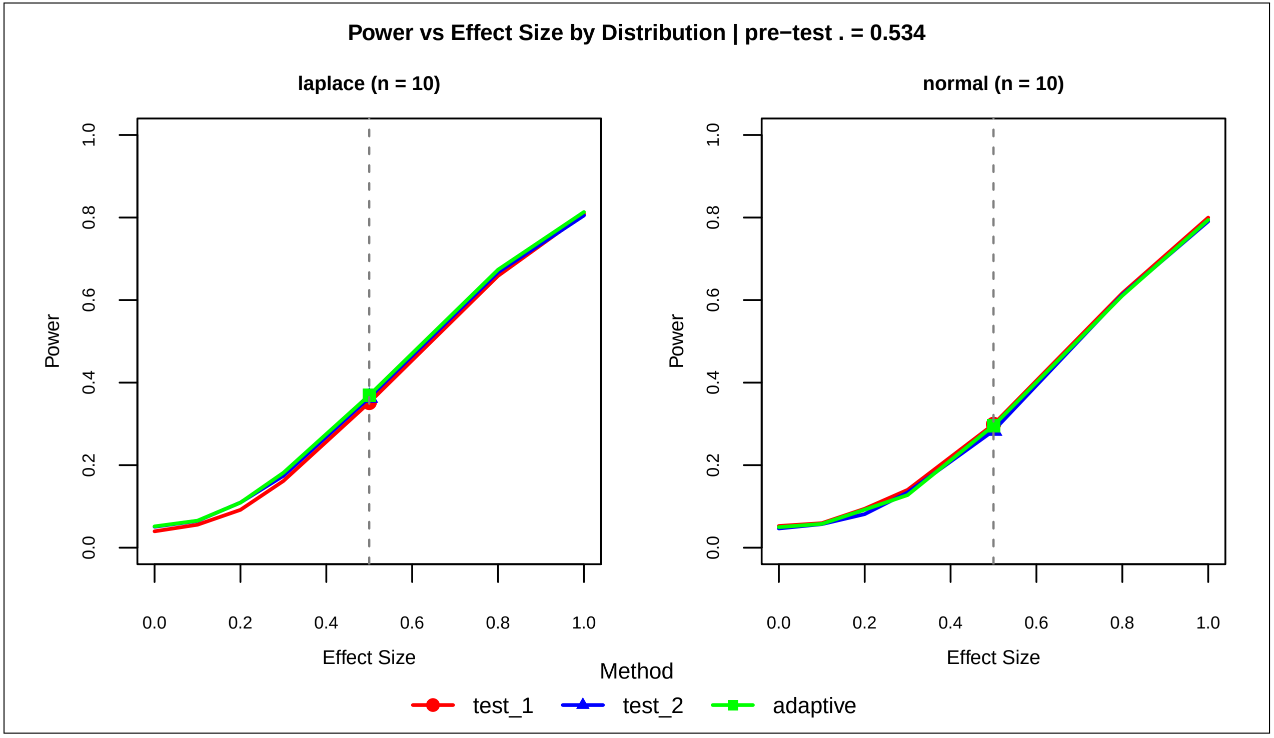


Figure 2.14: *Showing power versus effect size for samples of size 10 drawn from normal and exponential distributions.*

To investigate the relative effects of pre-testing on the power and Type I error of a downstream procedure for a given sample size, we plot the power against the Type I error rate to obtain an ROC-Like curve as shown in Figure 2.15. All test methods controlled type I error rates at the nominal alpha, and are similar in their performance across the given significance levels.

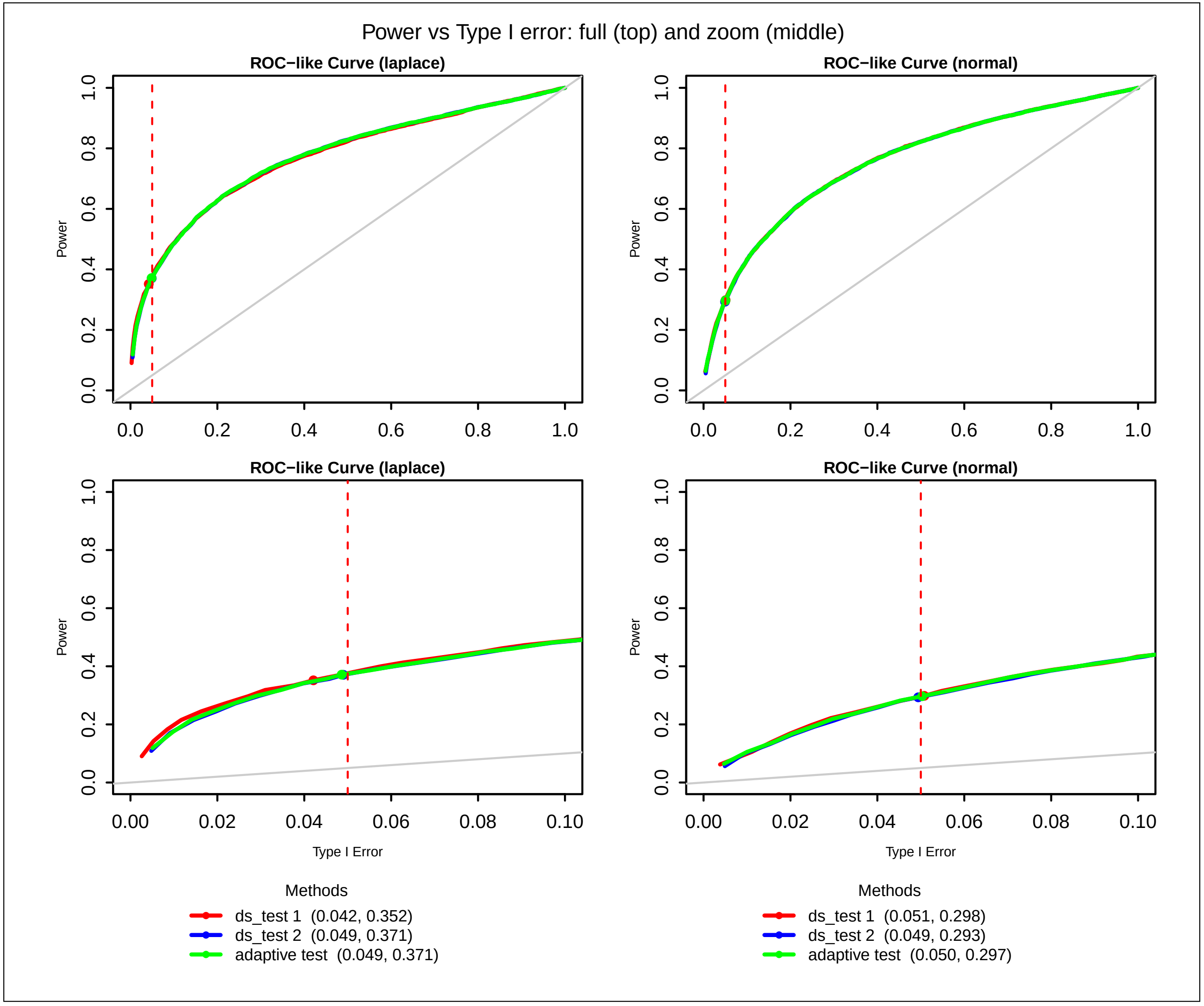


Figure 2.15: *Showing the power vs Type I error plot for t-test, one-sample permutation test and the adaptive test based on Shapiro-Francia test for normality in an ROC-Like curve for a given sample size* $n = 10$. *ds_test 1 and ds_test 2 mean test 1 and test 2 respectively.*

Next we compare the power and Type I error rates of the *t-test*, *one-sample permutation test* and the *adaptive procedure* across varying sample sizes, estimating the area under each curve as a measure of the overall performance of each method as shown in Figure 2.16. Note that we need to estimate α_{pre} for each sample size for this. This requires multiple iterations and so we assumed a uniform. Using the α_{pre} together with the researchers choice of effect size, we calculate the power as shown in figure 2.16 below.

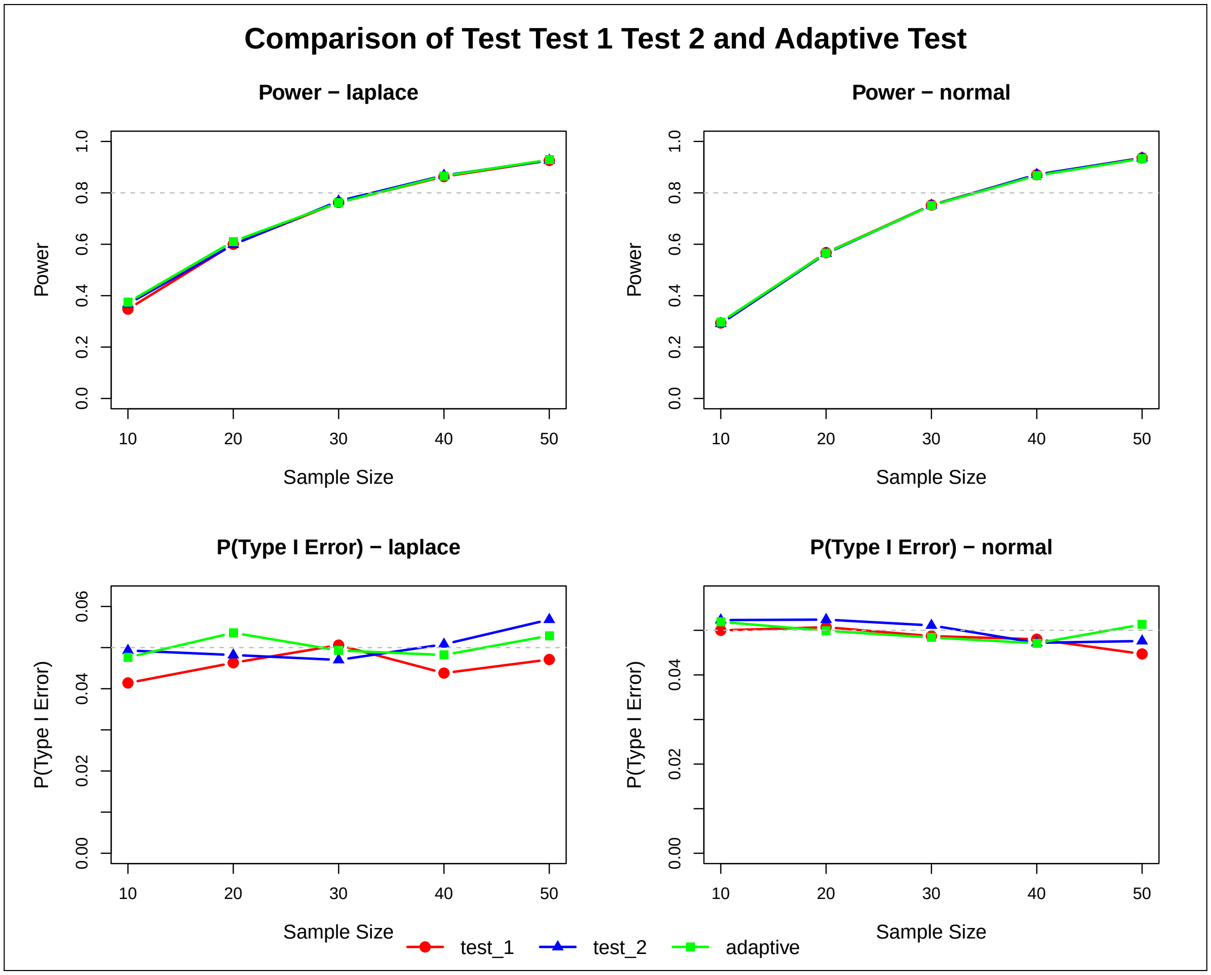


Figure 2.16: *Showing the area under the power curves and area under the Type I error curves for t-test, one-sample permutation test and the adaptive test by pre-testing for samples drawn from Laplace and normal distributions.*

Distribution	Method	AUC-Power	AUC-TypeI
Laplace			
	Parametric (t-test)	0.717 68	0.047 66
	Nonparametric (permutation)	0.720 67	0.049 15
	Adaptive	0.719 78	0.048 72
Normal			
	Parametric (t-test)	0.702 94	0.048 42
	Nonparametric (permutation)	0.699 45	0.049 67
	Adaptive	0.700 83	0.050 67

Table 2.4: *AUC for power and Type I error by method and distribution for one sample t, permutation, and adaptive tests for samples drawn from Laplace and normal distributions.*

Observations and Recommendations: The above analysis showed no real gains in using the adaptive procedure as it obtained power similar or lower than that test 1. This is primarily because test 1 and test 2 had very similar performance here. Hence pre-testing is not necessary in this case. One may proceed using either test 1 or test 2 since both are conservative, and have similar power.

2.6.5 Demonstration 4: One-sample t-test vs Box-Cox t-test:

Here we maintained the original t for test 1 and then use a modified t-test after Box-Cox transformation for test 2 and test for the median. See section 2.4.1 for details on Box-Cox t-test.

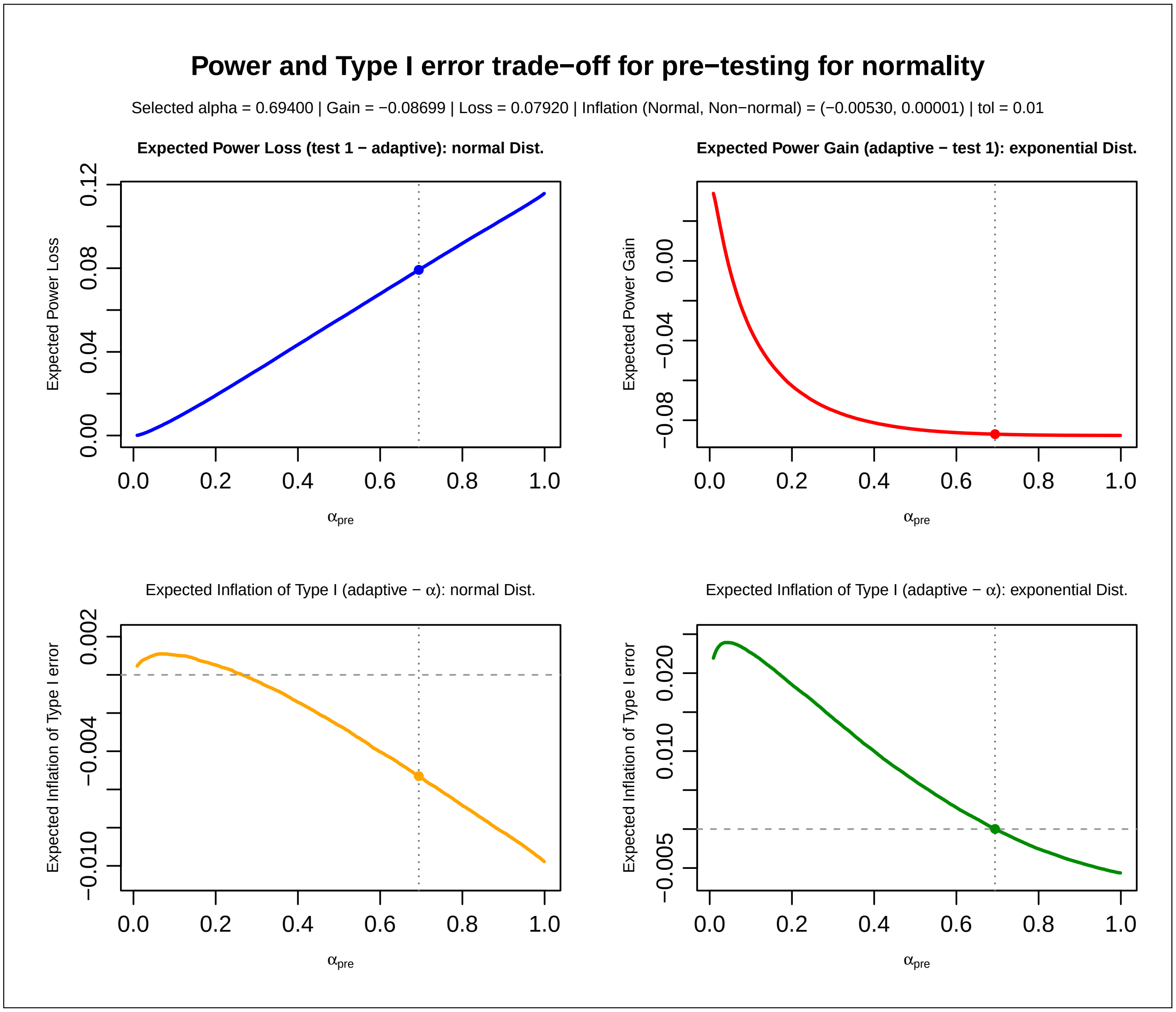


Figure 2.17: *Showing expected power loss for pretesting, left plot and the expected power gain for pre-testing, right plot in row 1. Row 2 displays expected inflation of Type I error rates for samples from normal and exponential distribution for sample of sizes 10 for t-test, box-cox t and the adaptive tests. The optimal α_{pre} accounting for 0.001 tolerance is about 0.694.*

In this particular situation, though type I error rates are controlled at the optimum alpha for pre-test, we loss power in both cases. This suggest that the Box-Cox t-test is less powerful compared to the classical t-test.

Next, for a given sample size n, and for any given distribution $\mathcal{D}$, we create a plot of power versus effect size. From Figure 2.18 below, there is a benefit using the adaptive adaptive test, instead of just the Box-Cox t-test.

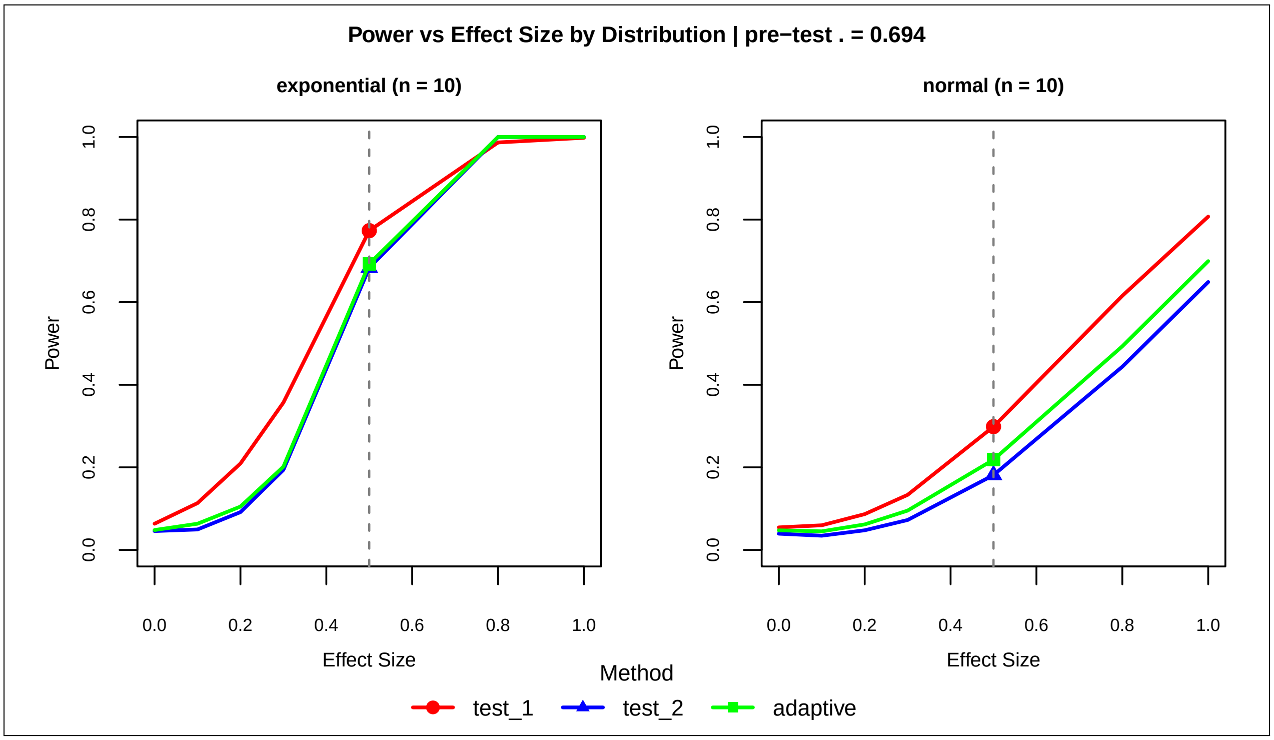


Figure 2.18: *Showing power versus effect size for samples of size 10 drawn from normal and exponential distributions.*

To investigate the relative effects of pre-testing on the power and Type I error of a downstream procedure for a given sample size, we plot the power against the Type I error rate to obtain an ROC-Like curve as shown in Figure 2.19. Test 1 and the adaptive test could not control type I error rates at the nominal alpha. The adaptive test however yielded higher power compared to the Box-Cox t-test signifying some potential benefits in using the adaptive test instead of the Box-Cox t-test.

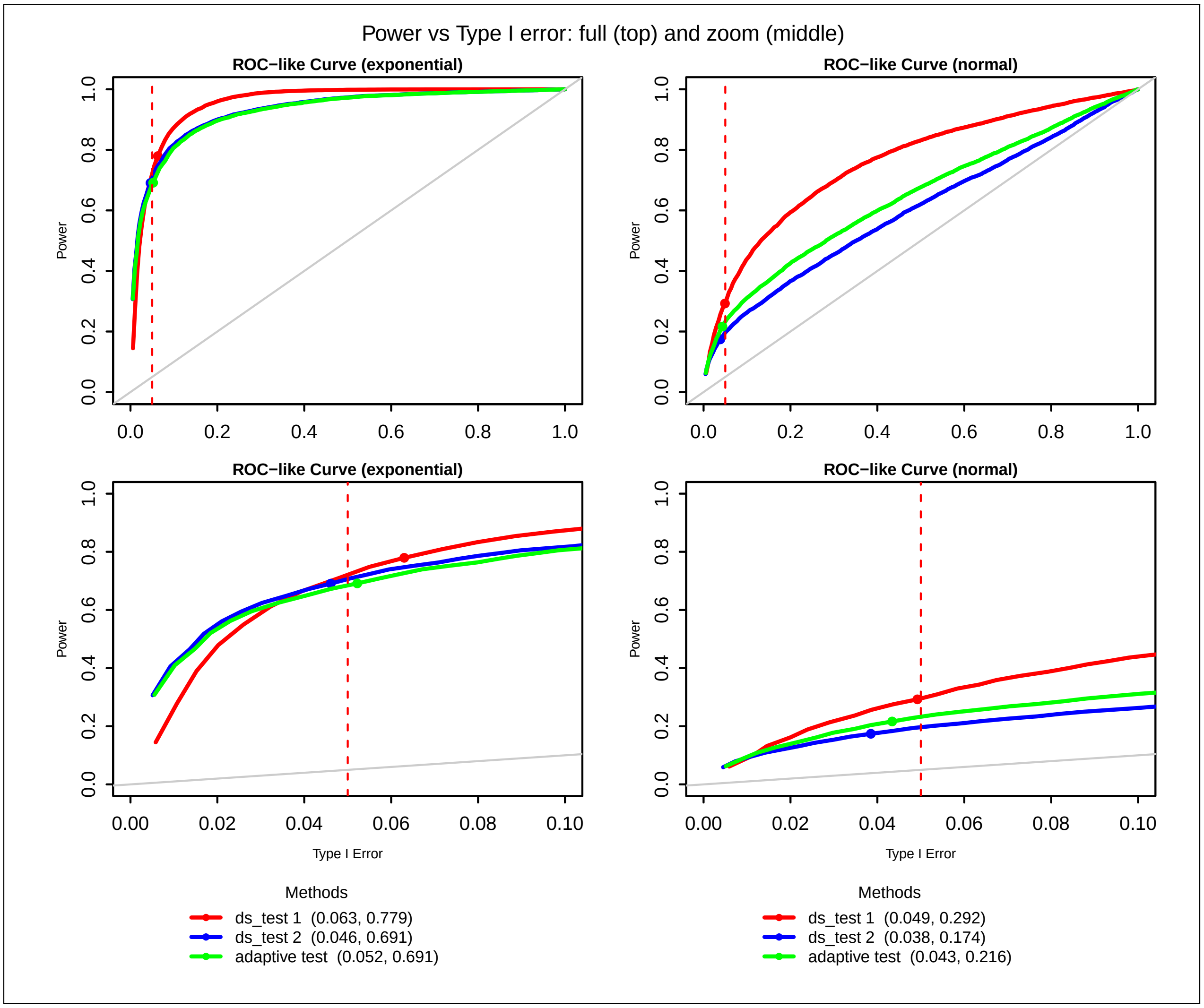


Figure 2.19: *Showing the power vs Type I error plot for t-test, box-cox t and the adaptive tests based on Shapiro-Wilk test for normality in an ROC-Like curve for a given sample size* $n = 10$. *ds_test 1 and ds_test 2 mean test 1 and test 2 respectively.*

Next we compare the power and Type I error rates of the *t-test*, *box-cox t-test* and the *adaptive procedure* across varying sample sizes, estimating the area under each curve as a measure of the overall performance of each method as shown in Figure 2.20. Again, the target parameter of the t-test is wrong when samples come from a non-normal distribution and hence the significant inflation of type I error rate.

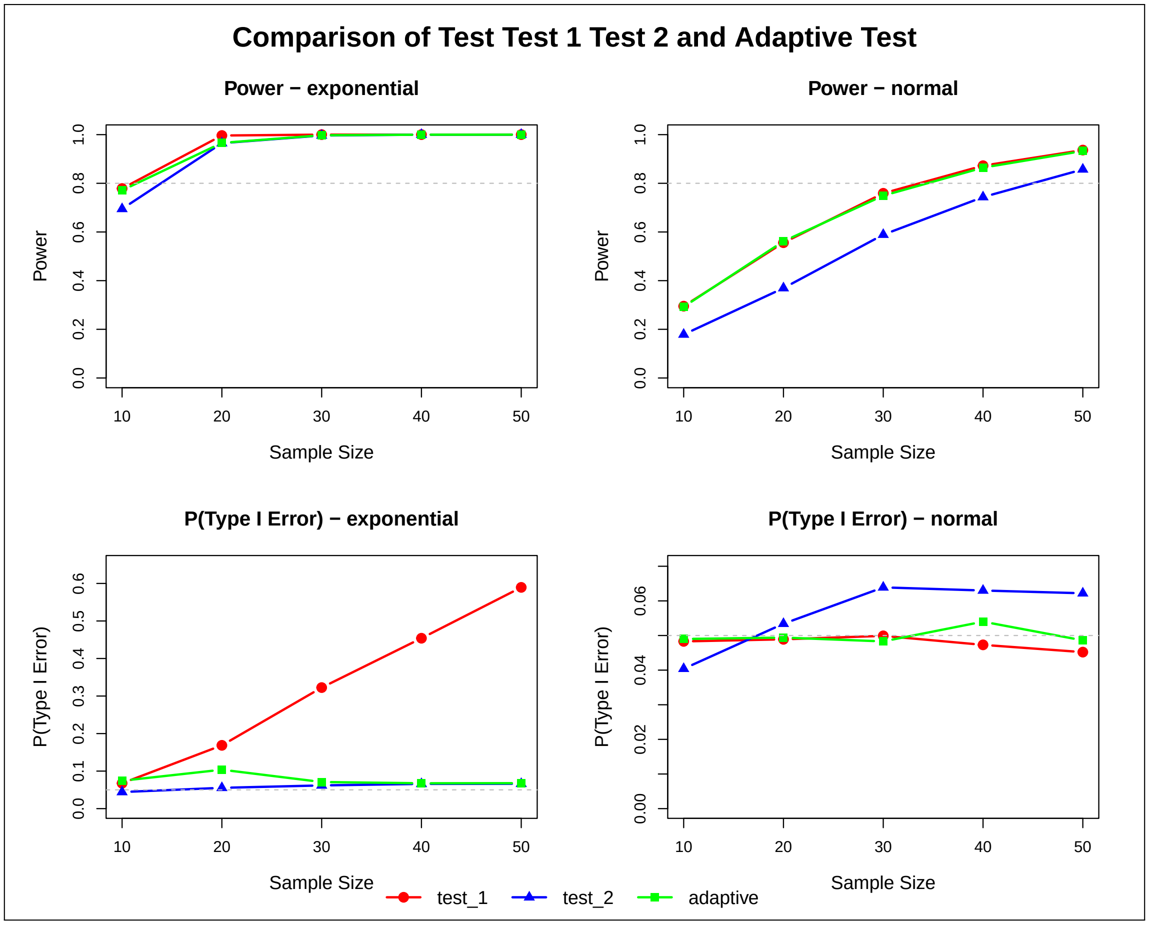


Figure 2.20: *Showing the area under the power curves and area under the Type I error curves for t-test, box-cox t-test and the adaptive test by pre-testing for samples drawn from exponential and normal distributions.*

Distribution	Method	AUC-Power	AUC-TypeI
Exponential			
	Parametric (t-test)	0.971 48	0.315 98
	Nonparametric (Box-Cox t)	0.952 05	0.058 43
	Adaptive	0.961 86	0.076 41
Normal			
	Parametric (t-test)	0.699 50	0.049 58
	Nonparametric (Box-Cox t)	0.560 37	0.056 58
	Adaptive	0.694 51	0.050 98

Table 2.5: *AUC for power and Type I error by method and distribution for one sample t, Box-Cox t, and adaptive tests.*

Observations and Recommendations: The above analysis showed some benefits in using the adaptive procedure in terms of controlling type I error rates compared to test 1(t-test). It also has higher power compared to test 2(Box-Cox t), hence, users may consider the adaptive approach for these gains in similar cases. However, if computation time will be an issue, one may opt for the Box-Cox t-test instead.

2.6.6 Comparison of Test Procedures for Exponential Distribution

Figure 2.21 compares the performance of test 1, test 2, and adaptive tests for the Sign, and Box-Cox t-tests across different sample sizes. Since the target parameter for the Sign and Box-Cox t-tests is the median, the standard t-test (test 1) is invalid under the non-normal Exponential distribution, leading to a severe inflation of the Type I error rate that worsens with larger samples. Except for the Sign test, all test 2 variants fail to maintain the nominal Type I error rate. Adaptive tests also perform poorly for small samples, even for the Sign test, where the normality tests have low power to detect non-normality.

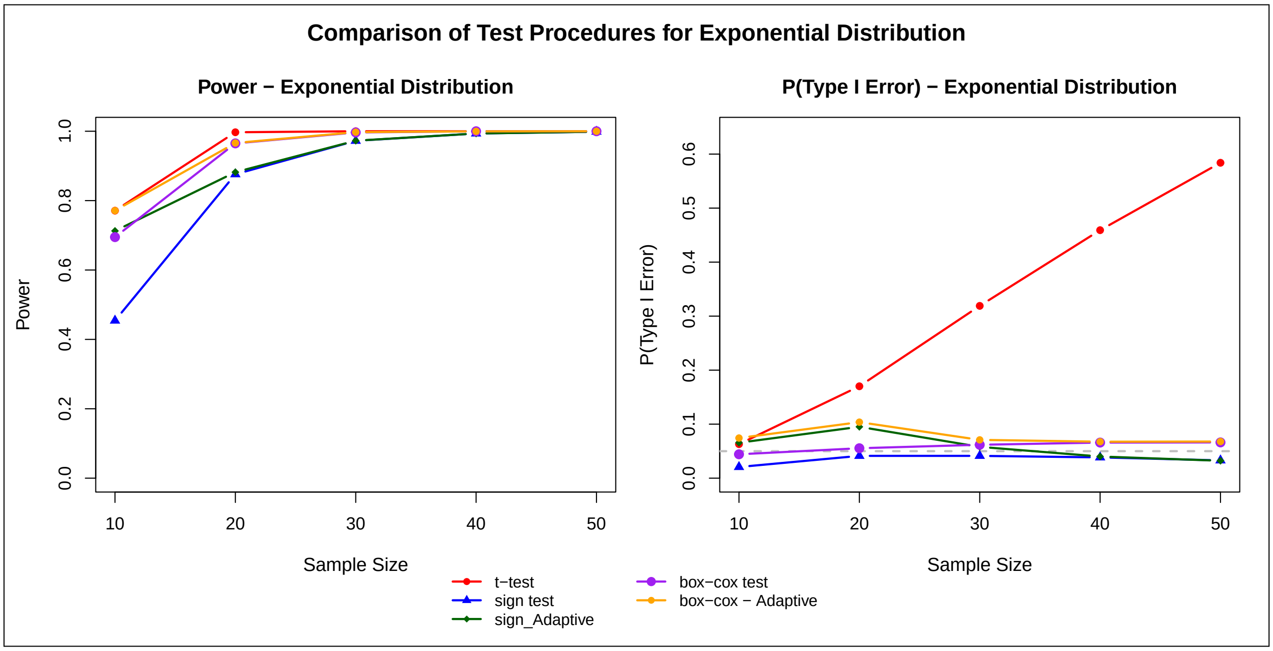


Figure 2.21: *Comparison of power and type I error rates of one sample Sign, Box-Cox t-test for test 1, test 2 and adaptive test for Exponential Distribution*

2.6.7 Two-Sample Location Test - The set up

For two-sample test, we require either both samples drawn from a normally distributed population or the residuals of collapsed samples, see Rochon et al. (2012). In this case, we test whether both samples come from a normally distributed population. Normality is considered satisfied if both samples passed normality test.

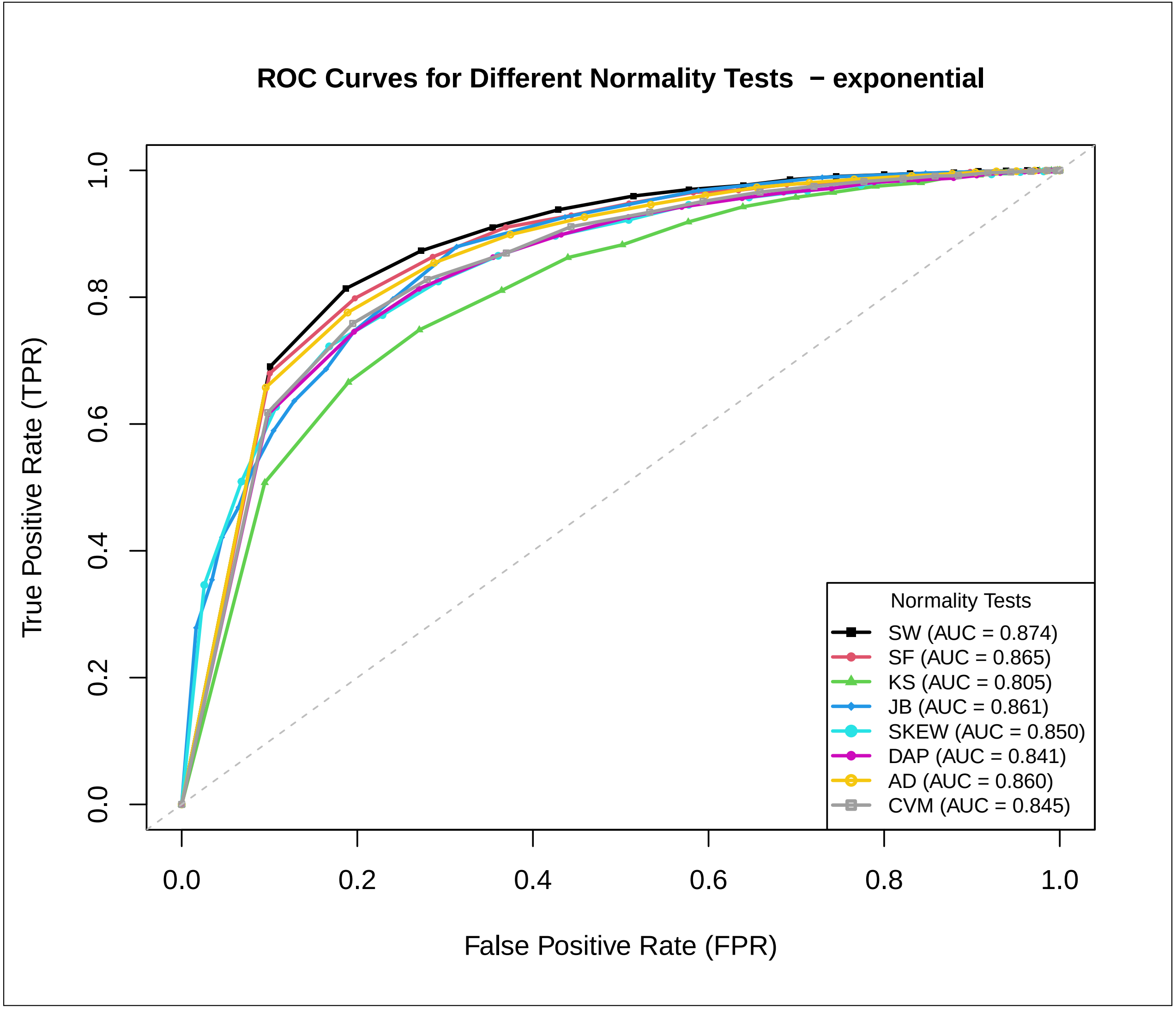


Figure 2.22: *Showing ROC curves for different normality test methods for pair of sample sizes 10 drawn from normal and exponential distribution.*

2.6.8 Demonstration 5: Two-Sample t-Test vs. Mann-Whitney Test

We set the target parameter so both tests, test 1 (two-sample t-test) and test 2 (Mann-Whitney U test) are valid under normality, testing whether the population medians are equal. Note this is true with additional assumptions that the two population are mutually independent for the Mann-Whitney U test (Hollander et al., 2013)

Using the Shapiro-Wilk test (N_T) for normality assessment, we evaluate expected gains and losses of pre-testing for normality across significance levels α_{pre} from Figure 2.23 below.

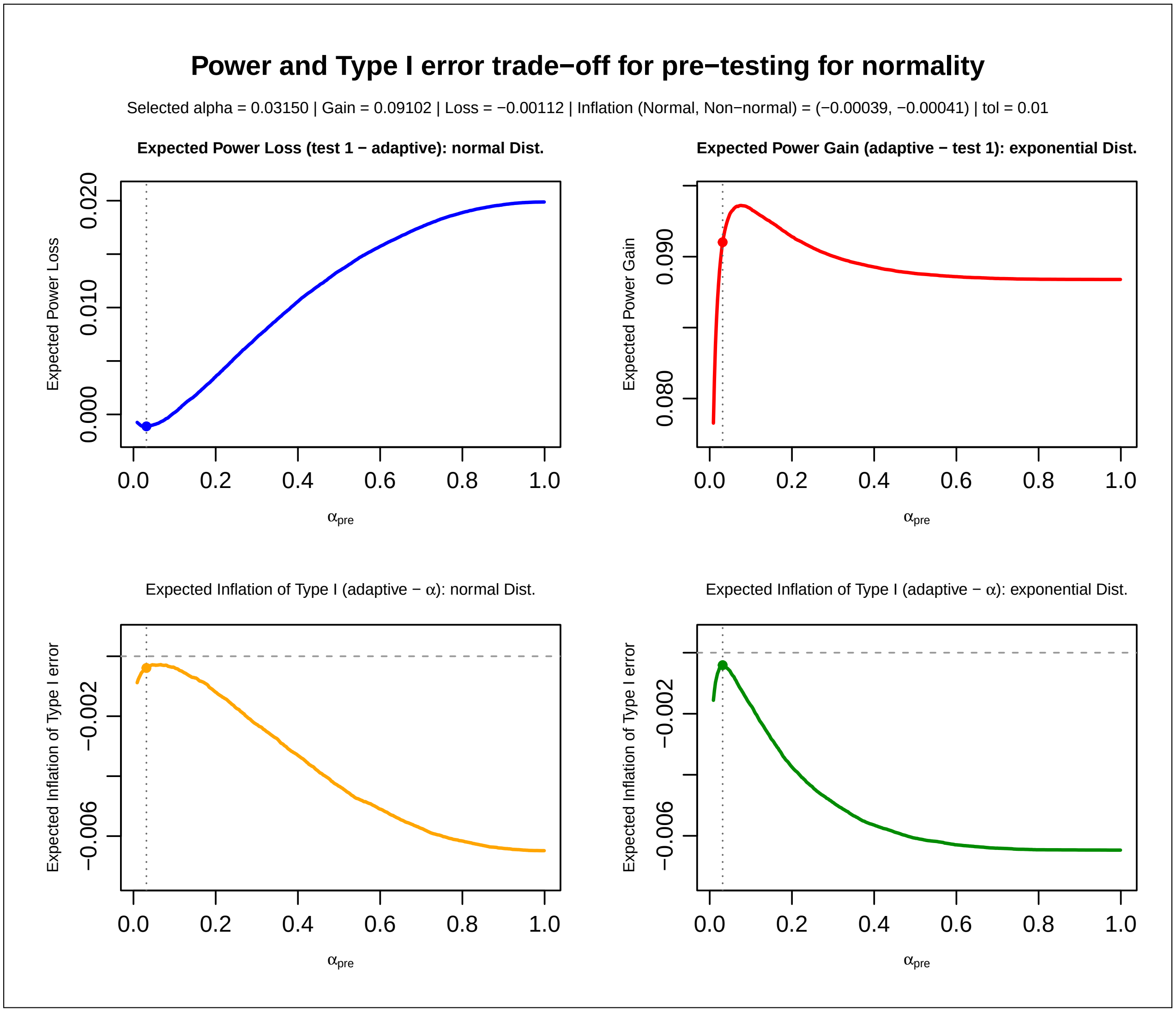


Figure 2.23: *Showing expected power loss for pretesting, left plot and the expected power gain for pre-testing, right plot in row 1. Row 2 displays expected inflation of Type I error rates for sample of sizes 10 each from normal and exponential distribution.*

Since the two-sample test is generally robust to non-normality, type I error rates remain controlled at the nominal level in all situations. This shifts the focus on the expected power loss and expected power gain to decide the optimal value of alpha. Thus, the value of alpha that yielded the lowest expected power loss and the highest power gain is about 0.0315 as shown in Figure 2.23 above.

Next, for a pair of given sample sizes $n = 10$ each and for any given distribution $\mathcal{D}$, we create a plot of power versus effect sizes as shown in Figure 2.24.

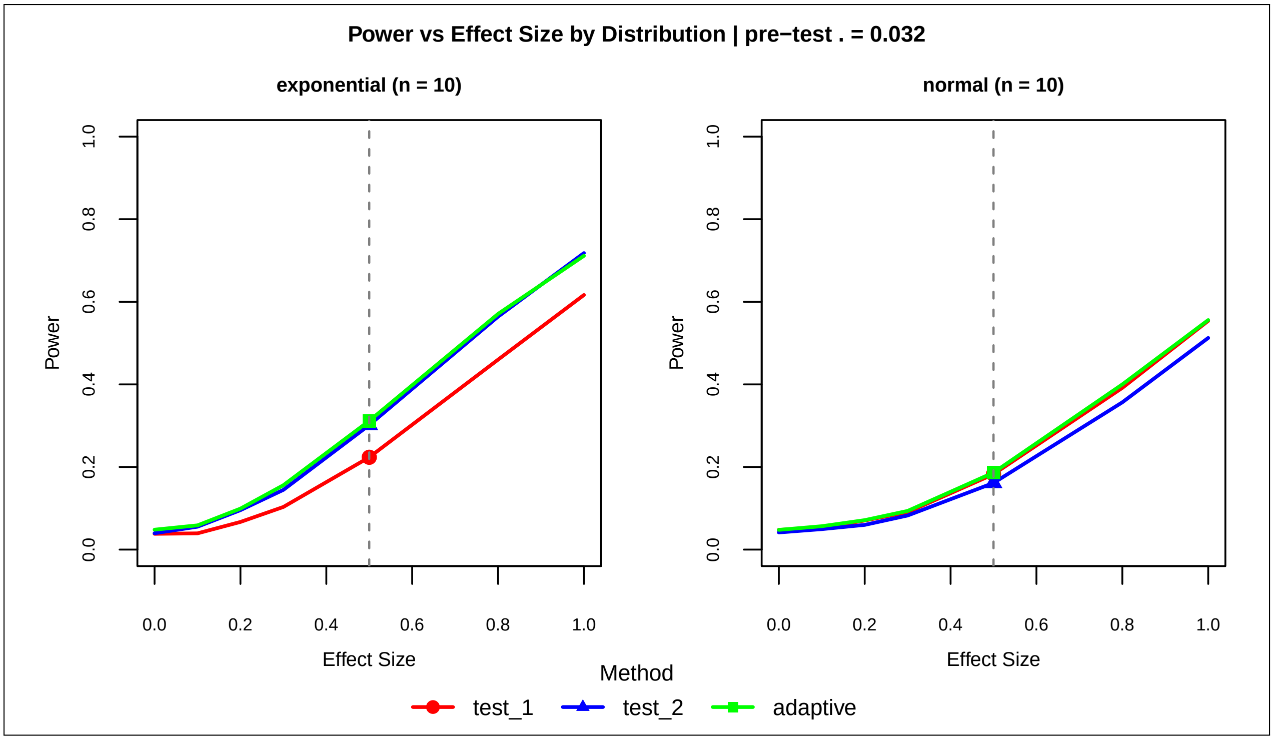


Figure 2.24: *Showing power versus effect size for samples of sizes 10 each drawn from normal and exponential distributions for t, Mann-Whitney U, and adaptive test.*

Next, we plot the power against the Type I error rate to obtain an ROC-Like curve. With α_{pre} from Figure 2.23, together with the researchers, choice of effect size, we created the ROC curve. Test 2 and the adaptive test produced higher than test 1 when samples come from a non-normal(exponential) distribution, however, for normal samples, test 1 and adaptive test are best, indicating some benefits in using the adaptive test as shown in Figure 2.25 below.

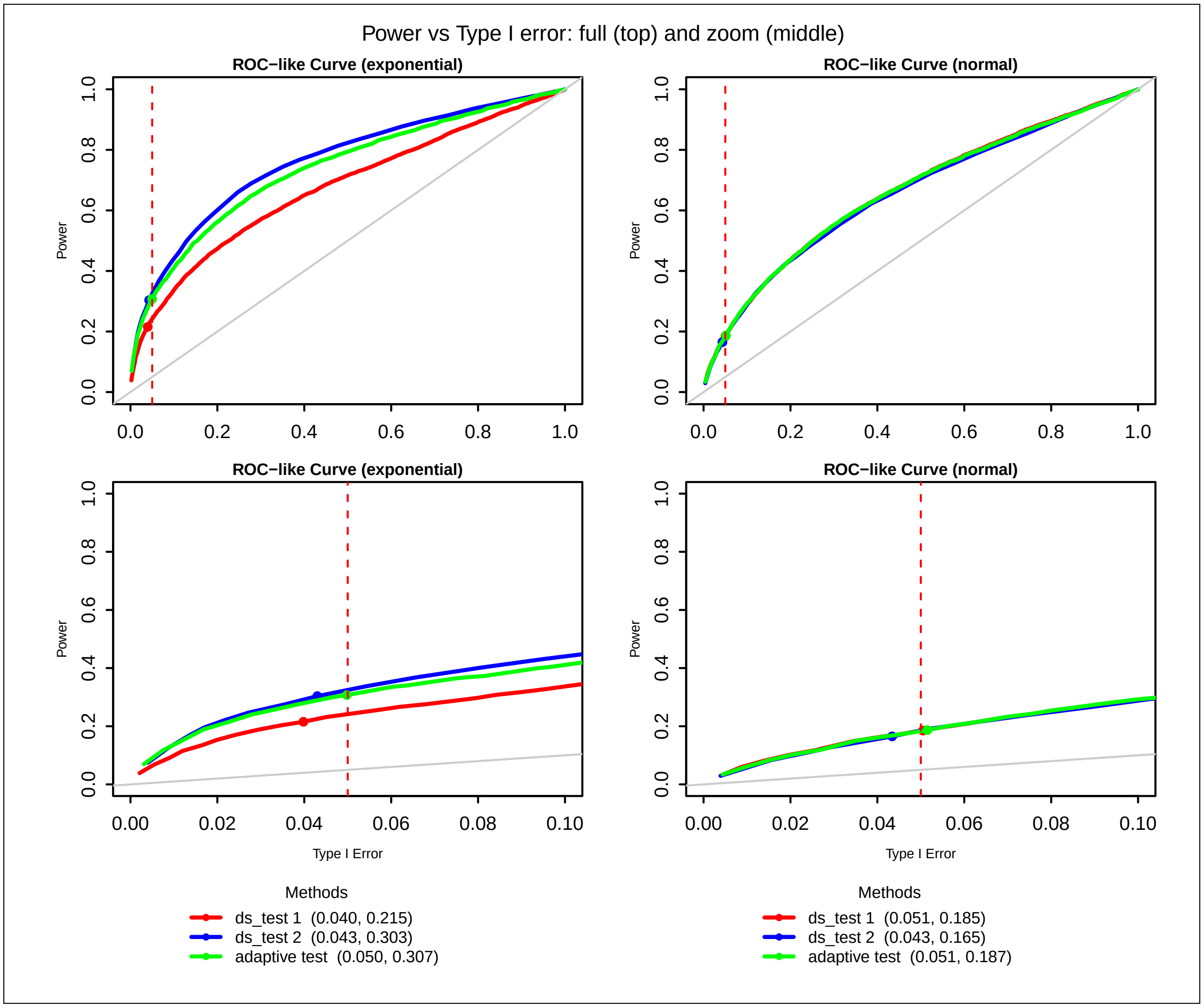


Figure 2.25: *Showing the power vs Type I error plot for two-sample t-test, Mann-Whitney U test test and the adaptive test based on Shapiro-Wilk test for normality in an ROC-Like curve for a given sample size* $n = 10$. *ds_test 1 and ds_test 2 mean test 1 and test 2 respectively.*

Next we compare the power and Type I error rates of the *t-test*, *Mann-Whitney U-test* and the *adaptive procedure* across varying sample sizes, estimating the area under each curve as a measure of the overall performance of each method as shown in Figure 2.26.

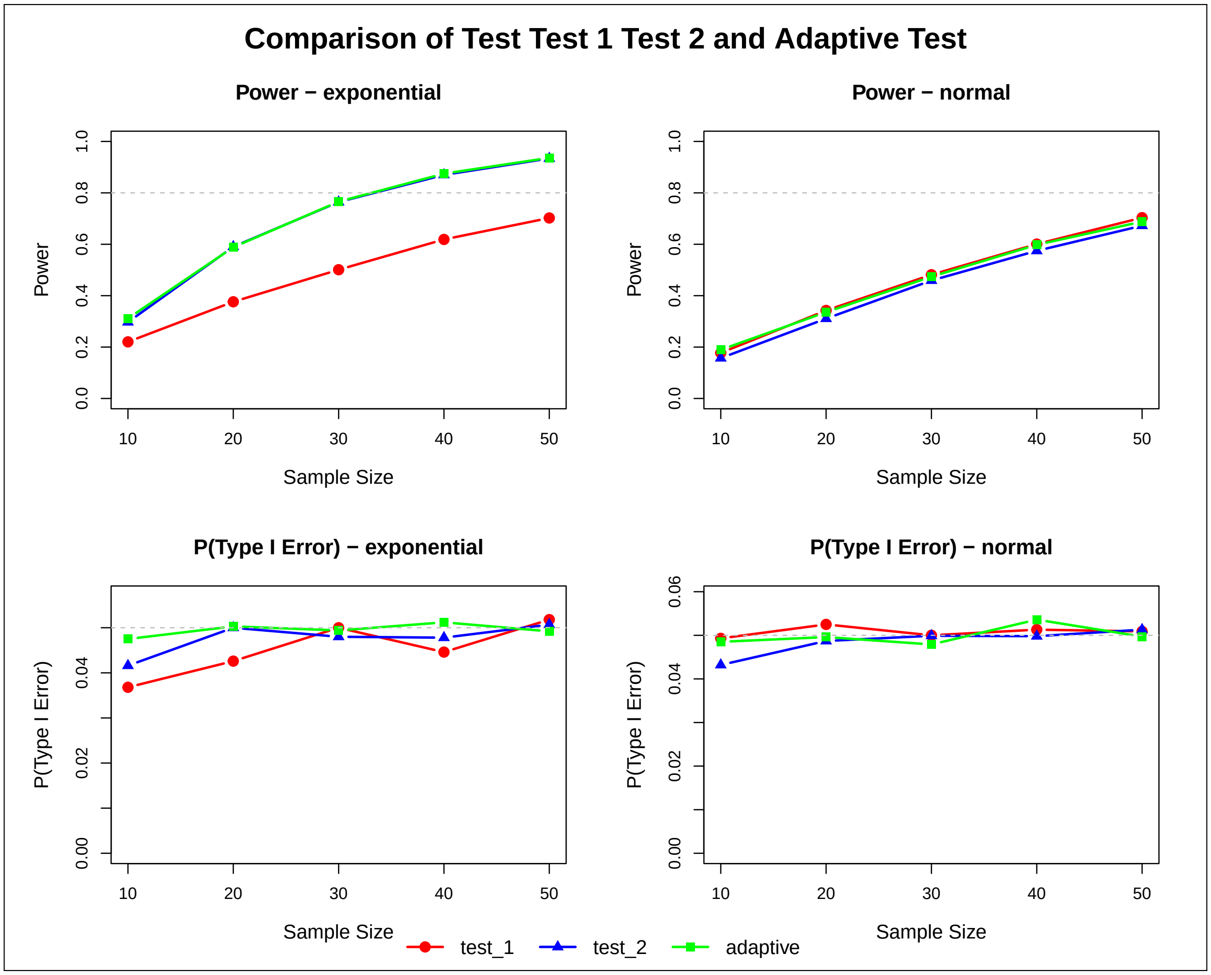


Figure 2.26: *Showing the area under the power curves and area under the Type I error curves for t-test, Mann-Whitney U-test and the adaptive test for samples drawn from exponential and normal distributions.*

Distribution	Method	AUC-Power	AUC-TypeI
Exponential			
	Parametric (t-test)	0.489 08	0.045 75
	Nonparametric (Mann-Whitney U test test)	0.711 89	0.049 04
	Adaptive	0.713 04	0.049 72
Normal			
	Parametric (t-test)	0.462 68	0.049 63
	Nonparametric (Mann-Whitney U test test)	0.443 59	0.048 75
	Adaptive	0.462 66	0.050 44

Table 2.6: *AUC for power and Type I error by method and distribution for two-sample t, Mann-Whitney U test, and adaptive tests.*

Observations and Recommendations: The above analysis showed that for two-sample location test with at least equal sample sizes, practitioners may consider using the nonparametric, Mann-Whitney U test test as it controls type I error rate with comparable power with the t and adaptive tests in both normal and non-normal sample cases. There is some gains in using the adaptive procedure, but considering computation time, the Mann-Whitney U test test should be preferred.

2.6.9 Demonstration 6: Two-sample t-Test vs Permutation Test

Here we compare the two-sample t-test to the two-sample permutation test procedure. The goal is to test whether the two population means are equal. The two-sample t-test statistic was used to conduct the permutation test. Again, with the chosen normality test method, N_T, Shapiro-Wilk, test we assess the expected gains and losses for pre-testing for normality at different normality alpha value, α_{pre}.

Type I error rates are controlled in both cases, and since expected power loss is already negative, we maximize expected power loss as much as possible. Thus, the value of alpha that yielded the highest expected power gain is about 0.5815 as shown in Figure 2.27 below.

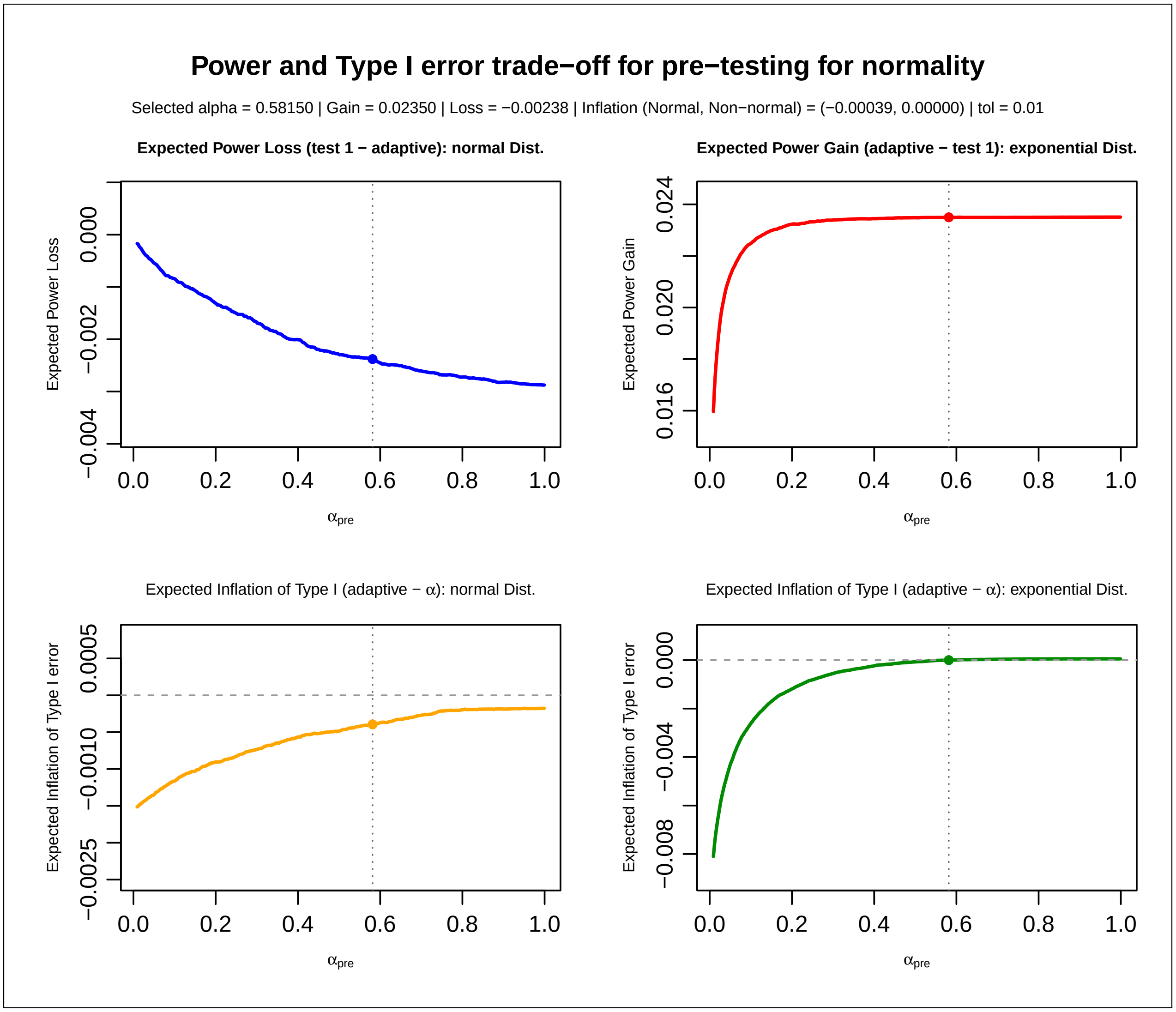


Figure 2.27: *Showing expected power loss for pretesting, left plot and the expected power gain for pre-testing, right plot in row 1. Row 2 displays expected inflation of Type I error rates for sample of sizes 10 each from normal and exponential distribution for the two-sample t, Mann-Whitney U, and adaptive tests.*

Next, for a given sample size n, and for any given distribution $\mathcal{D}$, we create a plot of power versus effect size as shown in Figure 2.28. The adaptive test and test 2 attained higher power compared to test 1 for non-normal samples, but all three tests have similar power when samples come from normal distribution.

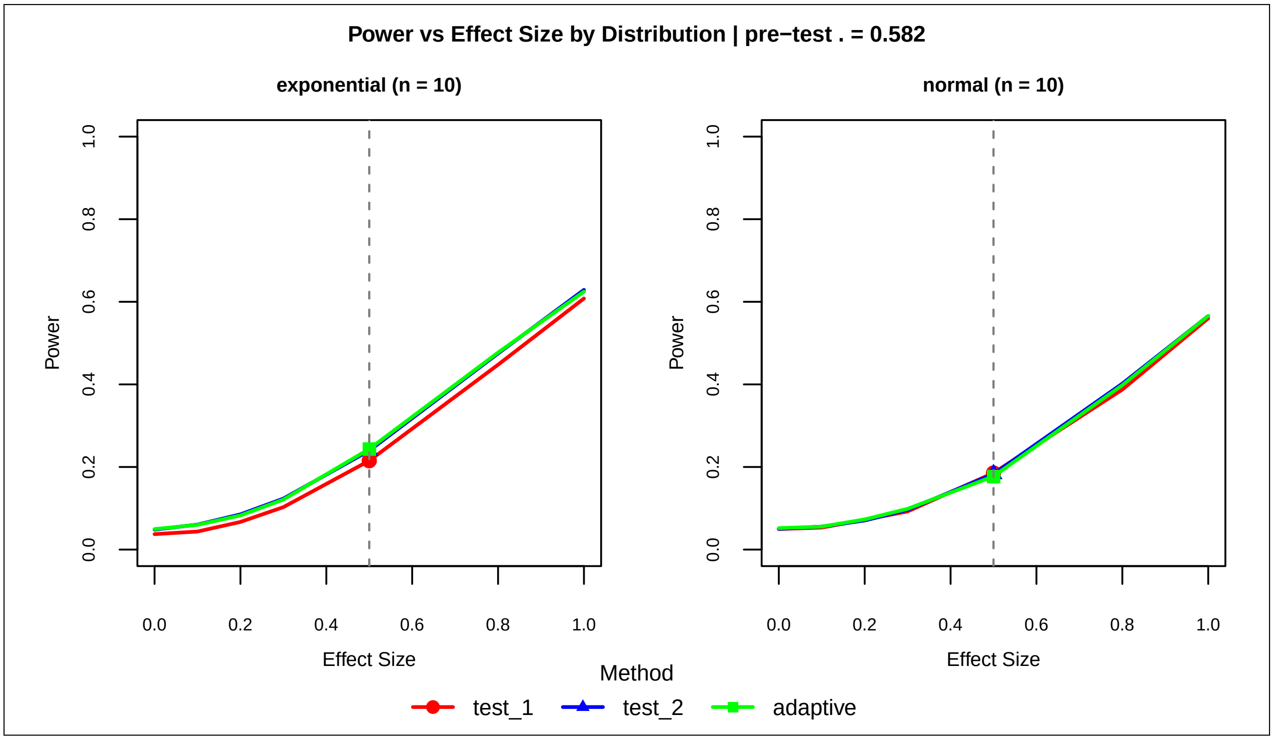


Figure 2.28: *Showing power versus effect size for samples of size 10 drawn from normal and exponential distributions.*

To investigate the relative effects of pre-testing on the power and Type I error of a downstream procedure for a given sample size, we plot the power against the Type I error rate to obtain an ROC-Like curve. Apparently, there is no difference in performance among the three procedures as shown in Figure 2.29 below.

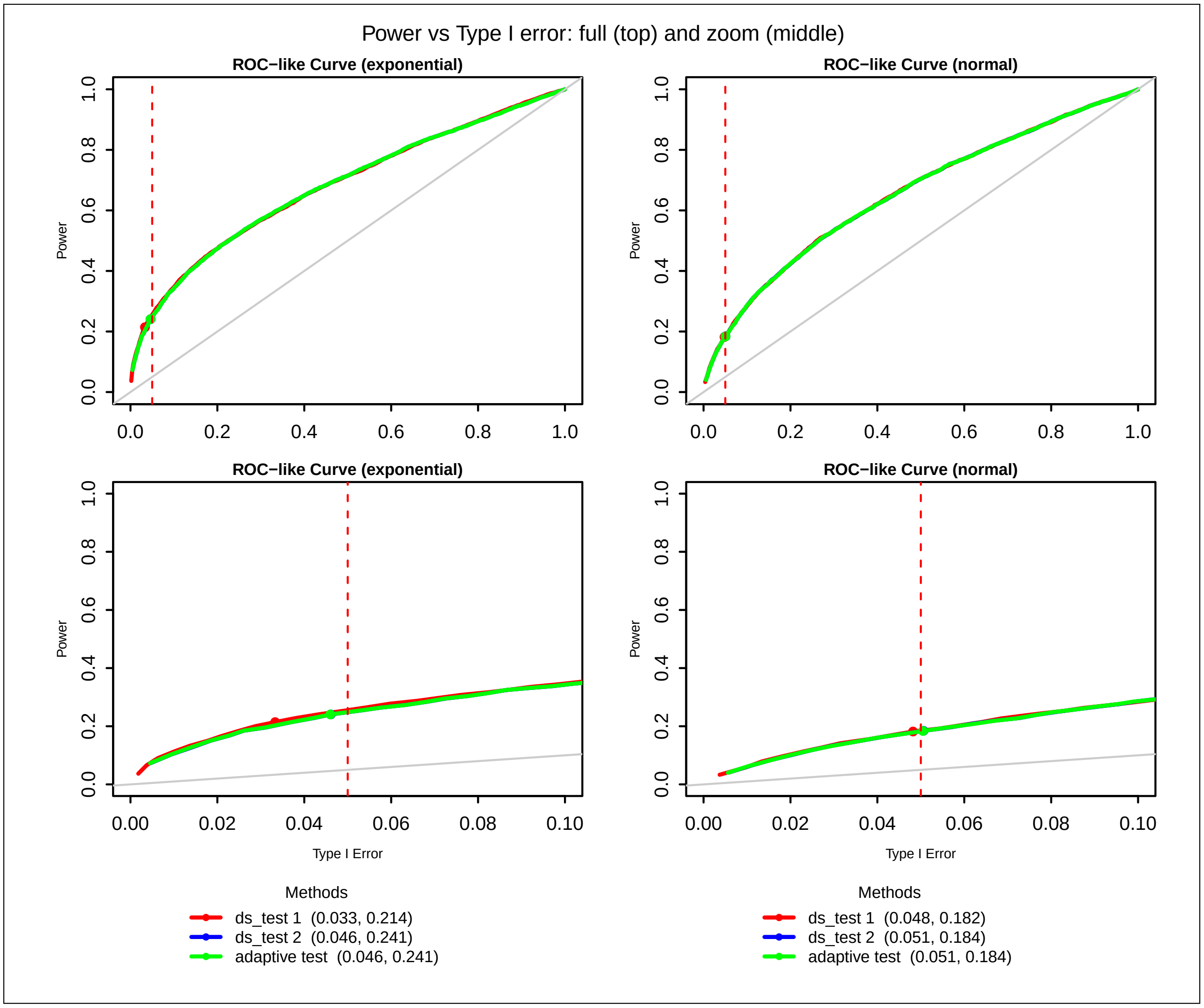


Figure 2.29: *Showing the power vs Type I error plot for two-sample t-test, two-sample permutation test and the adaptive test based on Shapiro-Wilk test for normality in an ROC-Like curve for a given sample size* $n = 10$. *ds_test 1 and ds_test 2 mean test 1 and test 2 respectively.*

Next we compare the power and Type I error rates of the two-sample t-test, two-sample permutation test and the *adaptive procedure* across varying sample sizes, estimating the area under each curve as a measure of the overall performance of each method. Type I error rates are generally controlled, with nearly identical overall power among all test procedures as shown in Figure 2.30.

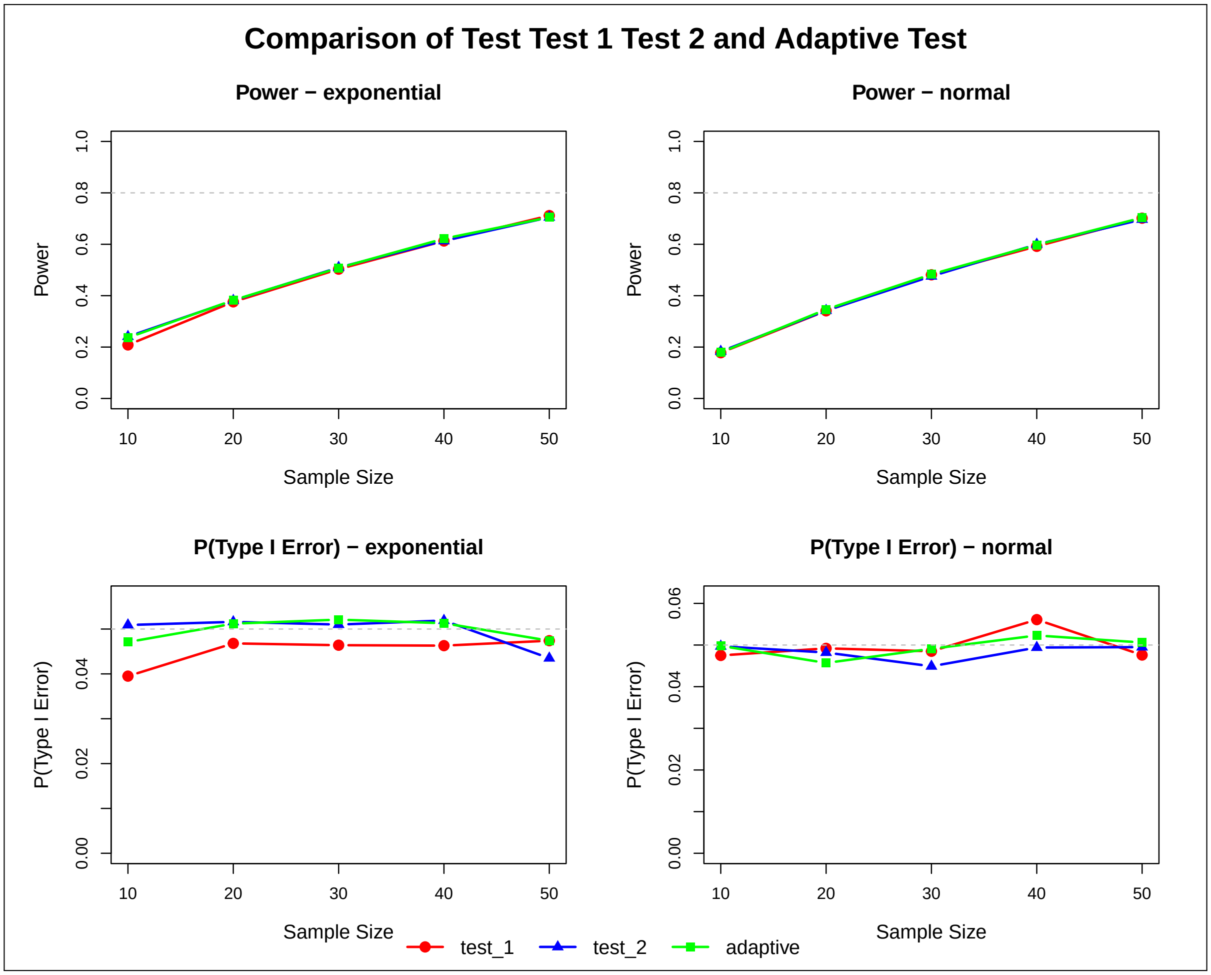


Figure 2.30: *Showing the area under the power curves and area under the Type I error curves for two-sample t-test,two-sample permutation test and the adaptive test by pre-testing for samples drawn from exponential and normal distributions.*

Distribution	**Method**	**AUC-Power**	**AUC-TypeI**
Exponential			
	Parametric (t-test)	0.489 95	0.045 47
	Nonparametric (permutation test)	0.496 18	0.050 08
	Adaptive	0.5001	0.048 61
Normal			
	Parametric (t-test)	0.463 54	0.049 26
	Nonparametric (permutation test)	0.464 58	0.050 77
	Adaptive	0.463 51	0.050 97

Table 2.7: *AUC for power and Type I error by method and distribution for two-sample t, Permutation, and adaptive tests.*

Observations and Recommendations: Based on the analysis above, all three test procedures controlled type I error rates with similar power, thus the adaptive procedure may not be a good choice, considering computation time and complications. We recommend the two-sample t-test in this case. Users should note that this is not a general recommendation.

2.6.10 One-Way ANOVA - The set up

For ANOVA, the residuals are what we require to be normally distributed, so we apply the normality test on the model's residuals, calculate the TPR and the FPR and then create the ROC as shown in figure 2.31 below.

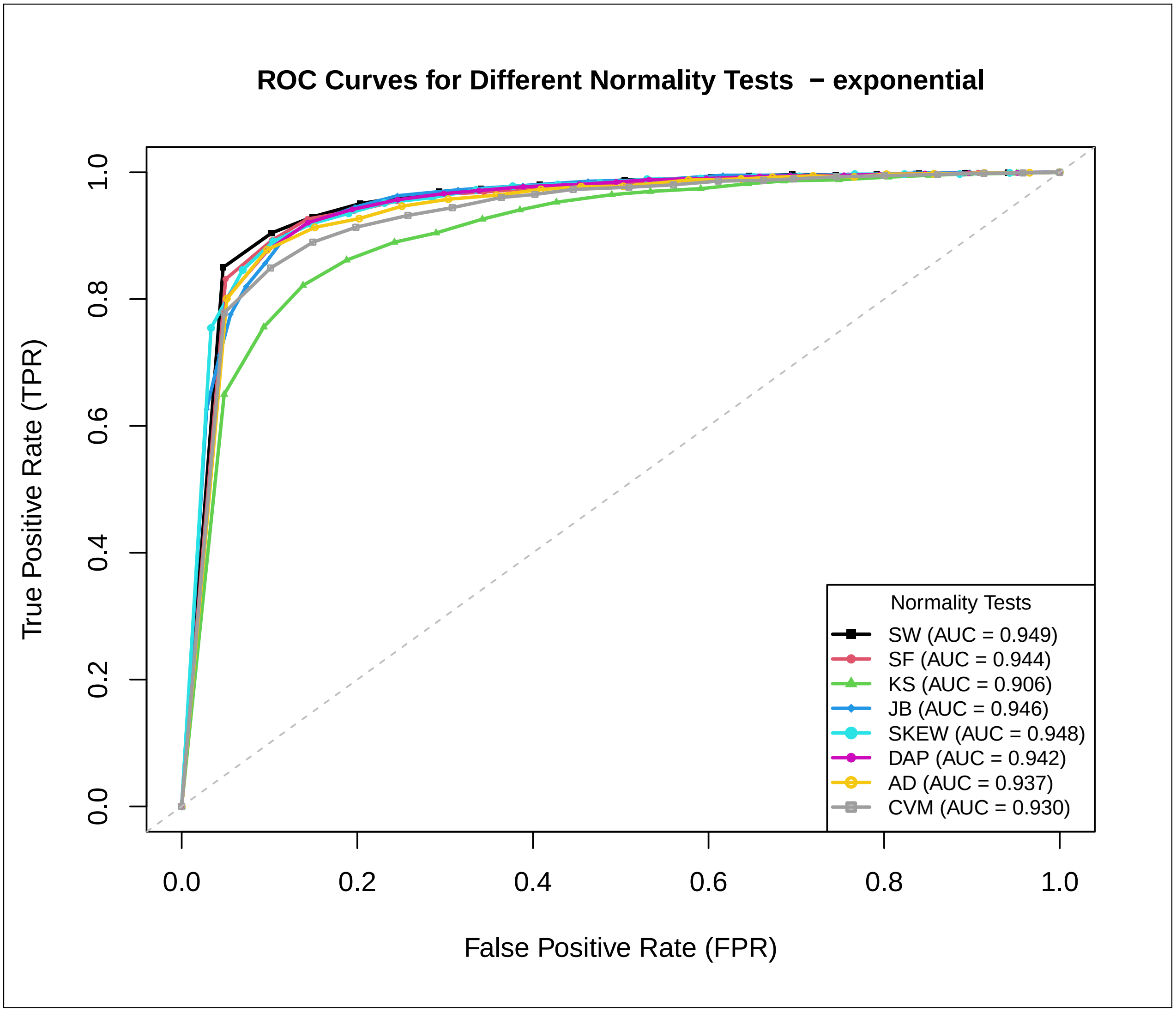


Figure 2.31: *Showing ROC curves for different normality test methods for sample size 10 drawn from normal and exponential distributions.*

2.6.11 Demonstration 7: One-Way ANOVA vs Kruskal-Wallis Test

We chose test 1 to be the ordinary One-Way ANOVA F-test and test 2 to be the nonparametric Kruskal-Wallis Test testing whether samples originate from the same distribution. With the chosen normality test method, N_T as the third moment (Skewness), we assess the expected gains and losses for pre-testing for normality at different normality alpha value, α_{pre}.

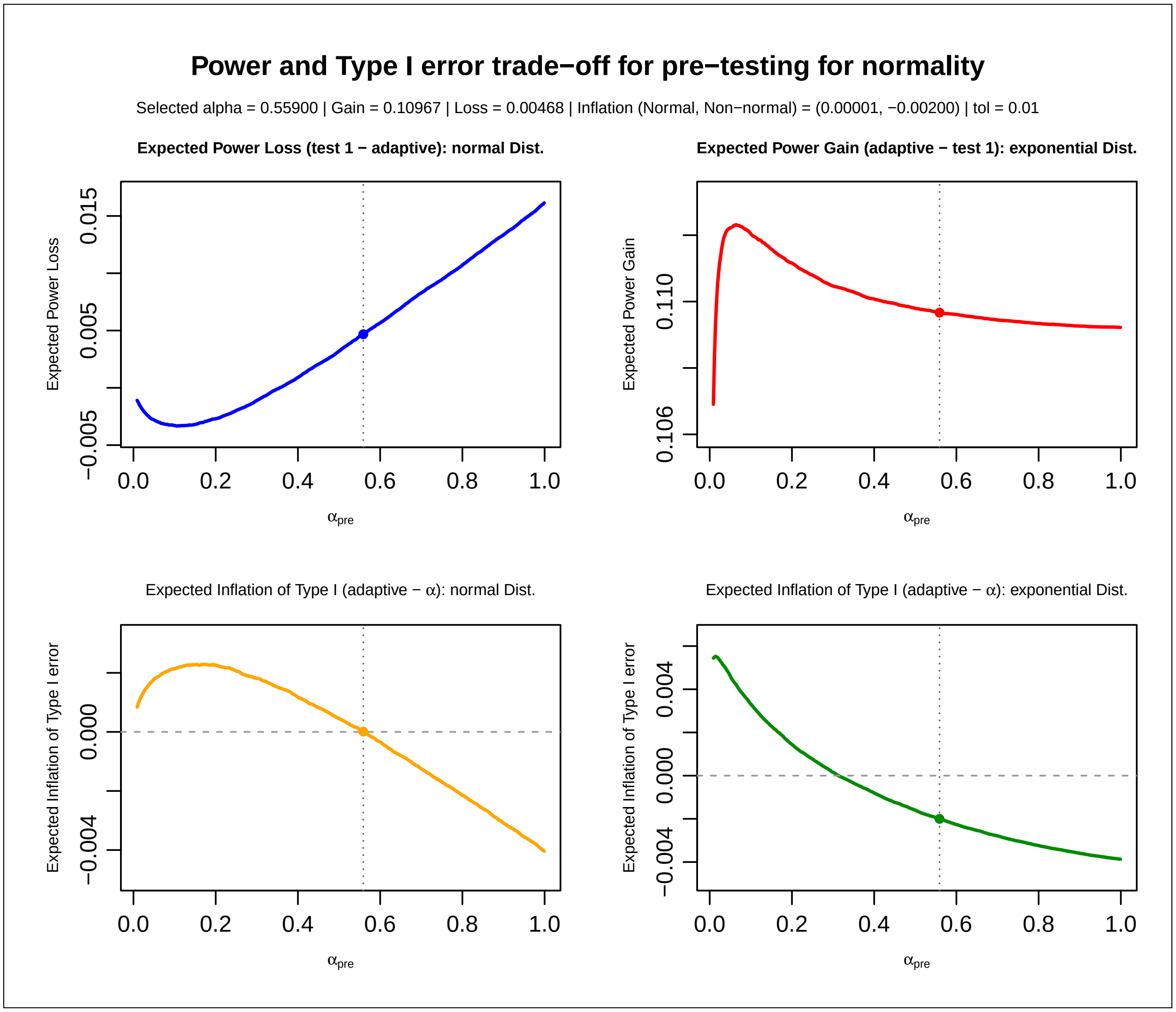


Figure 2.32: *Showing expected power loss for pretesting, left plot and the expected power gain for pre-testing, right plot in row 1. Row 2 displays expected inflation of Type I error rates for samples from normal and exponential distribution for sample of sizes 10. The optimal α_{pre} is about 0.559.*

With 0.01 tolerance, the optimal alpha for pre-testing for a sample of size 10 is about 0.559. With this pre-test alpha value, expected power loss is about 0.0045, expected power gain 0.1097, and nearly zero expected inflation in Type I error rate in both cases.

Next, for a given sample size n, and for any given distribution $\mathcal{D}$, we create a plot of power versus effect size as shown in Figure 2.33. For errors drawn from exponential distribution, the Kruskal-Wallis and adaptive test are more power, but for errors from normal distributions, the parametric ANOVA test has a slight edge compared to the Kruskal-Wallis and adaptive tests as shown in Figure 2.33 below.

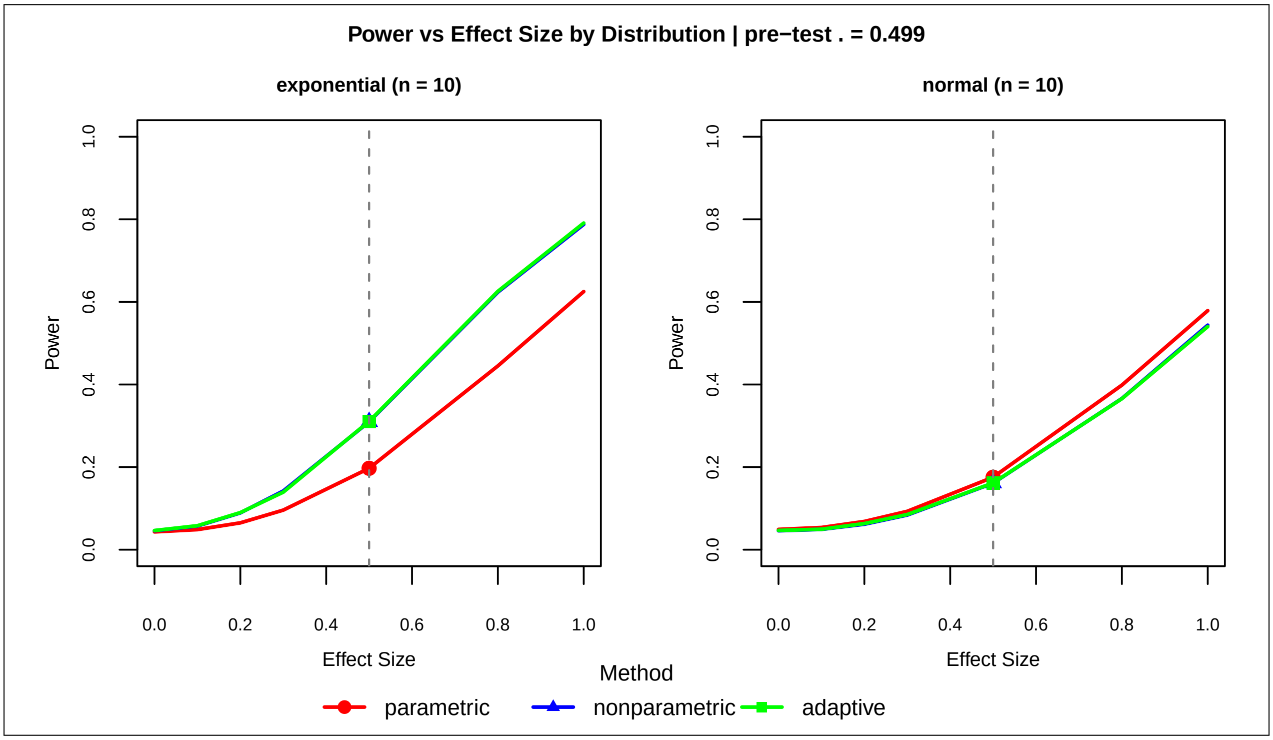


Figure 2.33: *Showing power versus effect size for three treatments of each of size 10 with model's errors drawn from normal and exponential distributions for ANOVA, Kruskal-Wallis and adaptive tests.*

To investigate the relative effects of pre-testing on the power and Type I error of a downstream procedure for a given sample size, we plot the power against the Type I error rate to obtain an ROC-Like curve. For errors drawn from exponential distribution, the Kruskal-Wallis and adaptive test are more powerful, but for errors from normal distributions, all three tests are comparable as shown in Figure 2.34 below.

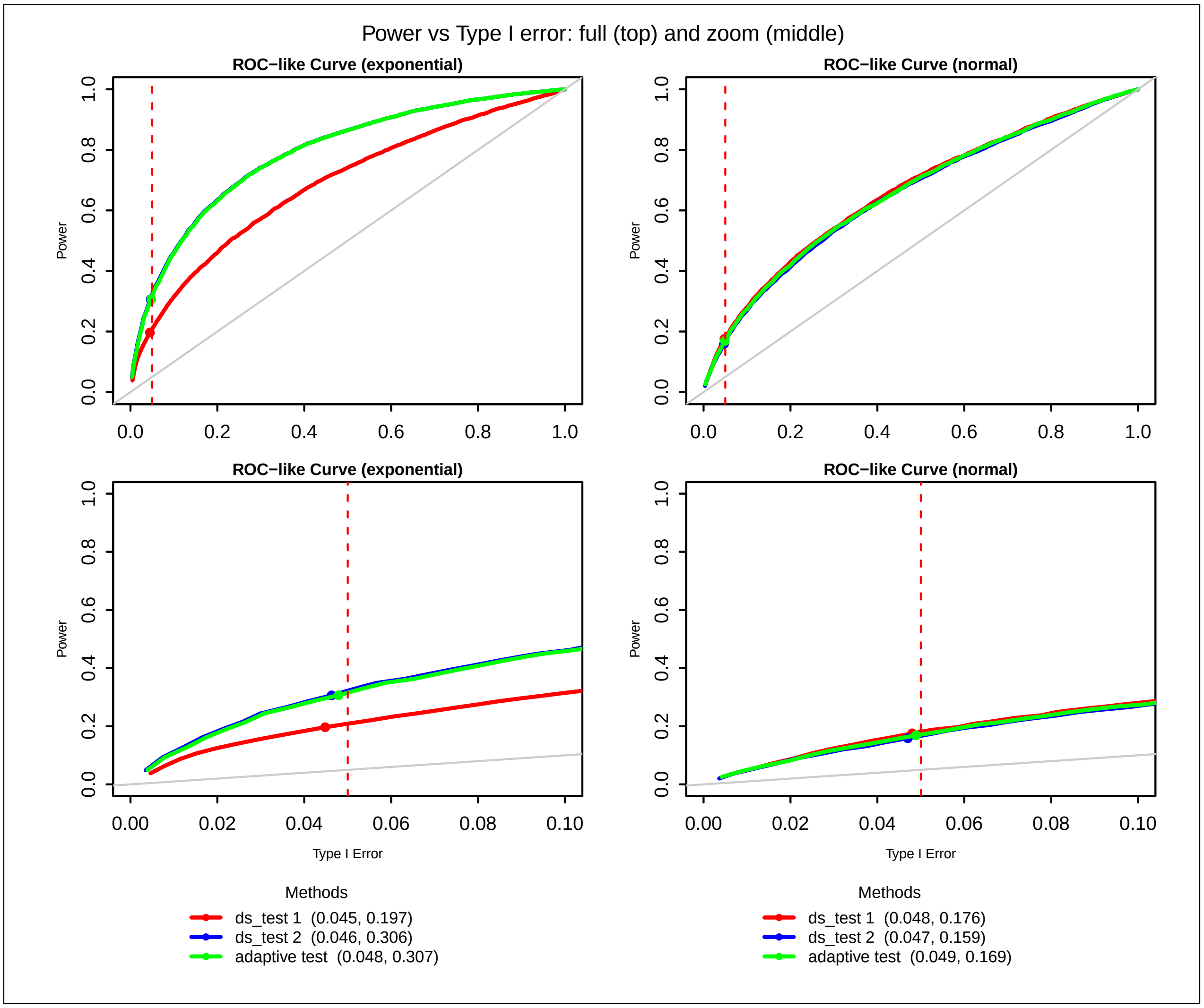


Figure 2.34: *Showing the power vs Type I error plot for the One-Way ANOVA, Kruskal-Wallis and the Adaptive tests. ds_test 1 and ds_test 2 mean test 1 and test 2 respectively.*

Next we compare the power and Type I error rates of the One-Way ANOVA, Kruskal-Wallis and the Adaptive tests across varying sample sizes, estimating the area under each curve as a measure of the overall performance of each method. Generally, Type I error rates are controlled for each test procedure for all situations. For errors drawn from exponential distribution, the Kruskal-Wallis and adaptive test are more power, but for errors from normal distributions, the parametric ANOVA have a slight edge compared to the Kruskal-Wallis and adaptive tests as shown in Figure 2.35 below.

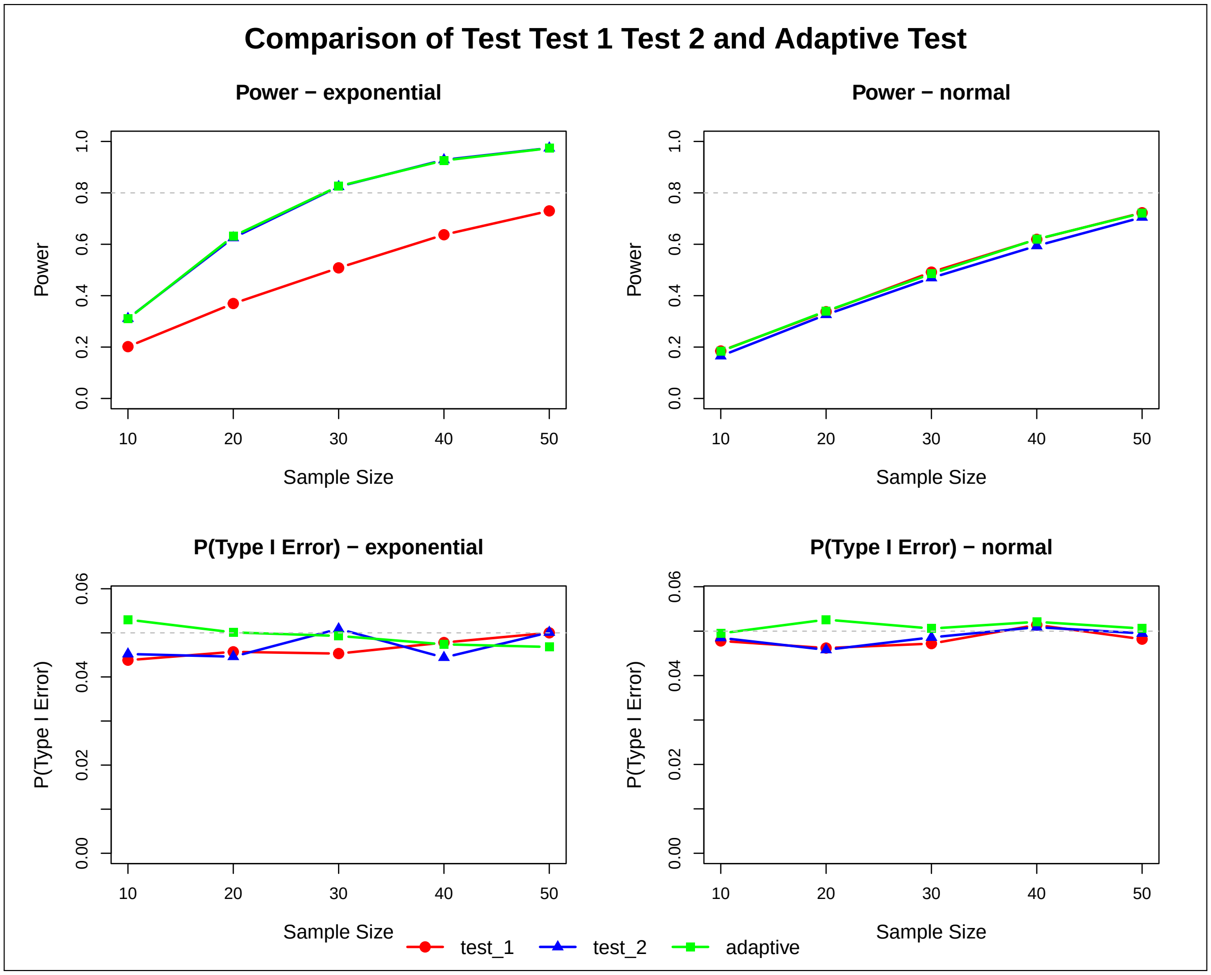


Figure 2.35: *Showing the area under the power curves and area under the Type I error curves for ANOVA, Kruskal Wallis and the adaptive test by pre-testing for errors drawn from exponential and normal distributions.*

Distribution	**Method**	**AUC-Power**	**AUC-TypeI**
Exponential			
	Parametric (ANOVA)	0.492 13	0.046 36
	Nonparametric (Kruskal Wallis)	0.756 08	0.048 28
	Adaptive	0.755 38	0.049 86
Normal			
	Parametric (ANOVA)	0.472 78	0.049 86
	Nonparametric (Kruskal Wallis)	0.452 49	0.048 60
	Adaptive	0.476 15	0.050 93

Table 2.8: *AUC for power and Type I error by method and distribution for ANOVA, Kruskal Wallis and adaptive tests.*

Observations and Recommendations: The above analysis showed that ANOVA Kruskal-Wallis and adaptive test procedures are conservative. The Kruskal Wallis and adaptive tests are consistently more powerful than the ANOVA in both cases, thus users should prefer the Kruskal Wallis test to the adaptive due to computational time and complications.

One-Way ANOVA vs Permutation ANOVA test Here, we compare the parametric One-Way ANOVA F-test (Test 1) against a permutation-based ANOVA (Test 2) that uses the F-statistic. Again, with the chosen normality test method, N_T, third moment(Skewness) we assess the expected gains and losses for pre-testing for normality at different normality alpha value, α_{pre}.

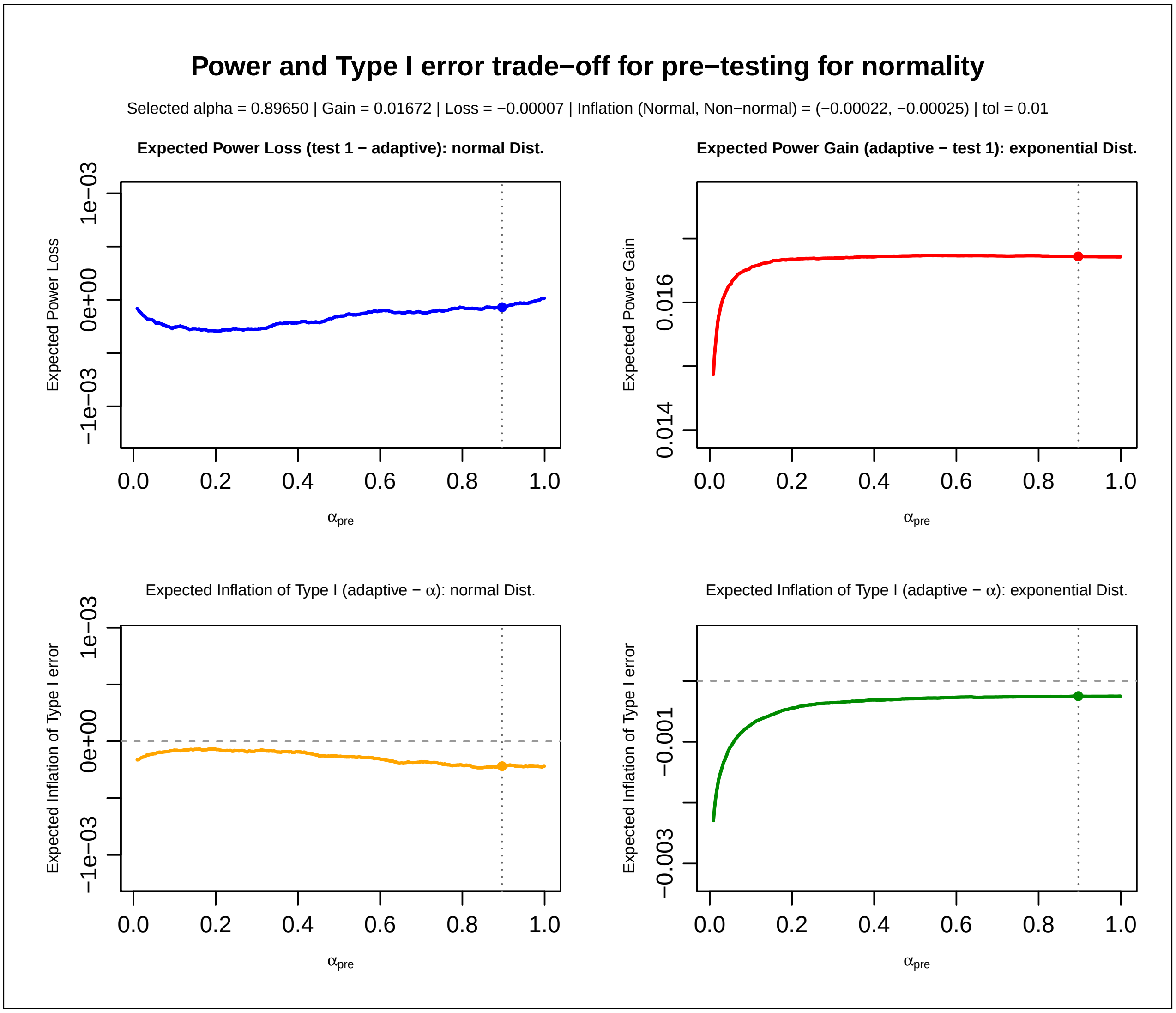


Figure 2.36: *Showing expected power loss for pretesting, left plot and the expected power gain for pre-testing, right plot in row 1. Row 2 displays expected inflation of Type I error rates for errors from normal and exponential distribution for sample of sizes 10. The optimal α_{pre} is about 0.8965.*

Type I error is barely inflated, so the algorithm focuses on the expected power loss and expected power gains to obtain the optimum alpha for pre-testing.

Next, with the optimum alpha value and for a given sample size n, and for any given distribution $\mathcal{D}$, we create a plot of power versus effect size. Power is identical for all three test procedures as shown in Figure 2.37.

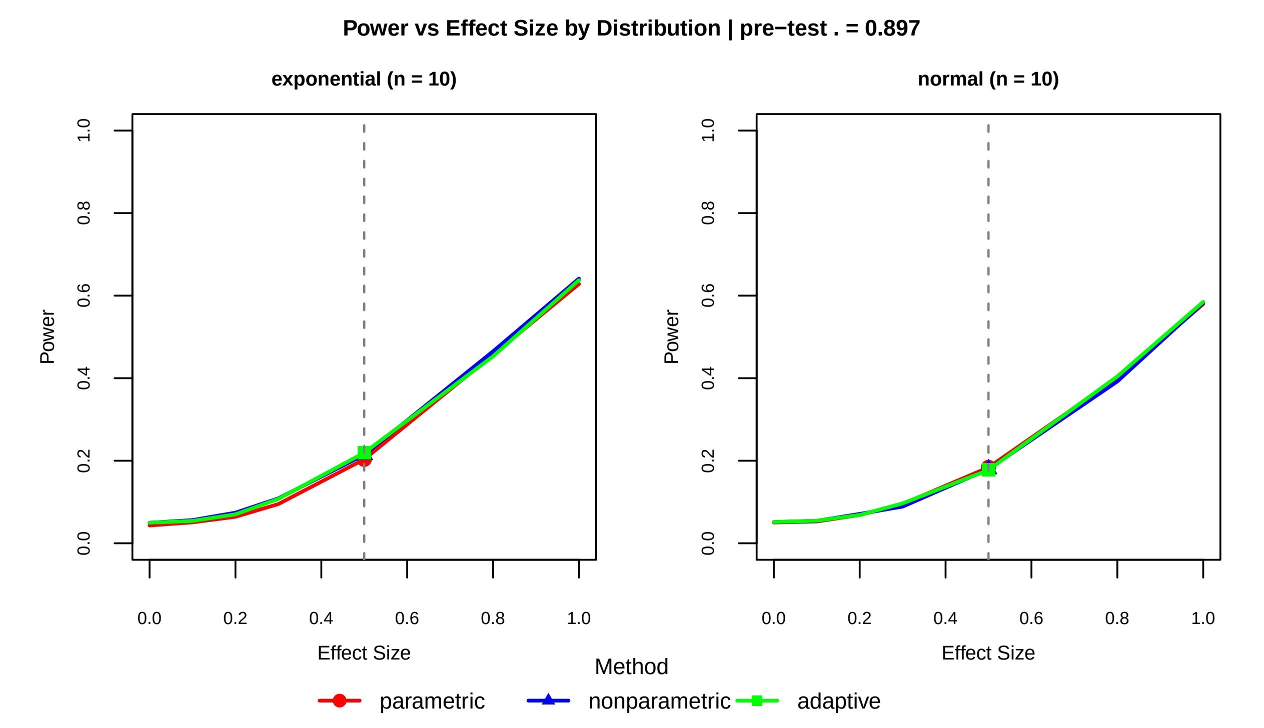


Figure 2.37: *Showing power versus effect size for errors drawn from normal and exponential distributions.*

To evaluate the performance trade-offs induced by normality pre-testing, we construct ROC-like curves relating statistical power to Type I error rates across different significance thresholds. These curves, presented in Figure 2.38, quantify how the adaptive testing procedure balances power with Type I error control for fixed sample sizes. In this case, Type I error rates are controlled at the nominal level of 0.05 for all test methods.

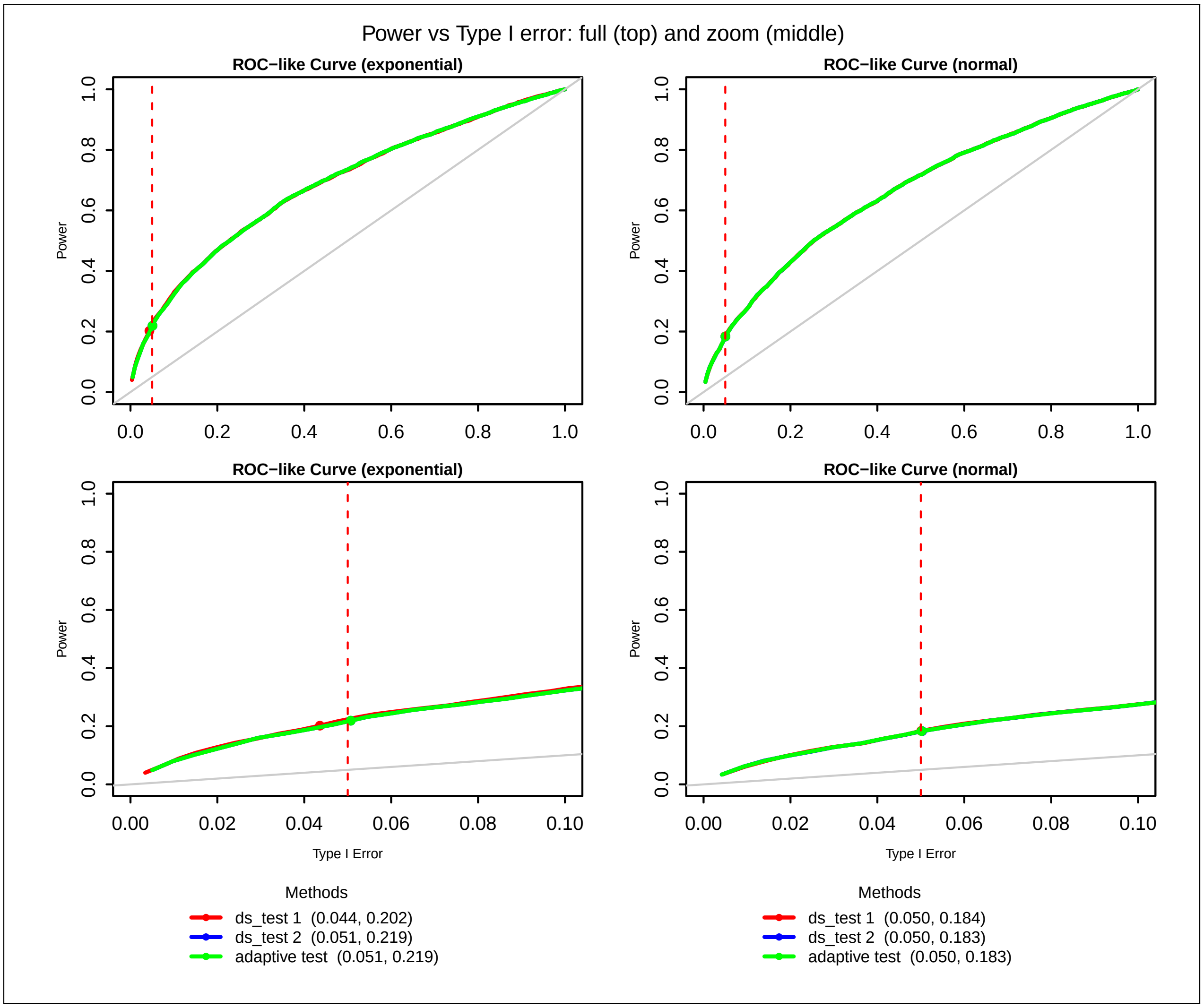


Figure 2.38: *Showing the power vs Type I error plot for the One-Way ANOVA, Permutation ANOVA, and the Adaptive tests. ds_test 1 and ds_test 2 mean test 1 and test 2 respectively.*

Next we compare the power and Type I error rates of the Onw-Way ANOVA, Permutation ANOVA and the Adaptive tests across varying sample sizes, estimating the area under each curve as a measure of the overall performance of each method as shown in Figure 2.39. Type I error rates are controlled for each test procedure and they all produce identical power as well in both cases.

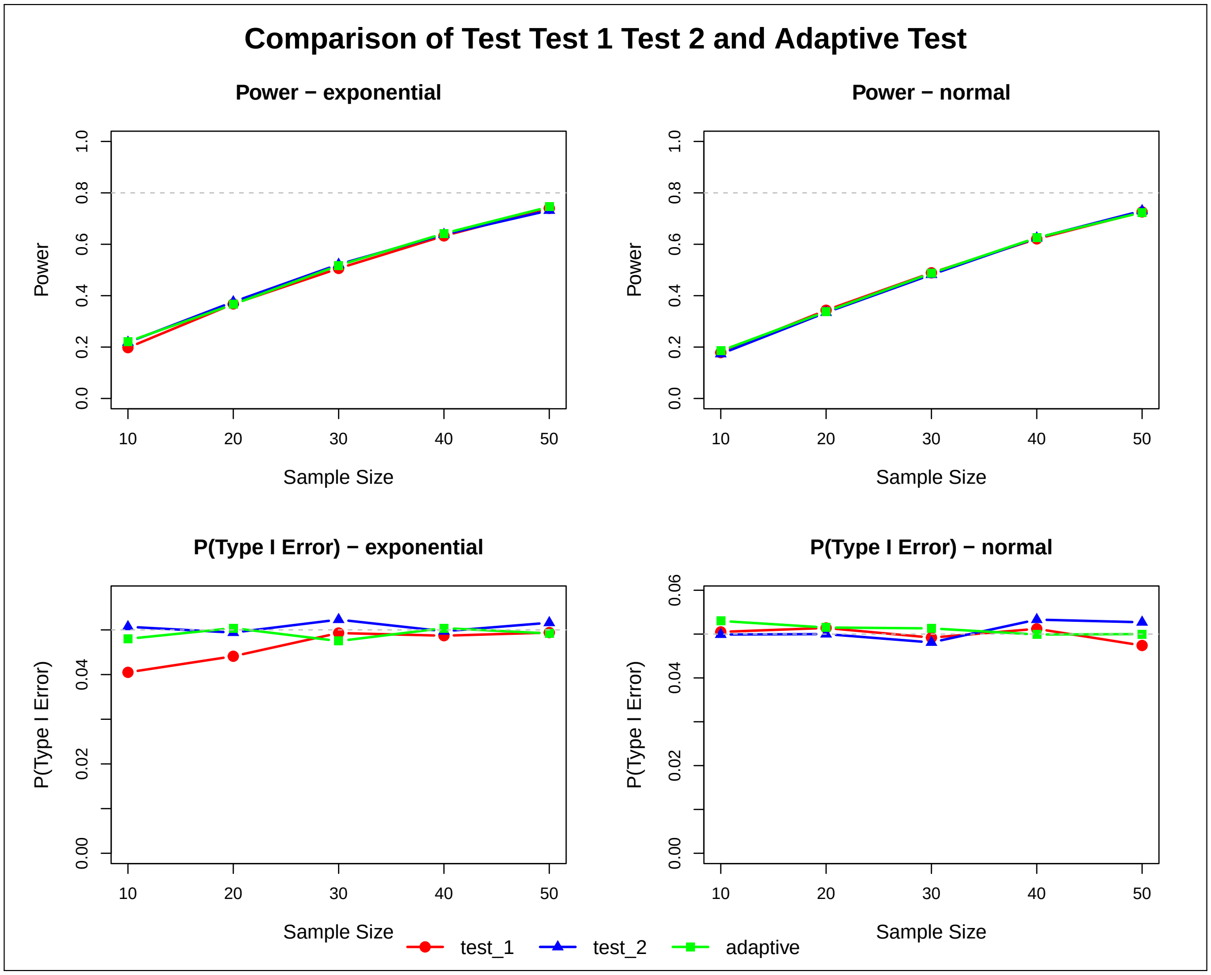


Figure 2.39: *Showing the area under the power curves and area under the Type I error curves for ANOVA, Permutation ANOVA and the adaptive test by pre-testing for samples drawn from exponential and normal distributions.*

Distribution	**Method**	**AUC-Power**	**AUC-TypeI**
Exponential			
	Parametric (ANOVA)	0.494 91	0.046 42
	Nonparametric (Permutation ANOVA)	0.506 90	0.049 92
	Adaptive	0.502 95	0.049 66
Normal			
	Parametric (ANOVA)	0.475 51	0.050 14
	Nonparametric (Permutation ANOVA)	0.474 54	0.050 45
	Adaptive	0.470 16	0.049 86

Table 2.9: *AUC for power and Type I error by method and distribution for ANOVA, Permutation ANOVA and adaptive tests.*

Observations and Recommendations: The above analysis showed that both parametric and nonparametric methods are conservative with identical power and so practitioners may consider the classical ANOVA since it requires lesser computation time than the permutation ANOVA.

2.6.12 Demonstration 8: Simple Linear Regression vs Bootstrap Regression

For Regression test, the residuals are what we require to be normally distributed, so we apply the normality test on the model's residuals, calculate the TPR and the FPR and then create the ROC as shown in figure 2.40 below.

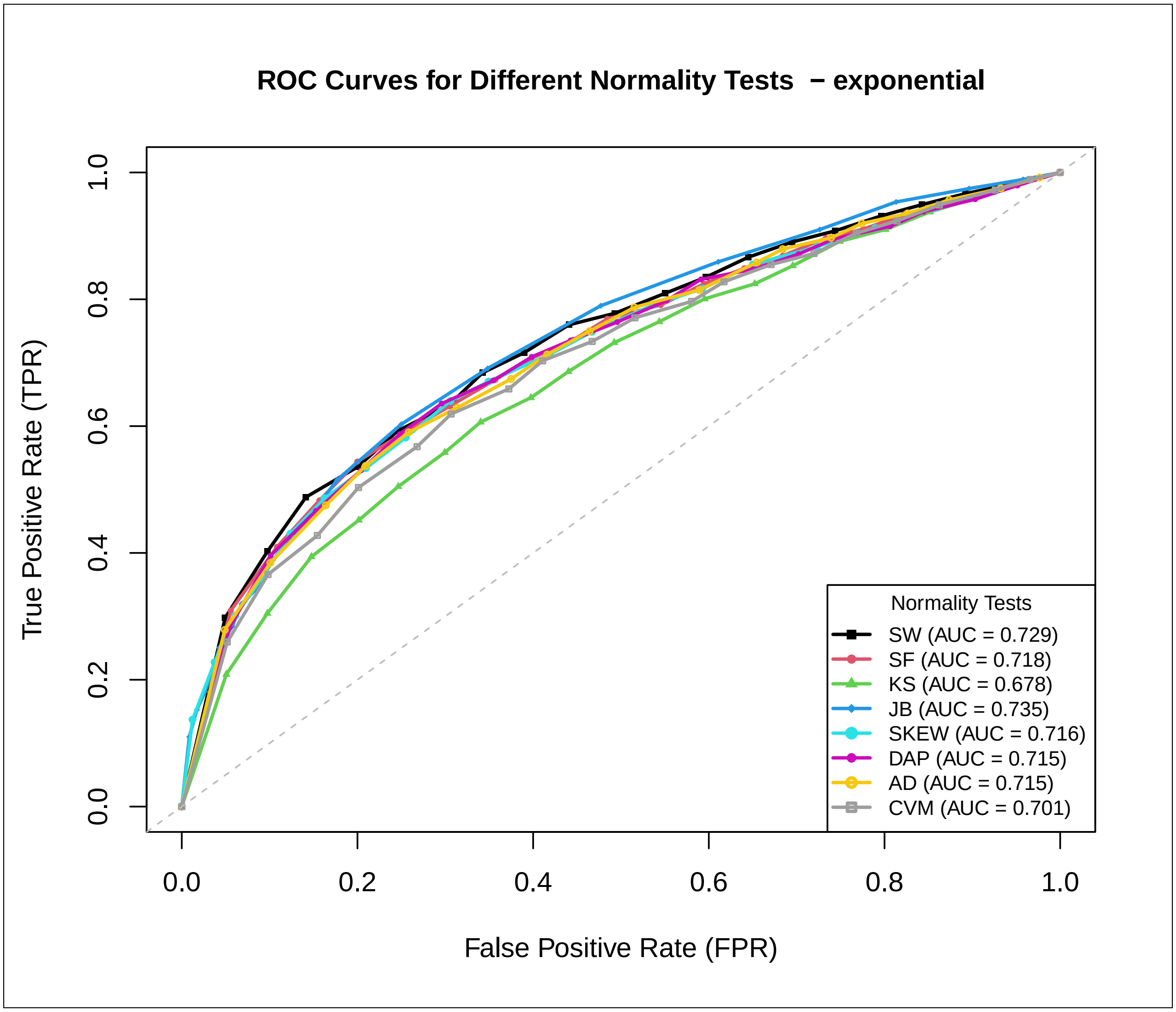


Figure 2.40: *Showing ROC curves for different normality test methods for sample size 10 drawn from normal and exponential distributions.*

We compare standard linear regression (Test 1) against bootstrap regression (Test 2). Both methods test for linear relationships, but the bootstrap approach does not require normal errors. Due to computation time, the simulations here are done for just N = 1000 iterations.

Again, with the chosen normality test method, N_T, Jarque-Bera-test we assess the expected gains and losses for pre-testing for normality at different normality alpha value, α_{pre}.

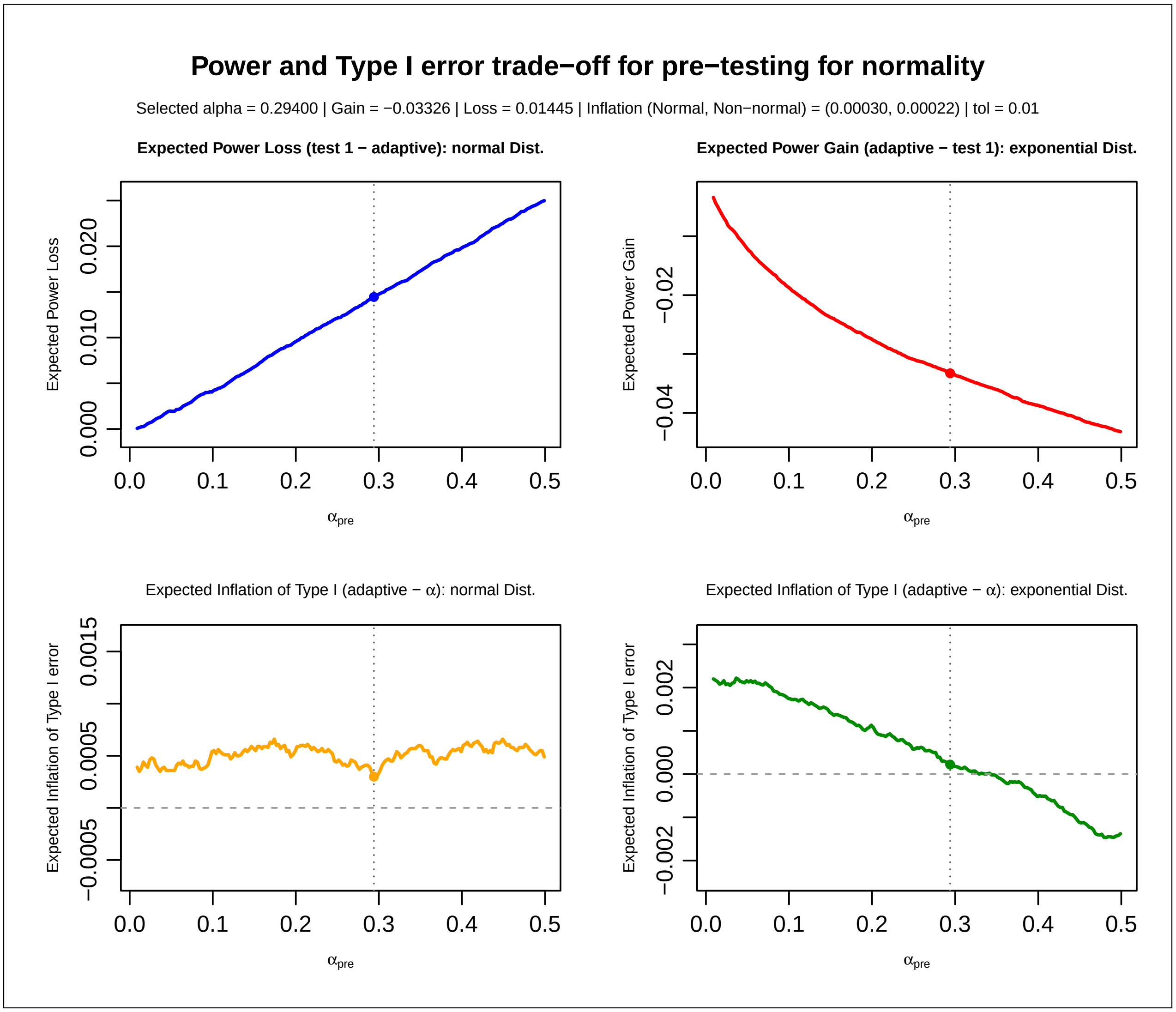


Figure 2.41: *Showing expected power loss for pretesting, left plot and the expected power gain for pre-testing, right plot in row 1. Row 2 displays expected inflation of Type I error rates for errors from normal and exponential distribution for sample of sizes 10. The optimal α_{pre} is about 0.294.*

With 0.01 tolerance, the optimal alpha for pre-testing for a sample of size 10 is about 0.294. There is no benefit for pre-testing and using the adaptive procedure in this case as we fail to control type I error rate and gain any power. The bootstrap regression as the nonparametric test here is not robust and less efficient.

Next, with the optimum alpha value and for a given sample size n, and for any given distribution

$\mathcal{D}$, we create a plot of power versus effect size. Power is identical for all three test procedures as shown in Figure 2.42.

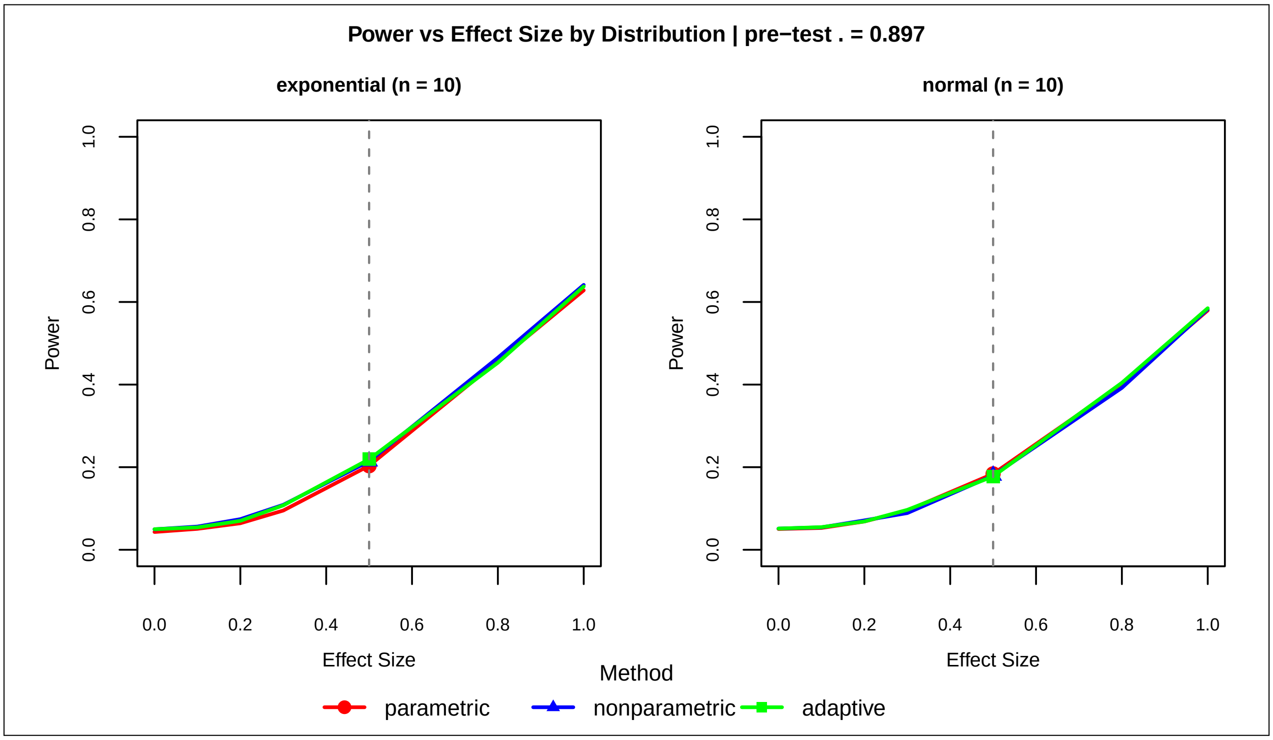


Figure 2.42: *Showing power versus effect size for errors drawn from normal and exponential distributions.*

To evaluate the performance trade-offs induced by normality pre-testing, we construct ROC-like curves relating statistical power to Type I error rates across different significance thresholds. These curves, presented in Figure 2.43, quantify how the adaptive testing procedure balances power with Type I error control for fixed sample sizes.

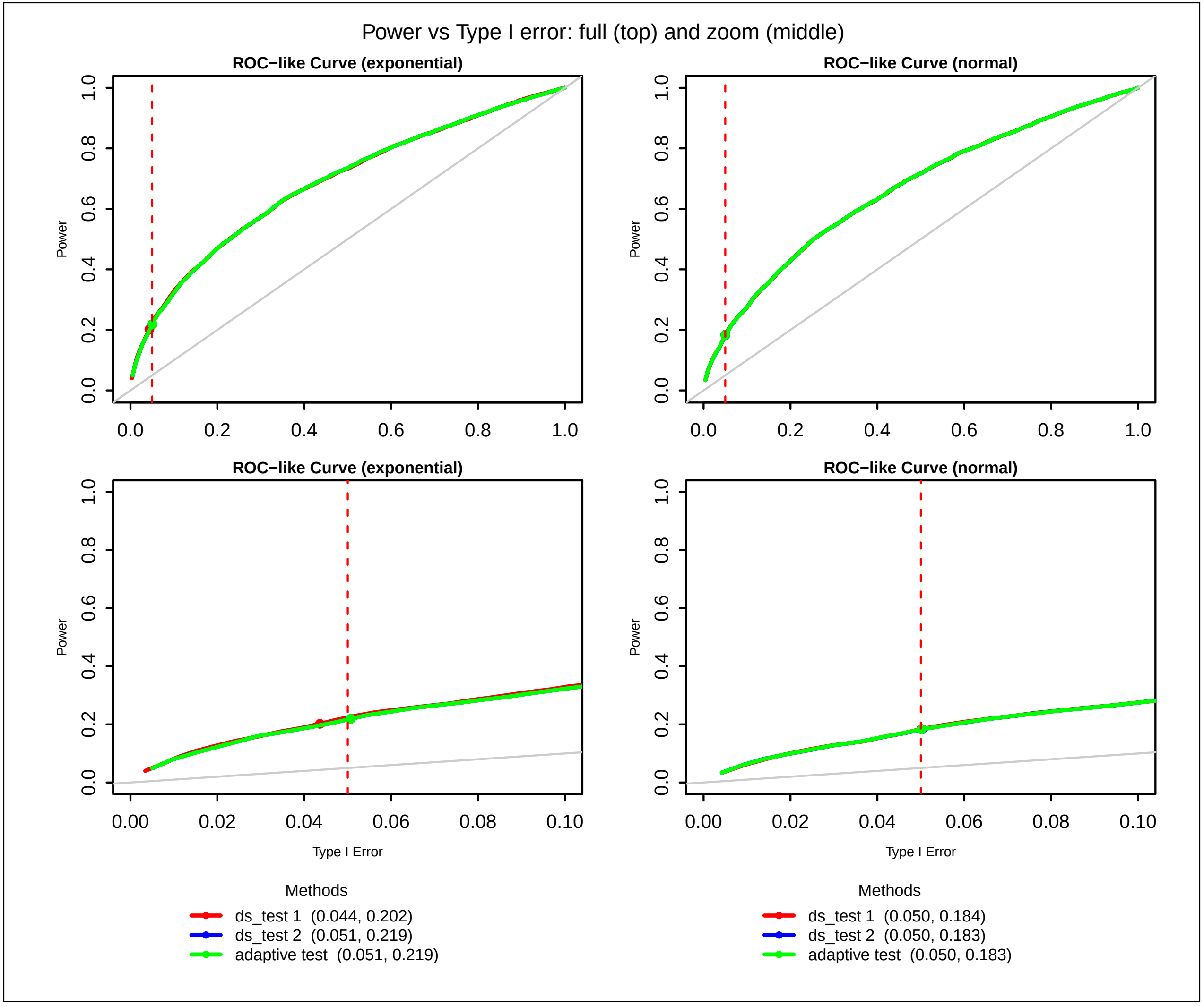


Figure 2.43: *Showing the power vs Type I error plot for the Simple linear regression, bootstrap regression, and the Adaptive tests. ds_test 1 and ds_test 2 mean test 1 and test 2 respectively.*

Next we compare the power and Type I error rates of the Simple linear regression, bootstrap regression, and the Adaptive tests across varying sample sizes, estimating the area under each curve as a measure of the overall performance of each method as shown in Figure 2.44.

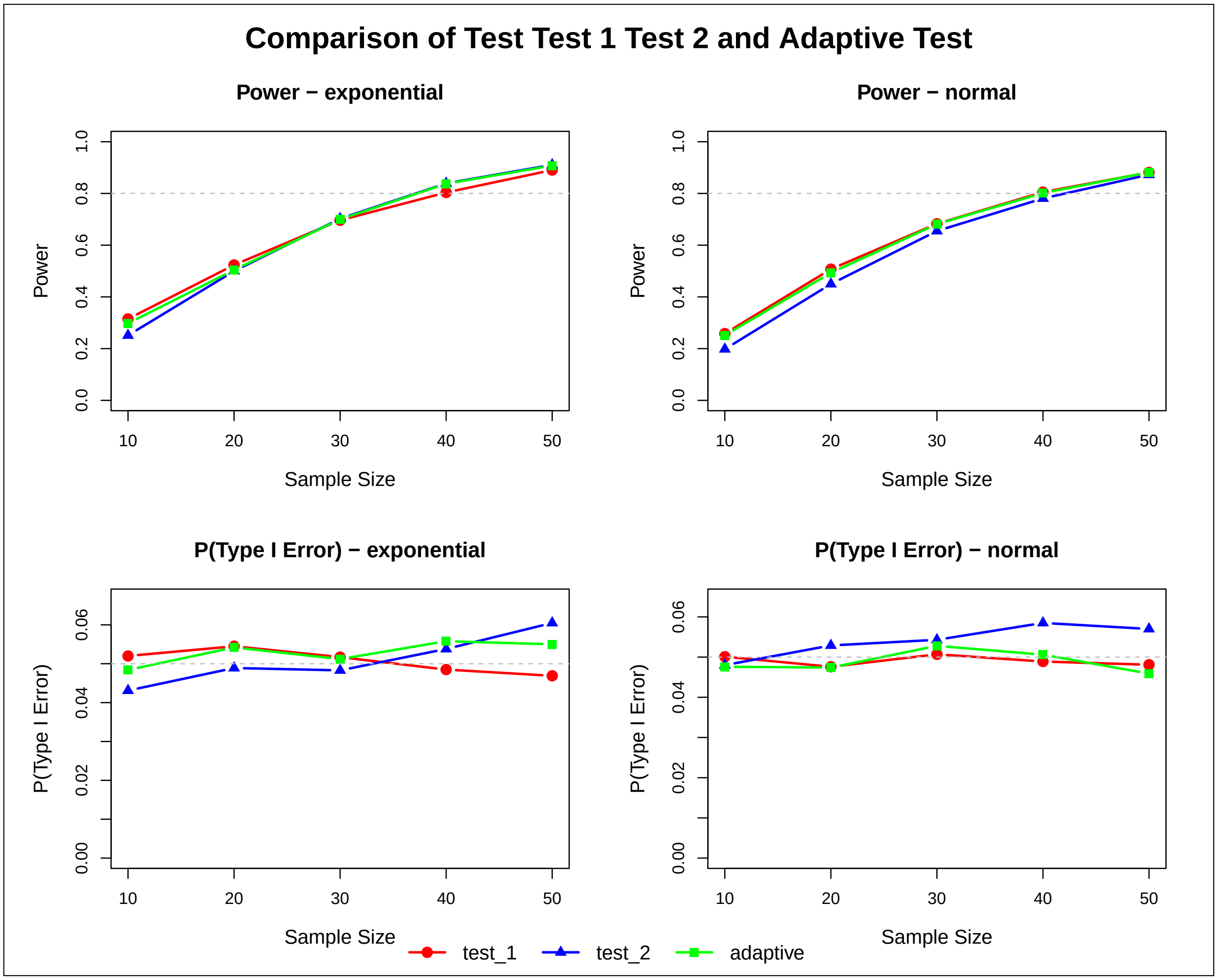


Figure 2.44: *Power and Type I error rate curves for Simple linear regression, bootstrap regression, and the Adaptive tests by pre-testing for errors drawn from exponential and normal distributions.*

Distribution	**Method**	**AUC-Power**	**AUC-TypeI**
Exponential			
	Parametric (Regression)	0.494 91	0.046 42
	Nonparametric (Bootstrap Regression)	0.506 90	0.049 92
	Adaptive	0.502 95	0.049 66
Normal			
	Parametric (Bootstrap)	0.475 51	0.050 14
	Nonparametric (Bootstrap Regression)	0.474 54	0.050 45
	Adaptive	0.470 16	0.049 86

Table 2.10: *AUC for power and Type I error by method and distribution for Simple linear regression, bootstrap regression, and the Adaptive tests*

Observations and Recommendations: The above analysis showed that the bootstrap regression is not an appropriate choice for test 2. Since the simple linear regression controls type I error rate and comparable power to the bootstrap and adaptive tests, we recommend choosing the classical simple linear regression in this situation.

2.7 Implementation of the Methods through an R-package

Our framework will be made available to the user through an R package as well as an R shiny application. It shall take user inputs such as sample size (n), number of Monte-Carlo simulations (N), effect sizes(d), significance levels of the downstream test (α), user defined functions such as test 1 and test 2 for downstream test etc and returns outputs such as normality assessment plots, power and type I error trade-off results, ROC Analysis results including the power vs type I error ROC-Like curves, power and type I error plots, area under power and type I error rates tables, etc. The first version of the shiny app is currently available for demonstrations. We will however need to update with changes to reflect exactly all that is contained in Section 2.6.

2.7.1 Key Functions in the package

`Generate_data`: Generate sample data from different probability distributions such as normal, exponential, that accepts different user specified parameters. Each sample data is centered and scaled using z-score function for uniformity.

`Generate_tests` Collection of classical normality test functions such as Shapiro-Wilk test, Anderson-Darling test etc. User may specify their own user-defined test too.

`User_defined_functions`: A collection of user defined function for the user frame work. It includes a specialized data generation function(`Gen_data`), a function to grab user specified parameters(`Get_parameters`), a function to extract the normality-test object, either the raw data or model residuals (`Fn_to_get_norm_obj`), two user defined functions for performing downstream test, test 1 & test 2(`Fn_for_ds_test_1` & `Fn_for_ds_test_2`)

`Normality_test` A universal normality test to pick the user morality test object to perform normality checks on the normality object using a user–specified method. Allows user to assess normality on data in different form and many different dimensions.

`Fn_for_roc_curve_for_norm_test` Compute ROC objects (TPR & FPR)for different classical normality tests methods for samples from normal distribution and any given alternative distribution for a given a sample size.

Generate_pval Generate p-values for normality pre-test and downstream tests methods.

Perform_analysis Compute Type I error and power for each method for tradeoff analysis.

Plot_power_error_tradeoff Create ROC-like curves and obtain the optimal pre-test α. Systematically balances tradeoff for pre-testing for normality in terms of expected power loss, power gains, and expected inflation of Type I error rates in a form of a grid search to obtain the optimal alpha for normality pre-testing.

Ds_test_function Generic downstream test wrapper (user supplies the test).

Perform_ds_func Run downstream tests; compute Type I error, power, and AUC for test 1, test 2, and the adaptive test.

Run_simulation runs the entire simulation pipeline.

2.8 Assessment

To be completed later - We will assess the performance of the framework in its ability to guide the user to the right direction. Some simulation-based assessment is currently being done, which we will add later.

2.9 Applications

We have currently included detailed demonstrations of our framework using several use cases, but more results on detailed applications of our framework for common use cases is pending and currently under progress. Note that so far we have chosen one arbitrary distributions for each example, used a limited number of specific normality test methods, arbitrarily chosen effect sizes, and so on. Some of the possible options include different alternative distributions, different normality test methods, etc., so that we can provide a general guideline to the statistical community regarding when to use normality pre-test and when not to use them for common use cases.

2.10 Discussion

This study provides a comprehensive framework for evaluating the practical utility of normality testing in statistical workflows. By moving beyond traditional power analyses and Type I error assessments of the normality tests themselves, we have developed a methodology that quantifies the real-world consequences of normality pre-testing decisions. Our findings offer both methodological contributions and practical insights for researchers navigating the complex landscape of statistical assumption checking.

2.10.1 Our Contributions

Our primary contribution lies in formalizing the concept of *expected power loss* and *Type I error inflation* as metrics for evaluating normality test utility. Rather than asking whether a normality test can detect non-normality which is the focus of most existing literature, we address the more relevant question: *Does using a normality test improve or degrade subsequent statistical inference?* This shift in perspective acknowledges that the ultimate goal of normality testing is not detection per se, but rather the improvement of downstream analytical decisions.

The conditional framework we developed reveals several nuanced insights. First, the cost of misapplying parametric tests when normality is violated is not symmetric with the cost of unnecessarily using non-parametric methods. Our simulations demonstrate that, in some cases, the former typically incurs greater inferential consequences, particularly in small samples where normality violations most severely impact parametric test performance. Second, the utility of normality testing exhibits strong dependency on sample size, the underlying distribution's departure from normality, and the choice of the downstream methods, creating a context-dependent judgment that researchers have to always make. Our framework enables the user to make this judgement for their particular use case.

2.10.2 Limitations and Challenges

Several limitations warrant consideration when interpreting our results. First, our framework assumes that researchers will follow the adaptive procedure consistently—using parametric methods when normality is not rejected and non-parametric methods otherwise. In practice, researchers may engage in additional data-dependent decisions or multiple testing that further complicate the selective inference problem.

A crucial assumption in our framework is the knowledge of the alternative distribution from which the data may be coming from. However, this is rarely known in practice. If the user has prior subject-matter knowledge to specify plausible alternative distributions, we could conduct evaluations across a range of the candidate distributions, or if the user has some pilot data, which sometimes happen in most experimental studies, we could sample from those to derive the empirical distribution of the alternative distribution and use that to assess our method.

Also, our criteria such as expected power loss and type I error inflation are estimated via Monte Carlo simulations. Though these estimates achieve stability for sufficiently large simulation sizes they should be consistent estimators of the true parameters. However there remain approximations and thus bears sampling errors that may limit exact inference. Nevertheless, our criteria serves as a reliable practical guide.

Finally, our analysis focuses primarily on hypothesis testing contexts. The impact of normality pre-testing on estimation precision, confidence interval coverage, and predictive modeling requires separate investigation. The selective inference consequences may differ substantially in these contexts, particularly for biased estimators or procedures with non-uniform error control.

2.10.3 Potential Future Research Directions

Although the demonstration of the framework is for evaluating normality tests, it can easily be adapted for any scenario involving pre-testing for any model assumptions such as constant variance, etc.

This work opens several promising avenues for future investigation. Most immediately, extending the conditional framework to Bayesian methods and robust statistical procedures could reveal alternative approaches to handling distributional uncertainty.

Developing practical decision rules based on our theoretical framework represents another important direction. Could we create simple heuristics or software tools that recommend when normality testing is likely to be beneficial based on sample size, effect size expectations, and research context? Such practical guidance would bridge the gap between methodological theory and daily statistical practice.

2.10.4 Concluding Remarks

Normality testing is controversial: widely taught and frequently used, yet its practical utility is very contentious. Our work suggests that this paradox stems from asking the wrong question. Instead of debating whether normality tests are ”good” or ”bad,” we should ask *when* and *how* they contribute to better scientific inference. The answer appears to be context-dependent and nuanced, reflecting the complex trade-offs inherent in statistical decision-making.

By providing a formal framework for evaluating these trade-offs, we hope to move the discussion beyond ideological debates toward evidence-based methodological choices. The ultimate utility of normality testing lies not in its ability to detect deviations from normality, but in its ability to guide researchers toward more accurate and reliable statistical conclusions.

Chapter 3

Machine Learning Approach for Normality Classification

3.1 Introduction

In recent years, machine learning (ML) has emerged as a promising alternative for many statistical analysis procedure. It is possible to use machine learning approaches in the context of normality tests by classifying the data generating distribution into 'normal' and 'non-normal'. Unlike traditional tests that rely on a single test statistic, ML classifiers can combine a number of distributional features–capturing shape, spread, entropy, and complexity into a unified, nonlinear model for classification. This has the potential to decrease mis-classification rates which may improve the performance of the downstream procedures.

The objective of this pre-testing step is not to formally control the Type-I error rate for the normality hypothesis itself, but to optimize the overall performance of the subsequent statistical procedure. In this context, a "misclassifications" is any decision that leads to suboptimal performance downstream. Therefore, the classifier should be evaluated and tuned based on its ability to maximize the final outcome, say estimation accuracy, or predictive power, rather than its adherence to a specific significance level for the pre-test. This pragmatic focus on the ultimate goal makes a strict hypothesis test for normality unnecessary.

3.2 Literature review

Pioneering work in this area was conducted by Wilson and Engel (1990), who employed an artificial neural network (ANN) trained on features like skewness and kurtosis. The model achieved performance comparable to the Lilliefors and Shapiro-Wilk tests but was limited to a fixed sample size of 30 and only seven non-normal distributions, raising concerns about its generalizability. Sigut et al. (2006) expanded on this by incorporating a broader set of input features, including the Shapiro-Wilk test statistic and measures from Vasicek (1976). Despite these improvements, the model's robustness to unseen distributions remained unclear.

A more recent advancement comes from Simić (2021), who used a Descriptor-Based Neural Network (DBNN) with an extended feature set, including measures of central tendency, dispersion, and quantiles. The model reportedly achieved high accuracy. However, its generalizability is also subject to contention, as the training and test sets appear to have been generated from similar families of distributions. Reported instances of perfect classification accuracy also suggest potential overfitting or data leakage, highlighting the ongoing challenges in applying ML to this domain.

These studies illustrate both the potential and the challenges of ML for normality testing. The primary advantage lies in its ability to learn complex, multi-feature patterns that no single traditional test has proven to be able to capture so far. The key challenges are ensuring model robustness, generalizability to truly unseen data, and avoiding overfitting. This dissertation aims to address these gaps by training and validating a suite of modern ML algorithms on a vast and diverse synthetic dataset, using a rigorous cross-validation framework that strictly separates the distributions used for training and testing. Furthermore, it will employ sensitivity analysis to interpret the models and rank feature importance, providing insights into the most telling characteristics of a distribution.

3.3 Methodology

To establish a robust foundation for this investigation, this section details the systematic methodology employed. The process is designed to leverage predictive analytics for classifying univariate data distributions as Normal or Non-Normal. The methodology encompasses several key stages: data generation and feature extraction, data preprocessing, predictive model development, model validation, sensitivity and performance analysis, and model deployment through R Shiny app. A high-level overview of this process is depicted in Figure 3.1. The following tentative machine learning models are considered in this study: Logistic Regression(LR), Random forest(RF), Artificial Neural Network(ANN), Gradient Boosting Model(GBM), Support Vector Machines(SVM), and K-Nearest Neighbor(KNN).

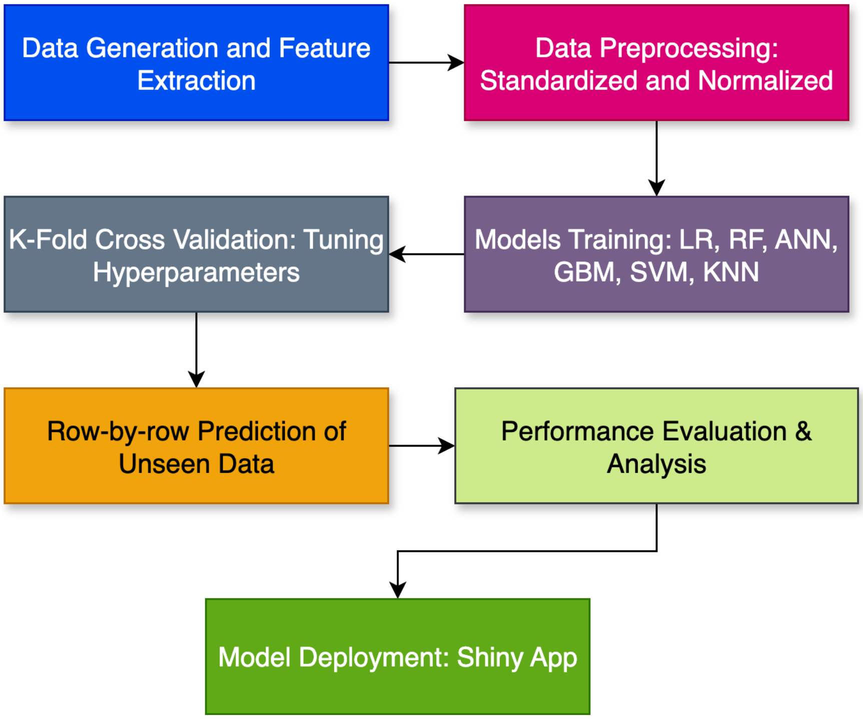

Data Generation and Feature Extraction
Data Preprocessing: Standardized and Normalized
Models Training: LR, RF, ANN, GBM, SVM, KNN
K-Fold Cross Validation: Tuning Hyperparameters
Row-by-row Prediction of Unseen Data
Performance Evaluation & Analysis
Model Deployment: Shiny App

Figure 3.1: A graphical overview of the research methodology.

3.3.1 Data Generation and Feature Extraction

To build a generalizable classifier for distribution types, we generated a large and diverse synthetic dataset through systematic simulation. For the **Normal** class, data was generated from a standard normal distribution and five normal distributions with increasing variance and centers to ensure the model learned the essence of normality beyond a single parameterization. For the **Non-Normal** class, data was generated from ten distinct distributions. This wide variety ensures the model is exposed to different forms of data characteristics. We obtain 1000 non-normal samples and 1000 normal samples to constitute a balanced set of classes for the training set. Data were then generated from completely different set of distributions from that of the training set to test the models as shown in Table 3.1

Training Set Distributions		
normal(0,1)	normal(5,2)	normal(100,25)
LogNormal(0,1)	normal(30,5)	normal(15,8)
f(6,15)	Chi-Square(7)	exponential(3)
Weibull(2,1)	Pareto(1,2)	Cauchy(0,1)
Uniform(0,1)	t(3)	Gamma(2,1)

Table 3.1: Training set distributions used to generate features for model fitting.

Feature Extraction: From each generated sample vector of n data points, a comprehensive set of 23 statistical features was calculated. These features were designed to capture various aspects of the sample's distribution. This process transformed each raw sample into a feature vector, creating a structured dataset suitable for machine learning. The complete list of extracted features is presented in Table A.1.

3.3.2 Data Preprocessing

Prior to model training, all features underwent a two-step preprocessing pipeline to enhance algorithmic stability and performance.

1. **Standardization:** Features were transformed to have zero mean and unit variance using Z-score normalization. This addressed scale differences between statistical measures (e.g., Variance vs. Skewness).
2. **Normalization:** Standardized features were rescaled to the [0, 1] interval using min-max transformation. This step proved particularly important for distance-based methods and neural networks sensitive to input magnitude.

The parameters for these transformations (means, standard deviations, mins, and maxs) were calculated exclusively on the training set within each cross-validation fold to prevent data leakage.

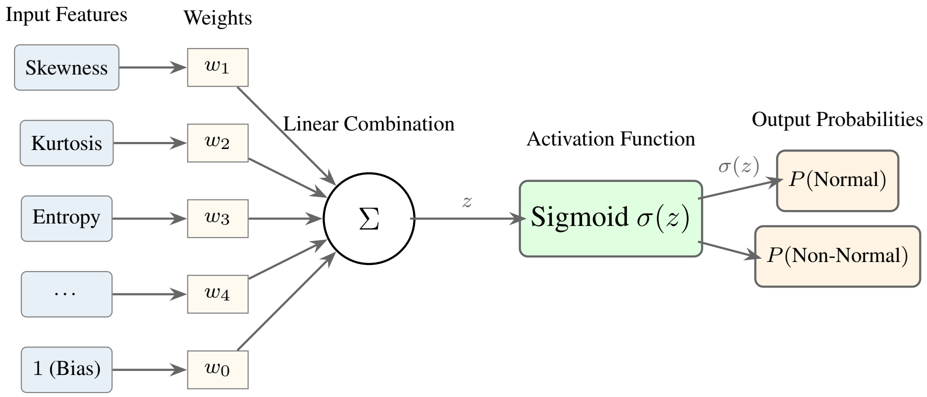
3.3.3 Prediction Methods

We treat the problem of testing for normality as a supervised binary classification task. For each sample, we use a feature vector with skewness, kurtosis, and entropy to predict whether the data comes from a normal distribution or not. We picked six different learning algorithms so there's a mix of linear and nonlinear approaches, and so we can balance easy interpretation with strong predictive ability (Hastie et al., 2009). The features were strategically selected to capture various characteristics of the distribution critical for normality assessment. Skewness and kurtosis quantify asymmetry and tail behavior, forming the foundation of classical tests like the D'Agostino–Pearson test (d'Agostino, 1971). Tail-sensitive departures are detected by tests such as Anderson–Darling, which emphasize discrepancies in the distribution extremes (An, 1933). Entropy serves as an information-theoretic measure reflecting the disorder in data; since the normal distribution maximizes entropy for a fixed variance, this feature is valuable for identifying deviations from normality (Vasicek, 1976). Together, these features provide a comprehensive and theoretically grounded basis for detecting departures from normality.

To ensure fair comparisons, all models were trained with the same cross-validation protocol.

3.3.4 Logistic Regression for Normality Classification

Logistic regression provides an interpretable baseline for classifying samples as normal or non-normal (Hosmer Jr et al., 2013). The model estimates the probability of the positive class using a linear predictor transformed by the logistic function, following the foundational formulation in Berkson (1944). Figure 3.2 shows how this method models the classification probabilities.



$$z = \sum_{j=1}^{n} w_j x_j + w_0 \quad \text{where} \quad P(\text{Non-Normal}) = \sigma(z) = \frac{1}{1+e^{-z}} > c \quad \text{and} \quad P(\text{Normal}) = 1 - \sigma(z) = \frac{e^{-z}}{1+e^{-z}} \leq c$$

Figure 3.2: *Architecture of the logistic regression model. The model computes a weighted sum of input features, applies the logistic sigmoid function, and produces probability estimates for binary classification.* c *is the decision boundary.*

3.3.5 Artificial Neural Networks for Normality Classification

Artificial neural networks (ANNs) offer a flexible, nonparametric approach to normality classification by learning complex relationships between distributional features (Goodfellow et al., 2016). This study implements a feedforward network with one hidden layer, consistent with theoretical work showing such architectures can approximate continuous functions (Hornik et al., 1989). The model was trained using the `nnet` package within the `caret` framework in R (Ripley, 2007; Kuhn, 2008).

The network processes statistical features through a hidden layer with ReLU activation (Nair and Hinton, 2010):

$$a_j^h = \max\left(0, \sum_{i=1}^{n} w_{ij}^{ih} x_i + b_j^h\right) \tag{3.1}$$

where w_{ij}^{ih} is the weight from the i-th input unit to the j-th hidden unit, and b_j^h is the bias for the

j-th hidden unit. The output layer then applies a softmax transformation to produce classification probabilities for the two classes (normal vs. non-normal). The output y_k for class k is given by:

$$y_k = \frac{\exp\left(\sum_{j=1}^{m} w_{jk}^{ho} a_j^h + b_k^o\right)}{\sum_{c=1}^{2} \exp\left(\sum_{j=1}^{m} w_{jc}^{ho} a_j^h + b_c^o\right)} \tag{3.2}$$

Model parameters were optimized by minimizing categorical cross-entropy loss using backpropagation (Rumelhart et al., 1986).

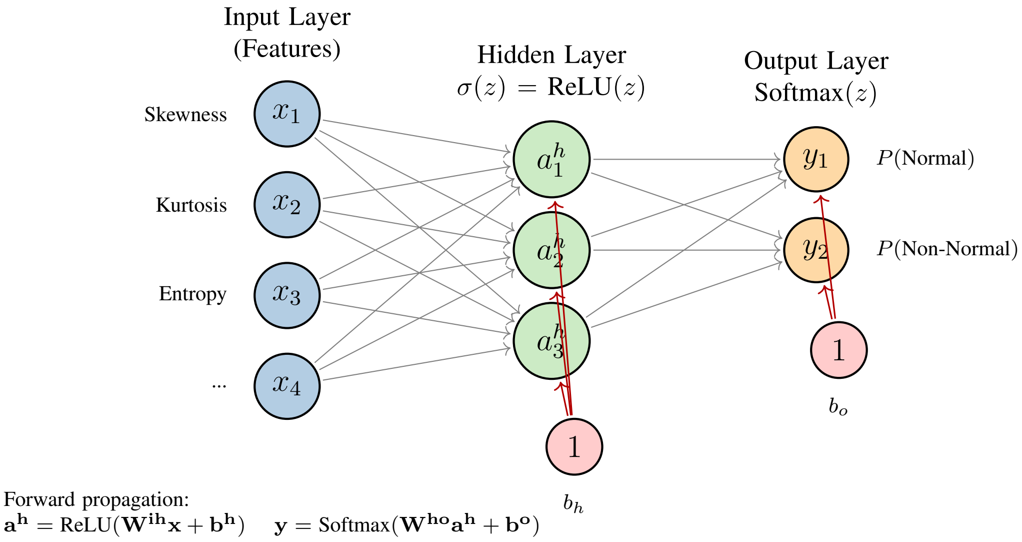


Figure 3.3: *Architecture of the multi-layer perceptron (MLP) neural network model. The network takes statistical features as input and produces classification probabilities for "Normal" and "Non-Normal" distributions through one hidden layer with ReLU activation.* b_h *and* b_0 *denote the bias units for the hidden layer and output layer, respectively.* b_h *and* b_o *are the bias vectors for the hidden layer the output layer respectively. Each output neuron has its own bias term.*

3.3.6 Random Forest for Normality Classification

Random Forest (RF) provides an ensemble approach to normality classification by combining multiple decision trees (Breiman, 2001). The method enhances prediction accuracy and controls overfitting through bootstrap aggregation and random feature selection (Ho, 1995). Our implementation uses the `randomForest` package in R (Liaw et al., 2002), with model training, parameter tuning, and performance assessment managed through the `caret` framework (Kuhn, 2008).

The algorithm constructs T decision trees , each trained on a bootstrap sample from the original data (Breiman, 2001). At each node split, a random subset of m features also known as "mtry" in the Random Forest literature is considered, where m is tuned from $\{2, 5, 10, 15\}$ via 10-fold cross-validation (Hastie et al., 2009). The optimal split maximizes the reduction in Gini impurity (Nembrini et al., 2018):

$$\Delta I(t) = I(t) - \frac{N_L}{N_t} I(t_L) - \frac{N_R}{N_t} I(t_R) \tag{3.3}$$

where $I(t)$

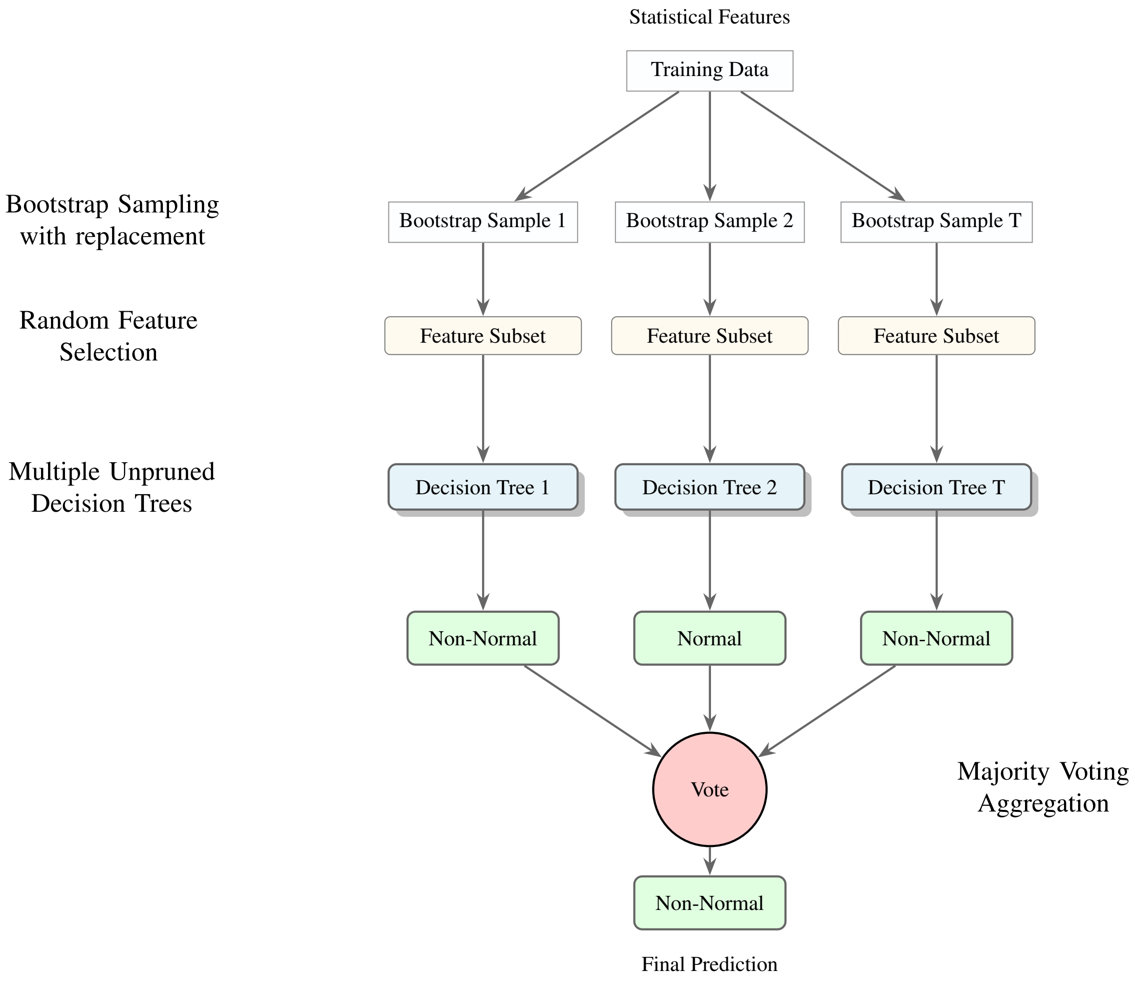
$$I(t) = 1 - \sum_{k=1}^{2} p_k^2(t) \tag{3.4}$$

measures node impurity with $p_k(t)$ denoting the proportion of class $k \in \{1, 2\}$ observations at node t, N_t the total number of samples at node t, and N_L and N_R representing the number of samples in the left and right branch nodes respectively (Nembrini et al., 2018).

Final predictions aggregate outputs from all T trees via majority voting:

$$\hat{P}(y = k|\mathbf{x}) = \frac{1}{T} \sum_{b=1}^{T} \mathbb{I}\big(\hat{y}_b(\mathbf{x}) = k\big), \tag{3.5}$$

where $\hat{y}_b(\mathbf{x})$ is the class predicted by the b^{th} tree for feature vector $\mathbf{x}$, and $\mathbb{I}(\cdot)$ is the indicator function. This ensemble approach leverages tree diversity to improve generalization while maintaining robust performance across different distributional characteristics (Breiman, 2001).



Training Configuration:
- 10-fold cross-validation for mtry parameter tuning
- Grid search: mtry $\in \{2, 5, 10, 15\}$
- Number of trees (ntree): T
- Variable importance: Enabled (Mean Decrease Gini)
- Bootstrap sampling with replacement
- Unpruned trees grown to maximum depth

Figure 3.4: Architecture of the Random Forest ensemble model with T trees. The model uses bootstrap sampling and random feature selection (mtry) to build diverse decision trees, with final predictions determined by majority voting. The optimal mtry parameter is determined through 10-fold cross-validation.

3.3.7 Extreme Gradient Boosting for Normality Classification

Extreme Gradient Boosting (XGBoost) implements a sequential ensemble method where each new decision tree corrects errors made by previous models (Chen and Guestrin, 2016). The algorithm builds upon gradient boosting principles (Friedman, 2001) while incorporating regularization to enhance generalization.

XGBoost optimizes a regularized objective function at each iteration k when adding tree $f_k(\mathbf{x})$:

$$\mathcal{L}^{(k)} = \sum_{i=1}^{n} l\left(y_i, F_{k-1}(\mathbf{x}_i) + f_k(\mathbf{x}_i)\right) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3.6}$$

where T represents the number of leaves, w_j are leaf weights, and γ, λ control regularization strength (Chen and Guestrin, 2016).

The ensemble prediction updates incrementally by:

$$F_k(\mathbf{x}) = F_{k-1}(\mathbf{x}) + \eta f_k(\mathbf{x}) \tag{3.7}$$

where η denotes the learning rate that controls each tree's contribution. This iterative refinement process enables XGBoost to capture complex patterns in distributional features while maintaining robustness through regularization (Chen and Guestrin, 2016; Friedman, 2001).

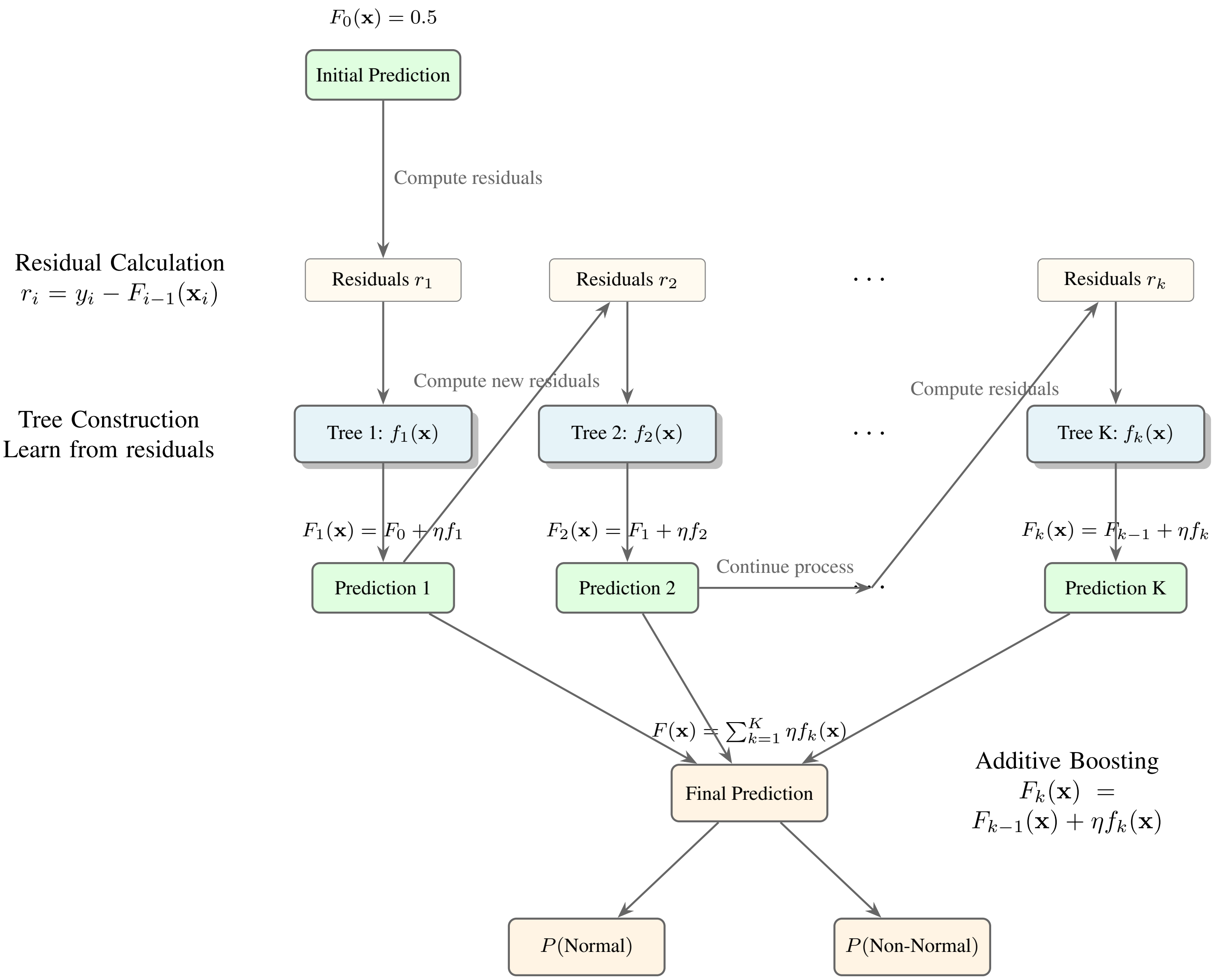


Figure 3.5: *Architecture of the XGBoost (Extreme Gradient Boosting) model. The algorithm builds trees sequentially, with each new tree learning from the residuals of the previous ensemble, and combines them additively for the final prediction.*

3.3.8 Support Vector Machines for Normality Classification

Support Vector Machines (SVM) construct an optimal separating hyperplane to discriminate between normal and non-normal distributions by maximizing the margin between classes (Cortes and Vapnik, 1995). Each sample is represented by a feature vector $\mathbf{x}$ containing statistical measures, with class labels $y \in \{-1, +1\}$ for non-normal and normal distributions, respectively.

The SVM defines the decision boundary $w^T\phi(\mathbf{x}) + b = 0$ where $\phi(\cdot)$ is a (potentially nonlinear) feature mapping, $w \in \mathbb{R}^d$ the weight vector, and $b \in \mathbb{R}$ the bias term. The classifier is trained by solving the following convex optimization problem (Cortes and Vapnik, 1995; Schölkopf and

Smola, 2002; Chen et al., 2025):

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \qquad \text{subject to} \quad y_i\left(w^T\phi(\mathbf{x}_i)+b\right) \geq 1-\xi_i, \quad \xi_i \geq 0 \tag{3.8}$$

where $C > 0$ is a user-specified regularization parameter that balances margin maximization with empirical classification error (penalized by the non-negative slack variables ξ_i) (Cortes and Vapnik, 1995).

For prediction, SVM uses kernel functions to handle nonlinear boundaries (Schölkopf and Smola, 2002; Chen et al., 2025):

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i\in SV}\alpha_i y_i K(\mathbf{x}_i,\mathbf{x})+b\right) \tag{3.9}$$

where $K(\mathbf{x}_i, \mathbf{x}_j) = \langle\phi(\mathbf{x}_i), \phi(\mathbf{x}j)\rangle$ is the kernel function and α_i are coefficients.

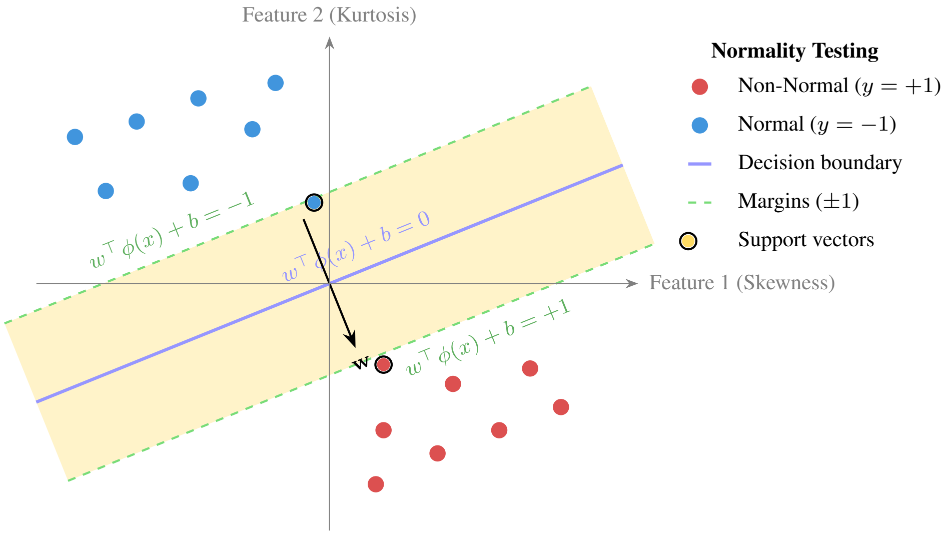


Figure 3.6: Support Vector Machine for **normality testing**. Red points are **Non-Normal** ($y = +1$); blue points are **Normal** ($y = -1$). The purple line is the decision boundary $w^\top\phi(x)+b = 0$; green dashed lines are the margins (-1 on the top, $+1$ on the bottom); the light yellow band is the margin region. Circled samples are the *support vectors*. The arrow $\mathbf{w}$ points toward the positive (Non-Normal) side.

3.3.9 K-Nearest Neighbors for Normality Classification

The K-Nearest Neighbors (KNN) algorithm classifies distributions by comparing them to their most similar instances in the training set (Cover and Hart, 1967). As an instance-based learning method, KNN stores all training examples and classifies new samples based on the predominant class among their k closest neighbors (Aha et al., 1991).

For a query sample $\mathbf{x}_q$, KNN computes distances to all training instances using the Euclidean metric as introduced by Cover and Hart (1967):

$$d(\mathbf{x}_q, \mathbf{x}_i) = \sqrt{\sum_{j=1}^{m}(x_{qj} - x_{ij})^2} \tag{3.10}$$

The algorithm then identifies the k nearest neighbors and assigns the majority class (Aha et al., 1991):

$$\hat{y}_q = \arg\max_c \sum_{i=1}^{k} \mathbb{I}(y_i = c) \tag{3.11}$$

Alternatively, distance-weighted voting (Dudani, 1976) can emphasize closer neighbors:

$$\hat{y}_q = \arg\max_c \sum_{i=1}^{k} \frac{1}{d(\mathbf{x}_q, \mathbf{x}_i)} \mathbb{I}(y_i = c) \tag{3.12}$$

The choice of k critically influences model performance, with smaller values capturing local patterns and larger values providing smoother decision boundaries. Optimal k selection through cross-validation balances this bias-variance tradeoff for normality classification.

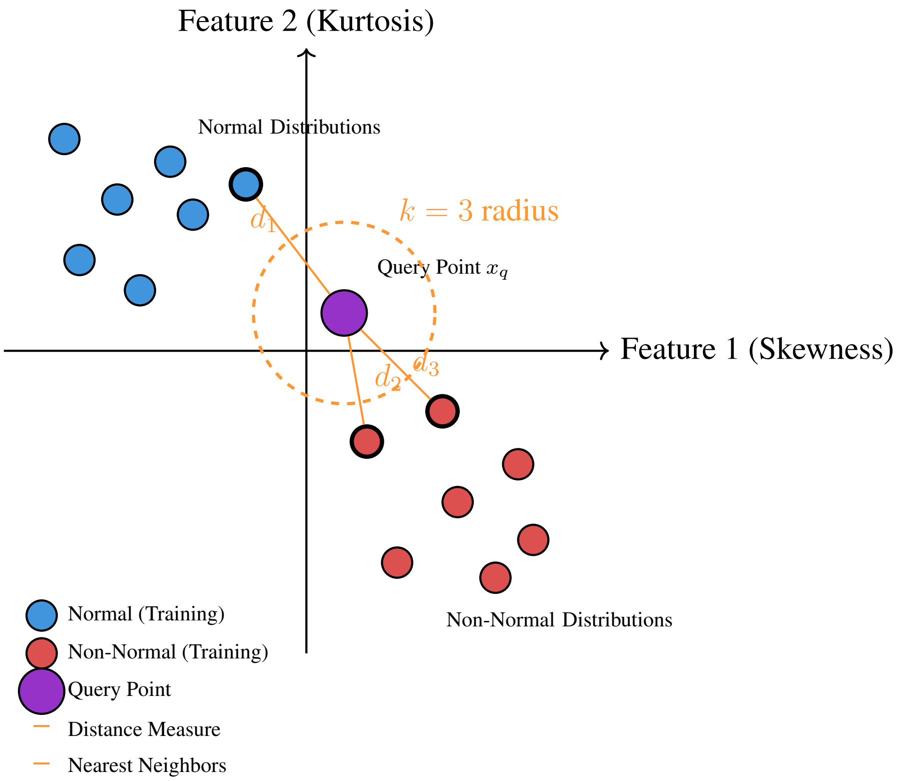


Figure 3.7: K-Nearest Neighbors algorithm for normality testing. The query point (new statistical sample) is classified based on the majority class among its $k = 3$ nearest neighbors in the feature space (Cover and Hart, 1967).

3.3.10 Models Training and Validation

A rigorous 10-fold cross-validation framework was employed to train and evaluate all models, ensuring reliable estimates of their generalization performance. A unified training control object was defined in R using the `caret` package to implement this strategy consistently across all six models, using accuracy as the primary metric for tuning and evaluation.

3.3.11 Evaluation Metrics

To comprehensively assess and compare the performance of the classification models, multiple evaluation metrics were calculated on completely new data sets generated from different distributions. The following metrics, derived from the confusion matrix (Table 3.2), were used (Suresh, 2015):

	Predicted Class	
Actual Class	Non_Normal	Normal
Non_Normal	True Positive (TP)	False Negative (FN)
Normal	False Positive (FP)	True Negative (TN)

Table 3.2: Structure of a confusion matrix for a binary classification problem.

- **Accuracy:** The overall proportion of correct predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.13}$$

- **Sensitivity (Recall/True Positive Rate):** The ability of the model to correctly identify Non_Normal distributions.

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.14}$$

- **Specificity (True Negative Rate):** The ability of the model to correctly identify Normal distributions.

$$Specificity = \frac{TN}{TN + FP} \tag{3.15}$$

- **Precision:** The proportion of predicted Non_Normal cases that are actually Non_Normal.

$$Precision = \frac{TP}{TP + FP} \tag{3.16}$$

- **F1 Score:** The harmonic mean of Precision and Sensitivity.

$$F1 = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \tag{3.17}$$

- **Area Under the ROC Curve (AUC):** An aggregate measure of performance across all classification thresholds.

3.3.12 Feature Importance and Model Evaluation

We evaluated model robustness and generalizability through a simulation study using distributions not seen during training. The study included both standard and challenging distributional cases to test model performance under varied conditions.

Feature importance was measured using the Random Forest's built-in permutation method, which estimates how much each feature contributes to prediction accuracy. This metric is based on the mean decrease in Gini impurity when a feature's values are permuted. Figure 3.8 presents the ranked importance of the most influential features.

Model performance was further assessed using the Area Under the Receiver Operating Characteristic (AUROC) curves obtained from the simulation results. We compared the top three machine learning models with three classical normality tests—Shapiro–Wilk, Jarque–Bera, and Anderson–Darling test across multiple sample sizes and alternative distributions. The choice of the classical normality tests is based on the literature and findings in our chapter 2 results. Shapiro-Wilk is considered to be generally most powerful in broad alternatives, Anderson-Darling is very sensitivity in the distribution tails, and Jarque-Bera good at detecting skewness and kurtosis departures from normality.

3.4 Results and Discussion

We reveal the results of the machine learning models, starting with the feature importance chart, highlighting the most important features in discriminating samples as either normal or non-normal, followed by a combined ROC curve plot, summary table of the metrics for all the models, then comparative ROC curves for the machine learning models and the classical normality test methods.

3.4.1 Variable Importance Chart

Variable importance charts highlight which statistical features most strongly influence the model's ability to distinguish normal from non-normal distributions. By ranking features such as skewness, kurtosis, and entropy based on their contribution to the classification, these charts translate complex model decisions into understandable terms. This helps clarify which aspects of the data drive the detection of normality departures and supports transparent interpretation and refinement of normality testing methods within this research.

An important observation from Figure 3.8 is that the importance of these features are context dependent, but nevertheless, they provide very important information as to which features from the data maybe be most important in classifying samples as either normal or non-normal. It is also evident from Figure 3.8 the weaknesses of the existing classical normality test procedures in small samples. With sample size 8, the features related to the classical tests are among the least important features, however, in large samples, 50, they are among the best features highlighting their sample size dependent nature. For small sample sample size, Median, Root Mean Squares(RMS), Energy, Moment Ratio, Tail Weight Ratio, and Coefficient of Variation(CV) are the five most important features but for larger sample size, Median, Moment Ratio, Energy, Anderson Darlin test, and

Shapiro-Wilk test are the five most important features in classifying samples as either normal or non-normal from the training set of distributions in Table 3.1 above.

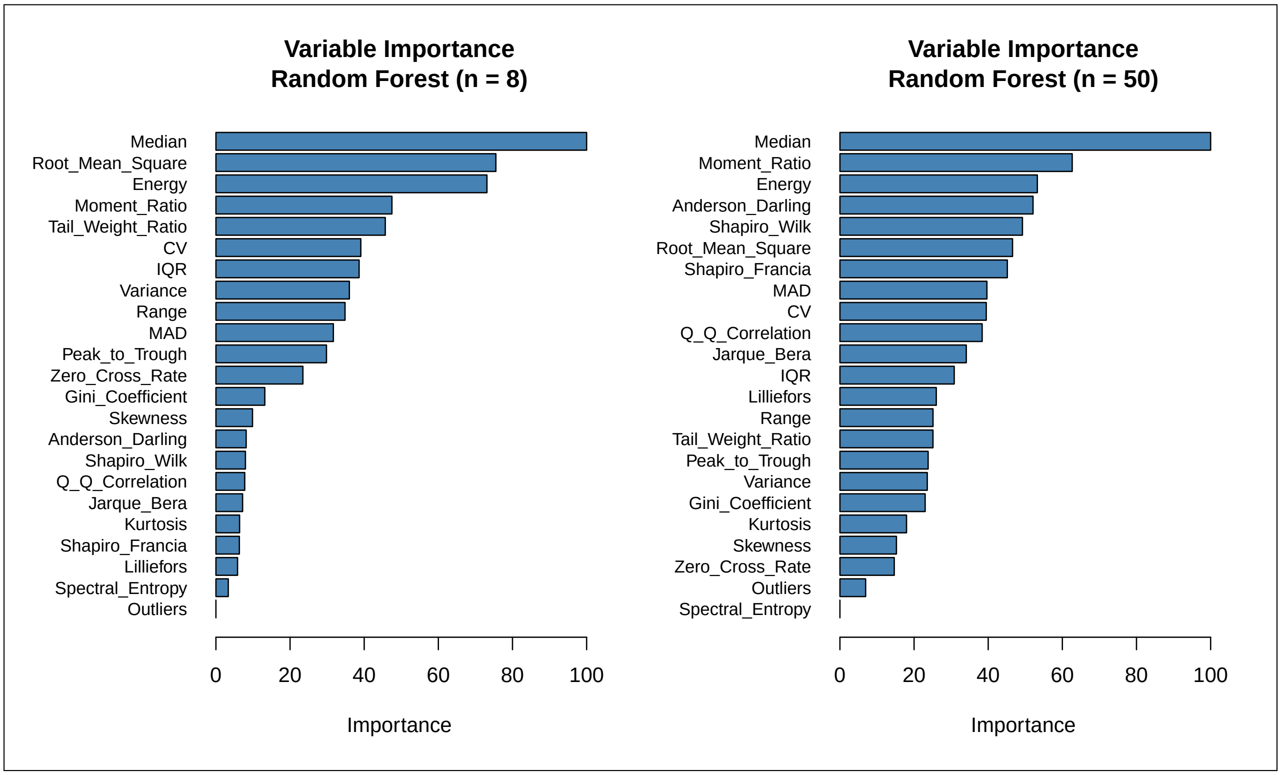


Figure 3.8: Variable importance from the Random Forest model for sample of sizes 8 generated from the training set of distributions in Table 3.1 with the most important features at the top.

3.4.2 Model Comparison Using AUROC

Model performance was compared using the Area Under the Receiver Operating Characteristic (AUROC) curve, which measures a classifier's ability to distinguish normal from non-normal distributions across all decision thresholds. Higher AUROC values indicate better discrimination. From Figure 3.9 below, Gradient Boosting and Random Forest are the best two performing models in the given scenario.

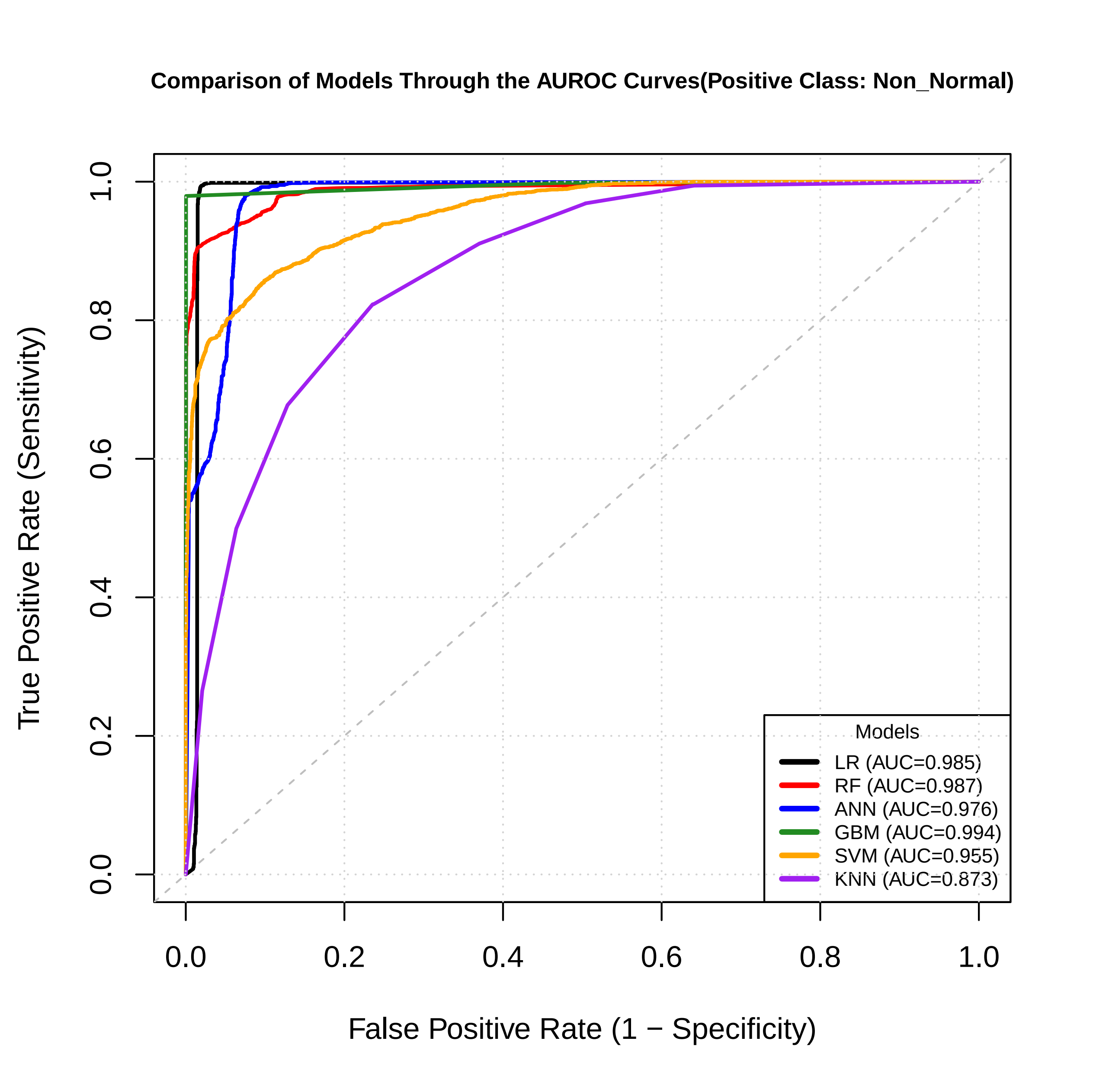


Figure 3.9: *Area Under the Curve(AUC) of Receiver Operating Characteristic (ROC) Curve for sample of sizes 8 generated from N(0, 5), N(50,15), N(25, 10), beta(2,5), Gumbel(0,1), and χ^2_3 distributions*

3.4.3 Model Evaluation Metrics

We also present a table of several key evaluation metrics to evaluate and compare the performance of the classification models. These metrics provide a comprehensive view of each model's accuracy, ability to correctly identify normal and non-normal samples, and overall discriminative power. We presented this for sample sizes 8, 30, and 50 to allow us see how each model performs with varying sample sizes as shown in Tables 3.3, 3.4, and 3.5 respectively.

Model Type	Accuracy	Sensitivity	Specificity	Precision	F1 Score	AUC
Logistic Regression	0.8950	0.8043	0.9857	0.9825	0.8845	0.985
Random Forest	0.8892	0.7793	0.9990	0.9987	0.8755	0.987
Artificial Neural Network	0.8473	0.7457	0.9490	0.9360	0.8301	0.976
Gradient Boosting Machines	0.9192	0.8387	0.9997	0.9996	0.9121	0.994
Support Vector Machines	0.8750	0.8187	0.9313	0.9226	0.8675	0.955
K-Nearest Neighbors	0.7933	0.8230	0.7637	0.7769	0.7993	0.873

Table 3.3: *Showing summary of important metrics for each model on samples of sizes 8 generated from N(0, 5), N(50,15), N(25, 10), beta(2,5), Gumbel(0,1), and χ^2_3 distributions*

Model Type	Accuracy	Sensitivity	Specificity	Precision	F1 Score	AUC
Logistic Regression	0.9335	0.8757	0.9913	0.9902	0.9294	0.933
Random Forest	0.9600	0.9200	1.0000	1.0000	0.9583	0.996
Artificial Neural Network	0.9012	0.8160	0.9863	0.9835	0.8920	0.992
Gradient Boosting Machines	0.9527	0.9053	1.0000	1.0000	0.9503	0.999
Support Vector Machines	0.9397	0.8980	0.9813	0.9796	0.9370	0.993
K-Nearest Neighbors	0.8930	0.8047	0.9813	0.9773	0.8826	0.978

Table 3.4: *Showing summary of important metrics for each model on samples of sizes 30 generated from N(0, 5), N(50,15), N(25, 10), beta(2,5), Gumbel(0,1), and χ^2_3 distributions*

Model Type	Accuracy	Sensitivity	Specificity	Precision	F1 Score	AUC
Logistic Regression	0.8070	0.6237	0.9903	0.9847	0.7637	0.807
Random Forest	0.9698	0.9403	0.9993	0.9993	0.9689	0.996
Artificial Neural Network	0.9285	0.8627	0.9943	0.9935	0.9235	0.996
Gradient Boosting Machines	0.9762	0.9523	1.0000	1.0000	0.9756	1.000
Support Vector Machines	0.9437	0.8997	0.9877	0.9865	0.9411	0.995
K-Nearest Neighbors	0.8988	0.8003	0.9973	0.9967	0.8878	0.959

Table 3.5: *Showing summary of important metrics for each model on samples of sizes 50 generated from N(0, 5), N(50,15), N(25, 10), beta(2,5), Gumbel(0,1), and χ^2_3 distributions*

Based on Tables 3.3, 3.4, 3.5, and Figure 3.9, Random Forest, Gradient Boosting Trees, and Artificial Neural Networks consistently rank as the top-performing models across all evaluation metrics among the six constructed models. However, it is important to acknowledge that these results are contingent on the specific set of test distributions used. Since the test data are generated from various, potentially differing distributions rather than a fixed subset, the relative performance ranking could shift if additional or alternative distributions are introduced. Therefore, while these three models currently demonstrate superior and stable performance, this selection remains adaptable.

Consequently, we focus on these models for comparison with the three classical normality tests in subsequent analyses.

3.4.4 Machine Learning Approach vs Classical Normality Test Methods

Figure 3.10 compares the performance of machine learning (ML) models with classical normality tests across different alternative distributions and sample sizes using AUROC curves. The results show that ML methods consistently outperform classical tests, particularly when sample sizes are small. This advantage stems from the ML models' flexibility and ability to learn complex distributional features without relying on restrictive assumptions inherent to classical methods. These findings highlight the potential of ML as a powerful alternative for normality assessment under diverse and challenging data conditions.

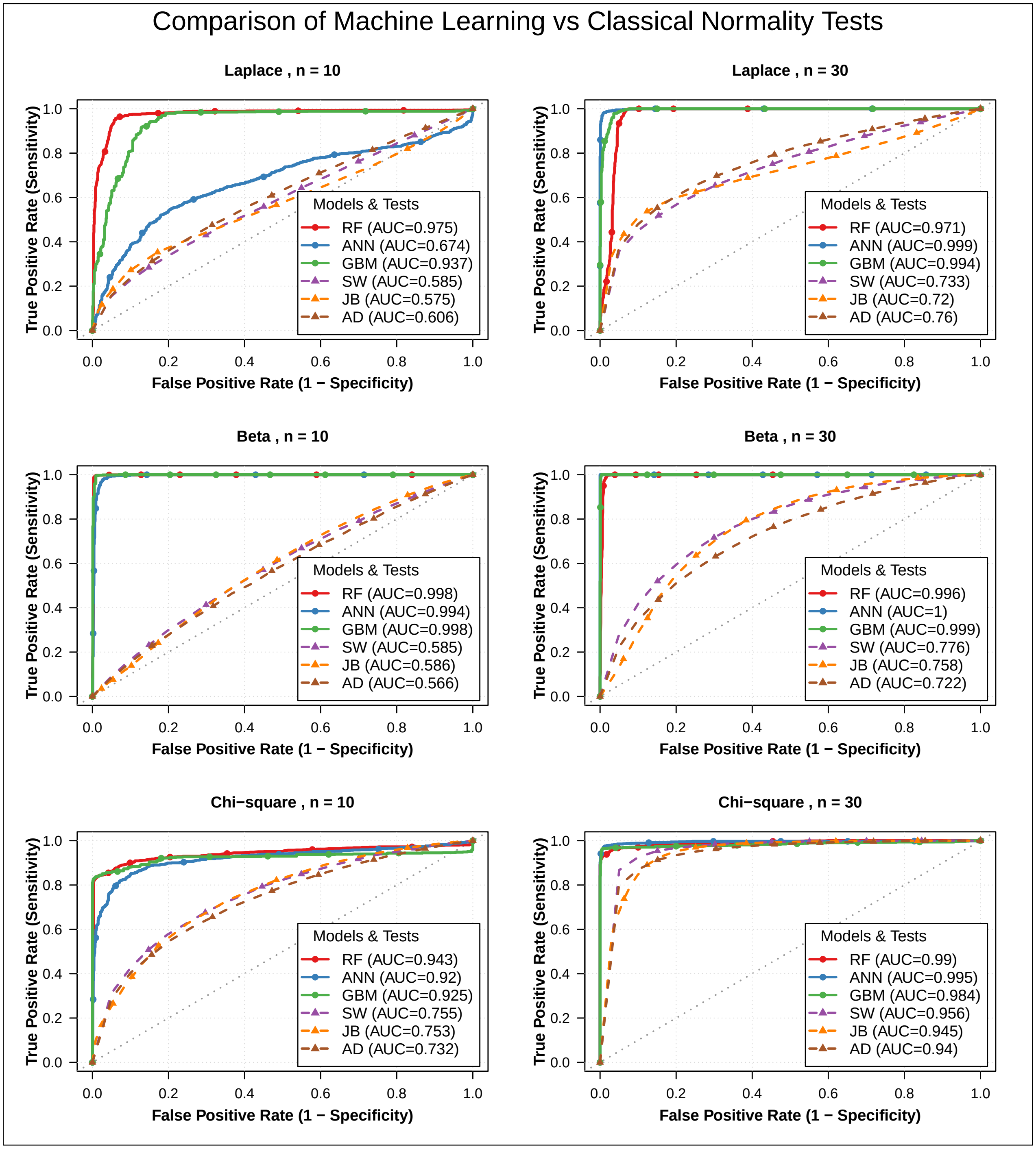


Figure 3.10: *Comparison of Machine Learning vs Classical Normality Tests for different alternative distributions: Laplace(0, 4),Beta(2,5) and χ^2_3 and sample sizes through the AUROC Curves.*

3.4.5 Discussion and Conclusion

The demonstration above highlights several significant advantages of the ML models over the classical normality test methods. First, the ML method outperformed all the “top-performing” classical normality test methods in all situations especially for smaller sample sizes. This is very important because we care more about non-normality when sample sizes are small. Secondly, unlike the classical normality test methods, ML methods are consistent across varied distributions and sample sizes, eliminating the confusion that practitioners often face in deciding which normality test method will be appropriate for a given situation. Finally, using the ML procedure, we can potentially reduce the amount of selection bias often associated with pre-testing for normality (explored further in the next chapter). The selection bias mostly arises due to the fact that normality pre-tests often have a fairly significant probability to mis-classify. Since the ML method is proven to be very effective at discriminating ‘normal’ from ‘non-normal’ samples, it can nearly eliminate the type I error and type II error in the pre-testing thereby ensuring correct downstream test choices and removing selection bias.

One potential challenge of using machine learning (ML) for normality testing is the need to train classifiers before applying them. However, this training step does not impose a burden on end users, as pre-trained ML models can be provided and used directly without requiring insight into the training process or model complexity.

The current study focused on univariate distributions and may not generalize to multivariate normality assessment. Additionally, while we evaluated a comprehensive set of distributions, real-world data may exhibit more complex patterns not captured in our simulation framework. Future work may aim to extend ML approaches to multivariate normality testing.

While classical tests will likely maintain their role in introductory statistics and quick diagnostic checks, ML approaches represent the future for rigorous normality assessment in research contexts where the consequences of failing to meet the assumptions are substantial.

Chapter 4

An Exploration of the Selective Inference Problem for Normality Pre-testing

4.1 Introduction

Selective inference addresses distortions that arise when statistical analyses are influenced by data-driven choices, such as conducting preliminary tests before main analyses. In the context of normality pre-testing, conditioning analytical decisions on pre-test outcomes can introduce selection bias and alter error rates, potentially undermining the validity and replicability of statistical conclusions. This chapter focuses on these issues, highlighting how data-dependent decision processes in normality assessment can compromise inference, a challenge that has been increasingly recognized and studied in recent statistical literature (Benjamini, 2020; Berk et al., 2013).

4.2 Literature Review

4.2.1 Selective Inference in Normality Pre-testing

One limitation of normality testing lies in its positioning within the analytical workflow. When the outcome of a normality test determines whether a parametric or non-parametric analysis is pursued, the resulting selection bias skews the inferential framework (Rochon et al., 2012). As Benjamini (2020) describe, this practice is a "silent killer" of replicability, violating the assumptions of underlying conventional statistical inference. The selection event conditions subsequent analysis on the data, changing the sampling distribution of test statistics and invalidating nominal Type I error rates (Rochon and Kieser, 2011).

Simulation studies further clarify the magnitude of this issue. Rochon and Kieser (2011); Schucany and Tony Ng (2006) reveal substantial inflation of Type I error rates, especially under skewed distributions, when analysis choices depend upon normality tests. The selection mechanism introduces systematic bias, particularly for one-sided hypotheses, altering both the direction and effect size estimates.

Historically, scholars have identified these challenges. Easterling and Anderson (1978) demonstrated through simulations that t-statistics, conditioned on passing normality tests, no longer conform to the assumed Student's t-distribution. Rochon et al. (2012) extended these findings to two-sample tests, showing how pre-testing distorts both Type I error and statistical power in downstream analyses.

4.2.2 Existing Strategies for Selective Inference Correction

Recognizing the challenge, statisticians have proposed several solutions to mitigate selection-induced bias. These methods, developed largely for variable selection, offer valuable perspectives for normality pre-testing.

Sample Splitting and Data Carving: Sample splitting divides the data into independent parts for selection and inference, restoring nominal error rates but somewhat diminishing statistical power (Cox, 1975; Rasines, 2023). The data carving technique, pioneered by Fithian et al. (2014), conditions on the selection event while exploiting the full dataset for inference, thereby optimally balancing power and validity.

Conditional Selective Inference: A mathematically rigorous framework by Lee et al. (2016) provides exact inference by conditioning directly on the selection event. This approach yields valid p-values and confidence intervals even after data-driven choices and has been extended to LASSO regression, stepwise selection, and principal component analysis (Tibshirani et al., 2016; G'Sell et al., 2016).

Simultaneous Inference Approaches: Simultaneous inference methods, such as Bonferroni and Scheffé's procedures, safeguard error rates across all selection outcomes by widening confidence intervals or raising significance thresholds (Kuchibhotla et al., 2022; Berk et al., 2013). Although robust, these methods trade improved validity for reduced statistical power.

False Discovery Rate Framework: In high-dimensional testing settings, controlling the false discovery rate (FDR) balances the risk of Type I errors across multiple comparisons (Benjamini and

Hochberg, 1995, 2000; Barber and Candes, 2015). While well-established for variable selection, the application of FDR methodologies to normality testing remains largely unexplored.

Applications Beyond Variable Selection: Selective inference principles extend to broader model diagnostics, affecting causal inference, change point detection, and traditional testing scenarios (Faraway, 1992; Efron, 2014; Taylor and Tibshirani, 2015; Lockhart et al., 2014).

4.2.3 Gaps in Application to Normality Testing

Despite robust progress in selective inference theory, applications to normality pre-testing remain underdeveloped. Existing work has documented the quantification of error inflation and inferential distortion associated with pre-testing (Schucany and Tony Ng, 2006; Rochon and Kieser, 2011), yet few studies offer practical frameworks to restore validity in this context.

Unique challenges—such as the continuous nature of selection, intricate dependencies between normality tests and analysis outcomes, and the distributional complexity of post-selection statistics—require tailored methodological innovation. Literature to date has concentrated on model and variable selection; assumption checking, particularly through normality testing, is not sufficiently addressed.

This gap matters because normality pre-testing is still widely used in research and practical applications (Rochon et al., 2012). We need to understand how selection effects from normality testing affect the validity of downstream statistical conclusions and whether our results will replicate in new studies.

In conclusion, normality testing creates a paradox. It is intended to protect against violations of distributional assumptions, yet the selection process itself introduces new problems that can be worse than the original concern. When researchers choose their analysis method based on a normality test, they alter the statistical properties of their subsequent tests in unpredictable ways. Moving forward, developing methods specifically tailored to handle selection effects in pre-testing would address a real problem in statistical practice and make research findings more trustworthy across different fields and applications.

In this chapter, we explore the issue of selective inference in the context of normality testing. We assess the impact of selective inference in some common use cases and propose some possible ways to mitigate these issues.

4.3 Demonstration of the selective inference problem in the context of normality pre-testing

As an illustration, consider the distribution of one and two sample location tests. In one case, we perform the t-test only when the normality pre-test is insignificant(conditional) and in the other case perform the t-test regardless(unconditional) and then calculate the test statistics from both situations, we obtain the true distribution of their respective test statistics. This reveals the extend to which pre-testing for normality distort the distribution of the test statistics leading to further inflation of type I error rates and loss of power. Figure 4.1, we see significant deviation of the distribution of conditional test statistic from the theoretical t-test statistics. This is more significant in the one sample case.

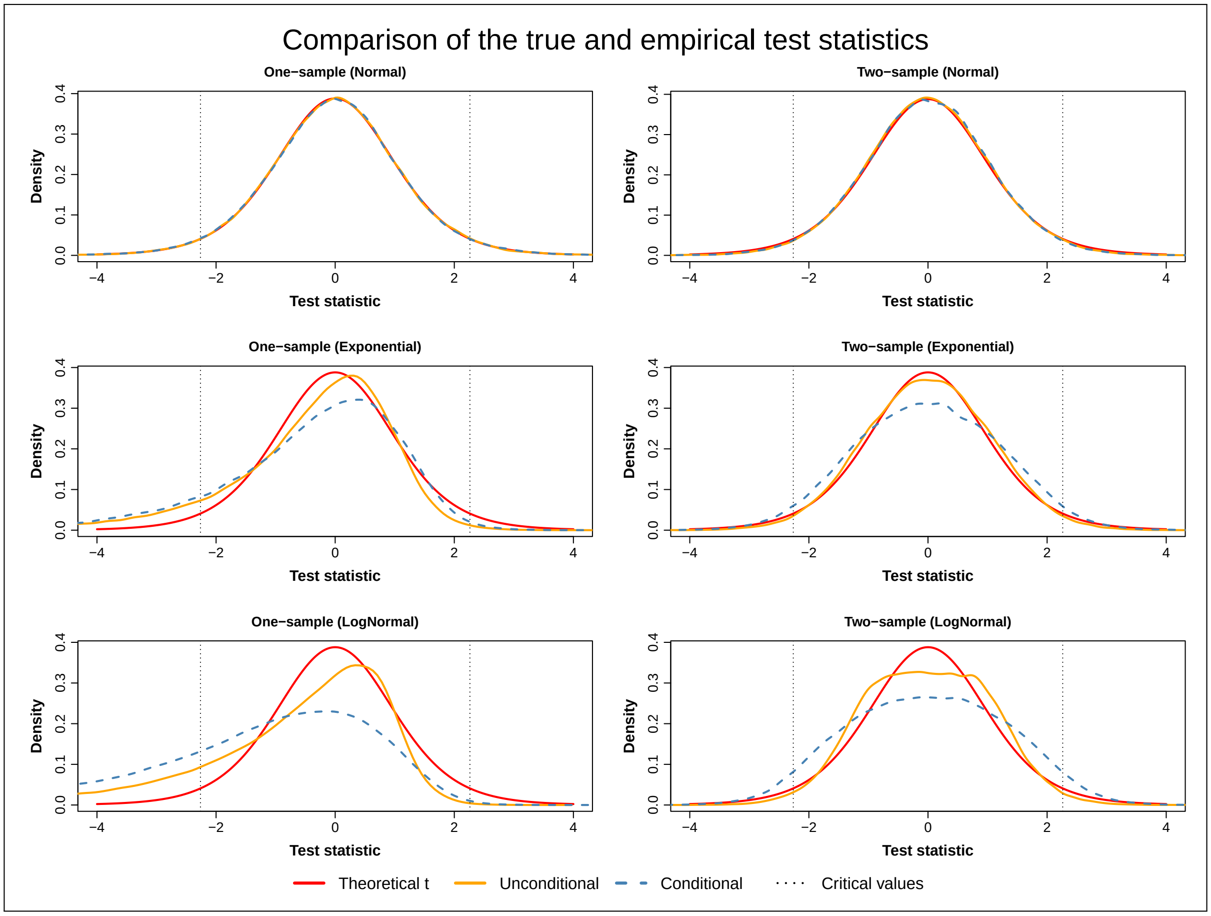


Figure 4.1: *Showing the impact of pretesting for normality on the distribution of one sample and two sample test statistics.*

From Figure 4.1, we estimate the empirical the Type I error probability for each procedure. Not all distributions are plotted in Figure 4.1 due to space. Obviously, there is significant inflation of conditional type I error rate compared to the unconditional type I error rates as shown in Table 4.1

below. For one sample case, the Type I error rate is inflated for asymmetric distributions like exponential, chi-squared, and lognormal. The inflation is more pronounced in the conditional t-test. The unconditional Type I error rate is controlled in two-sample t-tests, but the conditional Type I error rate is inflated. This emphasizes the effects of selection bias due to pre-testing for normality.

One-Sample t-Test			**Two-Sample t-Test**		
Distribution	**Unconditional**	**Conditional**	**Distribution**	**Unconditional**	**Conditional**
Normal	0.0518	0.047	Normal	0.0371	0.0435
Uniform	0.0541	0.0472	Uniform	0.0351	0.0387
t	0.0477	0.0463	t	0.0315	0.0444
Exponential	0.097	0.1353	Exponential	0.0310	0.1113
Chi-Squared	0.0821	0.1069	Chi-Squared	0.0332	0.0789
Lognormal	0.1602	0.4019	Lognormal	0.0213	0.1602

Table 4.1: *Showing estimated probability of Type I error rates for the conditional and unconditional t-test for both one-sample and two-sample tests for sample size of 10.*

4.4 Proposed solutions

To be completed

4.5 Assessment of the magnitude of the selective inference problem and the performance of the proposed solutions

To be completed

4.6 Discussion and future directions

To be completed

Declaration of Generative AI use

In this dissertation, I did refer to Generative AI to get ideas on how to structure and debug some R codes. I verified any R codes for correctness and then rewrite them to meet my purpose.

References

David W Aha, Dennis Kibler, and Marc K Albert. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991.

Sivan Aldor-Noiman, Lawrence D Brown, Andreas Buja, Wolfgang Rolke, and Robert A Stine. The power to see: A new graphical test of normality. *The American Statistician*, 67(4):249–260, 2013.

Kolmogorov An. Sulla determinazione empirica di una legge didistribuzione. *Giorn Dell'inst Ital Degli Att*, 4:89–91, 1933.

Theodore W Anderson and Donald A Darling. A test of goodness of fit. *Journal of the American statistical association*, 49(268):765–769, 1954.

Rina Foygel Barber and Emmanuel J Candes. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

Husam Awni Bayoud. Tests of normality: new test and comparative study. *Communications in Statistics-Simulation and Computation*, 50(12):4442–4463, 2021.

Yoav Benjamini. Selective inference: The silent killer of replicability. *Harvard Data Science Review*, 2(4), 2020.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Yoav Benjamini and Yosef Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of educational and Behavioral Statistics*, 25(1):60–83, 2000.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, pages 802–837, 2013.

Joseph Berkson. Application of the logistic function to bio-assay. *Journal of the American statistical association*, 39(227):357–365, 1944.

Peter J Bickel and Kjell A Doksum. An analysis of transformations revisited. *Journal of the american statistical association*, 76(374):296–311, 1981.

Douglas G Bonett and Edith Seier. A test of normality with high uniform power. *Computational statistics & data analysis*, 40(3):435–445, 2002.

George EP Box. Science and statistics. *Journal of the American Statistical Association*, 71(356): 791–799, 1976.

George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 26(2):211–243, 1964.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Raymond J Carroll and David Ruppert. Power transformations when fitting theoretical models to data. *Journal of the American Statistical Association*, 79(386):321–328, 1984.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Zhaokun Chen, Chaopeng Zhang, Xiaohan Li, Wenshuo Wang, Gentiane Venture, and Junqiang Xi. Driving style recognition like an expert using semantic privileged information from large language models. *arXiv preprint arXiv:2508.13881*, 2025.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

David R Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, pages 441–444, 1975.

Ralph B d'Agostino. An omnibus test of normality for moderate and large size samples. *Biometrika*, 58(2):341–348, 1971.

Sahibsingh A Dudani. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, (4):325–327, 1976.

RG Easterling and HE Anderson. The effect of preliminary normality goodness of fit tests on subsequent inference. *Journal of Statistical Computation and Simulation*, 8(1):1–11, 1978.

Bradley Efron. Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007, 2014.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

Julian J Faraway. On the cost of data analysis. *Journal of Computational and Graphical Statistics*, 1(3):213–229, 1992.

James J Filliben. The probability plot correlation coefficient test for normality. *Technometrics*, 17 (1):111–117, 1975.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

RS Francia. An approximate analysis of variance test for normality. *Journal of the American statistical Association*, 67(337):215–216, 1972.

David A Freedman and David A Freedman. A note on screening regression equations. *the american statistician*, 37(2):152–155, 1983.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Robert Charles Geary. Testing for normality. *Biometrika*, 34(3/4):209–242, 1947.

Asghar Ghasemi and Saleh Zahediasl. Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, 10(2):486, 2012.

Phillip Good. *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media, 2013.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Max Grazier G'Sell, Stefan Wager, Alexandra Chouldechova, and Robert Tibshirani. Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(2):423–444, 2016.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, et al. The elements of statistical learning, 2009.

Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric statistical methods*. John Wiley & Sons, 2013.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005.

Tanweer Ul Islam. Ranking of normality tests: An appraisal through skewed alternative space. *Symmetry*, 11(7):872, 2019.

Carlos M Jarque and Anil K Bera. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics letters*, 6(3):255–259, 1980.

A kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91, 1933.

William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

Arun K Kuchibhotla, John E Kolassa, and Todd A Kuffner. Post-selection inference. *Annual Review of Statistics and Its Application*, 9:505–527, 2022.

Max Kuhn. Building predictive models in r using the caret package. *Journal of statistical software*, 28:1–26, 2008.

Sang Gyu Kwak and Jong Hae Kim. Central limit theorem: the cornerstone of modern statistics. *Korean journal of anesthesiology*, 70(2):144, 2017.

Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44(3):907–927, 2016.

E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, 3rd edition, 2005.

Erich Leo Lehmann. *Elements of large-sample theory*. Springer, 1999.

Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3): 18–22, 2002.

Hubert W Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, 62(318):399–402, 1967.

Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.

Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

Zhen Meng and Zhenzhen Jiang. Cauchy combination omnibus test for normality. *Plos one*, 18 (8):e0289498, 2023.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

Stefano Nembrini, Inke R König, and Marvin N Wright. The revival of the gini importance? *Bioinformatics*, 34(21):3711–3718, 2018.

Egon S Pearson. A further development of tests for normality. *Biometrika*, pages 239–249, 1930.

D García Rasines. Splitting strategies for post-selection inference. *Biometrika*, 110(3):597–614, 2023.

Nik AK Razali and Yap Bee Wah. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011.

Brian D Ripley. *Pattern recognition and neural networks*. Cambridge university press, 2007.

Justine Rochon and Meinhard Kieser. A closer look at the effect of preliminary goodness-of-fit testing for normality for the one-sample t-test. *British Journal of Mathematical and Statistical Psychology*, 64(3):410–426, 2011.

Justine Rochon, Matthias Gondan, and Meinhard Kieser. To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC medical research methodology*, 12:1–11, 2012.

Walter A Rosenkrantz. Confidence bands for quantile functions: A parametric and graphic alternative for testing goodness of fit. *The American Statistician*, 54(3):185–190, 2000.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, Aaditya Ramdas, and Ryan J Tibshirani. A higher-order kolmogorov-smirnov test. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2621–2630. PMLR, 2019.

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

William R. Schucany and Hong Ng. Preliminary test estimation: An overview. *Statistical Papers*, 47:399–427, 2006.

William R Schucany and HK Tony Ng. Preliminary goodness-of-fit tests for normality do not validate the one-sample student t. *Communications in Statistics-Theory and Methods*, 35(12): 2275–2286, 2006.

George AF Seber and Alan J Lee. *Linear regression analysis*. John Wiley & Sons, 2003.

Robert J Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.

J Sigut, J Piñeiro, J Estévez, and P Toledo. A neural network approach to normality testing. *Intelligent Data Analysis*, 10(6):509–519, 2006.

Miloš Simić. Testing for normality with neural networks. *Neural Computing and Applications*, 33 (23):16279–16313, 2021.

Yeresime Suresh. *Software Fault Prediction and Test Data Generation Using Articial Intelligent Techniques*. PhD thesis, National Institute of Technology Rourkela, 2015.

Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

Henry C Thode. *Testing for normality*. CRC press, 2002.

Jun Yan Thomas Lumley. The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23:151–169, 2002.

Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

Oldrich Vasicek. A test for normality based on sample entropy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 38(1):54–59, 1976.

Sanford Weisberg. Yeo-johnson power transformations. *Department of Applied Statistics, University of Minnesota. Retrieved June*, 1(2003):8, 2001.

Bernard Lewis Welch. The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

Rand R Wilcox. *Introduction to robust estimation and hypothesis testing*. Academic press, 2012.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1(6):80–83, 1945.

M. B. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis of data. *Journal of the American Statistical Association*, 63(324):1063–1073, 1968.

Paul R Wilson and Alejandro B Engel. Testing for normality using neural networks. In *[1990] Proceedings. First International Symposium on Uncertainty Modeling and Analysis*, pages 700–704. IEEE, 1990.

Bee Wah Yap and Chiaw Hock Sim. Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, 81(12):2141–2155, 2011.

In-Kwon Yeo and Richard A Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000.

Donald W Zimmerman. A simple and effective decision rule for choosing a significance test to protect against non-normality. *British Journal of Mathematical and Statistical Psychology*, 64 (3):388–409, 2011.

Appendix A

Appendix

A.0.1 Algorithm Implementation

Algorithm 1 One-Sample Bootstrap Location Test

```
procedure BOOTSTRAPTEST($\mathbf{X}, \theta_0, T, B$, alternative)
    $t_{\text{obs}} \leftarrow T(\mathbf{X})$                          ▷ Compute observed test statistic
    $\mathbf{X}^0 \leftarrow \mathbf{X} - t_{\text{obs}} + \theta_0$              ▷ Center data under $H_0$
    bootstrap_stats $\leftarrow$ array of length $B$
    for $b = 1$ to $B$ do
        $\mathbf{X}^{*(b)} \leftarrow$ sample with replacement from $\mathbf{X}^0$
        bootstrap_stats$[b] \leftarrow T(\mathbf{X}^{*(b)})$
    end for
    if alternative $=$ ”two-sided” then
        $p \leftarrow \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\left(|\text{bootstrap\_stats}[b] - \theta_0| \geq |t_{\text{obs}} - \theta_0|\right)$
    else if alternative $=$ ”greater” then
        $p \leftarrow \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\left(\text{bootstrap\_stats}[b] \geq t_{\text{obs}}\right)$
    else
        $p \leftarrow \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\left(\text{bootstrap\_stats}[b] \leq t_{\text{obs}}\right)$
    end if
    return $p$
end procedure
```

Algorithm 2 Two-Sample Bootstrap Location Test

```
procedure TwoSampleBootstrapTest($\mathbf{X}, \mathbf{Y}, \delta_0, B$)
    $n_1 \leftarrow \text{length}(\mathbf{X}), \quad n_2 \leftarrow \text{length}(\mathbf{Y})$
    $\bar{X} \leftarrow \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \quad \bar{Y} \leftarrow \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j$
    $s_X^2 \leftarrow \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2, \quad s_Y^2 \leftarrow \frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_j - \bar{Y})^2$
    $T_{\text{obs}} \leftarrow \frac{(\bar{X}-\bar{Y})-\delta_0}{\sqrt{s_X^2/n_1+s_Y^2/n_2}}$        ▷ Compute observed test statistic
    $\bar{G} \leftarrow \frac{1}{n_1+n_2}(\sum X_i + \sum Y_j)$        ▷ Compute grand mean
    $\mathbf{X}^0 \leftarrow \mathbf{X} - \bar{X} + \bar{G} + \delta_0/2$        ▷ Center first sample
    $\mathbf{Y}^0 \leftarrow \mathbf{Y} - \bar{Y} + \bar{G} - \delta_0/2$        ▷ Center second sample
    for $b = 1$ to $B$ do
        $\mathbf{X}^{*(b)} \leftarrow$ sample with replacement from $\mathbf{X}^0$
        $\mathbf{Y}^{*(b)} \leftarrow$ sample with replacement from $\mathbf{Y}^0$
        $\bar{X}^{*(b)} \leftarrow \frac{1}{n_1}\sum_{i=1}^{n_1} X_i^{*(b)}$
        $\bar{Y}^{*(b)} \leftarrow \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j^{*(b)}$
        $s_X^{2*(b)} \leftarrow \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i^{*(b)} - \bar{X}^{*(b)})^2$
        $s_Y^{2*(b)} \leftarrow \frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_j^{*(b)} - \bar{Y}^{*(b)})^2$
        $T_b^* \leftarrow \frac{(\bar{X}^{*(b)}-\bar{Y}^{*(b)})-\delta_0}{\sqrt{s_X^{2*(b)}/n_1+s_Y^{2*(b)}/n_2}}$
    end for
    $p \leftarrow \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(|T_b^*| \geq |T_{\text{obs}}|)$        ▷ Two-sided p-value
    return $p$
end procedure
```

Algorithm 3 One-Sample Permutation Test for the Mean

```
Require: Sample data $\mathbf{X} = (X_1, \ldots, X_n)$, null value $\mu_0$, number of permutations $B$
Calculate observed test statistic:
    $T_{\text{obs}} \leftarrow \frac{\sqrt{n}(\overline{X} - \mu_0)}{s} \quad \text{where } s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$
for $j = 1$ to $B$ do
    Generate random sign vector $I_j \in \{-1, 1\}^n$
    Permute around the null: $X_i^* = \mu_0 + I_{j,i} \cdot (X_i - \mu_0)$ for all $i$
    Compute $\overline{X^*}$ and sample standard deviation $s^*$
    Compute permuted statistic: $T_j^* = \frac{\sqrt{n}(\overline{X^*} - \mu_0)}{s^*}$
end for
Compute p-value:
    $p = \frac{1}{B}\sum_{j=1}^{B} \mathbb{I}\left(|T_j^*| \geq |T_{\text{obs}}|\right)$
Return: p-value
```

Algorithm 4 Bootstrap Regression Slope Test

```
Require: Dataset $\{(x_i, y_i)\}_{i=1}^{n}$, number of bootstraps $B$
Fit OLS regression $y \sim x$ on observed data; let $b_{\text{obs}} \leftarrow$ estimated slope
for $b = 1$ to $B$ do
    Draw $n$ rows with replacement from the data to form a bootstrap sample $(x^*, y^*)$
    Fit OLS regression $y^* \sim x^*$; store bootstrap slope $b_b^*$
end for
$\bar{b^*} \leftarrow$ mean of $\{b_1^*, \ldots, b_B^*\}$
Centered slopes: $b_{\text{cent},b}^* \leftarrow b_b^* - \bar{b^*}$ for all $b$
$p \leftarrow \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(|b_{\text{cent},b}^*| \geq |b_{\text{obs}}|)$
Return: $p$ as two-sided bootstrap p-value
```

Algorithm 5 Permutation Regression Slope Test

```
Require: Dataset $\{(x_i, y_i)\}_{i=1}^{n}$, number of permutations $B$
Fit OLS regression $y \sim x$; let $b_{\text{obs}} \leftarrow$ observed slope estimate
for $b = 1$ to $B$ do
    Permute $y$ entries randomly to get $y^*$; keep $x$ fixed
    Fit OLS regression $y^* \sim x$; store permuted slope $b_b^*$
end for
$p \leftarrow \frac{1+\sum_{b=1}^{B}\mathbb{I}(|b_b^*| \geq |b_{\text{obs}}|)}{B+1}$          ▷ Two-sided permutation p-value with +1 correction
Return: $p$
```

Algorithm 6 Box-Cox T-Test for Transformed Median

```
Require: Sample data $\mathbf{X} = (X_1, \ldots, X_n)$, null value $\mu_0$
if $\texttt{min}(\mathbf{X}) \leq 0$ then
    Shift data: $\mathbf{X}_{\text{shift}} \leftarrow \mathbf{X} + |\min(\mathbf{X})| + \varepsilon$
    $\mu_0^{\text{shift}} \leftarrow \mu_0 + |\min(\mathbf{X})| + \varepsilon$
else
    $\mathbf{X}_{\text{shift}} \leftarrow \mathbf{X}$
    $\mu_0^{\text{shift}} \leftarrow \mu_0$
end if
Find optimal $\lambda$ maximizing normality (e.g., by profile likelihood method over grid $\Lambda$)
Box-Cox transform data:
if $\lambda = 0$ then
    $\mathbf{X}_{\text{bc}} \leftarrow \log(\mathbf{X}_{\text{shift}})$
    $\mu_{0,\text{bc}} \leftarrow \log(\mu_0^{\text{shift}})$
else
    $\mathbf{X}_{\text{bc}} \leftarrow \frac{\mathbf{X}_{\text{shift}}^{\lambda} - 1}{\lambda}$
    $\mu_{0,\text{bc}} \leftarrow \frac{(\mu_0^{\text{shift}})^{\lambda} - 1}{\lambda}$
end if
Perform one-sample t-test on $\mathbf{X}_{\text{bc}}$ against $\mu_{0,\text{bc}}$:
Compute t-statistic $t = \frac{\overline{\mathbf{X}_{\text{bc}}} - \mu_{0,\text{bc}}}{s_{\text{bc}}/\sqrt{n}}$        ▷ $s_{\text{bc}}$ is the sample standard deviation of $\mathbf{X}_{\text{bc}}$
Compute two-sided p-value
return p-value as test result
```

Feature Name	**Description**	**Type**
Median	Sample median of the series.	numeric
Variance	Sample variance.	numeric
IQR	Interquartile range (Q3–Q1).	numeric
MAD	Median absolute deviation from the median.	numeric
Range	Max minus min.	numeric
CV	Coefficient of variation, $\mathrm{sd}(x)/\mathrm{mean}(x)$.	numeric
Root_Mean_Square	$\sqrt{\mathrm{mean}(x^2)}$.	numeric
Skewness	Third standardized moment	numeric
Kurtosis	Fourth standardized moment	numeric
Jarque_Bera	Jarque–Bera normality test statistic.	numeric
Anderson_Darling	Anderson–Darling normality test statistic.	numeric
Shapiro_Wilk	Shapiro–Wilk W statistic.	numeric
Shapiro_Francia	Shapiro–Francia W' statistic.	numeric
Lilliefors	Lilliefors (KS–type) test statistic.	numeric
Zero_Cross_Rate	Fraction of sign changes in successive samples.	numeric
Gini_Coefficient	Gini index of $\lvert x - \min(x) \rvert$.	numeric
Outliers	$\sum x_i \notin ([Q1 - 1.5\,\mathrm{IQR},\ Q3 + 1.5\,\mathrm{IQR}])$.	integer
Peak_to_Trough	$\max(x)/\lvert\min(x)\rvert$.	numeric
Spectral_Entropy	$-\sum_{k=1}^{K} p_k \log p_k,\ p_k = \dfrac{I_k}{\sum_{j=1}^{K} I_j}$.	numeric
Energy	$\sum_i x_i^2$.	numeric
Q_Q_Correlation	Correlation of data quantiles with normal quantiles.	numeric
Tail_Weight_Ratio	$\mathrm{Quantile}_{0.95}/\mathrm{Quantile}_{0.05}$.	numeric
Moment_Ratio	$\mathrm{moment}_4/\mathrm{moment}_2^2$.	numeric

Table A.1: Features used as predictors in the machine–learning models.