Reassessing the Goodness-of-Fit tests in Actuarial Science decision-making: Can we rely on p-value reports?

Michael Okanta

April 1, 2025

1 Abstract

Goodness-of-fit (GoF) tests are widely used in statistical modeling, particularly in actuarial science, where choosing an appropriate model for loss distributions has direct implications for pricing, reservation, and risk management. Despite their popularity, increasing attention is being paid to the limitations of these tests, particularly with regard to their dependence on p-values. Although p-values are traditionally used to accept or reject null hypotheses, recent discussions in the literature have raised concerns about their interpretability, especially in finite-sample contexts or when testing complex models such as heavy-tailed or mixture distributions. This paper aims to revisit and reassess the performance of four commonly used GoF tests: the Kolmogorov-Smirnov (KS), Anderson-Darling (AD), Cramér-von Mises (CvM) and Chi-square tests, by evaluating their p-value distributions under the null hypothesis and their power under various alternatives.

A thorough simulation study is conducted using both mixture Burr and Lognormal distributions, which are known for their flexibility in modeling skewed and heavy-tailed data, as well as their ability to cover multimodality cases. The parameter estimates for the Burr and Lognormal distributions are derived from real-world financial datasets, ensuring the practical relevance of the results. The study investigates the empirical behavior of p-values generated under the null model and compares them with the distribution of p-values under slightly misspecified alternative models. In doing so, we highlight instances in which the tests may produce misleading p-values, even when the underlying assumptions are only slightly violated.

In addition to studying the shape and spread of the p-value distributions, the statistical power of each test is assessed in varying sample sizes and distributional settings. The results reveal significant differences in test performance, with some tests being overly conservative or liberal depending on the alternative and sample size. The findings suggest that no single test performs uniformly well across all scenarios and that relying solely on p-values, without a proper understanding of their distributional properties, can lead to flawed conclusions about model adequacy.

In general, the paper emphasizes the need for a more nuanced approach to model validation, one that integrates

both power analysis and p-value distribution assessments. We argue that such an approach is particularly important in actuarial applications, where model risk carries significant financial implications. Our conclusions contribute to the growing body of literature urging a re-evaluation of traditional GoF testing practices, and we offer practical recommendations for practitioners and researchers engaged in risk modeling and statistical inference in the insurance and finance sectors.

2 Introduction

GoF tests play a critical role in actuarial science and financial risk management, where accuracy of statistical models is essential for predicting uncertainties such as insurance claims, stock price fluctuations, and operational risks. The validity of these models directly influences decision-making, risk assessment, and financial stability. GoF tests assist practitioners in evaluating how well a chosen probability distribution represents observed data, with particular attention to both the central tendencies and tail behaviors. This evaluation is especially important in high-stakes applications, such as loss modeling and catastrophic risk analysis, where rare but extreme events can severely affect financial reserves and the effectiveness of risk mitigation strategies.

Several GoF tests are commonly used in actuarial science, with the most prominent being the Anderson-Darling (A-D), Kolmogorov-Smirnov (K-S), and Chi-squared tests. Each of these tests provides a unique way of comparing the empirical distribution function (EDF) with the theoretical cumulative distribution function (CDF). The A-D test, introduced in 1952 by Theodore Anderson and Donald Darling, is particularly sensitive to deviations in the tails of distributions. This sensitivity makes it highly suitable for heavy-tailed distributions, which are prevalent in actuarial contexts where rare but extreme events, such as large insurance claims, can disproportionately impact financial reserves. By placing more weight on the tails, the A-D test rigorously assesses these critical regions, making it a favored choice for models involving extreme losses (Engmann and Cousineau).

In contrast, the K-S test, first developed by Andrey Kolmogorov in the 1930s and later extended by Nikolai Smirnov, measures the maximum absolute difference between the EDF and CDF. While useful for detecting general discrepancies between distributions, the K-S test is less sensitive to tail behavior than the A-D test, which limits its usefulness in scenarios where accurate tail modeling is crucial (Koyuncu and Karahasan). The Chi-squared test, introduced by Karl Pearson in 1900, compares observed and expected frequencies across different intervals of data. Despite its simplicity and wide use, it is less reliable for continuous distributions and small sample sizes, and it performs poorly in identifying tail anomalies. As a result, practitioners often use it alongside more tail-sensitive tests to improve model validation, particularly for heavy-tailed distributions (Ma and Ma).

Heavy-tailed distributions are crucial in actuarial science because they capture tail risks linked to rare but impactful events, such as large insurance claims or operational losses. Accurately modeling these risks is key for calculating reserves and ensuring financial stability during catastrophic events. Recent studies have highlighted the importance of heavy-tailed distributions in actuarial models. For instance, Bakar et al. (2015) emphasized that

these distributions help capture rare, high-impact events that significantly affect financial models. They discussed several heavy-tailed models, including the Pareto, lognormal, Weibull, and gamma distributions, which are used to assess risks not only in insurance but also in financial and network data modeling Resnick.

Similarly, Adamu (2024) highlighted the importance of selecting appropriate distributions for accurate risk forecasts, especially in the insurance sector, where large claims can dramatically impact reserves. By modeling tail risks with heavy-tailed distributions, actuaries can better estimate the necessary reserves to cover catastrophic losses.

Operational risk management also requires accurate selection of distributions to account for rare but severe losses, such as system failures, fraud, or external catastrophic events. Feuerverger (2016) emphasized the significance of addressing these low-frequency, high-impact events, which can cause considerable operational disruptions and financial losses. In this context, GoF tests are essential for evaluating whether a chosen statistical distribution fits empirical data. These tests are especially valuable when dealing with distributions that exhibit heavy tails, as they allow for a rigorous assessment of tail behavior, which is crucial for loss modeling and operational risk analysis Wang and Zhu (2024).

Table 1 provides a categorized overview of key research studies that have utilized GoF tests in actuarial science, with a focus on insurance and loss modeling. The studies are grouped based on the application of similar GoF tests, including A-D, K-S, and Chi-squared tests, along with an additional test, Cramér-von Mises (CvM), which will be discussed in more detail later. Beyond the GoF tests, the table also highlights important details such as sample sizes, supplementary statistical methods (e.g., AIC, BIC), and whether p-values were reported, all of which are critical for assessing the robustness of the studies and the methodologies used to validate actuarial models.

The first group of studies in Table 1 focuses on the combined use of A-D, K-S, and Chi-squared tests, often supplemented with model selection criteria such as AIC, BIC, and log-likelihood. Miljkovic and Grün (2016) applied these GoF tests to analyze Danish fire insurance data comprising 2,492 observations. They compared finite mixture models with K components—estimated using the Expectation-Maximization (EM) algorithm, based on distributions such as Burr, Gamma, Inverse Burr, Inverse Gaussian, Log-normal, and Weibull, against previously established composite Weibull models that had been considered the best fit for this dataset. Their goal was to identify the model that best captured not only the body of the distribution but also its tail, as tail behavior is critical for assessing extreme losses. None of the GoF tests rejected the null hypothesis at the 5% significance level, indicating that the fitted distributions were appropriate representations of the population.

Similarly, Eling and Loperfido (2017) used a combination of A-D, K-S, and Chi-squared tests to analyze cyber risk data (2,266 observations) from the Privacy Rights Clearinghouse (PRC). Their study combined exploratory methods, such as multidimensional scaling, with confirmatory approaches to evaluate the fit of various distributions. For data breach frequency, they tested distributions such as the negative binomial, Poisson, and generalized Pareto distribution (GPD), with the negative binomial showing the best fit according to the K-S test. For breach severity, parametric distributions such as log-normal, exponential, and Weibull were assessed, with the log-normal

distribution emerging as the best fit based on the K-S and A-D tests. The Chi-squared test was used to compare observed and expected frequencies across categories, further validating the models. Their findings suggested that alternative distributions like the skew-normal may provide an even better fit for modeling breach severity.

The second group in Table 1 highlights studies that primarily used the A-D and K-S tests only, without reliance on the Chi-squared test. For example, Ma and Ma (2013) analyzed catastrophe loss data and explored the limitations of the Chi-squared test, particularly for small sample sizes (770 observations). They showed that in such cases, the K-S and A-D tests provide more reliable measures for assessing heavy-tailed distributions. By focusing on the global fit of the distribution without imposing strict distributional assumptions, the K-S test offered a non-parametric approach to evaluating model fit, while the A-D test's sensitivity to tail behavior made it an ideal tool for risk models involving heavy-tailed data.

Reynkens et al. (2017) extended this approach by using A-D and K-S tests to evaluate a splicing model for the Danish fire insurance dataset. The splicing model combined a mixed Erlang (ME) distribution for the body and a Pareto distribution for the tail, capturing both moderate and extreme losses. The model was tested using the K-S and A-D statistics, along with graphical tools like QQ-plots and survival plots, to compare the empirical and fitted distributions. A bootstrap approach was employed to compute p-values for the K-S and A-D tests, indicating that the models, including the ME-Pareto, fit the data well. Ultimately, the ME-Pareto model was recommended based on its superior trade-off between fit quality and parameter complexity. This method was further supported by Ahmad et al. (2020a), who demonstrated that the Z-Weibull model outperformed competing models for earthquake insurance data, making it a strong candidate for modeling heavy-tailed insurance claims.

The third group introduces the Cramér-von Mises (CvM) test, which provides a more balanced assessment across the entire distribution, complementing A-D and K-S tests. According to Laio (2004), Koyuncu and Karahasan (2024), the CvM test, introduced by Harald Cramér and Richard Edler von Mises between 1928 and 1930, measures the squared distance between the empirical and theoretical cumulative distribution functions. Unlike the K-S test, which focuses on the maximum deviation between the distributions, the CvM test offers a more balanced assessment across the entire distribution. This broader scope makes it a valuable complement to the tail sensitivity of the A-D test, providing a more comprehensive view of the model's overall fit.

Ahmad et al. (2020b) integrated the CvM test alongside A-D and K-S tests in their study of vehicle insurance loss data. The study proposed the Weighted T-X Weibull (WT-XW) distribution as a new heavy-tailed model for fitting insurance loss data, offering more flexibility than the traditional Weibull distribution. The model's performance was assessed using 500 Monte Carlo simulations with sample sizes ranging from 25 to 500, and parameters were estimated via Maximum Likelihood Estimation (MLE). The results demonstrated that the WT-XW distribution outperformed competing models in all GoF tests and selection criteria, providing the best fit for the vehicle insurance data. Additionally, actuarial measures such as Value at Risk (VaR), Tail Value at Risk (TVaR), and Tail Variance (TV) confirmed the WT-XW distribution's heavier tail, making it a strong candidate for modeling heavy-tailed insurance losses.

Afify et al. (2020) also used the CvM test in his analysis of the newly proposed Alpha Power Exponentiated Exponential (APExE) distribution for modeling heavy-tailed insurance data. The APExE model was applied to a dataset of monthly unemployment insurance metrics from Maryland, USA, consisting of 58 observations. A-D, K-S, and CvM tests were employed alongside discrimination measures such as AIC, BIC, HQIC, and CAIC to compare the APExE distribution against 15 competing models. Monte Carlo simulations were conducted to evaluate the performance of different parameter estimation methods, including MLE, OLSE, WLSE, ADE, CVME, and PE. The APExE model was found to outperform all other models, providing the best fit for the unemployment insurance dataset. Graphical tools such as PP-plots and histograms further supported these findings, demonstrating the APExE model's flexibility and accuracy in modeling heavy-tailed insurance data.

Brazauskas and Serfling (2003) also used the CvM, A-D, and K-S tests in their analysis. Their paper evaluated the performance of robust estimators, specifically Generalized Median (GM) and Trimmed Mean (T) types, for fitting Pareto models using three real datasets: Wind Catastrophes (40 observations), OLT Bodily Injury Liability Claims (from 1976), and Norwegian Fire Claims (142 observations from 1975). The study concluded that GM and T-type estimators outperformed traditional methods like MLE by consistently providing better model fits across all datasets. Estimators were ranked based on GoF statistics, and those achieving smaller ranks in at least two out of the three GoF tests were considered superior. GM-type estimators, especially with parameters k=3 and k=4, demonstrated the best overall performance, consistently surpassing both MLE and other estimation methods in terms of robustness and efficiency. Thus, the GM and T-type estimators were identified as the preferred choices for fitting Pareto models in real-world insurance datasets.

Finally, the last group of studies in Table 1 examines cases where a single GoF test was employed to assess model performance. For example, Keatinge (1999) used the Chi-squared test to evaluate the goodness-of-fit for the Mixed Exponential distribution compared to traditional models like the Pareto distribution. Simulations were conducted with sample sizes of 10, 50, and 250 observations to test the model's ability to capture heavy-tailed loss data. The results indicated that the Mixed Exponential distribution outperformed the Pareto distribution, particularly in modeling both moderate and extreme losses. The Chi-squared test confirmed the Mixed Exponential's superior flexibility, making it a more reliable choice for actuarial loss data.

Similarly, Huang and Meng (2020) employed the K-S test to validate a Bayesian nonparametric model designed to predict insurance losses across three datasets: 4,624 non-zero claim samples from an Australian dataset, 15,390 claims from the French MTPL dataset, and 2,151 policies with 376 claims from a Chinese UBI dataset. The model's performance was measured using the D-statistic, which calculates the deviation between the empirical and theoretical distributions across risk classes. Results showed that the Bayesian model consistently achieved the lowest D-statistic, indicating the best fit in 28 of 48 classes in the Australian dataset and 634 of 1,492 classes in the French dataset. This demonstrated the model's superior flexibility and accuracy in fitting complex, heavy-tailed insurance losses compared to other models.

Although goodness-of-fit (GoF) tests are widely used for model selection and validation in actuarial science, their heavy reliance on *p*-values has sparked substantial debate within the statistical community. Traditionally, *p*-values are used to determine statistical significance—often with the conventional threshold of $p < 0.05$—but this practice has been increasingly criticized for encouraging oversimplified interpretations. Concerns have been raised about issues such as *p*-hacking, selective reporting, and over-reliance on arbitrary thresholds, which can ultimately lead to misleading conclusions—even when *p*-values are derived from GoF tests assessing model adequacy.

As highlighted by Wasserstein and Lazar (2016), *p*-values are frequently misunderstood and misapplied, often treated as definitive evidence for or against a hypothesis. In response, the American Statistical Association (ASA) issued a landmark statement cautioning against the sole use of *p*-values for scientific decision-making. According to the ASA, *p*-values do not measure the probability that a hypothesis is true, nor do they reflect the magnitude or practical importance of an effect. Instead, they simply indicate the extent to which the observed data are consistent with a specified model. This position has led to a broader call for decision-making frameworks that extend beyond rigid *p*-value thresholds and incorporate richer inferential perspectives.

To address these limitations, researchers have proposed a variety of alternatives and enhancements to traditional *p*-value usage. For example, Benjamin and Berger (2019) recommend lowering the threshold for statistical significance from $p < 0.05$ to $p < 0.005$ in the context of new discoveries, arguing that stricter criteria can help reduce false positives. They also advocate the use of Bayes factor bounds to provide more context on the strength of evidence. Similarly, Blume et al. (2019) introduce the concept of Second-Generation *p*-values (SGPVs), which emphasize practical significance by identifying regions of indifference—thereby shifting focus from binary decisions to a broader interpretation of statistical evidence.

While some scholars push for reform, others argue for the careful retention of *p*-values. Murtaugh (2014), for instance, defends their use when applied appropriately and interpreted in conjunction with complementary tools like confidence intervals. Lakens (2021) echoes this view, suggesting that the real problem lies not in the *p*-values themselves but in how they are used. Both authors advocate for educational and methodological reforms that encourage responsible *p*-value interpretation rather than outright rejection.

At the more radical end of the spectrum, McShane et al. (2019) propose abandoning the conventional $p < 0.05$ threshold altogether. They argue that *p*-values should be treated as just one part of a broader inferential process—alongside considerations of study design, effect size, data quality, and prior evidence. This perspective reflects a growing consensus that statistical decision-making should be more holistic, especially in fields like actuarial science where model assumptions can significantly impact financial risk assessments.

To summarize the current landscape, **Table 2** groups recent research perspectives on *p*-value use into three main categories: those that advocate for retention, reform, or abandonment. The table outlines the various proposals, suggested thresholds, and key recommendations from leading statisticians. This synthesis highlights the ongoing evolution of *p*-value interpretation and reinforces the importance of adopting more nuanced approaches—particularly in actuarial contexts where data complexity and decision stakes are high.

Motivated by this debate, the present study critically reassesses the role of p-values in GoF testing for actuarial applications. Specifically, it evaluates the performance of the A-D, K-S, Chi-Squared, and CvM tests when applied to heavy-tailed distributions relevant to actuarial modeling. The central research question is: **Can we rely solely on p-value reports from GoF tests in actuarial decision-making, or is there a need for a more comprehensive approach to model validation?** By exploring how these tests behave under different distributional settings and sample sizes, this paper advocates for a more robust inferential framework that integrates power analysis and the distribution of p-values alongside traditional GoF procedures.

The remainder of the paper is organized as follows: Section 3 outlines the methodological framework; Section 4 presents the design and execution of simulation studies used to evaluate the power of different GoF tests under varying conditions, including different sample sizes and distributional complexities. Section 5 discusses the results and their implications for actuarial modeling. Finally, Section 6 concludes with a summary of findings, recommendations for practice, and directions for future research.

Author	Method(s) Used	*p*-value Reported?	Test(s) Used	Sample Size(s)	Comments
Miljkovic and Grün (2016)	NLL, AIC, BIC	Yes	K-S, A-D, χ^2	2492	All three GoF tests indicate an appropriate fit for the population.
Eling and Loperfido (2017)	Log-likelihood, AIC	Yes	K-S, A-D, χ^2	2,266	Process similar to loss modeling; Chi-squared mentioned but not used in analysis.
Ma and Ma (2013)	CAIC, BIC	Not specifically stated	K-S, A-D	770	KS and AD used to check goodness of fit.
Reynkens et al. (2017)	NLL, AIC, BIC	Yes	K-S, A-D	2167	Bootstrap approach used for parameter estimation.
Ahmad et al. (2020a)	AIC, BIC, HQIC, CAIC, parametric bootstrapping	Yes	K-S, A-D	N/A	K-S with bootstrapped *p*-values for validation.
Afify et al. (2020), Ahmad et al. (2020b)	AIC, BIC, HQIC, CAIC	Yes	A-D, K-S, CvM	58	GoF measures show large *p*-values for better model fit.
Brazauskas and Serfling (2003)	MLE	Yes	A-D, CvM, K-S	40 / 90 / 142	Used to rank estimators.
Gui et al. (2018)	GEM-CMM, CV, Log-likelihood, BIC	Yes	A-D, CvM, K-S	2167 / 1500	Model adequacy tested after GEM-CMM algorithm.
Keatinge (1999)	No specific method stated	Yes	χ^2	10 / 50 / 250	Chi-square deemed inappropriate for mixed exponential distribution due to variable parameters.
Huang and Meng (2020)	MAE, RMSE, WAIC	Not specifically stated	K-S	4,624 / 15,390 / 376	Bayesian model had lowest KS statistic; unique calculation of median and mean KS statistics.
Eling (2012)	Log-likelihood, AIC	Yes	K-S	1500 / 2167	KS used as a final check after AIC comparisons to test the fit of the model.

Table 1: Grouped Summary of Statistical Methods and GoF Tests Based on Similar Tests Used

Author	Keep, Abandon, or Modify?	Comments
Keep *p*-value		
Murtaugh (2014)	Keep	Defends the use of p-values, noting that they, along with confidence intervals and AIC, have their place depending on the application.
Lakens (2021)	Keep	Argues that p-values, when correctly used, remain a valuable tool in well-controlled experiments, especially when testing ordinal claims.
Pawitan (2020)	Keep	Defends the p-value against critics and argues that the proposal to ban p-values should also apply to Bayes factors if consistency is the goal.
Modify *p*-value Usage		
Benjamin et al.	Modify	Proposes lowering the p-value threshold from 0.05 to 0.005 for new discoveries.
Imbens (2021)	Modify	Does not support banning p-values but prefers emphasizing confidence intervals or Bayesian intervals over p-values alone.
Blume et al. (2019)	Modify	Advocates for Second-Generation p-values (SGPV), focusing on ranges rather than binary significance thresholds. Emphasizes practical significance.
Abandon *p*-value Usage		
Wasserstein and Lazar (2016)	Abandon	Criticizes p-value misuse; emphasizes that p-values should not be the sole criterion for decision-making. Advocates for more comprehensive frameworks that include effect size and other measures.
Hubbard and Lindsay (2008)	Abandon	Recommends replacing p-values with replication research focusing on sample statistics and effect sizes for better knowledge development.
Trafimow (2019)	Abandon	Advocates for a ban on null hypothesis significance testing, including p-values, in psychology journals.
McShane et al. (2019)	Abandon	Suggests abandoning the $p < 0.05$ threshold altogether and treating p-values as one piece of evidence among others (e.g., study design, effect size).

Table 2: A Summary of Key Contributions and Debates Regarding the Reliability and Future of p-values in Statistical Reporting

3 Methodology

3.1 Distribution of *p*-values

To establish a robust understanding of the role of *p*-values in statistical testing, particularly within the context of goodness-of-fit tests in actuarial science, it is crucial to contextualize these concepts with theoretical underpinnings. This section provides a comprehensive overview of *p*-values, their relationships with test statistics, the influence of sample size, and the role of effect size. *p*-values are central to hypothesis testing, reflecting the probability of observing a test statistic as extreme as, or more extreme than, the value observed, assuming that the null hypothesis is true. Mathematically, the *p*-value, P, is expressed as:

$$P = \Pr(T > t) = 1 - F(T),$$

where $F(T)$ denotes the cumulative distribution function (CDF) of the statistic T. Hung et al. states that under the null hypothesis, the distribution of *p*-values should uniformly cover the interval $[0, 1]$, regardless of sample size. This uniformity is demonstrated below. The derivation confirms that if the null hypothesis is true, *p*-values are expected to be uniformly distributed across $[0, 1]$, leading to valid inferential conclusions about statistical evidence:

$$\begin{aligned}
\Pr(P \leq p) &= \Pr(1 - F(T) \leq p) \\
&= \Pr(F(T) \geq 1 - p) \\
&= \Pr(T \geq F^{-1}(1 - p)) \\
&= 1 - F(F^{-1}(1 - p)) \\
&= 1 - (1 - p) \\
&= p.
\end{aligned}$$

However, the interpretation of *p*-values is nuanced by the sample size employed in statistical analysis. Specifically, larger sample sizes yield more reliable estimates due to the Law of Large Numbers (Lem). Altman and Bland mention that as sample size n increases for a fixed effect size, the standard error diminishes as a result of reduced variation. This typically leads to an increase in the test statistic value and, correspondingly, a decrease in the *p*-value. For instance, when examining a random sample of observations, $X_1, X_2, \ldots, X_n$, where $X_i \sim N(\mu_t, \sigma^2)$, the sample mean $\bar{X}$ converges towards the population mean μ_t as n grows.

In conducting a one-sided z-test to determine if the population mean is significantly greater than a specified value, the hypotheses are formulated as:

$$H_0 : \mu_t \leq \mu_f, \quad H_a : \mu_t > \mu_f.$$

The relevant test statistic, given that the alternative hypothesis holds, is described by:

$$Z = \frac{\bar{X} - \mu_f}{\frac{\sigma}{\sqrt{n}}} \sim N\left(\frac{(\mu_t - \mu_f)\sqrt{n}}{\sigma}, 1\right).$$

As the sample size n increases, the distribution of the test statistic shifts, leading to increasingly lower *p*-values. For practical illustration, if the true difference between the hypotheses is substantial, as n approaches infinity, *p*-values tend to converge towards zero with high probability. This behavior is crucial for understanding that *p*-values are not only a measure of significance but also reflect the power of the test, that is the likelihood of correctly rejecting a false null hypothesis.

Moreover, the role of effect size cannot be overlooked when interpreting *p*-values. Tomczak and Tomczak (2014) mention that the effect size quantifies the magnitude of the difference between groups and provides valuable context that *p*-values alone cannot offer. While a small *p*-value might indicate statistical significance, it does not convey the practical significance or relevance of the observed effect. For example, consider two studies: one with a small effect size but a very large sample, resulting in a highly significant *p*-value; and another reporting a larger effect size but a non-significant *p*-value due to a small sample size. Hence, relying exclusively on *p*-values may lead to misleading conclusions regarding the importance of the findings. Effect size offers a standardized measure of the strength of a phenomenon or effect, facilitating comparisons across different studies and contexts. Common measures of effect size include Cohen's d, which assesses the difference between two means in terms of standard deviation, and the odds ratio, which quantifies the odds of an event occurring in one group compared to another. In the context of goodness-of-fit tests, reporting effect sizes alongside *p*-values can enhance the interpretability of results and provide clearer insights for decision-making processes.

Furthermore, it is essential to acknowledge the concern of "*p*-value hacking," where researchers may manipulate data or testing methods to achieve desirable *p*-values (Head et al.). This practice has raised questions about the reliability of *p*-values as benchmarks for statistical significance. As highlighted in recent literature, reliance solely on *p*-values without considering other statistical evidence can be misleading, particularly when addressing complex datasets or multivariate analyses. Consequently, adopting a comprehensive methodological framework that incorporates power analysis, Type I error rate assessment, effect size estimation, and the appropriate application of GoF tests is essential for drawing valid conclusions.

In summary, rigorously applying these principles not only underscores the importance of both *p*-values and effect sizes in statistical inference but also fosters a clearer understanding of how these metrics function within the framework of GoF tests. Acknowledging the pivotal influence of sample size and employing robust methodologies allows actuarial scientists to make informed decisions based on statistical evidence that is both reliable and meaningful.

3.2 Kolmogorov-Smirnov Goodness-of-Fit Test

As earlier stated, the K-S test serves as a fundamental tool for evaluating how well a theoretical distribution matches observed data. This nonparametric test measures the largest difference between the empirical CDF of a sample and the CDF of the assumed distribution. Essentially, it allows us to see where our data diverges most from the hypothesized model, highlighting potential areas of misfit.

Given a sample $X = \{X_1, X_2, \ldots, X_n\}$ of size n, the empirical distribution function $F_n(x)$ is calculated as follows:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(X_i)$$

Here, $I_{(-\infty,x]}(X_i)$ is an indicator function that equals 1 if $X_i \leq x$ and 0 otherwise (Moscovich-Eiger et al. (2013)). This function $F_n(x)$ represents the proportion of sample values that fall below any given x. The K-S statistic, D_n, captures the largest absolute difference between $F_n(x)$ and the theoretical CDF $F(x)$:

$$D_n = \sup_x |F_n(x) - F(x)|$$

where $\sup_x$ represents the maximum deviation across all values of x in the sample. This difference, D_n, provides a single measure of how closely our sample data matches the theoretical model. That is, if it's large, we have a strong indication that the sample might not come from the specified distribution (Darling (1957)).

The hypotheses for the K-S test are as follows:

- **Null Hypothesis (H_0)**: The sample data comes from the specified distribution $F(x)$.
- **Alternative Hypothesis (H_a)**: The sample data does not come from the specified distribution $F(x)$.

In our study, the K-S test is used to compare the distribution of observed data against a proposed heavy-tailed model. If the calculated D_n exceeds a critical value (derived from the Kolmogorov distribution) or if the *p*-value falls below a set significance level (e.g., $\alpha = 0.05$), we reject H_0 and conclude that the sample likely does not fit the specified distribution. This approach provides an objective basis for assessing distributional fit, particularly useful in our context with heavy-tailed data where misfit often occurs in the tail regions—critical areas in actuarial applications.

The K-S test, while widely used due to its nonparametric nature and ability to apply across various distributional forms, has both strengths and limitations. Moscovich-Eiger et al. highlights that the K-S test enjoys several desirable properties: it is asymptotically consistent against any fixed alternative and generally has good power for detecting shifts in the central tendency of the distribution (Janssen (2000)). Furthermore, the test's simplicity allows for straightforward computation of *p*-values, which has contributed to its popularity in practice. However, they also point out a significant limitation: the K-S test has relatively low power for detecting deviations in the tails of the distribution. This limitation is particularly relevant in cases involving rare contamination, where only

a few observations in the sample differ from the assumed model, or in contexts such as high-dimensional variable selection or multiple hypothesis testing under sparsity assumptions. For these cases, alternative or modified versions of the K-S test that focus more on the tails may provide more robust insights (Donoho and Jin (2004); Cai and Wu (2014)).

Lanzante (2021) further discusses the use of the K-S test for distributional testing in scenarios where the data exhibit dependencies, such as spatial or temporal correlation. While the K-S test is nonparametric and makes minimal assumptions about the underlying distribution, it assumes that sample values are statistically independent. When this assumption is violated, as is common in fields like climate science where observations may be spatially or temporally coherent, the test may yield excessive rejections of H_0. In cases like these, additional steps are necessary to address dependencies, either by adjusting the sampling framework or employing advanced methods to handle spatial and temporal coherence Wilks (2016).

In this study, the K-S test is implemented in R using the `ks.test` function, which calculates both the K-S statistic and *p*-value for our sample. The test is specifically applied to examine the fit of a heavy-tailed distribution to loss data, providing insights not only into general differences between the observed and hypothesized distributions but also into potential misfits in the tail regions. Such accuracy is essential in actuarial contexts where tail behavior can significantly impact risk assessment and decision-making. Although we are aware of the K-S test's limitations, especially in detecting tail discrepancies, we have chosen it for its versatility and because it offers a clear, objective metric for initial distributional assessment.

3.3 Anderson-Darling Goodness-of-Fit Test

The A-D test, another nonparametric test, is used to assess how well a sample aligns with a specified theoretical distribution. Like the K-S test, the A-D test compares the EDF of the sample with the CDF of the proposed distribution. However, the A-D test places more emphasis on the tails (Jäntschi and Bolboacă), making it particularly useful for analyzing heavy-tailed data where accurate tail behavior is crucial, such as in actuarial and risk modeling.

For a sample $X = \{X_1, X_2, \ldots, X_n\}$ of size n, the empirical distribution function $F_n(x)$ is calculated as follows:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(X_i)$$

where $I_{(-\infty,x]}(X_i)$ is an indicator function that equals 1 if $X_i \leq x$ and 0 otherwise. This empirical function $F_n(x)$ represents the proportion of sample values less than or equal to x, providing a stepwise cumulative view of the sample distribution.

According to Anderson (2011), the A-D statistic A^2 can be calculated using two approaches: an integral form and a summation-based formula.

The integral form of the A-D statistic is defined as:

$$A^2 = n \int_{-\infty}^{\infty} \frac{[F_n(x) - F(x)]^2}{F(x) \cdot (1 - F(x))} \, dF(x)$$

Here:

- n is the sample size,
- $F_n(x)$ is the EDF of the sample, as reiterated by Lewis (1961), and
- $F(x)$ is the CDF of the specified theoretical distribution.

In this expression, the entire integrand $\frac{[F_n(x)-F(x)]^2}{F(x)\cdot(1-F(x))}$ serves as a weight, giving extra emphasis to differences near the tails, where $F(x)$ approaches 0 or 1. This weight increases the sensitivity of the A-D test to discrepancies in the tails, which is essential for assessing fit in heavy-tailed data.

For an ordered sample, the A-D statistic A^2 can also be calculated using a summation-based formula:

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i - 1) \left[\log(u_{(i)}) + \log(1 - u_{(n-i+1)}) \right]$$

where $F(X_i)$ represents the cumulative probability at each ordered data point X_i,

This summation-based approach applies a similar emphasis on tail values. The logarithmic terms magnify differences between the EDF and CDF at the extremes of the distribution, reflecting the A-D test's focus on tail behavior. The A-D test evaluates the same hypotheses as the K-S test.

A large value of A^2 or a low *p*-value suggests a poor fit, especially in the tails, indicating that the sample distribution may differ significantly from the hypothesized model. For this study, the A-D test was implemented in R using the `ad.test` function, which computes both the A-D statistic and the corresponding *p*-value. This test provides a powerful complement to the K-S test by emphasizing tail fit, which is crucial for accurately modeling risk in heavy-tailed data. By examining discrepancies in the tails, the A-D test offers a nuanced measure of how well the proposed model captures the extreme values in the data, which are particularly significant in actuarial applications.

3.4 Chi-Square Goodness-of-Fit Test

The Chi-square (χ^2) GoF test is another nonparametric method widely used to assess whether observed categorical data align with a theoretical distribution. Unlike the K-S and A-D tests, which are designed for continuous data, the χ^2 test requires data to be grouped into distinct categories (Rolke and Gongora). This grouping, known as binning, can lead to some information loss as finer details within the data are obscured. Consequently, the test has traditionally been considered less powerful than other GoF tests, particularly when used for continuous distributions (Cirrone et al.). However, the χ^2 test remains valuable for categorical data, where well-defined categories improve interpretability.

For a dataset divided into k categories, the χ^2 test compares the observed frequencies O_i in each category to the expected frequencies E_i under the hypothesized distribution (Shankar). The chi-square statistic χ^2 is calculated as:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where:

- O_i represents the observed frequency for category i,
- E_i is the expected frequency for category i, and
- k is the total number of categories.

This statistic measures the cumulative squared differences between observed and expected frequencies, with larger values indicating greater divergence from the expected distribution. If χ^2 surpasses a critical threshold, it suggests that the observed data might not follow the hypothesized distribution.

The hypotheses for the chi-square test are as follows:

- **Null Hypothesis (**H_0**)**: The observed frequencies align with the expected frequencies under the specified distribution.
- **Alternative Hypothesis (**H_a**)**: The observed frequencies differ from the expected frequencies, suggesting a departure from the hypothesized distribution.

The degrees of freedom for the chi-square test are calculated as $df = k-1$, where k is the number of categories. This degrees-of-freedom value, along with the calculated χ^2 statistic, is used to assess the test result against a critical value from the chi-square distribution.

Rolke and Gongora (2021) emphasize that the performance of the χ^2 test heavily depends on the binning strategy. While arbitrary binning can reduce power, modern approaches such as adaptive binning can substantially improve the test's performance. For instance, bins chosen to balance equal sizes and equal probabilities or those aligned with specific alternative hypotheses enhance the ability to detect deviations. Moreover, ensuring a minimum expected frequency ($E_i \geq 5$) in each bin stabilizes the test statistic, improving Type I error control. To ensure the validity of the chi-square test, expected frequencies are calculated based on the hypothesized distribution, providing a benchmark for comparison. In actuarial contexts, these might reflect historical data or projected risk factors across categories, such as different types of insurance claims.

Once the expected and observed frequencies are determined, the chi-square statistic χ^2 is calculated by summing the squared differences between observed and expected counts across all categories. Larger values of χ^2 suggest stronger deviations from the hypothesized distribution, which could signal a misalignment. The calculated χ^2 statistic is then compared to a critical value from the chi-square distribution table, determined by the

degrees of freedom. Alternatively, a *p*-value associated with χ^2 is computed. If χ^2 exceeds the critical value or the *p*-value falls below a chosen significance level (e.g., $\alpha = 0.05$), we reject H_0. This suggests that the observed data differ significantly from the hypothesized distribution, potentially reflecting changes in underlying patterns.

In this study, the chi-square test complements the K-S and A-D tests by addressing categorical distributions. While the K-S and A-D tests assess fit without requiring data binning, the chi-square test is particularly useful for examining discrete, interpretable categories. Recent advancements in binning strategies demonstrate that a well-designed binning approach can make the chi-square test competitive for continuous data. This layered approach provides a robust assessment of fit across both continuous and categorical data, ensuring comprehensive validation of distributional assumptions in actuarial applications.

3.5 Cramér–von Mises Goodness-of-Fit Test

The CvM test provides another approach to assessing how well a sample conforms to a specified distribution. Like the K-S and A-D tests, the CvM test compares the EDF of the sample to the CDF of the hypothesized distribution. However, while the K-S test focuses on the maximum deviation and the A-D test places extra weight on the tails, Laio mentions the CvM test assesses the GoF across the entire distribution by evaluating the squared differences between the EDF and CDF. This characteristic makes the CvM test sensitive to discrepancies throughout the distribution, rather than focusing only on extremes.

For a sample $X_1, X_2, \ldots, X_n$ with EDF $F_n(x)$ and a specified theoretical CDF $F(x)$, the CvM statistic W^2 is defined as:

$$W^2 = n \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 \, dF(x)$$

This statistic is very similar to the integral form of the A-D statistic stated above, but the denominator in this integrand is 1. In practice, Dickhaus and Dickhaus emphasizes this integral is approximated with the ordered data points $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ as:

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - F(X_{(i)}) \right)^2$$

where $X_{(i)}$ represents the i-th ordered observation. This formula captures the overall difference between the empirical and theoretical distributions, effectively averaging the discrepancies across the data range. The CvM test is structured around the same hypothese as the K-S and A-D tests. Darling (1957) mentions that a higher W^2 value indicates a stronger deviation from the theoretical distribution, suggesting that the observed data do not match the hypothesized model.

To perform the CvM test, we first compute the EDF $F_n(x)$ from the observed data. Then, using the ordered values $X_{(i)}$, we calculate the test statistic W^2 as described above. Finally, we compare W^2 to critical values from the CvM distribution or calculate the *p*-value. If the *p*-value falls below a chosen significance level (e.g.,

$\alpha = 0.05$), we reject the null hypothesis, concluding that the observed data do not align with the hypothesized distribution.

In this study, the CvM test is applied as part of a broader goodness-of-fit evaluation, complementing the K-S, A-D, and chi-square tests. By focusing on deviations across the entire distribution, the CvM test provides a balanced measure that highlights any inconsistencies between the observed and expected distributions without emphasizing any particular region of the data. This approach enhances our ability to validate the distributional assumptions for actuarial applications, offering a more nuanced assessment of fit alongside the other tests.

4 Simulation Study

This simulation study evaluates the performance of the K-S, A-D, CvM and Chi-squared GoF tests across two critical actuarial models: the one-component left-truncated Lognormal model and the two-component Burr mixture model. These models represent two key challenges in actuarial science: left truncation in loss distributions and modeling heavy-tailed behavior. The aim is to assess the reliability of these GoF tests in *p*-value-based decision making, specifically their ability to evaluate model adequacy and detect subtle misspecifications.

The study employs a simulation framework in which data are generated under two key conditions. Under the null hypothesis, the data is sampled from a correctly specified mixture model. Under the alternative hypothesis, data are generated from a slightly misspecified model, introducing minor deviations in the parameters. This setup allows us to examine the ability of each test to maintain Type I error rates under the null hypothesis while also evaluating its statistical power under the alternative. By testing how these GoF methods respond to model adequacy and misspecification, we aim to provide insights into their comparative strengths, particularly for tail-sensitive and left-truncated datasets encountered in actuarial applications.

4.1 Motivation for Parameter Deviations

In this study, the alternative hypothesis is introduced by making adjustments to the parameters of the 1-component Lognormal model and the 2-component Burr mixture model to simulate realistic model misspecifications. The goal is to evaluate the sensitivity of various GoF tests to deviations that might be encountered in practice, without resorting to artificially large changes that would trivialize model comparison. To guide the choice of parameter shifts, we adopt the framework of local model misspecification, where the true distribution differs only slightly from the assumed model. This approach is supported by Bonhomme and Weidner (2022), who emphasize the importance of designing inference procedures that remain valid under small, structured departures from the reference model.

In our case, ”minor deviations” refer to systematic parameter changes in the range of approximately 3–8%, affecting shape and scale parameters. These values were chosen to reflect a balance between being small enough to mimic routine estimation uncertainty (as typically observed in maximum likelihood estimation under moder-

ate sample sizes) and large enough to pose a realistic challenge to the sensitivity of GoF tests. This range is consistent with the magnitude of deviations examined in global sensitivity analysis Saltelli et al. (2008) and in simulation-based robustness studies such as Xiang et al. (2022), where moderate shifts are employed to reflect model uncertainty without introducing trivial detectability.

These deviations were tailored to probe dimensions critical to actuarial risk modeling. In the Burr mixture model, modifications to shape and scale parameters alter the heaviness of the tail and the concentration of extreme values, which are especially consequential for capital estimation and solvency analysis. Likewise, in the left-truncated Lognormal model, increasing the standard deviation from 0.5 to 1 introduces substantial tail expansion without significantly shifting the central tendency, which is precisely the kind of structural perturbation actuaries may face in practice. By simulating under these slightly misspecified conditions, we assess the ability of classical GoF tests to detect meaningful but non-obvious deviations.

Furthermore, the magnitude of parameter shifts was deliberately calibrated to keep the null and alternative distributions overlapping but distinguishable. This ensures that rejections of the null hypothesis, when they occur, are based on substantive distributional divergence rather than exaggerated or unrealistic discrepancies. In particular, we observe that some tests, such as the Anderson-Darling and Cramér-von Mises, exhibit greater sensitivity to tail-based deviations, while others such as the Kolmogorov-Smirnov are less responsive in these contexts. These patterns validate the use of moderate deviations to meaningfully differentiate the performance of competing GoF tests.

The design choice to focus on small, interpretable deviations ensures that our conclusions speak directly to real-world concerns around model adequacy and test selection. By grounding our parameter modifications in the literature on local misspecification and simulation-based sensitivity analysis, we provide a principled basis for understanding when and why specific GoF tests are likely to succeed—or fail—in detecting practically relevant model errors.

4.2 Model Specifications and Simulation Setup

The 2-component Burr mixture model used in this study is parameterized to capture the skewed and heavy-tailed characteristics typical of actuarial datasets. The probability density function (PDF) for the mixture model is defined as:

$$f(x|\theta) = \pi_1 f_{\text{Burr}}(x|\alpha_1, \theta_1, \gamma_1) + \pi_2 f_{\text{Burr}}(x|\alpha_2, \theta_2, \gamma_2),$$

where π_1 and $\pi_2 = 1 - \pi_1$ represent the weights (mixture proportions), and $f_{\text{Burr}}(x|\alpha, \theta, \gamma)$ is the PDF of the Burr distribution given by:

$$f_{\text{Burr}}(x|\alpha, \theta, \gamma) = \frac{\alpha\gamma \left(\frac{x}{\theta}\right)^{\gamma-1}}{\theta \left(1 + \left(\frac{x}{\theta}\right)^{\gamma}\right)^{\alpha+1}}.$$

This mixture model, including parameters, are adapted from the paper published by Miljkovic and Grün (2016) in which they propose modeling insurance losses using K-component finite mixture models. The first component is characterized by parameters $\alpha_1 = 0.207175$, $\theta_1 = 1.236993$, and $\gamma_1 = 7.047898$, with a weight of $\pi_1 = 0.397634$.The second component, on the other hand, uses parameters $\alpha_2 = 0.028161$, $\theta_2 = 0.856898$, and $\gamma_2 = 50.277542$, with a weight of $\pi_2 = 0.602366$. These parameter values are informed by the empirical dataset, the Danish Insurance Fire losses in R, using the package **SMPracticals**, which provides a realistic foundation for evaluating GoF tests. To generate random samples from the 2-component Burr mixture model, a two-step process was employed. First, a Bernoulli trial was performed for each observation using the `rbinom` function, where the probability of success (π_1) determined whether the sample was drawn from the first component. For both components, observations were generated from a Burr distribution using the parameters listed above. This approach ensured that the simulated data reflected the probabilistic composition of the mixture model. For the alternative hypothesis, slight parameter deviations were introduced. Specifically, the parameters for first component are altered to $\alpha_1 = 0.2$, $\theta_1 = 1.2$, and $\gamma_1 = 7.1$, while the second component's are modified to $\alpha_2 = 0.02$, $\theta_2 = 0.9$, and $\gamma_2 = 50.4$.

The 1-component left-truncated lognormal model, as used by Blostein and Miljkovic (2019) is the second model considered for our simulation study. The dataset used in the study is the Secura Belgian Re automobile claim dataset in R, using the package **ReIns**. This dataset consists of 371 claims from several European insurers for the period 1988 to 2001. The claim amounts are adjusted for inflation to reflect their 2002 Euro value and are left-truncated at 1.2 million euros. Using the `ltmix` package, the left-truncated one-component Lognormal model is fitted to the dataset. The CDF of the left-truncated data is given by:

$$g(x, l|\theta) = \begin{cases} 0 & \text{if } x < l, \\ \frac{f(x|\theta)}{1-F(l|\theta)} & \text{if } x \geq l, \end{cases}$$

where $f(x|\theta)$ and $F(x|\theta)$ are the probability density and cumulative density functions of the untruncated data, respectively, and l is the truncation point.

The PDF of the 1-component left-truncated Lognormal model is parameterized by:

$$h(x, l|\Phi) = \tau_1 g(x, l|\theta_1),$$

where $\tau_1 = 1$ since it is a 1-component model, and $\theta_1 = (\mu, \sigma)$ are the parameters of the Lognormal distribution. The parameters for the null hypothesis were estimated as $\mu = 14.3258849$ and $\sigma = 0.5014714$, while the slightly misspecified alternative model assumes $\mu = 14$ and $\sigma = 1$. Random samples are generated under the two conditions, just as stated above.

The simulation design spans a range of sample sizes, including $n = 10, 50, 100, 500, 1000, 2000,$ and 2500, to analyze the tests' performance across varying data volumes. For all configurations, significance levels are

maintained. To ensure robust estimates of Type I error and power, 2000 simulations are performed for each sample size and model condition. The K-S, A-D, CvM and Chi-squared tests are applied to each sample, and the resulting *p*-values are recorded. Significance levels of $\alpha = 0.01$, $\alpha = 0.05$, and $\alpha = 0.10$ are now considered. Type I error rate is defined as the proportion of simulations under the null hypothesis yielding *p*-values below the significance level, representing the tests' false positive rate when the model is correctly specified. Power, on the other hand, is the proportion of simulations under the alternative hypothesis where the test correctly rejects the null hypothesis, indicating its effectiveness in identifying model misspecification. Additionally, the minimum sample size required to achieve optimal power (Power $= 1$) is determined for each test at all significance levels.

5 Results and Analysis

The simulation results provide insights into the performance of each GoF test in terms of Type I error control and power. Figures and tables present the results, detailing Type I error rates and power across various sample sizes for the K-S, A-D, CvM and Chi-squared tests.

5.1 2-component Burr Mixture Model

To illustrate the distributional characteristics of the 2-component Burr mixture model, the density plot was generated for the individual components as well as the overall mixture distribution. As shown in Figure 1, the first component (blue dashed line) represents moderate losses, characterized by a broader, gentler peak and heavier tails, while the second component (red dashed line) captures extreme losses with a sharper peak and quicker decay.

The mixture distribution (black solid line), which is the weighted sum of the two components, integrates these distinct features, effectively capturing the variability in both moderate and extreme losses. The mixture skews closer to the second component due to its higher weight **(0.602366)**, yet retains the influence of the first component **(0.397634)**, resulting in a more comprehensive representation of the underlying data. This visualization highlights the flexibility and suitability of the Burr mixture model for analyzing heavy-tailed data commonly encountered in actuarial applications. It also provides a framework for understanding the underlying distribution that the GoF tests aim to assess, emphasizing the model's ability to capture both moderate and extreme loss behaviors.

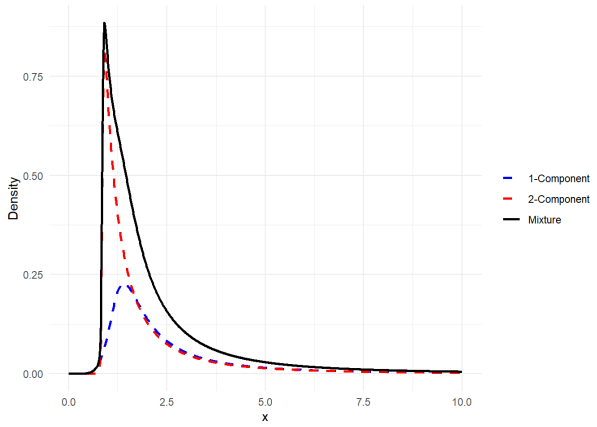

Figure 1: Density Plot of 1-Component, 2-Component and Mixture Burr

Kolmogorov Smirnov

As depicted in Figure 2, for smaller sample sizes ($n = 10, 100$), the *p*-values approximate a uniform distribution over the interval $[0, 1]$, aligning with theoretical expectations under the null hypothesis. This uniformity becomes more pronounced as sample sizes increase ($n = 500$ and above), indicating the test's robustness with larger datasets. The red vertical line at $p = 0.05$ represents the significance threshold. The proportion of *p*-values below this threshold corresponds to the nominal Type I error rate of 5%, confirming that the KS test maintains an appropriate error rate under the null hypothesis. For larger sample sizes, the histograms exhibit smoother distributions, further demonstrating the test's validity in adhering to its theoretical behavior. In contrast, Figure 3 shows the distribution of *p*-values under the alternative hypotheses shifted significantly toward zero. This shift indicates that the K-S test effectively rejects the null hypothesis when the alternative model is true. The clustering of *p*-values near zero becomes more pronounced as the sample size increases, reflecting the growing power of the test. For smaller sample sizes ($n = 10, 100$), there is more variability in the *p*-value distributions, suggesting lower power to detect deviations from the null hypothesis. However, as sample sizes increase ($n = 500, 1000, 2000, 2500$), all *p*-values fall below the critical threshold, with increased cluster near zero, confirming the test's increasing sensitivity with larger sample sizes.

Figure 4 provides a comprehensive visualization of the power curves for the K-S test across three significance levels ($\alpha = 0.01, 0.05, 0.1$) as a function of sample size. The power increases consistently with sample size for all alpha levels, eventually reaching 1 (perfect power). The rate of increase and the sample size required to achieve $power = 1$ depend on the alpha level. For $\alpha = 0.1$, the test achieves perfect power at $n = 650$, while $\alpha = 0.05$ and $\alpha = 0.01$ require $n = 710$ and $n = 840$, respectively. The power curves also highlight the trade-off between significance level and the sample size needed for effective detection of deviations from the null hypothesis. Higher alpha levels ($\alpha = 0.1$) result in steeper power curves, enabling faster detection of departures from the null hypothesis. However, they require careful consideration of the increased Type I error rates. These observations align with the known behavior of the K-S test, where the power to detect differences increases with

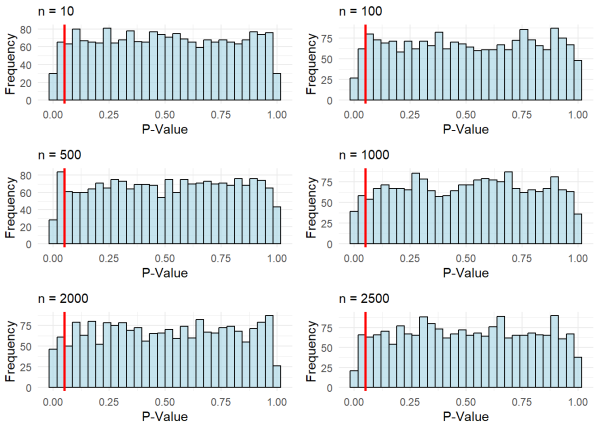

Figure 2: Sampling Distribution of KS test under Null Hypothesis

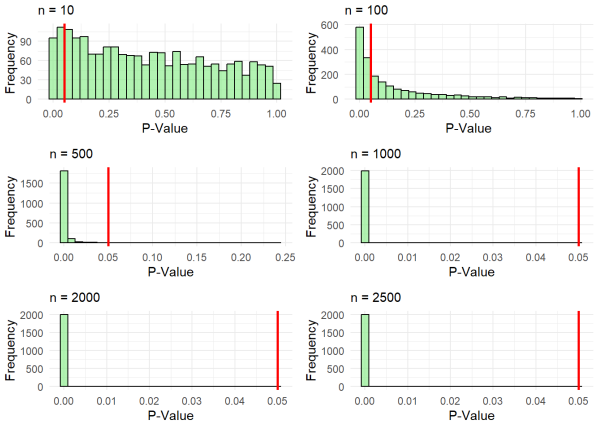

Figure 3: Sampling Distribution of KS test under Alternative Hypothesis

sample size and chosen significance level (Razali et al.).

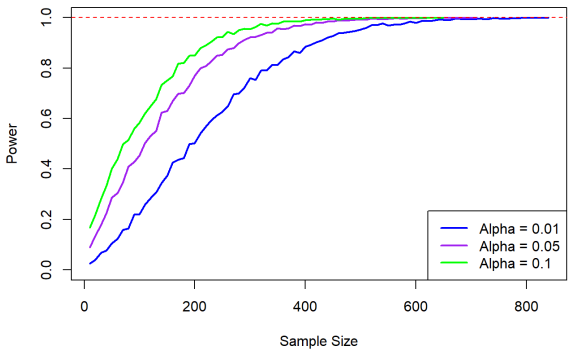

Figure 4: Power Curve of KS test at Different significance levels

Table 3 summarizes the Type I error rates and power for varying sample sizes at each alpha level. The Type I error rates remain close to their theoretical values under the null hypothesis, validating the K-S test's ability to maintain its expected behavior. The power increases consistently with sample size across all alpha levels. For smaller sample sizes ($n = 10, 50, 100$), the power is relatively low, but as the sample size increases ($n =$

$500, 1000, 2000, 2500$), the power approaches 1, particularly for higher alpha levels. These results further confirm the effectiveness of the KS test in rejecting the null hypothesis when the alternative is true.

Miljkovic and Grün (2016) reports a KS test statistic of **0.01591** with a *p*-value of **0.5537**, suggesting that the 2-component Burr mixture provides a satisfactory fit to the Danish dataset. In our simulation study, the sampling distribution of K-S test *p*-values under the null hypothesis consistently maintained Type I error rates close to the nominal 0.05, even as the sample size increased. This consistency reinforces the validity of the model fit observed in the paper. Additionally, the power results from our simulation under the alternative hypothesis show a clear trend of improved model sensitivity with larger sample sizes, supporting the reliability of the K-S test for detecting deviations from the null model. Together, these findings confirm the robustness of the conclusions reported in the paper when applied to the same Danish dataset with a large sample size of 2492 losses.
However, for the small sample size used by Afify et al. (2020) on unemployment insurance data (58 observations), the reliability of the K-S test findings is limited, as shown by our simulations. With a power of only 0.2705 at $n = 50$, the test struggles to detect differences under the alternative hypothesis, despite the Type I error rate staying near 0.05. This highlights the challenges of drawing robust conclusions from small datasets, where the K-S test has limited sensitivity. To address this, Afify et al. combines the K-S test with the CvM and AD tests to rank estimators. This integrated approach helps offset the limitations of individual tests by leveraging their different focuses, providing a more comprehensive evaluation despite the constraints of small sample sizes.

Anderson-Darling Test

Extending the procedure used and results observed for the K-S test, we perform the simulation for the A-D test. The trend is very similar to what was observed for the K-S test. Figure 5 illustrates the distribution of *p*-values generated by the A-D test under the null hypothesis. For smaller sample sizes ($n = 10, 100$), the *p*-values closely follow a uniform distribution over the interval $[0, 1]$, which aligns with theoretical expectations under the null hypothesis, just like for the K-S test. As the sample size increases ($n = 500$ and above), the uniformity of the *p*-value distribution also becomes more pronounced, and the histograms exhibit smoother and more consistent patterns. Larger datasets ($n = 1000, 2000, 2500$) reinforce the A-D test's robustness in maintaining its theoretical properties, confirming its adherence to the null hypothesis. This behavior is similar to that of the K-S test but is particularly noteworthy for the A-D test because it places greater weight on the tails of the distribution. This tail sensitivity does not disrupt the uniformity of the *p*-values under the null hypothesis, demonstrating the test's ability to maintain proper control of Type I errors while accounting for the extremes in the data. The proportion of *p*-values below the red line aligns with the expected 5% Type I error rate, further validating the A-D test's reliability in controlling false positives. These results, consistent across all sample sizes, align with existing research, which confirms the uniform distribution of *p*-values for goodness-of-fit tests under the null hypothesis, particularly for tests like AD that integrate a focus on the tails of the data.

Under the alternative hypothesis, the A-D test demonstrates significant shifts in the distribution of *p*-values

toward zero, effectively rejecting the null hypothesis when deviations exist. Figure 6 depicts the behavior of p-values for varying sample sizes. For small sample sizes ($n = 10, 100$), the A-D test exhibits a noticeable clustering of p-values below the critical threshold of $p = 0.05$. However, there remains some variability, reflecting limited power to detect deviations when data is sparse. This behavior is consistent with the K-S test but is mitigated more effectively by the A-D test due to its greater emphasis on tail behavior. As the sample size increases ($n = 500, 1000, 2000, 2500$), the p-value distributions shift heavily toward zero, with a substantial proportion of p-values falling below $p = 0.05$. This trend highlights the increasing power of the A-D test as more data becomes available. Compared to the K-S test, the A-D test demonstrates superior sensitivity, as is the general case, particularly in datasets where tail behavior is critical. By weighting the tails more heavily, the A-D test excels at identifying deviations in scenarios where extreme values carry significant implications, such as in modeling heavy-tailed actuarial data. These findings support existing studies emphasizing the A-D test's power advantages in detecting departures from the null hypothesis, especially when deviations are concentrated in the distribution's tails.

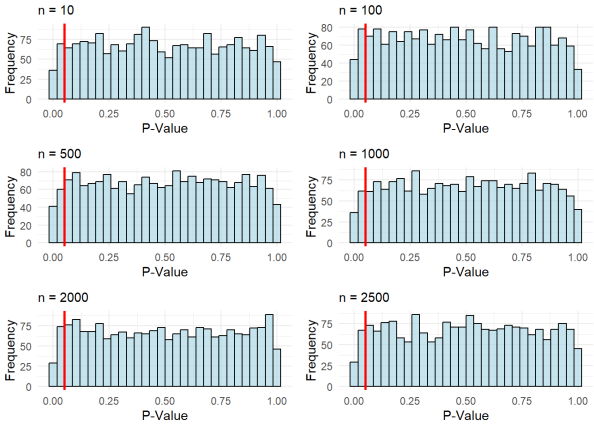

Figure 5: Sampling Distribution of AD test under Null Hypothesis

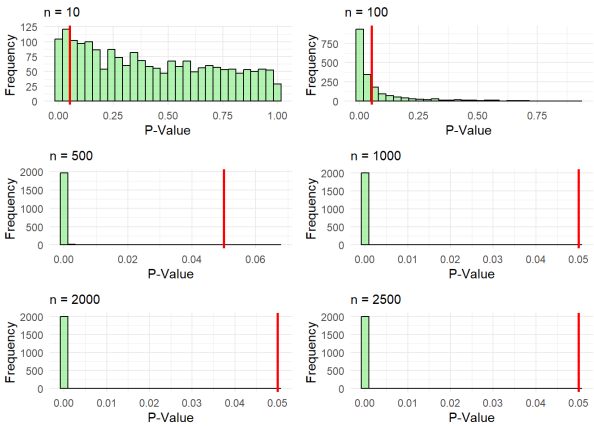

Figure 6: Sampling Distribution of AD test under Alternative Hypothesis

The power curves in Figure 7 further illustrate the superior performance of the A-D test across three significance levels ($\alpha = 0.01, 0.05, 0.1$) as sample size increases. Similar to the K-S test, the power of the A-D test rises

consistently with sample size, reaching 1.0 (perfect power) for large datasets. However, the A-D test achieves perfect power at smaller sample sizes compared to the K-S test, emphasizing its greater sensitivity. At $\alpha = 0.1$, the A-D test reaches perfect power at $n = 340$, significantly earlier than the K-S test ($n = 650$). At $\alpha = 0.05$, the test achieves perfect power at $n = 450$, compared to $n = 710$ for the K-S test. Finally, at $\alpha = 0.01$, perfect power is achieved at $n = 620$, compared to $n = 840$ for the K-S test. These results demonstrate that the A-D test requires fewer observations to achieve high power, making it a more efficient tool for detecting deviations.

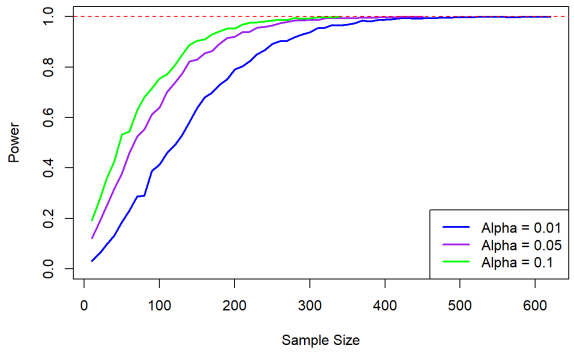

Figure 7: Power Curve of AD test at Different significance levels

Table 4 provide a detailed summary of the A-D test's Type I error rates and power across varying sample sizes and alpha levels. Under the null hypothesis, the A-D test consistently maintains Type I error rates near the nominal levels, validating its ability to control false positives. Under the alternative hypothesis, the A-D test demonstrates superior power compared to the K-S test at equivalent sample sizes, particularly for smaller datasets. For example, at $\alpha = 0.05$, the A-D test achieves a power of 0.3875 at $n = 50$ and 0.6665 at $n = 100$, compared to the K-S test's 0.2705 and 0.4570, respectively. Similarly, at $\alpha = 0.1$, the A-D test reaches power values of 0.5105 at $n = 50$ and 0.7595 at $n = 100$, outperforming the K-S test. For larger sample sizes ($n \geq 1000$), both tests converge to perfect power (1.0), indicating their effectiveness in rejecting the null hypothesis when deviations are present. However, the A-D test achieves this threshold with fewer observations, making it particularly useful in scenarios with moderate data availability.

The A-D test builds on the findings from the K-S test, demonstrating enhanced sensitivity and robustness, particularly in datasets where tail behavior is significant. Under the null hypothesis, the A-D test maintains uniform p-value distributions and effectively controls Type I error rates, similar to the K-S test. Under the alternative hypothesis, the A-D test surpasses the K-S test in sensitivity, achieving higher power at smaller sample sizes. The power curves highlight the A-D test's efficiency, requiring fewer observations to reach perfect power across all alpha levels.

These findings align closely with the conclusions drawn in Engmann and Cousineau (2011), which emphasize the A-D test's superior performance compared to the K-S test. The A-D test has been shown to maintain an exact Type I error rate of 0.05 while demonstrating greater power in detecting differences between samples from two

distinct distributions. In contrast, the K-S test tends to be overly conservative. Several key pieces of evidence further support the A-D test's advantages. First, the A-D test consistently identifies small variations in distribution parameters, such as shift, scale, and symmetry, more reliably than the K-S test, regardless of sample size. Second, it exhibits greater sensitivity to differences in the tails of distributions, even in cases of substantial overlap between the distributions or when sample sizes are small. Finally, the A-D test requires significantly less data to achieve sufficient statistical power, making it more efficient in practice. Given that the A-D test retains the same fundamental advantages as the K-S test while outperforming it in critical areas, this evidence strongly supports its use as a preferred tool for comparing distributions.

Cramervon Mises

The CvM test was also subjected to the same procedure as done for the first two tests. Figure 8 illustrates the distribution of *p*-values, showcasing a pattern consistent with theoretical expectations for goodness-of-fit tests. For small sample sizes ($n = 10, 100$), the *p*-values generated by the CvM test approximate a uniform distribution across $[0, 1]$, demonstrating the test's robustness in maintaining the expected Type I error rate under the null hypothesis. As the sample size increases ($n = 500$ and above), the uniformity of the *p*-value distribution becomes more pronounced, with histograms appearing smoother and more regular. This behavior indicates the CvM test's ability to maintain its theoretical properties even with larger datasets. While similar trends were observed for the K-S and A-D tests, the CvM test provides additional flexibility by assessing deviations across the entire distribution rather than focusing solely on the largest discrepancies (as in the K-S test) or emphasizing the tails (as in the A-D test). The proportion of *p*-values below the red aligns with the nominal Type I error rate of 5%, confirming that the CvM test effectively controls for false positives under the null hypothesis. These results reinforce the CvM test's reliability, especially for datasets where uniformity in *p*-values is critical. When analyzed under the alternative hypothesis, the CvM test exhibits significant shifts in *p*-value distributions toward zero as sample sizes increase. This is illustrated in Figure 9, where smaller sample sizes ($n = 10, 100$) show more variability in *p*-value distributions, reflecting lower power to detect deviations. However, even at these smaller sample sizes, the CvM test demonstrates improved performance over the K-S test, with a larger proportion of *p*-values falling below the $p = 0.05$ threshold. For moderate sample sizes ($n = 500$), the *p*-value distributions shift noticeably, with a growing cluster of *p*-values below the critical threshold. As sample sizes reach $n = 1000, 2000, 2500$, this clustering becomes more pronounced, reflecting the test's increased sensitivity to deviations from the null hypothesis. Compared to the K-S and A-D tests, the CvM test provides a balanced assessment of deviations, combining the global sensitivity of the K-S test with the tail sensitivity of the A-D test. This makes it particularly suitable for detecting discrepancies that are distributed across the entirety of the data, rather than concentrated in specific regions. These results align with existing studies emphasizing the CvM test's balanced approach to detecting GoF deviations, particularly in scenarios involving complex distributions with heavy tails.

The power curve in figure 10 shows a similar result to the first two tests. Power increases as sample size grows,

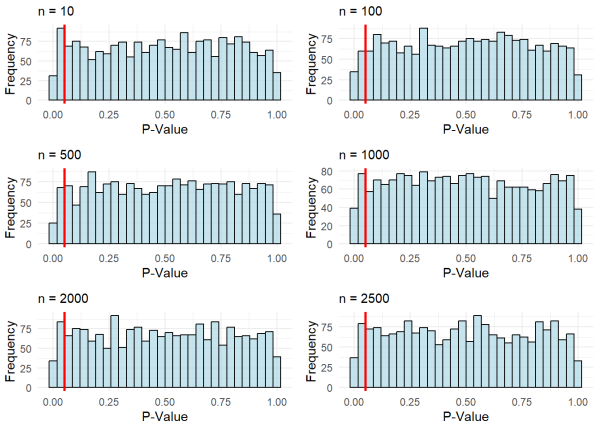

Figure 8: Sampling Distribution of CvM test under Null Hypothesis

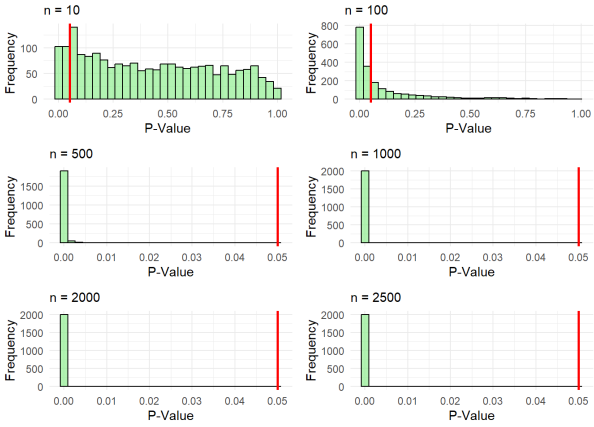

Figure 9: Sampling Distribution of CvM test under Alternative Hypothesis

eventually reaching perfect power (1.0). While this trend is observed across all three tests (CvM, K-S, and A-D), the CvM test strikes a middle ground in terms of efficiency and sensitivity. At $\alpha = 0.1$, the CvM test achieves perfect power at $n = 420$, which is earlier than the K-S test ($n = 650$) but slightly later than the A-D test ($n = 340$). Similarly, at $\alpha = 0.05$, the CvM test reaches perfect power at $n = 550$, earlier than the K-S test ($n = 710$) but requiring more data than the AD test ($n = 450$). For $\alpha = 0.01$, the CvM test attains perfect power at $n = 700$, earlier than the K-S test ($n = 840$) but slightly later than the A-D test ($n = 620$). These results highlight the CvM test's balanced efficiency in achieving high power with moderate sample sizes, demonstrating its robustness across varying significance levels.

The relationship between alpha levels and sample size is also evident in the power curves. Higher alpha levels ($\alpha = 0.1$) result in faster convergence to perfect power, reflecting greater sensitivity but at the cost of increased Type I error rates. Conversely, stricter alpha levels ($\alpha = 0.01$) require larger sample sizes to achieve comparable sensitivity, emphasizing the trade-off between significance level and detection power.

Table 5 summarizes the CvM test's Type I error rates and power for different sample sizes and alpha levels. The results reinforce the test's ability to control Type I errors under the null hypothesis and its effectiveness in detecting deviations under the alternative hypothesis. For example, at $\alpha = 0.05$, the CvM test achieves a power

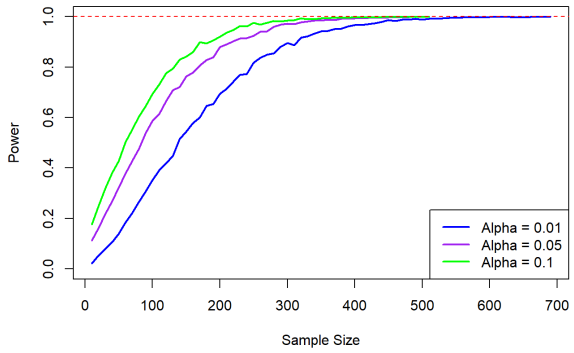

Figure 10: Power Curve of CvM test at Different significance levels

of 0.350 at $n = 50$ and 0.640 at $n = 100$, outperforming the K-S test (0.2705 and 0.4570, respectively) and approaching the A-D test (0.3875 and 0.6665, respectively). Similarly, at $\alpha = 0.1$, the CvM test reaches power values of 0.520 at $n = 50$ and 0.750 at $n = 100$, surpassing the K-S test while demonstrating comparable results to the A-D test.

For larger sample sizes ($n \geq 1000$), all three tests converge to perfect power (1.0), confirming their ability to detect deviations when sufficient data is available. However, the CvM test achieves this threshold more efficiently than the K-S test, requiring fewer observations, while the AD- test remains slightly more efficient in smaller datasets. The CvM test demonstrates a robust and balanced performance across all metrics, making it a strong contender alongside the KS and AD tests for evaluating goodness-of-fit. Under the null hypothesis, the CvM test maintains uniform *p*-value distributions and controls Type I error rates effectively, even with smaller datasets. Under the alternative hypothesis, the CvM test excels in detecting deviations across the entire distribution, striking a balance between the global sensitivity of the K-S test and the tail sensitivity of the A-D test.

The power curves and tabulated results emphasize the CvM test's efficiency, requiring fewer observations to achieve high power compared to the K-S test while offering comparable performance to the A-D test in many scenarios. These findings underscore the CvM test's versatility, particularly for heavy-tailed distributions like the Burr mixture models, where deviations may occur across both central and extreme regions of the data.

By integrating the CvM test into the suite of goodness-of-fit evaluations, practitioners can achieve a more comprehensive assessment of model fit, leveraging its balanced sensitivity and robust performance across diverse datasets.

Chi-squared

The Chi-Square (χ^2) test was evaluated following the CvM test to assess its performance across various sample sizes under both null and alternative hypotheses. As the dataset comprises losses, a continuous variable, the implementation of the χ^2 test required dynamic binning strategies to effectively handle the nature of the data and preserve the integrity of the test. The dynamic binning approach used in the code applies equal-frequency binning,

where bin edges are determined such that each bin contains approximately the same number of observations. This method adapts to the dataset's distribution, ensuring the test remains robust across varying sample sizes and data characteristics. Rolke and Gongora (2021) talks in detail about binning methods used for the chi-squared goodness of fit test.

The dynamic binning function used here calculates the number of bins using Sturges' formula:

$$\text{number of bins} = \lceil 1 + \log_2(\text{length of data}) \rceil,$$

with a minimum of 10 bins to avoid overfitting or instability. The quantile-based bin edges are then computed, ensuring that the bins are spaced according to the distribution of the data. This process is particularly effective for continuous datasets, where traditional fixed-width binning may fail to capture the underlying structure, leading to unreliable results. The adaptability of this binning method makes it well suited for the χ^2 test, allowing for improved sensitivity and stability, especially in small or highly variable datasets.

As evident in Figure 11, for small sample sizes ($n = 10, 50, 100$), the *p*-values exhibit a slight deviation from uniformity, a pattern that can be attributed to the instability of the χ^2 statistic when sample sizes are small. However as sample sizes increase ($n = 500, 1000, 2000, 2500$), the *p*-value distributions converge toward a uniform distribution, demonstrating the test's adherence to its theoretical Type I error rate. Notably, the proportion of *p*-values below the significance threshold ($\alpha = 0.05$) aligns closely with the nominal Type I error rate, confirming the χ^2 test's ability to control false positives effectively under the null hypothesis. This is evident in the progressively smoother and more consistent histograms for larger sample sizes, where the uniformity of *p*-value distributions becomes pronounced. The dynamic binning strategy played a pivotal role in maintaining this uniformity. By adjusting the bin edges to reflect the data distribution, the test avoids over- or under-estimation of observed frequencies, particularly in the tails of the distribution, which are more sensitive to binning errors. This adaptability ensures that the χ^2 test remains robust even with the challenges posed by the continuous nature of the dataset. Under the alternative hypothesis, the *p*-value distributions in Figure 12 reveal a pronounced shift toward zero as sample sizes increase, just as evident from the first few tests. For smaller sample sizes ($n = 10, 50, 100$), the distributions exhibit significant variability, reflecting the test's limited power to detect model misspecifications with insufficient data. However, as sample sizes grow ($n = 500, 1000, 2000, 2500$), the clustering of *p*-values below the critical threshold ($\alpha = 0.05$) becomes more evident, highlighting the test's increasing sensitivity. At larger sample sizes ($n = 2000, 2500$), the majority of *p*-values fall below the critical threshold, indicating near-perfect power in detecting deviations. This trend underscores the χ^2 test's reliability in identifying model inadequacies when sufficient data is available. The χ^2 test's power advantage under the alternative hypothesis can be attributed to its global sensitivity, which assesses discrepancies across the entire data range. The dynamic binning approach ensures that deviations are not masked by inappropriate bin widths, further enhancing the test's performance in detecting subtle differences between the null and alternative models.

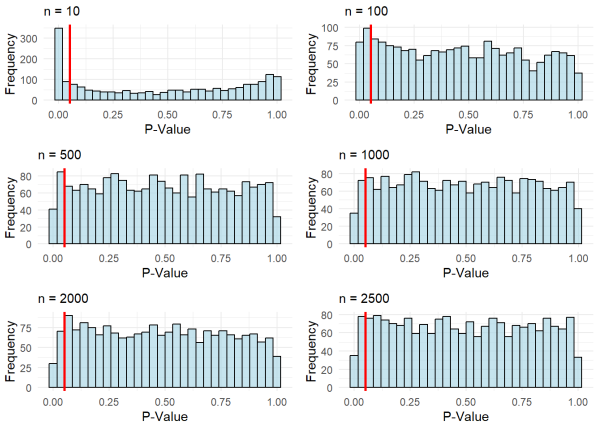

Figure 11: Sampling Distribution of Chi-squared test under Null Hypothesis

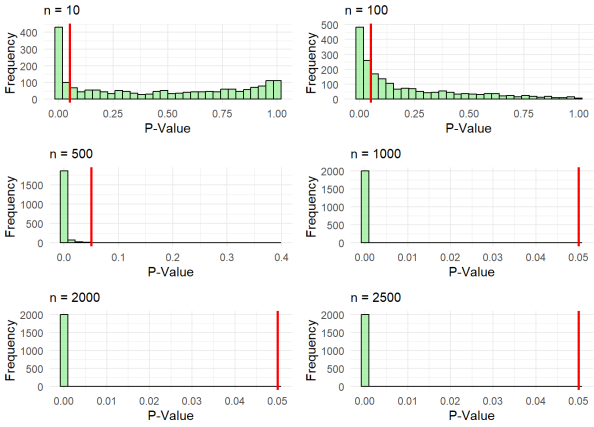

Figure 12: Sampling Distribution of Chi-squared test under Alternative Hypothesis

The power curve for the χ^2 test, shown in Figure 13, demonstrates a consistent increase in power as sample sizes grow, eventually reaching perfect power (1.0) for all α levels. For $\alpha = 0.05$, the χ^2 test achieves perfect power at $n = 730$, which is comparable to the K-S test ($n = 710$) but slightly less efficient than the CvM ($n = 550$) and A-D ($n = 450$) tests. Similarly, for $\alpha = 0.01$, the test requires $n = 790$ to reach perfect power, reflecting the trade-off between stricter significance levels and the required sample size. Conversely, for $\alpha = 0.1$, the test attains perfect power at $n = 670$, demonstrating its ability to detect deviations more effectively with relaxed significance thresholds.

The results are summarized in Table 6. The χ^2 test maintains its nominal Type I error rates across all sample sizes under the null hypothesis and demonstrates increasing power under the alternative hypothesis. For instance, at $n = 50$ and $\alpha = 0.05$, the test achieves a power of 0.26, which grows to 0.38 at $n = 100$ and reaches 1.0 at $n = 730$. However, the Type 1 error rates begin at a relatively higher rate even for small sample sizes and steadily decreases as the sample size increases until reaching a size after it fluctuates. This pattern is not as clear and distinct in the first few goodness of fit tests.

While the χ^2 test shares similarities with the CvM test in its global assessment of distributional discrepancies, it differs in its sensitivity due to the chosen binning strategy. Unlike the CvM and A-D tests, which evaluate

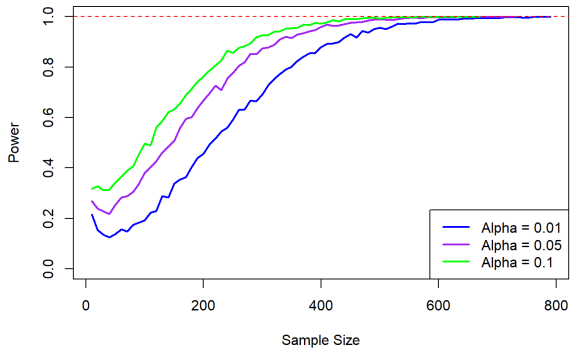

Figure 13: Power Curve of Chi-squared test at Different significance levels

GoF through CDFs, the χ^2 test relies on observed and expected frequencies across bins. This distinction makes it particularly suitable for datasets with diverse characteristics, as demonstrated by its robust performance with dynamically binned continuous data. Compared to the K-S and A-D tests, the χ^2 test exhibits a middle ground in efficiency and sensitivity. While it requires more observations to achieve perfect power compared to the A-D test, it outperforms the KS test in smaller sample sizes. Its global sensitivity also complements the local and tail-focused assessments provided by the KS and AD tests, respectively.

The χ^2 test, when implemented with dynamic binning, proves to be a reliable and flexible tool for evaluating goodness-of-fit in continuous datasets. Its robust control of Type I error rates under the null hypothesis and its growing power under the alternative hypothesis make it a valuable addition to the suite of goodness-of-fit tests. By integrating the χ^2 test alongside the K-S, CvM, and A-D tests, researchers can achieve a more comprehensive evaluation of model adequacy, leveraging its unique strengths to complement the sensitivity and efficiency of other tests.

Table 3: Type I Error Rates and Power across Different Sample Sizes for K-S Test

(a) Results for $\alpha = 0.01$

Sample Size	**Type I Error Rate**	**Power**
10	0.01	0.02
50	0.01	0.10
100	0.01	0.21
500	0.02	0.95
1000	0.01	1.00
2000	0.01	1.00
2500	0.01	1.00

(b) Results for $\alpha = 0.05$

Sample Size	**Type I Error Rate**	**Power**
10	0.05	0.08
50	0.06	0.27
100	0.05	0.46
500	0.05	0.99
1000	0.05	1.00
2000	0.05	1.00
2500	0.05	1.00

(c) Results for $\alpha = 0.1$

Sample Size	**Type I Error Rate**	**Power**
10	0.09	0.16
50	0.10	0.39
100	0.09	0.59
500	0.09	1.00
1000	0.10	1.00
2000	0.10	1.00
2500	0.10	1.00

Table 4: Type I Error Rates and Power across Different Sample Sizes for A-D Test

(a) Results for $\alpha = 0.01$

Sample Size	**Type I Error Rate**	**Power**
10	0.01	0.04
50	0.01	0.20
100	0.01	0.40
500	0.01	1.00
1000	0.01	1.00
2000	0.01	1.00
2500	0.01	1.00

(b) Results for $\alpha = 0.05$

Sample Size	**Type I Error Rate**	**Power**
10	0.05	0.13
50	0.05	0.39
100	0.05	0.67
500	0.04	1.00
1000	0.05	1.00
2000	0.06	1.00
2500	0.05	1.00

(c) Results for $\alpha = 0.1$

Sample Size	**Type I Error Rate**	**Power**
10	0.01	0.20
50	0.10	0.51
100	0.10	0.76
500	0.10	1.00
1000	0.11	1.00
2000	0.11	1.00
2500	0.10	1.00

Table 5: Type I Error Rates and Power across Different Sample Sizes for CvM Test

(a) Results for $\alpha = 0.01$

Sample Size	**Type I Error Rate**	**Power**
10	0.01	0.03
50	0.01	0.14
100	0.01	0.33
500	0.01	1.00
1000	0.01	1.00
2000	0.01	1.00
2500	0.01	1.00

(b) Results for $\alpha = 0.05$

Sample Size	**Type I Error Rate**	**Power**
10	0.06	0.11
50	0.05	0.34
100	0.05	0.57
500	0.05	1.00
1000	0.04	1.00
2000	0.05	1.00
2500	0.05	1.00

(c) Results for $\alpha = 0.1$

Sample Size	**Type I Error Rate**	**Power**
10	0.10	0.18
50	0.10	0.45
100	0.08	0.72
500	0.10	1.00
1000	0.10	1.00
2000	0.10	1.00
2500	0.10	1.00

Table 6: Type I Error Rates and Power across Different Sample Sizes for Chi-squared Test

(a) Results for $\alpha = 0.01$

Sample Size	**Type I Error Rate**	**Power**
10	0.17	0.20
50	0.06	0.13
100	0.03	0.19
500	0.01	0.96
1000	0.01	1.00
2000	0.01	1.00
2500	0.01	1.00

(b) Results for $\alpha = 0.05$

Sample Size	**Type I Error Rate**	**Power**
10	0.23	0.26
50	0.11	0.26
100	0.09	0.38
500	0.06	0.99
1000	0.05	1.00
2000	0.05	1.00
2500	0.05	1.00

(c) Results for $\alpha = 0.1$

Sample Size	**Type I Error Rate**	**Power**
10	0.30	0.33
50	0.18	0.33
100	0.14	0.49
500	0.11	0.99
1000	0.09	1.00
2000	0.11	1.00
2500	0.10	1.00

5.2 1-Component Lognormal

The 1-component lognormal distribution is a robust model for actuarial loss data, particularly when the data exhibits moderate skewness and lighter tails. Figure 14 illustrates the fitted density for the lognormal distribution,

which aligns closely with the observed left-truncated loss data (Secura Belgian Re dataset) of 371 automobile claims, adjusted for inflation and truncated at €1.2 million (Blostein and Miljkovic). The fitted parameters, mean log-scale ($\mu = 14.3258849$) and standard deviation ($\sigma = 0.5014714$), effectively capture the central tendency and variability of the dataset. This ensures that the lognormal model provides a reliable representation of the data while maintaining computational simplicity. Its utility is particularly evident in left-truncated data scenarios, a common feature in actuarial datasets. The lognormal distribution's adaptability to multiplicative processes and its ability to model a wide range of data distributions make it an ideal choice for assessing goodness-of-fit tests in actuarial science. Its straightforward structure enables clear interpretations, making it a foundational tool for evaluating model adequacy and performance.

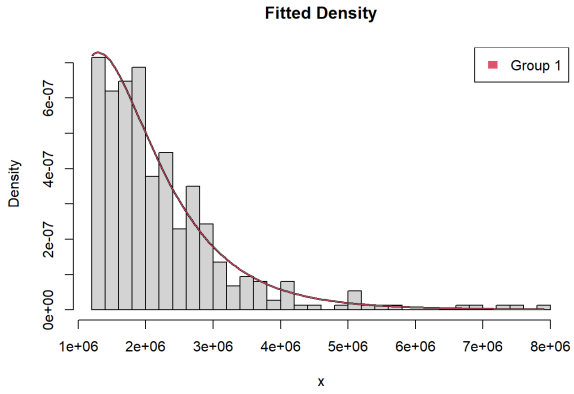

Figure 14: Fitted density plot for the 1-component lognormal distribution

Kolmogorov-Smirnov Test

Just as done for the Burr mixture model, we subject the lognormal model to the same scrutiny, providing insights into its behavior for simpler, non-heavy-tailed distributions. The lognormal distribution serves as a contrast to the more complex 2-component Burr mixture model, offering an opportunity to examine how distributional characteristics influence the effectiveness of goodness-of-fit tests. Under the null hypothesis, the distribution of p-values approximates a uniform distribution over the interval $[0, 1]$, as shown in Figure 15. In comparison to the Burr mixture model, the lognormal model exhibits more consistent uniformity in p-values at smaller sample sizes. This distinction highlights the reduced complexity of detecting goodness-of-fit deviations for simpler distributions, where the absence of heavy tails and mixture components allows for more straightforward model evaluation. Under the alternative hypothesis, where the parameters μ and σ are slightly perturbed, to 14 and 1 respectively, the p-value distributions shift markedly toward zero, as depicted in Figure 16. For smaller sample sizes ($n = 10, 100$), the variability in p-values is more pronounced, reflecting lower power to detect deviations. However, as the sample size increases ($n \geq 500$), the clustering of p-values near zero becomes increasingly prominent, signaling the test's growing sensitivity. This trend is consistent across all tested sample sizes and demonstrates the K-S test's ability to effectively detect deviations from the null hypothesis under the lognormal model.

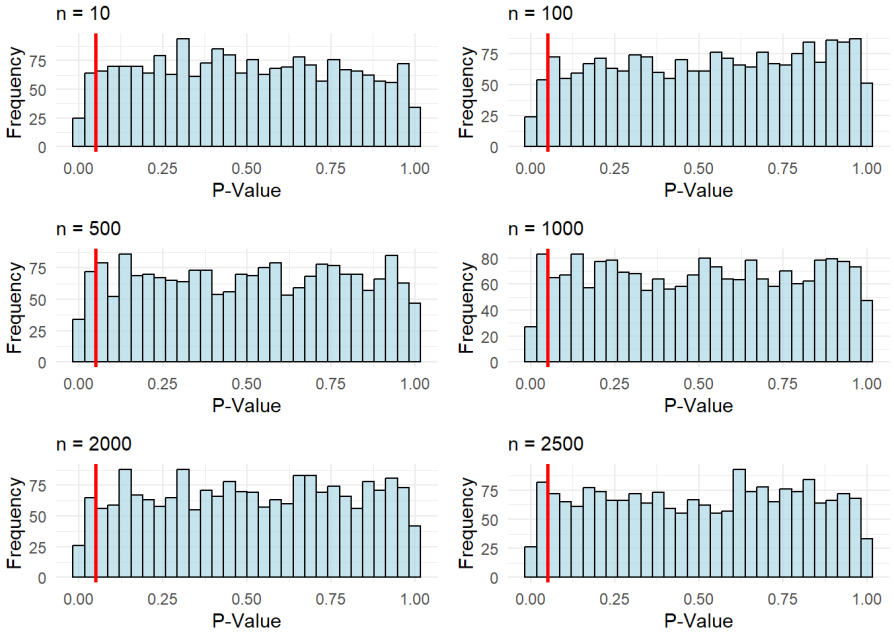

Figure 15: Sampling Distribution of KS Test Under Null Hypothesis for Lognormal Model

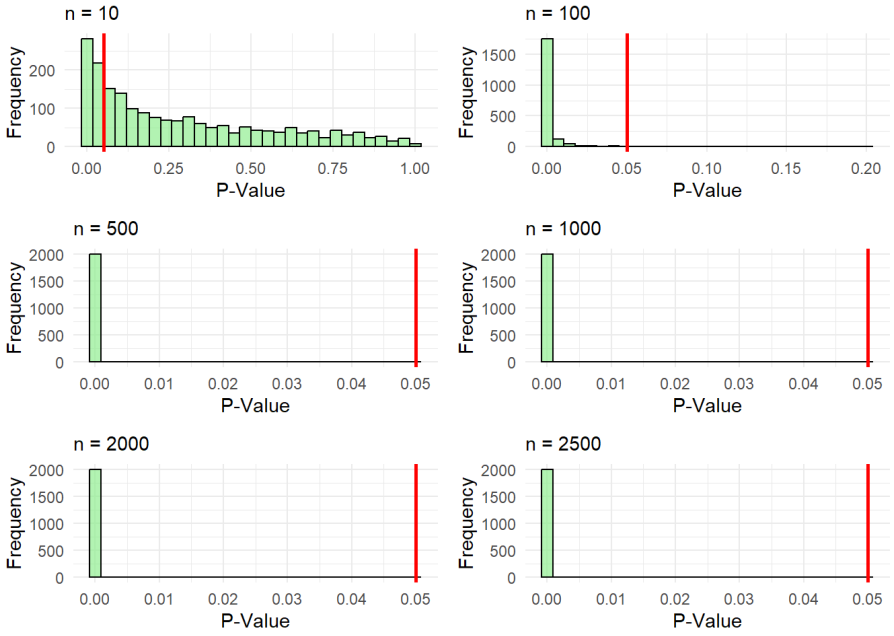

Figure 16: Sampling Distribution of KS Test Under Alternative Hypothesis for Lognormal Model

The power analysis of the K-S test, presented in Figure 17, reveals the relationship between sample size, significance level, and the test's ability to detect deviations. For the lognormal model, the test achieves perfect power at smaller sample sizes compared to the Burr mixture model. Specifically, at $\alpha = 0.1$, the test reaches perfect power at $n = 140$, while $\alpha = 0.05$ and $\alpha = 0.01$ require $n = 170$ and $n = 190$, respectively. In contrast, the Burr mixture model required significantly larger sample sizes ($n \geq 650$) to achieve comparable power due to the inherent complexity of its heavy-tailed nature.

The summarized results in Table 7 provide a detailed comparison of Type I error rates and power for varying

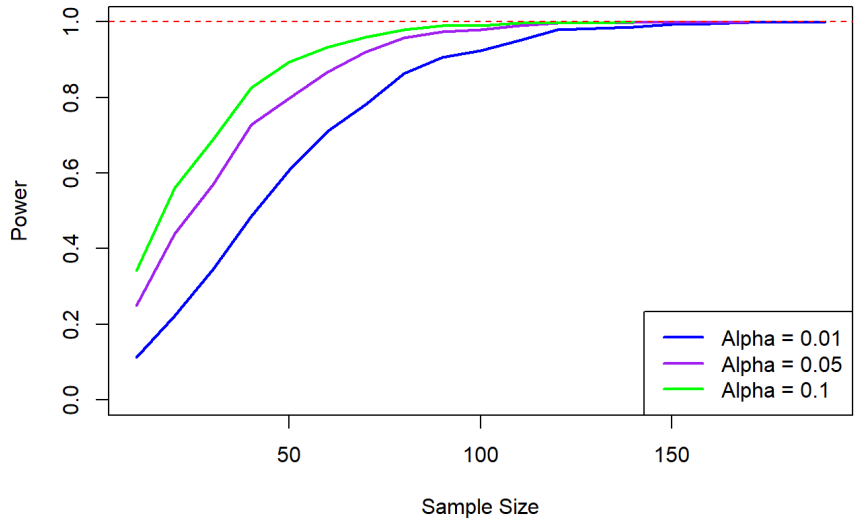

Figure 17: Power Curve of KS Test for Lognormal Model at Different Significance Levels

sample sizes. These findings underscore the K-S test's versatility in evaluating GoF for simpler models like the lognormal distribution, where the reduced complexity allows for more efficient detection of deviations. The comparison with the Burr mixture model highlights the challenges posed by heavy-tailed and multi-component distributions, reaffirming the need for tailored GoF approaches based on the underlying data characteristics.

Anderson-Darling Test

The A-D test extends the analysis of GoF for the 1-component lognormal model. Building on the results observed for the K-S test, the A-D test provides complementary insights into the distributional characteristics and deviations of this model. Under the null hypothesis, the A-D test demonstrates robust performance, maintaining uniformity in the distribution of p-values across sample sizes. Figure 18 illustrates the p-value distributions for increasing sample sizes ($n = 10, 100, 500, 1000, 2000, 2500$). Under the alternative hypothesis, the A-D test displays significant shifts in the p-value distributions toward zero, reflecting its superior ability to detect deviations. Figure 19 highlights these shifts across various sample sizes. The behavior mirrors that of the K-S test but is particularly significant for the A-D test due to its greater weighting of tail behavior. This trend illustrates the increasing power of the A-D test with larger datasets and demonstrates its heightened sensitivity compared to the K-S test.

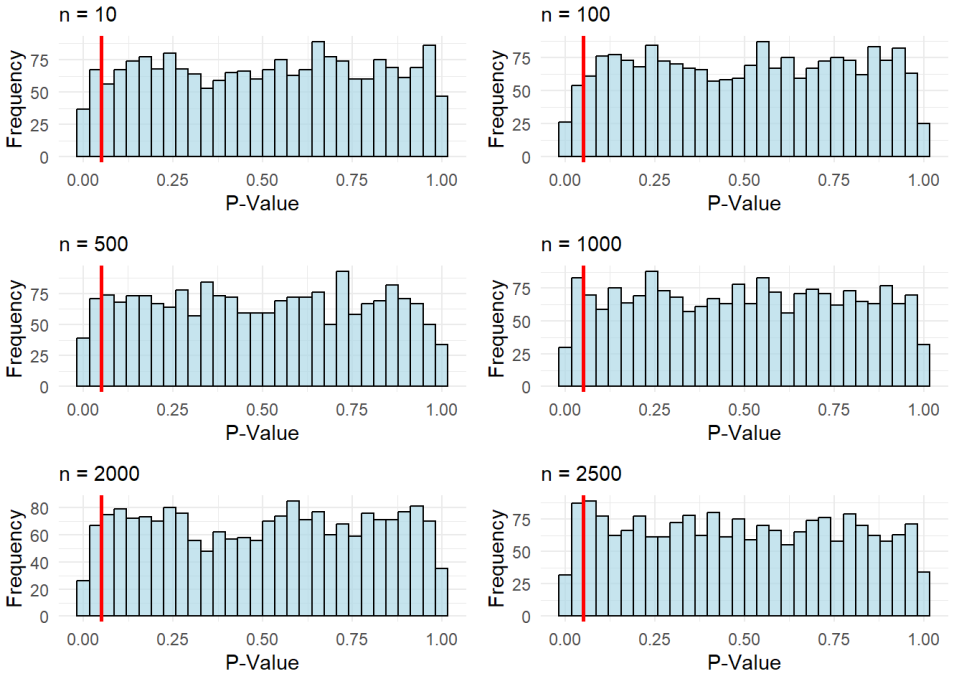

Figure 18: Sampling Distribution of AD Test under Null Hypothesis for Lognormal Model

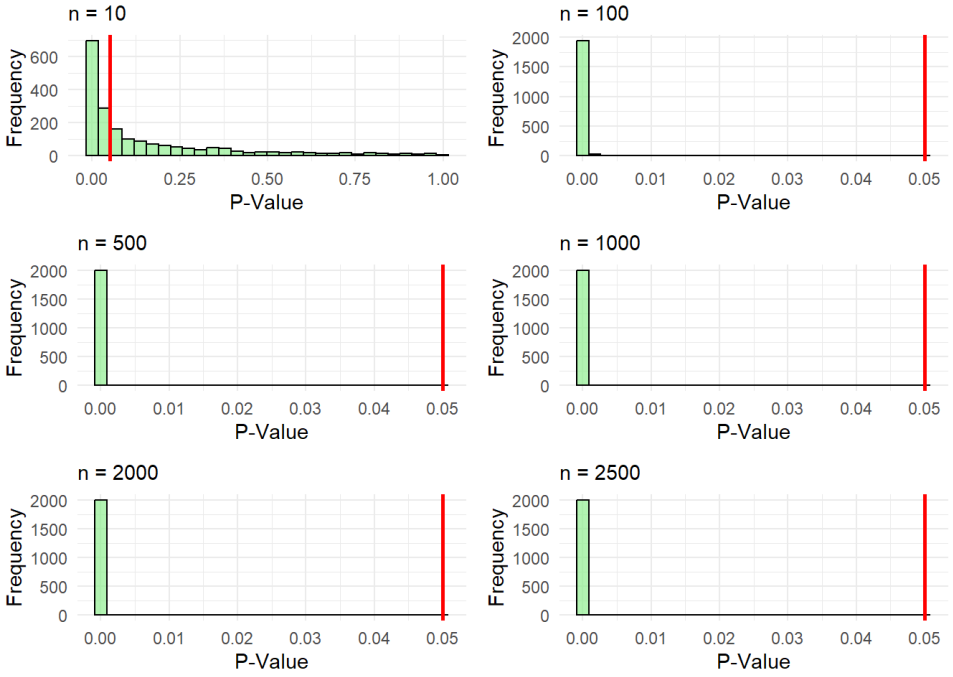

Figure 19: Sampling Distribution of AD Test under Alternative Hypothesis for Lognormal Model

The power curves, shown in Figure 20, further reinforce the A-D test's performance. At $\alpha = 0.1$, the A-D test achieves perfect power at $n = 100$, whereas $\alpha = 0.05$ and $\alpha = 0.01$ require $n = 110$ and $n = 140$, respectively. These results are consistent with the Burr mixture model which highlights the A-D test's efficiency, requiring fewer observations than the K-S test to achieve comparable power. This efficiency is particularly valuable for actuarial applications, where data availability can be constrained.

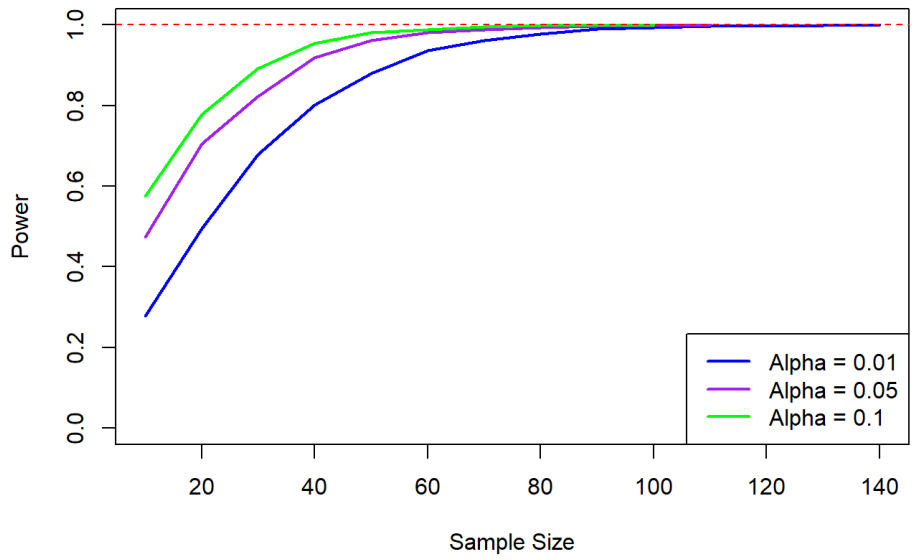

Figure 20: Power Curve of AD Test for Lognormal Model

The results for the A-D test are summarized in Table 8, which details the Type I error rates and power across varying sample sizes and significance levels. Under the null hypothesis, the Type I error rates consistently align with theoretical expectations, validating the test's reliability. Under the alternative hypothesis, the A-D test demonstrates superior power compared to the K-S test, achieving higher sensitivity at smaller sample sizes. For example, at $\alpha = 0.05$, the A-D test reaches a power of 0.9605 at $n = 50$, compared to the K-S test's 0.8135, and achieves perfect power (1.0) at $n = 100$.

Cramér-von Mises Test

The CvM test provides a balanced perspective on GoF assessments for the 1-component lognormal model, complementing the K-S and A-D tests. Known for its uniform weighting across the entire distribution, the CvM test builds on the findings from the K-S and A-D analyses, offering additional insights into the overall fit of the model. As shown in Figures 21 and 22, the behavior parallels the performance of both the K-S and A-D tests, yet it reflects the CvM test's unique focus on deviations across the entire distribution, rather than emphasizing the tails. It also demonstrates the CvM test's increasing sensitivity and power with relatively smaller datasets and simpler distributions.

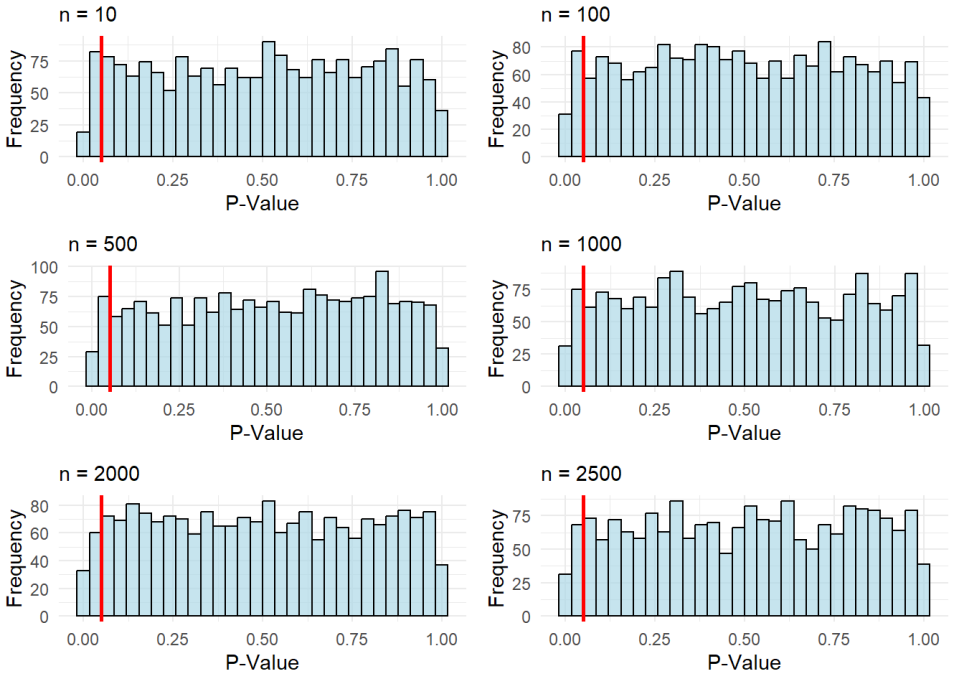

Figure 21: p-value distributions of the CvM test under the null hypothesis for varying sample sizes.

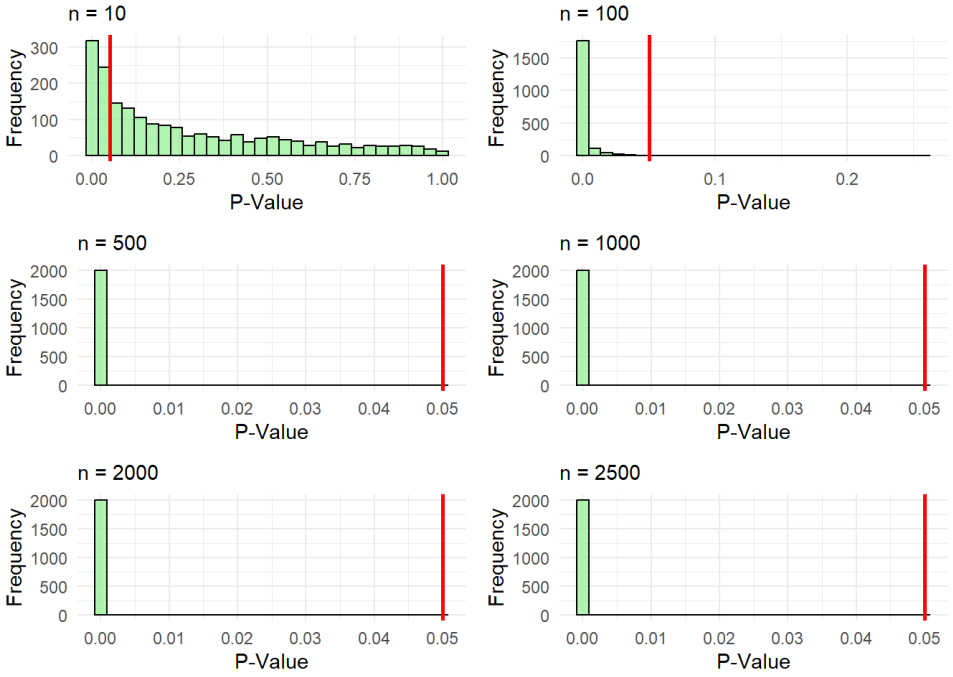

Figure 22: p-value distributions of the CvM test under the alternative hypothesis for varying sample sizes.

The CvM test's power curves, presented in Figure 23, further illustrate its performance across three significance levels ($\alpha = 0.01, 0.05, 0.1$). The test achieves perfect power at $n = 160$ for $\alpha = 0.1$, $n = 170$ for $\alpha = 0.05$, and $n = 210$ for $\alpha = 0.01$. Compared to the A-D and K-S tests, the CvM test requires slightly larger sample sizes to achieve equivalent power, reflecting its more uniform weighting of distributional deviations. However, this characteristic makes the CvM test particularly valuable when a holistic assessment of goodness-of-fit is required.

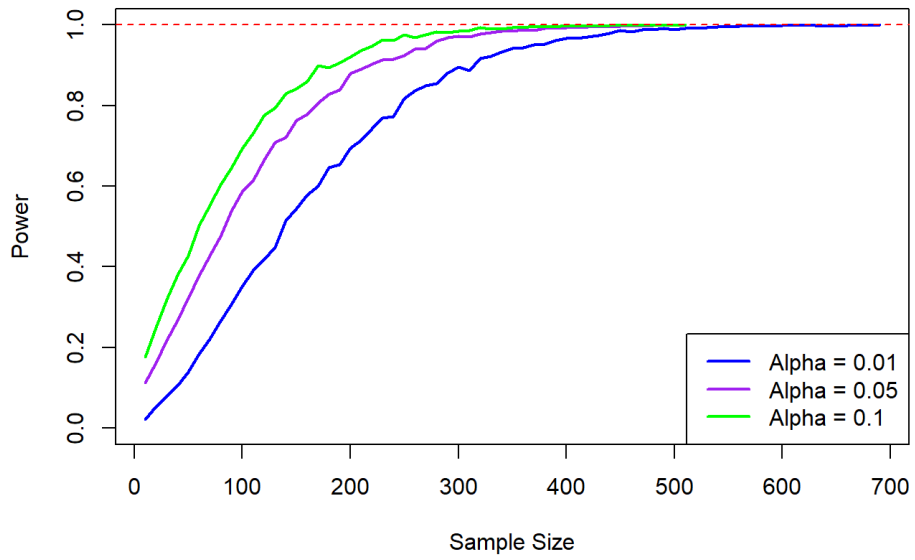

Figure 23: Power curves of the CvM test at different significance levels ($\alpha = 0.01, 0.05, 0.1$).

In comparison to the Burr mixture model, the CvM test's performance on the 1-component lognormal model emphasizes its flexibility in handling simpler distributions. The lognormal model's lack of a mixture structure reduces the complexity of tail behavior, allowing the CvM test to provide a broader evaluation of deviations across the entire distribution. While the A-D test remains superior in detecting tail-heavy deviations, the CvM test excels in scenarios where a more uniform sensitivity is desired.

As evident in table 9, under the null hypothesis, the CvM test consistently controls Type I error rates near their nominal values, validating its theoretical reliability. Under the alternative hypothesis, the test demonstrates robust power gains with increasing sample sizes. For instance, at $\alpha = 0.05$, the CvM test achieves a power of 0.8170 at $n = 50$, which increases to perfect power (1.0) at $n = 170$. Just as stated for the Burr mixture model, the findings for the CvM test highlight its complementary role to the K-S and A-D tests. Its ability to evaluate deviations uniformly across the entire distribution makes it a valuable tool for analyzing datasets such as the 1-component lognormal model. These attributes emphasize the importance of integrating multiple goodness-of-fit tests to derive comprehensive insights, particularly in actuarial applications where data characteristics vary widely.

Chi-square Test

The Chi-Square (χ^2) test offers a distinctive perspective on GoF by evaluating global discrepancies between observed and expected frequencies. For the 1-component lognormal model, this test provides valuable insights under both the null and alternative hypotheses, particularly when paired with a dynamic binning strategy that adapts to the continuous nature of the data. Just as seen in the Burr mixture model, Figure 24 shows similar results for smaller sample sizes ($n = 10, 50, 100$) with deviations from uniformity observed, which can be attributed to

the instability of the χ^2 statistic when data is sparse. The role of dynamic binning in this analysis once again cannot be overstated. By tailoring the bin edges to the data's distribution, the test avoids biases that can arise from fixed-width bins, especially in the tails. This approach ensures that the χ^2 test remains robust across a range of sample sizes, even when data availability or distributional characteristics vary. The resulting histograms reveal increasingly smooth and uniform patterns as sample sizes grow, reinforcing the test's theoretical validity. Figure 12 shows a similar pattern to what was observed for the Burr mixture model. By $n = 2000$, nearly all p-values fall below the $p = 0.05$ threshold, signifying near-perfect power to identify model misspecifications. This progression highlights the test's sensitivity and its ability to capitalize on larger datasets to detect deviations with precision.

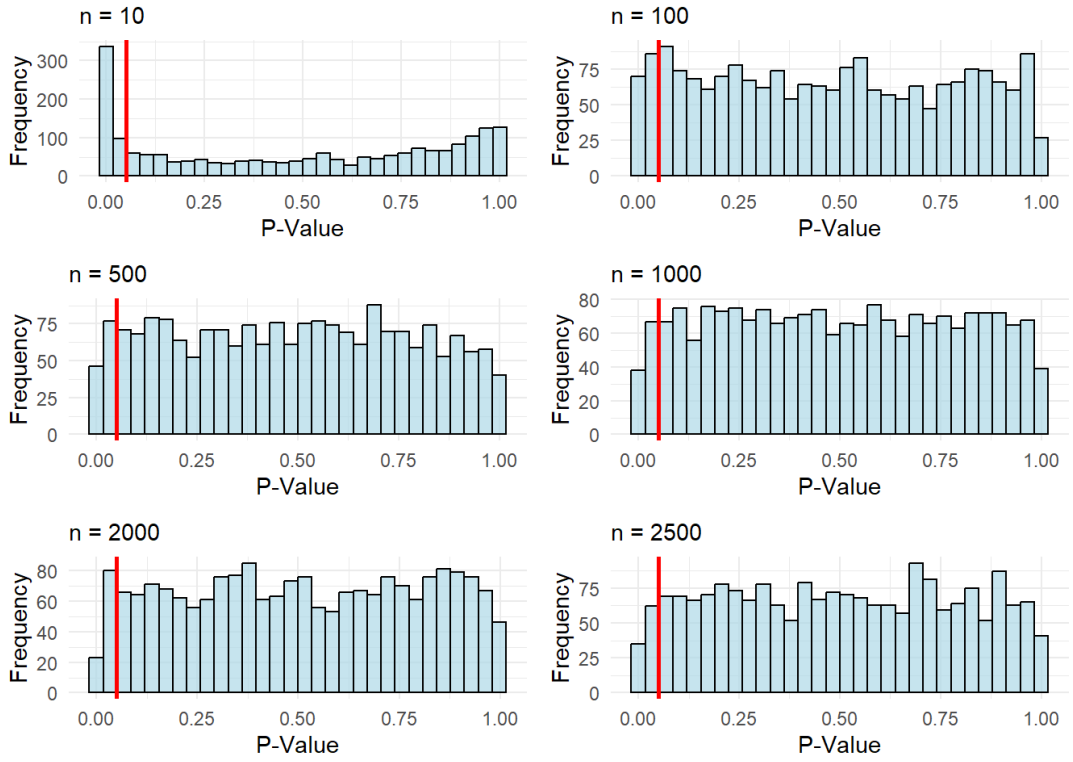

Figure 24: Sampling Distribution of the Chi-Square Test Under the Null Hypothesis

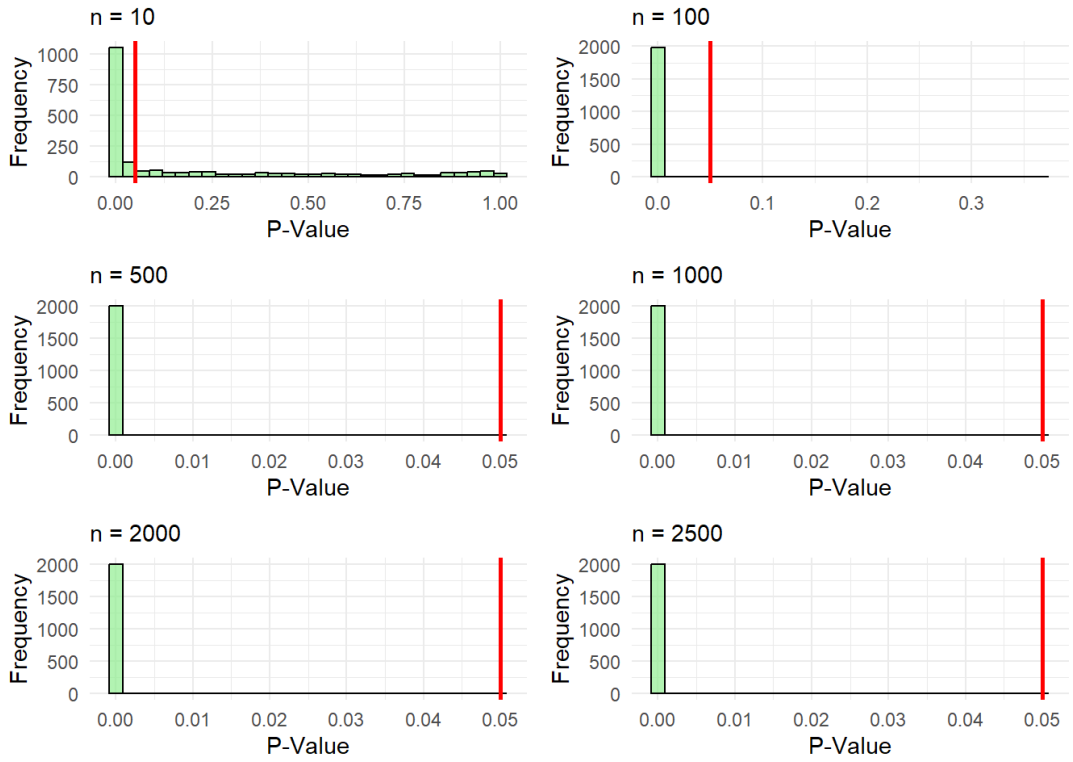

Figure 25: Sampling Distribution of the Chi-Square Test Under the Alternative Hypothesis

The power curve for the χ^2 test, presented in Figure 26 shows that as sample sizes increase, the test's power grows consistently, eventually reaching perfect power (1.0) at $n = 140$ for $\alpha = 0.01$, $n = 130$ for $\alpha = 0.05$, and $n = 120$ for $\alpha = 0.1$. This efficiency, especially when compared to the other tests, reflects the χ^2 test's ability to balance sensitivity with sample size requirements. While it requires slightly larger samples to achieve perfect power, the χ^2 test excels in providing a global view of distributional discrepancies, complementing the local and tail-sensitive insights offered by other methods.

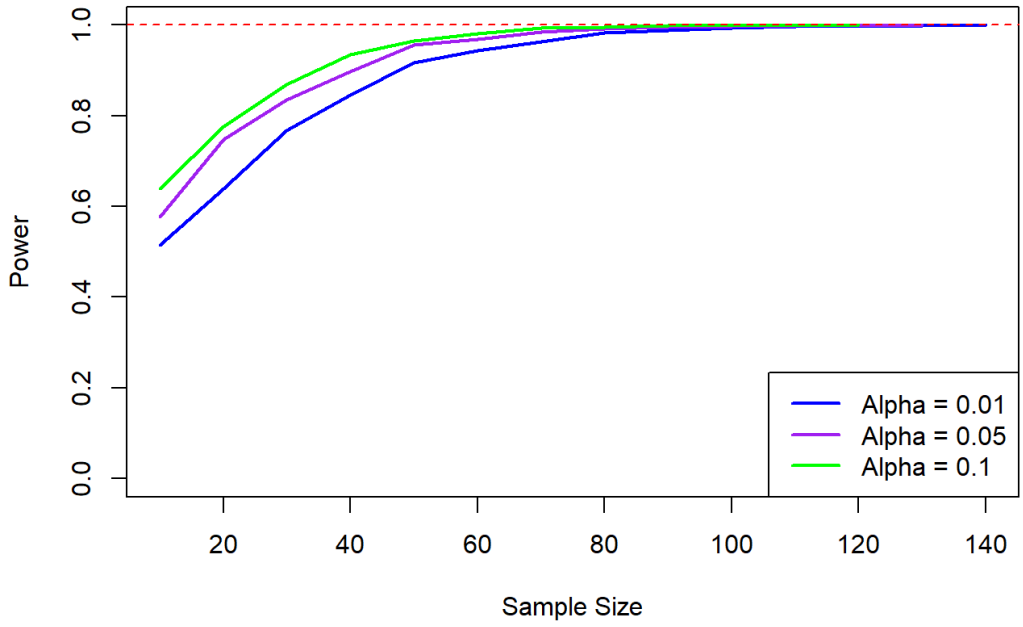

Figure 26: Power Curve of the Chi-Square Test for Different Significance Levels

In conclusion, the χ^2 test, when enhanced with dynamic binning, emerges as a robust and versatile tool for assessing goodness-of-fit in continuous datasets. Its ability to maintain Type I error rates under the null hypothesis and its growing sensitivity under the alternative hypothesis across all significance levels, as summarized in Table 10, make it an indispensable addition to the suite of statistical tests. By leveraging its strengths alongside the K-S, CvM, and A-D tests, researchers can achieve a comprehensive evaluation of model adequacy, addressing both global and specific distributional characteristics. These findings underscore the χ^2 test's pivotal role in statistical and actuarial applications, particularly when data complexity demands adaptability and precision.

Table 7: Type I Error Rates and Power across Different Sample Sizes for KS Test

(a) Results for $\alpha = 0.01$

Sample Size	**Type I Error Rate**	**Power**
10	0.01	0.10
50	0.01	0.60
100	0.01	0.93
500	0.01	1.00
1000	0.01	1.00
2000	0.02	1.00
2500	0.01	1.00

(b) Results for $\alpha = 0.05$

Sample Size	**Type I Error Rate**	**Power**
10	0.06	0.24
50	0.05	0.81
100	0.05	0.99
500	0.05	1.00
1000	0.05	1.00
2000	0.05	1.00
2500	0.04	1.00

(c) Results for $\alpha = 0.1$

Sample Size	**Type I Error Rate**	**Power**
10	0.10	0.37
50	0.10	0.89
100	0.10	1.00
500	0.10	1.00
1000	0.11	1.00
2000	0.10	1.00
2500	0.10	1.00

Table 8: Type I Error Rates and Power across Different Sample Sizes for AD Test

(a) Results for $\alpha = 0.01$

Sample Size	**Type I Error Rate**	**Power**
10	0.01	0.30
50	0.01	0.88
100	0.01	0.99
500	0.01	1.00
1000	0.01	1.00
2000	0.01	1.00
2500	0.01	1.00

(b) Results for $\alpha = 0.05$

Sample Size	**Type I Error Rate**	**Power**
10	0.05	0.47
50	0.05	0.96
100	0.05	1.00
500	0.06	1.00
1000	0.04	1.00
2000	0.05	1.00
2500	0.05	1.00

(c) Results for $\alpha = 0.1$

Sample Size	**Type I Error Rate**	**Power**
10	0.09	0.58
50	0.11	0.98
100	0.10	1.00
500	0.10	1.00
1000	0.09	1.00
2000	0.11	1.00
2500	0.09	1.00

Table 9: Type I Error Rates and Power across Different Sample Sizes for CvM Test

(a) Results for $\alpha = 0.01$

Sample Size	**Type I Error Rate**	**Power**
10	0.01	0.13
50	0.01	0.61
100	0.01	0.94
500	0.01	1.00
1000	0.01	1.00
2000	0.01	1.00
2500	0.01	1.00

(b) Results for $\alpha = 0.05$

Sample Size	**Type I Error Rate**	**Power**
10	0.04	0.27
50	0.05	0.82
100	0.04	0.99
500	0.06	1.00
1000	0.05	1.00
2000	0.05	1.00
2500	0.05	1.00

(c) Results for $\alpha = 0.1$

Sample Size	**Type I Error Rate**	**Power**
10	0.11	0.38
50	0.10	0.90
100	0.10	0.99
500	0.09	1.00
1000	0.10	1.00
2000	0.10	1.00
2500	0.10	1.00

Table 10: Type I Error Rates and Power across Different Sample Sizes for Chi-square Test

(a) Results for $\alpha = 0.01$

Sample Size	**Type I Error Rate**	**Power**
10	0.17	0.52
50	0.05	0.90
100	0.03	1.00
500	0.01	1.00
1000	0.01	1.00
2000	0.01	1.00
2500	0.01	1.00

(b) Results for $\alpha = 0.05$

Sample Size	**Type I Error Rate**	**Power**
10	0.22	0.57
50	0.11	0.95
100	0.09	1.00
500	0.06	1.00
1000	0.06	1.00
2000	0.05	1.00
2500	0.05	1.00

(c) Results for $\alpha = 0.1$

Sample Size	**Type I Error Rate**	**Power**
10	0.26	0.64
50	0.18	0.96
100	0.14	1.00
500	0.10	1.00
1000	0.12	1.00
2000	0.10	1.00
2500	0.10	1.00

6 Conclusion

This study critically examined the reliability of *p*-value-based decisions in actuarial model validation by assessing the performance of four widely used GoF tests; K-S, A-D, CvM, and Chi-Square (χ^2), through extensive simulation studies. These tests were evaluated under both null and alternative hypotheses using two actuarial models: a 1-component left-truncated Lognormal model and a 2-component Burr mixture model. The results underscore that GoF test performance is highly sensitive to distributional characteristics, parameter deviations, and sample size.

For the Lognormal model, the A-D test demonstrated superior sensitivity to deviations, particularly in the distribution tails, consistently outperforming the K-S and CvM tests. The CvM test offered more balanced detection across the entire distribution, while the K-S test, though maintaining nominal Type I error control, exhibited limited power in tail-focused scenarios. In the more complex 2-component Burr mixture model, which better reflects the heavy-tailed behavior often observed in insurance data, the A-D test again emerged as the most reliable in detecting subtle deviations. The χ^2 test showed weaker performance in small samples but improved significantly with larger datasets, affirming its utility for global model checks when binning is handled appropriately.

These findings reinforce the core question guiding this research: **Can actuaries rely solely on *p*-value reports in determining model adequacy?** Our results suggest not. *p*-values, while useful, are influenced by both test selection and sample size, which can lead to misleading conclusions, especially in actuarial contexts involving small datasets or distributions with extreme tail behavior. Non-significant *p*-values may mask genuine model inadequacy due to low power, while significant *p*-values may exaggerate trivial discrepancies in large samples. Thus, this study cautions against interpreting *p*-values in isolation and calls for a more comprehensive, context-sensitive approach.

6.1 Implications for Actuarial Practice, Insurance Model Validation, and Alternative Approaches

The simulation findings carry direct implications for actuarial practice and the development of model validation protocols within the insurance industry. Given the variable sensitivity of GoF tests, actuaries should employ multiple tests tailored to the specific modeling scenario. That is, using A-D and CvM for tail-heavy distributions, K-S for general fit, and supplementing statistical output with visual tools like QQ-plots (D'Agostino, 2017). These results highlight the importance of matching GoF tests not only to the data but also to the type of risk being modeled. This is especially relevant in pricing, reserving, and capital adequacy tasks that rely on robust tail modeling.

From a regulatory standpoint, the study supports evolving expectations for more transparent and rigorous validation frameworks (Stricker et al., 2014). Insurers and actuaries should be encouraged to disclose not only test results but also justifications for test selection, awareness of test limitations, and supporting diagnostics such as

power analysis and simulation evidence. Regulators may also consider codifying recommendations that distinguish test applicability by model type, for example, emphasizing AD and CvM in catastrophe or operational risk modeling.

In addition to methodological refinement within the classical framework, this study highlights the potential value of exploring alternative inferential procedures. One such alternative is ***Bayesian model comparison***, where models are evaluated using ***Bayes factors*** rather than p-values. Bayes factors provide a formal mechanism for weighing evidence between models while naturally incorporating prior uncertainty (Kass and Raftery, 1995), which is an advantage particularly useful when data are sparse or modeling stakes are high. Although not implemented in this paper, future research in actuarial science may benefit from integrating Bayesian methods alongside simulation-based validation, offering more flexible and interpretable decision tools.

In summary, this research calls for a reassessment of how model adequacy is determined in actuarial science. Rather than relying solely on p-value thresholds, practitioners and regulators should adopt validation strategies that reflect model complexity, test limitations, and practical significance. Doing so will enhance the credibility of actuarial models and improve the quality of risk-based decisions in an increasingly data-driven insurance landscape.

While this study offers meaningful insights, it is not without limitations. The analysis is grounded in simulated data generated from parametric models with controlled deviations, which although ideal for isolating the behavior of GoF tests, do not capture the full complexity of real-world insurance datasets. In practice, actuarial data are subject to features such as reporting delays, censoring, claim heterogeneity, structural shifts, and outliers, which can induce more severe and varied forms of model misspecification than those considered here. As a result, the performance of GoF tests observed under simulated conditions may differ when applied to operational datasets.

To bridge this gap, future research should incorporate real-world data drawn from various insurance lines, such as auto, property, or catastrophe claims. These applications would help validate the robustness of GoF tests under messy, high-stakes conditions and could guide best practices for model adequacy assessments. They would also offer an opportunity to explore how real-world data challenges interact with diagnostic tools, potentially informing refinements to test procedures or motivating the use of hybrid approaches that combine classical, simulation-based, and Bayesian methods.

Acknowledgment

I would like to express my sincere gratitude to Professor Tatjana Miljkovic, Actuarial Science Advisor and Associate Professor at Miami University, for her invaluable supervision and guidance throughout this research. I am also deeply thankful to my friend and coursemate, Simon Atoyire, for his insightful assistance during the development of the methodology section. Additionally, I extend my heartfelt appreciation to Isaiah, a Consultant at the Howe Writing Center at Miami University, whose guidance was instrumental in helping me shape and

structure the introduction and literature review. Their support and contributions have greatly enhanced the quality of this work.

References

Adamu, N.Y., 2024. A bayesian approach to weibull distribution with application to insurance claims data .

Afify, A.Z., Gemeay, A.M., Ibrahim, N.A., 2020. The heavy-tailed exponential distribution: risk measures, estimation, and application to actuarial data. Mathematics 8, 1276.

Ahmad, Z., Mahmoudi, E., Dey, S., Khosa, S.K., 2020a. Modeling vehicle insurance loss data using a new member of t-x family of distributions. Journal of Statistical Theory and Applications 19, 133–147.

Ahmad, Z., Mahmoudi, E., Kharazmi, O., 2020b. On modeling the earthquake insurance data via a new member of the t-x family. Computational intelligence and neuroscience 2020, 20.

Altman, D.G., Bland, J.M., 2005. Standard deviations and standard errors. Bmj 331, 903.

Anderson, T.W., 2011. Anderson-darling tests of goodness-of-fit. International encyclopedia of statistical science 1, 52–54.

Bakar, S.A., Hamzah, N.A., Maghsoudi, M., Nadarajah, S., 2015. Modeling loss data using composite models. Insurance: Mathematics and Economics 61, 146–154.

Benjamin, D., Berger, J., Johannesson, M., Nosek, B., . Wagenmakers, ej, berk, r.,... johnson, ve (2018). redefine statistical significance. nature human behaviour, 2 (1), 6-10. .

Benjamin, D.J., Berger, J.O., 2019. Three recommendations for improving the use of p-values. The American Statistician 73, 186–191.

Blostein, M., Miljkovic, T., 2019. On modeling left-truncated loss data using mixtures of distributions. Insurance: Mathematics and Economics 85, 35–46.

Blume, J.D., Greevy, R.A., Welty, V.F., Smith, J.R., Dupont, W.D., 2019. An introduction to second-generation p-values. The American Statistician 73, 157–167.

Bonhomme, S., Weidner, M., 2022. Minimizing sensitivity to model misspecification. Quantitative Economics 13, 907–954.

Brazauskas, V., Serfling, R., 2003. Favorable estimators for fitting pareto models: A study using goodness-of-fit measures with actual data. ASTIN Bulletin: The Journal of the IAA 33, 365–381.

Cai, T.T., Wu, Y., 2014. Optimal detection of sparse mixtures against a given null distribution. IEEE Transactions on Information Theory 60, 2217–2232.

Cirrone, G., Donadio, S., Guatelli, S., Mantero, A., Mascialino, B., Parlati, S., Pia, M., Pfeiffer, A., Ribon, A., Viarengo, P., 2004. A goodness-of-fit statistical toolkit. IEEE Transactions on Nuclear Science 51, 2056–2063.

D'Agostino, R., 2017. Goodness-of-fit-techniques. Routledge.

Darling, D.A., 1957. The kolmogorov-smirnov, cramer-von mises tests. The annals of mathematical statistics , 823–838.

Dickhaus, T., Dickhaus, T., 2018. Goodness-of-fit tests. Theory of Nonparametric Tests , 37–46.

Donoho, D., Jin, J., 2004. Higher criticism for detecting sparse heterogeneous mixtures .

Eling, M., 2012. Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models? Insurance: Mathematics and Economics 51, 239–248.

Eling, M., Loperfido, N., 2017. Data breaches: Goodness of fit, pricing, and risk measurement. Insurance: mathematics and economics 75, 126–136.

Engmann, S., Cousineau, D., 2011. Comparing distributions: the two-sample anderson-darling test as an alternative to the kolmogorov-smirnoff test. Journal of applied quantitative methods 6.

Feuerverger, A., 2016. On goodness of fit for operational risk. International Statistical Review 84, 434–455.

Gui, W., Huang, R., Lin, X.S., 2018. Fitting the erlang mixture model to data via a gem-cmm algorithm. Journal of Computational and Applied Mathematics 343, 189–205.

Head, M.L., Holman, L., Lanfear, R., Kahn, A.T., Jennions, M.D., 2015. The extent and consequences of p-hacking in science. PLoS biology 13, e1002106.

Huang, Y., Meng, S., 2020. A bayesian nonparametric model and its application in insurance loss prediction. Insurance: Mathematics and Economics 93, 84–94.

Hubbard, R., Lindsay, R.M., 2008. Why p values are not a useful measure of evidence in statistical significance testing. Theory & Psychology 18, 69–88.

Hung, H.J., O'Neill, R.T., Bauer, P., Köhne, K., 1997. The behavior of the p-value when the alternative hypothesis is true. Biometrics , 11–22.

Imbens, G.W., 2021. Statistical significance, p-values, and the reporting of uncertainty. Journal of Economic Perspectives 35, 157–174.

Janssen, A., 2000. Global power functions of goodness of fit tests. The Annals of Statistics 28, 239–253.

Jäntschi, L., Bolboacă, S.D., 2018. Computation of probability associated with anderson–darling statistic. Mathematics 6, 88.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the american statistical association 90, 773–795.

Keatinge, C.L., 1999. Modeling losses with the mixed exponential distribution, in: Proceedings of the Casualty Actuarial Society, pp. 654–698.

Koyuncu, A., Karahasan, M., 2024. A new goodness of fit test for complete or type ii right censored samples. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi 28, 240–250.

Laio, F., 2004. Cramer–von mises and anderson-darling goodness of fit tests for extreme value distributions with unknown parameters. Water Resources Research 40.

Lakens, D., 2021. The practical alternative to the p value is the correctly used p value. Perspectives on psychological science 16, 639–648.

Lanzante, J.R., 2021. Testing for differences between two distributions in the presence of serial correlation using the kolmogorov–smirnov and kuiper's tests .

Lem, S., 2015. The intuitiveness of the law of large numbers. ZDM 47, 783–792.

Lewis, P.A., 1961. Distribution of the anderson-darling statistic. The Annals of Mathematical Statistics , 1118–1124.

Ma, Z.G., Ma, C.Q., 2013. Pricing catastrophe risk bonds: A mixed approximation method. Insurance: Mathematics and Economics 52, 243–254.

McShane, B.B., Gal, D., Gelman, A., Robert, C., Tackett, J.L., 2019. Abandon statistical significance. The American Statistician 73, 235–245.

Miljkovic, T., Grün, B., 2016. Modeling loss data using mixtures of distributions. Insurance: Mathematics and Economics 70, 387–396.

Moscovich-Eiger, A., Nadler, B., Spiegelman, C., 2013. The calibrated kolmogorov-smirnov test. arXiv preprint arXiv:1311.3190 65, 694–706.

Murtaugh, P.A., 2014. In defense of p values. Ecology 95, 611–617.

Pawitan, Y., 2020. Defending the p-value. arXiv preprint arXiv:2009.02099 .

Razali, N.M., Wah, Y.B., et al., 2011. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. Journal of statistical modeling and analytics 2, 21–33.

Resnick, S.I., 2007. Heavy-tail phenomena: probabilistic and statistical modeling. volume 10. Springer Science & Business Media.

Reynkens, T., Verbelen, R., Beirlant, J., Antonio, K., 2017. Modelling censored losses using splicing: A global fit strategy with mixed erlang and extreme value distributions. Insurance: Mathematics and Economics 77, 65–77.

Rolke, W., Gongora, C.G., 2021. A chi-square goodness-of-fit test for continuous distributions against a known alternative. Computational Statistics 36, 1885–1900.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global sensitivity analysis: the primer. John Wiley & Sons.

Shankar, P.M., 2019. Pedagogy of chi-square goodness of fit test for continuous distributions. Computer Applications in Engineering Education 27, 679–689.

Stricker, M., Wang, S., Strommen, S.J., 2014. Model validation for insurance enterprise risk and capital models. Sponsored by CAS, CIA, SOA Joint Risk Management Section .

Tomczak, M., Tomczak, E., 2014. The need to report effect size estimates revisited. an overview of some recommended measures of effect size .

Trafimow, D., 2019. My ban on null hypothesis significance testing and confidence intervals, in: Structural Changes and their Econometric Modeling 12, Springer. pp. 35–48.

Wang, C., Zhu, H., 2024. Tests of fit for the power function lognormal distribution. Plos one 19, e0298309.

Wasserstein, R.L., Lazar, N.A., 2016. The asa statement on p-values: context, process, and purpose.

Wilks, D., 2016. "the stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. Bulletin of the American Meteorological Society 97, 2263–2273.

Xiang, X., Ao, T., Xiao, Q., Li, X., Zhou, L., Chen, Y., Bi, Y., Guo, J., 2022. Parameter sensitivity analysis of swat modeling in the upper heihe river basin using four typical approaches. Applied Sciences 12, 9862.