The PYSATL Experiment Framework

Alexey Mironov, Lev Golofastov, Viacheslav Gorikhovskii
Saint Petersburg State University
Saint Petersburg, Russia
{st056067, st102178}@student.spbu.ru, v.gorikhovskii@spbu.ru

***Abstract*—Goodness-of-fit testing is a statistical methodology used to assess whether a dataset conforms to a hypothesized theoretical distribution or model. This process is critical across scientific and industrial domains — from validating normality assumptions in medical research to evaluating financial risk models — as it ensures the reliability of subsequent analyses and conclusions. However, the effectiveness of such testing depends on the choice of criteria, which vary in their sensitivity to sample size, significance level, and alternative hypotheses.**

To address this challenge, we propose a flexible, open-source framework designed for systematic comparison of goodness-of-fit criteria. The framework enables researchers to configure experiments by adjusting parameters such as sample size, significance level, and alternative distributions, while offering modular integration under any criterion. Its architecture decouples data generation, criterion application, and result analysis, ensuring reproducibility and scalability.

Using this framework, we provide a comprehensive comparison of normality criteria, evaluating their performance under varying sample sizes and alternative distributions. The results demonstrate significant differences in criterion power and robustness, underscoring the importance of context-aware methodology selection. This work advances statistical practice and supports the development of new criteria.

I. Introduction

Trying to describe the main tasks of mathematical statistics in one sentence, we can say that we want to draw conclusions about the world around us based on observations. Various methods help us draw these conclusions, one of which is goodness-of-fit testing. This method is used to assess how well a given data set matches theoretical models or distributions. This type of testing plays an important role in various scientific disciplines and industries, including biology, medicine, physics, economics, engineering, marketing, and psychology, where verification of data compliance with a theoretical model is necessary to obtain reliable conclusions and ensure the correctness of subsequent analysis. [1] [2] [3]

To conduct such testing, goodness-of-fit criteria are used. They represent some rules, which become a base for a decision of rejecting the hypothesis. The criteria are based on the sample data statistics calculation. Statistics always have a limit distribution, but its analytical representation is not always known.

For example, there are many developed criteria for goodness-of-fit testing for a normal distribution. [4] The main question is which criterion is better to use in a particular situation, since they all give different results, depending on the parameters of the experiment, such as the significance level, the size of the input data, the alternative hypothesis and its parameters. In this regard, comparative studies are conducted with different lists of criteria.

Currently, there are several tools for goodness-of-fit testing, such as SciPy, R packages and others. [5] [6] Their main problem is the lack of flexibility in testing and the lack of functionality for comparing criteria. We will talk about this in detail in Section 2.

In addition, new criteria are developed over time. Consequently, it makes sense to compare them with well-known criteria in order to obtain information about the appropriateness of their application.

All of the above led us to the idea of creating a framework for conducting statistical experiments comparing the goodness-of-fit criteria. Designing the framework, the following requirements were highlighted:

1) Configurability of the experiment, taking into account the parameters mentioned above.
2) Easy adding a new criterion to the framework.
3) An option to execute criteria independently.

In this article, we present an open-source framework[1] for conducting statistical experiments based on goodness-of-fit criteria, as well as the results of a comparison of goodness-of-fit criteria powers for normal distribution carried out using the framework.

II. Related Work

The following section provides a structured overview of existing tools and frameworks related to goodness-of-fit testing and criterion comparison. This review highlights the limitations of current solutions, which the proposed framework aims to address.

1) **Classical Statistical Packages**
   - **R (stats, goftest, fitdistrplus)**
     The R programming language offers packages such as *stats*, which includes functions for common goodness-of-fit tests (e.g., Pearson's chi-square via *chisq.test*, Kolmogorov-Smirnov via *ks.test*, and Shapiro-Wilk for normality via *shapiro.test*). Packages like *goftest* and *fitdistrplus* extend functionality by implementing tests such as Anderson-Darling and providing visualization tools. However, these tools lack built-in functionality for criterion comparison or customizable experimental configurations (e.g., variable sample sizes, significance levels).

[1] https://github.com/PySATL/pysatl-experiment

- **Python (SciPy, statsmodels)**
  Libraries such as *scipy.stats* (Kolmogorov-Smirnov, chi-square criteria and others) and *statsmodels* (normality criteria like Lilliefors) enable goodness-of-fit testing but focus on individual tests rather than systematic criterion evaluation. For instance, comparing criteria for normality requires manual result aggregation and analysis, as no infrastructure exists for automated benchmarking.

2) **Commercial Platforms**
   - **MATLAB (Statistics and Machine Learning Toolbox)** Provides functions such as *chi2gof*, *kstest*, and *lillietest*, but similar to R/Python, it lacks native support for criteria comparison within a unified framework. Users must manually program loops to analyze criterion sensitivity.
   - **SAS (PROC UNIVARIATE)**
     Offers normality tests (e.g., Shapiro-Wilk, Kolmogorov-Smirnov) but is constrained by its proprietary interface and limited customization.
   - **JMP (Distribution Platform)**
     A graphical interface for distribution analysis with automated test applications. However, criterion comparisons require manual data export and post-processing.
3) **Research Tools**
   - **Minitab, SPSS**
     Commercial GUI-based software with basic goodness-of-fit tests (e.g., chi-square, Anderson-Darling). Their limitations include the absence of APIs for experiment automation and poor support for new criteria.
   - **HypothesisTests.jl (Julia)**
     A Julia package for goodness-of-fit testing with extensibility features. However, Julia's ecosystem is less widespread, and the tool lacks infrastructure for criterion benchmarking.

A. Drawbacks of existing solutions

1) **Lack of unified frameworks**
   Most tools implement individual criteria but there is no infrastructure for systematic comparison. Users must manually configure experiments and aggregate results.
2) **Limited flexibility**
   Even in robust environments like R or Python, integrating new criteria demands custom code development, pipeline integration, and validation.
3) **Insufficient support for experimental parameters**
   Many tools do not allow to variate following parameters easily:
   - Significance levels;
   - Sample sizes;
   - Alternative distribution parameters;
   - Data generation and visualization configuration.

B. How the proposed framework addresses these drawbacks

As outlined in the introduction, the proposed framework resolves the above issues through:

- **Modular architecture:** decoupled components for data generation, criteria, and analysis.
- **Configurability:** experiment parameters defined by user.
- **Extensibility:** simplified addition of new criteria via standardized interfaces.
- **Comparative analysis:** built-in functionality for evaluating power and result visualization.

III. Framework

In this section we describe the PYSATL Experiment framework.

A. Workflow, Architecture, and Implementation

The architecture of the proposed solution is illustrated in Fig. 1. The system is developed using Python and employs the following technical stack: SQLAlchemy as the ORM framework, Matplotlib for creating visualizations, Scipy for fundamental statistics algorithms, Pandas for data manipulation, and Numpy for scientific computing and optimizations. The core of the application is a pipeline-based experiment mechanism, with the start point interface. The system includes three primary modules:

1) Data Generator: This module is designed to synthesize datasets with diverse statistical properties and distributions. This module is responsible for generating synthetic data that adheres to specified probability distributions. The module supports the generation of data from a wide range of probability distributions, including but not limited to: Gaussian, Uniform Distribution, Exponential Distribution, Log-Normal Distribution.
2) Worker: This module calculates metrics using generated data obtained from the Data Generator. This module is designed to handle large-scale datasets efficiently and is a critical component in the data pipeline, enabling it to calculate actionable information from synthetic data.
3) Report Generator: This module is a comprehensive tool designed to automate the creation of detailed, structured, and visually appealing reports from raw data, obtained in worker. The module supports the use of customizable generators to ensure consistency in report formatting and styling.

The workflow of the application is structured as follows:

1) The experiment creation process begins with the configuring pipeline (threads count, sample sizes, goodness-of-fit tests, alternatives).
2) Then, data is generated according to the experiment configuration. This phase ensures that the data is representative, high-quality, and suitable for addressing the research question.
3) Generated data is processed by Worker, which calculates statistics of interest to the researcher.
4) Based on calculated results, a detailed report is generated.

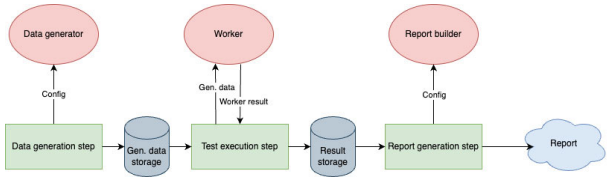

Fig. 1. The architecture of the PYSATL Experiment framework

B. Framework features

The PYSATL Experiment framework encompasses the following features, designed to enhance user experience and efficiency in conducting an experiment:

1) Pipeline customization: The framework allows users to customize every stage of the pipeline to meet their specific needs. This flexibility is essential for addressing the diverse requirements of modern data-driven research and applications. The framework's modular architecture enables users to design, modify, and optimize their analysis pipelines, ensuring that they can adapt to varying data types, research questions, and computational constraints.
2) Progress Monitoring: A progress bar is integrated into the framework, providing users with detailed information on the remaining workload. This feature displays the number of generated data left to process and the average time required per generating data, enabling users to accurately estimate the total time needed to complete the task.
3) Different databases support: Robust support for multiple data sources and databases, enabling users to seamlessly integrate and analyze data from diverse origins. The framework natively supports popular relational database management systems (RDBMS) such as MySQL, PostgreSQL, Oracle, SQLite and others.
4) Parallelization: The framework incorporates a specialized generation and calculation parallelization technique. Users can specify the number of threads for each step of the pipeline. This approach enables users to parallelize the generation and calculation process.
5) Efficiency Considerations: A key innovation in the proposed framework is the introduction of a shared store feature, which enables multiple users to collaboratively access and utilize precomputed critical values, significantly reducing computation time and resource usage. This feature is particularly valuable in scenarios where goodness-of-fit testing involves repeated calculations of critical values for large datasets or complex models, as it eliminates redundant computations and promotes efficiency across teams and projects.

IV. Power comparison experiment

In this work 36 goodness-of-fit tests of normality are used. Among these, some criteria are universal and can be applied to test not only normality but also other distributions, such as exponentiality, weibullness and others. Others are specifically designed for evaluating normality.

1) Chi-Square Test (CHI2) [7]
2) Kolmogorov–Smirnov (KS) [8]
3) Anderson–Darling (AD) [9]
4) Cramer–Von Mises (CVM) [10]
5) Shapiro–Wilk (SW) [11]
6) Skew (SKEW) [12]
7) Kurtosis (KURTOSIS) [12]
8) Lilliefors (LILLIE) [13]
9) D'Agostino (D) [14]
10) Shapiro–Francia (SF) [15]
11) D'Agostino–Pearson (DAP) [16]
12) Filliben (Filli) [17]
13) Martinez–Iglewicz (MI) [18]
14) Epps-Pulley (EP) [19]
15) Jarque-Bera (JB) [20]
16) Hosking $(HOSKING1 - HOSKING4)$ [21]
17) Cabaña-Cabaña (CC1) [22]
18) Chen-Shapiro (CS) [23]
19) Modified Shapiro-Wilk (SWRG) [24]
20) Doornik-Hansen (DH) [25]
21) Zhang Q (ZQ) [26]
22) Zhang Q^*(ZQS) [26]
23) Glen-Leemis-Barr (GLB) [27]
24) Bonett-Seier (BS) [28]
25) Bontemps-Meddahi (BM1, BM2) [15]
26) Zhang-Wu (ZWC, ZWA) [29]
27) Gel-Miao-Gastwirth (GMG) [30]
28) Robust Jarque-Bera (RJB) [30]
29) Looney-Gulledge (LG) [31]
30) Ryan-Joiner (RJ) [32]
31) Coin β_3^2 (COIN) [33]

A. Experiment organization

In this section the experiment is discussed. Experiment consists of 5 steps:

1) **Forms alternative hypothesis.** Three distribution groups are considered: symmetric distributions, asymmetric distributions and modified normal distributions.
2) **Data generation.** Generating 1000 samples of alternative hypothesis distributions with sample sizes 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000.
3) **Power calculation.** Calculate power for each alternative and significance levels 0.1, 0.05 and 0.01.
4) **Estimate performance.** Performance for each test is estimated and calculated mean, standard deviation, median and 0.95 percentile. To achieve this result each test was run 10000 times.
5) **Analyze results.** The obtained data was analyzed and recommendations of criteria usage were formed.

Symmetric distributions considered in this research are:

- three cases of the $Beta(a, b)$ distribution $Beta(0.5; 0.5)$, $Beta(1; 1)$, and $Beta(2; 2)$, where a and b are the shape parameters;

- three cases of the $Cauchy(t, s)$ distribution $Cauchy(0; 0.5)$, $Cauchy(0; 1)$, and $Cauchy(0; 2)$, where t and s are location and scale parameters;
- one case on the $Laplace(t, 5)$ distribution $Laplace(0; 1)$, where t and s are location and scale parameters;
- one case on the $Logistic(t, s)$ distribution $Logistic(2; 2)$, where t and s are the location and scale parameters;
- four cases on the $Student(v)$ distribution $Student(1)$, $Student(2)$, $Student(4)$, and $Student(10)$, where v is the number of degrees of freedom;
- five cases of the $Tukey(A)$ distribution $Tukey(0.14)$, $Tukey(0.5)$, $Tukey(2)$, $Tukey(5)$, and $Tukey(10)$, where A is the shape parameter;
- one case of the standard normal $N(0; 1)$ distribution.

Asymmetric distributions considered this research are:

- four cases of the $Beta(a, b)$ distribution $Beta(2; 1)$, $Beta(2; 5)$, $Beta(4; 0.5)$, and $Beta(5; 1)$;
- four cases of the $Chi - squared(v)$ distribution $\chi^2(1)$, $\chi^2(2)$, $\chi^2(4)$, and $\chi^2(10)$, where v is the number of degrees of freedom;
- six cases of the $Gamma(a, b)$ distribution $Gamma(2; 2)$, $Gamma(3; 2)$, $Gamma(5; 1)$, $Gamma(9; 1)$, $Gamma(15; 1)$, and $Gamma(100; 1)$, where a and b are the shape and scale parameters;
- one case of the $Gumbel(t, s)$ distribution $Gumbel(1; 2)$, where t and s are the location and scale parameters;
- one case of the $Lognormal(t, s)$ distribution $LN(0; 1)$, where t and s are the location and scale parameters;
- four cases of the $Weibull(a, b)$ distribution $Weibull(0.5; 1)$, $Weibull(1; 2)$, $Weibull(2; 3.4)$, and $Weibull(3; 4)$, where a and b are the shape and scale parameters.

Modified normal distributions considered in this research are:

- six cases of the standard normal distribution truncated at a and b $Trunc(a; b)$ $Trunc(-1; 1)$, $Trunc(-2; 2)$, $Trunc(-3; 3)$, $Trunc(-2; 1)$, $Trunc(-3; 1)$, and $Trunc(-3; 2)$;
- nine cases of a location-contaminated standard normal distribution, hereon termed $LoConN(p; a)$ $LoConN(0.3; 1)$, $LoConN(0.4; 1)$, $LoConN(0.5; 1)$, $LoConN(0.3; 3)$, $LoConN(04; 3)$, $LoConN(0.5; 3)$, $LoConN(0.3; 5)$, $LoConN(0.4; 5)$, and $LoConN(0.5; 5)$, which are referred 10 as NORMAL2;
- nine cases of a scale-contaminated standard normal distribution, hereon termed $ScConN(p; b)$ $ScConN(0.05; 0.25)$, $ScConN(0.10; 0.25)$, $ScConN(0.20; 0.25)$, $ScConN(0.05; 2)$, $ScConN(0.10; 2)$, $ScConN(0.20; 2)$, $ScConN(0.05; 4)$, $ScConN(0.10; 4)$, and $ScConN(0.20; 4)$;

TABLE I. COUNT TOP 5 MOST POWERFUL CRITERIA

Test	$n < 100$	$100 \leq n < 500$	$500 \leq$ n ≤ 1000
HOSKING1	66	65	63
CS	65	63	66
ZWA	65	61	63
GLB	65	60	63
DH	54	62	65
BM2	52	54	58
HOSKING2	44	46	55
BM1	43	56	62
CC1	43	47	56
GMG	36	32	35
BS	32	29	30
HOSKING4	27	37	43
MI	23	16	11
ZQS	21	11	11
D	17	17	18
SW	14	7	5
ZQ	11	7	5
SWRG	11	5	2
COIN	3	4	4
KS	1	1	0

- twelve cases of a mixture of normal distributions, hereon termed $MixN(p; a; b)$ $MixN(03; 1; 0.25)$, $MixN(04; 1; 0.25)$, $MixN(0.5; 1; 0.25)$, $MixN(0.3; 3; 0.25)$, $MixN(0.4; 3; 0.25)$, $MixN(0.5; 3; 0.25)$, $MixN(0.3; 1; 4)$, $MixN(04; 1; 4)$, $MixN(0.5; 1; 4)$, $MixN(0.3; 3; 4)$, $MixN(0.4; 3; 4)$, and $MixN(0.5; 3; 4)$.

B. Experiment result analysis

The statistical power of each criterion is calculated and aggregated in table[2].

From the table of powers was selected the top most powerful 5 criteria for each sample size, and counted the number of occurrences of each criterion in the list. Calculated number of occurrences was divided into three groups 30–90, 100–400, 500–1000. The result is presented in Table I.

It can be easily seen that HOSKING1 stands out as the most powerful criterion among the others in the range 30–90 of sampling sizes. Additionally, CS, ZWA, and GLB criteria exhibit considerable power, particularly when applied to sample sizes ranging from 30 to 90. For sample sizes ranging from 100 to 400, HOSKING1 continues to assert its dominance as the most powerful criterion. Similarly, the CS, ZWA, and GLB criteria maintain their status as formidable contenders in terms of power within this sample size range. In the context of sample sizes from 500 to 1000, a close examination of Table I reveals that the CS test emerges as the the most powerful criterion, closely followed by the DH criterion.

Despite the statistical power of goodness-of-fit criteria, it remains crucial for engineers to understand their practical effectiveness in real-world applications. Let us take a look at performance of the most powerful goodness-of-fit criteria from

[2]https://github.com/PySATL/pysatl-experiment/blob/result/normality/result/normality/report.pdf

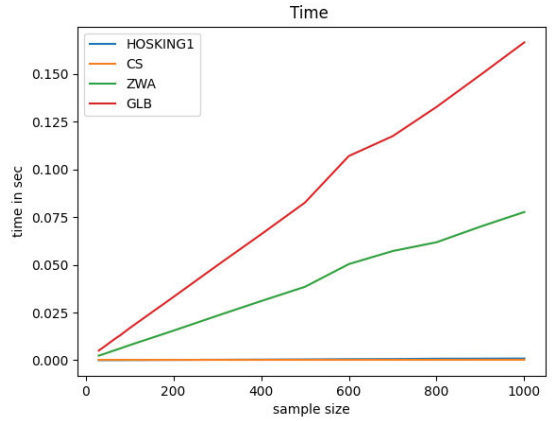

Fig. 2. Performance of HOSKING1, CS, ZWA, GLB criteria

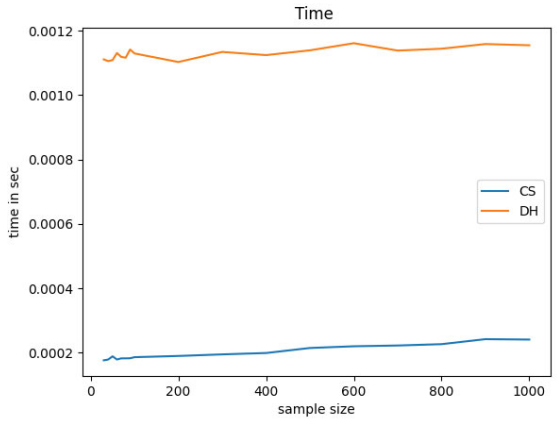

Fig. 3. Performance of CS and DH criteria

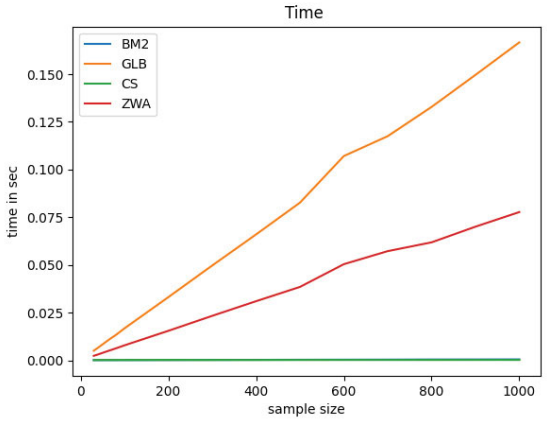

Fig. 4. Performance of BM2, CS, ZWA, GLB criteria

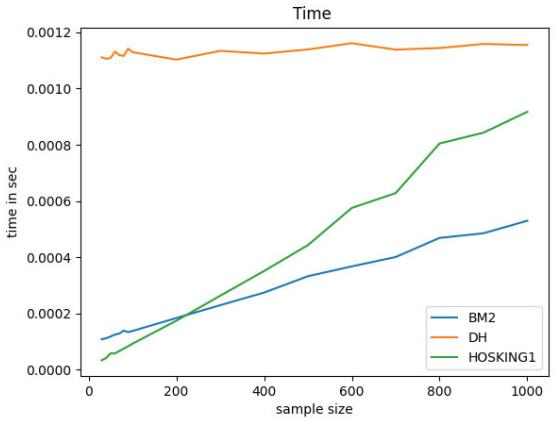

Fig. 5. Performance of HOSKING1, BM2, DH criteria

Table I: HOSKING1, CS, ZWA, GLB. Performance of these criteria is shown in Fig. 2. It can be easily seen that HOSKING1 criterion and CS criterion have better performance than ZWA, GLB.

Our following analysis will be aimed at studying each group of alternatives separately: symmetric distributions, asymmetric distributions and modified normal distributions.

The comparison result for the group of symmetric distributions is presented in Table II. It is evident that the criteria BM2, GLB, CS, and ZWA demonstrate the highest statistical power for sample sizes ranging from 30 to 90. Performance of BM2, GLB, CS, and ZWA is presented in Fig. 4. It is clear that BM2 and CS have better performance than GLB and ZWA. For larger sample sizes between 100 and 400, the criteria BM2, DH, and HOSKING1 emerge as the most powerful. Benchmarking of this criteria is presented in Fig. 5. Furthermore, the data presented in Table II indicate that BM2, GLB, and CS are the most effective criteria for sample sizes spanning 500 to 1000. Performance comparison for these criteria is shown in Fig. 6. These findings highlight the varying performance of goodness-of-fit criteria across different sample sizes, underscoring the importance of selecting appropriate criteria based on the scale of the dataset under analysis.

Another category of alternatives under investigation comprises asymmetric distributions. The results of the most powerful tests for this group are summarized in Table III. Analysis of the table reveals that HOSKING1, CS, ZWA, and DH consistently exhibit the highest statistical power across all sample sizes. Additionally, as illustrated in Fig. 2 and Fig. 3, HOSKING1 and CS demonstrate the fastest computational performance among the tested criteria. This combination of high power and efficiency makes them particularly suitable for researchers and engineers who require robust and rapid goodness-of-fit assessments, especially in time-sensitive or resource-constrained applications.

In certain scenarios, distinguishing between a standard normal distribution and a modified normal distribution is critical. Table IV presents the occurrences of the top five most powerful criteria, categorized by sample sizes, for this purpose. From Table IV, it is evident that HOSKING1, GLB, and CS are the most powerful tests for sample sizes ranging from 30 to 90, and they maintain their dominance for sample sizes between

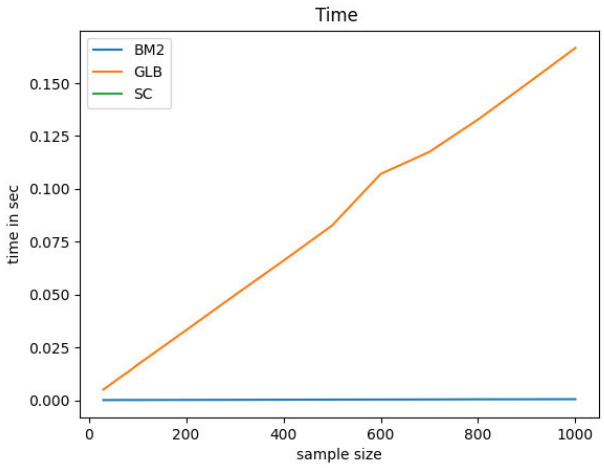

Fig. 6. Performance of BM2, CS, GLB criteria

TABLE II. COUNT TOP 5 MOST POWERFUL CRITERIA FOR SYMMETRIC

Test	$n < 100$	$100 \leq n < 500$	$500 \leq$ n ≤ 1000
BM2	14	14	15
GLB	14	13	15
CS	14	12	15
ZWA	14	12	14
DH	13	14	15
HOSKING1	13	14	14
BS	12	12	12
BM1	11	13	15
GMG	12	12	12
HOSKING2	9	9	12
CC1	8	8	13
D	5	5	7
MI	5	5	3
HOSKING4	3	7	9
COIN	3	3	3
ZQS	2	1	2
ZQ	2	1	2
SW	2	1	2
SWRG	1	1	1
KS	1	1	0

TABLE III. COUNT TOP 5 MOST POWERFUL CRITERIA FOR ASYMMETRIC

Test	$n < 100$	$100 \leq n < 500$	$500 \leq$ n ≤ 1000
HOSKING1	20	20	20
CS	20	20	20
ZWA	20	20	20
DH	19	20	20
CC1	19	20	18
GLB	19	16	18
BM2	16	15	17
BM1	14	18	19
HOSKING2	9	12	16
ZQS	8	4	4
GMG	5	5	7
HOSKING4	4	7	10
D	3	3	2
BS	3	2	4
SW	3	1	0
SWRG	3	0	0
MI	1	2	1
ZQ	1	0	1
COIN	0	1	1

TABLE IV. COUNT TOP 5 MOST POWERFUL CRITERIA FOR MODIFIED NORMAL DISTRIBUTION

Test	$n < 100$	$100 \leq n < 500$	$500 \leq$ n ≤ 1000
HOSKING1	33	31	29
GLB	32	31	30
CS	31	31	31
ZWA	31	29	29
HOSKING2	26	25	27
DH	22	28	30
BM2	22	25	26
HOSKING4	20	23	24
GMG	19	15	16
BM1	18	25	28
BS	17	15	14
MI	17	9	7
CC1	16	19	25
D	9	9	9
ZQS	9	6	5
SW	9	5	3
ZQ	8	6	2
SWRG	7	4	1

TABLE V. RESULT TABLE

Alternative	$n < 100$	$100 \leq n < 500$	$500 \leq$ n ≤ 1000
Symmetric	BM2, GLB, CS, ZWA	BM2, DH, HOSKING1	BM2, GLB, CS
Asymmetric	HOSKING1, CS	HOSKING1, CS	HOSKING1, CS
Modified	HOSKING1, GLB, CS	HOSKING1, GLB, CS	ZWA, DH

100 and 400. Upon closer examination of Table IV for the sample size range of 500 to 1000, it becomes apparent that these criteria, along with ZWA and DH, exhibit significantly higher statistical power compared to other criteria. This highlights their effectiveness in distinguishing between standard and modified normal distributions across a wide range of sample sizes.

Let us discuss the summary Table V.

Based on the analysis of the symmetric distribution group, the selection of the most appropriate goodness-of-fit criteria depends significantly on the sample size, as evidenced by the results presented in Table V. For small sample sizes (less than 100), the criteria BM2, GLB, CS, and ZWA demonstrate superior statistical power and are therefore recommended for use. These criteria are particularly effective in detecting deviations from symmetry in smaller datasets. For moderate sample sizes ranging from 100 to 500, the tests BM2, DH, and HOSKING1 emerge as the most powerful. Their robust performance in this range makes them well-suited for applications where sample sizes are neither too small nor excessively large. For larger sample sizes exceeding 500, the criteria BM2, GLB, and CS are identified as the most effective. These criteria exhibit high statistical power and reliability when applied to larger datasets, ensuring accurate assessments of distributional symmetry.

For the asymmetric group of distributions, the analysis of results presented in Table V indicates that the criteria HOSKING1 and CS consistently demonstrate superior statis-

tical power across all sample sizes. These criteria are highly effective in detecting deviations from asymmetry, regardless of whether the dataset is small, moderate, or large in size. The robustness of HOSKING1 and CS makes them particularly suitable for a wide range of applications, as they reliably identify distributional asymmetries even in challenging scenarios. This consistency in performance underscores their utility as preferred choices for goodness-of-fit testing when dealing with asymmetric distributions.

Based on the analysis of the modified normal distribution group, the optimal selection of goodness-of-fit criteria varies depending on the sample size, as detailed in Table V. For small sample sizes (less than 100), the criteria HOSKING1, GLB, and CS exhibit the highest statistical power and are therefore recommended for use. These criteria are particularly effective in identifying deviations from the modified normal distribution in smaller datasets. For moderate sample sizes ranging from 100 to 500, HOSKING1, GLB, and CS continue to demonstrate strong performance, making them reliable choices for this range. Their consistency in detecting deviations ensures accurate results for datasets of intermediate size. For larger sample sizes exceeding 500, the criteria ZWA and DH emerge as the most powerful. These criteria are better suited for handling the complexities of larger datasets, providing robust and accurate assessments of deviations from the modified normal distribution.

These findings, derived from Table V, offer clear guidance for researchers and practitioners in selecting the most appropriate goodness-of-fit criteria based on the scale of their data. This ensures both accuracy and efficiency in evaluating modified normal distributions across varying sample sizes.

V. Conclusion and Future Work

This study presents a comprehensive framework for conducting goodness-of-fit experiments in statistical analysis, designed to evaluate how well observed data aligns with theoretical models. The framework introduces a systematic approach to assessing model adequacy, incorporating robust statistical criteria and visualization tools to ensure accurate and interpretable results. To the best of our knowledge, this framework is uniquely tailored for flexibility and scalability, making it suitable for both small-scale and large-scale datasets. The framework is lightweight, customizable, and accessible, enabling researchers to perform goodness-of-fit analyses without relying on resource-intensive cloud solutions. Its modular design allows users to adapt the framework to specific research needs, including the selection of appropriate statistical criteria and the integration of additional diagnostic tools.

In the future, we plan to expand the framework by incorporating advanced features such as automated criterion selection, support for other state-of-art criteria, and enhanced visualization techniques for better interpretation of results. Additionally, we aim to integrate machine learning algorithms to improve the detection of subtle deviations between observed and expected data. These advancements will further solidify the framework's utility in a wide range of applications, from academic research to industry-specific analyses.

Acknowledgment

This work was supported by St. Petersburg State University (Pure ID 116636233).

References

[1] M. Steele, N. A. Smart, C. P. Hurst, and J. Chaseling, "Evaluating the statistical power of goodness-of-fit tests for health and medicine survey data," 2009.

[2] J. Chu, O. Dickin, and S. Nadarajah, "A review of goodness of fit tests for pareto distributions," *Journal of Computational and Applied Mathematics*, vol. 361, pp. 13–41, 2019.

[3] A. G. Sawyer and T. J. Page, "The use of incremental goodness of fit indices in structural equation models in marketing research," *Journal of Business Research*, vol. 12, no. 3, pp. 297–308, 1984.

[4] R. D. Xavier Romão and A. Costa, "An empirical power comparison of univariate goodness-of-fit tests for normality," *Journal of Statistical Computation and Simulation*, vol. 80, no. 5, pp. 545–591, 2010.

[5] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İlhan Polat, Y. Feng, E. W. Moore, J. Vanderplas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and P. van Mulbregt, "Scipy 1.0: fundamental algorithms for scientific computing in python," *Nature Methods*, vol. 17, pp. 261 – 272, 2019.

[6] J. A. V. Elizabeth González-Estrada, "An r package for testing goodness of fit: goft," *Journal of Statistical Computation and Simulation*, vol. 88, pp. 1–26, 2018.

[7] K. Pearson, *On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling*. New York, NY: Springer New York, 1992, pp. 11–28.

[8] A. Kolmogorov, "Sulla determinazione empírica di uma legge di distribuzione," *Giornale dell'Istituto Italiano degli Attuari*, vol. 4, pp. 83–91, 1993.

[9] T. W. Anderson and D. Darling, "Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes," *Annals of Mathematical Statistics*, vol. 23, pp. 193–212, 1952.

[10] R. A. L. V. Choulakian and M. A. Stephens, "Cramér-von mises statistics for discrete distributions," *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, vol. 22, no. 1, pp. 125–137, 1994.

[11] S. S. SHAPIRO and M. B. WILK, "An analysis of variance test for normality (complete samples)†," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, 12 1965.

[12] R. B. D'Agostino and A. J. Belanger, "A suggestion for using powerful and informative tests of normality," *The American Statistician*, vol. 44, pp. 316–321, 1990.

[13] H. W. Lilliefors, "On the kolmogorov-smirnov test for normality with mean and variance unknown," *Journal of the American Statistical Association*, vol. 62, pp. 399–402, 1967.

[14] M. M. Rahman and Z. Govindarajulu, "A modification of the test of shapiro and wilk for normality," *Journal of Applied Statistics*, vol. 24, pp. 219–236, 1997.

[15] F. Ahmad and R. A. Khan, "A power comparison of various normality tests," *Pakistan Journal of Statistics and Operation Research*, vol. 11, pp. 331–345, 2017.

[16] R. D'Agostino and E. S. Pearson, "Tests for departure from normality. empirical results for the distributions of b_2 and $\sqrt{b_1}$," *Biometrika*, vol. 60, no. 3, pp. 613–622, 1973.

[17] J. J. Filliben, "The probability plot correlation coefficient test for normality," *Technometrics*, vol. 17, pp. 111–117, 1975.

[18] J. Martinez and B. Iglewicz, "A test for departure from normality based on a biweight estimator of scale," *Biometrika*, vol. 68, no. 1, pp. 331–333, 1981.

[19] T. W. Epps and L. B. Pulley, "A test for normality based on the empirical characteristic function," *Biometrika*, vol. 70, no. 3, pp. 723–726, 1983.

[20] A. K. Bera and C. M. Jarque, "Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte carlo evidence," *Economics Letters*, vol. 7, pp. 313–318, 1981.
[21] J. R. M. Hosking, "L-moments: Analysis and estimation of distributions using linear combinations of order statistics," *Journal of the royal statistical society series b-methodological*, vol. 52, pp. 105–124, 1990.
[22] A. Cabana and E. M. Cabana, "Goodness-of-Fit and Comparison Tests of the Kolmogorov-Smirnov Type for Bivariate Populations," *The Annals of Statistics*, vol. 22, no. 3, pp. 1447 – 1459, 1994.
[23] L. Chen and S. S. Shapiro, "An alernative test for normality based on normalized spacings," *Journal of Statistical Computation and Simulation*, vol. 53, pp. 269–287, 1995.
[24] M. M. Rahman and Z. Govindarajulu, "A modification of the test of shapiro and wilk for normality," *Journal of Applied Statistics*, vol. 24, no. 2, pp. 219–236, 1997.
[25] J. A. Doornik and H. Hansen, "An Omnibus Test for Univariate and Multivariate Normality," *Oxford Bulletin of Economics and Statistics*, vol. 70, no. s1, pp. 927–939, December 2008.
[26] P. Zhang, "Omnibus test of normality using the q statistic," *Journal of Applied Statistics*, vol. 26, pp. 519–528, 1999.
[27] A. Glen, L. Leemis, and D. Barr, "Order statistics in goodness-of-fit testing," *IEEE Transactions on Reliability*, vol. 50, no. 2, pp. 209–213, 2001.
[28] D. G. Bonett and E. Seier, "A test of normality with high uniform power," *Computational Statistics & Data Analysis*, vol. 40, no. 3, pp. 435–445, 2002.
[29] J. Zhang and Y. Wu, "Likelihood-ratio tests for normality," *Computational Statistics & Data Analysis*, vol. 49, no. 3, pp. 709–721, 2005.
[30] Y. R. Gel, W. Miao, and J. L. Gastwirth, "Robust directed tests of normality against heavy-tailed alternatives," *Computational Statistics & Data Analysis*, vol. 51, no. 5, pp. 2734–2746, 2007.
[31] S. Looney and T. Gulledge, "Commentaries: Use of the correlation coefficient with normal probability plots," *American Statistician*, vol. 39, no. 1, pp. 75–79, Feb. 1985, copyright: Copyright 2016 Elsevier B.V., All rights reserved.
[32] T. A. Ryan and B. L. Joiner, "Normal probability plots and tests for normality," 1976.
[33] D. Coin, "A goodness-of-fit test for normality based on polynomial regression," *Computational Statistics & Data Analysis*, vol. 52, no. 4, pp. 2185–2198, 2008.