The Journal of Open Source Software

Phitter: A library designed to streamline the process of fitting and analyzing probability distributions

Sebastián José Herrera Monterrosa [1] **and Carlos Andrés Masmela Pinilla** [1]

1 Pontificia Universidad Javeriana

DOI: 10.21105/joss.07625

Software

- Review
- Repository
- Archive

Editor: Kanishka B. Narayan

Reviewers:

- @EwoutH
- @mdhaber

Submitted: 04 October 2024
Published: 02 June 2025

Summary

Phitter is an open-source Python library designed to streamline the process of fitting and analyzing probability distributions for applications in statistics, data science, operations research, and machine learning. It provides a comprehensive catalog of over 80 continuous and discrete distributions, multiple goodness-of-fit measures (Chi-Square, Kolmogorov-Smirnov, and Anderson-Darling), interactive visualizations for exploratory data analysis and model validation, and detailed modeling guides with spreadsheet implementations. By reducing the complexity of distribution fitting, Phitter helps researchers and practitioners identify distributions that best model their data.

Statement of Need

Fitting probability distributions to empirical data is a cornerstone of numerous scientific and engineering disciplines, underpinning applications such as stochastic modeling, risk assessment, and event simulation (Vose, 2008). While fitting certain distributions, like the normal distribution, is straightforward due to easily derived parameters from sample statistics, the task becomes considerably more complex when selecting the most suitable distribution from a large set of candidates or when estimating parameters for distributions requiring advanced techniques.

Commercial software packages that automate distribution fitting exist, but their cost and proprietary nature restrict access for many researchers and practitioners. Within the open-source scientific Python ecosystem, the `scipy.stats` module (Virtanen et al., 2020) provides a robust foundation, offering implementations of numerous probability distributions and parameter estimation capabilities. However, it requires users to individually select and fit each distribution, lacking a unified interface to systematically evaluate all distributions against a dataset. Each distribution is defined by shape, location, and scale parameters, and parameter estimation is typically performed using maximum likelihood estimation (MLE). While MLE is highly accurate, it is also computationally intensive, especially when applied across many candidate distributions.

To address these challenges, there is a clear need for an accessible, open-source Python library tailored for automated probability distribution fitting that:

- Provides implementations for a wide range of common distributions.
- Employs efficient parameter estimation techniques, prioritizing computational speed without sacrificing accuracy.
- Offers a user-friendly interface for fitting distributions, analyzing results, and visualizing outcomes.

Phitter has been developed to meet these requirements, offering the scientific community a specialized tool that streamlines and standardizes the process of probability distribution fitting.

By bridging the gap between general-purpose statistical libraries and proprietary software, Phitter enhances accessibility and usability for researchers and practitioners.

Key Features

1. Phitter implements probability distributions as standardized as possible according to the following sources: (McLaughlin, 2001; Walck & others, 1996; Wikipedia, 2025)
2. **Accelerated Parameter Estimation:** For a subset of implemented distributions, Phitter employs direct solutions to the distribution's parametric equations. This approach offers significant computational speed advantages compared to iterative methods like standard Maximum Likelihood Estimation (MLE), particularly beneficial for large datasets. Performance benchmarks detailing estimation times are provided Estimation Time Parameters for Continuous Distributions. Where direct parametric solutions are not implemented or feasible, Phitter seamlessly integrates with and leverages the well-established MLE implementation provided by the SciPy library (`scipy.stats.rv_continuous.fit` and related methods). This ensures broad applicability across a wide range of distributions while prioritizing speed where possible.
3. **Anderson Darling:** Phitter evaluates the Anderson-Darling distribution according to this article (Marsaglia & Marsaglia, 2004).
4. **Comprehensive Distribution Documentation:** Phitter is accompanied by detailed documentation for both continuous and discrete distributions Continuous Distributions and Discrete Distributions. This documentation outlines the mathematical formulations used and provides Excel and Google Sheet implementation for each distribution.
5. **Fit Characteristics:**

- **Extensive Distribution Library:** Access to over 80 continuous and discrete probability distributions.
- **Multiple Goodness-of-Fit Tests:** Offers a choice of Kolmogorov-Smirnov (K-S), Chi-Square, and Anderson-Darling (A-D) tests for quantitative fit evaluation.
- **Parallel Processing:** Utilizes multiprocessing for faster evaluation of multiple distributions, particularly effective for large datasets (e.g., 100K+ samples).
- **Integrated Visualizations:** Provides built-in plotting functions (Histogram/PDF overlay, ECDF vs. CDF, Q-Q plots) for visual assessment of distribution fits.
- **Automated Modeling Guides:** Generates detailed reports for best-fit distributions, including parameters, key formulas (PDF, CDF, PPF), usage recommendations, and implementation details.

6. **Simulation**: Phitter not only incorporates a robust set of functionalities for fitting and analyzing over 80 probability distributions, both continuous and discrete, but also offers capabilities for simulating processes and queues: FIFO, LIFO and PBS.

Comparison with Existing Tools

The process of fitting probability distributions to data is a fundamental step in various scientific and analytical disciplines. It allows for the modeling of random phenomena, enabling tasks such as statistical inference, forecasting, and simulation. Several Python libraries have been developed to facilitate this process, providing tools for identifying the best-fitting theoretical distribution for a given dataset. This section describes two such prominent libraries: `distfit` (Taskesen, 2020) and `fitter` (Cokelaer et al., 2024).

The `distfit` Library

The `distfit` library, created by Erdogan Taskesen and released in 2020, is a Python package designed for fitting probability density functions to univariate data. It can determine the best fit from 89 theoretical distributions using metrics like RSS/SSE, Wasserstein, KS, and Energy. Beyond parametric fitting, `distfit` also supports non-parametric methods (quantile and percentile) and discrete fitting using the binomial distribution. The library offers functionalities for predictions and a range of visualizations, including basic plots, QQ plots, and the ability to overlay multiple fitted distributions. Notably, `distfit` supports parallel computing to enhance performance and is available under the MIT License .

The `fitter` Library

The `fitter` library, developed by Thomas Cokelaer, is a Python tool for simplifying the process of fitting probability distributions to data. It automatically attempts to fit a dataset to around 80 distributions from the SciPy package, ranking them based on the sum of the square errors (SSE). `fitter` supports parallelism to speed up the fitting process, especially with larger datasets. It also provides a standalone command-line application for fitting distributions from CSV files. Users can manually specify a subset of distributions for fitting if desired. The library is under active development and is licensed under the GNU Library or Lesser General Public License (LGPL).

Speed Comparison: Phitter vs Distfit vs Fitter

The following table presents a performance comparison of the Phitter, Distfit, and Fitter libraries in terms of parameter estimation time using their default configurations. Each library was evaluated on normally distributed datasets of varying sizes: 100, 1,000, 10,000, 100,000, and 1,000,000 samples.

Library / Sample Size	100	1,000	10,000	100,000	1,000,000
Phitter	1.120	1.818	9.102	79.829	791.674
Distfit	2.604	5.279	28.575	299.398	2726.630
Fitter	37.252	30.380	31.522	401.644	1322.134

- **Phitter** tests 75 continuous probability distributions.
- **Distfit** evaluates 85 continuous distributions. See Distfit Parametric Distributions.
- **Fitter** iterates over all continuous distributions available in `scipy.stats`, automatically excluding those whose parameter estimation exceeds 30 seconds.

Goodness-of-Fit Comparison

- **Phitter** supports statistical goodness-of-fit tests including **Chi-Square**, **Kolmogorov–Smirnov**, and **Anderson–Darling**.
- **Distfit**, by default, relies on error-based metrics such as **RSS/SSE**, **Wasserstein distance**, and **energy distance**. It does not perform hypothesis testing unless explicitly instructed to use a function from `scipy.stats.goodness_of_fit`.
- **Fitter** always reports the **Kolmogorov–Smirnov** test statistic and p-value. However, its primary selection criterion is the minimization of the **sum of squared errors (SSE)**.

Documentation

Find the complete Phitter documentation here.

References

Cokelaer, T., Kravchenko, A., lahdjirayhan, msat59, Ferrari, V., Varma, A., L, B., Stringari, C. E., Brueffer, C., Broda, E., Pruesse, E., Singaravelan, K., Russo, S. A., Li, Z., padgham, mark, & negodfre. (2024). *Cokelaer/fitter: v1.7.1*. Zenodo. https://doi.org/10.5281/ZENODO.12514960

Marsaglia, G., & Marsaglia, J. (2004). Evaluating the Anderson-Darling distribution. *Journal of Statistical Software*, *9*, 1–5. https://doi.org/10.18637/jss.v009.i02

McLaughlin, M. P. (2001). *A compendium of common probability distributions*. Michael P. McLaughlin.

Taskesen, E. (2020). *distfit is a Python library for probability density fitting*. (Version 1.4.0). https://erdogant.github.io/distfit

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Vose, D. (2008). *Risk analysis: A quantitative guide*. John Wiley & Sons.

Walck, C., & others. (1996). *Hand-book on statistical distributions for experimentalists*. Stockholms universitet.

Wikipedia. (2025). *List of probability distributions — Wikipedia, the free encyclopedia*. http://en.wikipedia.org/w/index.php?title=List%20of%20probability%20distributions&oldid=1282462828.