A Solid Foundation for Statistics in Python with SciPy

Warren Weckesser, Matt Haberland

August 1, 2019
A Proposal for the Chan Zuckerberg Initiative:
Essential Open Source Software for Science

1 Introduction

SciPy is a Python library that provides fundamental building blocks for modeling and solving scientific problems. SciPy is the foundation on which more than 6750 higher level scientific libraries, including scikit-learn [1] and scikit-image [2], are built. Scientists and engineers around the world rely on SciPy for analyzing experimental data. For example, published scripts [3, 4] used in the analysis of gravitational waves [5, 6] import several subpackages of SciPy, and the M87 black hole imaging project directly cites SciPy [7]. Among SciPy's 3700 citations on GitHub are hundreds of biomedical studies on diseases including HIV [8, 9, 10], heart disease [11, 12, 13], diabetes [14], cancer [15, 16, 17], stroke [18], Alzheimer's and other neurological diseases [19, 20, 21], epilepsy [22], and more [23, 24, 25, 26, 27, 28].

As the use of Python for scientific computing continues to grow [29, 30], more scientists will look to SciPy for their statistical analysis, so it is essential that the library provides a solid collection of standard statistical tools. SciPy currently has many common statistical tests; for example, in the papers cited above we see the use of SciPy's Kolmogorov-Smirnov test [8], Mann-Whitney and Fisher's exact tests [17], and one-way ANOVA [20]. However, there are substantial gaps in SciPy's statistical functionality. The central goal of the proposed project is to fill those gaps.

There are more statistical libraries available in Python in addition to SciPy. For example, StatsModels [31, 32] describes itself as a "complement to SciPy" [33] for advanced statistical models, LinearModels [34] extends StatsModels with additional regression tools, and PyMC3 [35] provides powerful Monte Carlo methods. We are not proposing to duplicate these advanced libraries in SciPy. Rather, SciPy is a dependency of these libraries, providing fundamental statistical calculations on which they can build. The target functionality of SciPy includes the topics covered in classic texts such as Sokal and Rohlf [36], Zar [37] or Johnson [38], along with a library of probability distributions.

2 Goals

2.1 Develop SciPy Statistics

To identify gaps in the library, we surveyed its current capabilities, and we reviewed the bug reports, requests for enhancements, and pull requests on Github. We also considered classic texts on biostatistics [36, 37] and studied the syllabi of online biostatistics courses, such as those offered

by Johns Hopkins School of Public Health [39, 40, 41]. As a result of this review, we have defined the following specific objectives:

- provide fundamental and widely used hypothesis tests;
- where appropriate, include confidence intervals with the results of statistical tests;
- ensure that fitting a probability distribution to a data set is easily accomplished with an assessment of the quality of fit;
- add widely used probability distributions; and
- improve the accuracy of SciPy's probability distribution calculations.

Completion Criteria Each of these objectives is broken down into specific tasks in §3. We will know that this goal has been met when code for each of the tasks, complete with unit tests, API documentation, and benchmarks, has been merged into SciPy.

2.2 Support Other Developers

A second focus of the project is to reduce the backlog of unreviewed improvements to the statistical tools in SciPy. While there are at least three developers who work regularly on `scipy.stats` and several others who make occasional contributions, the backlog of issues and pull requests continues to grow. Currently, these volunteers can only work on SciPy in their spare time, and consequently, contributions may languish without review for weeks or months. Potentially valuable code is not being merged as fast as we would like, and perhaps even more importantly, the slow response is demotivating for new contributors. To ameliorate this situation, the developers funded by this project will dedicate 20% of their time to mentoring and reviewing the work of unfunded developers.

Completion Criteria In the past year, 230 new statistics issues have been created, but only about 190 have been closed, increasing the number of open issues by over 15%. We will know that our support of other developers has been successful if the number of open statistics issues *decreases* over the term of the project.

2.3 Diversity Outreach

The final goal is to increase the diversity of the SciPy contributor community. In collaboration with the Women in Software and Hardware club at Cal Poly, we will host a code development sprint that will introduce students to the open source community while they make meaningful contributions to the SciPy statistics tests and benchmarks.

Completion Criteria We will know that our outreach efforts have been successful if at least 20 students participate in the event, which is detailed in §3.3.7.

3 Work plan

3.1 Developer time allocation

Warren Weckesser will focus on code development. As a SciPy core developer who has merged, on average, over two PRs per month for the past nine years, Dr. Weckesser is an ideal choice for this position.

Matt Haberland's time will be balanced between code development, managing and mentoring a Cal Poly student assistant, and organizing the outreach sprint at Cal Poly. Dr. Haberland has been a core developer for SciPy for one year; among his contributions are major enhancements to SciPy's optimization subpackage [42, 43], thorough code reviews [44, 45], and an overhaul of the SciPy developer documentation [46]. After teaching the Python programming course he created at UCLA for three years and mentoring several successful undergraduate research experiences involving programming, Dr. Haberland is uniquely suited to this role.

Throughout the project, Dr. Weckesser and Dr. Haberland will also reserve 20% of their time for reviewing the pull requests (PRs) of other contributors. This will ensure that the review burden on volunteers is not increased by the proposed work.

The student assistant will write API documentation, create tutorials, and implement benchmark tests to evaluate the reliability, accuracy, and speed of SciPy's statistics functionality. The student slated for this work is Fletcher Easton, who has over three years of Python programming experience.

3.2 Proposed work and the SciPy roadmap

The goals described in §2.1 form one of six major components of SciPy's big-picture roadmap [47], and the proposed enhancements described below are also part of SciPy's detailed roadmap [48].

3.3 Specific Tasks

3.3.1 Expand tools for the analysis of variance

SciPy has only one basic test for *one-way analysis of variance* (ANOVA, see §8 of [36], §10.1 of [37], §12.1 of [38]). We propose adding tools for:

- *multiway* ANOVA (§11 of [36], §14 of [37]),
- *nested* ANOVA (§10 of [36], §15 of [37]), and
- *analysis of covariance* (§16.8 of [36], §12.6 of [38]).

3.3.2 New statistical tests

We will add:

- Alexander-Govern test [49, 50],
- Somers' D [51],
- Kendall's tau-c [52, 53],
- Page's L-test [54] (§14.11 of [36]), and
- Tukey-Kramer test (§11.1(a) of [37]).

3.3.3 Improve existing tests

Confidence intervals. There are 39 statistical tests in SciPy that return just a test statistic and a p-value. The potential for the misinterpretation and misuse of the p-value has been understood since the concept was invented. It is only in recent years, however, that the scientific community has been acting to enhance reproducibility and avoid "p-hacking" and other abuses of the p-value [55, 56, 57].

We propose to expand the results computed by these functions to include a confidence interval for the statistic (§7.10 of [36], §6.4 of [37], [58]). While the confidence interval by itself is not a solution to the problem of reproducibility, it is one of the tools recommended [55] to help avoid so-called "bright-line" hypothesis testing, where a fixed p-value, typically 0.05, is the threshold for "significance".

Options for one-sided p-values. Many of the existing tests compute only a two-sided p-value. Where appropriate, an option will be added to these tests for computing one-sided p-values.

Enhanced results for 2×2 contingency tables. In Fisher's exact test, we will add the *conditional maximum likelihood odds ratio* [59, 60] and the *relative risk* [61, 62, 63, 64].

3.3.4 Fitting probability distributions to data

The continuous distributions in SciPy all have a method for fitting the distribution to data using the method of maximum likelihood estimation (MLE). The proposed enhancements are:

- Where possible, use an analytical formula rather than a generic optimization procedure for greater speed and accuracy.
- Compute standard errors and confidence intervals of the fit parameters.
- Add the ability to fit the continuous probability distributions to *censored data* [65].
- Add fitting by the *method of moments*. Despite its limitations, it can provide a starting point for the numerical optimization routines used by other fitting methods.
- Add fitting by the method of *maximal product spacing* [66, 67, 68], an alternative to MLE. It has the advantage of always having a well-defined solution, unlike MLE, which may behave poorly for some distributions.

3.3.5 New probability distributions

We will add:

- noncentral hypergeometric distribution [69, 70],
- negative hypergeometric distribution [71],
- multivariate hypergeometric distribution [72],
- multivariate t distribution [73], and
- mixture distributions [74, 75].

3.3.6 Improve underlying code for PDF and CDF calculations

Some of the probability distributions in SciPy use the Fortran library CDFLIB [76] to compute their probability density (PDF) and cumulative distribution (CDF) functions. SciPy has already made

over a dozen patches to this code, and the heavy use of `GOTO` statements makes the remaining dozen issues challenging to fix.

By porting the vetted and actively maintained Boost C++ library [77], we will replace the legacy functions with new code that is more robust and easier to maintain.

3.3.7 Coding sprint with the Cal Poly Women in Software and Hardware

Cal Poly *Women in Software and Hardware* (WISH) [78] is a support group for female computing majors at Cal Poly that seeks to empower technical women and rectify the gender gap in the computing field. In Fall 2020, we will collaborate with WISH to host a one-day development sprint to inspire women to join the SciPy development team.

Before the sprint, the principal investigators will prepare English and mathematical descriptions of additional unit tests and benchmarks (e.g. textbook examples) that are still needed in the SciPy statistics subpackage. We will work with WISH club leadership before the event so that the *students* are prepared to lead the sprint by example. After setting up their development environment, each sprint participant will perform an assigned calculation using the appropriate SciPy functionality, compare the output of the code with the provided expected result, and make a pull request to add their code to the SciPy repository.

Besides making valuable technical contributions to the proposed work, participants will be introduced to the process of contributing to an open source project. The goal is for those who enjoy the experience to continue contributing to SciPy or other open source projects long after the event is complete.

3.4 Resources besides requested funding

SciPy has an established GitHub repository and CI suite, so no additional infrastructure is needed.

3.5 Maintenance plan

Despite the addition of new code, SciPy's statistics will be easier for volunteers to maintain after the proposed enhancements. First, the new Boost code will be more robust than the CDFLIB code it replaces, as explained in §3.3.6. Moreover, SciPy's statistics will be unified by the full-time attention of a ten-year veteran, and the result will be much more consistent than the patchwork assembled over the past two decades by developers of varying skill levels. Finally, Dr. Weckesser and Dr. Haberland intend to support their work as volunteers for the foreseeable future.

4 Existing support

SciPy has received several Small Development Grants from NumFOCUS [79] for specific projects:

Complete the SciPy Special Functions Documentation	$2500	2019
SciPy Development Documentation Overhaul	$4274	2019
An Efficient, High-Level Implementation of Linear Programming	$2000	2018
Maturing a Sparse Array Implementation for SciPy	$3000	2018

Starting June 2019, SciPy receives $2500 per month from Tidelift [80]. This funding is slated for the redesign of the SciPy website.

References

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[2] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, "scikit-image: image processing in Python," *PeerJ*, vol. 2, p. e453, 2014.

[3] A. Nitz, I. Harry, D. Brown, C. M. Biwer, J. Willis, T. D. Canton, L. Pekowsky, C. Capano, T. Dent, A. R. Williamson, S. De, M. Cabero, B. Machenschalk, P. Kumar, S. Reyes, T. Massinger, D. Macleod, A. Lenon, S. Fairhurst, A. Nielsen, shasvath, F. Pannarale, S. Khan, L. Singer, dfinstad, S. Babak, H. Gabbard, M. Tápai, P. Couvares, and L. M. Zertuche, "gwastro/pycbc: Pycbc v1.13.2 release," Nov. 2018.

[4] M. Vallisneri, J. Kanner, R. Williams, A. Weinstein, and B. Stephens, "The LIGO Open Science Center," *Journal of Physics: Conference Series*, vol. 610, no. 1, p. 012021, 2015.

[5] B. P. Abbott et al, "Gw150914: First results from the search for binary black hole coalescence with advanced ligo," *Phys. Rev. D*, vol. 93, p. 122003, Jun 2016.

[6] B. P. Abbott, R. Abbott, T. Abbott, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. Adhikari, V. Adya, *et al.*, "Gw170817: observation of gravitational waves from a binary neutron star inspiral," *Physical Review Letters*, vol. 119, no. 16, p. 161101, 2017.

[7] Event Horizon Telescope Collaboration, K. Akiyama, A. Alberdi, W. Alef, K. Asada, R. Azulay, A.-K. Baczko, D. Ball, M. Baloković, J. Barrett, and et al., "First M87 Event Horizon Telescope Results. III. Data Processing and Calibration," *The Astrophysical Journal Letters*, vol. 875, p. L3, Apr. 2019.

[8] R. A. Neher and T. Leitner, "Recombination rate and selection strength in hiv intra-patient evolution," *PLoS computational biology*, vol. 6, no. 1, p. e1000660, 2010.

[9] D. Sok, U. Laserson, J. Laserson, Y. Liu, F. Vigneault, J.-P. Julien, B. Briney, A. Ramos, K. F. Saye, K. Le, *et al.*, "The effects of somatic hypermutation on neutralization and binding in the pgt121 family of broadly neutralizing hiv antibodies," *PLoS pathogens*, vol. 9, no. 11, p. e1003754, 2013.

[10] W. Muranyi, S. Malkusch, B. Müller, M. Heilemann, and H.-G. Kräusslich, "Super-resolution microscopy reveals specific recruitment of hiv-1 envelope proteins to viral assembly sites dependent on the envelope c-terminal tail," *PLoS pathogens*, vol. 9, no. 2, p. e1003198, 2013.

[11] D. A. Liem, S. Murali, D. Sigdel, Y. Shi, X. Wang, J. Shen, H. Choi, J. H. Caufield, W. Wang, P. Ping, *et al.*, "Phrase mining of textual data to analyze extracellular matrix protein patterns

across cardiovascular disease," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 315, no. 4, pp. H910–H924, 2018.

[12] Y. Dabiri, K. L. Sack, S. Shaul, P. P. Sengupta, and J. Guccione, "Relationship of transmural variations in myofiber contractility to left ventricular ejection fraction: implications for modeling heart failure phenotype with preserved ejection fraction," *Frontiers in physiology*, vol. 9, p. 1003, 2018.

[13] A. B. Ramachandra, A. M. Kahn, and A. L. Marsden, "Patient-specific simulations reveal significant differences in mechanical stimuli in venous and arterial coronary grafts," *Journal of cardiovascular translational research*, vol. 9, no. 4, pp. 279–290, 2016.

[14] S. L. Charreyron, E. Gabbi, Q. Boehler, M. Becker, and B. J. Nelson, "A magnetically steered endolaser probe for automated panretinal photocoagulation," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. xvii–xxiii, 2018.

[15] S. Bickelhaupt, P. F. Jaeger, F. B. Laun, W. Lederer, H. Daniel, T. A. Kuder, L. Wuesthof, D. Paech, D. Bonekamp, A. Radbruch, *et al.*, "Radiomics based on adapted diffusion kurtosis imaging helps to clarify most mammographic findings suspicious for cancer," *Radiology*, vol. 287, no. 3, pp. 761–770, 2018.

[16] E. Gonçalves, M. Sciacovelli, A. S. Costa, M. G. B. Tran, T. I. Johnson, D. Machado, C. Frezza, and J. Saez-Rodriguez, "Post-translational regulation of metabolism in fumarate hydratase deficient cancer cells," *Metabolic engineering*, vol. 45, pp. 149–157, 2018.

[17] Y. S. Low, C. Blöcker, J. R. McPherson, S. A. Tang, Y. Y. Cheng, J. Y. Wong, C. Chua, T. K. Lim, C. L. Tang, M. H. Chew, *et al.*, "A formalin-fixed paraffin-embedded (ffpe)-based prognostic signature to predict metastasis in clinically low risk stage i/ii microsatellite stable colorectal cancer," *Cancer letters*, vol. 403, pp. 13–20, 2017.

[18] M. D. Schirmer, A. V. Dalca, R. Sridharan, A.-K. Giese, K. L. Donahue, M. J. Nardin, S. J. Mocking, E. C. McIntosh, P. Frid, J. Wasselius, *et al.*, "White matter hyperintensity quantification in large-scale clinical acute ischemic stroke cohorts–the mri-genie study," *NeuroImage: Clinical*, p. 101884, 2019.

[19] K. Young, A.-T. Du, J. Kramer, H. Rosen, B. Miller, M. Weiner, and N. Schuff, "Patterns of structural complexity in alzheimer's disease and frontotemporal dementia," *Human brain mapping*, vol. 30, no. 5, pp. 1667–1677, 2009.

[20] A. Tam, C. Dansereau, Y. Iturria-Medina, S. Urchs, P. Orban, H. Sharmarke, J. Breitner, P. Bellec, and A. D. N. Initiative, "A highly predictive signature of cognition and brain atrophy for progression to alzheimer's dementia," *GigaScience*, vol. 8, no. 5, p. giz055, 2019.

[21] L. Kalash, C. Val, J. Azuaje, M. I. Loza, F. Svensson, A. Zoufir, L. Mervin, G. Ladds, J. Brea, R. Glen, *et al.*, "Computer-aided design of multi-target ligands at a 1 r, a 2a r and pde10a, key proteins in neurodegenerative diseases," *Journal of cheminformatics*, vol. 9, no. 1, p. 67, 2017.

[22] C. E. Solórzano-Espíndola, B. Tovar-Corona, and Á. Anzueto-Ríos, "Pediatric seizure forecasting using nonlinear features and gaussian mixture hidden markov models on scalp eeg

signals," in *2018 15th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, pp. 1–6, IEEE, 2018.

[23] C. Lois, I. Gonzalez, D. Izquierdo-Garcia, N. R. Zurcher, P. Wilkens, M. L. Loggia, J. M. Hooker, and H. D. Rosas, "Neuroinflammation in huntington's disease: New insights with 11c-pbr28 pet/mri," *ACS chemical neuroscience*, vol. 9, no. 11, pp. 2563–2571, 2018.

[24] J. D. Warner, M. V. Irazabal, V. E. Torres, B. F. King, and B. J. Erickson, "Regional cyst concentration as a prognostic biomarker for polycystic kidney disease," in *Medical Imaging 2014: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 9038, p. 903804, International Society for Optics and Photonics, 2014.

[25] K. I. K. Cho, Y. B. Kwak, W. J. Hwang, J. Lee, M. Kim, T. Y. Lee, and J. S. Kwon, "Microstructural changes in higher-order nuclei of the thalamus in patients with first-episode psychosis," *Biological psychiatry*, vol. 85, no. 1, pp. 70–78, 2019.

[26] J. D. Zaremba, *Spatial memory in health and disease: Hippocampal stability deficits in the Df (16) A+/-mouse model of schizophrenia*. PhD thesis, Columbia University, 2017.

[27] A. Rokem, J. H. Yoon, R. E. Ooms, R. J. Maddock, M. Minzenberg, and M. A. Silver, "Broader visual orientation tuning in patients with schizophrenia," *Frontiers in human neuroscience*, vol. 5, p. 127, 2011.

[28] A. Sagar and J. D. Varner, "Dynamic modeling of the human coagulation cascade using reduced order effective kinetic models (open access)," tech. rep., Cornell University, Ithaca, NY, United States, 2015.

[29] T. Oliphant, "Python for scientific computing," *Computing in Science and Engineering*, vol. 9, no. 3, pp. 10–20, 2007.

[30] D. Robinson, "The incredible growth of python." `https://stackoverflow.blog/2017/09/06/incredible-growth-python/`, 2017.

[31] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[32] statsmodels developers, "statsmodels." `https://www.statsmodels.org/stable/index.html`.

[33] statsmodels developers, "Github: statsmodels." `https://github.com/statsmodels/statsmodels`.

[34] K. Sheppard, "linearmodels." `https://bashtage.github.io/linearmodels/`.

[35] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, "Probabilistic programming in Python using PyMC3," *PeerJ Computer Science*, vol. 2, p. e55, 2016.

[36] R. R. Sokal and F. J. Rohlf, *Biometry*. New York: Freeman, fourth ed., 2012.

[37] J. H. Zar, *Biostatistical Analysis*. Upper Saddle River, New Jersey: Prentice Hall, fifth ed., 2010.

[38] R. Johnson, *Miller & Freund's Probability and Statistics for Engineers*. Upper Saddle River, New Jersey: Pearson Prentice Hall, seventh ed., 2005.

[39] "Introduction to Biostatistics." Johns Hopkins School of Health Open Courseware, http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/IntroBiostats/coursePage/index/.

[40] "Methods in Biostatistcs I." Johns Hopkins School of Health Open Courseware, http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/MethodsInBiostatisticsI/coursePage/index/.

[41] "Methods in Biostatistcs II." Johns Hopkins School of Health Open Courseware, http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/MethodsInBiostatisticsII/coursePage/index/.

[42] M. Haberland, "Enh: added "interior-point" method for scipy.optimize.linprog." https://github.com/scipy/scipy/pull/7123.

[43] M. Haberland, "Enh: optimize: added "revised simplex" for scipy.optimize.linprog." https://github.com/scipy/scipy/pull/9263.

[44] W. Weckesser, "Enh: integrate: add args argument to solve_ivp." https://github.com/scipy/scipy/pull/10098.

[45] A. Nelson, "Enh: adds constraints for differential evolution." https://github.com/scipy/scipy/pull/10002.

[46] M. Haberland, "Development documentation overhaul #10286." https://github.com/scipy/scipy/pull/10286.

[47] SciPy developers, "Scipy roadmap." http://scipy.github.io/devdocs/roadmap.html.

[48] SciPy developers, "Detailed scipy roadmap." http://scipy.github.io/devdocs/roadmap-detailed.html#stats.

[49] R. A. Alexander and D. M. Govern, "A new and simpler approximation for ANOVA under variance heterogeneity," *Journal of Educational Statistics*, pp. 91–101, 1994.

[50] P. J. Schneider and D. A. Penfield, "Alexander and Govern's approximation: providing an alternative to ANOVA under variance heterogeneity," *The Journal of Experimental Education*, vol. 65, no. 3, pp. 271–286, 1997.

[51] R. H. Somers, "A new asymmetric measure of association for ordinal variables," *American Sociological Review*, no. 6, pp. 799–811, 1962.

[52] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[53] M. G. Kendall, "The treatment of ties in ranking problems," *Biometrika*, vol. 33, no. 3, pp. 239–251, 1945.

[54] E. B. Page, "Ordered hypotheses for multiple treatments: A significance test for linear ranks," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 216–230, 1963.

[55] R. Nuzzo, "Statistical errors," *Nature*, vol. 506, pp. 150–152, 13 February 2014.

[56] R. L. Wasserstein and N. A. Lazar, "The ASA's statement on p-values," *The American Statistician*, vol. 70, no. 2, pp. 129–133, 2016.

[57] S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman, "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations," *Eur J Epidemiol*, vol. 31, pp. 337–350, 2016.

[58] S. Nakagawa and I. C. Cuthill, "Effect size, confidence interval and statistical significance: a practical guide for biologists," *Biol Rev*, vol. 82, pp. 591–605, 2007.

[59] J. J. Gart, "Point and interval estimation of the common odds ratio in the combination of 2×2 tables with fixed marginals," *Biometrika*, vol. 57, no. 3, pp. 471–475, 1970.

[60] W. W. Hauck, "A comparative study of conditional maximum likelihood estimation of a common odds ratio," *Biometrics*, vol. 40, pp. 1117–1123, 1984.

[61] J. Zhang and K. Yu, "What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes," *JAMA*, vol. 280, no. 19, pp. 1690–1691, 1998.

[62] J. Deeks, "Odds ratios should be used only in case-control studies and logistic regression analyses," *BMJ (Letters)*, vol. 317, pp. 1155–1157, 1998.

[63] M. B. Bracken and J. C. Sinclair, "Avoidable systematic error in estimating treatment effects must not be tolerated," *BMJ (Letters)*, vol. 317, pp. 1155–1157, 1998.

[64] C. O. Schmidt and T. Kohlmann, "When to use the odds ratio or the relative risk?," *Int. J. Public Health*, vol. 53, pp. 165–167, 2008.

[65] Wikipedia editors, "Censoring (statistics)." `https://en.wikipedia.org/wiki/Censoring_(statistics)`.

[66] R. C. H. Cheng and N. A. K. Amin, "Estimating parameters in continuous univariate distributions with a shifted origin," *J. R. Statist. Soc. B*, vol. 45, no. 3, pp. 394–403, 1983.

[67] B. Ranneby, "The maximum spacing method. An estimation method related to the maximum likelihood method," *Scand. J. Statist.*, vol. 11, pp. 93–112, 1984.

[68] R. C. H. Cheng and M. A. Stephens, "A goodness-of-fit test using Moran's statistic with estimated parameters," *Biometrika*, vol. 76, no. 2, pp. 385–392, 1989.

[69] A. Fog, "Sampling methods for Wallenius' and Fisher's noncentral hypergeometric distributions," *Communications in Statistics–Simulations and Computation*, vol. 37, pp. 241–257, 2008.

[70] A. Fog, "Calculation methods for Wallenius' noncentral hypergeometric distribution," *Communications in Statistics–Simulations and Computation*, vol. 37, pp. 258–273, 2008.

[71] Wikipedia editors, "Hypergeometric distribution." `https://en.wikipedia.org/wiki/Hypergeometric_distribution`.

[72] Wikipedia editors, "Hypergeometric distribution." https://en.wikipedia.org/wiki/Hypergeometric_distribution#Multivariate_hypergeometric_distribution.

[73] Wikipedia editors, "Multivariate t-distribution." https://en.wikipedia.org/wiki/Multivariate_t-distribution.

[74] W. Seidel, "Mixture models," in *International Encyclopedia of Statistical Science* (M. Lovric, ed.), pp. 827–829, Heidelberg: Springer, 2010.

[75] Wikipedia editors, "Mixture distribution." https://en.wikipedia.org/wiki/Mixture_distribution.

[76] Numerous authors, "Netlib." http://www.netlib.org/.

[77] Numerous authors, "Boost C++ libraries." https://www.boost.org/.

[78] Women Involved in Software & Hardware, "WISH Cal Poly." https://web.calpoly.edu/~wish/.

[79] NumFOCUS. https://numfocus.org/.

[80] Tidelift. https://tidelift.com/.