




# Multiple Inference: A Python package for comparing multiple parameters

Dillon Bowen <sup>1</sup>

<sup>1</sup> Wharton School of Business, University of Pennsylvania

DOI: [10.21105/joss.04492](https://doi.org/10.21105/joss.04492)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

---

Editor: [Vissarion Fisikopoulos](#) 

## Reviewers:

- [@blakeaw](#)
- [@mattpitkin](#)
- [@nhejazi](#)

Submitted: 27 May 2022

Published: 19 July 2022

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

## Summary

Scientists often want to compare many parameters. For example, scientists often run randomized control trials to study the effects of many treatments, use observational data to compare many geographic regions, and study how public policy will impact many subgroups of people. `multiple-inference` implements many of the latest econometric and statistical tools for making such comparisons, including inference after ranking, simultaneous confidence sets, and Bayesian estimators. It uses a `statsmodels`-like API and provides template notebooks for ease of use. In just a few clicks, researchers can upload a `.csv` file of conventional estimates (e.g., ordinary least squares or instrumental variables estimates) to a Jupyter binder and click “run” to apply a suite of multiple inference tools to their data.

## Statement of need

Researchers often want to compare multiple parameters. We designed this package to help practitioners quickly, easily, and accurately perform such comparisons.

For example, there is a recent trend in social science to run large-scale studies and randomized control trials designed to test the effectiveness of many interventions ([Cortese, 2019](#)). Researchers have used large-scale field studies to test the effectiveness of many text messages reminding patients to get vaccinated ([Banerjee et al., 2021](#); [Milkman, Patel, et al., 2021](#); [Milkman et al., 2022](#)), behavioral nudges encouraging 24 Hour Fitness customers to exercise more often ([Milkman, Gromet, et al., 2021](#)), monetary and social incentives to exert effort ([DellaVigna & Pope, 2018](#)), behavioral interventions to decrease implicit racial bias ([Lai et al., 2014](#)), donation matching schemes to increase charitable giving ([Karlan & List, 2007](#)), and job training programs to increase employment among refugees in Jordan ([Caria et al., 2020](#)). Researchers also perform multiple comparisons using observational data. For example, economists often use observation data to compare many neighborhoods in terms of intergenerational mobility ([Chetty et al., 2014, 2018](#); [Chetty & Hendren, 2018](#)).

Researchers tend to ask the same set of questions when comparing many parameters.

1. Which parameters are significantly different from zero?
2. Which parameters are significantly better than the average (i.e., the average value across all parameters)?
3. Which parameters are significantly different from which other parameters?
4. What is the ranking of each parameter?
5. Which parameters might be the largest (i.e., the highest-ranked)?
6. What are the values of the parameters given their rank? e.g., What is the value of the parameter with the largest estimated value?
7. How are the parameters distributed?

Researchers often use conventional estimators like ordinary least squares (OLS) and instrumental

variables (IV) to answer such questions (Lai et al., 2014; Milkman, Patel, et al., 2021; Milkman, Gromet, et al., 2021; Milkman et al., 2022). Unfortunately, conventional estimators overestimate the value of the top-performing parameter (i.e., the parameter with the largest estimated value) and exaggerate the variability of the parameters (Andrews et al., 2022, 2019). These problems lead researchers to overstate the effectiveness of the top-performing treatments and the differences between treatment effects.

Statisticians and econometricians have advanced multiple inference tools in recent years. Recent publications describe new statistical techniques for inference after ranking (Andrews et al., 2022, 2019), multiple hypothesis testing (Romano & Wolf, 2005), rank estimation (Mogstad et al., 2020), and Bayesian estimation (Brown & Greenshtein, 2009; Cai et al., 2021; Dimmery et al., 2019; James & Stein, 1992; Stein, 1956). However, these techniques are mathematically complex and often inaccessible to all but professional statisticians.

multiple-inference solves this problem by implementing many of the latest multiple inference tools in an easy-to-use statsmodels-like API. Additionally, multiple-inference provides Jupyter binders with boilerplate code and narrative explanations to help researchers interpret the output. These binders allow researchers to upload a .csv file of their conventional estimates and click “run” to apply multiple inference tools to their data without downloading any software or writing a single line of code.

multiple-inference initially implemented the inference after ranking techniques in Andrews et al. (2019) and extended in Andrews et al. (2022). The latter paper uses multiple-inference to compare many United States commuting zones regarding intergenerational mobility. The World Bank Group is currently using multiple-inference to reanalyze the results of a multi-treatment study designed to improve tax collection in Poland (see Hernandez et al. (2017) for an earlier version of the paper).

## State of the field

multiple-inference’s defining features are inference after ranking, rank estimation, and hypothesis testing tools. Most importantly, multiple-inference contains the only implementation of the inference after ranking techniques developed in Andrews et al. (2019) and Andrews et al. (2022) in any language. These techniques correct for the winner’s curse when performing inference on top-performing parameters (e.g., the parameters that rank in the top five according to conventional estimates). Specifically, multiple-inference implements computationally efficient algorithms for computing quantile-unbiased point estimates and confidence intervals with correct coverage for parameters of specific ranks.

Mogstad et al. (2020) has an associated R package for estimating rankings. For example, it may estimate that a particular parameter has a 95% chance of being one of the three largest parameters. It also computes sets of parameters that contain all of the truly largest  $K$  parameters with 95% confidence. multiple-inference contains the only Python implementation of these techniques.

statsmodels implements multiple testing corrections that control the family-wise error rate based on p-values, such as the Holm-Bonferroni correction (Seabold & Perktold, 2010). multiple-inference implements a more powerful stepdown method based on simultaneous confidence sets for jointly Gaussian distributed estimates that also controls the family-wise error rate (Romano & Wolf, 2005). Additionally, multiple-inference implements a multiple testing procedure based on  $q$ -values that control the false discovery rate (Storey & Tibshirani, 2003).

Bayesian estimators are essential tools for multiple inference, and robust packages for Bayesian analysis already exist in Python. For example, statsmodels implements two Bayesian models (binomial and Poisson) with independent Gaussian priors (Seabold & Perktold, 2010). pymc3 is a comprehensive package for Bayesian inference (Salvatier et al., 2016). PosteriorStacker fits

parametric and nonparametric empirical Bayes models based on posterior samples from multiple datasets (Baronchelli et al., 2020). Additionally, Dimmery et al. (2019) implements a Gaussian prior Bayesian model fit using Stein-type estimation. It distinguishes itself by incorporating uncertainty about the estimates of the prior parameters into the posterior distribution.

multiple-inference aims to be a one-stop-shop for multiple inference and therefore includes parametric and nonparametric Bayesian estimators. Its Gaussian prior Bayesian estimator is most similar to the Stein-type estimator from Dimmery et al. (2019). However, Dimmery et al. (2019) does not account for correlated errors. For example, if we underestimate the prior mean and shrink all posterior estimates towards the estimated prior mean, we will underestimate many parameters. multiple-inference accounts for this correlated uncertainty in its James-Stein fit method of the Gaussian prior Bayesian model. Additionally, multiple-inference provides a maximum likelihood fit method for the Gaussian prior model, also accounting for correlated uncertainty about the prior parameters. Finally, multiple-inference's parametric Bayesian estimators provide an option to compute *robust confidence intervals* (Armstrong et al., 2020). Robust confidence intervals ensure that the confidence intervals have correct coverage even if the parametric assumptions are incorrect. While Armstrong et al. (2020) already has associated R, Stata, and Matlab packages for robust parametric empirical Bayes confidence intervals, multiple-inference contains only python implementation of this technique.

multiple-inference also implements several "intermediate products" that researchers can use in other applications. The most notable is a truncated normal distribution with two advantages over scipy's truncated normal distribution (Virtanen et al., 2020). First, multiple-inference uses an exponential tilting method to improve accuracy when the truncation set is far from the mean of the underlying normal (Z. I. Botev, 2017; Z. I. Botev & l'Ecuyer, 2015). multiple-inference uses the same exponential tilting method as the R package TruncatedNormal (Z. Botev & Belzile, 2021). (TruncatedNormal works for multivariate truncated normal distributions using convex truncation sets that are essentially hypercubes. multiple-inference's implementation applies the same exponential tilting method to univariate truncated normal distributions using both convex and concave truncation sets.) We can see the advantage of multiple-inference's implementation for the cumulative distribution function of a standard normal truncated to the interval (8,9) evaluated at 8.7.

```
>> from scipy.stats import truncnorm
>> truncnorm(8, 9).cdf(8.7)
1.0709836154559238
>> from conditional_inference.stats import truncnorm
>> truncnorm([(8, 9)]).cdf(8.7)
0.9978948153314305
```

Clearly, scipy's result is incorrect because a CDF cannot exceed 1 by definition. multiple-inference's result, by contrast, is less than 1.

Second, both scipy and TruncatedNormal require convex truncation sets, whereas multiple-inference accepts both convex and concave truncation sets. Z. I. Botev (2017) and Z. I. Botev & l'Ecuyer (2015) provide an accurate exponential tilting method for computing  $l(a, b) = Pr\{a < z < b\}$ , where  $z$  follows a standard normal distribution. We can use this method to calculate the probability density function (PDF) and CDF of a standard normal truncated to the interval  $(a, b)$ ,

$$f(z; (a, b)) = \frac{\phi(z)\mathbf{1}\{a < z < b\}}{l(a, b)},$$

and

$$F(z; (a, b)) = \frac{l(a, z)\mathbf{1}\{a < z < b\} + l(a, b)\mathbf{1}\{b \leq z\}}{l(a, b)},$$

where  $\phi$  is the PDF of a standard normal.

We can easily extend this to calculate the PDF and CDF of a standard normal truncated to the union of  $N$  non-overlapping intervals  $(a_1, b_1) \cup, \dots, \cup (a_N, b_N)$ ,

$$f(z, (a_1, b_1) \cup, \dots, \cup (a_N, b_N)) = \frac{\phi(z) \sum_{n=1}^N \mathbf{1}\{a_n < z < b_n\}}{\sum_{n=1}^N l(a_n, b_n)},$$

and

$$F(z, (a_1, b_1) \cup, \dots, \cup (a_N, b_N)) = \frac{\sum_{n=1}^N [l(a_n, z) \mathbf{1}\{a_n < z < b_n\} + l(a_n, b_n) \mathbf{1}\{b_n \leq z\}]}{\sum_{n=1}^N l(a_n, b_n)}.$$

## Acknowledgements

I would like to thank Sarah Reed and Christian Kaps for feedback on this paper. I would also like to thank Isaiah Andrews, Toru Kitagawa, Adam McCloskey, and Jeff Rowley for feedback on my early drafts of the software.

## References

- Andrews, I., Bowen, D., Kitagawa, T., & McCloskey, A. (2022). Inference for losers. *AEA Papers and Proceedings*, 112, 635–642. <https://doi.org/10.1257/pandp.20221065>
- Andrews, I., Kitagawa, T., & McCloskey, A. (2019). *Inference on winners*. National Bureau of Economic Research. <https://doi.org/10.3386/w25456>
- Armstrong, T. B., Kolesár, M., & Plagborg-Møller, M. (2020). Robust empirical bayes confidence intervals. *arXiv Preprint arXiv:2004.03448*. <https://doi.org/10.48550/arXiv.2004.03448>
- Banerjee, A., Chandrasekhar, A. G., Dalpath, S., Duflo, E., Floretta, J., Jackson, M. O., Kannan, H., Loza, F. N., Sankar, A., Schrimpf, A., & Shrestha, M. (2021). *Selecting the most effective nudge: Evidence from a large-scale experiment on immunization*. National Bureau of Economic Research. <https://doi.org/10.3386/w28726>
- Baronchelli, L., Nandra, K., & Buchner, J. (2020). Relativistic accretion disc reflection in AGN x-ray spectra at  $z = 0.5$ –4: A study of four chandra deep fields. *Monthly Notices of the Royal Astronomical Society*, 498(4), 5284–5298. <https://doi.org/10.1093/mnras/staa2684>
- Botev, Z. I. (2017). The normal law under linear restrictions: Simulation and estimation via min-imax tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1), 125–148. <https://doi.org/10.1111/rssb.12162>
- Botev, Z. I., & l'Ecuyer, P. (2015). Efficient probability estimation and simulation of the truncated multivariate student-t distribution. *2015 Winter Simulation Conference (WSC)*, 380–391. <https://doi.org/10.1109/wsc.2015.7408180>
- Botev, Z., & Belzile, L. (2021). *TruncatedNormal: Truncated multivariate normal and student distributions*. <https://CRAN.R-project.org/package=TruncatedNormal>
- Brown, L. D., & Greenshtein, E. (2009). Nonparametric empirical bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, 1685–1704. <https://doi.org/10.1214/08-aos630>

- Cai, J., Han, X., Ritov, Y., & Zhao, L. (2021). Nonparametric empirical bayes estimation and testing for sparse and heteroscedastic signals. *arXiv Preprint arXiv:2106.08881*. <https://doi.org/10.48550/arXiv.2106.08881>
- Caria, S., Gordon, G., Kasy, M., Quinn, S., Shami, S., & Teytelboym, A. (2020). *An adaptive targeted field experiment: Job search assistance for refugees in Jordan*. <https://doi.org/10.2139/ssrn.3689456>
- Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R., & Porter, S. R. (2018). *The opportunity atlas: Mapping the childhood roots of social mobility*. National Bureau of Economic Research. <https://doi.org/10.3386/w25147>
- Chetty, R., & Hendren, N. (2018). The impacts of neighborhoods on intergenerational mobility II: County-level estimates. *The Quarterly Journal of Economics*, 133(3), 1163–1228. <https://doi.org/10.3386/w23002>
- Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics*, 129(4), 1553–1623. <https://doi.org/10.3386/w19843>
- Cortese, M. J. (2019). The megastudy paradigm: A new direction for behavioral research in cognitive science. In *New methods in cognitive psychology* (pp. 67–85). Routledge. <https://doi.org/10.4324/9780429318405-4>
- DellaVigna, S., & Pope, D. (2018). What motivates effort? Evidence and expert forecasts. *The Review of Economic Studies*, 85(2), 1029–1069. <https://doi.org/10.3386/w22193>
- Dimmery, D., Bakshy, E., & Sekhon, J. (2019). Shrinkage estimators in online experiments. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2914–2922. <https://doi.org/10.1145/3292500.3330771>
- Hernandez, M., Jamison, J., Korczyk, E., Mazar, N., & Sormani, R. (2017). *Applying behavioral insights to improve tax collection*. <https://doi.org/10.1596/27528>
- James, W., & Stein, C. (1992). Estimation with quadratic loss. In *Breakthroughs in statistics* (pp. 443–460). Springer. [https://doi.org/10.1007/978-1-4612-0919-5\\_30](https://doi.org/10.1007/978-1-4612-0919-5_30)
- Karlan, D., & List, J. A. (2007). Does price matter in charitable giving? Evidence from a large-scale natural field experiment. *American Economic Review*, 97(5), 1774–1793. <https://doi.org/10.3386/w12338>
- Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E. E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., ... Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, 143(4), 1765. <https://doi.org/10.1037/a0036260>
- Milkman, K. L., Gromet, D., Ho, H., Kay, J. S., Lee, T. W., Pandiloski, P., Park, Y., Rai, A., Bazerman, M., Beshears, J., Bonacorsi, L., Camerer, C., Chang, E., Chapman, G., Cialdini, R., Dai, H., Eskreis-Winkler, L., Fishbach, A., Gross, J., ... Duckworth, A. (2021). Megastudies improve the impact of applied behavioural science. *Nature*, 600(7889), 478–483. <https://doi.org/10.1038/s41586-021-04128-4>
- Milkman, K. L., Patel, M. S., Gandhi, L., Graci, H. N., Gromet, D. M., Ho, H., Kay, J. S., Lee, T. W., Akinola, M., Beshears, J., Bogard, J., Bottenheim, A., Chabris, C., Capman, G., Choi, J., Dai, H., Fox, C., Goren, A., Hilchey, M., ... Duckworth, A. (2021). A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. *Proceedings of the National Academy of Sciences*, 118(20). <https://doi.org/10.1073/pnas.2101165118>

- Milkman, K. L., Patel, M. S., Gandhi, L., Graci, H. N., Gromet, D. M., Ho, H., Kay, J. S., Lee, T. W., Rothschild, J., Bogard, J., Brody, I., Chabris, C., Chang, E., Capman, G., Dannais, J., Goldstein, N., Goren, A., Hershfield, H., Hirsch, A., ... Duckworth, A. (2022). A 680,000-person megastudy of nudges to encourage vaccination in pharmacies. *Proceedings of the National Academy of Sciences*, 119(6). <https://doi.org/10.1073/pnas.2115126119>
- Mogstad, M., Romano, J. P., Shaikh, A., & Wilhelm, D. (2020). *Inference for ranks with applications to mobility across neighborhoods and academic achievement across countries*. National Bureau of Economic Research. <https://doi.org/10.3386/w26883>
- Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4), 1237–1282. <https://doi.org/10.1111/j.1468-0262.2005.00615.x>
- Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, 2, e55. <https://doi.org/10.7717/peerj-cs.55>
- Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*. <https://doi.org/10.25080/majora-92bf1922-011>
- Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1, 197–206. <https://doi.org/10.1525/9780520313880-018>
- Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16), 9440–9445. <https://doi.org/10.1073/pnas.1530509100>
- 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>