Pingouin: statistics in Python

Raphael Vallat¹

¹ Department of Psychology, University of California, Berkeley.

Summary

Python is currently the fastest growing programming language in the world, thanks to its ease-of-use, fast learning curve and its numerous high quality packages for data science and machine-learning. Surprisingly however, Python is far behind the R programming language when it comes to general statistics and for this reason many scientists still rely heavily on R to perform their statistical analyses.

In this paper, we present Pingouin, an open-source Python package aimed at partially filling this gap by providing easy-to-use functions for computing some of the main statistical tests that scientists use on an every day basis. This includes basics functions such as ANOVAs, ANCOVAs, post-hoc tests, non-parametric tests, effect sizes, as well as more advanced functions such as Bayesian T-tests (Rouder, Speckman, Sun, Morey, & Iverson, 2009), repeated measures correlations (Bakdash & Marusich, 2017), robust correlations (Pernet, Wilcox, & Rousselet, 2012) and circular statistics (Berens, 2009), to cite but a few. Pingouin is written in Python 3 and is mostly built on top of the Pandas (McKinney, 2010) library, therefore allowing a fluid integration within a data analysis pipeline. Pingouin comes with an extensive documentation and API as well as with several Jupyter notebook examples.

References


