The Journal of Open Source Software

resample: Simple Bootstrapping for Python

Robert M. Cenon[1] **and Dennis L. Sun**[1]

1 California Polytechnic State University, San Luis Obispo

DOI: 10.21105/joss.00788

Software

- Review
- Repository
- Archive

Submitted: 20 June 2018
Published: 24 June 2018

Summary

A central goal of statistics is to understand and to quantify uncertainty. For a few specific statistics, like the sample mean, there are formulas for estimating this uncertainty. For a general statistic though, no such formula exists.

The fundamental difficulty is that "uncertainty" describes what would happen if we were to observe repeated samples from the population, but typically we only have a single sample. The bootstrap (Efron and Tibshirani 1993) is a simulation-based technique that circumvents this issue by treating it as if it were the population and repeatedly sampling from it. The statistic can be re-calculated for each of these pseudo-samples, from which we can extract a measure of uncertainty from (Davison and Hinkley 1997). The beauty of the bootstrap is that it is a simple algorithm that works for almost any statistic (Hesterberg 1997).

`resample` is a Python package that creates a convenient framework for performing the bootstrap. This package allows the user to easily specify a variety of different bootstraps, thus serving as a counterpart to the popular `boot` library in R. The main features of `resample` include:

- bootstrap confidence intervals (including options to use Efron, Percentile, BCa)
- standard errors, point estimates, and bias
- plots of bootstrap distributions
- user defined estimators with checks that the bootstrap is valid
- integration with the `pandas` and `numpy` libraries
- specification of more complex bootstraps

Example

resample allows the user to define and run a bootstrap in one line. As an example, we will bootstrap the variance of the "mec" column from the student scores data set (Efron and Tibshirani 1993):

```
bootstrap = rs.boot(data["mec"], np.var)
```

This creates a `Results` object, which stores the original and bootstrapped statistics, in addition to having methods to calculate various estimates, plot the bootstrap distribution, etc. For example, a plot of the simulation can be obtained as follows:

```
bootstrap.plot()
```

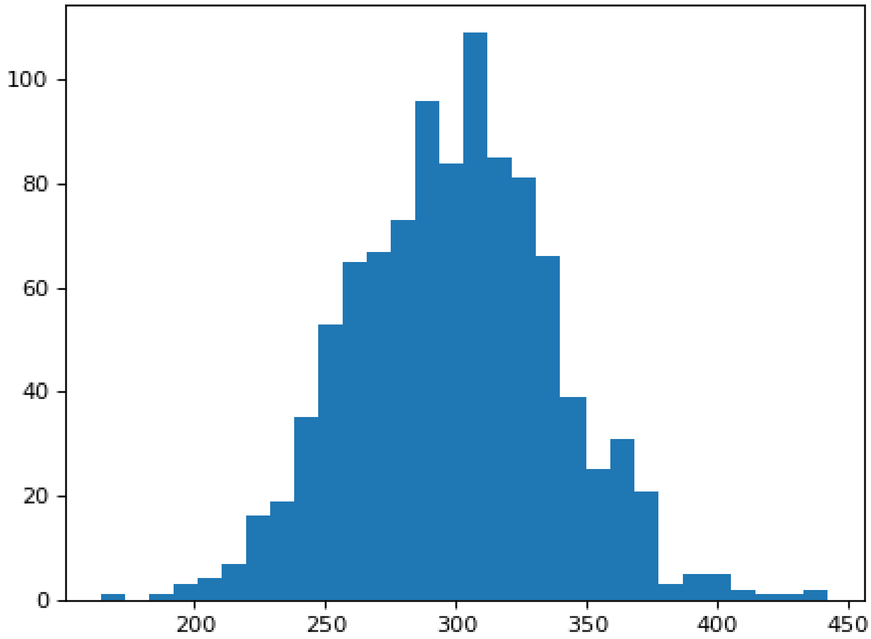

Figure 1: Example of basic plots

`resample` accommodates different confidence interval methods:

```
bootstrap.ci() #efron interval by default
bootstrap.ci(kind="loc_percentile")
bootstrap.ci(kind="BCa")
```

Full documentation with a tutorial can be accessed at https://cenonn.github.io/resample/.

References

Davison, A. C., and D.V. Hinkley. 1997. *Bootstrap Methods and Their Application.* Cambridge University Press.

Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap.*

Hesterberg, Tim. 1997. “What Teachers Should Know about the Bootstrap: Resampling in the Undergraduate Statistics Curriculum.” 1997. https://arxiv.org/abs/1411.5279.