Basic introduction to bootstrapping for confidence intervals

# What is bootstrapping?

Bootstrapping is the process of estimating the statistical properties of a dataset using random sampling without replacement. We can use bootstrapping to estimate the total population of a Rwandan district and its corresponding confidence intervals by sampling populations in each enumeration area many times and analysing their distribution. This document is a brief introduction to bootstrapping ‘from scratch’ in Python.

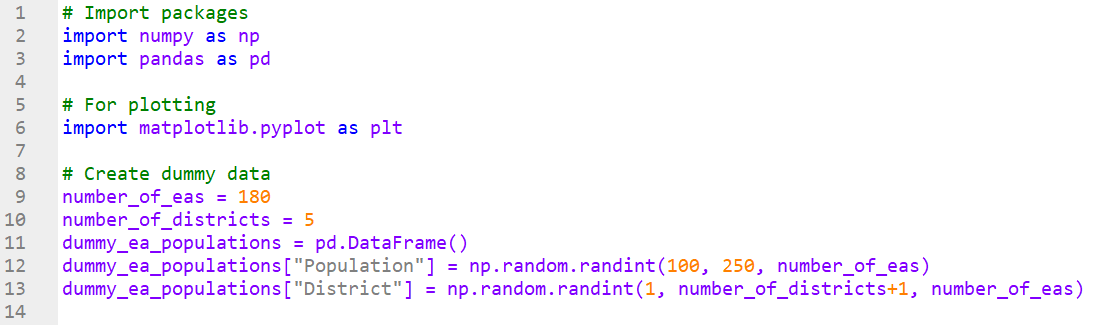
# Basic Bootstrapping for confidence intervals in Python

Our bootstrapping code can be divided into the following steps:

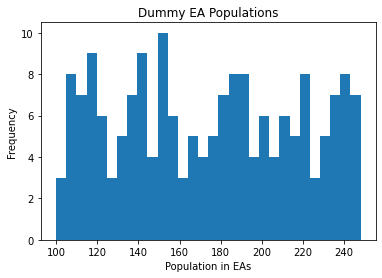
* Sampling with replacement of the dataset
* Calculating the summary statistic of interest for that sample e.g., mean or sum
* Repeat the above steps many times to create a distribution
* Use this distribution to calculate the confidence intervals

## Coded example of bootstrapping enumeration areas

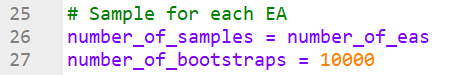
Some packages will bootstrap in one function but it’s very simple to create our own. I’m using Pandas here to create a dummy dataset. For each of the 5 districts I’m assigning a random population between 100 and 250 people.



The district populations in this dummy dataset have the following distribution:

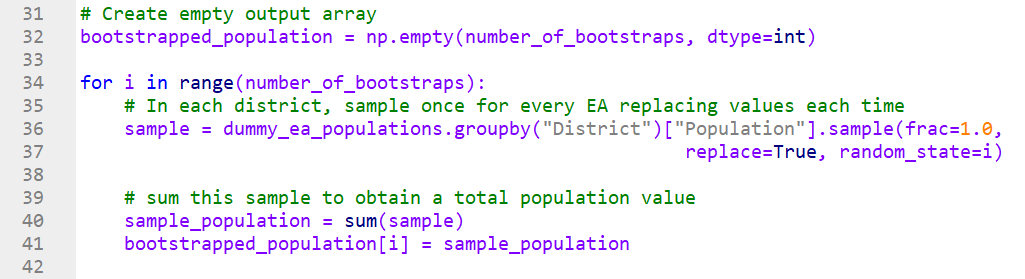


We can then define the number of samples we need to take and the number of bootstraps to perform. I’ve chosen 10,000 bootstraps here but if it’s a reasonably large number you should get similar results.

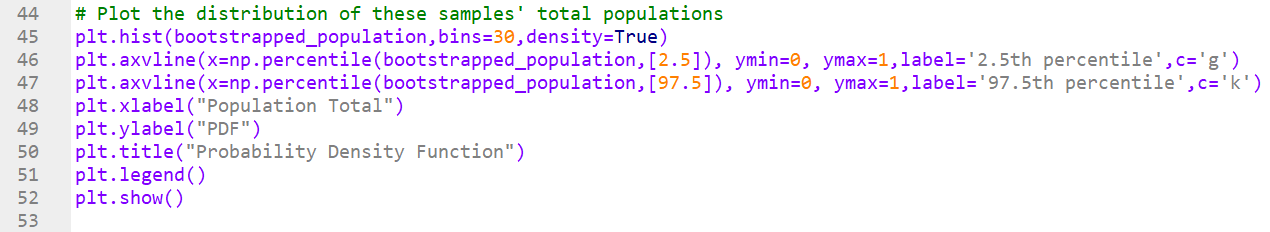


We create an empty NumPy array to return our population counts to for efficiency reasons, although with datasets this small appending to a list is fine. For each bootstrap we retrieve the same number of samples as there are EAs in each district i.e, if a district has 32 EAs we sample 32 populations, with replacement after each. The random\_state variable makes this reproducible and but isn’t necessary.

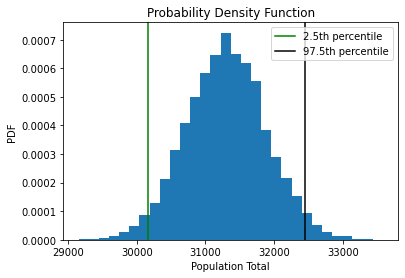
For each of our samples we then calculate the total population and insert that into the previously empty NumPy array.



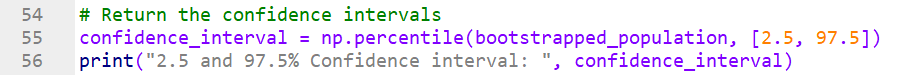
We can then plot the total populations of all our bootstraps with the below code to create a probability density function (PDF).



The PDF produced is shown below:



And finally calculate our confidence intervals with the below code:

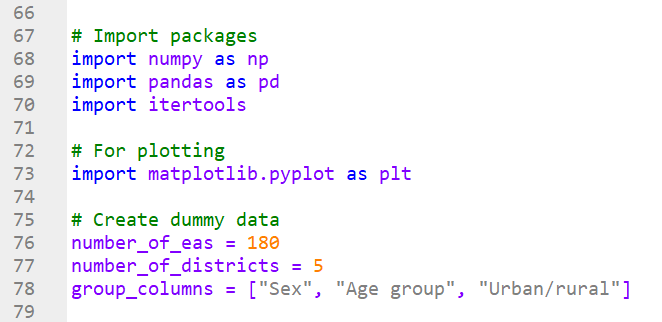




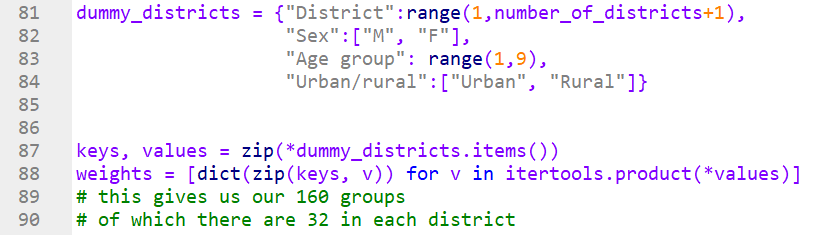
Full example code is available at <https://github.com/Alexdlewis1/ONS_Rwanda/blob/main/Bootstrapping%20example.py>

# Bootstrapping for district populations by group

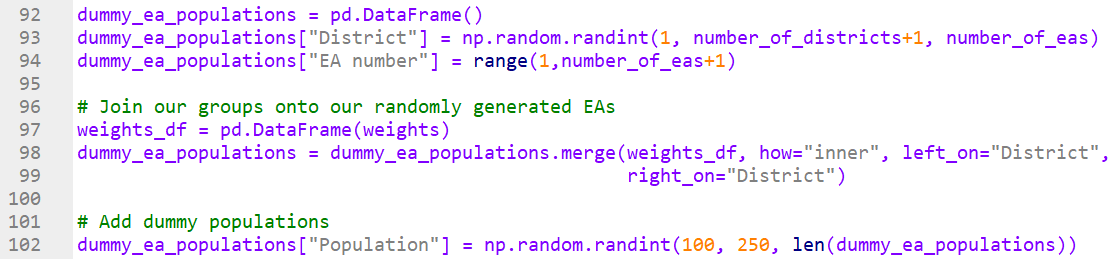
After importing the required packages, we define our groups: Sex, age group and Urban/rural. As there are 8 age groups, this gives us 32 groups per district for 160 in total.

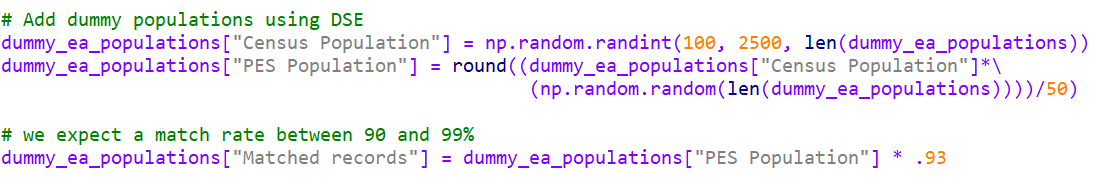


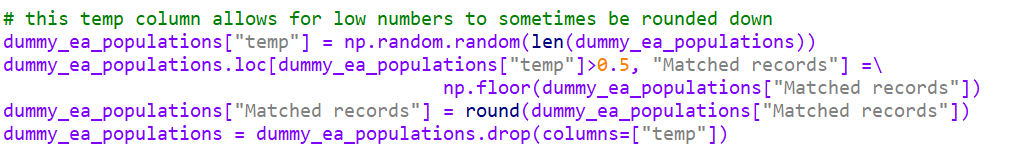
We can then dynamically create all 160 combinations using the itertools function ‘product’.



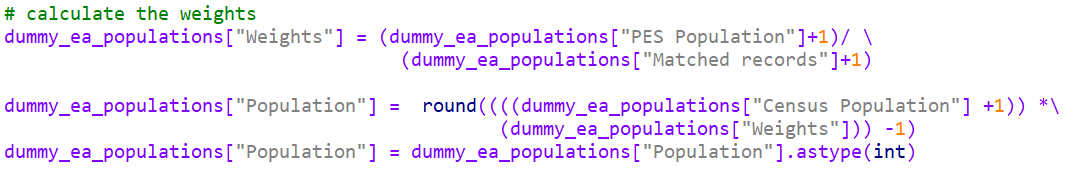
We then create our dummy data assigning census and PES populations to each group in each district. We will have this dataset already, so these steps won’t need to be recreated in the same way.



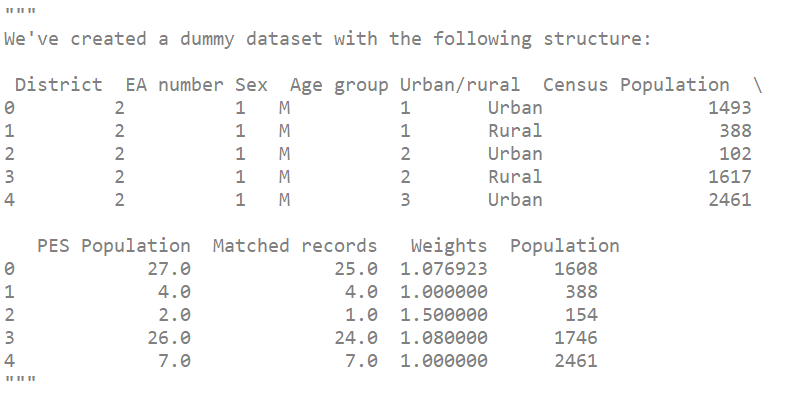


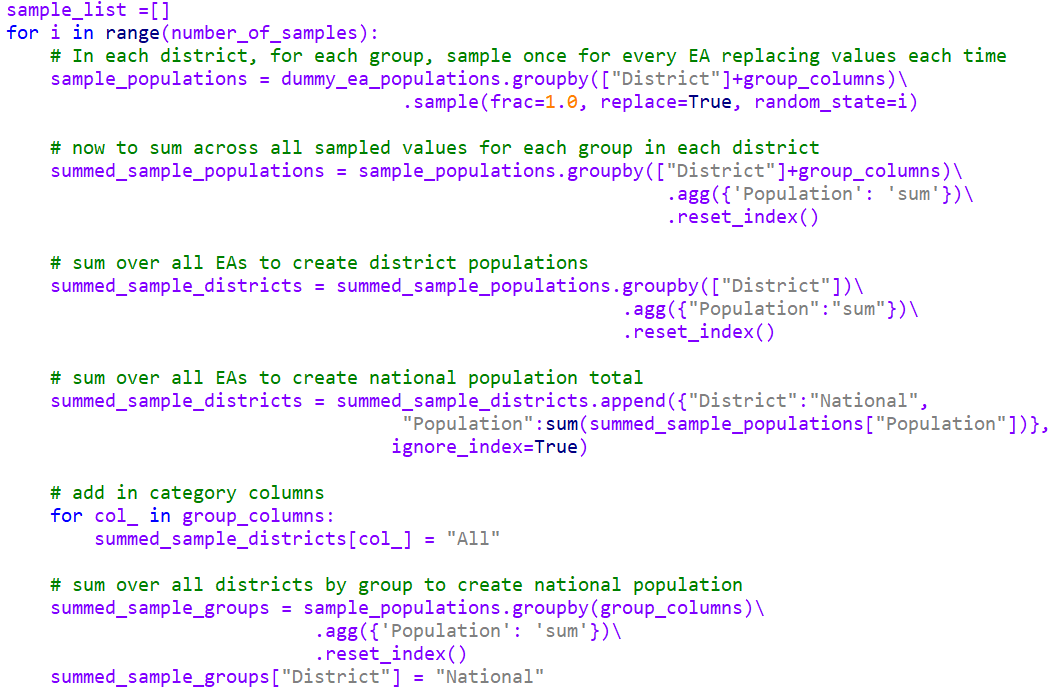
We can then estimate the number of matches between the census and PES. This looks more complicated because we are handling small numbers so a 93% match rate would often lead to small counts not being decreases correctly. 

We can finally create the weights using the census population and the number of matched records:

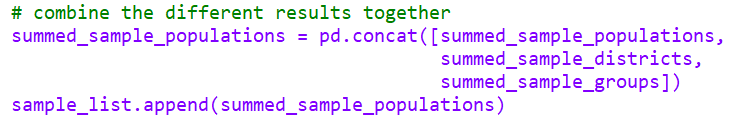


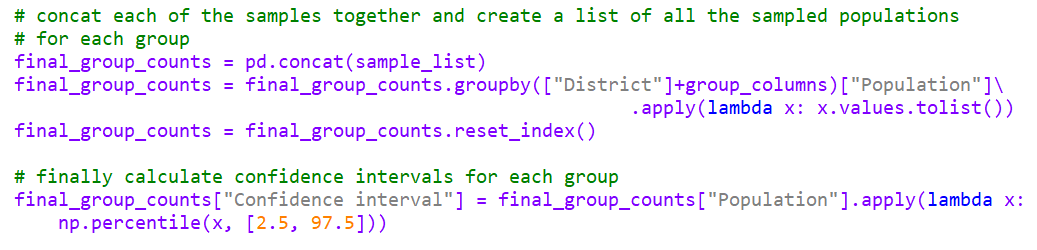
So we have dummy data with the following structure:



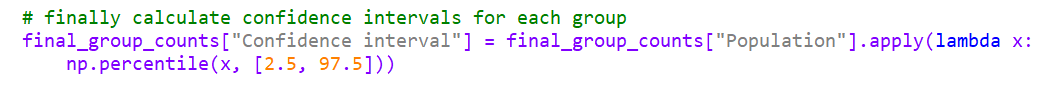
Now we can begin the bootstrapping. As before sample with replacement for each EA in each district and sum together populations in each group. We can also calculate the total populations in each district and nationally then concatenate the results together. We repeat this for each sample we need to take and collect the results into a list. 

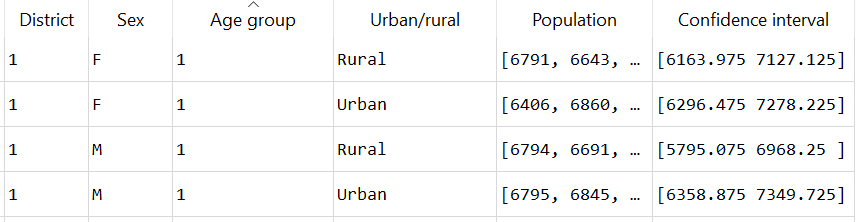
We then concatenate these dataframes in the list and collect the total populations for each group from all of the samples together into a list.





And finally use these lists to produce the confidence intervals for each group in each district:





Full example code is available at https://github.com/Alexdlewis1/ONS\_Rwanda/blob/main/Bootstrapping%20groups%20example.py