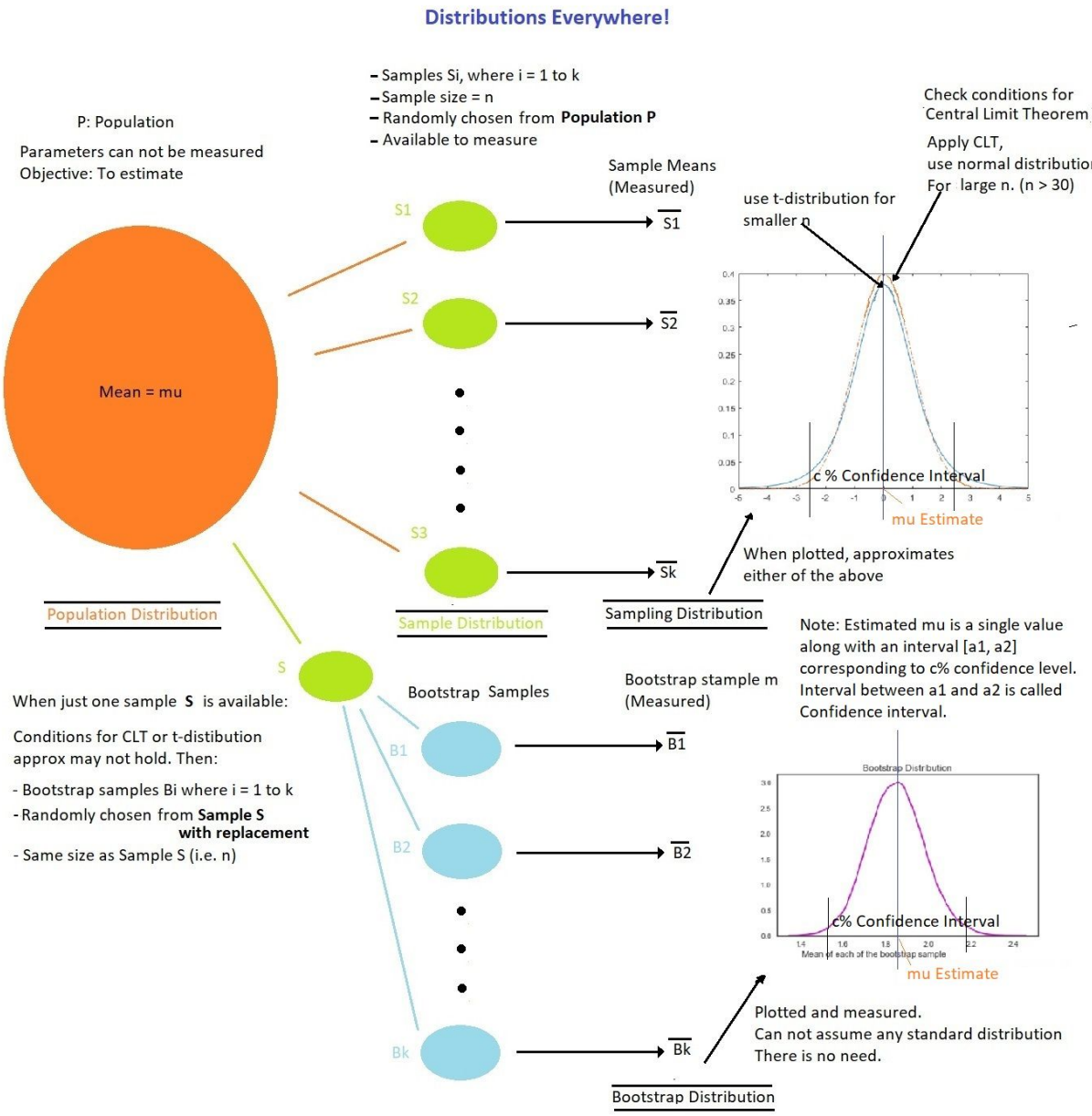
A note on Distributions: Population, Sample, Sampling and Bootstrap

By Aparna Shastry

It is very important to understand these terms individually and their relationships. This is helpful in understanding population parameter estimation.



Enlarged picture here

Population is very large and hence unknown. The objective of taking samples is to know it better.

There are two ways of taking samples from a population.

- With replacement: This is not covered in this notes. Currently not of my interest.
- Without replacement: Usually population is large enough so that samples can be taken without replacement, hence this is of my interest.

Relation between Population distribution and Sample distributions:

a) Shape:

Sample distribution has (almost) same shape as Population distribution, if

- Sample size is large enough: The larger the better it resembles the population
- Samples are independent: Samples taken without replacement for n less than 10% or preferably 5% of the population size

b) Center (mean/median/mode):

These parameters in these distributions will be close to each other with certain error. That is exactly why we use the samples to estimate them! How close can be quantified for a given sample size.

c) Spread (Variance/Standard deviation):

Often, this number for population and sample distributions are taken same to conduct the tests to determine the center. That is, measured standard deviation from sample distribution is used in the formulas for estimating the population parameters in b) above.

Relation between Sample distribution and Sampling distributions:

a) Shape:

If everything is good about choosing samples (I mean, if all conditions for Central Limit Theorem are met,) shapes of sample distribution and Sampling distributions are different!

While Sample distribution resembles the population distribution, Sampling distribution resembles normal distribution, regardless of shapes of sample or population distributions

b) Center:

Mean, Mode, Median are all equal for a normal distribution. Sampling distribution has a mean that is at the measured mean of mean of samples.

c) Spread:

Sampling distribution standard deviation is 1/sqrt(n) of that of measured sample standard deviation.

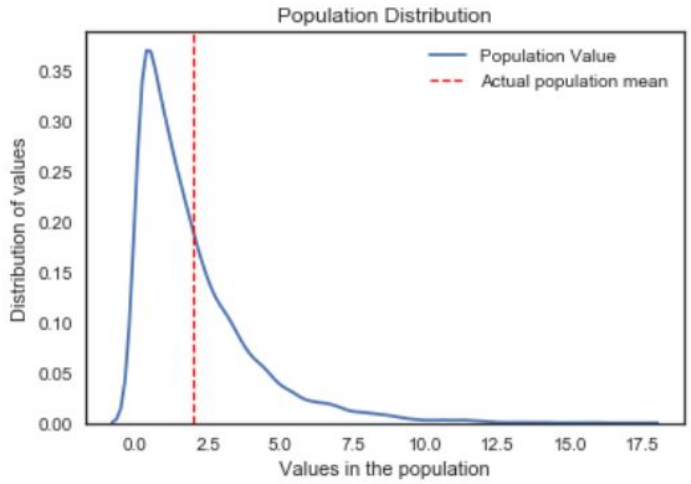
Relation between Sampling distribution and Bootstrap distributions:

Shape, Center and Spread of these two distributions are identical for distributions originated from the same population. Main difference is,

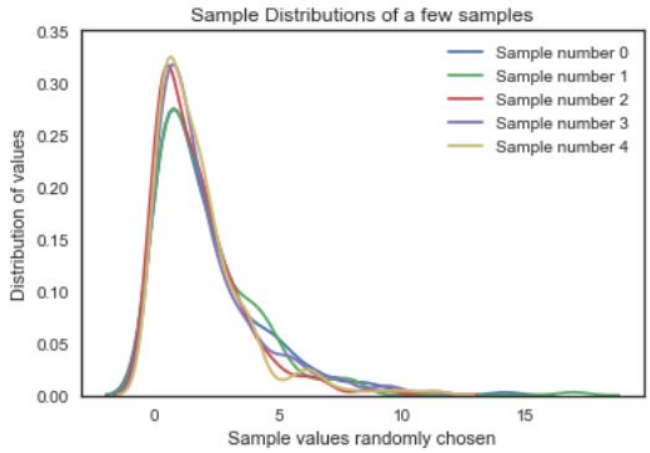
- Sampling Distribution is created by sampling with or without replacement from the population. The statistics under consideration is mean or standard deviation
- Bootstrap Distribution is created by sampling with replacement from one sample of the population. The statistics under consideration can be mean, median, proportion or standard deviation

All of these are easy to understand with an example. Note that in reality, we have access to either one or more samples from a population. Step 1 is absent in reality, instead we get to choose from part of such a huge population.

1. **Simulate a population:** Used a built-in function in python numpy.random. It is right skewed.
   Size N=10000
   Mean is = 1.9944
   Standard deviation = 2.005



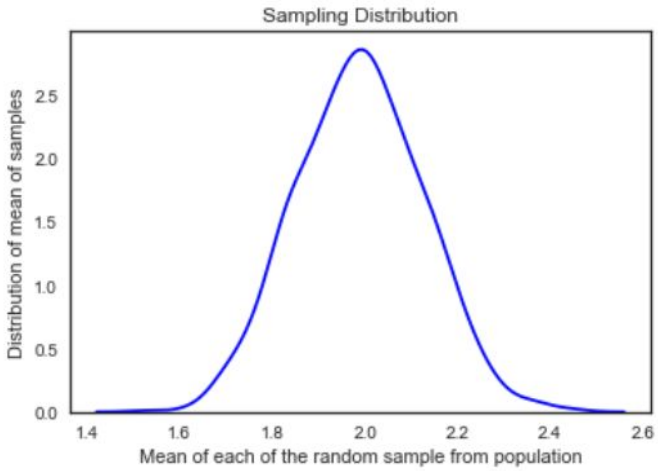
2. Plot a few Samples of the population: Next, take k=1000 samples of size n=200. First 5 of them are plotted below, just to demonstrate how they look similar to the population distribution, and to each other. One can see that they are less right skewed than population. Note that meaning of word ‘sample’ is a set of members taken from population and not an individual member from population, as normally mistaken by a beginner. Below is a set of plots showing how samples look.



Measured mean on these 5 samples:[1.81395203, 1.78721229, 1.99852289, 1.70612163, 1.91401734]
Measured standard deviation on these 5 samples: [1.65513274, 1.91820765, 1.92008107, 1.90149981, 1.89791581]

Note that there is a huge variability, but they all try to mimic the population.

3. Plotting mean(average value) of Samples: Compute mean of each of these, and plot the mean values. This is the sampling distribution of mean of the samples.



Measured mean = 1.989
Measured standard deviation = 0.13
This standard deviation is called Standard Error in Mean or (SEM)
Note that the mean of means is close to the real mean of (simulated) population, with much less variability. This is called point estimate of the population mean.
In practice, we don't know the population, hence measured mean is not equal to population mean with 100% probability. Then what is it? That is where confidence level and Confidence Interval helps. A c% confidence interval is the interval on either side of the point estimate, corresponding to c% confidence level. The area under the above curve within this confidence interval is c% of the total area under the curve. How to do calculations?

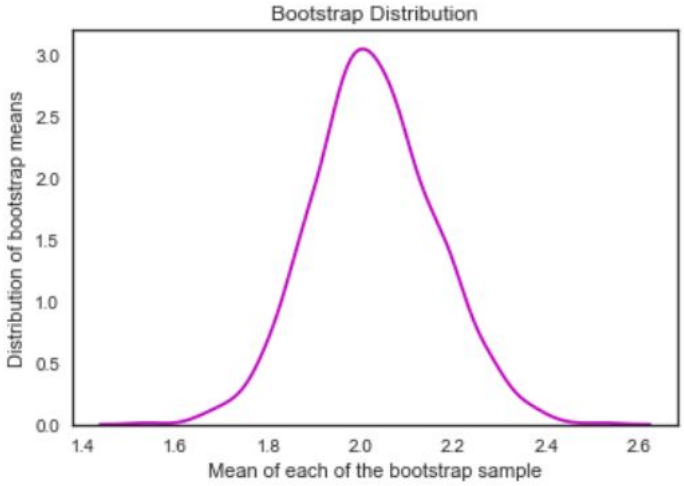
Conditions for Central Limit theorem (CLT):

1. Independence: Sample values are independent because 200 = 5%N, less than 10%N
2. Sample size 200 is large, as it is > 30, and samples look nearly normal. This can also be verified with scipy.stats package's normaltest

As the conditions are met, sample means are assumed to be normally distributed with population mean and sigma of approximately 1.9/sqrt(200) = 0.134.
Meaning the shape and parameters of the curve above closely resembles that of normal distribution.

Hence one sample z-test is to be used to evaluate a point estimate of population with a confidence interval. Normal practice is to use a 90%, 95% or 99% confidence. Higher the confidence, wider the interval.
If conditions for CLT are not met, other distributions will have to be used, most common is t-distribution with n-1 degrees of freedom. How to conduct the tests is out of scope.

4. Demonstrate Bootstrapping:
- Take one of these 1000 samples. Randomly pick one value out of 200 values at a time, but do not consider it deleted from the sample. It is again available. Do this 200 times and create one more sample of size 200. This is called one bootstrap sample (also called bootstrap replicate).
- Repeat previous step another 999 or 9999 times.
- Compute the means of each of the bootstrap samples drawn
- Plot and observe that it looks similar to the sampling distribution in this case



Measured mean = 2.0266
Measured standard deviation = 0.133

In practice, we end up doing either 3 or 4. Also bootstrapping is just one of the resampling methods. That is sufficient for a start.

Date: 01/19/2018