Exercise – Understanding the method of basic bootstrap

1. Aim of the exercise

The basic bootstrap method is a resampling technique used to estimate the sampling distribution of a statistic, such as the mean, median, or variance, by repeatedly sampling with replacement from the original dataset. This approach is especially useful when the population distribution is unknown or difficult to model analytically. The aim of this exercise is to understand how to correctly estimate the sampling distribution of the sample mean under such uncertainty. We compare two approaches: (i) bootstrap resampling from the sample, and (ii) repeated sampling from the full population. This comparison helps evaluate how well the bootstrap method approximates the true sampling distribution. While the bootstrap may not replicate the exact center (mean) of the population distribution, it often provides a reliable estimate of its variability, making it a powerful tool for statistical inference.

2. Theory

Consider a random sample $X_1, X_2, \ldots, X_n$ drawn independently and identically distributed (i.i.d.) from an unknown distribution with mean μ. The sample mean

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

serves as an estimator of the population mean μ. Since $\bar{X}_n$ is computed from random variables, it is itself a random variable and possesses a sampling distribution that reflects its variability across repeated samples of size n. In practice, the true sampling distribution of $\bar{X}_n$ is typically unobservable due to limited access to the full population. Nevertheless, understanding this distribution is essential for evaluating the reliability and precision of our estimate. The bootstrap method provides a powerful approach to approximate the sampling distribution of an estimator, such as the sample mean, by repeatedly resampling with replacement from the observed data. For a comprehensive treatment of bootstrap theory, refer to the exercise on bootstrap methods.

3. Set values for the parameters of the simulation

We begin by defining the number of simulations, which determines how many samples we will draw from the population to construct the true sampling distribution, as well as how many bootstrap samples we will generate from the initial sample. To ensure a fair comparison and avoid introducing simulation noise, we keep the number of samples the same for both the true and bootstrap distributions.

```
14 %% 3. Set values for the parameters of the simulation
15
16 % 3.1. Clear the memory
17 clear;
18
19 % 3.2. Set the number of simulations as the number of bootstrap samples
20 N_sim = 1000;
```

4. Generate the population data

Generate population data from a normal distribution with a known mean of 4 and a standard deviation of 5.

```
%% 4. Generate population data

% 4.1. Set the population size
N_obs_pop = 1000;

% 4.2. Generate population data
data_pop = random('Normal',4,5,[N_obs_pop,1]);
```

5. Draw samples from the population

To analyze the behavior of sample means, we draw repeated samples from the known population. First, we define the sample size for each draw. We then preallocate a matrix to store the samples and a vector to record their corresponding means. Using a loop, we draw samples from the population without replacement using the `datasample` function, meaning that each observation is selected at most once. Compute the mean in each sample. This process is repeated for a specified number of simulations, allowing us to examine the distribution of sample means and assess sampling variability.

```
%% 5. Draw samples from the population

% 5.1. Set the sample size
N_obs_sample = 100;

% 5.2. Preallocate matrix to store samples
data_samples_pop = NaN(N_obs_sample,N_sim);

% 5.3. Preallocate vector to store sample means
means_data_samples_pop = NaN(N_sim,1);

% 5.4. Draw samples from the population and compute their means
for i = 1:N_sim
    data_samples_pop(:,i) = datasample(data_pop,N_obs_sample, ...
        'Replace',false);
    means_data_samples_pop(i) = mean(data_samples_pop(:,i));
end
```

6. Pick an initial sample

Pick a sample among the samples already randomly drawn from the population and use it as the initial sample for bootstrap sampling.

```
%% 6. Pick an "initial" sample

% 6.1. Randomly pick one sample index
sample_index = randi(N_sim);
```

```
52
53 % 6.2. Pick one sample from the previously drawn samples
54 data_sample = data_samples_pop(:,sample_index);
```

7. Draw (bootsrap) samples from the initial sample

Perform bootstrap sampling by repeatedly drawing samples with replacement from the initial sample. For each resample, compute the mean and store it to approximate the sampling distribution and analyze the variability in sample means.

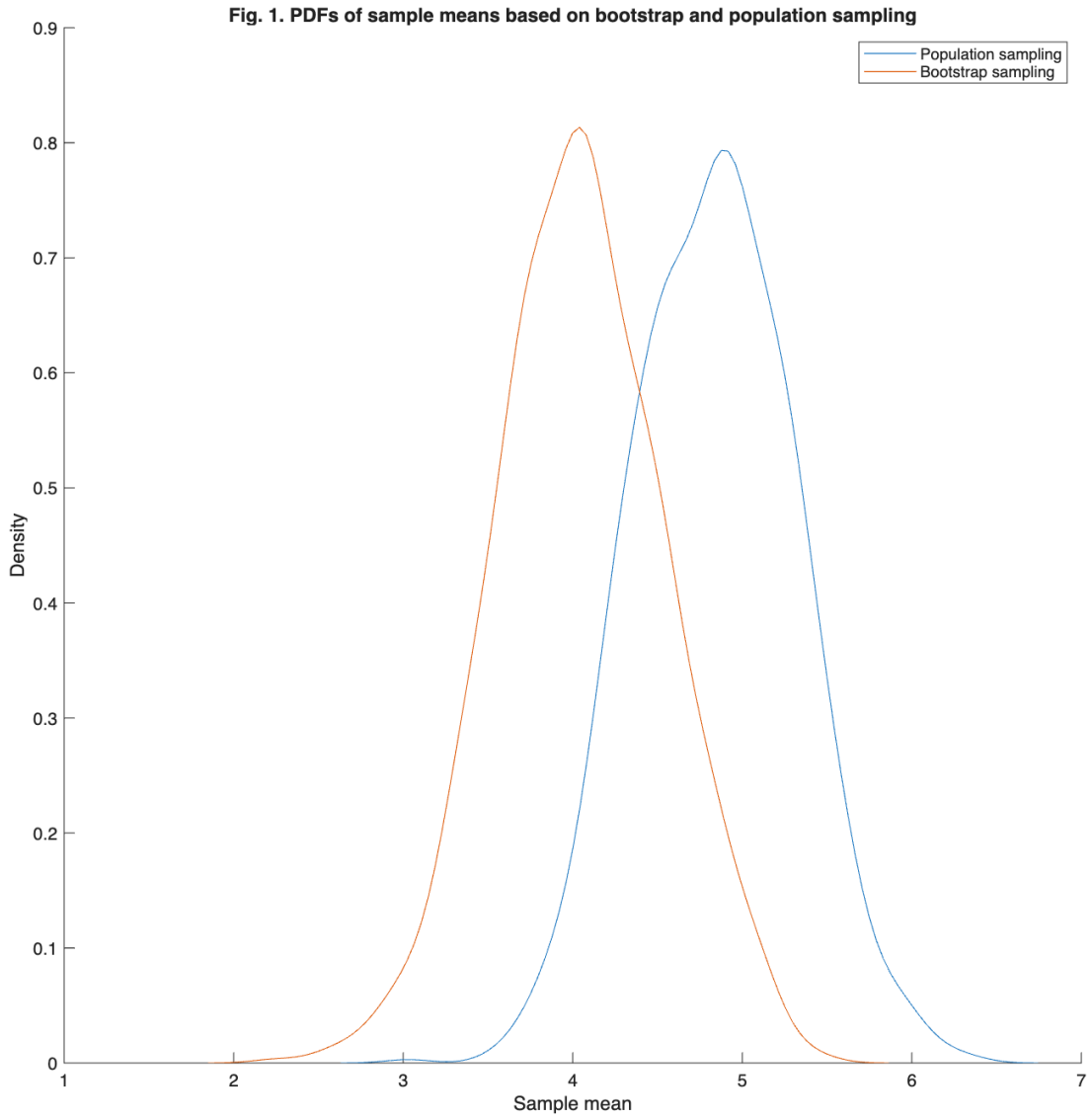
```
54 %% 7. Draw (bootsrap) samples from the initial sample
55
56 % 7.1. Preallocate vector to store (bootsrap) sample means
57 means_data_samples_boot = NaN(N_sim,1);
58
59 % 7.2. Draw samples from the initial sample and compute their means
60 time
61 for i = 1:N_sim
62     data_samples_boot = datasample(data_sample,N_obs_sample, ...
63         'Replace',true);
64     means_data_samples_boot(i) = mean(data_samples_boot);
65 end
```

8. Plot the PDFs of sample means from bootstrap and population sampling

This figure compares the probability density functions (PDFs) of sample means obtained through bootstrap sampling and direct sampling from the population. The key observation is that both distributions exhibit similar variability, as shown by the comparable spread and shape of the curves. This indicates that the bootstrap method effectively captures the uncertainty associated with sampling.

The means of the two distributions are not identical, but this difference is expected. The bootstrap distribution is centered around the mean of the original sample, whereas the population sampling distribution is centered around the true population mean. These means do not need to match for the bootstrap to be valid: its primary purpose is to approximate the variability of the sampling distribution, which it does well.

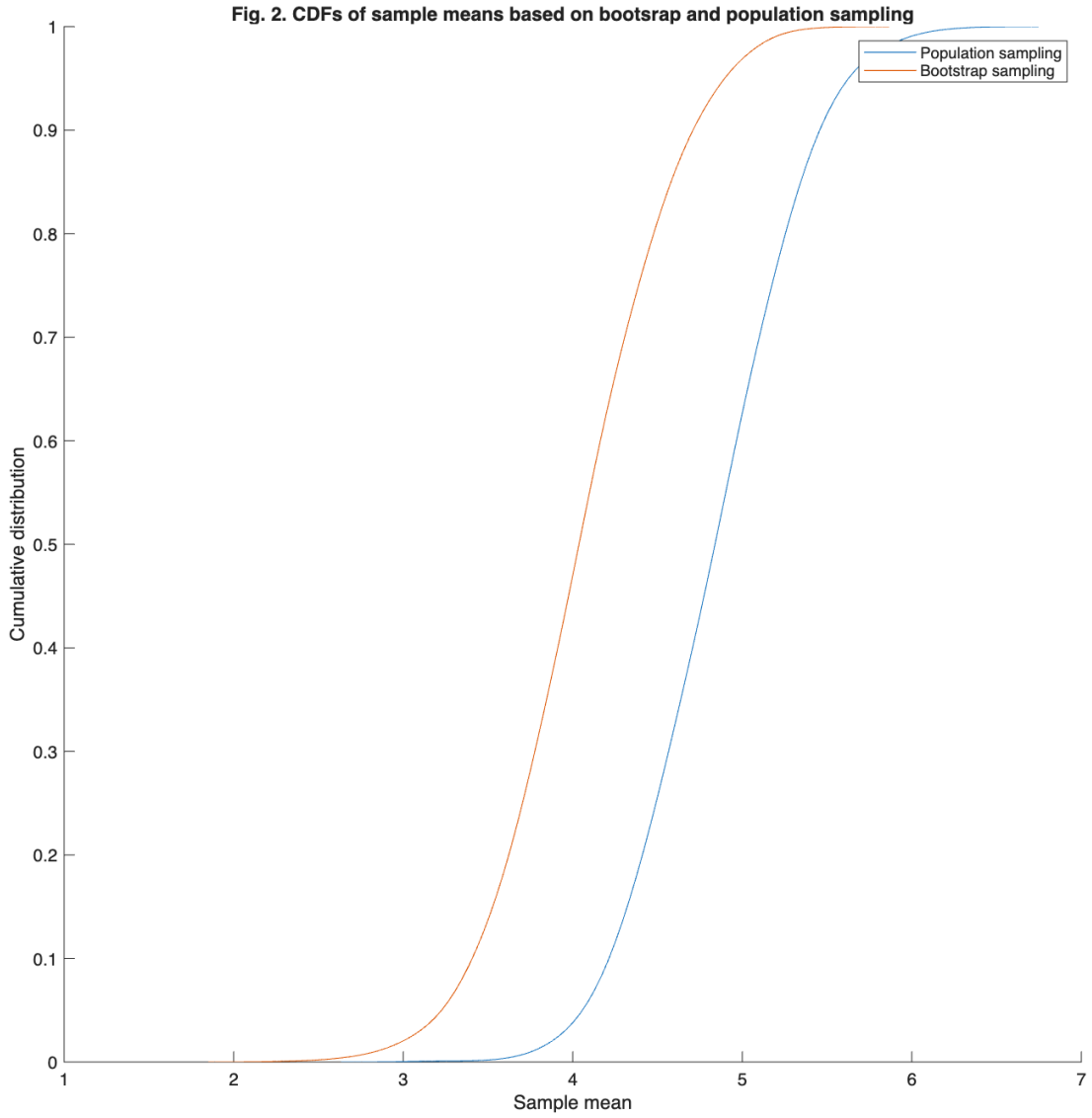
```
66 figure;
67 hold on
68 ksdensity(means_data_samples_boot,'function','pdf');
69 ksdensity(means_data_samples_pop,'function','pdf');
70 ylabel('Density');
71 xlabel('Sample mean');
72 legend('Population sampling','Bootstrap sampling');
73 title(['Fig. 1. PDFs of sample means based on bootstrap and ' ...
74     'population sampling']);
75 hold off
```



9. Plot the CDFs of sample means from bootstrap and population sampling

This figure compares the cumulative distribution functions (CDFs) of sample means from bootstrap and population sampling. The two curves show that the overall distribution shapes are similar, indicating that the bootstrap method closely approximates the cumulative behavior of sample means. Minor differences in location reflect the fact that the bootstrap is centered on the sample mean, not the true population mean. Again, this is expected and acceptable.

```
figure;
hold on
ksdensity(means_data_samples_boot,'function','cdf');
ksdensity(means_data_samples_pop,'function','cdf');
ylabel('Cumulative distribution');
xlabel('Sample mean');
legend('Population sampling','Bootstrap sampling');
title(['Fig. 2. CDFs of sample means based on bootsrap and ' ...
    'population sampling']);
hold off
```



10. Final notes

This file is prepared and copyrighted by Axel Zoons, Quinten Salomons, and Tunga Kantarcı. This file and the accompanying MATLAB file are available on GitHub and can be accessed using this link.