So far in the course, you have used the bootstrap to estimate multiple different parameters of a population such as the median and mean. You are now capable of building empirical distributions for these sample statistics. An **empirical distribution** for a sample statistic is usually obtained by repeatedly resampling and calculating the statistic for those resamples (i.e. via bootstrapping!).

Now we will introduce the **Central Limit Theorem** (CLT), which tells us more about the distribution of the sample mean: if you draw a **large** random sample **with replacement** from a population, then, regardless of the distribution of the population, the probability distribution for that sample's mean is roughly normal, centered at the population mean.

Furthermore, the *standard deviation* (spread) of the distribution of sample means is governed by a simple equation, shown below:

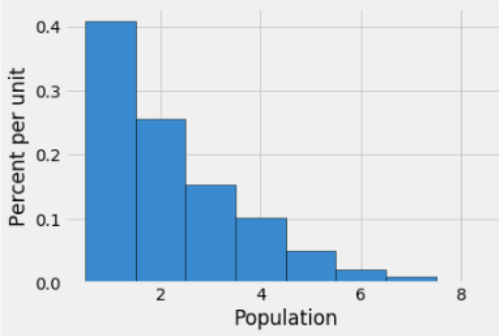
$$\text{SD of all possible sample means} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

"SD of the distribution of all sample means" is the same thing as saying "sample mean SD".

1. Sample Means

Note that in this question, the empirical distribution of the sample mean is made up from the means of **samples drawn from the population** (not from resamples from a single sample).

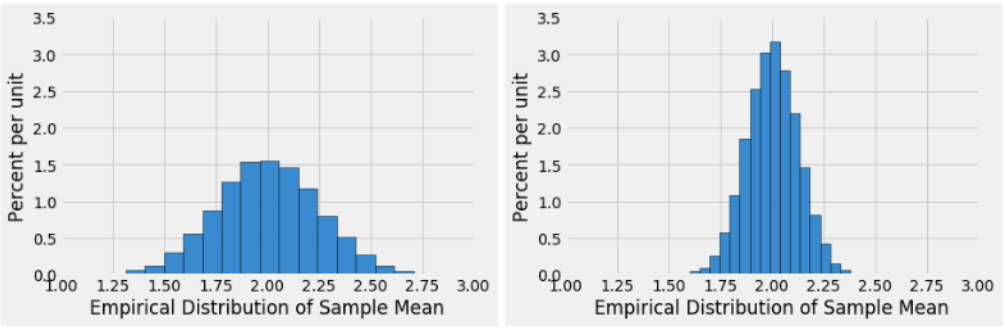
Assume that you have a certain population of interest whose histogram is below.



a. Sahand takes many large random samples **with replacement** from the population with the goal of generating an empirical distribution of the sample mean. What shape do you expect this distribution to have? Which value will it be centered around?

b. Why are we be able to use the CLT to reason about the empirical distribution of the sample mean's shape if the population data is skewed?

c. Suppose that Sahand creates two empirical distributions of sample means, with different sample sizes. Which distribution corresponds to a larger sample size? Why?



d. Based solely on the information in the histogram, what is an estimate for the standard deviation of the sample mean on the left? How did you determine this?

e. Suppose you were told that the distribution on the right was generated based on a sample size of 100 and has a standard deviation of 0.2. How big of a sample size would you need if you wanted the standard deviation of my distribution of sample means to be 10 times smaller?

2. How Tall are Berkeley Students?

You are working with Rithvik on constructing a confidence interval for the mean height of all Berkeley students. You take a random sample of 400 Berkeley students and compute the mean height of students in the sample; it is 170 cm. We also calculate the standard deviation of our sample to be 10 cm.

a. Rithvik claims that the distribution of all possible sample means is normal with SD 0.5 cm. Use this information to construct an approximate 68% confidence interval for the mean height of all Berkeley students.

b. If Rithvik hadn't told you what the SD of the sample mean was, could you estimate it from the data in the sample? If yes, how?

3. CLT with TLC

You are a superfan of the girl group TLC, and are interested in estimating the average amount of plays their songs have online. You generate an 80% confidence interval for this parameter to be [700000, 1200000] based on a random sample of 35 songs using the Central Limit Theorem. Are each of the following statements true or false?

a. Our population parameter changes depending on our sampling process.

b. The empirical distribution of any statistic we choose will be roughly Normal based on the Central Limit Theorem, but it requires our population to have a Normal distribution to begin with.

c. If we generate a 95% confidence interval using the same sample, the interval will be narrower than the original confidence interval because we are more certain of our results.

d. Using the same process to generate another 80% confidence interval, there is an 80% chance that the next confidence interval you generate will contain the true average number of plays for TLC songs.

e. 80% of TLC's songs have between 700000 and 1200000 plays.

f. The original sample mean you obtained was 950000 plays.

Helpful Note

In Q1, our large samples drawn with replacement were obtained directly from the population. As a result, our empirical distribution of the sample mean was centered around the population mean.

In contrast, in Q2, we only have access to a single sample. As a result, we use the sample's mean and standard deviation in place of the population's mean and standard deviation.

Tying this together, if you bootstrap (resample from one representative sample) and obtain the empirical distribution of the sample means, it'll be roughly centered around the original sample's mean, **not the population mean**. Though, it is helpful to note that as the sample size gets larger, your original sample's mean will likely get closer and closer to the population mean.