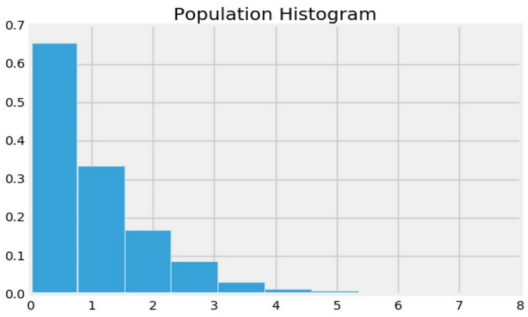
CS2065 Spring 2022
Discussion: Sample Means and Correlation (Lab 08)

So far in the course, you have studied multiple different statistics that you can calculate from a sample, including the maximum, median, and the mean. You are now capable of building *empirical distributions* of these different statistics. However, calculating the empirical distribution of the *sample mean* is unique. If you draw a large random sample **with replacement** from a population, then, regardless of the distribution of the population, the probability distribution of the sample 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:

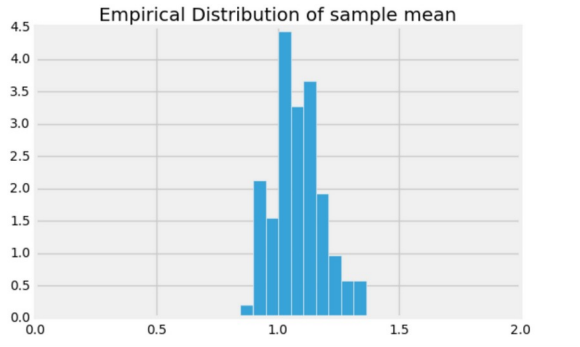
$$\text{SD of all possible sample means} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

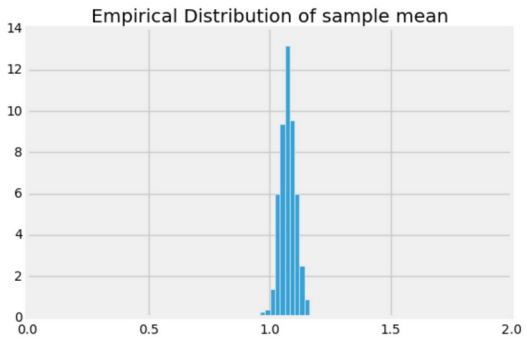
Question 1. Assume that you have a certain population of interest whose histogram is at right.



a) Caroline takes multiple 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) Suppose that Caroline creates two empirical distributions of sample means, with different sample sizes. Which distribution corresponds to a larger sample size? Why?





c) Suppose you were told that the distribution on the left has a standard deviation of 0.03 and was generated based on a sample size of 100. How big of a sample size would you need if you wanted the standard deviation of my distribution of sample means to be 0.003 instead?

Question 2. You are working with Colby on constructing a confidence interval for the mean height of all Berkeley students. Colby tells you that the empirical distribution of the mean height generated through bootstrapping a sample of size 200 is roughly normal with **mean 170 cm** and **SD 10 cm**. Use this information to construct an approximate 95% confidence interval.

Hint: If you know the empirical distribution is roughly normal, what do you know about the proportion of values that lie within a few SDs of its mean?

Correlation

An important aspect of data science is using data to make *predictions* about the future, using information that we currently possess. A question one might ask would be "Given the US GDP of every year of the previous decade, how can we predict the US GDP for next year?" In order to answer this question, we will investigate a method of using one variable to predict another by looking at the *correlation* between two variables.

Question 3. Why do we convert data to standard units?

Question 4. Write a function called `convert_su` which takes in an array of elements called `data` and returns an array of the values represented in standard units.

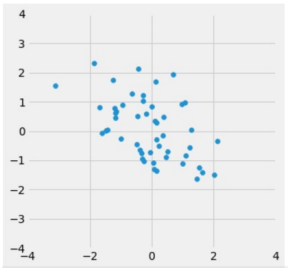
```
def convert_su(data):
```

Question 5. Now let's write a function called `correlation_coefficient` that takes in two arrays `x` and `y` of the same length, and returns the correlation coefficient between the two.
Hint: Feel free to use the function you wrote in the previous question.

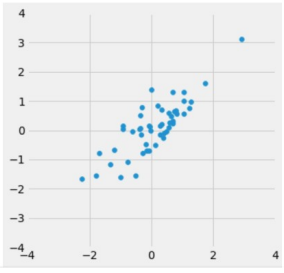
```
def correlation_coefficient(x, y):
```

Question 6. Look at the following four datasets. Rank them from least correlated to most correlated *in magnitude*.

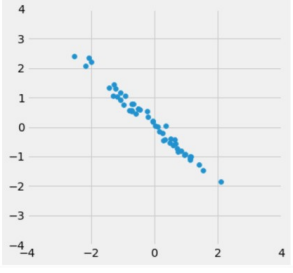
A



B



C



D