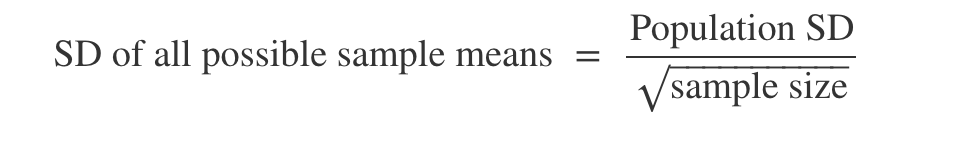
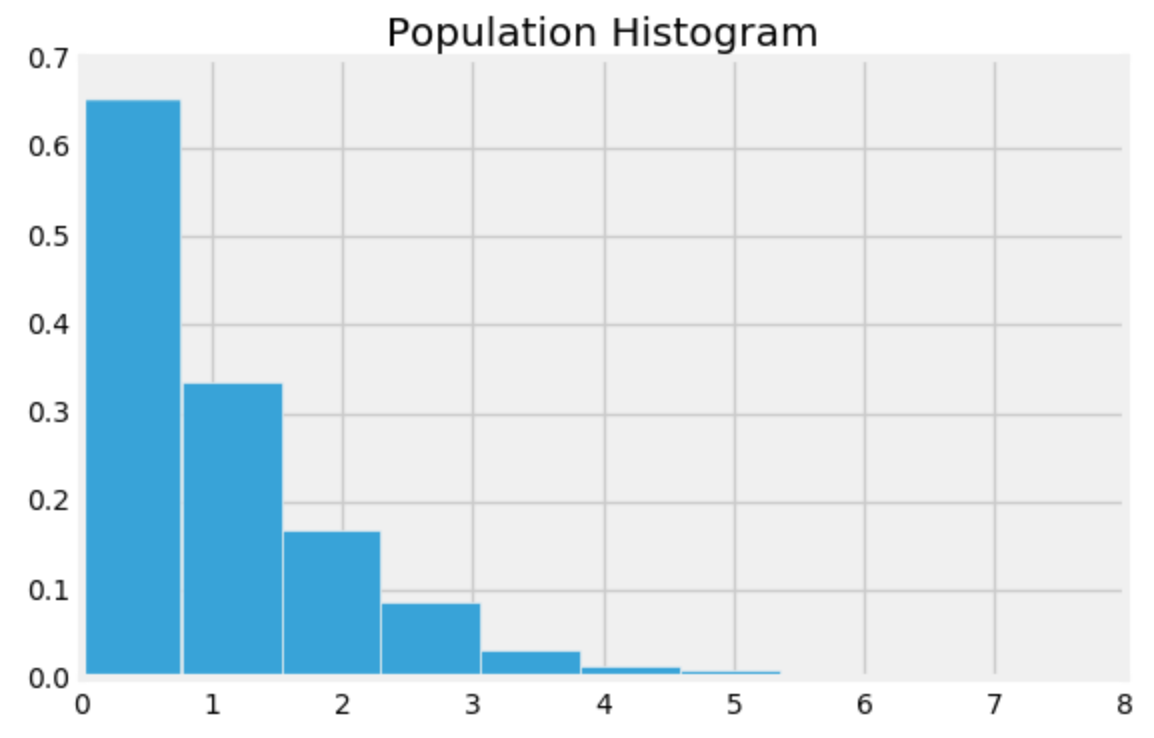
**Data 8 Spring 2020**

**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:

****

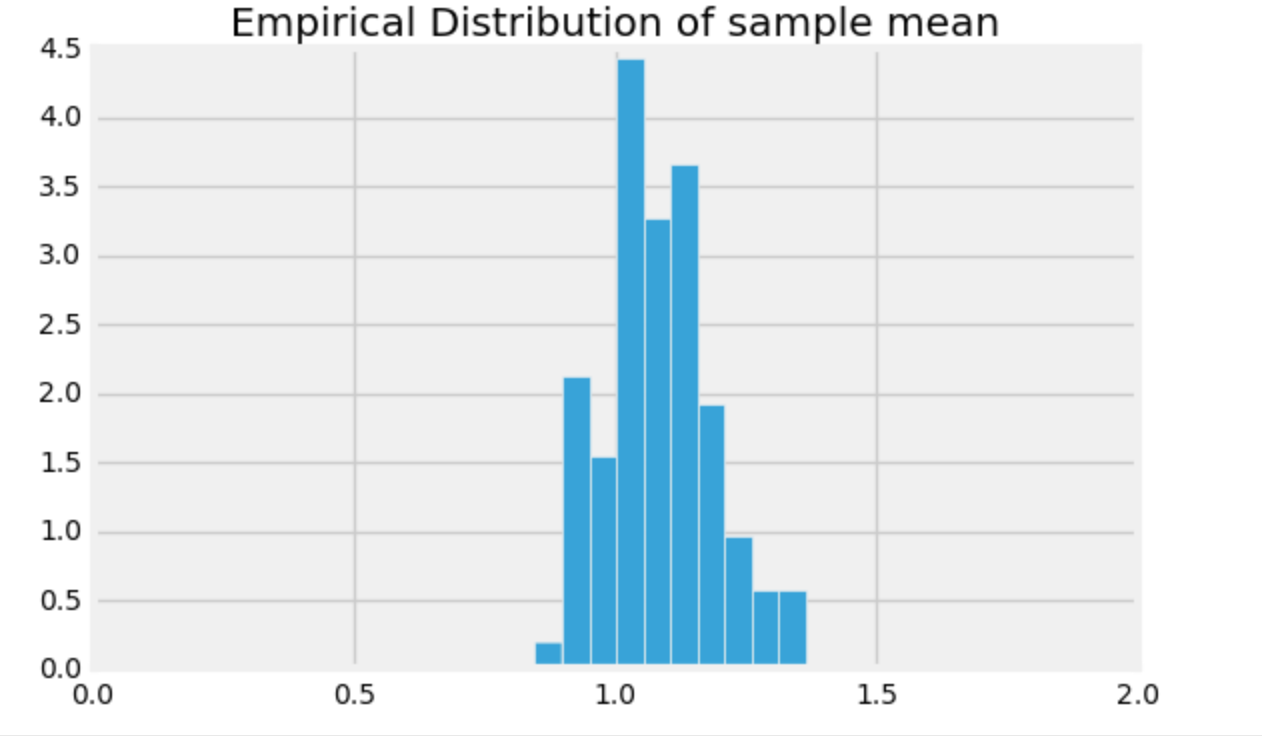
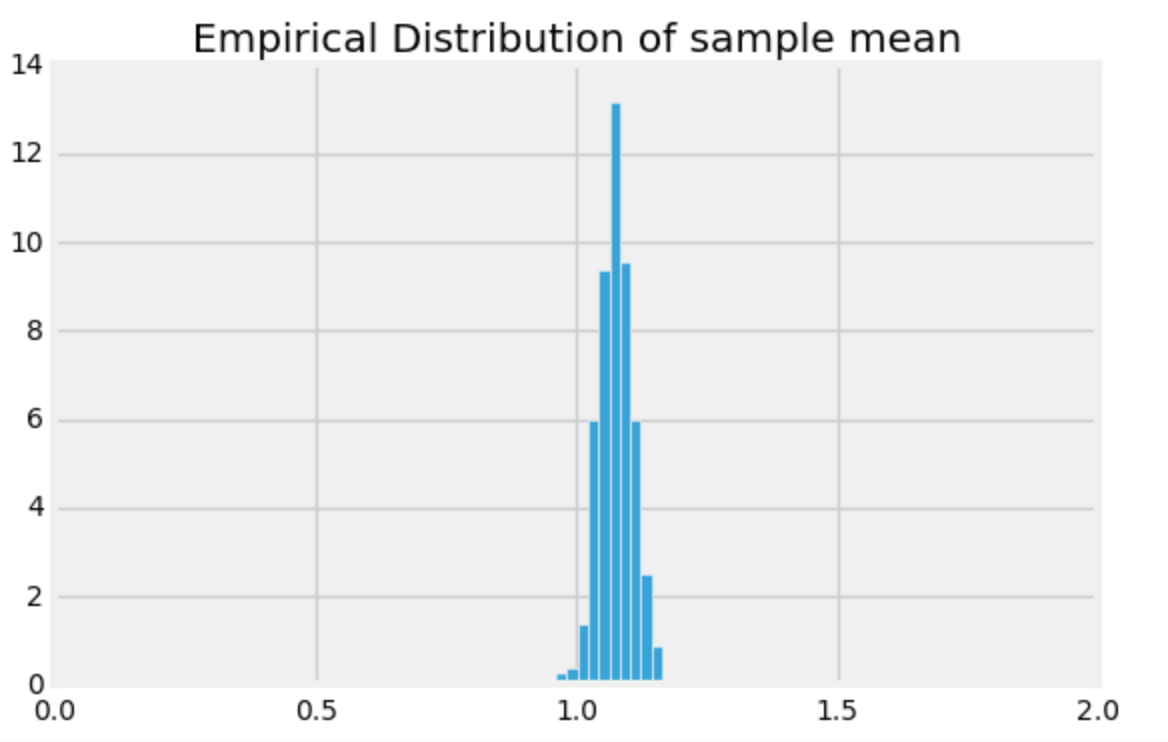


**Question 1.** Assume that you have a certain population of interest whose histogram is at right.

1. 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?

The distribution will look like a bell curve centered around the population mean, by the Central Limit Theorem.

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



Distribution to the right corresponds to a larger sample size compared to the one to the left. We can see it based on the spread of the two distributions. Smaller spread = larger sample size, as the larger sample size you take, the less variable the distribution of the sample mean becomes. You can see that increasing the sample size is increasing the denominator in calculating the SD of sample means, which decreases the standard deviation.

1. 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?

Need to have a sample size of 10,000 by the square root law.

.03 = PopSD / sqrt(100). PopSD is equal to .3

.003 = .3 / sqrt(newSampleSize)

newSampleSize = 10,000 = 100^2

To change by a factor of 10, we have to multiply by 10^2 which is 100!

**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?*

We know that the distribution is roughly normal and hence we know that 95% of the area under the normal curve (AKA 95% of the data) is contained within 2 SDs from the mean. So our confidence interval range will be (170 - 2\*10, 170 + 2\*10) = (150, 190).

**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?

We convert data to standard units in order to compare it to other data with different units and distributions: for example, if we wanted to compare the weights of cars, which are thousands of pounds, to the maximum speed of cars, which are in dozens of miles or kilometers per hour.

**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):

data\_mean = np.mean(data)

data\_sd = np.std(data)

return (data-data\_mean)/data\_sd

**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):

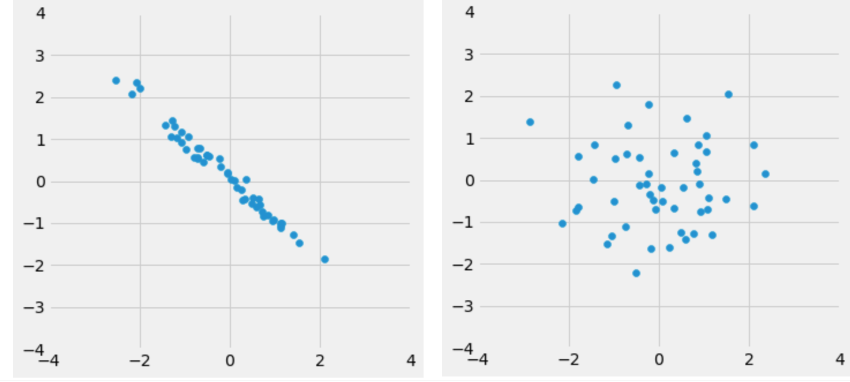
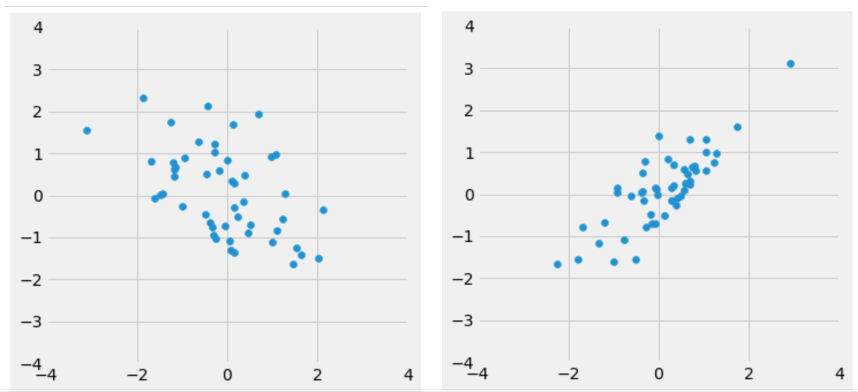
x\_su = convert\_su(x)

y\_su = convert\_su(y)

return np.mean(x\_su \* y\_su)

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

A B C D



D, A, B, C

D has almost no visible negative or positive trend as it is basically a blob, so it’s correlation is near 0

A has a negative correlation, but the points are not very tightly clustered around a straight line, so the strength of its correlation is smaller

B has a positive correlation, and the points are more tightly clustered around a positive sloping line, so its correlation strength is stronger than A

Lastly, C has a negative correlation, and the points are almost all along a straight line, so it has a very strong correlation in magnitude