CS 031

Project 2 Lab: Sample Means, Standard Units

So far in the course, you have used the bootstrap to estimate multiple different parameters of a population such as the maximum, median, and mean. You are now capable of building *empirical distributions* for these sample statistics. An empirical distribution for a sample statistic is obtained by repeatedly resampling and calculating the statistic for those resamples. However, there is special theory, namely the **Central Limit Theorem**, that tells us 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 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:

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

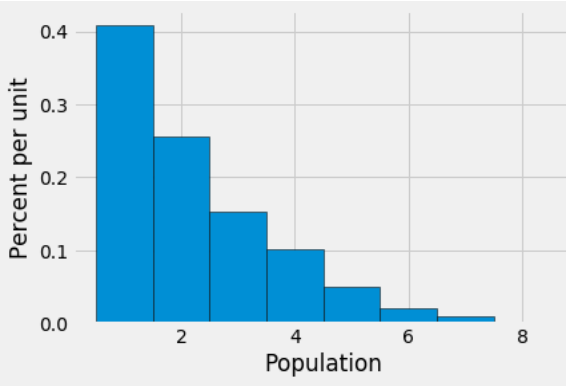
"*SD of the distribution of all sample means" is the same thing as saying "sample mean SD*".

Question 1. Sample Means: Assume that you have a certain population of interest whose histogram is to the right.

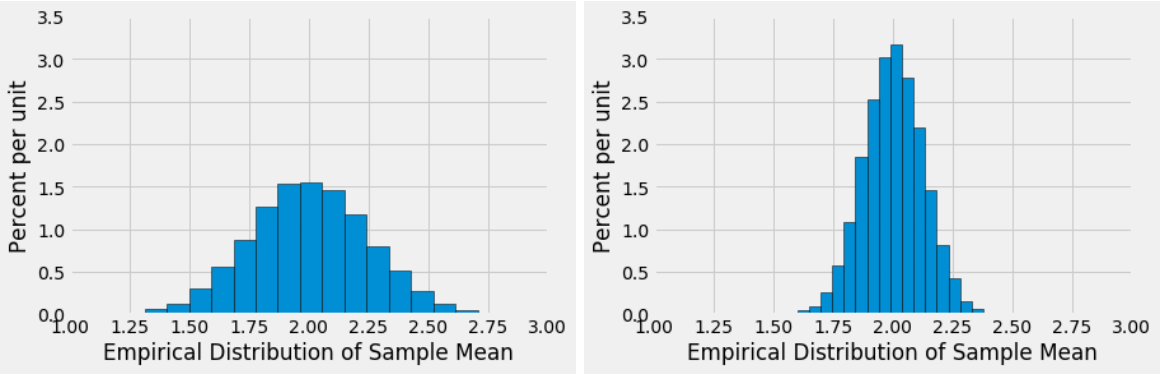
a) Aarushi takes many large random samples **with replacement** from the population with the goal of generating an empirical distribution of the sample mean. What shape do you expect this distribution to have? Which value will it be centered around?

I expect this distribution to be Gaussian, ie. it will have a bell curve shape.

It will be centered around the mean of the population, which would be a little over 1 but less than 2.



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



The second distribution corresponds to a larger sample size because the distribution is narrower and taller, therefore it is more accurate.

c) Suppose you were told that the distribution on the left has a standard deviation of 0.3 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 0.03 instead?

sample size = (SD range of confidence interval * SD of the population / width of the confidence interval)^2

sample_size = (4 * 0.03 / width of confidence interval)^2

sample_size is 100 x 100 (original sample size)

sample_size = 10000

Question 2. Confidence Intervals: You are working with Oscar 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) Oscar claims that the distribution of all possible sample means is normal with SD 0.5 cm. Use this information to construct an approximate 95% confidence interval for the mean height of all Berkeley students.

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

mean ± 2 SD = 170±20 cm

b) If Oscar 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?

Yes, one could estimate the SD of the sample mean with the formula SD_of_all_possible_sample_means = population_sd/sqrt(sample_size)

SD_of_all_possible_means = 10/sqrt(400) = 10/20 = 0.5 cm.

c) Does your answer from part (b) agree with what Oscar claims in part (a)?

Yes, my answer in part (b) agrees with what Oscar claims in part (a). The SD of all possible means is 0.5 cm.

Question 3: Standard Units and Correlation

a) When calculating the correlation coefficient, why do we convert data to standard units?

We use standard units to ensure that both axes are drawn in the same scale and so that we can compute other values that rely on standard units, such as r, the coefficient of correlation.

b) 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):
    sd = np.std(data)
    mean = np.mean(data)
    return (data-mean)/sd
```