Data 8 Spring 2022

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.

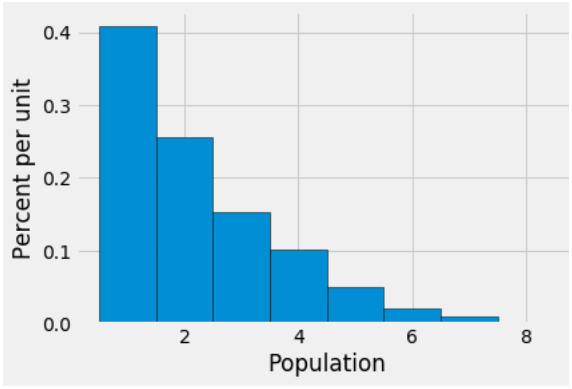
Furthermore, the *standard deviation* (spread) of the distribution of sample means is governed by a simple equation, shown below:

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

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

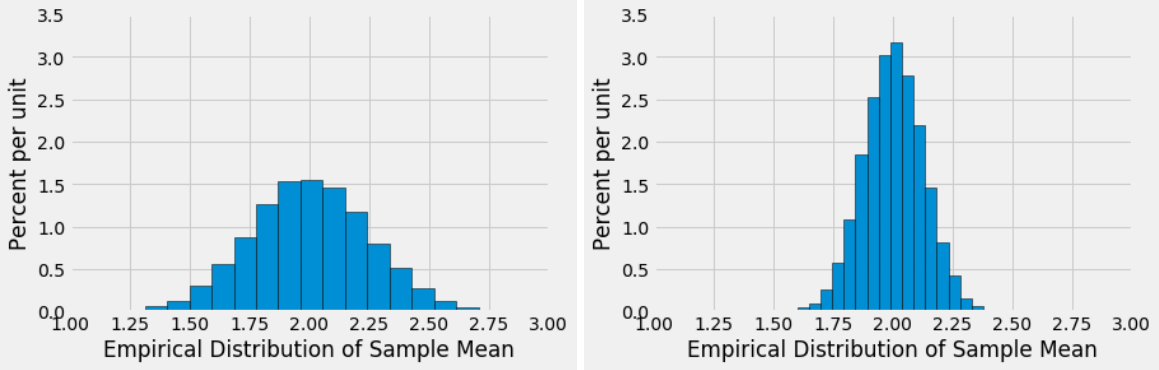
Question 1. Sample Means: Assume that you have a certain population of interest whose histogram is to the right.

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?



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

b) Suppose that Aarushi 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.

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?

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

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

0.03 = 3 / sqrt(newSampleSize)

newSampleSize = 10,000 = 100^2

To divide the SD of the sample means by a factor of 10, we need to multiply the sample size by 10^2, which is 100!

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?

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*0.5, 170 + 2*0.5) = (169, 171).

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?

We know that the sample size is 400 and that the standard deviation of our sample was 10 cm. Since the SD of sample means is equal to pop sd / sqrt(sample size), we can substitute the sample sd (i.e. 10 cm) in place of the pop sd in this formula (the sample is a large, representative sample from the population) and calculate the value!

SD of all possible sample means = 10 cm / sqrt (400) = 10 / 20 = 0.5

c) Does your answer from part (b) agree with what Oscar claims in part (a)?
Yes, Oscar claims that the SD of all possible sample means is 0.5 cm, which is what we calculated in part (b).

Question 3: Standard Units and Correlation

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

We convert data to standard units in order to compare it with 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.

Moreover, we have the following nice properties of *r* when using standard units:

1. *r* is a pure number with no units (because of standardization)
2. *r* is unaffected by changing the units on either axis (because of standardization)

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
```