Details: The Sampling Distribution of the Sample Mean

We might not have time to discuss this in class…or maybe we rushed through it too quickly for you to follow! Here are the details about the sampling distribution of the sample mean.

First of all, we need to get our heads straight about where $\bar{x}$ comes from. The sample mean is a random variable, defined as $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Note that each of those *x*-values (the x_i in the formula) is its own random variable. Thus, the sample mean is the sum of *n* (hopefully independent) identical random variables. Perhaps you should write it like this:

$$\bar{x} = \underbrace{\frac{1}{n}x + \frac{1}{n}x + \frac{1}{n}x + \ldots + \frac{1}{n}x}_{n \text{ of these}}.$$

Measuring the Center

From our study of random variables, we (should) know that $\mu_{X+Y} = \mu_X + \mu_Y$: the mean of the sum of random variables is the sum of the means. Also, $\mu_{kX} = k\mu_X$: the mean of a multiple of a random variable is just that multiple of the mean. Putting those together we get

$$\mu_{\bar{x}} = \mu_{\frac{1}{n}x} + \mu_{\frac{1}{n}x} + \mu_{\frac{1}{n}x} + \ldots + \mu_{\frac{1}{n}x} = \frac{1}{n}\mu_x + \frac{1}{n}\mu_x + \frac{1}{n}\mu_x + \ldots + \frac{1}{n}\mu_x$$

Since μ_x is a number, adding up *n* of them is $n \cdot \mu_x$ (this is not true for random variables!). Thus, the work above becomes

$$\mu_{\bar{x}} = \frac{1}{n}\mu_x + \frac{1}{n}\mu_x + \frac{1}{n}\mu_x + \ldots + \frac{1}{n}\mu_x = n \cdot \left(\frac{1}{n}\mu_x\right) = \mu_x$$

Measuring the Spread

Recall what we know about the *variance* of combined variables: $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$ and $\sigma^2_{kX} = k^2\sigma^2_X$. Putting these together, we get

$$\sigma^2_{\bar{x}} = \sigma^2_{\frac{1}{n}x} + \sigma^2_{\frac{1}{n}x} + \sigma^2_{\frac{1}{n}x} + \ldots + \sigma^2_{\frac{1}{n}x} = \frac{1}{n^2}\sigma^2_x + \frac{1}{n^2}\sigma^2_x + \frac{1}{n^2}\sigma^2_x + \ldots + \frac{1}{n^2}\sigma^2_x$$

Since σ^2_x is just a number, the sum of *n* instances of σ^2_x can be reduced to $n \cdot \sigma^2_x$. Thus, we get

$$\sigma^2_{\bar{x}} = \frac{1}{n^2}\sigma^2_x + \frac{1}{n^2}\sigma^2_x + \frac{1}{n^2}\sigma^2_x + \ldots + \frac{1}{n^2}\sigma^2_x = n\left(\frac{1}{n^2}\sigma^2_x\right) = \frac{1}{n}\sigma^2_x$$

This is variance; we want standard deviation. Take the square root!

$$\sigma^2_{\bar{x}} = \frac{1}{n}\sigma^2_x \Rightarrow \sigma_{\bar{x}} = \sqrt{\frac{1}{n}\sigma^2_x} = \sqrt{\frac{1}{n}}\sigma_x = \frac{1}{\sqrt{n}}\sigma_x = \frac{\sigma_x}{\sqrt{n}}$$

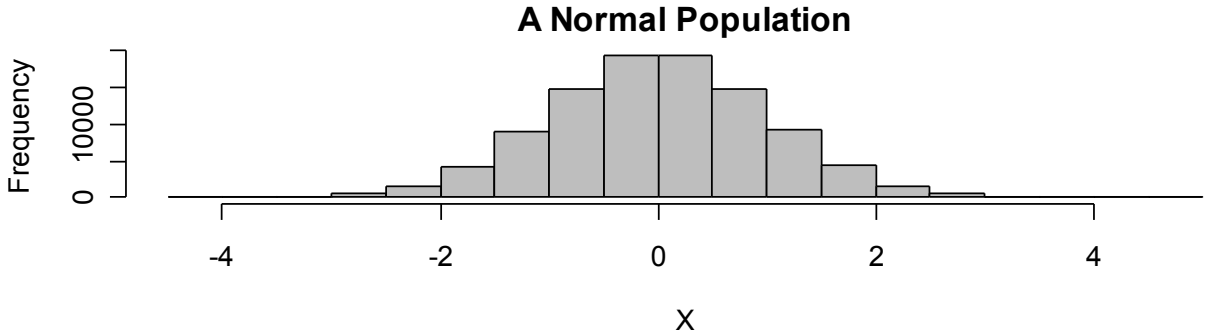
Finding the Shape of the Sampling Distribution

I'm going to take a simulation approach to the shape. Thus, we'll get lots of pictures!

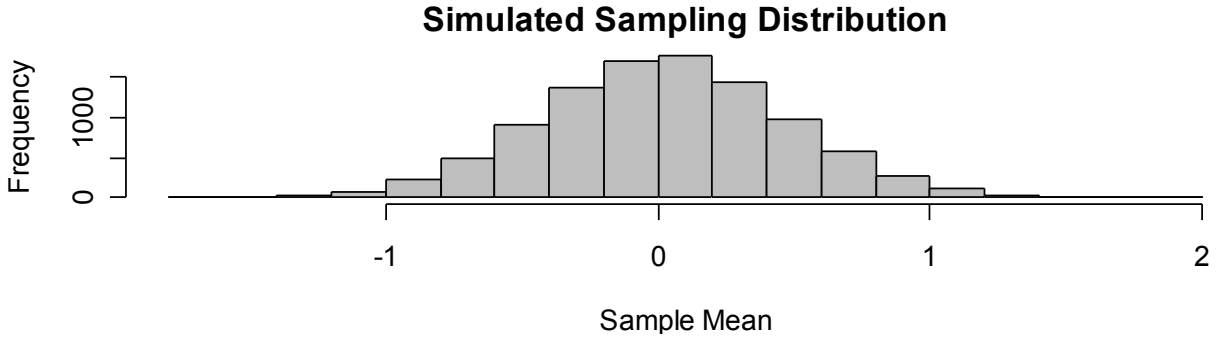
The idea is to create a population, then take lots (!) of samples with varying sample sizes to simulate the sampling distribution(s). What we'll look at is how the sample size and the shape of the population distribution affect the shape of the sampling distribution.

Case 1: Normal Population

Let's see how big a sample we need if the population is normal. First, a picture of my population distribution:



Now, let's take 10,000 samples with sample size 5. Here's the simulated sampling distribution:

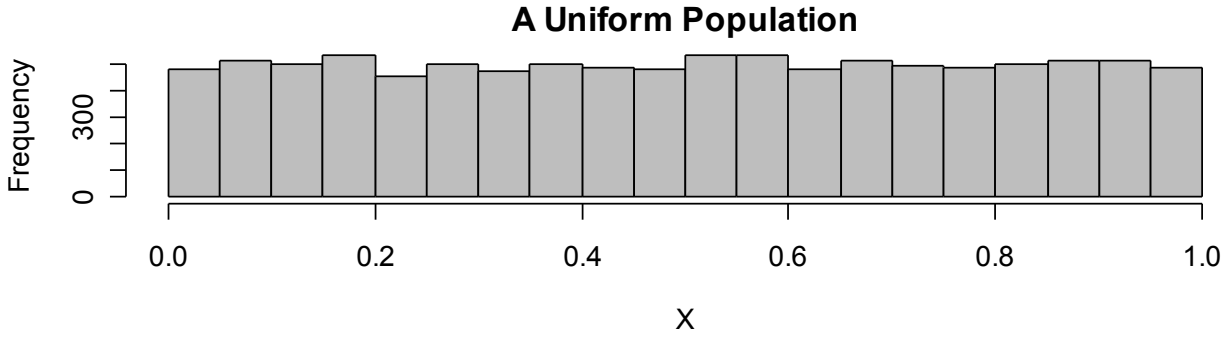


A sample size of 5 is awfully small—it looks like a normal distribution in the population gives a normal population in the sampling distribution for nearly every sample size.

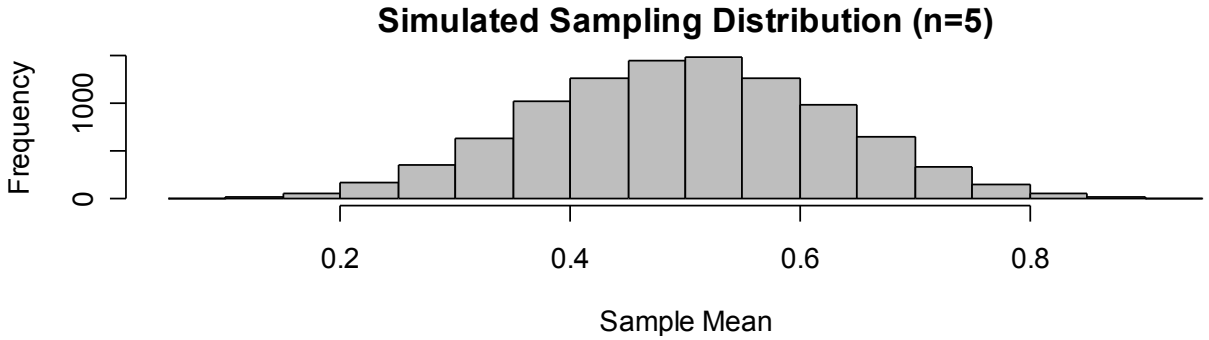
(I don't *really* have to draw the pictures for sample sizes of 2 and 3, do I?)

Case 2: Symmetric Population

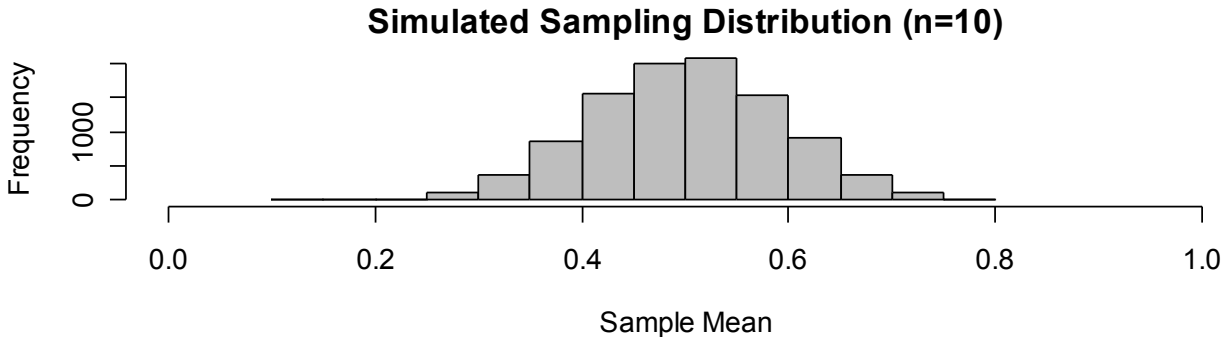
We'll repeat the process—this time with a uniform distribution.



Clearly symmetric. Now, a myriad (10,000) of samples of size 5:



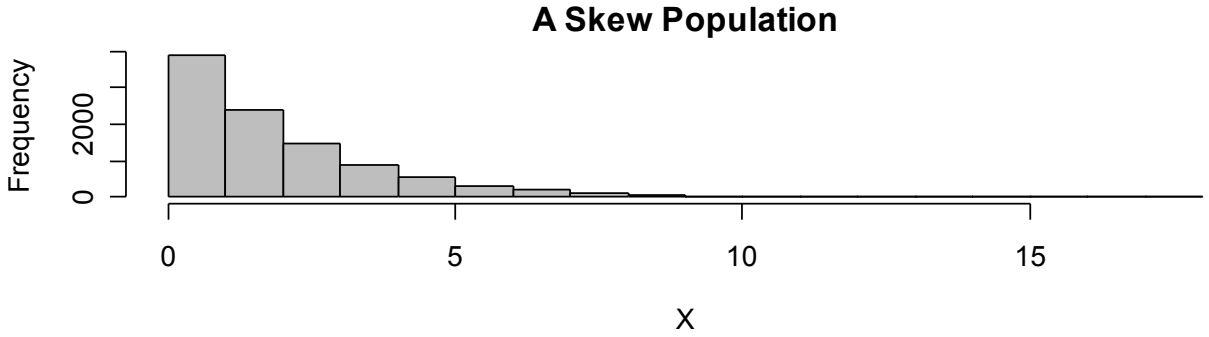
Hmmm. That's pretty normal—not as normal as the first case. Let's do it again with a sample size of 10:



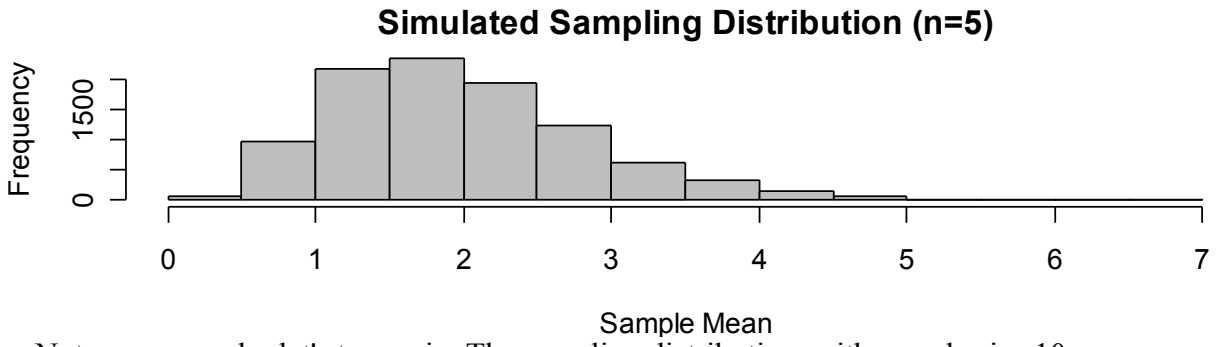
That's better. Note that a symmetric population provides an approximately normal sampling distribution for small sample sizes.

Case 3: Non-symmetric Population

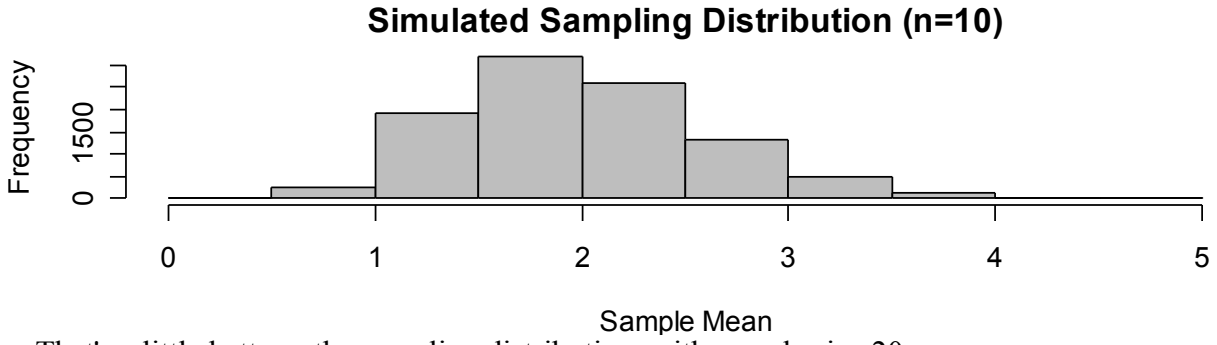
Now for a non-symmetric population.



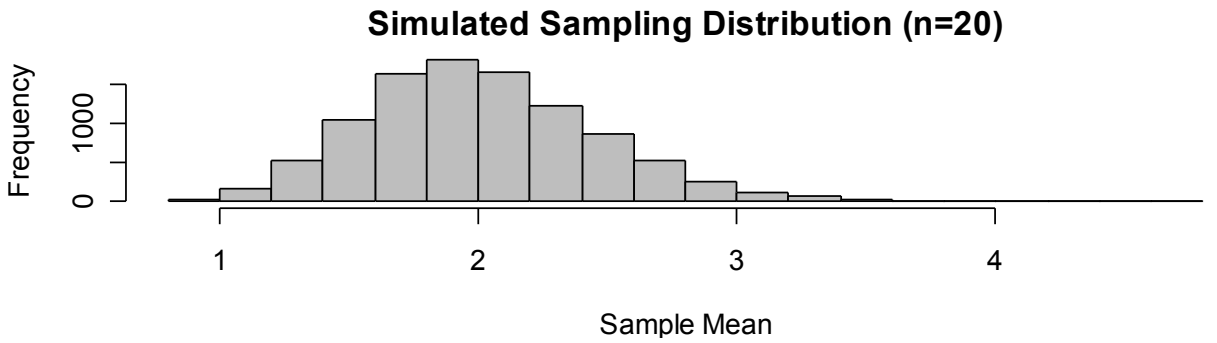
The sampling distribution, with sample size 5:



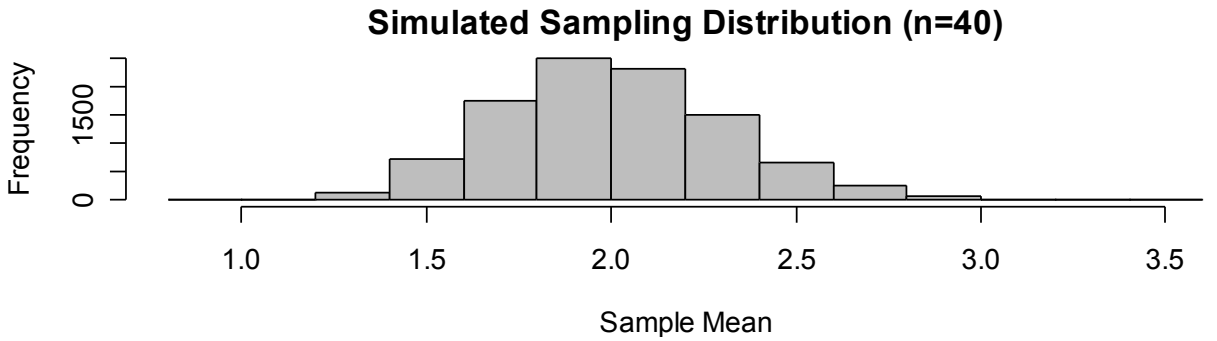
Not very normal…let's try again. The sampling distribution, with sample size 10:



That's a little better…the sampling distribution, with sample size 20:



Closer…the sampling distribution, with sample size 40:



Clearly getting better. The trend is obvious: the greater the sample size, the closer the sampling distribution gets to a normal distribution (can you say CLT?).

For a terribly skewed distribution, it looks like a sample size over 40 is needed to get an approximately normal sampling distribution.