Activity 16

We are going to use the ride share data again for this activity. We happen to have the entire census of ride shares (ie: we know the population). This is incredibly RARE in real life.

Sampling distribution and sample size

In general, how should a random sampling distribution relate to the population distribution?

The distribution for a random sample of observations from a population should be “representative” of the population distribution. That is, the random sample distribution should reflect or be similar to the population distribution.

Suppose we had a choice of taking a sample of size 10 or 300. Which one would you choose? Why?

Should pick a sample size of 300 because more data/information will provide us with a better picture of what is going on in the population (we can be more certain).

Let’s explore what happens to the sampling distribution of $\bar{x}$ as we change our sample size. We will use samples of size 5, 30, and 100. The number of repetitions will be 10,000. Sketch the distribution of each of the following and report the means and standard deviations.

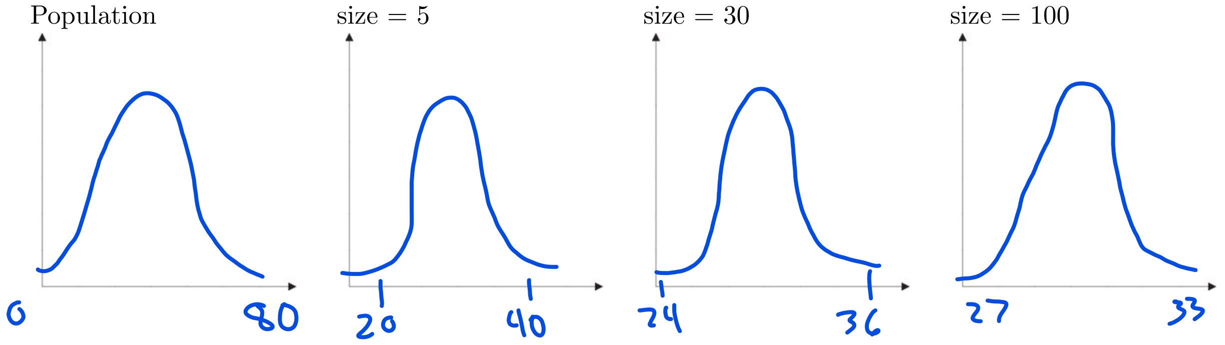
Duration

samples are random so your answers will be slightly different

Provide a description for the distribution of ‘duration‘ population.

pay attention to the x-axis!

The population is unimodal, symmetric, centered around 30.11 and spread of 9.88 in terms of standard deviation



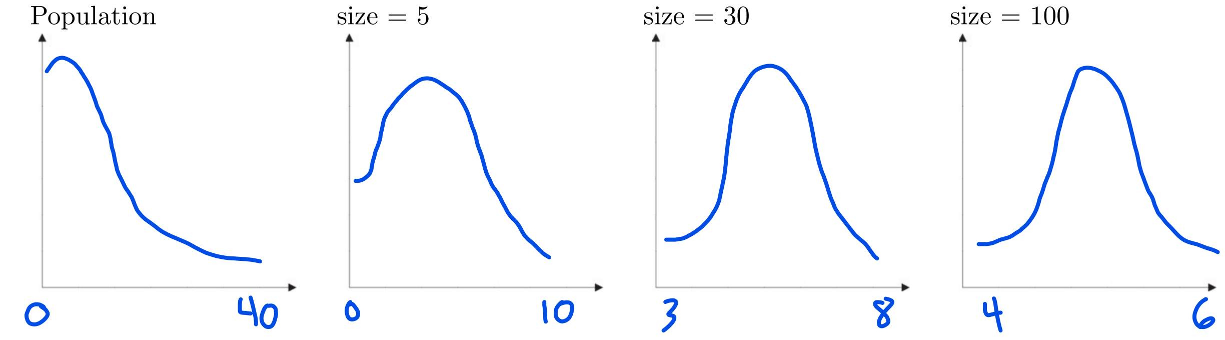
Population | size = 5 | size = 30 | size = 100

Population	size = 5	size = 30	size = 100
μ : 30.11	$\bar{x}$: 30.09	$\bar{x}$: 30.11	$\bar{x}$: 30.12
σ : 9.88	s : 4.41	s : 1.81	s : 0.98

Wait time

Provide a description for the distribution of ‘wait_time‘ population.

The population is unimodal, right skewed, centered around 19.5 and spread of 4.56



Population | size = 5 | size = 30 | size = 100

Population	size = 5	size = 30	size = 100
μ : 4.51	$\bar{x}$: 4.5	$\bar{x}$: 4.51	$\bar{x}$: 4.51
σ : 3.02	s : 1.34	s : 0.55	s : 0.3

Central Limit Theorem (CLT)

The **CLT** tells us that for a sufficient sample size, the distribution of the sample means will be approximately normally distributed even if the population distribution is not normal! More specifically, for a variable X with mean μ_x and standard deviation σ_x we have $\bar{x} \sim N\left(\mu_{\bar{x}} = \mu_x, SE(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}\right)$

Sampling Mean Check

What do you observe for the sampling distribution of the mean for ‘wait_time‘ as the sample size gets larger? What about for ‘duration‘? Does the CLT hold true?

The distribution centers around the population parameter!
Yes the CLT holds no matter the shape of the original distribution for a sample mean.

Standard Error Check

Using the σ’s calculated and the standard error formula from the CLT (see Table 9.6 as well), calculate the theoretical standard errors and compare them to the simulated standard errors (s) for the three sampling distributions for each variable. Are the theoretical and simulated values close?

Duration:

9.88/sqrt(5) = 4.42
9.88/sqrt(30) = 1.80
9.88/sqrt(100) = 0.988

Wait time

4.51/sqrt(5) = 2.02
4.51/sqrt(30) = 0.82
4.51/sqrt(100) = 0.451

The theoretical are very close to what we observed above!

Big Picture

Explain how the sampling distribution is related to a single sample.

A single sample is represented by a single data point in the sampling distribution.
For example, a single sample produces one sample mean which is one data point/observation in the sampling distribution (of means).

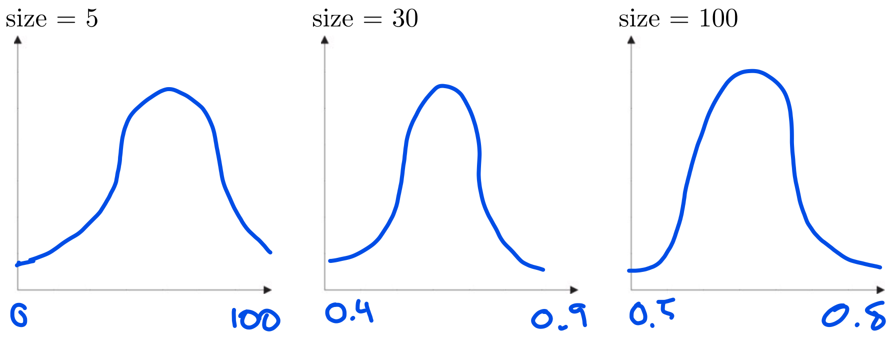
Explain how the sampling distribution is related to the population distribution.

The sampling distribution provides a way to evaluate an estimator (estimation procedure) for a population parameter (a numerical property of the population distribution). For example, consider the population mean, , which a measure of center for the population distribution. The sample mean, x¯, is an estimator for the population mean, . The sampling distribution allows us to determine is x¯ is unbiased and measure how precise it is. If the mean of the sampling distribution is equal to the population mean, then it is unbiased (on target). The standard error, standard deviation of the sampling distribution, allows us to determine how close/precise the estimator is to its mean/target.

Wait time under 5 minutes

The central limit theorem also applies to proportions. What is the population proportion (p) for ‘wait_time_under5‘?

0.686



size = 5

size = 30

size = 100

unimodal, symmetric

normally distributed!

$\hat{p}$:	0.69	$\hat{p}$:	0.69	$\hat{p}$:	0.69
se :	0.21	se :	0.08	se :	0.05

Standard Error Check

Using the population proportion and the standard error formula from Table 9.6, calculate the theoretical standard errors and compare them to the simulated standard errors (se) for each sampling distribution.

sqrt(0.6864*(1-0.6864)/5) = 0.207
sqrt(0.6864*(1-0.6864)/30) = 0.085
sqrt(0.6864*(1-0.6864)/100) = 0.046

The theoretical standard errors are all close to the observed sampling distribution standard errors.