This project continues our investigation into sampling distributions through simulation.

Recall, the main setup of this class: we wish to understand a population. Since this is complicated, we may settle for understanding a parameter, such as the mean, that summarizes the population. To do so, we take a sample (preferably random for reasons discussed in class) and from this we summarize the sample with a statistic. This statistic allows us to *infer* things about the parameter.

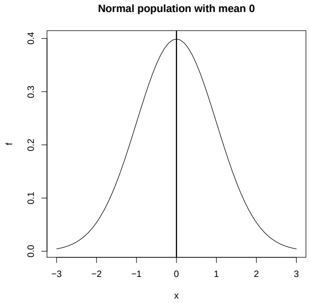

Figure 1: A population and its parameter (the mean μ)

In class we have the following three facts:

center The population mean is the same as the mean of the sample average, $\bar{X}$:

$$\mu_X = \mu_{\bar{X}}$$

Spread The spread for the sample average is less than the spread of the population, but related through the square root of the sample size:

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

Shape As long as the shape of the population is not completely crazy, *as n gets large* the shape of the sampling distribution of $\bar{X}$ becomes normally distributed.

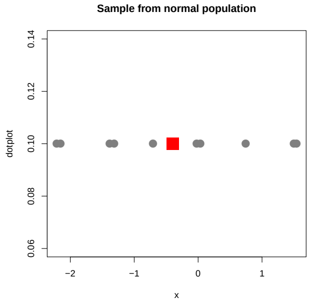

Figure 2: A sample from the normal poulation with mean $\mu = 0$ and a summary statistic $\bar{x}$ of the sample. Can the red box tell us something about the parameter μ?

The above are true for the sampling distribution of $\bar{X}$. They need not be the case for other statistics based on a sample. This project will look at a) verifying the above and b) investigating what happens for other statistics. As before we will use a helper to perform the simulations. Download the helper by typing (or copying and pasting) in this command exactly:

```
> source("http://www.math.csi.cuny.edu/verzani/tmp/simex.R")
```

To summarize its functionality:

No. of samples, sample size The number of samples controls how many different simulations are run to produce the sampling distributions through simulation. Even if just a few are shown, this many are used to estimate the density plot or boxplot.

Population This allows you to select one of several populations. You can edit the parameters if desired.

Statistic The allows you to specify the statitistic whose sampling distribution is investigated

Plot population If `TRUE` a layered graphic showing the population, 10 samples from the population, the 10 realizations of the statistic and a boxplot of the full number of realizations of the statistics is shown. Otherwise, a density estimate of the realizations of the statistic are shown.

The mean and normal populations

When the population is normal and the statistic is the mean, things are different and the same. Huh? Let's see. Change the population to the `rnorm` line and the statistic to `mean`.

Question 0.1. Run a few simulations. Explain how the diagram supports the formula

$$\mu_X = \mu_{\bar{X}}$$

Question 0.2. Select `Plot population` so that the population is drawn. Look at the relationship between the population and the sampling distribution (the boxplot) as you change the sample size from 2 through 128. What can you say about the spread as indicated by the boxplot?

Question 0.3. Now, deselect the `Plot population` option. Does the *shape* of the sampling distribution change if you change n from 2 through 128?

Does shape depend on statistic? (For normal populations)

The fact that the variability gets smaller as the sample size gets larger can prevent us from comparing across sample sizes. One way to compare different sample sizes is to standardize their values – that is we compare z-scores. As we don't generally know the value σ, we may subsitute s and see where that leaves us. That statistic is available in the line `(mean - 0)/(sd/sqrt(n))`. Select that.

We ask: *Were* we to use the true standard deviation, we would have a normal sampling distribution. Is this the case now?

Question 0.4. Look at the shape of the sampling distribution for this statistic when $n = 2, 4, 8, 16, 32, 64$. (Use `Plot population: FALSE`.)

Does the sampling distribution look normal for all values of n?

Question 0.5. When n is small, the estimate s for σ can be quite off, if the value is too small, then the resulting statistic is larger than is "should" be. This makes larger values than expected. Explain how you saw this in you previous explorations.

The mean and non-normal populations

Let's return to the case where the sampling distribution is the mean. We saw that normal populations yield normally distributed sampling distributions – although they have different spreads.

What about changing the population, will we still get a normal population? Will it depend on sample size?

Question 0.6.

1. Change the population to the skewed one and the statistic to the `mean`. For $n = 8$ and $n = 128$ is there a difference in the shape of the sampling distributions? Describe the difference using the language we've used in class (skew, symmetry, modes, ...).
2. Now change the population to the long tailed one and ask the same question.
3. Indicate in this table when the sampling distribution of the mean appears normal:

Population	small n	large n
normal		
skewed		
long-tailed		

Using one statistic over another

There are reasons we have more than one measure for the center and spread. We've talked about sensitivity to the presence of outliers as one. Lets compare their spreads.

For example, let's compare the spread of the sampling distribution for the standard deviation and the IQR for different populations. We can use a sample size of 32 and for fun, we will look at the population too with `Plot population: TRUE`.

Question 0.7.

1. For the normal population option, which is larger the spread for `sd` or the spread for `IQR` or are they about the same?
2. For the skewed population option, which is larger the spread for `sd` or the spread for `IQR` or are they about the same?
3. For the long-tailed population option, which is larger the spread for `sd` or the spread for `IQR` or are they about the same?

Some statistics are just not very good, despite being intuitive to use. For example, the range summarizes spread as does the IQR. Why would we prefer the IQR? Something about outliers.

Let's compare the range and the IQR by looking at the spread of their sampling distributions.

Question 0.8.

1. For the normal population option, which is larger the spread for the `range` or the spread for `IQR` or are they about the same?
2. For the skewed population option, which is larger the spread for the `range` or the spread for `IQR` or are they about the same?
3. For the long-tailed population option, which is larger the spread for the `range` or the spread for `IQR` or are they about the same?