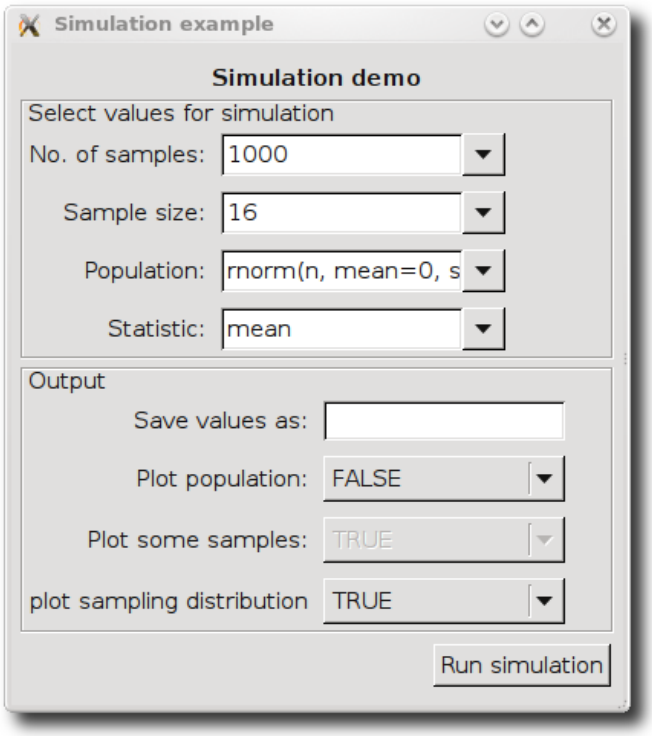

Figure 1: Window that allows one to simulate values of a statistic after specifying: the population, the size of the sample, and the statistic.

This project looks at simulating sampling distributions.

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.

Sounds easy, but there's a catch. By introducing randomness into the problem, we also introduce *sampling variation*. It is this variation that leads us to out key problem: is any difference between the observed value (the statistic) and its expected value (the parameter if unbiased) due to a real difference or due merely to sampling variability. To filter this out, the sampling distribution – the description of the variability of the statistic due to the randomness of the sample – is necessary. For certain statistics, we will see later this term that the sampling distribution is known under certain assumptions.

But for right now, we don't know this. All we know is a) for proportion the sampling distribution for large samples is more or less normally distributed and b) in general larger samples produce smaller sampling variability (at a rate constant divided by the square root of n).

To learn more we will *simulate* many different samples (m of them) each of a certain sample size n. From this we summarize each sample with a statistic. This leaves us with several realizations of the statistic, allowing us to visualize its variability. Not only that, unlike real life we know the population parameters so we can look at biasness.

So how we going to do this? Easily, but first we need to download a file. Run this command. Type it exactly. (You can do this within pmg, or from R's command line.)

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

This will open a window (Figure 1) allowing us to perform simulations.

0.1 The sample proportion

As mentioned, if we use a sample proportion for our statistic to summarize a poll, say, then when the sample size is large the sampling distribution will be approximately normally distributed.

To visualize this, do the following: change the sample size to 512; change the population to `rbinom` (which essentially tosses a coin with probability 1/2, although you can edit this to your liking); and change the statistic to `prop`. Then click the `Run simulation` button.

If you don't adjust the default plots, you should see an odd graphic. It will have the following visible:

The population shown with a density plot. In this case, the population is full of 0's and 1's, with nothing in between, so the density plot has two peaks. The relative height of the peaks indicates how likely a 0 or 1 will be. In this case, the probabilities are identical.

As well, 10 samples are shown. Again, in this case the samples are all either 0 or 1, so we don't see much, just some dots under the density curves. Each sample is shown on a different value of the y axix. Finally, each sample produces a statistic, which are visualized with red squares. These show the variability of 10 samples. The sampling distribution is indicated by a boxplot at the bottom of the graphic. With your graph, you should see that the variability is quite small (the IQR is about 0.03). This is due to the large sample size (n = 512).

To get a better view of the sampling distribution, change the plot population value to `FALSE` and rerun a simulation. Then a density plot of several samples is produced. If you don't want a density plot, then you can a) save the values to a variable and then b) plot that variable.

1. We understand that variability is affected by sample size. Verify this, by observing the change in variability when the sample size value is changed from 512 to 256, 128 and 64. Does variability increase or decrease as n decreases?

2. If n is big enough, then the sampling distribution of the proportion is *basically normal*. What if n is *not* big enough? Let's check:

   (a) For a sample size of 2, how many modes does the sampling distribution show. (Make `Plot population: FALSE`.)

   (b) Repeat with a sample size of 8

   (c) If the sample size if 64, is the shape now "bell-shaped?"

   (d) The actual sample size needed to get a "bell-shaped" curve depends on the parameter p. The default is $p = 1/2$, but this can be changed. Find `prob=1/2` and make it `prob=1/25` by tying in a 5. Ask the last question with sample size of 64. Is it different? How? (We will discuss a rule of thumb that np and $n(1-p)$ should be 10 or more.)

0.2 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`.

1. 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?

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

0.3 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.

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

1. 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?

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

2. 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.

0.4 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?

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 is normal:

Population	small n	large n
normal		
skewed		
long-tailed		

0.5 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`.

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.

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?