Instructor Materials for Law of Large Numbers Simulation

Main Ideas

- The sample mean approaches the true population mean as the sample size approaches infinity. This property is called the Law of Large Numbers.
- Fundamentally, the Law of Large Numbers works because of the way the mean finds the "balance point" of the sample observations. The mean seeks the center.
- The underlying population distribution affects the speed of convergence (though the Law of Large Numbers still holds, no matter what the population looks like).

Suggestions for Building Intuition

- Begin by selecting a population distribution. I encourage beginning with the uniform distribution. Remind students what the distribution shape implies about the values of the observations (i.e., in a uniform distribution, points are just as likely to be found on the extremes as close to the center).
- Select a sample of any size you like. Find the sample mean. Is the sample mean exactly equal to the population mean? Is it close?
- Try samples of various sizes. What can you say about the relationship between the sample size and the difference between the sample mean and the population mean?
- Now spend some time on very small samples. Take a sample of size one. What is the mean of that sample? Here emphasize that the mean is just the observation itself, of course. Now take a sample of size two. Notice that the mean finds the center, which is likely to be closer to the population mean. Keep increasing the sample size by one observation at a time. This step should help students understand *why* the Law of Large Numbers is true.
- Now allow students to run the final step, the simulation, which draws samples of increasing size. Ask students: what happens as the sample size gets larger? How large does the sample size have to be for the sample mean to be essentially indistinguishable from the population mean?
- Now ask students to try all the previous experiments with different population distributions. What happens to the speed of convergence? Are there any differences in the small samples?

Suggested Handout for Lab Activity or Self-Guided Work

Point your favorite browser to go.middlebury.edu/econsims

From the Menu, select "Law of Large Numbers." Select a type of population distribution ("Pick a Population Distribution"). The page will display the true population mean. Below the figure depicting the population distribution, you can begin selecting samples from the population, calculating the mean of each sample. Here we can examine the difference between the population mean and the sample mean.

Q1: Is the mean of the sample exactly equal to the mean of the population? What happens to the difference between the population mean and the sample mean as you increase the sample size?

The simulation at the bottom of the page collects many samples of size n.

Q2: What does the simulation show you? What happens to the difference between the sample mean and the population mean as the sample size gets large? Do we have a name for this property?

Q3: Now select another type of population distribution from the left panel and rerun Steps 1-5. Do you notice any differences if the population distribution is not normal? Is not symmetric?

Instructor Materials for Central Limit Theorem Simulation

Main Ideas

- The sample mean is a point estimator of the true population mean. The sample mean is a random variable, and like any random variable, it has a distribution, called a *sampling distribution*. We could theoretically take many samples, calculate the mean of each, and describe the mean, variance, distribution shape, etc. of those means.
- The distribution of the sample mean will be normally distributed if the underlying population from which you draw the sample observations is normally distributed.
- The distribution of the sample mean will be approximately normally distributed *for any population distribution* as long as the sample size is large enough. For some population distributions, samples of size *n=2* have sampling distributions that already appear approximately normal.
- Because the standard deviation of the sampling distribution is equal to $\sigma/\sqrt{n}$, as the sample size increases, the distribution of the sample mean becomes narrower. (In fact, as $n \to \infty$, $\sigma/\sqrt{n} \to 0$ and hence $\bar{x} \to \mu$, which gives us the Law of Large Numbers). The bell shape implies that observations of the sample mean are more likely to be located within a closer vicinity of the population mean.
- Taking more samples improves the resolution of the figure, i.e. it will begin to appear "more normal."

Suggestions for Building Intuition

- Ask the students to draw a sample of a reasonable size (*n=20* or *n=30*) and plot the sample mean. Spend a few moments reminding them of the definition of the sample mean, pointing out that we are plotting the mean of the points highlighted on the left graph. Plot a few more points, then many more. Ask the students: what will this picture look like if we take *many* samples and plot *many* means? Remind them that the graph shows us a *distribution*. Then, finally, show the students the app simulations clicking many times and collecting many samples, and run the simulation. You should see a beautiful bell-shaped plot.
- Probably the most powerful single exercise is the following: ask students to plot the means of many samples of size *n=1*. This should just give back the population distribution (check for understanding). Then, ask the students to plot the means of samples of size *n=2*. Already, the distribution of the sample means will begin to appear bell-shaped (Setting the Mystery Distribution as the population distribution produces especially stark results!) Allow them to continue to increase the sample size by 1 each iteration. This really drives home the central intuition: because the sample mean finds the center of the sample points in any sample, the sample means will be more likely to be located close to the population mean than far from the population mean. The bell-shaped curve implies that points are more likely to be located close to the center than far from the center.
- Ask the students to experiment with sample sizes. Ask: what happens to the shape of the distribution of the sample mean as the sample size increases? Why does this make sense, mathematically? Explain the role of the "normalize" button, which rescales the distribution of the sample mean to a distribution with mean 0 and standard deviation 1.
- Ask the students to repeat the above steps with different population distribution shapes. Ask: what happens to the shape of the distribution? What is the relationship among the sample size, distribution shape, and the shape of the distribution of the sample mean?

Suggested Handout for Lab Activity or Self-Guided Work

Point your favorite browser to go.middlebury.edu/econsims

From the Menu, select "Central Limit Theorem." Select a population distribution, and then follow the steps suggested. This simulation will plot a distribution of sample means.

Q1: *What is the shape of the resulting histogram of sample means? How does it compare to the shape of the population distribution? Try different population distribution shapes.*

Q2: What happens to the standard deviation of the distribution of the sample mean as the sample size increases?

Additional Suggested Questions for Discussion or Problem Sets

Q1: What is a sampling distribution? How can the sample mean have a distribution?

Q2: What is the shape of the distribution of the sample mean from samples drawn from a population that is normally distributed?

Q3: What is the shape of the distribution of the sample mean from any population distribution? Do we have a name for this property?

Q4: What happens to the shape of the sampling distribution of the sample mean as the sample size increases?

Instructor Materials for Joint Distributions

Main Ideas

- This module makes a transition from the focus on single random variables and their sampling distributions in the LLN and CLT modules to thinking about the joint distribution of two random variables.
- Covariance measures the extent to which two random variables move together, or "co-move" or "co-vary." Understanding covariance is an essential building block in learning about regression.
- Correlation is an alternative measure of how two variables co-move that is scaled by the standard deviation of the variables and ranges between -1 and 1.
- The mean and variance of each individual random variable factor into the calculation of covariance (or correlation), but pairs of random variables with same underlying distributions could have very different correlations.

Suggestions for Building Intuition

- Ask students to generate and observe the two histograms and scatterplot for the pre-loaded parameter values. Ask them to click on a point in the scatter plot and observe that a point lights up on each of the respective histograms. Ask them to observe the axes on the scatter plot and think about how each point on the scatter plot represents the combined information about a given parent-child pair.
- Ask students to change the mean and standard deviation of the individual random variables – maybe pick opposite extremes for parents and children. Generate the plots and ask them to think about how their choices affect the placement of the peak of each histogram (mean) and dispersion of the histogram (standard deviation). Ask them to explain how that same information is conveyed on the scatter plot.
- Now keeping the underlying means and standard deviations fixed, ask students to play with the correlation and regenerate the plots each time. Try a large positive correlation (close to or equal to 1), large negative. How about zero?
- Ask students to pay attention to the covariance that is displayed below the correlation box as they change each of the parameters. Note that the underlying means and standard deviations will affect the relevant ranges of covariance for this activity (which is also a learning point).

Suggested Handout for Lab Activity or Self-Guided Work

Point your favorite browser to go.middlebury.edu/econsims

Click "Start" and select the "Joint Distributions" module from the Menu. In this module, we will start thinking about the relationship between two random variables. Relationships between two or more variables is the crux of most empirical economic analysis – what is the relationship between price and quantity demanded? How does one's level of education affect one's earnings? How does the corporate tax rate affect R&D expenditure?

The example we will use in this module is the relationship between parents' and children's height. This may not see like an economics example. But, it is the classic example used by Sir Francis Galton who is the originator of the statistical terminology of "Regression."[1] And, some economists remain very interested in height as it pertains to economic development, or health and nutritional investments in children and how they affect economic outcomes like earnings in adulthood.

The mean and standard deviation of parents' and children's height are prepopulated on the page. Click the generate button.

Q1: Describe what is displayed in the first two plots.

Q2. Change the mean and standard deviation for both parents and children (make them quite different from each other). Explain in what ways the figure for parents is now different from that for their children. Can you come up with a story for the means and standard deviations you picked (economists have lots of ideas about this that relate to economic development, nutritional improvements, migration etc.)

Q3. Take a look at the third graph and click on any point. What does each point on this "scatter plot" represent?

Q4. Change the correlation and regenerate the plots. Describe the scatter plot under the following three scenarios: 1) a large correlation covariance (close to +1), 2) a large negative correlation (close to -1), and 3) co correlation variance=0. Tell me a story in words for at least one of these scenarios.

Q5. What is the equation for covariance? What is the equation for correlation?

[1] Galton, Francis. "Regression towards mediocrity in hereditary stature." *The Journal of the Anthropological Institute of Great Britain and Ireland* 15 (1886): 246-263.

Instructor Materials for Least Squares Module

Main Ideas

- The Ordinary Least Squares method estimates the intercept and slope of a line that "best fits" the observed data by minimizing the sum of the squared distances between the points and the line.
- We can construct a measure of how far a given line is from each point as the vertical distance between the point and the line. Squaring that distance allows us to account for both positive and negative distances without them canceling each other out.
- With a small number of points, we can work our way to best fit iteratively (i.e. guess and check), but thank goodness for the minimizing power of calculus.

Suggestions for Building Intuition

- After they have generated a random scatter of data points, ask students to guess which y-intercept and slope they think will plot a line that best fits (matches, describes) this cloud of points. Have them enter their guesses and plot the associated line by clicking the "Plot your guess button." Ask them to observe and discuss the squares appear for each point and the sum of squares calculation.
- Ask students to change their slope by at least one. What happens to the squares and the sum of squares? Similarly for the intercept.
- Now ask students to "play" with the slope and intercept until they think they have the best possible line. Then check how close they were by clicking the "Reveal the Least Squares Line" button.

Suggested Handout for Lab Activity or Self-Guided Work

Point your favorite browser to go.middlebury.edu/econsims

Click "Start" and select the "Least Squares" module from the Menu. In this module,

Q1:. What is the equation of a line?

Q2: Generate a set of points. What straight line do you think would fit most closely through this cloud (scatter) of points? Is the slope of that line positive or negative? Where would the line hit the y-axis?

Q3: Enter your guesses and plot the line. What is the sum of squares?

Q4: Try a much bigger or smaller slope. What happens to the squares on the screen? the sum of squares?

Q5: Play around and try to make the sum of squares as small as possible (remember you can switch both the slope and intercept). Write down your final guess.

*Q6: Click the "*Reveal the Least Squares Line*" button. How close were you? What is the minimum sum of squared distances for this data set?*

Q7: In class, we talked about a mathematical term for the distance between a point and the regression line. What did we call it?

Instructor Materials for Omitted Variable Bias Module

Main Ideas

- Omitted Variable Bias (OVB) is the bias in a regression estimator that arises when there is a variable (*V*) which is not included in the regression that is correlated with the regressor (*X*) and is a determinant of the outcome (*Y*).
- In the regression model $Y_i = \beta_0 + \beta_1 X_i + \delta V_i + \epsilon_i$ where *V* is omitted from the regression estimation, the OVB is described as the final term in the following expression[1]

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{\delta Cov(X,V)}{Var(X)}$$

- OVB is driven by two components:
  - $Cov(X,Z)$: The relationship between the omitted variable (*V*) and the regressor of interest (*X*).
  - δ : The relationship between the omitted variable and the outcome (*Y*).
  - If either of these components is zero, we will not have OVB
- We rarely know the exact values of these two components (if we did, it must mean we have data on *V* and we could stop omitting it from the regression) but we often have (economic) intuition about their signs. So we often can (and should) think hard about the sign of the bias.
- Getting one's head around the sign (and direction) of the bias can be tricky because there are actually three signs floating around this problem: 1) the sign of the population slope, 2) δ, and 3) $Cov(X,V)$. This module lets students play around with all three separately and visualize how they affect the regression estimates.

Suggestions for Building Intuition

- Start by explaining the example that is the basis of the module: the relationship between grades or test scores and the amount of time spent studying. In many real data sets we see a zero or even negative *correlation* between hours of studying and grades (Stinebrickner and Stinebrickner 2008 provide background on this topic and an interesting IV strategy that students usual enjoy hearing about – it involves video games!).
- Ask students to think about what they believe the *population* slope parameter (β_1) to be. Thinking about the population relationship can be a deep concept for students. It may help to ask students to step away from what they might *see* in the data and try to think about what they believe is the true causal relationship between increasing study hours and test scores. It is also helpful to ask them to imagine a randomized control trial where they could "hold everything else constant."
- Next, ask students to think through the relationship between the omitted variable (sleep) and test score ($Cov(X,V)$). Next, think about the relationship between hours of sleep and hours spent studying (δ). Ask them to write down these explanations in words without any jargon and then assign a "sign" to each.
- Have the students enter numbers for the three parameters ($\beta_1, Cov(X,V)$, and δ) with the chosen signs. Next, plot the data points and the naïve regression line based on a single variable regression of *Score* on *Study* by clicking the "Generate" button. The preloaded values assume a positive population relationship between study and score (which is intuitive to most students) but show a negative correlation due to negative correlation between sleep hours and study hours and

[1] Assuming that $E(\epsilon_i|X_i) = 0$

a positive relationship between sleep and scores. The simulation will be informative even if students have different beliefs about the parameters and discussing these difference and the resulting outcomes amongst themselves is often useful.

- Before plotting the corrected line, ask students to use the OVB equation and their chosen parameters to sign the bias. What do they expect will happen to the slope coefficient if the omitted variable is added to the regression? Click "Show the Corrected Regression Line" and see if their prediction is correct.
- Have students play around with different choices for the three parameters. It is particularly useful to have students try setting first $Cov(X, V) = 0$, and then $\delta = 0$.
- While the module is written for this particular example, students can use the functionality to help visualize OVB in other scenarios. Ask students to consider the example of the relationship between GDP Growth and Inequality, for example (an interesting and hot topic that is discussed in Forbes 2000). *Y*=GDP Growth, *X*=Inequality, *V*=Corruption, degree of capitalism, etc.

References:

Forbes, Kristin J. 2000. "A Reassessment of the Relationship Between Inequality and Growth." *American Economic Review* 90 (4): 869–87. https://doi.org/10.1257/aer.90.4.869.

Stinebrickner, Ralph, and Todd R Stinebrickner. 2008. "The Causal Effect of Studying on Academic Performance." *The B.E. Journal of Economic Analysis & Policy* 8 (1): 1–55.

Suggested Handout for Lab Activity or Self-Guided Work

Point your favorite browser to go.middlebury.edu/econsims

Click "Start" and select the "Omitted Variable Bias" module from the Menu. In this module,

In many real world data sets we see a zero or even negative correlation between the number of hours students spend studying and their grades. This may be seem surprising to you. Omitted variable bias is a potential explanation for this surprising observation.

Q1: What do you think the true causal relationship is between the number of hours a student spends studying and the score they get on a test? I do not mean what do we see in the data, I mean imagine we could hold all else equal and randomly assign some students to study more than others, would we see an increase or decrease in their grade (or would their grade not change at all – remember zero is always a possibility.). If you believe studying more increases your grade, you are positing a positive slope in the population regression function (β_1). Plug in a positive number in the relevant box. If you think studying more will decrease your score, then plug in a negative number.

Q2: I think that the amount of sleep students get may be related to both the amount of time they sleep and how well they do on a test. Write 1 to 2 sentences hypothesizing about the relationship between sleep and study ($Cov(X,V)$. Plug in a number with the corresponding sign in the relevant box. Next, think about the relationship between sleep and test scores (δ). Plug in a number with the corresponding sign in the relevant box.

Q3: Push the "Generate" button. Take a look at the corresponding scatter plot and regression line which is based on a single variable regression of Test Score on Study. How does the slope of this "naïve" regression line differ from the population slope you entered? Is it steeper? Flatter? Of a different sign?

Q4: Write down the omitted variable bias equation and use the values you entered for $Cov(X,V)$, and δ to sign the bias. Based on this calculation, what do expect to happen to the slope of the regression line if we were to add sleep to the regression equation rather than omitting it?

Q5: Try some different numbers for each of the parameters – specifically trying changing their signs. Regenerate the scatter and naïve regression and walk through the OVB equation to determine the sign of the bias under the new scenarios.

Q6: Now set a non-zero value for δ but set $Cov(X,V) = 0$. Compare the slope in the naïve regression line to the population slope you picked. What happens when you click "Show Corrected Regression Line"? Set a non-zero value for $Cov(X,V)$ but set $\delta = 0$. What happens?

Q7: Let's try some other examples. How about the relationship between GDP Growth and Inequality, for example. Y=GDP Growth, X=Inequality, Z=Corruption, degree of capitalism, etc. Repeat the 6 previous steps for this scenario.