

Topic 1 Introductory Notes: Notes about sampling distributions and standard errors

These notes are provided as a supplement to the lecture slides in BSM946 and are not expected to be used as your sole understanding of this topic. The main arguments are presented in the lecture; with extra notes here and the key reading to ensure your understanding of the topic is complete.

In term 1 you spent a lot of time building up to the OLS regression model, looking at why it is such an important model and what assumptions need to hold for OLS to be our preferred model or estimator. If these assumptions (known as the Gauss Markov assumptions) hold then OLS can be shown to be BLUE (the Best Linear Unbiased Estimator). This is a very strong result and one that we find students are often a little under-excited about. Why is this? Often students are not really sure what “Best” and “Unbiased” mean in BLUE and thus they struggle to truly understand this result. These notes look to explain the concept of a sampling distribution so that the idea of a best unbiased estimator can be fully understood. This is the main take away point from the Topic 1 lecture.

So our first step is ensuring we understand sampling distributions. When we are estimating a relationship between variables we are looking at subsamples of a population. Imagine we are interested in whether firms in the UK that spend more in training their workers are more productive. The population here would be all firms in the UK. We are never likely to have information on ALL of these firms. Instead we may use a questionnaire that has been sent to 10,000 companies. This is our subsample of the population. The last statistic I saw on the UK showed there were 5.2 million firms. Thus our subsample will be around one firm in every 520 or 0.2% of the population.

The first thing to note here is that we are trying to work out relationships for the population and thus if our subsample includes a larger percentage of the population our estimates are more likely to be accurate. Thus if we increased our sample size up to a million firms we would have around 20% of the population and thus would be expected to get better estimates.

How does this idea relate to sampling distributions? Well our subsample of 10,000 firms is just one of many possible subsamples that could have been drawn from the population of firms in the UK. Let's think about a very basic concept such as a mean. If we want to look at the mean firm size in the UK and we use our subsample we will get a given estimate. If we took a different subsample of 10,000 firms in the UK then we would likely get a different figure for the mean. We now have two different estimates of the mean firm size in the UK. Which one is right? Well.... Both of them and neither of them.

We could keep taking subsamples in this manner and we can plot out our values for the mean. Let's pretend we did this in Table 1 below.

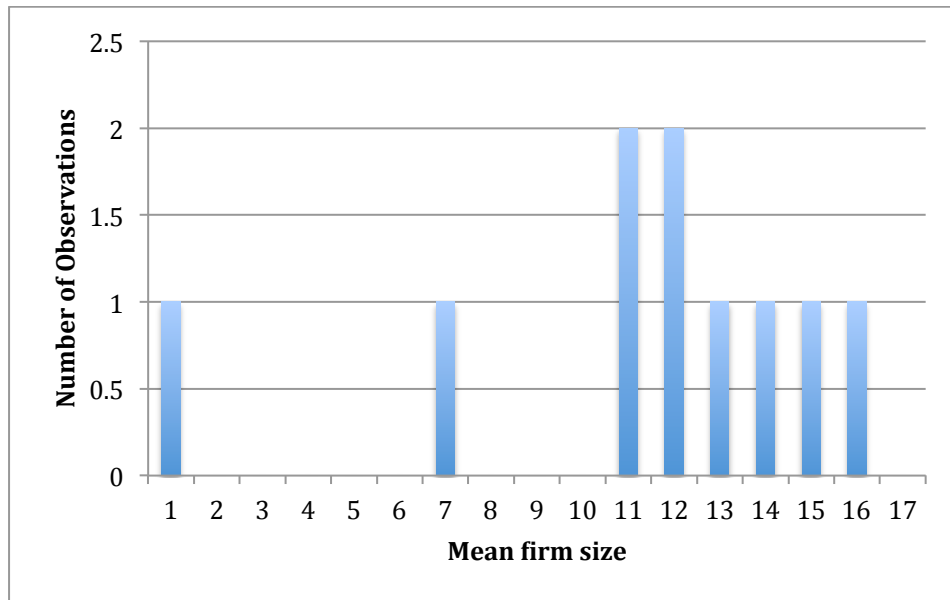
Table 1: Subsample means

Sample	Mean (Number of Employees)
Subsample 1	24
Subsample 2	28
Subsample 3	25
Subsample 4	24
Subsample 5	20
Subsample 6	14
Subsample 7	29
Subsample 8	25
Subsample 9	26

Subsample 10	27
Average:	24.2

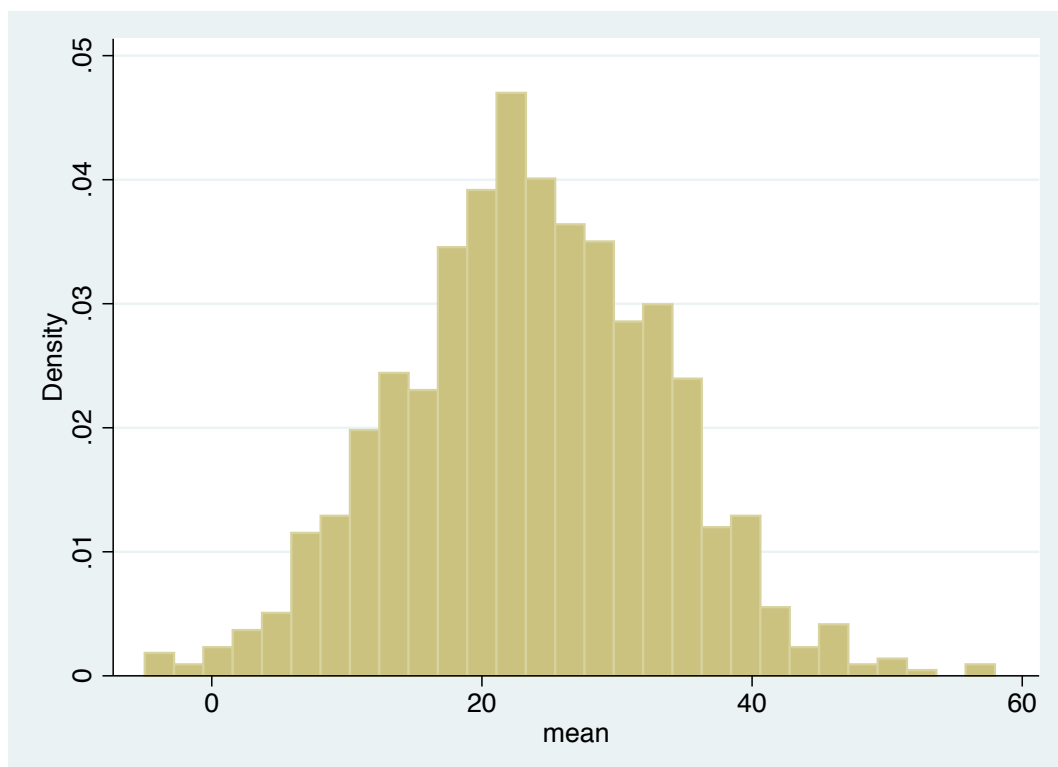
Each of our different subsamples from the 5.2 million firms is likely to give us a different answer for the mean number of employees. We could actually plot these to form a distribution of the means.

Figure 1: Sampling distribution of Table 1



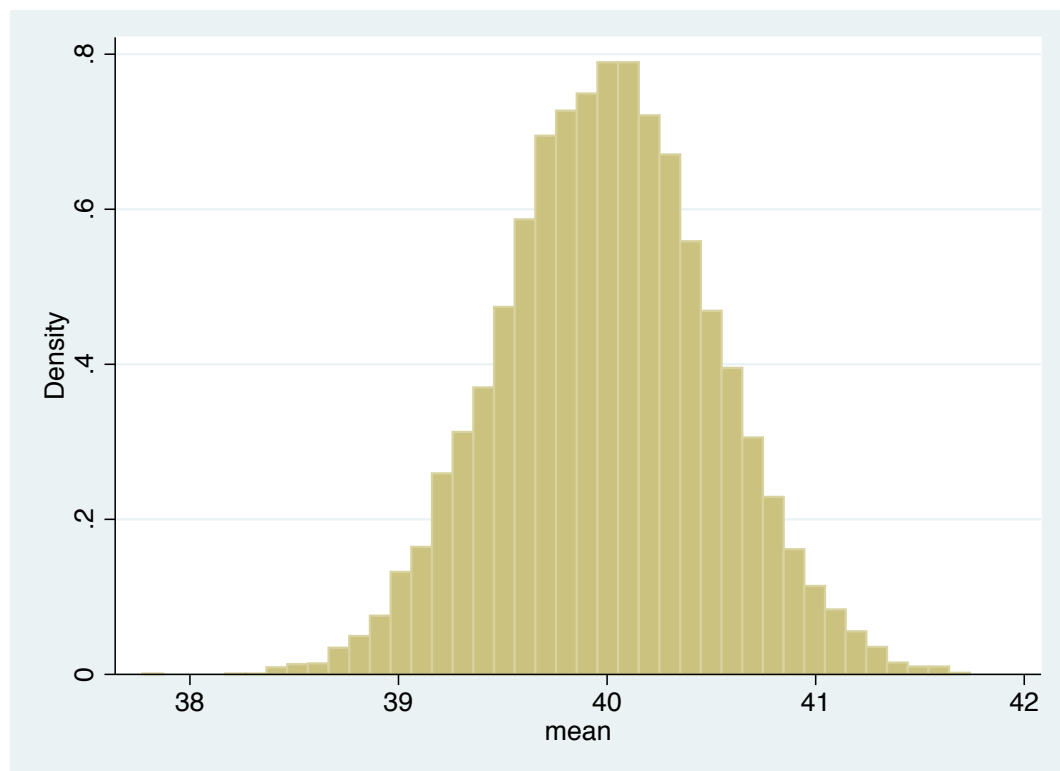
If we draw a larger number of subsamples and plot them together this is what we would call a sampling distribution for this statistic of interest. We can see this below in Figure 2.

Figure 2: Sampling distribution for a larger sample



In reality we never really do this when estimating a relationship, we just draw one sample from the population and use this to find our statistic of interest. By knowing the shape of the sampling distribution for a statistic however, we can work out how well our model works. Looking at the sampling distribution we can tell if the statistical model gives us the true population mean on average and whether it gives us a narrow variance. For example, if I knew the true mean number of employees in a firm is 40, and I saw the sampling distribution above in Figure 2, I may believe I am calculating the mean wrong as the majority of my samples are suggesting the size of firms is around 24.

Because of this, one of the key things econometricians are interested in is working out the sampling distributions for different estimators under different circumstances to understand when a particular model will work well, and when it will work badly. In terms of working well we are mainly interested in a sampling distribution that on average gives the true population value. Again if the true mean number of employees in a firm is 40 then the sampling distribution in Figure 2 is very poor as it is what we call “Biased”, it is not centered on the true population mean. We are also interested in what we call “Best” where the sampling distribution has a tighter fit around the true population average. Bearing this in mind, we would much prefer the estimator that gives Figure 3 to Figure 2. This now appears to be unbiased and have a tighter fit to the true value.



That is the basic concept of a sampling distribution and why they are so integral to estimation. This course will be focused on understanding why OLS is sometimes not the best model to use, i.e. the properties of its sampling distributions are not great, and which models will have better properties in these situations.

Thought of the week on standard errors:

Students seem to be more interested in coefficients when estimating relationships, often thinking about standard errors and p-values as matters of secondary importance. Without standard errors coefficients are largely uninteresting however, as we don't know how much noise there is around the point estimate and thus whether we should pay attention to it.

For example, if we obtain a coefficient of 10 from a regression, this does not mean much without considering the standard error around it. With a standard error of 1, then the estimate of 10 is very much different to zero and we should pay attention to this result. However, if the standard error is instead 1000, then a coefficient of 10 is meaningless and is likely just random noise. Thus the coefficients do not really matter until we start to consider their standard error.

We will look at a number of issues in estimation throughout this course, some of which will cause our coefficients to be biased, and some which will cause our standard errors to be wrong. Neither of these outcomes is good but standard errors are often much more difficult to calculate than coefficients and even more sensitive to correct specification. Biased standard errors can have the effect of making an outcome look less likely than it actually is (over-reject the null) or more likely than it actually is (under-reject the null). Biased coefficients are likely to make you under or over estimate the importance of a particular factor, but the bias is usually not large enough to lead you to find the wrong result (i.e. the opposite sign). In the case of standard errors we can quite easily find the wrong result (i.e. incorrectly reject the null etc.) if our standard errors are incorrect.

This is why we need to be careful with some of the simple underlying assumptions of OLS, running a quick check for heteroskedasticity (more on this next week!) or multicollinearity is important, as is thinking about the model you are running and its application in this particular situation.

As we move on through this course we will be looking at situations where the OLS assumptions do not hold and thus the classical linear regression model will not be BEST. To allow us to still perform reliable estimation and generate both unbiased coefficients and accurate standard errors we will look at a range of other models. Do catch up on OLS and its assumptions if you are a little weak on these as they will be important throughout this course.

That is all for this week, see you again soon!

David Morris