SAMPLE DESIGN OF RURAL ASER

The purpose of rural ASER is twofold:

(i) To get reliable estimates of the status of children's schooling and basic learning (reading, writing and math ability) at the district level; and

(ii) To measure the change in these basic learning and school statistics over time. Every year a core set of questions regarding schooling status and basic learning levels remains the same. However a set of new questions are added for exploring different dimensions of schooling and learning in the elementary stage. The latter set of questions is different each year.

Since one of the goals of ASER is to generate estimates of change in learning, a panel survey design would provide more efficient estimates of the change. However, given the large sample size of the ASER surveys and cost considerations, a rotating panel of villages (rather than children) is adopted. Each year, 10 villages from 2 years ago are dropped and 10 new villages added. For instance, in ASER 2010 the 10 villages from ASER 2007 were dropped, the 10 villages from 2008 and 2009 were retained and 10 new villages from the census village directory of 2001 were added.

The sampling strategy used generates a representative picture of each district. All rural districts are surveyed. The estimates obtained are then aggregated to the state and all-India levels.

Since estimates are generated at the district level, the minimum sample size calculations start at the district level. The sample size is determined by the following considerations:

- Incidence of what is being measured in the population. Since a survey of learning has never been done in India, the incidence of what we are trying to measure is unknown in the population.[1]
- Confidence level of estimates. The standard used is 95%.
- Precision required on either side of the true value. The standard degree of accuracy most surveys employ is between 5 and 10 per cent. An absolute precision of 5% along with a 95% confidence level implies that the estimates generated by the survey will be within 5 percentage points of the true values with a 95% probability. The precision can also be specified in relative terms --- a relative precision of 5% means that the estimates will be within 5% of the true value. Relative precision requires higher sample sizes.

Sample size calculations can be done in various ways, depending on what assumptions are made about the underlying population. With a 50% incidence, 95% confidence level and 5% absolute precision, the minimum sample size required in each strata[2] is 384.[3] This derivation assumes that the

[1] For the rural sector we can use the estimates from previous ASERs to get an idea of the incidence in the population.

[2] Stratification is discussed below.

population proportion is normally distributed. On the other hand, a sample size of 384 would imply a relative precision of 10%. If we were to require a 5% relative precision, the sample size would increase to 1600.[4] Note that all the sample size calculations require estimating the incidence in the population. In our case, we can get an estimate of the incidence from previous ASER surveys. However, incidence varies across different indicators --- so incidence of reading ability is different from incidence of dropouts. In addition, we often want to measure things that are not binary for which we need more observations.

Given these considerations, the sample size is set at 600 households in each district.[5] Note that at the state level and at the all-India level the survey has many more observations lending estimates at those levels much higher levels of precision.

ASER has a two-stage sample design. In the first stage, 30 villages are randomly selected using the village directory of the 2001 census as the sample frame.[6] In the second stage 20 households are randomly selected in each of the 30 selected villages in the first stage.

Villages are selected using the probability proportional to size (PPS) sampling method. This method allows villages with larger populations to have a higher chance of being selected in the sample. It is most useful when the sampling units vary considerably in size because it assures that those in larger sites have the same probability of getting into the sample as those in smaller sites, and vice verse.[7], [8]

In the selected villages, 20 households are surveyed. Ideally, a complete houselist of the selected village should be made and 20 households selected randomly from it. However, given time and resource constraints a procedure for selecting households is adopted that preserves randomness as much as possible. The field investigators were asked to divide the village into four parts. This is done

---

[3] The sample size with absolute precision is given by $\frac{z^2 p q}{d^2}$ where *z* is the standard normal deviate corresponding to 95% probability (=1.96), *p* is the incidence in the population (0.5), *q=(1-p)* and *d* is the degree of precision required (0.05).

[4] The sample size with relative precision is given by $\frac{z^2 q}{r^2 p}$ where *z* is the standard normal deviate corresponding to 95% probability (=1.96), *p* is the incidence in the population (0.5), *q=(1-p)* and *r* is the degree of relative precision required (0.1).

[5] Sample size calculations assume simple random sampling. However, simple random sampling is unlikely to be the method of choice in an actual field survey. Therefore, often a "design effect" is added to the sample size. A design effect of 2 would double the sample size. At the district level a 7% precision along with a 95% confidence level would imply a sample size of 196, giving us a design effect of approximately three. However, note that a sample size of 600 households gives us approximately 1000 – 1200 children per district.

[6] Of these 30 villages, 10 are from the ASER two years ago, 10 from ASER of the previous year and 10 are newly selected. These are selected independently from the same sample frame.

[7] Probability proportional to size (PPS) is a sampling technique in which the probability of selecting a sampling unit (village, in the case of ASER) is proportional to the size of its population. The method works as follows: First, the cumulative population by village calculated. Second, the total household population of the district is divided by the number of sampling units (villages) to get the sampling interval (SI). Third, a random number between 1 and the SI is chosen. This is referred to as the random start (RS). The RS denotes the site of the first village to be selected from the cumulated population. Fourth, the following series of numbers is formed: RS; RS+SI; RS+2SI; RS+3SI; …. The villages selected are those for which the cumulative population, contains the numbers in the series.

[8] Most large household surveys in India, like the National Sample Survey and the National Family Health Survey also use this two stage design and use PPS to select villages in the first stage.

because villages often consist of hamlets and a procedure that randomly selects households from some central location may miss out households on the periphery of the village. In each of the four parts, investigators are asked to start at a central location and pick every 5th household in a circular fashion till 5 households are selected. In each selected household, all children in the age group of 5-16 were tested.[9]

The survey provides estimates at the district, state and national levels. In order to aggregate estimates up from the district level households have to be assigned weights – also called inflation factors. The inflation factor corresponding to a particular household denotes the number of households that the sampled household represents in the population. Given that 600 households are sampled in each district regardless of the size of the district, a household in a larger district will represent many more households and, therefore, have a larger weight associated with it than one in a sparsely populated district.

The advantage of using PPS sampling is that the sample is self weighting at the district level. In other words, in each district the weight assigned to each of the sampled household turns out to be the same. This is because the inflation factor associated with a household is simply the inverse of the probability of it being selected into the sample times the number of households in the sample. Since PPS sampling ensures that all households have an equal chance of being selected at the district level, the weights associated with households in the same district are the same. Therefore, weighted estimates are exactly the same as the un-weighted estimates at the district level. However, to get estimates at the state and national levels, weighted estimates are needed since states have a different number of districts and districts vary by population.

Even though the purpose of the survey is to estimate learning levels among children, the household is chosen as the second stage sampling unit. This has a number of advantages. First, children are tested at home rather than in school, allowing all children to be tested rather than just those in school. This also ensures that children in different types of schools – government, private or religious – are captured in the survey. Further, testing children in school might create a self-selection bias since only children who are present in school on the day of the survey will end up being tested. Second, a household sample will generate an age distribution of children which can be cross-checked with other data sources, like the census and the NSS. Third, a household sample makes calculation of the inflation factors easier since the population of children is no longer needed.

Often household surveys are stratified on various parameters of interest. The reason for stratification is to get enough observations on entities that have the characteristic that is being studied. The ASER survey stratifies the sample by population in the first stage. No stratification is done at the second stage. Finally, if stratification is done on households with children in the 3-16 age group, the population of such households in the village is needed, which is not possible without a complete houselist of the village.

[9] In larger villages, the investigators increased the interval according to a rough estimate of the number of households in each part. For instance, if a village had 2000 households, each part in the village would have roughly 500 households. Selecting every 5th household would leave out a large chunk of the village un-surveyed. In such situations, investigators were asked to increase the interval between selected households.