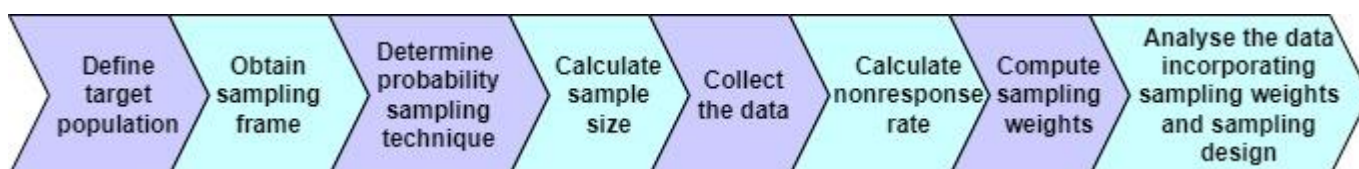


SAMPLING IS ESSENTIAL BUT HOW?

In a survey, we want to know certain characteristics of a large population, but we are almost never able to do a complete census of it. So we draw a sample—a subset of the population—and collect data on it. Then we generalize the results, with an allowance for sampling error (usually within 5% at 95% confidence level), to the entire population from which the sample was selected. If we interview people in a “convenience” sample—whoever easy to find—we cannot calculate the sampling error or quantify possible selection bias. This reduces the value of a survey to a qualitative study, where the results are limited to the respondents themselves, and therefore are only indicative and not representative of the target population. To have confidence in generalizing sample results to the target population requires a “probability sample” of the population.

Probability Sampling Process

- 1 Define target population** – the population of interest for the survey that realistically everyone can be reached and invited to participate in the survey if randomly selected. There are sometimes limitations (e.g., security, distance, etc.) that exclude some population of interest from the target survey population because they are unreachable and have a zero probability to be selected.
- 2 Obtain sampling frame** – a sampling frame is a complete list of everyone in the target population for calculating each respondent’s probability of selection of being selected for the survey. It is necessary a probability sampling design.
- 3 Determine probability sampling technique** – random selection methods that are appropriate for different sampling frames. Simple random sampling is the simplest but only feasible for a list frame. An area frame typically requires a multi-stage sampling.
- 4 Calculate sample size** – the formula differs by the sampling technique. Most online sample size calculators are based on the single-stage random sampling design.
- 5 Collect the data** – reach out to the selected participants to conduct the survey following the protocols (e.g., enumerating every residential household in a sampled village) of the chosen sampling technique.
- 6 Calculate nonresponse rate** – keep track of selected participants who did not respond to the survey and calculate the response rate.
- 7 Compute sampling weights** – the inverse probability of each respondent being selected adjusting for the nonresponse rate.
- 8 Analyse the data incorporating sampling weights and sampling design** – ignoring the sampling weights generates incorrect estimates, and ignoring the multi-stage sampling design generates incorrect standard errors for the estimates. Use a sampling software to analyse the data derived from a complex sampling design.



National Society's Role in the Sampling Process

Define the target population – The NS specifies the survey objectives and scope/target population. Only the NS knows their capacity and limitations to reach the population of interest. For example, some areas are unsafe or too far for their staff/volunteers to travel, hence excluded from the target survey population.

Obtain the sampling frame – The NS should consider what sampling frame with contact information (e.g., address, phone, email, etc.) of each person/household is available to them. In rare instances, a list of the entire population may be available, e.g., a list of all the migrant camps by location with the number of migrants and/or families living in each camp. All of them are reachable since the locations of the camps are known.

Ensure the field team to adhere to the data collection protocols – There are strict protocols to follow in order to fulfil a probability-sampling design. The protocols may include enumerating all eligible persons/households in the sampled village or city block, using a random number table to decide which person/house to interview, keeping track of the nonresponse of the originally sampled participants, etc. The sampling expert may have given the field data collection training, still, the NS is responsible to oversee that the field team is adhering to the protocols.

See the flowchart of a probability sampling design process at the end of this guide that shows the decisions that the NS needs to make, recommended sampling techniques for different scenarios, and resources associated with each step in the process.

Terminology

- **Probability sampling.** Each individual or household in the sampling frame has a known but not necessarily equal probability of selected to participate in the survey.
- **Convenience sampling.** A type of non-probability sampling where the sample is taken from a group of people easy to contact or to reach rather than a random selection from everyone in the target population. The data is not representative of the target population because the selection bias is unknown.
- **Quota sampling.** A type of non-probability sampling where a fixed number of participants from mutually exclusive subgroups are selected.
- **Simple random sampling.** The simplest form of probability sampling where everyone in the target population has an equal chance of being selected. This can be implemented using a random number generator.
- **Systematic sampling.** A probability sampling method where random starting points with fixed interval are used to select members from a larger population. This interval, called the sampling interval, is calculated by dividing the population size by the desired sample size.
- **Proportionate-to-size sampling.** A probability sampling method where the more populous segments of the target population have a higher chance of being selected that is proportional to their relative size.
- **Multi-stage sampling.** A probability sampling method where samples are drawn first from higher order groupings (e.g., provinces) and then from successively lower level groupings (e.g., districts within provinces, towns within districts) in order to make the data collection more practical.
- **Gridded sampling.** A probability sampling method where the total population areas are divided into small grid cells derived with a geo-statistical model using census or publicly available spatial datasets (e.g., Google maps).
- **Time-location sampling.** A probability sampling method of participants at specific times in set locations. The sampling frame consists as time-location units which represent the potential universe of locations, days and times.
- **List frame.** A list of target population individuals or households with contact information (address, phone, email, etc.) whereby they can be reached and invited to participate in the survey.
- **Area frame.** An area sampling frame consists of geographical units arranged hierarchically, which may include province, district, tract, ward and village (rural areas) or block (urban areas).
- **Random-walk.** Use a random number generator to determine the direction taken and the distance moved between sample points from the starting point (e.g., the middle of a sampled village or city block).
- **Stratification.** Dividing a target population into smaller groups based on the shared characteristics of the members in the group. Stratified sampling ensures each subgroup has an enough sample size to generate reliable estimates.
- **Cluster.** Naturally existing homogeneous groups (e.g., a village, a school or a class within a school).

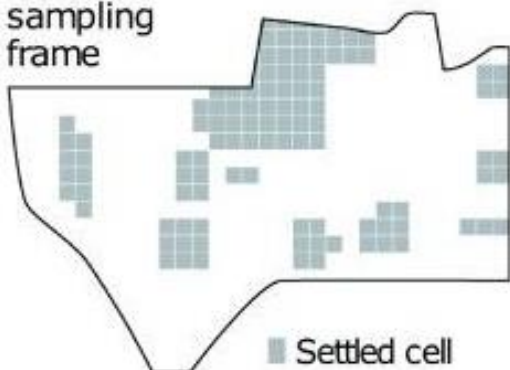
Gridded sampling design based on GIS maps – uses a gridded population dataset as the area sampling frame. Gridded population datasets are estimates of the total population in small grid cells derived with a geo-statistical model using spatial datasets. In gridded population sampling, grid cells are often aggregated into clusters of a desired population size, and used in place of census enumeration areas. [GridSample.org](https://www.grid-sample.org/) is a free web-based tool that provides a point-and-click interface, preloaded datasets, and guidance to enter parameters and select clusters for a gridded population survey. Preloaded datasets include WorldPop-Global 100×100 meter gridded population estimates, Global Administrative Areas (GADM) administrative boundaries (at the country, provincial, and district scale), and Global Human Settlement (GHS-SMOD) urban/rural boundaries. The main reason for using gridded population sampling is the lack of access to the census sample frame. The second reason is that standard survey methods struggle to sample nomadic and vulnerable households accurately. Spatial sampling covers all areas within a geographical region and hence the entire population are included in the sampling frame.

A Sampling frame

A1 Study area

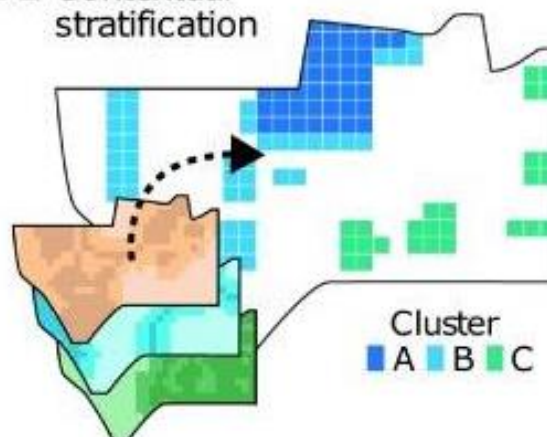


A2 Gridded sampling frame

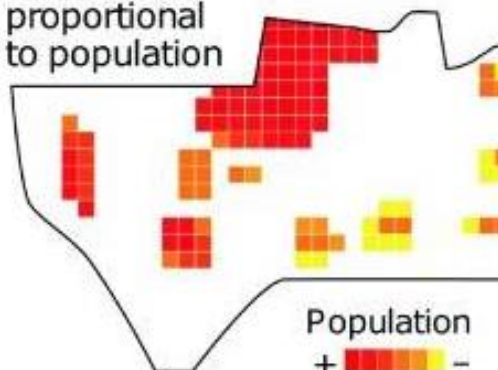


B Sampling design

B1 Contextual stratification

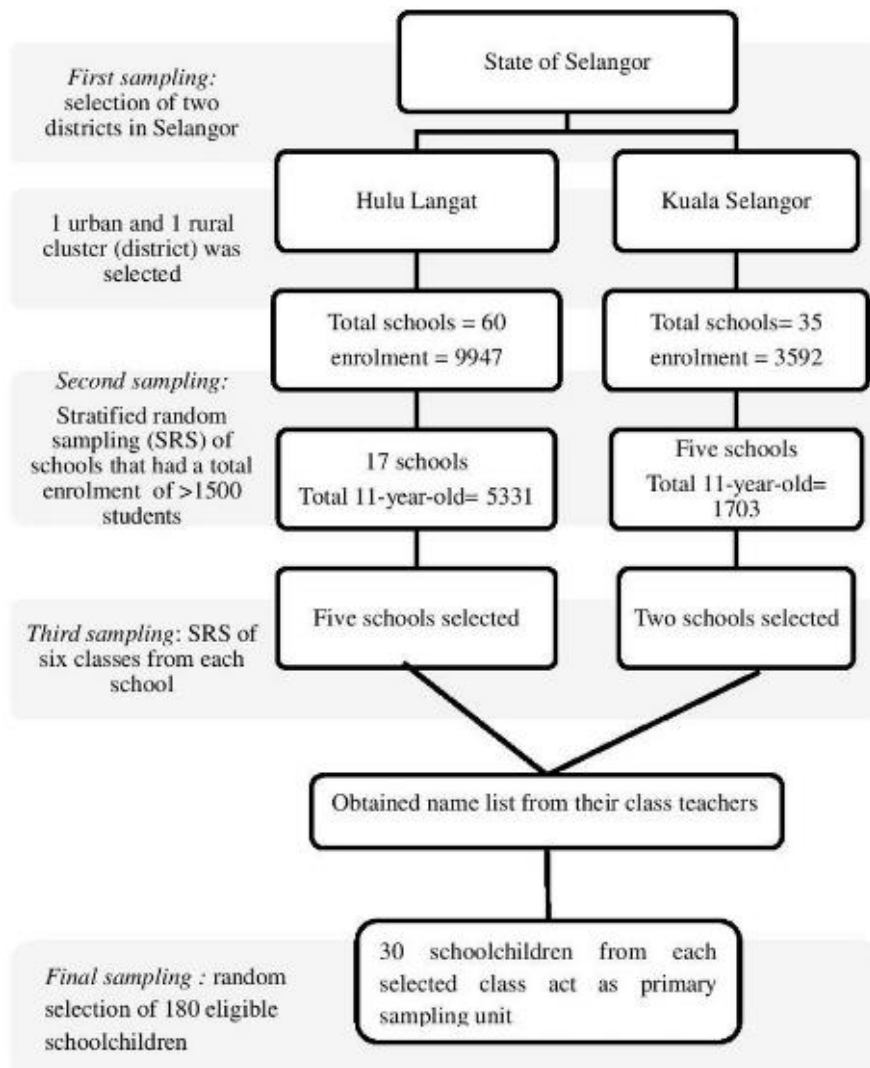


B2 Probability proportional to population



A grid-based sample design framework for household surveys.

Gianluca Boo et al., *Gates Open Research*, 27 January, 2020, doi: 10.12688/gatesopenres.13107.1



An example of the process of the multi-stage stratified random sampling method. Noor azhani Zakaria et al., "A Study on Parental Acceptance Towards the Use of Dental Therapists in Malaysian Private Sectors", *Malaysian Journal of Medicine and Health Sciences* 16(4):13-20



Malls



Metro stations



Community centres

Time-location sampling – a probabilistic method used to recruit members of a target population at specific times in set locations. The sampling frame consists as time-location units which represent the potential universe of locations, days and times (e.g., 8-10 am Mon-Sun). The field team visits the locations and prepares a list of time-location units which are considered potentially eligible on the basis of checking the number of people present. In addition, interviews are conducted with those in charge of location to ascertain affluence on certain days and at certain times. With this information, population size for each time-location unit, and the number eligible for each sample are estimated.

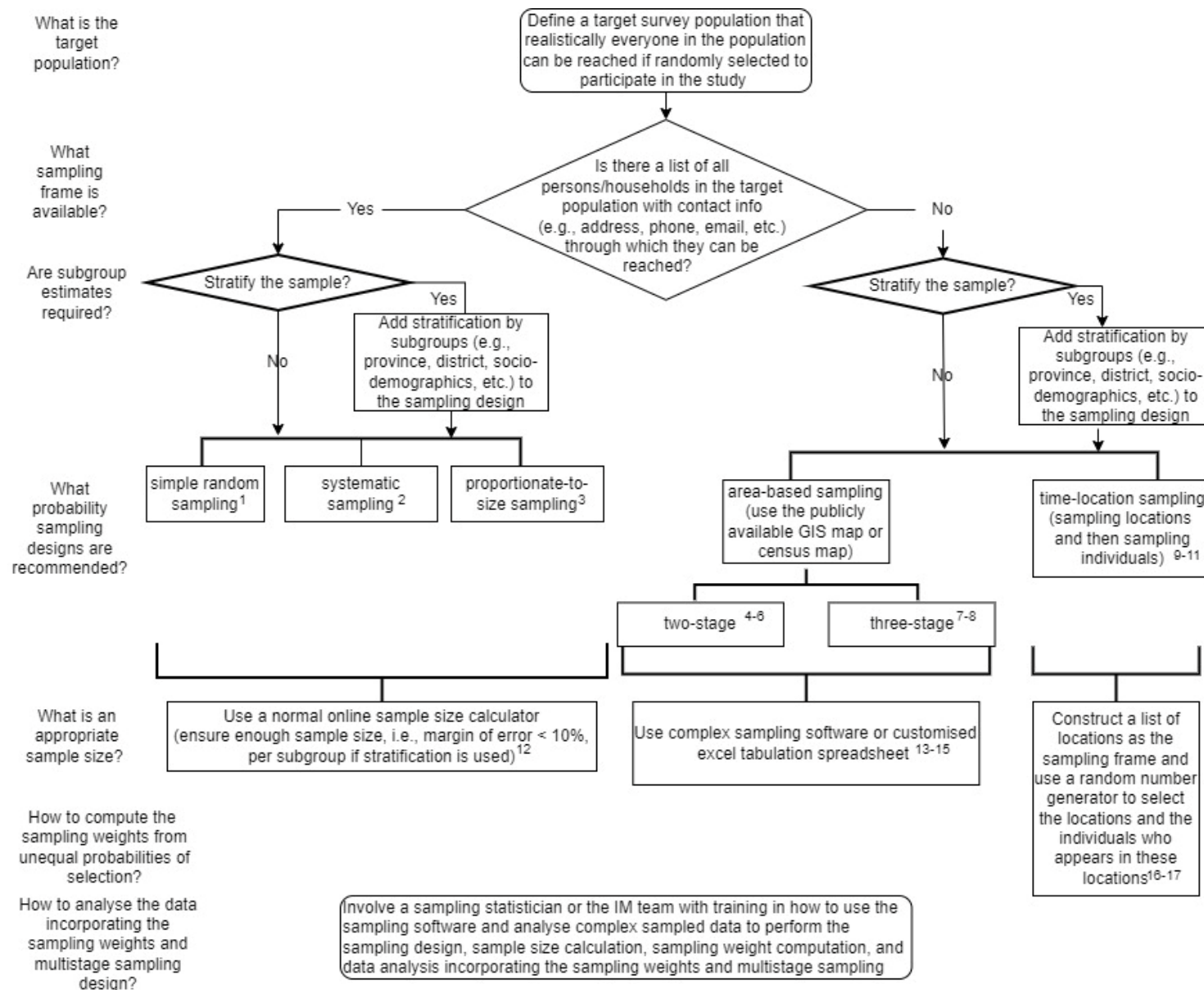
Location	Time-Location Units						
	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Community Centre 1	8:00-10:00 17:00-19:00			8:00-10:00 17:00-19:00			8:00-10:00 17:00-19:00
Community Centre 2		8:00-10:00 17:00-19:00			8:00-10:00 17:00-19:00		
Community Centre 3			8:00-10:00 17:00-19:00			8:00-10:00 17:00-19:00	

Common Mistakes that Jeopardize Probability Sampling

- Fail to list every eligible household or person when the enumerators visit the selected village, city block, etc.** – The census or GIS maps do not show the actual number of eligible households in a sampled village or city block (for example, see the satellite image of a sampled block). The field team must number each eligible (e.g., residential) household and interview the sampled household following the predetermined probability sampling protocol, and do not interview any households available at the time of visit.
- Substitute nonrespondents with someone who has not been selected in the original sample** – The nonresponse rate is part of deriving the final sampling probabilities of the survey participants. The enumerator should keep track of whether each originally sampled participant has responded or not, rather than substituting a nonrespondent with another participant who was not in the original sample.
- Within household non-random selection** – At the final stage of sampling, the enumerator may be instructed to interview only one of the eligible participants within the sampled household. It is crucial to use a random selection method to choose the final respondent and not conveniently interview anybody who is at home at the time. A pseudo within-household random sampling technique is interviewing the eligible person who will be the next celebrating a birthday.
- Replace household listing with random-walk selection** – Random-walk is often justified as a way to avoid costly and time-consuming listing of all households in the selected cluster in the absence of a sampling frame. As long as the starting point is selected randomly and probabilities of selection can be calculated (the number of households selected divided by the total number eligible for selection from the sampling frame), thus obtaining a probability sample. However, these two conditions are rarely met. First, by starting household selection from the center of the cluster, households near the center are more likely to be selected than outlying households. Second, the total number of households in the cluster is rarely known, as the number of eligible households in an area is not always available.
- Last-minute planning** – A probability sampling design process can take anywhere from a few days to a few weeks depending on the complexity of the design. Therefore, it is crucial to start the planning process at least one month before the estimated data collection field period. See the flowchart of a probability sampling design process at the end of this guide that shows the decisions that the NS needs to make, recommended sampling techniques for different scenarios, and resources associated with each step in the process.



A Flowchart of Probability Sampling Design Process



- <https://www.excel-demy.com/select-random-sample-in-excel/>
- <https://statisticsbyjim.com/basics/systematic-sampling/>
- <https://www.iedunote.com/pps-sampling>
- <https://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-11-12>
- https://www.cdc.gov/nceh/caspe/docs/CASPER-toolkit-3_508.pdf
- <https://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-10-785>
- <https://pubmed.ncbi.nlm.nih.gov/26844121/>
- <https://journals.sagepub.com/doi/10.1177/0002764220910223>
- <https://www.demographic-research.org/volumes/vol26/5/26-5.pdf>
- <https://iussp2009.princeton.edu/papers/93359>
- http://hivhub.ir/wp-content/uploads/2018/07/TLS_E_NG.pdf
- <http://www.raosoft.com/samplesize.html>
- <https://www.who.int/teams/noncommunicable-diseases/surveillance/systems-tools/steps/planning-sampling>
- https://cdn.who.int/media/docs/default-source/immunization/immunization-coverage/sample_size_calculator_survey.xlsx?sfvrsn=5a05341_6
- <https://app.gridsample.org/tool/>
- <https://globalhealthsciences.ucsf.edu/sites/globalhealthsciences.ucsf.edu/files/tls-res-guide-2nd-edition.pdf>
- <https://academic.oup.com/biostatistics/article/16/3/565/269802>