## Contents

**About**
- Why use GridSample? 3

**Before using GridSample**
- Survey Design 5

**Using GridSample**
- Sample selection 7
- How GridSample works 13
- GridSample output 14

**Implementing your survey**
- Approaches & tools 15

**After your survey**
- Sample weights 19
- Data cleaning & analysis 23

**Practice exercises**
- Exercise 1: 2014-15 Rwanda DHS 24
- Exercise 2: 2017 Kathmandu SUE 27

**Additional resources**

**Exercise answers** 33
Why use GridSample?

**Outdated or inaccurate census sample frame.** Standard “top-down” gridded population datasets, such as WorldPop-Global, reflect the population totals in census data. While census population totals may be inaccurate, the relative distribution of the population in WorldPop-Global may be more accurate than the original census data, and thus serve as a better sample frame. This is because gridded population datasets disaggregate population totals to areas with new population growth (reflected in new housing developments, roads, and changes in land cover) and in informal settlements not originally counted in the census.

(updated April 2019)
**Spatial oversampling to improve small area estimates.** In household surveys, it is common to collect latitude-longitude coordinates of clusters, and to use survey results to generate small area estimates in administrative areas smaller than strata. Small area estimates based on standard samples of the population distribution are prone to reduced accuracy in remote rural areas, and within highly heterogeneous urban populations. Spatial oversampling ensures spatial coverage of the sample and a base-level of accuracy in small area estimates across the study area.

**Custom sizes clusters.** In standard census-based sampling, cluster size is determined by the size of census enumeration areas, with each cluster containing approximately 500 people. When the sample frame is generated from a gridded population dataset such as WorldPop-Global, there is good reason to consider clusters of different sizes.

- **Small clusters (~75 people each)** can be generated to perform one-stage sampling. One-stage sampling in which all eligible households in the cluster are interviewed, is an attractive option in complex settings where large numbers of vulnerable and mobile populations live. One-stage sampling allows for the household listing and interviews to be conducted on the same day, removing the months, or even years, long delay between listing and interviewers faced in standard multi-stage sampling.

- **Medium clusters (~500 people each)** are approximately the size of a census enumeration area and can be used to replicate standard household survey designs and methods with a gridded population sample frame.

- **Large clusters (~1200 people each)** are an attractive option in settings where the census is extremely outdated, for example more than 15 years old. WorldPop-Global estimates are generally more accurate as its 100m X 100m grid cells are aggregated into larger units. The use of larger areas in the sample frame helps to smooth out errors in the gridded population dataset.
Before using GridSample

Survey Design

Before using GridSample, have a clear research question, decide the survey design and calculate sample size.

Several resources are available to assist with survey planning including materials produced by the World Health Organization (WHO) Vaccination Coverage Cluster Surveys, Multiple Indicator Cluster Surveys (MICS), and Demographic and Health Surveys (DHS).

The primary question affects the survey design, sample size, and budget. Identify the type of question that the survey aims to address.

Calculating a sample size that both meets the inferential goals of the survey and the budget constraints is an iterative process that must be negotiated within a survey steering group. Use the WHO Vaccination Coverage Survey sample size + budget spreadsheet to generate multiple side-by-side scenarios.

Checklist before using GridSample

- I know the coverage area of the survey
- I know the year the survey will be implemented
- I know whether the survey will be stratified, and the geographic boundaries of those strata
- I know the multi-stage cluster sampling design
- I know the target cluster size
- I know the target sample size (number of clusters, and households per cluster)
<table>
<thead>
<tr>
<th>Classification question</th>
<th>Descriptive or estimation question</th>
<th>Comparative or hypothesis-test question</th>
</tr>
</thead>
<tbody>
<tr>
<td>A classification survey labels groups as “pass”, “marginal”, or “fail” to inform programmatic decisions</td>
<td>An estimation survey results in quantitative estimates in one or more groups</td>
<td>A comparative survey make a quantitative estimate of difference between two groups, or of change between two time points</td>
</tr>
<tr>
<td>Example question: Which health districts have low coverage of antenatal care for pregnant women?</td>
<td>Example question: What percent of pregnant women receive one or more antenatal care visits?</td>
<td>Example question: Has the percent of pregnant women receiving antenatal care increased since the last survey?</td>
</tr>
<tr>
<td>Inferential goal:</td>
<td>Inferential goal:</td>
<td>Inferential goal:</td>
</tr>
<tr>
<td>• Pass/fail threshold</td>
<td>• Coverage estimate</td>
<td>• Minimum detectable difference between 2 groups</td>
</tr>
<tr>
<td>Uncertainty is reported as:</td>
<td>Uncertainty is reported as:</td>
<td>Uncertainty is reported as:</td>
</tr>
<tr>
<td>• Misclassification of pass</td>
<td>• Confidence interval</td>
<td>• Confidence interval</td>
</tr>
<tr>
<td>• Misclassification of fail</td>
<td>Design parameter:</td>
<td>Design parameter:</td>
</tr>
<tr>
<td></td>
<td>• ( \alpha ) (alpha) - the probability of Type 1 error - the hypothesis test declares the difference to be statistically significant when, in truth, it is not</td>
<td>• ( \alpha ) (alpha) - the probability of Type 1 error - the hypothesis test declares the difference to be statistically significant when, in truth, it is not</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• ( 1-\beta ) (beta) - the power of the test - this is the probability that the hypothesis correctly identified a statistically significant difference</td>
</tr>
<tr>
<td>Example parameters:</td>
<td>Example parameters:</td>
<td>Example parameters:</td>
</tr>
<tr>
<td>• Pass/fail threshold: 90%</td>
<td>• ( \alpha = 95% ) (95% confidence interval)</td>
<td>• difference = 10% or more</td>
</tr>
<tr>
<td>• Misclassify passes: 5%</td>
<td></td>
<td>• 2-sided hypothesis test</td>
</tr>
<tr>
<td>• Misclassify fails: 10%</td>
<td></td>
<td>• ( \alpha = 95% ) (95% confidence interval)</td>
</tr>
<tr>
<td>Generally requires 15+ PSUs per stratum</td>
<td></td>
<td>• power = 80% (( \beta = 20% ))</td>
</tr>
<tr>
<td>Generally requires 30+ PSUs per stratum</td>
<td></td>
<td>Sample size is highly dependent on the difference between groups and design parameters. 30-60 PSUs per stratum.</td>
</tr>
</tbody>
</table>
Using GridSample

Sample selection

Sample selection steps are demonstrated below with Exercise 1, reproducing the survey design of the 2015 Rwanda Demographic and Health Survey in GridSample.

1. Enter an email address to receive your sample output

2. Select “Start new GridSample” or open an existing GridSample job

3. Give your sample a name

4. Select the country

5. Define the survey coverage area

6. See the guidance text if needed

7. Select “Save and Next” when done defining coverage
1. Select the WorldPop dataset to use as the foundation of the sample frame
2. Define the sample frame unit
3. See the guidance text if needed
4. Calculate population per sample frame unit
5. Review the histogram(s) of sample frame unit population estimates
6. When you are satisfied with the gridded population-based sample frame, select “Save and Next”
1. Define whether the survey will have one or two stages of sampling
2. Select whether you will use stratification or sample oversampling
3. Optionally specify a “seed” number to replicate the sample
4. See the guidance text if needed
5. Select “Save and Next” when done flagging the sample design features
1. Define which boundaries will define strata
2. See the guidance text if needed
3. Calculate population per stratum
4. Review the population per stratum
5. If satisfied with the strata, select “Save and Next”
1. Give a descriptive label to your target population.

2. Enter the average number of target population members per household (e.g., from DHS report).

3. Enter the average household size.

4. See the guidance text if needed.

5. When the target population is defined, select “Save and Next.”

---

1. If sample is stratified, define how to allocate clusters to strata.

2. Enter total number of household to be sampled.

3. If a two-stage sample, specify the number of households to be sampled per cluster.

4. See the guidance text if needed.

5. When the sample size is entered, select “Save and Next.”
1. Review the sample summary
2. Optionally print the summary to store for your records
3. When ready, select “Confirm and Submit”
How GridSample works

Coverage

- Boundary
  - Underlying WorldPop-Global estimates
    - Frame unit: Single-cell
    - Frame unit: gridEZ
    - Frame unit: Own EA.shp
  - Histogram: pop per sample frame unit

Frame

- One-stage
  - Strata: Yes
    - Boundaries
      - Table: pop per stratum
    - Convert unit pops to target HHs
      - Equal or proportional allocation of clusters to strata
        - Total # HHs expected
          - Median # HHs/unit
            - # clusters
  - Strata: No
    - Boundaries
      - Table: pop per stratum
    - Convert unit pops to target HHs
      - Custom allocation of clusters to strata
        - Total # HHs desired
          - Median # HHs/unit
            - # clusters

Design

- Two-stage
  - Strata: Yes
    - Boundaries
      - Table: pop per stratum
    - Convert unit pops to target HHs
      - Equal or proportional allocation of clusters to strata
        - Total # HHs
          - # HHs per cluster
            - # clusters
  - Strata: No
    - Boundaries
      - Table: pop per stratum
    - Convert unit pops to target HHs
      - Custom allocation of clusters to strata
        - Total # HHs
          - # HHs per cluster
            - # clusters

Strata

- Target
  - Convert unit pops to target HHs
    - Equal or proportional allocation of clusters to strata
      - Total # HHs expected
        - Median # HHs/unit
          - # clusters
    - Custom allocation of clusters to strata
      - Median # HHs/unit
        - # clusters

Sample size

- Total # HHs
  - Median # HHs/unit
    - # clusters

Spatial

- Oversample scale (area)
  - Boundaries of clusters: Oversample
GridSample output

After a GridSample job is complete, the user receives an email with a link to download the following as a 7zip file:

**Shapefile** and **KML** file of the selected cluster boundaries
- Shapefile can be opened with ArcGIS or QGIS
- KML file can be opened with Google Earth

**Excel** spreadsheet with shapefile attributes
- Use attribute values to calculate sample weights

**PDF** report summarising the input files and parameters
- Summary of survey parameters
- Map of each input dataset, and links to original input data sources

**Key to shapefile and Excel attributes**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>cl_id:</td>
<td>Unique numeric ID of cluster</td>
</tr>
<tr>
<td>str_id:</td>
<td>Unique number ID of strata</td>
</tr>
<tr>
<td>cl_type:</td>
<td>Whether cluster was selected during &quot;main&quot; PPS sampling or spatial &quot;oversample&quot;</td>
</tr>
<tr>
<td>s_cl_st:</td>
<td>Number of selected clusters in stratum</td>
</tr>
<tr>
<td>s_cl_tot:</td>
<td>Number of selected clusters in coverage area</td>
</tr>
<tr>
<td>s_cl_pop:</td>
<td>Estimated population in selected cluster</td>
</tr>
<tr>
<td>s_cl_hh:</td>
<td>Estimated households in selected cluster = s_cl_pop / st_hhsiz</td>
</tr>
<tr>
<td>urb_rur:</td>
<td>GHS-SMOD classification where 3=high dense urban, 2=low dense urban, 1=rural, 0=unsettled</td>
</tr>
<tr>
<td>u_n:</td>
<td>Number of sample frame units in the survey coverage area</td>
</tr>
<tr>
<td>u_medpop:</td>
<td>Median population per sample frame unit</td>
</tr>
<tr>
<td>st_hhsiz:</td>
<td>Average household size specified in this stratum</td>
</tr>
<tr>
<td>st_pop_n:</td>
<td>Estimated population in stratum</td>
</tr>
<tr>
<td>st_pop_p:</td>
<td>Percent of population in the coverage area falling in this stratum</td>
</tr>
<tr>
<td>st hh n:</td>
<td>Estimated households in stratum = st_pop_n / st_hhsiz</td>
</tr>
<tr>
<td>st hh p:</td>
<td>Percent of households in the coverage area falling in this stratum</td>
</tr>
<tr>
<td>st_name:</td>
<td>Unique alphanumeric name of stratum</td>
</tr>
<tr>
<td>ori_id</td>
<td>Unit (cluster) ID from the original sample frame</td>
</tr>
</tbody>
</table>
Implementing your survey

Approaches & tools

Four approaches and various tools are available to implement your gridded population survey.

Cell

Grid cells of any size can be generated and sampled with GridSample. Cells will always be a multiple of 100m X 100m grid cells. This approach results in a sample frame of units that are uniform in size, but not in population. This approach is generally not used on its own for household survey sampling, but might be useful for other applications.

gridEZ

A sample frame of multi-cell units with approximately the same population in each unit and maximum area can be generated with the gridEZ algorithm. Multi-cell units can be sampled and used directly, segmented manually along natural features, or segmented automatically along 100m X 100m cell boundaries for field work.

gridEZ original boundaries. This approach was used in Nepal by Elsey et al (2016). It was also used in Mozambique by World Vision International (2018) and is a featured case study.

gridEZ, segment along natural boundaries. This approach was used in Myanmar by Munoz and Langeraar (2013) and in Somalia by Pape and Wollberg (2019). It was also used in Nepal by Elsey et al (2018) and is a featured case study.

gridEZ, segment along sub-cell boundaries. This approach was used by the World Food Programme (2018) in DR Congo and is a featured case study. Sub-cells were randomized and fully enumerated until the target number of households per cluster was achieved, allowing for one field visit and calculation of sample probability weights as a one-stage segmented survey design.
Own EA.shp boundaries

GridSample allows users to define custom sample frame boundaries by uploading a zipped shapefile. This approach is suitable if census enumeration areas or other boundaries are available, and population estimates need to be updated. The output from GridSample and fieldwork approach is identical to a standard household survey.

Simple random sample (SRS) or non-probability

Some researchers, particularly in urban areas or other densely settled areas (e.g. IDP camps) survey a random sample of households. A simple random sample of buildings can be selected via GridSample by first sampling 100m X 100m WorldPop cells, then using a technique such as random point placement, or random selection of mini grid cells in a GIS to identify one building at random.

This approach can also be used to identify a random starting point for non-probability sampling techniques such as “random walk” or “spin-the-pen”. Galway et al. (2012) used this approach in Iraq.

Higher-tech tools

Navigation: In urban areas where most roads are mapped in OpenStreetMap, MAPS.ME allows for offline navigation over very long distances. In areas where OpenStreetMap data are sparse, navigation based on GPS coordinates and place names may be necessary.

Within clusters, an app such as OSMAnd app can be used for offline navigation based on preloaded MBTiles, displaying a blue dot at the device location.

Mapping: Applications such as Vespucci allow for tablet-based updates to OpenStreetMap in the field. While this is a seemingly efficient way to reduce steps during field work, we have found that field-based tablet editing of OpenStreetMap is time-intensive and frustrating for staff working on small screens often in adverse weather conditions.

In our experience, survey teams almost unanimously prefer printed geographic maps in the field with OpenStreetMap or satellite imagery as a base layer. This is because paper maps can be marked up quickly in the field with a pencil, and edits to
OpenStreetMap can be made quickly after field work in the comfort of an office. Furthermore, field teams that have compared paper and tablet mapping say that paper maps produced in ArcGIS or QGIS helped to facilitate positive conversations with residents about the survey while editing maps on tablets in the field fuelled suspicion.

Whichever mapping approach you choose, it is wise to update roads and building footprints in OpenStreetMap using iD editor or similar tool before visiting the field. GridSample cluster boundaries can be visualized on top of OpenStreetMap as a GPX trace file, keeping the dataset private. QGIS and a number of free apps can be used to transform the GridSample shapefile or KML file of cluster boundaries to a GPX file.

**Listing:** A number of apps are available to collect household listing data. Many of these apps, including OpenMapKit, GeoODK, and KoBoToolbox, also allow for collection of spatial data. Other tablet-based apps include the World Bank’s Survey Solutions tool which enables monitoring of field workers.

**Questionnaire:** The same apps used for listing - OpenMapKit, GeoODK, Kobo Collect, Survey Solutions - can be used to administer questionnaires.

Tablet-based data collection requires a server, someone to design and configure the data collection form, and someone to set up, secure, and maintain all of the devices. All of the linked apps and programs are free, and most are open source.

The Surveys for Urban Equity guides for survey planners and field teams provide guidance to implement higher-tech field tools and methods for a gridded population survey.

**Lower-tech tools**

**Navigation:** If you are working in a context without power, in a team with limited technical skills, or your field staff face high security risks, you will likely opt to use lower-tech tools, and possibly avoiding tablets or GPS units in the field. Lower-tech navigation to clusters can be done with a travel map and asking for directions based on place names.

**Mapping:** Going lower-tech does not mean that field staff need to sacrifice geographically accurate maps. Two simple options
are available to produce field maps to navigate and update building footprints in the field.

In rural contexts, satellite imagery from Google Earth is generally a suitable base map. Simply double-click the KML file of cluster boundaries provided in the GridSample output to visualize cluster boundaries in Google Earth. Then zoom to each cluster and print, recording the Cluster ID on each map.

In dense urban contexts where buildings are attached, the OpenStreetMap base layer may be needed to distinguish buildings and walking paths. The Field Papers website can be used to generate a map with the OpenStreetMap or Bing imager base layer for each cluster.

Many current surveys require field staff to hand-sketch maps of roads, buildings, and points of interest on a blank piece of paper. This approach is decades old but does not result in a geographically accurate map, and it is time intensive for field staff.

**Listing and questionnaire**: Printed paper forms provide a low-tech solution to conduct the household listing and administer questionnaires.

The linked tools are free and open source.
**After your survey**

### Sample weights

Sample weights are necessary to make accurate estimates about the population from a survey with a complex design (e.g. cluster sampling). Sample weight calculations are described below, and an Excel template file is provided at GridSample.org/tutorial to calculate sample weights. Inputs come from three sources of information:

- **GridSample:** Cluster ID (cl_id), strata name (st_name), strata ID (str_id), number of households in strata (st.hh.n), and number of households per cluster (s_cl hh)
- **Mapping-listing:** Number of clusters visited, number of segments created in each cluster during pre-field and post-field (one-stage only) enumeration, number of households listed in each (segmented) cluster, and number of households selected for sampling in each (segmented) cluster (two-stage only)
- **Interview:** Number of responded households per cluster (household questionnaire), and number of non-responding individuals per cluster (individual questionnaire)

#### Household sample (design) weight.** The formulas use 2 indices: 1...k strata (or entire coverage area) and 1...i cluster. The household sample (design) weight – the probability that cluster i is selected – is given by:

\[
W_{hh,i} = \frac{G_k / g_{ik}}{n_k} \times \frac{M_{ik}}{m_{ik}} \times b_{ik}
\]

Where:

- \( n_k \) is the number of selected clusters in stratum k
- \( G_k \) is the estimated total population in stratum k from GridSample
- \( g_{ik} \) is the estimated population in cluster i in stratum k from GridSample
- \( m_{ik} \) is the number of households sampled in cluster i and stratum k during fieldwork
- \( M_{ik} \) is the number of total households enumerated in cluster i and stratum k during fieldwork
- \( b_{ik} \) is the number of segments (if segmentation was performed before and after enumeration, then \( b_{ik} = b_{ik,\text{before}} \times b_{ik,\text{after}} \))
**Household response weight.** Interviewers will list households and record household and individual response rates during fieldwork. After interviews are completed, calculate household response weight - the probability that cluster $i$ is found and sampled, and households are found and respond – is given by:

$$w_{hh.r} = \frac{n_k}{n_k^* \times \frac{m_k}{m_k^*}}$$

Where:

- $n_k$ is the number of selected clusters in stratum $k$
- $m_k$ is the number of households sampled in stratum $k$ during fieldwork
- $n_k^*$ is the number of found and sampled clusters in stratum $k$
- $m_k^*$ is the number of found and responded households in stratum $k$

**Household sample weight.** To calculate the raw household sample weight, multiply the sample design weight and household response weight like this:

$$w_{hh} = \frac{G_k / g_{ik}}{n_k} \times \frac{M_{ik}}{m_{ik}} \times \frac{n_k}{n_k^* \times \frac{m_k}{m_k^*}}$$

Note, in one-stage samples, $\frac{M_{ik}}{m_{ik}}$ is equal to 1, and in two-stage samples, $b_{ik}$ is usually equal to 1.

**Individual sample weight.** The individual sample weight includes four additional terms to account for the sampling of one respondent among all eligible respondents in the household, and the response rate of those respondents. The individual sample weight is given by:

$$w_{ind.s} = \frac{G_k / g_{ik}}{n_k} \times \frac{M_{ik}}{m_{ik}} \times b_{ik} \times \frac{n_k}{n_k^*} \times \frac{m_k}{m_k^*} \times \frac{U_{ik}}{u_{ik}} \times \frac{u_k}{u_k^*}$$

Where:

- $U_k$ is the number of eligible individuals in cluster $i$ and stratum $k$
- $u_k$ is the number of sampled individuals in cluster $i$ and stratum $k$
- $u_k^*$ is the number of responded individuals in stratum $k$
**Normalizing sample weights.** Household surveys are often “normalized” or “standardized” such that the sum of weighted respondents equals the sum of respondents. Conceptually, each observation in the sample represents slightly more or slightly less than 1 household or person. To normalize sample weights, apply the below formulas:

\[
w_{hh,\text{norm}} = w_{hh} \times \frac{\sum (m_{ik^*})}{\sum (w_{hh} \times m_{ik^*})}
\]

Where:
- \(m_{ik^*}\) is the number of respondents with a completed interview in cluster \(i\) in stratum \(k\)
- \(w_{hh}\) is the raw household sample weight in cluster \(i\) in stratum \(k\)

\[
w_{\text{ind, norm}} = w_{\text{ind}} \times \frac{\sum (u_{ik^*})}{\sum (w_{\text{ind}} \times u_{ik^*})}
\]

Where:
- \(u_{ik^*}\) is the number of individuals with a completed interview in cluster \(i\) in stratum \(k\)
- \(w_{\text{ind}}\) is the raw individual sample weight in cluster \(i\) in stratum \(k\)

**Calculating sample weights.** Calculate sample weights in the provided Excel template, or in a statistical software programme such as SPSS or Stata. The provided Excel template is available at GridSample.org/tutorial.

To calculate household sample weights in the template, copy corresponding GridSample values into the grey columns of the “hh_survey_weights” tab. Strata and cluster values are included in an Excel file along with other downloaded outputs from GridSample. After performing the household listing and interviews, enter field-generated information into the template’s orange columns.
The template’s blue columns with populate; the dark blue columns show the household sample weight and normalized household sample weight for each cluster. Analysts generally prefer to use the normalized weights, outlined in red in the template.

<table>
<thead>
<tr>
<th>m_k</th>
<th>m_k*</th>
<th>w_hh</th>
<th>w_hh</th>
<th>m_k</th>
<th>m_k*</th>
<th>w_hh × m_k</th>
<th>w_hh norm</th>
<th>w_hh norm × m_k*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1143.729114</td>
<td>1083.729114</td>
<td>192.172</td>
<td>202.812</td>
<td>1683.729</td>
<td>1471.225</td>
<td>90082.787</td>
<td>2.440</td>
<td>17.699</td>
</tr>
<tr>
<td>1143.729114</td>
<td>1083.729114</td>
<td>172.548</td>
<td>182.101</td>
<td>1083.729</td>
<td>1491.936</td>
<td>90082.787</td>
<td>2.191</td>
<td>17.949</td>
</tr>
<tr>
<td>1143.729114</td>
<td>1083.729114</td>
<td>322.610</td>
<td>340.471</td>
<td>1083.729</td>
<td>1333.566</td>
<td>90082.787</td>
<td>4.096</td>
<td>16.043</td>
</tr>
<tr>
<td>1143.729114</td>
<td>1083.729114</td>
<td>57.595</td>
<td>60.784</td>
<td>1083.729</td>
<td>1613.253</td>
<td>90082.787</td>
<td>0.731</td>
<td>19.468</td>
</tr>
<tr>
<td>1143.729114</td>
<td>1083.729114</td>
<td>46.289</td>
<td>48.852</td>
<td>1083.729</td>
<td>1625.185</td>
<td>90082.787</td>
<td>0.598</td>
<td>19.532</td>
</tr>
<tr>
<td>1143.729114</td>
<td>1083.729114</td>
<td>62.032</td>
<td>65.406</td>
<td>1083.729</td>
<td>1571.189</td>
<td>90082.787</td>
<td>0.788</td>
<td>18.962</td>
</tr>
<tr>
<td>1143.729114</td>
<td>1083.729114</td>
<td>43.383</td>
<td>45.784</td>
<td>1083.729</td>
<td>1628.252</td>
<td>90082.787</td>
<td>0.551</td>
<td>19.588</td>
</tr>
<tr>
<td>1143.729114</td>
<td>1083.729114</td>
<td>425.366</td>
<td>446.916</td>
<td>1083.729</td>
<td>1225.121</td>
<td>90082.787</td>
<td>5.401</td>
<td>14.799</td>
</tr>
<tr>
<td>1143.729114</td>
<td>1083.729114</td>
<td>44.204</td>
<td>46.652</td>
<td>1083.729</td>
<td>1627.385</td>
<td>90082.787</td>
<td>0.561</td>
<td>19.578</td>
</tr>
<tr>
<td>1143.729114</td>
<td>1083.729114</td>
<td>267.786</td>
<td>287.612</td>
<td>1083.729</td>
<td>1391.425</td>
<td>90082.787</td>
<td>3.400</td>
<td>16.739</td>
</tr>
</tbody>
</table>

If individuals were listed, sampled, and interviewed within each selected household, then use the “ind_survey_weights” tab in the Excel template to calculate individual sample weights. Household sample weights will be pre-populated; simply enter field-generated information about individuals in the template’s orange columns.
Data cleaning & analysis

Household listing data collected via paper forms should be entered into Excel or OpenOffice Calc. Counts from the household listing and household response rates are needed for sample weight calculations.

Questionnaire data collected via paper forms should be double entered and cleaned using a software such as CSPro.

Household listing and questionnaire data collected via tablet should be reviewed and quality checked throughout the data collection process. Data should then be imported into a statistical software programme, such as SPSS or Stata, for further processing.

All summary statistics generated from complex household surveys should account for unequal probability of selection by applying sample weights, clustering of observations, and if applicable, stratification. Here are several resources to conduct survey data analysis in SPSS, Stata, and other statistical software programmes:

- Population Survey Analysis
- UNC - Carolina Population Center
- UCLA Institute for Digital Research & Education
Practice exercises

Exercise 1: 2014-15 Rwanda DHS

The 2014-15 Rwanda Demographic and Health Survey interviewed all women age 15-49 in selected households about their own health and wellbeing and the health of their children. The final report describes the survey design on pages 7 and 8 as follows:

The 2014-15 RDHS followed a two-stage sample design and was intended to allow estimates of key indicators at the national level as well as for urban and rural areas, five provinces, and each of Rwanda’s 30 districts (for some limited indicators). The first stage involved selecting sample points (clusters) consisting of EAs delineated for the 2012 RPHC. A total of 492 clusters were selected, 113 in urban areas and 379 in rural areas.

The second stage involved systematic sampling of households. A household listing operation was undertaken in all of the selected EAs from July 7 to September 6, 2014, and households to be included in the survey were randomly selected from these lists. Twenty-six households were selected from each sample point, for a total sample size of 12,792 households. However, during data collection, one of the households was found to actually be two households, which increased the total sample to 12,793. Because of the approximately equal sample sizes in each district, the sample is not self-weighting at the national level, and weighting factors have been added to the data file so that the results will be proportional at the national level.
These tables from the 2014-15 RDHS final report provide additional information to reproduce this survey design in GridSample.
How would you enter parameters into GridSample to reproduce the 2014-14 RDHS?

**Coverage**

Subnational option (pick one):
- None (National survey)
- Urban only
- Rural only
- Admin 1 area(s):
- Admin 2 area(s):
- Admin 3 area(s):
- Admin 4 area(s):

**Frame**

WorldPop-Global Dataset (year):

Gridded multi-cell cluster with gridEZ algorithm (select one):
- Small (target 75 people, max areas 1km X 1km)
- Medium (target 500 people, max area 3km X 3km)
- Large (target 1200 people, max area 5km X 5km)

**Design**

Stages (select one):
- One-stage cluster sample
- Two-stage cluster sample

Stratification (select one):
- No
- Yes

Spatial oversample (select one):
- No
- Yes

**Strata**

Admin area boundaries (select one):
- Admin 1 (e.g. Umujyi wa Kigali, Amajyaruguru)
- Admin 2 (e.g. Musanze, Rulindo)
- Admin 3 (e.g. Busogo, Cyuve, Gacaca)
- Admin 4 (e.g. Gisesero, Kavumu, Nyagisozi)

**Target**

Target Population Label:

Average number of target population members per household: 

Average household size:

**Sample size**

Allocation of clusters to strata (select one):
- Equal
- Proportional
- Custom (describe )

Total number of households sampled:

Number of households sampled per urban cluster:

Number of households sampled per rural cluster:
Exercise 2: 2017 Kathmandu SUE

The 2017 Surveys for Urban Equity in Kathmandu Valley aimed to overcome several challenges that lead to exclusion of the urban poorest in standard household surveys as described on page 2 of the SUE survey protocol.

First, census data, which is used to select first-stage samples, is often outdated and undercounts informally settled households.

Second, by design, surveys typically exclude the homeless and institutional populations. Use of two-stage cluster sampling methods requires two visits to households over several months or years, resulting in underlisting or higher non-response by mobile and fragile households.

Third, underlisting and undersampling of poorer households can occur if standardised, detailed protocols are not used by enumerators to interact with residents during the household listing process. For example, multihousehold dwellings will be underlisted if the enumerator assumes one dwelling to be occupied by one household or poorer members of households, such as guards and servants may be excluded.

Furthermore, periurban communities frequently home to urban migrants and slum areas, maybe classified as rural.

The protocol includes the following figures demonstrating that population in the Kathmandu Valley extend well beyond the official city administrative boundaries, and are located within the “high-dense urban” area defined by GHS-SMOD. Thus GHS-SMOD is used to define the Kathmandu Valley survey coverage boundary.
The survey protocol goes on to describe the survey design as follows on page 4:

To compare the effectiveness of one-stage sampling compared with two-stage sampling, in the Kathmandu survey we will randomly allocate half of the clusters to each approach. [For the purposes of this exercise, let us assume that one-stage sampling will be conducted in all clusters.]

One-stage sampling of approximately 20 households in each sampling area will be facilitated by the use of WorldPop 100 m x 100m grid cells rather than much larger census enumeration areas as the sampling frame.

We will interview adults, 18 years and above…

We will aim for a sample size of 1200 in the Kathmandu survey…. This assumes… one eligible individual per household.

This sample population will be distributed across 60 clusters...

Additional information about average household size is needed to replicate this survey design in GridSample. For this information, we use the 2016 Nepal Demographic and Health Survey final report.

<table>
<thead>
<tr>
<th>Table 2.10 Household composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent distribution of households by sex of head of household and by household size, mean size of household, and percentage of households with orphans and foster children under age 18, according to residence, Nepal DHS 2016</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Urban</th>
<th>Rural</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household headship</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>68.3</td>
<td>69.3</td>
<td>68.7</td>
</tr>
<tr>
<td>Female</td>
<td>31.7</td>
<td>30.7</td>
<td>31.3</td>
</tr>
<tr>
<td>Total</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Number of usual members</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>6.8</td>
<td>5.9</td>
<td>6.4</td>
</tr>
<tr>
<td>2</td>
<td>14.9</td>
<td>14.9</td>
<td>14.9</td>
</tr>
<tr>
<td>3</td>
<td>21.0</td>
<td>16.0</td>
<td>19.1</td>
</tr>
<tr>
<td>4</td>
<td>21.6</td>
<td>19.2</td>
<td>20.7</td>
</tr>
<tr>
<td>5</td>
<td>15.2</td>
<td>17.0</td>
<td>15.9</td>
</tr>
<tr>
<td>6</td>
<td>9.3</td>
<td>11.7</td>
<td>10.2</td>
</tr>
<tr>
<td>7</td>
<td>5.5</td>
<td>6.5</td>
<td>5.9</td>
</tr>
<tr>
<td>8</td>
<td>2.4</td>
<td>3.8</td>
<td>3.0</td>
</tr>
<tr>
<td>9+</td>
<td>3.2</td>
<td>4.9</td>
<td>3.9</td>
</tr>
<tr>
<td>Total</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Mean size of households</td>
<td>4.1</td>
<td>4.4</td>
<td>4.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Percentage of households with orphans and foster children under age 18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double orphans</td>
</tr>
<tr>
<td>Single orphans</td>
</tr>
<tr>
<td>Foster children</td>
</tr>
<tr>
<td>Foster and/or orphan children</td>
</tr>
<tr>
<td>Number of households</td>
</tr>
</tbody>
</table>

Note: Table is based on de jure household members, i.e., usual residents.  
1 Includes children with one dead parent and an unknown survival status of the other parent  
2 Foster children are those under age 18 living in households with neither their mother nor their father present, and the mother and/or the father are alive.
Although GridSample does not provide individual metropolitan boundaries for specific cities, it does provide boundaries of all urban areas (defined as GHS-SMOD=high dense urban) and multiple levels of sub-national administrative units. A clever solution is possible in GridSample using the provided parameter options to limit the coverage area to urban area(s) in a sub-region of the country. Consider the below map, and see if you can figure it out.
How would you enter parameters into GridSample to reproduce the 2017 Kathmandu SUE?

**Coverage**
- Subnational option (pick one):
  - None (National survey)
  - Urban only
  - Rural only
  - Admin 1 area(s):
  - Admin 2 area(s):
  - Admin 3 area(s):
  - Admin 4 area(s):

**Frame**
- WorldPop-Global Dataset (year):
- Gridded multi-cell cluster with gridEZ algorithm (select one):
  - Small (target 75 people, max areas 1km X 1km)
  - Medium (target 500 people, max area 3km X 3km)
  - Large (target 1200 people, max area 5km X 5km)

**Design**
- Stages (select one):
  - One-stage cluster sample
  - Two-stage cluster sample
- Stratification (select one):
  - No
  - Yes
- Spatial oversample (select one):
  - No
  - Yes

**Strata**
- Admin area boundaries (select one):
  - Admin 1 (e.g. Central, East)
  - Admin 2 (e.g. Koshi, Mechi)
  - Admin 3 (e.g. Bhojpur, Dhankuta, Morang)
  - Admin 4 (e.g. Aamtep, Annapurna, Baikunthe)

**Target**
- Target Population Label:
- Average number of target population members per household:
- Average household size:

**Sample size**
- Allocation of clusters to strata (select one):
  - Equal
  - Proportional
  - Custom
- Number of clusters per stratum:

<table>
<thead>
<tr>
<th>Stratum Name</th>
<th>Number of clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Additional resources

Gridded population sampling manual


Gridded population survey implementations


Exercise answers

Coverage
Subnational option (pick one):
✓ None (National survey)
☐ Urban only
☐ Rural only
☐ Admin 1 area(s):
☐ Admin 2 area(s):
☐ Admin 3 area(s):
☐ Admin 4 area(s):

Frame
WorldPop-Global Dataset (year): 2019
Gridded multi-cell cluster with gridEZ algorithm (select one):
☐ Small (target 75 people, max areas 1km X 1km)
✓ Medium (target 500 people, max area 3km X 3km)
☐ Large (target 1200 people, max area 5km X 5km)

Design
Stages (select one):
☐ One-stage cluster sample
✓ Two-stage cluster sample

Stratification (select one):
☐ No
✓ Yes

Spatial oversample (select one):
✓ No
☐ Yes

Strata
Admin area boundaries (select one):
☐ Admin 1 (e.g. Umujyi wa Kigali, Amajyaruguru)
✓ Admin 2 (e.g. Musanze, Rulindo)
☐ Admin 3 (e.g. Busogo, Cyuve, Gacaca)
☐ Admin 4 (e.g. Gisesero, Kavumu, Nyagisozi)

Target
Target Population Label: Women 15-49
Average number of target population members per household: 1.1
Average household size: 4.3
146 × 4.3 = 628 pop per EA

Sample size
Allocation of clusters to strata (select one):
☐ Equal
☐ Proportional
✓ Custom (describe: 20 clusters in each of 3 Kigali districts, 16 clusters in each of other districts)

Total number of households sampled: 12,792
Number of households sampled per urban cluster: 26
Number of households sampled per rural cluster: 26
Coverage
Subnational option (pick one):
- None (National survey)
- Urban only
- Rural only
- Admin 1 area(s):____________________________
- Admin 2 area(s):Bagmati________________________
- Admin 3 area(s):____________________________
- Admin 4 area(s):____________________________

Frame
WorldPop-Global Dataset (year):__2019________________
Gridded multi-cell cluster with gridEZ algorithm (select one):
- Small (target 75 people, max areas 1km X 1km)
- Medium (target 500 people, max area 3km X 3km)
- Large (target 1200 people, max area 5km X 5km)

Design
Stages (select one):
- One-stage cluster sample
- Two-stage cluster sample
Stratification (select one):
- No
- Yes
Spatial oversample (select one):
- No
- Yes

Strata
Admin area boundaries (select one):
- Urban / rural
- Admin 1 (e.g. Central, East)
- Admin 2 (e.g. Koshi, Mechi)
- Admin 3 (e.g. Bhojpur, Dhankuta, Morang)
- Admin 4 (e.g. Aamtep, Annapurna, Baikunthe)

Target
Target Population Label:__Adults 18+_____________________
Average number of target population members per household: __1________
Average household size: __4.1____________________

Sample size
Allocation of clusters to strata (select one):
- Equal
- Proportional
- Custom
Number of clusters per stratum:

<table>
<thead>
<tr>
<th>Stratum Name</th>
<th>Number of clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban</td>
<td>60</td>
</tr>
<tr>
<td>Rural</td>
<td>0</td>
</tr>
</tbody>
</table>