SAMPLE DESIGN OPTIMIZER

EPIC IMMUNIZATION COSTING | IMMUNIZATION ECONOMICS.ORG

Sample Design Optimizer (SDO) Instrument: User Guide

V.2 – March 2019

Acknowledgements

This instrument was developed as part of the EPIC Project and ProVac Initiative, both of which were supported by grants from the Bill & Melinda Gates Foundation.

Introduction

This document is intended to provide guidance on how to use the Sample Design Optimizer (SDO) software. The SDO was designed to assist research teams develop cost-effective and efficient hierarchical clustered sample designs. Research teams will be able to use prior information to maximize expected precision within a data collection budget constraint. The primary application of this instrument is to design nationally representative data collections for immunization costing studies that draw from a sample of health facilities, district health offices or other hierarchical jurisdiction levels within a country.

Context

In a recent wave of studies examining the cost of delivering national immunization programs as part of the EPIC Project and ProVac Initiative's COSTVAC studies, which emphasized using randomized selection of study sites to ensure representativeness, local teams sought technical support with sample design. Frequently asked questions related to sample design included:

- *In a hierarchical design, how do we trade-off between the number of districts sampled and the number of facilities per district sampled?*
- *What is study's expected precision?*
- *How much will uncertainty be reduced by adding more units to our sample?*
- *What is an appropriate procedure for substituting units during fieldwork, if a sampled unit is determined to be non-reachable, non-functional, or non-responsive?*

These questions motivated the effort to design an instrument that could help teams visualize their sampling frame (the population of sites from which they could select a sample) and estimate the expected precision of results collected under different sample designs. While the SDO is not a replacement for the input of a statistician with expertise in sample design, it may help build intuition and lead to improved designs in situations where technical advice of a statistician is not available.

The SDO assumes that the user is conducting a study of an immunization program and that the outcome of interest is delivery cost. In theory, the instrument could be used in other contexts in which the outcome of interest is something else. But the user guide will assume the outcome of interest is immunization delivery cost.

Data Requirements

The SDO requires three main inputs: a dataset describing the Sample Frame, data collection unit costs and budget, and "prior" estimates of the outcome of interest (usually immunization program delivery cost).

Sample Frame

The SDO requires that users supply a data file describing the Sample Frame. For each organizational level that data will be collected from, a complete list of sites eligible for selection must be supplied. The minimum information required regarding each site is a unique identifier, a prior estimate of immunization delivery cost, and the 'parent' site at one level higher to which it belongs (for example a facility site record must indicate what district it belongs to). If the user wishes to employ stratification in

the sample design, then the stratum category of the site must be indicated. If the user wishes to employ a design in which the probability of a site being selected is proportional to its size, then a "size" variable must also be included in the Sample Frame data set.

Data Collection Unit Cost & Budget

In general, as samples get larger, precision of the resulting estimates improves. However, collecting data is costly, and these costs limit sample size. The SDO will help users find sample designs that maximize precision for a given data collection budget. In order to do this, the SDO requires that the user inputs estimates of data collection unit costs and a budget limit.

Prior Estimates of Outcome of Interest

Any sample size or precision calculation carried out before a study is conducted is necessarily based on assumptions about what the findings of the study will be. To estimate the expected precision of a particular design, the SDO relies on "prior" estimates of the main outcome of interest: immunization delivery cost incurred at health facilities, and higher-level units. These cost estimates can be supplied by the user of the instrument, if past costing studies have been performed. Alternatively, the cost estimates can generated based on findings in past EPIC studies. EPIC studies, and others, have consistently found a strong inverse relationship between facility volume (doses delivered) and delivery cost per dose. Other predictors that are observable prior to conducting a costing study include facility type, urbanicity, ownership type.

Terminology

A population of sites from which the sites are selected for the sample is known as the **sampling frame**. An algorithm (set of rules) for selecting a sample is a **sample design**. Sample design may vary according to the number of units they select, the stratification of units, and the probability of selection assigned to units. A **sample** is a list of selected units that results from applying a sample design to a sampling frame. This is the list that data collectors would take into the field.

Data collection unit cost refers to the cost of collecting data a selected unit. In the SDO, this cost is expressed as a money cost. But it is possible that time is the more relevant constraint limiting the scale of data collection. The SDO cannot currently calculate time requirements of data collection. However, users could use the instrument to find cost-feasible designs, and then manually estimate their time requirements to determine overall feasibility.

A "**size**" variable refers to a variable that the SDO will use in testing sample designs in which units have a probability of selection that is proportional to their size. Common size variables include: under-1 population size of a district, number of vaccine doses delivered in a facility in a prior year, number of outpatient visits at a facility in a prior year.

A "**stratification**" variable refers to a variable used to categories units in the sampling frame. For example health facilities might be stratified according to facility types (clinics, health centers, hospitals) or according to urbanicity (urban, peri-urban, rural), or some other feature thought to be associated with immunization delivery cost.

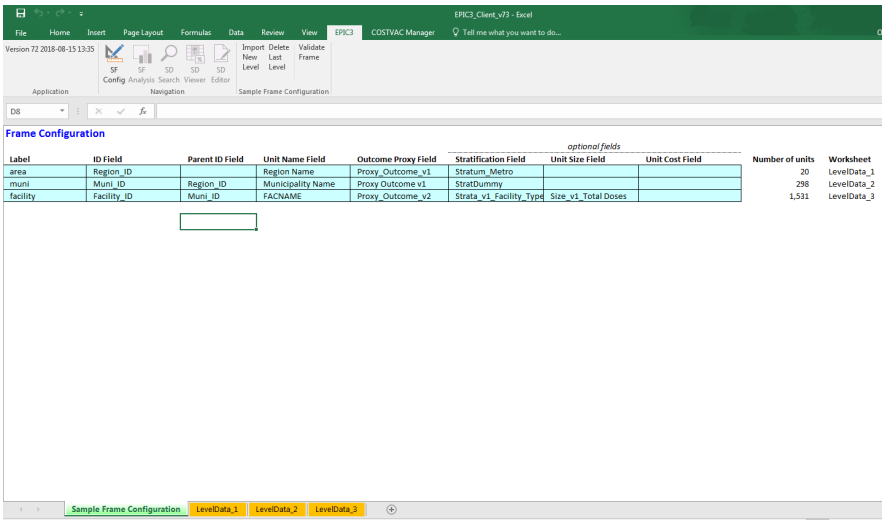
Process

The user interface for this application is an Excel workbook application; the back-end calculations for the optimizer run in a Java-based program.

The SDO workflow is fairly linear. First, a sampling frame dataset is uploaded into the instrument. Then the user can review a report describing the structure of the sampling frame. In the next step, the user can specify any design constraints it wants to enforce on the instrument. For example, users may wish to require that a minimum or maximum number of units be selected from a certain level. Then the SDO simulates millions of instances of sample selection from the range of possible designs and calculates the expected data collection cost and the expected precision in the outcome variable of interest. The user can observe these results in the Results Explorer and can compare different designs. Finally, if the user selects a specific design, the SDO will draw a random sample from the Sampling Frame using that design. The user can then proceed with data collection from that sample.

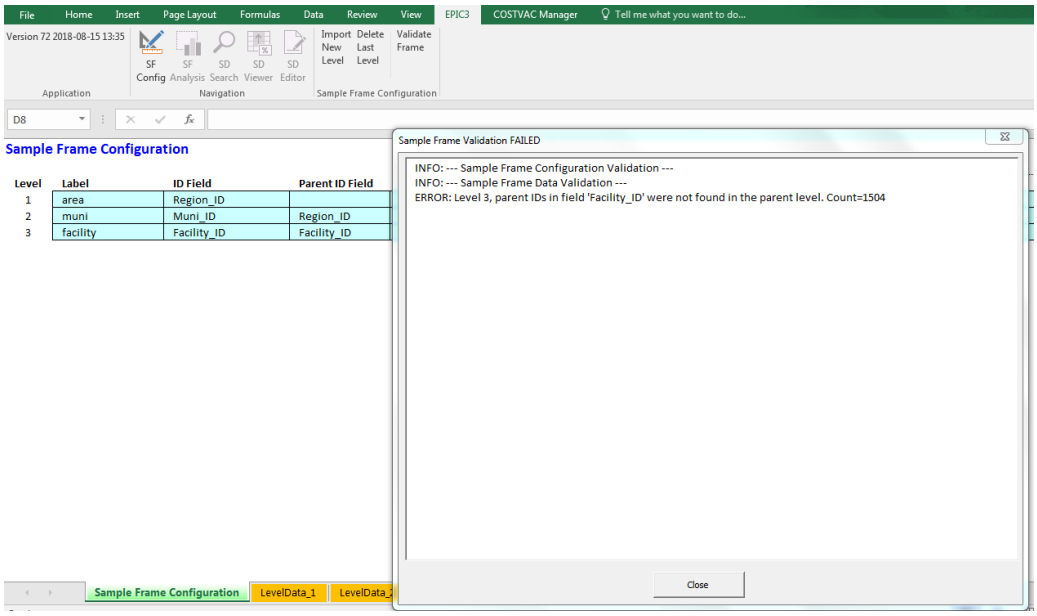
What follows is a step-by-step instructions for how to carry out this process and produce a sample that fits the project constraints and prioritizations of the analyst.

1. **Frame Data Import**
    - **STEP 1.** [Outside of SDO] Compile data on the sampling frame and organize it into the prescribed format (e.g. CSV or Excel workbook). This format can be observed in the demo files for the instrument. In short, the first row is reserved for headings indicating variable names, and the variables that must be included for each entry (row) are the unique ID, the Parent ID, the Unit Name, and the Outcome Proxy field. The outcome proxy will contain the prior estimate of the outcome of interest—usually immunization delivery costs at that unit. There will also be up to three optional fields for Stratification, Unit Size and Unit Cost. Unit Cost is an optional field for users wishing to supply a unit-specific *data collection* cost. In most cases, this will not be used.
    - **STEP 2. User loads data representing the sampling frame into SDO.** To start, open the SDO, “EPIC3_Client_v73” worksheet. When you initially open the application the only active navigation function should be “SF Config” (these functions will be validated once you have a validated sample frame).

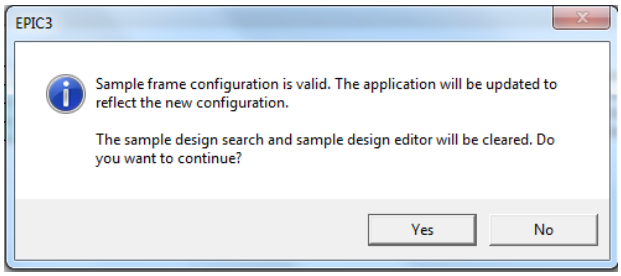


In the demo version, the Honduras dataset is already preloaded into the application. To remove an existing data level, go to “sample frame configuration” in the EPIC 3 ribbon and click “delete last level.” (If you do not see these functions, check the ribbon bar at the top of the window to make sure you are working within the “EPIC3” tab). To add a new data level click on “import new level.” This brings up a file chooser dialogue. Select the level data you would like to import. Hit open. It reads it in. Now the data is available as a level in your sampling frame. Import data levels in hierarchical order, starting with the highest jurisdictional level included in the study. For example, in the Honduras study, the CSV files in the project folder are organized as “L1_Region,” “L2_Municipality,” and “L3_Facility.” Importing a new data level will generate a new row in the sample configuration table.

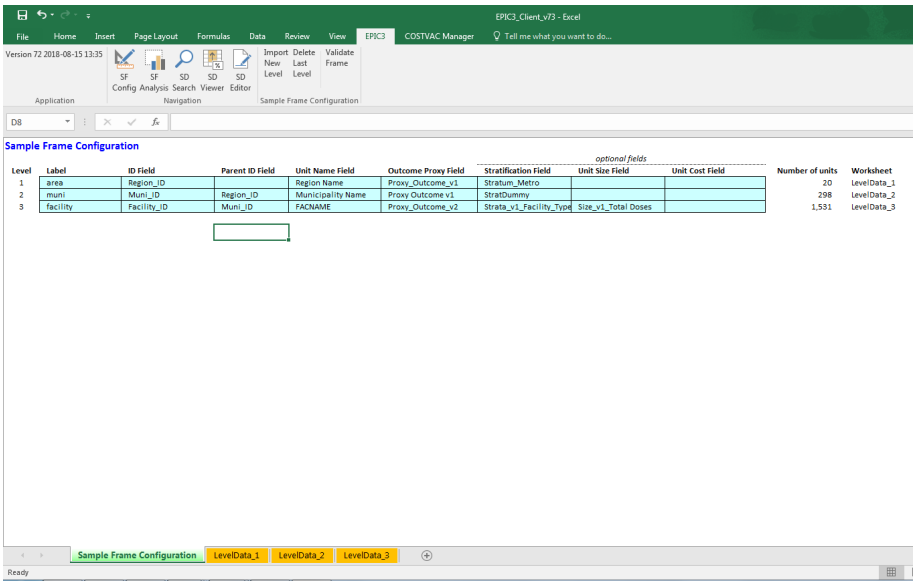
- **STEP 3**. **Fill in sample frame configuration table**. The sample frame configuration table fields should be populated with the corresponding field names from the imported data. The “ID Field” should be the name of the unique identifier variable for the jurisdiction level. The “Parent ID field” relates each sub-jurisdictional level to the level that precedes it in the hierarchy. In the Honduras example, the parent ID field for facility is the unique ID field for the municipality and the parent ID for the municipality is the unique ID for the region. Region does not have an entry for the parent ID field as that is the highest level of aggregation in the Honduras data frame. Moving across the table, the next column is the “Unit Name Field.” If you look back at your original imported dataset, this will be the actual name or abbreviation used to identify units within this level. The other required field is the “Outcome Proxy Field” which contains the prior estimate of immunization delivery cost. An error window pops up if you don’t select the correct field. There are restrictions on each of these fields and tests of data completeness.
- **STEP 4**. **Validate data frame**. Once you have imported your data and entered all required fields in the sample frame configuration table, the next step is to validate the sample frame. (Select: “Validate Frame”). If you specify the incorrect field name/column heading, and validate the date frame, an error message will appear, like the one below, with a log of the errors [In this example, I incorrectly specified the “Facility_ID” as the Parent ID field for Facility. The correct Parent ID field in this case is “Muni_ID”.]



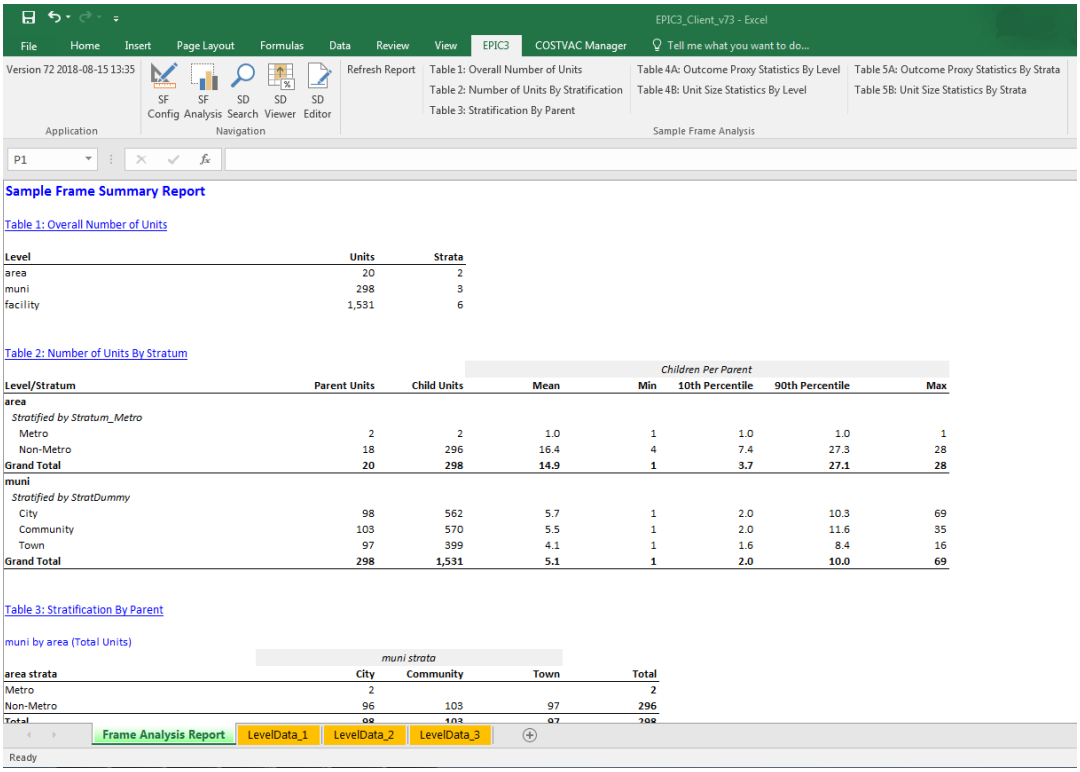
If the data frame is valid, you will receive the message in the window below. Because the SDO has a linear workflow, the downstream sample design search and sample design editor forms will be linked to the sample frame configuration. Therefore if you change the sample frame configuration, the downstream forms will be cleared. To retain past work, you can save multiple copies of the SDO workbook in different directories. When you are ready to save your sample frame configuration, click “Yes”.



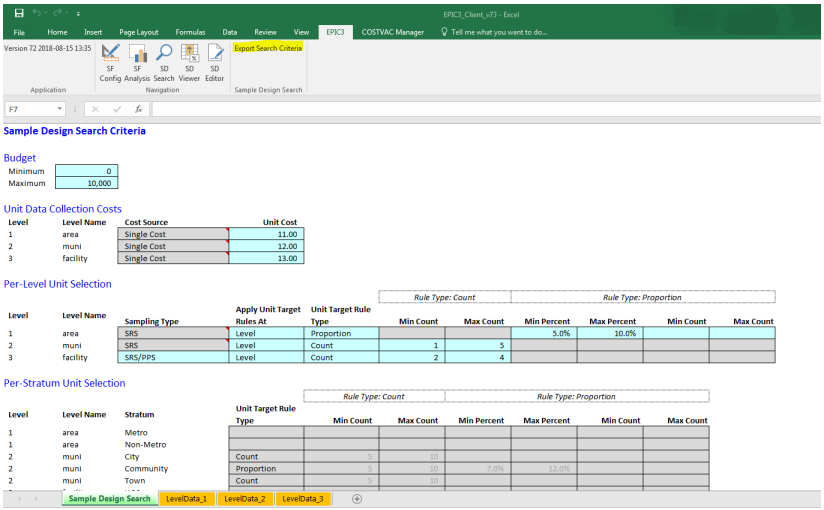
Once the data frame is validated, the other navigation functions become enabled.



- **Sample Frame Analysis:** The “Sample Frame Analysis” brings you to a page that provides a sample frame summary report**.** Anytime you make changes on the sample design configuration sheet you need to hit “refresh report” on the frame analysis report sheet (default: not current and reflections last sample frame configuration. The Sample Frame Analysis was programmed this way to allow the user to decide when to run. Depending on sample size, it can take a long time to generate a new report) .

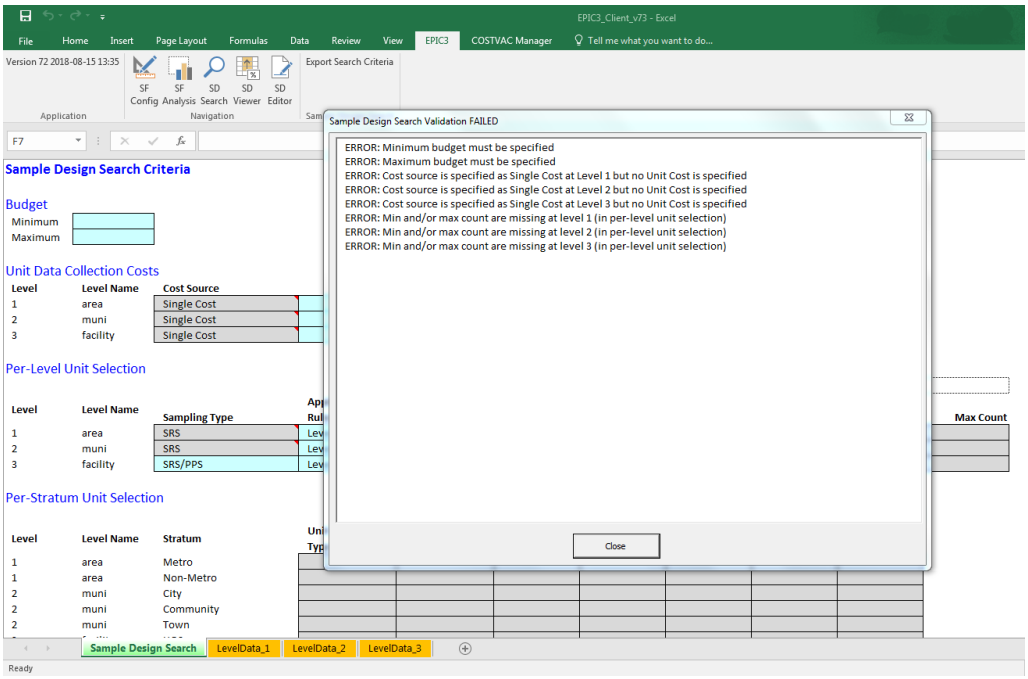


2. **Design Constraints**
   - **STEP 5**. In this step you will enter the input parameters in the "SD Search" that establish the constraints for the optimization function that will be performed by the Java Code. The next step is to set the constraints of your sample design. To do this go to the "SD Search" in the navigation menu (the frame data must be validated in order to advance to this next step). Scrolling down the "SD Search" window you will see a set of three tables that allow you to set constraints for the following criteria:
     1. Budget (min, max). This is where you indicate the total among of resources you have available for data collection. If desired, to approximate an unconstrained budget, you can set the maximum value very large.
     2. Unit cost data (data collection cost parameters: if you select "single cost", then you can enter a constant data collection cost per unit; if you select to use individual unit costs, then the SDO will look for this information in the Unit Cost field of the sample frame dataset that you imported in the earlier step.)
     3. Choice of simple random sampling (SRS) or probability proportional to size sampling (PPS). Application default is SRS if alternative isn't specified. To use PPS sampling, a "size" variable must be present in the sample frame dataset.
     4. Choice to use stratification of units at each level
     5. Lower or upper bound, or specific number of percentage of items to sample at each level.



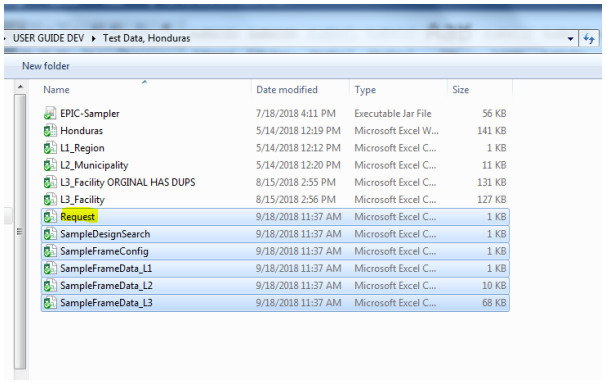
SDO validates that parameter set is consistent with the data in the frame files. E.g., if the parameter set says stratify on the variable 'City' then that variable must exist in the frame file.

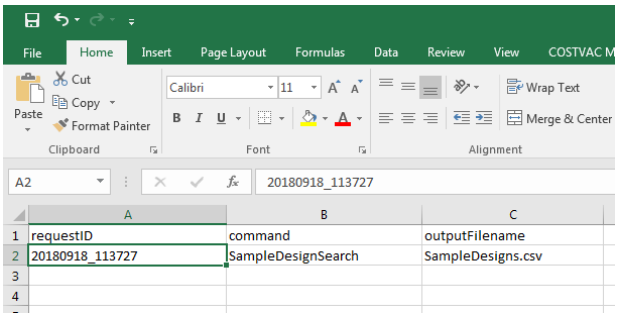
3. **Export Data to Optimizer**
   - **STEP 6.** Within "SD search," click the "export search criteria." If the search criteria you have entered does not validate then you will receive a dialogue box with an error log similar to the one displayed below.



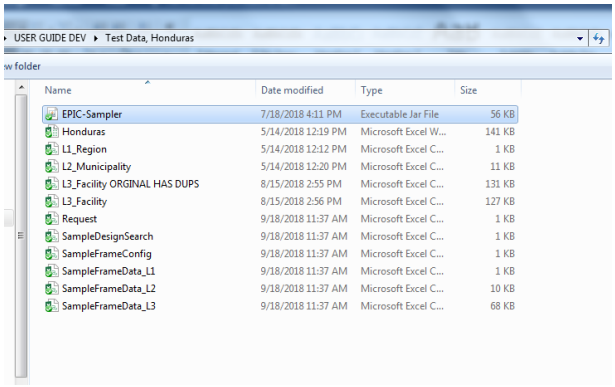
If the search criteria validates then you get a dialogue box, which requires you to establish the folder/location where you would like to export the data. Hit export.

Next go to the project folder you specified and find the 'Request' CSV. The request ID is simply a time stamp. It has a command that tells the sampler what to do. It has an output file name that tells that sample what to call the output file. There are five files that give the sampler what it needs in additional to the request itself.

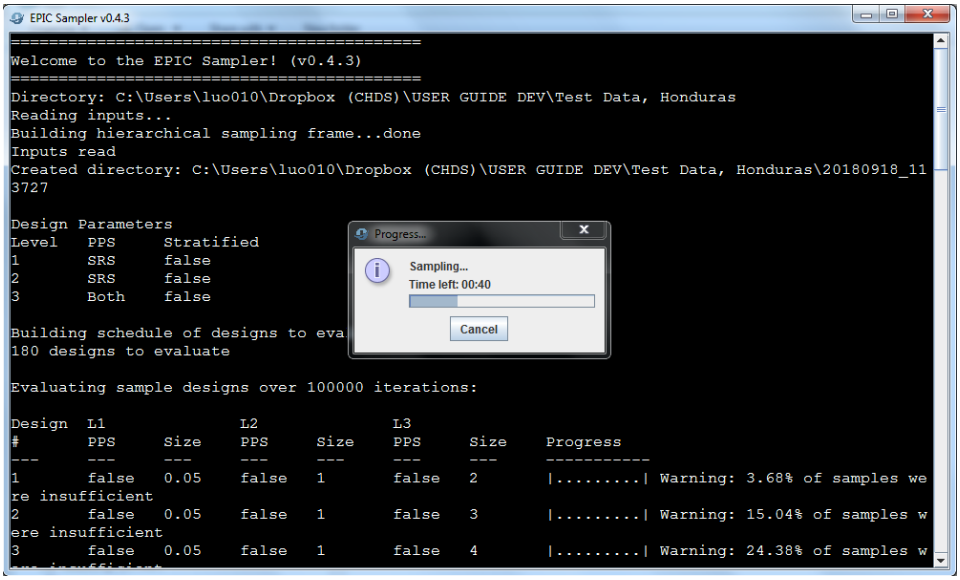




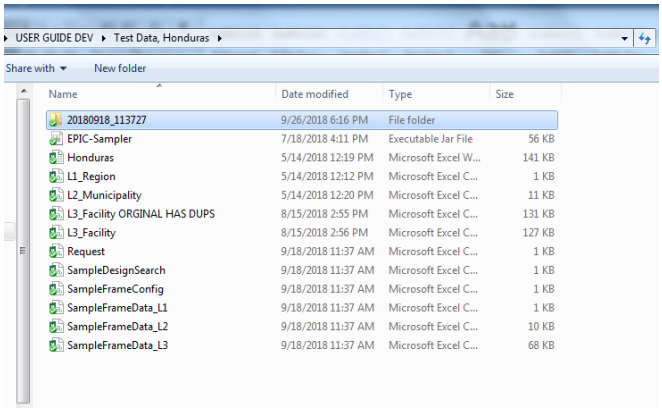
- **STEP 7.** Go to the “EPIC-Sampler”(this is a Jar file)[make sure this in your project folder where the parametrization for the search is contained]



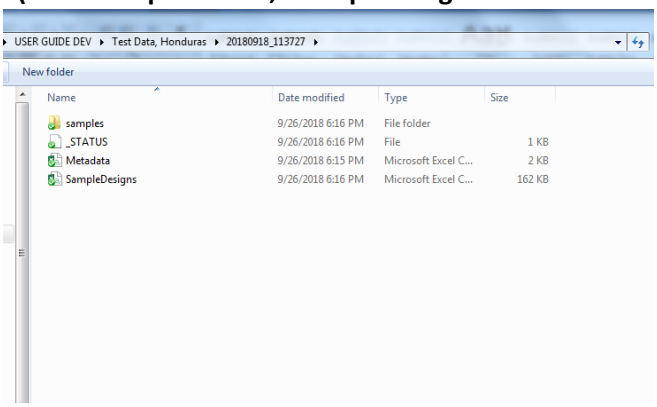
Select it and let the sampler run. Once it’s done, you can go ahead and close the sampler window.



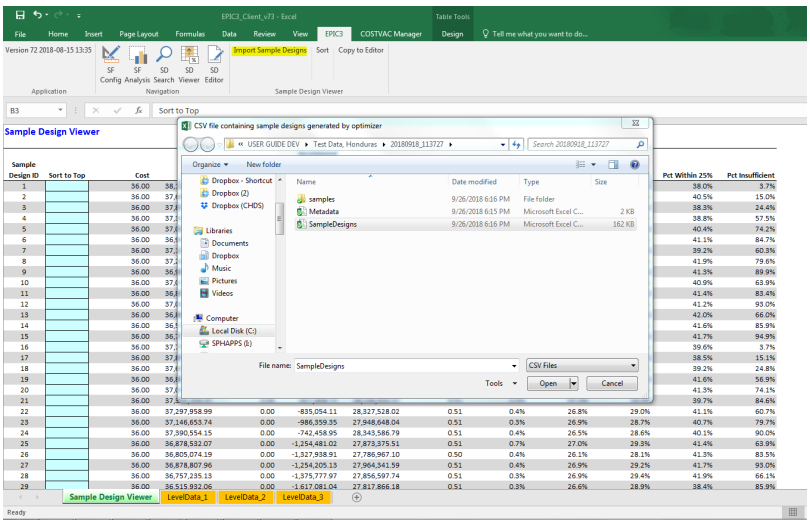
The results will appear in a new folder within you project folder. The name of the folder is the Request ID.



(Inside output folder; “SampleDesigns.csv” is the output file)



- **STEP 8. Import “SampleDesigns.csv” into the “SD viewer”** Back in the EPIC3 Excel application, you now want to go to the “SD viewer.” And import the SampleDesigns file that was generated by running the simulation program (EPIC-Sampler.jar). In the SD Viewer worksheet you will see the results for all the sample designs tested. You can sort them and filter them as you wish to find options that have the best precision without exceeding your data collection budget. The SD Viewer also reports other statistics for the design such as the percent of time the design could not meet a user-indicated constraint. For instance, if the user indicated that a minimum of 2 facilities must be sampled from each selected district, it may be that in some cases there are districts with only 1 facility present.
- If you wish to see the details of the sample design, simply select a cell in the row for your design of interest and click “Copy to Editor”. This will show you the sample design algorithm details in the SD Editor.
- Each Sample Design has a unique ID. If you wish to obtain a file with an actual sample drawn using that design, you can look in the simulation output folder for a folder called “samples” and find the file with the corresponding ID number.



If you have questions or issues to report, please email: epic@hsph.harvard.edu
and include "SDO" somewhere in the subject line