Kenya lactation rooms - interpreting the HFCs

For additional questions contact: flavia.ungarelli@studbocconi.it

The high frequency checks (HFCs) are updated every day during the interviewing period. Each day, a new folder is added to the main folder "HFC_lactationrooms_field". Each of these daily folders is identified by the date on which it was created and each one contains the following two excel files:

1. "hfc_output.xlsx"
2. "hfc_output_recent.xlsx"

This document proceeds, sheet by sheet, to explain the contents of these two excel files. If any of the sheets listed below are not present in the files it is because there is no issue to report. A more comprehensive set of checks will be monitored from the office, from which it will also be possible to investigate the flagged issues more in depth.

1 hfc_output.xlsx

1.1 id duplicates

This sheet lists all the surveys that report the same teacher ID, which should instead be unique. The column "key" refers to the code that is automatically generated by the survey platform, surveyCTO, when a survey is submitted. The sheet reports the duplicate teacher IDs under "unique_teacher_id", together with the enumerator name, the enumerator ID, the subcounty, the school code, the survey status (i.e. complete, incomplete, etc...) and the availability (whether the respondent was available to be surveyed or not). The list includes all respondents with whom the enumerators spoke multiple time. What is important is that there should not be more than one "Complete" survey with the same unique teacher ID.

1.2 other specify

This sheet lists all the text responses that are submitted when a respondent selects "other (specify)" to a multiple choice question that provides a list of options. The question text is listed under "parent label" and the response given is reported under "child value". Together with this information, the sheet reports the enumerator name, submission date and unique teacher id. This sheet provides information on the types of responses that are being specified, and if they make sense.

1.3 other specify (choices)

This sheets can be used to see the percentage of respondents who select "other(specify)" among the options provided. This is the percentage that is shown next to the value "-777" for each question. The "variable" column can be used to look for the text of the questions, by matching it to the "parent variable" which appears on the previous sheet (section 1.2).

1.4 form versions

This sheet lists the different form versions that have been submitted. The name of the survey version corresponds to the date on which it was created on the survey platform. The table lists

the count of forms that were submitted of each survey version, together with the date on which that version was first and last used. Forms are counted as "oudated" if they are employed after a new survey version has been created. Outdated forms arise when the survey is updated or corrected after the interviewing has already started and an enumerator does not update the version of the survey that he has downloaded.

1.5 duplicates

This sheet reports a list of observations that report the same survey key (skey), phone number or tsc number, which should instead be different for each respondent. The important thing is that, if the "unique_id_nodups" is different, then these variables are also different.

1.6 missing

This sheet reports the percentage of responses that should be present but are instead missing. Ideally, this should be 0% for all variables. The most important variables have a "yes" in the "priority" column. "# distinct" lists the number of distinct responses that have been provided by all respondents to that question. The variables ending in "_duration", "_endur" or "_startdur" refer to the time stamps that are generated by the survey platform automatically. If these are missing, it may be a sign that there where technical problems on the device used for that survey at that time.

1.7 weekly submissions

This sheet lists the number of surveys submitted each week. Alternatively, it may appear as "daily_submissions" during the first week of interviewing.

1.8 weekly submissions by enum

This sheet lists the same information as the previous sheet but divided according to the enumerator name. The last row lists the number of total submissions for that week.

1.9 enum_stats

This sheet reports the distribution of responses to some of the most important questions, by enumerator. For each enumerator, the first row reports the count of each answer; the second row reports the percentage of each response by enumerator (e.g. 50% of those interviewed by enumerator X replied that they are married); the third row reports the percentage reported by each enumerator by answer (e.g. 5% of those married were interviewed by enumerator X).

1.10 outliers

This sheet reports the list of outliers for all the numerical variables of the survey, together with the enumerator name, submission date, unique teacher ID and a label that explains what the variable refers to. The column "observed value" reports the response that was reported. "lower/upper fence" reports the minimum/maximum value that is flagged as "too low/high" (this corresponds to $Q1 - (1.5 * IQR)$ and $Q3 + (1.5 * IQR)$ with Q1 the 25^{th} percentile, Q3 the 75^{th} percentile and IQR the interquartile range). The variables that have no "label" are all linked to the length of the open-answer questions, these are not as much of a priority and they can be monitored more in depth from the office.

1.11 duration by enumerator

This sheet reports the number of "valid" surveys (i.e. distinct, "consented") that were submitted by each enumerator, together with the minimum, mean and maximum duration of these surveys. The last column "sd" reports the standard deviation of the duration.

1.12 constraints

This sheet lists the responses that defy the constraints that we would like to be met and that are explained in the "label" column, together with the enumerator name, school code, subcounty, enumerator ID, unique teacher ID and submission date. More stringent constraints are marked as "hard" constraints, while the "soft" constraint violations are less concerning. In general, "value" reports the actual numerical value that was reported in the survey by the respondent. Some of the variables that will appear refer to the length of the responses to open-answer questions, to check that they are not too short. Any ambiguous pattern can be further investigated from the office, by reading the full answers.

1.13 logic

This sheet lists the logical inconsistencies that were recorded, together with the enumerator name, enumerator ID, subcounty, school code, survey status, availability status, submission date, unique teacher ID, a "label" that explains the logical inconsistency that is being flagged.

1.14 respondent tracking by school

This sheet reports, for each school, the count and percentage of completed (1) male teacher (TC) surveys and (2) female teacher (TC) surveys and (3) head teacher (HT) surveys. (E.g. 100 means all surveys of that category have been completed for that school).

1.15 respondent tracking subcounty

This sheet reports, for each subcounty, the count and percentage of completed (1) male teacher (TC) surveys and (2) female teacher (TC) surveys and (3) head teacher (HT) surveys. Additionally, it reports the number of distinct schools that were visited in that subcounty.

2 hfc_output_recent.xlsx

This excel file considers only the observations that were recorded in the previous day of interviewing and repeats the following checks:

- id duplicates (see section 1.1)
- other specify (see section 1.2)
- other specify (choices) (see section 1.3)
- form versions (see section 1.4)
- duplicates (see section 1.5)
- missing (see section 1.6)
- enum_stats (see section 1.9)
- outliers (see section 1.10)
- constraints (see section 1.12)
- logic (see section 1.13)