CRA Weights Models Revised

Andres Ochoa Toasa- UNICEF Evaluation Office

October 17, 2020 - Updated on March 3, 2021

Overview

The following document details the standard weights model for UNICEF's Community Rapid Assessment survey, as implemented in 12 Eastern and Southern Africa (ESARO) and Southern Asia (ROSA) countries. This model aims to be a model of relatively ease of deployment to facilitate better representativeness of the surveyed samples of the study.

Given the rapid deployment nature of the exercise the model does not aim to be as comprehensive and thorough as more elaborate weights models used for forecasting; such limitations are noted at the end of this document. However, the model is still grounded on a solid sampling approximation to provide more accurate sample distibutions across all areas surveyed.

To be deployed, this model can be applied where the following conditions have been met.

a) The CRA follows a probabilistic sampling approach - For example, if it follows an IVR based design.
b) Total number of all calls made are available by the surveying company.
c) There is a comparable recent survey/census from which to extract demographic distributions to callibrate the weights to demographic charactheristics of the population - No more than 5 years before deployment of CRA.
d) It excludes all under 18 years respondents from the sample.

For the current model, references are included as hiperlynks where it corresponds. Each section is accompanied by brief description where needed and the applicable code to apply the weights.

For all the subsequent coding and operations in R the following sources where used to deploy the methodological choices on the dataset through the correct R coding procedure:

- Richard Valliant, Jill A. Dever, and Frauke Kreuter, **Practical Tools for Designing and Weighting Survey Samples**, Springer, Washington D.C. 2018
- Richard Valliant and Jill A. Dever, **Survey Weights: A Step-by-Step Guide to Calculation**, Stata Press, College Station, 2018

Other auxiliary coding sources used for extracting weights, control checks, advance weights analysis, and Latex are here hyperlinked if need to be accessed.

Model

The following model was constructed in 4 stages and includes a 5 methodological limitations point:

1. Base Weights Estimation
2. Post-Stratification Callibrating
3. Trimming of the weights
4. Design Effect
5. Limitations (Few Methodological Issues to Consider)

1. Base Weights

To develop the base weight formula we followed the model used by the Pew Research Center.The original formula is as follows:

$$BWt = \frac{1}{\left(\frac{S_{LL}}{U_{LL}} * \frac{LL}{AD}\right) + \left(\frac{S_{CP}}{U_{CP}} * CP\right) - \left(\frac{S_{LL}}{U_{LL}} * \frac{LL}{AD} * \frac{S_{CP}}{U_{CP}} * CP\right)}$$

Where

S_{LL} = Number of Numbers on the Sample of Landline Phones

U_{LL} = Universe of all available Landline Phones

S_{CP} = Number of Numbers on the Sample of Cellphones

U_{CP} = Universe of all available Landline Phones

LL = Factor 1 or 0 if respondent has landline

CP = Factor 1 or 0 if respondent has cellphone

AD = Number of individuals on household where respondent lives

To determine the Sample Size, Pew Research Center accounts for

$S = Calls_{Completed} + Calls_{Incomplete} + Calls_{Refused} + Calls_{NonAnswered} + Calls_{Errors}$

For any exercise, this is equivalent to

$S = \sum AllCallsMade$

Such model accounts for the probability of selection of respondents based on double coverage of cellphones and landlines and also for the number of individuals on a household. It also follows the standard model for dual frames that use both cellphones and landlines. This includes asking respondents on type of phone ownership and size of household.

Our sampling frame does not take into consideration landline phones and it is what would considere to be a single-frame phone survey. Currently, there is a vigorous debate about how to develop a a unique weight model for single frame phone surveys and as previously stated we aim to provide a simplified but statistically grounded approach.

For these reasons, for the CRA study we will adapt the Pew Research Model for a single frame mobile phone survey that also takes into consideration cellphone penetration rates.

Base Weights Formula - CRA

$$BWt_{CRA} = \frac{1}{\frac{S_{CP}}{U_{CP}}}$$

Where

S_{CP} = Number of Numbers on the Sample of Cellphones

U_{CP} = Universe of all available Landline Phones

This formula is similar to the Pew Research Center in where all Landline Factor is equal to zero and therefore the formula only accounts for the cellphoen related sample and universe. When including a factor to account for undercoverage, or cellphone penetration rates, the formula is:

$$BWt_{CRA} = \frac{1}{\frac{S_{CP}}{U_{CP}}} * \frac{1}{Teledensity_{CP}}$$

Where

S_{CP} = Number of Numbers on the Sample of Cellphones

U_{CP} = Universe of all available Landline Phones

$Teledensity_{CP}$ = 0-1 Cellphone Coverage (Subscriptions per 100 habitants) as reported by World Bank/ITU

```
##Dialed Numbers as Reported by Polling Company
Dial<- 216776

##Mobile/Telephone Cellullar Subscriptions (Depending if country calls only cellphones or all numbers)
Subs<- 54555497

##Initial Base Weight
BW1 <- Subs/Dial

##Cellphone Coverage (Subscriptions per 100 habitants) Adjustment as reported by World Bank/ITU
Cov <- 1.03769
BW2 <- BW1 * (1/Cov)

##Adding BW2 to Database
Kenya_Round1$BW2 <- BW2
```

2. Post-Stratified Weights Calibration (PsW)

Callibration of the base weights was done through post-stratification with selected demographic variables. For that purpose we use Sex, Urban/Rural Setting, Age, and Education (where available) as factors to control the demographic distribution of the sample. Education is not included on the Kenya sampling exercise.

For each available sample, a nationally representative survey or census datafile is be needed to complete calculations. The database needs to be recent or no more than 5 years prior and should be readable into R.

For this model we use the function `postStratify` of the package `Survey`. The model renders similar output to that of Raking and requires less data transformations facilitating deployment for multiple countries on a quicker fashion.

```
knitr::opts_chunk$set(echo = TRUE)

##Importing Comparison Survey - Kenya Integrated Household Budget Survey 2015-2016

KenyaBase <- read_dta("C:/Users/andre/OneDrive - UNICEF/CRA - Weights Model/HH_Members_Information_No_M
KenyaBase$U_ID <- paste(KenyaBase$clid,KenyaBase$hhid)#Generating an Unique ID
Kenya_HH_Information <- read_dta("C:/Users/andre/OneDrive - UNICEF/CRA - Weights Model/HH_Information.d
Kenya_HH <- Kenya_HH_Information[c("Set", "clid", "hhid")]
Kenya_HH$U_ID <- paste(Kenya_HH$clid, Kenya_HH$hhid)#Generating an Unique ID
Kenya_HH <- Kenya_HH[c("U_ID","Set")]
Kenya_SurveyForWeights <- merge(KenyaBase, Kenya_HH, by = "U_ID")

Kenya_SurveyForWeights$Sex1 <- Kenya_SurveyForWeights$Sex
Kenya_SurveyForWeights$Set1 <- Kenya_SurveyForWeights$Set
Kenya_SurveyForWeights$Age1 <- Kenya_SurveyForWeights$Age

KEN.PS <- xtabs(~Sex1 + Set1 + Age1, data = Kenya_SurveyForWeights)

# Subsetting database without minors
Kenya_Round1 <- subset(Kenya_Round1, Age != '<18 \nYrs Old')

###Preparing Survey Object as Numeric Factors from Sample Database
```

```r
Kenya_Round1$Sex1 <- 1 #Male
Kenya_Round1$Sex1[Kenya_Round1$Sex == 'Female'] <- 2
Kenya_Round1$Sex1 <- as.numeric(Kenya_Round1$Sex1)

Kenya_Round1$Set1 <- 1 #Urban
Kenya_Round1$Set1[Kenya_Round1$Set == 'Rural'] <- 2

Kenya_Round1$Age1 <- 1 #18-24 years old
Kenya_Round1$Age1[Kenya_Round1$Age == '25-34 \nYrs Old'] <- 2
Kenya_Round1$Age1[Kenya_Round1$Age == '35-44 \nYrs Old'] <- 3
Kenya_Round1$Age1[Kenya_Round1$Age == '45 Yrs \nOld & >'] <- 4



### Establishing Survey Object
KEN_WtdR1 <- svydesign( ids = ~0, strata = NULL, data = Kenya_Round1, weights = ~BW2) #No Clusters (ids

ps.dsgn <-postStratify(design = KEN_WtdR1, strata = ~Sex1 + Set1 + Age1, population = KEN.PS)

### Extracting Weights to Dataset

ps.weight <- ps.dsgn$postStrata[[1]] %>% attributes() %>% .[["weights"]]
Kenya_Round1$PsW <-ps.weight
```

3. Trimming

Weights are currently trimmed at the 95th percentile. Values above that level are winsorized to the 95th value, in order to reduce the design effect of having weighted the sample.

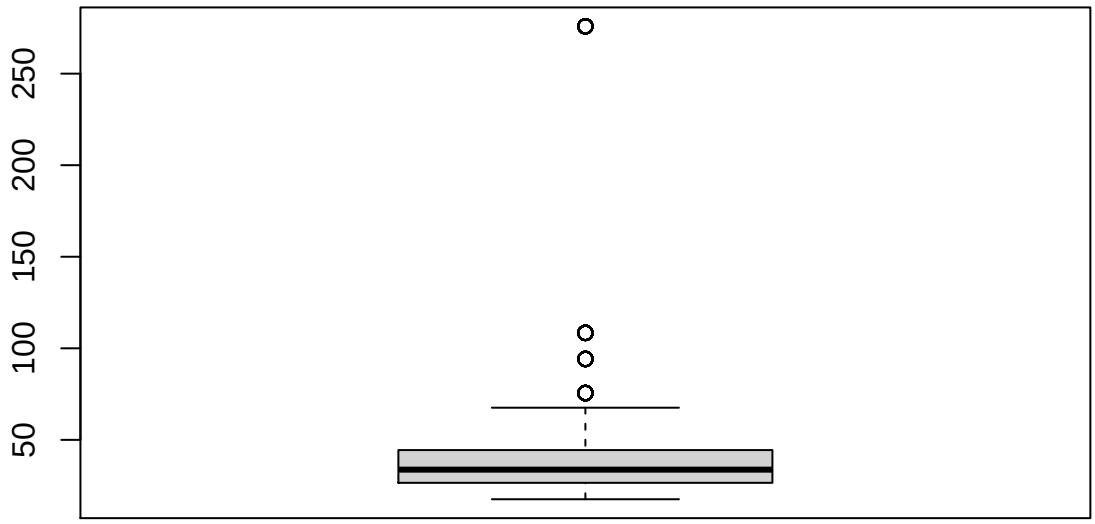
Below is graphed a distribution of the weights into boxplots drafted before and after trimming and windsoring to visualize the changes.

```r
# After Post-Stratification Weights

summary(Kenya_Round1$PsW)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.57   26.53   33.73   43.50   44.42  275.83
```

```r
boxplot(Kenya_Round1$PsW)
```



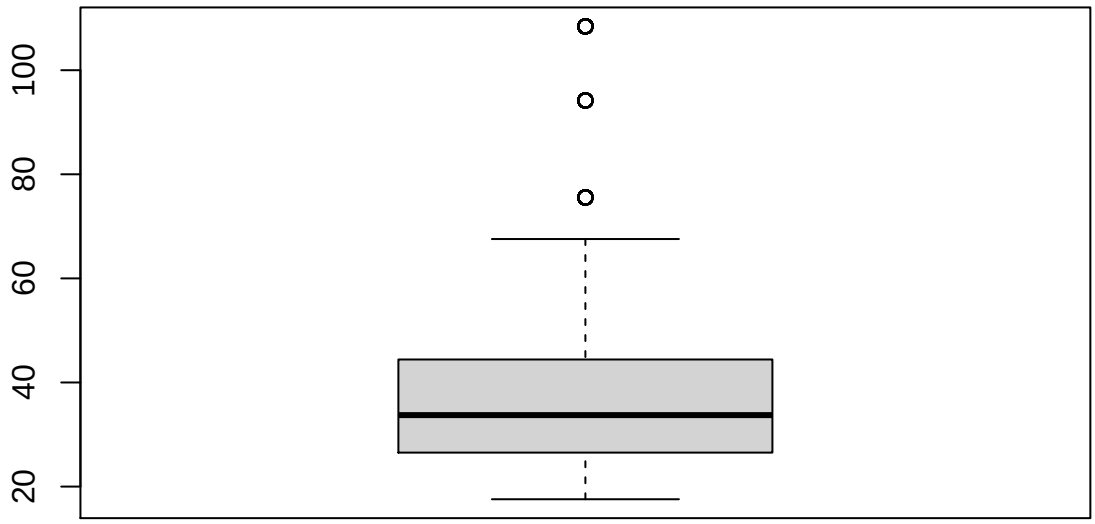

```
Kenya_Round1$Wt <- Kenya_Round1$PsW
Kenya_Round1$Wt[Kenya_Round1$Wt > (quantile(Kenya_Round1$PsW, c(.95)))] <- (quantile(Kenya_Round1$PsW,

# After Trimming and Windsoring the weights

summary(Kenya_Round1$Wt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.57   26.53   33.73   40.64   44.42  108.42
```

```
boxplot(Kenya_Round1$Wt)
```



4. Design Effect

To calculate the design effect of weighting, we used the design effect as postulated by Kish, where the design effect is the the ratio of the variance to the square of the mean of the relative weights, assuming that all stratum unit variances are equal.

We apply this approximation on the basis of pursuing an SRS sampling modeli and an approximation where individual weights are unequal to each other.

See https://rdrr.io/cran/PracTools/man/deffK.html#:~:text=for%20a%20sample-,Details,with%20equal%20weighting%20is%2

The formula used is:

$$Design_{Effect} = 1 + cv^2 = 1 + \frac{s^2_{Wt}}{\bar{x}^2_{Wt}}$$

For this document, we include both the design effect of trimmed and untrimmed weights.Please note how trimming reduces the overall design with out sample CRA dataset.

```
Deff_PSW= 1+(var(Kenya_Round1$PsW)/mean(Kenya_Round1$PsW)^2) #Without Trimming
Deff_PSW
```

```
## [1] 1.718587
```

```
Deff_Wt=1+(var(Kenya_Round1$Wt)/mean(Kenya_Round1$Wt)^2) # With Trimming
Deff_Wt
```

```
## [1] 1.304133
```

Model Limitations

Forecasting Estimates:

This weights model was constructed to provide a more accurate distribution of respondents when providing descriptive statistics for the Country Offices that are currently using the CRA. Since the CRA does not intend to provid forecasting modelling this weight model is not sufficient on its current form to be used for forecasting estimates as it lacks additional information on the universe of all phone users of a country, namely the number of phone lines per household and number of household members which is defined as an element of our starting Base Weights formula.

This also includes the need to have more and better information about non-respondents on each sampling exercise as well as how do the universe of phone users on each country compare to the country national population.

Regardless of this limitations, these weights still provide a solid approximation to reflect the samples closer to the population estimates, and to provide a probabilistic grounded approach on population based data where is non-existent, of very low quality, or based on self-selected samples.

Non-Response Adjustment

This model follows the same approach established by the Pew Research Center where non-response is accounted on the full sample count. However, there are other models that take into consideration non-response through screening variables. Current literature points out that single frame phoen surveys have a better fit than doble frame surveys even though non-response is not accounted through other factors. That is because the SE and design effects of single frames currently seem to be smaller than double frames designs as cellphone coverage increases.

Post-Stratification Weights

With the above mentioned limitations, the weighting model puts more emphasis on weighting the sample to national distributions of the population than to calculate probabilities of selection on every step. Base weights help to establish a more nuanced approximation to total number of calls made and to account for mobile phone penetraton rates on a given country but the respondents weights variation takes place on post-stratification.

Special care has been put to use the most recent available nationally representative population data to weight the surveys; including the use of recent Census data where available or nationally representative surveys fromnational statistical authorities, or methodologically sound surveys with national sampling strategies like the Demographic and Health Survey (DHS).

Exclusion of children under 18 years old

In order to correctly sample children a more specialized model is needed to account their participation on a mobile phone survey. Similarly, the original CRA design does not include children as sampling children over cellphones is still a very difficult endeavor as their access to cellphones is even more limited than the general population. Given the large segment of the population they represent, and their under-representation on the universe of cellphoen users, a mobile phone survey at the moment is not an appropiate medium to survey their self-reported behaviors.

All respondents under 18 years old when present haven been excluded and also the surveys used to weight the sample are also basd on population that is 18 years old or older.