# [Family Planning Panel Data Now Available from IPUMS PMA](#X2fd221be9553d7a441781cb29fb6de40d2723c9)

## [Background](#background)

Dating back to 2013, the original PMA survey design included high-frequency, **cross-sectional** samples of women and service delivery points collected from eleven countries participating in [Family Planning 2020](http://progress.familyplanning2020.org/) (FP2020) - a global partnership that supports the rights of women and girls to decide for themselves whether, when, and how many children they want to have. These surveys were designed to monitor annual progress towards [FP2020 goals](http://progress.familyplanning2020.org/measurement) via population-level estimates for several [core indicators](http://www.track20.org/pages/data_analysis/core_indicators/overview.php).

Beginning in 2019, PMA surveys were redesigned under a renewed partnership called [Family Planning 2030](https://fp2030.org/) (FP2030). These new surveys have been refocused on reproductive and sexual health indicators, and they feature a **longitudinal panel** of women of childbearing age. This design will allow researchers to measure contraceptive dynamics and changes in women’s fertility intentions over a **three year period** via annual in-person interviews.[1](#fn1)

Questions on the redesigned survey cover topics like:

* awareness, perception, knowledge, and use of contraceptive methods
* perceived quality and side effects of contraceptive methods among current users
* birth history and fertility intentions
* aspects of health service provision
* domains of empowerment

## [Sampling](#sampling)

PMA panel data includes a mixture of **nationally representative** and **sub-nationally representative** samples from eight participating countries. The panel study consists of three data collection phases, each spaced one year apart. IPUMS PMA has released data from the first *two* phases for countries where Phase 1 data collection began in 2019; we have released data from only the *first* phase for countries where Phase 1 data collection began in August or September 2020. Phase 3 data collection and processing is currently underway.

\*\* XXX - Insert blog post table here \*\*

PMA uses a multi-stage clustered sample design, with stratification at the urban-rural level or by sub-region. Geographically defined sample clusters - called [enumeration areas](https://pma.ipums.org/pma-action/variables/EAID#description_section) (EAs) – are provided by the national statistics agency in each country.[2](#fn2) These EAs are sampled using a *probability proportional to size* (PPS) method relative to the population distribution in each stratum.

At Phase 1, 35 household dwellings were selected at random within each EA. Resident enumerators visited each dwelling and invited one household member to complete a [Household Questionnaire](https://pma.ipums.org/pma/resources/questionnaires/hhf/PMA-Household-Questionnaire-English-2019.10.09.pdf)[3](#fn3) that includes a census of all household members and visitors who stayed there during the night before the interview. Female household members and visitors aged 15-49 were then invited to complete a subsequent Phase 1 [Female Questionnaire](https://pma.ipums.org/pma/resources/questionnaires/hhf/PMA-Female-Questionnaire-English-2019.10.09.pdf).[4](#fn4)

Questionnaires are administered in-person by **resident enumerators** visiting selected households in each EA. These are typically women over age 21 living in (or near) each EA and who hold at least a high school diploma.

One year later, resident enumerators visited the same dwellings and administered a Phase 2 Household Questionnaire. A panel member in Phase 2 is any woman still age 15-49 who could be reached for a second Female Questionnaire, either because:

* she still lived there, or
* she had moved elsewhere within the study area,[5](#fn5) but at least one member of the Phase 1 household remained and could help resident enumerators locate her new dwelling.[6](#fn6)

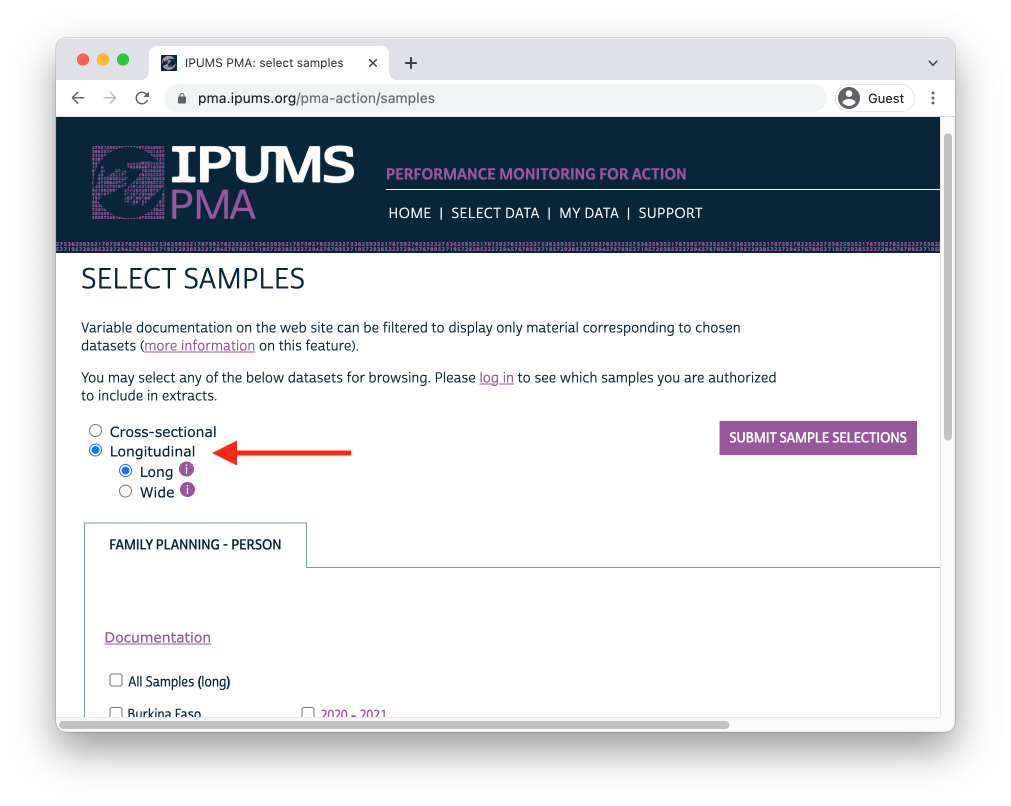
[SAMEDWELLING](https://pma.ipums.org/pma-action/variables/SAMEDWELLING#codes_section) indicates whether a Phase 2 female respondent resided in her Phase 1 dwelling or a new one.

Additionally, resident enumerators administered the Phase 2 Female Questionnaire to *new* women in sampled households who:

* reached age 15 after Phase 1
* joined the household after Phase 1
* declined the Female Questionnaire at Phase 1, but agreed to complete it at Phase 2

[PANELWOMAN](https://pma.ipums.org/pma-action/variables/PANELWOMAN#codes_section) indicates whether a Phase 2 household member completed the Phase 1 Female Questionnaire.

When you select the new **Longitudinal** sample option at checkout, you’ll be able to include responses from every available phase of the study. These samples are available in either “long” format (responses from each phase will be organized in separate rows) or “wide” format (responses from each phase will be organized in columns).

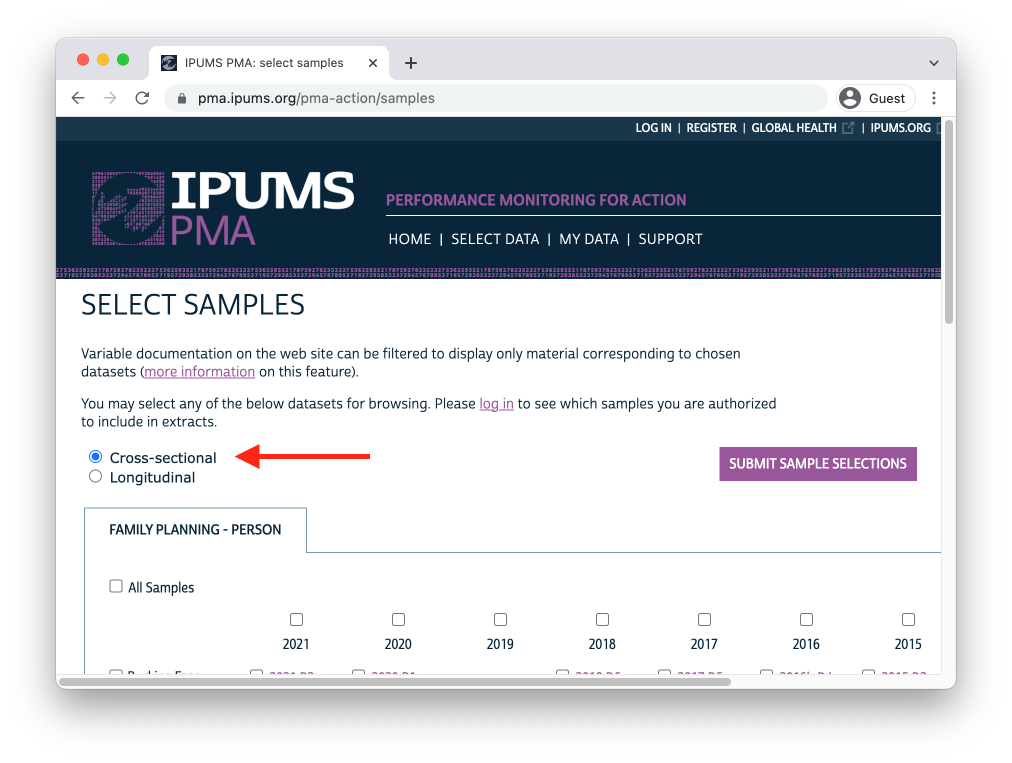


In addition to following up with women in the panel over time, PMA also adjusted sampling so that a cross-sectional sample could be produced concurrently with each data collection phase. These samples mainly overlap with the data you’ll obtain for a particular phase in the longitudinal sample, except that replacement households were drawn from each EA where more than 10% of households from the previous phase were no longer there. Conversely, panel members who were located in a new dwelling at Phase 2 will not be represented in the cross-sectional sample drawn from that EA. These adjustments ensure that population-level indicators may be derived from cross-sectional samples in a given year, even if panel members move or are lost to follow-up.

[CROSS\_SECTION](https://pma.ipums.org/pma-action/variables/CROSS_SECTION#codes_section) indicates whether a household member in a longitudinal sample is also included in the cross-sectional sample for a given year (every person in a cross-sectional sample is included in the longitudinal sample).

We'll cover **sample composition** in much greater detail in an upcoming chapter. \*\* XXX - Mention which chapter \*\*

You’ll find PMA cross-sectional samples dating back to 2013 if you select the **Cross-sectional** sample option at checkout.



## [Survey Design Elements](#survey-design-elements)

In upcoming chapters, we’ll demonstrate how to incorporate PMA sampling weights and information about its stratified cluster sampling procedure into your analysis.

Whether you intend to work with a new **Longitudinal** or **Cross-sectional** data extract, you’ll find the same set of sampling weights available for all PMA Family Planning surveys dating back to 2013.[7](#fn7)

* [HQWEIGHT](https://pma.ipums.org/pma-action/variables/HQWEIGHT#description_section) can be used to generate cross-sectional population estimates from questions on the Household Questionnaire.[8](#fn8)
* [FQWEIGHT](https://pma.ipums.org/pma-action/variables/FQWEIGHT#description_section) can be used to to generate cross-sectional population estimates from questions on the Female Questionnaire.[9](#fn9)
* [EAWEIGHT](https://pma.ipums.org/pma-action/variables/EAWEIGHT#description_section) can be used to compare the selection probability of a particular household with that of its EA.

A fourth Family Planning survey weight, [POPWT](https://pma.ipums.org/pma-action/variables/POPWT#description\_section), is currently available only for \*\*Cross-sectional\*\* data extracts and Phase 1 panel data.[10](#fn10)

Additionally, PMA created a new weight, [PANELWEIGHT](https://pma.ipums.org/pma-action/variables/PANELWEIGHT#description_section), which should be used in longitudinal analyses spanning multiple phases, as it adjusts for loss to follow-up. PANELWEIGHT is available only for **Longitudinal** data extracts.

For example, suppose we wanted to estimate the proportion of reproductive age women in Burkina Faso who were using contraception at the time of data collection for both Phase 1 and Phase 2. In a cross-sectional or “long” longitudinal extract, you’ll find this information in the variable CP. In a “wide” longitudinal extract, you’ll find it in CP\_1 for Phase 1, and in CP\_2 for Phase 2. We’ll be working with a “wide” extract loaded into Stata. This code restricts the dataset to the women we are interested in, and counts the number of respondents in who were known to be using contraceptives in both Phases 1 and 2.

Variable names in a "wide" extract have a numeric suffix corresponding with a data collection phase. CP\_1 is the Phase 1 version of [CP](https://pma.ipums.org/pma-action/variables/CP#codes_section), while CP\_2 comes from Phase 2.

cd "Q:/BMGF - PMA IPUMS FP Blog Posts/Blog1/Blog1 Stata Markdown"  
  
use "pma\_00126.dta", clear  
  
\* In BF Phase 1, and had female questionnaire at least partly completed &   
\* under age 49 & usual resident who slept here last night  
\* Note: To be eligible for Phase 2, the woman had to be < age 49 at Phase 1.  
gen of\_interest\_1 = sample\_1 == 85409 & ///  
 (resultfq\_1 == 1 | resultfq\_1 == 5 ) & ///  
 agehq\_1 < 49 & ///  
 resident\_1 != 21  
  
\* In BF Phase 2, and had female and household questionnaire completed &   
\* <= age 49 & usual resident or visitor who slept here last night  
gen of\_interest\_2 = sample\_2 == 85412 & ///  
 (resultfq\_2 == 1 & resulthq\_2 == 1 ) & ///  
 agehq\_2 <= 49 & ///  
 resident\_2 != 21 & resident\_2 != 31  
  
\* Of interest in both studies  
gen of\_interest\_both = of\_interest\_1 & of\_interest\_2  
  
label variable cp\_1 "Contraceptive user (Phase 1)"  
label variable cp\_2 "Contraceptive user (Phase 2)"  
  
keep if of\_interest\_both  
  
table ( cp\_1 ) ( cp\_2 ) (), nototals missing zerocounts

. table ( cp\_1 ) ( cp\_2 ) (), nototals missing zerocounts  
  
--------------------------------------------------------------  
 | Contraceptive user (Phase 2)   
 | no yes  
-----------------------------+--------------------------------  
Contraceptive user (Phase 1) |   
 no | 2,589 821  
 yes | 556 1,241  
 no response or missing | 5 0  
--------------------------------------------------------------

To go beyond counts and estimate a population percentage, we will need to tell Stata that we are working with a sample survey dataset and stipulate the sample design (specify which variables identify strata and clusters) and where to find the survey weights. This is accomplished with the [svyset](https://www.stata.com/manuals/svysvyset.pdf) command.

We use eaid\_1 as the cluster ID[13](#fn13) and strata\_1 as the stratum ID[14](#fn14) and `panelweight’ holds the survey weight.

We also generate a binary 0/1 variable that indicates which women were using contraception in both Phase 1 and 2.

gen cp\_both = cp\_1 == 1 & cp\_2 == 1 if cp\_1 < 90  
label variable cp\_both "Contraceptive user (Phases 1 & 2)"  
label define cp\_both 1 "Yes" 0 "No", replace  
label values cp\_both cp\_both  
  
\* This is a lean svyset call. We recall that the default vce option is   
\* vce(linearized) and the default singleunit option is (missing).   
\* Read the svyset documentation if you want to consider using other settings.   
\* For now, the defaults are fine.  
  
svyset eaid\_1, strata(strata\_1) weight(panelweight)   
svy: proportion cp\_both

. svyset eaid\_1, strata(strata\_1) weight(panelweight)   
  
Sampling weights: <none>  
 VCE: linearized  
 Single unit: missing  
 Strata 1: strata\_1  
 Sampling unit 1: eaid\_1  
 FPC 1: <zero>  
 Weight 1: panelweight  
  
. svy: proportion cp\_both  
(running proportion on estimation sample)  
  
Survey: Proportion estimation  
  
Number of strata = 2 Number of obs = 5,207  
Number of PSUs = 167 Population size = 5,215.6413  
 Design df = 165  
  
--------------------------------------------------------------  
 | Linearized Logit  
 | Proportion std. err. [95% conf. interval]  
-------------+------------------------------------------------  
 cp\_both |  
 No | .8122041 .012815 .7855839 .8362084  
 Yes | .1877959 .012815 .1637916 .2144161  
--------------------------------------------------------------

This is our first look at Stata’s output for estimating proportions. The top of the output table lists the number of strata and the number of PSUs in the dataset, along with the number of respondents in the sample and the sum of their weights (under the heading: Population size). The number of design degrees of freedom (df) is the number of PSUs minus the number of strata.[15](#fn15)

The lower portion of the table lists the values of the outcome variable, or in this case their value labels: No and Yes. It lists the proportion of the population that are estimated to have each outcome, that proportion’s standard error, and a two-sided survey-adjusted confidence interval for the proportion.

Stata’s default confidence interval is the so-called “logit interval” which is one of several possibilities.[11](#fn11) For now we will simply say that the default logit interval is a fine choice for most circumstances. To request a different kind of confidence interval, read about the options and specify what you want using the citype() option to the svy: proportion command (e.g., citype(wilson) or citype(exact)).

To describe this output in an English language sentence, we might say something like: “Based on this survey sample of 5,207 women from Burkina Faso, we estimate that if the surveys were free from bias then about 18.8% women who were eligible to be sampled in the PMA surveys would be self-reported users of contraception in both Phases 1 and 2 (95% CI: 16.4-21.4%).”

With survey data collected from using a complex sample design that employs strata and/or clusters, we sometimes like to report the **design effect** which is an index of the statistical precision penalty that we pay for using that sample design. In Stata, we can see the design effect by issuing the following post-estimation command [estat effects](https://www.stata.com/manuals/svysvypostestimation.pdf)

\* Calculate the design effect for the most recent estimation  
estat effects

.   
. \* Calculate the design effect for the most recent estimation  
. estat effects  
  
----------------------------------------------------------  
 | Linearized  
 | Proportion std. err. DEFF DEFT  
-------------+--------------------------------------------  
 cp\_both |  
 No | .8122041 .012815 5.6052 2.36753  
 Yes | .1877959 .012815 5.6052 2.36753  
----------------------------------------------------------  
  
.

We see that the design effect (DEFF) is 5.6, which we might interpret by saying “The confidence interval for this estimation is as wide as we would expect from a simple random sample of this sample size (5,207) divided by 5.6 or about 929 respondents.”

The DEFT is the square root of DEFF and we might use it in a sentence thus: “Because of the complex sample design and heterogeneity of survey weights, the confidence interval for this estimation is 2.4 times wider than we would expect from a simple random sample of size 5,207 respondents.”

The figure 929 is sometimes called the **effective sample size**.

Let’s take a moment and estimate proportions from two simple random samples where 18.8% of the respondents have the outcome: one where the sample size is 5,207 and one where the sample size is 929. We can do this by generating an empty dataset with the appropriate number of respondents and a binary variable named y.

\* Generate a dataset of a simple random sample of 5,207 respondents where  
\* 18.8% have the outcome and estimate the proportion;  
  
clear  
set obs 5207  
gen y = 0  
replace y = 1 if \_n < 0.188 \* 5207  
tab y  
svyset \_n  
svy: proportion y  
  
\* Generate a dataset of a simple random sample of 929 respondents where  
\* 18.8% have the outcome and estimate the proportion;  
  
clear  
set obs 929  
gen y = 0  
replace y = 1 if \_n < 0.188 \* 929  
tab y  
svyset \_n  
svy: proportion y

.   
. \* Generate a dataset of a simple random sample of 5,207 respondents where  
. \* 18.8% have the outcome and estimate the proportion;  
.   
. clear  
  
. set obs 5207  
Number of observations (\_N) was 0, now 5,207.  
  
. gen y = 0  
  
. replace y = 1 if \_n < 0.188 \* 5207  
(978 real changes made)  
  
. tab y  
  
 y | Freq. Percent Cum.  
------------+-----------------------------------  
 0 | 4,229 81.22 81.22  
 1 | 978 18.78 100.00  
------------+-----------------------------------  
 Total | 5,207 100.00  
  
. svyset \_n  
  
Sampling weights: <none>  
 VCE: linearized  
 Single unit: missing  
 Strata 1: <one>  
 Sampling unit 1: <observations>  
 FPC 1: <zero>  
  
. svy: proportion y  
(running proportion on estimation sample)  
  
Survey: Proportion estimation  
  
Number of strata = 1 Number of obs = 5,207  
Number of PSUs = 5,207 Population size = 5,207  
 Design df = 5,206  
  
--------------------------------------------------------------  
 | Linearized Logit  
 | Proportion std. err. [95% conf. interval]  
-------------+------------------------------------------------  
 y |  
 0 | .8121759 .0054131 .8013328 .8225583  
 1 | .1878241 .0054131 .1774417 .1986672  
--------------------------------------------------------------  
  
.   
. \* Generate a dataset of a simple random sample of 929 respondents where  
. \* 18.8% have the outcome and estimate the proportion;  
.   
. clear  
  
. set obs 929  
Number of observations (\_N) was 0, now 929.  
  
. gen y = 0  
  
. replace y = 1 if \_n < 0.188 \* 929  
(174 real changes made)  
  
. tab y  
  
 y | Freq. Percent Cum.  
------------+-----------------------------------  
 0 | 755 81.27 81.27  
 1 | 174 18.73 100.00  
------------+-----------------------------------  
 Total | 929 100.00  
  
. svyset \_n  
  
Sampling weights: <none>  
 VCE: linearized  
 Single unit: missing  
 Strata 1: <one>  
 Sampling unit 1: <observations>  
 FPC 1: <zero>  
  
. svy: proportion y  
(running proportion on estimation sample)  
  
Survey: Proportion estimation  
  
Number of strata = 1 Number of obs = 929  
Number of PSUs = 929 Population size = 929  
 Design df = 928  
  
--------------------------------------------------------------  
 | Linearized Logit  
 | Proportion std. err. [95% conf. interval]  
-------------+------------------------------------------------  
 y |  
 0 | .8127018 .0128073 .786262 .8365509  
 1 | .1872982 .0128073 .1634491 .213738  
--------------------------------------------------------------  
  
.

Now let’s compare the CI width from the simple random sample with N=929 with that from the complex sample with N=5,207.

\* Now examine the complex data 95% CI width divided by the   
\* simple random sample of 929 95% CI width and see that it is ~= DEFT  
  
di (.2144-.1638) / (.1987-.1774)

.   
. \* Now examine the complex data 95% CI width divided by the   
. \* simple random sample of 929 95% CI width and see that it is ~= DEFT  
.   
. di (.2144-.1638) / (.1987-.1774)  
2.3755869  
  
.

It can be disheartening to know that the teams did all the work to interview 5,207 respondents and yet for this estimation that sample only has the statistical precision of a simple random sample of 929 respondents. The statistical penalty is because of both a clustering effect – spatial heterogeneity in the outcome across PSUs – and because of heterogeneity in the survey weights. In some survey reporting contexts you will be expected to report either DEFF or DEFT, or both. Be clear about which one you are reporting. The design effect will vary across outcomes, across strata, and across PMA Phases, so if it is of interest, estimate it anew for each analysis. You can learn more about the survey design effect in materials on survey sampling statistics. **See Section [XXX].**

\*\* Consider adding a short section on organ pipe plots. \*\*

This syntax and svyset command worked well for Burkina Faso, but take note: the variable [STRATA](https://pma.ipums.org/pma-action/variables/STRATA#codes_section) is *not available* for samples collected from DRC - Kinshasa or DRC - Kongo Central. If your extract includes any DRC sample, you’ll need to amend this variable to include a unique numeric code for each of those regions.

For example, let’s look at a different wide extract, containing all of the samples included in this data release. Notice that STRATA\_1 lists the sample strata for all values of [COUNTRY](https://pma.ipums.org/pma-action/variables/COUNTRY#codes_section) *except* for DRC, where the variable is missing.

use "pma\_00153.dta", clear  
  
\* Phase 1 : had female questionnaire at least partly completed &   
\* under age 49 & usual resident who slept here last night  
gen of\_interest\_1 = (resultfq\_1 == 1 | resultfq\_1 == 5 ) & ///  
 agehq\_1 < 49 & resident\_1 != 21  
  
\* Phase 2 : had female and household questionnaire completed &   
\* <= age 49 & usual resident or visitor who slept here last night  
gen of\_interest\_2 = (resultfq\_2 == 1 & resulthq\_2 == 1 ) & ///  
 agehq\_2 <= 49 & resident\_2 != 21 & resident\_2 != 31  
  
\* Of interest in both studies  
gen of\_interest\_both = of\_interest\_1 & of\_interest\_2  
  
keep if of\_interest\_both  
  
table ( strata\_1 ) () ( country ) if of\_interest\_both, ///  
 nototals missing zerocounts

.   
. use "pma\_00153.dta", clear  
  
.   
. \* Phase 1 : had female questionnaire at least partly completed &   
. \* under age 49 & usual resident who slept here last night  
. gen of\_interest\_1 = (resultfq\_1 == 1 | resultfq\_1 == 5 ) & ///  
> agehq\_1 < 49 & resident\_1 != 21  
  
.   
. \* Phase 2 : had female and household questionnaire completed &   
. \* <= age 49 & usual resident or visitor who slept here last night  
. gen of\_interest\_2 = (resultfq\_2 == 1 & resulthq\_2 == 1 ) & ///  
> agehq\_2 <= 49 & resident\_2 != 21 & resident\_2 != 31  
  
.   
. \* Of interest in both studies  
. gen of\_interest\_both = of\_interest\_1 & of\_interest\_2  
  
.   
. keep if of\_interest\_both  
(12,453 observations deleted)  
  
.   
. table ( strata\_1 ) () ( country ) if of\_interest\_both, ///  
> nototals missing zerocounts  
  
pma country = burkina faso  
----------------------------------  
 | Frequency  
----------------------+-----------  
strata |   
 urban, burkina faso | 3,058  
 rural, burkina faso | 2,154  
----------------------------------  
  
pma country = congo, democratic republic  
-------------------  
 | Frequency  
-------+-----------  
strata |   
 . | 3,487  
-------------------  
  
pma country = kenya  
----------------------------------------  
 | Frequency  
----------------------------+-----------  
strata |   
 bungoma - urban, kenya | 153  
 bungoma - rural, kenya | 489  
 kakamega - urban, kenya | 133  
 kakamega - rural, kenya | 438  
 kericho - urban, kenya | 249  
 kericho - rural, kenya | 453  
 kiambu - urban, kenya | 214  
 kiambu - rural, kenya | 311  
 kilifi - urban, kenya | 170  
 kilifi - rural, kenya | 455  
 kitui - urban, kenya | 153  
 kitui - rural, kenya | 586  
 nairobi - urban, kenya | 494  
 nandi - urban, kenya | 260  
 nandi - rural, kenya | 711  
 nyamira - urban, kenya | 143  
 nyamira - rural, kenya | 382  
 siaya - urban, kenya | 130  
 siaya - rural, kenya | 437  
 west pokot - urban, kenya | 104  
 west pokot - rural, kenya | 474  
----------------------------------------  
  
pma country = nigeria  
-----------------------------  
 | Frequency  
-----------------+-----------  
strata |   
 lagos, nigeria | 1,088  
 kano - urban | 437  
 kano - rural | 561  
-----------------------------  
  
.

We can replace those values with numeric codes from the variable [GEOCD](https://pma.ipums.org/pma-action/variables/GEOCD#codes_section):

table ( geocd ) if country == 2, nototals missing zerocounts  
  
tab geocd  
tab geocd, nolabel

.   
. table ( geocd ) if country == 2, nototals missing zerocounts  
  
-------------------------------  
 | Frequency  
-------------------+-----------  
province, congo dr |   
 kinshasa | 1,973  
 kongo central | 1,514  
-------------------------------  
  
.   
. tab geocd  
  
 province, |  
 congo dr | Freq. Percent Cum.  
--------------+-----------------------------------  
 kinshasa | 1,973 56.58 56.58  
kongo central | 1,514 43.42 100.00  
--------------+-----------------------------------  
 Total | 3,487 100.00  
  
. tab geocd, nolabel  
  
 province, |  
 congo dr | Freq. Percent Cum.  
------------+-----------------------------------  
 1 | 1,973 56.58 56.58  
 2 | 1,514 43.42 100.00  
------------+-----------------------------------  
 Total | 3,487 100.00  
  
.

If GEOCD is not missing, we’ll use its numeric code in place of STRATA\_1. Otherwise, we’d like to leave STRATA\_1 unchanged. To avoid confusion with the original variable STRATA\_1, we’ll call our new variable STRATA\_RECODE.

\* Note that the values of geocd are distinct from the values of strata\_1  
sum strata\_1  
sum geocd  
  
\* So make a new variable named strata\_recode and set it to strata\_1   
\* everywhere except DRC and set it to geocd in DRC  
clonevar strata\_recode = strata\_1   
replace strata\_recode = geocd if country == 2  
  
\* Now copy the value label from strata\_1 into a new label named strata\_recode  
\* and update it with the labels from geocd  
  
label copy STRATA\_1 strata\_recode, replace  
label define strata\_recode 1 "Kinshasa, DRC" 2 "Kongo Central, DRC", modify  
  
\* Use the new value label with the new variable  
label values strata\_recode strata\_recode  
  
\* Note that the new variable is not missing for any women  
\* of interest from Phase 1 and Phase 2  
tab strata\_recode, m

.   
. \* Note that the values of geocd are distinct from the values of strata\_1  
. sum strata\_1  
  
 Variable | Obs Mean Std. dev. Min Max  
-------------+---------------------------------------------------------  
 strata\_1 | 14,237 59259.26 20596.78 40410 85402  
  
. sum geocd  
  
 Variable | Obs Mean Std. dev. Min Max  
-------------+---------------------------------------------------------  
 geocd | 3,487 1.434184 .4957204 1 2  
  
.   
. \* So make a new variable named strata\_recode and set it to strata\_1   
. \* everywhere except DRC and set it to geocd in DRC  
. clonevar strata\_recode = strata\_1   
(3,487 missing values generated)  
  
. replace strata\_recode = geocd if country == 2  
(3,487 real changes made)  
  
.   
. \* Now copy the value label from strata\_1 into a new label named strata\_recode  
. \* and update it with the labels from geocd  
.   
. label copy STRATA\_1 strata\_recode, replace  
  
. label define strata\_recode 1 "Kinshasa, DRC" 2 "Kongo Central, DRC", modify  
  
.   
. \* Use the new value label with the new variable  
. label values strata\_recode strata\_recode  
  
.   
. \* Note that the new variable is not missing for any women  
. \* of interest from Phase 1 and Phase 2  
. tab strata\_recode, m  
  
 strata | Freq. Percent Cum.  
----------------------------------------+-----------------------------------  
 Kinshasa, DRC | 1,973 11.13 11.13  
 Kongo Central, DRC | 1,514 8.54 19.67  
 bungoma - urban, kenya | 153 0.86 20.54  
 bungoma - rural, kenya | 489 2.76 23.30  
 kakamega - urban, kenya | 133 0.75 24.05  
 kakamega - rural, kenya | 438 2.47 26.52  
 kericho - urban, kenya | 249 1.40 27.92  
 kericho - rural, kenya | 453 2.56 30.48  
 kiambu - urban, kenya | 214 1.21 31.69  
 kiambu - rural, kenya | 311 1.75 33.44  
 kilifi - urban, kenya | 170 0.96 34.40  
 kilifi - rural, kenya | 455 2.57 36.97  
 kitui - urban, kenya | 153 0.86 37.83  
 kitui - rural, kenya | 586 3.31 41.14  
 nairobi - urban, kenya | 494 2.79 43.92  
 nandi - urban, kenya | 260 1.47 45.39  
 nandi - rural, kenya | 711 4.01 49.40  
 nyamira - urban, kenya | 143 0.81 50.21  
 nyamira - rural, kenya | 382 2.16 52.36  
 siaya - urban, kenya | 130 0.73 53.10  
 siaya - rural, kenya | 437 2.47 55.56  
 west pokot - urban, kenya | 104 0.59 56.15  
 west pokot - rural, kenya | 474 2.67 58.82  
 lagos, nigeria | 1,088 6.14 64.96  
 kano - urban | 437 2.47 67.43  
 kano - rural | 561 3.17 70.59  
 urban, burkina faso | 3,058 17.25 87.85  
 rural, burkina faso | 2,154 12.15 100.00  
----------------------------------------+-----------------------------------  
 Total | 17,724 100.00  
  
.

Now, we can use STRATA\_RECODE with the svyset command to obtain population estimates for each nationally representative or sub-nationally representative sample.

\* Generate cp\_both again for this wide dataset  
gen cp\_both = cp\_1 == 1 & cp\_2 == 1 if cp\_1 < 90  
label variable cp\_both "Contraceptive user (Phases 1 & 2)"  
label define cp\_both 1 "Yes" 0 "No", replace  
label values cp\_both cp\_both  
  
svyset eaid\_1, strata(strata\_recode) weight(panelweight)   
  
\* For Stata to estimate the proportion for each population,   
\* we will use the over(varname) option where varname needs to  
\* be an integer variable - preferably with a value label.   
  
\* So construct a new variable named pop\_numeric and give it a   
\* unique value for each PMA population.  
  
gen pop\_numeric = .  
replace pop\_numeric = 1 if country == 1 // Burkina Faso  
replace pop\_numeric = 2 if country == 2 & geocd == 1 // Kinshasa  
replace pop\_numeric = 3 if country == 2 & geocd == 2 // Kongo Central  
replace pop\_numeric = 4 if country == 7 // Kenya  
replace pop\_numeric = 5 if country == 9 & geong == 4 // Kano  
replace pop\_numeric = 6 if country == 9 & geong == 2 // Lagos  
  
label define pop\_numeric ///  
 1 "Burkina Faso" ///  
 2 "DRC-Kinshasa" ///  
 3 "DRC-Kongo Central" ///  
 4 "Kenya" ///  
 5 "Nigeria-Kano" ///  
 6 "Nigeria-Lagos", replace  
  
label values pop\_numeric pop\_numeric  
  
svy : proportion cp\_both , over(pop\_numeric)

.   
. \* Generate cp\_both again for this wide dataset  
. gen cp\_both = cp\_1 == 1 & cp\_2 == 1 if cp\_1 < 90  
(19 missing values generated)  
  
. label variable cp\_both "Contraceptive user (Phases 1 & 2)"  
  
. label define cp\_both 1 "Yes" 0 "No", replace  
  
. label values cp\_both cp\_both  
  
.   
. svyset eaid\_1, strata(strata\_recode) weight(panelweight)   
  
Sampling weights: <none>  
 VCE: linearized  
 Single unit: missing  
 Strata 1: strata\_recode  
 Sampling unit 1: eaid\_1  
 FPC 1: <zero>  
 Weight 1: panelweight  
  
.   
. \* For Stata to estimate the proportion for each population,   
. \* we will use the over(varname) option where varname needs to  
. \* be an integer variable - preferably with a value label.   
.   
. \* So construct a new variable named pop\_numeric and give it a   
. \* unique value for each PMA population.  
.   
. gen pop\_numeric = .  
(17,724 missing values generated)  
  
. replace pop\_numeric = 1 if country == 1 // Burkina Faso  
(5,212 real changes made)  
  
. replace pop\_numeric = 2 if country == 2 & geocd == 1 // Kinshasa  
(1,973 real changes made)  
  
. replace pop\_numeric = 3 if country == 2 & geocd == 2 // Kongo Central  
(1,514 real changes made)  
  
. replace pop\_numeric = 4 if country == 7 // Kenya  
(6,939 real changes made)  
  
. replace pop\_numeric = 5 if country == 9 & geong == 4 // Kano  
(998 real changes made)  
  
. replace pop\_numeric = 6 if country == 9 & geong == 2 // Lagos  
(1,088 real changes made)  
  
.   
. label define pop\_numeric ///  
> 1 "Burkina Faso" ///  
> 2 "DRC-Kinshasa" ///  
> 3 "DRC-Kongo Central" ///  
> 4 "Kenya" ///  
> 5 "Nigeria-Kano" ///  
> 6 "Nigeria-Lagos", replace  
  
.   
. label values pop\_numeric pop\_numeric  
  
.   
. svy : proportion cp\_both , over(pop\_numeric)   
(running proportion on estimation sample)  
  
Survey: Proportion estimation  
  
Number of strata = 28 Number of obs = 17,705  
Number of PSUs = 665 Population size = 17,691.26  
 Design df = 637  
  
------------------------------------------------------------------------  
 | Linearized Logit  
 | Proportion std. err. [95% conf. interval]  
-----------------------+------------------------------------------------  
 cp\_both@pop\_numeric |  
 No Burkina Faso | .8122041 .012815 .785736 .8360846  
 No DRC-Kinshasa | .6802513 .0163794 .647268 .711525  
 No DRC-Kongo Central | .7318119 .0287314 .6718062 .7843679  
 No Kenya | .6342298 .0083126 .6177575 .6503939  
 No Nigeria-Kano | .9463423 .0130503 .9141428 .9669031  
 No Nigeria-Lagos | .7065456 .0176703 .6706908 .7400099  
 Yes Burkina Faso | .1877959 .012815 .1639154 .214264  
 Yes DRC-Kinshasa | .3197487 .0163794 .288475 .352732  
Yes DRC-Kongo Central | .2681881 .0287314 .2156321 .3281938  
 Yes Kenya | .3657702 .0083126 .3496061 .3822425  
 Yes Nigeria-Kano | .0536577 .0130503 .0330969 .0858572  
 Yes Nigeria-Lagos | .2934544 .0176703 .2599901 .3293092  
------------------------------------------------------------------------  
  
.

## [Inclusion Criteria for Analysis](#inclusion-criteria-for-analysis)

In the remainder of this manual, we’ll be showcasing code you can use to reproduce key indicators included in the **PMA Longitudinal Brief** for each sample. In many cases, you’ll find separate reports available in English and French, and for both national and sub-national summaries. For reference, here are the highest-level population summaries available in English for each sample where Phase 2 IPUMS PMA data is currently available:

* [Burkina Faso](https://www.pmadata.org/sites/default/files/data_product_results/Burkina%20National_Phase%202_Panel_Results%20Brief_English_Final.pdf)
* [DRC - Kinshasa](https://www.pmadata.org/sites/default/files/data_product_results/DRC%20Kinshasa_%20Phase%202%20Panel%20Results%20Brief_English_Final.pdf)
* [DRC - Kongo Central](https://www.pmadata.org/sites/default/files/data_product_results/DRC%20Kongo%20Central_%20Phase%202%20Panel%20Results%20Brief_English_Final.pdf)
* [Kenya](https://www.pmadata.org/sites/default/files/data_product_results/Kenya%20National_Phase%202_Panel%20Results%20Brief_Final.pdf)
* [Nigeria - Kano](https://www.pmadata.org/sites/default/files/data_product_results/Nigeria%20KANO_Phase%202_Panel_Results%20Brief_Final.pdf)
* [Nigeria - Lagos](https://www.pmadata.org/sites/default/files/data_product_results/Nigeria%20LAGOS_Phase%202_Panel_Results%20Brief_Final.pdf)

Panel data in these reports is limited to the *de facto* population of women who completed the Female Questionnaire in both Phase 1 and Phase 2. This includes women who slept in the household during the night before the interview for the Household Questionnaire. The *de jure* population includes women who are usual household members, but who slept elsewhere that night. We will remove *de jure* cases recorded in the variable [RESIDENT](https://pma.ipums.org/pma-action/variables/RESIDENT#codes_section).

Missing data in RESIDENT\_2 represent women who were lost to follow-up in Phase 2.

For example, returning to our “wide” data extract for Burkina Faso, you can see the number of women who slept in the household before the Household Questionnaire for each phase reported in RESIDENT\_1 and RESIDENT\_2:

use "pma\_00126.dta", clear  
  
keep if sample\_1 == 85409  
  
table ( resident\_1 ) () (), nototals missing zerocounts  
table ( resident\_2 ) () (), nototals missing zerocounts

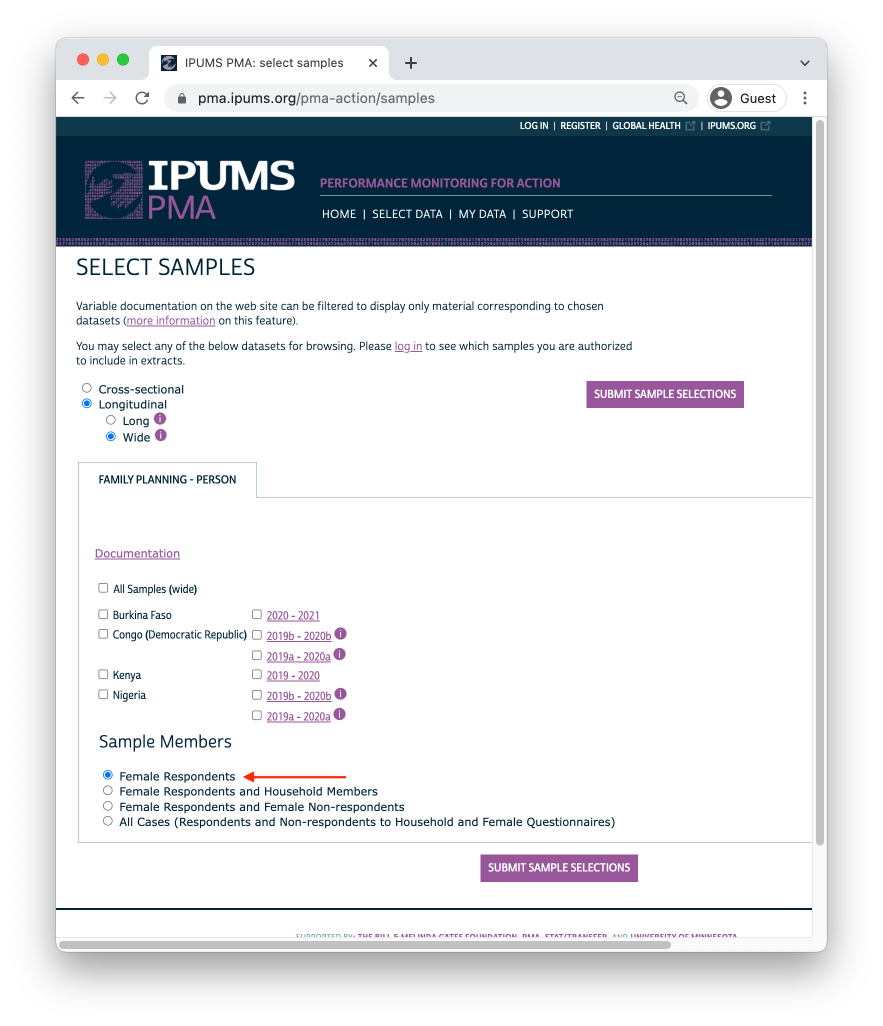
.   
. use "pma\_00126.dta", clear  
  
.   
. keep if sample\_1 == 85409  
(23,387 observations deleted)  
  
.   
. table ( resident\_1 ) () (), nototals missing zerocounts  
  
-----------------------------------------------------------  
 | Frequency  
-----------------------------------------------+-----------  
usual member of household |   
 visitor, slept in hh last night | 106  
 usual member, did not sleep in hh last night | 174  
 usual member, slept in hh last night | 6,510  
-----------------------------------------------------------  
  
. table ( resident\_2 ) () (), nototals missing zerocounts  
  
-------------------------------------------------------------------------  
 | Frequency  
-------------------------------------------------------------+-----------  
usual member of household |   
 visitor, slept in hh last night | 74  
 usual member, did not sleep in hh last night | 230  
 usual member, slept in hh last night | 5,993  
 slept in hh last night, no response if usually lives in hh | 1  
 . | 492  
-------------------------------------------------------------------------  
  
.

The *de facto* population is represented in codes 11 and 22. We will use an if statement or keep statement to include only those cases.

keep if inlist(resident\_1,11,22) & inlist(resident\_2,11,22)  
label variable resident\_1 "Resident type - Phase 1"  
label variable resident\_2 "Resident type - Phase 2"  
label define RESIDENT\_1 11 "Visitor" 22 "Usual", modify  
label define RESIDENT\_2 11 "Visitor" 22 "Usual", modify  
  
table ( resident\_1 ) ( resident\_2 ) (), nototals missing zerocounts

.   
. keep if inlist(resident\_1,11,22) & inlist(resident\_2,11,22)  
(823 observations deleted)  
  
. label variable resident\_1 "Resident type - Phase 1"  
  
. label variable resident\_2 "Resident type - Phase 2"  
  
. label define RESIDENT\_1 11 "Visitor" 22 "Usual", modify  
  
. label define RESIDENT\_2 11 "Visitor" 22 "Usual", modify  
  
.   
. table ( resident\_1 ) ( resident\_2 ) (), nototals missing zerocounts  
  
----------------------------------------------------  
 | Resident type - Phase 2   
 | Visitor Usual  
------------------------+---------------------------  
Resident type - Phase 1 |   
 Visitor | 56 39  
 Usual | 17 5,855  
----------------------------------------------------  
  
.

Additionally, these reports only include women who completed (or partially completed) both Female Questionnaires. This information is reported in [RESULTFQ](https://pma.ipums.org/pma-action/variables/RESULTFQ#codes_section). In our “wide” extract, this information appears in RESULTFQ\_1 and RESULTFQ\_2: if you select the “Female Respondents” option at checkout, only women who completed (or partially completed) the Phase 1 Female Questionnaire will be included in your extract.



We’ll further restrict our sample by selecting only cases where RESULTFQ\_2 shows that the woman also completed the Phase 2 questionnaire. Notice that, in addition to each of the value 1 through 10, there are several non-response codes numbered 90 through 99. You’ll see similar values repeated across all IPUMS PMA variables, except that they will be left-padded to match the maximum width of a particular variable (e.g. 9999 is used for INTFQYEAR, which represents a 4-digit year for the Female Interview).

use "pma\_00126.dta", clear  
  
keep if sample\_1 == 85409  
  
tab resultfq\_2, m  
  
label list RESULTFQ\_2

.   
. use "pma\_00126.dta", clear  
  
.   
. keep if sample\_1 == 85409  
(23,387 observations deleted)  
  
.   
. tab resultfq\_2, m  
  
 result of female questionnaire | Freq. Percent Cum.  
----------------------------------------+-----------------------------------  
 completed | 5,491 80.87 80.87  
 not at home | 78 1.15 82.02  
 postponed | 22 0.32 82.34  
 refused | 66 0.97 83.31  
 partly completed | 12 0.18 83.49  
 respondent moved | 15 0.22 83.71  
 incapacitated | 19 0.28 83.99  
 not interviewed (female questionnaire) | 4 0.06 84.05  
not interviewed (household questionnair | 192 2.83 86.88  
 niu (not in universe) | 399 5.88 92.75  
 . | 492 7.25 100.00  
----------------------------------------+-----------------------------------  
 Total | 6,790 100.00  
  
.   
. label list RESULTFQ\_2  
RESULTFQ\_2:  
 1 completed  
 2 not at home  
 3 postponed  
 4 refused  
 5 partly completed  
 6 respondent death  
 7 respondent moved  
 8 household moved  
 10 incapacitated  
 90 other  
 95 not interviewed (female questionnaire)  
 96 not interviewed (household questionnaire)  
 99 niu (not in universe)  
  
.

Possible **non-response codes** include:

* 95 Not interviewed (female questionnaire)
* 96 Not interviewed (household questionnaire)
* 97 Don’t know
* 98 No response or missing
* 99 NIU (not in universe)

A missing value in an IPUMS extract indicates that a particular variable is not provided for a selected sample. In a “wide” **Longitudinal** extract, it may also signify that a particular person was not included in the data from a particular phase. Here, an missing result in RESULTFQ\_2 indicates that a Female Respondent from Phase 1 was not found in Phase 2.

You can drop incomplete Phase 2 female responses as follows:

use "pma\_00126.dta", clear  
  
keep if sample\_1 == 85409  
  
keep if resultfq\_2 == 1  
  
tab resultfq\_1 resultfq\_2,m

.   
. use "pma\_00126.dta", clear  
  
.   
. keep if sample\_1 == 85409  
(23,387 observations deleted)  
  
.   
. keep if resultfq\_2 == 1  
(1,299 observations deleted)  
  
.   
. tab resultfq\_1 resultfq\_2,m  
  
 | result of  
 | female  
 | questionna  
 result of female | ire  
 questionnaire | completed | Total  
----------------------+-----------+----------  
 completed | 5,487 | 5,487   
 partly completed | 4 | 4   
----------------------+-----------+----------  
 Total | 5,491 | 5,491   
  
.

Generally, we will combine both filtering steps together in a single function like so:

use "pma\_00126.dta", clear  
  
keep if sample\_1 == 85409  
  
keep if inlist(resident\_1,11,22) & inlist(resident\_2,11,22) & resultfq\_2 == 1  
  
tab resultfq\_1 resultfq\_2,m

.   
. use "pma\_00126.dta", clear  
  
.   
. keep if sample\_1 == 85409  
(23,387 observations deleted)  
  
.   
. keep if inlist(resident\_1,11,22) & inlist(resident\_2,11,22) & resultfq\_2 == 1  
(1,578 observations deleted)  
  
.   
. tab resultfq\_1 resultfq\_2,m  
  
 | result of  
 | female  
 | questionna  
 result of female | ire  
 questionnaire | completed | Total  
----------------------+-----------+----------  
 completed | 5,208 | 5,208   
 partly completed | 4 | 4   
----------------------+-----------+----------  
 Total | 5,212 | 5,212   
  
.

In upcoming chapters, we’ll use the remaining cases to show how PMA generates key indicators for **contraceptive use status** and **family planning intentions and outcomes**. The summary report for each country includes measures disaggregated by demographic variables like:

* [MARSTAT](https://pma.ipums.org/pma-action/variables/MARSTAT#codes_section) - marital status
* [EDUCATT](https://pma.ipums.org/pma-action/variables/EDUCATT#codes_section) and [EDUCATTGEN](https://pma.ipums.org/pma-action/variables/EDUCATTGEN#codes_section) - highest attended level of education[16](#fn16)
* [AGE](https://pma.ipums.org/pma-action/variables/AGE#codes_section) - age[17](#fn17)
* [WEALTHQ](https://pma.ipums.org/pma-action/variables/WEALTHQ#codes_section) and [WEALTHT](https://pma.ipums.org/pma-action/variables/WEALTHT#codes_section) - household wealth quintile or tertile[18](#fn18)
* [URBAN](https://pma.ipums.org/pma-action/variables/URBAN#codes_section) and [SUBNATIONAL](https://pma.ipums.org/pma-action/variables/SUBNATIONAL#codes_section) - geographic location[19](#fn19)

Dean, Natalie, and Marcello Pagano. 2015. “Evaluating Confidence Interval Methods for Binomial Proportions in Clustered Surveys.” *Journal of Survey Statistics and Methodology* 3 (4): 484–503. <https://doi.org/10.1093/jssam/smv024>.

1. In addition to these three in-person surveys, PMA also conducted telephone interviews with panel members focused on emerging issues related to the COVID-19 pandemic in 2020. These telephone surveys are already available for several countries - see our series on [PMA COVID-19 surveys](../../index.html#category:COVID-19) for details.[↩︎](#fnref1)

1. [Displaced GPS coordinates](https://tech.popdata.org/pma-data-hub/posts/2021-10-15-nutrition-climate/PMA_displacement.pdf) for the centroid of each EA are available for most samples [by request](https://www.pmadata.org/data/request-access-datasets) from PMA. IPUMS PMA provides shapefiles for PMA countries [here](https://pma.ipums.org/pma/gis_boundary_files.shtml).[↩︎](#fnref2)

1. Questionnaires administered in each country may vary from this **Core Household Questionnaire** - [click here](https://pma.ipums.org/pma/enum_materials.shtml) for details.[↩︎](#fnref3)

1. Questionnaires administered in each country may vary from this **Core Female Questionnaire** - [click here](https://pma.ipums.org/pma/enum_materials.shtml) for details.[↩︎](#fnref4)

1. The “study area” is area within which resident enumerators should attempt to find panel women that have moved out of their Phase 1 dwelling. This may extend beyond the woman’s original EA as determined by in-country administrators - see [PMA Phase 2 and Phase 3 Survey Protocol](https://www.pmadata.org/data/survey-methodology) for details.[↩︎](#fnref5)

1. In cases where no Phase 1 household members remained in the dwelling at Phase 2, women from the household are considered lost to follow-up (LTFU). A panel member is also considered LTFU if a Phase 2 Household Questionnaire was not completed, if she declined to participate, or if she was deceased or otherwise unavailable.[↩︎](#fnref6)

1. For thorough discussion of the types of weights available in both R and Stata, we recommend [this blog post](https://notstatschat.rbind.io/2020/08/04/weights-in-statistics/) by Dr. Lumley.[↩︎](#fnref7)

1. HQWEIGHT reflects the [calculated selection probability](https://pma.ipums.org/pma/resources/documentation/weighting_memo.pdf) for a household in an EA, normalized at the population-level. Users intending to estimate population-level indicators for *households* should restrict their sample to one person per household via [LINENO](https://pma.ipums.org/pma-action/variables/LINENO#description_section) - see [household weighting guide](https://pma.ipums.org/pma/weightguide.shtml#hh) for details.[↩︎](#fnref8)

1. FQWEIGHT adjusts HQWEIGHT for female non-response within the EA, normalized at the population-level - see [female weighting guide](https://pma.ipums.org/pma/weightguide.shtml#female) for details.[↩︎](#fnref9)

1. POPWT can be used to estimate population-level *counts* - [click here](https://pma.ipums.org/pma/population_weights.shtml) or check out [this video](https://www.youtube.com/watch?v=GnCq26t4zgM) for details.[↩︎](#fnref10)

1. See Dean & Pagano ([2015](#ref-Dean-Pagano)) for discussion. If you estimate a proportion where the sample have either 0% or 100% of respondents with the outcome, then as of the time of this writing, neither Stata nor R's `survey` package will report a confidence interval. Here at Biostat Global Consulting, we have written programs in both Stata and R that yield meaningful confidence intervals for any proportion. Those programs are made freely available as part of software we have written for the World Health Organization. If you want to learn more about them, write to us at Dale.Rhoda@biostatglobal.com or Caitlin.Clary@biostatglobal.com.[↩︎](#fnref11)

1. See [Stata's help for the proportion command](https://www.stata.com/manuals/rproportion.pdf) for a complete list of citype() methods.[↩︎](#fnref12)

1. As we’ll see in an upcoming post, women are considered “lost to follow-up” if they moved outside the study area after Phase 1. Therefore, EAID\_1 and EAID\_2 are identical for all panel members: you can use either one to identify sample clusters.[↩︎](#fnref13)

1. As with [EAID](https://pma.ipums.org/pma-action/variables/EAID#codes_section), you may use either STRATA\_1 or STRATA\_2 if your analysis is restricted to panel members.[↩︎](#fnref14)

1. Some survey materials guide analysts to only report results for estimates or tests where the relative standard error (100 x standard error of the estimate / the estimate itself) is no greater than 30% or where there are at least twelve degrees of freedom. See the Centers for Disease Control and Prevention's [NHANES CMS tutorial](https://www.cdc.gov/nchs/tutorials/nhanes-cms/variance/variance.htm).[↩︎](#fnref15)

1. Levels in EDUCATT may vary by country; EDUCATTGEN recodes country-specific levels in four general categories.[↩︎](#fnref16)

1. Ages are frequently reported in five-year groups: 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, and 45-49.[↩︎](#fnref17)

1. Households are divided into quintiles/tertiles relative to the distribution of an asset [SCORE](https://pma.ipums.org/pma-action/variables/SCORE#description_section) weighted for all sampled households. For subnationally-representative samples (DRC and Nigeria), separate wealth distributions are calculated for each sampled region.[↩︎](#fnref18)

1. SUBNATIONAL includes subnational regions for all sampled countries; country-specific variables are also available on the [household - geography](https://pma.ipums.org/pma-action/variables/group?id=hh_geo) page.[↩︎](#fnref19)

### References