

CONTENTS

1 Introduction	1
1.1 IPUMS PMA data in Stata	2
1.2 PMA Background	3
1.3 Sampling	4
1.4 Inclusion Criteria for Analysis	8
1.5 Survey Design Elements	13
1.5.1 Set survey design	14
1.5.2 Design Effect	17
1.5.3 Sample strata for DRC	21
2 Longitudinal Data Extracts	28
2.1 Sample Selection	29
2.2 Variable Selection	33
2.2.1 Codes	35
2.2.2 Variable Description	38
2.2.3 Comparability Notes	39
2.2.4 Sample Universe	40
2.2.5 Availability Across Samples	41
2.2.6 Questionnaire Text	42
2.2.7 Checkout	43
2.3 Data for Stata Users	44
2.4 Long Data Structure	46
2.5 Wide Data Structure	49
2.6 Which format is best for me?	52

1 INTRODUCTION

Performance Monitoring for Action (PMA) uses innovative mobile technology to support low-cost, rapid-turnaround surveys that monitor key health and development indicators.

PMA surveys collect longitudinal data throughout a country at the household and health facility levels by female data collectors, known as resident enumerators, using mobile phones. The survey collects information from the same women and households over time for regular tracking of progress and for understanding the drivers of contraceptive use dynamics. The data are rapidly validated, aggregated, and prepared into tables and graphs, making results quickly available to stakeholders. PMA surveys can be integrated into national monitoring and evaluation systems using a low-cost, rapid-turnaround survey platform that can be adapted and used for various health data needs.

The PMA project is implemented by local partner universities and research organizations who train and deploy the cadres of female resident enumerators.

The purpose of this manual is to provide guidance on the analysis of **harmonized longitudinal data** for a panel of women age 15-49 surveyed by PMA and published in partnership with [IPUMS PMA](#). IPUMS provides census and survey products from around the world in an integrated format, making it easy to compare data from multiple countries. IPUMS PMA data are available free of charge, subject to terms and conditions: please [register here](#) to request access to the data featured in this guide.

PMA has also published a guide to cross-sectional analysis in both English and French.

This manual provides reproducible coding examples in the statistical software program [Stata](#). You can download .do files containing all of the code needed to reproduce these examples on our [GitHub page](#).

R users: a companion manual for IPUMS PMA longitudinal analysis is also available with coding examples written in R. Additionally, the [IPUMS PMA data analysis blog](#) includes an online version of each chapter and posts on a range of other topics updated every two weeks.

1.1 IPUMS PMA DATA IN STATA

The first two chapters of this manual introduce new users to [PMA longitudinal data](#) and the [IPUMS PMA website](#), respectively. After demonstrating how to obtain an IPUMS PMA data extract, the remaining chapters feature extensive data analysis examples written in Stata.

To follow along, you'll need to purchase and download the appropriate version of Stata for your computer's operating system at [stata.com](#). Discounted licences are available for students and for faculty and staff at participating institutions: learn more [here](#).



For a general introduction to analysis of IPUMS PMA data in Stata, visit the [IPUMS PMA Support](#) page, where you'll find links to video tutorials and data exercises written for Stata users. Similar resources are available for users of R, SPSS, and SAS.

Questions for Dale:

- Did you find that you needed a particular *version* to complete all of our exercises
- Are any supplementary packages needed?
- In the R version, I list some ways to get help with R. Do you have any favorite resources for getting help with Stata?

1.2 PMA 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** (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** via population-level estimates for several **core indicators**.

Beginning in 2019, PMA surveys were redesigned under a renewed partnership called **Family Planning 2030** (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

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

¹In 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 - the IPUMS PMA blog series on **PMA COVID-19 surveys** covers this topic in detail.

1.3 SAMPLING

PMA panel data includes a mixture of **nationally representative** and **sub-nationally representative** samples. The panel study consists of three data collection phases, each spaced one year apart.

As of this writing, IPUMS PMA has released data from the first *two* phases for four countries where Phase 1 data collection began in 2019; IPUMS PMA has released data from only the *first* phase for three countries where Phase 1 data collection began in August or September 2020. Phase 3 data collection and processing is currently underway.

Sample	Phase 1 Data Collection*	Now Available from IPUMS PMA		
		Phase 1	Phase 2	Phase 3
Burkina Faso	Dec 2019 - Mar 2020	x	x	
Cote d'Ivoire	Sep 2020 - Dec 2020	x		
DRC - Kinshasa	Dec 2019 - Feb 2020	x	x	
DRC - Kongo Central	Dec 2019 - Feb 2020	x	x	
India - Rajasthan	Aug 2020 - Oct 2020	x		
Kenya	Nov 2019 - Dec 2019	x	x	
Nigeria - Kano	Dec 2019 - Jan 2020	x	x	
Nigeria - Lagos	Dec 2019 - Jan 2020	x	x	
Uganda	Sep 2020 - Oct 2020	x		

*Each data collection phase is spaced one year apart

PMA uses a multi-stage clustered sample design, with stratification at the urban-rural level or by sub-region. Sample clusters - called **enumeration areas** (EAs) - are provided by the national statistics agency in each country.² These EAs are sampled using a *probability proportional to size* (PPS) method relative to the population distribution in each stratum.

Resident enumerators are women over age 21 living in (or near) each EA who hold at least a high school diploma.

²Displaced GPS coordinates for the centroid of each EA are available for most samples [by request](#) from PMA. IPUMS PMA provides shapefiles for PMA countries [here](#).

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**³ 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**.⁴

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 but at least one member of the Phase 1 household remained and could help resident enumerators locate her new dwelling.⁶

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

samedwelling
indicates whether a Phase 2 female respondent resided in her Phase 1 dwelling or a new one.

panelwoman
indicates whether a Phase 2 household member completed the Phase 1 Female Questionnaire.

³Questionnaires administered in each country may vary from this Core Household Questionnaire - [click here](#) for details.

⁴Questionnaires administered in each country may vary from this Core Female Questionnaire - [click here](#) for details.

⁵The “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](#) for details.

⁶In 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.

When you select the new **Longitudinal** sample option from IPUMS PMA, 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).

The screenshot shows a web browser window for 'IPUMS PMA: select samples'. The URL is pma.ipums.org/pma-action/samples. The page title is 'SELECT SAMPLES'. It features the IPUMS PMA logo and navigation links for HOME, SELECT DATA, MY DATA, and SUPPORT. A red arrow points to the 'Long' radio button under the 'Longitudinal' sample type. The 'SUBMIT SAMPLE SELECTIONS' button is visible on the right. Below the sample selection, there is a section for 'FAMILY PLANNING - PERSON' with a 'Documentation' link and filter options for 'All Samples (long)', 'Burkina Faso', and '2020 - 2021'.

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 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).

You'll find PMA cross-sectional samples dating back to 2013 if you select the **Cross-sectional** sample option from IPUMS PMA.

The screenshot shows a web browser window for 'IPUMS PMA: select samples'. The URL is pma.ipums.org/pma-action/samples. The page title is 'SELECT SAMPLES'. It features the IPUMS PMA logo and navigation links for HOME, SELECT DATA, MY DATA, and SUPPORT. A note says 'Variable documentation on the web site can be filtered to display only material corresponding to chosen datasets ([more information](#) on this feature)'. Below, it says 'You may select any of the below datasets for browsing. Please [log in](#) to see which samples you are authorized to include in extracts.' There are two radio button options: 'Cross-sectional' (selected, indicated by a red arrow) and 'Longitudinal'. A purple 'SUBMIT SAMPLE SELECTIONS' button is visible. A section titled 'FAMILY PLANNING - PERSON' contains a checkbox for 'All Samples' and a timeline from 2015 to 2021 with checkboxes for each year.

1.4 INCLUSION CRITERIA FOR ANALYSIS

Several chapters in this manual feature 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
- DRC - Kinshasa
- DRC - Kongo Central
- Kenya
- Nigeria - Kano
- Nigeria - Lagos

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'll remove *de jure* cases recorded in the variable `resident`.

We will demonstrate how to request and download an IPUMS PMA data extract in Chapter 2.

For example, let's consider a "wide" format data extract containing Phase 1 and Phase 2 respondents to the Female Questionnaire from Burkina Faso. You'll find 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  
  
table ( resident_1 ) () (), nototals missing zeroCounts
```

Variable names in a "wide" extract have a numeric suffix for their data collection phase. `resident_1` is the Phase 1 version of `resident`, while `resident_2` comes from Phase 2.

	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

This extract includes 174 women who are not members of the *de facto* population because they did not sleep in the sampled household during the night before the Phase 1 interview.

Let's turn to Phase 2:

```
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 extract also includes 230 women who are not members of the *de facto* population because they did not sleep in the sampled household during the night before the Phase 2 interview. Moreover, there are 492 blank values in `resident_2` representing women who were lost to follow-up after Phase 1.

The *de facto* population is represented in both variables by 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
```

	Resident type - Phase 2	
	Visitor	Usual
Resident type - Phase 1		
Visitor	56	39
Usual	17	5,855

Additionally, PMA reports only include women who completed (or partially completed) both Female Questionnaires. This information is reported in `resultfq`. 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
```

```
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 blank 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, a blank value appearing 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 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 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 questionnaire	ire completed	Total
completed	5,208	5,208	
partly completed	4	4	
Total	5,212	5,212	

In subsequent analyses, 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** - marital status
- **educatt** and **educattgen** - highest attended level of education⁷
- **age** - age
- **wealthq** and **wealtht** - household wealth quintile or tertile⁸
- **urban** and **subnational** - geographic location⁹

⁷Levels in **educatt** may vary by country; **educattgen** recodes country-specific levels in four general categories.

⁸Households are divided into quintiles/tertiles relative to the distribution of an asset **score** weighted for all sampled households. For subnationally-representative samples (DRC and Nigeria), separate wealth distributions are calculated for each sampled region.

⁹**subnational** includes subnational regions for all sampled countries; country-specific variables are also available on the **household - geography** page.

1.5 SURVEY DESIGN ELEMENTS

Throughout this guide, we'll demonstrate how to incorporate PMA sampling weights and information about its stratified cluster sampling procedure into your analysis. This section describes how to use survey weights, cluster IDs, and sample strata in Stata.

Let's return to the data extract described in the previous section, which includes Phase 1 and Phase 2 respondents to the Female Questionnaire from Burkina Faso. As a reminder: we'll drop women who are non members of the *de facto* population and those who did not complete all or part the Female Questionnaire in both phases.

We will demonstrate how to request and download an IPUMS PMA data extract in Chapter 2.

```
use "pma_00126.dta", clear  
keep if inlist(resident_1,11,22) & inlist(resident_2,11,22) & resultfq_2 == 1
```

(1,578 observations deleted)

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:

- **hqweight** can be used to generate cross-sectional population estimates from questions on the Household Questionnaire.¹¹
- **fqweight** can be used to generate cross-sectional population estimates from questions on the Female Questionnaire.¹²
- **eaweight** can be used to compare the selection probability of a particular household with that of its EA.

A fourth Family Planning survey weight, **popwt**, is currently available only for **Cross-sectional** data extracts.¹⁰

Additionally, PMA created a new weight, **panelweight**, which should be used in longitudinal analyses spanning multiple phases, as it adjusts for loss to follow-up.

¹⁰ POPWT can be used to estimate population-level counts - [click here](#) or view [this video](#) for details.

¹¹ HQWEIGHT reflects the calculated selection probability 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** - see [household weighting guide](#) for details.

¹² FQWEIGHT adjusts HQWEIGHT for female non-response within the EA, normalized at the population-level - see [female weighting guide](#) for details.

1.5.1 Set survey design

In the following example, we'll show how to use `panelweight` 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 the “wide” extract featured here, you'll find it in `cp_1` for Phase 1, and in `cp_2` for Phase 2.

Here is how to create an *unweighted* crosstab for `cp_1` and `cp_2`:

```
table ( cp_1 ) ( cp_2 ) (), nototals missing zero
```

		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 estimate a population percentage, we'll need to tell Stata that we are working with a sample survey dataset and stipulate the sample design (specify which variables identify survey weights, strata, and clusters). This is accomplished with the `svyset` command.

We use `eaid_1` as the cluster ID¹³ and `strata_1` as the stratum ID¹⁴ and `panelweight` holds the survey weight. We also make a binary variable indicating which women were using contraception in both phases.

```
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_1) weight(panelweight)
```

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.

¹³Because women are considered “lost to follow-up” if they moved outside the study area, `eaid_1` and `eaid_2` are identical for all panel members: you can use either one to identify sample clusters.

¹⁴As with `eaid`, you may use either `strata_1` or `strata_2` if your analysis is restricted to panel members

```

Sampling weights: <none>
      VCE: linearized
Single unit: missing
Strata 1: strata_1
Sampling unit 1: eaid_1
      FPC 1: <zero>
Weight 1: panelweight

```

Now, we can use this survey design information to obtain a population estimate for the proportion of women who used family planning in both phases.

```
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]	
<hr/>				
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 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

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.

¹⁵Some survey materials guide analysts to only report results for estimates or tests where the relative standard error ($100 \times$ 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](#).

Stata's default confidence interval is the so-called "logit interval" which is one of several possibilities.¹⁶ 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 `ci_type()` option to the `svy: proportion` command (e.g., `ci_type(wilson)` or `ci_type(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%)."

¹⁶See Dean & Pagano [-@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.

1.5.2 Design Effect

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`.

```
estat effects
```

		Linearized		
		Proportion	std. err.	DEFF
cp_both				DEFT
No		.8122041	.012815	5.6052
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`.

Here we create `y` for the larger, complex sample:

```
clear  
set obs 5207
```

Number of observations (_N) was 0, now 5,207.

```
gen y = 0  
replace y = 1 if _n < 0.188 * 5207  
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

	y	Linearized			Logit	
		Proportion	std. err.	[95% conf. interval]		
	0	.8121759	.0054131	.8013328	.8225583	
	1	.1878241	.0054131	.1774417	.1986672	

And here we create y for the smaller, simple sample:

```
clear
set obs 929
```

Number of observations (_N) was 0, now 929.

```
gen y = 0
replace y = 1 if _n < 0.188 * 929
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. That is: we'll divide the difference between the upper and lower limits of our 95% confidence interval from the complex data by that of the simple random sample. We'll see that it is approximately equal to 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](#).

1.5.3 Sample strata for DRC

This syntax and svyset command worked well for Burkina Faso, but take note: the variable **strata** 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. Here, we again include only panel members who completed all or part of the female questionnaire in both phases, and who slept in the household during the night before the interview:

```
use "pma_00153.dta", clear  
keep if inlist(resident_1,11,22) & inlist(resident_2,11,22) & resultfq_2 == 1
```

(12,453 observations deleted)

Notice that **strata_1** lists the sample strata for all values of **country** except for DRC, where the variable is missing.

```
table ( strata_1 ) () ( country ), 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		

strata	
lagos, nigeria	1,088
kano – urban	437
kano – rural	561

We can replace those values with numeric codes from the variable `geocd`:

```
table ( geocd ) if country == 2, nototals missing zeroCounts
```

		Frequency
province, congo dr		
kinshasa	1,973	
kongo central	1,514	

```
tab geocd
```

province,	Freq.	Percent	Cum.
congo dr			
kinshasa	1,973	56.58	56.58
kongo central	1,514	43.42	100.00

Total	3,487	100.00	
-------	-------	--------	--

```
tab geocd, nolabel
```

province,	Freq.	Percent	Cum.
congo dr			
1	1,973	56.58	56.58
2	1,514	43.42	100.00

Total	3,487	100.00	
-------	-------	--------	--

Note that the values of `geocd` are distinct from the values of `strata_1`: 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`.

```
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

clonevar strata_recode = strata_1

(3,487 missing values generated)

replace strata_recode = geocd **if** country == 2

(3,487 real changes made)

Copy the value labels from strata_1 into a new label strata_recode and update it with the labels from geocd. This leaves no blank values in strata_recode.

```

label copy STRATA_1 strata_recode, replace
label define strata_recode 1 "Kinshasa, DRC" 2 "Kongo Central, DRC", modify
label values strata_recode strata_recode
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.

First, we'll create 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, we 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

	cp_both@pop_numeric	Linearized		Logit
		Proportion	std. err.	[95% conf. interval]
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

2 LONGITUDINAL DATA EXTRACTS

This chapter provides a guided tour of the IPUMS PMA data extract system. While you may also access the original data directly from our partners at PMA, harmonized data from IPUMS have a few additional features. For instance, you can request an extract that:

- includes samples from multiple countries
- includes samples from multiple rounds of data collection
- are formatted in either **long** or **wide** format

IPUMS PMA also makes it easy to switch between multiple **units of analysis** covered in PMA surveys. In addition to the data featured in this guide, you'll find surveys representing:

- Service Delivery Points (SDPs)
- Client Exit Interviews conducted at SDPs
- Participants in special surveys covering topics like COVID-19, nutrition, and maternal & newborn health

To get started with a longitudinal data extract, you'll need to select the **Family Planning** topic under the **Person** unit of analysis.

Register here to access IPUMS PMA data at no cost. See our user guide for details.

A video tour of the longitudinal extract system is available here on the IPUMS PMA Youtube channel.

The screenshot shows a web browser window for the IPUMS PMA website. The URL in the address bar is <https://pma.ipums.org/pma-action/variables/group>. The page title is "IPUMS PMA: vars by group". On the left, there's a sidebar with a "SELECT SAMPLES" button and the text "Select samples and variables". The main content area has a header "CHOOSE THE TOPIC FOR DATA BROWSING". Below it is a grid of categories. The first row has three columns: "PERSON", "FAMILY PLANNING", and "NUTRITION". The "FAMILY PLANNING" button is circled in red. The second row has two columns: "CLIENT EXIT INTERVIEW" and "COVID-19". The third row has three columns: "SERVICE DELIVERY POINT", "FAMILY PLANNING", and "NUTRITION". The fourth row has two columns: "INFANT" and "MATERNAL AND NEWBORN HEALTH". To the right of the grid is a "DATA CART" section showing "0 VARIABLES" and "0 SAMPLES". At the bottom of the page is a footer with the URL again: https://pma.ipums.org/pma-action/variables/group?unit_of_analysis=person.

2.1 SAMPLE SELECTION

Once you've selected the **Family Planning** option, you'll next need to choose between cross-sectional or longitudinal samples. Cross-sectional samples are selected by default; these are nationally or sub-nationally representative samples collected each year dating backward as far as 2013.

The screenshot shows the IPUMS PMA website at pma.ipums.org/pma-action/samples. The top navigation bar includes links for LOG IN, REGISTER, GLOBAL HEALTH, and IPUMS.ORG. The main header features the IPUMS PMA logo and the text 'PERFORMANCE MONITORING FOR ACTION'. Below the header, there are links for HOME, SELECT DATA, MY DATA, and SUPPORT. The main content area is titled 'SELECT SAMPLES'. A note states: 'Variable documentation on the web site can be filtered to display only material corresponding to chosen datasets ([more information](#) on this feature). You may select any of the below datasets for browsing. Please [log in](#) to see which samples you are authorized to include in extracts.' A radio button group allows selecting 'Cross-sectional' (which is checked) or 'Longitudinal'. To the right is a 'SUBMIT SAMPLE SELECTIONS' button. The main section is titled 'FAMILY PLANNING - PERSON'. It contains a table where rows represent countries and columns represent years from 2015 to 2021. Each cell contains a checkbox for selecting a specific dataset. The table includes rows for Burkina Faso, Congo (Democratic Republic), and other countries like India and Uganda. A legend indicates that some datasets are Phase 2 (e.g., 2020 P2, 2019b P1, etc.) and others are Phase 1 (e.g., 2018 R6, 2017 R5, etc.).

Longitudinal samples are only available from 2019 onward, and they include all of the available phases for each sampled country (sub-nationally representative samples for DRC and Nigeria are listed separately). You'll only find longitudinal samples for countries where Phase 2 data has been made available; Phase 1 data for Cote d'Ivoire, India, and Uganda can currently be found under the Cross-sectional sample menu (Phase 2 data will be released soon!).

Annual cross-sectional samples are also available for each of the countries participating in the new PMA panel study. See our [last post](#) for details.

Clicking the Longitudinal button reveals options for either **long** or **wide** format. You'll find the same samples available in either case.

Important: if you decide to change formats after selecting variables, your Data Cart will be emptied and you'll need to begin again from scratch.

The screenshot shows the 'SELECT SAMPLES' page of the IPUMS PMA website. At the top, there are navigation links for 'LOG IN | REGISTER | GLOBAL HEALTH | IPUMS.ORG'. Below the header, the IPUMS PMA logo is displayed, followed by 'PERFORMANCE MONITORING FOR ACTION', 'HOME | SELECT DATA | MY DATA | SUPPORT', and a search bar labeled 'Guest'.

The main section is titled 'SELECT SAMPLES'. It contains a note about variable documentation and dataset filtering. Below this, a message says 'You may select any of the below datasets for browsing. Please [log in](#) to see which samples you are authorized to include in extracts.' A red oval highlights the 'Longitudinal' radio button, which is selected. Other options shown are 'Cross-sectional' (unchecked), 'Long' (unchecked), and 'Wide' (unchecked). To the right of these options is a 'SUBMIT SAMPLE SELECTIONS' button.

The 'FAMILY PLANNING - PERSON' section includes a 'Documentation' table with two columns. The first column lists sample types: 'All Samples (wide)', 'Burkina Faso', 'Congo (Democratic Republic)', 'Kenya', and 'Nigeria'. The second column lists corresponding years: '2020 - 2021', '2019b - 2020b', '2019a - 2020a', '2019 - 2020', '2019b - 2020b', and '2019a - 2020a'. An arrow points to the 'All Samples (wide)' checkbox. Below the table is a 'Sample Members' section with four radio button options: 'Female Respondents' (selected), 'Female Respondents and Household Members', 'Female Respondents and Female Non-respondents', and 'All Cases (Respondents and Non-respondents to Household and Female Questionnaires)'. A second 'SUBMIT SAMPLE SELECTIONS' button is located at the bottom of this section.

At the bottom of the page, there is a footer note: 'SUPPORTED BY: THE BILL & MELINDA GATES FOUNDATION, PMA, STAT/TRANSFER, AND UNIVERSITY OF MINNESOTA.' and a copyright notice: 'COPYRIGHT © MINNESOTA POPULATION CENTER, UNIVERSITY OF MINNESOTA.'

After you've selected one of the available longitudinal formats, choose one or more samples listed below. There are also several Sample Members options listed.

The screenshot shows a web browser window titled "IPUMS PMA: select samples". The URL is "pma.ipums.org/pma-action/samples". The page has a "Documentation" section with a list of checked boxes for "All Samples (wide)", "Burkina Faso", "Congo (Democratic Republic)", "Kenya", and "Nigeria", along with their corresponding years: "2020 - 2021", "2019b - 2020b", "2019a - 2020a", "2019 - 2020", "2019b - 2020b", and "2019a - 2020a". Below this is a "Sample Members" section with a red oval circling the first option: "Female Respondents". Other options in this section are "Female Respondents and Household Members", "Female Respondents and Female Non-respondents", and "All Cases (Respondents and Non-respondents to Household and Female Questionnaires)". At the bottom right is a purple "SUBMIT SAMPLE SELECTIONS" button. At the very bottom of the page, there is a small footer with the text "SUPPORTED BY: THE BILL & MELINDA GATES FOUNDATION, PMA, STAT/TRANSFER, AND UNIVERSITY OF MINNESOTA." and "COPYRIGHT © MINNESOTA POPULATION CENTER, UNIVERSITY OF MINNESOTA".

Female Respondents only includes women who completed *all or part* of a Female Questionnaire. This option selects all members of the panel study. In addition, it includes women who only participated in only one phase - we will demonstrate how to identify and drop these cases below.¹⁷

Female Respondents and Female Non-respondents includes all women who were eligible to participate in a Female Questionnaire. Eligible women are those age 15-49 who were listed on the roster collected in a Household Questionnaire. If an eligible woman declined the Female Questionnaire or was not available, variables associated with that questionnaire will be coded “Not interviewed (female questionnaire)”.

panelwoman indicates whether an individual is a member of the panel study.

eligible indicates whether an individual was eligible for the female questionnaire.

¹⁷Women who completed all or part of the Female Questionnaire in *more than one phase* of the study are considered **panel members**. Women who completed it only at Phase 1 are included in a longitudinal extract, but they are not **panel members**. Likewise, women who completed it for the first time at Phase 2 are included, but are not **panel members** if they 1) will reach age 50 before Phase 3, or 2) declined the invitation to participate again in Phase 3.

Female Respondents and Household Members adds records for all other members of a Female Respondent's household. These household members did not complete the Female Questionnaire, but were listed on the household roster provided by the respondent to a Household Questionnaire. Basic **demographic** variables are available for each household member, as are common **wealth**, **water**, **sanitation**, and other variables shared for all members of the same household.

All Cases includes all members listed on the household roster from a Household Questionnaire. If the Household Questionnaire was declined or if no respondent was available, any panel member appearing in other phases of the study will be coded "Not interviewed (household questionnaire)" for variables associated with the missing Household Questionnaire.

After you've selected samples and sample members for your extract, click the "Submit Sample Selections" button to return to the main data browsing menu.

resultfq indicates whether an individual completed the Female Questionnaire.

resulthq indicates whether a member of the individual's household completed the Household Questionnaire.

2.2 VARIABLE SELECTION

You can browse IPUMS PMA variables by topic or alphabetically by name, or you can search for a particular term in a variable name, label, value labels, or description.

The screenshot shows the IPUMS PMA website interface for variable selection. At the top, there is a navigation bar with links for 'LOG IN | REGISTER | GLOBAL HEALTH | IPUMS.ORG'. On the right, a 'DATA CART' section shows '0 VARIABLES' and '6 SAMPLES' with a 'VIEW CART' button. Below the navigation, the IPUMS PMA logo is displayed. A banner indicates 'CURRENTLY BROWSING: "FAMILY PLANNING - PERSON"' with a 'CHANGE' link. The main content area is titled 'SELECT VARIABLES' with tabs for 'TOPICS' (selected), 'A-Z', and 'SEARCH'. Under 'TOPICS', 'FAMILY PLANNING' is selected, showing a dropdown menu with options like 'FERTILITY PREFERENCES', 'SEXUAL BEHAVIOR', 'CURRENT OR RECENT FAMILY PLANNING USE', etc. Other topics listed include 'TECHNICAL', 'DEMOGRAPHICS (WOMEN)', 'HEALTH', 'ABORTION', 'HOUSEHOLD', 'WATER AND SANITATION', and 'COVID-19'. There are also sections for 'COPYRIGHT', 'MINNESOTA.', 'HELP', and 'COUNTRY ABBREVIATIONS'. A small note at the bottom left says 'Samples have been used in this analysis.'

In this example, we'll select the **Discontinuation of Family Planning** topic. The availability of each associated variable is shown in a table containing all of the samples we've selected.

- x indicates that the variable is available for *all phases*
- / indicates that the variable is available for *one phase*
- – indicates that the variable is not available for *any phase*

You can click the + button to add a variable to your cart, or click a variable name to learn more.

The screenshot shows the IPUMS PMA website interface. At the top, there's a navigation bar with links for LOG IN, REGISTER, GLOBAL HEALTH, and IPUMS.ORG. On the right, a "DATA CART" section shows 0 VARIABLES and 6 SAMPLES, with a "VIEW CART" button. Below the navigation, the IPUMS PMA logo is displayed. The main content area has a header "CURRENTLY BROWSING: 'FAMILY PLANNING - PERSON'" with a "CHANGE" link. There are three tabs: "SELECT VARIABLES" (highlighted), "DISPLAY OPTIONS", and "HELP". Under "SELECT VARIABLES", there are buttons for "TOPICS", "A-Z", and "SEARCH". A note says "AN 'X' INDICATES THE VARIABLE IS AVAILABLE IN THAT DATASET." The main table is titled "DISCONTINUATION OF FAMILY PLANNING VARIABLES (TOP)" and "LONGITUDINAL SAMPLES". It lists variables like FPSTOPMO, EPIMPREMOVEYR, and EPIMPRMYYUNAVAIL, along with their labels and availability across five datasets: BURKF, CONDR, CONDR, KENYA, NIGERA, and NIGERA. The table includes columns for Type (2020 - 2021, 2019a - 2020a, 2019b - 2020b, 2019 - 2020, 2019a - 2020a, 2019b - 2020b). At the bottom of the page, there's a footer with copyright information: "SUPPORTED BY: THE BILL & MELINDA GATES FOUNDATION, PMA, STAT/TRANSFER, AND UNIVERSITY OF MINNESOTA." and "COPYRIGHT © MINNESOTA POPULATION CENTER, UNIVERSITY OF MINNESOTA."

DISCONTINUATION OF FAMILY PLANNING VARIABLES (TOP)		LONGITUDINAL SAMPLES							
Add to cart	Variable	Variable Label	Type	BURKF 2020 - 2021	CONDR 2019a - 2020a	CONDR 2019b - 2020b	KENYA 2019 - 2020	NIGERA 2019a - 2020a	NIGERA 2019b - 2020b
+ FPSTOPMO	Month stopped using most recent method	P X / / X . X X							
+ FPSTOPYR	Year stopped using most recent method	P X X X . X X							
+ FPSTOPSECMC	Date stopped using recent method of FP in century month	P X X X . X X							
+ FPIMPREMOVEYR	Tried to remove implant in past 12 months	P X / / X / /							
+ EPIMPRMTRYLOC	Location of implant removal attempt	P X / / / /							
+ EPIMPRMYYCOST	Why implant not removed: Service cost	P X / / X / /							
+ EPIMPRMYYCOUND	Why implant not removed: Provider counseled against	P X / / X / /							
+ EPIMPRMYYCLOSED	Why implant not removed: Facility closed	P X / / X / /							
+ EPIMPRMYYOTH	Why implant not removed: Other	P X / / X / /							
+ EPIMPRMYYREFUSE	Why implant not removed: Provider refused	P X / / X / /							
+ EPIMPRMYYELSEWH	Why implant not removed: Referred elsewhere	P X / / X / /							
+ EPIMPRMYYRETURN	Why implant not removed: Told to return another day	P X / / X / /							
+ EPIMPRMYYTRAVEL	Why implant not removed: Travel cost	P X / / X / /							
+ EPIMPRMYYUNAVAIL	Why implant not removed: Qualified provider not available	P X / / X / /							
+ EPIMPRMYYUNSUCC	Why implant not removed: Failed attempt by provider	P X / / X / /							

2.2.1 Codes

Let's take a look at the variable **pregnant**. You'll find the variable name and label shown at the top of the page. Below, you'll see several tabs beginning with the **CODES** tab. For discrete variables, this tab shows all of the available codes and value labels associated with each response. You'll also see the same x, /, and – symbols in a table indicating the availability of each response in each sample.

“Case-count view” is not available for longitudinal samples, where each sample includes data from multiple phases. For cross-sectional samples, this option shows the frequency of each response.

The screenshot shows the IPUMS PMA website interface. At the top, there is a navigation bar with links for LOG IN, REGISTER, GLOBAL HEALTH, and IPUMS.ORG. On the left, the IPUMS PMA logo is displayed. In the center, the title "PREGNANT" is shown, along with links for ADD TO CART and CHANGE SAMPLES. To the right, a "DATA CART" section indicates 0 VARIABLES and 6 SAMPLES, with a VIEW CART button. Below the main title, there is a section titled "Codes and Frequencies". Under this section, there are two radio button options: "Category availability view" (selected) and "Case-count view (Unavailable for longitudinal samples)". A legend follows, listing five categories: Female Respondents (selected), Female Respondents and Household Members, Female Respondents and Female Non-respondents, All Cases (Respondents and Non-respondents to Household and Female Questionnaires), and Female Non-respondents. Below the legend, a note states: "An 'X' indicates the category is available for that sample". A table titled "LONGITUDINAL SAMPLES" is provided, showing availability across six countries: BURKF, CONDR, KENYA, NIGERA, and NIGERA. The table includes rows for codes 00 (No), 01 (Yes), 95 (Not interviewed (female questionnaire)), 96 (Not interviewed (household questionnaire)), 97 (Don't know), 98 (No response), and 99 (NIU (not in universe) or missing). The "Female Respondents" row is highlighted with a red oval. The "Case-count view" link is also highlighted with a red oval. At the bottom of the page, there is a footer with links for SUPPORTED BY: THE BILL & MELINDA GATES FOUNDATION, PMA, STAT/TRANSFER, AND UNIVERSITY OF MINNESOTA, and a copyright notice: COPYRIGHT © MINNESOTA POPULATION CENTER, UNIVERSITY OF MINNESOTA.

Above, there are no responses for “Not interviewed (female questionnaire)” and “Not interviewed (household questionnaire)”; this is because only samples members included in a “Female Respondents” extract are displayed by default. If we instead choose “All Cases”, this variable will include those response options because we’ll include every person listed on the household roster (even if the Household or Female Questionnaire was not completed).

PREGNANT

CATEGORIES

Category availability view

- Female Respondents
- Female Respondents and Household Members
- Female Respondents and Female Non-respondents
- All Cases (Respondents and Non-respondents to Household and Female Questionnaires)

An 'X' indicates the category is available for that sample

LONGITUDINAL SAMPLES						
Code	Label	BURKF	COND1	COND2	KENYA	NIGERA
20	- 21	X	X	X	X	X
19a	- 20a	X	X	X	X	X
19b	- 20b	X	X	X	X	X
19	- 20	X	X	X	X	X
19a	- 20a	X	X	X	X	X
19b	- 20b	X	X	X	X	X
95	Not interviewed (female questionnaire)	X	X	X	X	X
96	Not interviewed (household questionnaire)	X	X	X	X	X
97	Don't know	X	X	X	X	X
98	No response	X	/	:	X	X
99	NIU (not in universe) or missing	X	X	X	X	X

SUPPORTED BY: THE BILL & MELINDA GATES FOUNDATION, PMA, STAT/TRANSFER, AND UNIVERSITY OF MINNESOTA.

COPYRIGHT © MINNESOTA POPULATION CENTER, UNIVERSITY OF MINNESOTA.

The symbol / again indicates that a particular response is available for some - but not all - phases of the study. For PREGNANCY it indicates that one of the options was either unavailable or was not selected by any sample respondents in a particular phase. If a variable was not included in all phases of the study, all response options will be marked with this symbol. For example, consider the variable **covidconcern**, indicating the respondent's level of concern about becoming infected with COVID-19.

The screenshot shows the IPUMS PMA website interface. At the top, there is a navigation bar with links for LOG IN | REGISTER | GLOBAL HEALTH | IPUMS.ORG and a guest status indicator. On the right, there is a "DATA CART" section showing 0 VARIABLES and 6 SAMPLES, with a "VIEW CART" button. The main content area is titled "COVIDCONCERN". Below the title are buttons for "ADD TO CART" and "CHANGE SAMPLES". A sub-header says "Concerned about getting infected" and "Group: Perceptions around COVID". There is a legend at the top of the data table with tabs for CODES, DESCRIPTION, COMPARABILITY, UNIVERSE, AVAILABILITY, and QUESTIONNAIRE TEXT. The "CODES" tab is selected. The data table is titled "Codes and Frequencies" and includes a note: "An 'X' indicates the category is available for that sample". It has a header row for "LONGITUDINAL SAMPLES" with columns for BURKF, CONDR, CONDR, KENYA, NIGERA, and NIGERA. The data rows show the following availability across samples:

Code	Label	BURKF 20 - 21	CONDR 19a - 20a	CONDR 19b - 20b	KENYA 19 - 20	NIGERA 19a - 20a	NIGERA 19b - 20b
01	Not concerned	/	/	/	/	/	/
02	A little concerned	/	/	/	/	/	/
03	Concerned	/	/	/	/	/	/
04	Very concerned	/	/	/	/	/	/
05	Currently / previously infected with COVID-19	/	/	/	/	/	.
95	Not interviewed (female questionnaire)
96	Not interviewed (household questionnaire)
98	No response or missing	/	/	/	.	.	/
99	NIU (not in universe)

At the bottom of the page, there is a footer with the text "SUPPORTED BY: THE BILL & MELINDA GATES FOUNDATION, PMA, STAT/TRANSFER, AND UNIVERSITY OF MINNESOTA." and a copyright notice: "COPYRIGHT © MINNESOTA POPULATION CENTER, UNIVERSITY OF MINNESOTA."

Because Phase 1 questionnaires were administered prior to the emergence of COVID-19, this variable only appeared on Phase 2 questionnaires. The symbol / indicates limited availability across phases.

2.2.2 Variable Description

You'll find a detailed description for each variable on the **DESCRIPTION** tab. This tab also indicates whether a particular question appeared on the Household or Female Questionnaire.

The screenshot shows a web browser window for the IPUMS PMA website. The URL in the address bar is pma.ipums.org/pma-action/variables/PREGNANT#description_section. The page title is "IPUMS PMA: descr: PREGNANT". The top navigation bar includes links for LOG IN | REGISTER | GLOBAL HEALTH | IPUMS.ORG and a "Guest" account indicator. A "DATA CART" sidebar on the right shows 0 VARIABLES and 6 SAMPLES, with a "VIEW CART" button. The main content area is titled "PREGNANT" and describes it as a "Pregnancy status" variable under the "Core demographics" group. Below the title are buttons for "ADD TO CART" and "CHANGE SAMPLES". A horizontal navigation bar at the top of the content area includes tabs for CODES, DESCRIPTION (which is selected), COMPARABILITY, UNIVERSE, AVAILABILITY, and QUESTIONNAIRE TEXT. The "DESCRIPTION" tab contains the following text:

Description

PREGNANT indicates whether or not the woman was pregnant at the time of the interview.

The question associated with this variable was included in the female questionnaire.

At the bottom of the page, there is a footer note: "SUPPORTED BY: THE BILL & MELINDA GATES FOUNDATION, PMA, STAT/TRANSFER, AND UNIVERSITY OF MINNESOTA." and a copyright notice: "COPYRIGHT © MINNESOTA POPULATION CENTER, UNIVERSITY OF MINNESOTA."

2.2.3 Comparability Notes

The **COMPARABILITY** tab describes important differences between samples. Additionally, it may contain information about similar variables appearing in **DHS** samples provided by **IPUMS DHS**.

The screenshot shows a web browser window for the IPUMS PMA website. The URL is pma.ipums.org/pma-action/variables/PREGNANT#comparability_section. The page title is "IPUMS PMA: desc: PREGNANT". The top navigation bar includes links for LOGIN | REGISTER | GLOBAL HEALTH | IPUMS.ORG and a "Guest" account indicator. A "DATA CART" section on the right shows "YOUR DATA EXTRACT" with "0 VARIABLES" and "6 SAMPLES", with a "VIEW CART" button. The main content area is titled "PREGNANT" and shows "Pregnancy status" under "Group: Core demographics". Below this, there are tabs for CODES, DESCRIPTION, COMPARABILITY, UNIVERSE, AVAILABILITY, and QUESTIONNAIRE TEXT. The "COMPARABILITY" tab is selected, displaying the heading "Comparability" and the text: "There are minor universe differences among samples; see the Universe tab for more details." It also contains a section titled "Comparability with IPUMS-DHS" which states: "PREGNANT in IPUMS-PMA is similar to the variable PREGNANT in IPUMS-DHS. There may be differences in questionnaire text or the variable's universe; see the Survey Text and Universe Tab of the IPUMS-DHS variable for more information." At the bottom of the page, there is a footer note: "SUPPORTED BY: THE BILL & MELINDA GATES FOUNDATION, PMA, STAT/TRANSFER, AND UNIVERSITY OF MINNESOTA." and a copyright notice: "COPYRIGHT © MINNESOTA POPULATION CENTER, UNIVERSITY OF MINNESOTA."

2.2.4 Sample Universe

The **UNIVERSE** tab describes selection criteria for this question. In this case, there are some differences between samples:

- In DRC samples, all women aged 15-49 received this question.
- For all other samples, the question was skipped if any such woman previously indicated that she was menopausal or had a hysterectomy.

The screenshot shows a web browser window for the IPUMS PMA website. The URL is pma.ipums.org/pma-action/variables/PREGNANT#universe_section. The page title is "IPUMS PMA" and the sub-section is "PERFORMANCE MONITORING FOR ACTION". The main content area is titled "PREGNANT" and shows the "Universe" tab selected. Below the tabs, there is a list of survey descriptions for various countries and years, each specifying the selection criteria for the "PREGNANT" variable. The "Universe" tab contains the following text:

Universe

Burkina Faso 2020 Baseline/Phase 1 Longitudinal Survey: Women aged 15-49 who are not menopausal and have not had a hysterectomy.
Burkina Faso 2021 Phase 2 Longitudinal Survey: Women aged 15-49 who are not menopausal and have not had a hysterectomy.
Congo Democratic Republic (Kinshasa) 2019 Baseline/Phase 1 Longitudinal Survey: Women aged 15-49.
Congo Democratic Republic (Kongo Central) 2019 Baseline/Phase 1 Longitudinal Survey: Women aged 15-49.
Democratic Republic of the Congo (Kinshasa) 2020 Phase 2 Longitudinal Survey: Women aged 15-49.
Democratic Republic of the Congo (Kongo Central) 2020 Phase 2 Longitudinal Survey: Women aged 15-49.
Kenya 2019 Baseline/Phase 1 Longitudinal Survey: Women aged 15-49 who are not menopausal and have not had a hysterectomy.
Kenya 2020 Phase 2 Longitudinal Survey: Women aged 15-49 who are not menopausal and have not had a hysterectomy.
Nigeria 2019 (Kano) Baseline/Phase 1 Longitudinal Survey: Women aged 15-49 who are not menopausal and have not had a hysterectomy.
Nigeria 2019 (Lagos) Baseline/Phase 1 Longitudinal Survey: Women aged 15-49 who are not menopausal and have not had a hysterectomy.
Nigeria (Kano) 2020 Phase 2 Longitudinal Survey: Women aged 15-49 who are not menopausal and have not had a hysterectomy.
Nigeria (Lagos) 2020 Phase 2 Longitudinal Survey: Women aged 15-49 who are not menopausal and have not had a hysterectomy.

At the bottom of the page, there is a footer with the text: "SUPPORTED BY: THE BILL & MELINDA GATES FOUNDATION, PMA, STAT/TRANSFER, AND UNIVERSITY OF MINNESOTA." and "COPYRIGHT © MINNESOTA POPULATION CENTER, UNIVERSITY OF MINNESOTA."

2.2.5 Availability Across Samples

The **AVAILABILITY** tab shows all other samples (including cross-sectional samples) where this variable is available.

The screenshot shows the IPUMS PMA website interface. At the top, there is a navigation bar with links for LOG IN, REGISTER, GLOBAL HEALTH, and IPUMS.ORG. On the right side of the header, there is a "DATA CART" section indicating 0 VARIABLES and 6 SAMPLES, with a "VIEW CART" button. Below the header, the IPUMS PMA logo is displayed, along with the text "PERFORMANCE MONITORING FOR ACTION", "HOME | SELECT DATA | MY DATA | SUPPORT", and a "PREGNANT" category. Under the "PREGNANT" category, there are two buttons: "ADD TO CART" and "CHANGE SAMPLES". A horizontal menu bar below these buttons includes tabs for "CODES", "DESCRIPTION", "COMPARABILITY", "UNIVERSE", "AVAILABILITY" (which is currently selected), and "QUESTIONNAIRE TEXT". The main content area is titled "Availability" and lists the following countries and years: Burkina Faso: 2014-2018, 2020-2021; Congo (Democratic Republic): 2013-2020; Cote d'Ivoire: 2017-2018, 2020; Ethiopia: 2014-2019; Ghana: 2013-2017; India: 2016-2018, 2020; Indonesia: 2015-2016; Kenya: 2014-2020; Niger: 2015-2018; Nigeria: 2014-2020; Uganda: 2014-2020. At the bottom of the page, there is a footer note: "SUPPORTED BY: THE BILL & MELINDA GATES FOUNDATION, PMA, STAT/TRANSFER, AND UNIVERSITY OF MINNESOTA." and a copyright notice: "COPYRIGHT © MINNESOTA POPULATION CENTER, UNIVERSITY OF MINNESOTA."

2.2.6 Questionnaire Text

Finally, you'll find the full text of each question on the **QUESTIONNAIRE TEXT** tab. Each phase of the survey is shown separately, and you may click the "view entire document: text" link to view the complete questionnaire for a particular sample in any given phase.

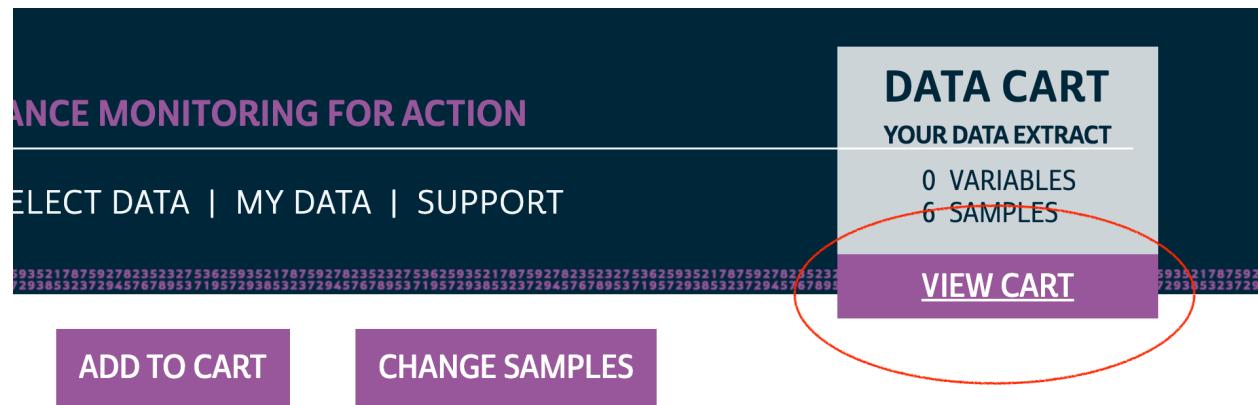
The screenshot shows a web browser window for the IPUMS PMA website. The URL is pma.ipums.org/pma-action/variables/PREGNANT#questionnaire_text_section. The page title is "IPUMS PMA" and the sub-section is "PERFORMANCE MONITORING FOR ACTION". The top navigation bar includes links for LOG IN | REGISTER | GLOBAL HEALTH | IPUMS.ORG and a "Guest" button. A "DATA CART" sidebar indicates 0 VARIABLES and 6 SAMPLES, with a "VIEW CART" button. The main content area is titled "PREGNANT" and shows a table with five columns: CODES, DESCRIPTION, COMPARABILITY, UNIVERSE, and AVAILABILITY. The "QUESTIONNAIRE TEXT" tab is selected. Below the table, the section "Questionnaire Text" lists three entries:

Burkina Faso 2020	Congo (Democratic Republic) Kenya 2019 2019b	Nigeria 2019b	
Burkina Faso 2021	Congo (Democratic Republic) Kenya 2020 2020a	Nigeria 2020a	
Congo (Democratic Republic)	Congo (Democratic Republic) Nigeria 2019a 2019b	Nigeria 2020b	

Below the table, there are two sections: "Burkina Faso 2020" and "Questionnaire form". The "Burkina Faso 2020" section contains the question "14. Are you pregnant now?" followed by four radio button options: Yes, No, Unsure, and No response. The "Questionnaire form" section contains a "view entire document: text" link. Similar sections are shown for Burkina Faso 2021 and Congo (Democratic Republic) 2019a, each with their own set of questions and response options.

2.2.7 Checkout

Use the buttons at the top of this page to add the variable to your Data Cart, or to “VIEW CART” and begin checkout.



2.3 DATA FOR STATA USERS

Your Data Cart shows all of the variables you've selected, plus several “preselected” variables that will be automatically included in your extract. Click the “CREATE DATA EXTRACT” button to prepare your download.

The screenshot shows the IPUMS PMA Data Cart interface. At the top right, there's a "DATA CART" section indicating "1 VARIABLE" and "6 SAMPLES". Below this, the main area is titled "DATA CART" and contains three buttons: "ADD MORE VARIABLES", "CREATE DATA EXTRACT" (which is circled in red), and "ADD MORE SAMPLES". A "Clear Data Cart" link is also present. The central part of the screen displays a table of variables and their characteristics across different countries and years. The table includes columns for "In cart", "Variable", "Variable Label", "Type", and country-specific columns (BURKF, CONDR, KENYA, NIGERA). The "CREATE DATA EXTRACT" button is highlighted with a red oval.

In cart	Variable	Variable Label	Type	BURKF 2020 - 2021	CONDR 2019a - 2020a	CONDR 2019b - 2020b	KENYA 2019 - 2020	NIGERA 2019a - 2020a	NIGERA 2019b - 2020b
<input checked="" type="checkbox"/>	SAMPLE	PMA sample number [preselected]	P	X	X	X	X	X	X
<input checked="" type="checkbox"/>	COUNTRY	PMA country [preselected]	P	X	X	X	X	X	X
<input checked="" type="checkbox"/>	YEAR	Year [preselected]	P	X	X	X	X	X	X
<input checked="" type="checkbox"/>	ELIGIBLE	Eligible female respondent [preselected]	P	X	X	X	X	X	X
<input checked="" type="checkbox"/>	EAIID	Enumeration area [preselected]	P	X	X	X	X	X	X
<input checked="" type="checkbox"/>	CONSENTFO	Female respondent provided consent to be interviewed [preselected]	P	X	X	X	X	X	X
<input checked="" type="checkbox"/>	FOINSTID	Unique ID for female questionnaire [preselected]	P	X	X	X	X	X	X
<input checked="" type="checkbox"/>	CONSENTHQ	Household respondent provided consent to be interviewed [preselected]	P	X	X	X	X	X	X
<input checked="" type="checkbox"/>	FOWEIGHT	Female weight [preselected]	P	X	X	X	X	X	X
<input checked="" type="checkbox"/>	STRATA	Strata [preselected]	P	X	.	.	X	X	X
<input checked="" type="checkbox"/>	PANELWOMAN	Panel woman interviewed in Phase 1	P	/	/	/	/	/	/

Before you submit an extract request, you'll have the opportunity to choose a "Data Format". **Stata users should select a Stata file (.dta)** - you'll notice that data formatted for R, SPSS, and SAS are also available. CSV files are provided, but not recommended. (If you wish to change Sample Members, you may do so again here.)

The screenshot shows the IPUMS PMA extract summary page. At the top, there are browser controls, a title bar with 'IPUMS PMA: extract summary', and a navigation bar with links for 'LOG IN | REGISTER | GLOBAL HEALTH | IPUMS.ORG'. Below the header is the IPUMS PMA logo. The main content area is titled 'EXTRACT REQUEST (HELP)'. It contains several input fields with 'Change' links: 'SAMPLES: 6', 'VARIABLES: 11', 'DATA FORMAT: .dta (Stata)', 'STRUCTURE: Rectangular (longitudinal - long)', 'SAMPLE MEMBERS: Female Respondents', and 'ESTIMATED SIZE: 12.8 MB'. Below these fields is a text area labeled 'Describe your extract' with a text input box. At the bottom is a purple 'SUBMIT EXTRACT' button. A red circle highlights the 'Change' link next to the 'DATA FORMAT' field.

Once the Stata option is selected, you may add a description and then proceed to the download page. After a few moments, you'll receive an email indicating that your extract has been created. Click the green "Download Stata" button to download your extract.

Extract Number	Date	Formatted Data	Fixed-width Text Files			Revise Extract	Resubmit Extract	Description (click to edit)	Hide selections Show all
			Data	Command Files	Codebook				
179	2022-10-26	Download Stata	-	-	-	Basic	DDI	revise	<input type="checkbox"/> Stata demo

Place both files in a folder that R can use as its [working directory](#). We **strongly recommend** using [RStudio projects](#) to manage all of the files and analysis scripts used for a particular research project. We'll place our files in a subfolder called "data" within our own RStudio project folder.

2.4 LONG DATA STRUCTURE

We've downloaded a **long** data extract (Female Respondents only), which we'll now load into Stata as follows:

```
use "pma_00119.dta", clear
```

In a **long** extract, data from each phase will be organized in *separate rows*. Here, responses from three panel members are shown:

```
sort fqinstid phase

list fqinstid phase age panelwoman ///
  if strmatch(fqinstid, "011*") | ///
  strmatch(fqinstid, "015*"), separator(8) noobs
```

fqinstid	phase	age	panelw~n
011W5S0HN91I4H4I3T9JCMBHB	baseline	29	.
011W5S0HN91I4H4I3T9JCMBHB	first fo	30	yes
015NP6FJTIA98FYCBBBS1F0F7	baseline	47	.
015NP6FJTIA98FYCBBBS1F0F7	first fo	48	yes
015WYNN02WXHH6JA4HA9PL1MR	baseline	20	.
015WYNN02WXHH6JA4HA9PL1MR	first fo	21	yes

Each panel member receives a unique ID shown in **fqinstid**. The variable **phase** shows that each woman's responses to the Phase 1 Female Questionnaire appears in the first row, while her Phase 2 responses appear in the second. **age** shows each woman's age when she completed the Female Questionnaire for each phase.

panelwoman indicates whether the woman completed all or part of the Female Questionnaire in a *prior* phase, and that she'd agreed to continue participating in the panel study at that time. The value **NA** appears in the rows for Phase 1, as **panelwoman** was not included in Phase 1 surveys.

We mentioned above that you'll also include responses from some non-panel members when you request an extract with Female Respondents. These include women who did not complete all or part the Female Questionnaire in a prior phase, as indicated by `panelwoman`. These women are not assigned a value for `fqinstid` - instead, you'll find an empty string:

```
gen non_panel = fqinstid == ""
label define fqinstid_blank 0 "fqinstid is not blank" 1 "fqinstid is blank"
label values non_panel fqinstid_blank
label variable panelwoman "Woman in the panel"
table (phase panelwoman) (non_panel), nototals missing
```

	non_panel
	fqinstid is not blank fqinstid is blank
longitudinal survey phase	
baseline	
Woman in the panel	
.	23,591
first follow up	
Woman in the panel	
no	6,586
yes	18,194

For most longitudinal analysis applications, you'll need to drop non-panel members together with any women who did not fully complete the Phase 2 Female Questionnaire. We'll demonstrate using a combination of `bysort` and `egen` to ensure that there is one row for every FQINSTID where PHASE == 1 and another row where PHASE == 2 & RESULTFQ == 1.

```
gen keep = 1 if phase == 1
replace keep = 1 if phase == 2 & resultfq == 1
bysort fqinstid : egen keep_both = sum(keep)
keep if keep_both == 2
drop keep keep_both
```

The PMA Longitudinal Briefs published for each sample also include only members of the *de facto* population. These are women who slept in the household during the night prior to the interview for each Household Questionnaire, such that `resident` takes the value 11 or 22. We can use a similar strategy to keep only *de facto* members who appear in both phases.

```

gen keep = 1 if phase == 1 & (resident == 11 | resident == 22)
replace keep = 2 if phase == 2 & (resident == 11 | resident == 22)
bysort fqinstid : egen keep_both = sum(keep)
keep if keep_both == 3
drop keep keep_both

```

Following these steps, you can check the size of each analytic sample like so:

```

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

table ( pop_numeric ) ( phase ) ( ), nototals missing

```

		longitudinal survey phase	
		baseline	first follow up
pop_numeric			
Burkina Faso		5,212	5,212
DRC-Kinshasa		1,973	1,973
DRC-Kongo Central		1,514	1,514
Kenya		6,939	6,939
Nigeria-Kano		998	998
Nigeria-Lagos		1,089	1,089

2.5 WIDE DATA STRUCTURE

We've also downloaded a **wide** data extract (Female Respondents only), which we'll load into Stata like so:

```
use "pma_00116.dta", clear
```

In a **wide** extract, all of the responses from one woman appear in the *same row*. The IPUMS extract system appends a numeric suffix to each variable name corresponding with the phase from which it was drawn. Consider our three example panel members again:

```
sort fqinstid

list fqinstid age_1 age_2 panelwoman_1 panelwoman_2 ///
if strmatch(fqinstid, "011*") | ///
strmatch(fqinstid, "015*"), separator(8) noobs
```

	fqinstid	age_1	age_2	panelw~1	panelw~2
	011W5S0HN91I4H4I3T9JCMHB	29	30	.	yes
	015NP6FJTIA98FYCBBBS1F0F7	47	48	.	yes
	015WYNN02WXHH6JA4HA9PL1MR	20	21	.	yes

Each panel member has one unique ID shown in **fqinstid**. However, **age** is parsed into two columns: **AGE_1** shows each woman's age at Phase 1, and **AGE_2** shows her age at Phase 2.

As we've discussed, **panelwoman** is not available for Phase 1, as it indicates whether the woman completed all or part of the Female Questionnaire in a *prior* phase. For this reason, all values in **PANELWOMAN_1** are blank. Most variables are copied once for each phase, even if they - like **PANELWOMAN_1** - are not available for all phases.

You might expect the total length of a **wide** extract to be half the length of a corresponding **long** extract. This is not the case! A **wide** extract includes one row for each woman who completed all or part of the Female Questionnaire *for any phase* - you'll find placeholder columns for phases where the interview was not conducted.

```
list resultfq_1 age_1 resultfq_2 age_2 ///
if fqinstid == "0C8VQU6B03BXLAVVZ8SB90EKQ", noobs
```

res~fq_1	age_1	res~fq_2	age_2
complete	31	not at h	not inte

In a **long** extract, rows for the missing phase are dropped. In this example, the woman was “not at home” for the Phase 2 Female Questionnaire. When we select a **long** extract containing only Female Respondents, her Phase 2 row is excluded automatically (it will be included if you request an extract containing Female Respondents and Female Non-respondents).

```
use "pma_00119.dta", clear
list phase age resultfq ///
if fqinstid == "0C8VQU6B03BXLAVVZ8SB90EKQ", noobs
```

phase	age	resultfq
baseline	31	complete

Again: for most longitudinal analysis applications, you’ll need to remove cases where women were not interviewed for Phase 1 or where the Phase 2 Female Questionnaire was not completed:

```
use "pma_00116.dta", clear
keep if resultfq_2 == 1 & resultfq_1 != .
```

The *de facto* population appearing in PMA Longitudinal Briefs is defined in **wide** extracts by cases where the values 11 or 12 appear in *both* RESIDENT_1 and RESIDENT_2:

```
keep if inlist(resident_1, 11, 22)
keep if inlist(resident_2, 11, 22)
```

Following these steps, each analytic sample contains the same number of cases shown in the final **long** format extract above.

```
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

table ( pop_numeric ) ( ), nototals missing
```

		Frequency
pop_numeric		
Burkina Faso		5,212
DRC-Kinshasa		1,973
DRC-Kongo Central		1,514
Kenya		6,939
Nigeria-Kano		998
Nigeria-Lagos		1,089

2.6 WHICH FORMAT IS BEST FOR ME?

The choice between **long** and **wide** formats ultimately depends on your research objectives.

Many data manipulation tasks, for example, are faster and easier to perform in the **wide** format. In the example above, we needed to identify women who completed a Female Questionnaire and were members of the *de facto* population in both phases. In the long format, we first had to use `bysort` and `egen` and keep to pare the dataset down to women with good data for both phases.

On the other hand, some of the longitudinal analysis commands require data to be in a long format - this includes both the suite of so-called `st` commands for time-to-event or survival analysis and the suite of so-called `xt` commands for analyzing panel data. Users who prefer the wide format for data cleaning and exploration can manually switch to long format with help from Stata's `reshape` command, for example:

```
use "pma_00116.dta", clear  
keep if resultfq_2 == 1 & resultfq_1 != .  
  
keep if inlist(resident_1, 11, 22)  
keep if inlist(resident_2, 11, 22)  
  
keep fqinstid age_1 pregnant_1 age_2 pregnant_2  
  
reshape long age_ pregnant_ , i(fqinstid) j(phase)
```

We will revisit
reshape when
analyzing PMA
contraceptive
calendar data in
Chapter 6.

(j = 1 2)

Data	Wide	→	Long
Number of observations	17,725	→	35,450
Number of variables	5	→	4
j variable (2 values)		→	phase
xij variables:			
	age_1 age_2	→	age_
	pregnant_1 pregnant_2	→	pregnant_

```
rename age_ age  
rename pregnant_ pregnant
```

Executing the `reshape` command with more variables takes practice, and we imagine many users will find it easier to simply work with data in the long format from the beginning. If you want to become adept at converting between long and wide formats, consult the [Stata documentation](#) or watch some of the numerous tutorials on the `reshape` command available on YouTube.

Fortunately, the updated IPUMS PMA extract system makes it easy to select the samples, sample members, and variables that matter to your particular research question. New choices for **long** and **wide** data formats save an additional data cleaning step, allowing you to jump into longitudinal analysis as quickly as possible.