CONTENTS

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

[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 - 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.

		Now Available from IPUMS PMA		
Sample	**Phase 1 Data Collection***	**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.

[2] Displaced 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.

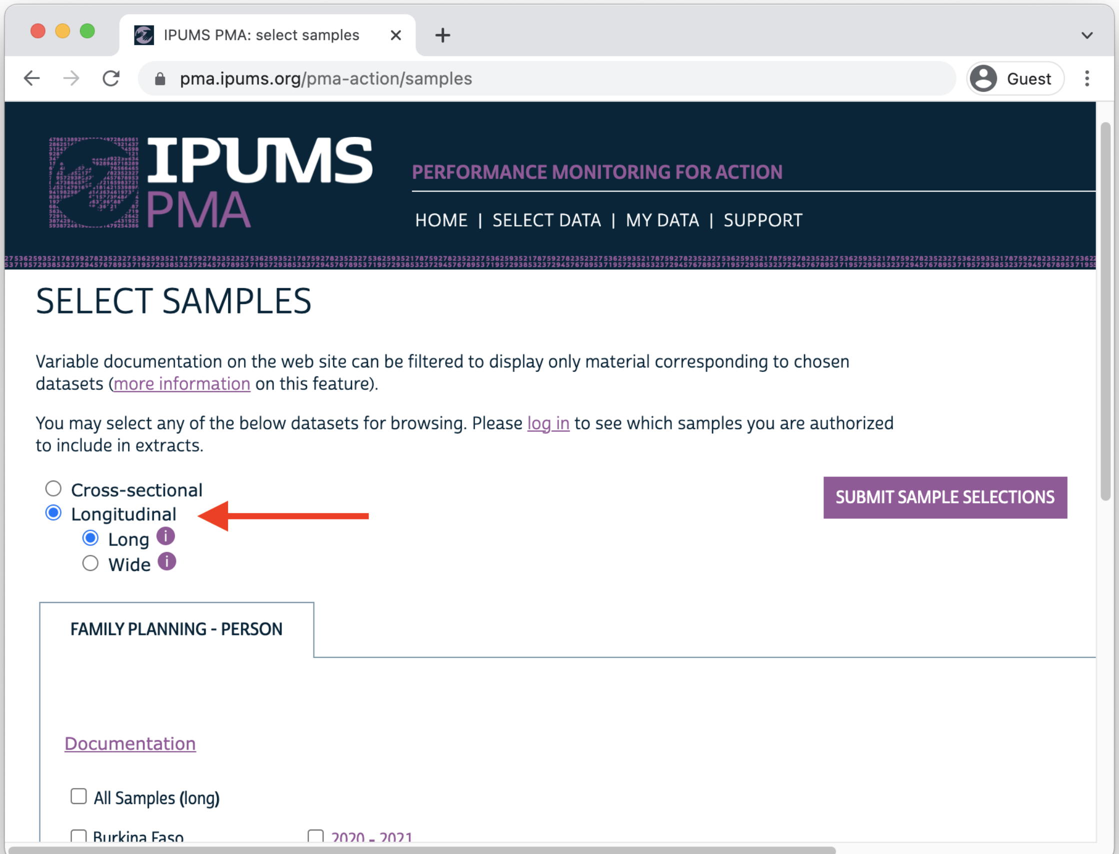
[3] Questionnaires administered in each country may vary from this **Core Household Questionnaire** - click here for details.

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

[5] 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 for details.

[6] 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.

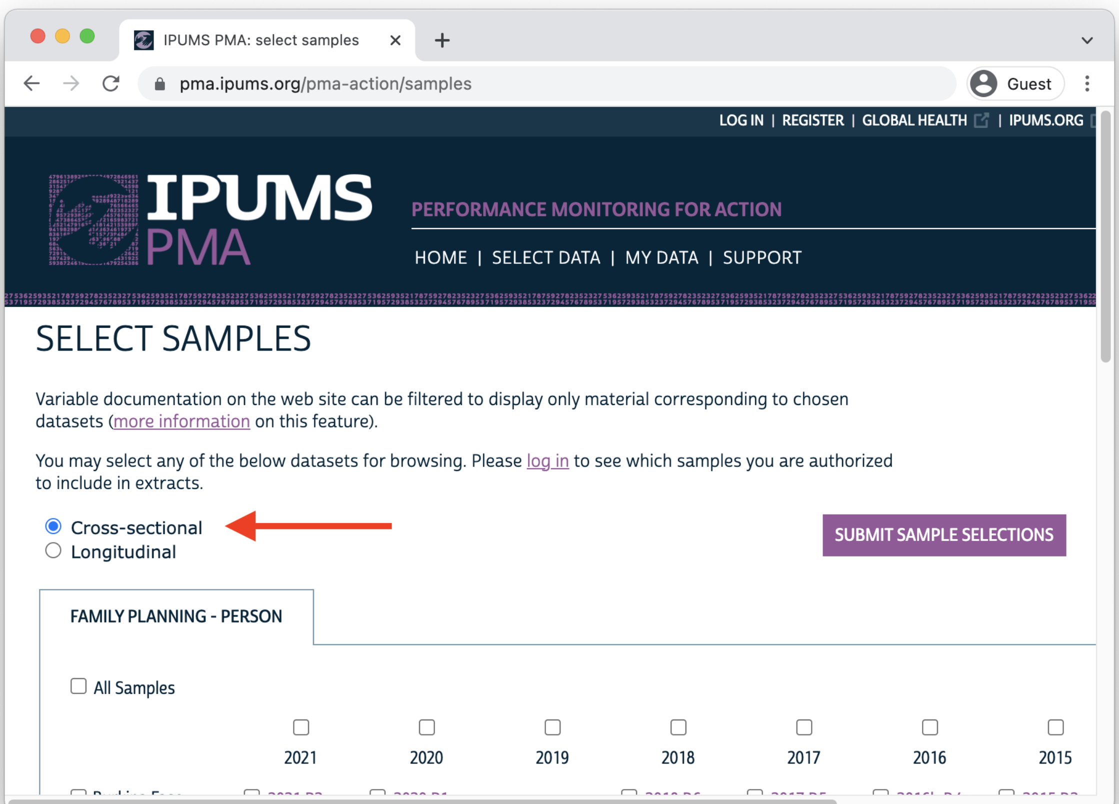
When you select the new **Longitudinal** sample option from IPUMS PMA, you'll be able to include responses from every available phase of the study. These samples are available in either "long" format (responses from each phase will be organized in separate rows) or "wide" format (responses from each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



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.



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:

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.

```
use "pma_00126.dta", clear

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
-----------------------------------------------------------
```

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
                     | questionna
     result of female |    ire
        questionnaire | 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 education[7]
- age - age
- wealthq and wealtht - household wealth quintile or tertile[8]
- urban and subnational - geographic location[9]

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

[8]Households 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 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 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 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 ID[13] and strata_1 as the stratum ID[14] 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.

[13] Because 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.

[14] As 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]
-------------+------------------------------------------------
     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.

[15] 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.

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 `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%)."

[16] See 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      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`.

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

--------------------------------------------------------------
             |             Linearized            Logit
             | Proportion   std. err.     [95% conf. interval]
-------------+------------------------------------------------
           y |
          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            |
  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, |
     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
```

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

-------------------------------------------------------------------------
                       |             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
-------------------------------------------------------------------------
```