IPUMS Data Training Exercise:

An introduction to IPUMS PMA

(Exercise 1 for R)

Learning goals

- Create and download an IPUMS PMA data extract
- Decompress data file and read data into R
- Analyze the data using sample code

Summary

In this exercise, you will gain an understanding of how IPUMS PMA is structured and how it can be leveraged to explore your research interests. This exercise will use the IPUMS PMA dataset to explore the relationship between family planning discontinuation and wealth. You will create data extracts that include the variables: LINENO, WEALTHQ, FPSTOPWHY, FPNOWUSPILL, FPAGE1STUSE, FP1STMETHOD, FQWEIGHT, POPWT, HQWEIGHT; then, you will use sample code to analyze these data.

R Code To Review

- This tutorial's sample code and answers use the so-called "tidyverse" style, but R has the blessing (and curse) that there are many different ways to do almost everything. If you prefer another programming style, please feel free to use it. But, for your reference, these are some quick explanations for commands that this tutorial will use:

Code	Purpose
%>%	The pipe operator helps make code with nested function calls easier to read. When reading code, it can be read as "and then". The pipe makes it so that code like "ingredients %>% stir() %>% cook()" is equivalent to cook(stir(ingredients)) (read as "take *ingredients* and then *stir* and then *cook* ").
as_factor	Converts the value labels provided for IPUMS data into a factor variable for R
summarize	Summarize a dataset's observations to one or more groups
group_by	Set the groups for the summarize function to group by
filter	Filter the dataset so that it only contains these values
mutate	Add on a new variable to a dataset
weighted.mean	Get the weighted mean of the variable
ggplot	Initializes a graphic object (histogram, box, plot, etc.)

Common Mistakes to Avoid

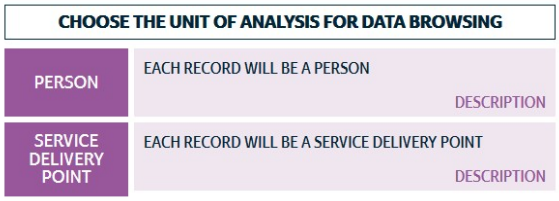
- Not changing the working directory to the folder where your data is stored.
- Mixing up = and ==; to assign a value in generating a variable, use "<-" (or "="). Use "==" to test for equality.

Registering with IPUMS

Go to http://pma.ipums.org, click on Register to Use IPUMS PMA on the left hand side of the screen. Click the Register for IPUMS PMA button and fill out the form to apply for access. You will have to wait for your account to be approved to access the data. Once you receive the approval email, click "Log In" at the top of the page and use your email and password.

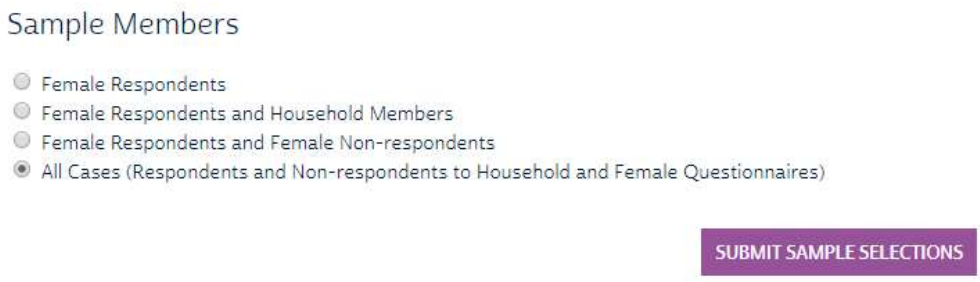
Select Samples

- Go to Select Data
- Choose the Person unit of analysis



- Click the Select Samples box, check the box for the Kenya 2016 R5

- Scroll to the bottom of the page and click the radio button option for All Cases. The default is Female Respondents
- Click the Submit Sample Selections box



Select Variables

- The search tool allows you to search for variables. Observe the options for limiting your search results by variable characteristics or variable type.

- You may add a variable to your cart by clicking on the plus sign in the "Add to Cart" column of the topical variable list, or list of search results.
- You may view information about the variable by clicking on the variable name, and navigating through the tabs that include a description of the variable, codes and value labels, the universe of persons asked the question, and information on the comparability of the variable among other pieces of information. If you are reviewing variable-specific information, you may click on the "Add to Cart" button near the top of the screen to add this variable to your data cart.
- Using the drop down menu or search feature, select the following variables:

  LINENO: Person number in sample unit

  WEALTHQ: Wealth quintile

  FPSTOPWHY: Reason for discontinuing family planning method

  FPNOWUSPILL: Whether birth control pills are currently used

  FPAGE1STUSE: Age first used family planning

  FP1STMETHOD: First method of family planning used

  FQWEIGHT: Female level weighting variable

  POPWT: Population expansion factor

  HQWEIGHT: Household level weighting variable

Review and submit your extract

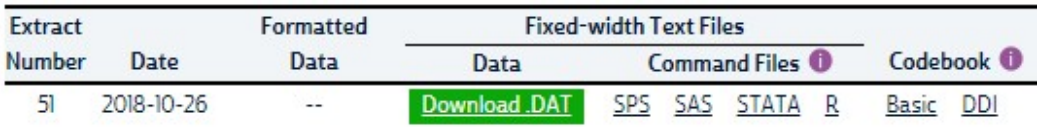
- Click the purple VIEW CART button under your data cart
- Review variable selection. Note that certain variables appear in your data cart even if you did not select them, and they are not included in the constantly updated count of variables in your data cart. The preselected variables are needed for weighting, for variance estimation, or to identify the year, country, and round of a sample.

- Click the Create Data Extract button
- Review the 'Extract Request Summary' screen, describe your extract and click Submit Extract
- You will get an email when the data is available to download.
- To get to the page to download the data, follow the link in the email, or follow the My Data Extracts link on the homepage.

Getting the data into your statistics software

Download the data

- Go to http://pma.ipums.org and click on My Data Extracts



Extract Number	Date	Formatted Data	Fixed-width Text Files						
			Data	Command Files				Codebook	
51	2018-10-26	--	Download .DAT	SPS	SAS	STATA	R	Basic	DDI

- Right-click on the data link next to extract you created
- Choose "Save Target As..." (or "Save Link As...")
- Save into "Documents" (that should pop up as the default location)
- Do the same thing for the DDI link next to the extract
- (Optional) Do the same thing for the R script
- You do not need to decompress the data to use it in R

Install and load packages for R

- Open R from the Start menu
- If you haven't already installed any of the following packages, type:

```
install.packages("ipumsr")
install.packages("dplyr")
install.packages("ggplot2")
```

- Next (or if you have already installed the packages on your computer), type:

```
library(ipumsr)
library(dplyr)
library(ggplot2)
options(tibble.print_max = Inf)
```

Read data into R

- Set your working directory to where you saved the data above by adapting the following command (Rstudio users can also use the "Project" feature to set the working directory. In the menubar, select File -> New Project -> Existing Directory and then navigate to the folder):

```
setwd("~/") # "~/" goes to your Documents directory on most computers
```

- Run the following command from the console, adapting it so it refers to the extract you just created (replace the #s below with the number of your extract):

```
ddi <- read_ipums_ddi("pma_000##.xml")
HHF <- read_ipums_micro(ddi)
```

- NOTE: To stay consistent with the exercises for other statistical packages, this exercise does not spend much time on the helpers to allow for translation of the way IPUMS uses labelled values to the way base R does. You can learn more about these in the value-labels vignette in the R package. From R run command: `vignette("value-labels", package = "ipumsr")`

Analyze the Sample

Part 1: Household Survey frequencies of WEALTHQ

1. According to the Description tab on the website, what does the variable WEALTHQ measure? _____

2. According to the Codes tab, what value labels and numeric codes apply to each wealth quintile? Include labels and codes that explain any "missing" values. _____

   _____

3. According to the Universe tab, who is in universe for WEALTHQ in the Kenya 2016 Round 5 sample? _____

   _____

4. What percentage of all *household members* are listed as living in a household in the lowest income quintile? _____

```
HHF%>%
 count(as_factor(WEALTHQ))%>%
 mutate(prop = prop.table(n))
```

5. What percentage of *households* are listed in the lowest income quintile?

   _____

```
HHF%>%
 filter(LINENO == 1)%>%
 count(as_factor(WEALTHQ))%>%
 mutate(prop = prop.table(n))
```


Part 2: Female Survey frequencies of FPSTOPWHY

6. According to the Description tab on the website, what does the variable FPSTOPWHY measure? It can be found in the Discontinuation of Family Planning variable group. _____

7. According to the Universe tab, who is in universe for FPSTOPWHY in the Kenya 2016 Round 5 sample? _____

8. Create a frequency table for the FPSTOPWHY variable for the Kenya 2016 Round 5 sample. What are the top two most common responses? What proportion of surveys provide these responses? _____

```
HHF%>%
  count(as_factor(FPSTOPWHY))%>%
  mutate(prop = prop.table(n))
```

9. Given the universe stated in question 7, only a small fraction of women aged 15 to 49 provided answers to FPSTOPWHY. Not in universe cases also include men and women outside of childbearing age. Suppose you wanted to know what proportion of *those who provided an answer* to FPSTOPWHY chose one of the two top answers; how does this answer differ to your response for question 8? _____

```
HHF%>%
  filter(FPSTOPWHY < 90)%>%
  count(as_factor(FPSTOPWHY))%>%
  mutate(prop = prop.table(n))
```

10. Of the women in the sample who stopped using a family planning method in the last year because of cost, what proportion lived in households in either of the lowest 2 income quintiles?

    _____

```
HHF%>%
  filter(FPSTOPWHY == 33)%>%
  count(as_factor(WEALTHQ))
```

Part 3: Using household, female, and population weights (HQWEIGHT, FQWEIGHT, and POPWT)

IPUMS PMA data require the use of weights to ensure that each sample is representative of the population from which it was drawn. However, because variables alternately reference either the population of households or the population of women aged 15-49 in each country, users must select weights that correspond to their intended unit of analysis.

For example: the frequency table for FPSTOPWHY from Part 2D indicates that, among the women aged 15-49 in Kenya 2016 Round 5 *who answered this question*, 30.85% stopped using family planning methods in the past year because they wanted to become pregnant. Because the relevant unit of analysis for FPSTOPWHY includes individual women aged 15-49 (rather than households), FQWEIGHT should be used to make inferences about that particular *population.*

11. Use FQWEIGHT to estimate the proportion of all Kenyan women aged 15-49 who stopped using family planning methods in order to become pregnant in 2016. _____

```
HHF%>%
  count(as_factor(FPSTOPWHY), wt=round(FQWEIGHT))%>%
  mutate(prop = prop.table(n))
```

While FQWEIGHT may be used to estimate proportions, it should not be used to generate representative population counts. Instead, POPWT combines information from FQWEIGHT with national population estimates from the United Nations Population Division to create an expansion factor appropriate for this purpose (see https://pma.ipums.org/pma/population_weights.shtml for details).

12. Use POPWT to estimate the total number of all Kenyan women aged 15-49 who stopped using family planning methods in order to become pregnant in 2016. _____

```
HHF%>%
  count(as_factor(FPSTOPWHY), wt=POPWT)
```

By contrast, variables pertaining to household characteristics should be weighted with HQWEIGHT, which accounts for each household's the probability of being sampled and for household non-response. Because each member of the same household shares the same value for HQWEIGHT, users should also ensure that each household is only represented once in their analysis.

For example: Part 1 revealed that 18.85% of households in the Kenya 2016 Round 5 sample fall into the lowest household income quintile. Because the variable WEALTHQ

reflects a household characteristic, HQWEIGHT should be used to make inferences about the population of households in Kenya in 2016.

13. Use HQWEIGHT to estimate the proportion of households in the lowest household income quintile in Kenya for 2016. Does the sample over- or under-represent the households in the lowest income quintile compared to the true population of households?

```
HHF%>%
  filter(LINENO == 1)%>%
  count(as_factor(WEALTHQ),wt=round(HQWEIGHT))%>%
  mutate(prop = prop.table(n))
```

Users should note that, much like FQWEIGHT, HQWEIGHT should only be used to generate proportion estimates for the population of households. There is currently no expansion factor for the purpose of estimating population frequency counts.

Part 4: Graphing

14. Examine a frequency table for the variable FPAGE1STUSE, which reflects the age a respondent first began use of any family planning method. Create a 35 bin histogram for FPAGE1STUSE, excluding any missing values.

```
HHF%>%
  filter(FPAGE1STUSE < 90)%>%
  group_by(FPAGE1STUSE)%>%
  count(FPAGE1STUSE)

HHF%>%
  filter(FPAGE1STUSE < 90)%>%
  ggplot(aes(x = FPAGE1STUSE)) + geom_histogram(bins = 35)))
```

15. Now consider the variable FP1STMETHOD, which classifies the first family planning method ever used by a respondent as either “modern” (values 100-199) or “traditional” (values 200-299). Create two 35 bin histograms: one for modern methods, one for traditional methods.

```
HHF%>%
  filter(FPAGE1STUSE <50 & FP1STMETHOD < 200)%>%
  group_by(age = FPAGE1STUSE)%>%
  count()%>%
  ggplot(aes(x = age, y = n)) + geom_col()


HHF%>%
  filter(FPAGE1STUSE <50 & FP1STMETHOD > 200 & FP1STMETHOD
<900)%>%
  group_by(age = FPAGE1STUSE)%>%
  count()%>%
  ggplot(aes(x = age, y = n)) + geom_col()
```

ANSWERS

Part 1: Household Survey frequencies of WEALTHQ

1. According to the Description tab on the website, what does the variable WEALTHQ measure?

   For all households with complete survey results, WEALTHQ refers to the relative wealth of the household where the woman lives, divided into quintiles from the poorest (Lowest quintile) to the richest (Highest quintile).

2. According to the Codes tab, what value labels and numeric codes apply to each wealth quintile? Include labels and codes that explain any "missing" values.

   01 - Lowest quintile

   02 - Lower quintile

   03 - Middle quintile

   04 - Higher quintile

   05 - Highest quintile

   96 - Not interviewed (household questionnaire)

   98 - No response or missing

3. According to the Universe tab, who is in universe for *WEALTHQ* in the Kenya 2016 Round 5 sample? All persons.

4. What percentage of all *household members* are listed as living in a household in the lowest income quintile? 22.7%

```
> HHF%>%
+   count(as_factor(WEALTHQ))%>%
+   mutate(prop = prop.table(n))
# A tibble: 7 x 3
  `as_factor(WEALTHQ)`                          n     prop
  <fct>                                     <int>    <dbl>
1 Lowest quintile                            5900 0.227
2 Lower quintile                             5418 0.208
3 Middle quintile                            5178 0.199
4 Higher quintile                            4598 0.177
5 Highest quintile                           4642 0.178
6 Not interviewed (household questionnaire)   265 0.0102
7 No response or missing                       17 0.000653
```


5. What percentage of *households* are listed in the lowest income quintile? 19.35%

```
> HHF%>%
+   filter(LINENO == 1)%>%
+   count(as_factor(WEALTHQ))%>%
+   mutate(prop = prop.table(n))
# A tibble: 7 x 3
  `as_factor(WEALTHQ)`                          n    prop
  <fct>                                     <int>   <dbl>
1 Lowest quintile                            1228 0.194
2 Lower quintile                             1142 0.180
3 Middle quintile                            1172 0.185
4 Higher quintile                            1258 0.198
5 Highest quintile                           1273 0.201
6 Not interviewed (household questionnaire)   265 0.0418
7 No response or missing                        7 0.00110
```

Part 2: Female Survey frequencies of FPSTOPWHY

6. According to the Description tab on the website, what does the variable FPSTOPWHY measure? It can be found in the Discontinuation of Family Planning variable group. For women who recently used a method of family planning to delay or avoid pregnancy but are no longer using it, FPSTOPWHY reports the reason why they stopped.
7. According to the Universe tab, who is in universe for FPSTOPWHY in the Kenya 2016 Round 5 sample? Women age 15-49 who used a family planning method to delay or avoid pregnancy in the last 12 months, but who are not currently using.
8. Create a frequency table for the FPSTOPWHY variable for the Kenya 2016 Round 5 sample. What are the top two most common responses? What proportion of surveys provide these responses? NIU (not in universe) (97.23%); Wanted to become pregnant (0.46%).

```
> HHF%>%
+   count(as_factor(FPSTOPWHY))%>%
+   mutate(prop = prop.table(n))
# A tibble: 20 x 3
   `as_factor(FPSTOPWHY)`                      n      prop
   <fct>                                   <int>     <dbl>
 1 Became pregnant while using                19 0.000730
 2 Wanted a more effective method              5 0.000192
 3 Health concerns                            48 0.00184
 4 Fear of side effects                       41 0.00158
 5 Interferes with body's processes           23 0.000884
 6 Inconvenient to use                         8 0.000307
 7 Infrequent sex/husband away                88 0.00338
 8 Husband/partner disapproved                 6 0.000231
 9 Wanted to become pregnant                 120 0.00461
10 Difficult to get pregnant/menopausal        4 0.000154
11 Fatalistic                                  1 0.0000384
12 No method available                         1 0.0000384
13 Lack of access / too far                    2 0.0000769
14 Costs too much                              3 0.000115
15 Other                                      15 0.000577
16 Not interviewed (female questionnaire)     68 0.00261
17 Not interviewed (household questionnaire) 264 0.0101
18 Don't know                                  4 0.000154
19 No response or missing                      1 0.0000384
20 NIU (not in universe)                   25297 0.972
```

9. Given the universe stated in 7, only a small fraction of women aged 15 to 49 provided answers to FPSTOPWHY. Not in universe cases also include men and women outside of childbearing age. Suppose you wanted to know what proportion of *those who provided an answer* to FPSTOPWHY chose one of the two top answers; how does this answer differ to your response for 8? Wanted to become pregnant (31.25 %) Infrequent sex / husband away (22.92%). These proportions now have a more appropriate denominator for analysis.

```
> HHF%>%
+   filter(FPSTOPWHY < 90)%>%
+   count(as_factor(FPSTOPWHY))%>%
+   mutate(prop = prop.table(n))
# A tibble: 15 x 3
   `as_factor(FPSTOPWHY)`                 n    prop
   <fct>                              <int>   <dbl>
 1 Became pregnant while using           19 0.0495
 2 Wanted a more effective method         5 0.0130
 3 Health concerns                       48 0.125
 4 Fear of side effects                  41 0.107
 5 Interferes with body's processes      23 0.0599
 6 Inconvenient to use                    8 0.0208
 7 Infrequent sex/husband away           88 0.229
 8 Husband/partner disapproved            6 0.0156
 9 Wanted to become pregnant            120 0.312
10 Difficult to get pregnant/menopausal   4 0.0104
11 Fatalistic                             1 0.00260
12 No method available                    1 0.00260
13 Lack of access / too far               2 0.00521
14 Costs too much                         3 0.00781
15 Other                                 15 0.0391
```

10. Of the women in the sample who stopped using a family planning method in the last year because of cost, what proportion lived in households in either of the lowest 2 income quintiles? 2 of 3 women, or roughly 66%.

```
> HHF%>%
+   filter(FPSTOPWHY == 33)%>%
+   count(as_factor(WEALTHQ))
# A tibble: 3 x 2
  `as_factor(WEALTHQ)`     n
  <fct>                <int>
1 Lowest quintile          1
2 Lower quintile           1
3 Middle quintile          1
```

Part 3: Using household, female, and population weights

11. Use FQWEIGHT to estimate the proportion of all Kenyan women aged 15-49 who stopped using family planning methods in order to become pregnant in 2016. 2.19%

```
> HHF%>%
+   count(as_factor(FPSTOPWHY), wt=round(FQWEIGHT))%>%
+   mutate(prop = prop.table(n))
# A tibble: 20 x 3
   `as_factor(FPSTOPWHY)`                          n     prop
   <fct>                                       <dbl>    <dbl>
 1 Became pregnant while using                    21 0.00335
 2 Wanted a more effective method                  5 0.000798
 3 Health concerns                                47 0.00751
 4 Fear of side effects                           45 0.00719
 5 Interferes with body's processes               27 0.00431
 6 Inconvenient to use                             6 0.000958
 7 Infrequent sex/husband away                    96 0.0153
 8 Husband/partner disapproved                     6 0.000958
 9 Wanted to become pregnant                     137 0.0219
10 Difficult to get pregnant/menopausal            5 0.000798
11 Fatalistic                                      1 0.000160
12 No method available                             1 0.000160
13 Lack of access / too far                        2 0.000319
14 Costs too much                                  3 0.000479
15 Other                                          12 0.00192
16 Not interviewed (female questionnaire)          0 0
17 Not interviewed (household questionnaire)       0 0
18 Don't know                                      4 0.000639
19 No response or missing                          1 0.000160
20 NIU (not in universe)                        5843 0.933
```

12. Use POPWT to estimate the total number of all Kenyan women aged 15-49 who stopped using family planning methods in order to become pregnant in 2016.

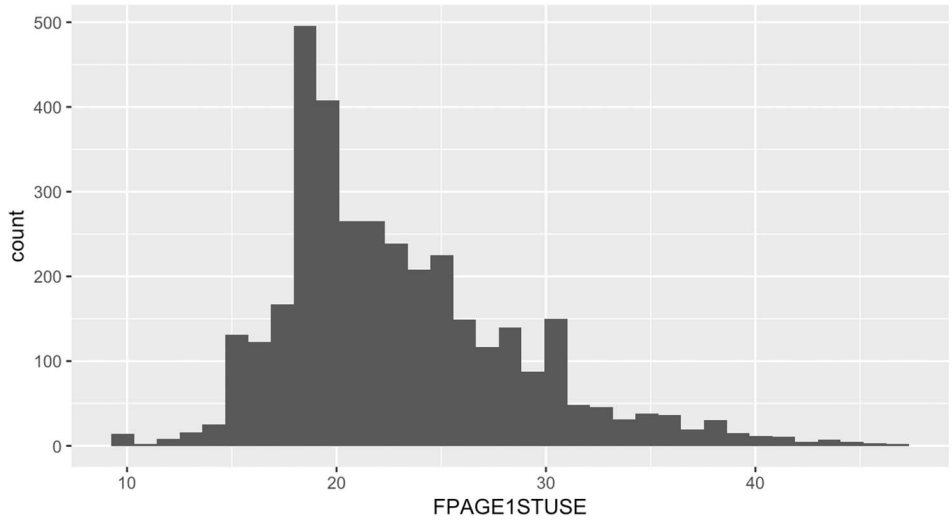
    273,016

```
> HHF%>%
+   count(as_factor(FPSTOPWHY), wt=POPWT)
# A tibble: 20 x 2
   `as_factor(FPSTOPWHY)`                          n
   <fct>                                       <dbl>
 1 Became pregnant while using                 40598
 2 Wanted a more effective method               9000
 3 Health concerns                             97002
 4 Fear of side effects                        91275
 5 Interferes with body's processes            57194
 6 Inconvenient to use                         13496
 7 Infrequent sex/husband away                182977
 8 Husband/partner disapproved                 12228
 9 Wanted to become pregnant                  273016
10 Difficult to get pregnant/menopausal        10939
11 Fatalistic                                   1772
12 No method available                          1288
13 Lack of access / too far                     3992
14 Costs too much                               4043
15 Other                                       26250
16 Not interviewed (female questionnaire)          0
17 Not interviewed (household questionnaire)       0
18 Don't know                                   5336
19 No response or missing                       1334
20 NIU (not in universe)                    11424445
```

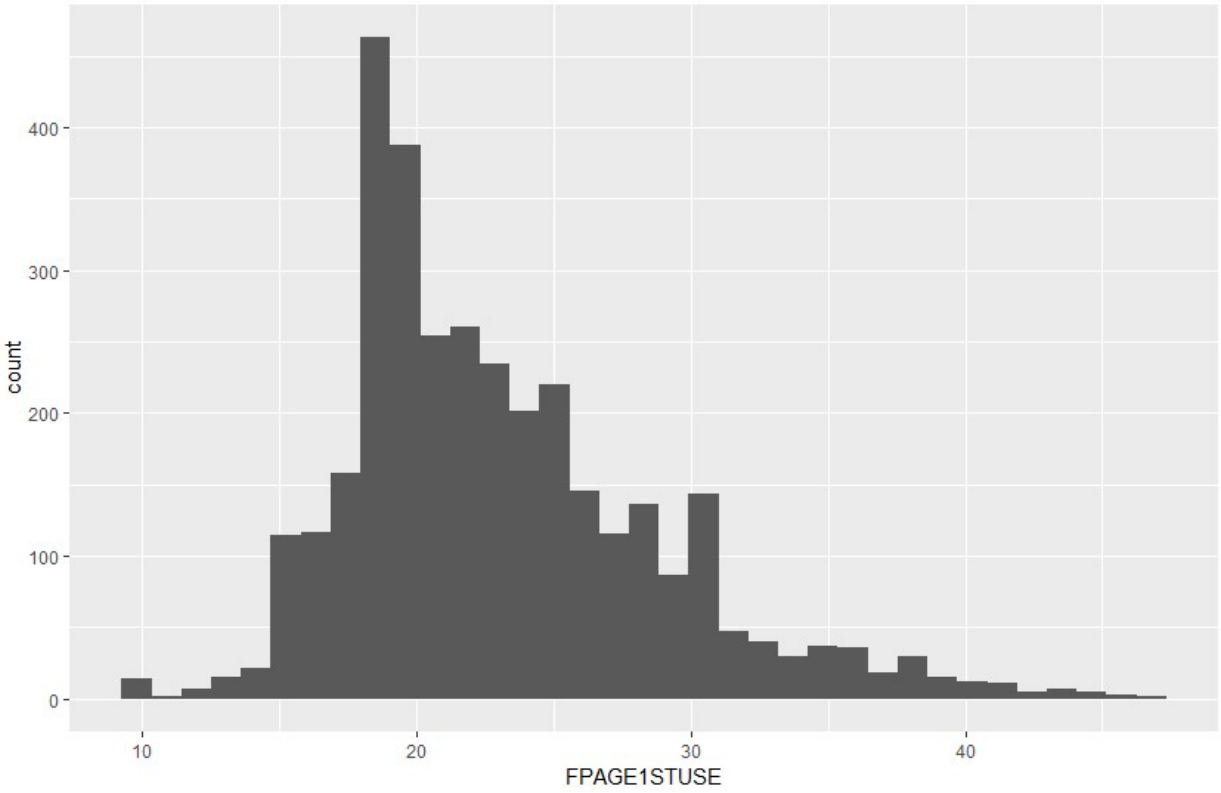

13. Use HQWEIGHT to estimate the proportion of households in the lowest household income quintile in Kenya for 2016. Does the sample over- or under-represent the households in the lowest income quintile compared to the true population of households? 18.9%. The lowest income quintile is slightly over-represented in the sample, compared to the true population of households.

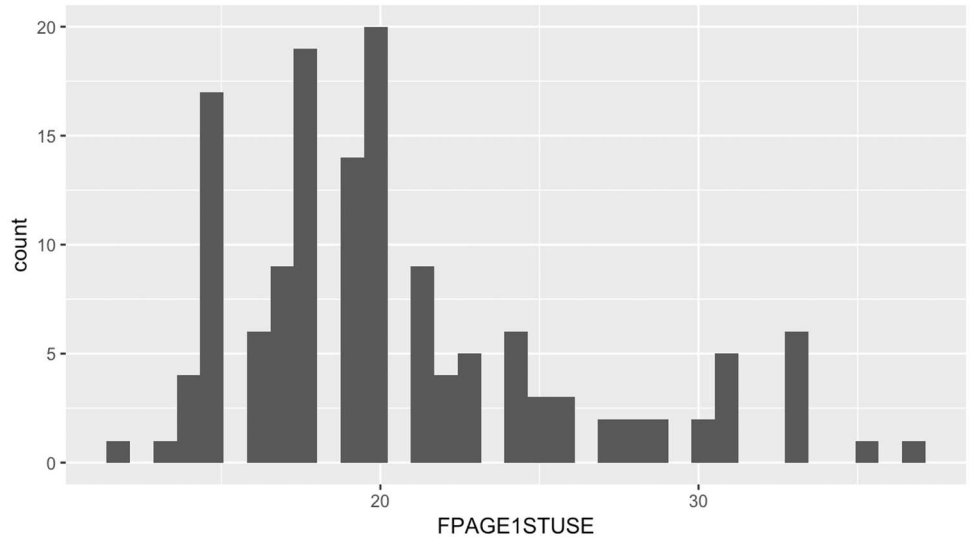
Part 4: Graphing

14. Examine a frequency table for the variable FPAGE1STUSE, which reflects the age a respondent first began use of any family planning method. Create a 35 bin histogram for FPAGE1STUSE, excluding any missing values.



15. Now consider the variable FP1STMETHOD, which classifies the first family planning method ever used by a respondent as either "modern" (values 100-199) or "traditional" (values 200-299). Create a new variable that divides the method type into modern and traditional methods based on the codes and frequencies presented on the PMA website. Recreate the histogram above twice, restricting to only modern or only traditional methods. Specify 35 bins for each using the bin(35) option.


400
300
200
100
0
count
10
20
30
40
FPAGE1STUSE


20
15
10
5
0
count
20
30
FPAGE1STUSE