Describing, Transforming, and Analyzing Data Using Stata

An Introductory Course

Gracie Rosenbach

International Food Policy Research Institute

Kigali, Rwanda

March 1-3, 2022

Lesson 1 – Introduction and Describing the Data

1. Introduction to Stata and the Integrated Household Living Conditions Survey 5 (EICV5)

a. Background to this Stata training manual

This manual describes how to use Stata statistical software to describe, transform, and analyze data. The emphasis is on the analysis of household and person data, but Stata can be used with any database.

This manual was created based on release 16 of Stata. If one is using an older version of Stata, it is not certain that all commands discussed in this training manual will be part of earlier releases of the program. However, a user should be able to determine how to obtain the results from a more recently included command using older commands by search for help on-line, as there are extensive resources to assist Stata users on-line.

The training course for which this manual is used is not a lecture course, but rather it is a semi-structured hands-on workshop in which trainees will use Stata on computers to learn different methods of analyzing data. Thus, active participation of the trainees is necessary to maximize the benefit from the training.

b. Background to the dataset (from NISR website) [1]

The Fifth Integrated Household Living Conditions Survey or Enquête Intégrale sur les Conditions de Vie des ménages (EICV 5) in French, provides information on changes in the well-being of the population such as poverty, inequality, employment, living conditions, education, health and housing conditions, household consumption, among others in 2016/17. The EICV5 dataset can be accessed through http://microdata.statistics.gov.rw/index.php/catalog/82

c. Sampling methodology (from NISR EICV5 Main Indicators Report)[2]

The EICV5 cross-sectional survey is designed to represent the current household-based population of Rwanda. The NISR national master sampling frame was used for selecting the sample villages in each district. This master sample was based on the 2012 Rwanda Census frame. The villages were selected from the Master Sample, stratified by district. Within each district the sample villages were selected systematically with probability proportional to size (PPS), where the measure of size was based on the number of households in each village from the 2012 Census frame. Within each district the villages in the master sampling frame were not explicitly stratified by urban and rural areas. However, the frame of villages within each district was ordered by urban and rural codes, and the systematic selection of the sample villages (with PPS) provides an implicit stratification of the Master Sample by urban and rural areas within each district, with a proportional allocation of the sample villages to each stratum. Similar to the EICV4 cross-sectional survey methodology, a nationally-representative sample of clusters was assigned for the EICV5 data collection each cycle out 10 cycles, so that the sample is geographically representative over time. This process

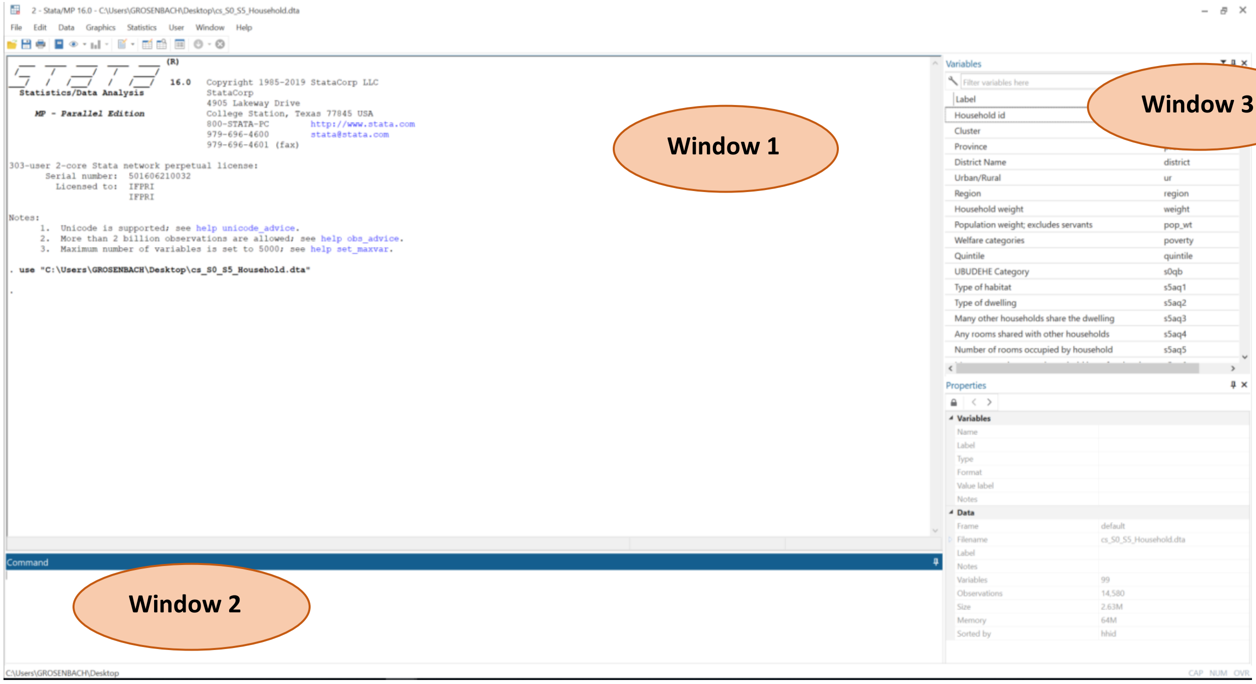
[1] NISR 2017. https://www.statistics.gov.rw/datasource/integrated-household-living-conditions-survey-5-eicv-5
[2] NISR 2017. https://www.statistics.gov.rw/publication/eicv-5-main-indicators-report-201617

ensured that the final distribution of the sample clusters to cycles and sub-cycles was geographically representative within each district.

2. Exploring Stata and the data

a. Let’s explore some data on household characteristics – ***double-click to open file cs_S0_S5_Household.dta*** – the following Stata screen should appear.
The *main Stata screen* consists of 3 windows:

Figure 1.1: Main Stata screen



```
                                                    (R)
  ___  ____  ____  ____  ____
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   16.0   Copyright 1985-2019 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     MP - Parallel Edition            College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

303-user 2-core Stata network perpetual license:
       Serial number:  501606210032
         Licensed to:  IFPRI
                       IFPRI

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  More than 2 billion observations are allowed; see help obs_advice.
      3.  Maximum number of variables is set to 5000; see help set_maxvar.

. use "C:\Users\GROSENBACH\Desktop\cs_S0_S5_Household.dta"

.
```

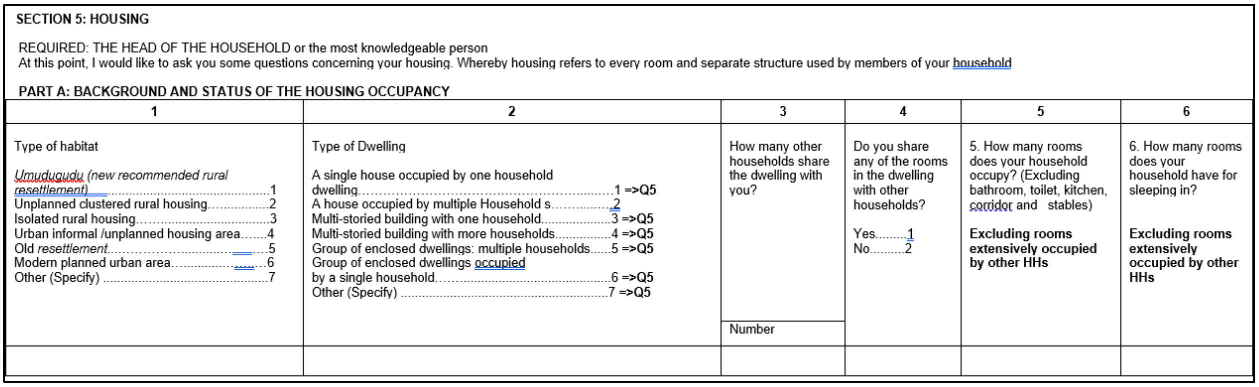
Label	
Household id	
Cluster	
Province	
District Name	district
Urban/Rural	ur
Region	region
Household weight	weight
Population weight; excludes servants	pop_wt
Welfare categories	poverty
Quintile	quintile
UBUDEHE Category	s0qb
Type of habitat	s5aq1
Type of dwelling	s5aq2
Many other households share the dwelling	s5aq3
Any rooms shared with other households	s5aq4
Number of rooms occupied by household	s5aq5

Variables	
Name	
Label	
Type	
Format	
Value label	
Notes	
Data	
Frame	default
Filename	cs_S0_S5_Household.dta
Label	
Notes	
Variables	99
Observations	14,580
Size	2.63M
Memory	64M
Sorted by	hhid

i. Window 1: Review and Results – shows the commands entered and output generated from these commands
ii. Window 2: Command – where you enter a Stata command
iii. Window 3: Variables – lists all of the variables and labels in the dataset

b. The datafile cs_S0_S5_Household corresponds to Sections 0 and 5 in the EICV5, for example, see Figure 1.2:

Figure 1.2: Section 5 Part A (Housing occupancy) from the EICV5 Questionnaire



SECTION 5: HOUSING

REQUIRED: THE HEAD OF THE HOUSEHOLD or the most knowledgeable person
At this point, I would like to ask you some questions concerning your housing. Whereby housing refers to every room and separate structure used by members of your household

PART A: BACKGROUND AND STATUS OF THE HOUSING OCCUPANCY

1	2	3	4	5	6
Type of habitat *Umudugudu (new recommended rural resettlement)*……………1 Unplanned clustered rural housing……………2 Isolated rural housing……………3 Urban informal /unplanned housing area…….4 Old *resettlement*……………5 Modern planned urban area……………6 Other (Specify) ……………7	Type of Dwelling A single house occupied by one household dwelling……………1 =>Q5 A house occupied by multiple Household s……………2 Multi-storied building with one household……………3 =>Q5 Multi-storied building with more households……………4 =>Q5 Group of enclosed dwellings: multiple households……5 =>Q5 Group of enclosed dwellings occupied by a single household……………6 =>Q5 Other (Specify) ……………7 =>Q5	How many other households share the dwelling with you? Number	Do you share any of the rooms in the dwelling with other households? Yes………1 No……….2	5. How many rooms does your household occupy? (Excluding bathroom, toilet, kitchen, corridor and stables) **Excluding rooms extensively occupied by other HHs**	6. How many rooms does your household have for sleeping in? **Excluding rooms extensively occupied by other HHs**

- The questions and their numbers from the questionnaire should match the variable names and labels found in Window 3 (Variables Window)

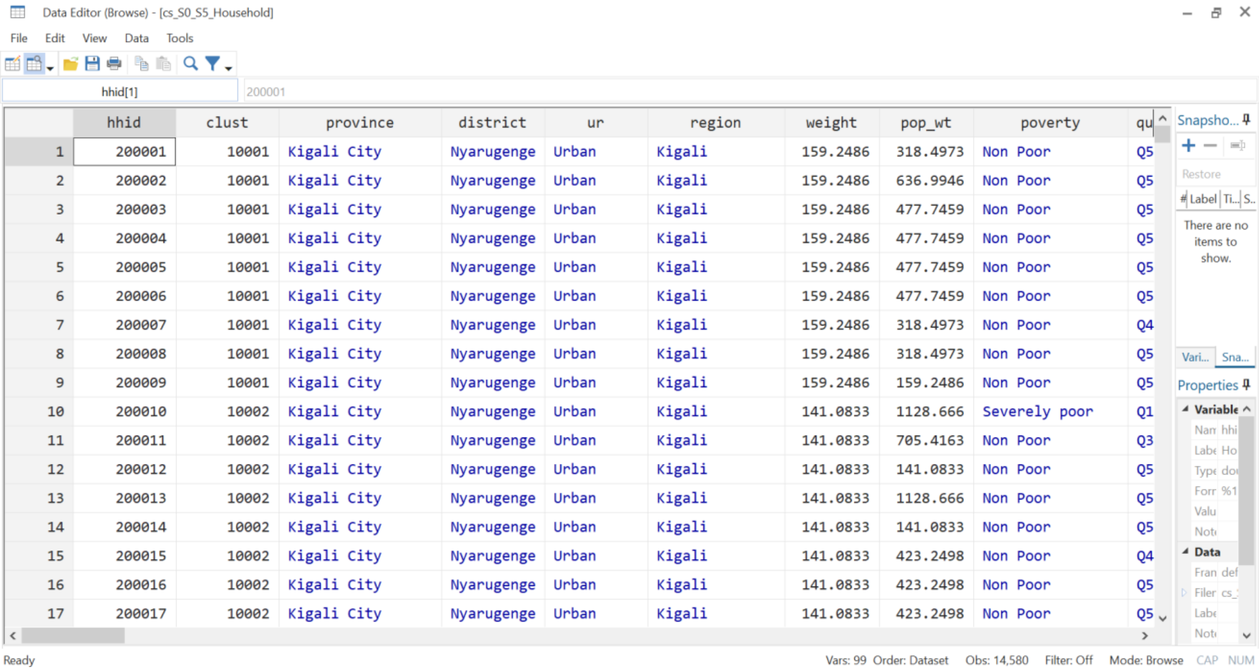
c. There are many separate Stata screens that can be opened. One of these screens is the *data browser* in which you can look at the data currently loaded in Stata. There are 3 different ways to access the *data browser:*

   i. Command Window: Type *"browse"* into Window 2 (Command Window) and press *enter*

   ii. Drop downs: Select *Data -> Data Editor -> Data Editor (Browse)*

   iii. Icon in top ribbon: Click this button

Figure 1.3: Stata data browser screen



Data Editor (Browse) - [cs_S0_S5_Household]

hhid[1] 200001

	hhid	clust	province	district	ur	region	weight	pop_wt	poverty	qu
1	200001	10001	Kigali City	Nyarugenge	Urban	Kigali	159.2486	318.4973	Non Poor	Q5
2	200002	10001	Kigali City	Nyarugenge	Urban	Kigali	159.2486	636.9946	Non Poor	Q5
3	200003	10001	Kigali City	Nyarugenge	Urban	Kigali	159.2486	477.7459	Non Poor	Q5
4	200004	10001	Kigali City	Nyarugenge	Urban	Kigali	159.2486	477.7459	Non Poor	Q5
5	200005	10001	Kigali City	Nyarugenge	Urban	Kigali	159.2486	477.7459	Non Poor	Q5
6	200006	10001	Kigali City	Nyarugenge	Urban	Kigali	159.2486	477.7459	Non Poor	Q5
7	200007	10001	Kigali City	Nyarugenge	Urban	Kigali	159.2486	318.4973	Non Poor	Q4
8	200008	10001	Kigali City	Nyarugenge	Urban	Kigali	159.2486	318.4973	Non Poor	Q5
9	200009	10001	Kigali City	Nyarugenge	Urban	Kigali	159.2486	159.2486	Non Poor	Q5
10	200010	10002	Kigali City	Nyarugenge	Urban	Kigali	141.0833	1128.666	Severely poor	Q1
11	200011	10002	Kigali City	Nyarugenge	Urban	Kigali	141.0833	705.4163	Non Poor	Q3
12	200012	10002	Kigali City	Nyarugenge	Urban	Kigali	141.0833	141.0833	Non Poor	Q5
13	200013	10002	Kigali City	Nyarugenge	Urban	Kigali	141.0833	1128.666	Non Poor	Q5
14	200014	10002	Kigali City	Nyarugenge	Urban	Kigali	141.0833	141.0833	Non Poor	Q5
15	200015	10002	Kigali City	Nyarugenge	Urban	Kigali	141.0833	423.2498	Non Poor	Q4
16	200016	10002	Kigali City	Nyarugenge	Urban	Kigali	141.0833	423.2498	Non Poor	Q5
17	200017	10002	Kigali City	Nyarugenge	Urban	Kigali	141.0833	423.2498	Non Poor	Q5

Ready Vars: 99 Order: Dataset Obs: 14,580 Filter: Off Mode: Browse CAP NUM

d. There are 3 different ways that data can be stored in Stata, and each appears as a different color in the *data browser:*

i. Numeric data – appears in **black**. An example is ***s5aq5***, for which respondents gave a numeric answer to "How many rooms does your household occupy? (Excluding bathroom, toilet, kitchen, corridor and stables)"

ii. Categorical data – appears in **blue**. Categorical data are stored as numbers, but each number value has a non-numeric label assigned to it. An example is ***province***, for which respondents selected which province they live in. Their selections are stored as numbers, but are assigned the labels of the answer that they chose. For example if you click on a cell that says "Kigali City", you will see the number "1" appear in the top bar, indicating that the number 1 is assigned the value "Kigali City".

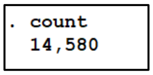
iii. String data – appears in **red**. String data are non-numeric. There are no examples in this dataset, but frequency these are used when a respondent selects, "Other, specify" to a question, and then provides an open-ended response. For example, question s5cq1 asks "Main source of drinking water" and has an option for "Other, specify". Bottled water was not one of the options, so someone could have selected "Other, specify" and then responded with "bottled water".

3. Top descriptive commands

a. Count – reports the number of observations in the dataset.

i. Practice: How many observations are in this dataset? *"count" – 14,580*

Figure 1.4: Stata command and output for "count"



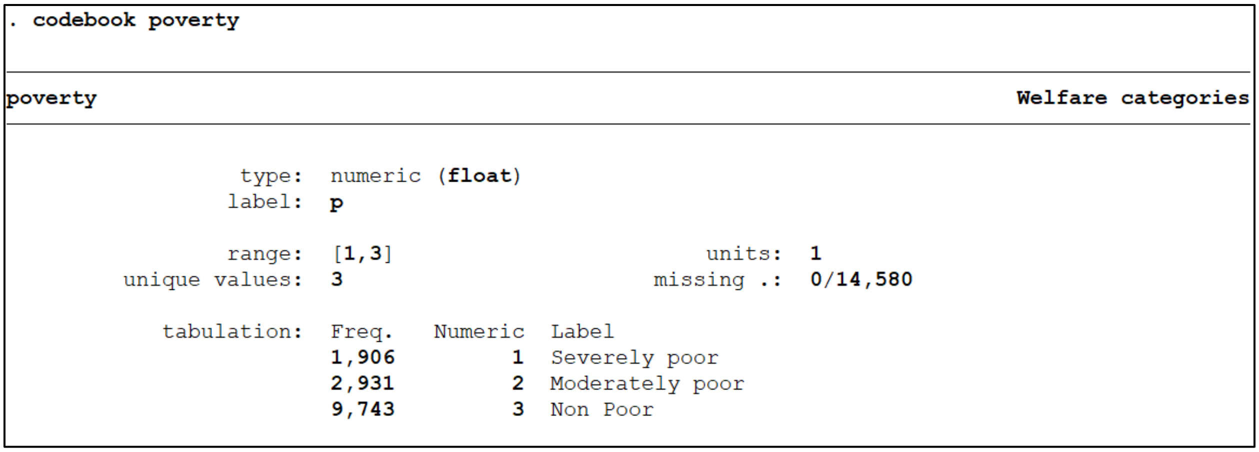
```
. count
  14,580
```

b. Codebook – another way to explore or describe the data; you can do it generally, or for a specific variable.

i. Code: *codebook [variable name]*

ii. Practice: What kind of information do we have on poverty? Type *"codebook poverty"* into Window 2 (Command Window) and press enter. The following output will appear in Window 1 (Review and Results Window). This output tells us:

Figure 1.5: Stata command and output for "codebook poverty"



```
. codebook poverty

-------------------------------------------------------------------------------
poverty                                                      Welfare categories
-------------------------------------------------------------------------------

                  type:  numeric (float)
                 label:  p

                 range:  [1,3]                        units:  1
         unique values:  3                        missing .:  0/14,580

            tabulation:  Freq.   Numeric  Label
                         1,906         1  Severely poor
                         2,931         2  Moderately poor
                         9,743         3  Non Poor
```

a) What is the variable *poverty* showing? (The variable label)? *Welfare categories*

b) What type of data is it? *Numeric (float)*

c) What is the range of the data? *1 to 3*

d) How many missing observations are there? *0*

e) How many unique values are there? *3*

f) What is one of the labels assigned to a value? *1 is Severely poor (for example)*

c. Single tabulations – tell us the frequency of each response

i. Code: ***tab**ulate [variable name]*

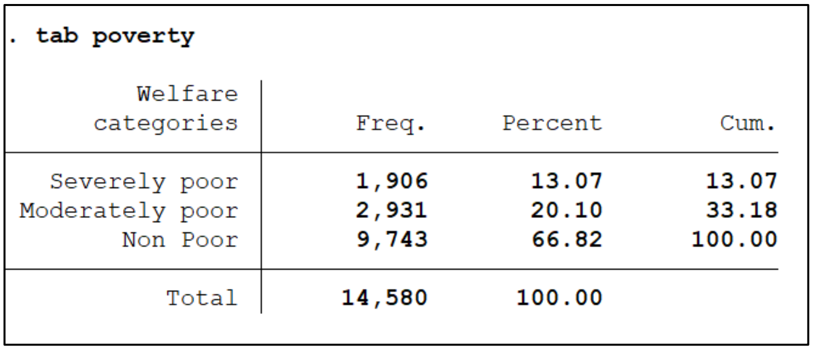
a) Many Stata codes have “shorthand” versions – you only have to type a shortened version of the code and Stata will recognize the full command.

b) The shorthand for *tabulate* is *“tab”* – e.g. *tab [varname]*

c) This manual will **bold** the shorthand for each code when the code is introduced (see above for ***tab**ulate*)

ii. Practice: How many households are severely poor? *“tabulate poverty”* or *“tab poverty”.* This output tells us:

Figure 1.6: Stata command and output for “tab poverty”



```
. tab poverty

         Welfare │
      categories │      Freq.     Percent        Cum.
─────────────────┼───────────────────────────────────
   Severely poor │      1,906       13.07       13.07
 Moderately poor │      2,931       20.10       33.18
        Non Poor │      9,743       66.82      100.00
─────────────────┼───────────────────────────────────
           Total │     14,580      100.00
```

a) “Freq.” - the number of observations (HHs) who responded with each answer

- Question: Looking at the “Freq.” column in Figure 1.6, how many households are severely poor in the sample? *1,906 households*

b) “Percent” – the percent of observations (HHs) who responded with each answer

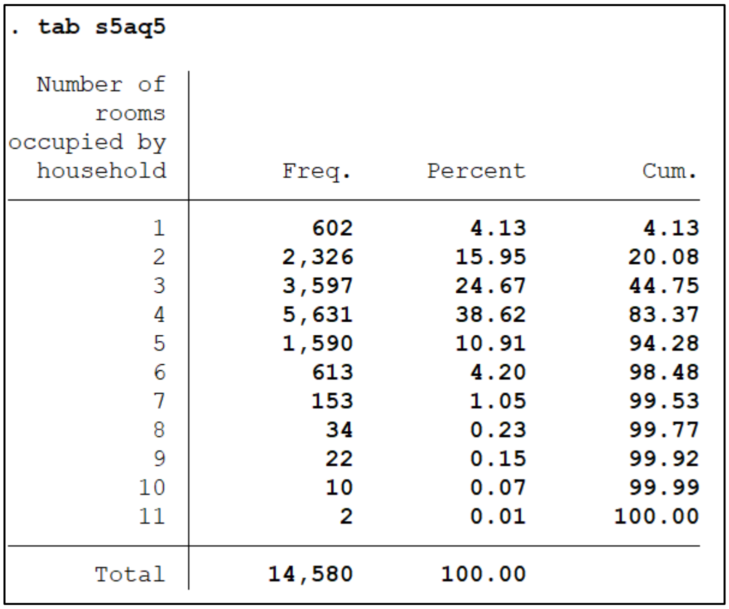
- Question: Looking at the “Percent” column in Figure 1.6, what percent of the households in the sample are severely poor? *13.07% of all surveyed households*

c) “Cum.” – the cumulative percent of the answers across all of the observations (HHs)

d) Question: In Figure 1.6, how many households in the sample are non-poor? *9,743 households*

iii. Question: How many rooms do most households have in their homes? *“tab s5aq5” – the most common number of rooms is 4 – 5,631 households (38.62%) have 4 rooms*

Figure 1.7: Stata command and output for "tab s5aq5"



```
. tab s5aq5

  Number of |
      rooms |
occupied by |
  household |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        602        4.13        4.13
          2 |      2,326       15.95       20.08
          3 |      3,597       24.67       44.75
          4 |      5,631       38.62       83.37
          5 |      1,590       10.91       94.28
          6 |        613        4.20       98.48
          7 |        153        1.05       99.53
          8 |         34        0.23       99.77
          9 |         22        0.15       99.92
         10 |         10        0.07       99.99
         11 |          2        0.01      100.00
------------+-----------------------------------
      Total |     14,580      100.00
```

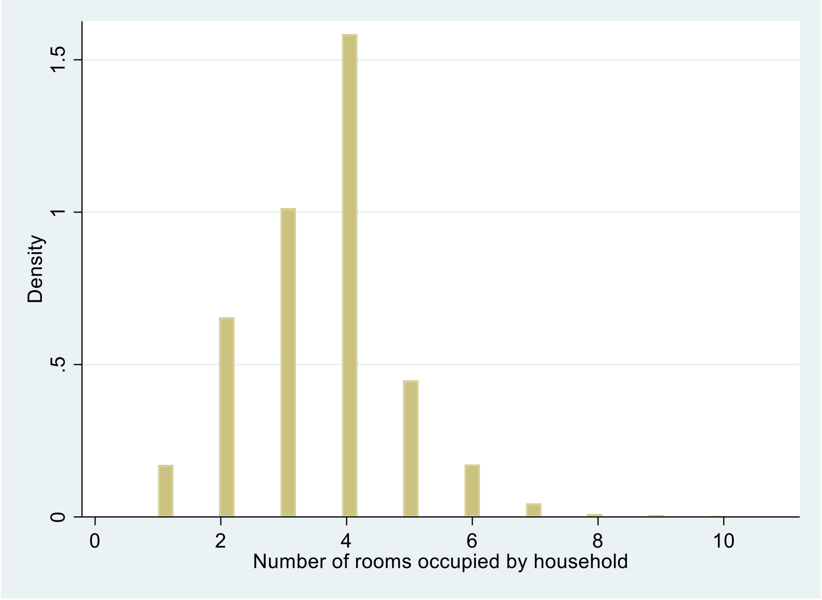
a) What would be more helpful to know? *Mean, median, etc.*

d. Histogram – produces a bar graph of one variable, where the height of each bar is the frequency of the variable at specific values

i. Code: ***hist****ogram [varname]*

ii. Practice: Let's visualize the distribution of the data for variable s5aq5 (number of rooms in each household) by typing "*hist s5aq5*"

Figure 1.8: Stata histogram output for "hist s5aq5"



a) Question: What is the mode (most common answer)? *4 rooms*

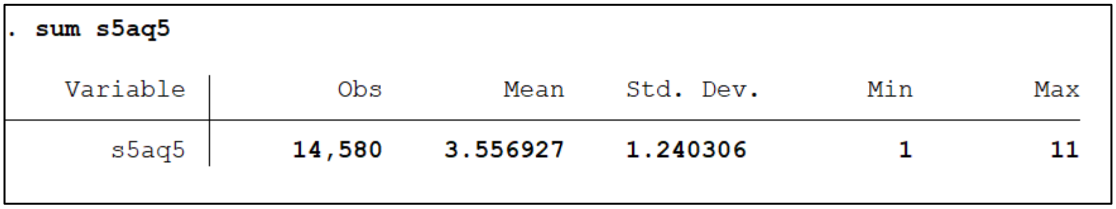
- This figure suggests that most houses have between 1-6 rooms, while very few have more than 6 (and none have 12+).

e. Summarize – outputs the number of observations, average (mean), standard deviation, minimum, and maximum of a numeric variable

i. Code: ***sum**marize [varname]*

ii. Practice: What if we want to know the average number of rooms in households in our sample? *"sum s5aq5"*

Figure 1.9: Stata command and output for "sum s5aq5



```
. sum s5aq5

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       s5aq5 |     14,580    3.556927    1.240306          1         11
```

iii. Challenge: What other code have we used before that gave us the minimum and maximum values of a variable? *codebook*

iv. What is the average number of rooms in the households in our sample? *3.56 rooms*

v. What is the median number of rooms? *This output doesn't tell us the median!*

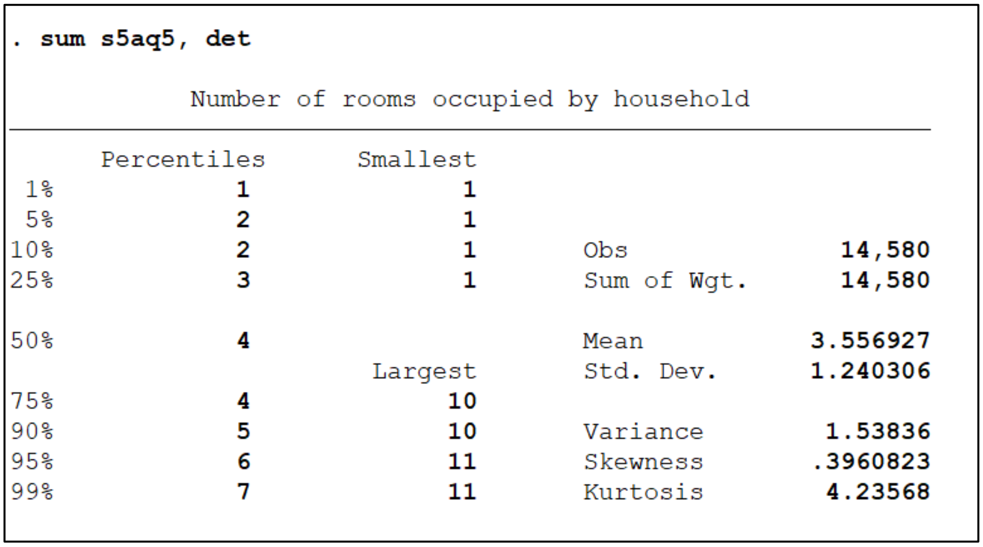
a) We know from the "tab" that very few HHs have a lot of rooms, and we know from this "sum" output that the maximum number is 11 even though the average is 3.56, so the mean could biased or skewed.

b) Most commands have different *options* to add to or adjust the results depending on your research question. *Options* are added to a code by first typing a comma, then writing the code for the options.

c) One of the options for *summarize* is "*detail"* which displays additional statistics

vi. Practice: What is the median number of rooms? "***sum**marize s5aq5, **det**ail"* – this output shows the following:

Figure 1.10: Stata command and output for "sum s5aq5, det"



```
. sum s5aq5, det

              Number of rooms occupied by household
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              1
 5%            2              1
10%            2              1       Obs              14,580
25%            3              1       Sum of Wgt.      14,580

50%            4                      Mean           3.556927
                        Largest       Std. Dev.      1.240306
75%            4             10
90%            5             10       Variance        1.53836
95%            6             11       Skewness       .3960823
99%            7             11       Kurtosis        4.23568
```

a) Percentiles and their ranges

b) The median (50%)

c) Variance, skewness, and kurtosis

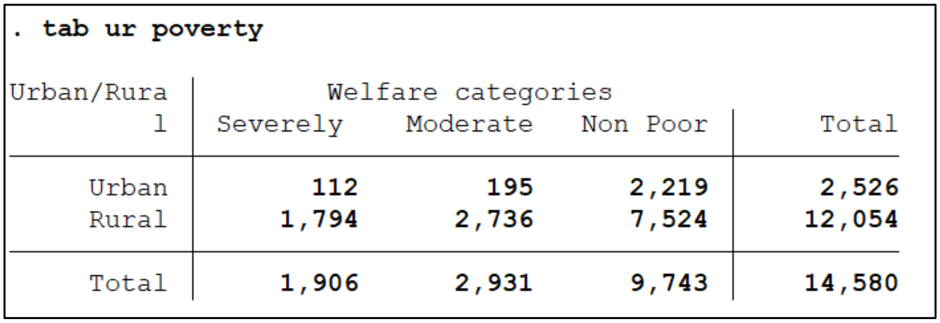
vii. What is the median number of rooms? *4 rooms*

f. Double tabulation – shows the responses to two variables at the same time by creating a two-way table of frequencies

i. <u>Code:</u> ***tab**ulate [varname1] [varname2]*

ii. <u>Practice:</u> Are non-poor households more likely to live in rural or urban areas? *"tab ur poverty"*

Figure 1.11: Stata command and output for "tab ur poverty"



```
. tab ur poverty

Urban/Rura |         Welfare categories
         l |  Severely   Moderate   Non Poor |     Total
-----------+---------------------------------+----------
     Urban |       112        195      2,219 |     2,526
     Rural |     1,794      2,736      7,524 |    12,054
-----------+---------------------------------+----------
     Total |     1,906      2,931      9,743 |    14,580
```

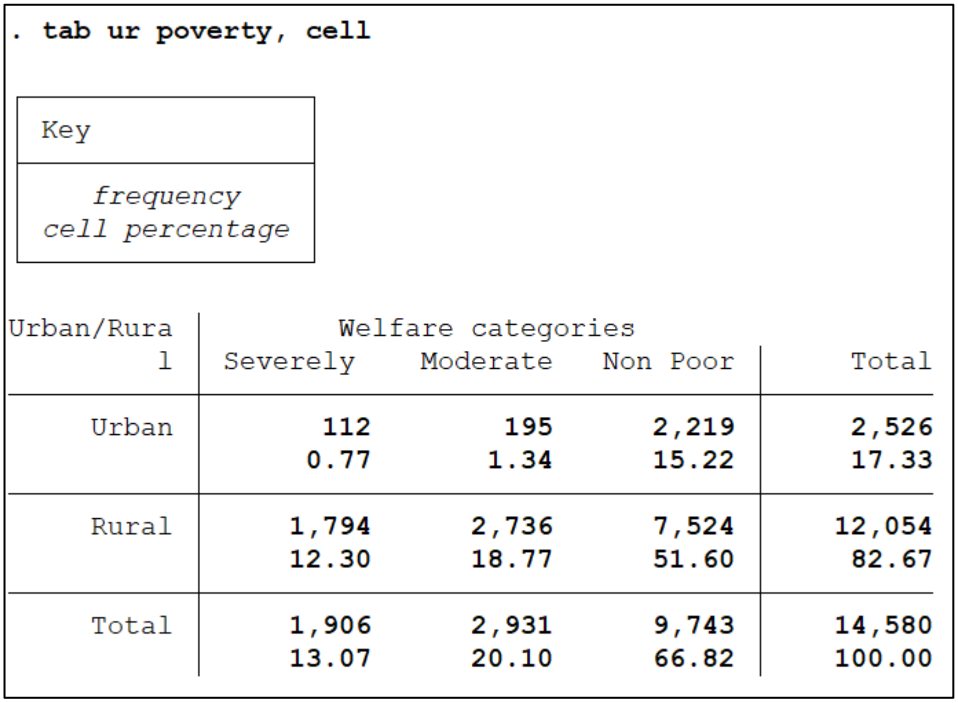
a) This output tells us the number of observations for each combination of responses between the two variables.

b) <u>Question:</u> How many households in rural areas are non-poor? *7,524 households*

iii. What if we want to know what *percent* of households who are in rural areas and are non-poor? - There are many *options* that we can add to the two-way tabulation command to get various types of percentages in the output.

iv. <u>Practice:</u> We want to know what percent of households **in the sample** live in rural areas and are non-poor – *"tab ur poverty, cell"*

Figure 1.12: Stata command and output for "tab ur poverty, cell"



```
. tab ur poverty, cell

+-----------------+
| Key             |
|-----------------|
|    frequency    |
| cell percentage |
+-----------------+

Urban/Rura |         Welfare categories
         l |  Severely   Moderate   Non Poor |     Total
-----------+---------------------------------+----------
     Urban |       112        195      2,219 |     2,526
           |      0.77       1.34      15.22 |     17.33
-----------+---------------------------------+----------
     Rural |     1,794      2,736      7,524 |    12,054
           |     12.30      18.77      51.60 |     82.67
-----------+---------------------------------+----------
     Total |     1,906      2,931      9,743 |    14,580
           |     13.07      20.10      66.82 |    100.00
```

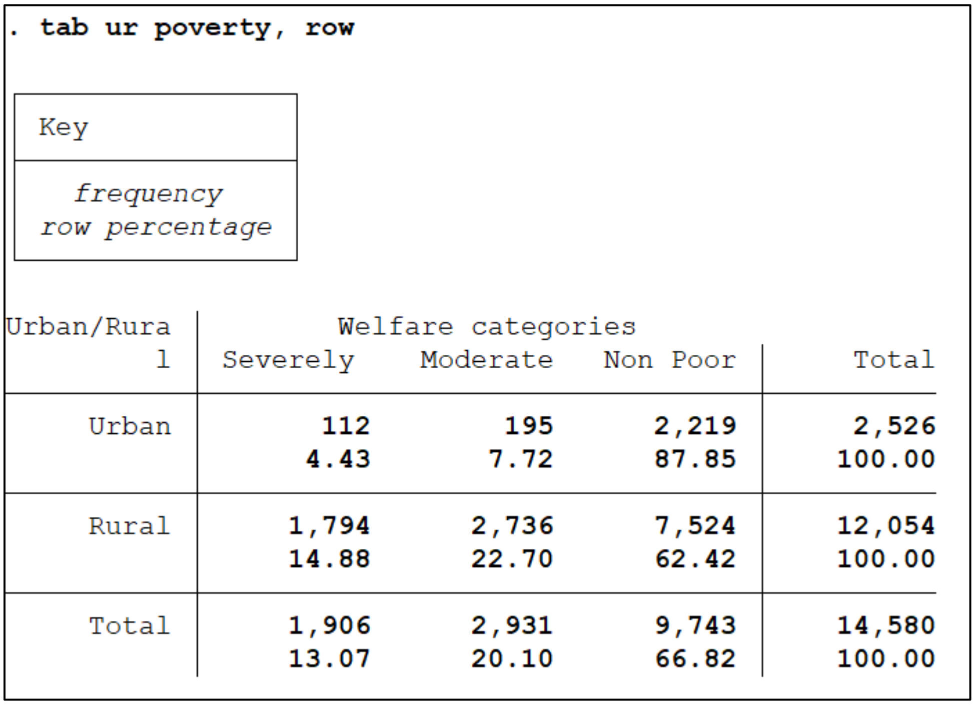
a) The option “*cell*” will tell us the percentage of households in the sample for each combination of responses.

b) All of the percentages in the cells will sum to 100

c) What percentage of households **in the sample** live in rural areas and are non-poor? *51.60% of the households **in the sample** live in rural areas and are non-poor.*

v. Practice: What if we want to know the percentage of households **who live in rural areas** who are non-poor? – *“tab ur poverty, row”*

Figure 1.13: Stata command and output for “tab ur poverty, row”



```
. tab ur poverty, row

+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
+-------------------+

Urban/Rura |        Welfare categories
         l |  Severely   Moderate   Non Poor |     Total
-----------+---------------------------------+----------
     Urban |       112        195      2,219 |     2,526 
           |      4.43       7.72      87.85 |    100.00 
-----------+---------------------------------+----------
     Rural |     1,794      2,736      7,524 |    12,054 
           |     14.88      22.70      62.42 |    100.00 
-----------+---------------------------------+----------
     Total |     1,906      2,931      9,743 |    14,580 
           |     13.07      20.10      66.82 |    100.00 
```

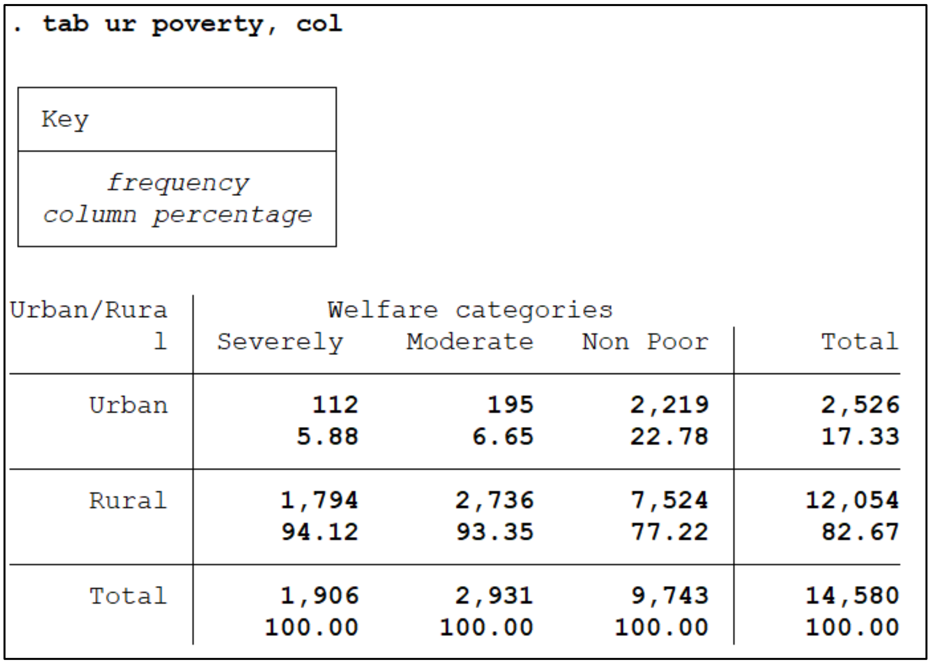
a) The option of “row” will tell us the row percentages – the percentages *in each row* will sum to 100.

b) In this specific code (where *ur* is typed before *poverty*), it will tell us: of the households who live in rural or urban areas, what percentage of them are poor or non-poor.

c) What percentage of households **who live in rural areas** are non-poor? *62.42% of households **who live in rural areas** are non-poor.*

vi. Practice: What if we want to know of the households **who are non-poor** who live in rural areas? – *“tab ur poverty, **col**umn”*

Figure 1.14: Stata command and output for “tab ur poverty, col”



```
. tab ur poverty, col

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

Urban/Rura |        Welfare categories
         l |  Severely   Moderate   Non Poor |     Total
-----------+---------------------------------+----------
     Urban |       112        195      2,219 |     2,526 
           |      5.88       6.65      22.78 |     17.33 
-----------+---------------------------------+----------
     Rural |     1,794      2,736      7,524 |    12,054 
           |     94.12      93.35      77.22 |     82.67 
-----------+---------------------------------+----------
     Total |     1,906      2,931      9,743 |    14,580 
           |    100.00     100.00     100.00 |    100.00 
```

a) The option “*column*” (the shorthand is “*col*”) will tell us the column percentages – the percentages *in each column* will sum to 100.

b) In this specific code (where *ur* is typed before *poverty*), it will tell us: of the households who are poor or non-poor, what percentage of them live in rural or urban areas.

c) What percentage of households **who are non-poor** live in rural areas?
*77.22% of households **who are non-poor** live in rural areas.*

4. Using “if” – a way to limit your output to certain observations that meet your defined criteria

a. When using “if”, we first need to know the logical operators for Stata:

~	not
==	equal
~=	not equal
!=	not equal
>	greater than
>=	greater than or equal to
<	less than
<=	less than or equal to
&	and
|	or

b. We can use “if” to answer one of the questions we just asked (What percentage of households who live in rural areas are non-poor?)

c. How might we write an expression to *tabulate* poverty status if the household lives in a rural area?

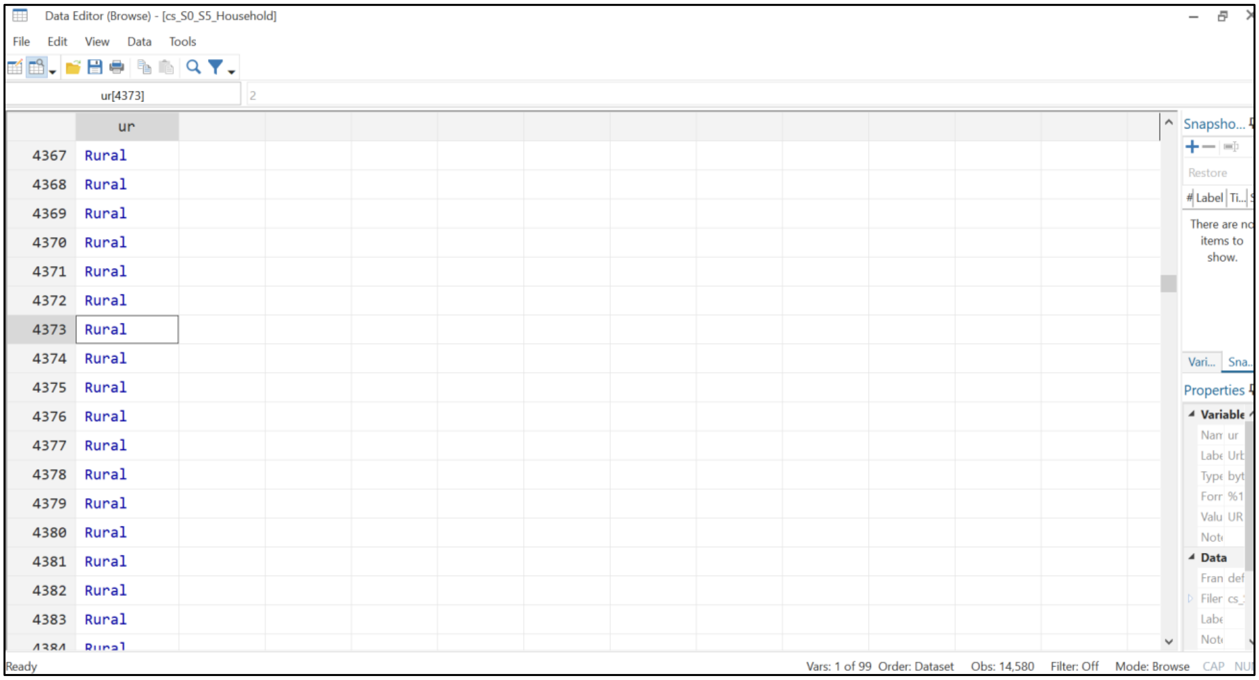
i. First, we know that urban/rural (*ur*) is a categorical variable (meaning “urban” and “rural” are labels assigned to a number value), so we need to know what number means “rural”. There are three easy ways to check this:

a) Practice: *"codebook ur"* – shows us that "rural"' is the label for 2

b) Practice: *"tab ur"* followed by *"tab ur, nolab"* – this will first show the tabulation using the labels, then it utilizes the tabulate option ***"nolabel"*** which will show the same output with the numerical values in place of labels

c) Practice: *"browse ur"* – click on a cell that says "rural" and look at what number is in the top bar (see Figure 1.15 below)

Figure 1.15: Stata data browser window for "browse ur"

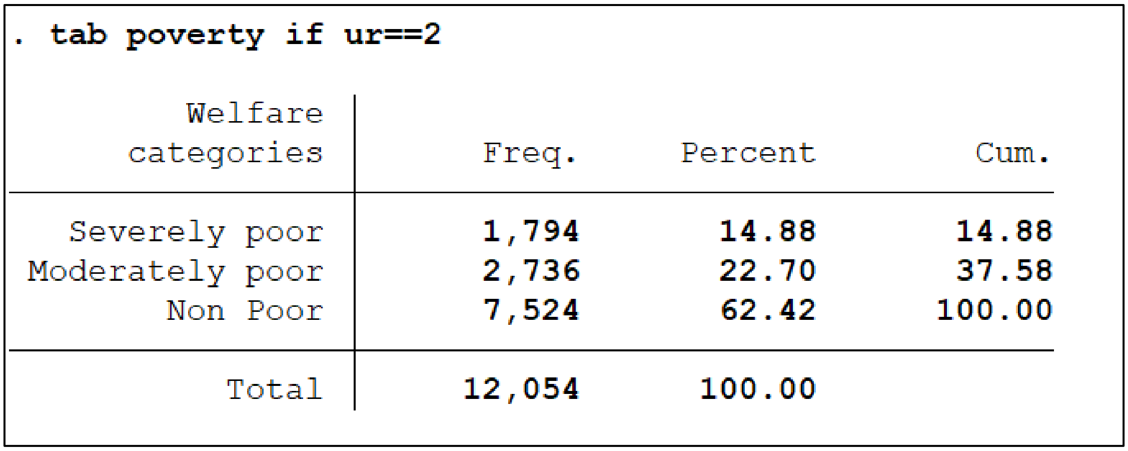


```
ur[4373]        2

        ur
4367    Rural
4368    Rural
4369    Rural
4370    Rural
4371    Rural
4372    Rural
4373    Rural
4374    Rural
4375    Rural
4376    Rural
4377    Rural
4378    Rural
4379    Rural
4380    Rural
4381    Rural
4382    Rural
4383    Rural
4384    Rural
```

```
Ready          Vars: 1 of 99   Order: Dataset   Obs: 14,580   Filter: Off   Mode: Browse   CAP   NU
```

ii. Question: Now that we know that ur equals 2 for "rural", how can we write the tabulation command using "if" to answer the question? (What percentage of households who live in rural areas are non-poor?) - *"tab poverty if ur==2"*

Figure 1.16: Stata command and output for "tab poverty if ur==2"

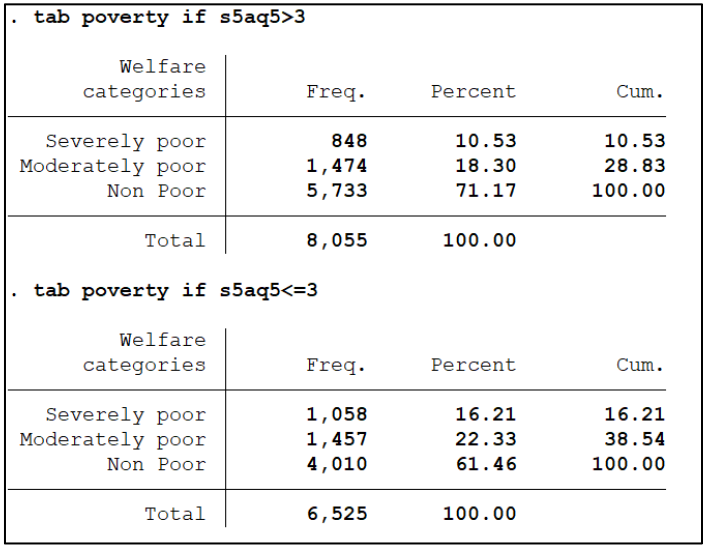


```
. tab poverty if ur==2

          Welfare |
       categories |      Freq.     Percent        Cum.
------------------+-----------------------------------
    Severely poor |      1,794       14.88       14.88
  Moderately poor |      2,736       22.70       37.58
         Non Poor |      7,524       62.42      100.00
------------------+-----------------------------------
            Total |     12,054      100.00
```

        a) Now we see this single tabulation of poverty *only* for the households that live in rural areas. Again, we see that of the households who live in rural areas, 62.42% are non-poor.
        b) Notice that the Total number under "Freq." (the number of observations included in this output) is smaller than the number if our dataset (this output shows only 12,054 households while our dataset has 14,580). This is because 12,054 households live in rural areas, and that is what we wanted to restrict this single tabulation to.

d. Let's try another tabulation using "if" with a different logical operator. Maybe we want to know if there are any patterns between the number of rooms in a household and their poverty status.
    i. Question: What variable tells us the number of rooms in a household? *s5aq5*
    ii. Question: What variable tells us the poverty status? *poverty*
    iii. Question: How could we see the poverty status for households that have *more than 3 rooms*? *"tab poverty if s5aq5>3" OR "tab poverty if s5aq5>=4"*
    iv. Question: How about the poverty status for households with *exactly 3 rooms*? *"tab poverty if s5aq5==3"*
    v. Question: How about the poverty status for households with *3-5 rooms*? *"tab poverty if s5aq5>=3 & s5aq5<=5" OR "tab poverty if s5aq5>2 & s5aq5<6"*
    vi. Question: How about the poverty status for households with l*ess than 3 rooms or greater than 5 rooms*? *"tab poverty if s5aq5<3 | s5aq5>5" OR "tab poverty if s5aq5<=2 | s5aq5>=6"*
    vii. Question: Which households are more likely to be non-poor? Households with more than 3 rooms? Or households with 3 rooms or less? What codes will you run to show this?
        a) *"tab poverty if s5aq5>3" and "tab poverty if s5aq5<=3"*
        b) *71.17% of households with more than 3 rooms are non-poor, but only 61.46% of households with 3 rooms or less are non-poor. So households with more rooms are more likely to be non-poor.*

Figure 1.17: Stata command and output for "tab poverty if s5aq5>3" and "tab poverty if s5aq5<=3"

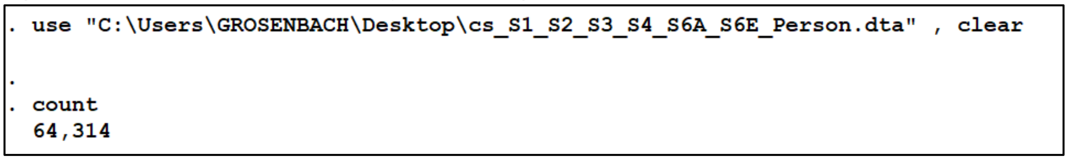


```
. tab poverty if s5aq5>3

        Welfare |
     categories |      Freq.     Percent        Cum.
----------------+-----------------------------------
  Severely poor |        848       10.53       10.53
Moderately poor |      1,474       18.30       28.83
       Non Poor |      5,733       71.17      100.00
----------------+-----------------------------------
          Total |      8,055      100.00

. tab poverty if s5aq5<=3

        Welfare |
     categories |      Freq.     Percent        Cum.
----------------+-----------------------------------
  Severely poor |      1,058       16.21       16.21
Moderately poor |      1,457       22.33       38.54
       Non Poor |      4,010       61.46      100.00
----------------+-----------------------------------
          Total |      6,525      100.00
```

Lesson 2 – Transforming Data

1. **Review of Lesson 1 – Describing the new dataset (cs_S1_S2_S3_S4_S6A_S6E_Person – household roster)**
   a. The datafile we are working with in Lesson 2 is the household roster – it will give us basic information about the people in each household. Remember, in Lesson 1 we looked at household characteristics, so each observation in the data was one household.
      i. In a household survey, where a "household" is the main observation unit, any data at the *household-level* will be called ***wide*** data – it has one unique identifier (HHID – household identifier) per observation
      ii. Because the household roster is at the *person-level* and has many *person* observations for each household, it is called ***long*** data – you need more than one identifier to identify each observation/person because there are multiple observations of the household identifier (e.g. you need HHID and person ID or PID to identify individuals)
         a) Another example of ***long*** data in this questionnaire is the crop production module. This module asked many questions about *each crop* grown by the household, and so this module is at the *crop-level*, and there are many observations for each household.
         b) It is possible to transform ***long*** data into ***wide*** data, mainly by creating summary statistic variables from the ***long*** data at the observation level in the ***wide*** data (e.g. household level). We will do so later in Lesson 2, Part 5, using the "collapse" command.
   b. Do you think we will have more or less observations in our dataset today? *More*
   c. Question: How can we see how many observations we have? *"count"*

Figure 2.1: Stata command and output for using the new dataset and "count"

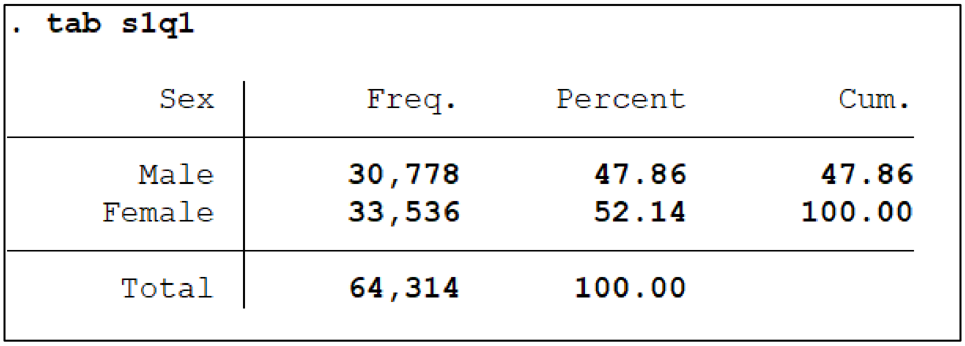


```
. use "C:\Users\GROSENBACH\Desktop\cs_S1_S2_S3_S4_S6A_S6E_Person.dta" , clear

.
. count
  64,314
```

      i. How many observations are in this dataset? *63,314*
      ii. There are more observations because this is the household roster – it has information for all of the household members in the sample. So it is at the *person* level, instead of the *household* level.
   d. Let's look at how many men vs women are in this dataset:
      i. Looking at Window 3 (the variables window), which variable will tell us about the number of men vs women? *s1q1 – "Sex"*
         - When looking for this variable, did you notice that there is no "name" variable? Any identifying information should always be removed from public and shared datasets to keep the respondents anonymous.
      ii. Question: What code(s) can we use to see the number of men vs women in the dataset?
         a) *tabulate ("tab s1q1")*

b) *codebook ("codebook s1q1")* would also work, but tabulate is the best option because it will also tell us the percentage rather than just the frequency

Figure 2.2: Stata command and output for "tab s1q1"



```
. tab s1q1

        Sex |      Freq.     Percent        Cum.
------------+-----------------------------------
       Male |     30,778       47.86       47.86
     Female |     33,536       52.14      100.00
------------+-----------------------------------
      Total |     64,314      100.00
```

iii. Question: How many males are in this dataset? *30,778*

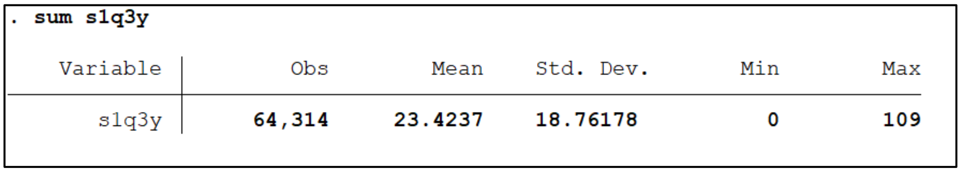
iv. Question: What percentage of the people in this dataset are males? *47.86%*

e. Let's look at the average age of everyone in this dataset:

i. Looking at Window 3, which variable will tell us about the age of the household members? *s1q3y – "Age _ Years_?"*

ii. Question: What code can we use to see the average age of all people in this dataset? *Summarize ("sum s1q3y")*

Figure 2.3: Stata command and output for "sum s1q3y"



```
. sum s1q3y

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       s1q3y |     64,314     23.4237    18.76178          0        109
```

iii. What is the average age of people in this dataset? *23.42 years*

iv. What is the maximum age of people in this dataset? *109 years – this is quite high. For the purposes of this training, let's assume any age over 100 is a mistake. We'll clean these data in the next section.*

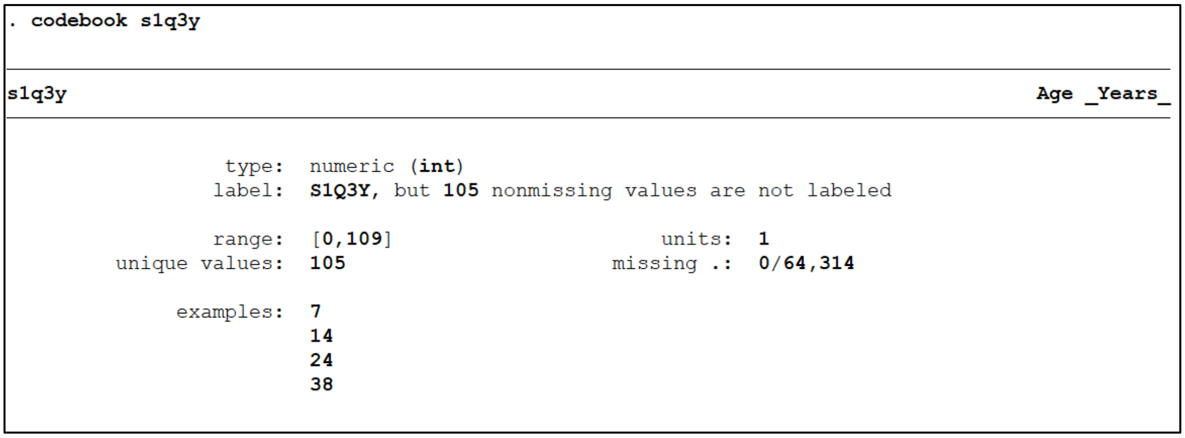
2. Transforming Existing Data

a. Missing values

i. Which command tells us how many observations have a missing value for a variable? *Codebook*

a) Question: How can we see if the age variable (s1q3y) has any missing values currently? *"codebook s1q3y"*

Figure 2.4: Stata command and output for "codebook s1q3y"



```
. codebook s1q3y

-------------------------------------------------------------------------------------
s1q3y                                                                    Age _Years_
-------------------------------------------------------------------------------------

                  type:  numeric (int)
                 label:  S1Q3Y, but 105 nonmissing values are not labeled

                 range:  [0,109]                      units:  1
         unique values:  105                      missing .:  0/64,314

              examples:  7
                         14
                         24
                         38
```

b) How many missing values are there currently in s1q3y? *0 missing values*

c) Now we need to change all values greater than 100 to missing values.

ii. Missing values appear in two different ways in Stata, depending on the variable's data type:

a) Numeric variables are coded as a period (.) for missing values

- To change values in a numeric variable to a missing value, you would type: *replace [varname]=. if …*

b) String variables are coded as a blank ("") for missing values

- To change values in a string variable to a missing value, you would type: *replace [varname]= " " if …*

c) Question: What type of data is s1q3y? *"codebook s1q3y" – numeric*

d) Question: So what type of missing value do we want to change the values greater than 100 to? *A period (.)*

b. Recoding values – there are two ways that we can change these values.

i. Code: *recode [varname] [original_value]=[new_value]*

- Practice: *"recode s1q3y 101 102 103 104 105 106 107 108 109=."*

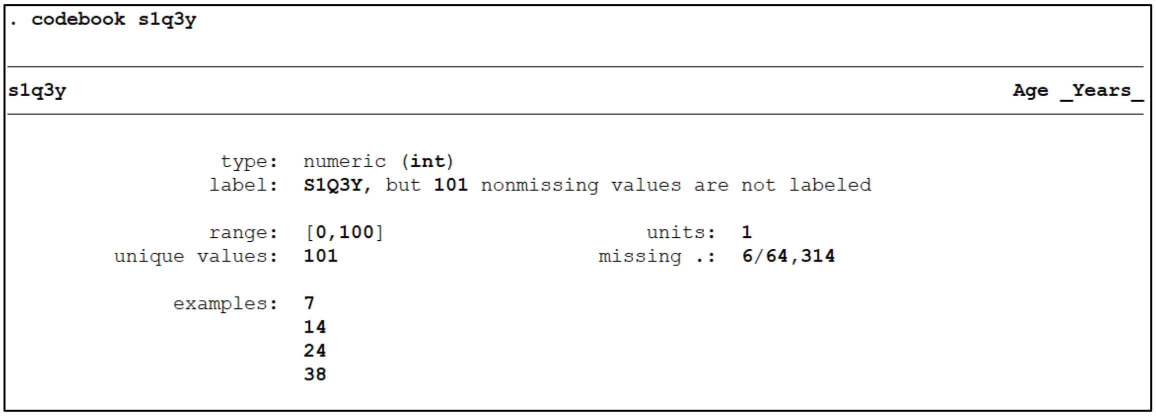
ii. Code: *replace [varname]=[new_value] if [varname]==[old_value]*

- Practice: *"replace s1q3y=. if s1q3y>100"*

iii. Question: After using one of these codes, how many missing values are there now for a4? *"codebook s1q3y"; 6 missing values*

iv. Now that we know that (.) means missing, we can also type "*tab s1q3y, **missing**"* to see how many missing values we have. This tabulate option *"**missing**"* includes the missing values in the tabulation

Figure 2.5: Stata command and output for "codebook s1q3y", after recoding values greater than 100 to missing

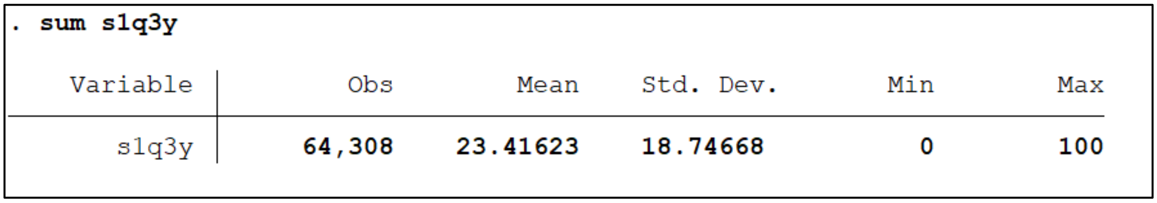


```
. codebook s1q3y

------------------------------------------------------------------------------------------
s1q3y                                                                          Age _Years_
------------------------------------------------------------------------------------------

                  type:  numeric (int)
                 label:  S1Q3Y, but 101 nonmissing values are not labeled

                 range:  [0,100]                      units:  1
         unique values:  101                      missing .:  6/64,314

              examples:  7
                         14
                         24
                         38
```

v. Question: Has the average age changed much due to our cleaning?

a) *"sum s1q3y"*

Figure 2.6: Stata command and output for "sum s1q3y", after recoding values greater than 100 to missing



```
. sum s1q3y

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       s1q3y |     64,308    23.41623    18.74668          0        100
```

b) *The average (mean) age is still 23.42 because only 6 observations were more than 100 (although the average is still slightly lower than before). This average does not take into account the 6 missing values (Obs is now 64,308 instead of 64,314)*

c. Changing labels – let's change the name and the label of our age variable to be more intuitive

i. Renaming a variable

a) Code: ***ren**ame [old_varname] [new_varname]*

b) Practice: *"**ren**ame s1q3y age"* – renames the variable to "age" (more intuitive than "s1q3y")

ii. Changing/adding a variable label (see in Window 3 – the Variable Window)

a) Code: ***lab**el **var**iable [varname] ["label"]*

b) Practice: *"**lab**el **var**iable age "Age of household member"* – changes the variable label

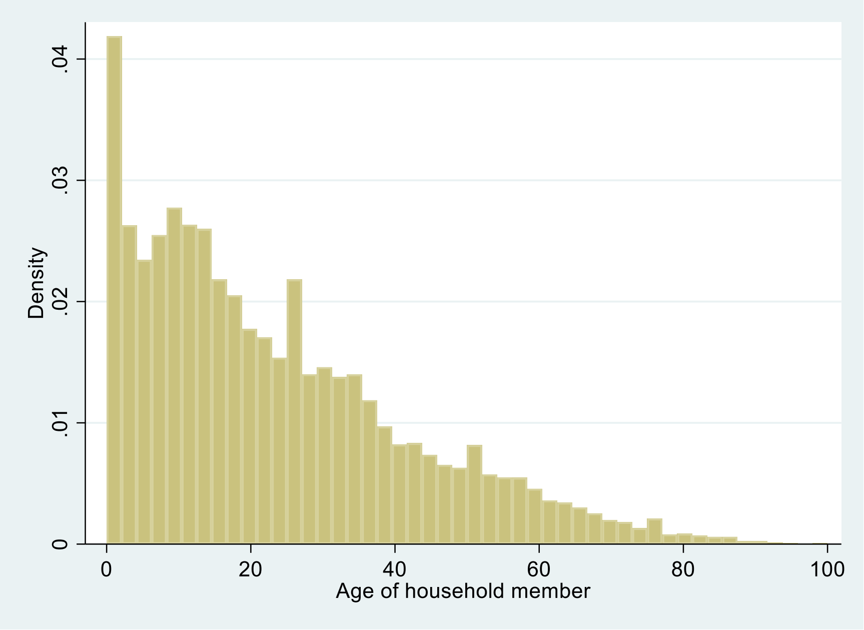
- The label has to be in quotations
- Remember, now the variable is named "age" instead of "s1q3y", so we have to use its new name when writing commands

d. Visualizing the data

i. Now, let's visualize our newly cleaned *age* variable. What command did we learn in Lesson 1 to view a bar figure of the data? *histogram*

ii. Question: How would we write it to view this newly cleaned *age* variable? *"hist age"*

Figure 2.7: Stata histogram output for "hist age"



iii. What can we learn from this figure?
   a) *The majority of people in the dataset are 30 years old or younger*
   b) *There are very few people in the dataset older than 80*

3. Creating New Variables

a. Dummy variables

i. A dummy variable (also known as an indicator variable or a binary variable) takes the value 0 or 1 to indicate the absence or presence of some categorical effect

ii. A dummy variable is a type of categorical variable – it is saved in Stata as numbers (0 and 1), but each number has a label assigned to it ("No" and "Yes")

iii. For example, it may be useful to have a variable that easily indicates whether or not someone on the household roster is a child (15 years or younger)

iv. Why might a variable like this be helpful? What could it help to easily show us?
   a) *How many children are in the sample*
   b) *What percent of children are in school*

b. Generating a new variable - let's make a dummy variable for whether a household member is a child (15 years old or younger). We will be creating a categorical variable (with two categories – "no" and "yes") from a numeric/continuous variable (age).

i. Code: ***gen**erate [new_varname]=[value]*

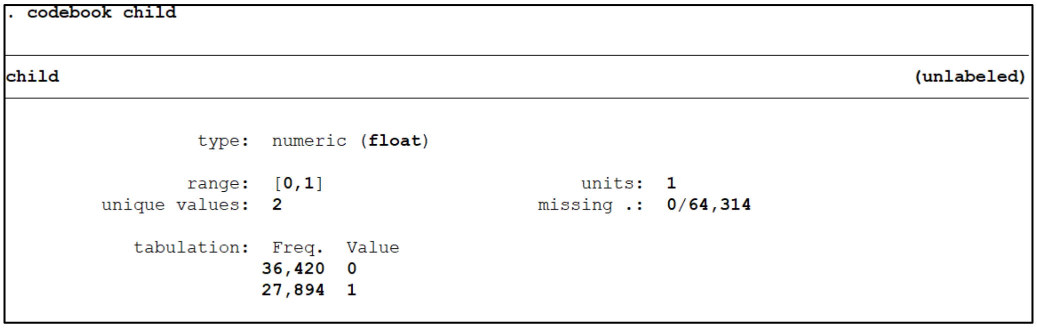
ii. Practice: *"**gen**erate child=."* – creates the new variable named "child", and makes all observations missing.

iii. Question: Now we want to change all of the observations to 1 if the person is 15 years old or younger. How do we change values? *"replace child=1 if age<=15 OR replace child=1 if age<16"*

iv. Question: How can we change all observations to 0 if the person is an adult (16+ years old)? *"replace child=0 if age>15 OR replace child=0 if age>=16"*

v. Let's look at our new variable so far. How can we see an overview of the new variable? *"codebook child"*

Figure 2.8: Stata command and output for “codebook child”



```
. codebook child

---------------------------------------------------------------------------
child                                                          (unlabeled)
---------------------------------------------------------------------------

                  type:  numeric (float)

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/64,314

            tabulation:  Freq.  Value
                        36,420  0
                        27,894  1
```

vi. How many missing values are there for our “child” variable? *0 – this is a problem! We changed the ages of 6 people from greater to 100 to missing because we don’t know their ages, but now every person is either characterized as a child or an adult…*

vii. Stata treats missing values in numeric variables as “infinite”, and so when we use > or >=, all missing values are included.

viii. <u>Question:</u> How can we recode our “child” variable to be missing if we don’t know the age of someone?

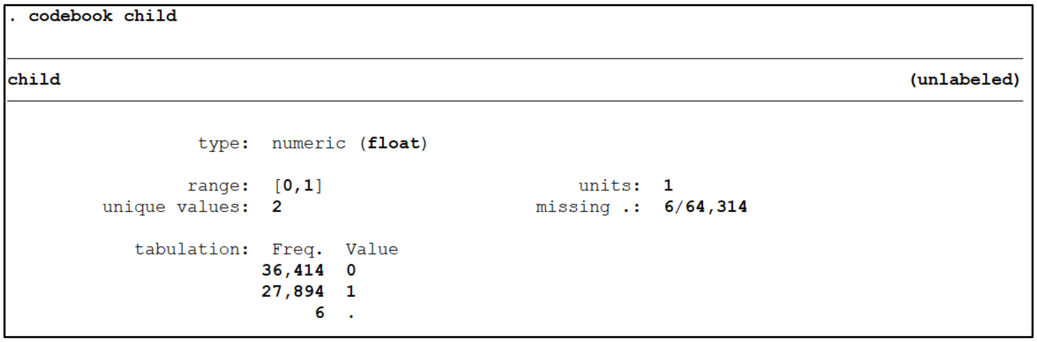
a) *“replace child=. if age==.”*

b) Note the use of the single vs. double equals sign

c. Labeling the new variable and its values:

i. Now, let’s see what our new variable looks like again. How can we get an overview of our new variable? *“codebook child”*

Figure 2.9: Stata command and output for “codebook child” after assigning missing values



```
. codebook child

---------------------------------------------------------------------------
child                                                          (unlabeled)
---------------------------------------------------------------------------

                  type:  numeric (float)

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  6/64,314

            tabulation:  Freq.  Value
                        36,414  0
                        27,894  1
                             6  .
```

a) What are we still missing? *Labels!*

- Our variable has no variable label, so we don’t know what it’s telling us or what “child” is defined as
- Also, our variable is just a bunch of 0s and 1s, we have to apply labels to the 0s and 1s to provide meaning to the categories

ii. <u>Question:</u> How can we assign a variable label to our new variable? *“**lab**el **var**iable child “Is the household member 15 years or younger?””*

iii. To assign value labels, first we have to *define* a set of value labels

a) <u>Code:</u> ***lab**el **def**ine [value_label_definition] # [“label”] # [“label”]*

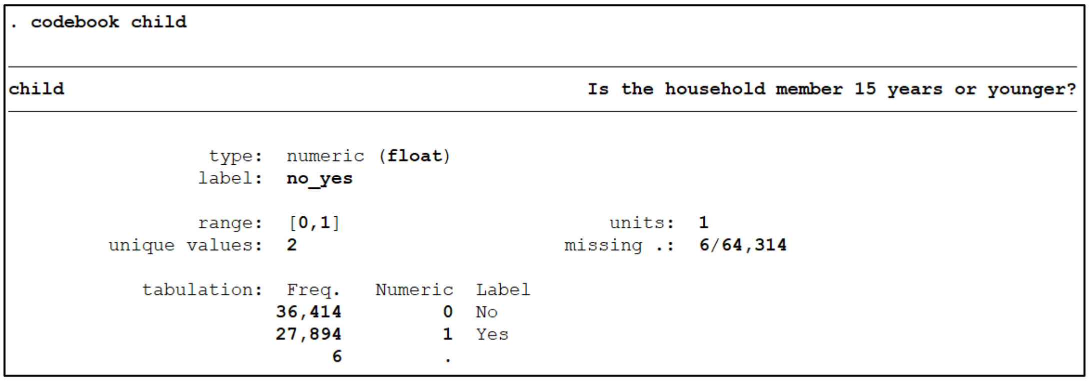
b) <u>Practice:</u> *“**lab**el **def**ine no_yes 0 “No” 1 “Yes””* – Creates a value label definition called “no_yes”; can be applied to any dummy variable, for which 0s are coded as “no” and 1s are coded as “yes”; right now, this value label definition is just saved in Stata’s memory, it has not been applied to any variables yet.

iv. Now, we have to apply our new value label definition to our variable

- a) <u>Code:</u> ***lab**el **val**ues [varname] [value_label_definition]*
- b) <u>Practice:</u> *“**lab**el **val**ues child no_yes”* – applies our newly created “no_yes” value label definition to our variable “child”

v. Let’s take one final look at the overview of our new variable, and see if we labeled everything – *“codebook child”*

Figure 2.10: Stata command and output for “codebook child” after adding labels



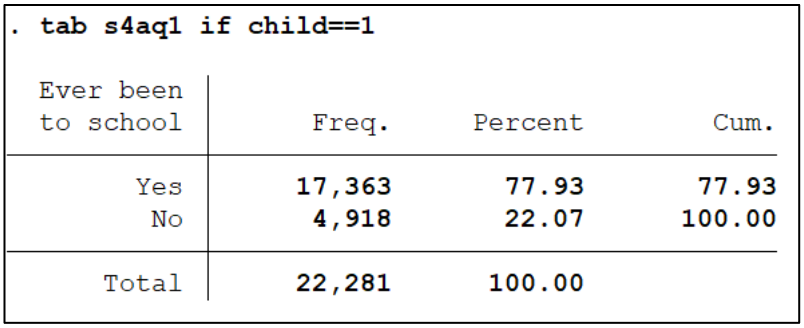
```
. codebook child

-------------------------------------------------------------------------------------------
child                                               Is the household member 15 years or younger?
-------------------------------------------------------------------------------------------

                  type:  numeric (float)
                 label:  no_yes

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  6/64,314

            tabulation:  Freq.   Numeric  Label
                        36,414         0  No
                        27,894         1  Yes
                             6         .
```

vi. How many children are in our dataset? *27,894*

vii. <u>Challenge:</u> What percent of children have ever been to school?

- a) *“tab s4aq1 if child==1”*

Figure 2.11: Stata command and output for “tab s4aq1 if child==1”



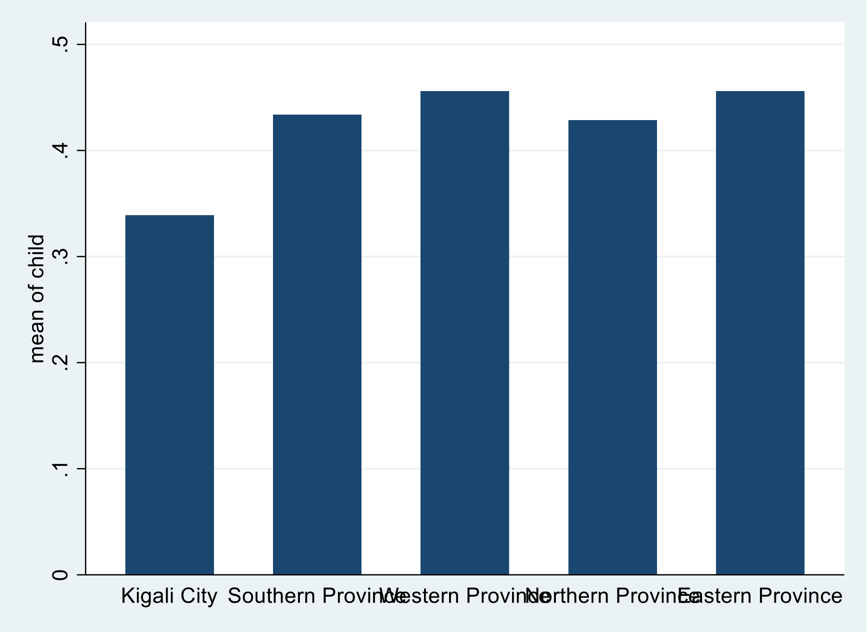
```
. tab s4aq1 if child==1

  Ever been |
  to school |      Freq.     Percent        Cum.
------------+-----------------------------------
        Yes |     17,363       77.93       77.93
         No |      4,918       22.07      100.00
------------+-----------------------------------
      Total |     22,281      100.00
```

- b) *77.93% of children 15 and under have ever been to school. (Keep in mind that this includes children ages 0-15 years, so it makes sense that not all will have been to school yet, since some are still too young).*

viii. <u>Challenge:</u> Which province has the highest percentage of children? Let’s use a bar graph to find out!

- a) <u>Code:</u> *graph bar [varname], over([groupvar])*
  - Note that the “over([groupvar])” is an option (it comes after a comma) and is not necessary to the code. This option will create different bars for the different categories in the [groupvar]
- b) <u>Practice:</u> “*graph bar child, over(province)”*

Figure 2.12: Stata bar graph output for "graph bar child, over(province)"



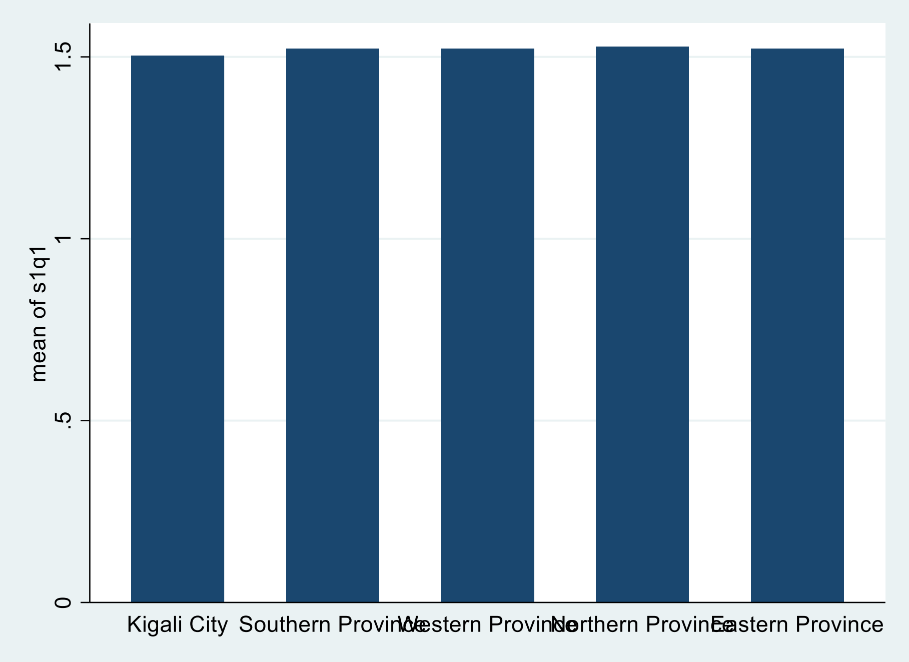
c) Because the values for *child* are 0 and 1, a bar graph (ranging from 0 to 1) shows the prevalence of the dummy variable (in the graph, 1=100%).
   - Similarly, we can also find the prevalence of a dummy variable by calculating its average. Which command tells us the average? *sum*
   - Note: The figures section of the manual discusses how to add and format labels to avoid overlap

d) *Western and Eastern Provinces have the highest prevalence/percentage of children aged 15 and under*

ix. Now, let's create the same graph, but instead showing the prevalence of women by province

a) Practice: "*graph bar s1q1, over(province)*"

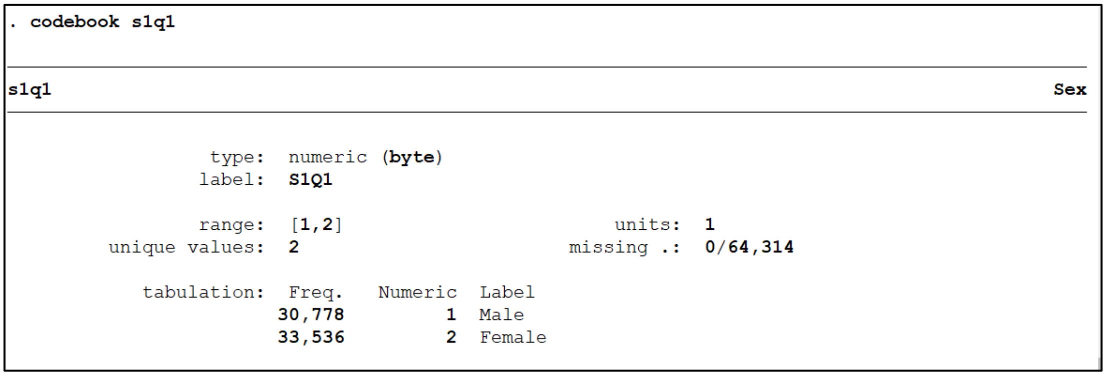
Figure 2.13: Stata bar graph output for "graph bar s1q1, over(province)"



b) Question: Does this show the percent of women in each province? *No! All of the bars are over 1*

c) Practice: Why is this figure different than the child figure? Let's look again at the gender variable. *"codebook s1q1"*

Figure 2.14: Stata command and output for "codebook s1q1"



```
. codebook s1q1

-------------------------------------------------------------------------------
s1q1                                                                        Sex
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  S1Q1

                 range:  [1,2]                        units:  1
         unique values:  2                        missing .:  0/64,314

            tabulation:  Freq.   Numeric  Label
                        30,778         1  Male
                        33,536         2  Female
```

d) The value labels are 1 and 2, instead of 0 and 1 – s1q1 is not a dummy variable!

x. Practice: Let's create a new dummy variable for whether or not a household member is a woman.

a) *"gen woman=."* – creates a new variable named *woman*, and sets all values to missing

b) *"replace woman=1 if s1q1==2"* – changes all values of the *woman* variable to 1 if the household member is a woman (coded as 2 in the variable s1q1)

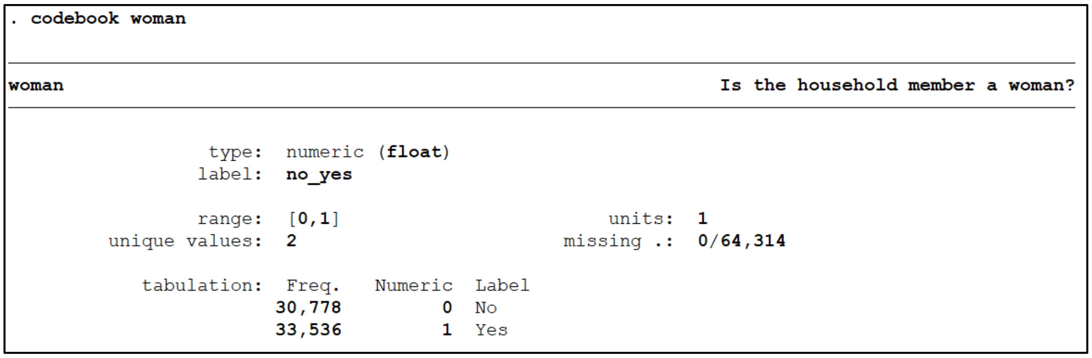
c) "*replace woman=0 if s1q1==1"* – changes all values of the *woman* variable to 0 if the household member is a man (coded as 1 in the variable s1q1)

d) *"lab var woman "Is the household member a woman?""* – labels the new variable

e) *"lab val woman no_yes"* – applies the value label definition that we created earlier, named "no_yes" to our new variable *woman*

f) Let's look at our new variable – *"codebook woman"*

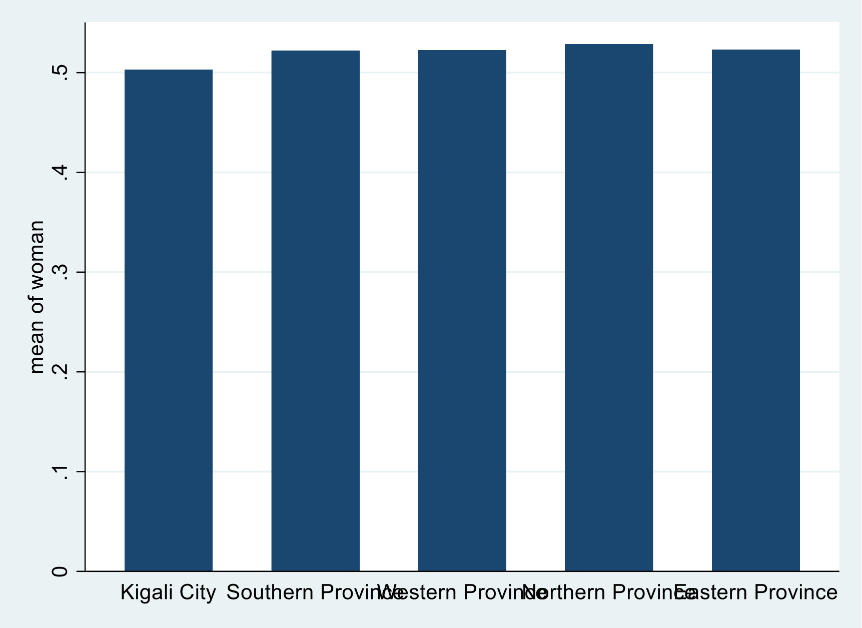
Figure 2.15: Stata command and output for "codebook woman"



```
. codebook woman

-------------------------------------------------------------------------------------------------
woman                                                        Is the household member a woman?
-------------------------------------------------------------------------------------------------

                  type:  numeric (float)
                 label:  no_yes

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/64,314

            tabulation:  Freq.   Numeric  Label
                        30,778         0  No
                        33,536         1  Yes
```

g) Let's look at the bar graph again with our new variable instead of s1q1 – *"graph bar woman, over(province)"*

Figure 2.16: Stata bar graph output for "graph bar woman, over(province)"



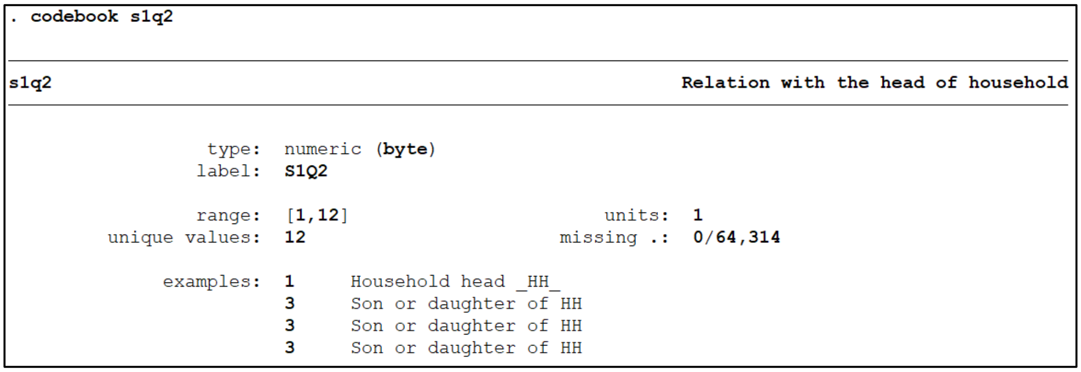
h) Question: Does it show prevalence now? *Yes!*

i) Question: Which province has the *lowest* prevalence of women in the sample? *Kigali City*

xi. Practice: Let's create a final new dummy variable for whether the household head is a woman

a) First we have to see how household head is coded. "codebook s1q2"

Figure 2.17: Stata command and output for "codebook s1q2"



```
. codebook s1q2

------------------------------------------------------------------------------
s1q2                                         Relation with the head of household
------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  S1Q2

                 range:  [1,12]                       units:  1
         unique values:  12                       missing .:  0/64,314

              examples:  1     Household head _HH_
                         3     Son or daughter of HH
                         3     Son or daughter of HH
                         3     Son or daughter of HH
```

b) The value 1 has the label "Household head _HH_"

c) *"gen hhh_female=0"* – creates a new variable named *hhh_female*, and sets all values to 0

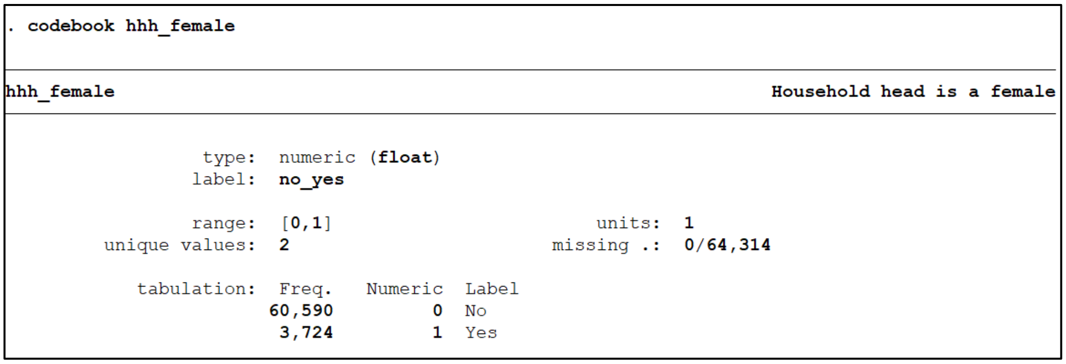
d) *"replace hhh_female=1 if woman==1 & s1q2==1"* – changes all values of the *hhh_female* variable to 1 if the household member is a woman (using our new dummy variable) ***and*** if the relationship to the household head is "household head" (coded as 1 in the variable s1q2)

e) *"lab var hhh_female "Household head is a female"* – labels the new variable

f) *"lab val hhh_female no_yes"* – applies the value label definition that we created earlier, named "no_yes" to our new variable *hhh_female*

g) Let's look at our new variable – "*codebook hhh_female"*

Figure 2.18: Stata command and output for "codebook hhh_female"



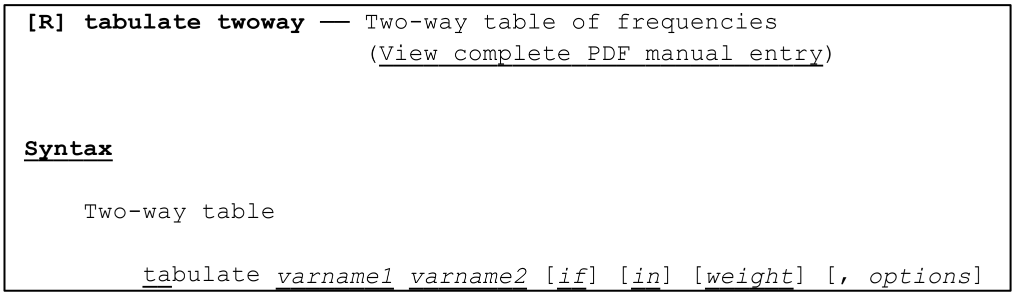
```
. codebook hhh_female

------------------------------------------------------------------------------
hhh_female                                          Household head is a female
------------------------------------------------------------------------------

                  type:  numeric (float)
                 label:  no_yes

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/64,314

            tabulation:  Freq.   Numeric  Label
                        60,590         0  No
                         3,724         1  Yes
```

h) How many people in the sample are women *and* the household head? *3,724*

4. Help Window – can type "help [command]" to pull up a window describing how to use that command

a. Try *"help tabulate twoway"*

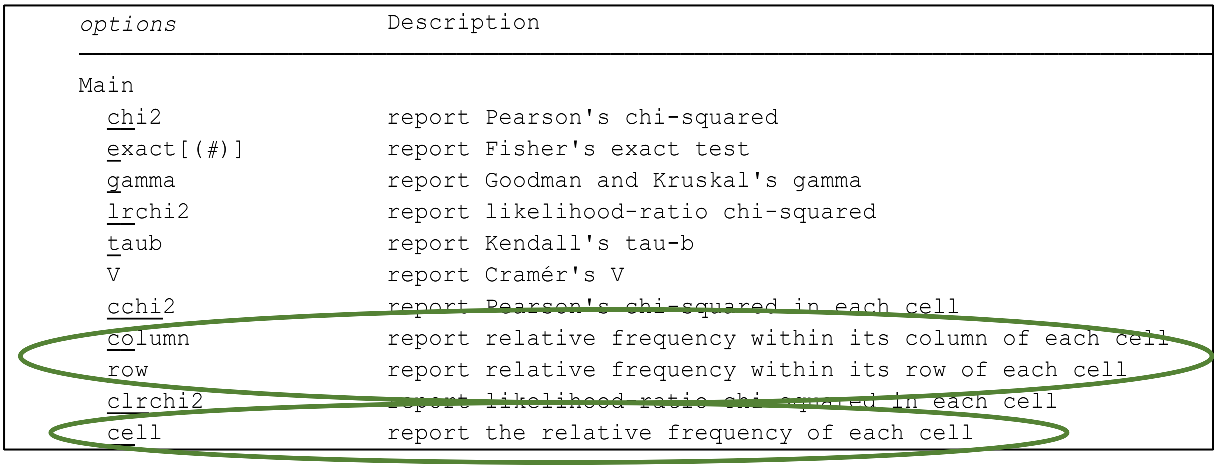
Figure 2.19: Stata help window for *tabulate twoway* – syntax



```
[R] tabulate twoway —— Two-way table of frequencies
                        (View complete PDF manual entry)


Syntax

    Two-way table

        tabulate varname1 varname2 [if] [in] [weight] [, options]
```

i. Shows the syntax for how to write and construct the command

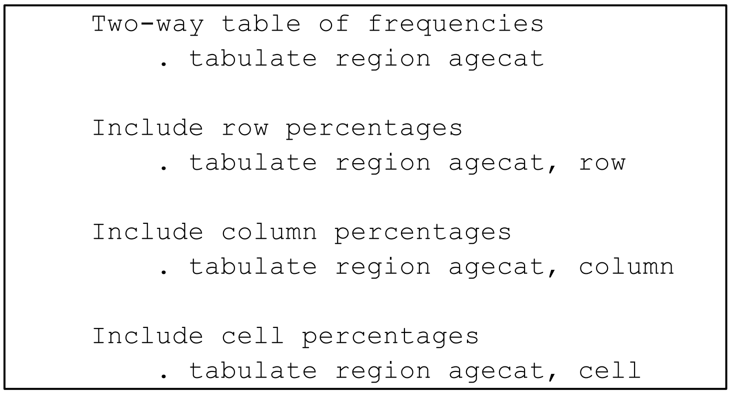
Figure 2.20: Stata help window for *tabulate twoway* – options



```
    options                Description
    ------------------------------------------------------------------
    Main
      chi2                 report Pearson's chi-squared
      exact[(#)]           report Fisher's exact test
      gamma                report Goodman and Kruskal's gamma
      lrchi2               report likelihood-ratio chi-squared
      taub                 report Kendall's tau-b
      V                    report Cramér's V
      cchi2                report Pearson's chi-squared in each cell
      column               report relative frequency within its column of each cell
      row                  report relative frequency within its row of each cell
      clrchi2              report likelihood-ratio chi-squared in each cell
      cell                 report the relative frequency of each cell
```

ii. Can see different options and what they do: “, cell” “, row” etc.

Figure 2.21: Stata help window for *tabulate twoway* – examples

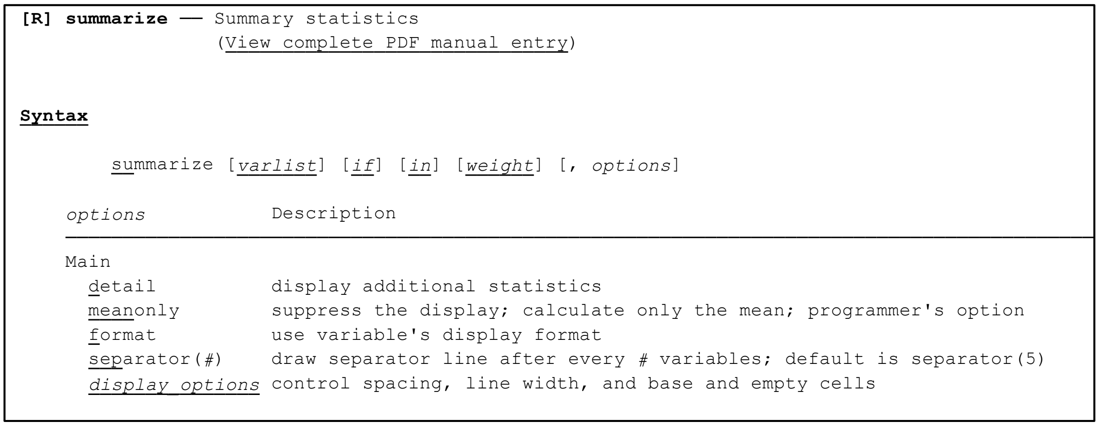


```
    Two-way table of frequencies
        . tabulate region agecat

    Include row percentages
        . tabulate region agecat, row

    Include column percentages
        . tabulate region agecat, column

    Include cell percentages
        . tabulate region agecat, cell
```

iii. Can see different examples: “tabulate region agecat, cell”

b. Try *“help summarize”*

Figure 2.22: Stata help window for *summarize* – syntax and options



```
[R] summarize -- Summary statistics
                (View complete PDF manual entry)


Syntax

        summarize [varlist] [if] [in] [weight] [, options]

    options            Description
    -----------------------------------------------------------------------
    Main
      detail           display additional statistics
      meanonly         suppress the display; calculate only the mean; programmer's option
      format           use variable's display format
      separator(#)     draw separator line after every # variables; default is separator(5)
      display_options  control spacing, line width, and base and empty cells
```

i. Can see different options and what they do: ", detail"

ii. Can see different examples: "sum mpg weight"

5. Collapse and Merge Datasets

a. What if we want to look at some of this information at the household level?

  i. Maybe we want to know the age of the youngest person in each household, the number of people in each household, the percent of women and children in each household, and the percent of households with a female household head.

  ii. We can do this with the "collapse" command which makes a dataset of the summary statistics that you specify.

b. Collapsing

  i. Examples of summary statistics that you can specify are: mean, median, sum, count, max, and min.

  ii. Which summary statistic and variable would we use to find the number of people in each household? *(count) pid*

  iii. Which summary statistic and variable would we use to find the youngest person in each household? *(min) age*

  iv. Which summary statistic and variables would we use to find the percent of children and the percent of women in each household? *(mean) child woman – "mean" works for these two because they are dummy variables (values of 0="No" and 1="Yes")*

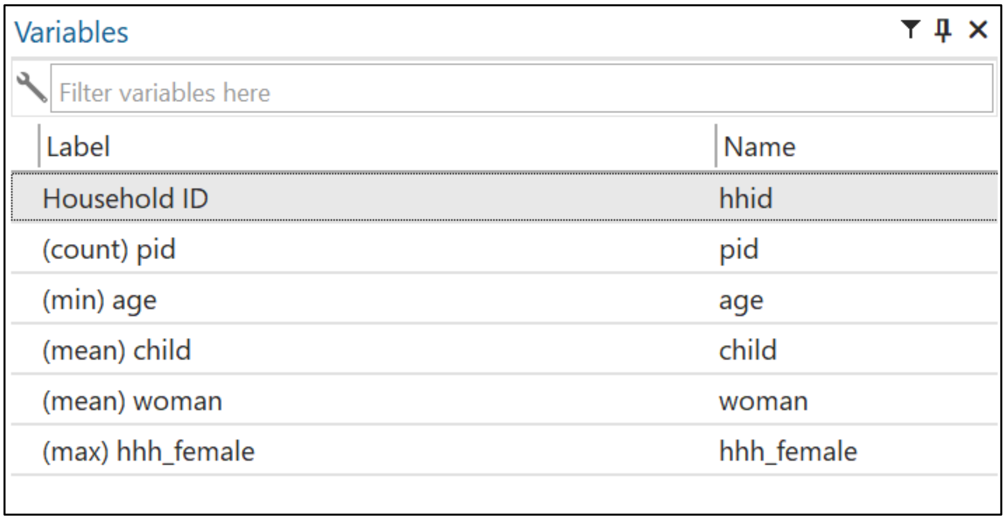
  v. Which summary statistic and variable would we use to find the percent of households with a female household head? *(max) hhh_female* – "max" works because right now our hhh_female variable identifies the women who are household heads. If we used "mean", it would tell us the average number of female household heads *in each household.* Rather, we want to know which household *have a female household head.* What other summary statistic would work? *sum (adding up all of the values for each household)*

  vi. Practice: *"collapse (count) a_pid (min) age (mean) child woman (max) hhh_female, by(hhid)"*

  vii. How many observations are there now in the dataset? *"count" – 14,580*

  viii. Now let's look at what happened to Window 3 (Variable Window) after the collapse

Figure 2.23: Variable window after collapsing pid, age, child, woman, and hhh_female by household



Label	Name
Household ID	hhid
(count) pid	pid
(min) age	age
(mean) child	child
(mean) woman	woman
(max) hhh_female	hhh_female

a) Variable names remained the same

b) Variable labels now show only the summary statistic and the name of the variable

ix. Practice: Let's create more meaningful variable names and labels

a) What does the variable *pid* tell us now?

- The count of people in the household = the household size
- *"ren pid hhsize"*
- *"lab var hhsize "Household size""*

b) What does the variable *age* tell us now?

- The minimum age out of the ages of each person in the HH = the age of the youngest person in the household
- *"ren age age_youngest"*
- *"lab var age_youngest "Age of the youngest household member""*

c) What does the variable *child* tell us now?

- The average/mean of the dummy variable "child" = the percent of children in each household
- *"ren child perc_children"*
- "*lab var perc_children "Percent of children 15 years and younger in the household""*

d) What does the variable *woman* tell us now?

- The average/mean of the dummy variable "woman" = the percent of women in each household
- *"ren woman perc_women"*
- "*lab var perc_ women "Percent of women in the household""*

e) What does the variable hhh_female tell us now?

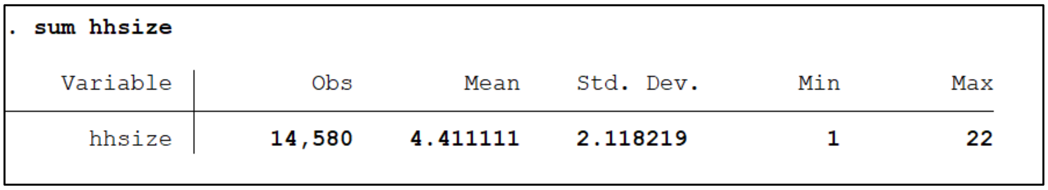
- Whether or not the household head is a female. So the variable name is okay.
- *"lab var hhh_female "Household head is female""*

- Is this a dummy variable? *Yes* – so we need to label the values again
- Let's make more intuitive value labels for this important variable: *"lab def hhhfemale 0 "Male headed household" 1 "Female headed household"* defines the label **then** *"lab val hhh_female hhhfemale"* applies the label to the variable.

x. Question: What is the average household size?

   a) *"sum hhsize"*

Figure 2.24: Stata command and output for "sum hhsize"



```
. sum hhsize

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      hhsize |     14,580    4.411111    2.118219          1         22
```

   b) *4.41 people per household*

xi. Question: What is the average age of the youngest household member?

   a) *"sum age_youngest"*

   b) *10.73 years old*

xii. Question: What is the average prevalence of *children* in the households?

   a) *"sum perc_children"*

   b) *37% children*

xiii. Question: What is the average prevalence of *women* in the households?

   a) *"sum perc_women"*

   b) *52% women*

xiv. Question: What percent of households are female headed?

   a) *"sum hhh_female"*

   b) *26% of households are female headed*

c. Merge Datasets

i. What if we want to know if the percent of children in a household is associated with a household's poverty status?

   a) We have percent of children in this dataset, but poverty status is in the dataset that we worked with in Lesson 1

   b) We can merge the two datasets together – and now they both have the same number of observations/households (14,580)

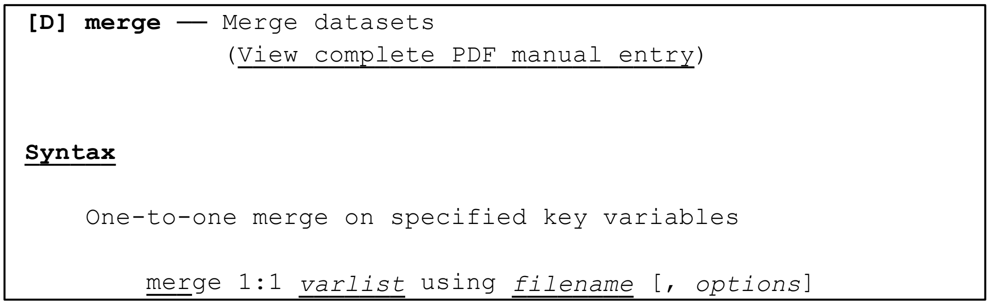
ii. Merging datasets with the same number of observations (the observations across the datasets represent the same levels of data: e.g. both datasets are at the household level)

Figure 2.25: Example of a 1:1 merge, shown with only one household identifier (hhid)

cs_S1_S2_S3_S4_S6A_S6E_Person (collapsed; n=14,580)			**cs_S0_S5_Household (n=14,580)**			**merged data (n=14,580)**		
hhid	perc_children		hhid	poverty		hhid	perc_children	poverty
214528	.6		214528	Non-poor		214528	.6	Non-poor

   a) Practice: *"help merge"*

Figure 2.26: Stata help window for *merge*



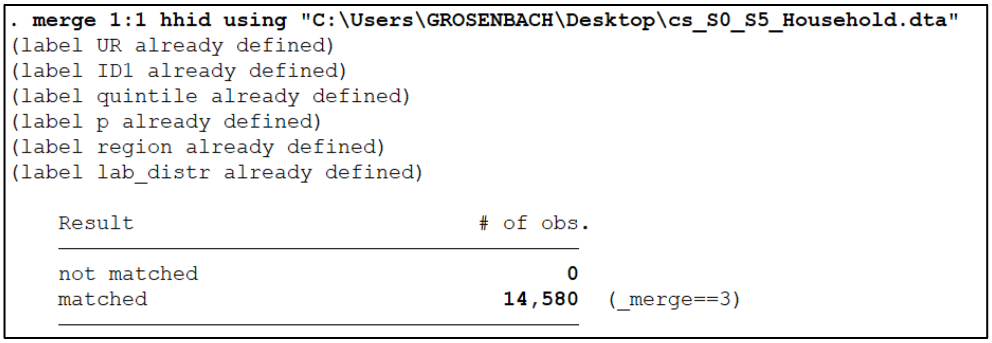
```
[D] merge —— Merge datasets
             (View complete PDF manual entry)


Syntax

    One-to-one merge on specified key variables

        merge 1:1 varlist using filename [, options]
```

- We are merging one-to-one because we now have a dataset with 14,580 unique households and we are merging it to another dataset with 14,580 unique households

b) What variable will we merge on? (Meaning, which variable should Stata use to match the two datasets to each other?) *hhid (Household ID)*

c) Practice: *“merge 1:1 hhid using “"F:\cs_S0_S5_Household.dta""*

Figure 2.27: Stata command and output for “merge 1:1 hhid using …”



```
. merge 1:1 hhid using "C:\Users\GROSENBACH\Desktop\cs_S0_S5_Household.dta"
(label UR already defined)
(label ID1 already defined)
(label quintile already defined)
(label p already defined)
(label region already defined)
(label lab_distr already defined)

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            14,580  (_merge==3)
    -----------------------------------------
```

d) The output after the merge tells us how many observations were and were not matched. How many were matched? *14,580 (all of them!)*

e) The “merge” command automatically creates a new variable called “_merge”

- _merge equals 1 in observations that were not matched from the *master* data file (the one that you started with). For example, if you had 14,581 observations in the collapsed household roster, and merged it to 14,580 observations in the household characteristics file, then the one extra observation would be _merge==1
- _merge equals 2 in observations that were not matched from the *using* data file (the one listed in the “merge” code). For example, if you had 14,580 observations in the collapse household roster, and merged it to 14,581 observations in the household characteristics file, then the one extra observation would be _merge==2
- _merge equals 3 in matched observations. Because our two data files had exactly the same households, each observation is _merge==3

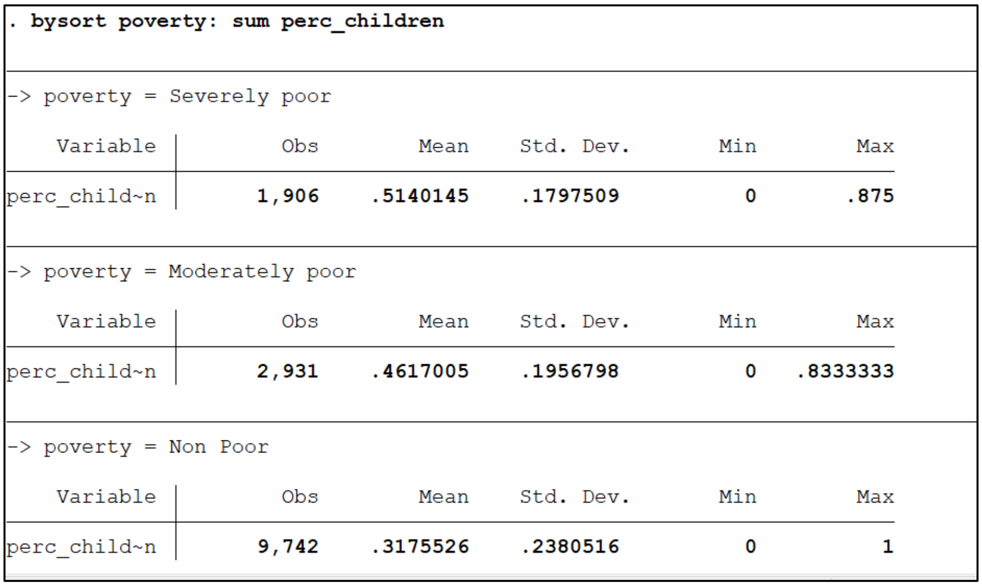
f) If you wanted to merge multiple datasets together, you will have to drop this new _merge variable, otherwise another merge will not work because Stata will tell you that the variable _merge is already defined.

- Practice: *“drop _merge”*

g) Challenge: On average, do households who are non-poor have more children?

- Option 1: two sum…if codes
  - *“sum perc_children if poverty==1 | poverty==2”* – on average in households that are below the poverty line (either moderately or severely poor), 48% of the household members are aged 15 years or younger
  - *“sum perc_children if poverty==3”* – on average in households that are non-poor, 32% of the household members are aged 15 years or younger
  - Descriptives suggest that households with below the poverty line have a higher percentage of children than households above the poverty line.
- Option 2: bysort. “bysort” repeats a Stata command on a subset of the data. We can repeat the “summarize perc_children” command, on the different values of poverty.
  - Code: ***bys**ort [varname1]: stata_command [varname2]*
  - *“**bys**ort poverty: sum perc_children”*

Figure 2.28: Stata command and output for “bysort poverty: sum perc_children”



```
. bysort poverty: sum perc_children

-> poverty = Severely poor

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
perc_child~n |      1,906    .5140145    .1797509          0       .875

-> poverty = Moderately poor

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
perc_child~n |      2,931    .4617005    .1956798          0   .8333333

-> poverty = Non Poor

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
perc_child~n |      9,742    .3175526    .2380516          0          1
```

  - Shows the same results as Option 1 except it keeps severely poor and moderately poor separate, and it only requires one line of code to run
  - Which households have the highest percent of children? *The households in severe poverty*

iii. What if we want to know how many *people* in our dataset are non-poor?

a) We would want to look at the poverty variable at the *person-level*

b) We can merge the housing characteristics dataset to the household roster dataset, maintaining the dataset at the *person-level (64,314 people)*

iv. Merging datasets with different numbers of observations (the observations across the datasets represent different levels of data: e.g. one dataset is household level and the other dataset is person level)

Figure 2.29: Example of a 1:m merge, shown with only one household identifier (hhid)

cs_S0_S5_Household (n=14,580)		cs_S1_S2_S3_S4_S6A_S6E_Person (n=64,314)			merged data (n=64,314)			
hhid	poverty	hhid	pid	s1q1	hhid	pid	s1q1	poverty
214528	Non-poor	214528	1	Female	214528	1	Female	Non-poor
		214528	2	Female	214528	2	Female	Non-poor
		214528	3	Male	214528	3	Male	Non-poor
		214528	4	Female	214528	4	Female	Non-poor
		214528	5	Female	214528	5	Female	Non-poor

a) Let's now start over and open the household characteristics dataset (cs_S0_S5_Household.dta).

- Close out of Stata, and double-click on this datafile to open it
- What is the level of observations in this dataset? *Household-level*

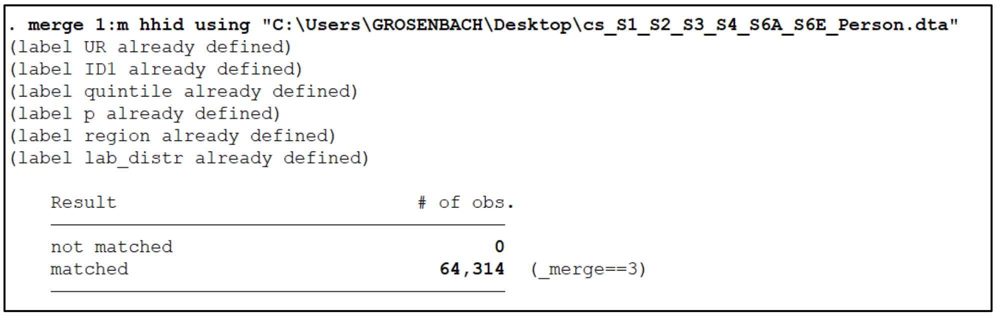
b) We are going to merge it with the household roster dataset we were just working with (cs_S1_S2_S3_S4_S6A_S6E_Person). What was the level of observations in that dataset? *Person-level*

c) So what part of the merge command do you think we have to change for this? *Change 1:1 to 1:m. We are matching 1 household in the cs_S0_S5_Household data to many (m) household observations in the cs_S1_S2_S3_S4_S6A_S6E_Person data set.*

d) Practice: *"merge 1:m hhid using "F:\ cs_S1_S2_S3_S4_S6A_S6E_Person.dta""*

- Note: The file path might be different on your computer depending on where the data is saved and how your computer is organized.

Figure 2.30: Stata command and output for "merge 1:m hhid using …"



```
. merge 1:m hhid using "C:\Users\GROSENBACH\Desktop\cs_S1_S2_S3_S4_S6A_S6E_Person.dta"
(label UR already defined)
(label ID1 already defined)
(label quintile already defined)
(label p already defined)
(label region already defined)
(label lab_distr already defined)

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            64,314  (_merge==3)
    -----------------------------------------
```

- Did all of the observations match? *Yes, all 64,314 matched*
- So what are the values of the _merge variable? *All are _merge==3*
- Question: How can we double check the values of the _merge variable? *"tab _merge"*

e) Question: How many *people* in our dataset are non-poor?

- *"tab poverty"* – now our data is at the person-level instead of the household-level (like in Lesson 1), so this code will now tell us the number/percent of *people* instead of the number/percent of *households* who are non-poor

- *39,378 people (61.23% of people in the sample are non-poor)*

f) <u>Challenge:</u> How many *women* in our dataset are non-poor? (Two ways):

- Tab…if
    - First we need to remember how women is coded in the gender variable. How can we check this? *"codebook s1q1"; 2 = female*
    - Now we can run the tab…if. What would it look like? *"tab poverty if s1q1==2"*
- Twoway tabulation:
    - *"tab poverty s1q1, col"*
    - *"tab s1q1 poverty, row"*
- *20,326 women are non-poor (60.61% of women in the sample are non-poor)*

6. Saving, Exporting, and Importing Data

a. Saving Data

i. Since we've modified this dataset by merging the household characteristics to the person roster, let's save the data so that it's ready to use in the future (and we don't have to run all of the same code again).

ii. The code to save the data is easy: "save "*filepath\filename.*dta", replace"

iii. It's good to put the 'replace' option, in case you modify and rerun the code later, and want to keep that new modified data file instead of the outdated one.

iv. *"save "F:\eicv_merged_data.dta", replace"*

v. This code will save the merged data file as a "dta" file. "dta" files are the names of the data files used in Stata, so you will only be able to open this file using the Stata software.

b. Exporting Data to Excel

i. Exporting and saving data in an Excel format is also very useful in case you want to share the data with someone who does not have Stata.

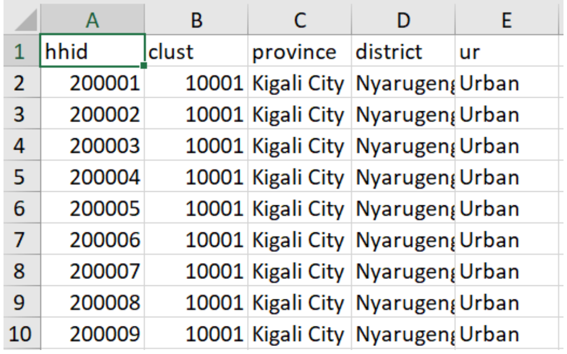
ii. The code to export data to Excel is: "export excel using "*filepath\filename*.xlsx", firstrow(variables) replace"

iii. The option 'firstrow(variables)' tells Stata to save the variable names in the first row of the Excel sheet. Unfortunately there is no option to save the variable labels, so these will be lost when saved.

iv. *"export excel using "F:\eicv_merged_data.xlsx", firstrow(variables) replace"*

v. Let's open the saved Excel file and see what it looks like.

Figure 2.31: Excerpt of exported Excel file



	A	B	C	D	E
1	hhid	clust	province	district	ur
2	200001	10001	Kigali City	Nyarugeng	Urban
3	200002	10001	Kigali City	Nyarugeng	Urban
4	200003	10001	Kigali City	Nyarugeng	Urban
5	200004	10001	Kigali City	Nyarugeng	Urban
6	200005	10001	Kigali City	Nyarugeng	Urban
7	200006	10001	Kigali City	Nyarugeng	Urban
8	200007	10001	Kigali City	Nyarugeng	Urban
9	200008	10001	Kigali City	Nyarugeng	Urban
10	200009	10001	Kigali City	Nyarugeng	Urban

vi. Because Excel does not have all of the same functionalities of Stata, the variable labels and value labels are not able to be saved. Notice that the ‘ur’ data is now saying “Urban” and “Rural” instead of using numbers to represent each category.

vii. This is a big limitation of using Excel to share the data. In order to inform other users of how to use the data when sharing via Excel, it’s important to always include a data dictionary (a separate Word or Excel document) which links the variable names to the variable labels (so users will know what ‘ur’ means, for example).

viii. An example data dictionary for the variables in Figure 2.31 could look like:

Figure 2.32: Example data dictionary using the Stata variable labels

Variable	Variable label	Variable type
hhid	Household id	Numeric
clust	Cluster	Numeric
province	Province	Categorical
district	District	Categorical
ur	Urban/Rural	Categorical

c. Importing Data from Excel

i. Now, let’s import this same Excel file back into Stata.

ii. The code to export data to Excel is: “import excel “*filepath\filename*.xlsx”, sheet(“Sheet1”) firstrow clear”

iii. You have to specify which sheet of the Excel file you want to import, using the ‘sheet()’ option.

iv. To tell Stata that you want to import the first row of the Excel file as the variable names, you use the ‘firstrow’ option.

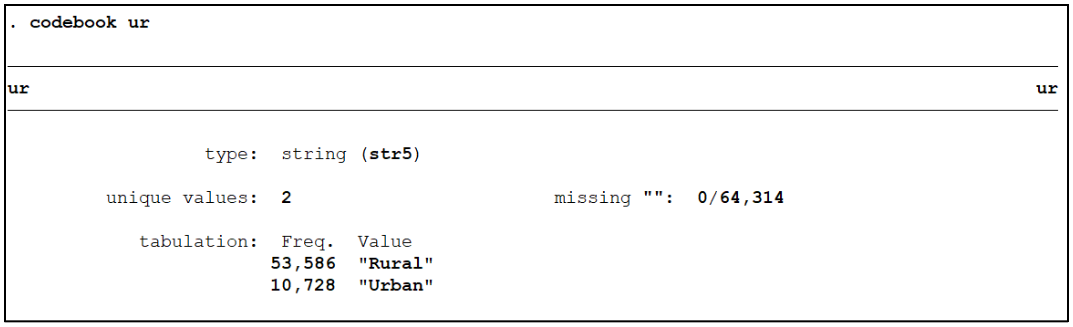
v. Use the ‘clear’ option to tell Stata that it’s okay to replace any data that was previously loaded.

vi. Does the imported data have variable labels? *No – the variable labels are identical to the variable names.*

vii. How can we check if the data have value labels? *Use codebook for one of the categorical variables.*

viii. Let’s type *“codebook ur*” and see what the variable looks like now. (Remember, this used to be a categorical variable, where 1=Urban and 2=Rural)

Figure 2.33: Stata command and output for "codebook ur" after importing data from Excel



```
. codebook ur

-------------------------------------------------------------------------------
ur                                                                           ur
-------------------------------------------------------------------------------

                  type:  string (str5)

         unique values:  2                        missing "":  0/64,314

            tabulation:  Freq.  Value
                        53,586  "Rural"
                        10,728  "Urban"
```

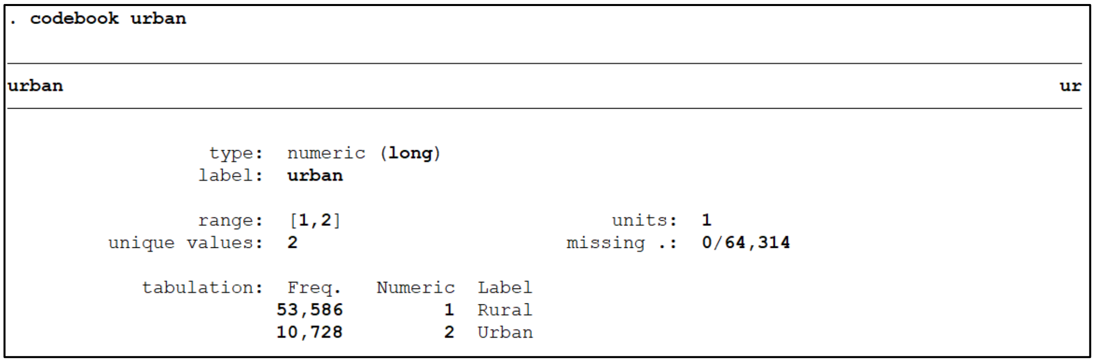
ix. Now, this variable is a string (the data are entered as text), rather than a categorical variable.

x. To create a categorical variable from *ur* , we can use the code "*encode ur, generate(urban)"*

xi. When using *encode*, we have to use the option 'generate(*new variable name*)' to tell Stata to create a new categorical variable based on the existing string variable.

xii. Now, when we use "*codebook urban"*, we see that it is a categorical variable again (1=Rural and 2=Urban). Notice that this is different from how *ur* was coded before (in which 1=Urban). This is because when Stata runs the 'encode' command, it codes them in alphabetical order.

Figure 2.34: Stata command and output for "codebook urban"

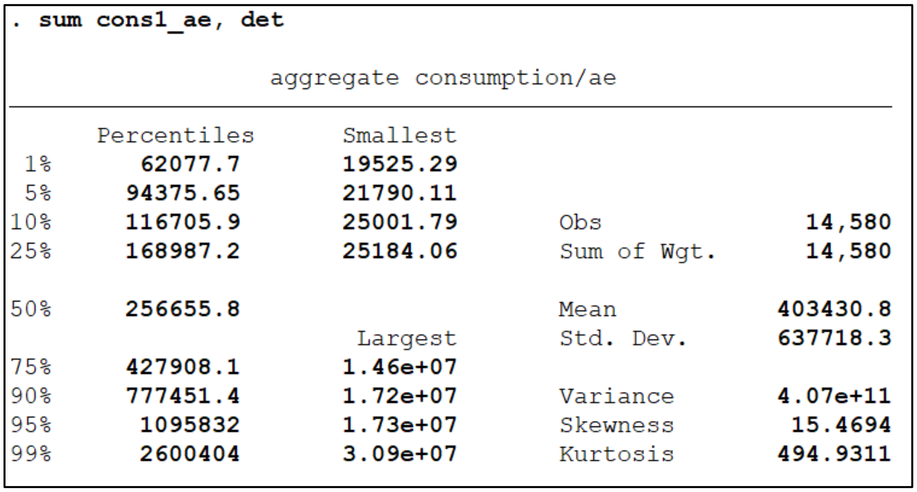


```
. codebook urban

-------------------------------------------------------------------------------
urban                                                                        ur
-------------------------------------------------------------------------------

                  type:  numeric (long)
                 label:  urban

                 range:  [1,2]                        units:  1
         unique values:  2                        missing .:  0/64,314

            tabulation:  Freq.   Numeric  Label
                        53,586         1  Rural
                        10,728         2  Urban
```

Lesson 3 – Analyzing Data

1. **Review of Lessons 1 and 2 – Describing and transforming the new dataset ("EICV5_Poverty_file" – compilation of key information from many modules)**
    a. The datafile we are working with in Lesson 3 is NISR's 'poverty file', in which they calculate monthly household expenditures per adult equivalent, and compare that to the national poverty lines to determine a household's poverty status.
    b. Question: How can we see how many observations we have? *"count"*
        i. How many observations are in this dataset? *14,580*
        ii. What level is this data? (What does each observation represent?) *Household-level (each observation is one unique household)*
    c. Let's quickly take a look at some of the new variables in this dataset:
        i. Question: What is the average household size? *"sum member" – 4.41 individuals per household*
        ii. Question: What is the average household size, *using adult equivalents*? *"sum ae" – 3.9998 adult equivalents*
        iii. The 'adult equivalent' measure assumes that different household members consume different amounts. Adult men are given a value of 1, adult women are given a value of 0.7, and children are given a value of 0.5 (assuming children consume about half of what an adult man consumes). This measure provides a more accurate way to compare consumption across households. For example, 100,000 RWF in a household of 3 adult men is not comparable to 100,000 RWF in a household of 1 adult woman and 2 children. The household size of each household (*member* variable) would be 3. The adult equivalent value (*ae* variable) would also be 3 for the 3 men (1 + 1 + 1 = 3), but the adult equivalent value for the household with 1 woman and 2 children would be 1.7 (0.7 + 0.5 + 0.5 = 1.7). And so, consumption *per capita* (aka per person) would be 33,333 RWF for both households. However consumption *per adult equivalent* would be 33,333 RWF (100,000 RWF divided by 3 ae) for the first household and 58,823 RWF (100,000 RWF divided by 1.7 ae) for the second household. Therefore, by the consumption per adult equivalent measure, the second household is better off because they have the same consumption level, but the members do not need to consume as much as the first household.
        iv. Question: Will the adult equivalent variable *always* have a smaller average than the household size variable? *Yes, because when calculating adult equivalent, many household members may be given values of less than 1*
        v. Monthly consumption per adult equivalent is going to be our key outcome variable today. This variable was constructed by NISR using Section 8 in the survey – "Household Expenditure and Consumption"
            1. Question: How can you check the minimum, maximum, median, 99th percentile and average values for this variable?
                a. *"sum cons1_ae, detail"*

Figure 3.1: Stata command and output for “hist cons1_ae, det”



```
. sum cons1_ae, det

                    aggregate consumption/ae
-------------------------------------------------------------
      Percentiles      Smallest
 1%      62077.7       19525.29
 5%     94375.65       21790.11
10%     116705.9       25001.79       Obs              14,580
25%     168987.2       25184.06       Sum of Wgt.      14,580

50%     256655.8                      Mean           403430.8
                        Largest       Std. Dev.      637718.3
75%     427908.1       1.46e+07
90%     777451.4       1.72e+07       Variance       4.07e+11
95%      1095832       1.73e+07       Skewness        15.4694
99%      2600404       3.09e+07       Kurtosis       494.9311
```

b. *The minimum is 19,525 RWF per person per month*

c. *The maximum is 30,882,483 RWF per adult equivalent per month*

d. *The median is 256,656 RWF per adult equivalent per month*

e. *The 99th percentile is 2,600,404 RWF per adult equivalent per month*

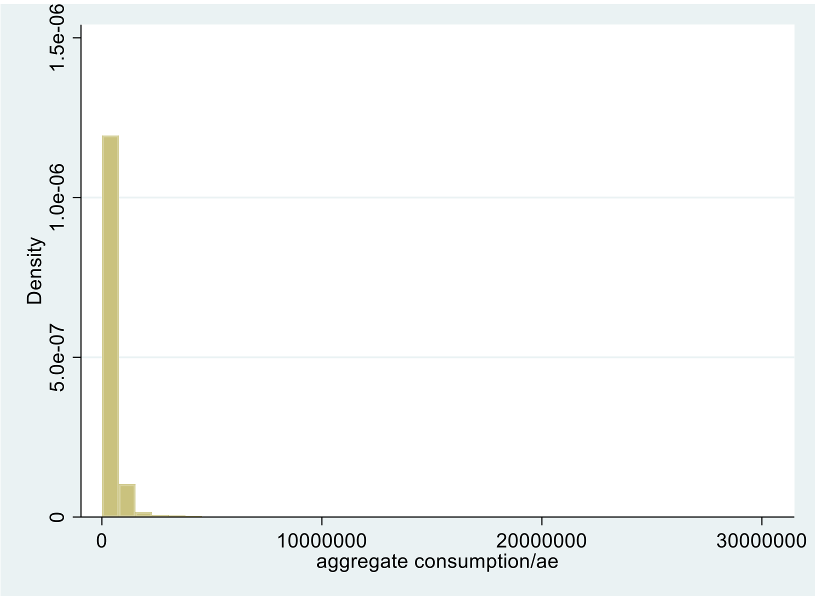
f. *The mean is 403,431 RWF per adult equivalent per month*

g. *The mean is much larger than the median (because the maximum is very high)*

2. Question: How can we look at a figure showing the distribution of this variable?

   a. “*hist cons1_ae*”

   Figure 3.2: Stata histogram output for “hist cons1_ae”



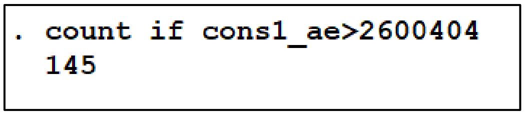
   b. What can we learn from this histogram? *There are a very few very large values which are skewing the distribution. If all values fell within a more reasonable range, we would be able to see more of* *the shape of the distribution. Instead, now it looks like most people consume almost 0 (which we know isn’t true from the last code)*

3. Continuous data can be messy and contain notable **outliers**. An outlier is an observation that is very different from all other observations. For example, we know that 99% of households have a consumption per adult equivalent under 2,600,404, but the maximum value *much* higher at 30,882,482 (an outlier). There are 3 main reasons an outlier could exist in the data:
    a. Input errors by the data collectors (e.g. entered than an egg costs 1,000 RWF instead of 100 by mistake)
    b. Confusion about the questions (e.g. a respondent double counted some of the food his household consumed, thinking that was what the data collector was asking)
    c. Best guesses (e.g. a respondent does not know how much his eggs would sell for at the market since he consumed them all, so he guesses 1,000 RWF each)
4. Usually we will change extreme outliers to either a missing value or a more reasonable value based on the distribution. Outliers may or may not be incorrect or need to be changed – each researcher has a different preference for how to deal with outliers, which usually depends on the question they are trying to answer and the variable of interest. Let’s learn how to change the outliers in cons1_ae to something more reasonable.

d. Everyone has a different preference for how to handle outliers. For now, let’s say that all values greater than the 99th percentile (2,600,404 – we know from “sum cons1_ae, det”) should be changed to the median (256,655.8). The 99th percentile tells us that 99% of all observations fall below, or are less than, 2,600,404, so this is a good cutoff to use to decide what constitutes being an outlier. (Another popular ‘cutoff’ is 3 standard deviations away from the mean).

   i. Question: First, let’s see how many observations are greater than the 99th percentile (2,600,404). How might we check this? (HINT: We can combine “if” with one of the commands that we know). *“count if cons1_ae>2,600,404” – 145 observations (or 1% of the number of observations)*

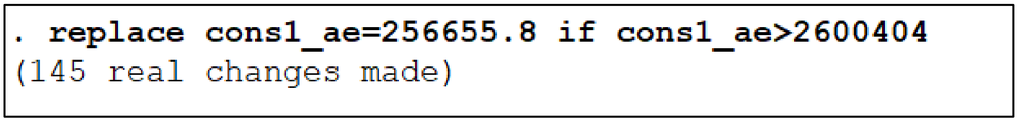
   Figure 3.3: Stata command and output for “count if cons1_ae>2,600,404”

```
. count if cons1_ae>2600404
  145
```

   ii. Question: How might we change these 145 observations to the median (256,655.8)? *“replace cons1_ae=256655.8 if cons1_ae>2600404”*

   Figure 3.4: Stata command and output for “replace cons1_ae=256655.8 if cons1_ae>2600404”

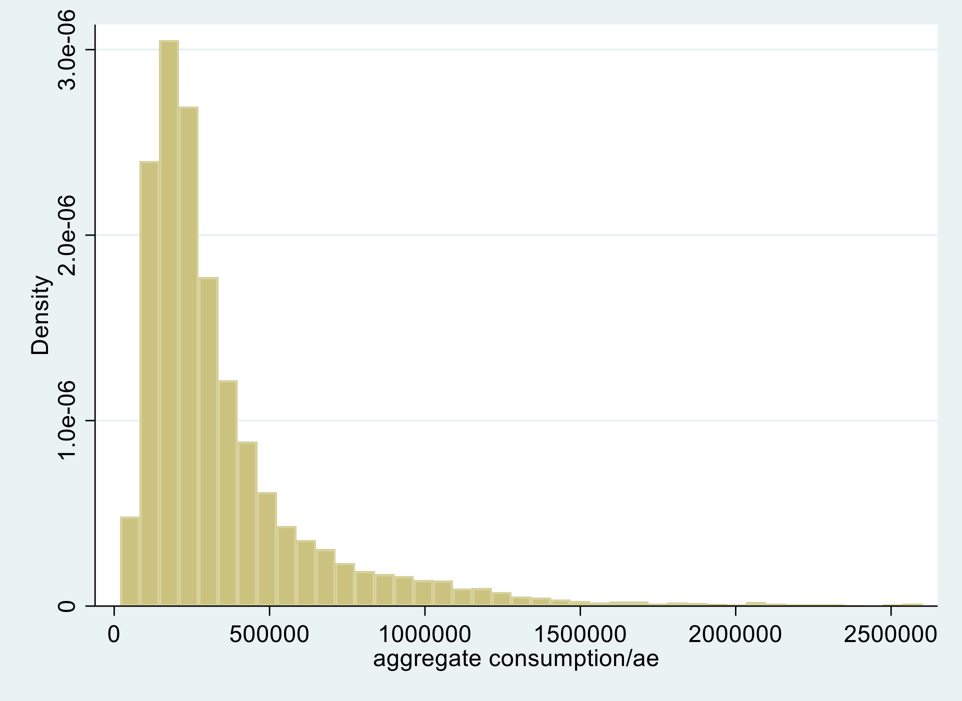
   

```
. replace cons1_ae=256655.8 if cons1_ae>2600404
(145 real changes made)
```

   iii. Stata tells us how many observations it changed with our command in the output. How many were changed? *145 – the same amount that the “count…if” command told us*

iv. <u>Question:</u> Now how can we look at the distribution of our newly cleaned *landholdings* variable? *“hist cons1_ae”*

Figure 3.5: Stata histogram output for “hist cons1_ae”, after cleaning outliers



1. Now the positive skew is much less
2. The figure only goes up to 2,500,000 RWF instead of more than 30 million like the first one!

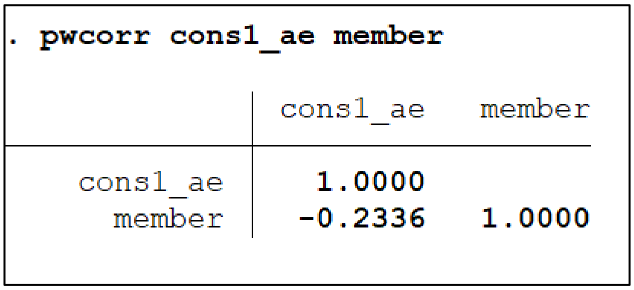
2. Correlations

a. Let’s see how correlated our variable of interest (cons1_ae) is with household size:

i. <u>Code:</u> *pwcorr [varname] [varname]*

ii. <u>Practice:</u> *“pwcorr cons1_ae member”*

Figure 3.6: Stata command and output for “pwcorr cons1_ae member”



```
. pwcorr cons1_ae member

             |  cons1_ae   member
-------------+------------------
    cons1_ae |    1.0000 
      member |   -0.2336   1.0000
```

iii. This only tells us the correlation coefficient between the two variables.

iv. Remember that a correlation coefficient equal to 0 is the weakest linear relationship, and a correlation coefficient equal to 1 or -1 is the strongest linear relationship.

1. What is the correlation coefficient between cons1_ae and member? *-0.2336*
2. Is this strong or weak? *Weak*

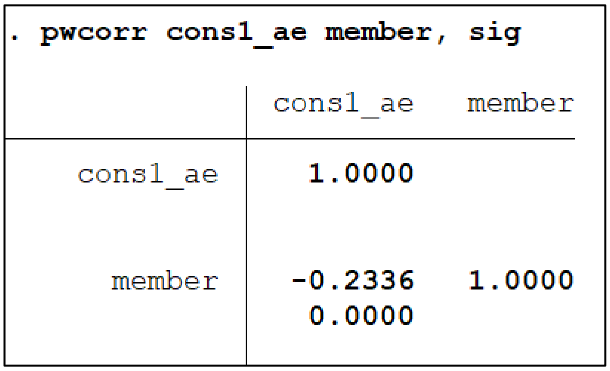
v. Also remember that a positive correlation coefficient means that as one variable increases, the other increases; and a negative correlation coefficient means that as one variable increases, the other decreases.

1. Is the correlation coefficient between cons1_ae and member positive or negative? *Negative*
2. Even though the correlation coefficient is very weak, how can you interpret this/explain this in common terms? *Households with more household members are associated with consuming less per adult equivalent per month.*

b. We can also look at how significant the correlation coefficient is, by adding the option “sig” (short for significance) to our code

i. Practice: *“pwcorr cons1_ae member, sig”*

Figure 3.7: Stata command and output for “pwcorr cons1_ae member, sig”



```
. pwcorr cons1_ae member, sig

             | cons1_ae   member
-------------+------------------
    cons1_ae |   1.0000
             |
             |
      member |  -0.2336   1.0000
             |   0.0000
             |
```

ii. This new number below the correlation coefficient is the p-value:

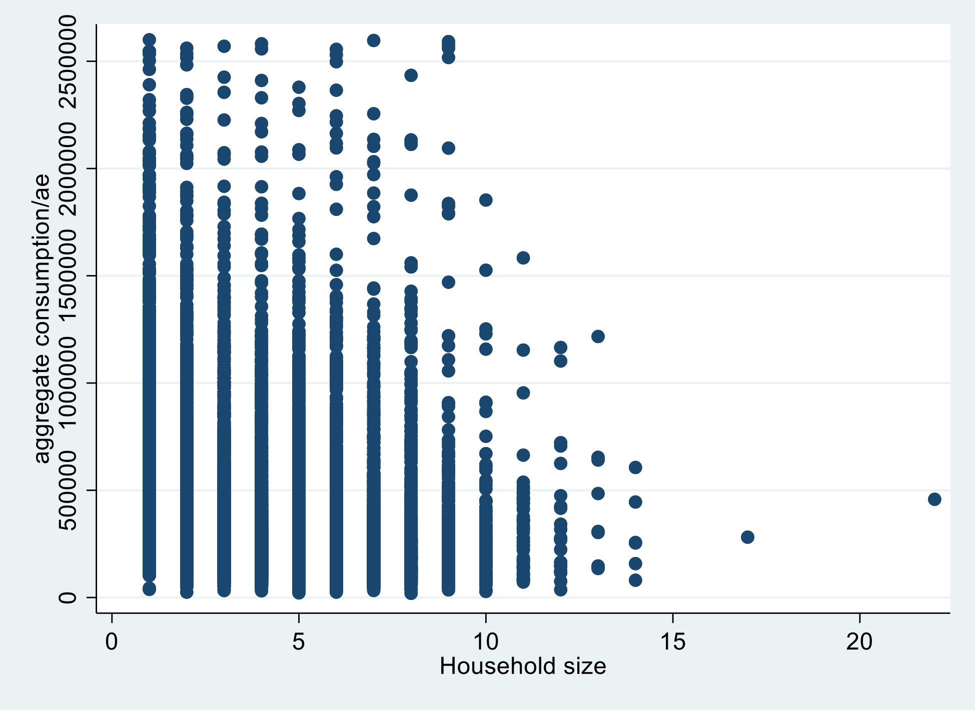
1. The p-value tells us the percent with which we are confident that the two variables are associated. You find this percent by subtracting the p-value by 1 (e.g. 1-0.0000 equals about 1 – so we can say with nearly 100% confidence that this correlation is significant)
2. The usual p-value cut-offs for stating that something is significant are:
    a. 90% confidence (p-value<0.1)
    b. 95% confidence (p-value<0.05)
    c. 99% confidence (p-value<0.01)
3. What is the p-value for this correlation? *0.0000*
4. Note that the p-value is not actually equal to 0 (it’s nearly impossible for a p-value to be equal to 0). Instead, the p-value is so small that when rounded to the ten-thousandth decimal place, it rounds to 0. So we can say that p<0.01).
5. Is this significant? *Yes*

c. Another way to look at this relationship between two continuous variables is to create a scatter plot

i. Code: *scatter [varname] [varname]*

ii. Practice: *“scatter cons1_ae member”*

Figure 3.8: Stata scatterplot output for "scatter cons1_ae member"



iii. It looks like there may be a ***negative*** relationship (as the values for household size increase the values for consumption per adult equivalent decrease)

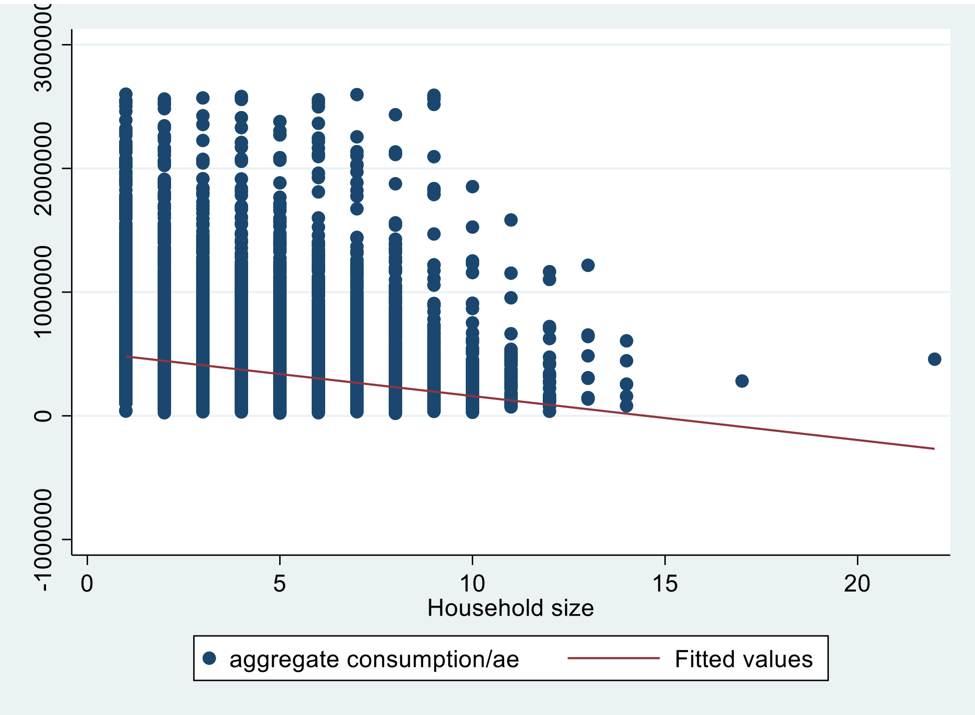
iv. We can also create a 'line of best fit' to see how negative and strong the relationship is

d. Let's add a line of best fit to this figure, to better see the trend and how strongly associated the variables are

i. Code: *twoway (scatter [varname1] [varname2]) (lfit [varname1] [varname2])*

ii. Practice: *"twoway (scatter cons1_ae member) (lfit cons1_ae member)"*

Figure 3.9: Stata scatterplot output for "twoway (scatter cons1_ae member) (lfit cons1_ae member)



iii. You can see the very slight negative slope on the line of best fit, which is consistent with our correlation results

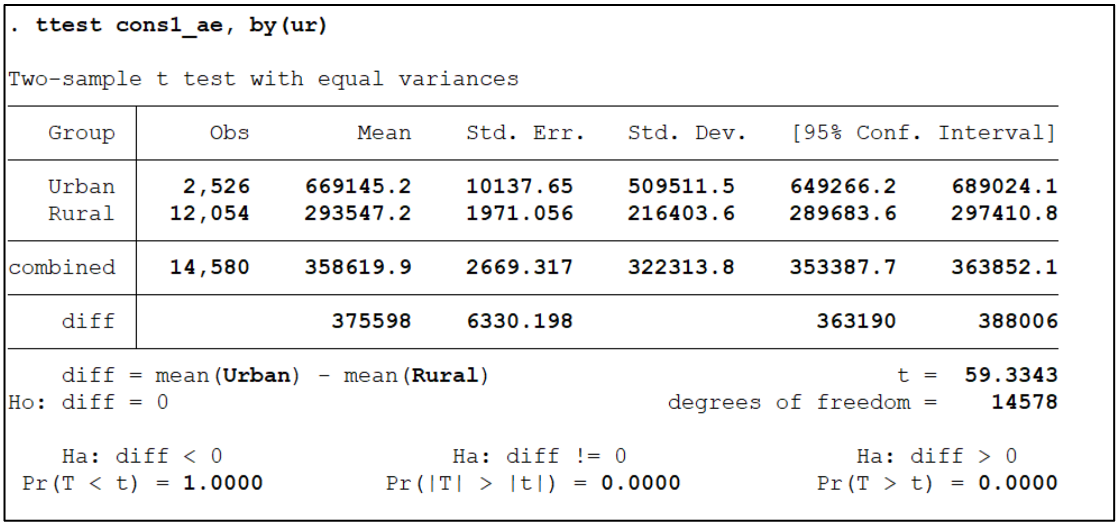
3. **T-Tests**

   a. The most frequently used t-tests are two-sample t-tests: these tell us whether one variable (e.g. cons1_ae) is significantly different between two groups in the data (e.g. whether a household lives in a rural or urban area)

   b. <u>Code:</u> *ttest [varname], by([groupvar])*

   c. <u>Practice:</u> *"ttest cons1_ae, by(ur)"* – output tells us:

Figure 3.10: Stata command and output for "ttest cons1_ae, by(ur)"



```
. ttest cons1_ae, by(ur)

Two-sample t test with equal variances

   Group      Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]

   Urban     2,526    669145.2    10137.65    509511.5    649266.2    689024.1
   Rural    12,054    293547.2    1971.056    216403.6    289683.6    297410.8

combined    14,580    358619.9    2669.317    322313.8    353387.7    363852.1

    diff                375598    6330.198                  363190      388006

    diff = mean(Urban) - mean(Rural)                              t =  59.3343
Ho: diff = 0                                     degrees of freedom =    14578

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

i. Number of cons1_ae observations in each of the two groups. How many observations live in urban areas? *2,526 households live in urban areas*

ii. Average monthly consumption per adult equivalent of each group. Which group has a higher consumption? *Households in urban areas have a higher monthly consumption per adult equivalent*

iii. Standard error, standard deviation, and 95% confidence interval of the consumption per adult equivalent of each group. Do the two confidence intervals overlap? *No, the confidence intervals for cons1_ae do not overlap between households in rural and urban areas*

iv. T-statistic, degrees of freedom, and p-values (3 values on the bottom). The p-values on the left/right are for whether the difference between the two means is less than or greater than 0 (one-sided t-test). The p-value in the middle is for whether the difference between the two means is not equal to 0 (two-sided t-test). We most commonly use two-sided t-tests.

v. What is the p-value that this difference in means is less than 0? And how can we interpret this? *P-value is 0.0000. We can say with more than 99% confidence that the mean monthly consumption per adult equivalent for households in urban areas is different than the mean monthly consumption per adult equivalent for households in rural areas*

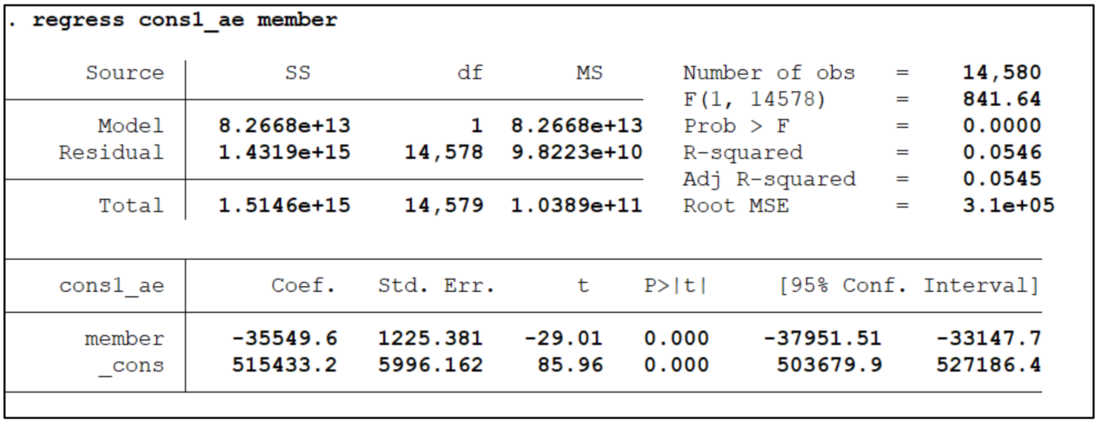
4. **Ordinary Least-Squares (OLS) Linear Regressions**

   a. Let's just start with one independent variable (the one that we used in our correlations – cons1_ae).

      i. Code: ***reg**ress [dependent_var] [independent_var1] [independent_var2] ….*

      ii. Practice: *"regress cons1_ae member"* – the output gives us a lot of information:

Figure 3.11: Stata command and output for "regress cons1_ae member"

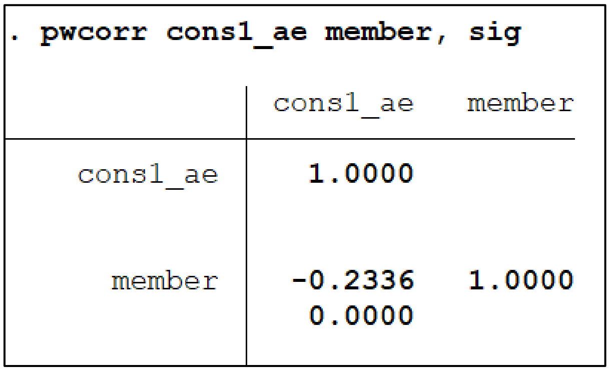


```
. regress cons1_ae member

      Source |       SS           df       MS      Number of obs   =    14,580
-------------+----------------------------------   F(1, 14578)     =    841.64
       Model |  8.2668e+13         1  8.2668e+13   Prob > F        =    0.0000
    Residual |  1.4319e+15    14,578  9.8223e+10   R-squared       =    0.0546
-------------+----------------------------------   Adj R-squared   =    0.0545
       Total |  1.5146e+15    14,579  1.0389e+11   Root MSE        =   3.1e+05

------------------------------------------------------------------------------
    cons1_ae |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      member |   -35549.6   1225.381   -29.01   0.000    -37951.51    -33147.7
       _cons |   515433.2   5996.162    85.96   0.000     503679.9    527186.4
------------------------------------------------------------------------------
```

1. Model specification information: Number of observations, degrees of freedom, F-statistic, R-squared, etc.
2. The bottom table shows us the associations between the independent variables (only 'member' in this output) and the dependent variable (cons1_ae): coefficient, standard error, t-statistic, p-value, and 95% confidence interval. The p-value should look familiar from our earlier analysis.

      iii. Practice: Let's run our correlation again, with significance: *"pwcorr cons1_ae member, sig"*

Figure 3.12: Stata command and output for "pwcorr cons1_ae member, sig"



```
. pwcorr cons1_ae member, sig

             |  cons1_ae   member
-------------+-------------------
    cons1_ae |    1.0000
             |
             |
      member |   -0.2336   1.0000
             |    0.0000
             |
```

iv. Question: Do you notice anything similar across the two outputs? *The p-values are the same*

v. An OLS regression with only two variables is basically showing the same thing as a correlation – you are not controlling for any other variables, and so the significance of the association between the two variables is the same.

b. Now, let's try some more variables. What else might be associated with consumption per adult equivalent?

i. Let's try to run a regression where consumption per adult equivalent is still the dependent variable, and the independent variables are the household size, the percent of consumption spent on food, and whether a household lives in an urban or rural area.

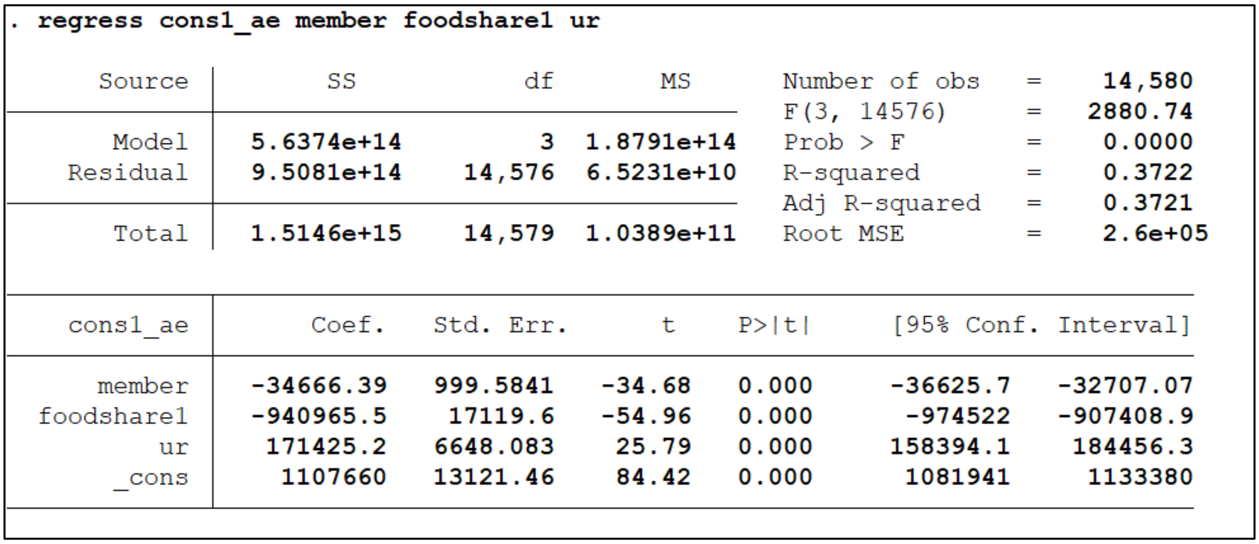
ii. Question: What types of variables are the independent variables in this regression? (Dummy variables and continuous variables are generally good to go into a regression, without any modifications)

1. *"codebook foodshare1"* – continuous; okay for regression
2. *"codebook member"* – continuous; okay for regression
3. *"codebook ur"* – categorical variable (1/2); **not** a dummy, so **not** okay for regression

iii. Practice: We need to recode *ur* so that it's a dummy variable. What code can we use? *"recode ur 2-=0";* now this variable is okay for the regression.

iv. <u>Practice: </u>*"regress cons1_ae member foodshare1 ur"*

Figure 3.13: Stata command and output for "regress cons1_ae member foodshare1 ur"



```
. regress cons1_ae member foodshare1 ur

      Source |       SS           df       MS      Number of obs   =    14,580
-------------+----------------------------------   F(3, 14576)     =   2880.74
       Model |  5.6374e+14         3  1.8791e+14   Prob > F        =    0.0000
    Residual |  9.5081e+14    14,576  6.5231e+10   R-squared       =    0.3722
-------------+----------------------------------   Adj R-squared   =    0.3721
       Total |  1.5146e+15    14,579  1.0389e+11   Root MSE        =    2.6e+05

------------------------------------------------------------------------------
    cons1_ae |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      member |  -34666.39   999.5841   -34.68   0.000     -36625.7   -32707.07
  foodshare1 |  -940965.5    17119.6   -54.96   0.000      -974522   -907408.9
          ur |   171425.2   6648.083    25.79   0.000     158394.1    184456.3
       _cons |    1107660   13121.46    84.42   0.000      1081941     1133380
------------------------------------------------------------------------------
```

v. Are any of these variables significant? *They are all significant – all p-values are 0.000*

vi. What does the coefficient on *ur* tell us? *Because ur is* ***now*** *a dummy variable where urban=1, the coefficient says that if the household lives in an urban area, their monthly consumption per adult equivalent will <u>increase</u> by 171,425 RWF compared to households in rural areas, holding household size and food share constant. (The coefficient is 171,425)*

vii. What does the coefficient on *member* tell us? *Because member is a continuous variable, the coefficient says that the marginal effect of one additional person in the household <u>decreases</u> the monthly consumption per adult equivalent by 34,666 RWF. (The coefficient is -34,666)*

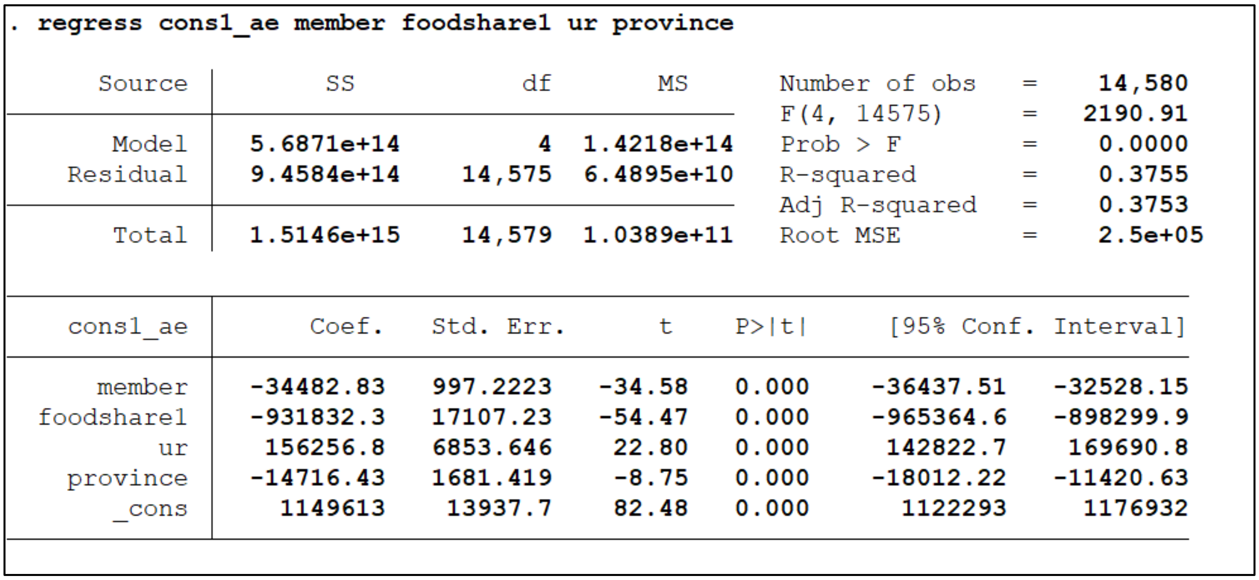
viii. What does the coefficient on *foodshare1* tell us? *Households that spend more of their budget on food tend to have <u>smaller</u> monthly consumption per adult equivalent (that is, lower income households are spending more of their incomes on food compared to higher income households)*

c. Let's add province to our regression.

i. What type of variable is province? *"codebook province" – categorical (5 different categories/values for the 4 different provinces and Kigali City)*

ii. Practice: *"regress cons1_ae member foodshare1 ur province"*

Figure 3.14: Stata command and output for "regress cons1_ae member foodshare1 ur province"



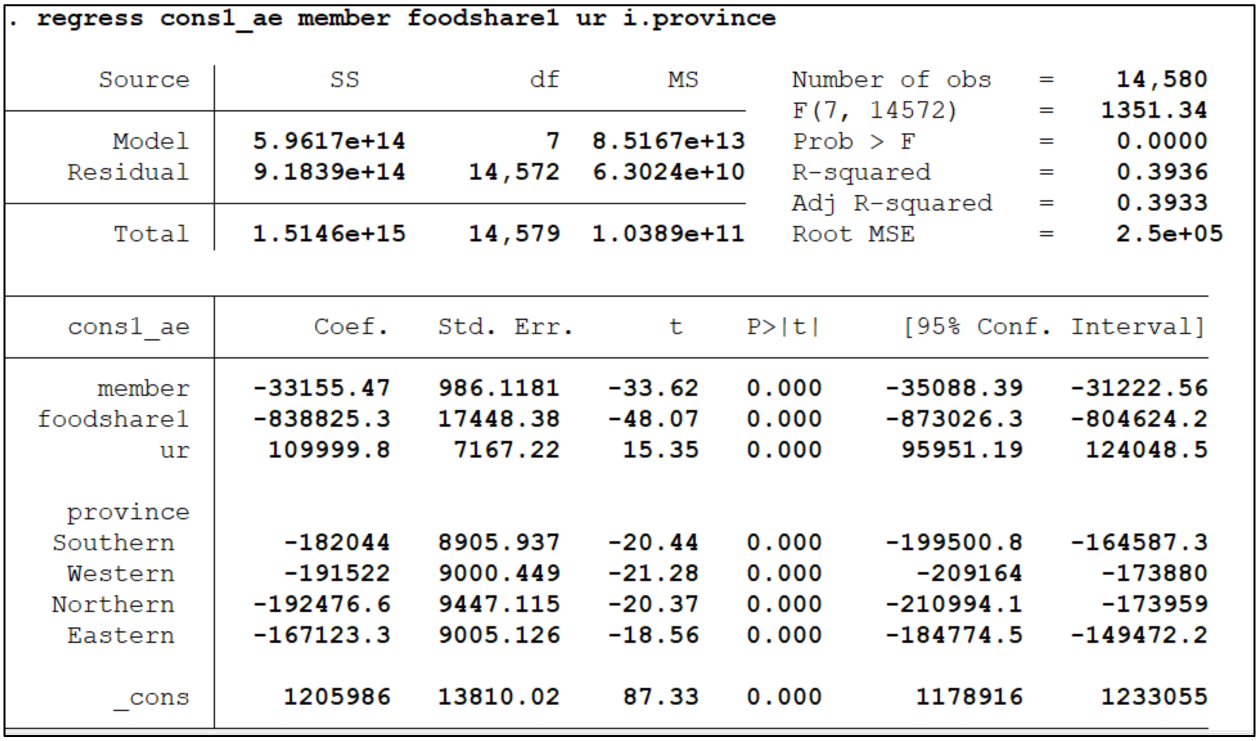
```
. regress cons1_ae member foodshare1 ur province

      Source |       SS           df       MS      Number of obs   =    14,580
-------------+----------------------------------   F(4, 14575)     =   2190.91
       Model |  5.6871e+14         4  1.4218e+14   Prob > F        =    0.0000
    Residual |  9.4584e+14    14,575  6.4895e+10   R-squared       =    0.3755
-------------+----------------------------------   Adj R-squared   =    0.3753
       Total |  1.5146e+15    14,579  1.0389e+11   Root MSE        =    2.5e+05

------------------------------------------------------------------------------
    cons1_ae |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      member |  -34482.83   997.2223   -34.58   0.000    -36437.51   -32528.15
  foodshare1 |  -931832.3   17107.23   -54.47   0.000    -965364.6   -898299.9
          ur |   156256.8   6853.646    22.80   0.000     142822.7    169690.8
    province |  -14716.43   1681.419    -8.75   0.000    -18012.22   -11420.63
       _cons |    1149613    13937.7    82.48   0.000      1122293     1176932
------------------------------------------------------------------------------
```

iii. Question: Is province significant? How can we interpret the coefficient on "province"? *Province is significant. For every one unit increase in province, the monthly consumption per adult equivalent decreases by 14,716 RWF… This doesn't make sense! Province isn't a continuous or dummy variable….*

iv. So instead, we can put "i." in front of *province* (or any categorical non-dummy variables). Let's try again:

v. Practice: *"regress cons1_ae member foodshare1 ur i.province"*

Figure 3.15: Stata command and output for "regress cons1_ae member foodshare1 ur i.province"



```
. regress cons1_ae member foodshare1 ur i.province

      Source |       SS           df       MS      Number of obs   =    14,580
-------------+----------------------------------   F(7, 14572)     =   1351.34
       Model |  5.9617e+14         7  8.5167e+13   Prob > F        =    0.0000
    Residual |  9.1839e+14    14,572  6.3024e+10   R-squared       =    0.3936
-------------+----------------------------------   Adj R-squared   =    0.3933
       Total |  1.5146e+15    14,579  1.0389e+11   Root MSE        =    2.5e+05

------------------------------------------------------------------------------
    cons1_ae |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      member |  -33155.47   986.1181   -33.62   0.000    -35088.39   -31222.56
  foodshare1 |  -838825.3   17448.38   -48.07   0.000    -873026.3   -804624.2
          ur |   109999.8    7167.22    15.35   0.000     95951.19    124048.5
             |
    province |
   Southern  |    -182044   8905.937   -20.44   0.000    -199500.8   -164587.3
    Western  |    -191522   9000.449   -21.28   0.000      -209164     -173880
   Northern  |  -192476.6   9447.115   -20.37   0.000    -210994.1     -173959
    Eastern  |  -167123.3   9005.126   -18.56   0.000    -184774.5   -149472.2
             |
       _cons |    1205986   13810.02    87.33   0.000      1178916     1233055
------------------------------------------------------------------------------
```

vi. This now turns province into 4 dummy variables for the regression. Stata automatically chooses the category with the lowest value (here 1=Kigali City) to drop to be the comparison group. So now, each other province is being compared to Kigali City.

vii. Are any of the provinces significant? *All provinces have p-values=0.000.*

viii. This means that, for example, holding these other variables constant, households in Southern Province have an average monthly consumption per adult equivalent that is 182,044 RWF *lower* than households in Kigali City.

ix. Similarly, holding these other variables constant, on average households in Western Province are 191,522 RWF lower, households in Northern province are 192,476 RWF lower, and households in Eastern Province are 167,123 RWF lower than Kigali City.

5. Conclusions and Next Steps

This concludes the Stata Introductory Course on Describing, Transforming, and Analyzing Data. This training simply provided an overview of the most common Stata commands, best practices to construct and use them, and how to interpret their output. However, this is only the beginning of all of the data cleaning and analytic capabilities that Stata can provide! You are encouraged to explore the Stata software more, especially through the “help” function and through online resources, to see what other tools are available.

However, it is important to remember that Stata is just a tool (one of many!), and the most important thing is to understand your data and your analysis objectives. It is important to always choose the appropriate Stata codes and statistical techniques to conduct your analysis, in order to adequately answer your research questions. The first step is to always know your data! It is best practice to first describe and clean your data, before beginning your analysis. There are numerous resources available online to provide further information on how to best use Stata to achieve your research goals.

Creating Figures in Stata

1. Bar Graphs (1 variable) – Poverty categories

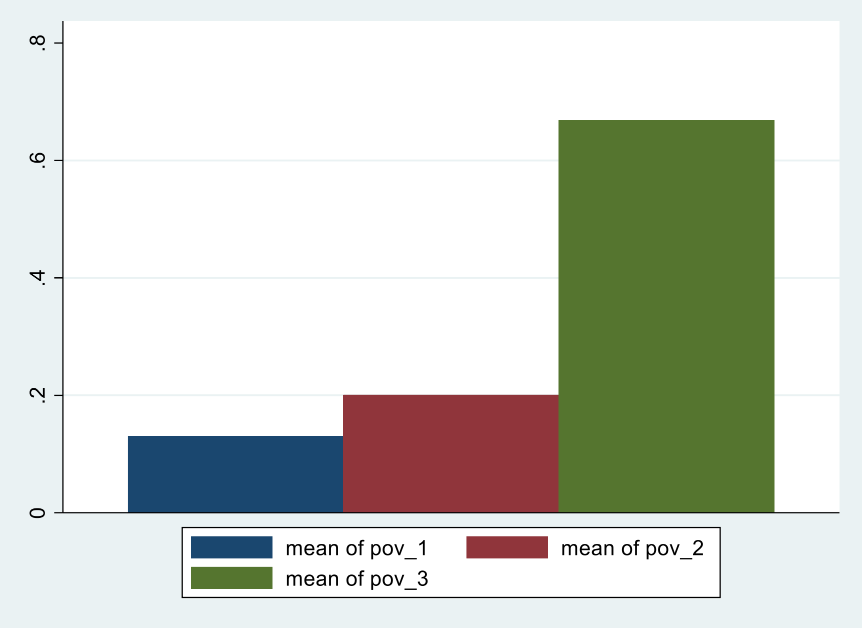
Data: "cs_S0_S5_Household.dta"

a. Figure A is a simple bar graph, showing the percent of households in each poverty category, across the whole sample. For categorical variables (such as poverty category), each category needs to be turned into a separate dummy variable (taking the values 0="No" and 1="Yes") so that the bar graph will show the prevalence of each.

   i. Code: To create separate dummy variables – *tab [varname], gen(XX_)* – where the XX is whatever variable name stem you want. This code will create a new dummy variable for each category of the original variable.

   ii. For poverty categories: *tab poverty, gen(pov_)* – since the poverty variable has 3 categories, this creates 3 variables ranging from pov_1 to pov_3

   iii. Code: *"graph bar pov_1 pov_2 pov_3"*



   iv. What does this figure tell us?

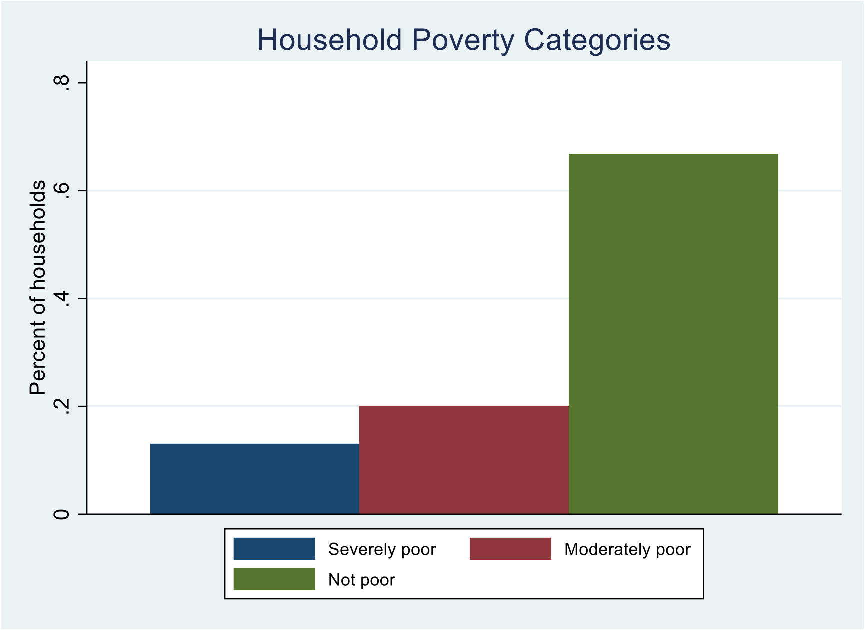
      1. pov_3 is the most common category

   v. What does this figure not tell us?

      1. What pov_1, pov_2, and pov_3 mean
      2. What the y-axis means
      3. What the title of the figure is

b. Figure B takes the same simple bar graph from above, but adds these three missing elements

i. Code: *“graph bar pov_1 pov_2 pov_3, legend(size(small) order(1 "Severely poor" 2 "Moderately poor" 3 "Not poor")) ytitle("Percent of households") title("Household Poverty Categories")”*

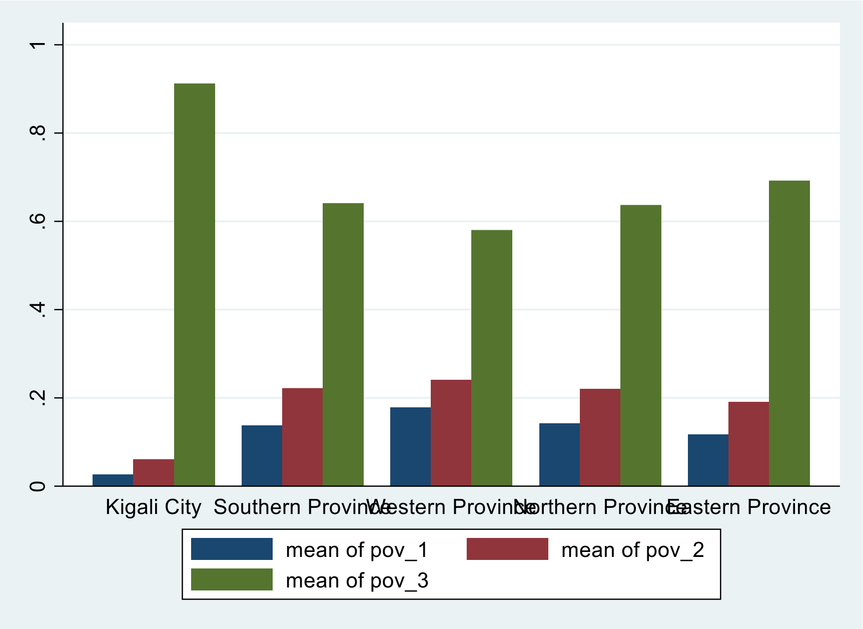


ii. What is now added to this figure?
   1. Labels for pov_1, pov_2, and pov_3 – *“legend(size(small) order(1 "Severely poor" 2 "Moderately poor" 3 "Not poor"))”*
   2. Y-axis label – *“ytitle("Percent of households")”*
   3. Title – *“title("Household Poverty Categories")”*

iii. What can we learn from this figure?
   1. Most households are non-poor

2. **Bar Graphs (2 variables) – Poverty Status by Province**
Data: “cs_S0_S5_Household.dta” (keep the modified data from #1 in which we generated separate dummy variables for each poverty category)
   a. Figure A is like the simple bar graph created above, but it is separated into 5 different groups (one for each province and Kigali City)

i. Code: *"graph bar pov_1 pov_2 pov_3, over(province)"*



ii. What does this figure tell us?

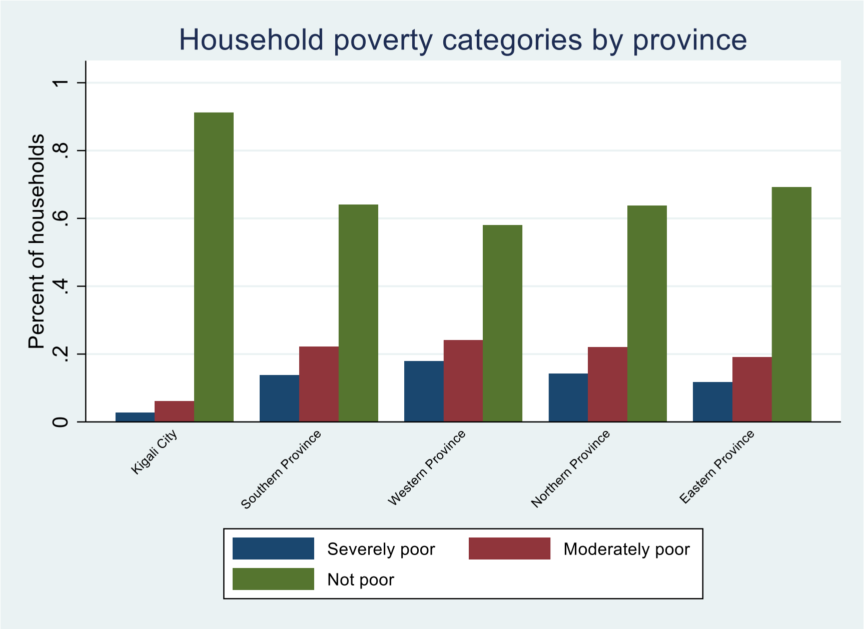
1. pov_3 is the highest in Kigali City and the lowest in Western Province

iii. What does this figure not tell us?

1. What pov_1, pov_2, and pov_3 mean
2. What the y-axis means
3. What the title of the figure is

b. Figure B takes the same bar graphs by province from above, but adds these three missing elements

i. Code: *“graph bar pov_1 pov_2 pov_3, over(province, label(labsize(vsmall) angle(45))) legend(size(small) order(1 "Severely poor" 2 "Moderately poor" 3 "Not poor")) ytitle("Percent of households") title("Household poverty categories by province")”*



ii. What is now added to this figure?

1. Labels for pov_1, pov_2, and pov_3 – *“legend(size(small) order(1 "Severely poor" 2 "Moderately poor" 3 "Not poor"))”*
2. Y-axis label – *“ytitle("Percent of households")”*
3. Title – *“title("Household poverty categories by province")”*
4. Fixed province names so they don’t overlap – *“over(province, label(labsize(vsmall) angle(45)))”*

iii. What can we learn from this figure?

1. Households in Kigali City are most likely to be non-poor
2. Households in Western Province are least likely to be non-poor

3. Line Graphs (1 variable – kernel density) – Consumption per Adult Equivalent

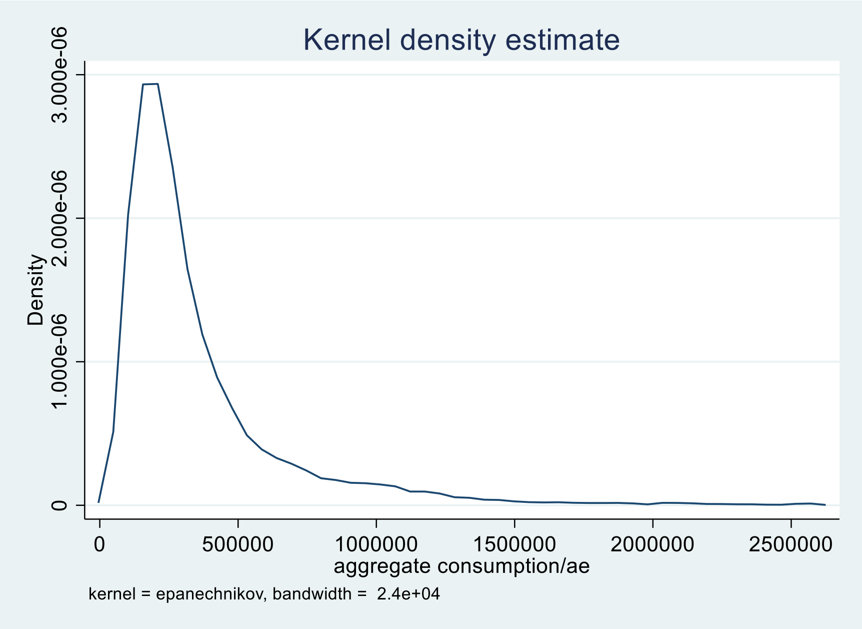
Data: “EICV5_Poverty_file.dta”

a. Figure A shows consumption per adult equivalent across the whole sample – because consumption is a numeric, continuous variable, and we are not looking at it compared to any other variable, we will use a kernel density line graph (kdensity). The kdensity code produces a smooth line graph of the density of one variable (a univariate kernel density estimation).

i. Let’s first replace the outliers again, as we did in Lesson 3

1. “sum cons1_ae, det”
2. “replace cons1_ae=256655.8 if cons1_ae>2600404”

ii. Code: *"kdensity cons1_ae"*

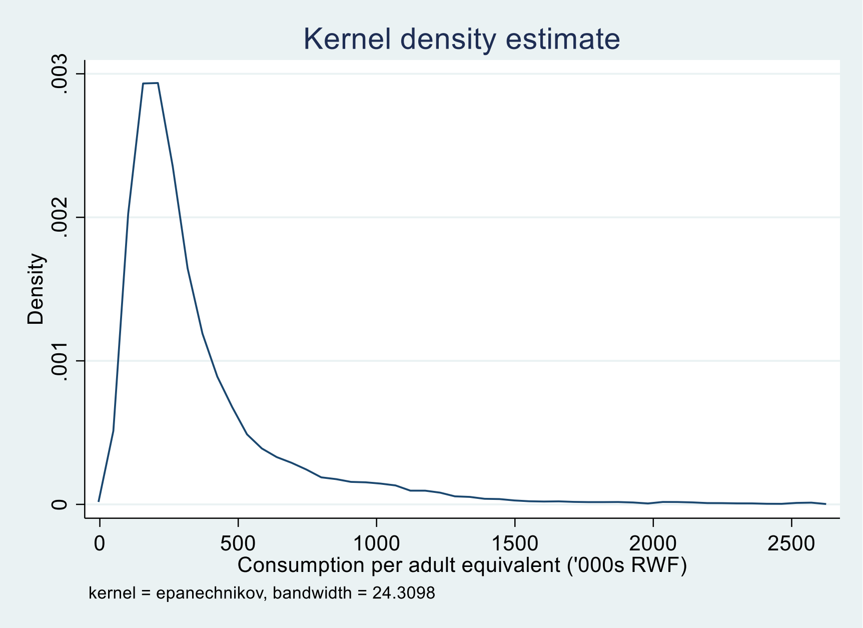


iii. What does this figure show? *Across the whole sample, most households consume between 0-500,000 RWF per adult equivalent per month*

b. Figure B changes the scale of the consumption variable, so the graph is not so busy with all of the big consumption numbers on the x-axis

i. "replace cons1_ae=cons1_ae/1000"

ii. Code: *"kdensity cons1_ae, xtitle("Consumption per adult equivalent ('000s RWF)")"*



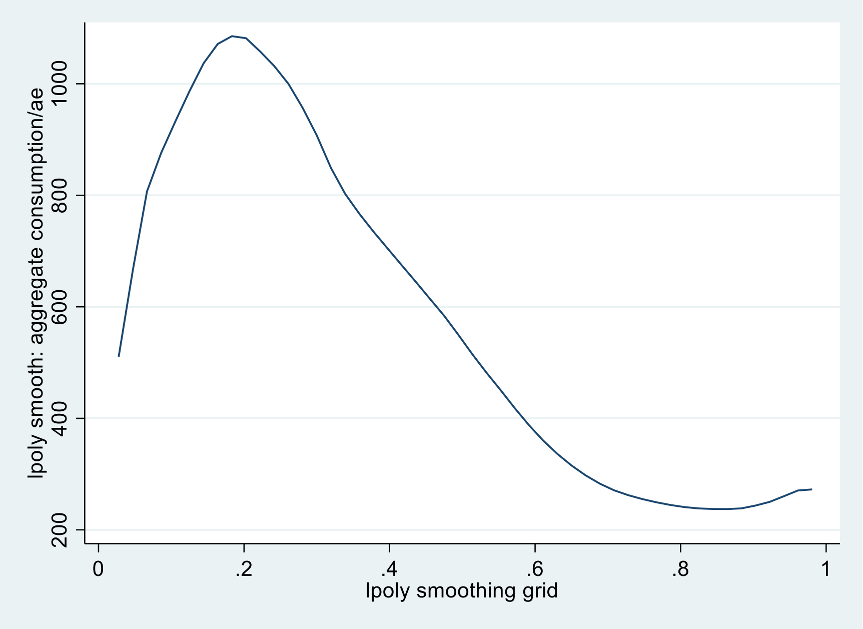
4. Line Graphs (2 variables – local polynomial) – Consumption per Adult Equivalent and Household Size

Data: "EICV5_Poverty_file.dta" (keep the modified data from #3 in which we changed the scale of cons1_ae)

a. Figure A creates a simple line graph (local polynomial) showing monthly consumption per adult equivalent by the percent of household budget spent on food

   i. Because both variables (cons1_ae and foodshare1) are numeric/continuous, we will make a line graph to show this (specifically, a local polynomial)

   ii. Code: *"twoway (lpoly cons1_ae foodshare1)"*

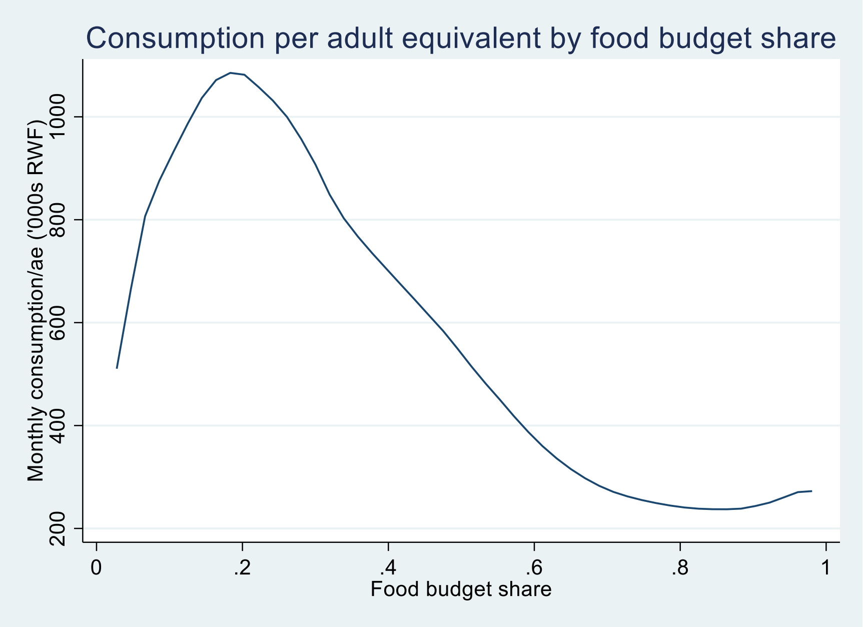


   iii. What is this figure missing?

      1. *X-axis title*
      2. *Title*

b. Figure B adds titles to the figure above

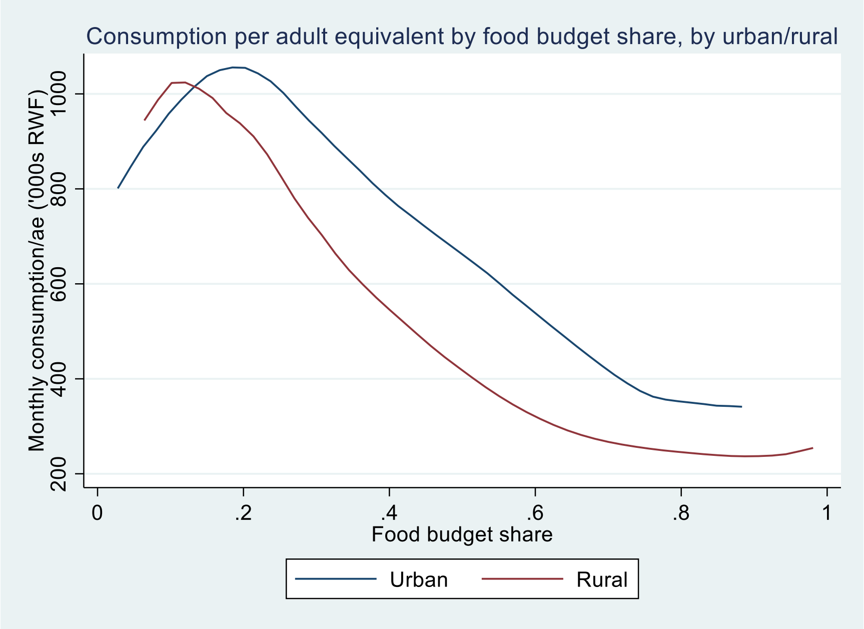
   i. Code: *"twoway (lpoly cons1_ae foodshare1), xtitle("Food budget share") ytitle("Monthly consumption/ae ('000s RWF)") title("Consumption per adult equivalent by food budget share")"*



ii. What happens to consumption as the food budget share increases?
Consumption per adult equivalent decreases as the food budget share increases (meaning households with lower consumption spend a larger percent of that consumption on food)

c. Consumption and food budget shares by rural/urban: Figure C creates a similar figure to Figure B, but creates separate lines for urban and rural

i. Code: *"twoway (lpoly cons1_ae foodshare1 if ur==1) (lpoly cons1_ae foodshare1 if ur==2), legend(lab(1 "Urban") lab(2 "Rural")) xtitle("Food budget share") ytitle("Monthly consumption/ae ('000s RWF)") title("Consumption per adult equivalent by food budget share, by urban/rural", size(smallmed))"*



ii. What does this figure show? *Overall, rural households have smaller consumption than urban households, but the trend between consumption and food budget share is similar across both groups of households*

IFPRI Contact Information

Gracie Rosenbach

Rwanda Country Program Manager

g.rosenbach@cgiar.org

Gilberthe Benimana

Research Analyst

g.benimana@cgiar.org

David J Spielman

Rwanda Program Leader

d.spielman@cgiar.org