INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
IFPRI

Australian Government
Department of Foreign Affairs and Trade

Australian National University

Describing, Transforming, and Analyzing Data Using Stata

An Introductory Course

Emily Schmidt, Peixun Fang, Lucia Carrillo

International Food Policy Research Institute

November - December, 2021

Virtual Training

Lesson 1 – Introduction and Describing the Data

1. Introduction to Stata and the PNG rural household survey

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 15 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 research and data

In May – July, 2018, the International Food Policy Research Institute (IFPRI) implemented a rural household survey to investigate the food systems of rural households in Papua New Guinea (PNG) and how they assure sufficient food to meet the nutritional needs of their household members.

The survey set out to achieve two goals:

1) Use the household survey results to inform inclusive policy dialogue centered on agricultural productivity, enhanced food security, and improved nutrition policies in PNG;

2) Collect baseline data for recently implemented rural development activities that are currently being supported by World Vision (WV).

These two objectives largely determined how the household survey sample was designed. It is important to note that the survey is not nationally representative, however it does represent a successful initiative to test the feasibility of implementing a comprehensive household survey in select rural areas of PNG.

c. Sampling methodology

Prior to survey implementation, IFPRI undertook scoping visits to the survey communities. We observed little variability in livelihood activities, production systems, and food systems ***within*** community sites, although there are differences ***across*** communities.

This lack of apparent variance ***within*** communities in local areas presents a challenge for planning a representative household survey to examine food systems and food security.

As a consequence of the lack of significant variance that we perceive in key variables for our study ***within*** communities, but likely significantly greater variance ***across*** communities, it is important that the study population we sample for the survey is as spatially extensive as possible in order to capture:

- Differences across communities in access to markets and, consequently, in the livelihood strategies that they might pursue;
- Perceived human capital – particularly, educational attainment – that varies across (and possibly within) communities;
- Agroecological conditions which vary as the distance between communities increases.

d. Sample Selection

Based on observations made during the scoping, we constructed the sample based on the following criteria:

1. Defined the study population based on a distance/time-of-travel measurement (using GIS analysis that considered road and walking times) from World Vision program communities.
2. The survey sample was drawn from the study populations defined within each of the four sites that were in the districts in which World Vision's programs were operating:
    a. West Sepik – Nuku District
    b. East Sepik – Maprik
    c. Madang – Middle Ramu
    d. Autonomous Region of Bougainville – South Bougainville
3. Within each sample area, 20 communities were randomly selected. The sample of communities was split into World Vision and non- World Vision sub-samples:
    a. Within each community, 15 households were randomly selected to be interviewed;
    b. In addition to 15 household questionnaires per community, a community questionnaire was administered to a group of identified community leaders (4 to 5 representatives) in each community.

e. Questionnaire development

Over 80% of the population in PNG is dependent on rain-fed subsistence farming for a significant component of their livelihoods, and more than three-quarters of the food consumed in the country is locally grown. Thus, it was important that the survey questionnaire captured agricultural activities, land ownership, and assets in order to understand rural livelihood structure.

In addition, we assumed that income (measured from consumption and expenditure data) is associated with household livelihood activities and household decisions, so understanding the income structure of each household by including detailed consumption and expenditure data was important to collect in order to classify households into expenditure categories

Finally, high child stunting rates, insufficient protein intake among rural populations, and substantial food shortages experienced during significant El Niño–Southern Oscillation (ENSO) events suggest that updated information on child nutrition is necessary in order to understand how to sufficiently meet the dietary needs of rural communities.

Thus, the primary objective of the survey investigated the food systems of rural households and how they assure sufficient food to meet the nutritional needs of their household members. Simultaneously, the survey aims to evaluate the nutritional status of children.

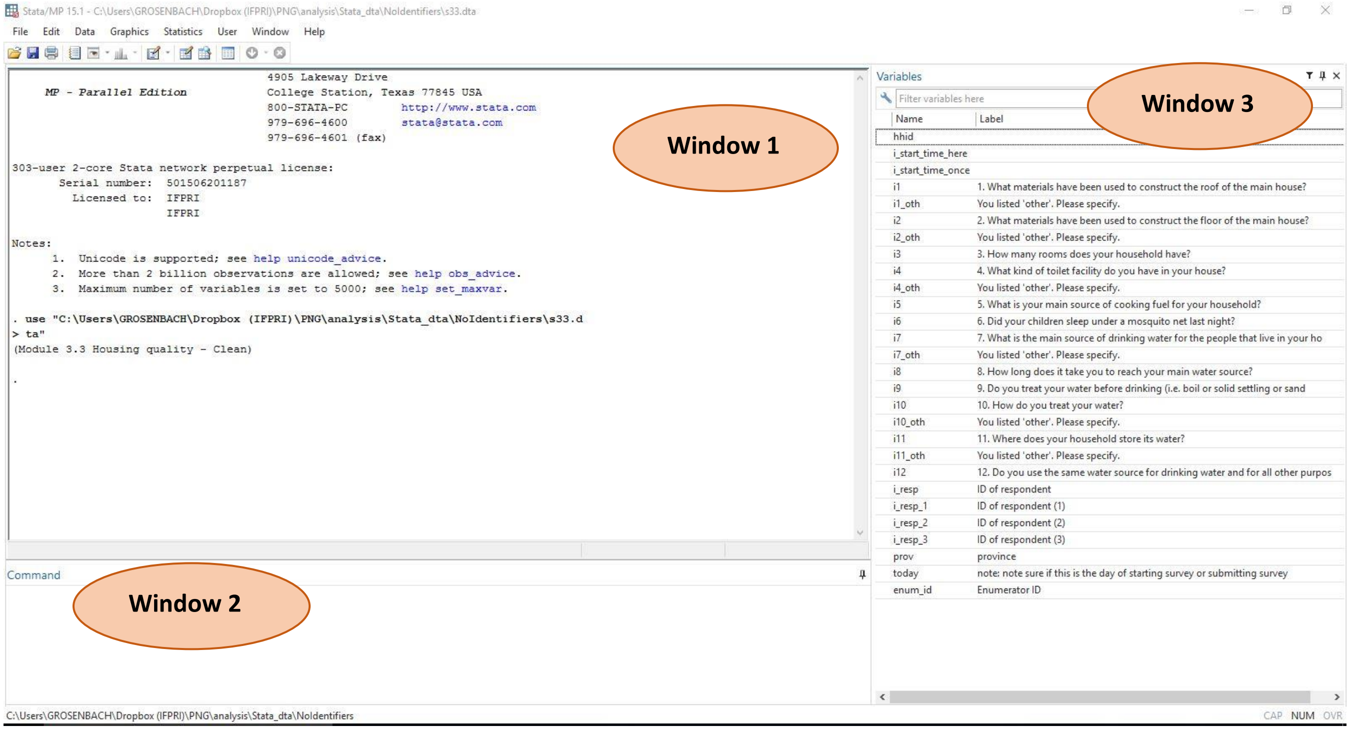
The household questionnaire was asked using tablets, which exported the responses into Excel (which we then converted to Stata files for analysis). The survey included modules on:

1. Household characteristics (demographics, education, migration, etc.)
2. Agricultural production (crop production, use of household labor, and agricultural extension)
3. Household assets (production equipment and consumer durables, livestock ownership, and housing quality)
4. Income apart from own agricultural activities (wage employment, own business activities, and income transfers and gifts)
5. Consumption and expenditures (nonfood expenditures, food consumption, and dietary diversity)
6. Economic shocks, household well-being, and food insecurity (shocks and coping strategies, perceptions of poverty, recent food insecurity, and health and nutrition extension)
7. Gender roles and social expectations (asked separately for men and women)
8. Female health (pregnancy care and breastfeeding practices)
9. Child health (healthcare, vaccinations, and anthropometry)

2. Exploring Stata and the data

a. Let’s explore some data on household characteristics – ***double-click to open file s33.dta*** – the following Stata screen should appear. The *main Stata screen* consists of 3 windows:

Figure 1.1: Main Stata screen


Stata/MP 15.1 - C:\Users\GROSENBACH\Dropbox (IFPRI)\PNG\analysis\Stata_dta\NoIdentifiers\s33.dta
File
Edit
Data
Graphics
Statistics
User
Window
Help
Window 1
Window 2
Window 3
Variables
Filter variables here
Command
C:\Users\GROSENBACH\Dropbox (IFPRI)\PNG\analysis\Stata_dta\NoIdentifiers
CAP
NUM
OVR

```
                                  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:  501506201187
         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\Dropbox (IFPRI)\PNG\analysis\Stata_dta\NoIdentifiers\s33.d
> ta"
(Module 3.3 Housing quality - Clean)

.
```

Name	Label
hhid	
i_start_time_here	
i_start_time_once	
i1	1. What materials have been used to construct the roof of the main house?
i1_oth	You listed 'other'. Please specify.
i2	2. What materials have been used to construct the floor of the main house?
i2_oth	You listed 'other'. Please specify.
i3	3. How many rooms does your household have?
i4	4. What kind of toilet facility do you have in your house?
i4_oth	You listed 'other'. Please specify.
i5	5. What is your main source of cooking fuel for your household?
i6	6. Did your children sleep under a mosquito net last night?
i7	7. What is the main source of drinking water for the people that live in your ho
i7_oth	You listed 'other'. Please specify.
i8	8. How long does it take you to reach your main water source?
i9	9. Do you treat your water before drinking (i.e. boil or solid settling or sand
i10	10. How do you treat your water?
i10_oth	You listed 'other'. Please specify.
i11	11. Where does your household store its water?
i11_oth	You listed 'other'. Please specify.
i12	12. Do you use the same water source for drinking water and for all other purpos
i_resp	ID of respondent
i_resp_1	ID of respondent (1)
i_resp_2	ID of respondent (2)
i_resp_3	ID of respondent (3)
prov	province
today	note: note sure if this is the day of starting survey or submitting survey
enum_id	Enumerator ID

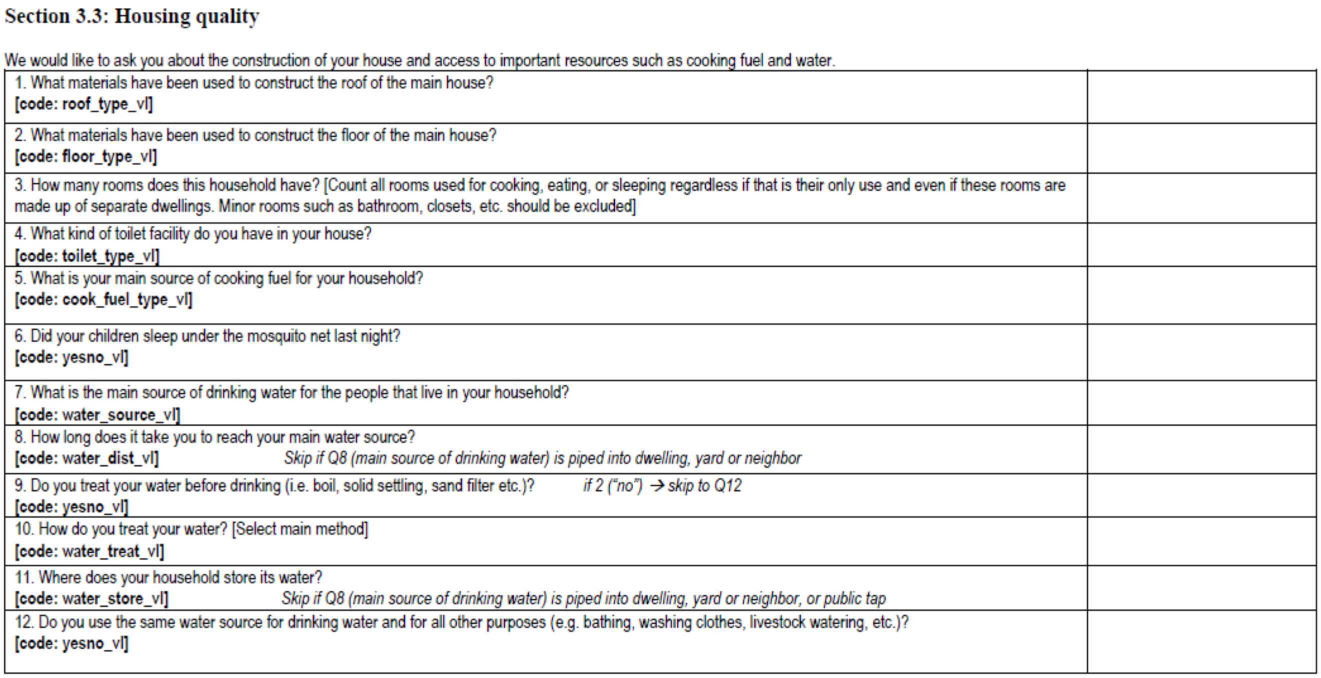
i. Window 1: Review and Results – shows the commands entered and output generated from these commands

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 s33 corresponds to Module 3.3 in the PNG Household Survey on Food Systems Questionnaire:

Figure 1.2: Module 3.3 (Housing quality) from PNG Household Survey on Food Systems Questionnaire



Section 3.3: Housing quality

We would like to ask you about the construction of your house and access to important resources such as cooking fuel and water.

1. What materials have been used to construct the roof of the main house? **[code: roof_type_vl]**	
2. What materials have been used to construct the floor of the main house? **[code: floor_type_vl]**	
3. How many rooms does this household have? [Count all rooms used for cooking, eating, or sleeping regardless if that is their only use and even if these rooms are made up of separate dwellings. Minor rooms such as bathroom, closets, etc. should be excluded]	
4. What kind of toilet facility do you have in your house? **[code: toilet_type_vl]**	
5. What is your main source of cooking fuel for your household? **[code: cook_fuel_type_vl]**	
6. Did your children sleep under the mosquito net last night? **[code: yesno_vl]**	
7. What is the main source of drinking water for the people that live in your household? **[code: water_source_vl]**	
8. How long does it take you to reach your main water source? **[code: water_dist_vl]** *Skip if Q8 (main source of drinking water) is piped into dwelling, yard or neighbor*	
9. Do you treat your water before drinking (i.e. boil, solid settling, sand filter etc.)? *if 2 ("no") → skip to Q12* **[code: yesno_vl]**	
10. How do you treat your water? [Select main method] **[code: water_treat_vl]**	
11. Where does your household store its water? **[code: water_store_vl]** *Skip if Q8 (main source of drinking water) is piped into dwelling, yard or neighbor, or public tap*	
12. Do you use the same water source for drinking water and for all other purposes (e.g. bathing, washing clothes, livestock watering, etc.)? **[code: yesno_vl]**	

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

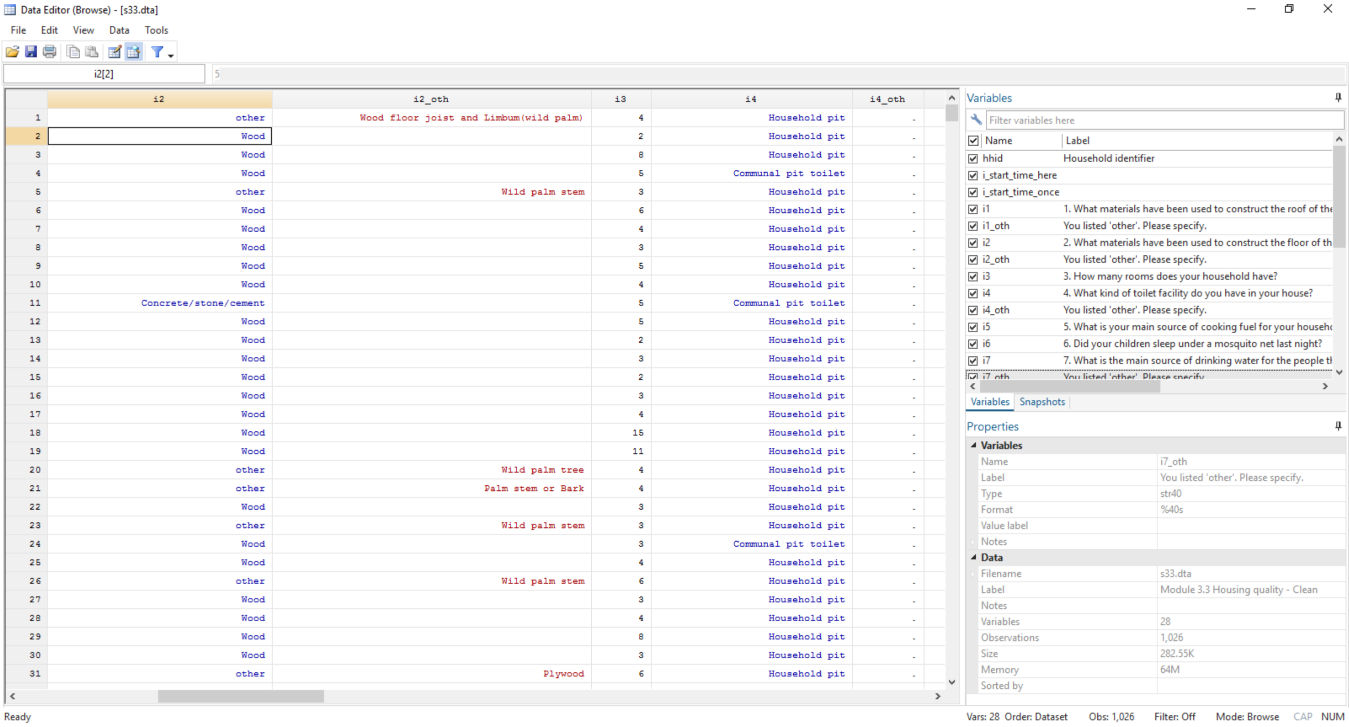
c. There are many separate Stata screens that can be opened. One of these screens is the *data browser* in which you can look at the data currently loaded in Stata. There are 3 different ways to access the *data browser:*

   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



	i2	i2_oth	i3	i4	i4_oth
1	other	Wood floor joist and Limbum(wild palm)	4	Household pit	.
2	Wood		2	Household pit	.
3	Wood		8	Household pit	.
4	Wood		5	Communal pit toilet	.
5	other	Wild palm stem	3	Household pit	.
6	Wood		6	Household pit	.
7	Wood		4	Household pit	.
8	Wood		3	Household pit	.
9	Wood		5	Household pit	.
10	Wood		4	Household pit	.
11	Concrete/stone/cement		5	Communal pit toilet	.
12	Wood		5	Household pit	.
13	Wood		2	Household pit	.
14	Wood		3	Household pit	.
15	Wood		2	Household pit	.
16	Wood		3	Household pit	.
17	Wood		4	Household pit	.
18	Wood		15	Household pit	.
19	Wood		11	Household pit	.
20	other	Wild palm tree	4	Household pit	.
21	other	Palm stem or Bark	4	Household pit	.
22	Wood		3	Household pit	.
23	other	Wild palm stem	3	Household pit	.
24	Wood		3	Communal pit toilet	.
25	Wood		4	Household pit	.
26	other	Wild palm stem	6	Household pit	.
27	Wood		3	Household pit	.
28	Wood		4	Household pit	.
29	Wood		8	Household pit	.
30	Wood		3	Household pit	.
31	other	Plywood	6	Household pit	.

Variables

Filter variables here

☑ Name	Label
☑ hhid	Household identifier
☑ i_start_time_here	
☑ i_start_time_once	
☑ i1	1. What materials have been used to construct the roof of the
☑ i1_oth	You listed 'other'. Please specify.
☑ i2	2. What materials have been used to construct the floor of th
☑ i2_oth	You listed 'other'. Please specify.
☑ i3	3. How many rooms does your household have?
☑ i4	4. What kind of toilet facility do you have in your house?
☑ i4_oth	You listed 'other'. Please specify.
☑ i5	5. What is your main source of cooking fuel for your househo
☑ i6	6. Did your children sleep under a mosquito net last night?
☑ i7	7. What is the main source of drinking water for the people th
☑ i7_oth	You listed 'other'. Please specify.

Variables Snapshots

Properties

Variables	
Name	i7_oth
Label	You listed 'other'. Please specify.
Type	str40
Format	%40s
Value label	
Notes	
Data	
Filename	s33.dta
Label	Module 3.3 Housing quality - Clean
Notes	
Variables	28
Observations	1,026
Size	282.55K
Memory	64M
Sorted by	

Ready

Vars: 28 Order: Dataset Obs: 1,026 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 ***i3***, for which respondents gave a numeric answer to "How many rooms does your household have?"

   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 ***i2***, for which respondents selected an option from a multiple-choice list for the question "What materials have been used to construct the floor of the main house?". 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 "Wood", you will see the number "5" appear in the top bar, indicating that the number 5 is assigned the value "Wood".

   iii. String data – appears in **red**. String data are non-numeric. An example in this dataset is ***i2_oth***, for which data collectors typed in respondents' "other" responses to question I2.

3. Top descriptive commands

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

   i. <u>Practice:</u> How many observations are in this dataset? *"count" – 1,026*

Figure 1.4: Stata command and output for "count"

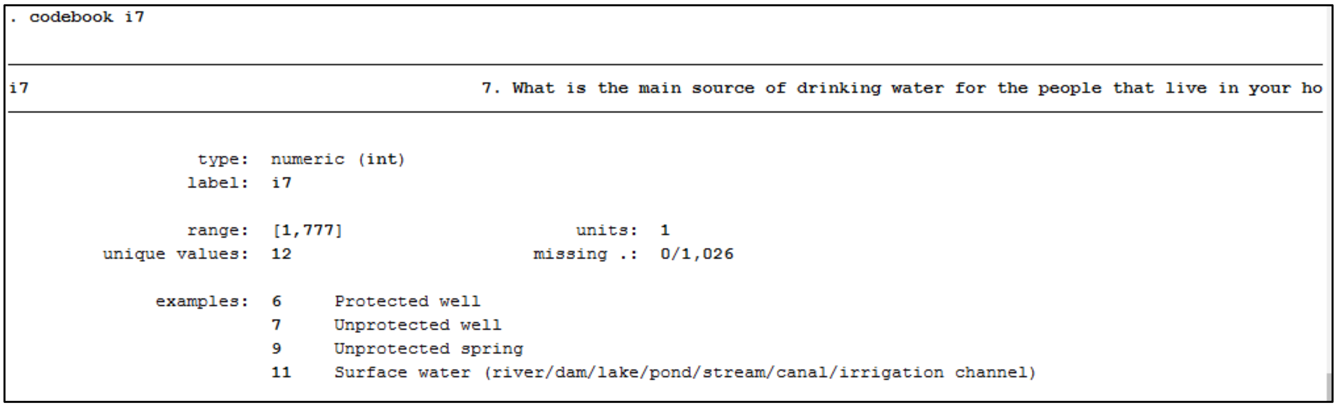
```
count
1,026
```

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

   i. <u>Code:</u> *codebook [variable name]*

   ii. <u>Practice:</u> What kind of information do we have on drinking water? Type *"codebook i7"* 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 i7"



```
. codebook i7

-------------------------------------------------------------------------------------------------------------
i7                                          7. What is the main source of drinking water for the people that live in your ho
-------------------------------------------------------------------------------------------------------------

                  type:  numeric (int)
                 label:  i7

                 range:  [1,777]                      units:  1
         unique values:  12                       missing .:  0/1,026

              examples:  6     Protected well
                         7     Unprotected well
                         9     Unprotected spring
                         11    Surface water (river/dam/lake/pond/stream/canal/irrigation channel)
```

a) What is the question that was asked in i7? (The variable label)? *7. What is the main source of drinking water for the people that live in your ho?*

- The label was cut off after 80 characters, but we can look at the questionnaire if we are unsure about the whole question.

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

c) What is the range of the data? *1 to 777*

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

e) How many unique values are there? *12*

f) What is one of the labels assigned to a value? *6 is protected well (for example)*

- We know that there are 12 unique values, but only 4 (6=protected well; 7=unprotected well; 9=unprotected spring; 11=surface water) are shown here as examples of what the data look like

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

   i. <u>Code:</u> ***tab**ulate [variable name]*

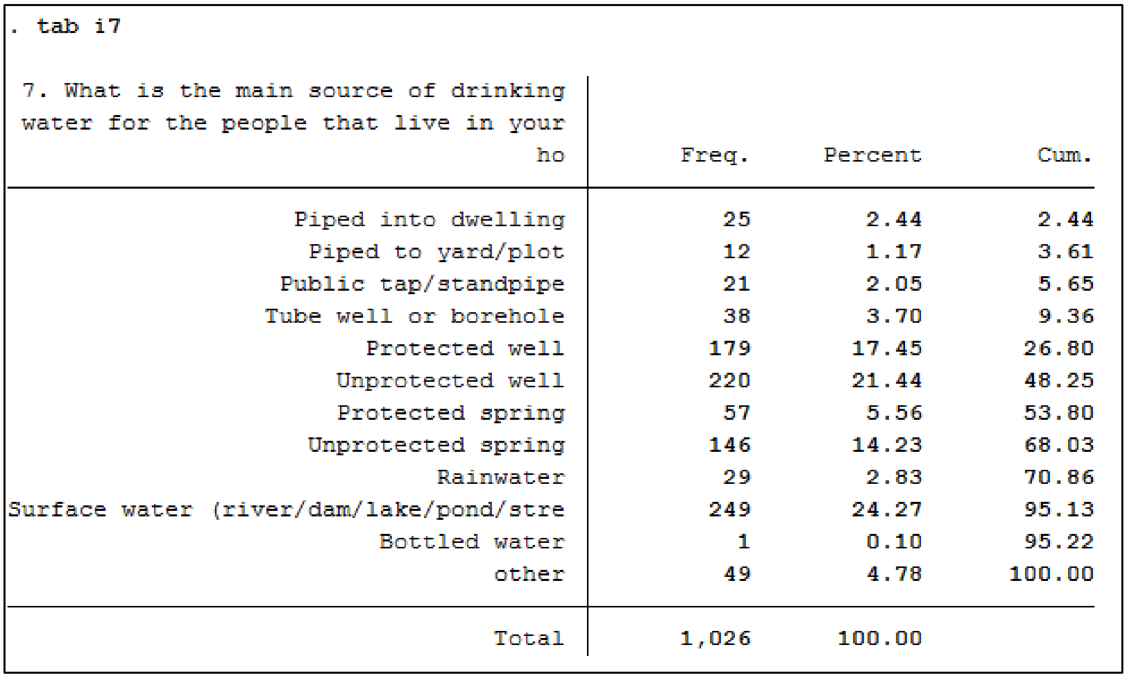
   a) Many Stata codes have "shorthand" versions – you only have to type a shortened version of the code and Stata will recognize the full command.

   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. <u>Practice:</u> How many households get their drinking water from a protected well? *"tabulate i7"* or *"tab i7".* This output tells us:

Figure 1.6: Stata command and output for "tab i7"



```
. tab i7

 7. What is the main source of drinking |
 water for the people that live in your |
                                     ho |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
                   Piped into dwelling |         25        2.44        2.44
                     Piped to yard/plot |         12        1.17        3.61
                  Public tap/standpipe |         21        2.05        5.65
                   Tube well or borehole |         38        3.70        9.36
                         Protected well |        179       17.45       26.80
                       Unprotected well |        220       21.44       48.25
                       Protected spring |         57        5.56       53.80
                     Unprotected spring |        146       14.23       68.03
                              Rainwater |         29        2.83       70.86
Surface water (river/dam/lake/pond/stre |        249       24.27       95.13
                          Bottled water |          1        0.10       95.22
                                  other |         49        4.78      100.00
----------------------------------------+-----------------------------------
                                  Total |      1,026      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 get their water from a protected well? *179 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 sample get their water from a protected well? *17.45% 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 people responded "other"? *49*

iii. What did the households that responded "other" write in as their water source?

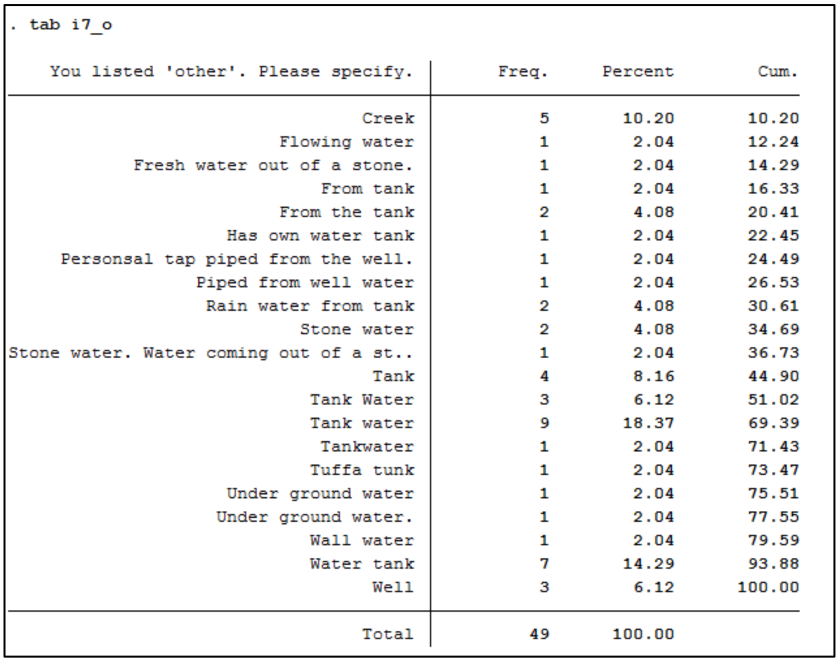
a) What kind of a variable do you think i7_oth is? Numeric, categorical, or string? *String*

b) How can we find out? *"codebook i7_oth" OR see what color it is when we "browse"*

c) Practice: Type *"codebook i7_oth"* and see if we are right

d) Practice: Type *"tab i7_oth"* to see what responses were typed in

Figure 1.7: Stata command and output for "tab i7_oth"



```
. tab i7_o

   You listed 'other'. Please specify. |      Freq.     Percent        Cum.
---------------------------------------+-----------------------------------
                                 Creek |          5       10.20       10.20
                         Flowing water |          1        2.04       12.24
                Fresh water out of a stone. |          1        2.04       14.29
                             From tank |          1        2.04       16.33
                         From the tank |          2        4.08       20.41
                     Has own water tank |          1        2.04       22.45
      Personsal tap piped from the well. |          1        2.04       24.49
                  Piped from well water |          1        2.04       26.53
                   Rain water from tank |          2        4.08       30.61
                           Stone water |          2        4.08       34.69
Stone water. Water coming out of a st.. |          1        2.04       36.73
                                  Tank |          4        8.16       44.90
                            Tank Water |          3        6.12       51.02
                            Tank water |          9       18.37       69.39
                             Tankwater |          1        2.04       71.43
                            Tuffa tunk |          1        2.04       73.47
                     Under ground water |          1        2.04       75.51
                    Under ground water. |          1        2.04       77.55
                            Wall water |          1        2.04       79.59
                            Water tank |          7       14.29       93.88
                                  Well |          3        6.12      100.00
---------------------------------------+-----------------------------------
                                 Total |         49      100.00
```

iv. Question: How many households treat their water? *"tab i9" – 239 households (23.29%) treat their water*

v. Question: How many rooms do households have?

a) *"tab i3"* – shows us how many HHs have each option of number of rooms

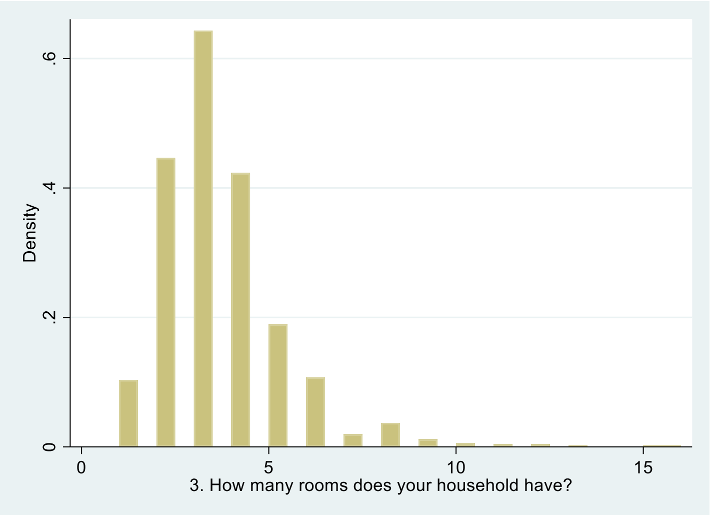
b) What would be more helpful to know? *Mean, median, etc.*

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 i3 (number of rooms in each household) by typing *"hist i3"*

Figure 1.8: Stata histogram output for "hist i3"



a) Question: What is the mode (most common answer)? *3 rooms*

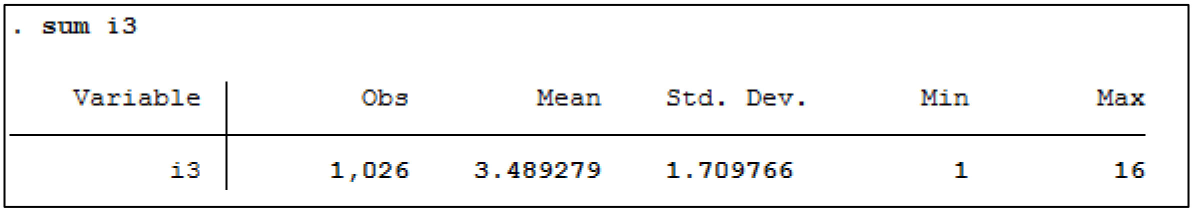
- This figure suggests that most houses have between 1-5 rooms, while very few have more than 5 (and none have 20+).

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

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

ii. <u>Practice:</u> What if we want to know the average number of rooms in households in our sample? *“sum i3”*

Figure 1.9: Stata command and output for “sum i3”



```
. sum i3

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          i3 |      1,026    3.489279    1.709766          1         16
```

iii. <u>Challenge:</u> 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? *3.49 rooms*

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

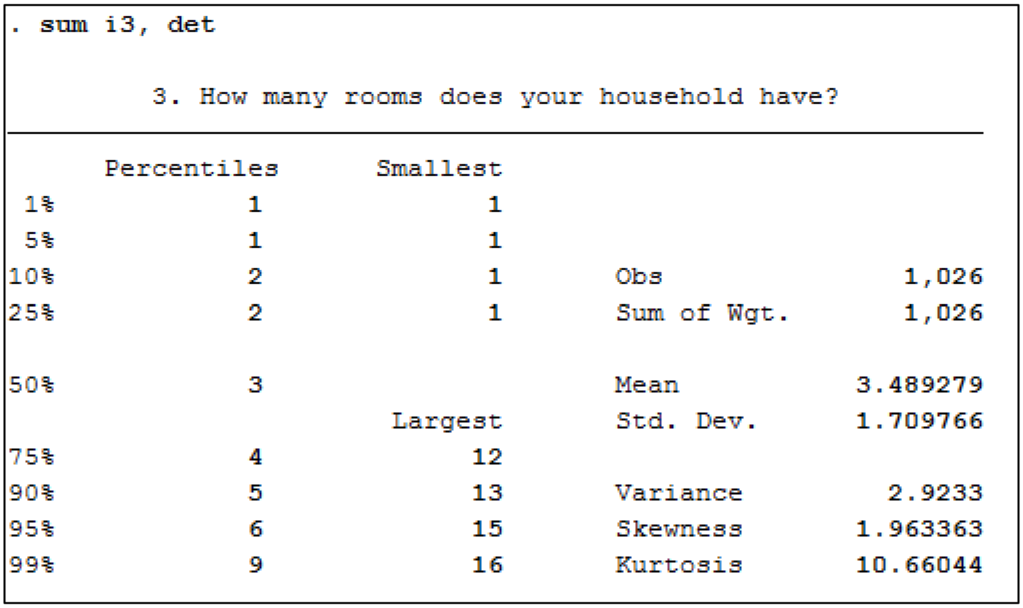
a) We know from the “tab” that very few HHs have a lot of rooms, and we know from this “sum” output that the maximum number is 16 even though the average is 3.49, so the mean is likely positively biased due to this skew.

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. <u>Practice:</u> What is the median number of rooms? “***sum**marize i3, **det**ail”* – this output shows the following:

Figure 1.10: Stata command and output for "sum i3, det"



```
. sum i3, det

         3. How many rooms does your household have?
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              1
 5%            1              1
10%            2              1       Obs               1,026
25%            2              1       Sum of Wgt.       1,026

50%            3                      Mean           3.489279
                        Largest       Std. Dev.      1.709766
75%            4             12
90%            5             13       Variance         2.9233
95%            6             15       Skewness       1.963363
99%            9             16       Kurtosis       10.66044
```

a) Percentiles and their ranges
b) The median (50%)
c) Variance, skewness, and kurtosis

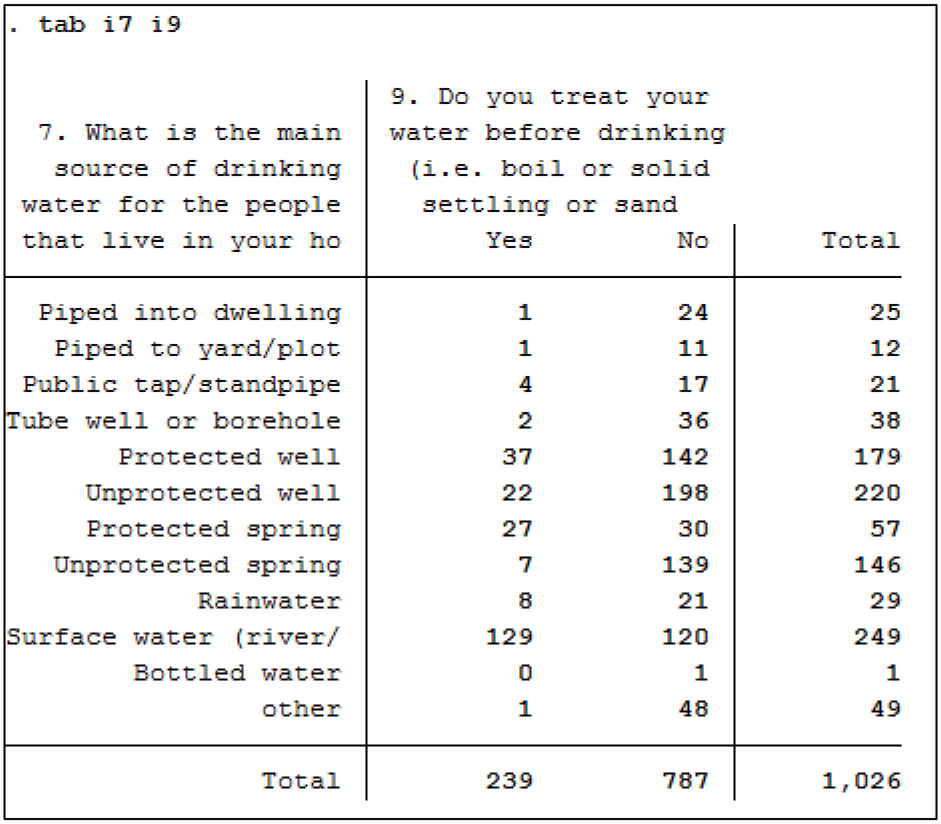
vii. What is the median number of rooms? *3 rooms*

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

i. Code: ***tab**ulate [varname1] [varname2]*

ii. Practice: Do households who get their water piped into their dwelling treat it before drinking? *"tab i7 i9"*

Figure 1.11: Stata command and output for “tab i7 i9”



```
. tab i7 i9

                      |  9. Do you treat your
  7. What is the main | water before drinking
   source of drinking |  (i.e. boil or solid
 water for the people |   settling or sand
 that live in your ho |       Yes         No |     Total
----------------------+----------------------+----------
  Piped into dwelling |         1         24 |        25
   Piped to yard/plot |         1         11 |        12
 Public tap/standpipe |         4         17 |        21
Tube well or borehole |         2         36 |        38
       Protected well |        37        142 |       179
     Unprotected well |        22        198 |       220
     Protected spring |        27         30 |        57
   Unprotected spring |         7        139 |       146
            Rainwater |         8         21 |        29
Surface water (river/ |       129        120 |       249
        Bottled water |         0          1 |         1
                other |         1         48 |        49
----------------------+----------------------+----------
                Total |       239        787 |     1,026
```

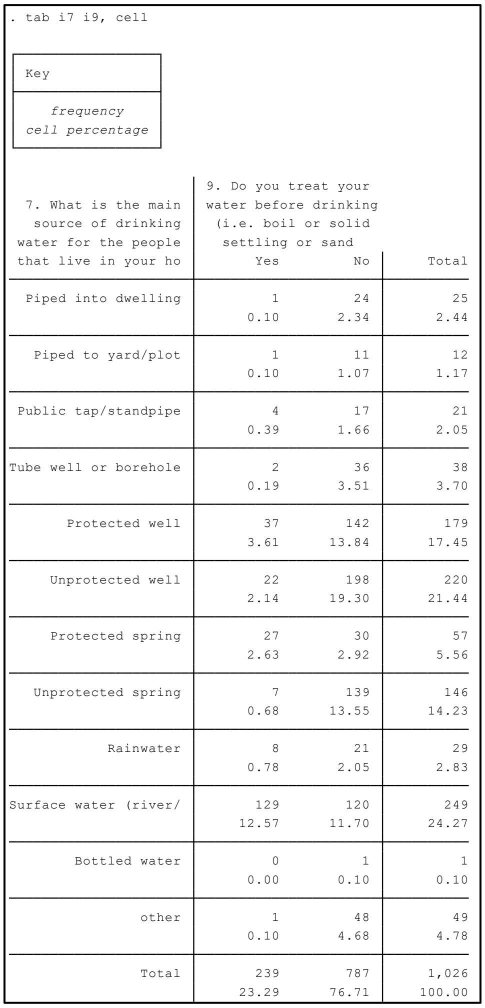
a) This output tells us the number of observations for each combination of responses between the two variables.

b) Question: How many households get their water piped into their dwelling, and do *not* treat it? *24 households*

iii. What if we want to know what *percent* of households get their water piped into their dwelling and do not treat it? - There are many *options* that we can add to the two-way tabulation command to get various types of percentages in the output.

iv. Practice: We want to know what percent of households **in the sample who** get their water piped into their dwelling and do not treat it – *“tab i7 i9, cell”*

Figure 1.12: Stata command and output for "tab i7 i9, cell"



```
. tab i7 i9, cell

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

                      |  9. Do you treat your
  7. What is the main | water before drinking
   source of drinking |  (i.e. boil or solid
 water for the people |   settling or sand
 that live in your ho |       Yes         No |     Total
----------------------+----------------------+----------
  Piped into dwelling |         1         24 |        25
                      |      0.10       2.34 |      2.44
----------------------+----------------------+----------
   Piped to yard/plot |         1         11 |        12
                      |      0.10       1.07 |      1.17
----------------------+----------------------+----------
 Public tap/standpipe |         4         17 |        21
                      |      0.39       1.66 |      2.05
----------------------+----------------------+----------
Tube well or borehole |         2         36 |        38
                      |      0.19       3.51 |      3.70
----------------------+----------------------+----------
       Protected well |        37        142 |       179
                      |      3.61      13.84 |     17.45
----------------------+----------------------+----------
     Unprotected well |        22        198 |       220
                      |      2.14      19.30 |     21.44
----------------------+----------------------+----------
     Protected spring |        27         30 |        57
                      |      2.63       2.92 |      5.56
----------------------+----------------------+----------
   Unprotected spring |         7        139 |       146
                      |      0.68      13.55 |     14.23
----------------------+----------------------+----------
            Rainwater |         8         21 |        29
                      |      0.78       2.05 |      2.83
----------------------+----------------------+----------
Surface water (river/ |       129        120 |       249
                      |     12.57      11.70 |     24.27
----------------------+----------------------+----------
        Bottled water |         0          1 |         1
                      |      0.00       0.10 |      0.10
----------------------+----------------------+----------
                other |         1         48 |        49
                      |      0.10       4.68 |      4.78
----------------------+----------------------+----------
                Total |       239        787 |     1,026
                      |     23.29      76.71 |    100.00
```

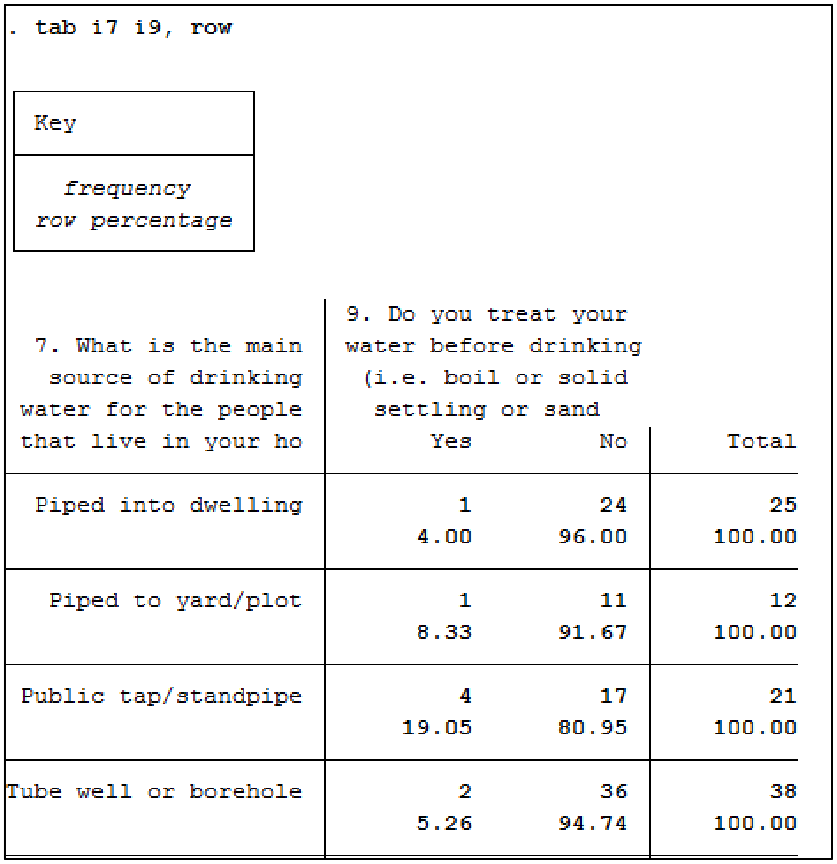
a) The option "*cell*" will tell us the percentage of households in the sample for each combination of responses.

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

c) What percentage of households **in the sample** get their water piped into their dwelling and do *not* treat it? *2.34% of the households* ***in the sample*** *get their water piped into their dwelling and do not treat it.*

v. <u>Practice:</u> What if we want to know the percentage of households **who get their water piped into their dwelling** do not treat it? – *"tab i7 i9, row"*

Figure 1.13: Stata command and output for "tab i7 i9, row"



```
. tab i7 i9, row

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

                      |  9. Do you treat your
 7. What is the main  | water before drinking
  source of drinking  |  (i.e. boil or solid
water for the people  |   settling or sand
that live in your ho  |       Yes         No |     Total
----------------------+----------------------+----------
 Piped into dwelling  |         1         24 |        25 
                      |      4.00      96.00 |    100.00 
----------------------+----------------------+----------
  Piped to yard/plot  |         1         11 |        12 
                      |      8.33      91.67 |    100.00 
----------------------+----------------------+----------
Public tap/standpipe  |         4         17 |        21 
                      |     19.05      80.95 |    100.00 
----------------------+----------------------+----------
Tube well or borehole |         2         36 |        38 
                      |      5.26      94.74 |    100.00 
----------------------+----------------------+----------
```

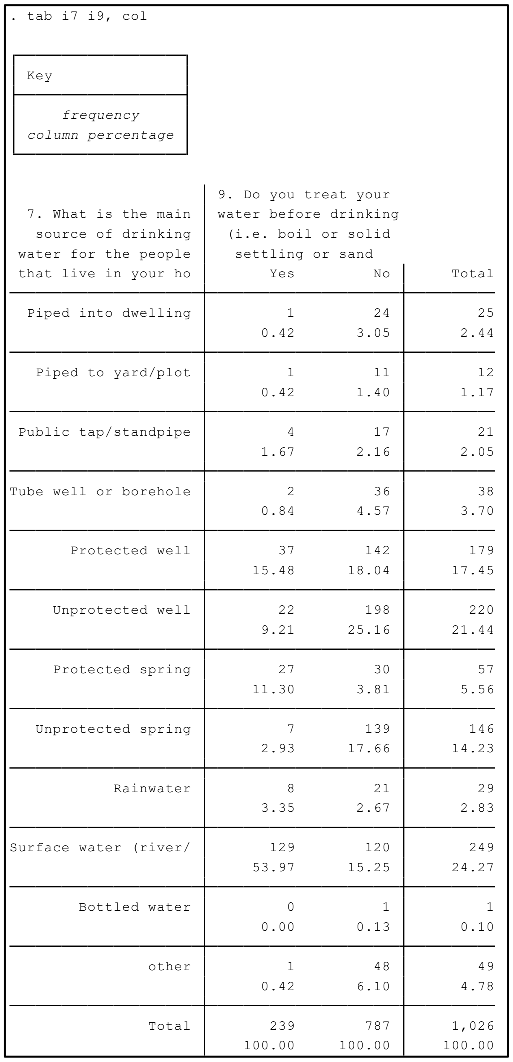
a) The option of "row" will tell us the row percentages – the percentages *in each row* will sum to 100.

b) In this specific code (where i7 is typed before i9), it will tell us: of the households **who receive their water from each source**, what percentage of them treat/don't treat their water.

c) What percentage of households **who pipe their water into their dwelling** do not treat their water? *96% of households **who pipe their water into their dwelling** do not treat their water.*

vi. Practice: What if we want to know of the households **who do not treat their water**, what percent get their water piped into their dwelling? – *"tab i7 i9, **col**umn"*

Figure 1.14: Stata command and output for “tab i7 i9, col”



```
. tab i7 i9, col

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

                      |  9. Do you treat your
  7. What is the main | water before drinking
   source of drinking |  (i.e. boil or solid
 water for the people |   settling or sand
 that live in your ho |       Yes         No |     Total
----------------------+----------------------+----------
  Piped into dwelling |         1         24 |        25 
                      |      0.42       3.05 |      2.44 
----------------------+----------------------+----------
    Piped to yard/plot |         1         11 |        12 
                      |      0.42       1.40 |      1.17 
----------------------+----------------------+----------
  Public tap/standpipe |         4         17 |        21 
                      |      1.67       2.16 |      2.05 
----------------------+----------------------+----------
Tube well or borehole |         2         36 |        38 
                      |      0.84       4.57 |      3.70 
----------------------+----------------------+----------
        Protected well |        37        142 |       179 
                      |     15.48      18.04 |     17.45 
----------------------+----------------------+----------
      Unprotected well |        22        198 |       220 
                      |      9.21      25.16 |     21.44 
----------------------+----------------------+----------
      Protected spring |        27         30 |        57 
                      |     11.30       3.81 |      5.56 
----------------------+----------------------+----------
    Unprotected spring |         7        139 |       146 
                      |      2.93      17.66 |     14.23 
----------------------+----------------------+----------
             Rainwater |         8         21 |        29 
                      |      3.35       2.67 |      2.83 
----------------------+----------------------+----------
Surface water (river/ |       129        120 |       249 
                      |     53.97      15.25 |     24.27 
----------------------+----------------------+----------
         Bottled water |         0          1 |         1 
                      |      0.00       0.13 |      0.10 
----------------------+----------------------+----------
                 other |         1         48 |        49 
                      |      0.42       6.10 |      4.78 
----------------------+----------------------+----------
                 Total |       239        787 |     1,026 
                      |    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 i7 is typed before i9), it will tell us: of the households who treat/don’t treat their water, what percentage of them get their water from various sources

c) What percentage of households **who do not treat their water** get their water piped into their dwelling? *3.05% of households **who do not treat their water** get their water piped into their dwelling.*

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 the question we just asked (What percentage of households who do not treat their water get their water piped into their dwelling?)

c. How might we write an expression to *tabulate* water source if the household does not treat their water?

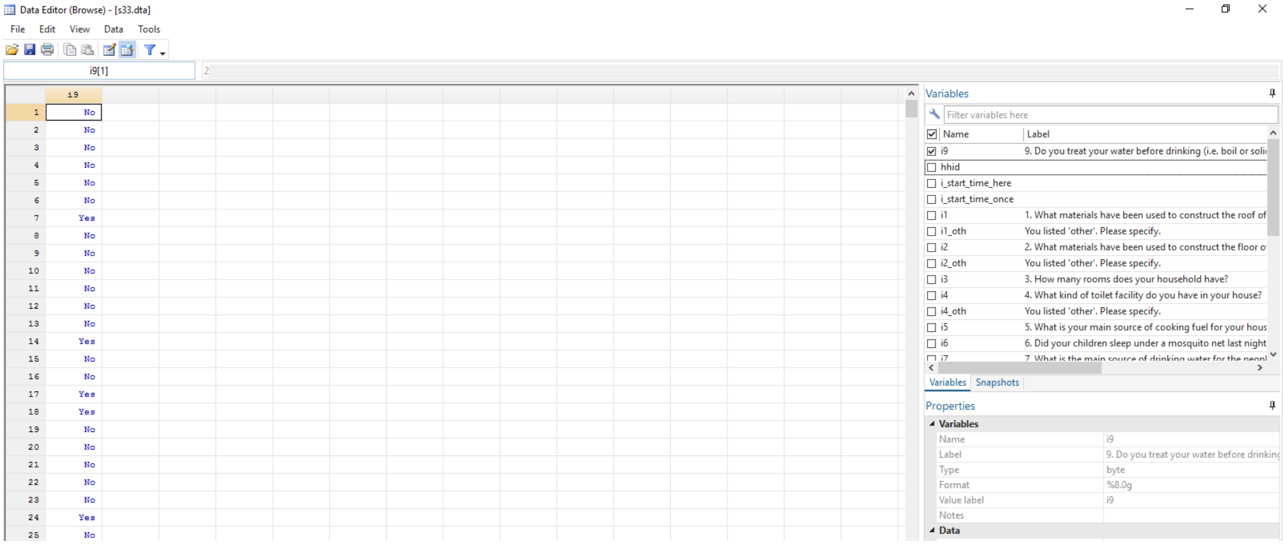
   i. First, we know that water treatment is a categorical variable (meaning "yes" and "no" are labels assigned to a number value), so we need to know what number means "no". There are three easy ways to check this:

      a) Practice: *"codebook i9"* – shows us that "no"' is the label for 2

      b) Practice: *"tab i9"* followed by "*tab i9, 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 i9"* – click on a cell that says "no" and look at what number is in the top bar (see Figure 1.15 below)

Figure 1.15: Stata data browser window for "browse i9"



Data Editor (Browse) - [s33.dta]

i9[1] 2

	i9
1	No
2	No
3	No
4	No
5	No
6	No
7	Yes
8	No
9	No
10	No
11	No
12	No
13	No
14	Yes
15	No
16	No
17	Yes
18	Yes
19	No
20	No
21	No
22	No
23	No
24	Yes
25	No

Variables

Filter variables here

☑ Name	Label
☑ i9	9. Do you treat your water before drinking (i.e. boil or soli
☐ hhid	
☐ i_start_time_here	
☐ i_start_time_once	
☐ i1	1. What materials have been used to construct the roof of
☐ i1_oth	You listed 'other'. Please specify.
☐ i2	2. What materials have been used to construct the floor o
☐ i2_oth	You listed 'other'. Please specify.
☐ i3	3. How many rooms does your household have?
☐ i4	4. What kind of toilet facility do you have in your house?
☐ i4_oth	You listed 'other'. Please specify.
☐ i5	5. What is your main source of cooking fuel for your hous
☐ i6	6. Did your children sleep under a mosquito net last night
☐ i7	7. What is the main source of drinking water for the peopl

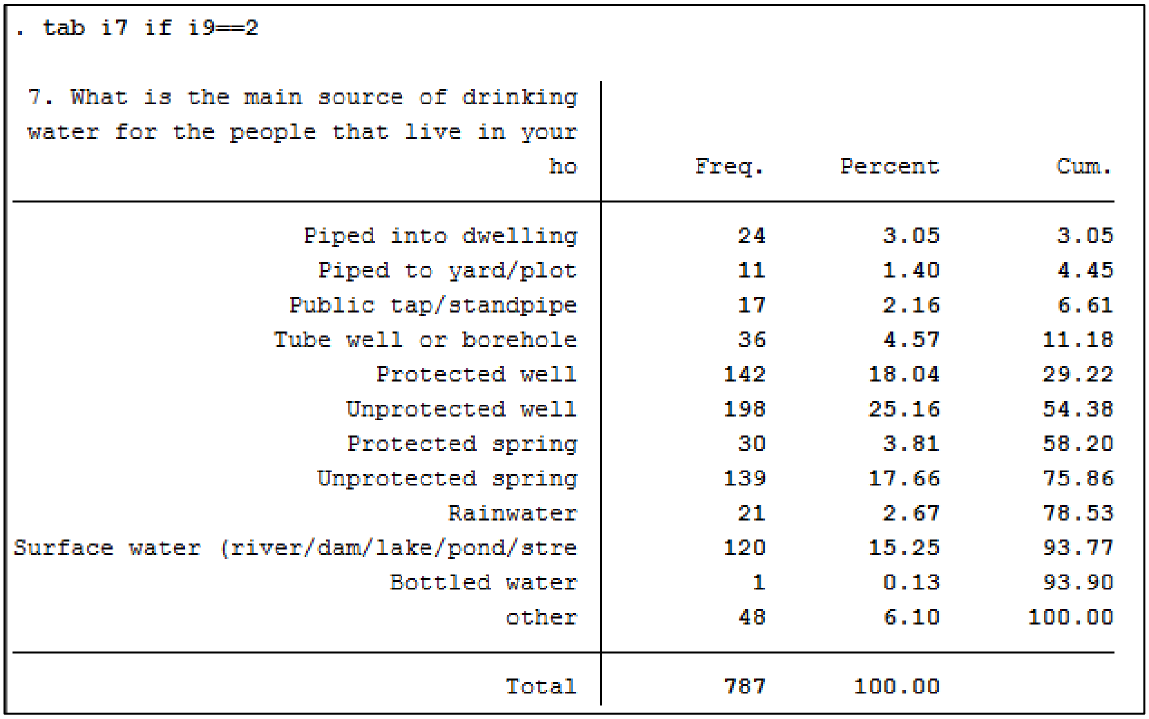
Properties

Variables	
Name	i9
Label	9. Do you treat your water before drinking
Type	byte
Format	%8.0g
Value label	i9
Notes	

Data

   ii. Question: Now that we know that i9 equals 2 for "no", how can we write the tabulation command using "if" to answer the question? (What percentage of households who do not treat their water get their water piped into their dwelling?) - *"tab i7 if i9==2"*

Figure 1.16: Stata command and output for "tab i7 if i9==2"



```
. tab i7 if i9==2

  7. What is the main source of drinking |
  water for the people that live in your |
                                      ho |      Freq.     Percent        Cum.
-----------------------------------------+-----------------------------------
                    Piped into dwelling |         24        3.05        3.05
                     Piped to yard/plot |         11        1.40        4.45
                   Public tap/standpipe |         17        2.16        6.61
                  Tube well or borehole |         36        4.57       11.18
                         Protected well |        142       18.04       29.22
                       Unprotected well |        198       25.16       54.38
                       Protected spring |         30        3.81       58.20
                     Unprotected spring |        139       17.66       75.86
                              Rainwater |         21        2.67       78.53
Surface water (river/dam/lake/pond/stre |        120       15.25       93.77
                          Bottled water |          1        0.13       93.90
                                  other |         48        6.10      100.00
-----------------------------------------+-----------------------------------
                                  Total |        787      100.00
```

a) Now we see this single tabulation of water source *only* for the households that do not treat their water. Again, we see that of the households who do not treat their water, 3.05% get their water piped into their dwelling.

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 787 households while our dataset has 1,026). This is because 787 households do *not* treat their water, 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 roof material.

i. Question: What variable tells us the number of rooms in a household? *i3*

ii. Question: What variable tells us the type of roof? *i1*

iii. Question: How could we see the type of roof for households that have *more than 3 rooms*? *"tab i1 if i3>3" OR "tab i1 if i3>=4"*

iv. Question: How about the type of roof for households with *exactly 3 rooms*? *"tab i1 if i3==3"*

v. Question: How about the type of roof for households with *3-5 rooms*? *"tab i1 if i3>=3 & i3<=5" OR "tab i1 if i3>2 & i3<6"*

vi. Question: How about the type of roof for households with l*ess than 3 rooms or greater than 5 rooms*? *"tab i1 if i3<3 | i3>5" OR "tab i1 if i3<=2 | i3>=6"*

vii. Question: Which households more frequently have a corrugated metal roof? Households with more than 3 rooms? Or households with 3 rooms or less? What codes will you run to show this?

a) *“tab i1 if i3>3” and “tab i1 if i3<=3”*

b) *28.74% of households with more than 3 rooms have a corrugated metal roof, but only 7.68% of households with 3 rooms or less have a corrugated metal roof. So households with more rooms are more likely to have a corrugated metal roof.*

Lesson 2 – Transforming Data

1. Review of Lesson 1 – Describing the new dataset (s11_long – 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\s11_long.dta"
(Module 1.1 Household demographics, current household members [long] - Clean)

. count
  6,057
```

   i. How many observations are in this dataset? *6,057*

   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? *a3 – "3. What is the sex of ${a1}?"*

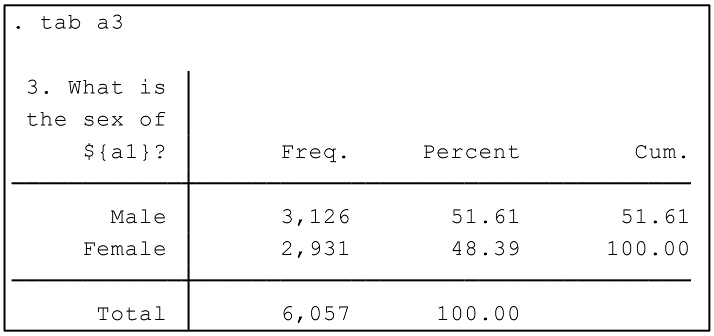
      - The part of the label "${a1}" means that when respondents were asked this question, the tablets filled in the response to the question *"a1"* here. Question *"a1"* asked the name of the household member. (This information has been removed from this dataset to keep respondents anonymous).

   ii. Question: What code(s) can we use to see the number of men vs women in the dataset?

      a) *tabulate ("tab a3")*

b) *codebook (“codebook a3”)* 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 a3”



```
. tab a3

 3. What is |
 the sex of |
    ${a1}? |      Freq.     Percent        Cum.
------------+-----------------------------------
       Male |      3,126       51.61       51.61
     Female |      2,931       48.39      100.00
------------+-----------------------------------
      Total |      6,057      100.00
```

iii. Question: How many males are in this dataset? *3,126*

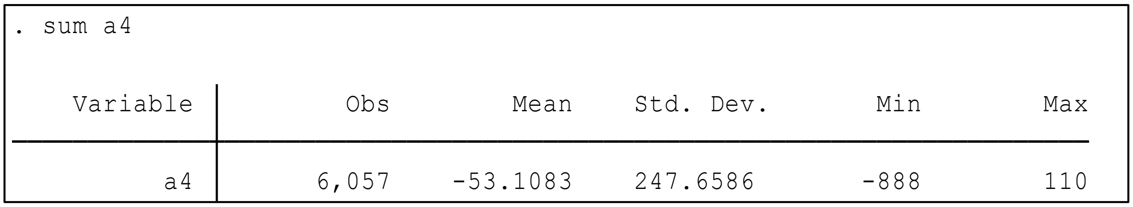
iv. Question: What percentage of the people in this dataset are males? *51.61%*

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? *a4 – “4. What is the age of ${a1}?”*

ii. Question: What code can we use to see the average age of all people in this dataset? *Summarize (“sum a4”)*

Figure 2.3: Stata command and output for “sum a4”



```
. sum a4

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          a4 |      6,057    -53.1083    247.6586       -888        110
```

iii. What is the average age of people in this dataset? *-53! Not a real age!*

iv. Look at all of the output from “sum a4” – where is the problem? *Minimum age is a impossibly large and negative number, -888.*

v. -888 is what data collectors were told to input if the respondent did not know the age of an individual. We will have to clean/transform this data before we can find the average age!

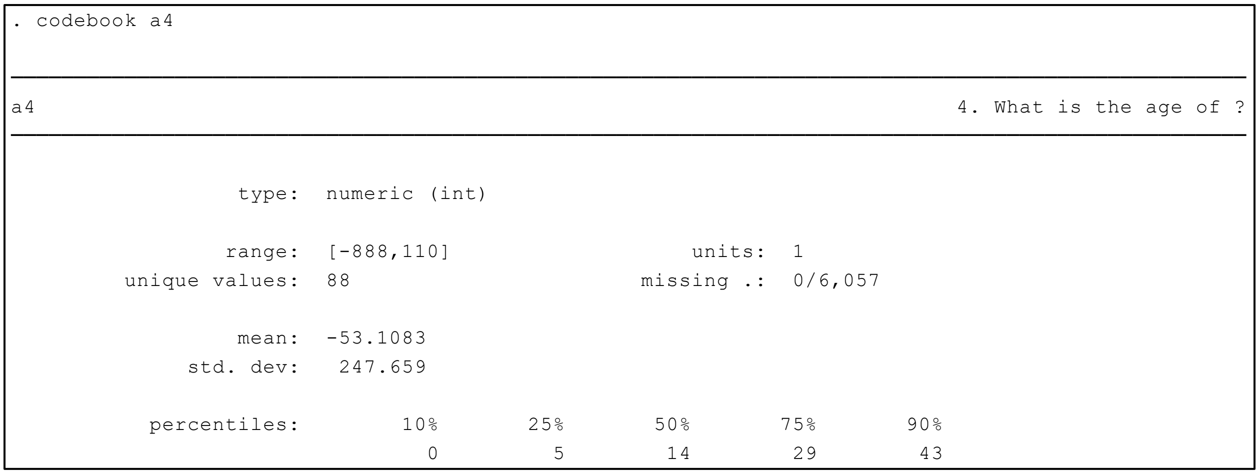
2. Transforming Existing Data

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 (a4) has any missing values currently? *“codebook a4”*

Figure 2.4: Stata command and output for "codebook a4"



```
. codebook a4

-----------------------------------------------------------------------------------------------------
a4                                                                         4. What is the age of ?
-----------------------------------------------------------------------------------------------------

                  type:  numeric (int)

                 range:  [-888,110]                     units:  1
         unique values:  88                         missing .:  0/6,057

                  mean:  -53.1083
             std. dev:    247.659

           percentiles:        10%       25%       50%       75%       90%
                                 0         5        14        29        43
```

b) How many missing values are there currently in a4? *0 missing values*

c) Now we need to change the -888 values 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 a4? *"codebook a4" – numeric*

d) Question: So what type of missing value do we want to change the -888 to? *A period (.)*

b. Recoding values – there are two ways that we can change these -888 values.

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

- Practice: *"recode a4 -888=."*

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

- Practice: *"replace a4=. if a4==-888"*

iii. Question: After using one of these codes, how many missing values are there now for a4? *"codebook a4"; 488 missing values*

iv. Now that we know that (.) means missing, we can also type "*tab a4, **m**issing*" to see how many missing values we have. This tabulate option *"**m**issing"* includes the missing values in the tabulation

Figure 2.5: Stata command and output for “codebook a4”, after recoding -888 values to missing

```
. codebook a4

------------------------------------------------------------------------------------------------
a4                                                                         4. What is the age of ?
------------------------------------------------------------------------------------------------

                  type:  numeric (int)

                 range:  [0,110]                      units:  1
         unique values:  87                       missing .:  488/6,057

                  mean:   20.0515
              std. dev:   16.2954

           percentiles:        10%       25%       50%       75%       90%
                                 3         7        16        30        44
```

v. Question: Now what is the average age in the dataset?

  a) *“sum a4”*

Figure 2.6: Stata command and output for “sum a4”, after recoding -888 values to missing

```
. sum a4

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          a4 |      5,569    20.05154    16.29536          0        110
```

  b) *Now the average (mean) age is 20.05. This average does not take into account the 488 missing values (Obs is now 5,569 instead of 6,057)*

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 a4 age”* – renames the variable to “age” (more intuitive than “a4”)

  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

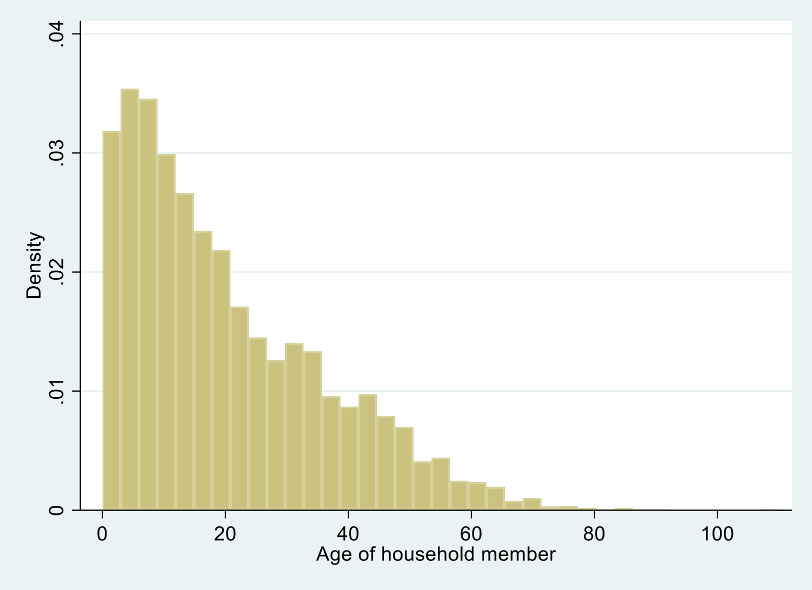
      - The label has to be in quotations
      - Remember, now the variable is named “age” instead of “a4”, so we have to use its new name when writing commands

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 25 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]*

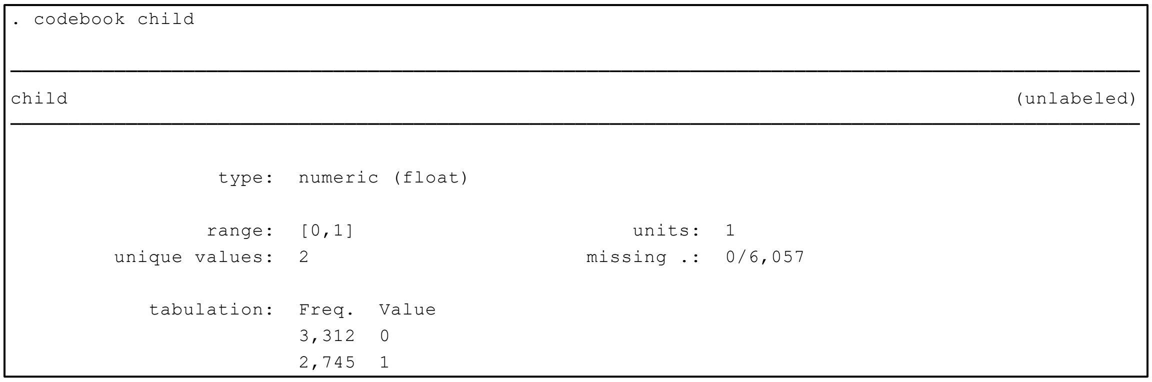
   ii. Practice: *"**gen**erate child=."* – creates the new variable named "child", and makes all observations missing.

   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 .:  0/6,057

            tabulation:  Freq.  Value
                         3,312  0
                         2,745  1
```

vi. How many missing values are there for our “child” variable? *0 – this is a problem! We changed the ages of 488 people from -888 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?

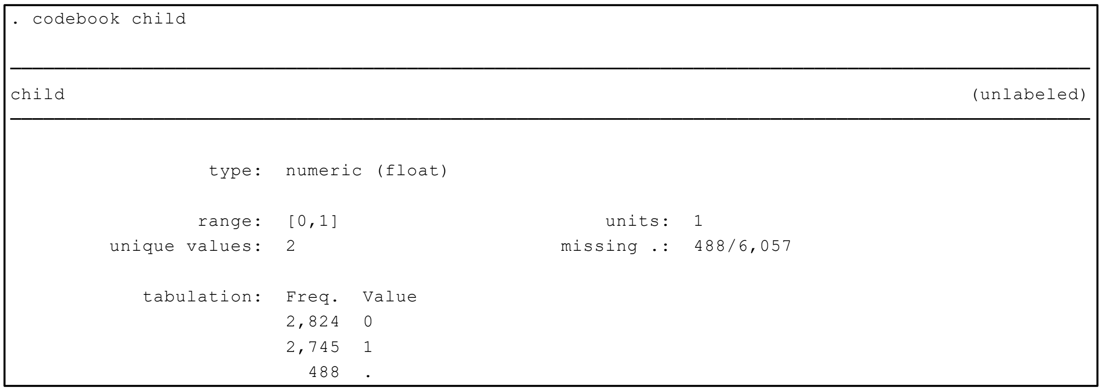
a) *“replace child=. if age==.”*

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 .:  488/6,057

            tabulation:  Freq.  Value
                         2,824  0
                         2,745  1
                           488  .
```

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.

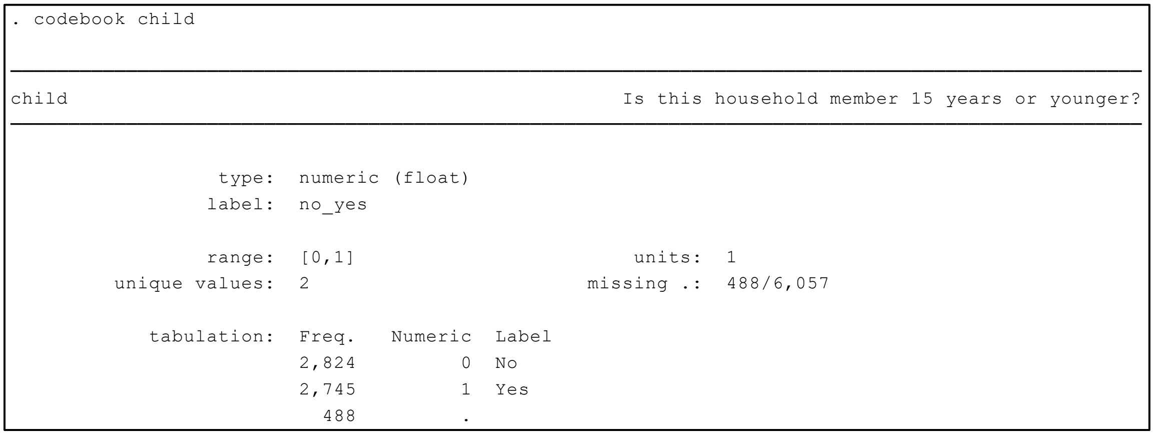
iv. Now, we have to apply our new value label definition to our variable

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



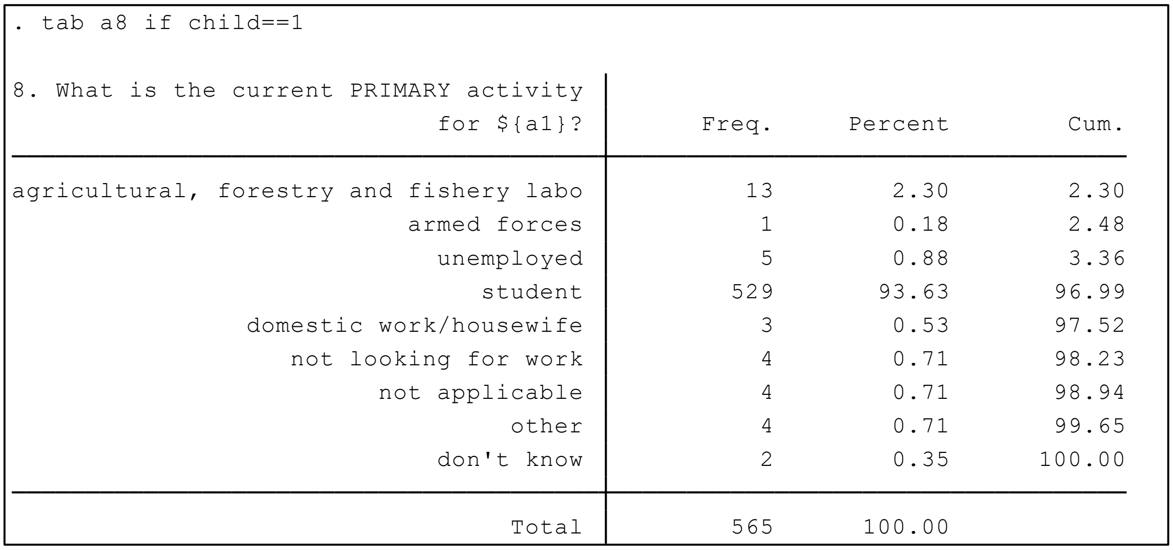
```
. codebook child

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

                  type:  numeric (float)
                 label:  no_yes

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  488/6,057

            tabulation:  Freq.   Numeric  Label
                         2,824         0  No
                         2,745         1  Yes
                           488         .
```

vi. How many children are in our dataset? *2,745*

vii. <u>Challenge:</u> What percent of children have “student” listed as their primary activity?

a) *“tab a8 if child==1”*

Figure 2.11: Stata command and output for “tab 8 if child==1”



```
. tab a8 if child==1

8. What is the current PRIMARY activity |
                               for ${a1}? |      Freq.     Percent        Cum.
------------------------------------------+-----------------------------------
agricultural, forestry and fishery labo  |         13        2.30        2.30
                            armed forces |          1        0.18        2.48
                              unemployed |          5        0.88        3.36
                                 student |        529       93.63       96.99
                  domestic work/housewife |          3        0.53       97.52
                     not looking for work |          4        0.71       98.23
                           not applicable |          4        0.71       98.94
                                   other |          4        0.71       99.65
                               don't know |          2        0.35      100.00
------------------------------------------+-----------------------------------
                                   Total |        565      100.00
```

b) *93.63% of children 15 and under have “student” listed as their primary activity*

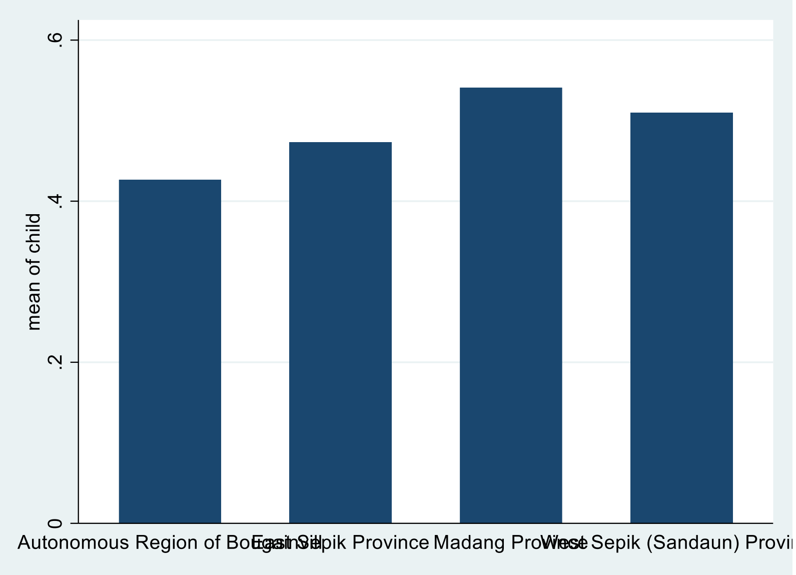
viii. <u>Challenge:</u> Which province has the highest percentage of children? Let’s use a bar graph to find out!

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) Practice: “*graph bar child, over(prov)*”

Figure 2.12: Stata bar graph output for “graph bar child, over(prov)”



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%).

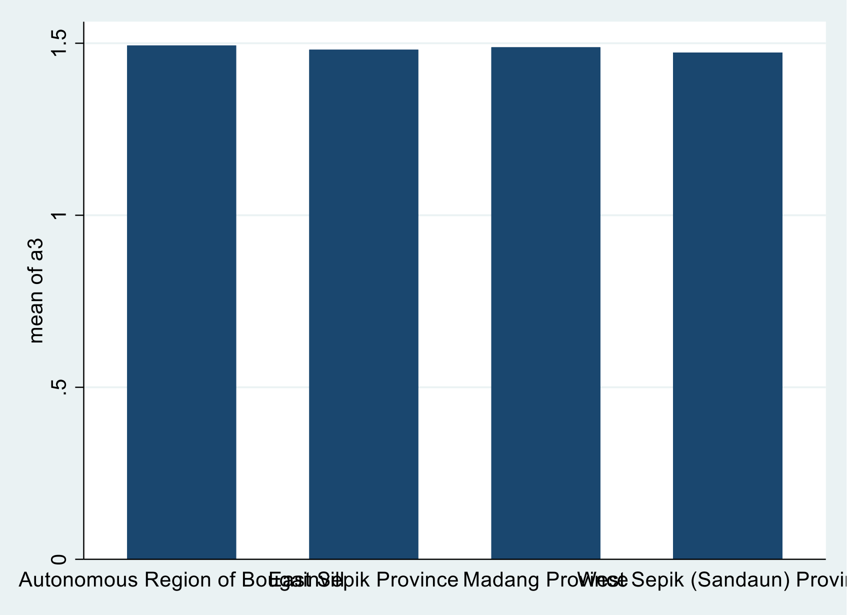
- Similarly, we can also find the prevalence of a dummy variable by calculating its average. Which command tells us the average? *sum*
- Note: The figures section of the manual discusses how to add and format labels to avoid overlap

d) *Madang has 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 a3, over(prov)*”

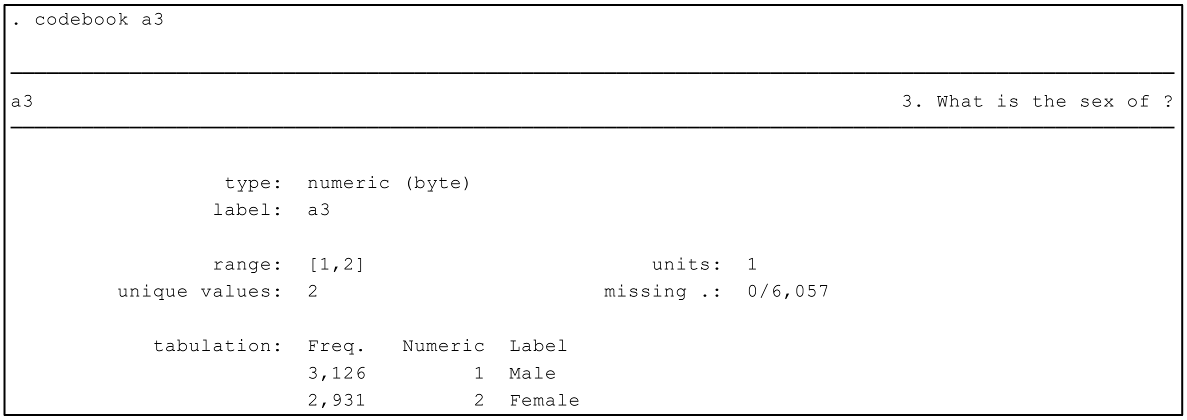
Figure 2.13: Stata bar graph output for “graph bar a3, over(prov)”



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 a3”*

Figure 2.14: Stata command and output for “codebook a3”



```
. codebook a3

-------------------------------------------------------------------------------------------------
a3                                                                  3. What is the sex of ?
-------------------------------------------------------------------------------------------------

                 type:  numeric (byte)
                label:  a3

                range:  [1,2]                        units:  1
        unique values:  2                        missing .:  0/6,057

           tabulation:  Freq.   Numeric  Label
                        3,126         1  Male
                        2,931         2  Female
```

d) The value labels are 1 and 2, instead of 0 and 1 – a3 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 a3==2”* – changes all values of the *woman* variable to 1 if the household member is a woman (coded as 2 in the variable a3)

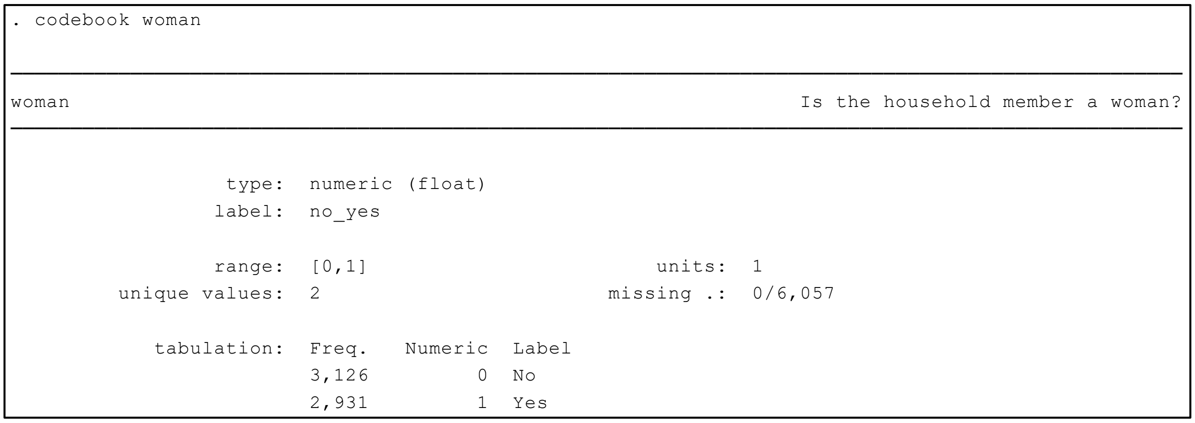
c) “*replace woman=0 if a3==1”* – changes all values of the *woman* variable to 0 if the household member is a man (coded as 1 in the variable a3)

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"*

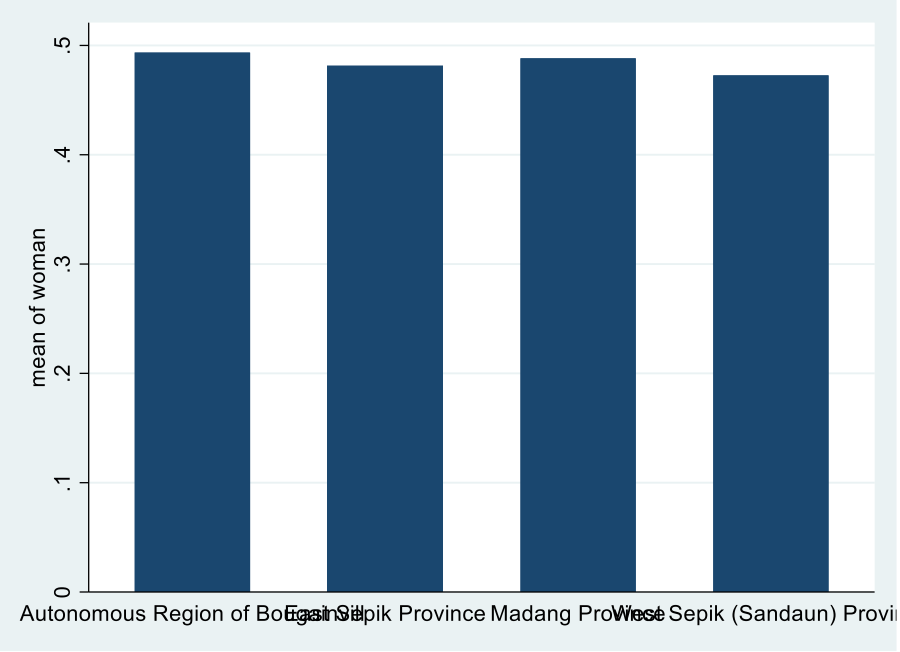
Figure 2.15: Stata command and output for "codebook woman"



```
. codebook woman

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


                  type:  numeric (float)
                 label:  no_yes

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/6,057

            tabulation:  Freq.   Numeric  Label
                         3,126         0  No
                         2,931         1  Yes
```

g) Let's look at the bar graph again with our new variable instead of a3 – *"graph bar woman, over(prov)"*

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



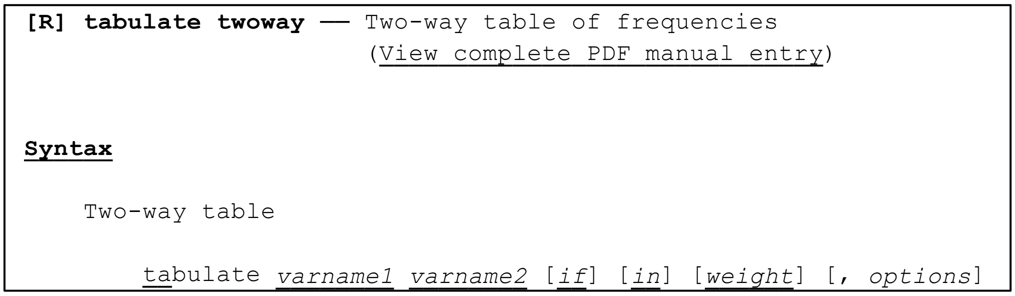
h) <u>Question:</u> Does it show prevalence now? *Yes!*

i) <u>Question:</u> Which province has the highest prevalence of women in the sample? *ARoB*

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

a. Try *"help tabulate twoway"*

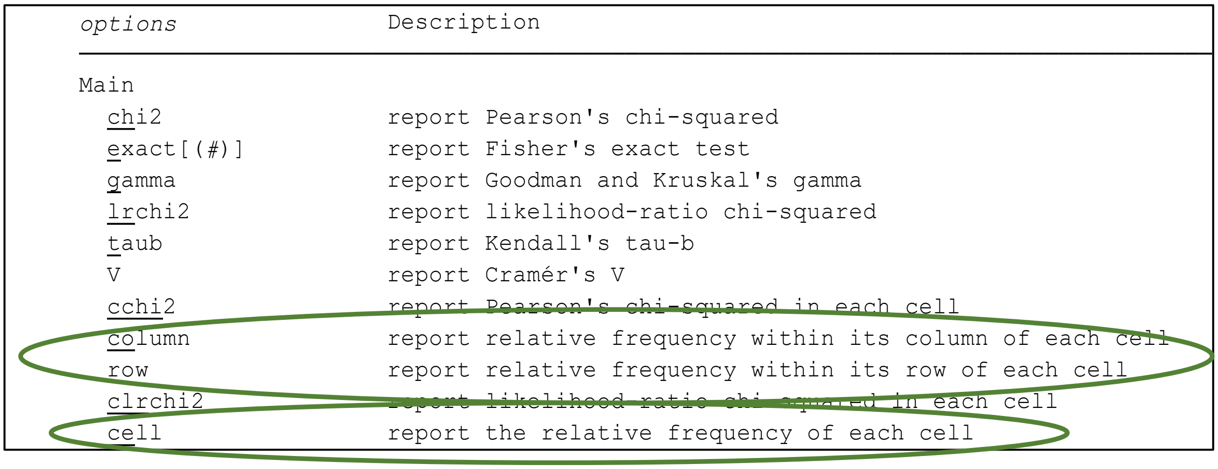
Figure 2.17: Stata help window for *tabulate twoway* – syntax



```
[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

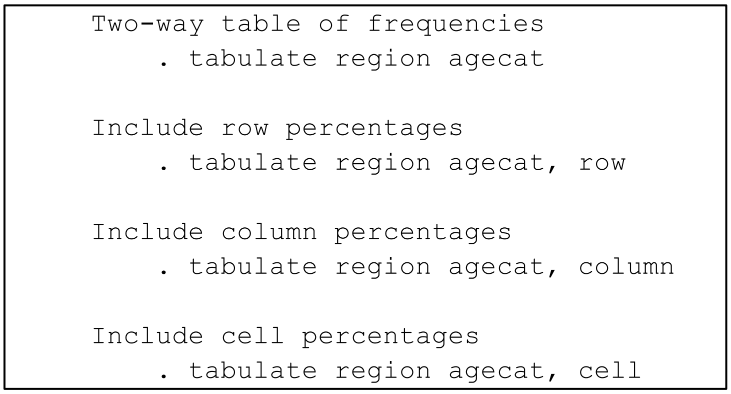
Figure 2.18: Stata help window for *tabulate twoway* – options



```
    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 2.19: Stata help window for *tabulate twoway* – examples

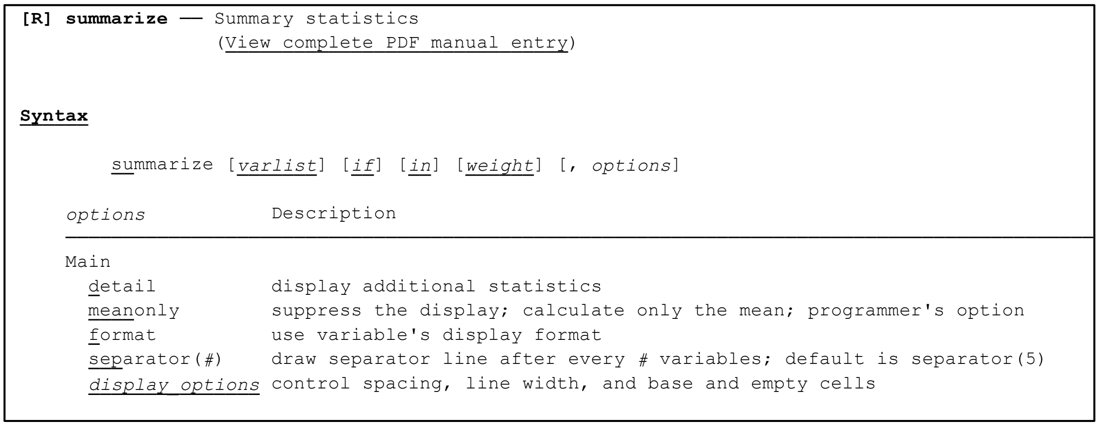


```
    Two-way table of frequencies
        . tabulate region agecat

    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 2.20: 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, and the percent of women and children in each household.

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) a_pid*

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

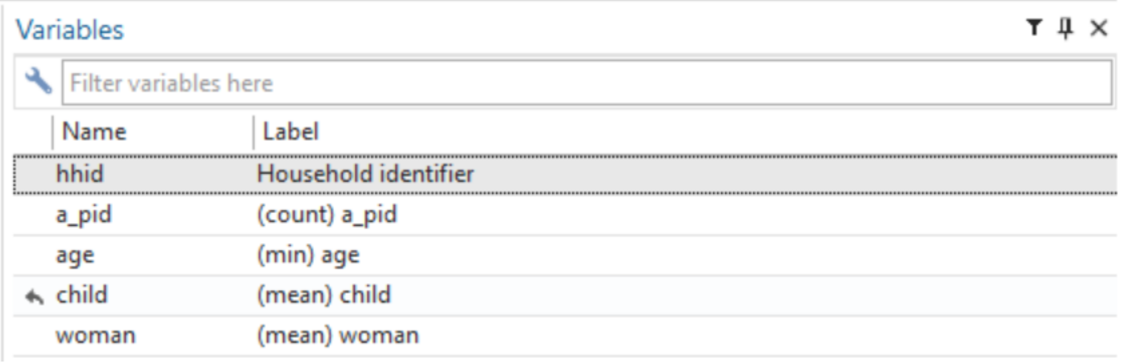
iv. Which summary statistic and variables would we use to find the percent of children and the percent of women in each household? *(mean) child woman – "mean" works for these two because they are dummy variables (values of 0="No" and 1="Yes")*

v. <u>Practice:</u> *"collapse (count) a_pid (min) age (mean) child woman, by(hhid)"*

vi. How many observations are there now in the dataset? *"count" – 1,026*

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

Figure 2.21: Variable window after collapsing a_pid, age, child, and woman by household



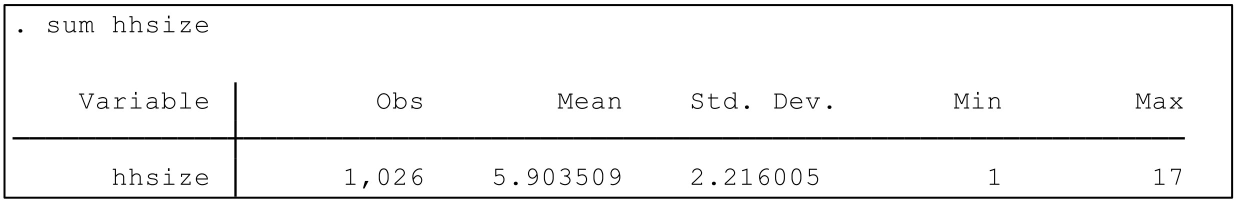
Name	Label
hhid	Household identifier
a_pid	(count) a_pid
age	(min) age
child	(mean) child
woman	(mean) woman

        a) Variable names remained the same
        b) Variable labels now show only the summary statistic and the name of the variable

    viii. <u>Practice:</u> Let’s create more meaningful variable names and labels
        a) What does the variable *a_pid* tell us now?
            - The count of people in the household = the household size
            - *“ren a_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””*

    ix. <u>Question:</u> What is the average household size?
        a) *“sum hhsize”*

Figure 2.22: Stata command and output for “sum hhsize”



```
. sum hhsize

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      hhsize |      1,026    5.903509    2.216005          1         17
```

        b) *5.9 people per household*

    x. <u>Question:</u> What is the average age of the youngest household member?
        a) *“sum age_youngest”*
        b) *5.8 years old*

    xi. <u>Question:</u> What is the average prevalence of *children* in the households?
        a) *“sum perc_children”*
        b) *48% children*

    xii. <u>Question:</u> What is the average prevalence of *women* in the households?
        a) *“sum perc_women”*
        b) *49% women*

c. Merge Datasets

i. What if we want to know if household size is associated with the type of roof that a household has?

   a) We have household size in this dataset, but roof type 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 (1,026)

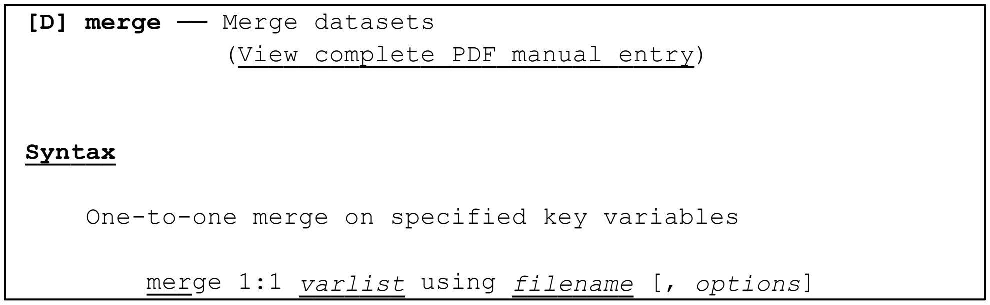
ii. Merging datasets with the same number of observations (the observations across the datasets represent the same levels of data: e.g. both datasets are at the household level)

Figure 2.23: Example of a 1:1 merge, shown with only one household identifier (hhid)

s11_long (collapsed; n=1,026)			**s33 (n=1,026)**			**merged data (n=-1,206)**		
hhid	age_youngest		hhid	i1		hhid	age_youngest	i1
0504_0940_UWMUO	2		0504_0940_UWMUO	metal roof		0504_0940_UWMUO	2	metal roof

   a) Practice: *"help merge"*

   Figure 2.24: Stata help window for *merge*



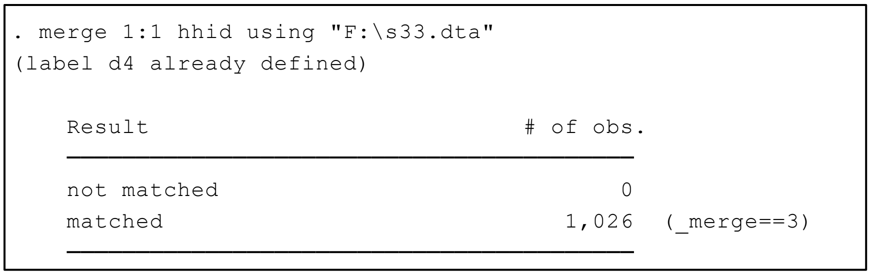
```
[D] merge —— Merge datasets
             (View complete PDF manual entry)


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 1,026 unique households and we are merging it to another dataset with 1,026 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:\s33.dta"*

   Figure 2.25: Stata command and output for "merge 1:1 hhid using …"



```
. merge 1:1 hhid using "F:\s33.dta"
(label d4 already defined)

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                             1,026  (_merge==3)
    -----------------------------------------
```

   d) The output after the merge tells us how many observations were and were not matched. How many were matched? *1,026 (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 1,027 observations in the collapsed household roster, and merged it to 1,026 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 1,026 observations in the collapse household roster, and merged it to 1,027 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

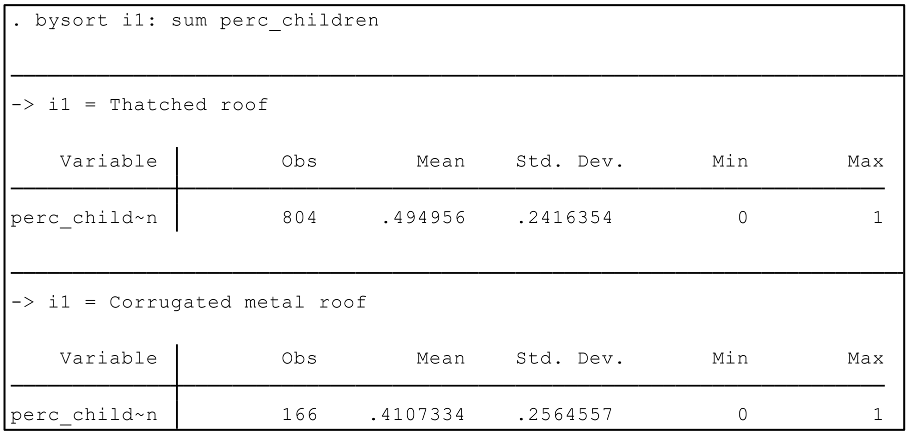
f) If you wanted to merge multiple datasets together, you will have to drop this new _merge variable, otherwise another merge will not work because Stata will tell you that the variable _merge is already defined.

- Practice: *"drop _merge"*

g) Challenge: On average, do households with a thatched roof or with a corrugated metal roof have more children?

- Option 1: two sum…if codes
  - *"sum child if i1==1"* – on average in households with thatched roofs, 49% of the household members are aged 15 years or younger
  - *"sum child if i1==2"* – on average in households with corrugated metal roofs, 41% of the household members are aged 15 years or younger
  - Descriptives suggest that households with thatched roofs have a higher percentage of children than households with corrugated metal roofs.
- Option 2: bysort. "bysort" repeats a Stata command on a subset of the data. We can repeat the "summarize child" command, on the different values of i1.
  - Code: ***bys**ort [varname1]: stata_command [varname2]*
  - *"**bys**ort i1: sum child"*

Figure 2.26: Stata command and output for "bysort i1: sum perc_children" (abridged)



```
. bysort i1: sum perc_children

-------------------------------------------------------------------------------
-> i1 = Thatched roof

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
perc_child~n |        804     .494956    .2416354          0          1

-------------------------------------------------------------------------------
-> i1 = Corrugated metal roof

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
perc_child~n |        166    .4107334    .2564557          0          1
```

        - Shows the same results as Option 1, but only requires one code

iii. What if we want to know how many people in our dataset live in a house with a thatched roof?

   a) We would want to look at the roof type 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 (6,057 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 2.27: Example of a 1:m merge, shown with only one household identifier (hhid)

s33 (n=1,026)		s11_long (n=6,057)			merged data (n=6,057)			
hhid	i1	hhid	a_pid	a4	hhid	a_pid	a4	i1
0504_0940_UWMUO	metal roof	0504_0940_UWMUO	1	M	0504_0940_UWMUO	1	M	metal roof
		0504_0940_UWMUO	2	F	0504_0940_UWMUO	2	F	metal roof
		0504_0940_UWMUO	3	M	0504_0940_UWMUO	3	M	metal roof
		0504_0940_UWMUO	4	M	0504_0940_UWMUO	4	M	metal roof
		0504_0940_UWMUO	5	M	0504_0940_UWMUO	5	M	metal roof

   a) Let’s now start over and open the household characteristics dataset (s33.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 (s11_long). 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 s33 data to many (m) household observations in the s11_long data set.*

   d) Practice: *“merge 1:m hhid using "F:\s11_long.dta"”*

Figure 2.28: Stata command and output for “merge 1:m hhid using …”

```
. merge 1:m hhid using "F:\s11_long.dta"
(label d4 already defined)

    Result                           # of obs.
    -----------------------------------------
    not matched                              0
    matched                              6,057  (_merge==3)
    -----------------------------------------
```

      - Did all of the observations match? *Yes, all 6,057 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 live in a house with a thatched roof?

    - *"tab i1"* – 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* with a thatched roof
    - *4,828 people (79.71% of people in the sample live in houses with a thatched roof)*

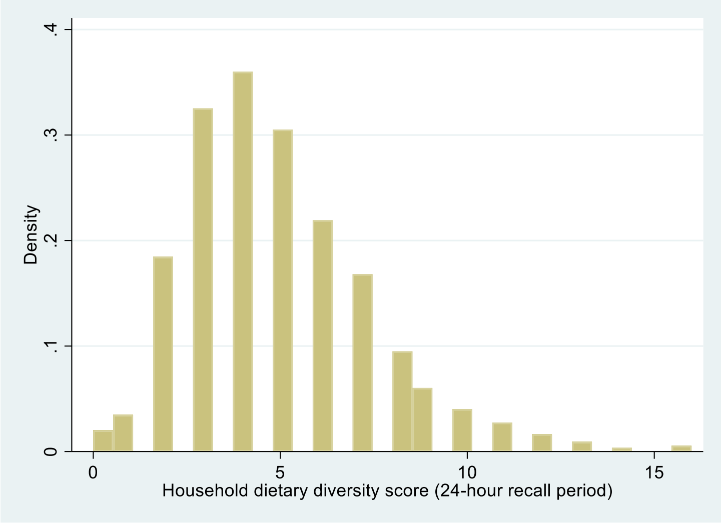
f) Challenge: How many *women* in our dataset live in a house with a thatched roof? (Two ways):

    - Tab…if
        - First we need to remember how women is coded in the gender variable. How can we check this? *"codebook a3"; 2 = female*
        - Now we can run the tab…if. What would it look like? *"tab i1 if a3==2"*
    - Twoway tabulation:
        - *"tab i1 a3, col"*
        - *"tab a3 i1, row"*
    - *2,339 women live in a house with a thatched roof (79.8% of women in the sample live in a house with a thatched roof)*

Lesson 3 – Analyzing Data

1. **Review of Lessons 1 and 2 – Describing and transforming the new dataset ("lesson3" – compilation of key information from many modules)**
    a. The datafile we are working with in Lesson 3 is a compilation of key household information from many different survey modules.
    b. Question: How can we see how many observations we have? *"count"*
        i. How many observations are in this dataset? *1,026*
        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: How many households have a female household head? *"tab hhh_female" – 104 households (10.14%)*
        ii. Question: We looked at roof and floor in Lesson 1. These variables have been cleaned more than when we first looked at them. How many categories are there now in the roof variable? *"codebook roof" – 3 unique values or categories (thatched, metal, or other)*
        iii. Question: What does the floor variable tell us now? What are the different values? "*codebook floor_wood" – tells us whether or not a household has a wood floor (no or yes); most households (963) have a wood floor*
        iv. Household Dietary Diversity is going to be our key outcome variable today. This variable was constructed in Module 5.4 in the survey – let's look at that module.
            1. Households were asked whether they ate any foods in 16 different food groups in the day before ("yesterday").
            2. The household dietary diversity score (HDDS) is calculated by counting how many different food groups were consumed by the household the day before. A higher score indicates higher household dietary diversity.
            3. What do you think the minimum and maximum values are for this variable? *0 and 16*
            4. Question: How can you check the minimum, maximum, and average values for this variable?
                a. *"codebook hdds"*
                b. *"sum hdds"*
                c. *The minimum is 0, the maximum is 16, and the mean is 4.96*
                d. *This means that on average, households consumed 4.96 different food groups the day before they were surveyed*
            5. Question: How can we look at a figure showing the distribution of this variable?
                a. "*hist hdds"*

Figure 3.1: Stata histogram output for "hist hdds"



b. What can we learn from this histogram? *The most frequent HDDS scores are between 2-6. The variable is positively skewed*

d. We have some continuous data in this dataset: for example, landholdings, household size (hhsize) and total livestock unit (TLU)
   i. Household size – we calculated this variable in Lesson 2 (it is the count of the number of people from the roster in each household)
   ii. Total livestock unit (TLU) – this variable was constructed based on the livestock module. Each type of livestock was given a value (e.g. larger livestock like cows were given larger values and smaller livestock like chicken were given smaller values). The values were then summed together in order to create a "livestock equivalent" that allows us to compare livestock ownership across all households, regardless of the types of livestock they own. Households with a higher TLU have a higher livestock equivalent. For example, a household with 2 cows will have a higher livestock equivalent than a household with 2 chickens.
   iii. Landholdings – this is the **raw data** (reported directly from the respondents), summing up all of their agricultural plot land areas. Frequently, raw continuous data can be messy because of:
      1. Input errors by the data collectors (e.g. put 100 hectares instead of 10 by mistake)
      2. Confusion about the questions (e.g. a respondent double counted some of his/her agricultural plots, thinking that was what the data collector was asking)
      3. Best guesses (e.g. a respondent does not know how many hectares his/her plot is, so he/she gives a best guess – says a plot is 5 hectares when in reality it is 2)

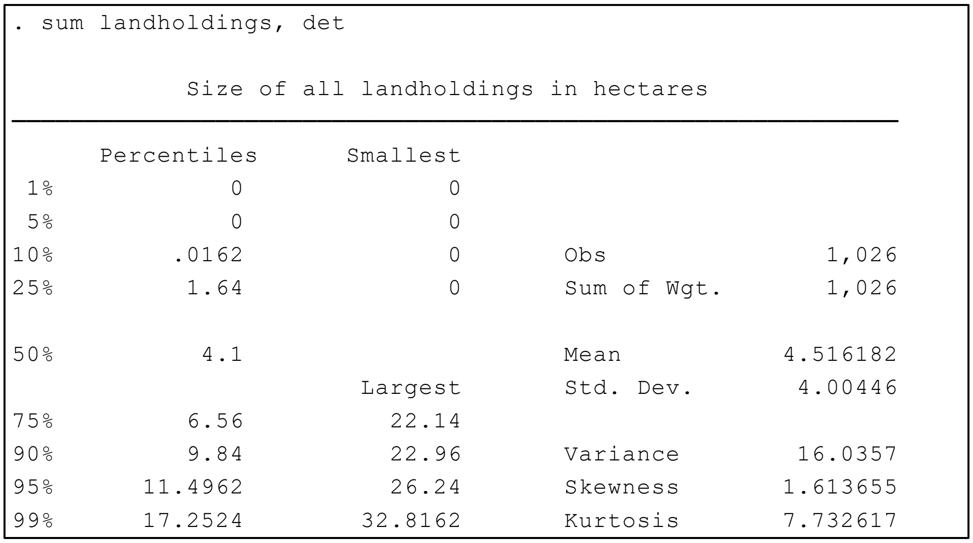
e. Many times there are notable **outliers** in raw, continuous data, for the 3 reasons listed above
   i. An outlier is an observation that is very different from all other observations
   ii. For example, perhaps all landholdings values fall in-between 0 and 30 hectares, but one household has a value of 100 hectares; that 100 hectares value is an outlier.

iii. 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.

f. What's a good way to check the *landholdings* variable for outliers?

i. <u>Practice:</u> *"sum landholdings, det"*

Figure 3.2: Stata command and output for "sum landholdings, det"

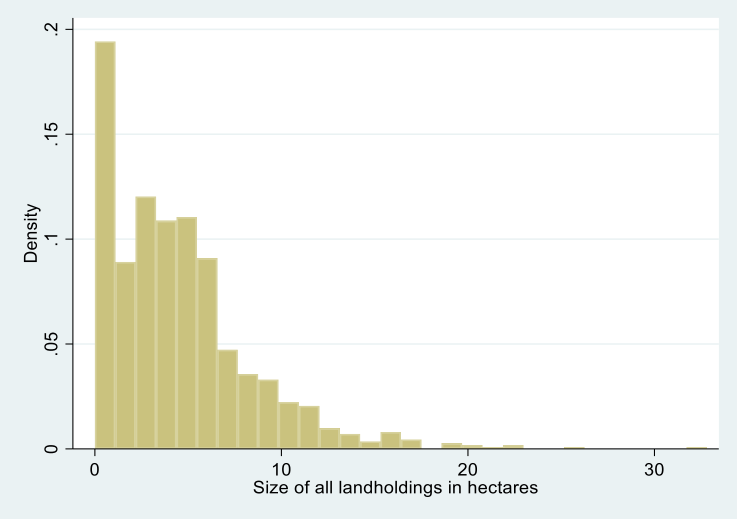


```
. sum landholdings, det

            Size of all landholdings in hectares
-------------------------------------------------------------

      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%        .0162              0       Obs               1,026
25%         1.64              0       Sum of Wgt.       1,026

50%          4.1                      Mean           4.516182
                        Largest       Std. Dev.       4.00446
75%         6.56          22.14
90%         9.84          22.96       Variance        16.0357
95%      11.4962          26.24       Skewness       1.613655
99%      17.2524        32.8162       Kurtosis       7.732617
```

1. First, check the mean and median: *the mean is 4.5 hectares and the median is 4.1 – these aren't too different from each other, which suggests a relatively normal distribution*
2. Then check the minimum and maximum: *the minimum is 0 (this makes sense because it's impossible to have negative land holdings), but the maximum is 32.8162!*
3. Because the distance from the maximum to the mean/median is so much larger than the distance from the minimum to the mean/median, there may be some outliers we want to address

ii. <u>Practice:</u> *"histogram landholdings"*

Figure 3.3: Stata histogram output for "hist landholdings"



1. Shows that there is a strong positive skew for this variable (a few observations very far on the right-hand side compared to the rest of the data)
2. This strong positive skew also shows the outliers that we may want to change

g. Everyone has a different preference for how to handle outliers. For now, let's say that all values greater than the 99th percentile (17.3 – we know from "sum landholdings, det") should be changed to the median (4.1). The 99th percentile tells us that 99% of all observations fall below, or are less than, 17.3, so this is a good cutoff to use to decide what constitutes being an outlier.

i. Question: First, let's see how many observations are greater than the 99th percentile (17.3). How might we check this? (HINT: We can combine "if" with one of the commands that we know). *"count if landholdings>17.3" – 10 observations*

Figure 3.4: Stata command and output for "count if landholdings>17.3"

```
. count if landholdings>17.3
  10
```

ii. Question: How might we change these 10 observations to the median (4.1)? *"replace landholdings=4.1 if landholdings>17.3"*

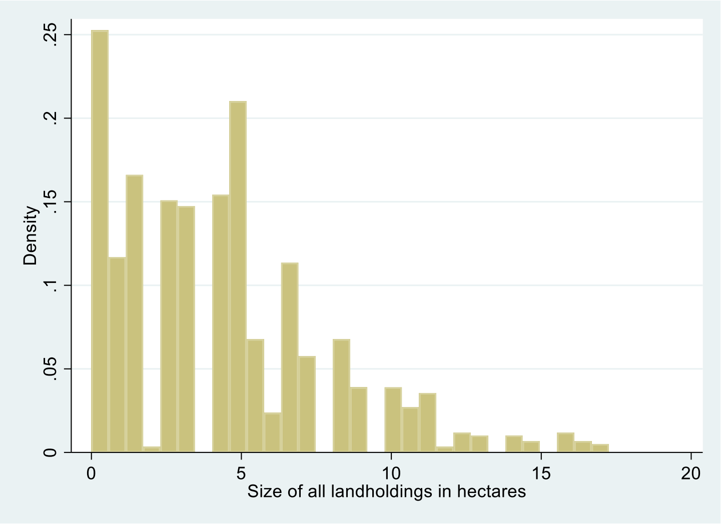
Figure 3.5: Stata command and output for "replace landholdings=4.1 if landholdings>17.3"

```
. replace landholdings=4.1 if landholdings>17.3
(10 real changes made)
```

iii. Stata tells us how many observations it changed with our command in the output. How many were changed? *10 – the same amount that the "count…if" command told us*

iv. Question: Now how can we look at the distribution of our newly cleaned *landholdings* variable? *"hist landholdings"*

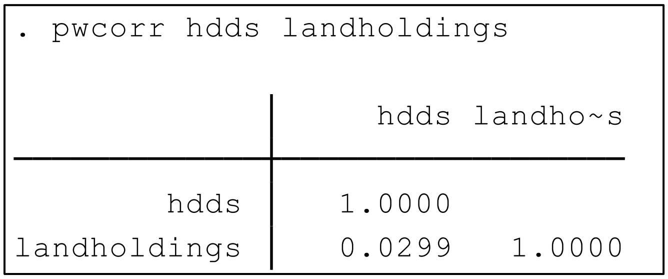
Figure 3.6: Stata histogram output for "hist landholdings", after cleaning outliers



1. Now the positive skew is much less
2. The figure only goes up to 20 hectares, instead of more than 30 like the first one!

2. **Correlations**
    a. Let's see how correlated our variable of interest (hdds) is with our newly cleaned continuous variable for land holdings:
        i. <u>Code:</u> *pwcorr [varname] [varname]*
        ii. <u>Practice:</u> *"pwcorr hdds landholdings"*

Figure 3.7: Stata command and output for "pwcorr hdds landholdings"



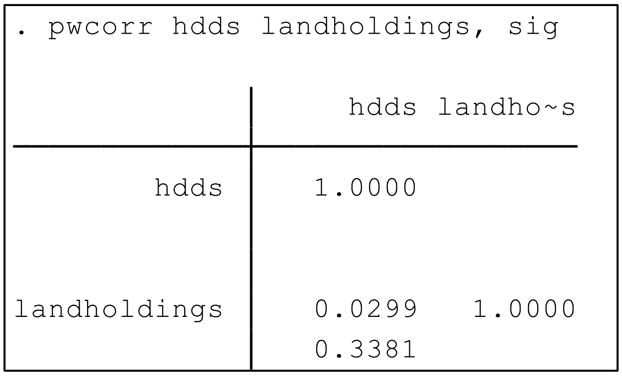
```
. pwcorr hdds landholdings

             |     hdds landho~s
-------------+------------------
        hdds |   1.0000 
landholdings |   0.0299   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 hdds and landholdings? *0.03*
            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 hdds and landholdings positive or negative? *Positive*
            2. Even though the correlation coefficient is very weak, how can you interpret this/explain this in common terms? *Households with <u>more</u> land are associated with eating <u>more</u> food groups.*

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 hdds landholdings, sig”*

Figure 3.8: Stata command and output for “pwcorr hdds landholdings, sig”



```
. pwcorr hdds landholdings, sig

             |     hdds landho~s
-------------+------------------
        hdds |   1.0000
             |
             |
landholdings |   0.0299   1.0000
             |   0.3381
```

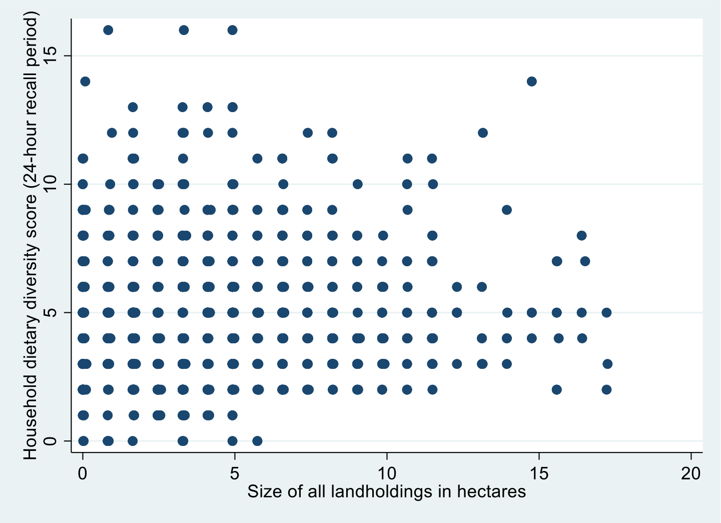
   ii. This new number below the correlation coefficient is the p-value:

      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. 1-0.3381 equals about .66 – so we can say with 66% 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? *0.3393*
      4. Is this significant? *No*

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 hdds landholdings”*

Figure 3.9: Stata scatterplot output for “scatter hdds landholdings”



iii. It looks like there may be a positive relationship (as the values for hdds increase the values for landholdings also increase)

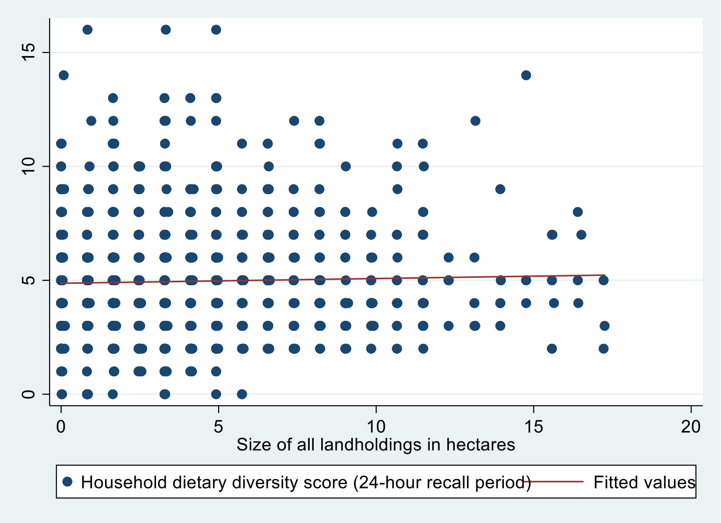
iv. We can also create a 'line of best fit' to see how positive and strong the relationship is

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 hdds landholdings) (lfit hdds landholdings)"*

Figure 3.10: Stata scatterplot output for "twoway (scatter hdds landholdings) (lfit hdds landholdings)



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

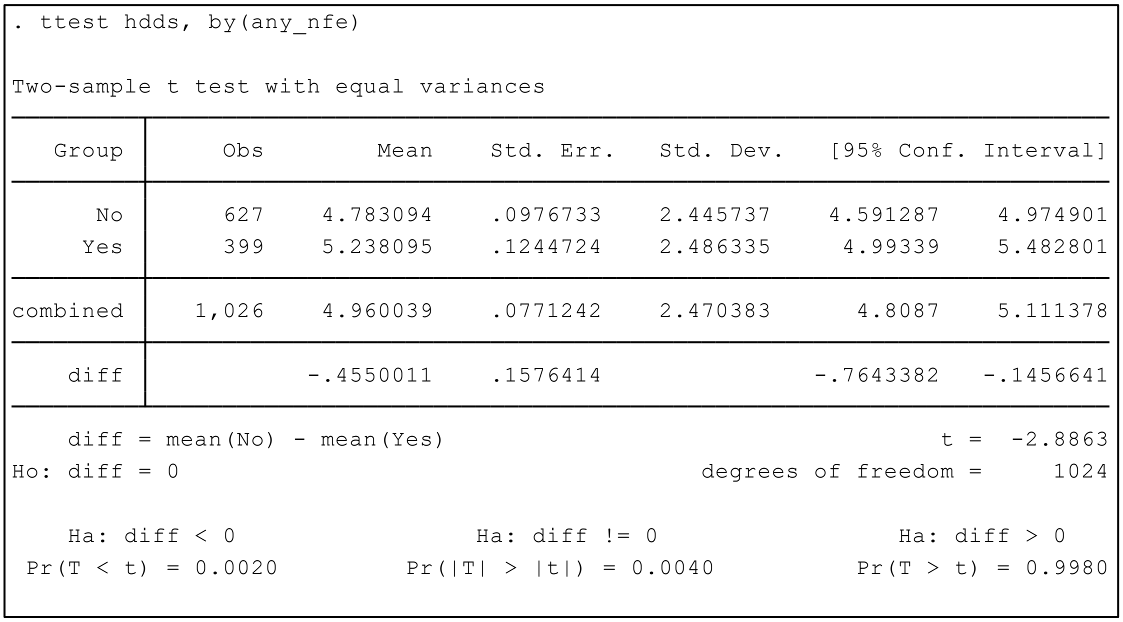
3. T-Tests

a. The most frequently used t-tests are two-sample t-tests: these tell us whether one variable (e.g. hdds) is significantly different between two groups in the data (e.g. whether or not a household has a non-farm enterprise)

b. Code: *ttest [varname], by([groupvar])*

c. Practice: *"ttest hdds, by(any_nfe)"* – output tells us:

Figure 3.11: Stata command and output for “ttest hdds, by(any_nfe)”



```
. ttest hdds, by(any_nfe)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      No |     627    4.783094    .0976733    2.445737    4.591287    4.974901
     Yes |     399    5.238095    .1244724    2.486335     4.99339    5.482801
---------+--------------------------------------------------------------------
combined |   1,026    4.960039    .0771242    2.470383      4.8087    5.111378
---------+--------------------------------------------------------------------
    diff |           -.4550011    .1576414               -.7643382   -.1456641
------------------------------------------------------------------------------
    diff = mean(No) - mean(Yes)                                   t =  -2.8863
Ho: diff = 0                                     degrees of freedom =     1024

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0020         Pr(|T| > |t|) = 0.0040          Pr(T > t) = 0.9980
```

i. Number of hdds observations in each of the two groups. How many observations do not have a non-farm enterprise? *627 households do **not** have a non-farm enterprise*

ii. Average HDDS of households each group. Which group has a higher HDDS? *Households with a non-farm enterprise have a higher HDDS*

iii. Standard error, standard deviation, and 95% confidence interval of the HDDS of each group. Do the two confidence intervals overlap? *No, the confidence intervals for HDDS do not overlap between households with and without a non-farm enterprise*

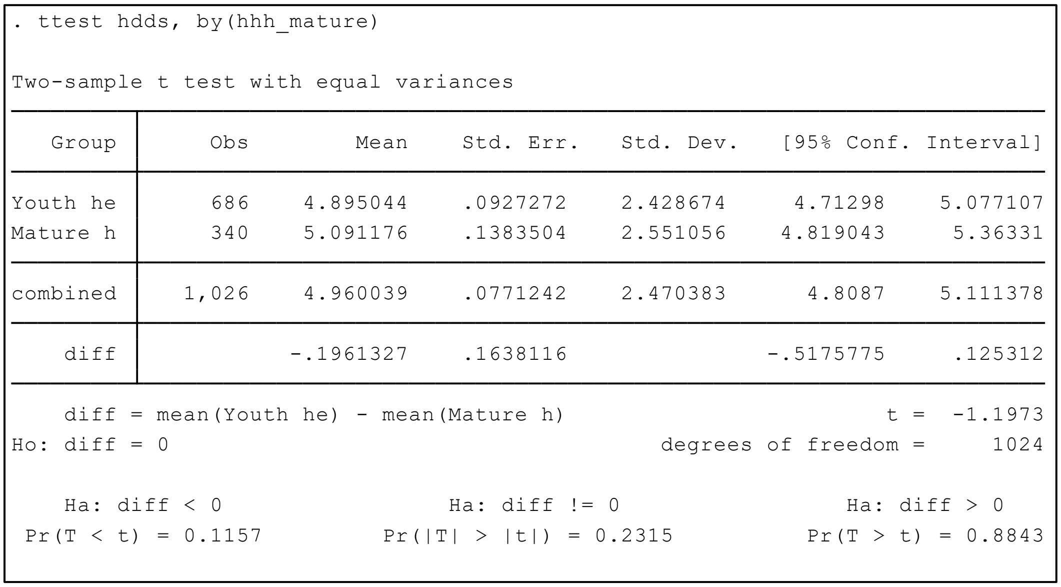
iv. T-statistic, degrees of freedom, and p-values (3 values on the bottom). The p-values on the left/right are for whether the difference between the two means is less than or greater than 0 (one-sided t-test). The p-value in the middle is for whether the difference between the two means is not equal 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 0.0020. We can say with more than 99% confidence that the mean household dietary diversity score of households with a non-farm enterprise is larger than the mean household dietary diversity score of households without a non-farm enterprise*

d. Now let’s see if households with a youth household head have significantly different HDDS than households with older household heads.

i. Question: How would we write this t-test? *“ttest hdds, by(hhh_mature)”*

Figure 3.12: Stata command and output for "ttest hdds, by(hhh_mature)"



```
. ttest hdds, by(hhh_mature)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Youth he |     686    4.895044    .0927272    2.428674     4.71298    5.077107
Mature h |     340    5.091176    .1383504    2.551056    4.819043     5.36331
---------+--------------------------------------------------------------------
combined |   1,026    4.960039    .0771242    2.470383      4.8087    5.111378
---------+--------------------------------------------------------------------
    diff |           -.1961327    .1638116               -.5175775     .125312
------------------------------------------------------------------------------
    diff = mean(Youth he) - mean(Mature h)                        t =  -1.1973
Ho: diff = 0                                     degrees of freedom =     1024

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.1157         Pr(|T| > |t|) = 0.2315          Pr(T > t) = 0.8843
```

ii. Which group has a higher average HDDS? *Mature-headed households have a higher average HDDS than youth-headed households*

iii. What is the p-value that the difference in means is less than 0? Is the difference in means significant? *P-value=0.1157; this is not significant*

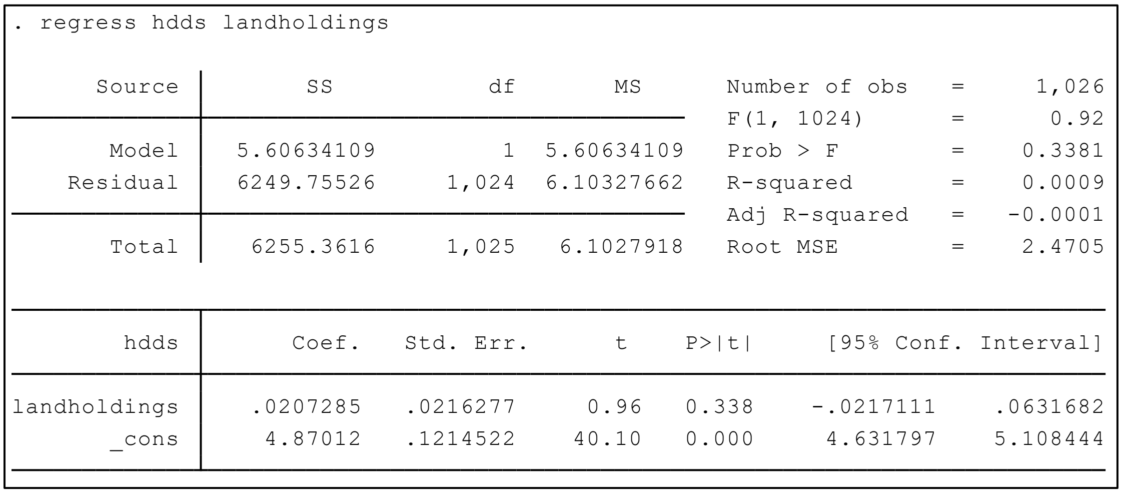
4. **Ordinary Least-Squares (OLS) Linear Regressions**

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

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

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

Figure 3.13: Stata command and output for "regress hdds landholdings"

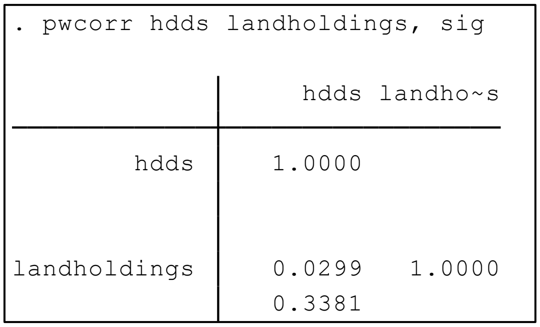


```
. regress hdds landholdings

      Source |       SS           df       MS      Number of obs   =     1,026
-------------+----------------------------------   F(1, 1024)      =      0.92
       Model |  5.60634109         1  5.60634109   Prob > F        =    0.3381
    Residual |  6249.75526     1,024  6.10327662   R-squared       =    0.0009
-------------+----------------------------------   Adj R-squared   =   -0.0001
       Total |   6255.3616     1,025   6.1027918   Root MSE        =    2.4705

------------------------------------------------------------------------------
        hdds |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
landholdings |   .0207285   .0216277     0.96   0.338    -.0217111    .0631682
       _cons |    4.87012   .1214522    40.10   0.000     4.631797    5.108444
------------------------------------------------------------------------------
```

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 landholdings in this output) and the dependent variable (hdds): 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 hdds landholdings, sig”*

Figure 3.14: Stata command and output for “pwcorr hdds landholdings, sig”



```
. pwcorr hdds landholdings, sig

             |     hdds landho~s
-------------+------------------
        hdds |   1.0000
             |
             |
landholdings |   0.0299   1.0000
             |   0.3381
```

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 household dietary diversity?

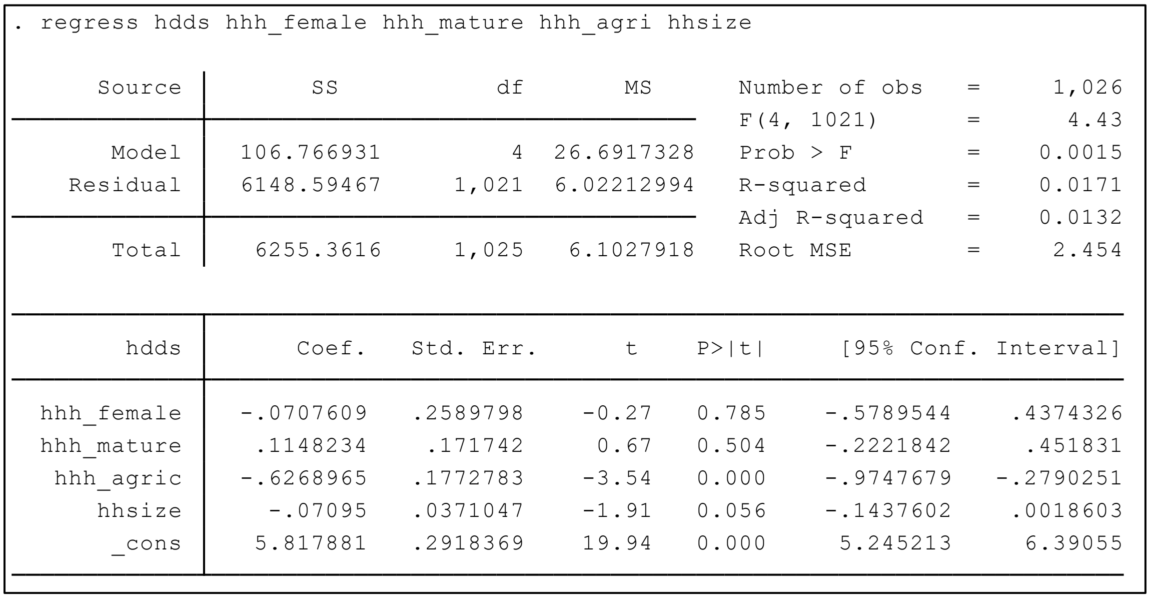
i. Let’s try to run a regression where HDDS is still the dependent variable, and the independent variables are whether or not the household head is female, mature, or has agriculture as his/her primary activity, and the household size

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 hhh_female”* – categorical dummy variable (0/1)
2. *“codebook hhh_mature”* – categorical dummy variable (0/1)
3. *“codebook hhh_agric”* – categorical dummy variable (0/1)
4. *“codebook hhsize”* – continuous

iii. Practice: *"regress hdds hhh_female hhh_mature hhh_agri hhsize"*

Figure 3.15: Stata command and output for "regress hdds hhh_female hhh_mature hhh_agri hhsize"



```
. regress hdds hhh_female hhh_mature hhh_agri hhsize

      Source |       SS           df       MS      Number of obs   =     1,026
-------------+----------------------------------   F(4, 1021)      =      4.43
       Model |  106.766931         4  26.6917328   Prob > F        =    0.0015
    Residual |  6148.59467     1,021  6.02212994   R-squared       =    0.0171
-------------+----------------------------------   Adj R-squared   =    0.0132
       Total |   6255.3616     1,025   6.1027918   Root MSE        =     2.454

------------------------------------------------------------------------------
        hdds |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  hhh_female |  -.0707609   .2589798    -0.27   0.785    -.5789544    .4374326
  hhh_mature |   .1148234    .171742     0.67   0.504    -.2221842     .451831
   hhh_agric |  -.6268965   .1772783    -3.54   0.000    -.9747679   -.2790251
      hhsize |   -.07095    .0371047    -1.91   0.056    -.1437602    .0018603
       _cons |   5.817881   .2918369    19.94   0.000     5.245213     6.39055
------------------------------------------------------------------------------
```

iv. Are any of these variables significant? *Whether or not the household head's primary activity is agriculture is significant – the p-value is 0.000*

v. What does the coefficient on this variable (hhh_agric) tell us? *Because hhh_agric is a dummy variable, the coefficient says that if the household head's primary activity is agriculture, the HDDS will decrease by 0.63. (The coefficient is -0.63)*

vi. Which variable is almost significant at the 95% level? *Household size – the p-value is 0.056*

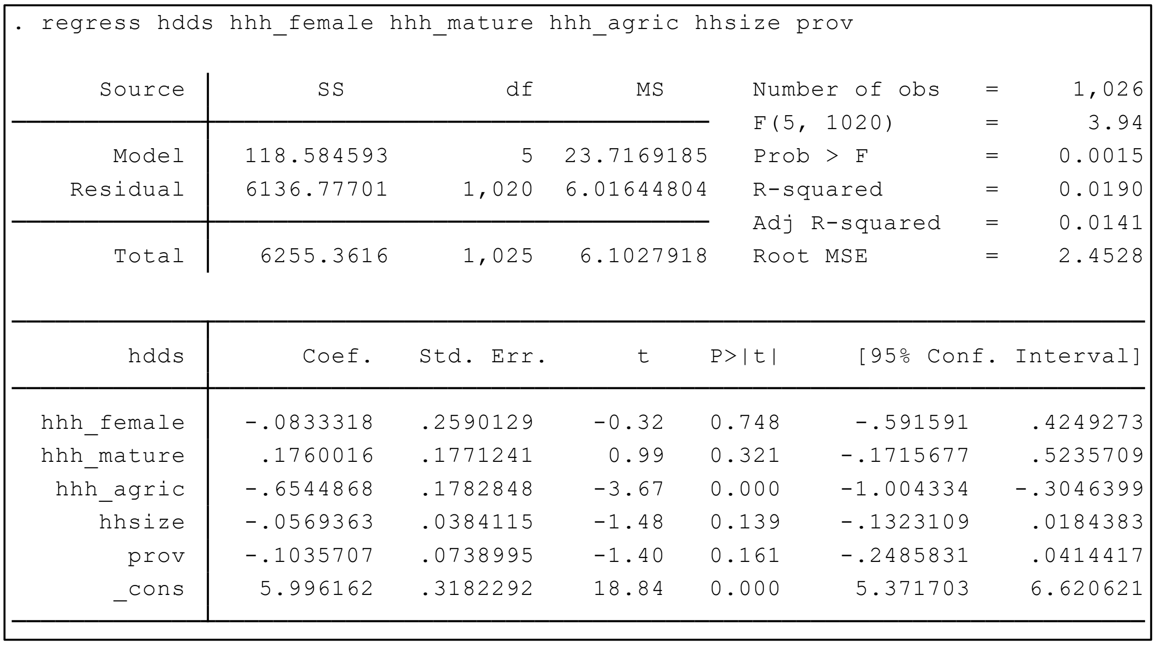
vii. What does this coefficient on hhsize tell us? *Because hhsize is a continuous variable, the coefficient says that the marginal effect of one additional person in the household decreases the HDDS by 0.07. (The coefficient is -0.07)*

c. Let's add province to our regression.

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

ii. Practice: *“regress hdds hhh_female hhh_mature hhh_agri hhsize prov”*

Figure 3.16: Stata command and output for “regress hdds hhh_female hhh_mature hhh_agri hhsize prov”



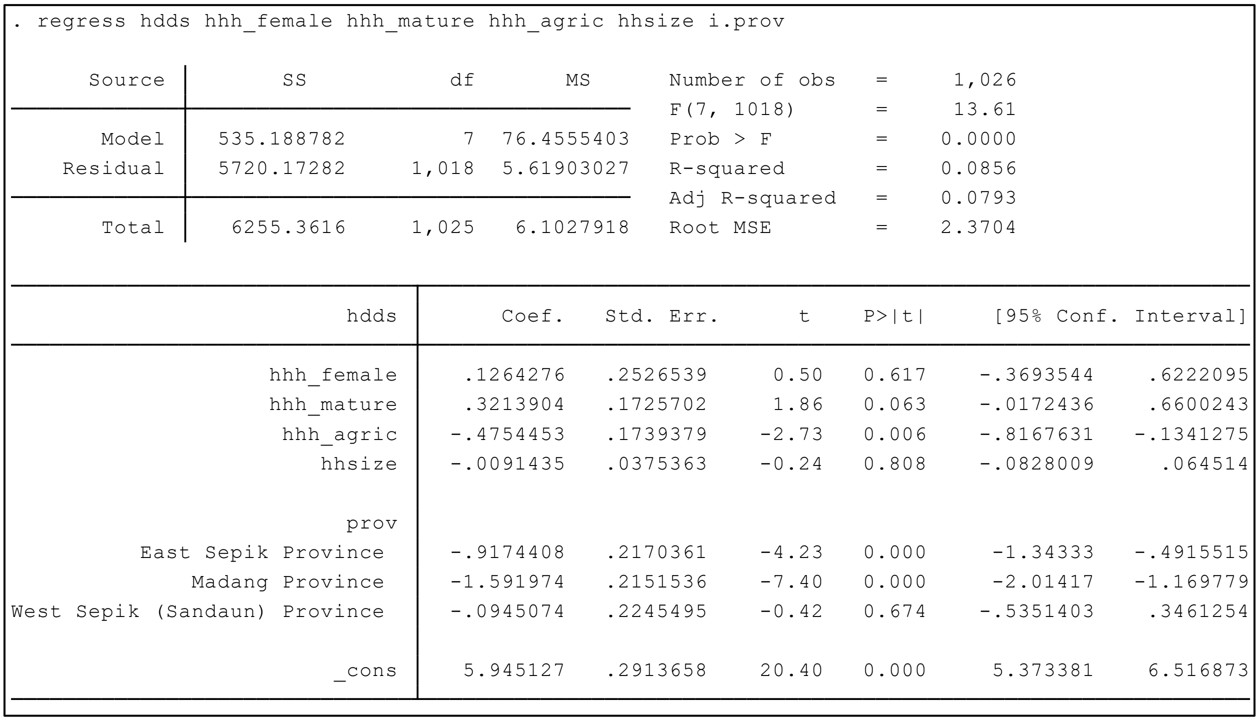
```
. regress hdds hhh_female hhh_mature hhh_agric hhsize prov

      Source |       SS           df       MS      Number of obs   =     1,026
-------------+----------------------------------   F(5, 1020)      =      3.94
       Model |  118.584593         5  23.7169185   Prob > F        =    0.0015
    Residual |  6136.77701     1,020  6.01644804   R-squared       =    0.0190
-------------+----------------------------------   Adj R-squared   =    0.0141
       Total |   6255.3616     1,025   6.1027918   Root MSE        =    2.4528

------------------------------------------------------------------------------
        hdds |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  hhh_female |  -.0833318   .2590129    -0.32   0.748     -.591591    .4249273
  hhh_mature |   .1760016   .1771241     0.99   0.321    -.1715677    .5235709
   hhh_agric |  -.6544868   .1782848    -3.67   0.000    -1.004334   -.3046399
      hhsize |  -.0569363   .0384115    -1.48   0.139    -.1323109    .0184383
        prov |  -.1035707   .0738995    -1.40   0.161    -.2485831    .0414417
       _cons |   5.996162   .3182292    18.84   0.000     5.371703    6.620621
------------------------------------------------------------------------------
```

iii. Question: Is province significant? How can we interpret the coefficient on “prov”? *Province is not significant. For every one unit increase in prov, the HDDS decreases by 0.1… This doesn’t make sense! Province isn’t a continuous or dummy variable….*

iv. So instead, we can put “i.” in front of prov (or any categorical non-dummy variables). Let’s try again:

v. Practice: *“regress hdds hhh_female hhh_mature hhh_agri hhsize i.prov”*

Figure 3.17: Stata command and output for “regress hdds hhh_female hhh_mature hhh_agri hhsize i.prov”



```
. regress hdds hhh_female hhh_mature hhh_agric hhsize i.prov

      Source |       SS           df       MS      Number of obs   =     1,026
-------------+----------------------------------   F(7, 1018)      =     13.61
       Model |  535.188782         7  76.4555403   Prob > F        =    0.0000
    Residual |  5720.17282     1,018  5.61903027   R-squared       =    0.0856
-------------+----------------------------------   Adj R-squared   =    0.0793
       Total |   6255.3616     1,025   6.1027918   Root MSE        =    2.3704

-----------------------------------------------------------------------------------------------
                         hdds |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------------------+----------------------------------------------------------------
                   hhh_female |   .1264276   .2526539     0.50   0.617    -.3693544    .6222095
                   hhh_mature |   .3213904   .1725702     1.86   0.063    -.0172436    .6600243
                    hhh_agric |  -.4754453   .1739379    -2.73   0.006    -.8167631   -.1341275
                       hhsize |  -.0091435   .0375363    -0.24   0.808    -.0828009     .064514
                              |
                         prov |
         East Sepik Province  |  -.9174408   .2170361    -4.23   0.000     -1.34333   -.4915515
             Madang Province  |  -1.591974   .2151536    -7.40   0.000     -2.01417   -1.169779
West Sepik (Sandaun) Province |  -.0945074   .2245495    -0.42   0.674    -.5351403    .3461254
                              |
                        _cons |   5.945127   .2913658    20.40   0.000     5.373381    6.516873
-----------------------------------------------------------------------------------------------
```

   vi. This now turns province into 3 dummy variables for the regression. Stata automatically chooses the category with the lowest value (here 1=ARoB) to drop to be the comparison group. So now, each other province is being compared to ARoB.

   vii. Are any of the provinces significant? *East Sepik and Madang both have p-values=0.000.*

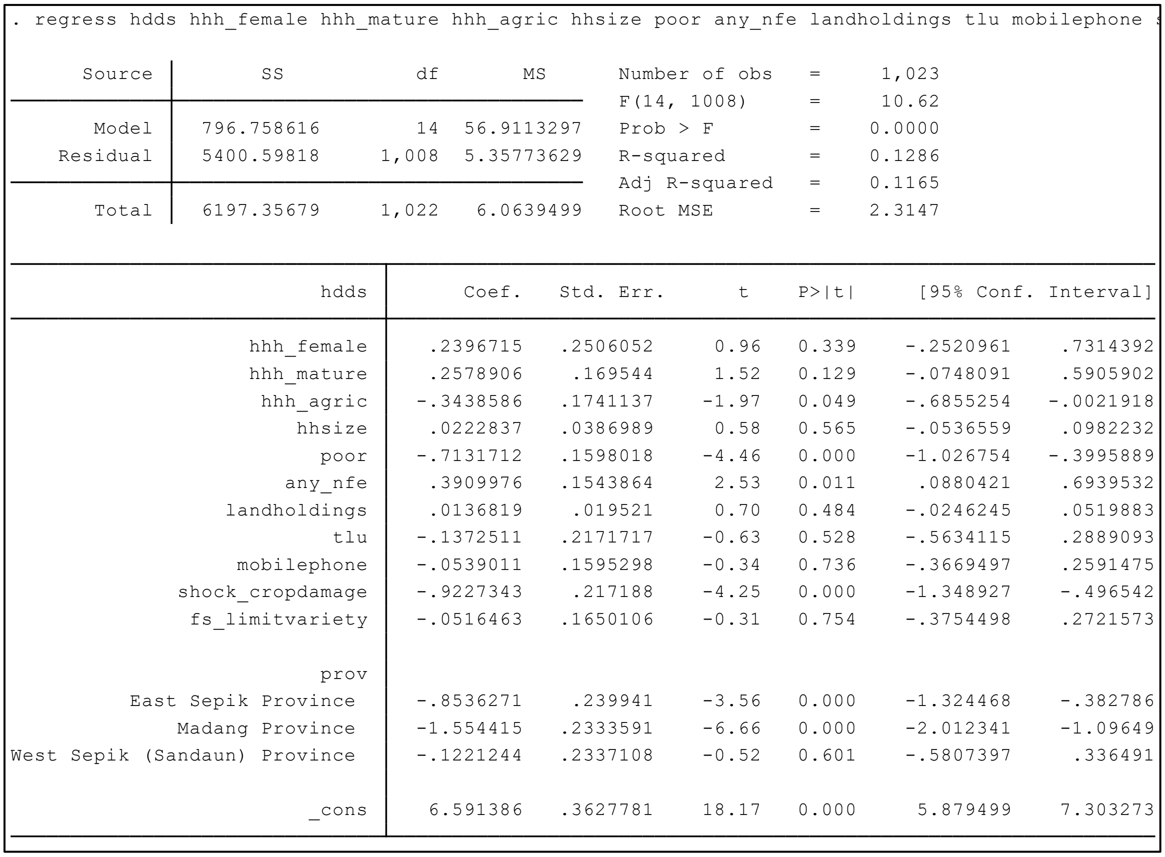
   viii. This means that holding these other variables constant, households in East Sepik have an average HDDS that is 0.92 *lower* than households in ARoB.

   ix. Similarly, holding these other variables constant, households in Madang have an average HDDS that is 1.59 *lower* than households in ARoB.

d. Refining the regression – let's run one more regression, with more independent variables

   i. Practice: *"regress hdds hhh_female hhh_mature hhh_agric hhsize poor any_nfe anymigrant landholdings tlu mobilephone shock_crop shock_earth shock_flood shock_price fs_limit i.prov"*

Figure 3.18: Stata command and output for "regress hdds hhh_female hhh_mature hhh_agri…"



```
. regress hdds hhh_female hhh_mature hhh_agric hhsize poor any_nfe landholdings tlu mobilephone s

      Source |       SS           df       MS      Number of obs   =     1,023
-------------+----------------------------------   F(14, 1008)     =     10.62
       Model |  796.758616        14  56.9113297   Prob > F        =    0.0000
    Residual |  5400.59818     1,008  5.35773629   R-squared       =    0.1286
-------------+----------------------------------   Adj R-squared   =    0.1165
       Total |  6197.35679     1,022  6.0639499    Root MSE        =    2.3147

-------------------------------------------------------------------------------------------------
                           hdds |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------------------------+----------------------------------------------------------------
                     hhh_female |   .2396715   .2506052     0.96   0.339    -.2520961    .7314392
                     hhh_mature |   .2578906    .169544     1.52   0.129    -.0748091    .5905902
                      hhh_agric |  -.3438586   .1741137    -1.97   0.049    -.6855254   -.0021918
                         hhsize |   .0222837   .0386989     0.58   0.565    -.0536559    .0982232
                           poor |  -.7131712   .1598018    -4.46   0.000    -1.026754   -.3995889
                        any_nfe |   .3909976   .1543864     2.53   0.011     .0880421    .6939532
                   landholdings |   .0136819    .019521     0.70   0.484    -.0246245    .0519883
                            tlu |  -.1372511   .2171717    -0.63   0.528    -.5634115    .2889093
                    mobilephone |  -.0539011   .1595298    -0.34   0.736    -.3669497    .2591475
               shock_cropdamage |  -.9227343    .217188    -4.25   0.000    -1.348927    -.496542
                fs_limitvariety |  -.0516463   .1650106    -0.31   0.754    -.3754498    .2721573
                                |
                           prov |
            East Sepik Province |  -.8536271    .239941    -3.56   0.000    -1.324468    -.382786
                Madang Province |  -1.554415   .2333591    -6.66   0.000    -2.012341    -1.09649
West Sepik (Sandaun) Province   |  -.1221244   .2337108    -0.52   0.601    -.5807397     .336491
                                |
                          _cons |   6.591386   .3627781    18.17   0.000     5.879499    7.303273
-------------------------------------------------------------------------------------------------
```

   ii. What is significant in this regression? *Poor, shock_cropdamage, East Sepik, and Madang*

   iii. How can we interpret these significant coefficients?

      1. *Poor households have an average HDDS 0.71 lower than non-poor households*

      2. *Households who experienced crop damage in the past 5 years have an average HDDS 0.92 lower than households who did not experience crop damage*

3. *Households in East Sepik and Madang provinces have average HDDS lower than households in ARoB (by 0.85 and 1.55, respectively)*

5. Conclusions and Next Steps

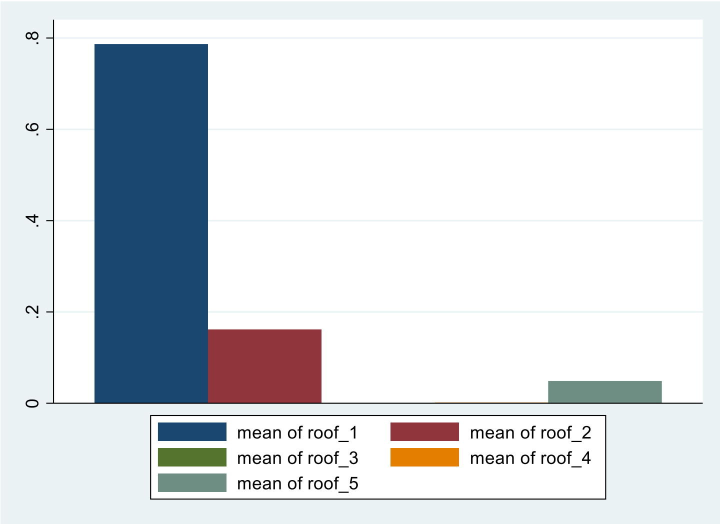
This concludes the Stata Introductory Course on Describing, Transforming, and Analyzing Data. This training simply provided an overview of the most common Stata commands, best practices to construct and use them, and how to interpret their output. However, this is only the beginning of all of the data 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.

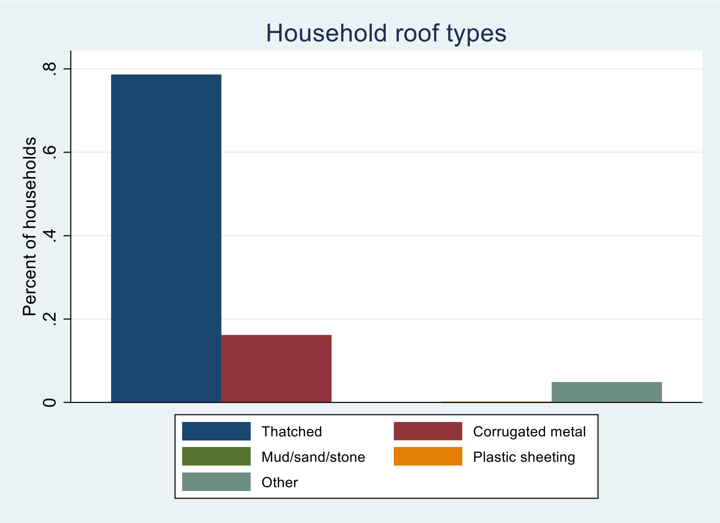
Creating Figures in Stata

Lesson and Presentation 1 – Household Characteristics – Bar graphs
Stata file: "PNG_StataTraining_P1_HouseholdCharacteristicsFigures.do"

1. Roof types across the whole sample
    a. Figure A is a simple bar graph, showing the percent of households with each roof type, across the whole sample. For categorical variables (such as roof type), 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: graph bar roof_1 roof_2 roof_3 roof_4 roof_5



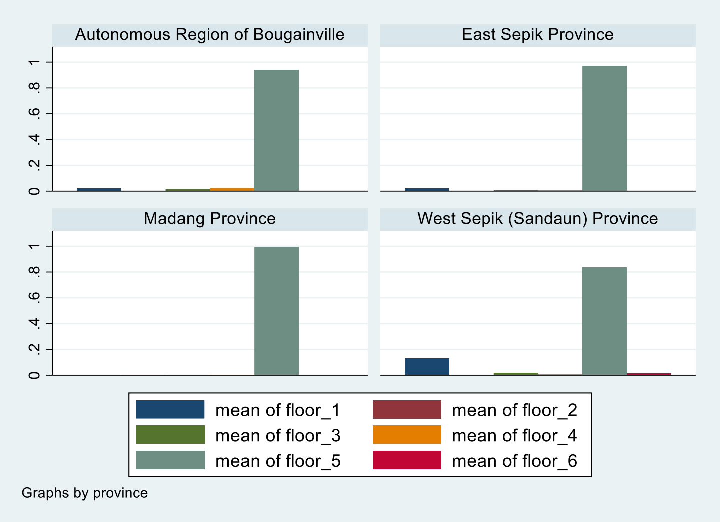
        ii. What does this figure tell us?
            1. roof_1 is the most common roof type
            2. Barely any households have roof_3 or roof_4
        iii. What does this figure not tell us?
            1. What roof_1, roof_2, etc. mean
            2. What the y-axis means
            3. What the title 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 roof_1 roof_2 roof_3 roof_4 roof_5, legend(size(small) order(1 "Thatched" 2 "Corrugated metal" 3 "Mud/sand/stone" 4 " Plastic sheeting" 5 "Other")) ytitle("Percent of households") title("Household roof types")



ii. What is now added to this figure?
   1. Labels for roof_1, roof_2, etc. – *"legend(size(small) order(1 "Thatched" 2 "Corrugated metal" 3 "Mud/sand/stone" 4 " Plastic sheeting" 5 "Other"))"*
   2. Y-axis label – *"ytitle("Percent of households")"*
   3. Title – *"title("Household roof types")"*

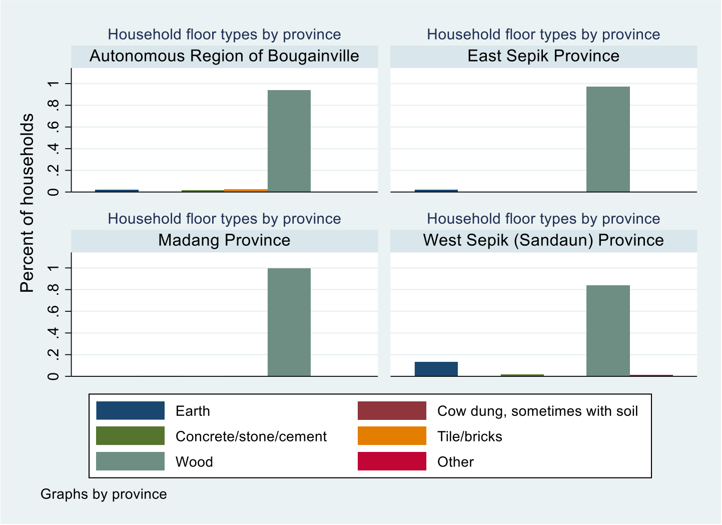
iii. What can we learn from this figure?
   1. Thatched roofs are the most common, followed by corrugated metal
   2. Mud/sand/stone and plastic sheeting roofs are nearly nonexistent in our sample

2. Floor types by province (separate graphs per province)
   a. Figure A is like the simple bar graph created above, but it is separated into 4 different figures (one for each province)
      i. Code: graph bar floor_1 floor_2 floor_3 floor_4 floor_5 floor_6, by(prov)

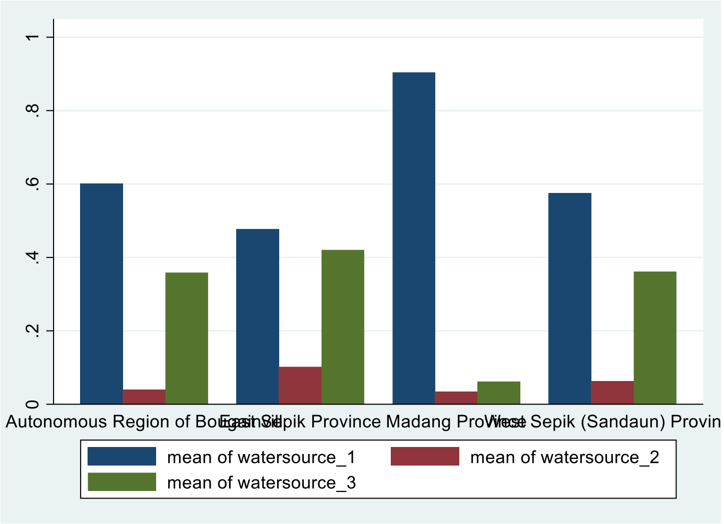


ii. What does this figure tell us?
   1. Floor_5 is the most common floor in each province

        2. West Sepik has the highest diversity of floors compared to the other provinces (although floor_5 is still overwhelmingly the most common)
    iii. What does this figure <u>not</u> tell us?
        1. What floor_1, floor_2, etc. 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. <u>Code:</u> graph bar floor_1 floor_2 floor_3 floor_4 floor_5 floor_6, by(prov) legend(size(small) order(1 "Earth" 2 "Cow dung, sometimes with soil" 3 "Concrete/stone/cement" 4 "Tile/bricks" 5 "Wood" 6 "Other")) ytitle("Percent of households") title("Household floor types by province", size(medsmall))



    ii. What is now added to this figure?
        1. Labels for roof_1, roof_2, etc. – *“legend(size(small) order(1 "Earth" 2 "Cow dung, sometimes with soil" 3 "Concrete/stone/cement" 4 "Tile/bricks" 5 "Wood" 6 "Other"))”*
        2. Y-axis label – *“ytitle("Percent of households")”*
        3. Title – *“title("Household floor types by province", size(medsmall))”*
    iii. What can we learn from this figure?
        1. Households in all provinces mainly have wood floors
        2. West Sepik is more likely to have earth floors compared to the other provinces
3. Water source by province (all on the same graph)
  a. Figure A is like the simple bar graphs created above for each province, but instead all of the provinces will be on the same graph
    i. <u>Code:</u> graph bar watersource_1 watersource_2 watersource_3, ***over***(prov)

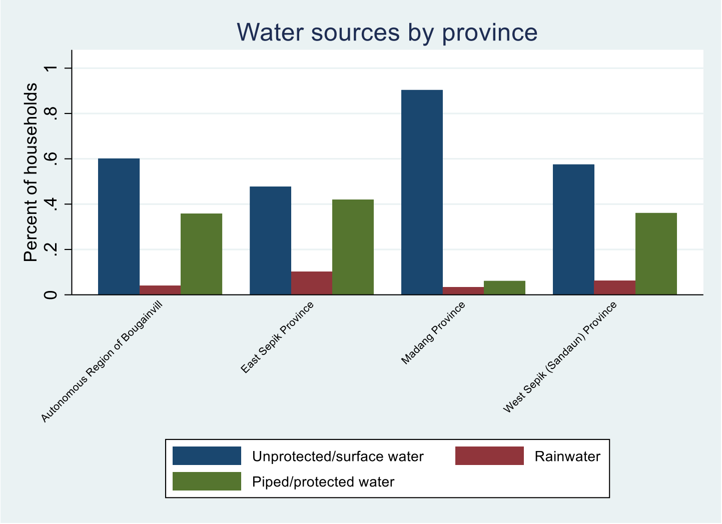


ii. How can we make this graph better?

   1. Adjust the province labels so they are not overlapping
   2. Add labels to what the different color bars mean
   3. Add a y-axis title
   4. Add a title for the figure

b. Figure B adds these 4 elements to improve upon Figure A

   i. <u>Code:</u> graph bar watersource_1 watersource_2 watersource_3, over(prov, label(labsize(vsmall) angle(45))) legend(size(small) order(1 "Unprotected/surface water" 2 "Rainwater" 3 "Piped/protected water")) ytitle("Percent of households") title("Water sources by province")



   ii. What is now added to the figure?

      1. Adjusted the province labels so they are not overlapping – *"over(prov, label(labsize(vsmall) angle(45)))"*
      2. Added labels to what the different color bars mean – *"legend(size(small) order(1 "Unprotected/surface water" 2 "Rainwater" 3 "Piped/protected water"))"*
      3. Added a y-axis title – *"ytitle("Percent of households")"*

4. Added a title for the figure – *“title("Water sources by province")”*

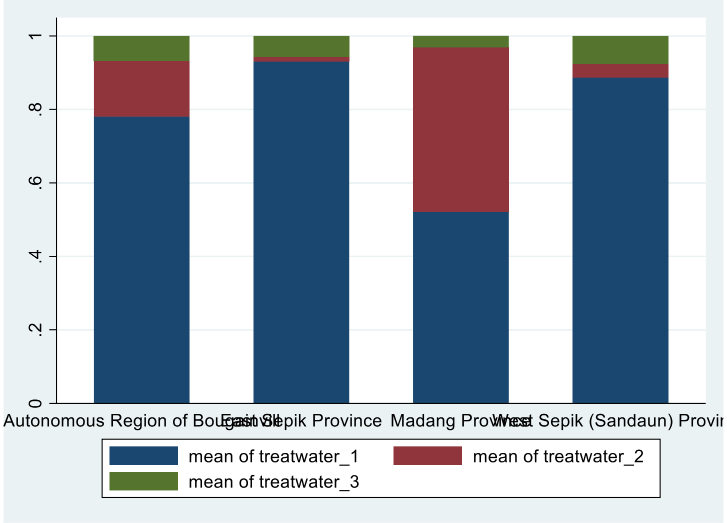
iii. What can we learn from this figure?

1. Households in Madang are the most likely to drink from unprotected sources
2. Households in East Sepik are (only slightly) the most likely to drink from protected sources

4. Water treatment by province (all on the same graph with ***stacked*** bars)

a. Figure A is similar to the figures above, except it stacks the colored bars on top of each other, so that each province has one bar that sums to 1 (100%), meaning that each bar shows the different colors, and in total shows all of the types of water treatment reported by the sampled households in the province

i. Code: graph bar treatwater_1 treatwater_2 treatwater_3, over(prov) ***stack***

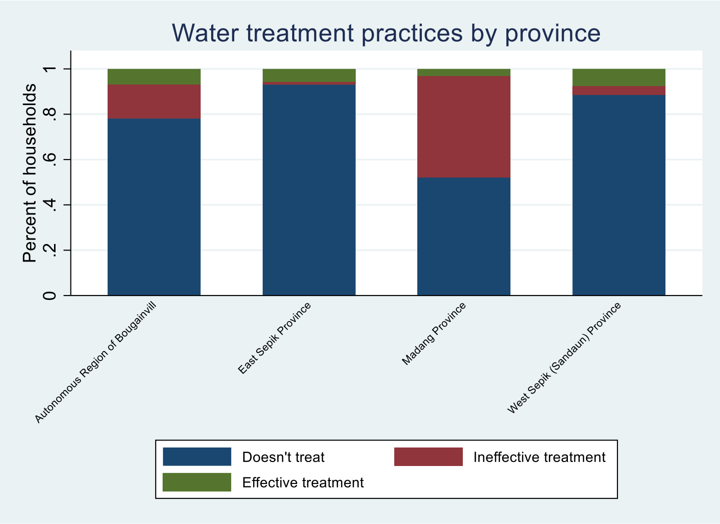


ii. How can we make this graph better?

1. Adjust the province labels so they are not overlapping
2. Add labels to what the different color bars mean
3. Add a y-axis title
4. Add a title for the figure

b. Figure B adds these 4 elements to improve upon Figure A

i. Code: graph bar treatwater_1 treatwater_2 treatwater_3, over(prov, label(labsize(vsmall) angle(45))) stack legend(size(small) order(1 "Doesn't treat" 2 "Ineffective treatment" 3 "Effective treatment")) ytitle("Percent of households") title("Water treatment practices by province")



ii. What is now added to the figure?

1. Adjusted the province labels so they are not overlapping – *"over(prov, label(labsize(vsmall) angle(45)))"*
2. Added labels to what the different color bars mean – *"legend(size(small) order(1 "Doesn't treat" 2 "Ineffective treatment" 3 "Effective treatment"))"*
3. Added a y-axis title – *"ytitle("Percent of households")"*
4. Added a title for the figure – *"title("Water treatment practices by province")*

iii. What can we learn from this figure?

1. East and West Sepik households are the least likely to treat their water.
2. Madang households are the most likely to report treating their water, but their treatment methods are mostly ineffective
3. Barely any households use effective treatment methods

Lesson and Presentation 2 – Anthropometry – Bar and line graphs

<u>Stata file:</u> "PNG_StataTraining_P2_AnthropometryFigures.do"

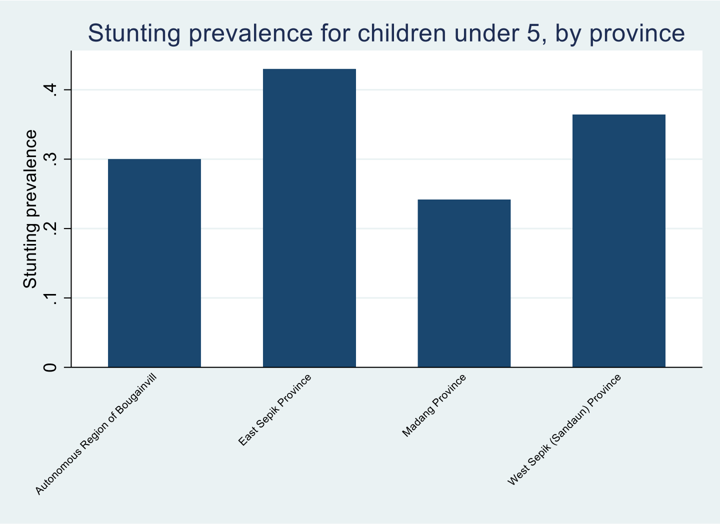
1. Stunting prevalence by province

   a. First, let's make a bar graph, where each bar is the stunting prevalence for a specific province

      i. Remember – to show prevalence on a bar graph, we have to make sure the variable of interest is a dummy variable!

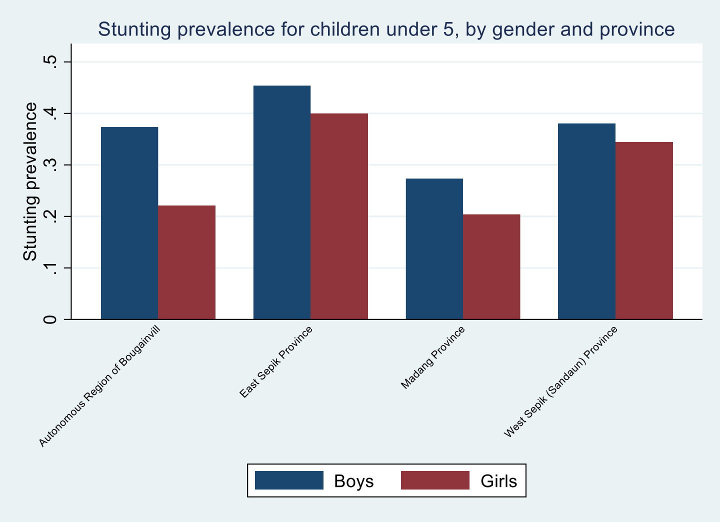
      ii. *"codebook stunted"* – yes, this is a dummy variable (0="No" and 1="Yes")

   b. <u>Code:</u> graph bar stunted, over(prov, label(labsize(vsmall) angle(45))) ytitle("Stunting prevalence") title("Stunting prevalence for children under 5, by province")



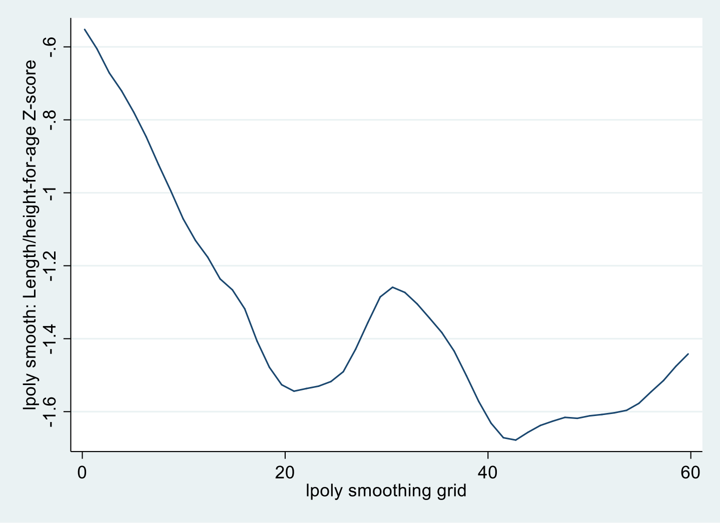
c. Which province has the highest stunting rate? *East Sepik*
d. Which province has the lowest stunting rate? *Madang*

2. Stunting prevalence by gender and province
    a. Next, we will create a similar figure, but now each province will have 2 bars – one showing the stunting rate of boys and the other showing the stunting rate of girls
    b. <u>Code:</u> graph bar stunt_boys stunt_girls, over(prov, label(labsize(vsmall) angle(45))) ytitle("Stunting prevalence") title("Stunting prevalence for children under 5, by gender and province", size(smallmed)) legend(order(1 "Boys" 2 "Girls"))



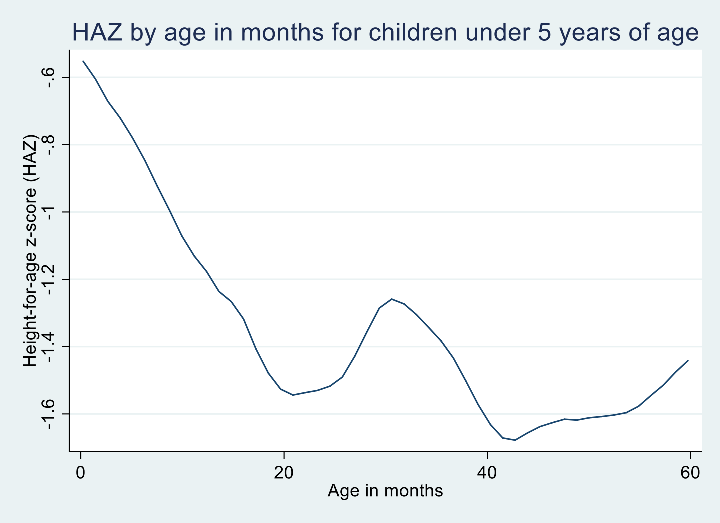
    c. Are girls or boys more likely to be stunted? *Boys*

3. Height-for-age z-scores (HAZ) by age in months, children 0-59 months
    a. Figure A creates a simple line graph (local polynomial) showing HAZ by age in months for children aged 0-59 months
        i. Because age in months (final_age) is numeric/continuous, we will make a line graph to show this (specifically, a local polynomial)
        ii. <u>Code:</u> twoway (lpoly haz final_age)



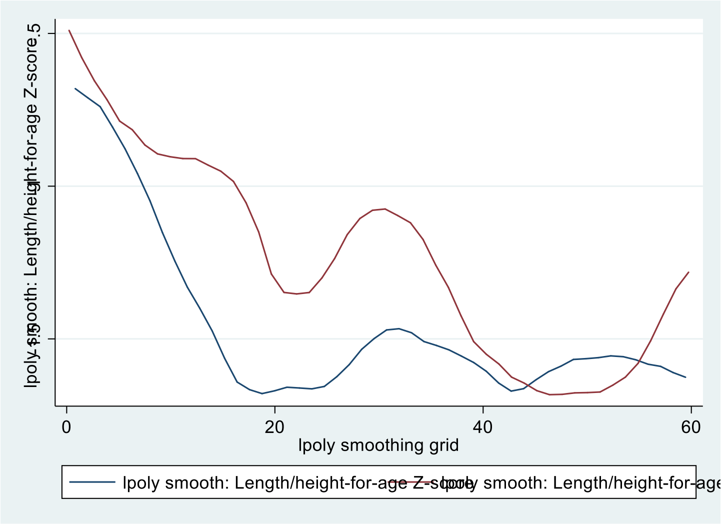
iii. What is this figure missing?
   1. *X-axis title*
   2. *Title*

b. Figure B adds titles to the figure above
   i. Code: twoway (lpoly haz final_age), xtitle("Age in months") ytitle("Height-for-age z-score (HAZ)") title("HAZ by age in months for children under 5 years of age")



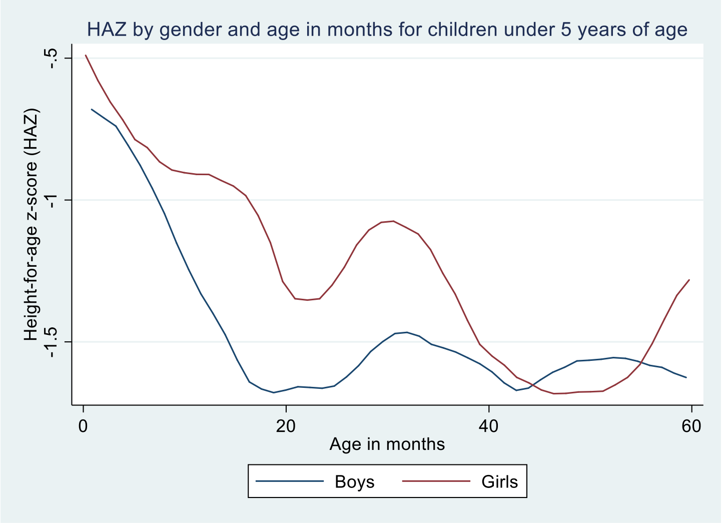
   ii. What happens to HAZ as age increases? *HAZ decreases as age increases*

4. Height-for-age z-scores (HAZ) by gender and age in months, children 0-60 months
   a. Figure A creates a similar figure to Figure A from above (number 3), but creates separate lines for boys and girls
      i. Code: twoway (lpoly haz final_age if gender==0) (lpoly haz final_age if gender==1)



ii. What is this figure missing?
   1. *X-axis title*
   2. *Title*
   3. *Meaningful labels for what the two color lines mean*

b. Figure B adds this missing elements
   i. Code: twoway (lpoly haz final_age if gender==0) (lpoly haz final_age if gender==1), legend(lab(1 "Boys") lab(2 "Girls")) xtitle("Age in months") ytitle("Height-for-age z-score (HAZ)") title("HAZ by gender and age in months for children under 5 years of age", size(smallmed))



   ii. What does this figure show? *For both boys and girls, HAZ decreases as age increases, however boys' HAZ decreases faster than girls'*
   iii. This is consistent with our bar graphs on stunting prevalence by gender, which showed that more boys are stunted than girls

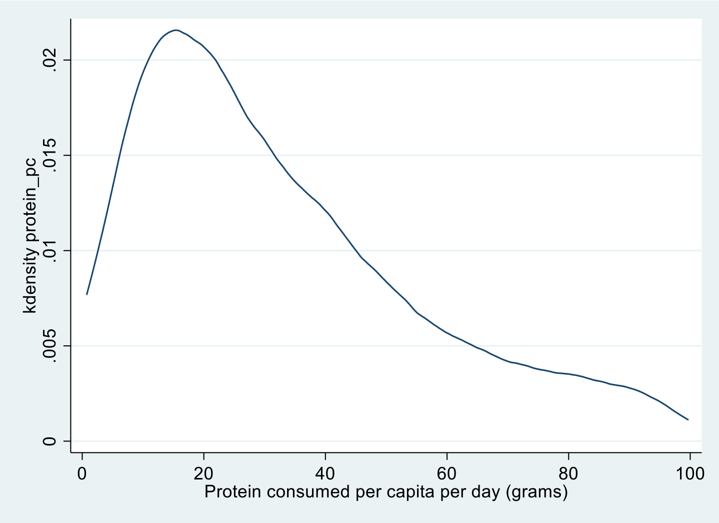
Lesson and Presentation 3 – Non-farm Enterprises – Bar graphs, line graphs, and box plots

Stata file: "PNG_StataTraining_P3_NFEFigures.do"

1. Protein consumption per capita

a. Figure A shows protein consumption per capita across the whole sample – because protein 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). We are also going to limit the figure to only showing 100 grams of protein (so that it does not show the outliers and to get a better look at the majority of the observations).

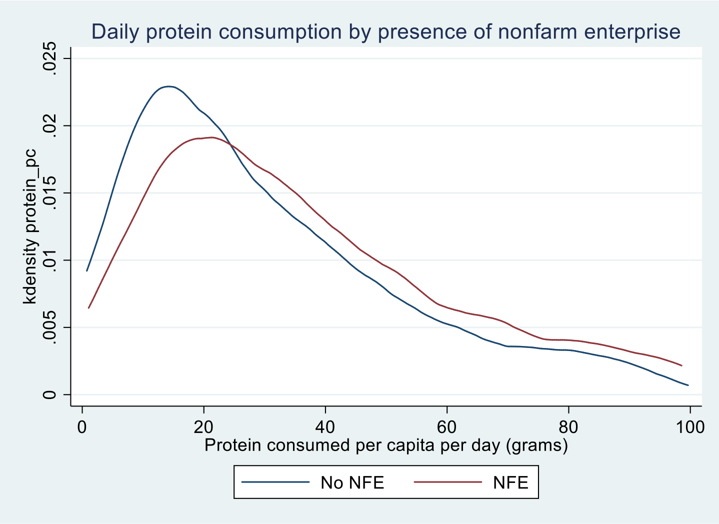
   i. Code: twoway (kdensity protein_pc if protein_pc<=100), xtitle(“Protein consumed per person per day (grams)”)



   ii. What does this figure show? *Across the whole sample, most people consume between 0-40 grams of protein per day*

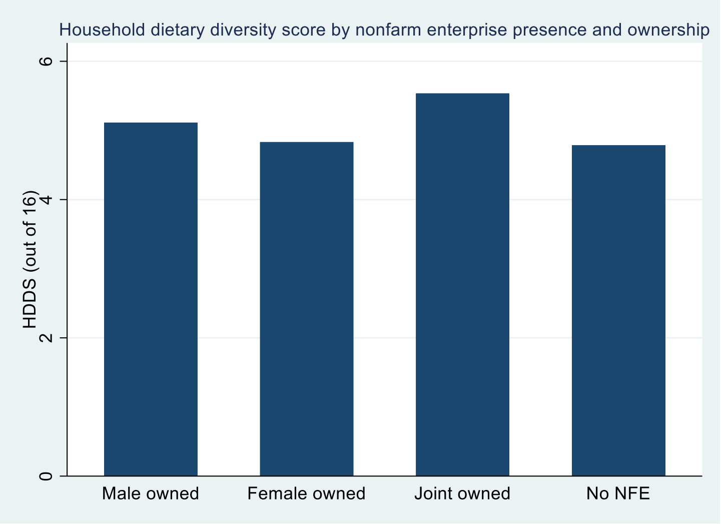
b. Figure B separates protein consumption per capita by households with and without non-farm enterprises – because we are keeping protein consumption on the x-axis, and density on the y-axis, we will separate the 2 groups of households into two lines by using two kdensity codes in the same twoway command.

   i. Code: twoway (kdensity protein_pc if any_nfe==0 & protein_pc<=100) (kdensity protein_pc if any_nfe==1 & protein_pc<=100), legend(lab(1 "No NFE") lab(2 "NFE")) xtitle("Protein consumed per capita per day (grams)") title("Daily protein consumption by presence of nonfarm enterprise", size(medlarge))

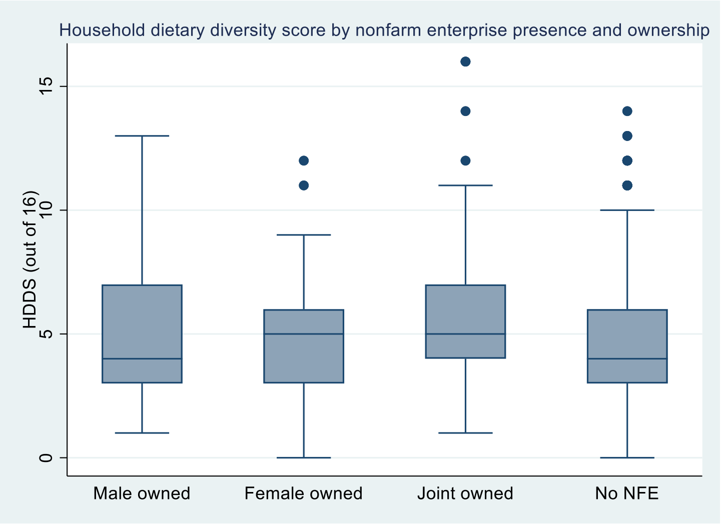


ii. What does this figure show? *The line for HHs with NFEs is above (has a higher density) than the line for "No NFE" after about 30 grams, meaning that households with NFEs are more likely to have higher protein consumption.*

2. Bar and box plots showing household dietary diversity scores (HDDS) by presence and ownership of a household enterprise
   a. Figure A shows the bar graph for this information
      i. <u>Code:</u> graph **bar** hdds, over(ownership) ytitle("HDDS (out of 16)") title("Household dietary diversity score by nonfarm enterprise presence and ownership", size(medsmall))



      ii. What does this figure show? *The average HDDS across the 4 categories of NFE ownership – Joint-owned have the highest HDDS, followed by male-owned.*
   b. Figure B shows the box plot for this information. The codes are **identical**, except for one word (changing bar to box)
      i. <u>Code:</u> graph **box** hdds, over(ownership) ytitle("HDDS (out of 16)") title("Household dietary diversity score by nonfarm enterprise presence and ownership", size(medsmall))



ii. What does this figure show? *The median, 25th and 75th percentiles, and lower and upper limits (lower limit is Q1 − 1.5 × IQR, and upper limit is Q3 + 1.5 × IQR). Dots are outliers, or anything above 1.5 x IQR. IQR stands for Inter-Quartile Range, which equals the value at the 75th percentile minus the value at the 25th percentile.*

iii. The box plots give us more information about the range and distribution of HDDS in each category.

IFPRI Contact Information

Emily Schmidt

Research Fellow

e.schmidt@cgiar.org

Rachel Gilbert

Research Analyst

r.gilbert@cgiar.org

Gracie Rosenbach

Research Analyst

g.rosenbach@cgiar.org