



Australian
National
University



Describing, Transforming, and Analyzing Data Using Stata

An Introductory Course

Harry Gimiseve, Rishabh Mukerjee, and Emily Schmidt

International Food Policy Research Institute

University of Papua New Guinea

March 2025

Port Moresby, PNG

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 18 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 – December 2023, 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 three goals:

- 1) Collect socio-economic data on rural households and ensure that analysis provides relevant information to policymakers.
- 2) Inform dialogue on food systems and household resilience and welfare.
- 2) Collaborate with diverse set of PNG departments, research organizations, and development community.

These 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

The sampling strategy was designed using “sentinel sites” as higher-level sample clusters to provide a rich and focused data set for policy analysis. To ensure a diversity of rural livelihood profiles, the country was characterized into five defined agroecological areas, based on rainfall seasonality and elevation namely:

- a. Seasonal Highlands
- b. Non-Seasonal Highlands
- c. Seasonal Lowlands
- d. Non-Seasonal Lowlands
- e. Islands

Within the five agroecological classifications, the sub-district local-level government (LLG) areas of rural PNG were used as administrative areas for randomly selecting sentinel sites. The following factors were considered in compiling comprehensive list of potential sentinel sites:

1. A rural but not remote location: Both urban and remote rural areas were excluded from potential selection. LLG areas close to provincial or other urban centers were excluded from the final set of potential sentinel sites. Similarly, LLG areas in which only a small number of people live within two hours’ travel time from the nearest motorable access point (road or river) were excluded from the selection set.
2. Sufficient population: Potential LLGs must have a minimum population of 4,000 people based on a criterion that the sample size of 150 households per site is no more than one-fifth of the household population of a potential sentinel site.
3. Elevation: Prior to selecting the sentinel sites, we distinguish lowland (areas below 1,000 meters above sea level) from highland (areas 1,000 meters or more above sea level) LLG areas based on median elevation in the LLG.
4. Rainfall seasonality: Prior to selecting sentinel sites, we distinguish areas of the country that experience large seasonal variation in rainfall (heavy to light, depending on the season) classified as seasonal, areas that experience moderate to continuously heavy rainfall throughout the year, classified as nonseasonal.

Within each LLG, 15 communities were randomly selected. Within each community, 10 households were randomly selected to be interviewed.

In addition to 10 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:

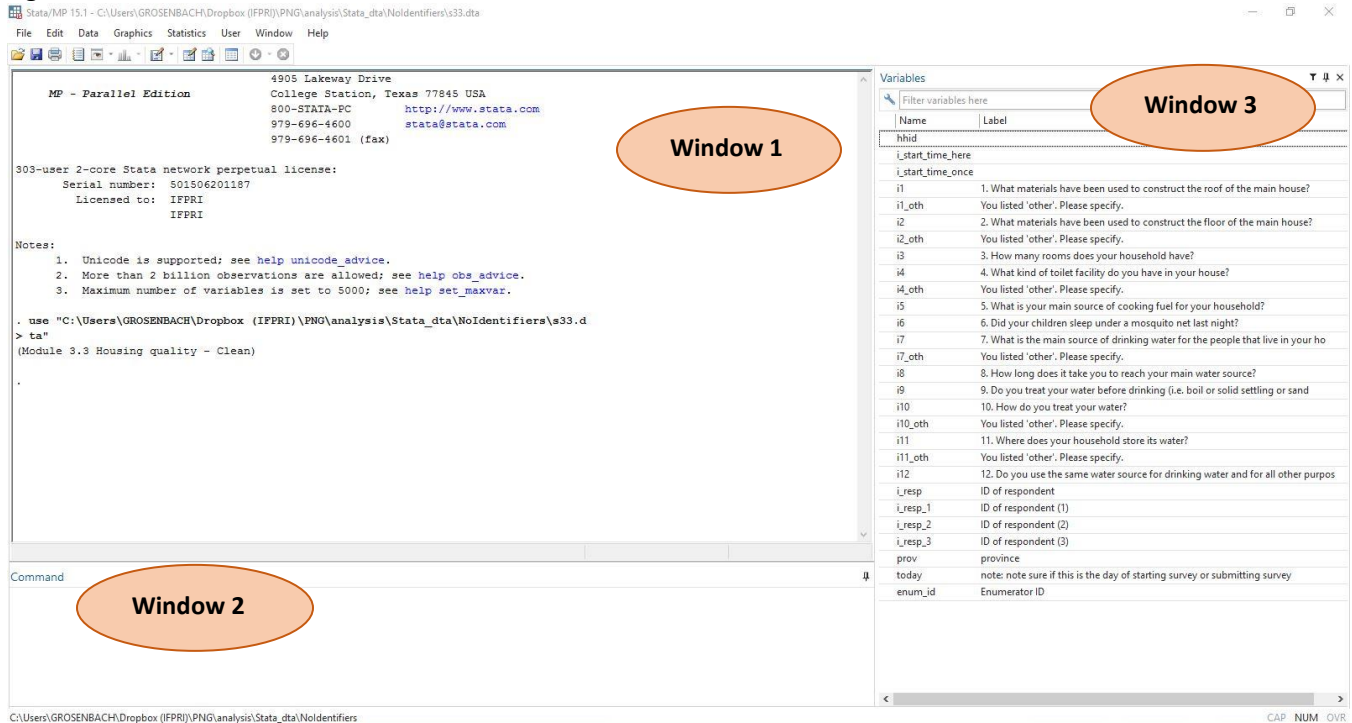
1. Household characteristics (demographics, education, migration, etc.)
2. Agricultural production (crop production, use of household labor, preference to preserve forest, and agricultural extension)
3. Household assets (production equipment and consumer durables, livestock ownership, chicken and egg 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 engagement with food market)
6. Economic shocks, household well-being, and food insecurity (shocks and coping strategies, perceptions of poverty, and recent food insecurity)
7. Dietary quality and female health (diet diversity for mothers and children under 5 years of age, women's pregnancy care, and access to health and nutrition extension)

8. Mother and 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 hh_3_4.dta**
– the following Stata screen should appear. The *main Stata screen* consists of 3 windows:

Figure 1.1: Main Stata screen



- 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 hh_3_4 corresponds to Module 3.4 in the 2023 PNG Rural Household:

Figure 1.2: Module 3.4 (Housing quality) from 2023 PNG Rural Household Survey

Section 3.4: 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_v1]			
2. What materials have been used to construct the floor of the main house? [code: floor_type_v1]			
3. What type of electricity does the main house have? [code: electricity_type_v1] [Select multiple]			
4. 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] [INTEGER]			
5. Where do members of the household usually go to the toilet? [code: toilet_type_v1]			
6. What is your main source of cooking fuel for your household? [code: cook_fuel_type_v1]			
7a. What share of the windows in your home have insect screens on them? [code: likert_share_v1]			
7b. Did your children sleep under a mosquito net last night? [code: yesno_v1] <i>Skip if there are no children under 12 in the household</i>			
8a. What is the main source of drinking water for the people that live in your household? [code: water_source_v1]			
8b-c. How long does it take to go to the water source, get water, and come back right now? [b. integer] [c. UNIT code: water_dist_v1]		Integer	Unit
9. Do you treat your water in any way to make it safe to drink? <i>if 2 ("no") → skip to Q11</i> [code: yesno_v1]			
10. What do you usually do to the water to make it safe to drink? [Select all methods used] [code: water_treat_v1]			
11. Where does your household store its water? [code: water_store_v1]			
12a. Did you wash your hands yesterday, anytime during the day or night? [code: yesno_v1]			
12b. Did your children wash their hands yesterday, anytime during the day or night? [code: yesno_v1] <i>[SKIP IF HH DOESN'T HAVE CHILDREN]</i>			
12c. What material do members of your household use with water to wash their hands? [Select all that apply] [code: handwash_v1]			

- 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*:


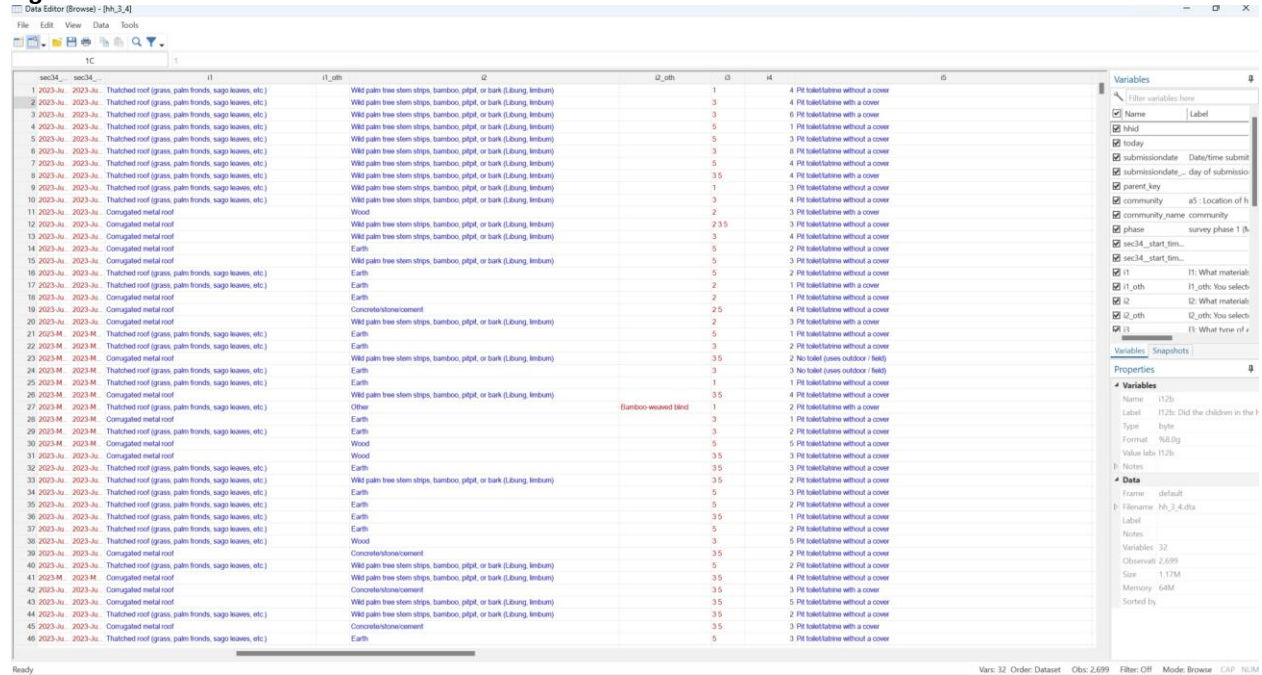
- Command Window: Type “*browse*” into Window 2 (Command Window) and press *enter*
- Drop downs: Select *Data* -> *Data Editor* -> *Data Editor (Browse)*
- Icon in top ribbon: Click this button 

Figure 1.3: Stata data browser screen



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 *i4*, 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 “6” appear in the top bar, indicating that the number 6 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. Practice: How many observations are in this dataset? “count” – 2,699

Figure 1.4: Stata command and output for “count”

```
count
2,699
```

- b. Codebook – another way to explore or describe the data; you can do it generally, or for a specific variable.
 - i. Code: `codebook [variable name]`

- ii. Practice: What kind of information do we have on drinking water? Type “**codebook i8a**” 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 i8a”

```
. codebook i8a

i8a          l8a: What is the main source of drinking water for the people that live in your

          Type: Numeric (int)
          Label: water

          Range: [1,12]          Units: 1
Unique values: 12          Missing .: 0/2,699

Examples: 4   Unprotected dug well
          6   Unprotected spring
          7   Rainwater collection
          8   Surface water
```

- a) What is the question that was asked in i8a? (The variable label)? *7. What is the main source of drinking water for the people that live in your?*
- 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 12*
- 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 unprotected spring (for example)*
- We know that there are 12 unique values, but only 4 (4 = unprotected dug well; 6=unprotected spring; 7=rainwater collection 8=surface water) are shown here as examples of what the data look like
- c. Single tabulations – tell us the frequency of each response
- i. Code: **tabulate** [variable name]
- Many Stata codes have “shorthand” versions – you only have to type a shortened version of the code and Stata will recognize the full command.
 - The shorthand for *tabulate* is “*tab*” – e.g. *tab [varname]*
 - This manual will **bold** the shorthand for each code when the code is introduced (see above for **tabulate**)
- ii. Practice: How many households get their drinking water from a protected dug well? “*tabulate i8a*” or “*tab i8a*”. This output tells us:

Figure 1.6: Stata command and output for “tab i8a”

. tab i8a			
18a: What is the main source of drinking water for the people that live in your	Freq.	Percent	Cum.
Public tap/standpipe	129	4.78	4.78
Tube well, borehole	86	3.19	7.97
Protected dug well	184	6.82	14.78
Unprotected dug well	206	7.63	22.42
Protected spring	282	10.45	32.86
Unprotected spring	282	10.45	43.31
Rainwater collection	691	25.60	68.91
Surface water	777	28.79	97.70
Bottled water	3	0.11	97.81
Piped to neighbor	16	0.59	98.41
Piped into yard or plot	29	1.07	99.48
Piped into dwelling	14	0.52	100.00
Total	2,699	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 dug well? *184 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 dug well? *6.82% of all households surveyed*
- c) “Cum.” – the cumulative percent of the answers across all of the observations (HHs)
 - Question: In Figure 1.6, how many people responded, “surface water”? *777*
- iii. Now we want to look at - what did the households that responded “other” write in when asked about floor material of their house?
 - a) What kind of a variable do you think i2_oth is? Numeric, categorical, or string? *String*
 - b) How can we find out? *“codebook i2_oth” OR see what color it is when we “browse”*
 - c) Practice: Type *“codebook i2_oth”* and see if we are right
 - d) Practice: Type *“tab i2_oth”* to see what responses were typed in

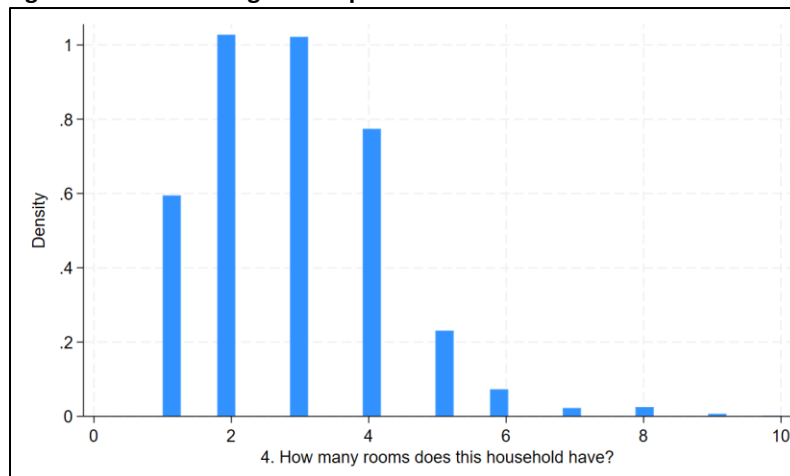
Figure 1.7: Stata command and output for “tab i2_oth”

```
. tab i2_oth
```

i2_oth: You selected 'other', please specify.			
-----	Freq.	Percent	Cum.
BAMBU	1	1.22	1.22
Bamboo	13	15.85	17.07
Bamboo Flooring	1	1.22	18.29
Bamboo Stem	1	1.22	19.51
Bamboo floor	1	1.22	20.73
Bamboo flooring	10	12.20	32.93
Bamboo-weaved blind	3	3.66	36.59
Bambu	1	1.22	37.80
Betelnut palm	1	1.22	39.02
Black Palm Skin and Plywood	1	1.22	40.24
Black palm	2	2.44	42.68
Bline	1	1.22	43.90
Both wood and limbum	1	1.22	45.12
Flatten or adzed wood.	1	1.22	46.34
Ground	2	2.44	48.78
Gum tree bark.	1	1.22	50.00
Iron	1	1.22	51.22
Layer Stones	1	1.22	52.44
Metal post	1	1.22	53.66
On the Ground	1	1.22	54.88
Pandanus	1	1.22	56.10
Ply wood	1	1.22	57.32
Plywood	3	3.66	60.98
River stones	1	1.22	62.20
Rough milled timber flooring	1	1.22	63.41
Rough sawn timber	1	1.22	64.63

- iv. Question: How many households treat their water? “tab i9” – 445 households (16.49%) treat their water
- v. Question: How many rooms do households have?
 - a) “tab i4” – 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: **histogram** [varname]
 - ii. Practice: Let’s visualize the distribution of the data for variable i4 (number of rooms in each household) by typing “hist i4”

Figure 1.8: Stata histogram output for “hist i4”



- a) Question: What is the mode (most common answer)? *2 rooms*
 - This figure suggests that most houses have between 1-5 rooms, while very few have more than 5.
- e. Summarize – outputs the number of observations, average (mean), standard deviation, minimum, and maximum of a numeric variable
 - i. Code: *summarize [varname]*
 - ii. Practice: What if we want to know the average number of rooms in households in our sample? “sum i4”

Figure 1.9: Stata command and output for “sum i4”

. sum i4					
Variable	Obs	Mean	Std. dev.	Min	Max
i4	2,699	2.86884	1.361818	1	10

- iii. Challenge: What other code have we used before that gave us the minimum and maximum values of a variable? *codebook*
- iv. What is the average number of rooms in the households in our sample? *2.86 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 10 even though the average is 2.86, 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. Practice: What is the median number of rooms? “*summarize i4, detail*” – this output shows the following:

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

```
. sum i4, detail
```

4. How many rooms does this household have?

Percentiles		Smallest		
1%	1	1		
5%	1	1		
10%	1	1	Obs	2,699
25%	2	1	Sum of wgt.	2,699
50%	3		Mean	2.86884
		Largest	Std. dev.	1.361818
75%	4	9		
90%	4	9	Variance	1.854548
95%	5	9	Skewness	.8535383
99%	7	10	Kurtosis	4.495453

- 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: `tabulate [varname1] [varname2]`
 - ii. Practice: Do households who get their water piped into their dwelling treat it before drinking? “`tab i8a i9`”

Figure 1.11: Stata command and output for “tab i8a i9”

```
. tab i8a i9
```

8a. What is the main source of drinking water for the people that live in your h	9. Do you treat your water in any way to make it safe to drink?		Total
	Yes	No	
Public tap/standpipe	21	108	129
Tube well, borehole	23	63	86
Protected dug well	28	156	184
Unprotected dug well	36	170	206
Protected spring	66	216	282
Unprotected spring	33	249	282
Rainwater collection	142	549	691
Surface water	91	686	777
Bottled water	1	2	3
Piped to neighbor	1	15	16
Piped into yard or pl	0	29	29
Piped into dwelling	3	11	14
Total	445	2,254	2,699

- 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 from rainwater collection, and do *not* treat it? *549 households*
- iii. What if we want to know what *percent* of households get their water from rainwater collection 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 from rainwater collection and do not treat it – “*tab i8a i9, cell*”

Figure 1.12: Stata command and output for “tab i8a i9, cell”

8a. What is the main source of drinking water for the people that live in your h	9. Do you treat your water in any way to make it safe to drink?		Total
	Yes	No	
Public tap/standpipe	21 0.78	108 4.00	129 4.78
Tube well, borehole	23 0.85	63 2.33	86 3.19
Protected dug well	28 1.04	156 5.78	184 6.82
Unprotected dug well	36 1.33	170 6.30	206 7.63
Protected spring	66 2.45	216 8.00	282 10.45
Unprotected spring	33 1.22	249 9.23	282 10.45
Rainwater collection	142 5.26	549 20.34	691 25.60

- 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 from rainwater collection and do *not* treat it? *20.34% of the households in the sample get their water piped into their dwelling and do not treat it.*
- v. Practice: What if we want to know the percentage of households **who get their water from rainwater collection** do not treat it? – “*tab i8a i9, row*”

Figure 1.13: Stata command and output for “tab i8a i9, row”

8a. What is the main source of drinking water for the people that live in your h	9. Do you treat your water in any way to make it safe to drink?		Total
	Yes	No	
Public tap/standpipe	21 16.28	108 83.72	129 100.00
Tube well, borehole	23 26.74	63 73.26	86 100.00
Protected dug well	28 15.22	156 84.78	184 100.00
Unprotected dug well	36 17.48	170 82.52	206 100.00
Protected spring	66 23.40	216 76.60	282 100.00
Unprotected spring	33 11.70	249 88.30	282 100.00
Rainwater collection	142 20.55	549 79.45	691 100.00

- 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 i8a 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 get their water from rainwater collection** do not treat their water? *79.45% of households get their water from rainwater collection and 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 from rainwater collection? – “tab i8a i9, column”

Figure 1.14: Stata command and output for “tab i8a i9, col”

8a. What is the main source of drinking water for the people that live in your h	9. Do you treat your water in any way to make it safe to drink?		Total
	Yes	No	
Public tap/standpipe	21 4.72	108 4.79	129 4.78
Tube well, borehole	23 5.17	63 2.80	86 3.19
Protected dug well	28 6.29	156 6.92	184 6.82
Unprotected dug well	36 8.09	170 7.54	206 7.63
Protected spring	66 14.83	216 9.58	282 10.45
Unprotected spring	33 7.42	249 11.05	282 10.45
Rainwater collection	142 31.91	549 24.36	691 25.60
Surface water	91 20.45	686 30.43	777 28.79
Bottled water	1 0.22	2 0.09	3 0.11

- The option “column” (the shorthand is “col”) will tell us the column percentages – the percentages *in each column* will sum to 100.
- In this specific code (where i8a 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
- What percentage of households **who do not treat their water** get their water from rainwater collection? *24.36% of households who do not treat their water get their water from rainwater collection.*

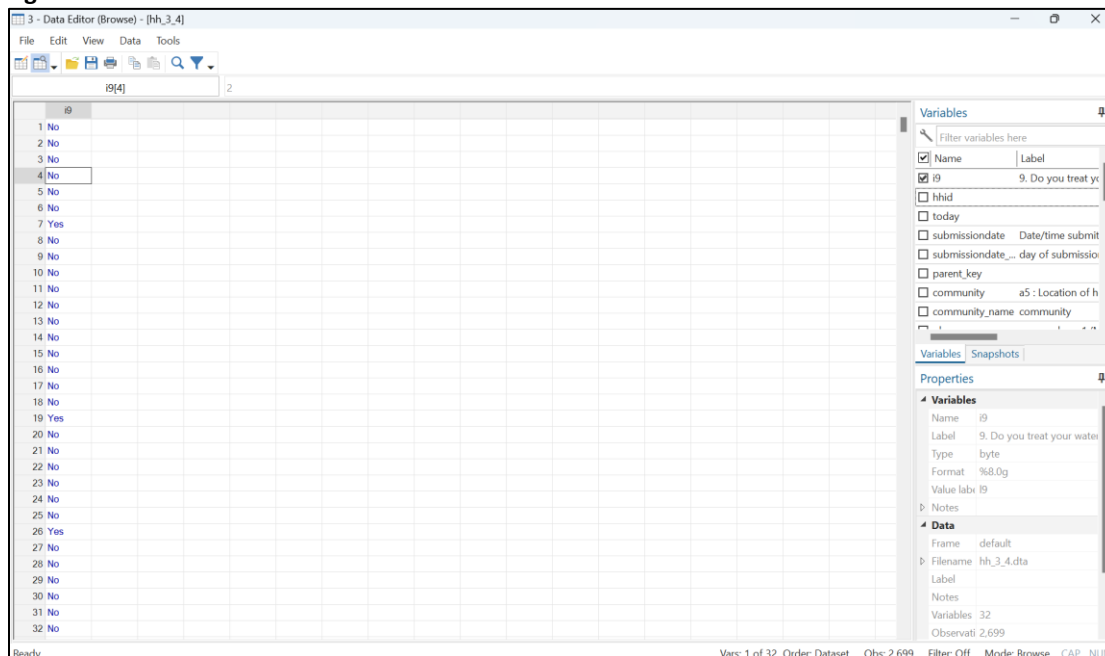
4. Using “if” – a way to limit your output to certain observations that meet your defined criteria

- 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 from rainwater collection?)
- 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 “nolab” 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”



- 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 from rainwater collection?) - “tab i8a if i9==2”

Figure 1.16: Stata command and output for “tab i8a if i9==2”

```
. tab i8a if i9 == 2
```

8a. What is the main source of drinking water for the people that live in your h	Freq.	Percent	Cum.
Public tap/standpipe	108	4.79	4.79
Tube well, borehole	63	2.80	7.59
Protected dug well	156	6.92	14.51
Unprotected dug well	170	7.54	22.05
Protected spring	216	9.58	31.63
Unprotected spring	249	11.05	42.68
Rainwater collection	549	24.36	67.04
Surface water	686	30.43	97.47
Bottled water	2	0.09	97.56
Piped to neighbor	15	0.67	98.23
Piped into yard or plot	29	1.29	99.51
Piped into dwelling	11	0.49	100.00
Total	2,254	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, 24.36% get their water from rainwater collection.
- b) Notice that the Total number under “Freq.” (the number of observations included in this output) is smaller than the number of our dataset (this output shows only 2,254 households while our dataset has 2,699). This is because 2,254 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? *i4*
 - 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 i4>3” OR “tab i1 if i4>=4”
 - iv. Question: How about the type of roof for households with *exactly 3 rooms*? “tab i1 if i4==3”
 - v. Question: How about the type of roof for households with *3-5 rooms*? “tab i1 if i4>=3 & i4<=5” OR “tab i1 if i4>2 & i4<6”
 - vi. Question: How about the type of roof for households with *less than 3 rooms or greater than 5 rooms*? “tab i1 if i4<3 | i4>5” OR “tab i1 if i4<=2 | i4>=6”
 - vii. Question: Which households more frequently have a corrugated metal roof?
 - a) Households with more than 3 rooms? Or
 - b) households with 3 rooms or less? What codes will you run to show this?

- a) *“tab i1 if i4>3” and “tab i1 if i4<=3”*
- b) *58.27% of households with more than 3 rooms have a corrugated metal roof, but only 39.65% of households with 3 rooms or less have a corrugated metal roof. So, households with more than 3 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”

```
. count
16,192
```

- i. How many observations are in this dataset? *16,192*
 - 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 ?	Freq.	Percent	Cum.
Male	8,459	52.24	52.24
Female	7,733	47.76	100.00
Total	16,192	100.00	

- iii. Question: How many males are in this dataset? *8,459*
- iv. Question: What percentage of the people in this dataset are males? *52.24%*
- 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 – “4a. 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 a4a”)*

Figure 2.3: Stata command and output for “sum a4a”

```
. sum a4a
```

Variable	Obs	Mean	Std. dev.	Min	Max
a4a	16,192	125.095	279.0221	0	888

- iii. What is the average age of people in this dataset? *125.095. Average age appears to be too high!*
- iv. Look at all of the output from “sum a4” – where is the problem? *Maximum age is an impossibly large 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 a4a”*

Figure 2.4: Stata command and output for “codebook a4a”

```
. codebook a4a
```

a4a	4a. What is the age of
-----	------------------------

```

Type: Numeric (int)
Range: [0,888]
Unique values: 97
Mean: 125.095
Std. dev.: 279.022
Units: 1
Missing.: 0/16,192
Percentiles: 10% 25% 50% 75% 90%
              5 10 22 45 888
  
```

- b) How many missing values are there currently in a4a? *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 a4a? *"codebook a4a" – 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 a4a 888=."*
 - ii. Code: *replace [varname]=[new_value] if [varname]==[old_value]*
 - Practice: *"replace a4a=. if a4a==888"*
 - iii. Question: After using one of these codes, how many missing values are there now for a4? *"codebook a4a"; 1904 missing values*
 - iv. Now that we know that (.) means missing, we can also type *"tab a4, missing"* to see how many missing values we have. This tabulate option *"missing"* includes the missing values in the tabulation

Figure 2.5: Stata command and output for "codebook a4a", after recoding 888 values to missing

```
. codebook a4a
```

a4a		4a. What is the age of
------------	--	-------------------------------

```

Type: Numeric (int)
Range: [0,106]
Unique values: 96
Mean: 23.4313
Std. dev.: 18.0899
Percentiles: 10% 25% 50% 75% 90%
              4   9   18   34   51
Units: 1
Missing .: 1,904/16,192

```

- v. Question: Now what is the average age in the dataset?
 - a) *"sum a4a"*

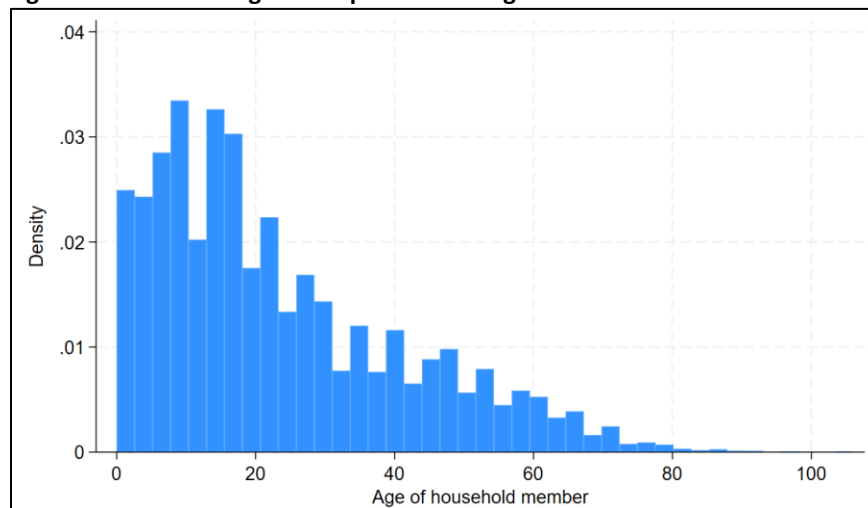
Figure 2.6: Stata command and output for "sum a4a", after recoding 888 values to missing

```
. sum a4a
```

Variable	Obs	Mean	Std. dev.	Min	Max
a4a	14,288	23.43127	18.08986	0	106

- b) *Now the average (mean) age is 23.43. This average does not take into account the 1,904 missing values (Obs is now 14,288 instead of 16,192)*
- 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: `rename [old_varname] [new_varname]`
 - b) Practice: “`rename a4a age`” – renames the variable to “age” (more intuitive than “a4a”)
 - ii. Changing/adding a variable label (see in Window 3 – the Variable Window)
 - a) Code: `label variable [varname] ["label"]`
 - b) Practice: “`label variable 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 “a4a”, 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*
 - c) *The sample comprise of more young population*

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 percentage 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: **generate** [new_varname]=[value]
 - ii. Practice: **“generate 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		//There are no missing values	
child		(unlabeled)	
Type: Numeric (float)			
Range: [0,1]		Units: 1	
Unique values: 2		Missing .: 0/16,192	
Tabulation: Freq. Value			
	10,132	0	
	6,060	1	

- vi. How many missing values are there for our “child” variable? *0 – this is a problem! We changed the ages of 1,904 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. Question: 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]
Unique values: 2
Units: 1
Missing ..: 1,904/16,192

Tabulation: Freq. Value
            8,228 0
            6,060 1
            1,904 .
```

- 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. Question: How can we assign a variable label to our new variable? *“label variable 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) Code: *label define [value_label_definition] # [“label”] # [“label”]*
 - b) Practice: *“label define 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) Code: *label values [varname] [value_label_definition]*
 - b) Practice: *“label values child no_yes”* – applies our newly created “no_yes” value label definition to our variable “child”
- v. Let’s take one final look at the overview of our new variable, and see if we labeled everything – *“codebook child”*

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

```
. codebook child

child Is the household member 15 years or younger?

Type: Numeric (float)
Label: no_yes
Range: [0,1]
Unique values: 2
Units: 1
Missing ..: 1,904/16,192

Tabulation: Freq. Numeric Label
            8,228 0 No
            6,060 1 Yes
            1,904 .
```

- vi. How many children are in our dataset? *6,060*
- vii. Challenge: 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
```

b8 : What is the current PRIMARY activity / occupation for \${b_name2}? -----	Freq.	Percent	Cum.
Any agricultural (farmer, gardener, sub	125	10.37	10.37
Forestry or fishery labor	1	0.08	10.46
Transport and storage laborers	1	0.08	10.54
Services and sales worker	1	0.08	10.62
Security guard	1	0.08	10.71
Student	1,041	86.39	97.10
Domestic work/housewife	7	0.58	97.68
Unemployed and NOT looking for work	6	0.50	98.17
Unemployed and ACTIVELY looking for wor	1	0.08	98.26
Community leader (counselor, peace medi	1	0.08	98.34
Not applicable	18	1.49	99.83
Other	2	0.17	100.00
Total	1,205	100.00	

b) 86.39% of children 15 and under have “student” listed as their primary activity

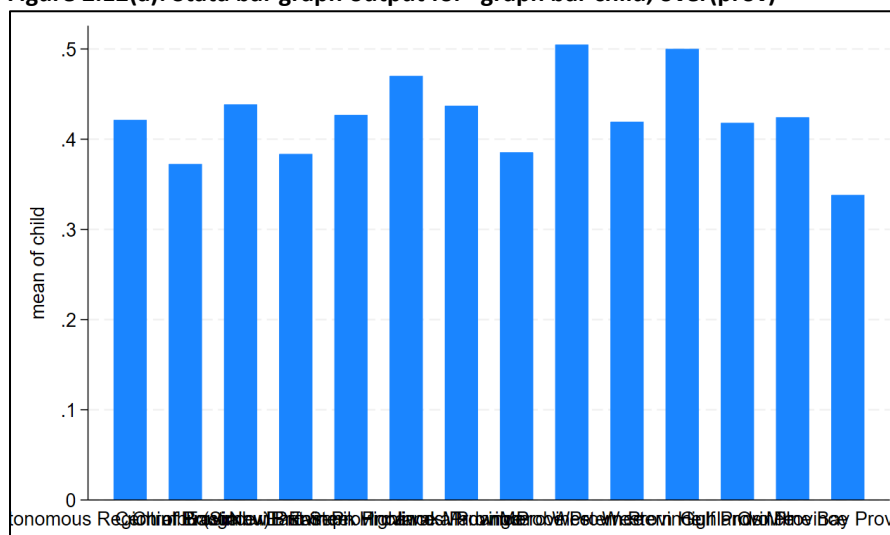
viii. Challenge: Which province has the highest percentage of children? Let’s use a bar graph to find out!

a) Code: `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(a): Stata bar graph output for “`graph bar child, over(prov)`”

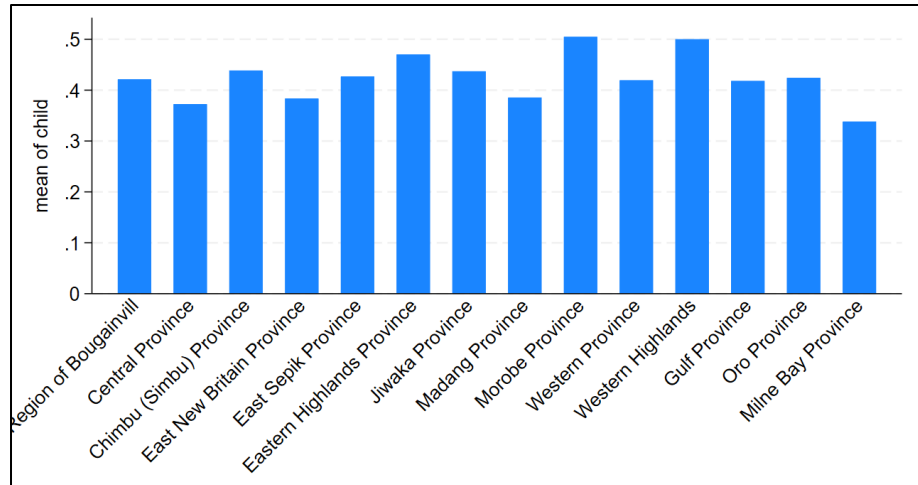


c) Notice that the bar labels on x-axis are on top of each other. To produce a neat graph, we can use the following code to avoid the overlap:

`graph bar child, over(prov, label(angle(45)))`

Figure 2.12(b): Stata bar graph output for “graph bar child, over (prov, label(angle(45)))”

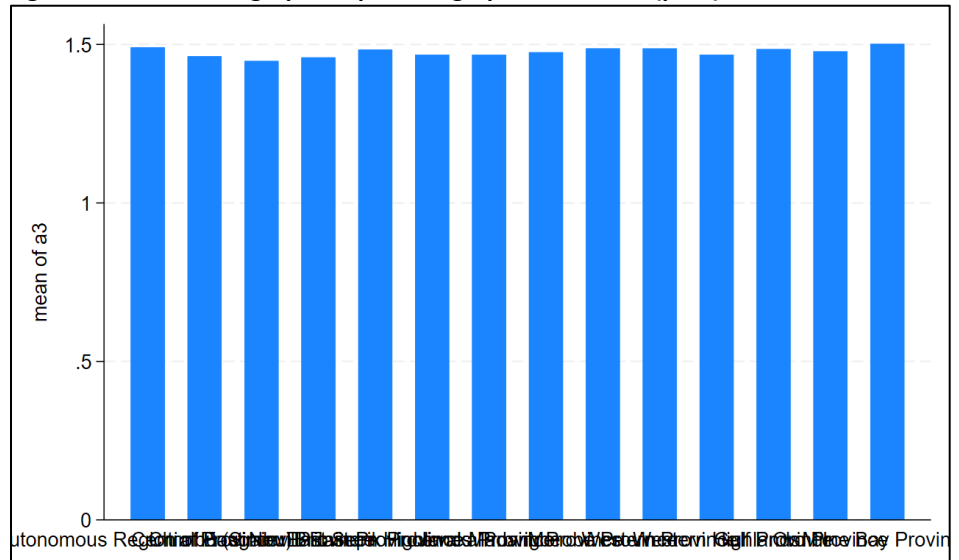
1



- d) 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*
- e) *Morobe 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: b3

      Range: [1,2]
Unique values: 2                    Units: 1
                                      Missing .: 0/16,192

      Tabulation: Freq.  Numeric  Label
                   8,459      1  Male
                   7,733      2  Female
```

- d) The value labels are 1 and 2, instead of 0 and 1. Hence, 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.
- “gen woman=.”* – creates a new variable named *woman*, and sets all values to missing
 - “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)
 - “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)
 - “lab var woman “Is the household member a woman?””* – labels the new variable
 - “lab val woman no_yes”* – applies the value label definition that we created earlier, named “no_yes” to our new variable *woman*
 - Let’s look at our new variable – “codebook *woman*”

Figure 2.15: Stata command and output for “codebook woman”

```
woman                                Is the household member a woman?

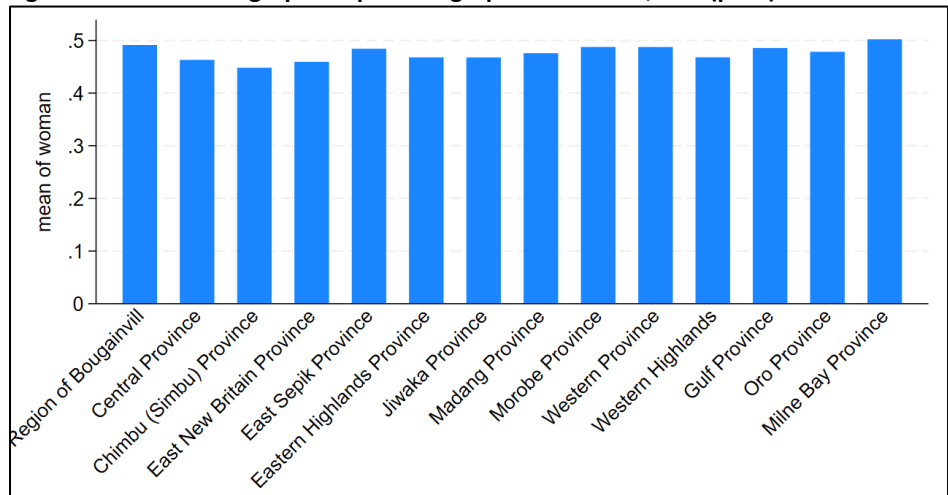
      Type: Numeric (float)
      Label: no_yes

      Range: [0,1]
Unique values: 2                    Units: 1
                                      Missing .: 0/16,192

      Tabulation: Freq.  Numeric  Label
                   8,459      0  No
                   7,733      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) Question: Does it show prevalence now? *Yes!*
- i) Question: Which province has the highest prevalence of women in the sample? *Milne Bay Province*

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	
<code>chi2</code>	report Pearson's chi-squared
<code>exact[<i>(#)</i>]</code>	report Fisher's exact test
<code>gamma</code>	report Goodman and Kruskal's gamma
<code>lrchi2</code>	report likelihood-ratio chi-squared
<code>taub</code>	report Kendall's tau-b
<code>V</code>	report Cramér's V
<code>cchi2</code>	report Pearson's chi-squared in each cell
<code>column</code>	report relative frequency within its column of each cell
<code>row</code>	report relative frequency within its row of each cell
<code>clrchi2</code>	report likelihood-ratio chi-squared in each cell
<code>cell</code>	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*) *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. Practice: *“collapse (count) pid (min) age (mean) child woman, by(hhid)”*
- vi. How many observations are there now in the dataset? *“count” – 2,699*
- vii. Now let’s look at what happened to Window 3 (Variable Window) after the collapse

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

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

- a) Variable names remained the same
 - b) Variable labels now show only the summary statistics and the name of the variable
- viii. Practice: Let’s create more meaningful variable names and labels
- a) What does the variable *pid* tell us now?
 - The count of people in the household = the household size
 - *“rename 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
 - *“rename 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
 - *“rename 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
 - *“rename woman perc_women”*
 - *“lab var perc_women “Percent of women in the household””*
- ix. Question: What is the average household size?
- a) *“sum hhsize”*

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

. sum hhsz

Variable	Obs	Mean	Std. dev.	Min	Max
hhsz	2,699	5.999259	2.603812	1	22

b) *5.9 people per household*

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

a) *“sum age_youngest”*

b) *8.7 years old*

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

a) *“sum perc_children”*

b) *40.8% children*

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

a) *“sum perc_women”*

b) *47.2% 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 (2,699)

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=2,699)		hh_3_4 (n=2,699)		merged data (n=2,699)		
hhid	age_youngest	hhid	i1	hhid	age_youngest	i1
0100RSPN	11	0100RSPN	metal roof	0100RSPN	11	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 2,699 unique households and we are merging with another dataset with 2,699 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 “copy the path where you have saved hh_3_4.dta in your system”

Figure 2.25: Stata command and output for “merge 1:1 hhid using ...”

```
. merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh Mukerjee\Stata_
> raining_PNG2025\Stata_training_png_March2025\Datasets\hh_3_4.dta" //Merge the dataset with
> the s33 dataset from Lesson 1, matching based on hhid (household ID)
(label community already defined)
```

Result	Number of obs
Not matched	0
Matched	2,699 (_merge==3)

- d) The output after the merge tells us how many observations were and were not matched. How many were matched? *2,699 (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 2,700 observations in the collapsed household roster, and merged it to 2,699 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 2,699 observations in the collapse household roster, and merged it to 2,700 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 perc_children if i1==1” – on average in households with thatched roofs, 43% of the household members are aged 15 years or younger
 - “sum perc_children if i1==2” – on average in households with corrugated metal roofs, 38% 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: `bysort [varname1]: stata_command [varname2]`
 - “`bysort 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 (grass, palm fronds, sago leaves, etc.)
```

Variable	Obs	Mean	Std. dev.	Min	Max
perc_child~n	1,387	.43039	.2671725	0	1

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

Variable	Obs	Mean	Std. dev.	Min	Max
perc_child~n	1,209	.3842051	.2353648	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)

Hh_3_4 (n=2,699)		s11_long (n=16,192)			merged data (n=6,057)			
hhid	i1	hhid	a_pid	a4	hhid	a_pid	a4	i1
0100RSPN	metal roof	0100RSPN	1	M	0100RSPN	1	M	metal roof
		0100RSPN	2	F	0100RSPN	2	F	metal roof
		0100RSPN	3	F	0100RSPN	3	F	metal roof
		0100RSPN	4	F	0100RSPN	4	F	metal roof
		0100RSPN	5	M	0100RSPN	5	M	metal roof
		0100RSPN	6	M	0100RSPN	6	M	metal roof
		0100RSPN	7	M	0100RSPN	7	M	metal roof

- a) Let’s now start over and open the household characteristics dataset (hh_3_4.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 h_3_4 data to many (m) household observations in the s11_long data set.*
- d) Practice: “merge 1:m hhid using “copy the folder path where you saved s11_long.dta””

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

Result	Number of obs
Not matched	0
Matched	16,192 (_merge==3)

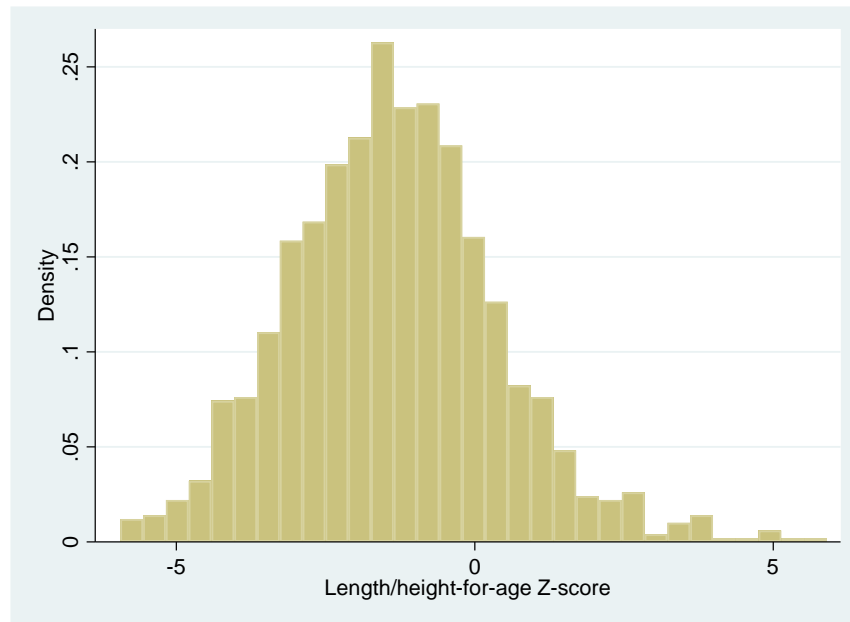
- Did all of the observations match? *Yes, all 16,192 matched*
 - So what are the values of the _merge variable? *All are _merge==3*
 - Question: How can we double check the values of the _merge variable? *“tab _merge”*
- e) Question: How many **people** in our dataset live in a house with a **thatched 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
 - *8,328 people (51.43% 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”*
 - *3,961 women live in a house with a thatched roof (51.22% 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? *1334*
 - ii. What level is this data? (What does each observation represent?) *household level (each observation is one unique household, see hhid: household id)*
 - 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” – 92 households (6.90%)*
 - 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” – tells us the type of material used to construct the floor of the main house); most individuals (568) have a floor made from Wild palm tree stem/strips, bamboo and pitipit/blind; followed by wood (494)*
 - iv. The Height for Age Z (HAZ) score of children under 5 years is going to be our key outcome variable today. This variable was constructed in Module 8.1 in the survey – let’s look at that module.
 1. To better understand nutrition and health outcomes of children in the survey sample, the 2023 PNG Rural Household Survey collected anthropometric measurements (height and weight) for all children under 5 years old in the survey sample. In doing so, the survey data allow for estimation of under-5 child statistics on undernutrition including stunting, wasting, and underweight indicators.
 2. We will focus on household and individual factors that are associated with improved child growth or reduced stunting.
 3. Stunting, or low height forage, is a measurement of long-term, chronic malnutrition and is associated with short and long-term health and development consequences.
 4. A child is considered stunted if his or her Height for Age Z scores (calculated using height measurements and age) is more than 2 standard deviations below the global standards set by World Health Organization.
 5. To remove biologically impossible scores (extreme outliers), a child haz-score greater than 6 or less than -6 was changed to missing, in accordance with World Health Organization’s (WHO) standards.
 6. What do you think the minimum and maximum values are for the variable haz? *-5.94 and 5.9*

7. Question: How can you check the minimum, maximum, and average values for this variable?
 - a. "codebook haz"
 - b. "sum haz"
 - c. *The minimum is -5.94, the maximum is 5.9, and the mean is -1.35*
 - d. *This means that the mean haz-score of children under 5 years in the surveyed area is not stunted. In the 2023 Rural Household Survey Report, a child aged 5 years or less is considered stunted if they have an extreme deviation from WHO growth standard, that is, their haz score is below -2 standard deviation from the 0 (WHO's standard).*
8. Question: How can we look at a figure showing the distribution of this variable?
 - a. "hist haz"

Figure 3.1: Stata histogram output for "hist haz"



- b. What can we learn from this histogram? *Majority of children aged 5 or less in the surveyed areas have haz scores that falls below 0 (WHO growth standard). A negative haz score indicates that a child is shorter than WHO growth standard. (We have a bell-shaped curve)*
- c. We have some continuous data in this dataset: for example, landholdings and household size (hhsz)
- v. Household size – we calculated this variable in Lesson 2 (it is the count of the number of people from the roster in each household)
- vi. 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)

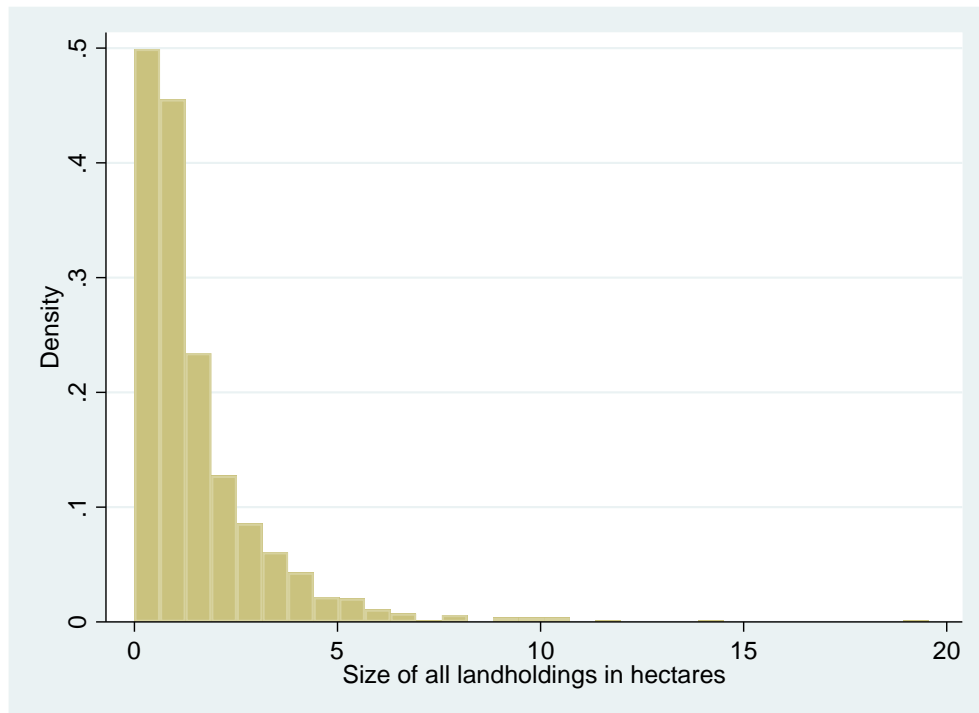
2. Confusion about the questions (e.g. a respondent double counted some of his/her agricultural plots, thinking that was what the data collector was asking)
 3. Best guesses (e.g. a respondent does not know how many hectares his/her plot is, so he/she gives a best guess – says a plot is 5 hectares when in reality it is 2)
- d. 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.
- e. What’s a good way to check the *landholdings* variable for outliers?
- i. Practice: “*sum landholdings, det*”

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

<code>. sum landholdings, det</code>				
Size of all landholdings in hectares				
	Percentiles	Smallest		
1%	0	0		
5%	.0162	0		
10%	.0324	0	Obs	1,334
25%	.3743	0	Sum of Wgt.	1,334
50%	1.05		Mean	1.455915
		Largest	Std. Dev.	1.688366
75%	2	10.7		
90%	3.5	11.9	Variance	2.850578
95%	4.55	14.4	Skewness	3.122489
99%	8.0824	19.55	Kurtosis	20.5761

1. First, check the mean and median: *the mean is 1.46 hectares and the median is 1.05 – 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 19.55!*
 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. Practice: “*hist 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 positive skew also shows the outliers that we may want to change
- f. Everyone has a different preference for how to handle outliers. For now, let’s say that all values greater than the 99th percentile (8.0824 – we know from “sum landholdings, det”) should be changed to the median (1.05). The 99th percentile tells us that 99% of all observations fall below, or are less than, 8.0824, 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 (8.0824). How might we check this? (HINT: We can combine “if” with one of the commands that we know). *“count if landholdings>8.0824” – 12 observations*

Figure 3.4: Stata command and output for “count if landholdings>8.0824”

```
. count if landholdings>8.0824
12
```

- ii. Question: How might we change these 12 observations to the median (1.05)? *“replace landholdings=1.05 if landholdings>8.0824”*

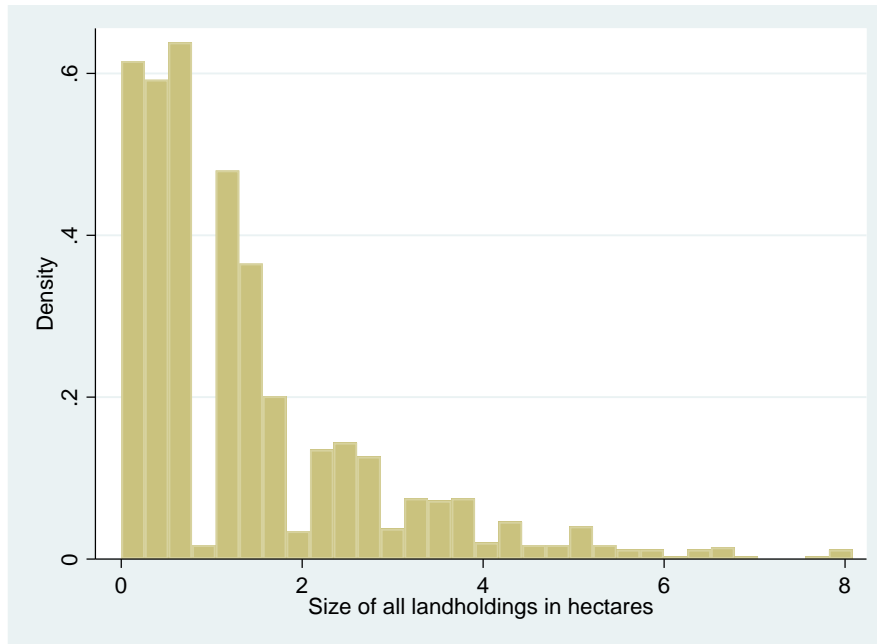
Figure 3.5: Stata command and output for “replace landholdings=1.05 if landholdings>8.0824”

```
. replace landholdings=1.05 if landholdings>8.0824
(12 real changes made)
```

- iii. Stata tells us how many observations it changed with our command in the output. How many were changed? *12 – 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 8 hectares, instead of more than 20 like the first one (Figure 3.3)

2. Correlations

- a. Let's see how correlated our variable of interest (*haz*) is with our cleaned continuous variable for real daily consumption expenditure i.e. *ex_r* (a proxy for income):
 - i. Code: `pwcorr [varname] [varname]`
 - ii. Practice: "`pwcorr haz ex_r`"

Figure 3.7: Stata command and output for "pwcorr haz ex_r"

<code>. pwcorr haz ex_r</code>		
	haz	ex_r
haz	1.0000	
ex_r	0.1194	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 *haz* and *ex_r*? *0.1194*
 2. Is this strong or weak? *Weak*

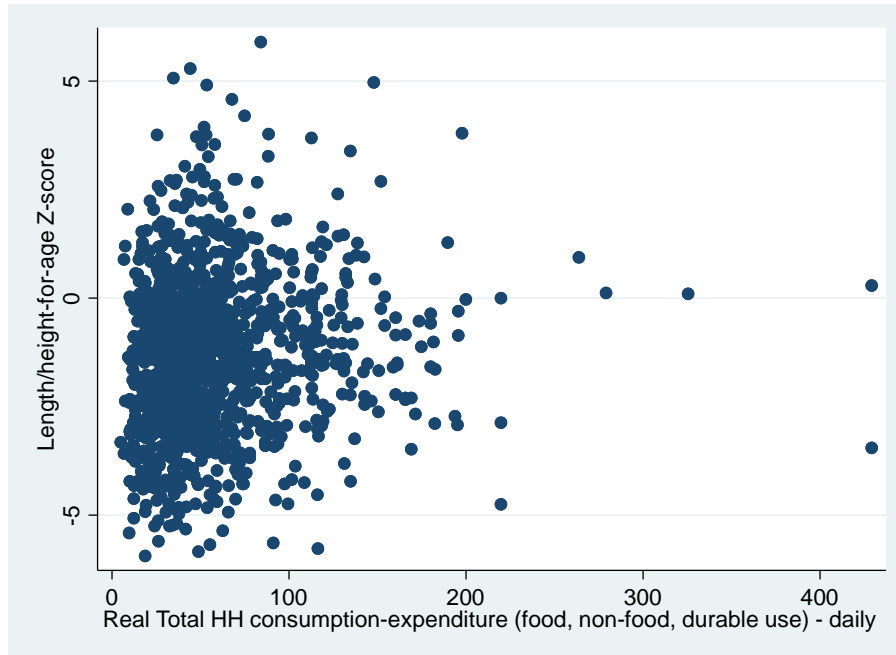
- v. Also remember that a positive correlation coefficient means that as one variable increases, the other increases; and a negative correlation coefficient means that as one variable increases, the other decreases.
 1. Is the correlation coefficient between haz (which is a child stunting indicator) and ex_r (real daily consumption expenditure) 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 higher consumption expenditure are associated with children being less likely to be too short for their age.*
- 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 haz ex_r, sig*”

Figure 3.8: Stata command and output for “pwcorr haz ex_r, sig”

<code>. pwcorr haz ex_r, sig</code>		
	haz	ex_r
haz	1.0000	
ex_r	0.1194 0.0000	1.0000

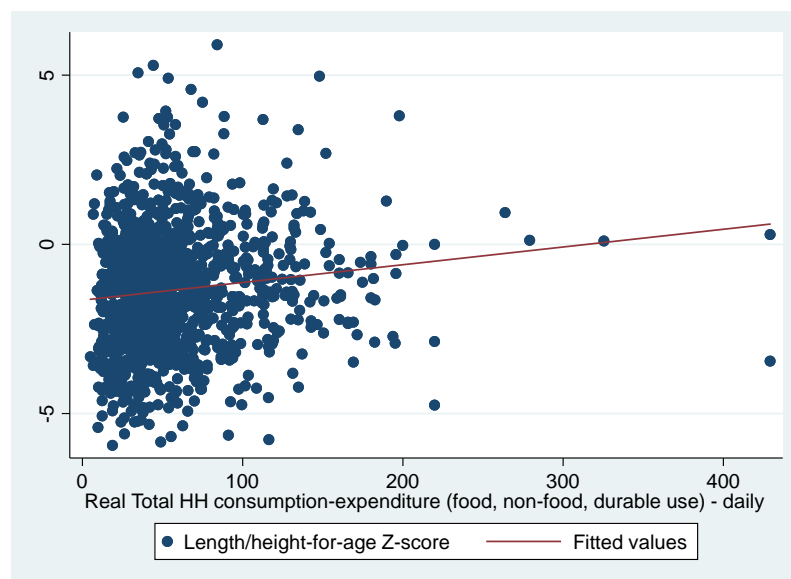
- ii. This new number below the correlation coefficient is the p-value:
 1. The p-value tells us the percent with which we are confident that the two variables are associated. You find this percent by subtracting the p-value by 1 (e.g., 1-0.000 equals about 1 – so we can say with 100% confidence that this correlation is significant)
 2. The usual p-value cut-offs for stating that something is significant are:
 - a. 90% confidence (p-value<0.1)
 - b. 95% confidence (p-value<0.05)
 - c. 99% confidence (p-value<0.01)
 3. What is the p-value for this correlation? *0.0000*
 4. Is this significant? *yes*
- c. Another way to look at this relationship between two continuous variables is to create a scatter plot
 - i. Code: *scatter [varname] [varname]*
 - ii. Practice: “*scatter haz ex_r*”

Figure 3.9: Stata scatterplot output for “scatter haz ex_r”



- iii. It looks like there may be a positive relationship, as the values for ex_r (daily real consumption expenditure) increase, the height-for-age z-score for children under 5 years also increases.
- iv. We can also create a 'line of best fit' to see how positive or negative and strong the relationship is
- d. Let's add a line of best fit to this figure, to better see the trend and how strongly associated the variables are
 - i. Code: `tway (scatter [varname1] [varname2]) (lfit [varname1] [varname2])`
 - ii. Practice: `"tway (scatter haz ex_r) (lfit haz ex_r)"`

Figure 3.10: Stata scatterplot output for `"tway (scatter haz ex_r) (lfit haz ex_r)"`



- iii. You can see the very 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. haz) is significantly different between two groups in the data (e.g. whether or not a child aged 2 to 5 years has eaten any meat, poultry or fish served on his/her plate in the last 24 hours)
- b. Code: `ttest [varname], by([groupvar])`
- c. Practice: `"ttest haz, by(meat_poul_fish_C)"` – output tells us:

Figure 3.11: Stata command and output for "ttest haz, by(meat_poul_fish_C)"

<code>. ttest haz, by(meat_poul_fish_C)</code>						
Two-sample t test with equal variances						
Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
No	224	-1.950223	.0990268	1.482097	-2.145371	-1.755075
Yes	374	-1.637834	.0827858	1.601001	-1.80062	-1.475049
combined	598	-1.754849	.0639408	1.56361	-1.880426	-1.629273
diff		-.312389	.1315951		-.5708356	-.0539424
diff = mean(No) - mean(Yes)				t = -2.3739		
Ho: diff = 0				degrees of freedom = 596		
Ha: diff < 0		Ha: diff != 0		Ha: diff > 0		
Pr(T < t) = 0.0090		Pr(T > t) = 0.0179		Pr(T > t) = 0.9910		

- i. Number of haz observations in each of the two groups. How many observations of children aged 2 to 5 years in the sampled areas do not eat fish, meat or poultry served on their plate in the last 24 hours? *224 children aged 2-5 years did not eat meat, fish or poultry in the last 24 hours.*
 - ii. Average haz score of children aged 2-5 years in each group. Which group has a higher haz-score of children aged 2 to 5 years? *Children aged 2 to 5 years who eat meat, poultry or fish have a higher average haz-score.*
 - iii. 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.
 - iv. What is the p-value that this difference in means is less than 0? And how can we interpret this? *P-value is 0.0090. We can say with more than 99% confidence that the mean height-for-age z-score of children aged 2 to 5 years who eat meat, poultry or fish in the last 24 hours is larger than the mean height-for-age z-score of children aged 2 to 5 years who do not eat meat, poultry or fish in the last 24 hours.*
- d. Now let's see if households with a youth household head have significantly different child height-for-age z-score than households with older household heads.

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: *"pworth haz ex_r, sig"*

Figure 3.14: Stata command and output for "pworth haz ex_r, sig"

. pworth haz ex_r, sig		
	haz	ex_r
haz	1.0000	
ex_r	0.1194	1.0000
	0.0000	

- iv. Question: Do you notice anything similar across the two outputs? *The p-values are the same!(0.000 is almost equal to 0.00524)*
- 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 child height-for-age z-score?
- i. Let's try to run a regression where the height-for-age z-score (haz) of children under 5 years in the sample areas is still the dependent variable, and the independent variables are real daily consumption expenditure of the household that the child under five years belong, whether or not the child under 5 years belong to a household that treats their water before drinking, whether or not the household that the child under 5 years belong to have access to improved toilet infrastructure, whether or not the household that the child under 5 years belong to receive extension on appropriate nutrition for pregnant women, the size of land owned by the household that the child under 5 years belong to, whether or not the head of the household that the child under 5 years belong to is matured and whether or not the head of the household that the child under 5 years belong to is female.
 - 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 ex_r" – numeric/continuous*
 2. *"codebook safewater" – categorical dummy variable (0/1)*
 3. *"codebook improvedtoilets_vs1" - categorical dummy variable (0/1)*
 4. *"codebook advice2" - categorical dummy variable (0/1)*
 5. *"codebook landholdings" – numeric/continuous*
 6. *"codebook hhh_mature" - categorical dummy variable (0/1)*
 7. *"codebook hhh_female" - categorical dummy variable (0/1)*
 - iii. Practice: *"regress haz ex_r safewater improvedtoilets_vs1 advice2 landholdings hhh_mature hhh_female"*

Figure 3.15: Stata command and output for "regress haz ex_r safewater improvedtoilets_vs1 advice2 landholdings hhh_mature hhh_female"

```
. regress haz ex_r safewater improvedtoilets_vs1 advice2 landholdings hhh_mature hhh_female
```

Source	SS	df	MS	Number of obs	=	1,305
Model	120.572361	7	17.2246229	F(7, 1297)	=	5.75
Residual	3882.54028	1,297	2.99347747	Prob > F	=	0.0000
				R-squared	=	0.0301
				Adj R-squared	=	0.0249
Total	4003.11264	1,304	3.06987166	Root MSE	=	1.7302

	haz	Coefficient	Std. err.	t	P> t	[95% conf. interval]
ex_r		.0046134	.0012492	3.69	0.000	.0021627 .0070641
safewater		.0789408	.1091627	0.72	0.470	-.135214 .2930956
improvedtoilets_vs1		.3604228	.100654	3.58	0.000	.1629602 .5578853
advice2		.1907452	.0968147	1.97	0.049	.0008146 .3806757
landholdings		-.065619	.0351863	-1.86	0.062	-.1346472 .0034093
hhh_mature		.0412775	.1060595	0.39	0.697	-.1667895 .2493445
hhh_female		.010044	.1911317	0.05	0.958	-.3649172 .3850051
_cons		-1.800801	.1246846	-14.44	0.000	-2.045406 -1.556195

iv. Are any of these variables significant? *Real daily consumption expenditure of the household and whether or not the household uses improved toilet infrastructure is significant – the p-value of these variables are 0.000. We also observe that whether households receive nutrition extension for pregnant women and plot size are significant, with p-values of less than 0.05 and 0.1, respectively.*

v. What does the coefficient of these variables tell us?

1- *Because ex-r is a continuous variable, the coefficient says that the marginal effect of one additional increase in daily real consumption expenditure increases the child's haz-score by 0.046, keeping all other factors constant.*

2- *Because improvetoilets_vs1 is a dummy variable, the coefficient says that if children under 5 years belong to a households that have access to improved toilet infrastructures, the haz-score will increase by 0.36. (The coefficient is 0.36).*

vi. Which variable is almost significant at the 95% level? *advice2– the p-value is 0.049*

vii. What does this coefficient on advice2 tell us? *Because advice2 is a dummy variable, the coefficient says that if children under 5 years belonged to a household that received extension on appropriate nutrition for pregnant women, their haz-score will increase by 0.190. (The coefficient is 0.1907452)*

c. Let's add province to our regression.

i. What type of variable is province? *"codebook province" – Numeric (byte)- (14 different categories/values for the 14 different provinces)*

- ii. Practice: “regress haz ex_r safewater improvedtoilets_vs1 advice2 landholdings hhh_mature hhh_female province”

Figure 3.16: Stata command and output for “regress haz ex_r safewater improvedtoilets_vs1 advice2 landholdings hhh_mature hhh_female province”

```
. regress haz ex_r safewater improvedtoilets_vs1 advice2 landholdings hhh_mature hhh_female provin
> ce
```

Source	SS	df	MS	Number of obs	=	1,305
Model	221.125425	8	27.6406782	F(8, 1296)	=	9.47
Residual	3781.98722	1,296	2.91820001	Prob > F	=	0.0000
				R-squared	=	0.0552
				Adj R-squared	=	0.0494
Total	4003.11264	1,304	3.06987166	Root MSE	=	1.7083

	haz	Coefficient	Std. err.	t	P> t	[95% conf. interval]
ex_r		.0041256	.0012362	3.34	0.001	.0017005 .0065508
safewater		-.0140756	.10894	-0.13	0.897	-.2277936 .1996425
improvedtoilets_vs1		.327707	.0995365	3.29	0.001	.1324366 .5229774
advice2		.1516445	.0958214	1.58	0.114	-.0363376 .3396266
landholdings		-.0137426	.0358475	-0.38	0.702	-.084068 .0565829
hhh_mature		.0103881	.1048496	0.10	0.921	-.1953055 .2160817
hhh_female		.116171	.1895773	0.61	0.540	-.255741 .4880829
province		.063776	.0108647	5.87	0.000	.0424617 .0850903
_cons		-2.290256	.1486872	-15.40	0.000	-2.58195 -1.998562

- iii. Question: Is province significant? How can we interpret the coefficient on “province”? province *is significant (p value =0.000). For every one unit increase in province, the child’s haz-score increases by 0.063... This doesn’t make sense; province isn’t a continuous or dummy variable....*
- iv. So instead, we can put “i.” in front of province (or any categorical non-dummy variables). Let’s try again:
- v. Practice: “regress haz ex_r safewater improvedtoilets_vs1 advice2 landholdings hhh_mature hhh_female i.province”

Figure 3.17: Stata command and output for “regress haz ex_r safewater improvedtoilets_vs1 advice2 landholdings hhh_mature hhh_female i.province”

```
. regress haz ex_r safewater improvedtoilets_vs1 advice2 landholdings hhh_mature hhh_female i.province
```

Source	SS	df	MS	Number of obs	=	1,305
Model	396.27377	20	19.8136885	F(20, 1284)	=	7.05
Residual	3606.83887	1,284	2.80906454	Prob > F	=	0.0000
				R-squared	=	0.0990
				Adj R-squared	=	0.0850
Total	4003.11264	1,304	3.06987166	Root MSE	=	1.676

haz	Coefficient	Std. err.	t	P> t	[95% conf. interval]
ex_r	.0020303	.0013157	1.54	0.123	-.0005508 .0046114
safewater	.0192869	.1093072	0.18	0.860	-.1951534 .2337272
improvedtoilets_vs1	.2902912	.1019296	2.85	0.004	.0903243 .4902581
advice2	.0962266	.1068725	0.90	0.368	-.1134372 .3058905
landholdings	.0405974	.0379195	1.07	0.285	-.0337936 .1149885
hhh_mature	-.0140334	.1038584	-0.14	0.893	-.2177842 .1897175
hhh_female	.0317556	.1871575	0.17	0.865	-.3354125 .3989236
province					
Central Province	.5943186	.2475298	2.40	0.016	.1087114 1.079926
Chimu (Simbu) Province	.4455552	.2231952	2.00	0.046	.007688 .8834225
East New Britain Province	.9316815	.24666	3.78	0.000	.4477806 1.415582
East Sepik Province	.3376934	.2644979	1.28	0.202	-.1812021 .8565889
Eastern Highlands Province	.7891199	.2534324	3.11	0.002	.2919328 1.286307
Jiwaka Province	.5610519	.2476679	2.27	0.024	.0751739 1.04693
Madang Province	.7052731	.2313612	3.05	0.002	.2513857 1.15916
Morobe Province	-.4243679	.2527177	-1.68	0.093	-.9201529 .0714171
Western Province	1.267225	.1785732	7.10	0.000	.9168979 1.617553
Western Highlands	.5562394	.2187539	2.54	0.011	.1270851 .9853938
Gulf Province	1.293717	.2343655	5.52	0.000	.833936 1.753499
Oro Province	1.167069	.2215291	5.27	0.000	.7324708 1.601668
Milne Bay Province	1.148766	.2669933	4.30	0.000	.6249753 1.672557

- vi. This now turns province into 14 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 provinces is being compared to ARoB.
- vii. Are any of the province significant? *East New Britain, Milne Bay, Oro, Gulf, Western province have p-values=0.000.*
- viii. This means that holding the effect of these other variables constant, children under 5 years in Western Province have an average HAZ-score that is 1.30 *higher* than children under 5 years in ARoB.
- ix. Similarly, holding these other variables constant, children under 5 years in Gulf province have an average HAZ-score that is 1.29 higher than children under 5 years in ARoB, children under 5 years in Milne Bay province have average HAZ-score that is 1.14 higher than children under 5 years in ARoB and children under 5 years in Oro province have a HAZ-score that is 1.16 higher than children under 5 years in ARoB.

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.

IFPRI Contact Information

Emily Schmidt

Senior Research Fellow

e.schmidt@cgiar.org

Rishabh Mukerjee

Research Analyst

r.mukerjee@cgiar.org

Harry Gimiseve

Research Consultant

gimiseveharry29@gmail.com