

Creating Descriptive Tables in Stata

A short training manual for IFPRI and NARI analysts

October 2026

About this manual

This is a shortened, hands-on manual that trains IFPRI and NARI analysts to build descriptive tables and cross-tabulations in Stata. It keeps every descriptive and data-preparation step from *Describing, Transforming, and Analyzing Data Using Stata: An Introductory Course*.

Who is it for. Analysts preparing summary tables from the producer survey. No prior Stata experience is assumed.

How to use it. This is a workshop manual, not a lecture. Type each command in the Command window as you read. Questions are marked **Practice**, **Question**, or **Challenge**, and the expected answer follows in *italics*.

Stata version. Written for Stata 18. Older versions may lack some commands; type `help [command]` or search online for the older equivalent.

Practice datasets. The examples use the 2023 PNG Rural Household Survey files from the original course, so the printed answers can be checked:

File	Level (one row per...)	Used in
hh_3_4.dta (Module 3.4, housing quality)	Household (2,699)	Lessons 1–3, 5
s11_long.dta (household roster)	Person (16,192)	Lessons 4–5
lesson3.dta (compiled household file)	Household (1,334)	Lesson 4, outliers

The same commands apply directly to the producer survey; only the file and variable names change.

Lesson 1 – Getting to know Stata and your data

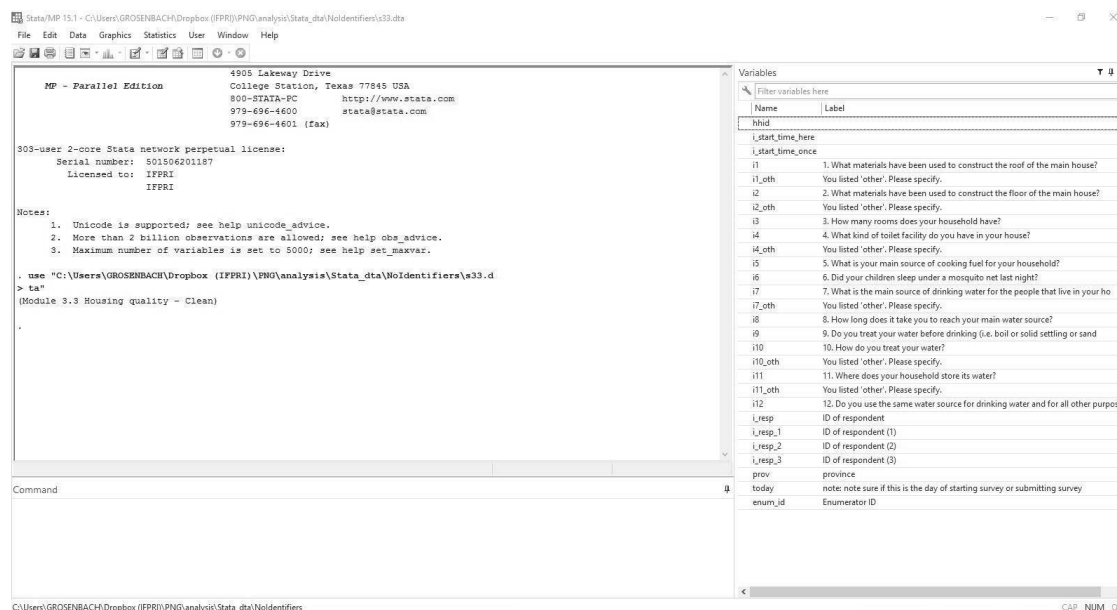
Before building any table, know what is in the file: how many rows, what each variable means, and how it is stored. Open `hh_3_4.dta` by double-clicking it.

1.1 The Stata screen

The main screen has three windows:

Window	Name	What it does
1	Review and Results	Shows the commands you ran and their output
2	Command	Where you type a command and press Enter
3	Variables	Lists every variable name and its label

Figure 1.1: Main Stata screen



1.2 Match variables to the questionnaire

hh_3_4 corresponds to Module 3.4 (Housing quality) of the questionnaire. Question numbers and wording in the questionnaire match the variable names and labels in the Variables window. Keep the questionnaire open while you work: labels are cut off after 80 characters, and the questionnaire shows the full question and its skip rules.

Figure 1.2: Module 3.4 (Housing quality) from the 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] [SKIP IF HH DOESN'T HAVE CHILDREN]			
12c. What material do members of your household use with water to wash their hands? [Select all that apply] [code: handwash_v1]			

1.3 Browse the data

The Data Browser shows the data as a spreadsheet. Open it any of three ways:

1. Type `browse` in the Command window and press Enter.
2. Menu: **Data** → **Data Editor** → **Data Editor (Browse)**.

3. Click the Data Browser icon in the toolbar:

To view only some variables, list them: `browse i9.`

Figure 1.3: Stata data browser screen



1.4 Know how each variable is stored

The storage type decides which table you can make from a variable. The Data Browser shows each type in a different color.

Type	Color in browser	Example	What it holds	Typical table
Numeric	Black	i4 – number of rooms	Numbers with no labels	summarize (mean, median)
Categorical	Blue	i2 – floor material	Numbers with a text label attached (6 = “Wood”)	tabulate (frequencies, %)
String	Red	i2_oth – “other” floor material, typed in	Free text	tabulate to read the responses, then recode

Click a blue cell such as “Wood” and the number behind it (6) appears in the bar at the top of the browser.

1.5 Count the observations

count reports the number of observations (rows).

Practice: How many observations are in this dataset? count → *2,699 households*.

Figure 1.4: Stata command and output for “count”

```
count
2,699
```

1.6 Describe a variable with codebook

codebook [varname] gives a one-variable overview: label, storage type, range, unique values, missing values, and example values.

Practice: What do we know about drinking water? Type codebook i8a.

Figure 1.5: Stata command and output for “codebook i8a”

```
. codebook i8a
```

i8a	18a: 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
```

- What question was asked (the variable label)? *8a. What is the main source of drinking water for the people that live in your household?* The label is cut off; check the questionnaire for the full wording.
- What type of data is it? *Numeric (int), with value labels, so it is categorical.*
- What is the range? *1 to 12.*
- How many observations are missing? *0.*
- How many unique values? *12.*
- What does one value mean? *6 = Unprotected spring.* Only four of the 12 values are shown as examples; tab shows them all.

Lesson 2 – One-variable descriptive tables

Use `tabulate` for categorical variables (counts and percentages) and `summarize` for numeric variables (mean, spread, median). Dataset: `hh_3_4.dta`.

2.1 Frequency tables with `tabulate`

Code: `tabulate [varname]`, shorthand `tab [varname]`.

Many commands have a shorthand; this manual gives it the first time a command appears.

Practice: How many households get drinking water from a protected dug well? `tab i8a`

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	

Column	Meaning	Example answer
Freq.	Number of households giving each answer	<i>184 households use a protected dug well</i>
Percent	Share of all households giving each answer	<i>6.82% use a protected dug well</i>
Cum.	Running total of the percentages, top to bottom	<i>97.70% use surface water or a source listed above it</i>

Question: How many households use surface water? 777.

Question: How many households treat their water? tab i9 → 445 households (16.49%).

2.2 Read the “other, specify” responses

String variables hold the typed-in answers when a respondent chose “other”. Tabulate them to see what was written, which shows whether they should be recoded into existing categories.

Question: What kind of variable is i2_oth? *String*. Check with codebook i2_oth, or see that it shows in red in browse.

Practice: tab i2_oth. The list includes “BAMBU”, “Bamboo”, “Bambu”, “Bamboo floor” and “Bamboo flooring”: the same answer spelled several ways. These must be cleaned before they can go into a table.

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

```
. tab i2_oth
```

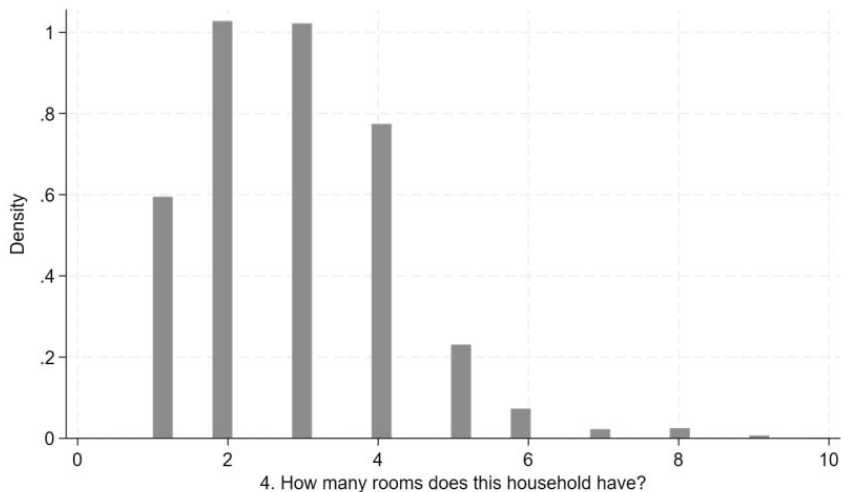
12_oth: You selected 'other', please specify. -----			
-----	Freq.	Percent	Cum.
BAMBU	1	1.22	1.22
Bamboo	13	15.85	17.07
Bamboo Flooring	1	1.22	18.29
Bamboo Stem	1	1.22	19.51
Bamboo floor	1	1.22	20.73
Bamboo flooring	10	12.20	32.93
Bamboo-weaved blind	3	3.66	36.59
Bambu	1	1.22	37.80
Betelnut palm	1	1.22	39.02
Black Palm Skin and Plywood	1	1.22	40.24
Black palm	2	2.44	42.68
Blinc	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

2.3 Visualize a distribution with histogram

Code: histogram [varname], shorthand hist. Bar height shows how common each value is.

Practice: hist i4 (number of rooms).

Figure 1.8: Stata histogram output for “hist i4”



Question: What is the most common number of rooms (the mode)? *2 rooms*. Most houses have 1–5 rooms; very few have more than 5.

`tab i4` also works but gives one row per value. For a count variable like rooms, the mean and median are usually more useful.

2.4 Summary statistics with `summarize`

Code: `summarize [varname]`, shorthand `sum`. Reports observations, mean, standard deviation, minimum and maximum.

Practice: Average number of rooms? `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

- Average number of rooms? *2.87*.
- **Challenge:** Which earlier command also gave the minimum and maximum? *codebook*.
- What is the median? *This output does not show it*. The maximum (10) is far above the mean, so a few large houses pull the mean up. The median is a better central value for skewed data.

2.5 Add options: `summarize`, `detail`

Most commands accept **options**, typed after a comma. The `detail` option (shorthand `det`) adds percentiles, the median, variance, skewness and kurtosis.

Practice: `sum i4, detail`

Figure 1.10: Stata command and output for “sum i4, detail”

```
. 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

Question: What is the median number of rooms? *3 rooms* (the 50% row).

Lesson 3 – Cross-tabulations

A cross-tabulation shows two categorical variables at once. The key skill is choosing the right percentage: cell, row or column each answer a different question. Dataset: hh_3_4.dta.

3.1 Two-way tables

Code: tabulate [varname1] [varname2]. The first variable forms the rows, the second the columns.

Practice: Do households treat their water, by source? 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?		
	Yes	No	Total
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

Each cell counts the households with that combination of answers.

Question: How many households collect rainwater and do not treat it? 549.

3.2 Choose the right percentage

Add an option after a comma to show percentages under each count. Read each one as “of [the group that sums to 100%], what share...”.

Option	Sums to 100% across	Answers the question	tab i8a i9 example (rainwater, not treated)
, cell	The whole table	What share of all households collect rainwater and do not treat it?	20.34% (549 of 2,699)
, row	Each row	Of households that collect rainwater , what share do not treat it?	79.45% (549 of 691)
, column (col)	Each column	Of households that do not treat their water, what share collect rainwater?	24.36% (549 of 2,254)

Practice: Run all three and find the rainwater/No cell in each.

tab i8a i9, cell

tab i8a i9, row

tab i8a i9, col

In each output, the top number in a cell is the count and the bottom number is the percentage.

Figure 1.12: Stata command and output for “tab i8a i9, cell” (first rows)

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?		
	Yes	No	Total
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

Figure 1.13: Stata command and output for “tab i8a i9, row” (first rows)

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

Figure 1.14: Stata command and output for “tab i8a i9, col” (first rows)

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

Rule of thumb for descriptive tables: put the grouping variable (region, sex, farm type) in the rows and use row percentages, so each row reads as a profile of that group. The order of the variables matters; tab i9 i8a, col gives the same numbers as tab i8a i9, row.

3.3 Restrict a table with if

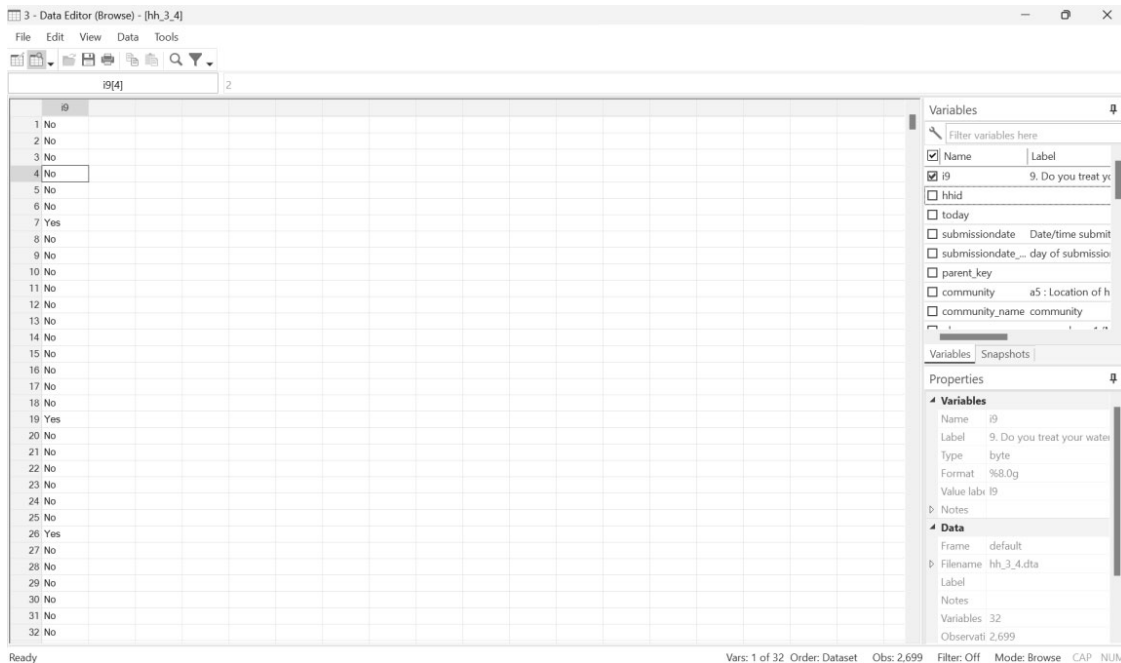
if limits a command to observations that meet a condition. It uses these logical operators:

Operator	Meaning
==	equal to (two equals signs when comparing)
~= or !=	not equal to
> / >=	greater than / greater than or equal to
< / <=	less than / less than or equal to
&	and
\	or
~	not

Categorical variables are stored as numbers, so you need the number behind a label before writing the condition. Three ways to find it for “No” in i9:

1. codebook i9 → shows that 2 = No.
2. tab i9 then tab i9, nolabel → the nolabel option (shorthand nolab) prints the numbers instead of the labels.
3. browse i9 → click a cell that says “No” and read the number in the top bar.

Figure 1.15: Stata data browser window for “browse i9”



Question: Using if, what share of households that do not treat their water use rainwater?
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	

24.36%, the same as the column percentage above. The total is 2,254, not 2,699, because only households that do not treat their water are included.

3.4 Practice: roof type by number of rooms

Roof type is i1; number of rooms is i4.

Question	Command
Roof type for households with more than 3 rooms	tab i1 if i4>3 or tab i1 if i4>=4
...with exactly 3 rooms	tab i1 if i4==3
...with 3 to 5 rooms	tab i1 if i4>=3 & i4<=5 or tab i1 if i4>2 & i4<6
...with fewer than 3 or more than 5 rooms	tab i1 if i4<3 \ i4>5 or tab i1 if i4<=2 \ i4>=6

Challenge: Which households more often have a corrugated metal roof: those with more than 3 rooms, or those with 3 or fewer?

```
tab i1 if i4>3
tab i1 if i4<=3
```

58.27% of households with more than 3 rooms have a corrugated metal roof, against 39.65% of households with 3 rooms or fewer.

Caution: Stata treats a missing numeric value (.) as larger than any number. if x>3 therefore includes missing values. Add & x<. (or & !missing(x)) when the variable has missing values; Lesson 4 shows why this matters.

Lesson 4 – Preparing variables for tables

Raw survey variables rarely go straight into a table. Missing-value codes distort means, unlabeled variables are unreadable, and many table rows need new indicator variables. Dataset: `s11_long.dta` (household roster), then `lesson3.dta` for outliers.

4.1 Household-level and person-level data

- **Wide (household-level) data** has one row per household, identified by `hhid`. `hh_3_4` is wide.
- **Long data** has several rows per household: one per person, per crop or per plot. You need two identifiers, such as `hhid` and a person ID (`pid`). The roster `s11_long` is long.

Question: Will the roster have more or fewer observations than `hh_3_4`? *More. count → 16,192 people.*

Figure 2.1: Stata command and output for “count” in the roster

```
. count  
16,192
```

Practice: How many men and women? Sex is `a3`, labeled 3. What is the sex of `${a1}`? The `${a1}` part is the person’s name, filled in by the tablet and removed for anonymity. `tab a3 → 8,459 male (52.24%) and 7,733 female (47.76%)`. `tab` is better than `codebook` here because it also gives percentages.

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	

4.2 Find and fix missing-value codes

Practice: Average age? Age is `a4a`. `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

A mean of 125 years is impossible. The maximum, 888, is the code enumerators entered when the respondent did not know the age. codebook a4a reports *0 missing*, so Stata treats 888 as a real age.

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]					Units: 1				
Unique values: 97					Missing .: 0/16,192				
Mean: 125.095									
Std. dev.: 279.022									
Percentiles:		10%	25%	50%	75%	90%			
		5	10	22	45	888			

Stata marks missing values differently by storage type:

Type	Missing shown as	Set to missing with
Numeric	. (a period)	replace [varname]=. if ...
String	"" (blank)	replace [varname]=" " if ...

a4a is numeric, so 888 becomes .. Use either command:

```
recode a4a 888=.
```

```
* or
```

```
replace a4a=. if a4a==888
```

Question: How many missing values now? codebook a4a → 1,904. Or use tab a4a, missing: the missing option adds a row for missing values in any tabulation.

Figure 2.5: “codebook a4a” after recoding 888 to missing

```
. codebook a4a
```

a4a					4a. What is the age of				
Type: Numeric (int)									
Range: [0,106]					Units: 1				
Unique values: 96					Missing .: 1,904/16,192				
Mean: 23.4313									
Std. dev.: 18.0899									
Percentiles:		10%	25%	50%	75%	90%			
		4	9	18	34	51			

Question: Average age now? sum a4a → 23.43 years, from 14,288 observations. The 1,904 people with unknown age are excluded.

Figure 2.6: “sum a4a” after recoding 888 to missing

```
. sum a4a
```

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

Check the questionnaire or enumerator manual for every special code (888, 999, -99 and so on) before tabulating.

4.3 Rename and label variables

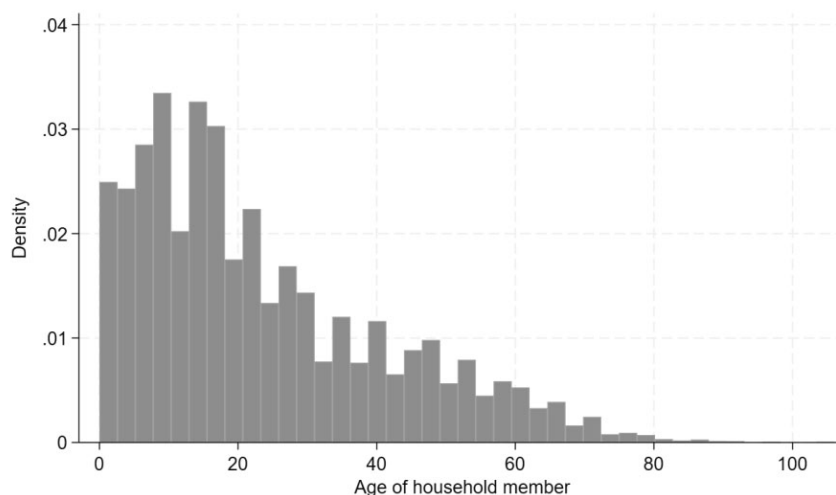
Tables print variable names and labels, so make them readable.

Task	Code	Practice
Rename a variable	<code>rename [old] [new] (ren)</code>	<code>rename a4a age</code>
Label a variable	<code>label variable [varname]</code> <code>"label" (lab var)</code>	<code>label variable age "Age of household member"</code>

The label must be in quotation marks. After renaming, use the new name (age) in every command.

Practice: `hist age`. *Most people in the sample are 25 or younger, and very few are over 80: a young population.*

Figure 2.7: Stata histogram output for “hist age”



4.4 Create dummy (indicator) variables

A dummy variable equals 1 if a condition is true and 0 if not. Its mean is the share of observations where the condition holds, which makes dummies the building blocks of most “% of households...” rows.

Practice: Create `child = 1` if the person is 15 or younger.

Step	Code (shorthand gen)
Create the variable, all missing	generate child=.
Set to 1 for children	replace child=1 if age<=15
Set to 0 for adults	replace child=0 if age>15

Question: codebook child. How many missing values? 0, which is a problem. The 1,904 people with unknown age were all set to 0 (adult), because . counts as larger than 15.

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

```

Fix: replace child=. if age==. Note the single = (assign a value) versus the double == (test a value). Better still, write the adult step as replace child=0 if age>15 & age<. so the error never happens.

Figure 2.9: “codebook child” after assigning missing values

```
. codebook child
```

```
child (unlabeled)
```

```

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

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

```

4.5 Add value labels

codebook child now shows 0, 1 and 1,904 missing, but no label says what 0 and 1 mean.

1. Label the variable: label variable child "Is the household member 15 years or younger?"
2. Define a set of value labels once: label define no_yes 0 "No" 1 "Yes". This is saved in memory and can be reused for any 0/1 variable.
3. Attach it to the variable: label values child no_yes (shorthand lab val).

Question: How many children are in the dataset? codebook child → 6,060 (8,228 adults, 1,904 unknown).

Figure 2.10: “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       .

```

Challenge: What share of children list “student” as their primary activity? `tab a8 if child==1` → 86.39% (1,041 of 1,205 children with a reported activity).

Figure 2.11: Stata command and output for “tab a8 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	

4.6 Bar graphs of prevalence by group

Code: `graph bar [varname], over([groupvar])`. For a 0/1 variable, bar height is the share with value 1.

Challenge: Which province has the highest share of children?

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

Figure 2.12(a): “graph bar child, over(prov)”

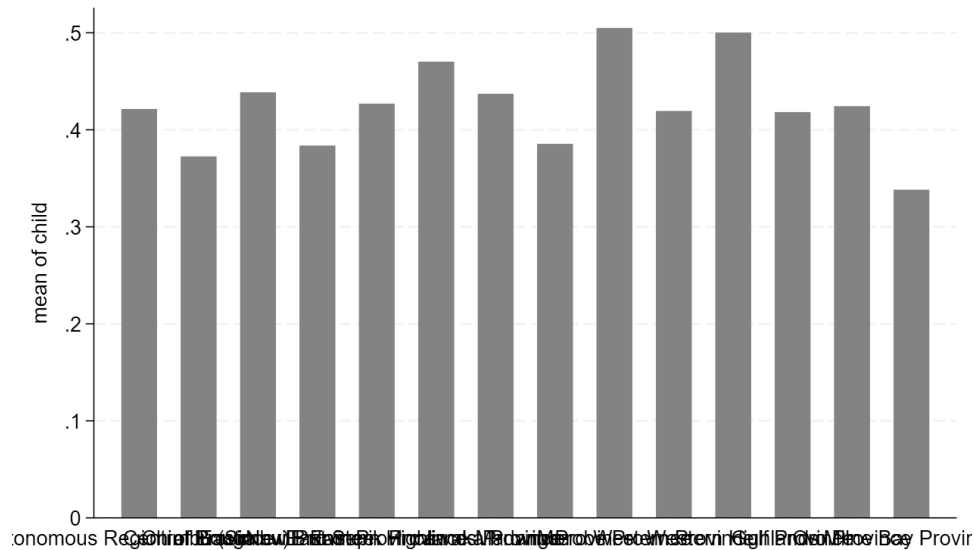
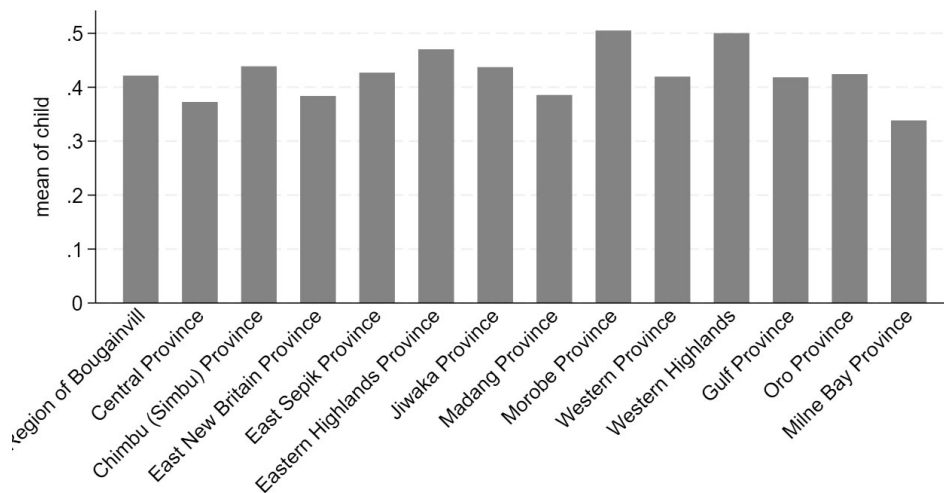


Figure 2.12(b): “graph bar child, over(prov, label(angle(45)))”



The second line tilts the province names so they do not overlap. *Morobe has the highest share of children aged 15 and under.* sum child gives the same share as a number.

Practice: Share of women by province: graph bar a3, over(prov). *Every bar is above 1, so this is not a share.* codebook a3 shows why: 1 = Male, 2 = Female. A 1/2 variable is not a dummy.

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

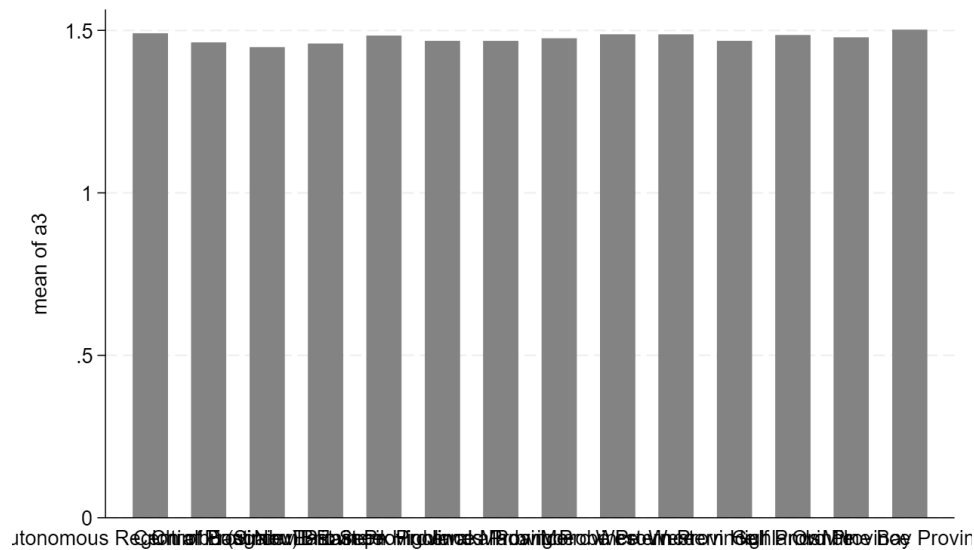


Figure 2.14: Stata command and output for “codebook a3”

```
. codebook a3
```

```

a3                                     3. What is the sex of ?

Type: Numeric (byte)
Label: b3

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

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

```

Practice: Build a proper dummy:

```

gen woman=.
replace woman=1 if a3==2
replace woman=0 if a3==1
lab var woman "Is the household member a woman?"
lab val woman no_yes
codebook woman
graph bar woman, over(prov)

```

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

woman

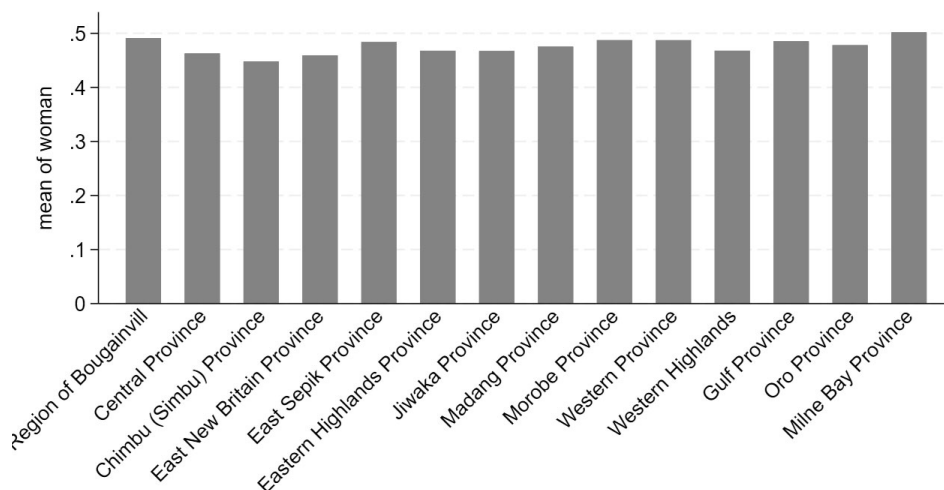
Is the household member a woman?

```
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
```

Figure 2.16: “graph bar woman, over(prov)”



Milne Bay has the highest share of women in the sample.

4.7 Check continuous variables for outliers

Open lesson3.dta (1,334 households). Self-reported continuous variables such as land area often contain outliers from data-entry errors, misunderstood questions, or rough guesses. Outliers inflate means in a table.

Practice: sum landholdings, det

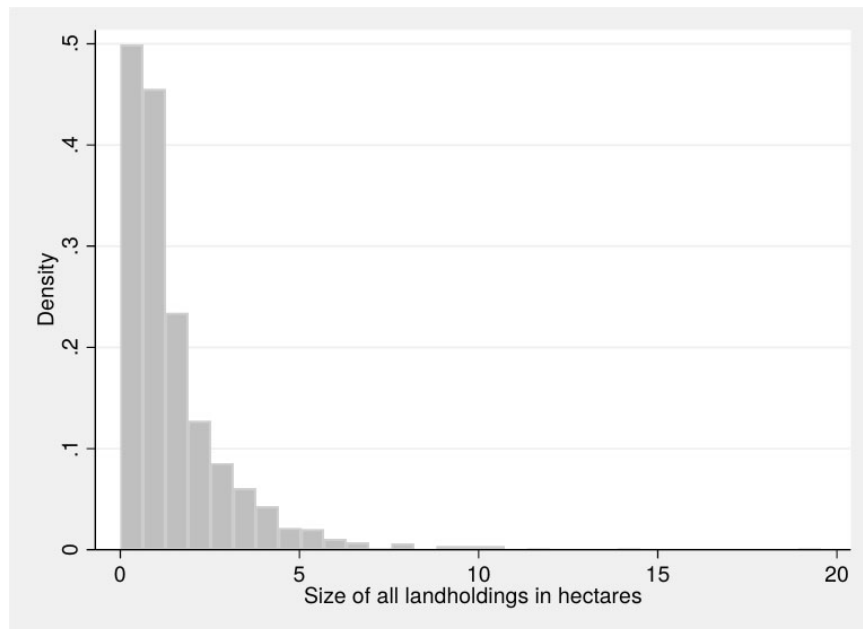
Figure 3.2: Stata command and output for “sum landholdings, det”

```
. sum landholdings, det
```

Size of all landholdings in hectares				
Percentiles		Smallest		
1%	0	0		
5%	.0162	0		
10%	.0324	0	Obs	1,334
25%	.3743	0	Sum of Wgt.	1,334
			Mean	1.455915
50%	1.05		Std. Dev.	1.688366
		Largest		
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. Compare mean and median: 1.46 vs 1.05 hectares; the mean is about 40% higher, a sign of a right skew.
2. Check minimum and maximum: 0 is plausible, but 19.55 hectares is far from the median.
3. Visualize: `hist landholdings` shows a long right tail.

Figure 3.3: Stata histogram output for “hist landholdings”



How to treat outliers is a judgment call that depends on the question. In this exercise, values above the 99th percentile (8.0824) are replaced with the median (1.05):

```
count if landholdings>8.0824          // 12 observations
replace landholdings=1.05 if landholdings>8.0824  // (12 real changes made)
hist landholdings
```

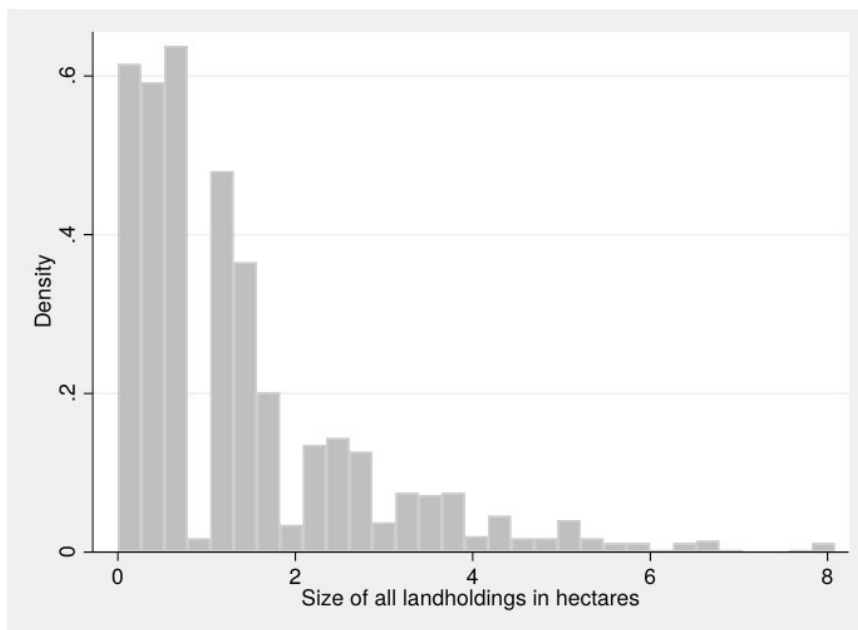
Figure 3.4: Stata command and output for “count if landholdings>8.0824”

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

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)
```

Figure 3.6: “hist landholdings” after cleaning outliers



The new histogram ends at 8 hectares instead of 20. Whatever rule your team uses, apply it the same way across tables and record it in the table notes.

Lesson 5 – Changing the unit of analysis: collapse and merge

A table describes one unit: households, people, plots or crops. collapse turns person- or plot-level data into household-level summaries; merge brings variables from different modules into one file so they can be cross-tabulated. Start from `s11_long.dta` with the age, child and woman variables built in Lesson 4.

5.1 Collapse long data to the household level

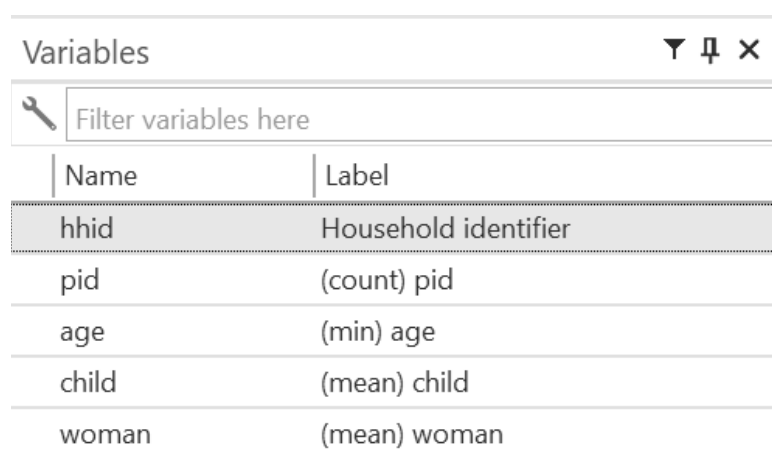
collapse replaces the data in memory with one row per group, holding the summary statistics you request: mean, median, sum, count, max, min. Save your person-level file first, because collapse overwrites it.

Household-level indicator	Statistic and variable
Number of people in the household	(count) pid
Age of the youngest member	(min) age
Share of members who are children; share who are women	(mean) child woman (the mean of a 0/1 dummy is a share)

Practice:

```
collapse (count) pid (min) age (mean) child woman, by(hhid)
count      // 2,699 households
```

Figure 2.21: Variables 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

The variables keep their old names, and their labels become “(count) pid”, “(min) age” and so on. Rename and relabel them:

```
rename pid hhsiz
lab var hhsiz "Household size"
rename age age_youngest
lab var age_youngest "Age of the youngest household member"
rename child perc_children
lab var perc_children "Share of household members aged 15 or younger"
rename woman perc_women
lab var perc_women "Share of household members who are women"
```

Question	Command	Answer
Average household size	sum hhsiz	6.0 people (range 1–22)
Average age of the youngest member	sum age_youngest	8.7 years
Average share of children per household	sum perc_children	40.8%
Average share of women per household	sum perc_women	47.2%

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

```
. sum hhsize
```

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

(mean) child skips members with missing age, so perc_children is the share among members whose age is known.

5.2 Merge two household-level files (1:1)

Is household size related to roof type? Household size is in the collapsed roster; roof type (i1) is in hh_3_4. Both have one row per household, so each household matches exactly one row in the other file: a one-to-one merge.

Figure 2.23: Example of a 1:1 merge, shown for one household (hhid)

File	hhid	age_youngest	i1
s11_long (collapsed; n=2,699)	0100RSPN	11	
hh_3_4 (n=2,699)	0100RSPN		metal roof
Merged data (n=2,699)	0100RSPN	11	metal roof

Code: merge 1:1 [key variable] using "[file path]". See help merge for the full syntax.

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
<code>merge 1:1 <u>varlist</u> using <u>filename</u> [, options]</code>

Practice: Merge on the household identifier hhid:

```
merge 1:1 hhid using "C:\...\hh_3_4.dta"
```

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

<pre>. merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh Mukerjee\Stata_1 > 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)</pre>	
Result	Number of obs
Not matched	0
Matched	2,699 (_merge==3)

All 2,699 households matched. Merge creates a variable `_merge`:

<code>_merge</code>	Meaning
1	In the master file (the one already open) only
2	In the using file (the one named in the command) only
3	Matched in both

Always check tab `_merge` before building tables, and investigate any 1s or 2s. Run `drop _merge` before the next merge, or Stata will refuse because `_merge` already exists.

5.3 Summary statistics by group with `bysort`

Challenge: Do households with thatched roofs or with corrugated metal roofs have a higher share of children?

Option 1, one sum ... if per group:

```
sum perc_children if i1==1    // thatched: 43%
sum perc_children if i1==2    // corrugated metal: 38%
```

Option 2, one command for all groups. `bysort [groupvar]: [command]` repeats the command for each value of the group variable:

```
bysort i1: sum perc_children
```

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

Households with thatched roofs have a higher share of children (43%) than those with corrugated metal roofs (38%). `bysort` gives the same numbers with less typing, and works with `tab`, `sum` and most other commands.

5.4 Merge household data onto person-level data (1:m)

How many **people** live in a house with a thatched roof? This needs roof type on every person’s row. Start again: open `hh_3_4.dta` (household level) and merge the roster `s11_long.dta` (person level) onto it. One household matches many people, so the merge is one-to-many.

Figure 2.27: Example of a 1:m merge, shown for one household (hhid)

hh_3_4 (n=2,699) has one row for household 0100RSPN with i1 = metal roof. s11_long (n=16,192) has seven rows for it, one per person. The merged data (n=16,192) copies the roof type onto every person:

hhid	a_pid	Sex	i1 (added by merge)
0100RSPN	1	M	metal roof
0100RSPN	2	F	metal roof
0100RSPN	3	F	metal roof
0100RSPN	4	F	metal roof
0100RSPN	5	M	metal roof
0100RSPN	6	M	metal roof
0100RSPN	7	M	metal roof

```
merge 1:m hhid using "C:\...\s11_long.dta"  
tab _merge      // all 16,192 observations: _merge==3
```

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)

If the person-level file is the one already open, the same merge is written `merge m:1 hhid using "...\hh_3_4.dta"`.

Question: How many people live in a house with a thatched roof? `tab i1` → *8,328 people (51.43%)*. The data are now person-level, so `tab` counts people, not households.

Challenge: How many women live in a house with a thatched roof? First confirm the coding (codebook a3: 2 = female), then use either approach:

```
tab i1 if a3==2  
tab i1 a3, col  
tab a3 i1, row
```

3,961 women (51.22% of women in the sample).

The unit of the data decides the unit of the table. Before running any table, ask: “one row per what?” and label the table accordingly (“% of households” vs “% of people”).

Getting help and quick reference

Type `help [command]` to open the manual page for any command. Each page shows three things:

- **Syntax:** how to write the command. For example, `help tabulate twoway` shows `tabulate varname1 varname2 [if] [in] [weight] [, options]`. Underlined letters are the shortest accepted abbreviation.
- **Options:** what each option does, such as `cell`, `row`, `column`, `missing`, `nolabel`.
- **Examples:** ready-to-adapt commands, such as `tabulate region agecat, row`.

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]
```

Figure 2.18: Stata help window for tabulate twoway – options

options	Description
Main	
<u>chi2</u>	report Pearson's chi-squared
<u>exact</u> [(#)]	report Fisher's exact test
<u>gamma</u>	report Goodman and Kruskal's gamma
<u>lrchi2</u>	report likelihood-ratio chi-squared
<u>taub</u>	report Kendall's tau-b
<u>V</u>	report Cramér's V
<u>cchi2</u>	report Pearson's chi-squared in each cell
<u>column</u>	report relative frequency within its column of each cell
<u>row</u>	report relative frequency within its row of each cell
<u>clrch</u> i2	report likelihood-ratio chi-squared in each cell
<u>cell</u>	report the relative frequency of each cell

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
```

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]`

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

Try `help tabulate twoway`, `help summarize`, `help collapse` and `help merge`.

Command reference

Task	Command	Shorthand	Useful options
Look at the data	<code>browse [varlist]</code>	<code>br</code>	
Count observations	<code>count [if]</code>		
Describe one variable	<code>codebook [varname]</code>		
Frequency table	<code>tabulate [var]</code>	<code>tab</code>	<code>missing, nolabel</code>
Cross-tabulation	<code>tabulate [var1] [var2]</code>	<code>tab</code>	<code>cell, row, column, missing</code>
Summary statistics	<code>summarize [varlist]</code>	<code>sum</code>	<code>detail</code>
Statistics by group	<code>bysort [groupvar]: [command]</code>		
Histogram	<code>histogram [var]</code>	<code>hist</code>	
Bar graph of a share by group	<code>graph bar [dummy], over([groupvar])</code>		<code>over(..., label(angle(45)))</code>
Restrict to a subgroup	<code>... if [condition]</code>		
Recode a value	<code>recode [var] [old]=[new]</code>		
Create a variable	<code>generate [new]=[value]</code>	<code>gen</code>	
Change values	<code>replace [var]=[value] if ...</code>		
Rename	<code>rename [old] [new]</code>	<code>ren</code>	
Variable label	<code>label variable [var] "text"</code>	<code>lab var</code>	
Define value labels	<code>label define [name] 0 "No" 1 "Yes"</code>	<code>lab def</code>	

Task	Command	Shorthand	Useful options
Attach value labels	label values [var] [name]	lab val	
Summarize to one row per group	collapse (stat) [vars], by([id])		mean, median, sum, count, min, max
Combine files	merge 1:1 / 1:m / m:1 [id] using "[file]"		then tab _merge, drop _merge
Delete a variable	drop [var]		

Good practice for descriptive tables

1. Know your data first: run count, codebook and tab before building anything.
2. Recode special codes (888, 999) to missing before computing any mean.
3. Use 0/1 dummies with value labels for every "% of..." row.
4. Write & var< . whenever an if condition uses > or >=.
5. Be explicit about the unit: households, people, plots or crops.
6. Choose row, column or cell percentages to match the question the table answers.
7. Check _merge after every merge.