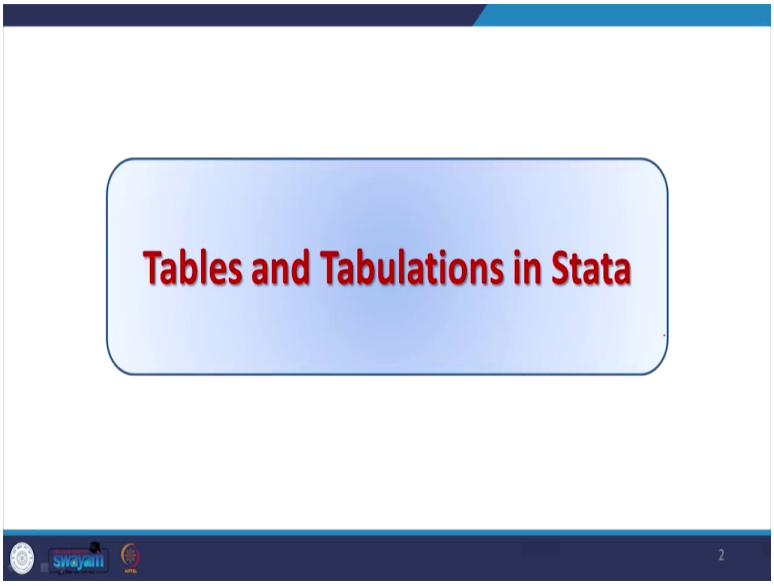
Handling Large-Scale Unit Level Data Using STATA
Professor Pratap C. Mohanty
Department of Humanities and Social Sciences,
Indian Institute of Technology, Roorkee
Lecture 17
Tabulation and Creation of New Variables in Stata – I

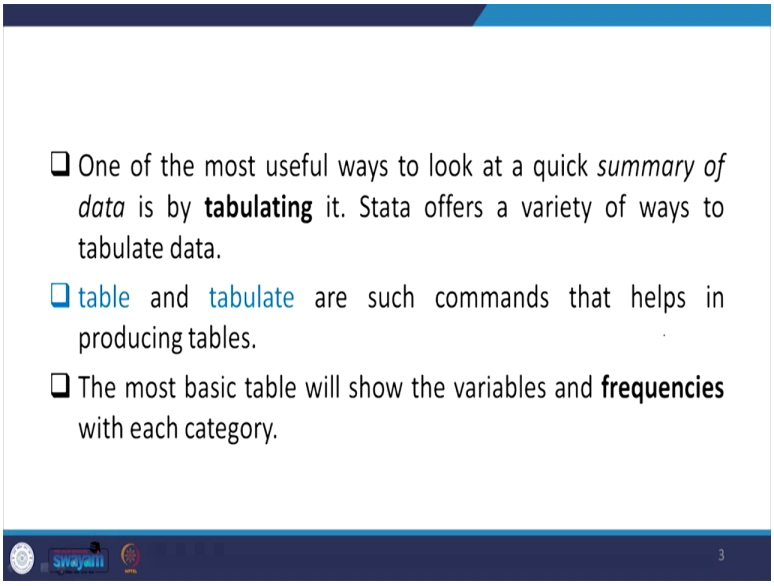
Welcome dear learner once again to the NPTEL MOOC module on Handling Large-Scale Data Using Stata. In the previous lecture we discussed about questionnaire design, the understanding of a good questionnaire for Stata. As a researcher, here in the week trying to give you the prerequisites of understanding unit level data with Stata.

(Refer Slide Time: 01:11)



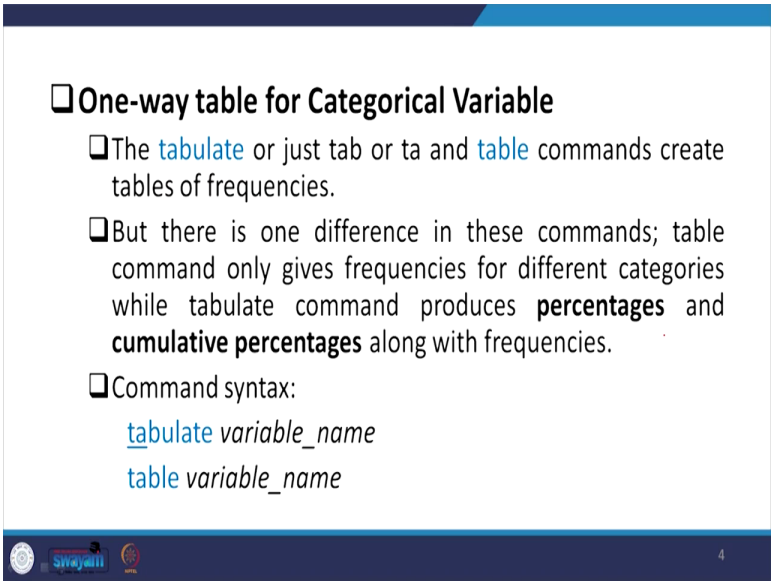
This particular lecture is targeted to understand the tabulation and creation of new variables in Stata. So we are focusing on tables and tabulation in Stata. What do you mean by tables and what do you mean by tabulation?

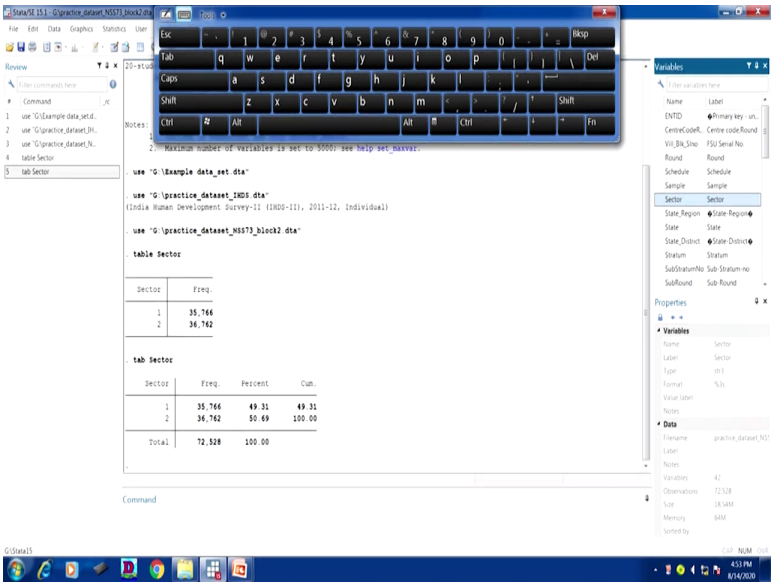
(Refer Slide Time: 01:17)



So, one of the most useful ways to look at a quick summary of data is tabulating it. Stata offers a variety of ways to tabulate data. There are different ways, we have already clarified in last to last lecture on different types of coding, replacing then also some forms of other operations of Stata. we are like summarizing data we discussed, codebook we discussed, I am focusing on difference between table and tabulate. Table and tabulate are such commands that helps in producing tables. The most basic table will show the variables and frequencies with each category.

(Refer Slide Time: 02:22)





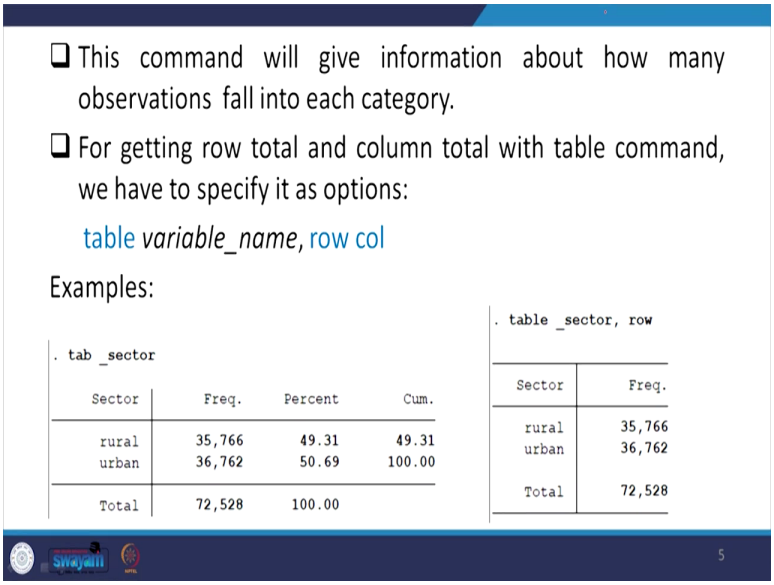
So, for the researcher, they do require one-way table, two-way table, three-way table, sometimes larger order tables as well depending upon the kind of analysis they wanted to target. The tabulate or just tab or ta in short or table commands create tables of frequencies, but there is one difference in these commands. Table command only gives frequencies for different categories, while tabulate command produces percentages as well as cumulative percentages along with their respective frequencies.

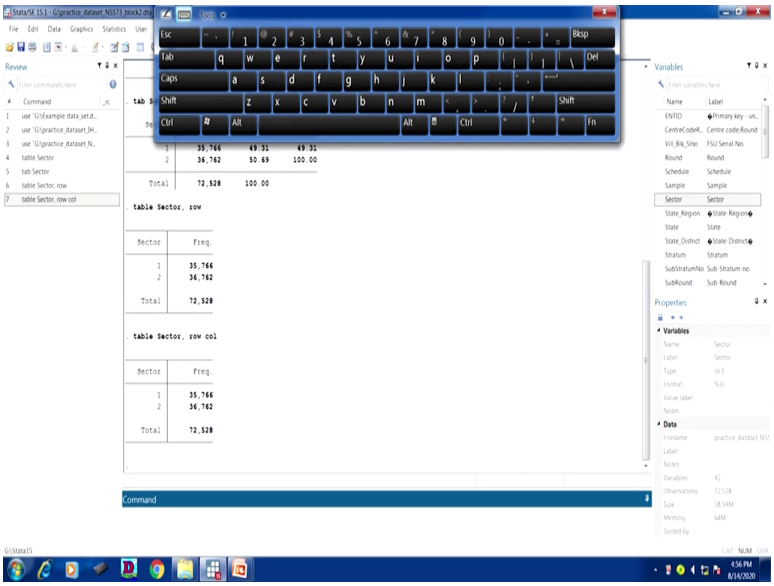
So the command syntax here is tabulate or table. So we are going to operate these two here as well. So let me start with table here. I am just going to type table, then I simply click on the variable, let it be sector, this simply enter. It gives you only frequencies. let us type ta or tabulate or tab, instead of table, then with the same variable. this gives different result, frequencies along with their percentages, respective percentages and cumulative percentages as well.

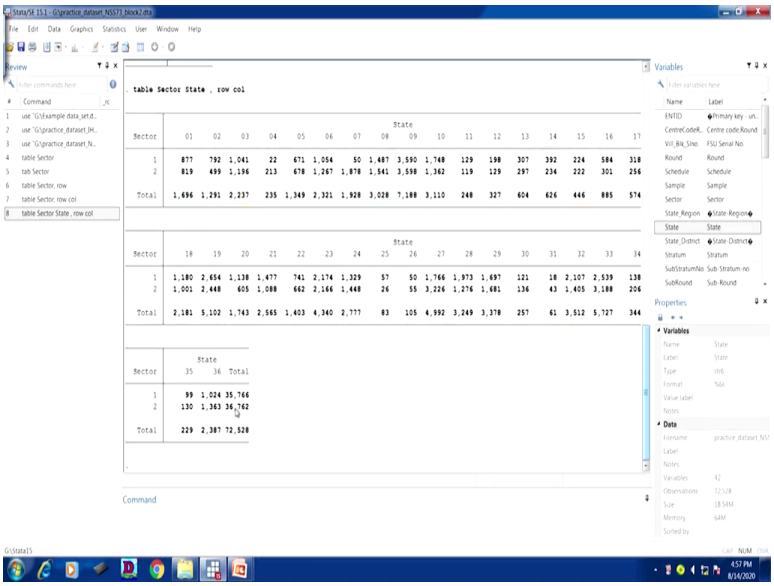
Is not it interesting for you? In this context this is the data we opened for you is from the sample data from NSS 73rd round that is on unorganized enterprises in India. The sample we have taken that we have discussed in the previous lectures on the sample be restricted to 72,528 observations. In the sector rural and urban we are going to show you how to code is also. If it is given like one and two, what one stands for here and what two stands for here. One maybe for rural and two maybe for urban. So we need to clarify it through a labeling. We are going to understand variable labeling as well in this lecture.

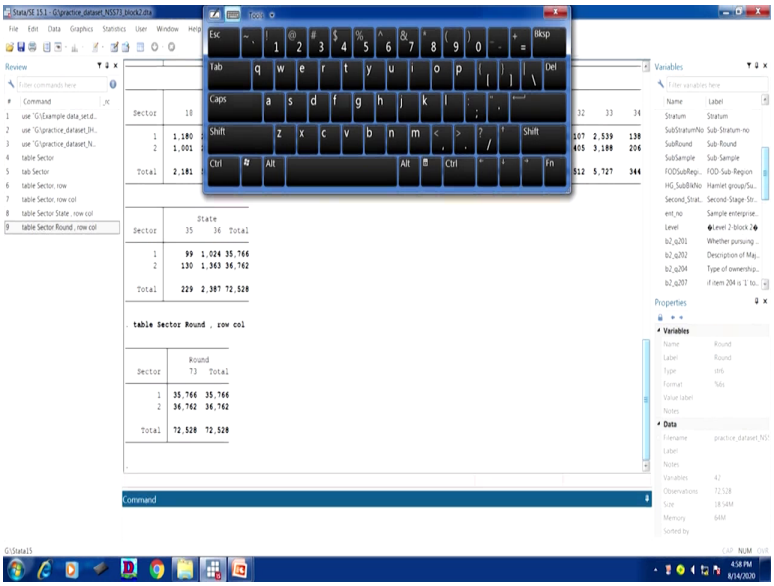
So let me proceed once again to the understanding we did. So table and tabulate, tabulate gives only frequencies, while tab or tabulate gives you frequencies along with percentages of their frequency and cumulative percentages, so this is clarified.

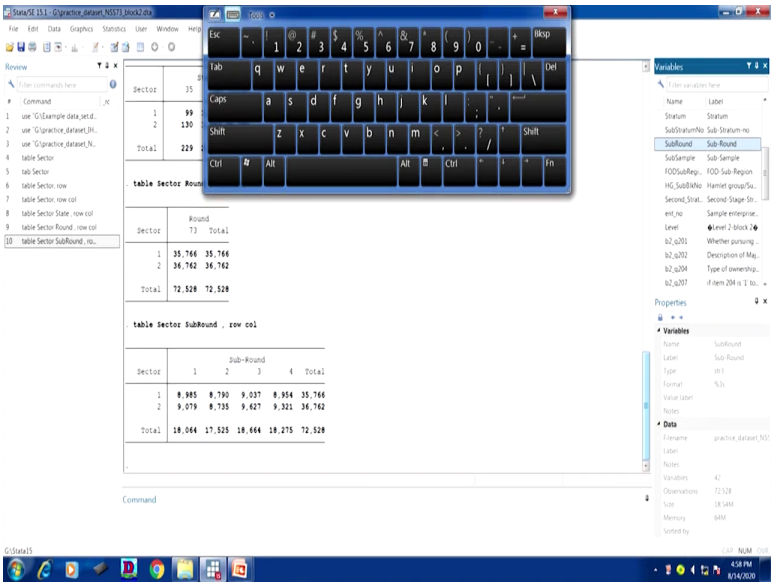
(Refer Slide Time: 05:35)






Stata/SE 15.1 - G:\practice_dataset_NSS73_block2.dta
File Edit Data Graphics Statistics User Window Help
Review
Variables
Properties
Command
. table Sector State , row col
. table Sector Round , row col
G:\Stata15
4:57 PM
8/14/2020


Stata/SE 15.1 - G:\practice_dataset_NSS73_block2.dta
File Edit Data Graphics Statistics User Window Help
Review
Variables
Properties
Command
. table Sector State , row col
. table Sector Round , row col
G:\Stata15
4:58 PM
8/14/2020



This command will give information about how many observations fall into each category and also for getting row total or column total with table command we have to specify it as options like these, table and variable name then comma row or column, like this. So you have to go back to the previous command table here it automatically comes when I click on my keyboard that page up. If I just click on the page up, it automatically comes.

Then you have to add a command like this comma then space row, you have to type row wise or column wise. Suppose I simply enter with row. It gives row wise total. Suppose I enter column wise, look at this, row and column both. It gives information of the same, because we are only taking one variable. If you are taking two variables simultaneously, then we have better clarity.

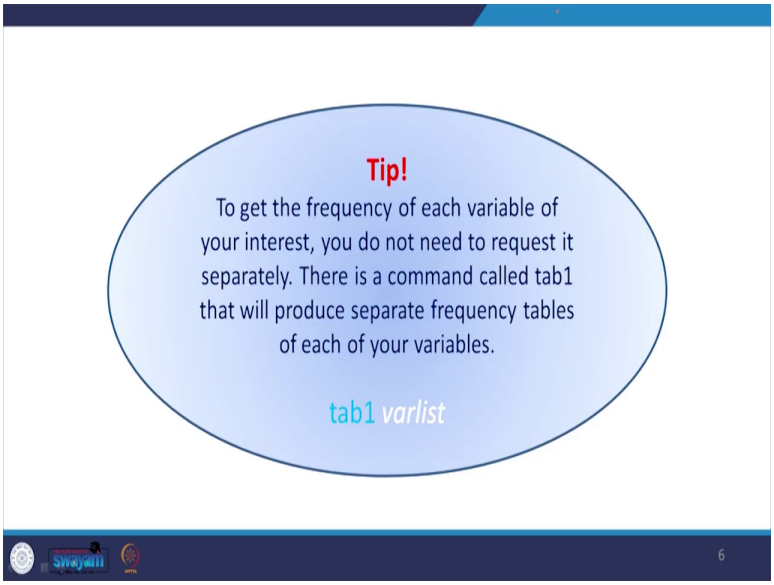
If I just add two variables once again, so sector then also I am adding another variable here, let it be state enter. It gives me at the end like this. It gives me information like this. I will show you. This is at the extreme right here. This side it gives total, this is row wise total and at the bottom all the states and their total is also coming that is column wise. So all the columns so far since this is, the window is not kept within one, within the visibility, so it gives a breaking structure.

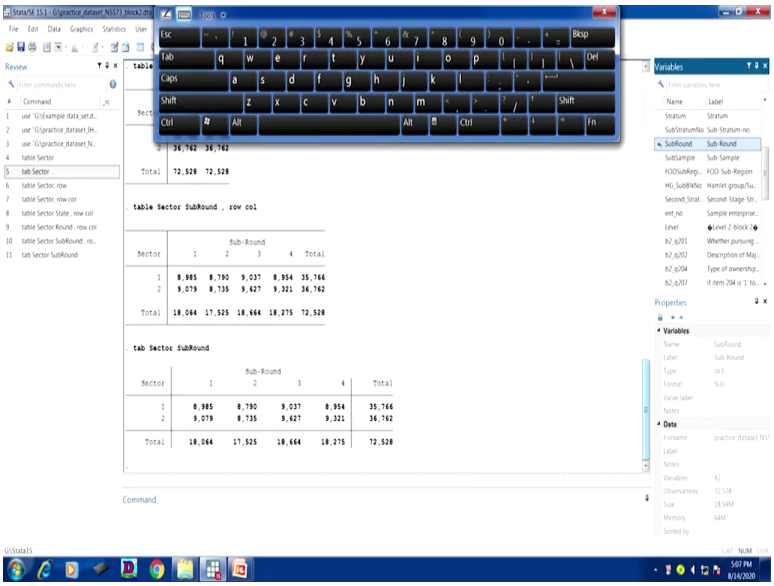
And instead of that, if I just change state to another variable, let it be round, instead of sector, it gives us the result like this. So round 73rd since it is of one round so more information is not there for us. We can clarify from here there are some other categorical information. You can get better result out of it. There are sub-rounds, let me test with the sub-rounds if it gives information. there are one, two, three, four sub-rounds and on the left hand side you have taken

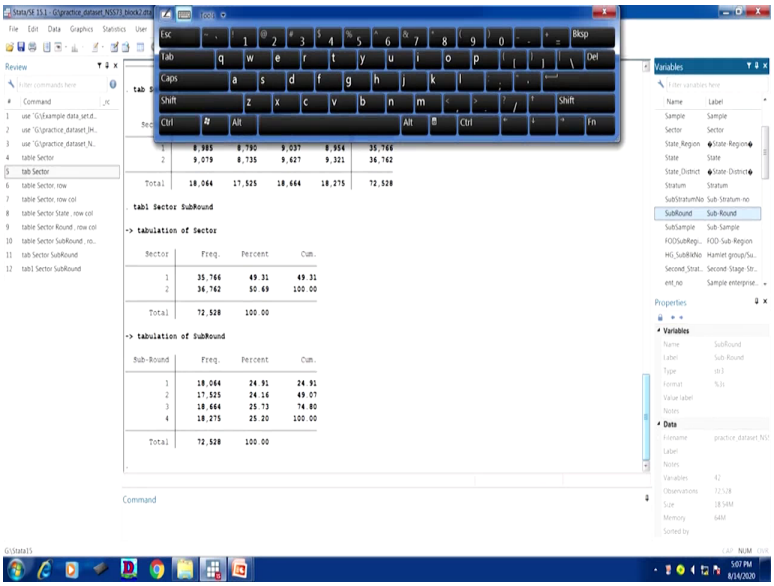
sector. So it gives a clear picture, right hand total and the bottom total. So row wise total and column wise total are resulted because of that particular command we have given to the Stata.

Let us come to our further details. In this particular example one is for rural and two is for urban. Rural total frequency which is shown in the data we have already shown to you, by number 35,766, I think that has already come here and for the rural area and for urban it is of 36,762.

(Refer Slide Time: 09:49)







So what we should also understand that in the table command, we are supposed to derive so many tables together and their frequencies, but if you are interested in the frequency separately, like to get the frequency of each variable of your interest, you do not need to request it separately. So there is a command called tab1 then that will produce separate frequency tables of each of your variables.

For example, here let me just do one thing here for you, like if I go by tab here, tab command which I did, if I just click on sub-round along with that it gives me cross-tabulation. It is not giving the frequencies of each separately, but I am interested in sector and sub-round and their frequencies only. So what I will do, I have to enter tab1 command. Tab1 then you can enter these things together sector, then if you enter sub-round, you can also get that, sub-round here. So your result will be like this. So it gives the sector frequency as well as it gives frequency of sub-rounds. That is also important where the tab command, ta command by taking together the variables, it does not give frequencies. Tab1 command is going to be useful more.

(Refer Slide Time: 12:05)

❑Two way tables for Categorical Variables

- ❑**Cross tabulation** is used to display the common distribution of two variables.
- ❑Syntax of table command:

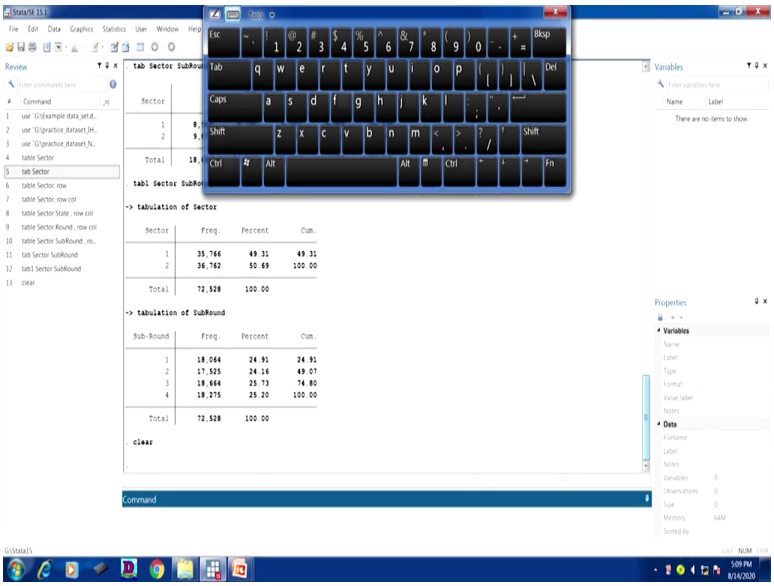
  table *var1 var2*

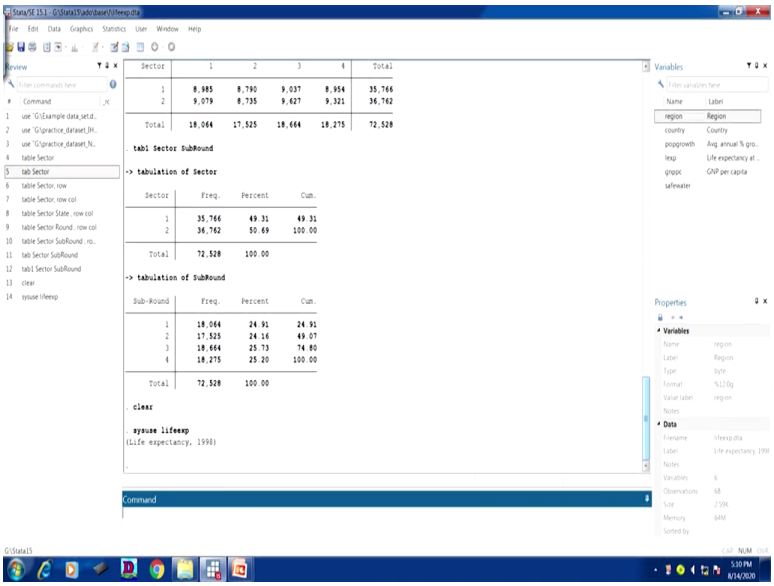
- ❑Syntax of tabulate command :

  tab *var1 var2*

- ❑The table will usually be easier to read if the variable with the most unique values is listed first.

Note!
The first variable you list will be placed in *rows* and the second in *columns*.





Sector	1	2	3	4	Total
1	8,985	8,790	9,037	8,954	35,766
2	9,079	8,735	9,627	9,321	36,762
Total	18,064	17,525	18,664	18,275	72,528

```
. tabl Sector SubRound
```

```
-> tabulation of Sector
```

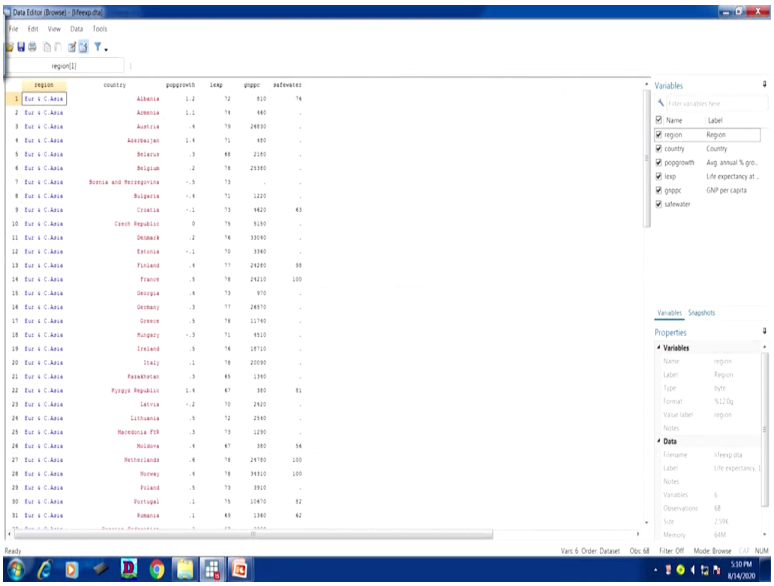
Sector	Freq.	Percent	Cum.
1	35,766	49.31	49.31
2	36,762	50.69	100.00
Total	72,528	100.00	

```
-> tabulation of SubRound
```

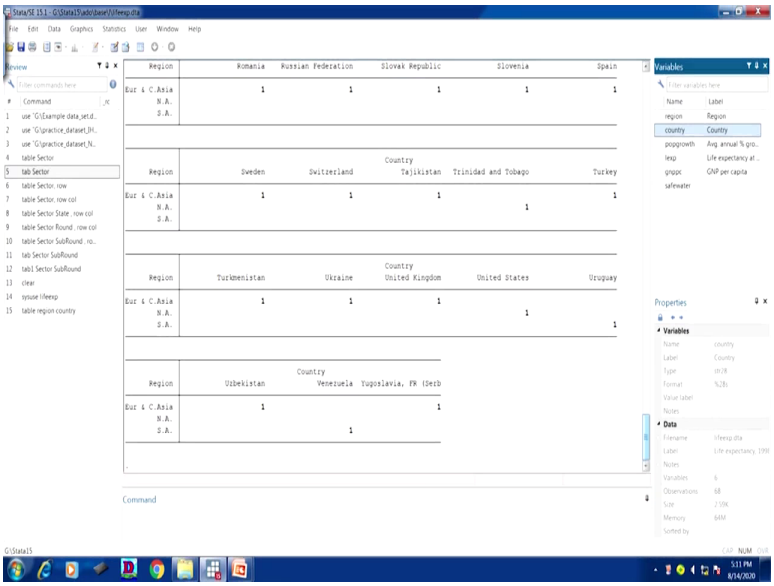
Sub-Round	Freq.	Percent	Cum.
1	18,064	24.91	24.91
2	17,525	24.16	49.07
3	18,664	25.73	74.80
4	18,275	25.20	100.00
Total	72,528	100.00	

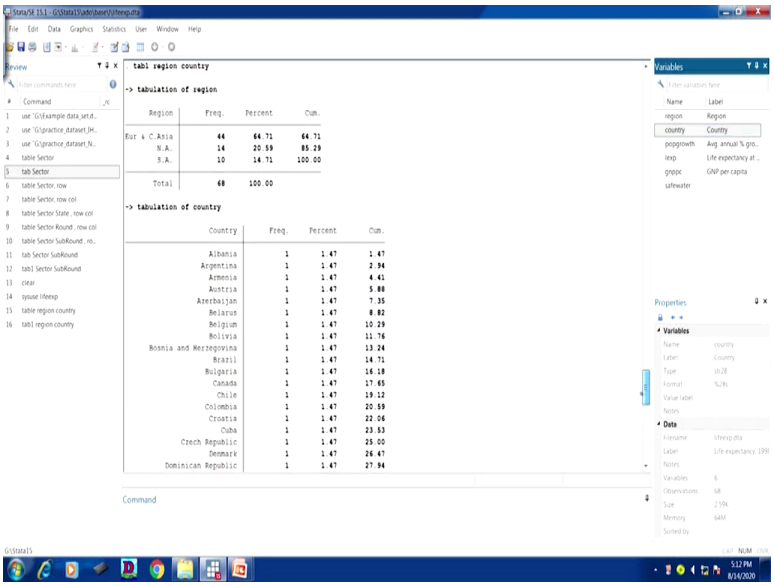
```
. clear

. sysuse lifeexp
(Life expectancy, 1998)
```



	region	country	popgrowth	lexp	gnppc	safewater
1	Eur & C.Asia	Albania	1.2	72	810	76
2	Eur & C.Asia	Armenia	1.1	74	460	.
3	Eur & C.Asia	Austria	.4	79	24830	.
4	Eur & C.Asia	Azerbaijan	1.4	71	480	.
5	Eur & C.Asia	Belarus	.3	68	2180	.
6	Eur & C.Asia	Belgium	.2	78	25380	.
7	Eur & C.Asia	Bosnia and Herzegovina	-.5	73	.	.
8	Eur & C.Asia	Bulgaria	-.4	71	1220	.
9	Eur & C.Asia	Croatia	-.1	73	4620	63
10	Eur & C.Asia	Czech Republic	0	75	5150	.
11	Eur & C.Asia	Denmark	.2	76	33040	.
12	Eur & C.Asia	Estonia	-.1	70	3360	.
13	Eur & C.Asia	Finland	.4	77	24280	98
14	Eur & C.Asia	France	.5	78	24210	100
15	Eur & C.Asia	Georgia	.4	73	970	.
16	Eur & C.Asia	Germany	.3	77	24570	.
17	Eur & C.Asia	Greece	.5	78	11760	.
18	Eur & C.Asia	Hungary	-.3	71	4510	.
19	Eur & C.Asia	Ireland	.5	76	18710	.
20	Eur & C.Asia	Italy	.1	78	20090	.
21	Eur & C.Asia	Kazakhstan	.3	65	1360	.
22	Eur & C.Asia	Kyrgyz Republic	1.4	67	380	81
23	Eur & C.Asia	Latvia	-.2	70	2420	.
24	Eur & C.Asia	Lithuania	.5	72	2560	.
25	Eur & C.Asia	Macedonia FYR	.3	73	1290	.
26	Eur & C.Asia	Moldova	.4	67	380	56
27	Eur & C.Asia	Netherlands	.6	78	24780	100
28	Eur & C.Asia	Norway	.4	78	34310	100
29	Eur & C.Asia	Poland	.5	73	3910	.
30	Eur & C.Asia	Portugal	.1	75	10470	82
31	Eur & C.Asia	Romania	.1	69	1360	62
32	Eur & C.Asia	Russian Federation	-.1	67	2680	.





More interestingly for the researcher to understand is the two-way table. Often the case we are interested in finding two-way table. How to get the two-way table? let me give you some examples for your interest. Like in this enterprise survey, we have information on gender and we have also information on sector.

That is sector, we mean rural and urban. Rural how much women represent and urban how much women represent or male represent in the unorganized enterprises can be extracted together from this information so that is called two-way tabulation. So, basically, on one-way you are including sector, on another way you are including your gender. So that is very much possible.

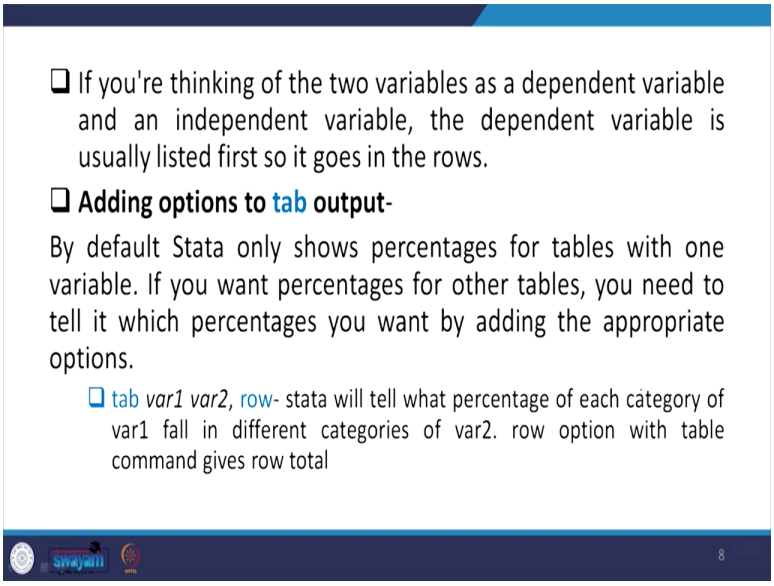
From our previous example if I operate like this. If I simply say, I am just going to clear the data from here, like if I just type clear, I can load previous data also, it has cleared our data from the Stata. You look at the variable window, there is no data. You cannot operate now. I am going to load with the lifeexp data or the system data which we discussed in the previous week lectures. Sysuse, auto directory whichever you do it will give you the data, lifeexp, if I enter now it has loaded the system use data before us.

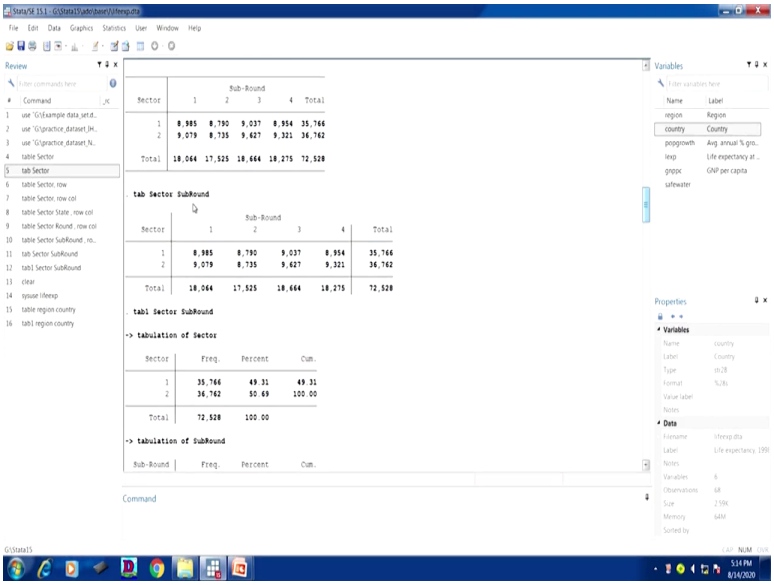
What I wanted to check, which kind of data it is. Let us check. I am going for the data browse. the country is in red color, so it is a string data. As I told you, numeric or mathematical calculation is not possible unless you convert that particular data to numeric number. The blue one is called labeled numeric and the black one is purely numeric, where we can able to do certain mathematical operations on it. let me close the data.

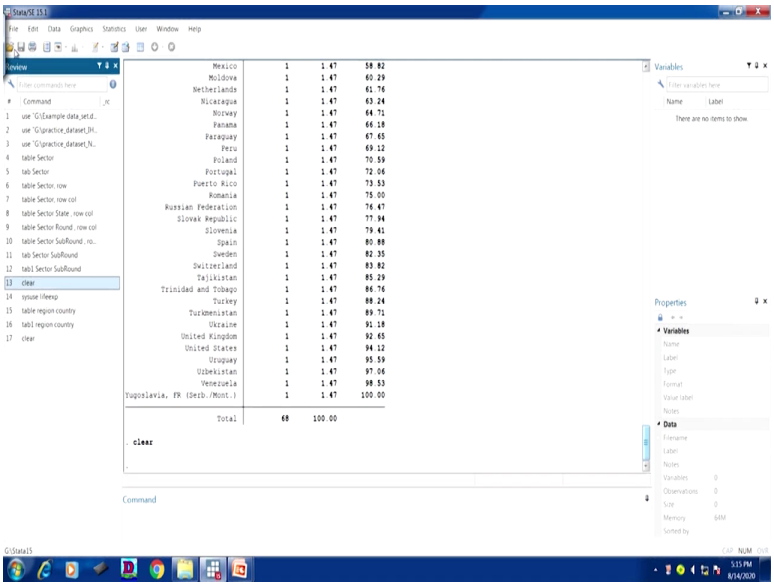
And I have to do, this is country, let me have a table command now which I did, table then region then country. If it is coded it automatically come. It is like this. but if it is table, so it has given information like this, all table region wise, then country wise. So cross-tab data has been popped up. But if I go by tab1 of these two together region and country, it gives the frequency information. Just tabulate together that gives a two-way table. But here we are getting the frequency of each, if you have entered tab1 command together.

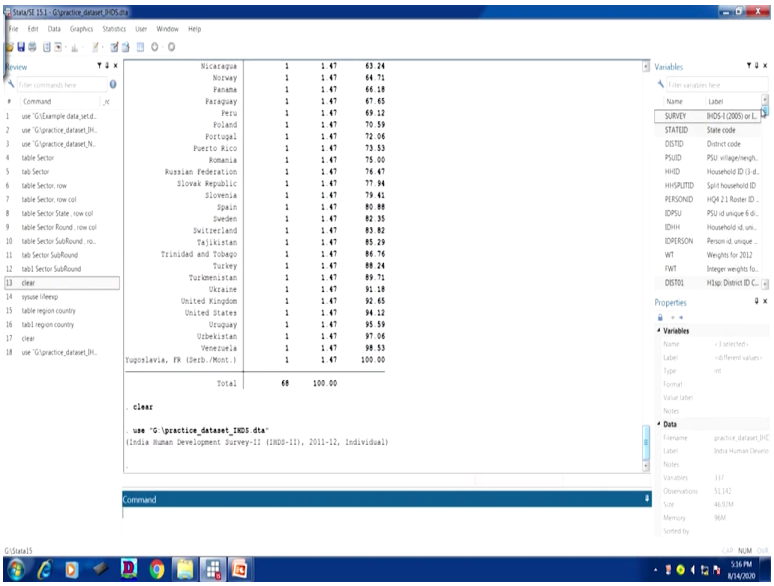
Cross-tabulation is important because we can get the information if there are more relationship. In most of our research papers, we do take the help of cross-tabulation. And this is useful to display the common distribution of two variables. And the syntax of this one I have already shown you that is either you go by table variable 1 and 2 or you go by tab variable 1 and 2. Then the table will usually be easier to read if the variable with the most unique values is listed.

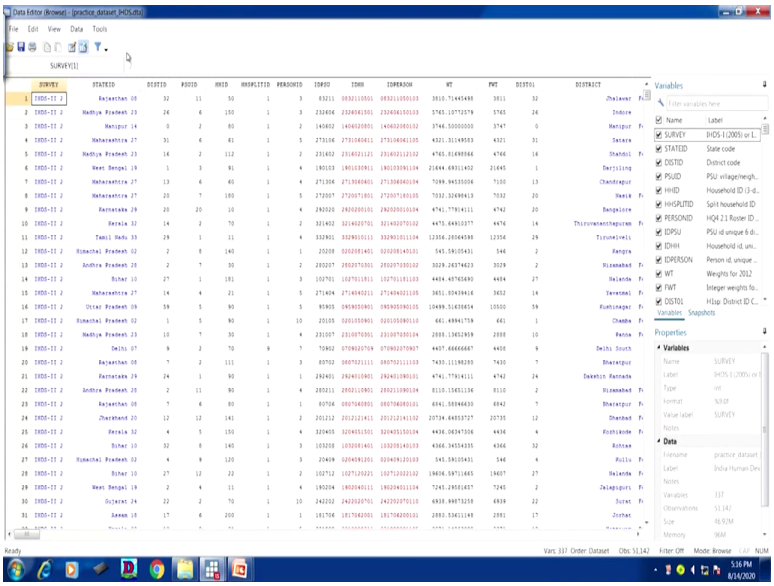
(Refer Slide Time: 16:56)

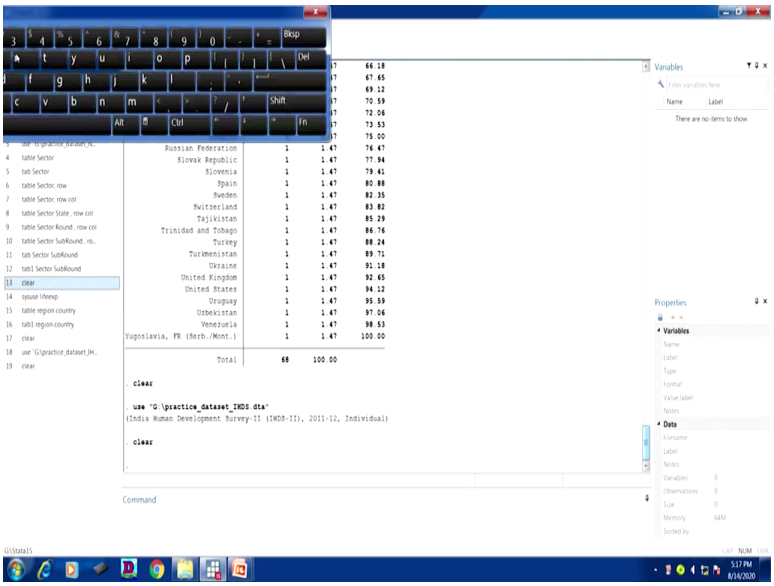


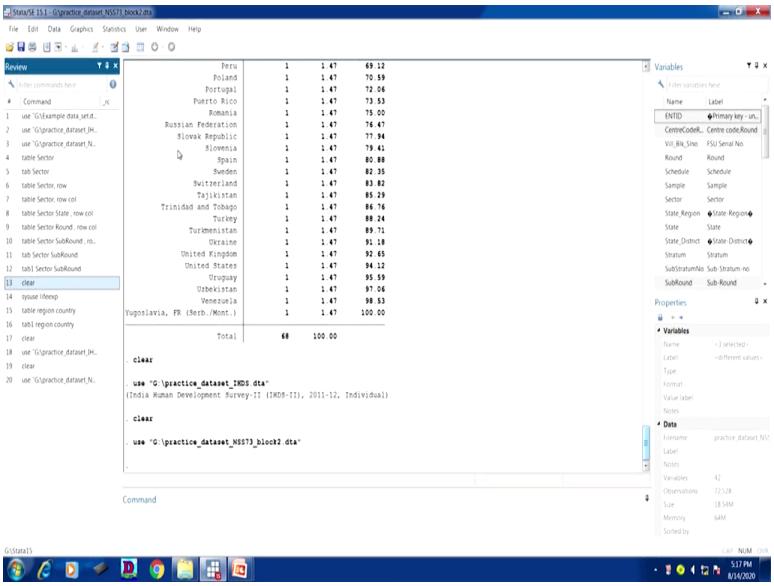



Stata/SE 15.1
File Edit Data Graphics Statistics User Window Help
Review
1 use "G:\Example data_set.d...
2 use "G:\practice_dataset_IH...
3 use "G:\practice_dataset_N...
4 table Sector
5 tab Sector
6 table Sector, row
7 table Sector, row col
8 table Sector State , row col
9 table Sector Round , row col
10 table Sector SubRound , ro...
11 tab Sector SubRound
12 tab1 Sector SubRound
13 clear
14 sysuse lifeexp
15 table region country
16 tab1 region country
17 clear
Mexico 1 1.47 58.82
Moldova 1 1.47 60.29
Netherlands 1 1.47 61.76
Nicaragua 1 1.47 63.24
Norway 1 1.47 64.71
Panama 1 1.47 66.18
Paraguay 1 1.47 67.65
Peru 1 1.47 69.12
Poland 1 1.47 70.59
Portugal 1 1.47 72.06
Puerto Rico 1 1.47 73.53
Romania 1 1.47 75.00
Russian Federation 1 1.47 76.47
Slovak Republic 1 1.47 77.94
Slovenia 1 1.47 79.41
Spain 1 1.47 80.88
Sweden 1 1.47 82.35
Switzerland 1 1.47 83.82
Tajikistan 1 1.47 85.29
Trinidad and Tobago 1 1.47 86.76
Turkey 1 1.47 88.24
Turkmenistan 1 1.47 89.71
Ukraine 1 1.47 91.18
United Kingdom 1 1.47 92.65
United States 1 1.47 94.12
Uruguay 1 1.47 95.59
Uzbekistan 1 1.47 97.06
Venezuela 1 1.47 98.53
Yugoslavia, FR (Serb./Mont.) 1 1.47 100.00
Total 68 100.00
. clear
.
Command
Variables
Filter variables here
Name Label
There are no items to show.
Properties
Variables
Data
G:\Stata15
CAP NUM OVR
5:15 PM
8/14/2020

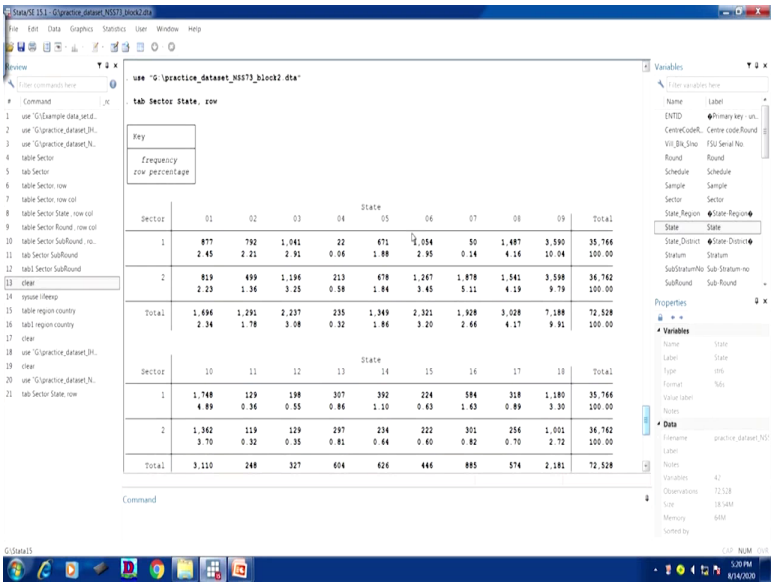

Stata/SE 15.1 - G:\practice_dataset_IHDS.dta
File Edit Data Graphics Statistics User Window Help
Review
1 use "G:\Example data_set.d...
2 use "G:\practice_dataset_IH...
3 use "G:\practice_dataset_N...
4 table Sector
5 tab Sector
6 table Sector, row
7 table Sector, row col
8 table Sector State , row col
9 table Sector Round , row col
10 table Sector SubRound , ro...
11 tab Sector SubRound
12 tab1 Sector SubRound
13 clear
14 sysuse lifeexp
15 table region country
16 tab1 region country
17 clear
18 use "G:\practice_dataset_IH...
Nicaragua 1 1.47 63.24
Norway 1 1.47 64.71
Panama 1 1.47 66.18
Paraguay 1 1.47 67.65
Peru 1 1.47 69.12
Poland 1 1.47 70.59
Portugal 1 1.47 72.06
Puerto Rico 1 1.47 73.53
Romania 1 1.47 75.00
Russian Federation 1 1.47 76.47
Slovak Republic 1 1.47 77.94
Slovenia 1 1.47 79.41
Spain 1 1.47 80.88
Sweden 1 1.47 82.35
Switzerland 1 1.47 83.82
Tajikistan 1 1.47 85.29
Trinidad and Tobago 1 1.47 86.76
Turkey 1 1.47 88.24
Turkmenistan 1 1.47 89.71
Ukraine 1 1.47 91.18
United Kingdom 1 1.47 92.65
United States 1 1.47 94.12
Uruguay 1 1.47 95.59
Uzbekistan 1 1.47 97.06
Venezuela 1 1.47 98.53
Yugoslavia, FR (Serb./Mont.) 1 1.47 100.00
Total 68 100.00
. clear
. use "G:\practice_dataset_IHDS.dta"
(India Human Development Survey-II (IHDS-II), 2011-12, Individual)
.
Command
Variables
Filter variables here
Name Label
SURVEY IHDS-I (2005) or I...
STATEID State code
DISTID District code
PSUID PSU village/neigh...
HHID Household ID (3 d...
HHSPLITID Split household ID
PERSONID HQ4.2.1 Roster ID ...
IDPSU PSU id unique 6 di...
IDHH Household id, uni...
IDPERSON Person id, unique ...
WT Weights for 2012
IWT Integer weights fo...
DIST01 H1sp: District ID C...
Properties
Variables
Data
G:\Stata15
CAP NUM OVR
5:16 PM
8/14/2020


Data Editor (Browse) - [practice_dataset_IHDS.dta]
File Edit View Data Tools
SURVEY[1]
SURVEY STATEID DISTID PSUID HHID HHSPLITID PERSONID IDPSU IDHH IDPERSON WT FWT DIST01 DISTRICT
1 IHDS-II 2 Rajasthan 08 32 11 50 1 3 83211 083211050101 083211050103 3810.71445498 3811 32 Jhalawar
2 IHDS-II 2 Madhya Pradesh 23 26 6 150 1 3 232406 232406150101 232406150103 5765.10772579 5765 26 Indore
3 IHDS-II 2 Manipur 14 0 2 80 1 2 140602 140602080101 140602080102 3746.50000000 3747 0 Manipur
4 IHDS-II 2 Maharashtra 27 31 6 41 1 5 273106 273106060111 273106060105 4321.31149583 4321 31 Satara
5 IHDS-II 2 Madhya Pradesh 23 16 2 112 1 2 231602 231602112101 231602112102 4765.81498066 4766 16 Shahdol
6 IHDS-II 2 West Bengal 19 1 3 91 1 4 190103 190103090111 190103090104 21644.69311402 21645 1 Darjiling
7 IHDS-II 2 Maharashtra 27 13 6 40 1 4 271306 271306040101 271306040104 7099.94535006 7100 13 Chandrapur
8 IHDS-II 2 Maharashtra 27 20 7 180 1 5 272007 272007180105 272007180105 7032.32498413 7032 20 Wasik
9 IHDS-II 2 Karnataka 29 20 20 10 1 4 292020 292020010101 292020010104 4741.77914111 4742 20 Bangalore
10 IHDS-II 2 Kerala 32 14 2 70 1 2 321402 321402070101 321402070102 4475.64910377 4476 14 Thiruvananthapuram
11 IHDS-II 2 Tamil Nadu 33 29 1 11 1 4 332901 332901011111 332901011104 12356.28064598 12356 29 Tirunelveli
12 IHDS-II 2 Himachal Pradesh 02 2 8 140 1 1 20208 020208140101 020208140101 545.59105431 546 2 Kangra
13 IHDS-II 2 Andhra Pradesh 28 2 7 90 1 2 282027 282027030101 282027030102 3029.26374623 3029 2 Nizamabad
14 IHDS-II 2 Bihar 10 27 1 181 1 3 102701 102701181111 102701181103 4404.48745690 4404 27 Nalanda
15 IHDS-II 2 Maharashtra 27 14 4 21 1 5 271404 271404021105 271404021105 3651.80439416 3652 14 Yavatmal
16 IHDS-II 2 Uttar Pradesh 09 59 5 90 1 5 95905 095905090105 095905090105 10499.51430854 10500 59 Kushinagar
17 IHDS-II 2 Himachal Pradesh 02 1 5 90 1 10 20105 020105090101 020105090110 661.49941759 661 1 Chamba
18 IHDS-II 2 Madhya Pradesh 23 10 7 30 1 4 231007 231007030101 231007030104 2888.13652959 2888 10 Panna
19 IHDS-II 2 Delhi 07 9 2 70 9 7 70902 070902070909 070902070907 4407.66666667 4408 9 Delhi South
20 IHDS-II 2 Rajasthan 08 7 2 111 1 3 80702 080702111103 080702111103 7430.11198280 7430 7 Bharatpur
21 IHDS-II 2 Karnataka 29 24 1 90 1 1 292401 292401090101 292401090101 4741.77914111 4742 24 Dakshin Kannada
22 IHDS-II 2 Andhra Pradesh 28 2 11 90 1 4 282011 282011090101 282011090104 8110.15651136 8110 2 Nizamabad
23 IHDS-II 2 Rajasthan 08 7 6 80 1 1 80706 080706080101 080706080101 6841.58046630 6842 7 Bharatpur
24 IHDS-II 2 Jharkhand 20 12 12 141 1 2 201212 201212141101 201212141102 20734.64853727 20735 12 Dhanbad
25 IHDS-II 2 Kerala 32 4 5 150 1 4 320405 320405150101 320405150104 4436.06747306 4436 4 Kozhikode
26 IHDS-II 2 Bihar 10 32 8 140 1 3 103208 103208140101 103208140103 4366.34554335 4366 32 Rohtas
27 IHDS-II 2 Himachal Pradesh 02 4 9 120 1 3 20409 020409120101 020409120103 545.59105431 546 4 Kullu
28 IHDS-II 2 Bihar 10 27 12 22 1 2 102712 102712022101 102712022102 19606.59711465 19607 27 Nalanda
29 IHDS-II 2 West Bengal 19 2 4 11 1 4 190204 190204011111 190204011104 7245.29581457 7245 2 Jalpaiguri
30 IHDS-II 2 Gujarat 24 22 2 70 1 10 242202 242202070101 242202070110 6938.99873258 6938 22 Surat
31 IHDS-II 2 Assam 18 17 6 200 1 1 181706 181706200101 181706200101 2880.53411148 2881 17 Jorhat
Variables
Filter variables here
Name Label
SURVEY IHDS-I (2005) or I...
STATEID State code
DISTID District code
PSUID PSU village/neigh...
HHID Household ID (3-d...
HHSPLITID Split household ID
PERSONID HQ4.2.1 Roster ID ...
IDPSU PSU id unique 6 di...
IDHH Household id, uni...
IDPERSON Person id, unique ...
WT Weights for 2012
FWT Integer weights fo...
DIST01 H1Isp District ID C...
Variables Snapshots
Properties
Variables
Name SURVEY
Label IHDS-I (2005) or I
Type int
Format %9.0f
Value label SURVEY
Notes
Data
Filename practice_dataset...
Label India Human Devel...
Notes
Variables 137
Observations 51,142
Size 46.92M
Memory 96M
Ready
Vars: 137 Order: Dataset Obs: 51,142 Filter: Off Mode: Browse CAP NUM
5:16 PM
8/14/2020


Bksp Del Shift Alt Ctrl Fn
1 table Sector
2 tab Sector
3 table Sector, row
4 table Sector, row col
5 table Sector State , row col
6 table Sector Round , row col
7 table Sector SubRound , ro...
8 tab Sector SubRound
9 tabl Sector SubRound
10 clear
11 sysuse lifeexp
12 table region country
13 tabl region country
14 clear
15 use "G:\practice_dataset_IH...
16 clear
Russian Federation 1 1.47 76.47
Slovak Republic 1 1.47 77.94
Slovenia 1 1.47 79.41
Spain 1 1.47 80.88
Sweden 1 1.47 82.35
Switzerland 1 1.47 83.82
Tajikistan 1 1.47 85.29
Trinidad and Tobago 1 1.47 86.76
Turkey 1 1.47 88.24
Turkmenistan 1 1.47 89.71
Ukraine 1 1.47 91.18
United Kingdom 1 1.47 92.65
United States 1 1.47 94.12
Uruguay 1 1.47 95.59
Uzbekistan 1 1.47 97.06
Venezuela 1 1.47 98.53
Yugoslavia, FR (Serb./Mont.) 1 1.47 100.00
Total 68 100.00
. clear
. use "G:\practice_dataset_IHDS.dta"
(India Human Development Survey-II (IHDS-II), 2011-12, Individual)
. clear
Command
Variables
Filter variables here
Name Label
There are no items to show.
Properties
Variables
Data
G:\Stata15
CAP NUM OVR
5:17 PM
8/14/2020

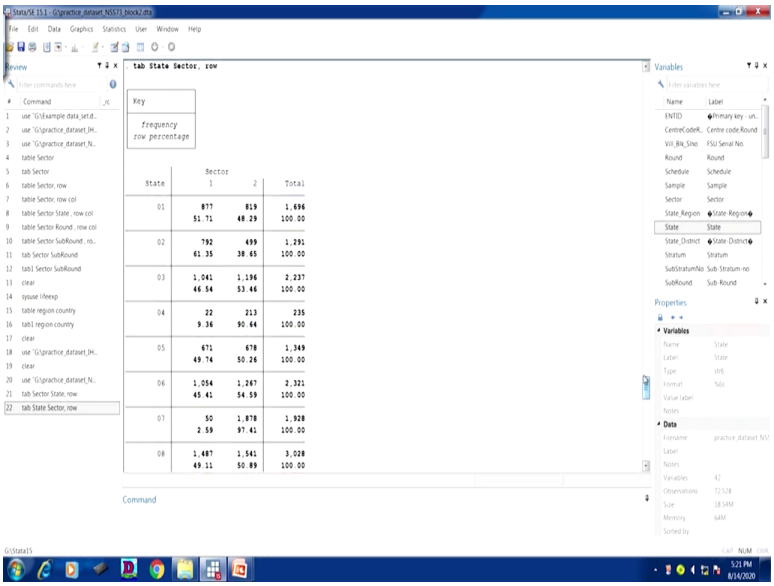


Peru	1	1.47	69.12
Poland	1	1.47	70.59
Portugal	1	1.47	72.06
Puerto Rico	1	1.47	73.53
Romania	1	1.47	75.00
Russian Federation	1	1.47	76.47
Slovak Republic	1	1.47	77.94
Slovenia	1	1.47	79.41
Spain	1	1.47	80.88
Sweden	1	1.47	82.35
Switzerland	1	1.47	83.82
Tajikistan	1	1.47	85.29
Trinidad and Tobago	1	1.47	86.76
Turkey	1	1.47	88.24
Turkmenistan	1	1.47	89.71
Ukraine	1	1.47	91.18
United Kingdom	1	1.47	92.65
United States	1	1.47	94.12
Uruguay	1	1.47	95.59
Uzbekistan	1	1.47	97.06
Venezuela	1	1.47	98.53
Yugoslavia, FR (Serb./Mont.)	1	1.47	100.00
Total	68	100.00	



Sector	State 01	02	03	04	05	06	07	08	09	Total
1	877	792	1,041	22	671	1,054	50	1,487	3,590	35,766
	2.45	2.21	2.91	0.06	1.88	2.95	0.14	4.16	10.04	100.00
2	819	499	1,196	213	678	1,267	1,878	1,541	3,598	36,762
	2.23	1.36	3.25	0.58	1.84	3.45	5.11	4.19	9.79	100.00
Total	1,696	1,291	2,237	235	1,349	2,321	1,928	3,028	7,188	72,528
	2.34	1.78	3.08	0.32	1.86	3.20	2.66	4.17	9.91	100.00

Sector	State 10	11	12	13	14	15	16	17	18	Total
1	1,748	129	198	307	392	224	584	318	1,180	35,766
	4.89	0.36	0.55	0.86	1.10	0.63	1.63	0.89	3.30	100.00
2	1,362	119	129	297	234	222	301	256	1,001	36,762
	3.70	0.32	0.35	0.81	0.64	0.60	0.82	0.70	2.72	100.00
Total	3,110	248	327	604	626	446	885	574	2,181	72,528



If you are thinking of the two variables as dependent variable and an independent variable, the dependent variable is usually listed first. So that it goes in the rows. Likewise, I will tell you, like in our result, it goes like this. since I first enter sector here then sub-round, sector goes to our row and on the column we have sub-round, if this is independent variable, but in this case this is not independent one.

So in our case suppose you are trying to understand the credit access of those from the 73rd round, if you are interested in trying to find out the credit access for male and female to those unorganized enterprises, so your credit access is considered to be the dependent variable in this context. So you first entered credit access as your first variable and then you entered your gender variable on the second variable so that it will give you the cross-tabulation correctly.

Let us come to adding, some adding options to the tab output we did. By default, Stata only shows percentages for tables with one variable. If you want percentages of another variable, you need to tell it which percentages you want by adding the appropriate options. Suppose what we have done so far, we only shown you the frequencies through table command or through tabulate command. But we are going to show you the percentages row wise or column wise.

Like I will tell what we have done in the data. Let me again go back to our 73rd round data. So I have to write or I have to once again clear this data and load the data, the 73rd data here. So the 73rd data is here for us or even we can also load the IHDS data for our operation, Stata operation it is displayed here. So this is the data I have loaded. So, I am going to just test the data to check

whether I am operating correctly or not from the beginning. So this is the IHDS 2 data, not the 73rd round data. Look at there are blue entries as well as black entries, red entries, there is so many information in the result, in the data file.

I will operate this for our understanding of those command for better interpretation. There are so many variables like this. But one problem in this data, here it requires labeling. I will tell you what do we mean labeling and how to do the labeling correctly and so that it gets easier for us to operate. Since the labeling has not been done correctly, I will prefer to open the 73rd data. So when I require labeling of the command, I will open this data once again. So I have to clear once again this as well. I am going to open 73rd data for you.

So, 73rd is here for us. We have already filtered that. I am trying to explain sector and since these are in categorical numbers, so what I do let me open here like this, tab I want to explain sector with the states. Suppose I wanted to do that, how it comes. How the sector wise distribution of enterprises across location. If it is this, what I will do, I will add percentage information instead of frequency. I have already shown you state and sector frequencies, but I am interested to look at or observe the data through their percentages.

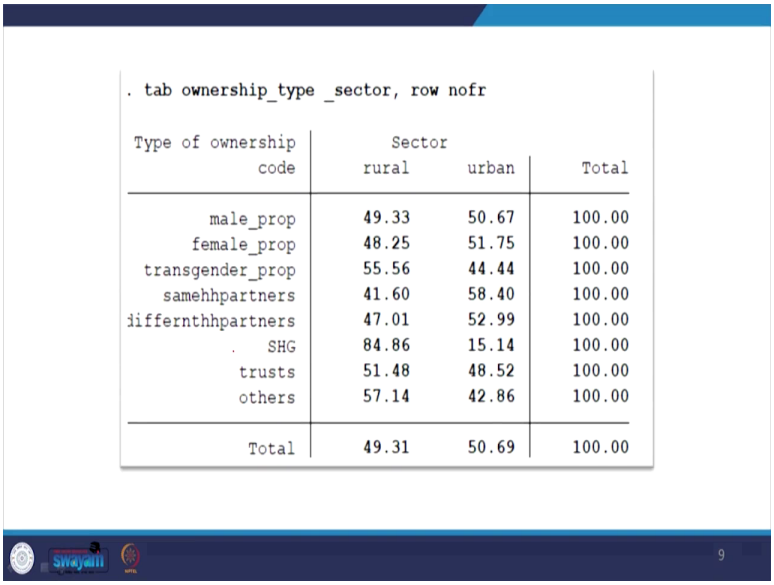
So what I will do, I will add, for example, row wise frequency, so enter. So if I have entered row, it has given me information like this. Look at carefully how it looks. Sector is on the left hand and state is on the right hand. So state is on the column and sector is on the row. Basically I am going to look at, sector on the left is given, so I am going to look at the percentage by state. So here under the state code 1, 2, 3, 4, 5, 6, 7, 8, 9, then 10, 11, 12, 13, 14, likewise other states information is there. And 1 is the sector, how much percentage of rural enterprises, unorganized enterprises are in state 1, state 5 is usually Uttarakhand.

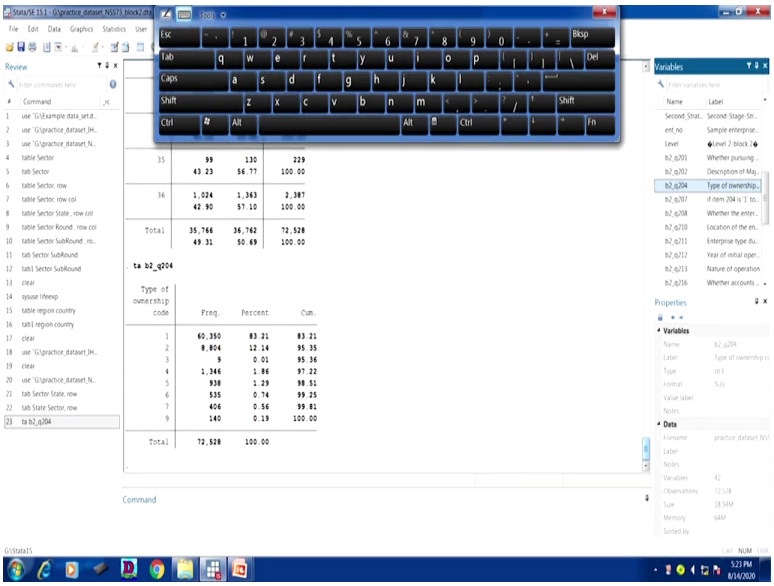
So Uttarakhand its percentage we have not yet labeled. We can label it. I will show you the command. But Uttarakhand percentage for the unorganized enterprises of women, we have already filtered this data for women. Women percentage is 1.88 out of the total enterprises and that too in our sample data. Similarly, if I just reverse this command, I will take state here. I have to change that. I am just deleting state from there. I will start with a state here. entering it. Now look at how it looks like.

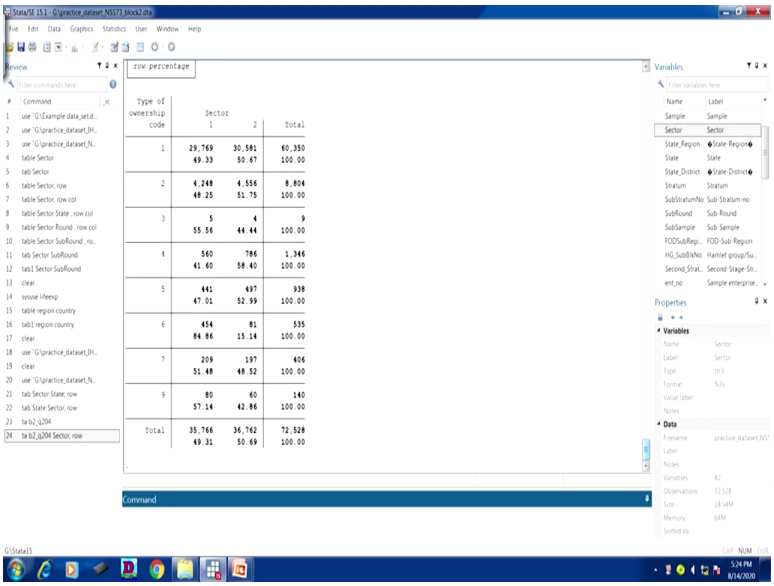
Now state is in the column, look at this, and sector is on the row. Basically if you go by every row, you have different states. Whereas each column represents sector. So how it looks like? it looks like first state, its total, have I taken percentages? yes, I have taken percentages. On the first state, out of the total 100 in the first state, the percentage of rural is of 51.71, whereas the urban representation is of 48.29. So the rural percentage is higher. Look at 05 is the code for the state Uttarakhand, 49.74 is from rural, whereas 50% from the Uttarakhand itself is from urban areas. So this is there.

I am going to clarify further. By default, Stata only shows percentage of tables with one variable, as I already told you. So we have already operated row wise percentages. So Stata will tell what percentage of each category of variable 1 fall in different categories of variable 2. So row option with table command gives row total, we have already discussed.

(Refer Slide Time: 26:12)







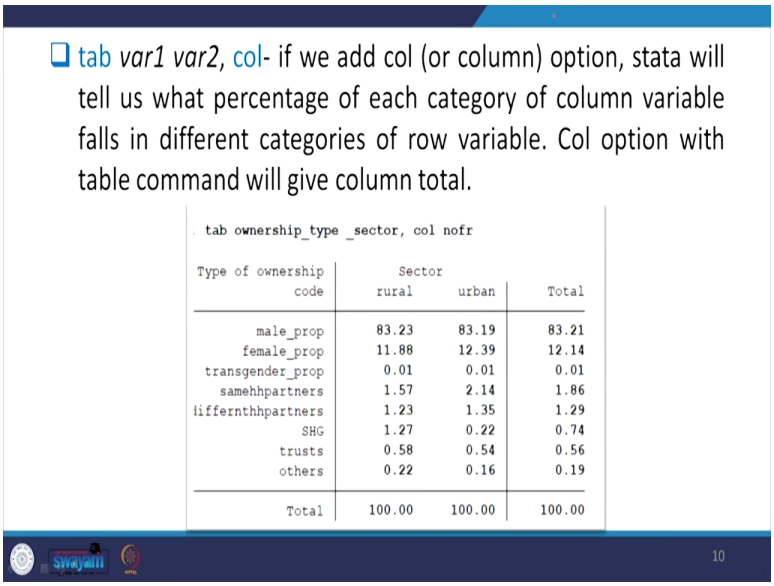
Like in our original data, not the filtered one, if I check the proprietorship, for the type of ownership if I just check whether it is there or not. let me have a look, type of ownership is there. Let me just have a check of the same by taking tab of type of ownership. I just wanted to check the frequency distribution. Yes, it is there. So we can find out the result that is available in our PPT.

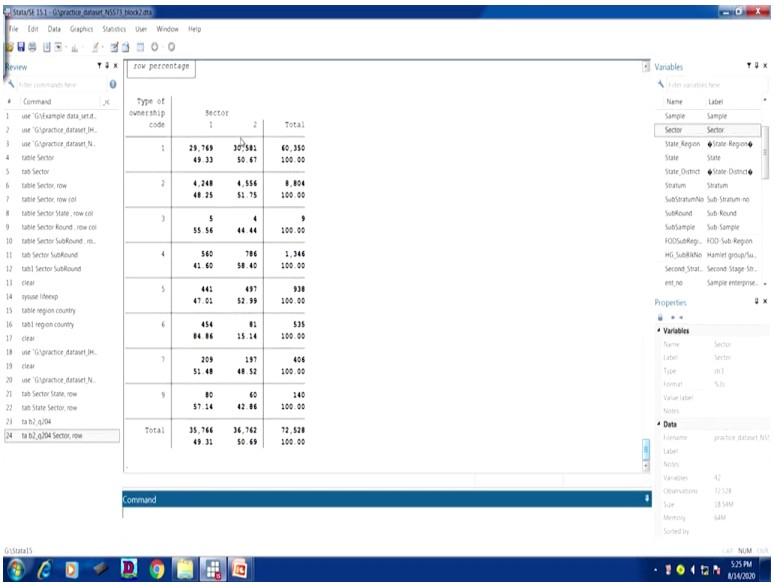
Like if I just go by the result we derived in our displayed result in this PPT, I can show you also. So what I do, we will type tab or only ta, then you click type of ownership first, then you click sector that is rural and urban. you add row and column percentages, suppose we start with row

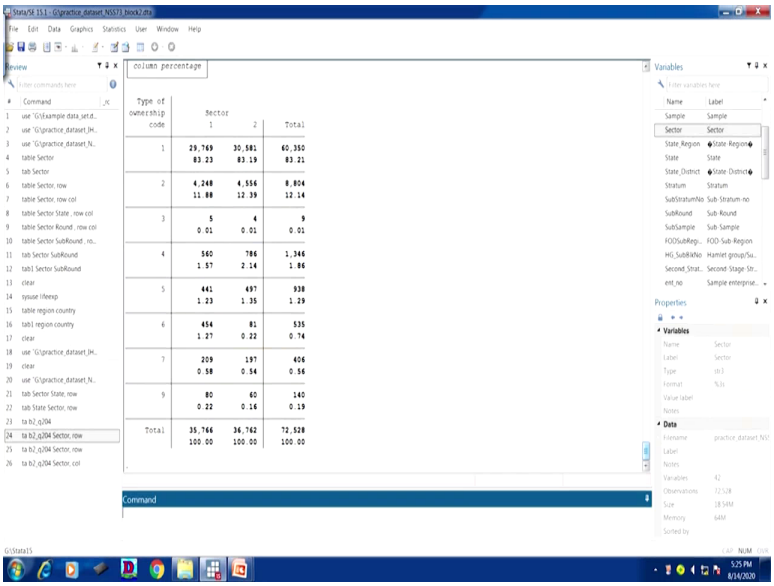
percentage and you simply enter. It gives you all the type of ownership, proprietary type or joint partnership type. There are different indicators of one to nine regarding the ownership. So it is now classified into rural and urban.

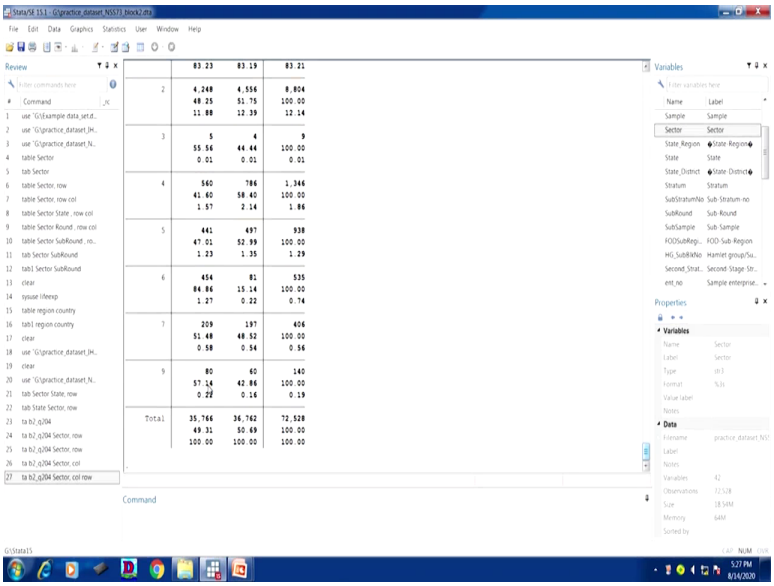
So in our original data one stands for male proprietary, two stands for female proprietary, then three stands for transgender proprietary, then same household partners are the owner, then different partner, if he is the owner then SHG Group as the owner, or the trust or others categories are there. We can find out the results accordingly. Let us make a move. I think we have already derived that. This is the one we have already displayed to you. So let me proceed.

(Refer Slide Time: 28:46)




Stata/SE 15.1 - G:\practice_dataset_NSS73_block2.dta
File Edit Data Graphics Statistics User Window Help
Review
Command
1 use "G:\Example data_set.d...
2 use "G:\practice_dataset_IH...
3 use "G:\practice_dataset_N...
4 table Sector
5 tab Sector
6 tabie Sector, row
7 tabie Sector, row col
8 tabie Sector State , row col
9 tabie Sector Round , row col
10 tabie Sector SubRound , ro...
11 tab Sector SubRound
12 tab1 Sector SubRound
13 clear
14 syuse hfeexp
15 table region country
16 tab1 region country
17 clear
18 use "G:\practice_dataset_IH...
19 clear
20 use "G:\practice_dataset_N...
21 tab Sector State, row
22 tab State Sector, row
23 ta b2_q704
24 ta b2_q704 Sector, row
row percentage
Type of ownership code
Sector
1
2
Total
1 29,769 30,581 60,350
49.33 50.67 100.00
2 4,248 4,556 8,804
48.25 51.75 100.00
3 5 4 9
55.56 44.44 100.00
4 560 786 1,346
41.60 58.40 100.00
5 441 497 938
47.01 52.99 100.00
6 454 81 535
84.86 15.14 100.00
7 209 197 406
51.48 48.52 100.00
9 80 60 140
57.14 42.86 100.00
Total 35,766 36,762 72,528
49.31 50.69 100.00
Command
Variables
Properties
G:\Stata15
5:25 PM
8/14/2020


Stata/SE 15.1 - G:\practice_dataset_NSS73_block2.dta
File Edit Data Graphics Statistics User Window Help
Review
Command
1 use "G:\Example data_set.d...
2 use "G:\practice_dataset_IH...
3 use "G:\practice_dataset_N...
4 tabie Sector
5 tab Sector
6 tabie Sector, row
7 tabie Sector, row col
8 tabie Sector State , row col
9 tabie Sector Round , row col
10 tabie Sector SubRound , ro...
11 tab Sector SubRound
12 tab1 Sector SubRound
13 clear
14 syuse hfeexp
15 table region country
16 tab1 region country
17 clear
18 use "G:\practice_dataset_IH...
19 clear
20 use "G:\practice_dataset_N...
21 tab Sector State, row
22 tab State Sector, row
23 ta b2_q704
24 ta b2_q704 Sector, row
25 ta b2_q704 Sector, row
26 ta b2_q704 Sector, col
column percentage
Type of ownership code
Sector
1
2
Total
1 29,769 30,581 60,350
83.23 83.19 83.21
2 4,248 4,556 8,804
11.88 12.39 12.14
3 5 4 9
0.01 0.01 0.01
4 560 786 1,346
1.57 2.14 1.86
5 441 497 938
1.23 1.35 1.29
6 454 81 535
1.27 0.22 0.74
7 209 197 406
0.58 0.54 0.56
9 80 60 140
0.22 0.16 0.19
Total 35,766 36,762 72,528
100.00 100.00 100.00
Command
Variables
Properties
G:\Stata15
5:25 PM
8/14/2020

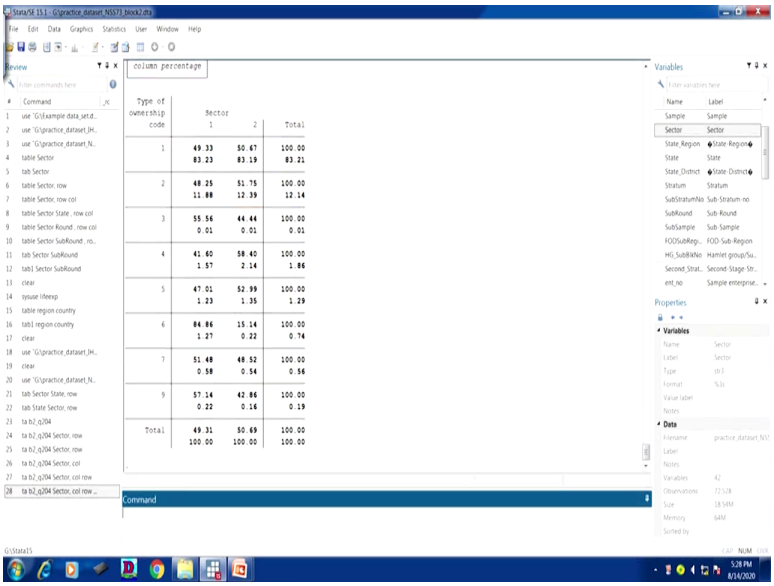


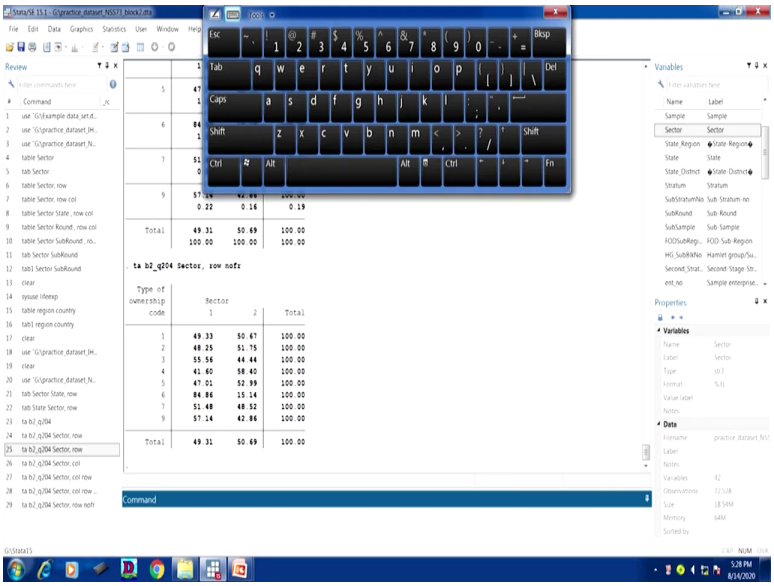
- nofrequency (or nofr) option- if you want to see only percentage, then add nofr option.
- nolabel- By default tab shows value labels for any variable that has them. If you need to see the actual values, add the nolabel option.
- missing- to check the number of observations with missing values.

- tab *var1 var2*, cell- If we add the cell option, Stata will tell us what percentage each combination of variable1 and variable2 is of the total sample.

```
. tab ownership_type _sector, cell nofr
```

Type of ownership code	Sector rural	urban	Total
male_prop	41.04	42.16	83.21
female_prop	5.86	6.28	12.14
transgender_prop	0.01	0.01	0.01
samehhpartners	0.77	1.08	1.86
iiffernthhpartners	0.61	0.69	1.29
SHG	0.63	0.11	0.74
trusts	0.29	0.27	0.56
others	0.11	0.08	0.19
Total	49.31	50.69	100.00





This is what we have shown you the row wise percentage. Basically, out of the total row, what is their individual percentages. So row wise 100% is shown here. Out of the 100%, how much representation they have. If I simply change to column, col, column wise percentages you just mark the differences carefully. Instead of the row total now here it is not 100, it is by column total that is 100. Out of 100 here, what is the individual share in rural are. Rural total or urban total is coming here.

So if I just give a command, another command here like row and column both, I am adding another one called row and column both. Look at this. It gives both the information. It gives information of row as well as column. row wise total, you can mark it from 100 here. The first mark is from, the second one is out of 100. So, 48.25 and 51.75 are added to 100. So second one is by row wise, the first one is of the frequencies and the last one is column wise. If you add these then it will add it to the required percentages, we have. I think there are other, here.

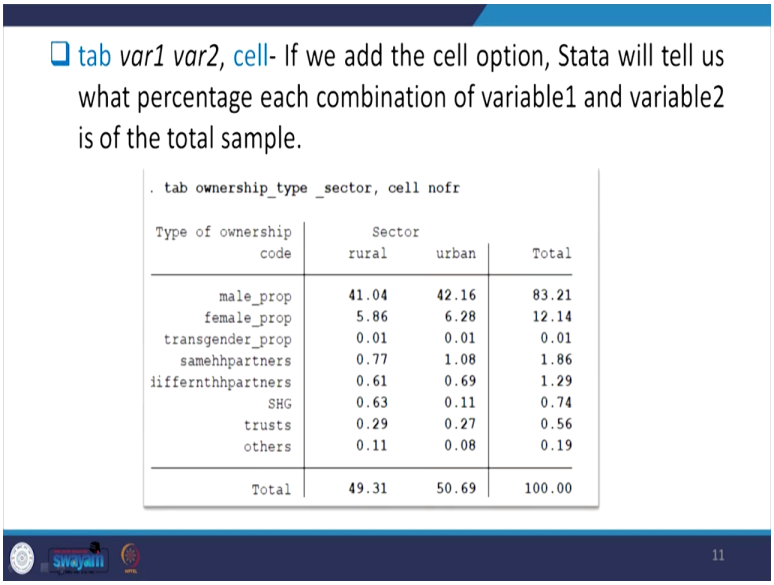
Let me add it to total, here 83.25, 11.88 then 0.01, 1.57, 1.23, 1.27, 0.58 and 0.22 this has to be added to this 100. So we can, in a single command, we can derive row wise as well as column wise percentage and for our better interpretation.

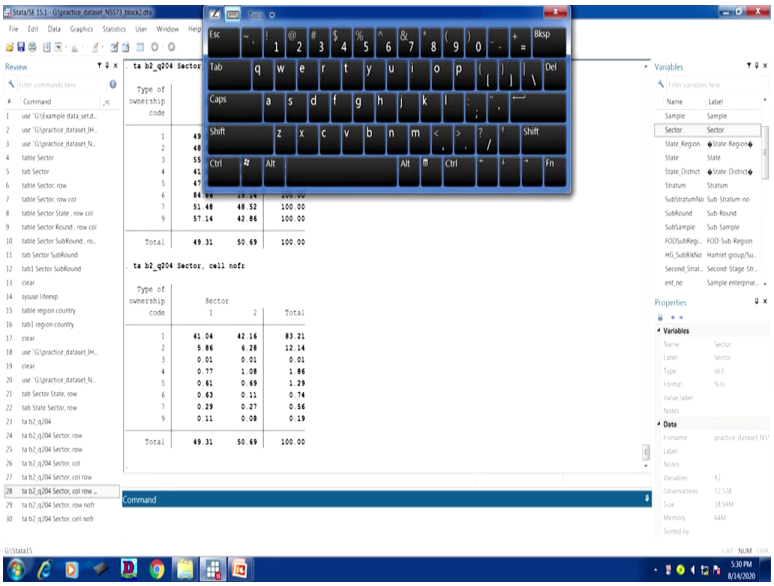
if you are interested not in frequencies, it gives a confusing figures also. Like it is giving frequencies, first entries frequency then rows wise percentage then column wise percentage, if you are only interested in percentages, you simply add nofr. I will come to the cell right now. Like nofr stands for no frequencies. Only percentages. So let me come back to the command like

this. What I did, instead of row column, what do I have to specify nofr. So nofr gives me, I have added nofr, there is no frequency. Frequency table is not there.

Similarly, for column or row also, suppose row wise, you have entered nofr, it only gives you the percentages. No frequencies given. You can simply copy this and then you copy the result to word file and you need not edit much and that will be very helpful for your answer.

(Refer Slide Time: 32:45)



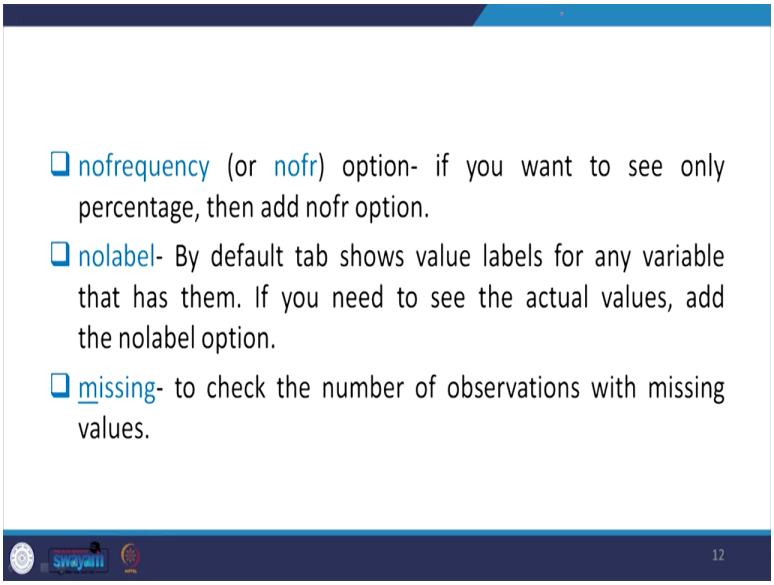


I have already clarified row wise and column wise. Another important part is called cell percentage at any particular cell and its percentage total. What is the percentages of that

particular cell? Not by row, not by column, we are saying out of the total and how much it contains. That is out of the gross total how much percentage the particular cell contains. So basically we need to specify the cell information here.

Like nofr I already did, nofr here. Here instead of that what I do, here I will delete row wise or column wise, I will only add cell. this has given me the individual cell and its percentages to the total 100. Individual 100 or row wise 100 and column wise 100 is not there. And this is also important for the researcher for their use. So this tell us what percentage of each combination of variable 1 and variable 2 is of the total sample.

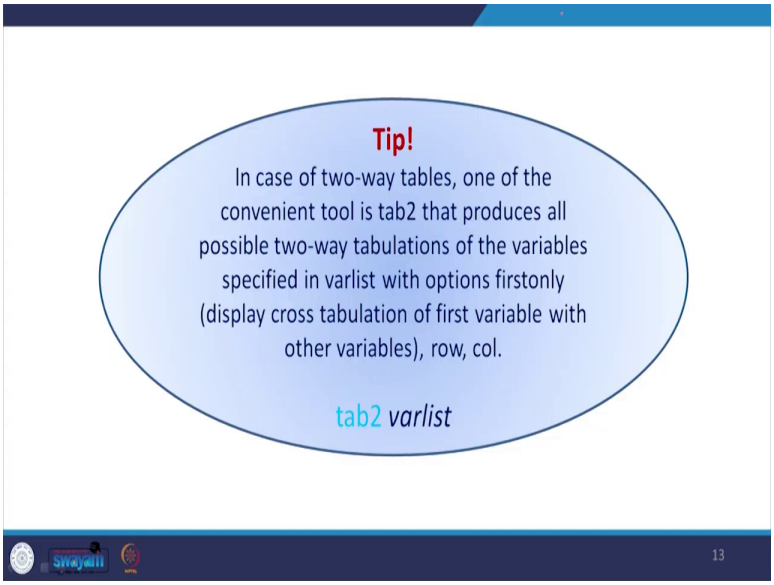
(Refer Slide Time: 34:03)



So nofr, I have already clarified, nofr avoids the frequencies. If you want to see only percentage, then add nofr that we told you. Similarly, no label, if you do not want to add the label in the command, then no label can be added. Like labels are attached, labels if you do not go by label. I will tell you label. What do you mean by label? You can operate later. By default, tab shows value labels for any variable that has them. If you need to see the actual values and add the no label option, then you can go for it.

Similarly, to check the number of observations with missing values, you can also do that. So missing information, we have already guided in the previous lecture. Through the codebook, you can also track the missing. From the summarize also you can get the missing information. So there are many other aspects to it. Let us proceed. We can operate.

(Refer Slide Time: 34:57)



One tip for you is here that in case of two-way table, one of the convenient tool is tab2 option that produces all possible two-way tabulations of the variables specified in the variable list with options first only, display cross tabulation of first variable with other variables that is through row and column. So tab2, similarly, tab1 I told you, you can go through tab2 variable list, you will also get those details correctly.

So, with this, I think I will stop here. We will continue other details regarding variable labeling in our next class. Thank you.