ICSSR
DATA SERVICE

HTTP://WWW.ICSSRDATASERVICE.IN/

ICSSR Data Service

Indian Social Science Data Repository

Stata: User Guide

Indian Council of Social Science Research

Stata 12.1: User Guide

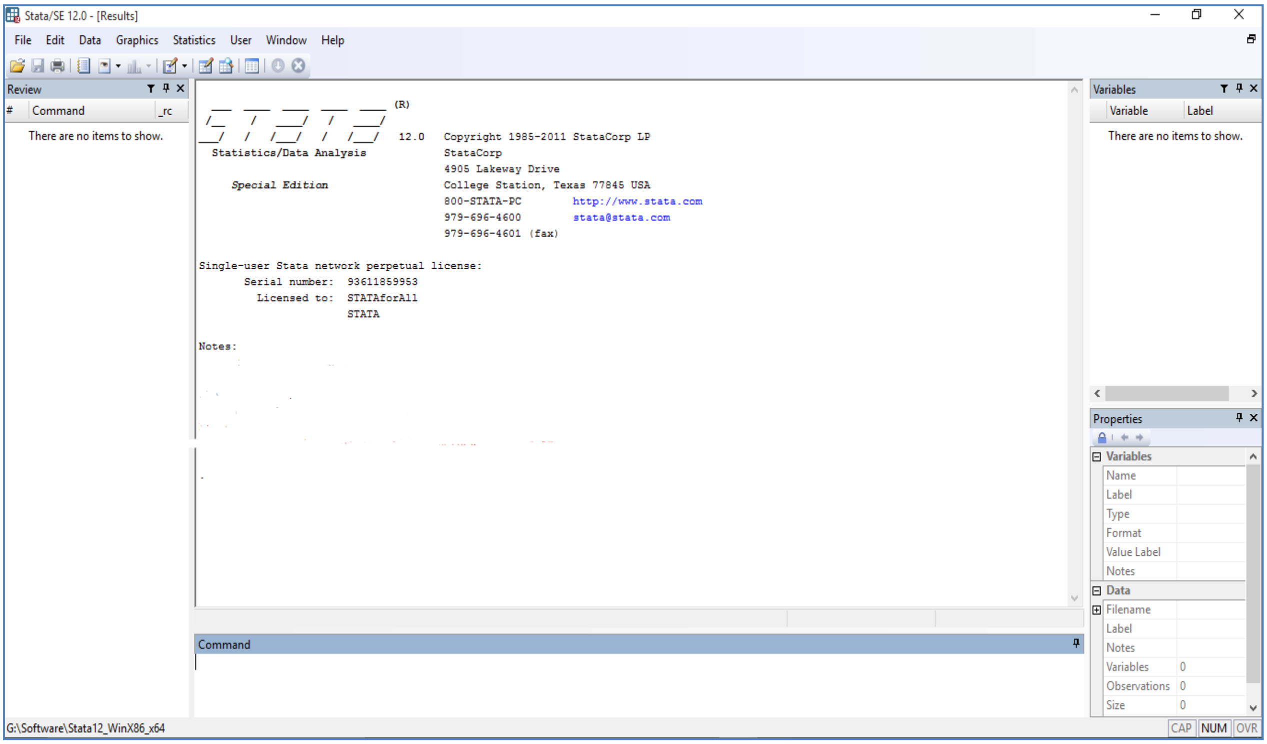
Contents:

1. Introduction

Stata is a general-purpose integrated statistical software package created in 1985 by StataCorp LP. It is a powerful statistical software that enables users to analyze, manage, and produce graphical visualizations of data. The software enables users to manipulate, analyse and produce data in final forms, like tables, graphs, figures, etc. Stata can be used either through dropdown menus or using commands. Since Stata is a programmable statistical package, new features are adding continuously by its users and developers. This manual describes features and functionalities of Stata version 12.1.

Stata improves the processing speed of system by holding the data in memory, instead of accessing it from hard drive. Moreover, it automatically adjusts the memory for the opened data file. As such, there is no need to allocate memory manually while using Stata. However, if you are using earlier versions of Stata, you may still be required to allocate memory manually.

On opening Stata, the screen that appears is shown in Fig. 1.



```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   12.0   Copyright 1985-2011 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

Single-user Stata network perpetual license:
       Serial number:  93611859953
         Licensed to:  STATAforAll
                       STATA

Notes:
```

There are no items to show.

There are no items to show.

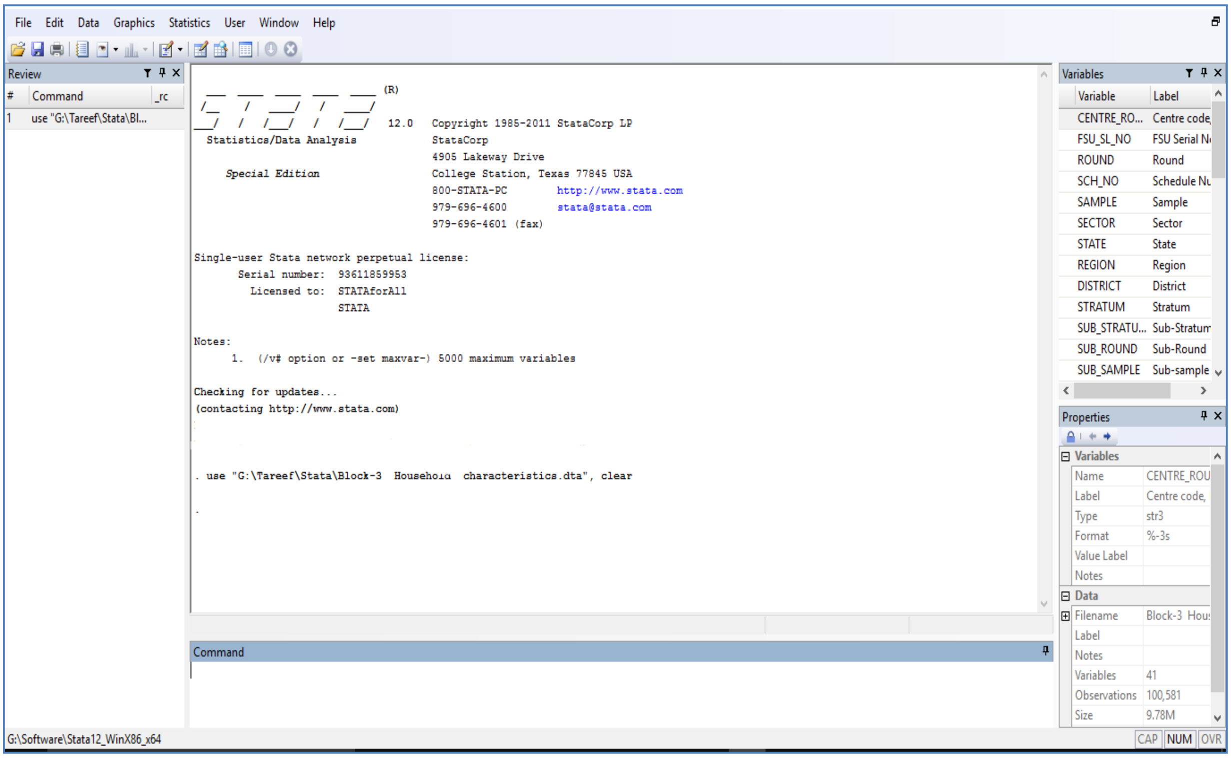
Fig. 1. Opening Screen of Stata

At shown in Fig. 1, opening screen of the Stata has five distinct panels, i.e. i) "**Command**" panel at the bottom where you need to use commands for functions and calculations supported by Stata; ii) Main panel where results of command is displayed; iii) "**Review**" panel at the left lists all the commands used during the calculation and analysis; iv) "**Variable**" panel, at the right side of screen, displays all the variables in the dataset; and v) "**Properties**", at the lower panel shows properties of each variables.

2. Opening Dataset

Example of NSSO round "Schedule 25.2: Participation and Expenditure in Education, 64thRound" dataset is used in this manual. This survey was conducted during the period of July 2007 to June 2008. Use the following dropdown function in Stata to open a dataset: **File > Open**

The screen displayed on opening the file is given in Fig. 2. You can see list of variables used in the dataset is displayed in the variables box at the right side of screen.



#	Command	_rc
1	use "G:\Tareef\Stata\Bl...	

```
                                          (R)
                               12.0   Copyright 1985-2011 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

Single-user Stata network perpetual license:
       Serial number:  93611859953
         Licensed to:  STATAforAll
                       STATA

Notes:
      1.  (/v# option or -set maxvar-) 5000 maximum variables

Checking for updates...
(contacting http://www.stata.com)


. use "G:\Tareef\Stata\Block-3  Household  characteristics.dta", clear

.
```

Variables

Variable	Label
CENTRE_RO...	Centre code,
FSU_SL_NO	FSU Serial N
ROUND	Round
SCH_NO	Schedule Nu
SAMPLE	Sample
SECTOR	Sector
STATE	State
REGION	Region
DISTRICT	District
STRATUM	Stratum
SUB_STRATU...	Sub-Stratum
SUB_ROUND	Sub-Round
SUB_SAMPLE	Sub-sample

Properties

Variables	
Name	CENTRE_ROU
Label	Centre code,
Type	str3
Format	%-3s
Value Label	
Notes	
Data	
Filename	Block-3 Hou
Label	
Notes	
Variables	41
Observations	100,581
Size	9.78M

Fig. 2. Screenshot of Opening a File in Stata

You can also open a dataset in Stata using the following command in the command panel available at the bottom of the screen: **use <filename>, clear**

This command will also lead to the opening of file as shown in Fig. 2. At the end of command, clear means that the existing data stored in the memory will be cleared.

3. Changing Directory

Use "**Changing Directory**" command, to specify the location to save the files. For example, if you want to save the files in **H:\NSSO** folder, use the following command: **cd H:\NSSO**

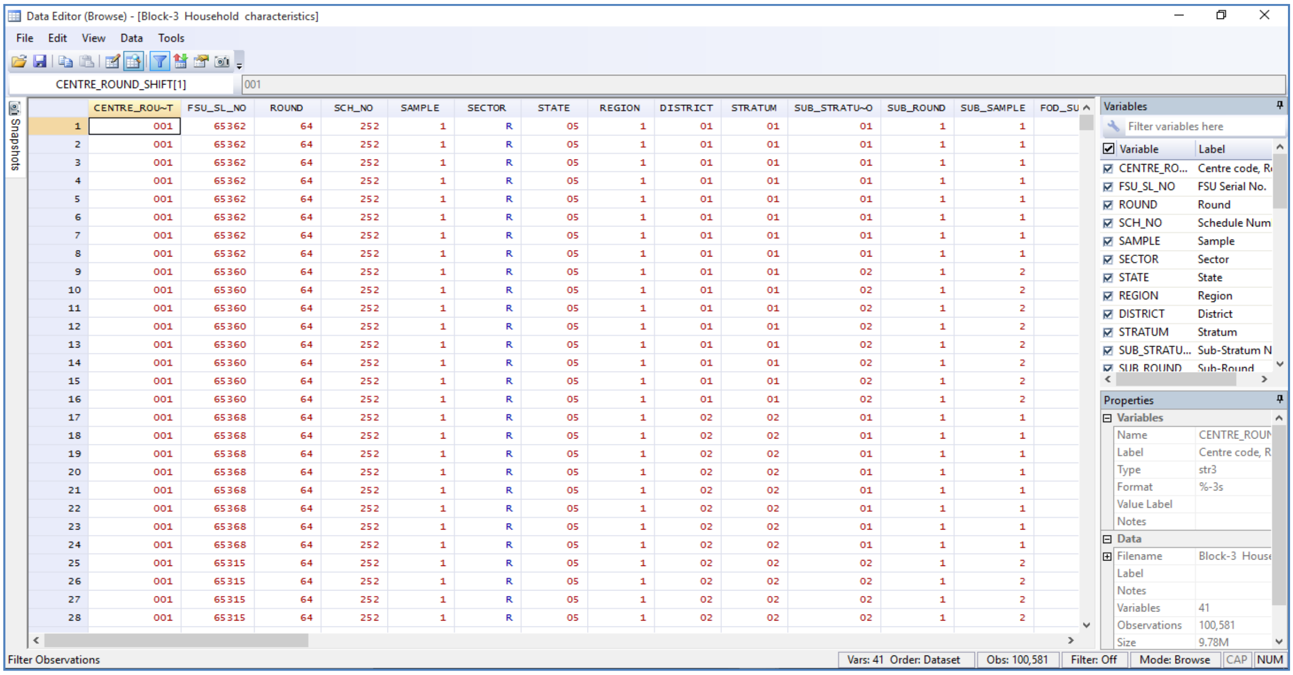
Once this command is executed, all the save and load options will work from this location only. There is no need to provide file directory specifications at the time of opening, closing or saving the dataset.

4. Viewing Data

Stata provides two options to view data, i.e. Data Editor (Browse) and Data Editor (Edit). Unlike other statistical software, data does not appear in the main window in Stata. A user is required to choose one of the two options from the tool bars to view the data in Stata.

4.1 Data Editor (Browse)

Click at **Data Editor (Browse)** from the **Tool Bars** to view or browse the uploaded data in Stata. The screen displayed on clicking at "Data Editor (Browse)" is given in Fig. 3.



CENTRE_ROUND_SHIFT[1] 001

	CENTRE_ROU~T	FSU_SL_NO	ROUND	SCH_NO	SAMPLE	SECTOR	STATE	REGION	DISTRICT	STRATUM	SUB_STRATU~O	SUB_ROUND	SUB_SAMPLE	FOD_SU
1	001	65362	64	252	1	R	05	1	01	01	01	1	1	
2	001	65362	64	252	1	R	05	1	01	01	01	1	1	
3	001	65362	64	252	1	R	05	1	01	01	01	1	1	
4	001	65362	64	252	1	R	05	1	01	01	01	1	1	
5	001	65362	64	252	1	R	05	1	01	01	01	1	1	
6	001	65362	64	252	1	R	05	1	01	01	01	1	1	
7	001	65362	64	252	1	R	05	1	01	01	01	1	1	
8	001	65362	64	252	1	R	05	1	01	01	01	1	1	
9	001	65360	64	252	1	R	05	1	01	01	02	1	2	
10	001	65360	64	252	1	R	05	1	01	01	02	1	2	
11	001	65360	64	252	1	R	05	1	01	01	02	1	2	
12	001	65360	64	252	1	R	05	1	01	01	02	1	2	
13	001	65360	64	252	1	R	05	1	01	01	02	1	2	
14	001	65360	64	252	1	R	05	1	01	01	02	1	2	
15	001	65360	64	252	1	R	05	1	01	01	02	1	2	
16	001	65360	64	252	1	R	05	1	01	01	02	1	2	
17	001	65368	64	252	1	R	05	1	02	02	01	1	1	
18	001	65368	64	252	1	R	05	1	02	02	01	1	1	
19	001	65368	64	252	1	R	05	1	02	02	01	1	1	
20	001	65368	64	252	1	R	05	1	02	02	01	1	1	
21	001	65368	64	252	1	R	05	1	02	02	01	1	1	
22	001	65368	64	252	1	R	05	1	02	02	01	1	1	
23	001	65368	64	252	1	R	05	1	02	02	01	1	1	
24	001	65368	64	252	1	R	05	1	02	02	01	1	1	
25	001	65315	64	252	1	R	05	1	02	02	02	1	2	
26	001	65315	64	252	1	R	05	1	02	02	02	1	2	
27	001	65315	64	252	1	R	05	1	02	02	02	1	2	
28	001	65315	64	252	1	R	05	1	02	02	02	1	2	

☑ Variable	Label
☑ CENTRE_RO...	Centre code, R
☑ FSU_SL_NO	FSU Serial No.
☑ ROUND	Round
☑ SCH_NO	Schedule Num
☑ SAMPLE	Sample
☑ SECTOR	Sector
☑ STATE	State
☑ REGION	Region
☑ DISTRICT	District
☑ STRATUM	Stratum
☑ SUB_STRATU...	Sub-Stratum N
☑ SUB_ROUND	Sub-Round

⊟ Variables	
Name	CENTRE_ROUN
Label	Centre code, R
Type	str3
Format	%-3s
Value Label	
Notes	
⊟ Data	
⊞ Filename	Block-3 House
Label	
Notes	
Variables	41
Observations	100,581
Size	9.78M

Vars: 41 Order: Dataset | Obs: 100,581 | Filter: Off | Mode: Browse | CAP | NUM

Fig. 3. Display of Data using Data Editor (Browse)

The **Data Editor (Browse)** option allows its users to view the data. However, users are not allowed to change the values of data. In Fig. 3, each column represents data for various variable and each row represents data of different household for given variables.

4.2 Data Editor (Edit)

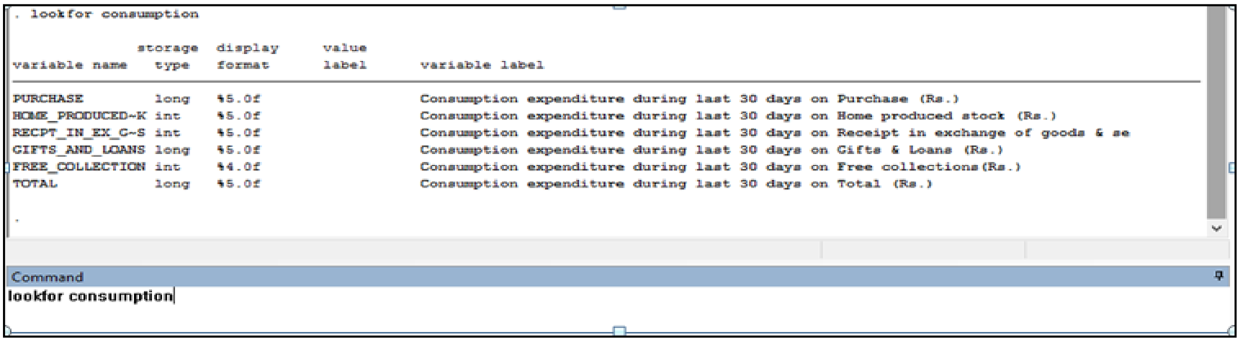
To **view** and edit the data, click at **Data Editor (Edit)** on the toolbar. Screen similar to Fig. 3 will appear .However, users can edit the values of data, numbers of variables, their properties, etc.

5. Basic Commands

This section covers basic descriptive statistics commandused in Stata for displaying descriptive statistics about type, nature and properties of data. These commands include lookfor, describe, summarize, etc.

5.1 Lookfor

Variables in a dataset can be searched using their names and descriptions using "**lookfor"** command. For example, if you want to search all variables dealing with consumption in the dataset, use "**lookfor consumption"** command. Resultantly, all variables dealing with consumptions is displayed as shown in Table 1. Table 1 shows that there are six categories of consumption variables available in the dataset. Likewise, one can search other variables by typing their names after "**lookfor**".



```
. lookfor consumption

                  storage  display     value
variable name      type    format      label       variable label
--------------------------------------------------------------------------------------------------------
PURCHASE          long     %5.0f                   Consumption expenditure during last 30 days on Purchase (Rs.)
HOME_PRODUCED~K   int      %5.0f                   Consumption expenditure during last 30 days on Home produced stock (Rs.)
RECPT_IN_EX_G~S   int      %5.0f                   Consumption expenditure during last 30 days on Receipt in exchange of goods & se
GIFTS_AND_LOANS   long     %5.0f                   Consumption expenditure during last 30 days on Gifts & Loans (Rs.)
FREE_COLLECTION   int      %4.0f                   Consumption expenditure during last 30 days on Free collections(Rs.)
TOTAL             long     %5.0f                   Consumption expenditure during last 30 days on Total (Rs.)

.
```

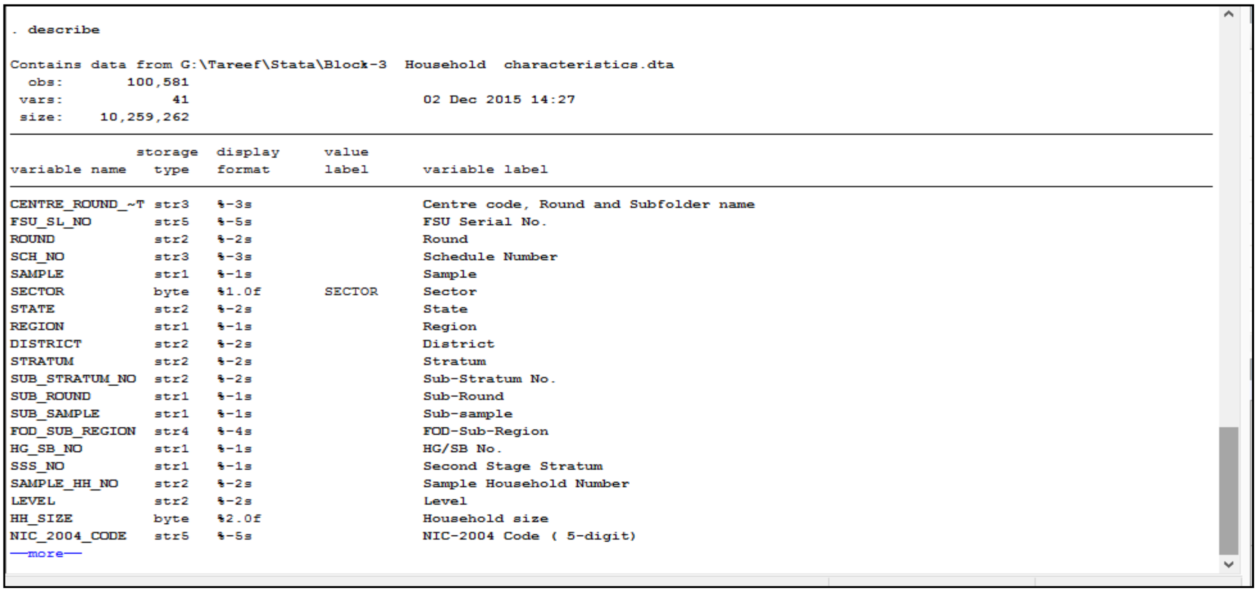
```
lookfor consumption
```

Table 1. Six Categories of Consumption Variables Displayed using "Lookfor" Command

5.2 Describe

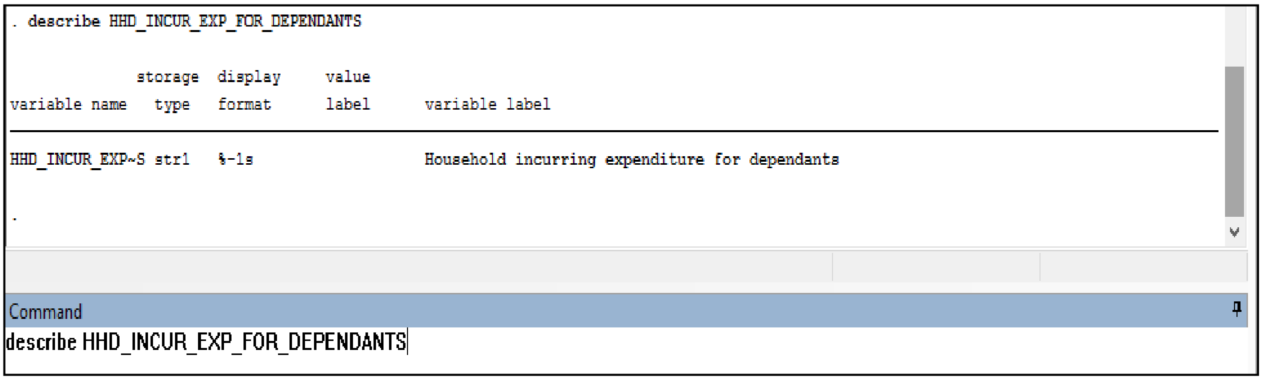
"**Describe" command**" is another way to look at the variables in the dataset. For describing all the variables in a dataset, use "**describe"** command in a given dataset and press "**enter**". Table 2 given below is displayed. However, if you wish to describe a particular variable, type "**name of variable"** after "**describe".** Resultant output is shown in Table 3. In this example, the variable "household expenditure on dependents" is described using the command given below:

describe HHD_INCUR_EXP_FOR_DEPENDANTS



```
. describe

Contains data from G:\Tareef\Stata\Block-3  Household  characteristics.dta
  obs:       100,581
 vars:            41                        02 Dec 2015 14:27
 size:    10,259,262
------------------------------------------------------------------------------------------

               storage  display     value
variable name    type   format      label      variable label
------------------------------------------------------------------------------------------
CENTRE_ROUND_~T str3    %-3s                   Centre code, Round and Subfolder name
FSU_SL_NO       str5    %-5s                   FSU Serial No.
ROUND           str2    %-2s                   Round
SCH_NO          str3    %-3s                   Schedule Number
SAMPLE          str1    %-1s                   Sample
SECTOR          byte    %1.0f       SECTOR     Sector
STATE           str2    %-2s                   State
REGION          str1    %-1s                   Region
DISTRICT        str2    %-2s                   District
STRATUM         str2    %-2s                   Stratum
SUB_STRATUM_NO  str2    %-2s                   Sub-Stratum No.
SUB_ROUND       str1    %-1s                   Sub-Round
SUB_SAMPLE      str1    %-1s                   Sub-sample
FOD_SUB_REGION  str4    %-4s                   FOD-Sub-Region
HG_SB_NO        str1    %-1s                   HG/SB No.
SSS_NO          str1    %-1s                   Second Stage Stratum
SAMPLE_HH_NO    str2    %-2s                   Sample Household Number
LEVEL           str2    %-2s                   Level
HH_SIZE         byte    %2.0f                  Household size
NIC_2004_CODE   str5    %-5s                   NIC-2004 Code ( 5-digit)
—more—
```

Table 2: Description of All Variables in a Dataset using Describe Command



```
. describe HHD_INCUR_EXP_FOR_DEPENDANTS


              storage  display     value
variable name   type   format      label      variable label
----------------------------------------------------------------------------------------------
HHD_INCUR_EXP~S str1   %-1s                   Household incurring expenditure for dependants

.
```

Command

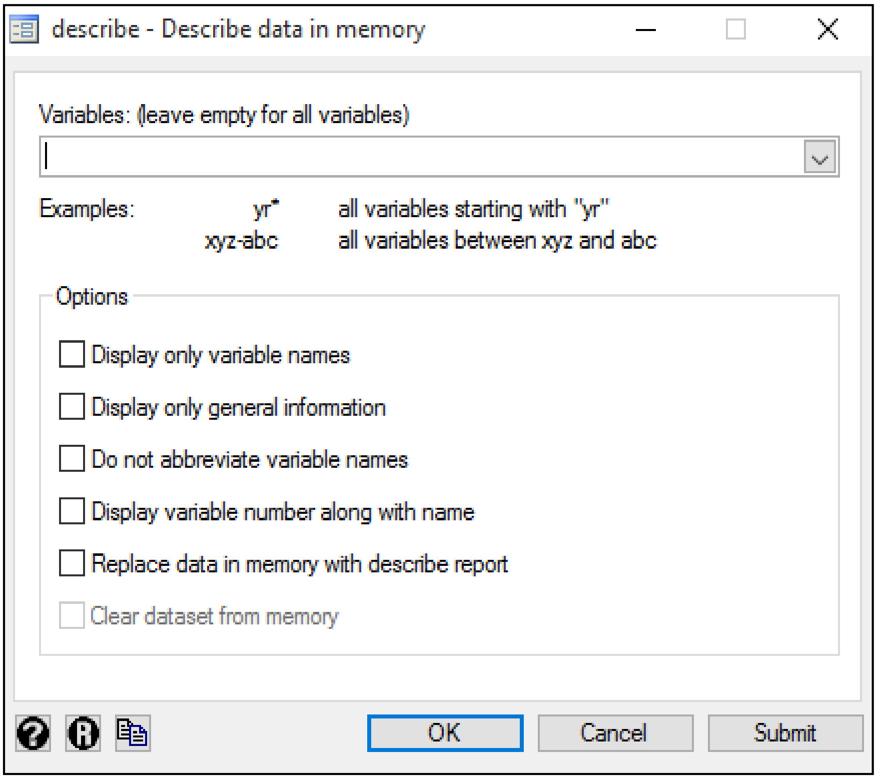
```
describe HHD_INCUR_EXP_FOR_DEPENDANTS
```

Table 3: Description of Variable “Household Expenditure on Dependents” using Describe Command

Likewise, more than one variable can be described by adding the names of variables after "**describe**" command.

The “**describe**” function in Stata can also be executed using dropdown menus as mentioned below:
Data > Describe Data in Memory

On navigating to "**describe function**" through menu system, the screen given in Fig. 4 will appear. Select desired variable(s) from the drop-down panel (see Fig. 4) or else leave the variable list empty and click on "OK". Table 2 will be displayed, if you choose to leave the variables box empty, or else Table 3 will appear, if you select a particular variable like “household expenditure on dependent”.



describe - Describe data in memory

Variables: (leave empty for all variables)

Examples:
yr* all variables starting with "yr"
xyz-abc all variables between xyz and abc

Options

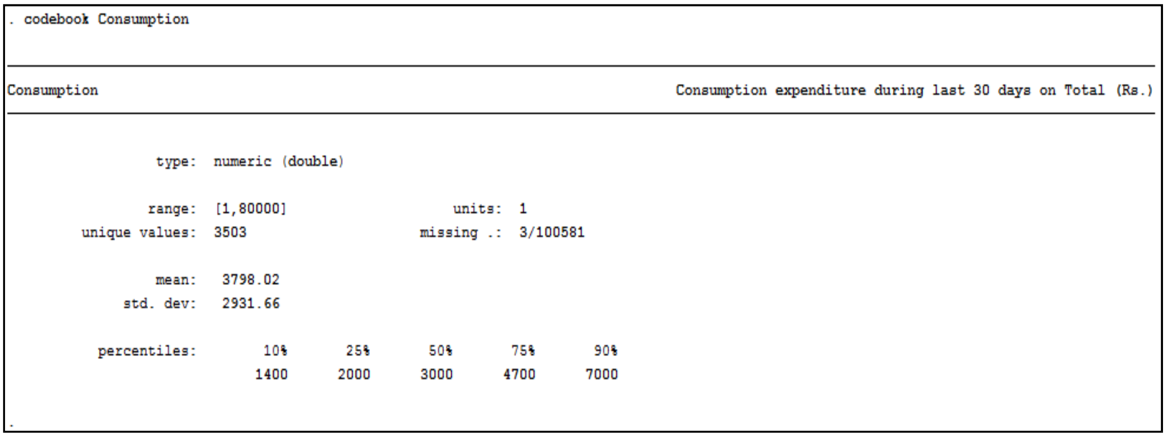
□ Display only variable names
□ Display only general information
□ Do not abbreviate variable names
□ Display variable number along with name
□ Replace data in memory with describe report
□ Clear dataset from memory

Fig. 4: "Describe" Function through Menu System

5.3 Codebook

The "**codebook**" command provides more detailed information about the variables than "**describe**" command. For example, if you want to get detailed information about a variable, namely, "Consumption expenditure" from a dataset, use the following command: **codebook consumption.**

The "**codebook**" command provides mean, standard deviation, percentiles, etc. of the "consumption expenditure" as shown in Table 4. Similarly, detailed information can also be obtained on all variables by just writing "codebook".

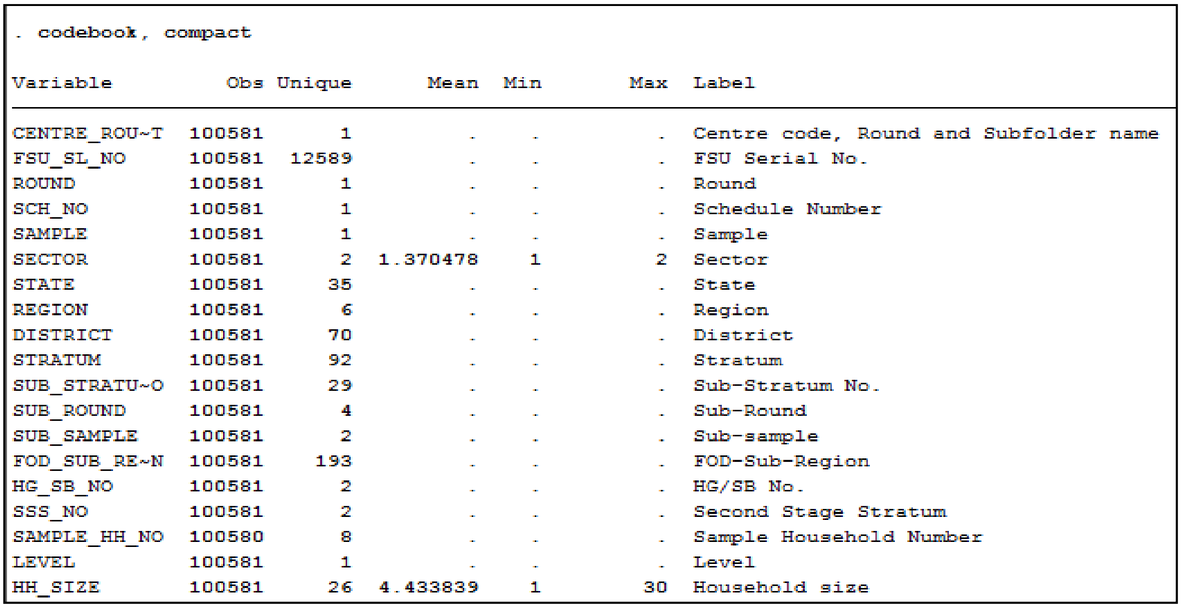


```
. codebook Consumption

------------------------------------------------------------------------------------------
Consumption                        Consumption expenditure during last 30 days on Total (Rs.)
------------------------------------------------------------------------------------------

                  type:  numeric (double)

                 range:  [1,80000]                    units:  1
         unique values:  3503                     missing .:  3/100581

                  mean:   3798.02
              std. dev:   2931.66

           percentiles:        10%       25%       50%       75%       90%
                              1400      2000      3000      4700      7000

.
```

Table 4: Details of Variable Consumption Expenditure using Codebook Command

To get compact information on variables, use the following command: **codebook, compact**

This command will produce the information provided in Table 5.

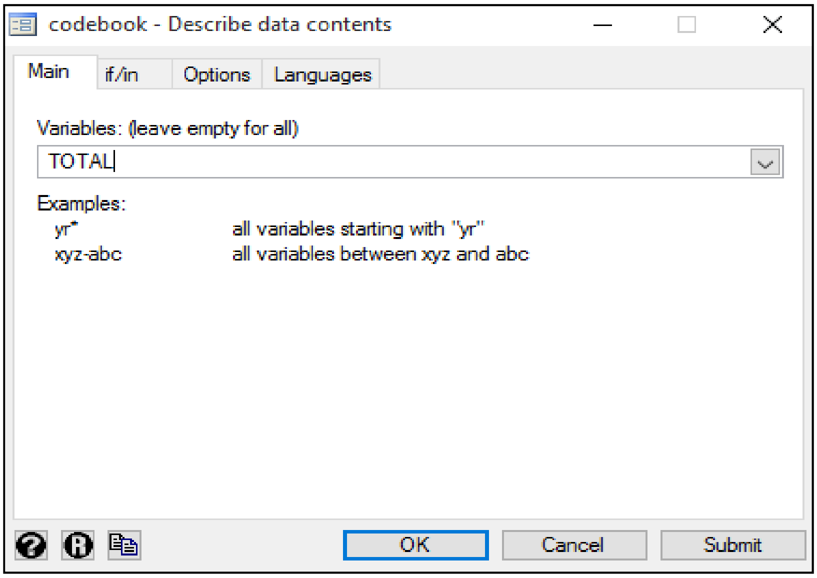


```
. codebook, compact

Variable          Obs Unique      Mean  Min       Max  Label
--------------------------------------------------------------------------------------
CENTRE_ROU~T   100581      1         .    .         .  Centre code, Round and Subfolder name
FSU_SL_NO      100581  12589         .    .         .  FSU Serial No.
ROUND          100581      1         .    .         .  Round
SCH_NO         100581      1         .    .         .  Schedule Number
SAMPLE         100581      1         .    .         .  Sample
SECTOR         100581      2  1.370478    1         2  Sector
STATE          100581     35         .    .         .  State
REGION         100581      6         .    .         .  Region
DISTRICT       100581     70         .    .         .  District
STRATUM        100581     92         .    .         .  Stratum
SUB_STRATU~O   100581     29         .    .         .  Sub-Stratum No.
SUB_ROUND      100581      4         .    .         .  Sub-Round
SUB_SAMPLE     100581      2         .    .         .  Sub-sample
FOD_SUB_RE~N   100581    193         .    .         .  FOD-Sub-Region
HG_SB_NO       100581      2         .    .         .  HG/SB No.
SSS_NO         100581      2         .    .         .  Second Stage Stratum
SAMPLE_HH_NO   100580      8         .    .         .  Sample Household Number
LEVEL          100581      1         .    .         .  Level
HH_SIZE        100581     26  4.433839    1        30  Household size
```

Table5: Compact Information on Variables using Codebook Command

The "**codebook**"function in Stata can also be executed using dropdown menus as mentioned below:
Data > Describe Data > Describe data contents (codebook)

On selection of “codebook” from the dropdown menu, screenshot of option box given in Fig. 5 will appear. From the option box, select desired variable. In this example, variable "total consumption expenditure" is selected. On clicking at "OK", Table 4 will be displayed.



Variables: (leave empty for all)
TOTAL

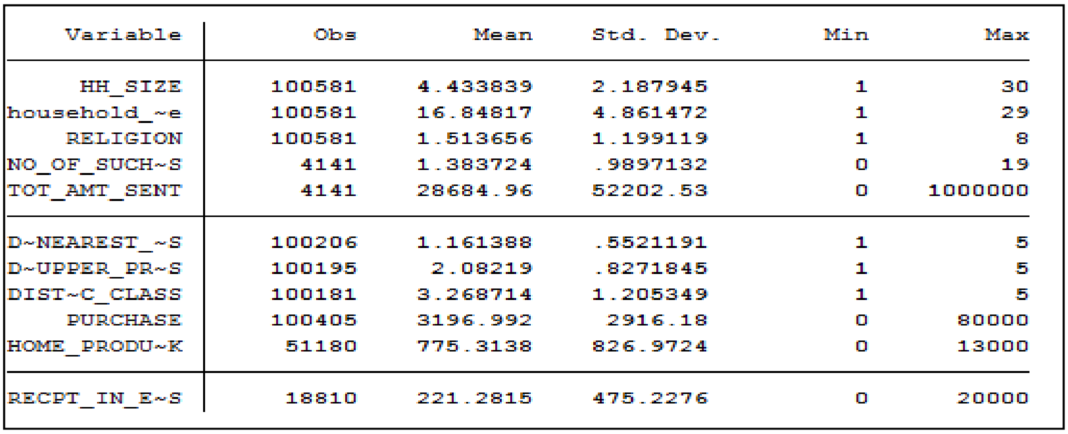
Examples:
yr* all variables starting with "yr"
xyz-abc all variables between xyz and abc

Fig. 5: Option Box for Selection of Variable (Codebook)

5.4 Summarize

The “**summarize**” command in Stata produces some basic characteristics of data, e.g. number of observations, mean, standard deviation, minimum and maximum, etc. Use the following command to generate a summary: **summarize**

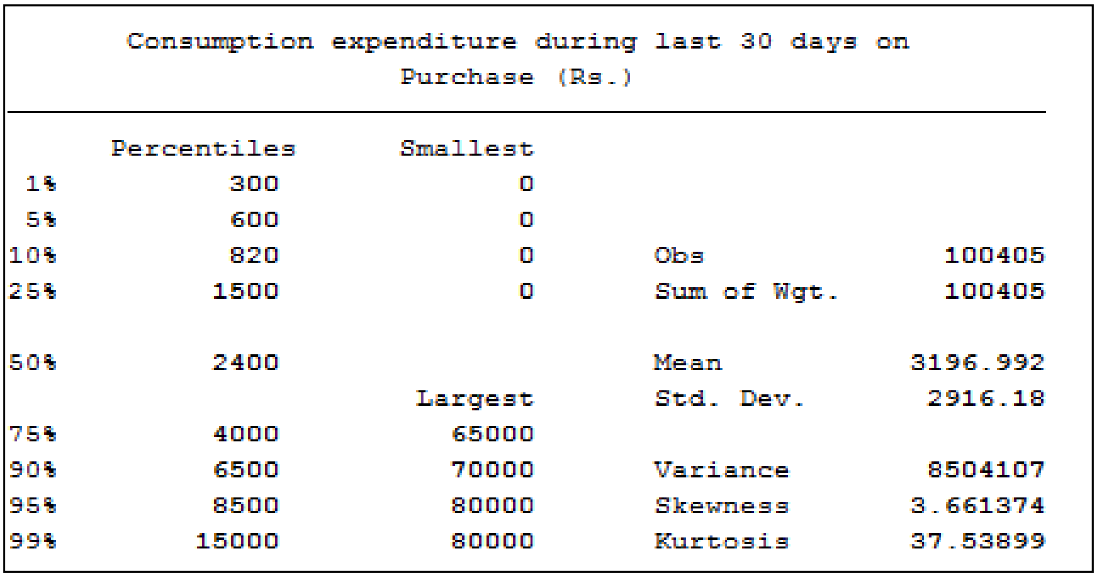
This command will produce the result as shown in Table 6.



```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     HH_SIZE |    100581    4.433839    2.187945          1         30
household_~e |    100581    16.84817    4.861472          1         29
    RELIGION |    100581    1.513656    1.199119          1          8
NO_OF_SUCH~S |      4141    1.383724    .9897132          0         19
TOT_AMT_SENT |      4141    28684.96    52202.53          0    1000000
-------------+--------------------------------------------------------
D~NEAREST_~S |    100206    1.161388    .5521191          1          5
D~UPPER_PR~S |    100195     2.08219    .8271845          1          5
DIST~C_CLASS |    100181    3.268714    1.205349          1          5
    PURCHASE |    100405    3196.992     2916.18          0      80000
HOME_PRODU~K |     51180    775.3138    826.9724          0      13000
-------------+--------------------------------------------------------
RECPT_IN_E~S |     18810    221.2815    475.2276          0      20000
```

Table 6: Summary of Basic Characteristics of Data using Summarize Command

You can also get the detailed statistics by typing detail after summarize i.e. "**summarize, detail"** command. This will produce Table 7 given below.

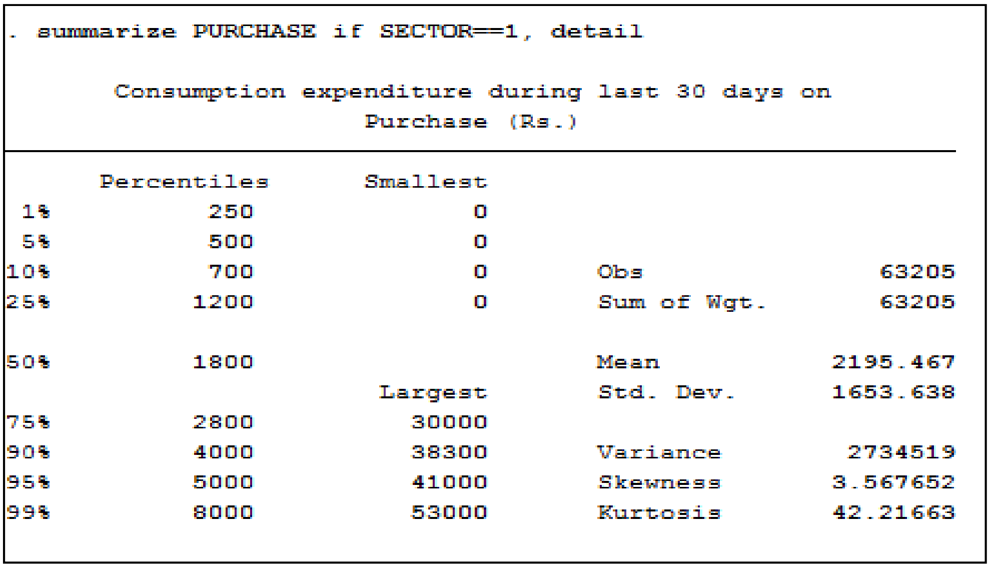


```
         Consumption expenditure during last 30 days on
                        Purchase (Rs.)
-------------------------------------------------------------
      Percentiles      Smallest
 1%          300              0
 5%          600              0
10%          820              0       Obs              100405
25%         1500              0       Sum of Wgt.      100405

50%         2400                      Mean           3196.992
                        Largest       Std. Dev.       2916.18
75%         4000          65000
90%         6500          70000       Variance        8504107
95%         8500          80000       Skewness       3.661374
99%        15000          80000       Kurtosis       37.53899
```

Table 7: Detailed Statistics Produced using "summarize, detail" Command

You can also obtain the detailed summary statistics for sub-group using **"IF"** function in summarize command. For example, if you are interested in obtaining the consumption expenditure for rural sector only, use the following command: **summarize PURCHASE if SECTOR==1, detail**

This command will produce Table 8, where the summary statistics of consumption expenditure in rural areas is shown. Sector code 1 here represents rural areas.

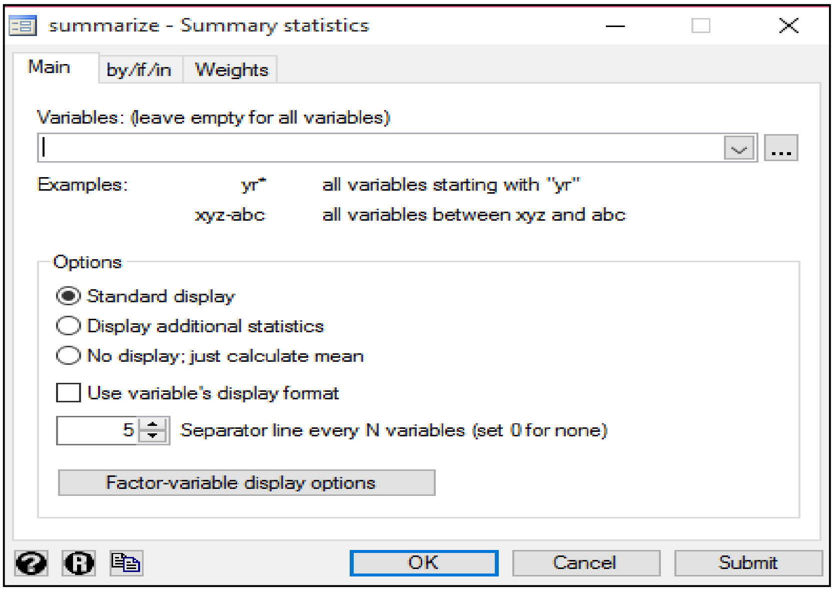


```
. summarize PURCHASE if SECTOR==1, detail

         Consumption expenditure during last 30 days on
                        Purchase (Rs.)
-------------------------------------------------------------
      Percentiles      Smallest
 1%          250              0
 5%          500              0
10%          700              0       Obs               63205
25%         1200              0       Sum of Wgt.       63205

50%         1800                      Mean           2195.467
                        Largest       Std. Dev.      1653.638
75%         2800          30000
90%         4000          38300       Variance        2734519
95%         5000          41000       Skewness       3.567652
99%         8000          53000       Kurtosis       42.21663
```

Table 8: Summary Statistics of Consumption Expenditure in Rural Areas Produced using Summarize Command for Sector 1 (Rural Area)

The "**summarize**” function in Stata can also be executed using dropdown menus as mentioned below: **Data > Describe data > Summary Statistics**

On selection of “**Summary Statistics**” from the dropdown menu, screenshot of option box given in Fig. 6 will appear. Select desired variables from the drop-down variable list or else leave it blank, if all variables are to be summarised. On clicking at "OK", Table 6 will be displayed as result.



Variables: (leave empty for all variables)

Examples: yr* all variables starting with "yr"
xyz-abc all variables between xyz and abc

Options

● Standard display
○ Display additional statistics
○ No display; just calculate mean
□ Use variable's display format
5 Separator line every N variables (set 0 for none)

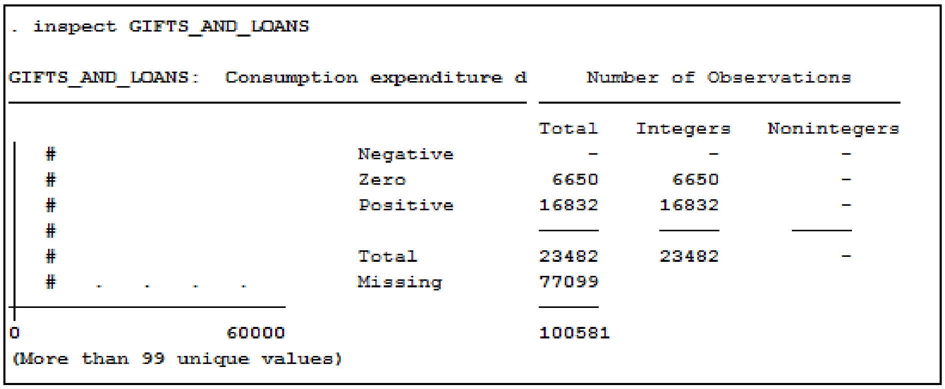
Factor-variable display options

Fig. 6: Option Box for Selection of Variable (Summarize)

5.5 Inspect Variables

The "**inspect**" command in Stata provides detailed information on the numeric variables. It provides negative, positive, zero, missing, unique values, integer and non-integer values and a histogram of the variables. For inspecting the variable "**GIFTS_AND_LOANS**", for example, use the following command: **inspect GIFTS_AND_LOANS**

This command will produce Table 9, where the essential statistics of "consumption expenditure" is shown.

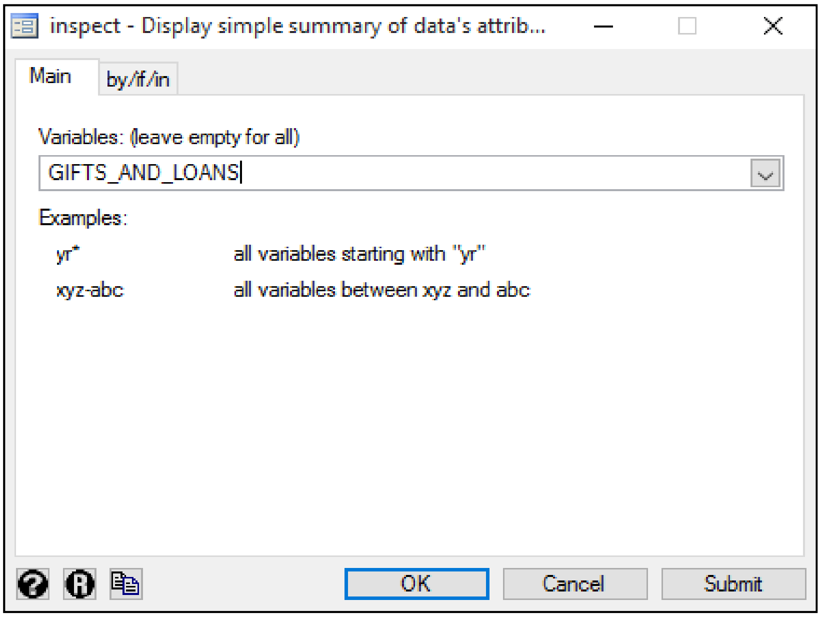


```
. inspect GIFTS_AND_LOANS

GIFTS_AND_LOANS:  Consumption expenditure d    Number of Observations
-------------------------------------------   ------------------------------
                                               Total   Integers   Nonintegers
|  #                        Negative               -          -             -
|  #                        Zero                6650       6650             -
|  #                        Positive           16832      16832             -
|  #                                           -----      -----        ------
|  #                        Total              23482      23482             -
|  #   .   .   .   .        Missing            77099
+-----------------------                       -----
0                  60000                      100581
 (More than 99 unique values)
```

Table 9: Essential Statistics of Consumption Expenditure using Inspect Command

The “Inspect variables” function in Stata can also be executed using dropdown menus as mentioned: **Data > Describe data > Inspect Variables**

On selection of “Inspect Variables” from the dropdown menu, screenshot of option box given in Figure 7 will appear. Select variable "**GIFTS_AND_LOANS**” from the option box. On clicking at "submit" and "OK" button, Table 9 will be displayed. The similar inspect function can also be used through dropdown menus, shown in Figure 7.


inspect - Display simple summary of data's attrib...

Variables: (leave empty for all)

GIFTS_AND_LOANS

Examples:

yr* all variables starting with "yr"

xyz-abc all variables between xyz and abc

Fig.7: Option Box for Selection of Variable (Inspect Variable)

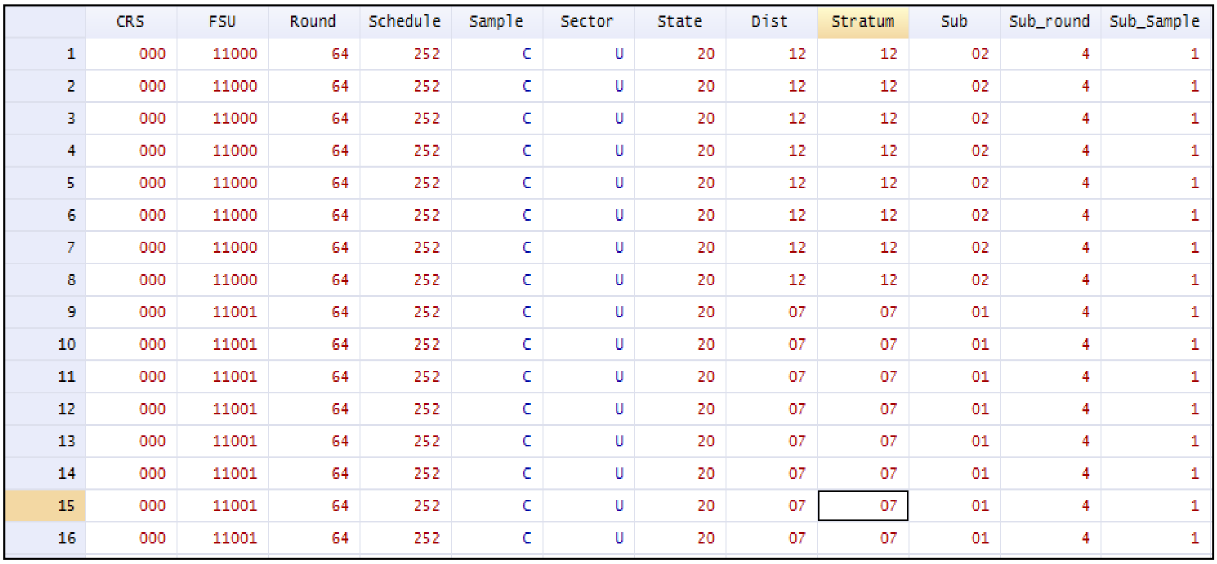
6. Importing Txt. File (Fixed Width Data)

Very often, important datasets carry textual information on each household, individual, or firm. In Stata, one can import the “txt. data” using the following command:
infix *specifications* using <filename>

In this example, for importing the text file provided by the NSSO, i.e. “Schedule 25.2: Participation and expenditure in Education, 64th Round" (Block 1&2), the following command is used:

```
infix CRS 1-3 FSU 4-8 Round 9-10 Schedule 11-13 Sample 14 Sector 15 State 16-18 Dist 19-20 Stratum 21-22 Sub 23-24 Sub_round 25 Sub_Sample 26 FOD 27-30 HG 31 Second_Stage_Str 32 Sample_HH_No 33-34 level 35-36 filler 37-41 Informant_sl_no 42-43 response_code 44 survey_code 45 subst_code 46 using "H:\NSS 64th Round-Participation and Exp in Education\Nss64_25.2\Data\AH1C25.TXT"
```

The above specifications like: CRS 1-3, FSU 4-8, Round 9-10, and Schedule 11-13, etc. are obtained from the layout file provided by the NSSO. As a result of the above command, data will be imported in Stata as shown in the Fig. 8.

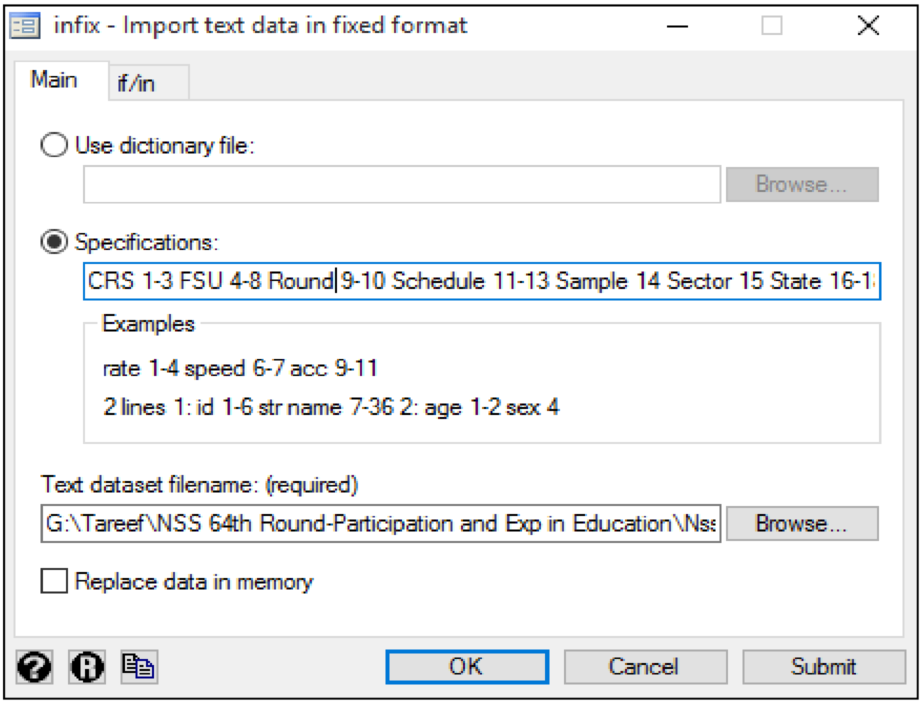


	CRS	FSU	Round	Schedule	Sample	Sector	State	Dist	Stratum	Sub	Sub_round	Sub_Sample
1	000	11000	64	252	C	U	20	12	12	02	4	1
2	000	11000	64	252	C	U	20	12	12	02	4	1
3	000	11000	64	252	C	U	20	12	12	02	4	1
4	000	11000	64	252	C	U	20	12	12	02	4	1
5	000	11000	64	252	C	U	20	12	12	02	4	1
6	000	11000	64	252	C	U	20	12	12	02	4	1
7	000	11000	64	252	C	U	20	12	12	02	4	1
8	000	11000	64	252	C	U	20	12	12	02	4	1
9	000	11001	64	252	C	U	20	07	07	01	4	1
10	000	11001	64	252	C	U	20	07	07	01	4	1
11	000	11001	64	252	C	U	20	07	07	01	4	1
12	000	11001	64	252	C	U	20	07	07	01	4	1
13	000	11001	64	252	C	U	20	07	07	01	4	1
14	000	11001	64	252	C	U	20	07	07	01	4	1
15	000	11001	64	252	C	U	20	07	07	01	4	1
16	000	11001	64	252	C	U	20	07	07	01	4	1

Table 10: Display of Result for Imported Data from txt.file

The importing text data can also be performed in Stata using dropdown menus as mentioned below:
File > Import > Text data in fixed format

On selection of "**Text data in fixed format**" from the dropdown menu, screenshot of option box given in Fig. 9 will appear. Select "**Specifications**" from the Option Box (Fig. 9) and provide the specifications i.e. CRS 1-3, FSU 4-8, etc. given in layout file of NSSO. After providing the specifications details, select a text file to be imported and click on "Submit" and "OK" button. Table 10 will be displayed.



infix - Import text data in fixed format

○ Use dictionary file:

● Specifications:

CRS 1-3 FSU 4-8 Round 9-10 Schedule 11-13 Sample 14 Sector 15 State 16-1

Examples

rate 1-4 speed 6-7 acc 9-11

2 lines 1: id 1-6 str name 7-36 2: age 1-2 sex 4

Text dataset filename: (required)

G:\Tareef\NSS 64th Round-Participation and Exp in Education\Nss

□ Replace data in memory

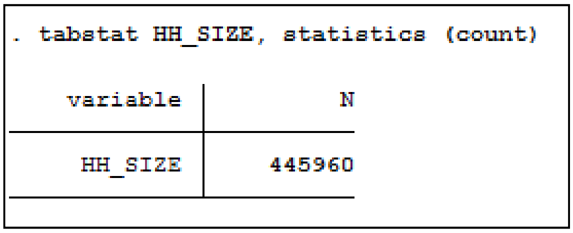
Fig. 9: Option Box for Selection of Specifications and Uploading of Data file (Import Text Data)

7. Tabulation

Stata can produce a wide range of tables with required statistics. For example, in this round of NSSO: 25.2, a large number of persons were interviewed during the period of survey at all India level.

Use the following command to get the data on number of interviews conducted: **tabstat <varname>, statistics (count)**

In this example, in place of "varname", "**HH_SIZE"**is used followed by "count" to count number of persons interviewed. Table 11 is produced as a result of this command.

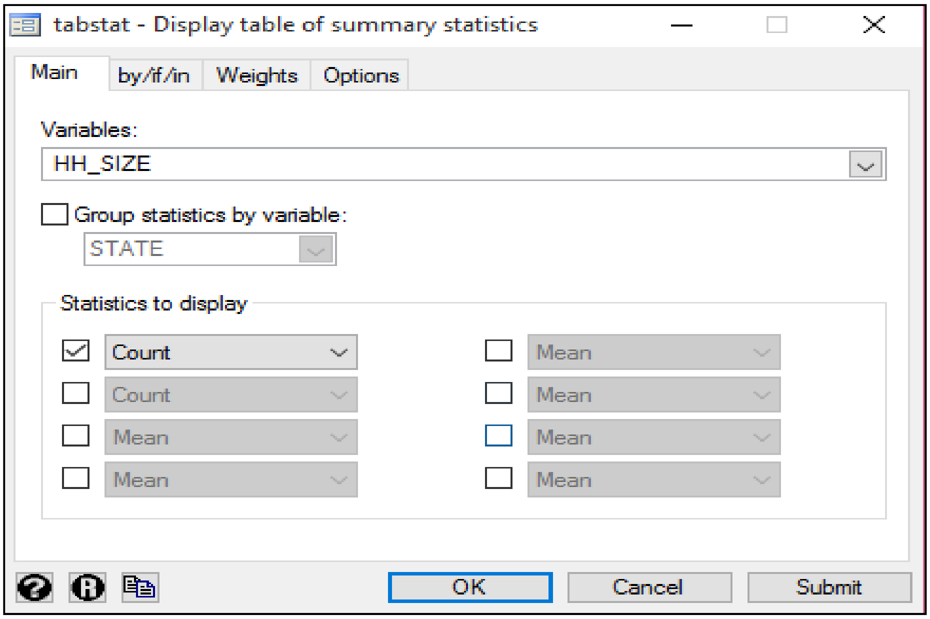


```
. tabstat HH_SIZE, statistics (count)

    variable |         N
-------------+----------
     HH_SIZE |    445960
------------------------
```

Table 11: Statistics on No. of Persons using Tabstat Command

The "**Tabstat**" command in Stata can also be executed using dropdown menus as mentioned below: **Statistics > Summaries, Tables and Tests > Tables > Table of Summary Statistics (tabstat)**

On selection of "Table of Summary Statistics" from the dropdown menu, screenshot of option box given in Fig. 10 will appear. Select variable from the "variable" drop down pan given in option box.In this example, we have selected variable "HH_SIZE" from the variable pan to get its summary statistics. On clicking at "submit" and "OK" button, Table 10 will be displayed.



tabstat - Display table of summary statistics

Variables: HH_SIZE

□ Group statistics by variable: STATE

Statistics to display

☑ Count　　□ Mean

□ Count　　□ Mean

□ Mean　　□ Mean

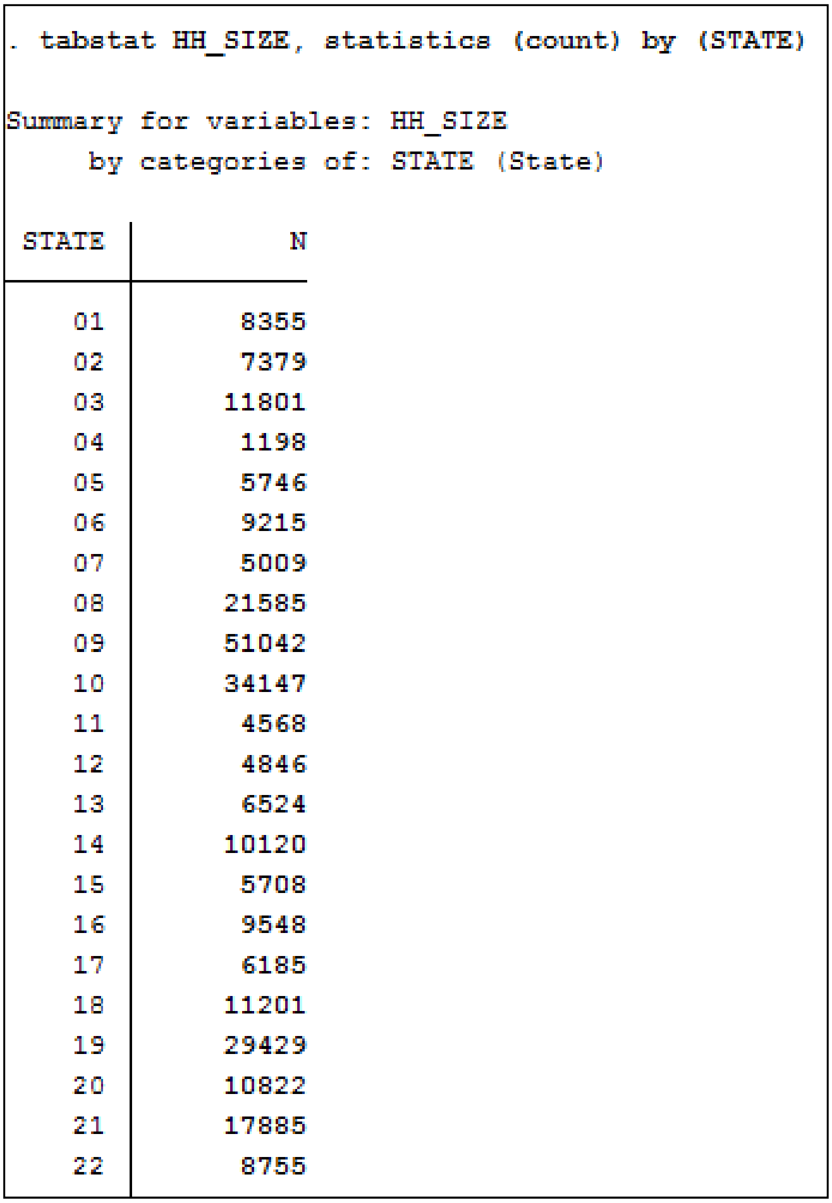
□ Mean　　□ Mean

Fig. 10: Option Box for Selection of Variable (Table of Summary Statistics)

7.1 Tabulation with One Categorical Variable

In the example given above, number of persons were counted for one variable at India level. Likewise, if number of persons are to be counted at state level, use the command given below:
tabstat <varname>, statistics (count) by (STATE)

This command will produce Table 12 where state-wise numbers of persons are depicted. In the "STATE" column, state codes are revealed and in column "N", numbers of persons are displayed. This function in Stata can also be executed using dropdown menus as shown in previous example. However, in addition, check the "**Group Statistics by Variables"** and select "**STATE"** in the option box. Table 12 will be produced as a result.



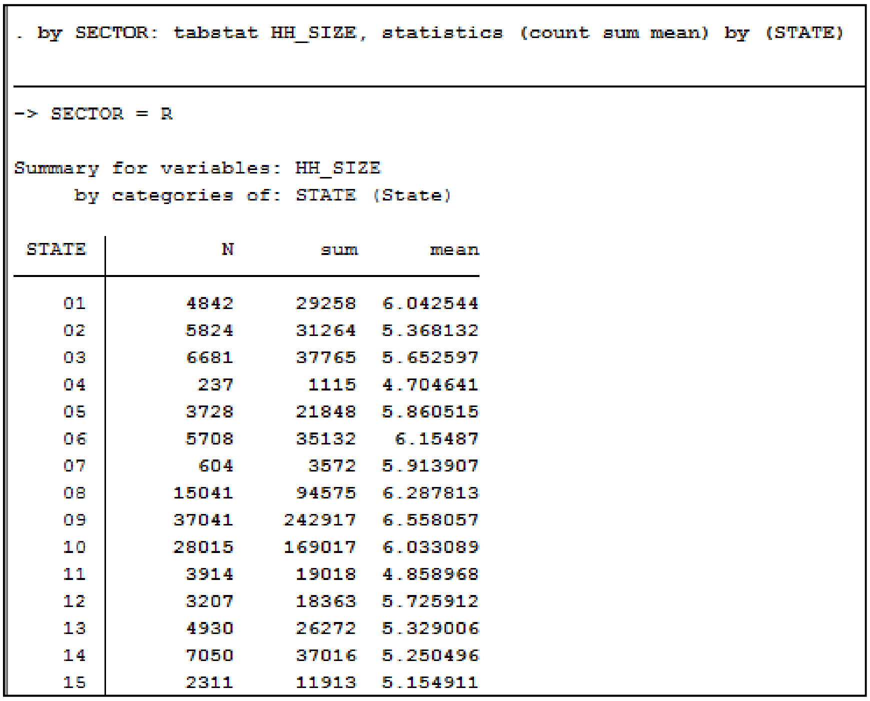
```
. tabstat HH_SIZE, statistics (count) by (STATE)

Summary for variables: HH_SIZE
     by categories of: STATE (State)

 STATE |         N
-------+----------
    01 |      8355
    02 |      7379
    03 |     11801
    04 |      1198
    05 |      5746
    06 |      9215
    07 |      5009
    08 |     21585
    09 |     51042
    10 |     34147
    11 |      4568
    12 |      4846
    13 |      6524
    14 |     10120
    15 |      5708
    16 |      9548
    17 |      6185
    18 |     11201
    19 |     29429
    20 |     10822
    21 |     17885
    22 |      8755
```

Table 12: Summary Statistics: State wise No. of Persons

7.2 Tabulation with Two Categorical Variables

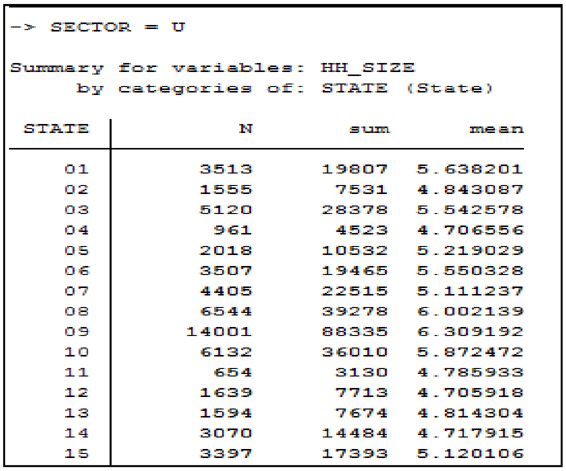
Further, using "**tabstat**" function, you can also calculate a variable with two categorical variables. In this example, numbers of persons from different Indian States in rural and urban areas are calculated using the following command: **by SECTOR: tabstat <HH_SIZE>, statistics (count sum mean) by (STATE)**

As shown in Table 12.1 and Table 12.2, the above mentioned command will produce numbers of persons in different Indian states for two different sectors viz. rural and urban. While Table 12.1 represents the rural sector, Table 12.2 represents the urban sector. In these two tables, one can also calculate "sum" and "mean" of number of persons in addition to the total counts.



```
. by SECTOR: tabstat HH_SIZE, statistics (count sum mean) by (STATE)

-> SECTOR = R

Summary for variables: HH_SIZE
     by categories of: STATE (State)

 STATE |         N       sum      mean
-------+------------------------------
    01 |      4842     29258  6.042544
    02 |      5824     31264  5.368132
    03 |      6681     37765  5.652597
    04 |       237      1115  4.704641
    05 |      3728     21848  5.860515
    06 |      5708     35132   6.15487
    07 |       604      3572  5.913907
    08 |     15041     94575  6.287813
    09 |     37041    242917  6.558057
    10 |     28015    169017  6.033089
    11 |      3914     19018  4.858968
    12 |      3207     18363  5.725912
    13 |      4930     26272  5.329006
    14 |      7050     37016  5.250496
    15 |      2311     11913  5.154911
```

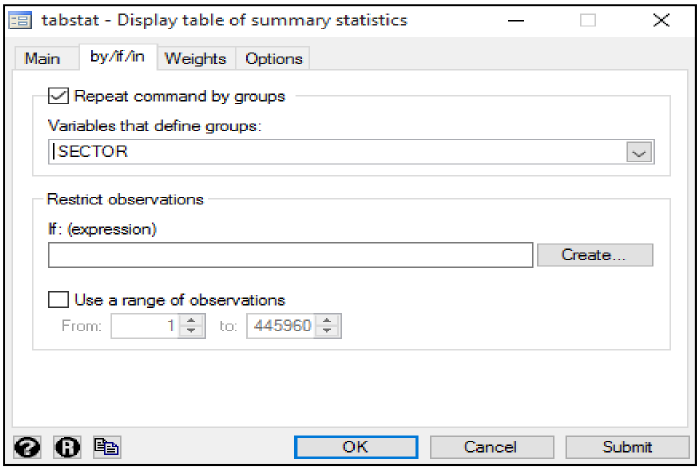
Table 12.1: No. of Persons in Different Indian States in Rural Sector (Tabstat Command-Two Variables)



```
-> SECTOR = U

Summary for variables: HH_SIZE
     by categories of: STATE (State)

 STATE |         N       sum      mean
-------+------------------------------
    01 |      3513     19807  5.638201
    02 |      1555      7531  4.843087
    03 |      5120     28378  5.542578
    04 |       961      4523  4.706556
    05 |      2018     10532  5.219029
    06 |      3507     19465  5.550328
    07 |      4405     22515  5.111237
    08 |      6544     39278  6.002139
    09 |     14001     88335  6.309192
    10 |      6132     36010  5.872472
    11 |       654      3130  4.785933
    12 |      1639      7713  4.705918
    13 |      1594      7674  4.814304
    14 |      3070     14484  4.717915
    15 |      3397     17393  5.120106
```

Table 12.2: No. of Persons in Different Indian States in Urban Sector (Tabstat Command-Two Variables)

The "**tabstat**” function in Stata can also be executed using dropdown menus. Click at “**by/if/in”** in the Menu bar as shown in Fig. 10 below. On clicking ay “by/if/an”, screenshot of option box given in Fig. 12 will appear. Check **“Repeat command by groups ”**and selected **SECTOR.** These selections will also produce Table: 12.1 and Table: 12.2.


tabstat - Display table of summary statistics

☑ Repeat command by groups

Variables that define groups: SECTOR

Restrict observations

If: (expression)

□ Use a range of observations

From: 1 to: 445960

Fig.12: Option Box for Selection of Variable (Tabstat)

7.3 Comparing Two Variables (Two Way Table)

Stata performs cross-tabulation function for two or more categorical variables. In this example, land possessions by the male and female is compared. Use the following command:
table LAND_POSSESSED_CODE SEX, row col

Here, "land possession" is taken as first variable and "sex" as second variable, whereas row and column represent the total numbers of male and female on the basis of their land possessed. On use of this command, Table 13 is produced wherein "land possessed" is given in acres.

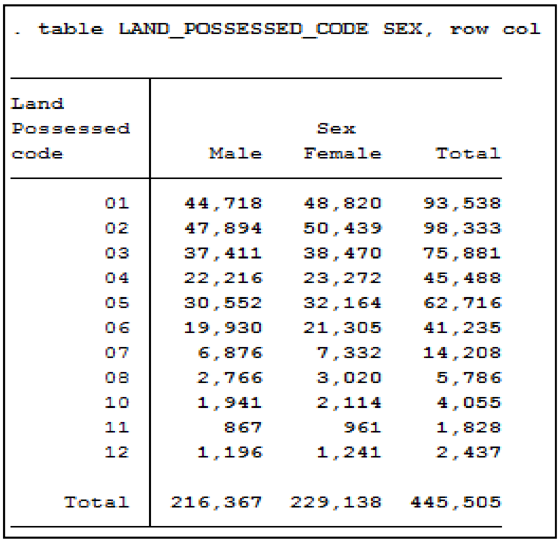
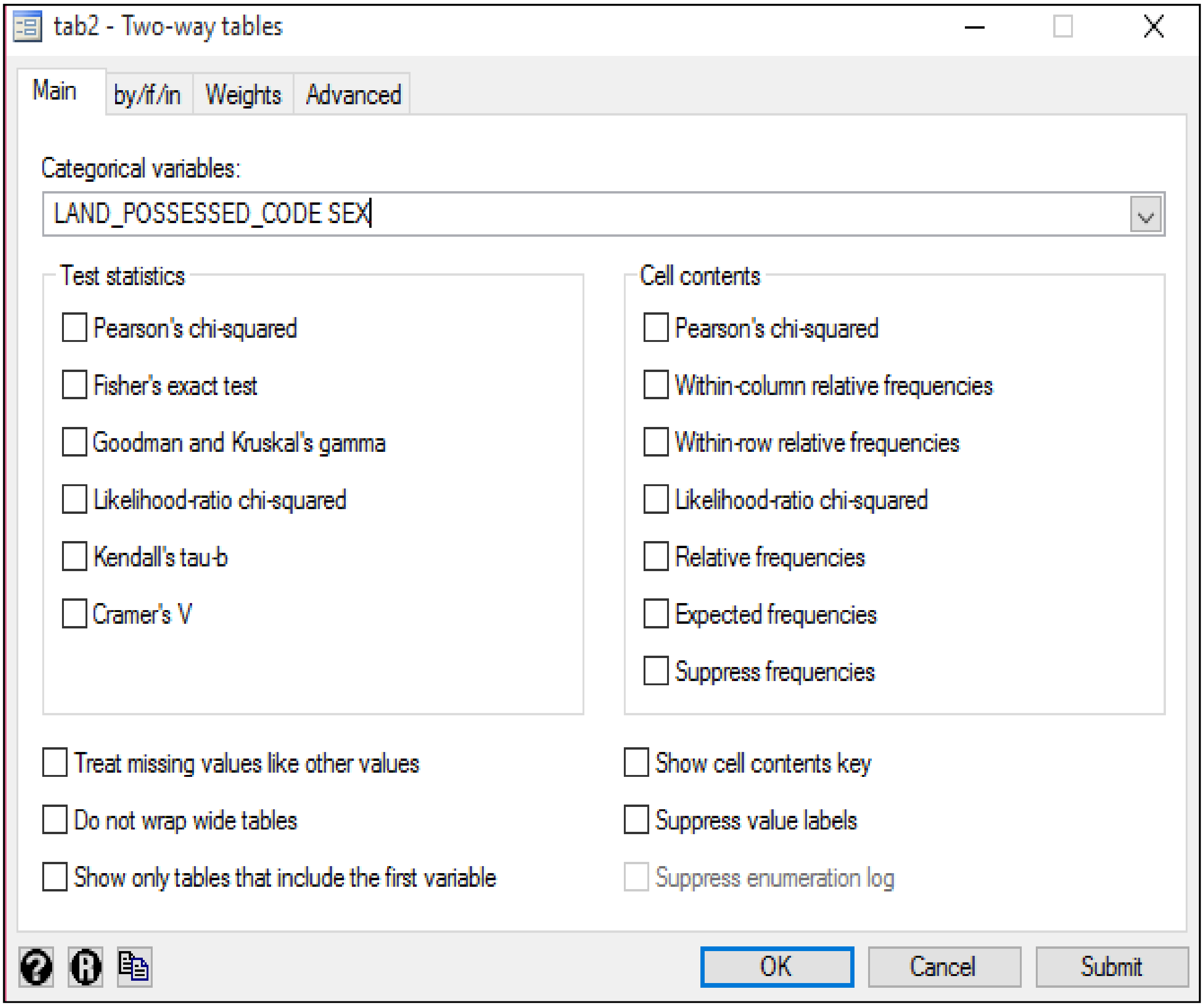

```
. table LAND_POSSESSED_CODE SEX, row col

----------------------------------------------
Land       |
Possessed  |             Sex
code       |     Male    Female     Total
-----------+----------------------------------
        01 |   44,718    48,820    93,538
        02 |   47,894    50,439    98,333
        03 |   37,411    38,470    75,881
        04 |   22,216    23,272    45,488
        05 |   30,552    32,164    62,716
        06 |   19,930    21,305    41,235
        07 |    6,876     7,332    14,208
        08 |    2,766     3,020     5,786
        10 |    1,941     2,114     4,055
        11 |      867       961     1,828
        12 |    1,196     1,241     2,437
           |
     Total |  216,367   229,138   445,505
----------------------------------------------
```

Table 13: Cross-Table Comparison for Two Variables

Cross-table comparison can also be performed in Stata using dropdown menus as mentioned below:
Statistics > Summaries, Tables and Tests > Tables >All Possible two-way Tabulations

On selection of options mentioned above from the dropdown menu, screenshot of option box given in Fig. 13 will appear. Select desired variables (in this case, "Land possessed" and "sex") from the drop down categorical pan, click at "submit" button. These selections will also produce Table 13.


tab2 - Two-way tables

Main | by/if/in | Weights | Advanced

Categorical variables:
LAND_POSSESSED_CODE SEX

Test statistics
- □ Pearson's chi-squared
- □ Fisher's exact test
- □ Goodman and Kruskal's gamma
- □ Likelihood-ratio chi-squared
- □ Kendall's tau-b
- □ Cramer's V

Cell contents
- □ Pearson's chi-squared
- □ Within-column relative frequencies
- □ Within-row relative frequencies
- □ Likelihood-ratio chi-squared
- □ Relative frequencies
- □ Expected frequencies
- □ Suppress frequencies

- □ Treat missing values like other values
- □ Do not wrap wide tables
- □ Show only tables that include the first variable
- □ Show cell contents key
- □ Suppress value labels
- □ Suppress enumeration log

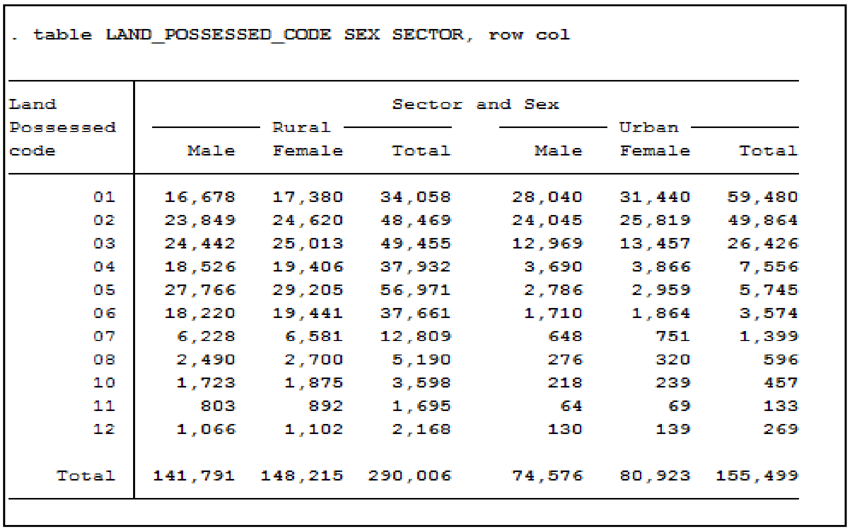
Fig. 13: Option Box for Selection of Categorical Variable (Two-way Tabulations)

7.4 Three Way Table

The three way table depicts relation between two variables with one categorical variable. For instance, in the previous example, the relationship between land possession and sex was examined. Let us now include one more categorical variable i.e. sector using the command given below:

```
table LAND_POSSESSED_CODE SEX SECTOR, row col
```

The above command will produce Table 14 given below.



```
. table LAND_POSSESSED_CODE SEX SECTOR, row col

----------------------------------------------------------------------------
Land       |                          Sector and Sex
Possessed  | ---------- Rural -----------    ---------- Urban -----------
code       |     Male   Female     Total        Male   Female     Total
-----------+----------------------------------------------------------------
        01 |   16,678   17,380    34,058      28,040   31,440    59,480
        02 |   23,849   24,620    48,469      24,045   25,819    49,864
        03 |   24,442   25,013    49,455      12,969   13,457    26,426
        04 |   18,526   19,406    37,932       3,690    3,866     7,556
        05 |   27,766   29,205    56,971       2,786    2,959     5,745
        06 |   18,220   19,441    37,661       1,710    1,864     3,574
        07 |    6,228    6,581    12,809         648      751     1,399
        08 |    2,490    2,700     5,190         276      320       596
        10 |    1,723    1,875     3,598         218      239       457
        11 |      803      892     1,695          64       69       133
        12 |    1,066    1,102     2,168         130      139       269
           |
     Total |  141,791  148,215   290,006      74,576   80,923   155,499
----------------------------------------------------------------------------
```

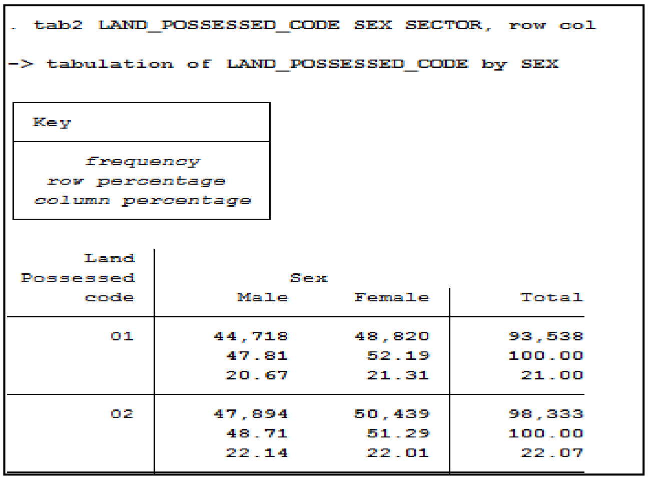
Table 14: Three-way Table Comparison

Besides, two way and three way tables, Stata also provides for four ways, five ways (and so forth) comparison of tables.

7.5 Tab2 Command

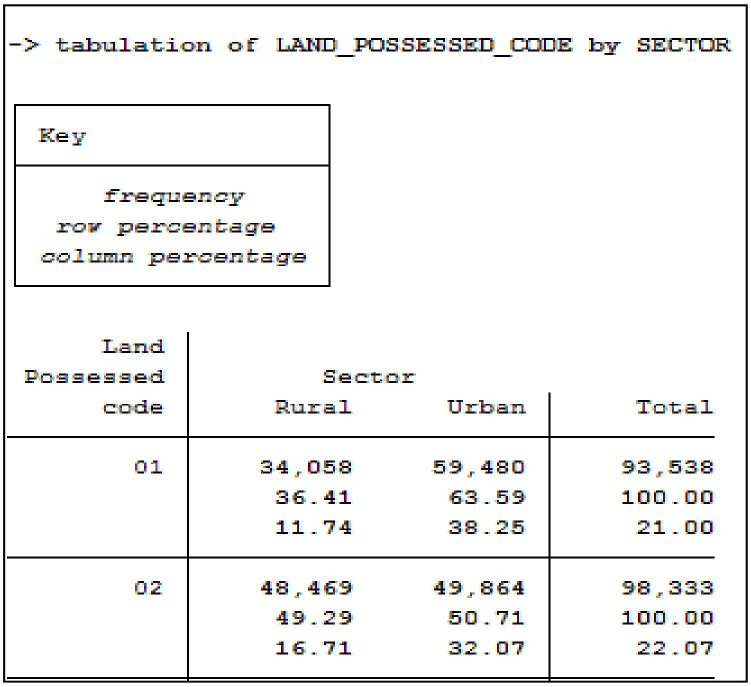
The "**tab2**" command basically provides all possible cross tabulation among variables. Carrying forward from the previous examples, here we have taken land possession, sex and sector as the variables for cross tabulation. Use the command given below:
tab2 LAND_POSSESSED_CODE SEX SECTOR, row col

Use of above mentioned command will produce three tables given below, namely Table 15.1, Table 15.2 and Table 15.3 respectively.



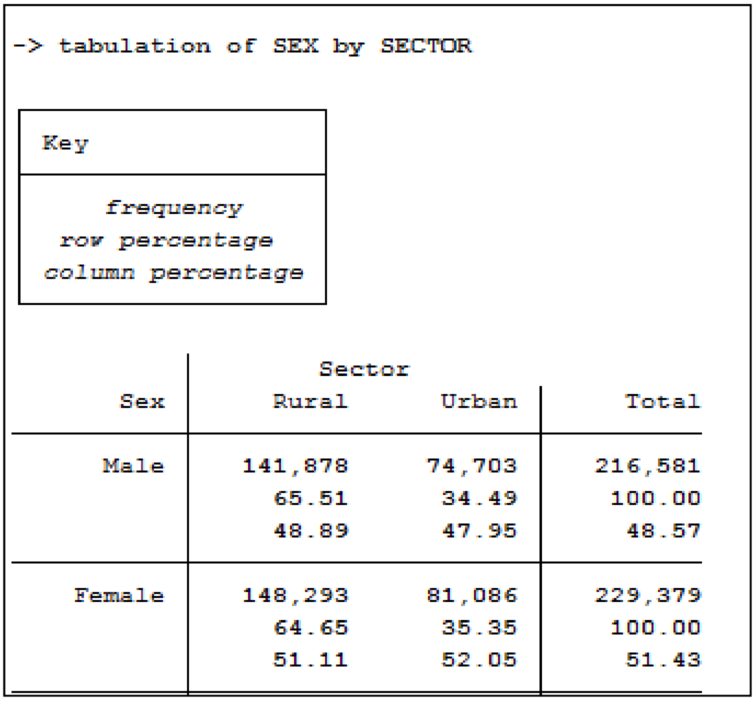
```
. tab2 LAND_POSSESSED_CODE SEX SECTOR, row col

-> tabulation of LAND_POSSESSED_CODE by SEX

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

      Land |
 Possessed |          Sex
      code |      Male     Female |     Total
-----------+----------------------+----------
        01 |    44,718     48,820 |    93,538
           |     47.81      52.19 |    100.00
           |     20.67      21.31 |     21.00
-----------+----------------------+----------
        02 |    47,894     50,439 |    98,333
           |     48.71      51.29 |    100.00
           |     22.14      22.01 |     22.07
```

Table 15.1. Land Possession by Males and Females in Rural and Urban Areas (Tab2 Command)



```
-> tabulation of LAND_POSSESSED_CODE by SECTOR

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

      Land |
 Possessed |        Sector
      code |     Rural      Urban |     Total
-----------+----------------------+----------
        01 |    34,058     59,480 |    93,538 
           |     36.41      63.59 |    100.00 
           |     11.74      38.25 |     21.00 
-----------+----------------------+----------
        02 |    48,469     49,864 |    98,333 
           |     49.29      50.71 |    100.00 
           |     16.71      32.07 |     22.07 
-----------+----------------------+----------
```

Table15.2: Land Possession in Rural and Urban Areas by Males and Females (Tab2 Command)



```
-> tabulation of SEX by SECTOR

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

           |        Sector
       Sex |     Rural      Urban |     Total
-----------+----------------------+----------
      Male |   141,878     74,703 |   216,581 
           |     65.51      34.49 |    100.00 
           |     48.89      47.95 |     48.57 
-----------+----------------------+----------
    Female |   148,293     81,086 |   229,379 
           |     64.65      35.35 |    100.00 
           |     51.11      52.05 |     51.43 
-----------+----------------------+----------
```

Table 15.3: Land Possession in Rural and Urban Areas by Males and Females (Tab2 Command)

The above three tables (Tables 15.1, 15.2 and 15.3) provide all possible cross tabulation among the variables, whereas rows and columns in the command represents the row percentage and column percentage.

8. Weight Data

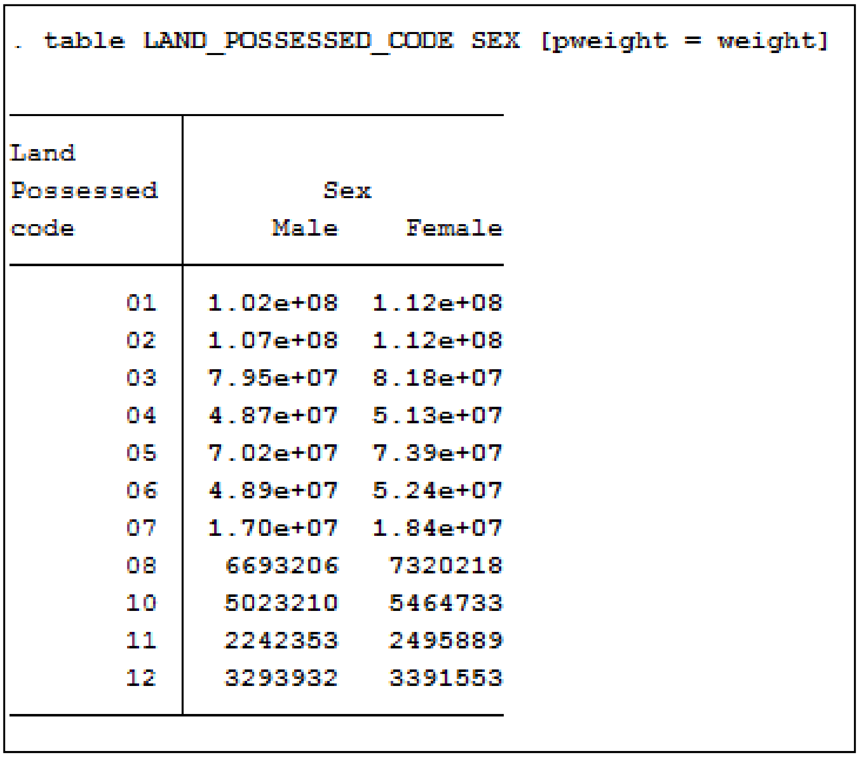
There are four sorts of weight Stata can assign, which are given below:

- Fweights (frequency weights)
- Pweights (sampling weights)
- Aweights (analytic weights)
- Iweights (importance weights)

Any of these weights can be used depending upon requirements.

To use "Sampling weight" while doing the cross tabulation between land possession and sex, use the following command: **table LAND_POSSESSED_CODE SEX [pweight = weight]**

Use of this command will produce Table 16 given below.



```
. table LAND_POSSESSED_CODE SEX [pweight = weight]


----------------------------------
Land       |
Possessed  |          Sex
code       |      Male     Female
-----------+----------------------
        01 |  1.02e+08   1.12e+08
        02 |  1.07e+08   1.12e+08
        03 |  7.95e+07   8.18e+07
        04 |  4.87e+07   5.13e+07
        05 |  7.02e+07   7.39e+07
        06 |  4.89e+07   5.24e+07
        07 |  1.70e+07   1.84e+07
        08 |   6693206    7320218
        10 |   5023210    5464733
        11 |   2242353    2495889
        12 |   3293932    3391553
----------------------------------
```

Table 16: Land Possession by Males and Females (Sampling Weighted Scores)

It may be noted that the values shown in Table 16 are much larger than the values of non-weighted cross tabulation. Because, usually weight is generated to gross up the sample values up to population values.

Weighted data can also be derived using dropdown menus as mentioned below:
Statistics > Summaries, Tables and Tests > Tables > Table of Summary Statistics (table)(wrong)

On selection of above mentioned menu item from the dropdown menu, screenshot of option box given in Fig. 14 will appear.

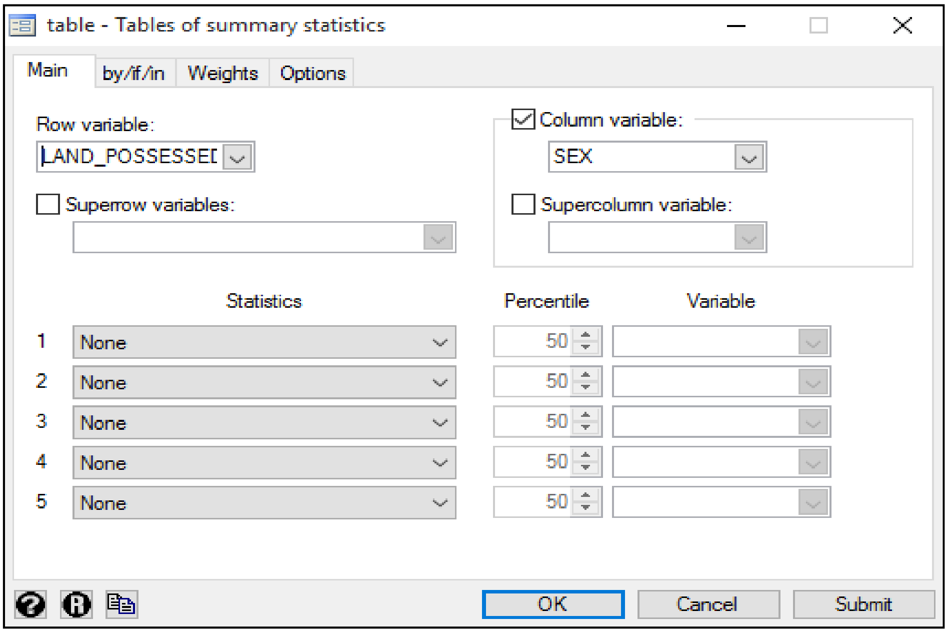


table - Tables of summary statistics

Row variable: LAND_POSSESSEI

☑ Column variable: SEX

□ Superrow variables:

□ Supercolumn variable:

	Statistics	Percentile	Variable
1	None	50	
2	None	50	
3	None	50	
4	None	50	
5	None	50	

Fig. 14: Option Box for Selection of Variable and Weights

Select variable “Land Possession” and click at “Weight” from the Option Box given in Fig. 14. Fig. 14.1 will appear with option to select required weight. Select "sampling weights", and click at "submit" button. Table 16 will be produced.

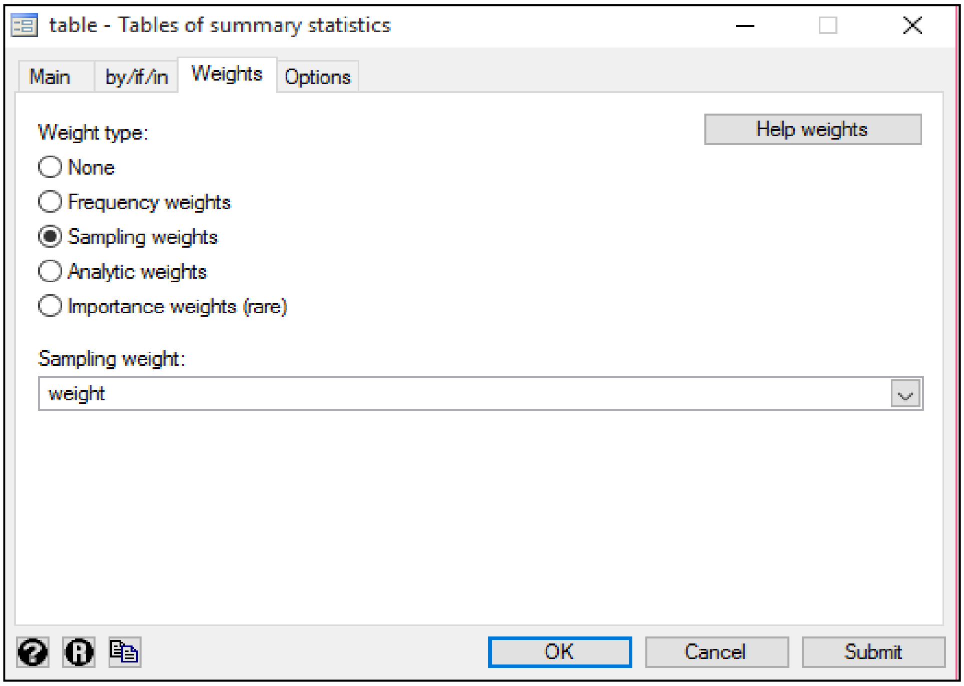


table - Tables of summary statistics

Weight type:

○ None

○ Frequency weights

● Sampling weights

○ Analytic weights

○ Importance weights (rare)

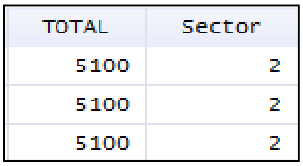
Sampling weight: weight

Fig. 14.1: Option Box for Selection of Weighted Score

9. Generation of New variables

In Stata, new variables can be generated from the existing variables using "**generate command**". For example, to create a new variable from the "SECTOR" variable, use the command given below:
gen Sector=SECTOR

The command given above will generate new variable in the last column with the name of Sector as shown in Table 17.



TOTAL	Sector
5100	2
5100	2
5100	2

Table 17: Generation of New Variable using Generate Command

After generating new variable, you can assign labels and values to the newly generated variable. Use following command to label new variable: **label variable Sector “Sector”**

After assigning label to the newly generated variable "Sector", provide the values to that variable, viz. 1 for rural and 2 urban. Use the following command for assigning value:

label define Sector 1 “rural” 2 “urban”

Using above mentioned command, value 1 is assigned to rural sector and value 2 is assigned to the urban sector. New variable based on mathematical functions, e.g. square, multiplication, addition and subtraction, etc. can also be generated. For example, square of household size can be obtained using the following command: **gen HH_SIZEsquared= HH_SIZE^2**

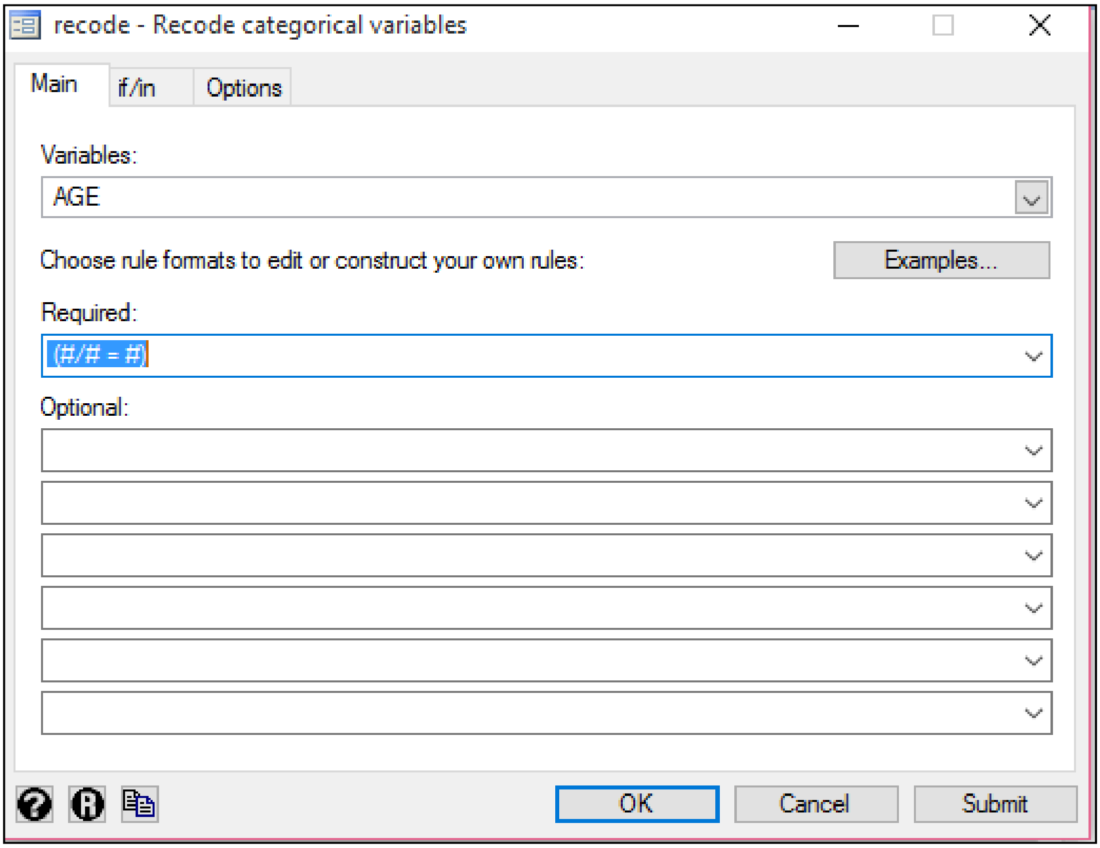
This command will generate new variable namely, "HH_SIZEsquared" where household size is squared. Similarly, new variables can be calculated with other mathematical functions.

10. Recoding

Recoding of a variable lead to collapsing of chosen categories in few or lesser categories. In this example, age of individual in different categories, viz. age group 1 to 10; 11 to 15; 16 to 24; and 24 and above has been recoded. To recode the variables based on above categories the following command is to be written: **recode AGE 1/10=1 11/15=2 16/24=3 *=4**

The above mentioned command will categorise the age variable in four parts. Here "*****" implies all other values. The "**recoding**” function in Stata can also be executed using dropdown menus as mentioned below:
Data > Create or Change Data > other variable transformation commands > recode categorical variable

On selection of “recode categorical variable” from the dropdown menu, screenshot of option box given in Fig. 15 will appear. Select the variable which needs to be recoded. In this example, "AGE" variable is selected from “Variables" dropdown pan. Additionally, the kind of recoding needs to be selected from "Required" dropdown pan given in the option box as shown in the Fig. 15.



Variables: AGE

Choose rule formats to edit or construct your own rules:

Required: (#/# = #)

Optional:

Fig.15: Option Box for Selection of Recoding of Categorical Variables

11. Percentile Calculation

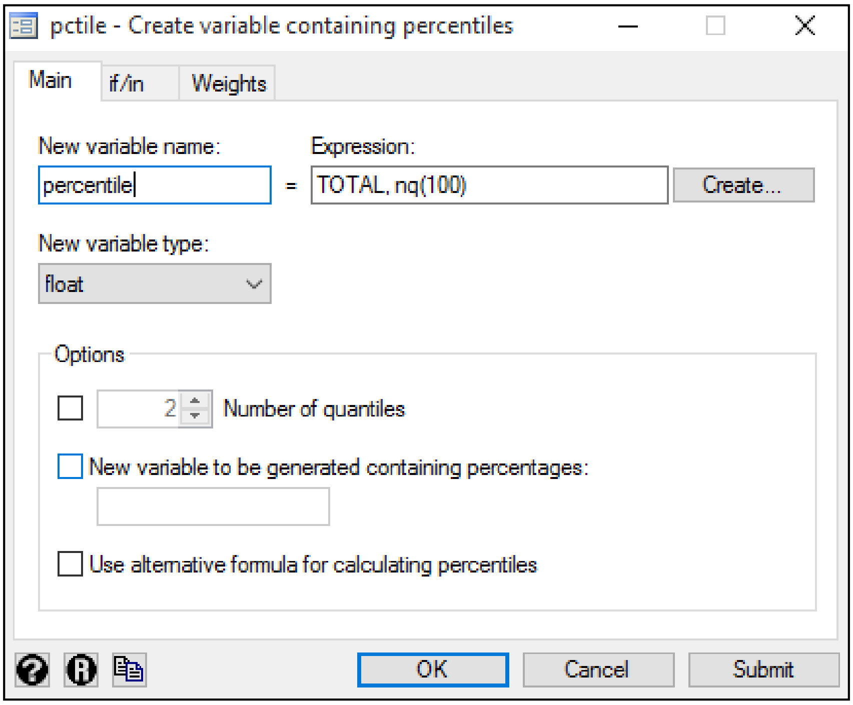
Stata can categories a distribution in different formats, e.g. percentile, quartiles, decile, etc.

To find out percentile of a variables, use the command given below:
pctile <varname> = <variable>, nq (100)

To calculate the percentile of consumption, use the command given below:
pctile percentile_consum = CONSUMPTION, nq (100)

This command will generate a new variable with the values corresponding to percentiles. Percentile can also be calculated using dropdown menus as mentioned below:
Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Create variable of percentiles

On making above-mentioned selections from the dropdown menu, screenshot of option box given in Fig. 16 will appear.



New variable name: percentile = Expression: TOTAL, nq(100)

New variable type: float

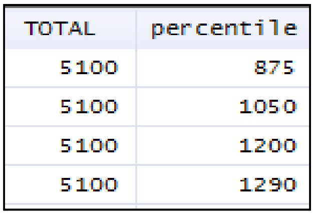
□ 2 Number of quantiles

□ New variable to be generated containing percentages:

□ Use alternative formula for calculating percentiles

Fig. 16: Option Box for Selection of New Variable (Percentile)

In the Option Box given in Fig. 16, write the name of new variable and expressions. Here, name of the new variable is "percentile"; and in Expression, the name of variable is written as "nq(100)" which mean percentile. This will produce a new variable, as shown in Table. 18.



TOTAL	percentile
5100	875
5100	1050
5100	1200
5100	1290

Table. 18: Generating New Variable (Percentile)

Similarly, you can also calculate quartile or decile using the following command:

To calculate the quintile value of a variable, user may use the following command:
pctile percentile_consum = CONSUMPTION, nq (25)

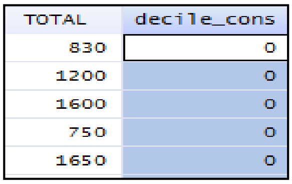
To calculate the decile values of variable the following command is to be used:
pctile percentile_consum = CONSUMPTION, nq (10)

Number 10 in the parenthesis (10), will produce the decile values as new variable.

It is to be noted that above commands dealing with percentile, quintile or decile provide us the values corresponding to percentile or decile; and do not categorize the whole distribution in e.g. 10 equal parts. The report provided by NSSO on 64th round explains some variables based on the decile values of distribution, however the above-mentioned command do not fulfil this requirement. As such, to categorize the distribution in 10 equal parts, use the command given below:

egen decile_cons=cut (TOTAL), group (10)

Use of above-mentioned command will provide a new variable which categorizes the distribution in 10 equal parts as shown in Table 19.



TOTAL	decile_cons
830	0
1200	0
1600	0
750	0
1650	0

Table 19: Distribution of Variables in 10 Equal Parts

Note: Calculation can be done based on decile values of consumption. For example, the education level of household based on decile class of consumption of households is shown in report provided by NSSO.

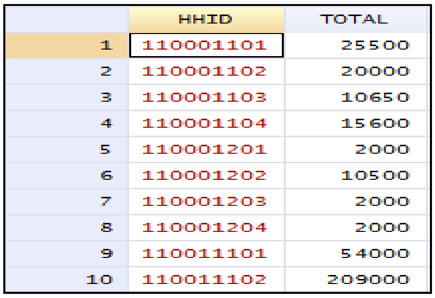
12. Data Aggregation

Data aggregation function in Stata provides the users to aggregate data from disaggregated data. For example, data is provided at individual (unit) level in NSSO datasets, and it is to be aggregated from individual data to household level. Data aggregation function can be used to achieve this. Use the command given below to aggregating the data:

collapse (sum) <variable>, by (grouping variable name)

To aggregate consumption expenditure from individual level to household level, use the below command : **collapse (sum) TOTAL, by (HHID)**

Here, "**HHID**" is the identification number of households and "**TOTAL**" is the total consumption expenditure. In this example, individual consumption data at household level is aggregated by summing up the consumption of individuals belonging to the similar households. The result of aggregation command is shown in Table 20.



	HHID	TOTAL
1	110001101	25500
2	110001102	20000
3	110001103	10650
4	110001104	15600
5	110001201	2000
6	110001202	10500
7	110001203	2000
8	110001204	2000
9	110011101	54000
10	110011102	209000

Table 20: Aggregation of Individual Consumption Data at Household Level

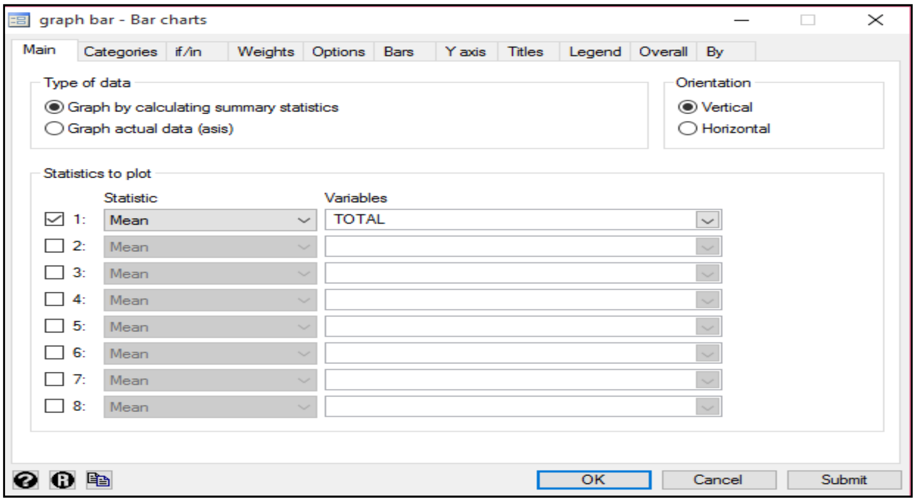
13. Graphics in Stata

Stata can produce a wide variety of graphs, viz. bar charts, pie charts, histograms, scatter plots, etc.

13.1 Bar Charts

To generate Bar charts, select the following options from the dropdown menus: **Graphics > Bar Charts**

Option Box given in Fig. 17 will appear on above mentioned selection.



Type of data
● Graph by calculating summary statistics
○ Graph actual data (asis)

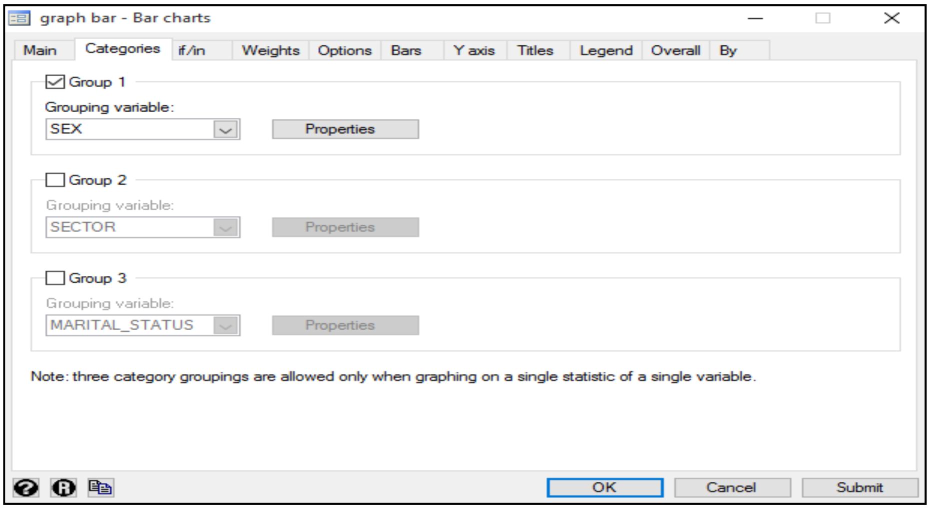
Orientation
● Vertical
○ Horizontal

Statistics to plot

	Statistic	Variables
☑ 1:	Mean	TOTAL
□ 2:	Mean	
□ 3:	Mean	
□ 4:	Mean	
□ 5:	Mean	
□ 6:	Mean	
□ 7:	Mean	
□ 8:	Mean	

Fig. 17: Option Box for Drawing a Bar Chart

Here, bar chart is created with the mean values of total consumption for different sex viz. male and female. To do so, "total consumption" is selected from the variable list and mean is selected as the statistics option. Further, to select the "Sex" variable, click on "**Categories"** as shown in Fig. 18, the screenshot of "Option Box" that appears. Then select "SEX" variable in the Group 1, and clicked on "OK".



☑ Group 1
Grouping variable: SEX

□ Group 2
Grouping variable: SECTOR

□ Group 3
Grouping variable: MARITAL_STATUS

Note: three category groupings are allowed only when graphing on a single statistic of a single variable.

Fig. 18: Option Box for Creating Bar Chart

The resultant bar chart that would be generated is shown in Fig. 19 depicting the mean values of consumption among male and female.

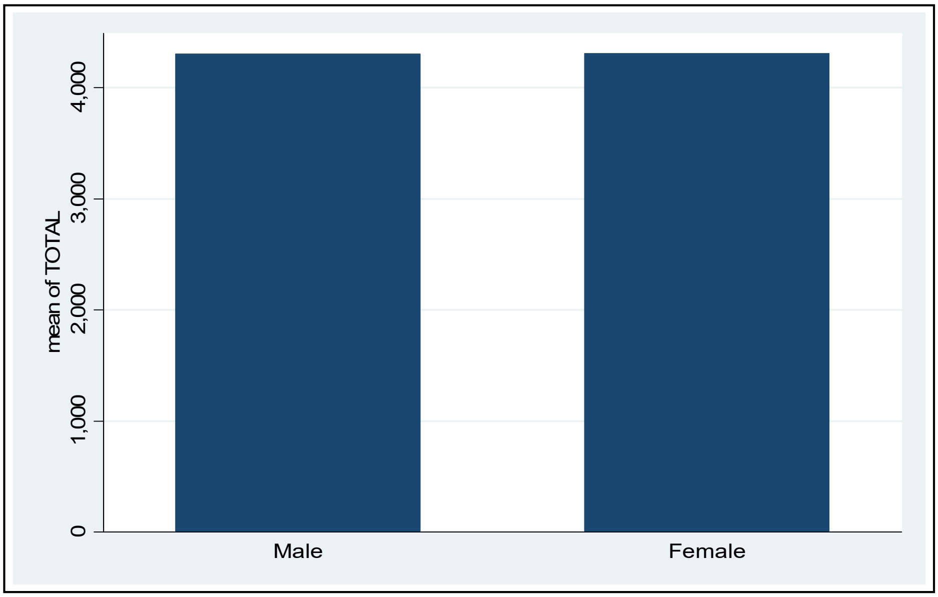


Fig. 19: Mean Values of Consumption among Male and Female

To add one more categorical variable to the bar chart, i.e. sector, click on "categories" tab and check at "Group 2" and select "Sector" as depicted in Fig. 18. The chart generated with additional inputs is shown in Fig. 20.

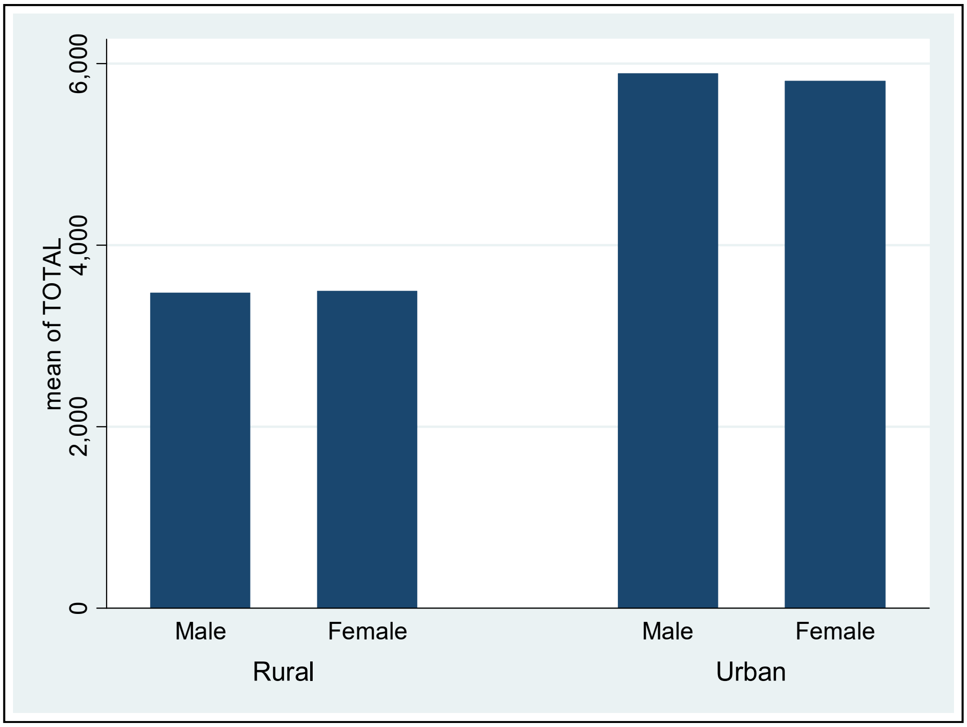
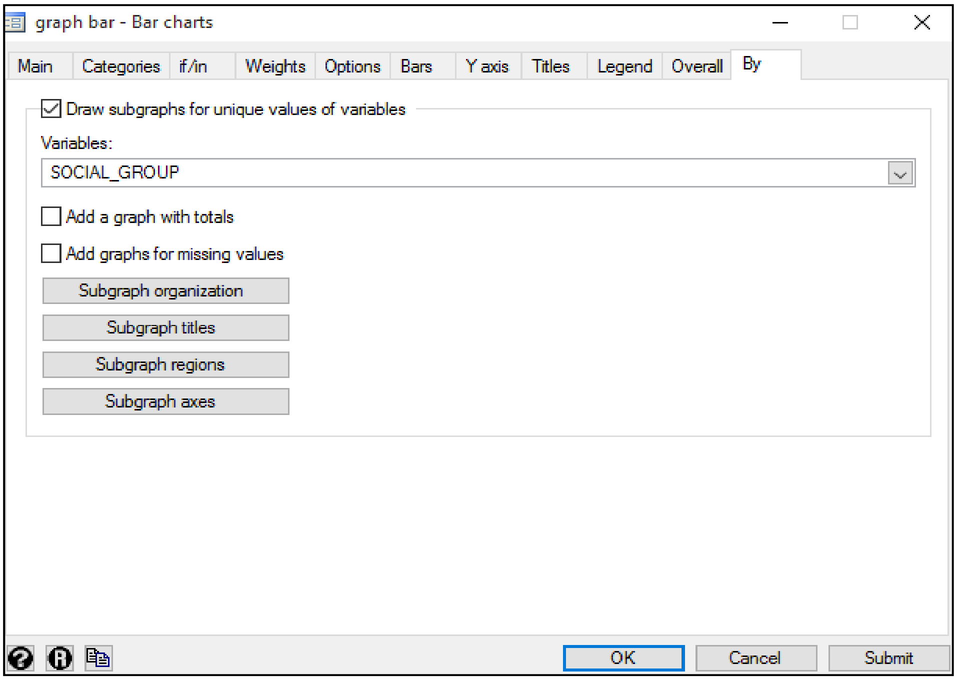


Fig. 20: Total Consumption by Males and Females in Rural and Urban Areas

Further, expenditure on total consumption can also be plotted with one or more variable and for new graphs can be created. This can be performed by using "**by"** function in Stata. Click at "**By"** as shown in Fig. 17. A new screen will be displayed as shown in Fig. 21, wherein an additional variable can be selected for depicting the mean values of total consumption. In this example, "social group" is selected which has produced the bar charts as shown in Fig. 22.



☑ Draw subgraphs for unique values of variables

Variables:

SOCIAL_GROUP

□ Add a graph with totals

□ Add graphs for missing values

Fig. 21: Option Box for Selecting Additional Variables

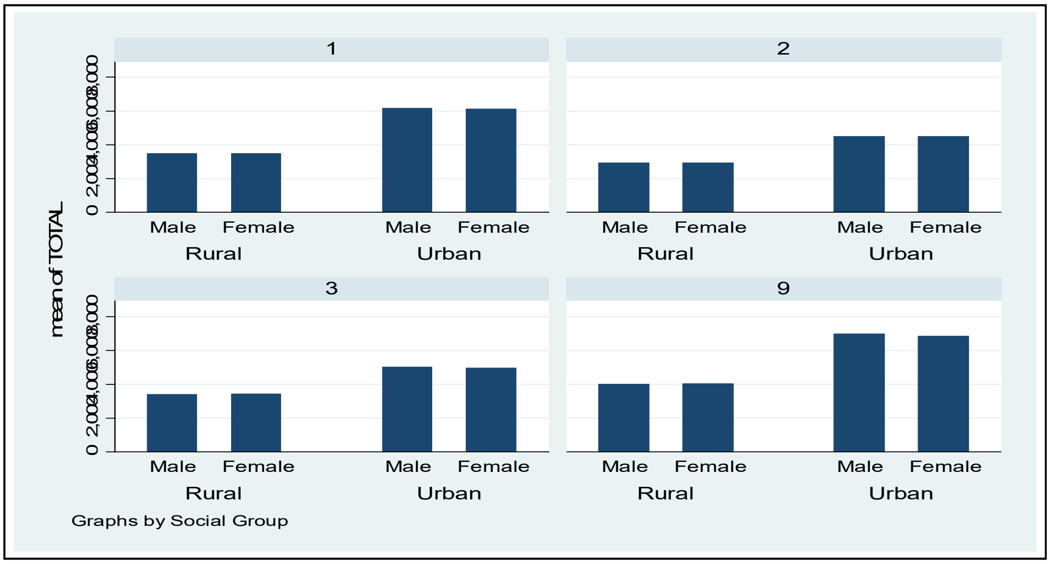
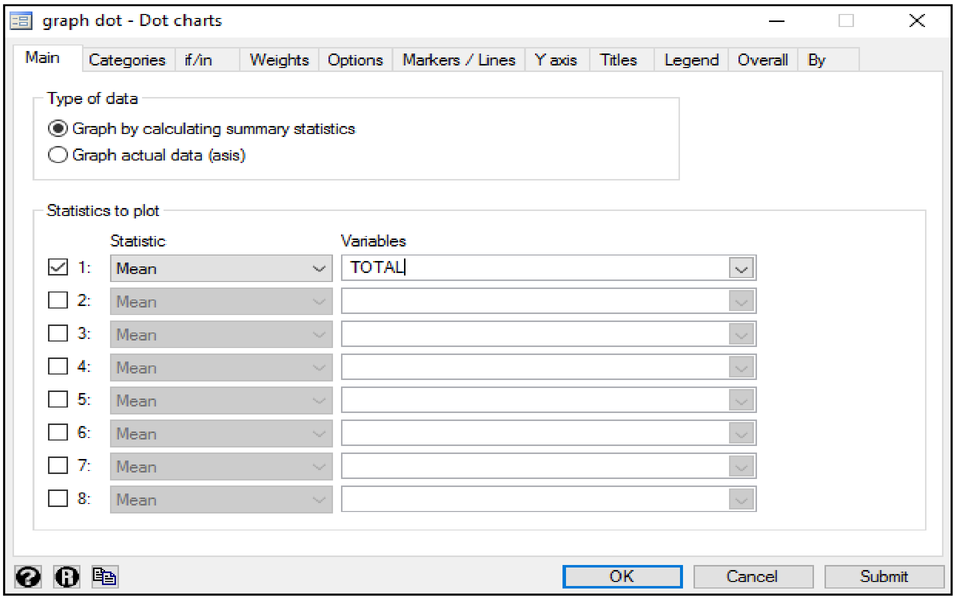


Fig. 22: Average Expenditure on Total Consumption for Different Social Group, Gender &Sector

Fig. 22 shows the average values of total consumption expenditure for different social group, different sex and sector.

13.2 Dot Charts

To generate "Dot Charts" in Stata, select **Graphics > Dot Charts** from the dropdown menus. After that, the following screen having option boxes will appear as shown in Fig. 23.


graph dot - Dot charts

Type of data

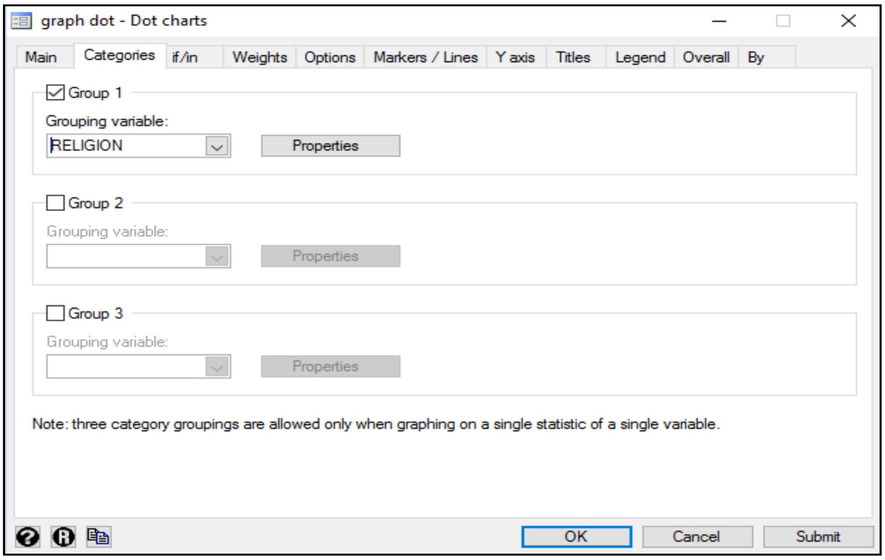
● Graph by calculating summary statistics
○ Graph actual data (asis)

Statistics to plot

	Statistic	Variables
☑ 1:	Mean	TOTAL
□ 2:	Mean	
□ 3:	Mean	
□ 4:	Mean	
□ 5:	Mean	
□ 6:	Mean	
□ 7:	Mean	
□ 8:	Mean	

Fig. 23: Opening Screen with Option Box for Drawing Dot Chart

You can select variables in the option boxes as per your requirement to generate dot charts, as here the "total consumption" variable with the name of "TOTAL" has been selected as shown in screenshot Fig. 23. After that you need to click on "categories" button where as a result Fig. 24 will appear. In that you can select the required variable in the "Grouping variable" option box as "religion" variable is selected here and click on "OK".


graph dot - Dot charts

☑ Group 1
Grouping variable: RELIGION

□ Group 2
Grouping variable:

□ Group 3
Grouping variable:

Note: three category groupings are allowed only when graphing on a single statistic of a single variable.

Fig. 24: Option Box for Drawing Dot Chart

Now the dot chart will be generated on the basis of "Mean Values of Total Consumption for Different Religion" as shown in below Fig. 25.

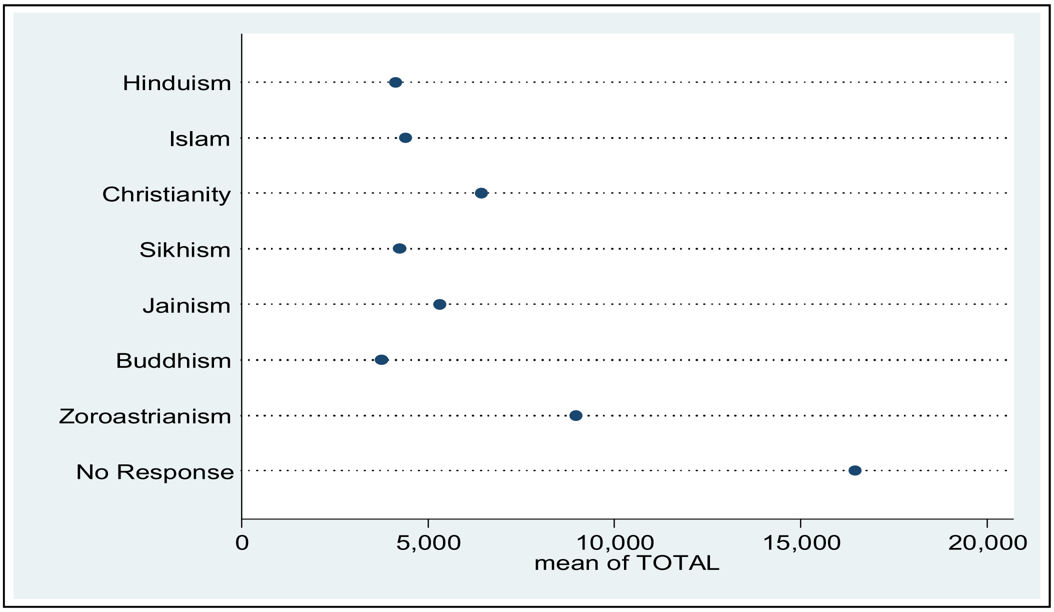


Fig. 25: Mean Values of Total Consumption for Different Religion

Further, dot charts can also be drawn for "rural and urban population", in addition to" expenditure on consumption based on religion". To add this additional variable, click at last tab **"By"** available on the screen (as shown in Fig. 24) and select your required variable and click "submit" to generate chart. As a result, the chart shown in Fig. 26 will be produced.

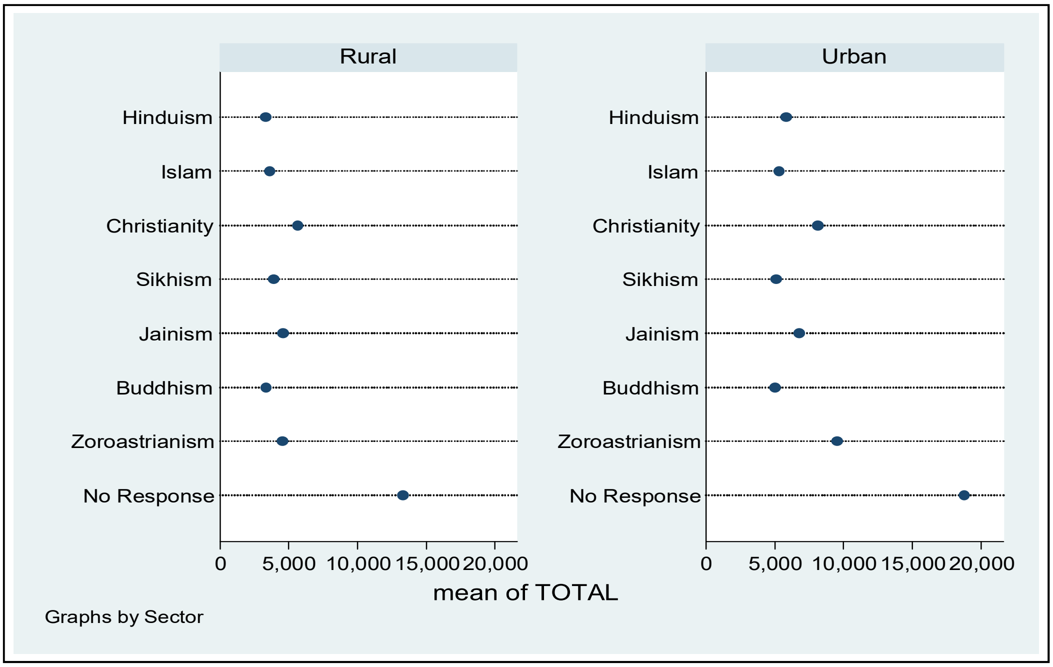


Fig. 26: Mean Values of Total Consumption in Different Religion for Rural and Urban Population

13.3 Pie Charts

Pie chart can be created to depict percentage share. To create pie chart in Stata, select the following from drop down menu: **Graphics > Pie Charts**

Screen with Option Boxes will appear consecutively as similar to Fig. 17, 18, 21, 23 and 24. The steps to generate pie charts in Stata are as same as drawing bar charts and dot charts. In the example given here, variable "religion" is selected. On submission, a pie-chart is produced as shown in Fig. 27 depicting religion-wise sample households.

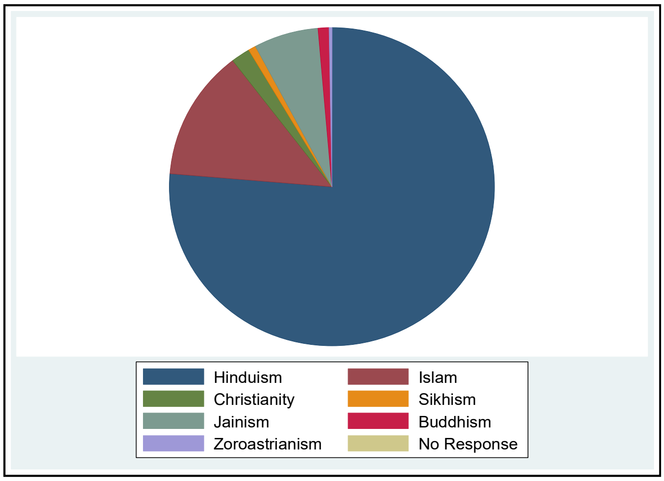


Fig. 27: Religion-wise Sample Households

Likewise, sample for male and female population can also shown as a pie chart in addition to religion-wise population. For this, you need to click **"By"** tab and select the required variable and click on "submit". This will produce a pie chart as shown in Fig. 28.

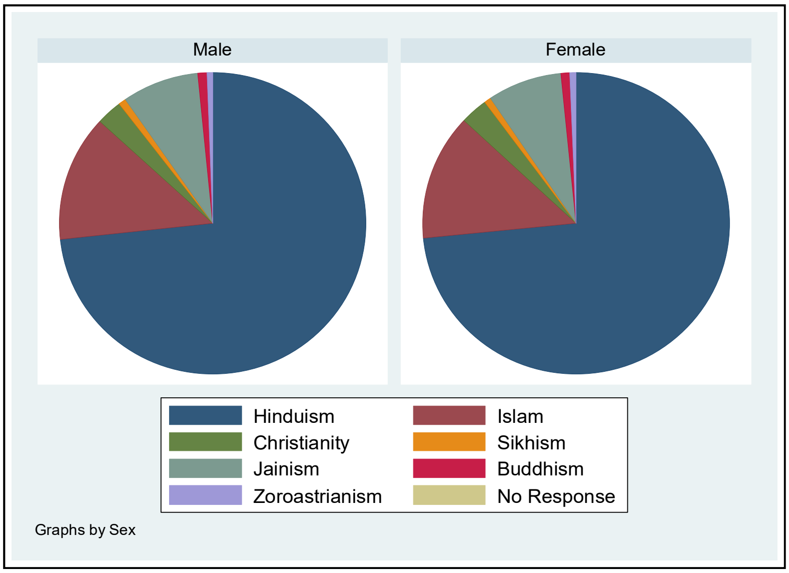
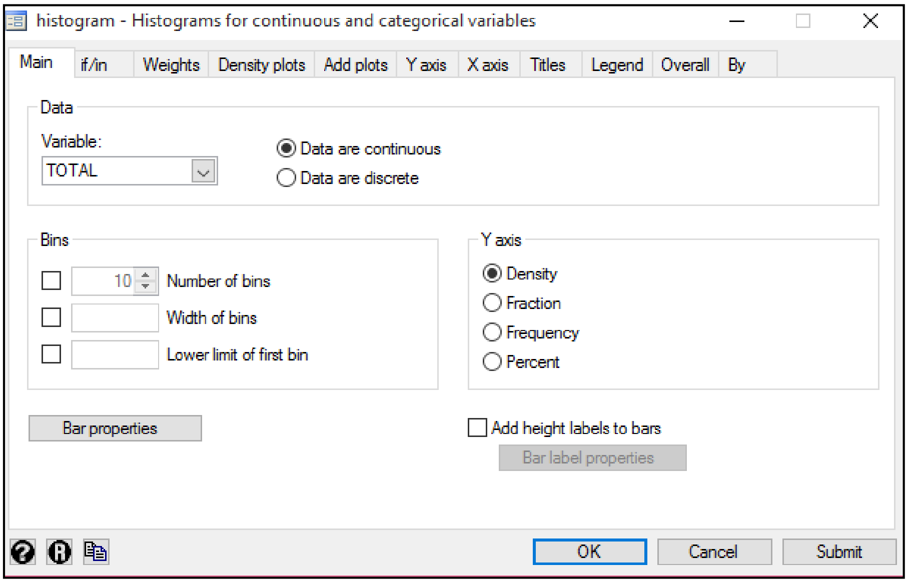


Fig. 28: Male and Female Population for Different Religion

13.4 Histogram

A histogram can also be drawn in Stata using the following selections from the dropdown menus: **Graphics > Histogram**

On selection of “Histogram” from the dropdown menu, a screen will appear as depicted in below screenshot Fig. 29.



histogram - Histograms for continuous and categorical variables

Data
Variable: TOTAL
● Data are continuous
○ Data are discrete

Bins
□ 10 Number of bins
□ Width of bins
□ Lower limit of first bin

Y axis
● Density
○ Fraction
○ Frequency
○ Percent

□ Add height labels to bars

Fig. 29: Option Box for Creating a Histogram

From that screen, select the variable of which "histogram" is to be created. In this example, "Total Consumption" is selected. On clicking at "submit”, histogram shown in Fig. 30 will be generated.

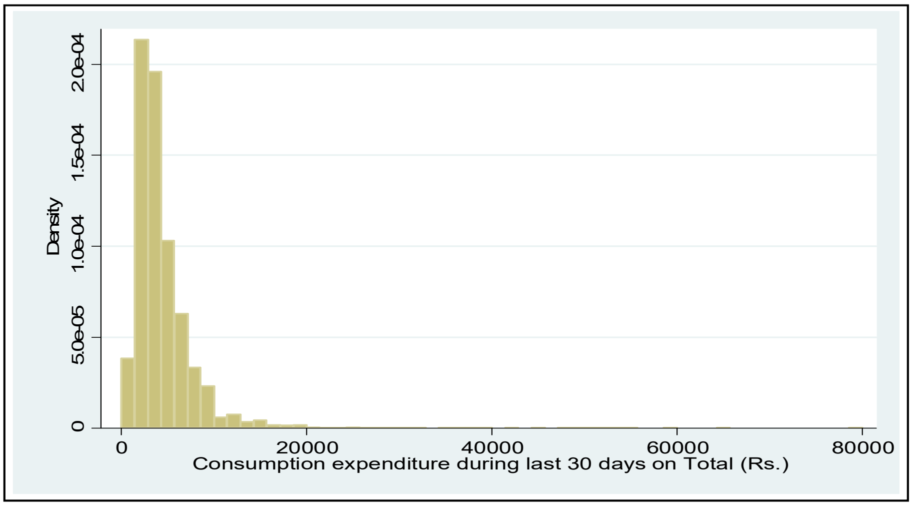


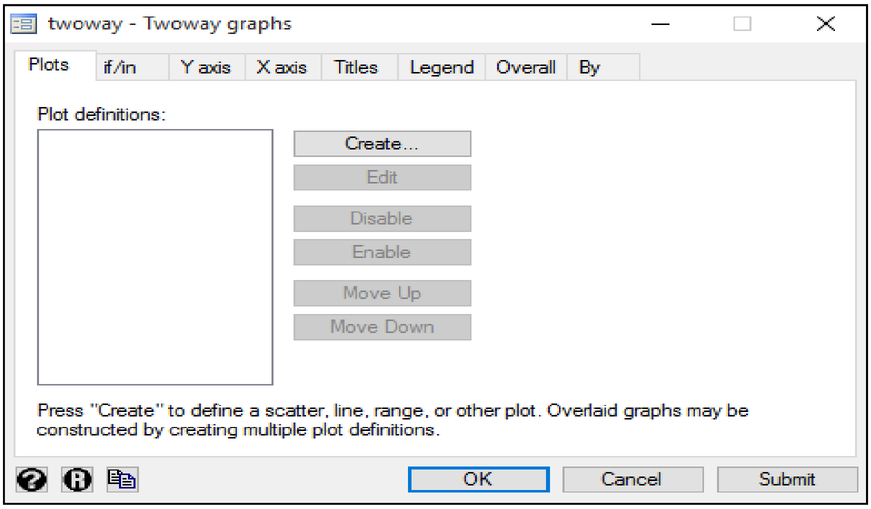
Fig. 30: Consumption Expenditure during Last 30 Days

Similarly, one may also create histogram for categorical variables, as shown in case of pie chart.

13.5 Two-way Scatter Plot

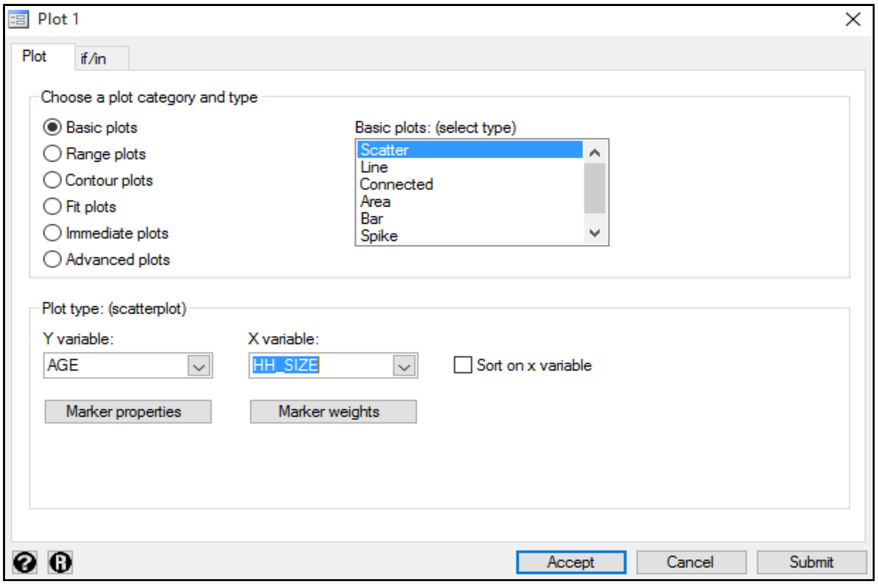
Two way scatter plot shows relationship between two variables which help in finding out the outliers i.e. the values which are not similar to others. In order to prepare a two way scatter plot, select the following options from the dropdown menus: **Graphics > Two Way Graphs (Scatter, Line, etc.)**

On selection of “Two Way Graph” from the dropdown menu, screenshot of option box given in Fig. 31 will appear. Click at "**Create...**" to create a two way scatter plot. On clicking at Fig. 31, Fig. 32 will appear wherein users can select Y variable and X variable in the respective boxes. In this example, "Age" is selected in the Y variable box and "HH_Size" in selected in X variable box as shown in Fig. 32. On clicking at "submit" button, a two-way scatter plot is generated as shown in Fig. 34 which shows relationship between age and household size.



Press "Create" to define a scatter, line, range, or other plot. Overlaid graphs may be constructed by creating multiple plot definitions.

Fig. 31: Option Box for Creation of Plots / Graphs



● Basic plots

Basic plots: (select type) Scatter

Plot type: (scatterplot)

Y variable: AGE

X variable: HH_SIZE

□ Sort on x variable

Fig. 32: Option Box for Creation of Plots / Graphs and Selection of X and Y Variables

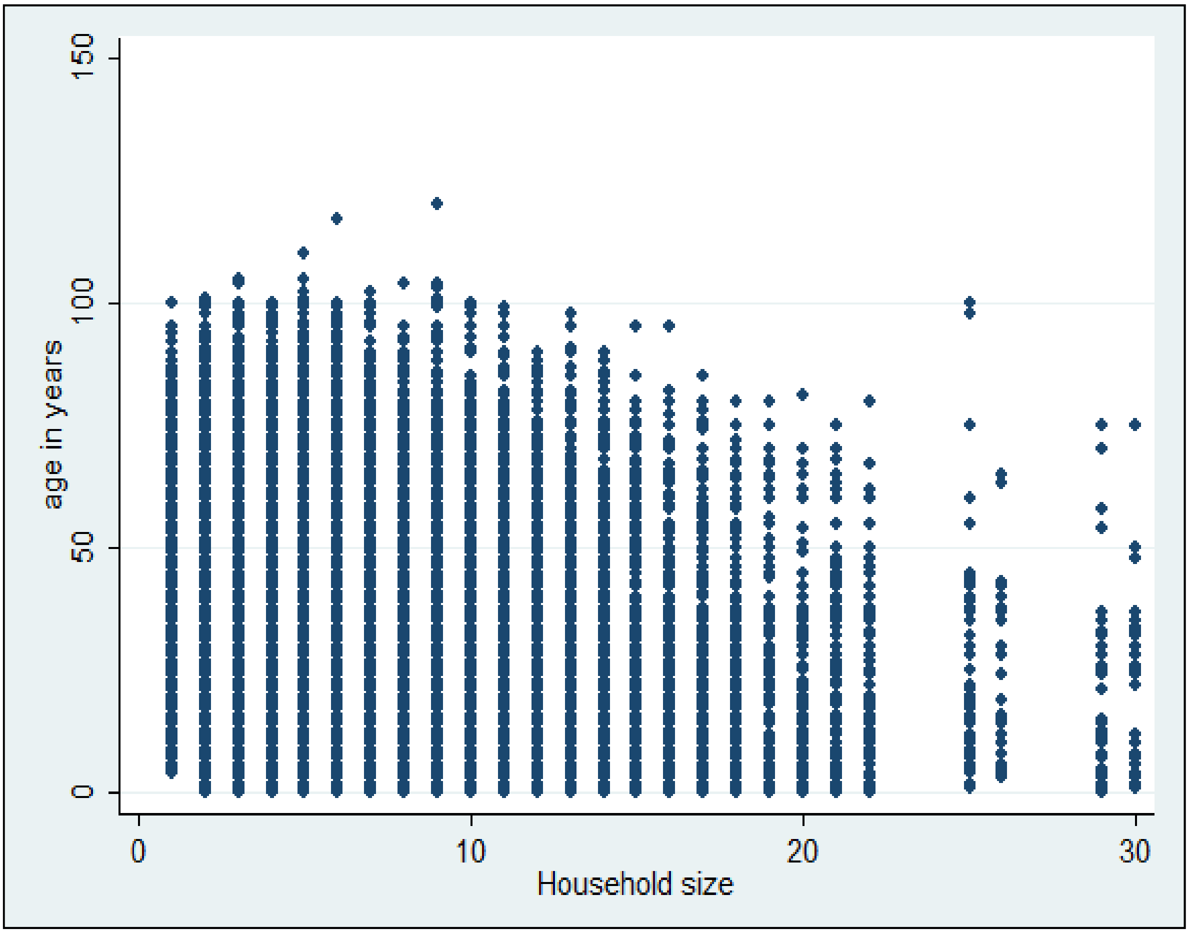


Fig. 33: Total Expenditure on Consumption: Relationship between Age and Household Size

Similarly, different two way charts can be created based on requirements, i.e. two way bar charts, time series charts and so on by selecting required plot types as depicted in Fig. 32.

14. Using Log in Stata

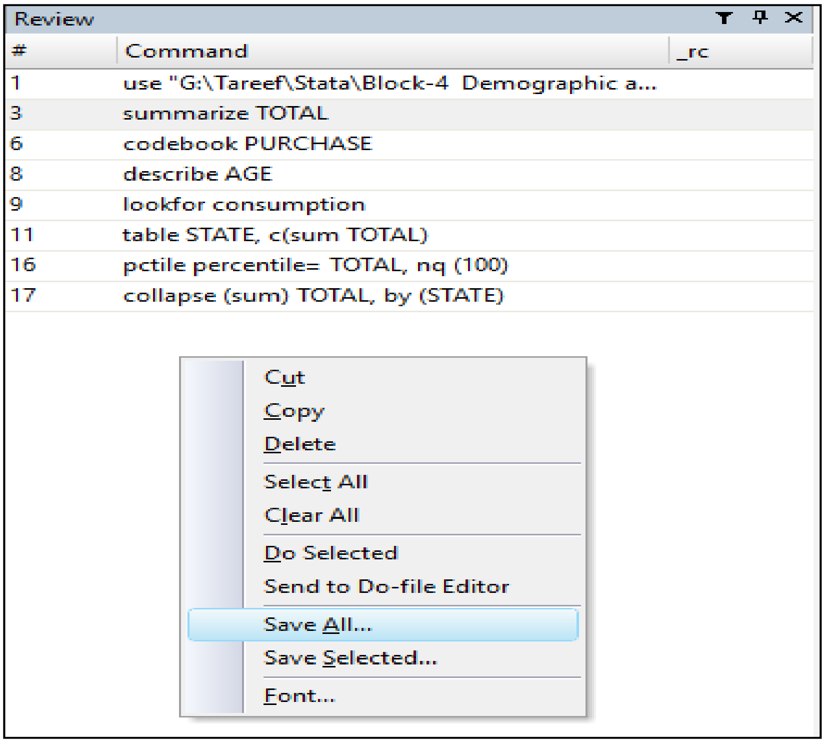
"Log function" in Stata provides the facility of saving the main window in the system. The main window of Stata (shown in Fig. 2), where all the results appear can be saved as SMCL file in the system. Use the following command:
log using <location and file name>

On using the above-command, all the work done during the session will automatically be saved in the assigned file that can be opened for later use. Log function in Stata can also be executed using following command from dropdown menus as mentioned below:
File > Log > Begin

Likewise, you can close, suspend and resume the log anytime in between the use of Stata for different datasets.

15. Do file in Stata

"Do file" in Stata provides the facilities of saving the command which have been used by the user. These saved command can be used in the future. Select Review option of Stata, click and select "save all" option as shown in Fig. 34.



#	Command	_rc
1	use "G:\Tareef\Stata\Block-4 Demographic a...	
3	summarize TOTAL	
6	codebook PURCHASE	
8	describe AGE	
9	lookfor consumption	
11	table STATE, c(sum TOTAL)	
16	pctile percentile= TOTAL, nq (100)	
17	collapse (sum) TOTAL, by (STATE)	

Fig. 34: Saving Previously Used Commands in Stata using Review Option

You may also open a new "do file" and use command in that file, you may also run the command from the do file. To open a new "do file", click on "new do file" editor from the toolbar, a new window will appear as shown in Fig. 35.

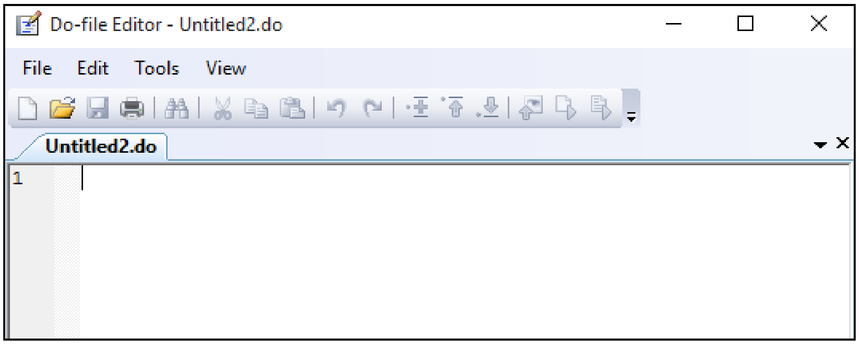


Fig. 35: Do-file Editor

In Fig. 35, write command and click on to execute (do) button from the toolbar. These commands will run in Stata automatically.

16. Merging Data

Merging data command is one of the most useable commands in Stata. This command provides linking or merging of two or more files in one file. Merging data command is used at the time of comparing two variables that are in two different files. There are two types of merging data namely, i) adding variable; and ii) adding cases.

16.1 Adding Variables

Adding variables imply that the merging of two files with addition of variables in one file from other file(s). Merging of two files does not involve much complexities when the number of cases are same and they are at the same level of measurement i.e. both of the files are either at household level or individual level. But, it becomes more complex when the two datasets are at different measurement level i.e. one is at household level and other one is at individual level.

These two sorts of adding variables have been elaborated in the subsequent sections.

16.1.1 Adding Variables with Same Level of Measurement

Adding variables with the same level of measurement implies that to files are being merged at the same level of measurement, i.e. both files are individual files. In order to merge two files, follow the steps given below:

a) Generation of Common ID;
b) Sort both files based on common ID; and
c) Merging Data

In order to generate a common ID, use the following command:
gen ID=variable1+variable2+variable3

In NSSO 64th round (Schedule 25.2) these variables include FSU_SL_NO, HG_SB_NO, SSS_NO, SAMPLE_HH_NO, and PERSON_SL_NO.

After generating common ID for both files, sort both files with common ID using "**sort ID"** command. After generating and sorting the data, one can merge both the files using the following command:
joinby ID using <filename>

This command merge two files based on one common ID, i.e. ID. In this example, the following command is used: (common ID???)
joinby ID using "H:\User\Stata\Block-6 Particulars of private expenditure for those aged 5-29 years who are currently attending at primary level and above.dta".

Here, while Block-5 file is kept opened, Block-6 file is merged into the Block-5 file. To check, whether two files have merged or not, "**describe command"** is used which reveals the list of variables.

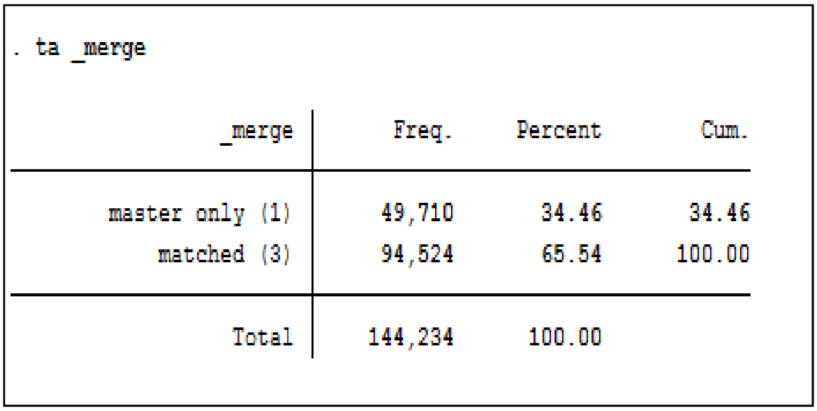
16.1.2 Adding Variables with Different Level of Measurement

For merging two files having different level of measurement, i.e. one at individual level and other at household level, first sort the data by the common ID. After sorting data, merge the two files by using either "one to many" or "many to one "command in Stata.

While merging data from individual level to household level, use the following command:
merge 1: m ID using <individual file name>

While merging data from household to individual level, use the following command:
merge m: 1 ID using <individual file name>

After using the above commands, check whether the files are merged or not using "**ta _merge"** command. If the value turns out to be 3, it implies that the files are merged. For example, in Table 21 the values 94, 524 are merged, as revealed by 3 in parenthesis.



```
. ta _merge

         _merge |      Freq.     Percent        Cum.
----------------+-----------------------------------
master only (1) |     49,710       34.46       34.46
    matched (3) |     94,524       65.54      100.00
----------------+-----------------------------------
          Total |    144,234      100.00
```

Table 21: Checking Merged Files

16.2 Adding Cases

In stata user may also add cases in the same variable through the following command:
append using "file name"

Here, file name implies that a user has to specify the location of file and name of file e.g. "C:\Users\admin\Desktop\second.dta".

17. Help and Find It

"Help and find it” is a very useful command in Stata. It provides help on a given command. For example, to get help on describe command, use: **help describe**.

Similarly, you can also use “**find it**” command, if you do not know the full command for a given function, e.g. use the following command to find function of “describe” command: **findit describe**

For further information please contact

Dr. Upendra Choudhury
Member-Secretary
Indian Council of Social Science Research
Aruna Asaf Ali Marg
New Delhi - 110067
E-mail: ms@icssr.org

Dr. Jagdish Arora
Director
Information and Library Network Centre
An IUC of UGC
Infocity, Gandhinagar -382007, Gujarat
E-mail: director@inflibnet.ac.in

Indian Council of Social Science Research
JNU Institutional Area, Aruna Asaf Ali Marg
New Delhi - 110067 (INDIA)

For more information : http://www.icssrdataservice.in/.