INTRODUCTION STATA AND THE DATA

OBJECTIVES: By the end of this lab, you should have:

1. Found where supplementary documentation for the data sets is stored (e.g.: the survey questionnaires) and looked at some typical questions from these surveys
2. Learned how to open a log-file and save it (and why log files are important!)
3. Found and accessed the data sets we will be working with; become familiar with the way that survey data can be captured and displayed in the statistical package we will use, called STATA.
4. Started to learn some of the basic commands for getting around in STATA (e.g.: the **`help`** command)

1. SUPPLEMENTARY DOCUMENTATION FOR THE DATA SETS

We will be using Statistics South Africa's (Stats SA) 2002 General Household Survey (GHS) which has been constructed from information collected from household surveys. We are going to have a look at what types of questions a typical household survey might ask, before looking at the 'answers' as they are stored in the data sets. The questionnaire for each of these surveys, as well as other material explaining how certain variables were created and coded (called the 'metadata') are available.

If you open the questionnaire at the top of the document, you should be able to read: "General Household Survey". This first page is a Cover Page, and captures information that is not usually available in the data set eg: the name of the interviewer and the address of the household.

What types of questions do household surveys ask?
Typically, a wide range of topics may be covered in such surveys:

- Personal Descriptive (demographic) questions - Surveys always include questions about personal-social characteristics of respondents such as sex, age, occupation, income, and education. This kind of information helps researchers to understand relationships between variables they may be interested in. For example, we could look at how many people in the age group 25-35 have a tertiary education, and what types of jobs they have, compared to others in the same age group with less education.
- Behavioural questions - Many survey questions relate to people's actions or behavior in various social areas. In economic research, for example, patterns of spending and saving have been studied intensively using surveys.
- Attitudinal questions - Some survey questions concern people's beliefs, opinions, attitudes, or expectations. For example, there could be questions about how safe people feel in their homes, compared to an earlier period. These are areas in which it is unlikely that data would be available from non-survey sources.
- Living Environment questions – These questions gather data on the circumstances in which respondents live, and may include information about the neighborhood, the adequacy of living quarters, membership in groups and organizations, and so on.

Let's have a look at some of the questions asked in the GHS survey. Scroll down to page 2 of the questionnaire where the first questions for the household roster appear. You can see the different types of questions that appear: some collect information which will not be released in the data (e.g.: name of the person), some collect information that must be coded (e.g.: if you answer YES to question B, your answer is recorded as 1), and still others must be written in as heard (for example: how old is .....?). Many of the questions will have answers that are coded: for example, if you are the spouse of the head of household, your 'answer' to question 1.1 (on page 3) will be recorded as 2.

There are hundreds of questions in any household survey, and thus many codes representing different answers. In the data sets that we will look at, sometimes the data will be labelled clearly (so when you look at whether an observation is male or female, the gender will be clear from how the data is labelled); but often, variables will not be labelled. For example, when we look at data on relationship to head, this

information may have only been captured as a list of numbers, rather than as "head", "spouse", "daughter" etc. Having the survey at hand means that we can always go back to the original question and check what the codes stand for. Before you use a variable, it is also important to clarify how a question has been phrased in the original questionnaire and how it has been coded, because your data may not be labelled as clearly as you would like.[1]

> See if you can find the section that asks questions about each household members' education, and look at the types of questions that appear in this section.
>
> See if you can find the section that asks questions about the household's access to land. Which question would you be directed to if you answered YES to question 4.28?

2. GETTING STARTED IN STATA: LEARNING ABOUT LOG FILES AND OTHER ISSUES

STATA is one type of statistical program that economists use to manage, look at and analyse economic data. It is particularly suited to the analysis of complex survey data – in our case, data from household surveys. It has a library of pre-programmed commands, which are useful for creating summary statistics from the data, and for performing more detailed statistical analyses. STATA is also programmable (we can write programs to tell STATA to perform a specific task), but we will largely not be learning how to program in STATA in this course.

Open up STATA 8 by navigating to

→ Start| Programs | Economics Software | STATA SE 8

You should now have the STATA window open. Note that there is a set of pull down menus, as well as 4 smaller windows, named: Review, Variables, STATA Command, and STATA Results. When we want to tell STATA to do something (execute some command), we will type it into the STATA Command line window. The results of our command will show up in the STATA Results window. In any session, all commands that we give to the program will show up in the Review window once we execute them. The Review window is useful, because it allows us to recall commands quickly (by clicking on the relevant command in the Review window it will reappear in the STATA Command line) and can be saved, for editing later into a do-file.[2]

Right now the Variables window is empty – this is because we have not loaded any data. Whenever you have a data set open, the variable names will show up in this window.

If you want to record anything that you do in a STATA session so that you can look at results later, you need to open a log-file. A log file is simply a record of all the commands you enter into STATA and the output from those commands. The key is to make sure you have a log file open at the BEGINNING of a STATA session, and to CLOSE it once you have finished, and BEFORE you close STATA.

Download person.dta from the web if you have not already done so.

NB you will need to replace F: with the relevant path for the filespace (directory) where you are storing the downloaded files..

There are three ways you might open a log file:

→ You can open a log file by typing:

[1] Something else that is useful for checking the way that data has been coded is the 'metadata' for a survey. This is a document written about how the data were captured, and so has a lot of detail which may be useful to you. See if you can open the Metadata file for GHS now (**`metadata.pdf`**). Scrolling down, you should be able to see how some of the questions have been coded. The metadata is useful when trying to figure out how answers like 'not specified'; 'not applicable' and 'did not reply' were coded.

[2] A STATA do-file is a short program file, which runs a set of commands at once.

log using F:*yourlastname*_LAB1.log

in the STATA Command window. If you want to create your log file by saving over an old one, you can type

log using F:*yourlastname*_LAB1.log, replace

If you wanted to add to that file instead of replacing it with a new one, you could write **append** instead of **replace**.

log using F:*yourlastname*_LAB1.log, append

Be sure to save your log file with the extension **.log**, NOT **.smcl**. This will make it easier for you to edit, cut and paste your log in any text editor. To verify that it is a .log file and not a .smcl file, look at the output that Stata shows you:

```
. log using "F:\hoek_LAB1.log"
-------------------------------------------------------------------------------
----------
      log:  F:\hoek_LAB1.log
 log type:  text
opened on:  16 Oct 2003, 15:48:53
```

If it says smcl instead of text under log type, then you have done it incorrectly. Close the log file and do it again.

Now that you have a log file open, we can start our STATA session.

3. LOOKING AT SOME OF THE DATA, AND HOW IT IS CAPTURED IN STATA

Before you open up your data you need to make sure that enough memory is allocated. STATA allocates 10m as a default. While this may be sufficient for small data sets often you will need more memory than 10m. To allocate 30m of memory type

set mem 30m

If you do not allocate enough memory before you try to open a data set STATA will issue the error message 'no room to add more observations'.

Now that you have allocated memory, open up the data set that you have downloaded, **person.dta**, by choosing the FILE drop down menu, and then choosing OPEN.

First let's take a more detailed look at what variables are stored in the data set: at the STATA command line, type in:

describe

You should see a list of what variables are contained in the data, how many observations there are, what kind of variables they are (eg. 'byte', 'float', 'int', 'str' or 'long' variables), and any comments which are attached.[3] Now type in:

browse

This command directs you to a spreadsheet, where the data appears. You should note the following:

→ each OBSERVATION (in this case, each person in the household for which there is recorded information) appears on a separate ROW of the spreadsheet. For example, the first person has UqNr = 1011001002201 (the unique household id number), PersonNr=1 (their person number within the household), Prov=1 (they are from the Western Cape) and C_Gender=2 (they are female). If you move along the row, you can see the rest of his data recorded for each question

[3] If you see --**more**-- at the bottom of your screen, you need to press the space bar to continue scrolling.

→ each VARIABLE appears in a separate COLUMN of the spreadsheet. What is the variable for 'What is the highest level of education that you have completed?' called? What is the variable for the question 'Is ....'s mother still alive?' called?

EXAMINING OBSERVATIONS
Using the **browse** command, find the 20136th observation.
What is the person number of this observation? What is their relationship to the household head? Their gender? Their education? Can they read in at least one language?

A quicker way to do this is to type in :

```
list
```

You should see ALL the data for the first observation being displayed in the Results window. If you continue to press the space bar, you will scroll through the entire data set – which is not what we want to do! A really useful key to use here is the **break** key, the red button with the white cross mark, at the right hand side of the tool bar at the top. If you click on this button, you interrupt the command that is currently being executed, and return control to yourself.

If you type **list in 20136**, you will see only the record for the observation that we want. Note that the household id number (UqNr) for this person is 2582180010101, while the person code (PersonNr) is 13. Together, these numbers UNIQUELY identify the person. So, another way to display only the record for this observation is to type:

```
list if UqNr==2582180010101 & PersonNr==13
```

Note that we typed in TWO equal signs when we wanted to tell STATA to fetch the observation for which the household id was 'equal to' some number. This is standard STATA syntax.

If we wanted to see only this person's data in the spreadsheet format, we would type in
```
browse if UqNr==2582180010101 & PersonNr==13
```

Using the **browse** command again, pull up the data for all the people in the household with UqNr 2582180010101. We're going to spend some time looking inside a specific household, to get a feel for what type of information is being collected in the GHS at the individual level.

Can you describe the composition of the household? Which VARIABLES would you look at to describe this household?

Has anyone in this household suffered from any injuries or illness in the past month?

What did person number 6 in the household answer to the question 1.11: "Is ____ currently attending school or any other educational institution?" What about person number 2?

How many children under 18 years of age are there in the household?

Can the household head read in at least one language?

Now consider the questions that we have looked at above, for a household of your choice. Use the **browse** command to find a household, and jot down a description based on the few variables of interest we have discussed. Make sure you distinguish between information about individuals, and information about the household. Be prepared to describe your household to the class.

EXAMINING VARIABLES

We've looked at values of particular variables for a few individuals and a few households. But now we want to look at these values more generally – for the entire data set. Close the spreadsheet and return to the STATA command window. There is an efficient way to find the names of variables that you are interested in. Type in:

`lookfor educ`

This should give you a list of all the variables which have 'educ' in their name.

Suppose you are interested in the variable **`age.`**

→ get back into the spreadsheet using **`browse,`** locate the relevant column, and take a look at the entries for this variable. Not a very useful way to learn about the variable, right?

→ now get back to the STATA command line, and type in

`codebook D_Age`

This command gives us more summarised information about the variable: (among other things), it tells us the range of values captured, the mean and standard deviation and the number of missing values.

Go back to the GHS questionnaire. Pick any question about education which you would be interested in finding out answers for. Find out what the variable name associated with this question is. Now suppose you want to look at this variable in conjunction with a few others. For example, if you type in

`browse UqNr PersonNr D_Age C_Gender` ***`thevariableyouchose`***

the spreadsheet will only show you data from these variables.

We have just worked from the questionnaire to the data set, to find variables that we are interested in. However, there may be some cases where we need to work in the opposite direction: for example, when we find a variable that has no clear value labels. In this case, you will want to work from the data set back to the questionnaire, to clarify (1) which question the variable captures information for; and (2) how the variable was coded.

SIMPLE SUMMARSING COMMANDS

A frequency distribution table lists all the observed values for a given variable and the number of observations that take on each of these values. **`tabulate`** produces one- and two-way tables of frequency counts along with various measures of association (relationships between variables). In this first class, we will use this command to produce summary tables of data, to answer the general question: how much of our data set falls into different variable categories. We can see what **`tab`** does easily by example.

So, if we wanted to know how many people in our data set were classified as 'coloured', then we could type:

`tab E_Race`

and we would get a table specifying the number of observations of each race-type, and the percentage of the data set taking on each of the variable values. From the table, what is the answer to our question above? What percentage of the data set is Indian?

To get a frequency distribution of the highest level of education in the sample, you could type:
`tab Q110HiEd`

For how many people is the highest level of education unspecified? If you look at the metadata for the GHS you will notice that Stats SA generally uses 9 (or 99, 999 etc) for "Unspecified" and 8 (or 88, 888 etc) for

"Not applicable". In STATA a missing value is usually recorded as a full stop/period. When we start to analyse the data we will need to recode the 9's to missing (see Section xxx for how to do this).

The tabulate command can be used to create tables of two variables, and may be restricted by using the qualifier, comparison or logical operators. These operators are listed below:

Qualifying operators		Comparison operators		Logical operators	
if	Qualifies when a command is to be executed	==	Equal to	|	or
		~= and !=	Not equal to	&	and
in	Qualifies which observation should have the command applied to it.	>	Greater than	~	not
		<	Less than		
		>=	Greater than/equal to		
		<=	Less than/equal to		

To generate a frequency table for gender and education, type
`tab Q110HiEd C_Gender`

To generate a frequency table for race and ability to read in at least on language, type
`tab E_Race Q15Read`

Note that we have a two way table with the number of observations in each of the following categories: Africans who can read in at least one language; Africans who can't read; Africans for whom Q1.5 is missing; Coloureds who can read etc. however, the percentages have disappeared. If we are interested in the percentages, we need to specify which type of percentages we want:

`tab E_Race Q15Read, row` will give a frequency table which generates the percentage of the ROW observations which are in each COLUMN category. Thus, 74% of the African observations in our sample can read in at least one language and 9% of Whites cannot read. At the bottom of each column, note that you will see the percentage of the entire data set that fall into that particular column category.

`tab E_Race Q15Read, column` will give a frequency table which generates the percentage of the COLUMN observations which are in each ROW category. Thus, 85% of those who cannot read in our sample are African. At the end of each row, you will find the percentage of the entire data set that fall into that particular row category.

`tab E_Race Q15Read, row column` will give a frequency table which combines the information in the above two tables. You just need to be careful about which percentages you are reading from this table!

`tab E_Race Q15Read, cell` will give a frequency table which gives the percentage of the sample in each possible cell. For example, 8.7% of the sample are Coloured who can read, and 20.5% are Africans who cannot read. The cell option can be combined with the row and column options.

Use the commands you have learnt so far to answer the following:
- What percentage of the sample have no schooling?
- What percentage of the sample over 18 years of age have no schooling?
- What percentage of the female population in the Western Cape have no schooling?
- How many women who are 30 years of age have not completed primary schooling?
- Which province has the largest proportion of men who have no education?

4. GETTING HELP FROM THE HELP MENU, AND CLOSING YOUR LOG FILE

The help command in STATA is a useful way to learn about new commands, or remind yourself how to use commands. For example:

`help log` (tells you how to open and close logs, and why we use log files)
`help browse` (reminds you about the browse command)

At the end of you STATA session, you generally need to do at least 3 of the following 4 tasks:

→ close your log file: type **`log close`**

→ clear your data set: type **`clear`**

→ OR save your data set: using the FILE menu, choose SAVE AS, find a space in your folder and give the data file a name, with file extension '.dta'

→ close STATA, by typing **`exit`**

Now that you have saved and closed your log file let's take a look at it. In MS Word or any other text editor open your log file. You should see something similar to this:

```
-------------------------------------------------------------------------------------------
---------------
       log:  C:\Documents and Settings\Cal\My Documents\Teaching\Survey
data\Labs\lab1.log
  log type:  text
 opened on:  13 Jul 2004, 11:05:16

. set mem 30m

Current memory allocation

                    current                                 memory usage
    settable          value     description                 (1M = 1024k)
    ------------------------------------------------------------------
    set maxvar         5000     max. variables allowed           1.733M
    set memory           30M    max. data space                 30.000M
    set matsize          400    max. RHS vars in models          1.254M
                                                              ----------
                                                                32.987M

. use "C:\Documents and Settings\Cal\My Documents\Teaching\Survey
data\GHS2002\Data\person"
(General Household Survey July 2002 -Person data -Statistics South Africa -Transl)

. describe

Contains data from C:\Documents and Settings\Cal\My Documents\Teaching\Survey
data\GHS2002\Data\person.d
> ta
  obs:       102,461                          General Household Survey July
                                                2002 -Person data -Statistics
                                                South Africa -Transl
 vars:            80                          26 Feb 2004 15:31
 size:    10,451,022 (66.8% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
UqNr            double %13.0f                 Unique Number
PersonNr        byte   %10.0g                 Person Number
Prov            byte   %10.0g                 Province
C_Gender        byte   %10.0g                 1=Male 2=Female
D_Age           int    %10.0g                 Age
```

```
E_Race          byte    %10.0g                Race
```

If you log file looks very different and is hard to read then you probably saved it as an scml file and not a text file. It is very important to save it as a text file so that you can look at the output once your STATA session is complete. If you are not sure how to save a log file as a text file read the section on opening log files at the top of page 3 again. Notice that all commands are preceded by a period. Log files store all the commands AND the output in a STATA session.

5. QUICK INTRODUCTION TO DO FILES

Open Stata and type the following commands

```
set mem 30m
use g:\eco427s\ghs2002\data\person
tab E_Race Q15Read
tab C_Gender
clear
```

Often, we will want to execute commands repetitively in STATA, or we'll want to save the set of commands for a particular procedure so that we can run them on a different data set. We can write simple programs in STATA's DO-file editor which will run a set of commands at once.

As an example, suppose we wanted to repeat the process of looking at the variables above. We can save the commands we have already used using the Review window. Click on the box in the upper left corner of the Review window, and click on "Save Review Contents." Save these to a file in your home directory called "Example.do."

Open the DO-file editor, clicking on the envelope-button on the toolbar. Open the file you just created called **`example.do`**.

In the DO file editor, click on the button with the arrow that says "Do Current File." If you just want to run some of the commands in a DO file then highlight those commands and click on the "Do Current File" button.

You can copy and paste and write new commands in the DO file, and then save the DO file to run again on this data set or other data sets. If you made any mistakes in your Stata session the do file will end and an error message will appear. You will need to edit the do file by correcting or deleting invalid commands.

You can also create a special log file of all the commands that you type in a Stata session. This has an advantage over saving the contents in the Review window as there is a limit to the number of commands that are stored in the Review window. When you start your Stata session type

```
cmdlog using F:\yourlastname_LAB1.do
```

and when you end your Stata session type

```
cmdlog close
```

The DO file is just a text file, and can be created, edited, and saved using Word or any editor, as well as the internal Stata DO editor. Note that if you generate new variables in a DO file, you cannot run the DO file again without dropping those variables or loading the data set again.

6. KEEPING A RECORD OF YOUR WORK WITH LOG FILES

Before you start any lab session, your problem sets or your assignment remember to open BOTH a log file (to record your results) and a cmdlog file to record your commands. If you want to continue from where you left off in a previous Stata session follow these steps:

Re-run the previous session by running your command log that you saved in the previous session:

```
do F:\yourlastname_LAB1.do
```

Then open your existing log and command log files. You will need to specify the **append** option so that all new commands and outputs will be appended to the bottom of the files:

```
log using F:\yourlastname_LAB1.log, append
cmdlog using F:\yourlastname_LAB1.do, append
```

Now you can continue your session.

7. HOW TO GET HELP FOR STATA

There are several ways you can learn more about STATA, or simply refresh your memory about what we have done in class:

- the STATA **help** command
- the STATA website – accessible from STATA, if you click on the HELP pull-down menu, and choose STATA web-site. This web site does tend to be somewhat more technical though.

LIST OF STATA COMMANDS INTRODUCED:

break	browse	browse if
clear	codebook	cmdlog close
cmdlog using	describe	do
exit	help	list
list if	list in	log close
log using	lookfor	tab
tab,cell	tab,column	tab,row
tab,row column	use	