COMPLEX SAMPLE SURVEYS

OBJECTIVES: In this module, you will learn how to:
1. Draw a random sample
2. Use weights
3. Use the **svy** commands
4. Merge household level and individual level data

1. STARTING UP

Before you start this lab session remember to open BOTH a log file (to record your results) and a cmdlog file to record your commands. Remember to create your log file as a text file so that you can easily read it in Word. Stata's default is to save log files as scml files which are very long and impossible to read in a text editor. The simplest way is to open your log file using the command line and specifying a log extension. For example

```
log using F:\yourlastname_LAB3.log
```

Start a command log file by typing:

```
cmdlog using F:\yourlastname_LAB3.do
```

Open up the data set **house.dta**.

2. CREATE YOUR VARIABLES

Generate a dummy variable called *electricity* that equals one if the household is connected to the mains electricity supply and 0 if not. Also generate a dummy variable called *hungry* that equals one if any adult in the household EVER went hungry because there wasn't enough food.

3. DRAWING A SAMPLE

We can use STATA to draw a random sub-sample from our data set. Before we start we must preserve the full data set so that we can restore it at a later stage. Type

```
preserve
```

Generate confidence intervals for the electricity and hungry variables.

```
ci electricity hungry
```

To draw a 10% sample from our data, type:

```
sample 10
```

Look at the confidence intervals for electricity and hungry again.

```
ci electricity hungry
```

Assuming that the full data set is the "population" and therefore the proportion of households with electricity in the full data sets is the population proportion (the parameter). Do your confidence intervals for your 10% sample include the population parameter?

Restore your data set, by typing:

```
restore
preserve
```

Draw another 10% sample and obtain new estimates for electricity and hungry.

```
sample 10
ci electricity hungry
```

Do the estimates differ from the previous sample?

Restore your data set, by typing:

```
restore
preserve
```

Now draw a 50% sample and obtain new estimates for electricity and hungry.

```
sample 50
ci electricity hungry
```

Are the confidence intervals narrower?

Restore your data set, by typing:
```
restore
```

4. WEIGHTS

Each household (person) interviewed for the survey represents some larger group of households (people) in the total population. To make your results representative of the population, you tell STATA to use the weight provided in the data set. STATA uses this weight to weigh some observations more heavily than others. Most STATA commands can deal with weighted data. STATA distinguishes between four types of weights – we concentrate on two of these – frequency weights and probability weights. Frequency weights are weights that indicate the number of duplicated observations – they are inflation weights. Probability weights are the weights that denote the inverse of probability that the observation is included due the sampling design. The weights in the GHS are both probability and frequency weights[1]. Some STATA commands require probability weights while others only allow frequency weights. We will use both. The syntax for producing weighted results is the same in most STATA commands: you specify the weight variable inside square brackets at the end of the command but *before the comma*. Below are some examples:

```
tab E_Race [w=weight]
tab E_Race [w=weight], sum(hungry)
ci electricity [w=weight]
tab Prov [w=weight]
```

Compare these summary details to the unweighted details:

```
tab E_Race
tab E_Race, sum(hungry)
ci electricity
tab Prov
```

Which provinces were oversampled? Which provinces were undersampled?

5. USING THE SVY COMMANDS

STATA has developed a whole range of commands for summarising and analysing complex sample data – the **svy** commands. Type

```
help svy
```

to see a description of these commands.

In order to use the **svy** commands we first have to let STATA know which variables specify the sampling design. We use the **svyset** command to specify the weight variable (house_Wgt), stratum variable (Stratum) and cluster variable (PSU). Type

```
svyset [pw=house_Wgt], strata(Stratum) psu(PSU)
```

We are now ready to use the svy commands. For a summary of the sampling design type

```
svydes
```

Previously we calculated a confidence interval for the number of households with electricity using the **ci** command. We were assuming that the sample was an SRS. We then introduced weights to our **ci** command to adjust for differential probabilities of selection, non-response and post-stratification. While using the weights meant we had unbiased estimates the standard errors were still calculated on the assumption that the sample was an SRS. We use the **svy** command **svymean** to take the complex sample design into account in calculating our standard errors. Type

```
svymean electricity
```

[1] The weight variable is **house_Wgt**. In STATA frequency weights must be an integer so a new weight was created using **gen weight=int(house_Wgt)** for use with commands that require frequency weights.

Notice how the estimated proportion of households with electricity is the same as when we used the **ci** command with weights[2] but the standard errors are quite different. The design effect (deff) of 6.25 is the ratio of the variance calculated using the complex sample design to the variance calculated assuming SRS. There are 26213 households in our sample. The design effect of 6.25 means that due to our complex sample design our precision is only as good as an SRS of 26213/6.25 = 4194.

To examine the impact of the various aspects of the complex sample design on our estimates we will look at each component in turn. Firstly we will clear the survey design variables. Type

```
svyset, clear
```

Let us first look at the affect of weights. Type

```
svyset [pw=house_Wgt]
svymean electricity hungry
```

Look at the design effects. What do they tell you about the impact of the weights on the precision of the estimates?

Next let's examine the strata. Type

```
svyset, clear
svyset, strata(Stratum)
svymean electricity hungry
```

Notice how the design effects are now less than 1 indicating a gain in precision. For which variable is there a larger gain? Why?

Third let's examine the effect of clustering. Type

```
svyset, clear
svyset, psu(PSU)
svymean electricity hungry
```

What are the design effects? For which variable is there a larger loss in precision? Why? Can you think of examples of other variables where clustering would have a large impact?

Finally let's put it altogether

```
svyset, clear
svyset [pw=house_Wgt], strata(Stratum) psu(PSU)
svymean electricity hungry
```

If we want to look at the frequency distribution of households among races and provinces type

```
svytab Prov E_Race
```

Suppose we wanted to estimate the total number of households in South Africa with electricity. Then we would type

```
svytotal electricity
```

Assuming we had an SRS design with epsem sampling (and no non-response etc.), what would the inflation weight per household be? Using this weight how many households would you estimate have electricity? What is the difference?

6. MERGING HOUSEHOLD LEVEL AND INDIVIDUAL LEVEL DATA

Most household surveys have questions that are asked at the household level and questions that are asked about each member of the household (individual level). Household level and individual level data are normally stored in separate files. Sometimes we want to do an analysis at the individual level but would like to use some household characteristics in our analysis. In the previous lab we ran regressions of years of completed education on age, sex and race. We may want to include some household level variables such as rural/urban or even access to electricity. To do this we would need to merge the individual level and household level files. In order to merge files you need to use an identifier that is common to both files and unique in at least one. In the GHS data set the field **UqNr** is in both the household level and

[2] There is a slight difference as the **ci** command requires integer weights.

individual level data and is a unique number for each household. We will use this field to merge. Before we merge we need to make sure that both files are sorted on the linking field.

Open up the data set **house.dta**.
Type:
```
sort UqNr
save f:\house.dta, replace
```

Open up the data set **person.dta**.
```
sort UqNr
merge UqNr using f:\house
```

The two data sets are now merged. If you scroll down the variable window you will see that all the person level variables appear first followed by the household level variables. When we merge two files Stata creates a new field named **_merge**. Let's look at this field.
```
tab _merge
```

Use the Stata help to find out what the codes mean.

Let's practice merging by combing the person and worker files. Individuals in the GHS are uniquely identified by their household number, **UqNr**, and their person code, **PersonNr**.

Open up the data set **worker.dta**.
Type:
```
sort UqNr PersonNr
save f:\worker, replace
```

Open up the data set **person.dta**.
```
sort UqNr PersonNr
merge UqNr PersonNr using f:\worker
```

Check the merge
```
tab _merge
```

Make sure that you understand what the codes mean.

LIST OF NEW STATA COMMANDS INTRODUCED IN MODULE 3

ci	merge	sample
svydes	svymean	svyset
svytab	svytotal	