**Stata Commands for Session (1) Intro to Stata**

**TIPS**

* Syntax elements within square brackets [ ] are optional – they extend or change the default setting of the command. The square brackets are **not** part of the code, do not type them.
* *var* means “variable”. So, v*arlist* means ‘enter list of variable names here’ and varname means ‘enter variable name here’. If you do not provide *varname* or *varlist* Stata’s default is to, where applicable, execute the command on **all** variables in the dataset and, where not applicable, the command will not work.
* stata code/syntax is case-sensitive. Commands should be in lower case. If a command isn’t working, check you haven’t used capital letters.
* Within command lines, if listing more than one variable, separate variable names with a space.
* To annotate your do files, type an Asterix at the start of the line of text. (the text will turn green). If you don’t do this, Stata will try to treat your notes as commands.

**BASIC COMMANDS**

1. **To set the working directory:**

cd “filepath\foldername”

1. **To import a .csv file into Stata:**

import delimited “filename.csv”

1. **To save a stata dataset (.dta file):**

save “filename.csv” [, replace]

1. **To load/open a stata dataset (.dta file):**

use “filename.csv” [, replace]

1. **To clear the console (i.e., remove dataset from stata memory):**

clear

1. **To see how many observations are in the dataset:**

count

1. **To get a quick overview of variables in the dataset:**

describe

describe *varlist*

codebook

codebook *varlist*

1. **To label a variable:**

label variable*varname* "varlabel"

1. **To label the values of a categorical variable (note, both commands are needed):**

label define *labelname* 0"*label*" 1"*label*" 2”*label*”

label values *varname* *labelname*

1. **To obtain basic summary statistics (e.g., mean, sd) for continuous variables:**

summarize *varlist*

summarize *varlist*, detail

1. **To obtain summary statistics (i.e., frequency and %) for categorical variables:**

tabulate *varname*

tabulate *varname1 varname2*

tabulate *varname1 varname2 [, row col]*

1. **To generate a new variable**

generate *newvarname = expression*

1. **To recode values within a variable:**

replace *varname* = *expression*

1. **To replace contents of a variable:**

recode *varname* (*rule*)

1. **To restrict commands to subsets of the data, use ‘if’ or ‘in’ after the command. e.g., to obtain the frequency and percentage of boys and girls in ethnic group 1:**

tabulate sex if ethnic==1

1. **To repeat commands over subgroups, type this before the command:**

bysort *varname*:

e.g., to obtain the mean mental health difficulties score (sdq) for boys and girls separately, type:

bysort sex: summarize sdq

1. **To drop a variable:**

drop *varname*

**EXTENSION (2)**

**Practical**

1. How many observations are in the REACH dataset? How many observations (participants) are there within each of the three schools?
2. What is the mean mental health difficulties (sdq) score among those receiving free school meals (low SES)? And among those not receiving free school meals (high SES)?
3. The mental health difficulties (sdq) score can be categorised into 4 groups:

Close to average: 0-14

Slightly raised: 15-17

High: 18-19

Very high: 20-40

Create a new variable called sdq4, which groups participants according to these thresholds.

Label the new variable.

Label the values of the new variable.

What percentage of participants are in each of the four groups?

1. Plot a histogram of sdq. Are the scores normally distributed? (hint: use the drop-down menu to find the command)
2. Explore the conduct variable. Is it a categorical or continuous variable? Obtain the appropriate summary statistics for this variable (i.e., mean and sd or frequency and percentage). How many missing values are there for the conduct variable?