Statistics for Mental Health Research

**Introduction to Stata, Data Preparation   
(Practical 1)**

# Introduction to Stata

The data set we are going to use in this practical session, **practice1.dta**, contains data on the positive symptom scores of people with schizophrenia. It uses variables from INTREPID II study but all information is randomly generated. The variables are as follows:

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| sub\_id | Participant id number |
| age | Age in years |
| DOB | Date of Birth |
| gender | Gender (1 = Male; 2 = Female) |
| ethnicity | Broad ethnic groups |
| site\_scr | Site where the participant is screened |
| case\_control\_rel | Participant Type |
| source | source of info collection |
| panss\_p | PANSS Positive Subscale Score |
| include | whether would include sample in analysis |

Create a folder for the statistics module. Download the practical files from Email and save them to the folder you have just created.

1. Open Stata.
2. Open a Do file and save it to your statistics module folder.

**Navigate the operation interface of STATA**

1. Change Stata’s working directory to the same folder using the following command as an example:

**cd “C:\Users\Joseph\Desktop\INTREPID MRC WORK\Teaching Fellows\STATA Teaching Materials\Session 1”**

**cd = change directory**

1. Open the Stata data file **practice.dta** using the following Stata command:

**use practice1.dta , clear**

**(or double-click the .dta file to open)**

1. Explore the data using the **browse**, **describe**, **summarize** and **codebook** commands.
2. Display the data for people who are female using the following Stata command:

**list if gender == 1**

1. Display the data for people who have Tamil ethnicity using the following Stata command:

**browse if ethnic == 1**

1. List the data just for people who have Igbo ethnicity.
2. Make sure you have a record of all the key commands you used in your Stata Do file and then save it.
3. In Stata the double equals sign (**==**) means equals to. The sign (!=) means not equal to. These are the operators which can be used with **if** in Stata. Use the different ways of obtaining help in Stata to find out which other operators can be used with **if**.

# Preparing data for analysis

1. Create a new binary variable for friends to indicate whether participants have missing age

**generate age\_miss = .**

**replace age\_miss = 1 if age ==.**

**replace age\_miss = 0 if age !=.**

**(Think of other codes that could achieve the same result)**

1. Browse the data to see what effect the above code has had.
2. Create a variable label for the new binary friends variable.

**Label define age\_bi 1”Missing Age”0”Not Missing Age”**

**(why didn’t it work? Remember STATA is case-sensitive!)**

1. Create and apply a value label for the new binary friends variable.

**lab val age\_miss age\_bi**

**lab var age\_miss “missing age”**

1. Save the changes you have made to the data set.
2. Find out what data is missing in gender variable.

# Replace and IF

1. Use codebook to learn about the include variable

2. We would **only include cases and controls** if their **panss positive score is higher or equal (>=) to 14.** Write the code to mark the included participants as “1” in include variable, and “0” if to be excluded.

**Replace include = 1 if**

3. We noticed some variables have missing data. Find those variables.

**Age and gender**

4. We would only like to include participants with no missing data in the above variables into the analysis. Please make those changes. Generate a new variable called “nomiss”.

5. We would like to investigate a sub-group of data, where only participants between the age of 18 and 24, that has panss positive score over 30. Please create a variable to mark these participants, using the variable name “include\_2”.

Extra Challenges

1. Try to use their DOB to calculate their current age.
2. Learn about the “tabulate” function, using help.
3. Categorise age into different groups (18-24, 25-34, 35-44, 45-54, 55-64), and tabulate the frequency and percentage of each group in the included sample.

Credits: This tutorial is modified from the notes by Rebecca Jones (2018-2019 UCL MSc Statistics in Mental Health Module).

Disclaimer: This tutorial is created for internal use only. Please do not share the data\_file or this tutorial with unauthorised personnel.