Guide to the Python Program, subsample_iid_test_setup.py

This program was written by Lucy Wan for the extraction of subsamples of simulated data in Section 6 of the paper by Birnbaum and Wan (2020). Questions about this guide and the code should be sent to Lucy Wan at the following email:
lucy.w@berkeley.edu

The program can be downloaded from the SJDM Website as a text file with extension of ".txt". You should change the extension to ".py" before using the program.

Anaconda Distribution

I recommend using the Spyder Integrated Development Environment (IDE) of the Anaconda Distribution of Python. (Other installations of Python should also work, such as that available at https://jupyter.org/try)

Use this link to install the recommended Anaconda Distribution:

https://www.anaconda.com/distribution/#download-section

Install Python 3.7

Working Folder

Now create a folder for the program and input files. You might name this folder, for example, subsample_results. Place the following 3 files in this folder:

subsample_iid_test_setup.py,
test_iid.txt,
MARTER_data.xlsx.

Note that for this purpose, the text file that contains the R program, test_iid.txt, has an extension of .txt. This Python program will revise that program and rename it with an extension of .R.

Open the Anaconda Navigator App and the Spyder IDE. Make sure all the input files are in the same folder and set it as the working directory. In this case, you can set the working directory by clicking on the folder icon in the upper, right corner of the Spyder window and choose the name of the folder, subsample_results.

Using subsample_iid_test_setup.py

A person could manually extract subsets of data from the data file created by MARTER_sim.htm. First, use the "text to columns" feature of Excel to separate the two response patterns in each row of the original data into 6 individual responses (6 columns), and then copy and paste blocks of responses from Excel into a text editor such as Notepad, and then save them individually to create separate files for each simulated subject. However, this tedious process can be automated by means of the Python program called *subsample_iid_test_setup.py*.

The program, subsample_iid_test_setup.py, takes results of MARTER_sim.htm, which have been saved in an Excel file, to create text files representing simulated "subjects" and it creates the corresponding R scripts to run in R, with test_iid.R, The Excel file of data in this case is *MARTER.data.xlsx*. The program, test_iid.R, computes two statistical tests of independence; it is described in Birnbaum (2012) and discussed further in Birnbaum (2013).

The program generates folders which include text files of the subsampled data in a format which allows it to be run using the modified version of test_iid.R program. These folders will automatically be labeled with the generating model name and the number of reps and subs. Within each folder will be the appropriate number (n_subs) of text files with each text file representing a consecutive portion of the data in the Excel files.

The program runs in Python 3.7 and the Python packages numpy, pandas, os, and shutil are needed for the program to run. Numpy and pandas should already be installed since you installed Anaconda but os and shutil need to be manually installed. To install packages, refer here: https://docs.python.org/3/installing/index.html.

It also runs on an Excel file with each sheet containing the data of a model. The data in the Excel file must be in two columns and be headed with "response1" and "response2". In other words, the first cell of the first columns must read "response1" while the first cell of the second column must read "response2". This is what you get if you copy data directly from MARTER_sim.htm into Excel and use the "Text to Columns" feature to create two columns of the response patterns.

Note: Examples of the Excel input file and the code to run the program are below.

Running the Program with Spyder

1) Assign Variables: The variables that need to be inputted into the program are excel_file, lis_model, lis_reps, n_subs, and lis_start.
excel_file is the name of the Excel file where all the data are stored.
lis_model is a list of the worksheet names of the Excel files that you would like to analyze. In this case, the worksheet names correspond to the generating model names, "Trans1", "Trans2", etc.
lis_reps is a list of the numbers of blocks of data that you would like to extract for each "subject".

n_subs is the number of simulated “subjects” you would like to create.
lis_start is a dictionary which indicates the locations of the starting points for the extracted values of each folder, which are created with names that indicate the worksheet, the number of blocks of data per subject, and the number of extracted subjects; e.g., Trans3_rep_20_subs_20. In the example below, the first element of the value of lis_start corresponds to 10 reps, the second corresponds to 20 reps, and the third corresponds to 100 reps. In this example below, the first line of data sampled is line 5001 for Subject No. 1 with 10 blocks of data (lis_reps = 10). Also make sure to add test_iid.txt to your working directory.

Then run the following code to assign variables, as in the following example:

```
excel_file = 'MARTER_data.xlsx'
lis_model = ['Trans1', 'Trans2', 'Trans3', 'Intrans1', 'Intrans2', 'iid_data']
lis_reps = ['10','20','100']
n_subs = 20
lis_start= {"Trans1":[5001,5201,5601], "Trans2":[5001,5201,5601],
"Trans3":[5001,5201,5601], "Intrans1":[5001,5201,5601], "Intrans2":[5001,5201,5601],
"iid_data":[5001,5201,5601]}
```

2. Import and call the program: Copy the following code into the editor window of the Spyder (left) to import

```
import subsample_iid_test_setup as sits
sits.subsample_iid_test_setup(lis_model, lis_reps, n_subs, excel_file, lis_start)
```

3. Now select the code with your cursor and press “shift” and “enter” to run the code.

Example files used in this program:
MARTER_data.xlsx
subsample_iid_test_setup.py
test_iid.txt
(The text file version of test_iid.R with specific adjustments to work in subsample_iid_test_setup)

4. When the program has run successfully with this example setup, you should see 18 folders with names that match the datasets (names of the worksheets of data in the Excel files) and include the number of reps (blocks of data extracted per subject) and the number of subjects. Inside each folder, there will be 20 separate text files showing the extracted data for the 20 “subjects”, and there will also be a suitably set-up version of iid_test.R that is ready to use. In order to run the small sample tests of iid, one must then open R, set the workspace to one of these folders, and source (i.e., run) the iid_test.R program that is in each folder. See Birnbaum (2012) for more detail on that program.