KnowHow
Learning Resources for the KnowFife Partnership Hub

Knowledge
Learning
Skills
Best Practice
Researchers' Alliance

SPSS File Operations

Contact: kerry.humphries@nhs.net Version: 1.0 Date: March 2018

SPSS is a powerful Statistical Package for the Social Sciences. This guide focuses on the file operations of merging and aggregating datasets within SPSS

Merging Datasets

Datasets can be merged to-:

- **Add Cases** (ADD FILES) **–** adds two or more datasets with the same variables.
- **Add Variables** (MATCH FILES) – adds variables from another dataset through matching on a key variable.

Add Cases

This operation is useful for joining the same data for different time periods or when two areas record the same data operationally and reporting requires the combined dataset.

With the dataset you wish to add cases to active, select from the Data menu > Merge Files > Add Cases to open the add cases dialog box.

- Select the open dataset or external SPSS sav file (not open) that you wish to join with your current dataset. Continue.
- SPSS will have paired the variables with the same name, data type and character length. You can rename a variable by clicking it and then clicking 'rename' then you can send it to the 'variables in the active dataset'. If you discover variables with different types or character lengths cancel the merge files and amend the details to match in the data editor- variable view before repeating the above actions.
- Once the 'New Active Dataset' is prepared you have the choice to action immediately or paste. If this function will be required as part of a regular process it will save time to build up a program of syntax that actions all of the procedures and will only require minimal amendments prior to running in future.

- Check your active dataset now contains the records from the original file as the top rows with the added records in the rows below.

Add Variables

This operation combines datasets in the same was as joins work for databases, choosing a one to one match or a one to many match, this operation requires the datasets you wish joined to have the same key field or for one to one matches two datasets with the same number of records in the same order may be joined without a key. It is useful when data stored about the same items, using the same ID are saved in separate files and this operation is also used as a file 'lookup' to add extra variables onto a record. It should be used with care, as not all variables that can be joined should be joined e.g. A file has datazone, Interzone and Locality you match SIMD quintile via the key variable datazone. The SIMD quintile only relates to datazone and doesn't give you SIMD of Interzone or Locality.

With the dataset you wish to add cases to active, select from the Data menu, Merge Files > Add Variables to open the add variables dialog box.

- Select the open dataset or external SPSS sav file (not open) that you wish to join with your current dataset. Continue.
- Check your excluded variables (* =Active dataset, + =other dataset) if you want to include any of these, rename them and send to the 'variables in the active dataset'. Move any variable you wish to exclude from the active dataset to the excluded list.
- If matching on a keyed variable(s), click to check the match cases on a key variable. The same variable(s) with the same name must exist in both datasets, rename where required and send the variables from both datasets to the key variables box via the excluded variables box.
- If not matching on a key variable, both datasets require to have the same number of records in exactly the same order, leave the match cases on a key variable blank.
- The keyed table is the 'lookup' dataset and must only have once case per keyed ID or IDs. Ensure the correct dataset is checked as the keyed table.
- If the datasets are already sorted you may check the box, this is particularly useful in saving time with very large datasets.
- Now choose whether to action immediately or paste. If this function will be required as part of a regular process it will save time to build up a program of syntax that actions all of the procedures and will only require minimal amendments prior to running in future.

Aggregating Data

Data can be aggregated to create a new dataset, add new variables to the existing dataset or, replace the existing dataset.

This is used when you have a file where the individual records have a field within the record that they can be grouped together with e.g. a file with datazone and interzone and you wish to see the data at the grouped level.

With the dataset you wish to add cases to active, select from the Data menu > Aggregate to open the aggregate data dialogue box.

- Select the Break variable or variables in the correct order sending the field(s) to the 'Break variable(s), in the example above this would be interzone as it's the field you wish to group on.
- Select the fields you wish to see at the aggregate level and send them to 'Summary of variable(s)' or if you just wish a count of cases per each category within the new grouping tick the 'Number of cases' check box instead. You can rename the field anything you wish, N_BREAK is the default.
- When you wish summary variables, in most cases you will need to click the function box and change how you wish each variable summarised, e.g. mean, sum, first etc. You can highlight all the variables and change them all to the one function if this suits your data.
- In the save section select your preferred option for your aggregated data, if you choose to add to the active dataset, keep in mind the data will still be at the lower level but you will have aggregated variables tagged onto each record. Remember to type a new dataset name is that's what you choose. When writing to a new data file, click File to browse to the preferred location and type the filename. The file will not automatically open for use.
- Select whether the file is already sorted on the break variable(s) or requires sorting.
- You now have the choice to action immediately or paste. If this function will be required as part of a regular process it will save time to build up a program of syntax that actions all of the procedures and will only require minimal amendments prior to running in future.

Further information

- IBM SPSS Statistics 24 Brief Guide
  ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/24.0/en/client/Manuals/IBM_SPSS_Statistics_Brief_Guide.pdf
- SPSS for Dummies
  http://www.academia.dk/BiologiskAntropologi/Epidemiologi/PDF/SPSS_For_Dummies_2ndEd.pdf
- SPSS Survival Manual
  https://www.mheducation.co.uk/openup/chapters/0335208908.pdf