What is SPSS 20 for Windows?

UK Data Service
Author: UK Data Service
Updated: May 2014
Version: 2.0

We are happy for our materials to be used and copied but request that users should:

- link to our original materials instead of re-mounting our materials on your website
- cite this as an original source as follows:

Contents
1. What is SPSS 20 for Windows? 2
2. Viewing data and output in SPSS 2
   2.1 Data View and Variable View 2
   2.2 Changing settings in SPSS 5
   2.3 The Statistics Viewer 6
3. Exploratory analysis in SPSS 20 7
   3.1 Weighting your data in SPSS 7
   3.2 Creating a one-way frequency table and bar chart 7
   3.3 Dealing with missing values 10
   3.4 Filtering the data to select certain cases 12
   3.5 Comparing two variables 14
   3.6 Graphing two categorical variables 16
4. Data manipulation in SPSS 21
   4.1 Recoding variables 21
   4.2 Computing new variables 24
5. Using hierarchical data in SPSS 27
   5.1 Selecting one individual per household 27
   5.2 Summarising characteristics of groups in hierarchical data 31
   5.3 Attaching household data to an individual level file 34
6. Linking and merging files in SPSS 39
   6.2 Linking multiple files at the same level of measurement 39
   6.3 Attaching household level data to individuals 43
   6.4 Merging files with different cases but the same variables 43
1. What is SPSS 20 for Windows?

SPSS 20 is a software package for data analysis. Because SPSS is a Windows programme, you can view and edit your data by pointing and clicking the movable windows, dropdown menus and dialogue boxes. SPSS 20 can be used to enter, manipulate and analyse data. You can also use SPSS 20 to produce graphics of your data.

2. Viewing data and output in SPSS

The dataset used in the following examples is the British Social Attitudes Survey, 2011. The British Social Attitudes Survey asks over 3,000 people what it’s like to live in Britain and what they think about how Britain is run. This dataset can be downloaded from the UK Data Service website, after completing a short registration.

2.1 Data View and Variable View

Datasets in SPSS 20 are most commonly saved as .sav SPSS data files. Open SPSS 20 and use File > Open to open your dataset. There are two ways to view the data: in the Data View or the Variable View tabs. The tabs to switch between the two are at the bottom left of the screen.
Click on the **Data View** tab (bottom left of screen). In **Data View**:

- Each column represents a variable in the survey. This is often a response to a question or derived from answers to a question or several questions. Hovering your cursor over the variable name at the top of a column will show the longer variable label.

- Each row represents an individual respondent, this might be a person (as in the British Social Attitudes Survey), or a household, or family unit, or other unit. These rows are often referred to as **cases** (or **observations**).

- It is common for datasets to have a unique identifier towards the beginning of the dataset, often the first column. It is possible to reorder the data in SPSS, so a unique identifier is required to identify particular cases, even if the data has been reordered. In the 2011 British Social Attitudes Survey, the data is uniquely identified by “Serial Number :Q1” (Serial), which is in the first column.

Click on the **Variable View** tab (bottom left of screen). In **Variable View**:

- Each row represents something that varies between respondents (known as a variable) and each column provides information about the variable including the name, label and coding information in the **Values** column.
In **Variable View**, you can see more information about the type of variable, the name (the short name used to identify the variable), the label (the longer name, often the full question from the survey questionnaire), whether it is numeric, or string (text variables), or a date, for example.

In the **Values** column, you can see the values that have been assigned to the different values of a variable, for example which values represent which area in the variable “Government office region 2003 version: Q18” (GOR 2).

The **Missing** column shows any values which have been allocated as missing in SPSS. Any cases with missing values will not be included in the analysis.

If you want to view the values or the missing values of a variable from the **Variable View**, clicking on the right hand of the values you want to see will bring up the **Values** window.

**Measure** shows whether a variable is *scale*, such as people’s age, *nominal*, an unranked categorical variable (e.g. Country), or *ordinal*, a ranked categorical variable (e.g. a Likert scale of agreement)

1 – Agree strongly
2 – Agree
3 – Neither agree nor disagree
4 – Disagree
5 – Disagree Strongly

You can also find out more about the variables by using the **Variables** window. Use **Utilities > Variables...**, or click the **Variables** tool button ![Variables tool button](image) to open the following dialogue box:
Click on any variable in the left-hand source-list to see more information, including the Value Labels. These are responses to questions. When viewing your data in Data View, you can click on the Value Labels tool button to toggle between the numerical value of a variable and the Value Labels.

2.2 Changing settings in SPSS
Before you start analysing data, it is worth understanding some of SPSS’s settings. If you are seeing the Label of the variables in the left-hand column of the Variables window (or other dialogue box), and would prefer to see the Names, you can change this via Edit > Options and the following window. Just ensure that Display names is checked, and then click Apply then OK.
2.3 The Statistics Viewer

In both Data View and Variable View, you can use the dropdown menus at the top of the page (i.e. File, Edit, View, Data, Analyze, Graphs etc.) to manipulate the data and to do analyses.

When you conduct an analysis, you see the results in the Statistics Viewer window, initially titled Output 1. In the Statistics Viewer, the pane on the left summarises the analyses conducted and the pane on the right contains the results.

You can move between the Statistics Viewer window and the data in Data View or Variable View, by clicking on the tab in the taskbar (the bar that is always at the bottom of the screen).

You can also click on the Go to databutton to return to the data from the Statistics Viewer.
3. Exploratory analysis in SPSS 20

The dataset used in the following examples is the British Social Attitudes Survey 2011. The British Social Attitudes Survey asks over 3,000 people what it’s like to live in Britain and what they think about how Britain is run. This dataset can be downloaded from the UK Data Service website, after completing a short registration.

3.1 Weighting your data in SPSS
Before you begin your analysis, you will need to weight the data to ensure that your sample is representative of the population. See our What is weighting? guide for further information on weighting datasets. In the 2011 British Social Attitudes Survey, the “Final BSA weight:Q25” is WtFactor. To apply this, use Data > Weight Cases, and select Weight cases by, then select WtFactor and move it into the box on the right by clicking on the arrow, and clicking OK.

3.2 Creating a one-way frequency table and bar chart
From the menu bar at the top of the page, use Analyze > Descriptive Statistics > Frequencies. From the Frequencies screen, select the variable of interest from the list. In this example, we are using “Marital status <5 categories> dv :Q145” (MarStat). You can scroll through the list to find the variable you want, or you can click on any variable in the list and type the first few letters of the variable you want. This will select the next variable on the list that starts with those letters. This may be the variable you want, or there may be others with the same starting letters that come first. Continue to type in the first few letters of the variable you want until you find the correct variable. Move your variable into the right hand box by clicking on the arrow. Click OK or Paste.
Note that if you would like to view the variable Labels rather than the Names you can right click on any variable in the dropdown list and select Display Variable Labels.

If you click OK, SPSS runs the command as requested. If you click Paste, this opens the Syntax Editor window. Clicking Paste saves the instructions for the command in a syntax file (syntax is the name of the computing language that can run commands in SPSS without using the menu). Saving this syntax allows you to re-run your analysis more easily in future. This is useful if the data changes, or if you make a mistake in your data manipulation or analysis.

Each command in the Syntax Editor window ends with a dot ".". To run the most recent command, select all of the most recent command and press the large green arrow button. To run all of the analysis, select all (Control-A) and then press the green arrow.
It is useful to keep the **Syntax Editor** window open so that you can go back to it after every command and run that command. You can also save the syntax to use again or make changes. Experienced SPSS users may find syntax quicker and easier than the menus, as they have learned the programming language. For example, the above command can be typed in as:

FREQUENCIES MarStat
You can access help about any syntax command by using the syntax help button within the **Syntax Editor**

Running your command/s will open the **Statistics Viewer**. This will display your frequency table:

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td>1676</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
</tr>
<tr>
<td>Married</td>
<td>1676</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
</tr>
<tr>
<td>Living as married</td>
<td>368</td>
<td>11.7</td>
<td>11.7</td>
<td>62.3</td>
</tr>
<tr>
<td>Separated or divorced after marrying</td>
<td>321</td>
<td>9.7</td>
<td>9.7</td>
<td>72.0</td>
</tr>
<tr>
<td>Widowed</td>
<td>224</td>
<td>6.8</td>
<td>6.8</td>
<td>78.8</td>
</tr>
<tr>
<td>Not married</td>
<td>700</td>
<td>21.2</td>
<td>21.2</td>
<td>100.0</td>
</tr>
<tr>
<td>Don’t know</td>
<td>1</td>
<td>0.0</td>
<td>0.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Refusal</td>
<td>1</td>
<td>0.0</td>
<td>0.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>3311</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

To view the variable as a graph, return to the **Frequencies** dialogue box. A shortcut to opening a previously used dialogue box is by using the **Dialogue Recall** button on the toolbar and clicking on the dropdown command:

Click on the Charts tool button. You can select a Bar chart or a Pie chart. In this example, a bar chart shows the frequencies of the variable. Click **Continue**, then **OK** or **Paste** from the **Frequencies** dialogue box.
Running the command will display the bar chart as output in the **Statistics Viewer**.

Double-clicking on the chart in the **Statistics Viewer** opens the **Chart Editor**, which allows you to make changes in the way your graph is displayed. There are many possible amendments that can be made, including labelling changes, changing the colour of the chart or the order of the bars.

### 3.3 Dealing with missing values

It may be preferable to perform your analyses to produce valid percentages; a percentage that excludes certain responses that are not of interest, such as “unknown” responses. You can do this by setting these responses to missing.
The frequency table and graph above show two responses “Don’t know”, and “Refusal”. It may be preferable to exclude these responses from analysis, by setting them to missing. Missing values are set in the **Variable View**. Go back to the data, either via the windows at the bottom of the screen, or via the Go To Data Button 🔄. Click the **Variable View** button:

![Variable View](image)

Use the **Variables** tool button to find the variable you want to modify, and note down the values of the responses you want to set to missing. Using "Marital status <5 categories> dv :Q145” (MarStat) in this example, the values to set to missing would be 8 = “Don’t know” and 9 = “Refusal”. Open the dropdown menu in the **Missing** column of the variable. This opens the **Missing Values** window. Click **Discrete missing values** and enter the values to set them to missing:

![Missing Values](image)

After setting these values to missing, running the frequency table shows the missing values separately, and reports a **Valid Percent**, which is the percentage based on only the valid (non-missing) data:

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>1676</td>
<td>50.6</td>
<td>50.6</td>
<td>50.6</td>
</tr>
<tr>
<td>Living as married</td>
<td>598</td>
<td>11.7</td>
<td>11.7</td>
<td>62.4</td>
</tr>
<tr>
<td>Separated or divorced after marrying</td>
<td>321</td>
<td>9.7</td>
<td>9.7</td>
<td>72.1</td>
</tr>
<tr>
<td>Widowed</td>
<td>224</td>
<td>6.6</td>
<td>6.6</td>
<td>78.9</td>
</tr>
<tr>
<td>Not married</td>
<td>700</td>
<td>21.2</td>
<td>21.2</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>3311</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Don’t know</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Refusal</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3311</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>
There are some variables in the 2011 British Social Attitudes Survey with missing values already assigned. For example, the variable “Did you have to retire because of your employer’s policy on retirement age? :Q337” (RRetPlyc), has the values -4 = “Skip, never worked” and -1 = “skip, not retired” set to missing. These will not show up in analysis.

Some datasets may have no information available for a variable for a particular case, in which case SPSS defaults to system missing”. It is preferable to assign a value to missing information so its status is clearly defined, as with the variable RRetPlyc, where the information is missing for distinct reasons. You can use any value, but minus values stand out as different and are commonly used.

3.4 Filtering the data to select certain cases
It may be that you wish to analyse subgroups of your data. In order to do this, you can filter your data to just include the groups of interest. This means that your analysis will use responses by some and not by others. For example, you may be interested in the responses by people who are married. Filtering does not delete data from the dataset; it just removes them from analyses while the filter is on.

From the menu bar at the top of the page, use Data > Select Cases. Select If condition is satisfied and click on the If button:
Then click on "Marital status <5 categories> dv :Q145" (MarStat) and use the arrow to move it to the box at the top, then type "= 1", and press Continue, then OK. This selects only the responses for individuals who are married.

You can see that the filtering has worked by looking at the data in the Data View. The rows with lines through the numbers will not be included in the analyses.

It also says "Filter On" near the bottom right of the SPSS Data View and Variable View screens.
You should also run a frequency table of MarStat to see if the filtering has worked correctly.

Don’t forget to remove the filter if you wish to analyse the entire dataset, use **Data > Select Cases > All cases.**

### 3.5 Comparing two variables

Cross-tabulations show the relationship between two (or more) categorical variables. For example, you may be interested in frequency of local bus travel by car drivers. To run this cross-tabulation, use **Analyze > Descriptive Statistics > Crosstabs** Select “How often nowadays do you usually travel by local bus:Q405” (TRAVEL3) as the row variables, and “May I just check, do you yourself drive a car at all these days? :Q396” (DRIVE) as the column variable.

![Crosstabs](image)

Click on the **Cells** button and select **Column Percentages.** Click **Continue**, and **OK.**
The first row in this cross-tabulation can be interpreted as 2.7 per cent of people who drive, travel by local bus every day or nearly every day, in comparison to 19.9 per cent of people who don't drive. If you were only interested in people who travelled by local bus at any point, you may wish to set the response “Never nowadays” and “Don’t Know” to missing.

If the crosstab is run to show the Row Percentages rather than the Column Percentages, the table would be as follows:
The first row of data in this table shows that 23.3 per cent of people who travel by local bus every day or nearly every day are drivers, and 76.7 per cent are not. This is based on the same information as the table with column percentages. Whether you choose row or column percentages will depend on your research question.

### 3.6 Graphing two categorical variables

A clustered bar chart is often used to visually illustrate the relationship between two categorical variables. Use **Graphs > Chart Builder**. This will open a dialogue box giving you the option to set your measurement level (i.e. nominal, ordinal, scale). If you are confident that your variables are correctly defined, press **OK**. At this point, you can also choose to tick the box **Don’t show this dialog again**.
This will open the **Chart Builder** window:

To create a bar chart, ensure that Bar is selected from the Gallery tab in the **Chart Builder**. The **Chart Builder** works on a drag and drop basis. From the Gallery tab, select the chart type you are interested in (in this instance the clustered bar chart), and drag and drop it onto the main window. This will open the **Element Properties** window.

To visually represent the data of local bus travel by car drivers, select “How often nowadays do you usually travel by local bus:Q405” (TRAVEL3) from the Variables menu in the **Chart Builder** and drag it across to the X-Axis. To cluster this by whether people drive a car or not,
select “May I just check, do you yourself drive a car at all these days? :Q396” (DRIVE) from the Variables menu, and drag this across to Cluster on X: set color. Within the Element Properties window, under Statistic, select Percentage(), and click Set Parameters...

This will open the Element Properties: Set Parameters window. For this data, it makes sense to set the Y-Axis as showing the percentage of people in each response to the bus travel question. To do that, select Total for Each Legend Variable Category (same fill color) from the dropdown menu and click Continue. You can also use the Element Properties window to set Axis labels. Click Apply in the Element Properties window and then on OK to run the graph.
Depending on what you want your data to show, you could choose to transpose these two variables in the **Chart Builder**. You could show separate bars for frequency of bus travel, clustered by car driving, which would make your graph look very different:

**Chart Builder** can be used for a variety of graphs. The table below shows a summary of some basic analyses and graphs that can be used to summarise a single variable and then the relationship between two variables of different measurement levels.

<table>
<thead>
<tr>
<th>Level of measurement variable</th>
<th>To summarize</th>
<th>With a categorical: nominal or ordinal variable</th>
<th>With a scalar/continuous variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical: nominal</td>
<td>One-way frequency</td>
<td>Cross-tab (two-way)</td>
<td>Mean (average)</td>
</tr>
</tbody>
</table>

ukdataservice.ac.uk
<table>
<thead>
<tr>
<th>Type</th>
<th>Table</th>
<th>Frequency Table</th>
<th>Histogram</th>
</tr>
</thead>
<tbody>
<tr>
<td>or ordinal</td>
<td>Pie chart</td>
<td>Clustered bar chart</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bar chart</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Median for ordinal variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mode for nominal variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scalar/continuous</td>
<td>Mean (average)</td>
<td>Mean (average)</td>
<td>Scatterplot</td>
</tr>
<tr>
<td></td>
<td>Histogram</td>
<td>Histogram</td>
<td></td>
</tr>
</tbody>
</table>
4. Data manipulation in SPSS

The dataset used in the following examples is the British Social Attitudes Survey, 2011. The British Social Attitudes Survey asks over 3,000 people what it’s like to live in Britain and what they think about how Britain is run. This dataset can be downloaded from the UK Data Service website, after completing a short registration.

It is often necessary to derive a new variable for analysis by grouping certain categories of existing variables together (recoding variables), or combining information from more than one existing variable (computing new variables).

4.1 Recoding variables

If you were interested in whether people thought the standard of care in the NHS had improved or declined in the last five years, you may wish to combine some of the categories of the variable "how much better or worse the general standard of health care on the NHS over the last five years? :Q363" (NHS5yrs).

Select **Transform > Recode into Different Variables**... You have the option to recode into the same variable. However, this means that you cannot check your recoding against the original variable. It is advisable to recode into different variables so that you can check your work and recode again if required. This opens a new window. Select NHS5yrs from the variable list, and move it across into the **Numeric Variable** box using the arrow. Type in the name of your new variable in the **Output Variable Name** box, then enter the description in the **Label** box underneath. Click **Change**.
Click Old and New Values to specify the values of new variable. In the new window that opens, click on Range: and enter the values to be combined, and the new value in the Value box under New Value. The original values for this variable were as follows:

<table>
<thead>
<tr>
<th>Original coding</th>
<th>Coding for new variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2 = “not asked this version”</td>
<td></td>
</tr>
<tr>
<td>1 = “Much better”</td>
<td>1 = “better”</td>
</tr>
<tr>
<td>2 = “Better”</td>
<td>2 = “the same”</td>
</tr>
<tr>
<td>3 = “About the same”</td>
<td>3 = “worse”</td>
</tr>
<tr>
<td>4 = “Worse”</td>
<td></td>
</tr>
<tr>
<td>5 = “Much worse”</td>
<td></td>
</tr>
<tr>
<td>8 = “Don’t know”</td>
<td></td>
</tr>
<tr>
<td>9 = “Refusal”</td>
<td></td>
</tr>
</tbody>
</table>

As this is recoding into a different variable, there is no need to worry about the values that aren’t carried across, as they will automatically be set to system missing. For the purposes of this example, we want a variable with three outcomes – “Better”, “The same”, “Worse”, and so will combine the values as follows.

Make sure that you have added the last change by clicking on Add, and clicking Continue and then OK.
To add values (i.e. “Better”, “The same”, “Worse”), go to the *Variable View* and find the new variable. Newly created variables will be added at the end of the dataset. In the *Values* column, click to the right to open the *Value Labels* window, and define the values as follows, and click *OK*:

![Value Labels window](image)

At this stage, it is a good idea to run a crosstabulation of your variable and the original to check that the recode has worked. Select *Analyze > Descriptive Analysis > Crosstabs...* This will show whether the recode has done what you expected it to.

![Crosstabs output](image)

The syntax for this process is as follows:

```
RECODE
   NHSSyrs (3=2) (1 thru 2=1) (4 thru 5=3)
   INTO NHSSstandard.
VARIABLE LABELS  NHSSstandard 'Standard of NHS improve decline stay same'.
EXECUTE.
```
ADD VALUE LABELS NHSstandard
1 "Better"
2 "About the same"
3 "Worse"

4.2 Computing new variables
Computing new variables allows for combining data from different variables. For example, you may be interested in the opinions of individuals aged 50 and over who are unemployed, using the 2011 British Social Attitudes Survey. Computing this variable means combining information from the variables "Person 1 age last birthday :Q50" (Rage), and "Respondent economic activity in last week< Priority coded> :Q693" (REconAct):

<table>
<thead>
<tr>
<th>Valid</th>
<th>Unemployed and registered at a JobCentre or JobCentre Plus</th>
<th>Unemployed, &quot;not&quot; registered, but actively looking for a job (of at least 10 hrs a week)</th>
<th>Unemployed, wanting a job (of at least 10 hrs per week) but &quot;not&quot; actively looking for a job</th>
<th>Permanently sick or disabled</th>
<th>Wholly retired from work</th>
<th>Looking after the home (Doing something else) (WRITE IN)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>In full-time education (not paid for by employer, including on vacation)</td>
<td>In paid work (or away temporarily) for at least 10 hours in week</td>
<td>Waiting to take up paid work already accepted</td>
<td>Valid</td>
<td>Percent</td>
<td>Valid Percent</td>
</tr>
<tr>
<td>Valid</td>
<td>66</td>
<td>2.1</td>
<td>2.1</td>
<td>2.1</td>
<td>10</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Select **Transform > Compute Variable**... This opens the **Compute Variable** window. Name the target variable `unemp_50` and then build up the expression to select people aged 50 and over, and unemployed. Click **Ok**.

In the numeric expression, "Rage > 49" denotes that we only want to select individuals aged over the age of 49. The rest of the expression is in brackets as we want to combine the information on unemployment from several values. (REconAct = 5 | REconAct = 6 | REconAct = 7) denotes that we want to select the individuals who have the code of "5" OR "6" OR "7" from the variable REconAct.

This process produces a variable which takes the value 1 when the logical expression is true, and 0 otherwise. It is a good idea to check whether the variable has computed correctly by running a crosstab of all three variables (the two originals and your computed variable). You can run a three-way crosstab by selecting a third variable as a **layer** variable using **Analyze > Descriptive Statistics > Crosstabs**, or by adding a second "by" statement in the syntax.
Don’t forget to define variable labels and value labels for the new variable “unemp_50” in the Variable Viewer once you have run this command.

The syntax for computing this variable is as follows. The first line sets a default value of -1 to explicitly handle information which is Not applicable (e.g. if you have any missing data through non-response). Where the logical condition is satisfied as true, this syntax asks SPSS to code the variable as 1 (and 0 where false). The variable label and value labels are also defined.

```
COMPUTE unemp_50 = -1.
if (RAge > 49 & (REconAct = 5 | REconAct = 6 | REconAct = 7)) unemp_50 = 1.
If (Rage < 50 or (REconAct < 5 | REconAct >7)) unemp_50 = 0.
VARIABLE LABELS unemp_50 Unemployed over 50.
EXECUTE.
```

ADD VALUE LABELS unemp_50

```
-1 Not applicable
0 No
1 Yes
```

This code can be reused by editing the variable names and logical expressions. Remember that SPSS can automatically generate syntax for you; until you become more experienced, all you have to do is edit it for future use.
5. Using hierarchical data in SPSS

The dataset used in the following examples is the English Housing Survey 2011-2012: Household Data (EHS). This survey asked over 13,000 households about their housing and local environment. This dataset can be downloaded from the UK Data Service website, after completing a short registration.

Hierarchical datasets consist of data at more than one level of measurement where lower level data are nested in one or more higher levels. For example, in a survey in which individuals within households are interviewed, both individual and household variables may be available. To link individuals to household, there will be a household identifier.

Hierarchical data are stored either as a single file containing variables at multiple levels (e.g. an individual level file containing individual and household variables) or in a multiple file format with one (or more) file at each level of measurement (e.g. an individual level file and a separate household level file). The information contained in the two formats is identical and Sections 5.1 and 5.3 of this chapter in effect show how to move between the two formats.

The examples in this guide use two level hierarchical data from the EHS as an example but the methods shown also apply to more complex hierarchical data.

5.1 Selecting one individual per household

Suppose you have individual and household variables contained in a single dataset. Analyses conducted using this data file are at the individual level by default. If you want to conduct household level analyses, you can select on individual in each household to create a household level dataset.

Use the file `people.sav` from the interview folder in the EHS. This is an individual level file that contains individual and household variables (e.g. `sex` at the individual level and `dvhhsiz`, household size, at the household level) and contains a household identifier `aancode`.
If you want to use these data to do household level analyses, you can select one person in each household\(^1\) by selecting only the Household Representative Person (HRP).

**Household Representative Person (HRP)**

The *Household Reference Person* (HRP) is a term commonly used in survey data to indicate a member of the household who is equivalent to a head of household. Where there are joint householders, the HRP is defined by the Office for National Statistics (who produce many of the UK large-scale surveys) as the individual with the highest income. If the joint householders have the same income, the eldest is selected. The HRP can be used in analysis to ‘represent’ the household in terms of income, say, or to select a single member of each household.

The way in which the HRP’s identity is stored varies from dataset to dataset. For example you may have a:

- binary variable that indicates whether the respondent is the HRP or not
- a variable which indicates the person number of the HRP, you will then need to test whether the respondent’s own person number is the same as that of the HRP
- variable that indicates the respondent’s relationship to the HRP

See the documentation that comes with your dataset to find a variable that identifies the HRP or some other variable that will allow you to select one individual per household.

Open the file *people.sav*. To select the HRP only via the menus, use **Data> Select Cases...**

---

\(^1\) Note that the EHS contains household level datasets that are more appropriate for household analysis.
Then select the option *If condition is satisfied* and click on the *If...* button.
The variable HRP is coded so that HRP=1 if the individual is the HRP and 2 if the individual is not the HRP. In the Select Cases: If dialogue box, either type in \texttt{hrp=1} or select \texttt{hrp} from the list of variables in the left-hand box, use the arrow to move it to the box at the top and the calculator buttons to add ‘=1’. Press \textit{Continue}.

It is advisable to check that this has worked as intended by doing a frequency of the variable \textit{hrp}. Analyses in SPSS will be conducted only on this selection of the data until all the data are selected again or some other selection is made. To select all the data, use Data> Select Cases... and choose ‘All cases’.

\textbf{Syntax to filter cases}

Using syntax, you can either use the \texttt{filter} command to temporarily select cases (until the filter is removed or changed):

\begin{verbatim}
COMPUTE filter_\$=(hrp=1).

VARIABLE LABELS filter_\$ 'hrp=1 (FILTER)'.

VALUE LABELS filter_\$ 0 'Not Selected' 1 'Selected'.

FORMATS filter_\$ (f1.0).

FILTER BY filter_\$.
\end{verbatim}
EXECUTE.

Syntax to temporarily drop cases

Alternatively, you can use the select command but note that unless preceded by the temporary command, the select command permanently drops all unselected cases. The temporary command applies until the next executable command – e.g. most processes which produce output. The syntax to temporarily select the HRP and conduct some simple analyses is below:

TEMPORARY.

***note that the temporary command means that the following selection is only in place until the next executable command.

SELECT IF (HRP=1).
DESC SEX persno.

***desc is an executable command.

TEMPORARY.

SELECT IF (HRP = 1).
FREQ SEX persno.

Syntax to permanently drop cases

Instead of using a temporary filter, it is possible to drop cases selecting only one person per household then save the file. In SPSS this can be achieved using the select command without the temporary command.

SELECT IF HRP = 1.
SAVE OUTFILE='newfilename'

/COMPRESSED.

5.2 Summarising characteristics of groups in hierarchical data

Suppose you have individual and household variables in a single individual level file and you wish to create a new summary household-level variable (e.g. the age of the oldest person in the household) using individual level data. The example below shows a simple example of
computing the age of the oldest person. It assumes that you have a household ID and age variable.

<table>
<thead>
<tr>
<th>Individual data</th>
<th>Summary variable age_max in household</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person 1 Household 1: age = 47</td>
<td>age_max = 47</td>
</tr>
<tr>
<td>Person 2 Household 1: age = 43</td>
<td></td>
</tr>
<tr>
<td>Person 3 Household 1: age = 18</td>
<td></td>
</tr>
<tr>
<td>Person 1 Household 2: age = 74</td>
<td>age_max = 74</td>
</tr>
</tbody>
</table>

Using the individual level dataset **people.sav**, from the drop-down menus, select **Data > Aggregate.**
Put the household ID aacode in the Break Variable(s) box and choose the variable age and move it into the aggregated variables box. Click the function button to choose the appropriate function to put in the Summaries of Variable(s) box. In this case, choose ‘maximum’. The new variable is named ‘age_max’ by default. In the Summaries of Variable(s) box, you should see age_max=MAX(age).

You have a choice as to how to store your resulting variable:

By default SPSS selects Add aggregated variable to active dataset which will match the new variable back on to the open file so that every person in the original file will contain the new household summary variable.

Alternatively to can produce a new household level file which will have one case per household and contain only the new variable(s) and the household ID variable, choose either Create a new dataset containing on the aggregated variable or Write a new dataset containing on the aggregated variable. The latter writes a new file to memory but does not open it in the current session.

The commands below create a household level file with the oldest person in the household where the new variable is added into the existing individual dataset.

Here is the syntax for this:

********************************************************************************
* SPSS syntax to generate household file with oldest person in a household.
********************************************************************************

*now create a file with one case per value of the household ID aacode.

AGGREGATE
/OUTFILE=* MODE=ADDVARIABLES
/BREAK=aacode
/age_max=MAX(age).

********************************************************************************

Or, to save the new variable in a household level data, use:
5.3 Attaching household data to an individual level file

Suppose you have hierarchical data held in two files, one with individual level variables and one with household level variables and in which each contains the household ID aacode.

To attach the household and individual level datasets to create an individual level dataset, first sort all files to be combined by the household ID variable aacode, save and close them. To sort the data by aacode, use Data> Sort Cases..., then select aacode and move it into the Sort by: box.
Using a household level file from the EHS *generalfs11.sav* (in the derived folder when the data are downloaded), from the menu, choose Data > Merge Files > Add Variables, and select the dataset: *people.sav* (in the interview folder when the data are downloaded) to add. In the new window, tick *Match cases on key variables in sorted files* and choose *Active data set is keyed table*. Then click on the ID variable *aacode* in the left-hand window and click on the arrow to move it to the *Key Variable* box as shown below:

![Add Variables window](image)

Press *OK*. Look at the new data in *Data View*. The variable *aacode* now contains multiple rows with the same value representing different people in the same household. Finally save the data under a new name.
<table>
<thead>
<tr>
<th>aacode</th>
<th>rumorph</th>
<th>aagth11</th>
<th>paired</th>
<th>casecat</th>
</tr>
</thead>
<tbody>
<tr>
<td>J0011101</td>
<td>1</td>
<td>1152.38</td>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>J0011102</td>
<td>1</td>
<td>796.28</td>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>J0011103</td>
<td>1</td>
<td>902.35</td>
<td>0</td>
<td>4.00</td>
</tr>
<tr>
<td>J0011104</td>
<td>1</td>
<td>1637.99</td>
<td>0</td>
<td>5.00</td>
</tr>
<tr>
<td>J0011105</td>
<td>1</td>
<td>1500.43</td>
<td>0</td>
<td>4.00</td>
</tr>
<tr>
<td>J0011202</td>
<td>1</td>
<td>1155.30</td>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>J0011202</td>
<td>1</td>
<td>1155.30</td>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>J0011203</td>
<td>1</td>
<td>1894.57</td>
<td>0</td>
<td>4.00</td>
</tr>
<tr>
<td>J0011203</td>
<td>1</td>
<td>1894.57</td>
<td>0</td>
<td>4.00</td>
</tr>
<tr>
<td>J0011204</td>
<td>1</td>
<td>1206.46</td>
<td>0</td>
<td>4.00</td>
</tr>
<tr>
<td>J0011204</td>
<td>1</td>
<td>1206.46</td>
<td>0</td>
<td>4.00</td>
</tr>
<tr>
<td>J0011207</td>
<td>3</td>
<td>1232.20</td>
<td>0</td>
<td>5.00</td>
</tr>
<tr>
<td>J0011207</td>
<td>3</td>
<td>1232.20</td>
<td>0</td>
<td>5.00</td>
</tr>
<tr>
<td>J0011207</td>
<td>3</td>
<td>1232.20</td>
<td>0</td>
<td>5.00</td>
</tr>
</tbody>
</table>
Check that the merge has worked as intended

If you merged all the household variables into the individual level dataset, the same information should be in the new file as there was in your two original files. Look at the data:

- does it look right? Is the merged dataset at the correct level of measurement? Are there lots of missing data that weren’t in the original datasets? If so, should there be?

- do frequency tables of the same variables in both the old and merged individual level files. The results should be the same.

- select only the hrp in the merged file and do frequency tables of some of the household level variables. Then do the same using the original household file. The results should be the same.

Using syntax to match files

When matching, it is important that the files are sorted by the variables that they are to be matched upon. You can achieve this by using the `sort` command.

The `match` command is a flexible command which can be used to match cases, either by matching case by case, or on the basis of an identifier. Where you have files at different levels, the upper level file is treated as a ‘lookup table’ from which values, uniquely defined by the ID, can be looked up and distributed to all cases in the main (lower level) file which have the ID variable value. Accordingly, if you seek to distribute household level data to the individual level, the household file will be the “table” and the individual file will be the “file”.

Note that you can replace a filename for either the lookup table or file with * when that file is already open. In this example, the household level file is already open, so it has been replaced with a ‘*’.

The execute command is optional; however you should use this (or another executable command such as frequencies) in order to execute the command as the command will not execute on its own.
The syntax for this is as follows, where `<householdfile.sav>` is the household level file with path and `<individualfile.sav>` is the individual level file with path:

GET
FILE='`<householdfile.sav>`'.
SORT CASES BY aocode (A).
MATCH FILES /TABLE=*  
   /FILE='`<individualfile.sav>`'
   /BY aocode.
EXECUTE.
SAVE OUTFILE='newname.sav'
   /COMPRESSED.
6. Linking and merging files in SPSS

The dataset used in the following examples is the English Housing Survey 2011-2012: Household Data (EHS). This survey asked over 13,000 households about their housing and local environment. This dataset can be downloaded from the UK Data Service website, after completing a short registration.

Many datasets come as a single file but others are supplied in multiple files. These files may contain:

1. Information about the same cases at the same level of measurement (e.g. two files containing information about the same households but with different variables in each)
2. Hierarchical data: data that contain more than one level of measurement nested within another (e.g. individuals within households in individual level file(s) and household file(s)).
3. Data with the same variables but different cases (e.g. the same survey conducted in England and Wales with the data for each country in its own file)

Combining together (merging) multiple files that contain the same cases involves using one or more matching variables that uniquely identify each case (e.g. an individual ID variable and/or household ID etc.) to link the data from the same case in different files.

6.2 Linking multiple files at the same level of measurement

In the EHS, the generalfs11.sav file contains the weight and interviewfs11.sav contains the answers to the household interview. Both files are at the household level and contain all the cases. To analyse the interview data using the household weight, these two files must be merged.

To do this, first order the matching variable (aacode) in the same way in both data sets if in by right-clicking on the title of the aacode column and select Sort Ascending. Save and close both files.
Open either one of the files. You can open and use either file first in this instance because there is a 1:1 match of households between the two files so it makes no difference whether you add the interviewfs11.sav to the generalfs11.sav or the other way round.

Open `generalfs11.sav`. To add a file to the open file, use the menu: Data> Merge Files> Add Variables to get to the following screen:

Browse to find the second data set: `interviewfs11.sav` then press Continue.

In the new window, select Match cases on key variables in sorted files, and Non active dataset is a keyed table. Then click on `aacode` in the left-hand window and click on the arrow to move it to the Key Variable box as shown below:
Press OK to merge the files. The new merged data set is now displayed: you should see that there were 19 variables in total before the merge, and now there are 126 variables. The new data set should be saved under a new name.

**Merging files when one file contains information about some of the cases only**

You can use the same method to merge two files with the same level of measurement but for which one file contains a subset of the cases contained in the other. Because this is not a 1:1 correspondence between the cases as in the example above the choice of data to open first does affect the resulting merged dataset.

If you open the file with all the cases first and add the subset file to it, you will obtain a file with all the cases in it. The dataset will contain values where they exist and missing values (system missing blanks) for the cases not contained in the subset.

If you open the subset file first and add the other file to it, you will obtain a file with all the variables in both datasets but only for the subset of cases.
The syntax to sort and merge two household files is below, where `<householdfile1.sav>` is the file name with file path for the first file and `<householdfile2.sav>` is the file name with file path for the second file and `<merge1.sav>` is the saved name and path of the new dataset.

******************************************************************************

* Opens, sorts and saves the two household files and closes `<householdfile2.sav>`.
******************************************************************************

GET
  FILE=`<householdfile1.sav>`.
DATASET NAME DataSet1 WINDOW=FRONT.
SORT CASES BY aacode (A).
DATASET ACTIVATE DataSet1.
SAVE OUTFILE=`<householdfile1.sav>`
  /COMPRESSED.
GET
  FILE=`<householdfile2.sav>`.
DATASET NAME DataSet2 WINDOW=FRONT.
SORT CASES BY aacode (A).
DATASET ACTIVATE DataSet2.
SAVE OUTFILE=`<householdfile2.sav>`
  /COMPRESSED.
DATASET ACTIVATE DataSet1.
DATASET CLOSE DataSet2.
******************************************************************************

* Merges the two household files and saves the new merged file as merge.sav.
******************************************************************************

MATCH FILES /FILE=* 
  /TABLE=`<householdfile2.sav>`.
  /BY aacode.
EXECUTE.
SAVE OUTFILE=`<merge1.sav>`
  /COMPRESSED.
6.3 Attaching household level data to individuals

Datasets sometimes come with multiple files because the files contain data at different levels (e.g. individual, and household).

When the data are of the following format, you may add the higher level data to the lower level file (e.g. add household variables to an individual level dataset):

- A higher level file (e.g. household level)
- A lower level file (e.g. individual level) where the lower level data are nested within the higher level data (e.g. individuals within households)
- Both files contain a matching variable to allow the higher level to be identified in each file e.g. both contain a household ID variable.

The data described are an example of hierarchical data. See Section 5.3 in Chapter 5 about using hierarchical data in SPSS for how to attach higher level data to lower level data in a hierarchical dataset.

6.4 Merging files with different cases but the same variables

If the two datasets have the same variables but different cases you may want to combine the files. For example, when using a survey conducted in both England and Wales with the same questions with the data contained in 2 files, you may want to create a dataset for England and Wales together.

In this instance, there is no matching variable because there is no linking of cases. They are different cases in each file.

To achieve this in SPSS, use Data > Merge files > Add Cases

One thing to remember when doing this is that you may not have a unique person identifier any more once you add files together so you might want to consider making a new person identifier.