2
DATA MANAGEMENT

INTRODUCTION

Once the data are entered into SPSS, it is inevitable that some degree of data management or manipulation will be necessary before analysing data. It is much better do this using SPSS than 'by hand'. It is inevitable that mistakes, such as errors in entering the data or transposing between columns will occur if data management is done 'by hand'. Moreover, if the dataset is large, carrying out such manipulations will be time consuming. Additionally, if syntax is being utilised, a record of the data management is made so that the process can be replicated in the future and others using the dataset know the logic behind the changes made. Data management functions may include recoding variables that are already present in the dataset, deriving new variables from variables already in the dataset, selecting specific cases for analysis or merging datasets. This chapter includes a number of data management functions that are useful in a variety of circumstances.

This chapter uses data from the student breast cancer awareness study. Data were collected in a cross sectional survey from female staff and students aged less than 50 years at a UK university. This chapter also utilises the student obesity dataset, which asked about risk factors for and opinions on obesity.

THE AIMS OF THIS CHAPTER ARE:

- To demonstrate a number of data management functions within SPSS to enable more succinct data analysis.

PRELIMINARIES TO THIS CHAPTER

Dialog boxes are invoked when the majority of menu items are selected. They allow specific choices related to the procedure to be made.

This arrow  is used by SPSS to transfer variables or functions around dialog boxes.

RECODING VARIABLES

Sometimes it may be necessary to recode variables. For example, it may be necessary to collapse categories within a variable. For example, ethnicity may have been collected in more categories that it is practical to use for analysis. Alternately, numerical codes may need changing to facilitate statistical modelling. For example, when doing logistic regression (described further in Chapter 12), the dependent (outcome) variable should be coded 0 versus 1, but may not have been coded so when data were originally entered. Recoding can also be used to categorise a variable which was collected as a continuous variable, but will be analysed as a categorical variable. For example, BMI score may be categorised to underweight, normal range, overweight and obese using standard cut-offs.

When recoding, it is advisable to recode into a new variable so that the original variable can still be used if necessary.

For example, using the student breast cancer awareness survey, ethnicity was collected in a large number of categories. For analysis purposes, it is beneficial to recode these to three categories: 'white', 'black' and 'other' because some of the categories had very small frequencies within them (for example, there was only one Bangladeshi woman), therefore it would be difficult to make inferences about those groups. The original variable was coded as shown in Figure 2.1.

To recode data into a different variable click on Transform → Recode into Different Variables … to get the Recode into Different Variables dialog box shown in Figure 2.2.

The variable to be recoded should be transferred to the Input Variable → Output Variable: box and a new variable name declared in the Output Variable Name: box. It is also possible (although not compulsory, this can also be done in Variable View

Ethnicity

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	White British	61	37.7	38.9	38.9
	White Irish	1	.6	.6	39.5
	White other	3	1.9	1.9	41.4
	Black British	22	13.6	14.0	55.4
	Black Caribbean	11	6.8	7.0	62.4
	Black African	13	8.0	8.3	70.7
	Mixed (B&W)	6	3.7	3.8	74.5
	Mixed other	3	1.9	1.9	76.4
	Indian	14	8.6	8.9	85.4
	Pakistani	4	2.5	2.5	87.9
	Bangladeshi	1	.6	.6	88.5
	Chinese	6	3.7	3.8	92.4
	Other	12	7.4	7.6	100.0
	Total	157	96.9	100.0	
Missing	System	5	3.1		
Total		162	100.0		

FIGURE 2.1 FREQUENCIES OF ETHNICITIES OF WOMEN WHO PARTICIPATED IN A BREAST AWARENESS SURVEY

Note: further explanation of frequencies is given in Chapters 5 and 6.

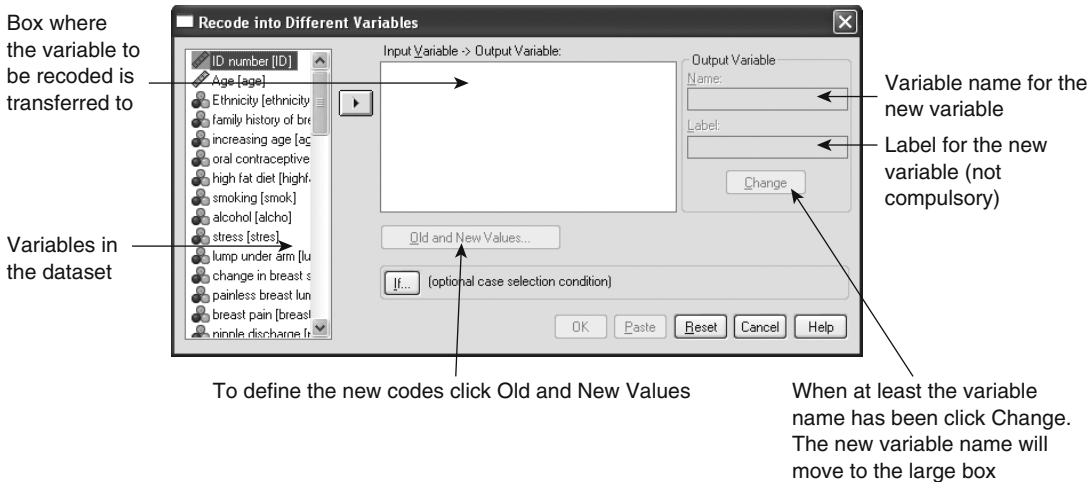

FIGURE 2.2 RECODE INTO DIFFERENT VARIABLES DIALOG BOX

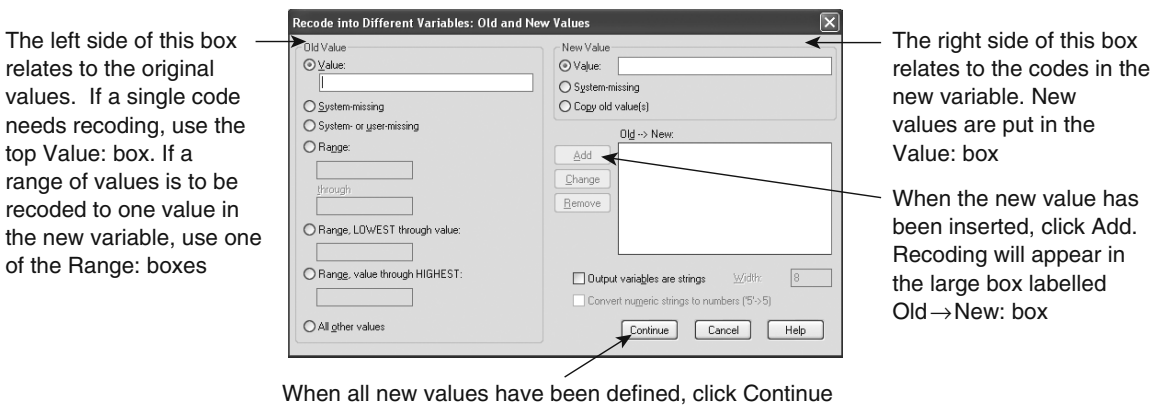

FIGURE 2.3 RECODE INTO DIFFERENT VARIABLES: OLD AND NEW VALUES DIALOG BOX

once the variable has been defined) to label the new variable by filling in the Label: box. When these have been completed, click the Change button and the new variable name will be transferred to the Input Variable → Output Variable: box. Next click Old and New Values… to give a dialog box like the one in Figure 2.3.

Looking at Figure 2.1, the aim of this recoding is to code the three white ethnicities together, the black ethnicities plus mixed black and white together and finally all other categories together, so that there are three categories ('white', 'black' and 'other'). Missing data in the original variable will remain missing with the new variable. To recode a variable, it will probably be necessary to refer to Variable View so that the right numerical codes are recoded. In the ethnicity example, the three white ethnicities are coded 1, 2 and 3; with the new variable they will be recoded to 1. On the left hand side of the dialog box (under Old Value) in Figure 2.3, click on the first Range: radio button to activate the boxes below; put 1 in the top box and 3 in the box below 'through'. On the right hand side of the dialog box (under New Value), in the Value box put 1. Then click Add. The changes that will occur will then be

Ethnic group

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	white	65	40.1	41.4	41.4
	black	52	32.1	33.1	74.5
	other	40	24.7	25.5	100.0
	Total	157	96.9	100.0	
Missing	System	5	3.1		
Total		162	100.0		

FIGURE 2.4 ETHNICITY VARIABLE FOLLOWING RECODING

Note: the value labels were defined using Variable View before Figure 2.4 was produced.

shown in the large white box on the right hand side of the dialog box, headed Old → New:. The 'black' and 'other' groups are recoded in a similar way. When all coding has been declared then press Continue, to return to the Recode into Different Variables dialog box (Figure 2.2). On that dialog box click OK. New variables are placed at the far right of the dataset in Data View, and at the bottom of the variable list in Variable View. The attributes of the variable can be altered in Variable View once the variable has been created in the same way as setting up a datasheet for data entry (Chapter 1). For example, for a categorical variable, it may be beneficial to define the Values as well as change the Measure (the default is Scale).

To make sure the recoding carried out produced the expected results, it is advisable to compare the frequencies of the new variable with the old variable (Figure 2.1). For the new variable, the frequency table shows 65 of 157 (41%) women were white, 52 (33%) were black and 40 (26%) were other ethnicity (Figure 2.4). As there should be, there are the same number of missing values in the original coding (Figure 2.1) of ethnicity and the recoded variable (Figure 2.4).

As the Recode into Different Variables: Old and New Values dialog box (Figure 2.3) shows, there are a number of options about how to specify the old values depending on how the original variable was coded in relation to how the new variable is to be coded. For example, if the aim was to change the coding of a variable for logistic regression from 1 and 2 to 1 and 0 (recoding 2 to 0) then 2 would be placed in the Value: box under Old Value and 0 placed in the Value: box under New Value. As previously the Add button has to be clicked to register the changes. Changes are only made when the Continue button has been clicked on the Recode into Different Variables: Old and New Values dialog box (Figure 2.3), followed by OK on the Recode into Different Variables dialog box (Figure 2.2).

Another method that could have been used to define some of the old values of ethnicities in the student breast cancer awareness study is to define the values in terms as Range, LOWEST through value: since the white group was coded 1, 2 and 3; and there were no codes below one, 3 could have been placed in the box next to Range, LOWEST through value: and as before 1 placed in the Value: box on the New Value side of the dialog box (Figure 2.5). A similar principle applies for the Range, value through HIGHEST: box.

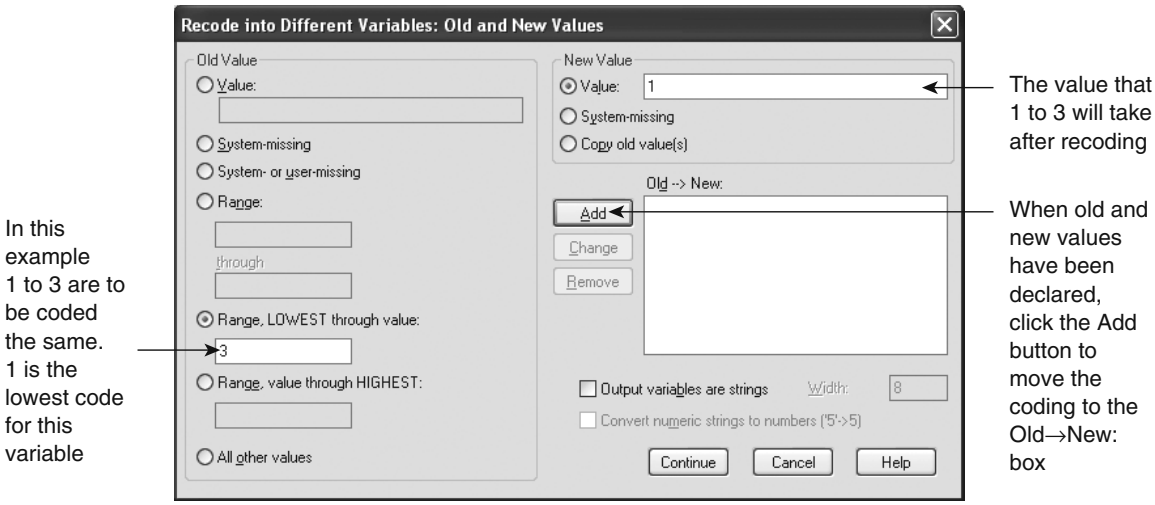

FIGURE 2.5 RECODE INTO DIFFERENT VARIABLES: OLD AND NEW VALUES, EXAMPLE OF LOWEST THROUGH VALUE

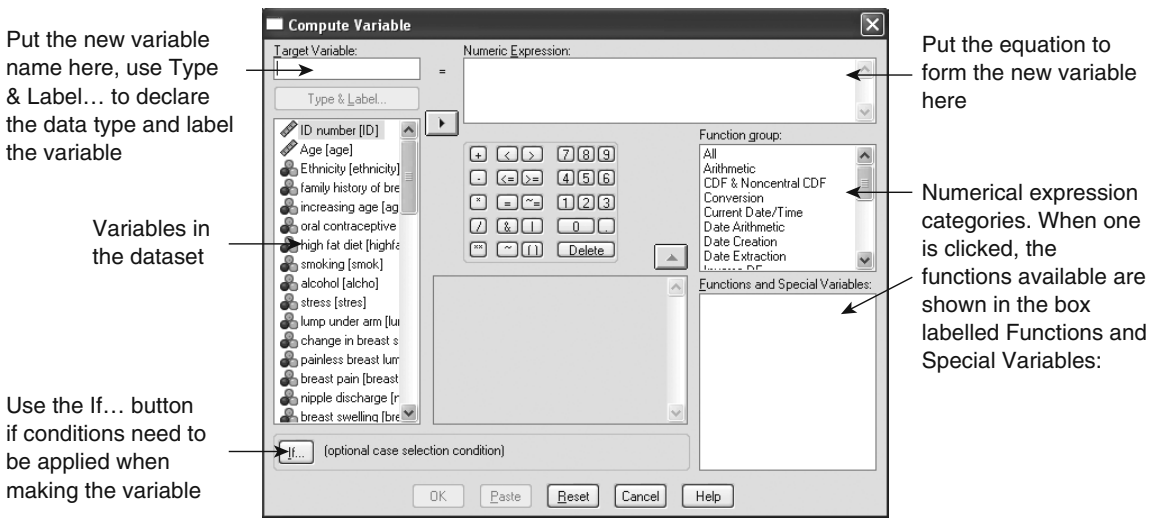

FIGURE 2.6 COMPUTE VARIABLE DIALOG BOX

CREATING NEW VARIABLES

It is often necessary to create new variables. This may be to create dummy variables for linear regression analysis (explained in Chapter 11); to combine data from more than one related variables to create a new variable or to manipulate the data using a function such as natural logarithms.

Whatever the reason for creating a new variable, the procedure starts by clicking on Transform → Compute Variable… to give the Compute Variable dialog box (Figure 2.6). Into the Target Variable: box, type the new variable name. Once a variable name has been added, the Type & Label… button becomes functional. This can

be used to declare the type of data and to label the variable. Both of these can also be declared or edited in Variable View after the variable has been computed.

The first example of computing variables shown will be to create dummy variables which can then be used as independent variables in linear regression. Linear regression is explained further in Chapter 11.

Dummy (indicator) variables

Where a categorical variable has more than two categories, the user has to make dummy (sometimes referred to as indicator) variables so that a linear regression coefficient can be calculated for each category within a variable. Dummy variables are a series of mutually exclusive dichotomous variables that represent all categories within the original variable. These are usually coded 0 indicating without the characteristic in question and 1 indicating with the variable in question. This textual example will use categorical age data from the obesity dataset (as this variable is used in multiple linear regression in Chapter 11).

In the original variable 1 = less than 20 years, 2 = 20 to 30 years, 3 = 31 to 40 years, 4 = 41 to 50 years and 5 = 51 to 60 years. There were no participants aged less than 20 years in the dataset. When constructing dummy variables, one category has to be designated the reference category, that is, the one which the other categories are compared to. In the example shown in Table 2.1 category 2 (age 20 to 30 years) will be the reference category. This means that it will not be necessary to create a variable representing this age group. For the other age categories, it is necessary to create new variables that equal 0 if the participant is not a member of the age category in question and 1 if they are. For example, looking at Table 2.1, the original coding of the variable is shown in the Age category column, with the following three columns being the dummy variables (Age31–40 representing age 31 to 40, Age41–50 representing age 41 to 50 and Age5160 representing age 51 to 60. Looking at the variable Age31–40, the only occasions where it takes 1 is where the value in the Age category variable is 3 (indicating 31 to 40 years in the original variable). Likewise, this is repeated in Age41–50 and Age51–60, with these variables taking 1 where Age category equals 4 and 5 respectively. Dummy variables can be constructed by recoding into a different variable using SPSS.

TABLE 2.1 CODING OF DUMMY VARIABLES – AN EXTRACT FROM THE AGE CATEGORY FROM THE OBESITY DATASET

Age category	Age 31–40	Age 41–50	Age 51–60
2	0	0	0
4	0	1	0
2	0	0	0
2	0	0	0
2	0	0	0
2	0	0	0
2	0	0	0
3	1	0	0
4	0	1	0
3	1	0	0
5	0	0	1
2	0	0	0

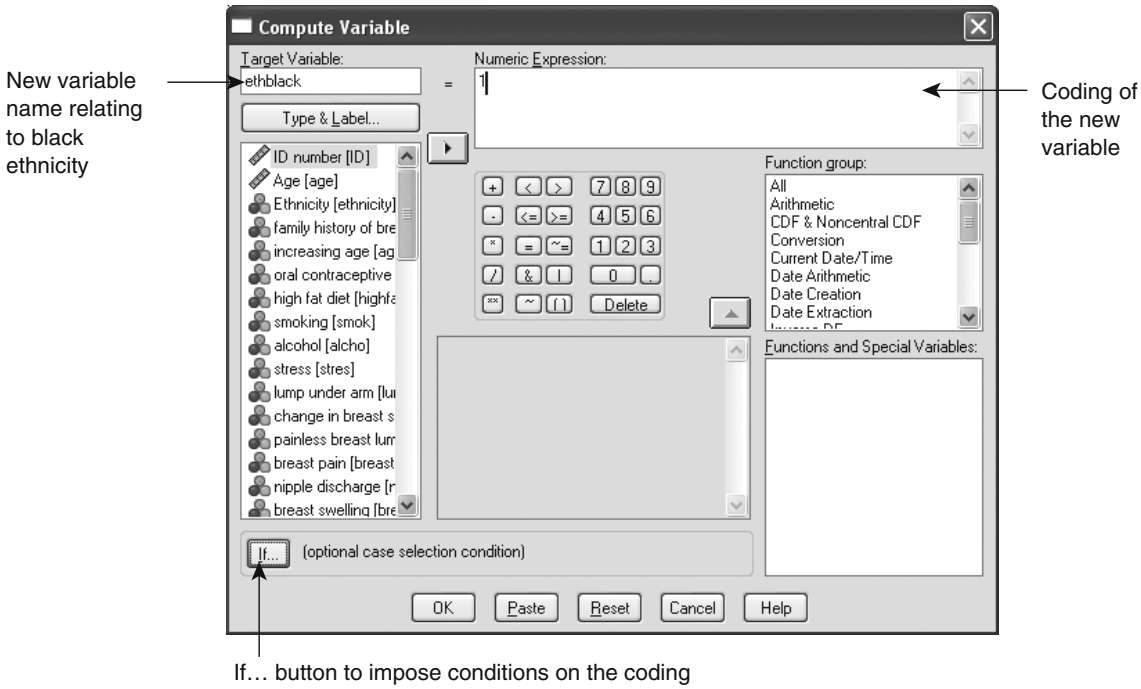

FIGURE 2.7 COMPUTE VARIABLE DIALOG BOX TO CREATE DUMMY VARIABLES

Figure 2.7 shows the student breast cancer awareness dataset. The variable to be made into a dummy variable is ethnicity, which has already been recoded (from the large number of categories the data were collected in) to three categories (white, black and other) earlier in this chapter. Two variables need to be created: black versus not black and other ethnicity versus not other ethnicity. White will be the reference category, so that if a participant is coded as not black and not other ethnicity, then as long as there is not missing data for that participant they will be white. Figure 2.7 shows the target variable (new variable name) is ethblack; this variable will be the dummy relating to black ethnicity. 1 has been placed in the Numeric Expression: box as that will indicate that the participant is black. However, not all women in this study were black, the If… button should be clicked to give the Compute Variable: If Cases dialog box shown in Figure 2.8. This is used so that only the women who are black are coded 1 with the dummy variable. Therefore in the Compute Variable: If Cases dialog box (Figure 2.8), click the radio button next to Include if case satisfies condition: then in the white box at the top of the dialog box (Figure 2.8), put the original ethnicity variable = the required coding of the original variable. For example, Figure 2.8 shows ethgroup = 2 because black is coded 2 in that variable. When that has been completed click Continue to return to the Compute Variable dialog box (Figure 2.7), then click OK.

The new variable will be situated at the far right of the dataset. It contains values where the coding is 1 (corresponding to the black participants). To fill in the remainder of the variable (where there is information on ethnicity, but the participant is not black) to 0 click on Transform → Recode into Same Variables… to give the Recode into Same Variables dialog box as shown in Figure 2.9. Move the variable to be recoded to the Numeric Variables: box. In this example, this is ethblack. Then click on

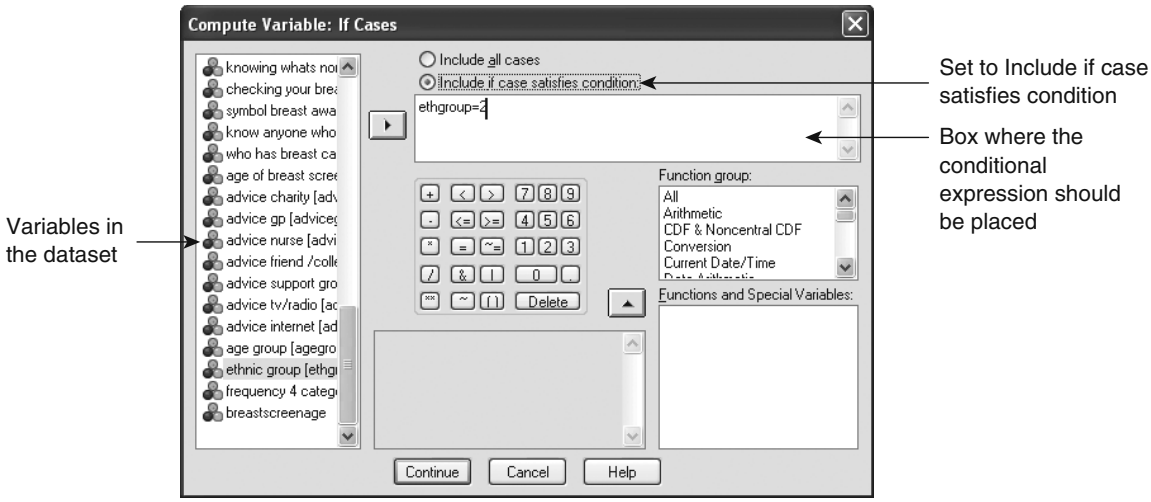

FIGURE 2.8 COMPUTE VARIABLE: IF CASES DIALOG BOX

If… to give the Recode into Same Variables: If Cases dialog box (Figure 2.10). Within this, a condition to only include those participants where there is not missing data in the original variable will be set up. Therefore change the radio button to select Include if case satisfies condition:, then enter the appropriate expression in the large white box (in Figure 2.10 this equates to the variable ethgroup not having missing data). When the expression has been entered click Continue to return to the Recode into Same Variables dialog box (Figure 2.9). The condition constructed will then appear next to the If… button, where (optional case selection condition) is shown in Figure 2.9. Then click on the Old and New Values… button to give the Recode into Same Variables: Old and New Values dialog box (Figure 2.11). On the Old Value side of the dialog box selected the System-Missing radio button and on the new values side of the dialog box put 0 in the Value: box. Then click Add followed by Continue to return to the Recode into Same Variables dialog box (Figure 2.9), and finally click OK. It is good practice to compare the frequencies of the original variable with the

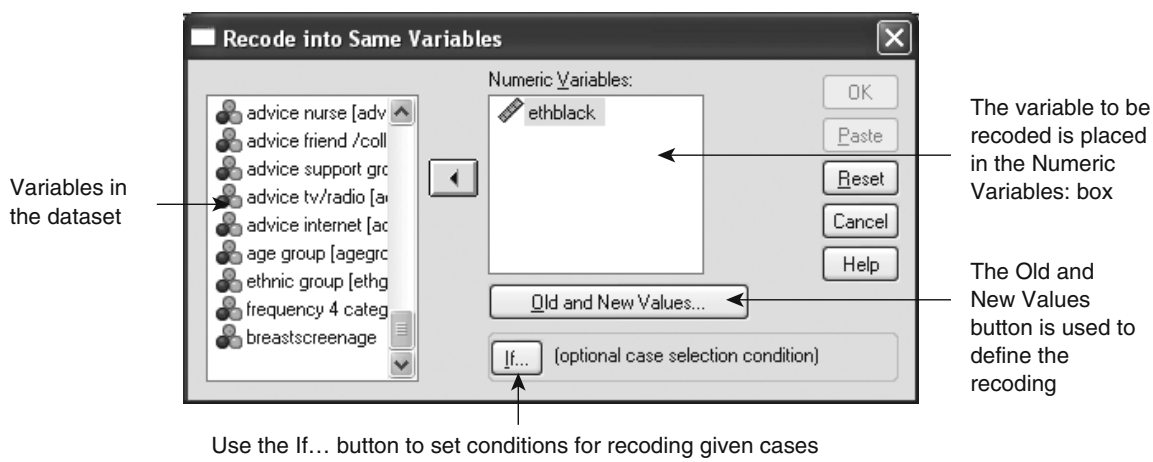

FIGURE 2.9 RECODE INTO SAME VARIABLES DIALOG BOX

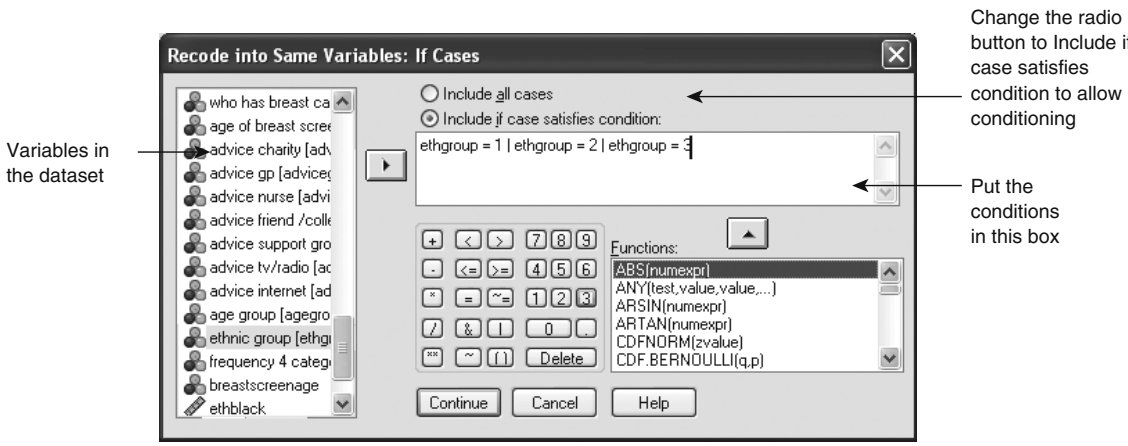

FIGURE 2.10 RECODE INTO SAME VARIABLES: IF CASES DIALOG BOX

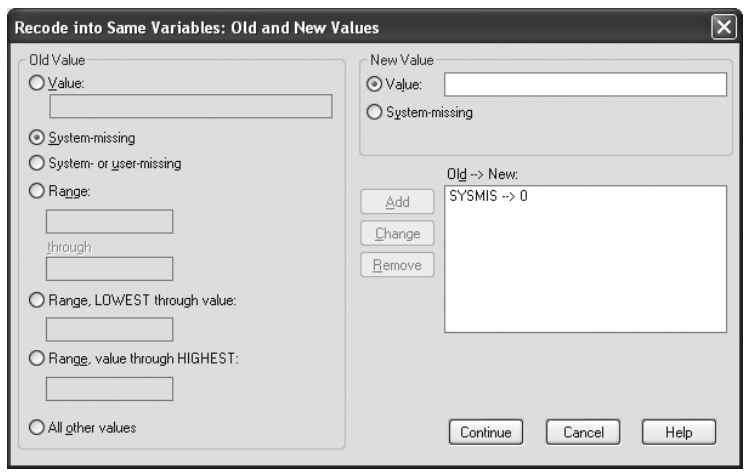

FIGURE 2.11 RECODE INTO SAME VARIABLES: OLD AND NEW VALUES DIALOG BOX

new variable to make sure the new variable is giving the same frequencies (in terms of missing and present data). These are shown in Figure 2.12; where the new variable has the same number of black women and the same amount of missing data. As previously value labels can be declared in Variable View.

The second example of the use of Compute Variable is to create a variable containing the natural logarithm of an existing continuous variable. This may be necessary when a variable is not Normally distributed as transformation can (but does not always) normalise skewed variables. However, it should be noted that if the original variable includes values of 0 (which may occur in variables representing health or quality of life scales), 0 is unable to be transformed to the natural logarithm scale (it would appear as missing data). This can be resolved by adding a small amount (0.5 or less) to each observation then remembering to subtract the amount added when back transforming for interpretation purposes.

If a skewed variable is transformed it may be possible to use parametric methods for analysis. Figure 2.13 shows a histogram of BMI from the student obesity dataset;

Ethnic black

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	105	64.8	66.9	66.9
	1	52	32.1	33.1	100.0
	Total	157	96.9	100.0	
Missing	System	5	3.1		
Total		162	100.0		

Ethnic group

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	White	65	40.1	41.4	41.4
	Black	52	32.1	33.1	74.5
	Other	40	24.7	25.5	100.0
	Total	157	96.9	100.0	
Missing	System	5	3.1		
Total		162	100.0		

FIGURE 2.12 FREQUENCIES OF THE BLACK ETHNIC VARIABLE AND ETHNIC GROUP VARIABLE

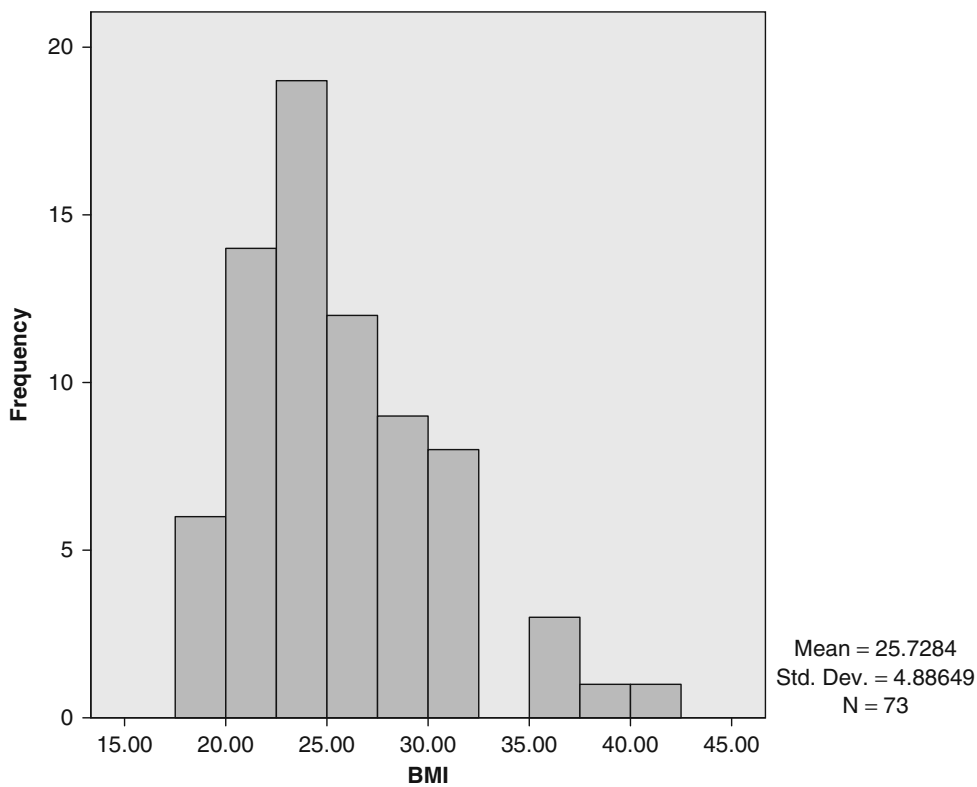

FIGURE 2.13 HISTOGRAM OF BMI FROM THE STUDENT OBESITY DATASET

it shows BMI is a little right skewed with more observations towards the lower end of the range with a small number of participants with a BMI of 35 or more. The mean BMI is 25.7 (SD 4.9). As the skew is not severe in this variable, it would be at the discretion of the user (and also considering whether the assumptions of statistical tests had been met) whether to transform this variable.

As with the previous example the new variable name has to be defined by completing the Target Variable: box (Figure 2.14) in the Compute Variable dialog box before variables and functions can be added to the Numeric Expression: box. Then select the LN function from the Functions and Special Variables: list from the Compute Variable dialog box and transfer it to the Numeric Expression: box using the upward pointing arrow. The insertion point should be between the brackets (if not, it should be moved to between the brackets); this is where the existing variable should be transferred from the variable list so that the Compute Variable dialog box appears like the one shown in Figure 2.14. When this has been completed click OK; the new variable has been created and will be situated at the right of the dataset. The distribution of the new variable is shown in Figure 2.15. This can be seen to be more Normally distributed than the original variable (Figure 2.13).

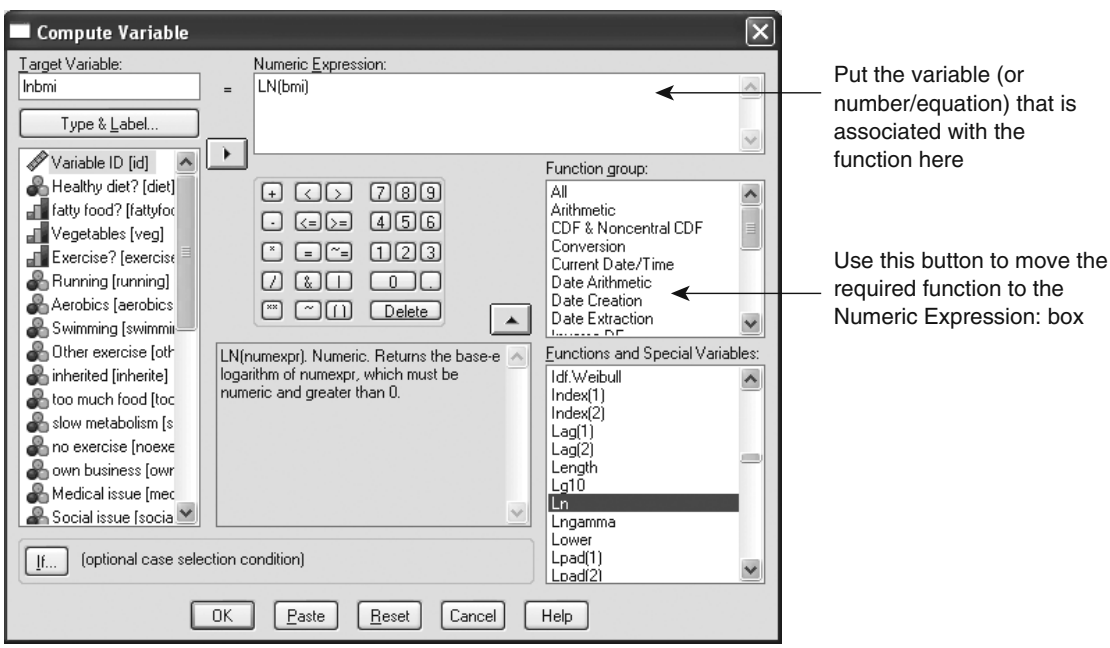

FIGURE 2.14 COMPUTE VARIABLE DIALOG BOX SHOWING LN (NATURAL LOGARITHMS) EXAMPLE

SELECTING CASES

Sometimes it may be necessary to exclude some data on the basis of the responses to a particular variable. For example, there may an interest in characteristics of one gender or ethnicity only.

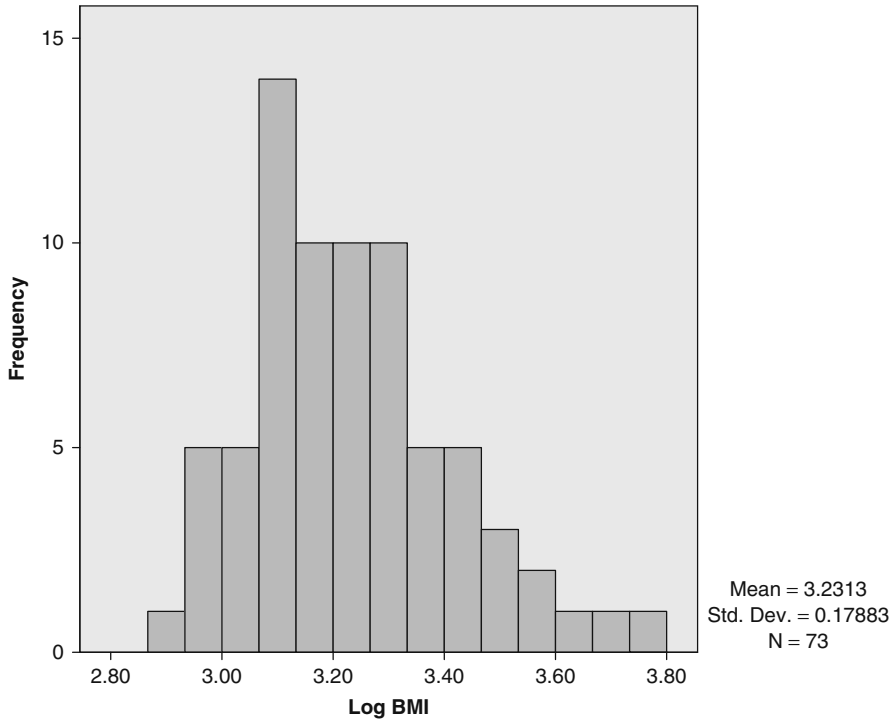

FIGURE 2.15 HISTOGRAM OF THE LOGARITHM OF BMI FROM THE STUDENT OBESITY DATASET

Click on Data → Select cases… . The Select Cases dialog box (Figure 2.16) will appear. To only use cases with specific characteristics, click on the If condition is satisfied radio button, then click the If… button to get the Select Cases: If dialog box (Figure 2.17). In this dialog box, the algorithm representing the condition should be

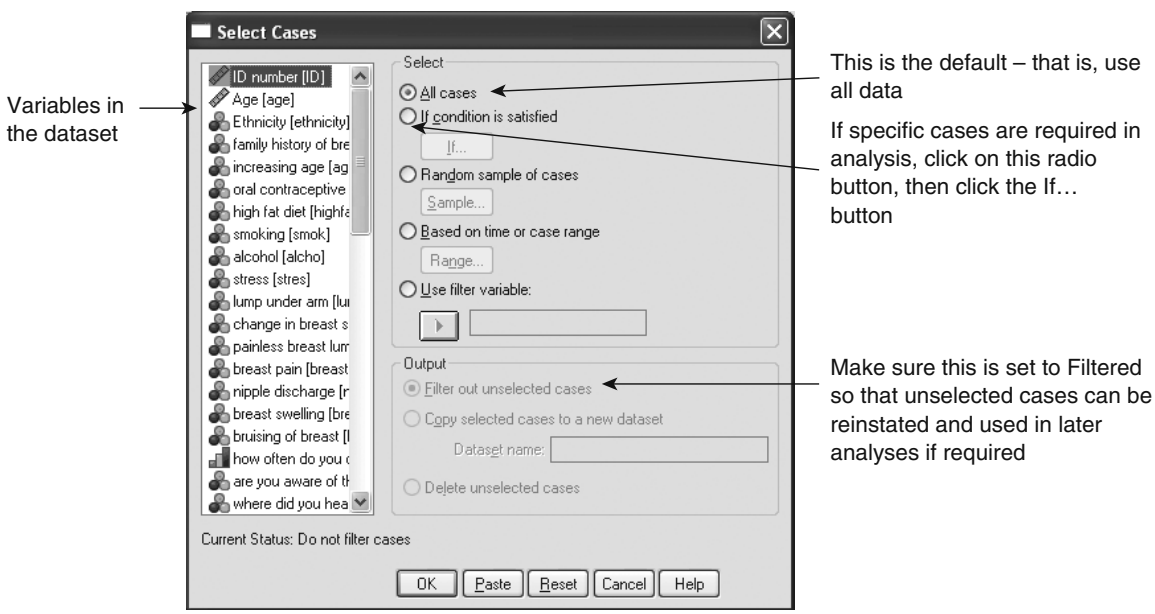

FIGURE 2.16 SELECT CASES DIALOG BOX

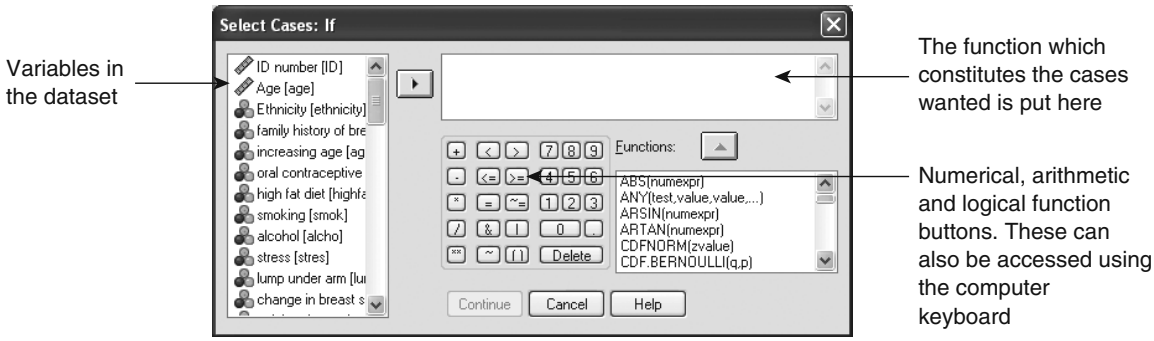

FIGURE 2.17 SELECT CASES: IF DIALOG BOX

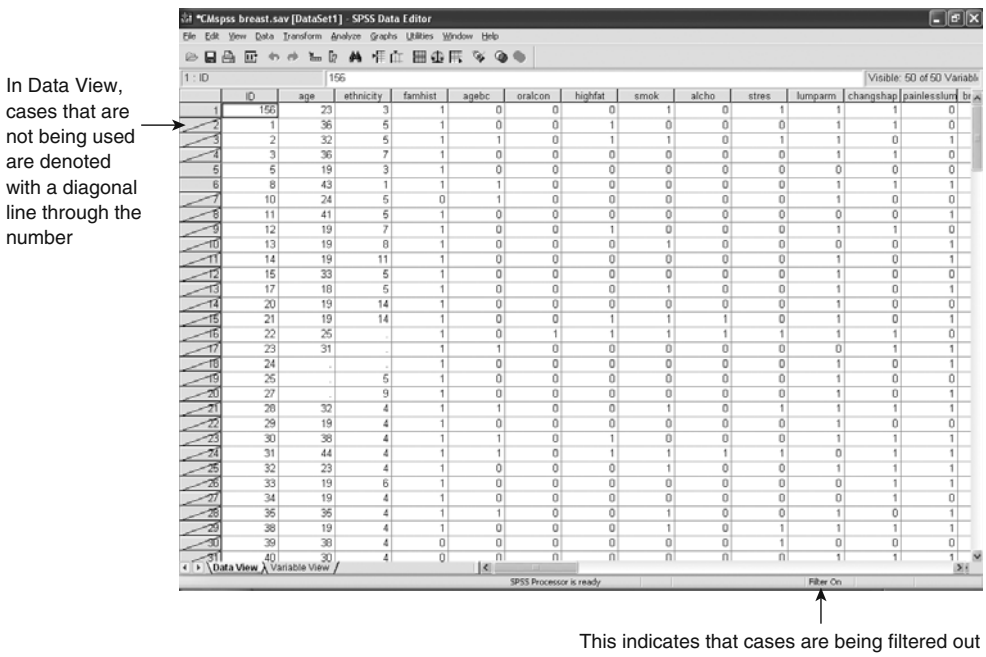

FIGURE 2.18 SCREENSHOT SHOWING HOW NON-SELECTED CASES ARE INDICATED IN DATA VIEW

placed. When the condition has been constructed (in the example shown in Figure 2.18 this is to select cases if ethgroup=1 in the breast cancer awareness dataset), click Continue, to return to the Select Cases dialog box (Figure 2.16) then click OK. In Figure 2.18 the observations that have been filtered out are signified by a diagonal line through the SPSS numerical identifier.

In addition, the selecting cases procedure creates a new variable filter_$, which takes the values 0 for not selected and 1 for selected. This is shown in Variable View (Figure 2.19).

When the whole dataset is required after using Select Cases, from the Select Cases dialog box (Figure 2.16) select the All cases radio button then click OK.

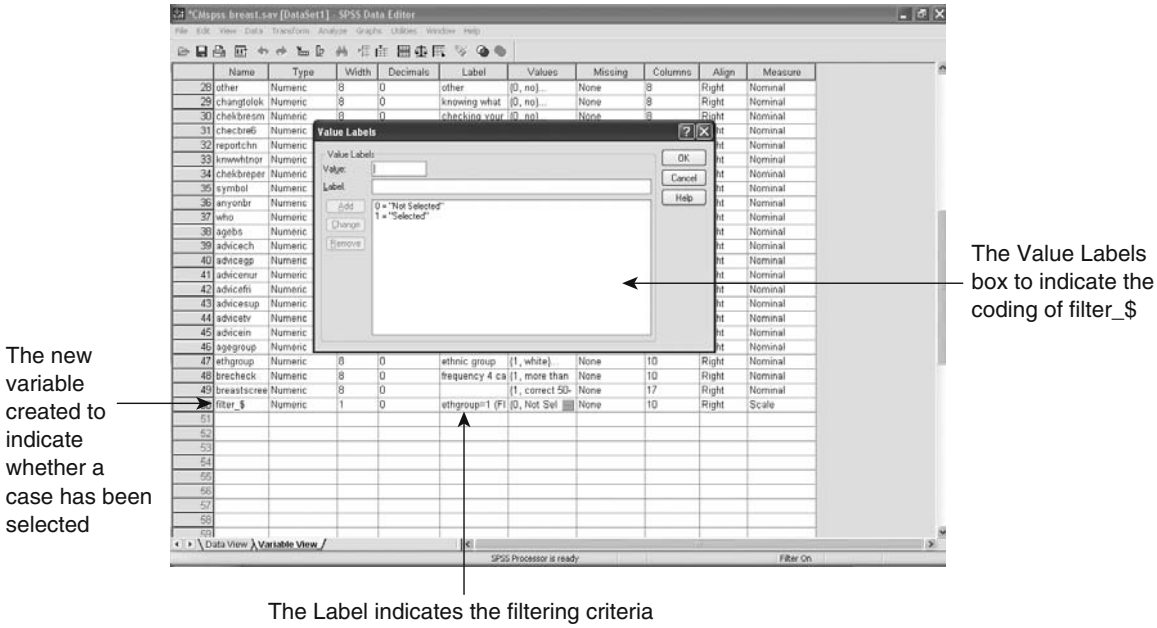

FIGURE 2.19 SCREENSHOT SHOWING HOW SELECTED AND NON-SELECTED CASES ARE INDICATED IN VARIABLE VIEW

SPLIT FILE

This is used if statistics are required separately by a given variable. For example, you might want to look at characteristics of members of a dataset by social class or age group. Summary statistics and other analyses can then be carried out on each defined group.

To split a file click on Data → Split File… to get the Split File dialog box shown in Figure 2.20. Select the radio button Organize output by groups. One or more variables in the dataset then have to be transferred to the Groups Based on: box. When this has been done, click OK. There will be no indicators that the file has been split when looking at the dataset in Data View or Variable View. It will only be apparent when data are analysed. For example, in Figure 2.21, the student breast cancer awareness data has been split by ethnic group then frequencies are shown for the variable 'Is increasing age a risk factor for breast cancer?' Frequencies for categorical data are explained further in Chapter 6.

Interpretation

The first set of statistics is for the group where ethnicity is missing (5 participants) and would not usually be reported. Following that it can be seen that 22% of white women thought increasing age was a risk factor for breast cancer and likewise 23% for black women and 13% for women from other ethnic groups.

To reverse splitting the file (so that all data are used again) click on Data → Split File… and select the radio button Analyze all cases, do not create groups, then click OK.

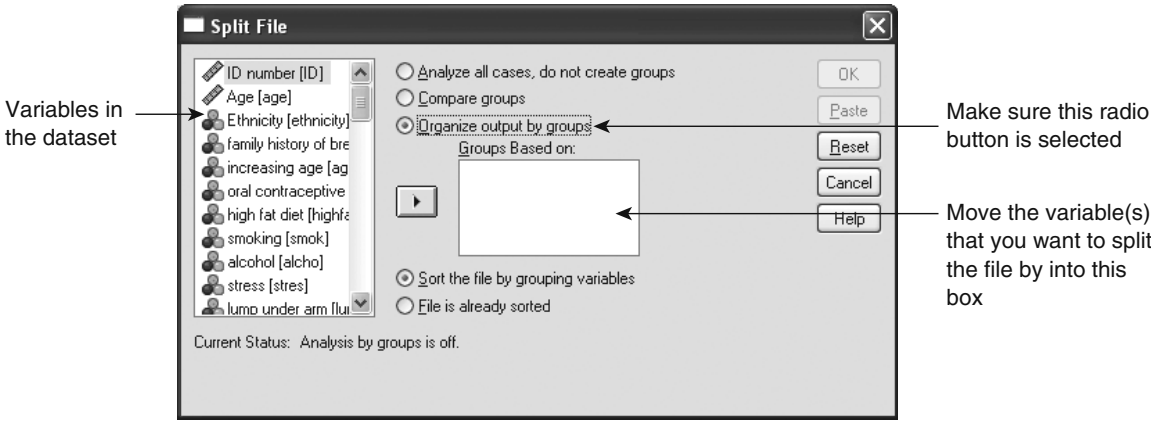

FIGURE 2.20 SPLIT FILE DIALOG BOX

ethnic group = .

Statistics[a]

increasing age

N	Valid	5
	Missing	0

[a]ethnic group =

increasing age[a]

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	3	60.0	60.0	60.0
	yes	2	40.0	40.0	100.0
	Total	5	100.0	100.0	

[a]ethnic group =

ethnic group = white

Statistics[a]

increasing age

N	Valid	65
	Missing	0

[a]ethnic group = white

increasing age[a]

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	51	78.5	78.5	78.5
	yes	14	21.5	21.5	100.0
	Total	65	100.0	100.0	

[a]ethnic group = white

FIGURE 2.21 (*Continued*)

ethnic group = black

Statistics[a]

increasing age[a]

N	Valid	52
	Missing	0

[a]ethnic group = black

increasing age[a]

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	40	76.9	76.9	76.9
	yes	12	23.1	23.1	100.0
	Total	52	100.0	100.0	

[a]ethnic group = black

ethnic group = other

Statistics[a]

increasing age

N	Valid	40
	Missing	0

[a]ethnic group = other

increasing age[a]

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	35	87.5	87.5	87.5
	yes	5	12.5	12.5	100.0
	Total	40	100.0	100.0	

[a]ethnic group = other

FIGURE 2.21 'IS INCREASING AGE A RISK FACTOR FOR BREAST CANCER?' BY ETHNIC GROUP AFTER SPLITTING THE FILE

SORTING DATA

Sometimes it is necessary for a given variable to be in numerical order, either ascending or descending. This may be to find an extreme value within a variable so that it can be checked or because merging to add variables requires the variable to be matched on to be sorted in ascending order. This is illustrated with the student breast cancer awareness study. To sort data in SPSS click on Data → Sort Cases… to give the Sort Cases dialog box shown in Figure 2.22. The variable(s) that are to be

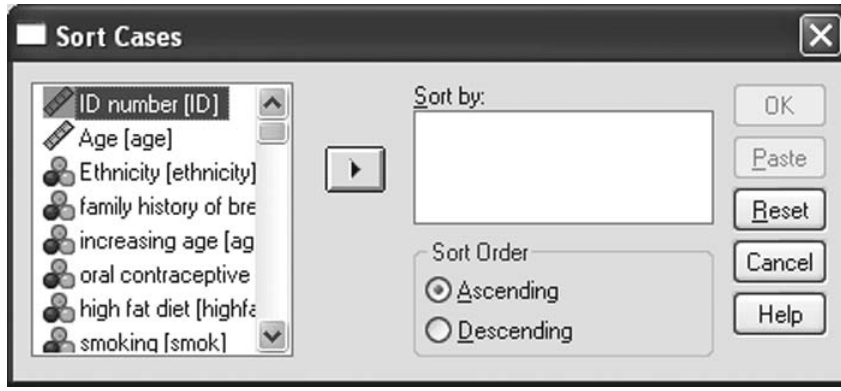

FIGURE 2.22 SORT CASES DIALOG BOX

merge, it is likely that the merge will be on ID number, so this would be transferred to the Sort by: box. When the sort by variables and their ordering has been declared, click OK to return to the sorted dataset.

MERGING

This is useful when there are two datasets containing either the same variables or the same participants, and their information needs to be combined to make one dataset for analysis purposes. For example, data may be collected at more than one time point, often analysis uses data collected at both (all) time points. This occurs when longitudinal datasets, such as the Millennium Cohort Study are being analysed, whereby data are supplied in files according to the time period they were collected in. The case where additional participants are added to a dataset occurs less frequently, but may occur when colleagues have been collecting the same data from different participants and have recorded it in different SPSS datasets, which need to be merged before analysis can take place.

The case where new variables are added is to be explored first. This example will use the student breast cancer awareness data. For this example the dataset has been bisected with most variables in one dataset and a few additional ones in another dataset. Before beginning the merge process, make sure that both datasets are sorted in ascending order on the variable which links the two datasets (in this dataset it is the ID number) otherwise the merge will not be executed. To execute a merge of datasets in SPSS to add variables, with the main dataset open, click on Data → Merge Files → Add Variables… to get the Add Variables to [open dataset] dialog box shown in Figure 2.23.

When a second dataset containing the additional variables has been selected, either through the datasets open or through browsing; click Continue to go to the Add Variables from [second dataset] dialog box (Figure 2.24). In this box, the majority of variables appear in the New Active Dataset: box, showing which variables will appear in the new dataset. Any variables that appear in both datasets (with exactly the same variable name) will be in the Excluded Variables: box. In this example this applies to ID. However, this variable is not to be excluded as this is the variable that is used to

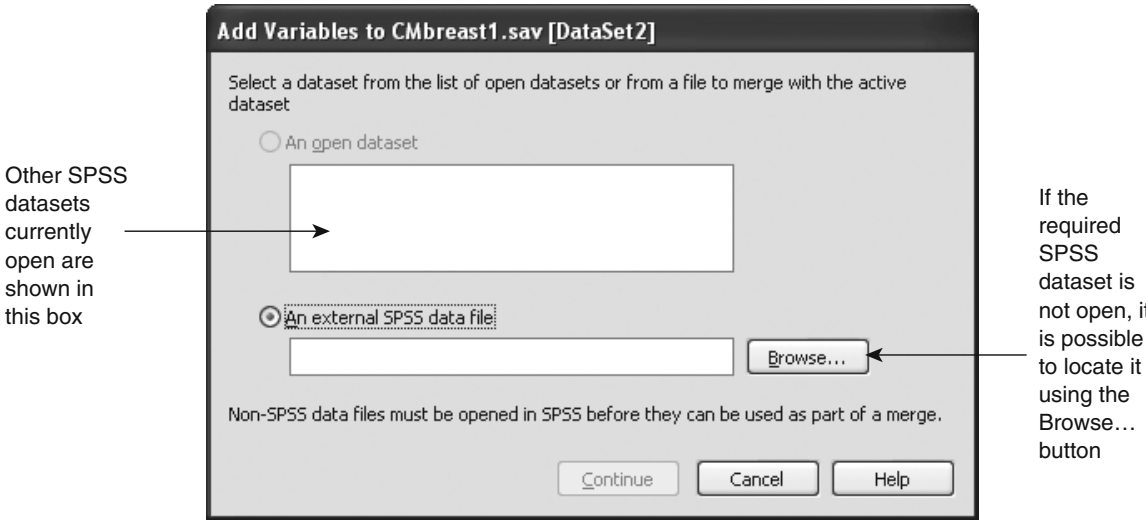

FIGURE 2.23 ADD VARIABLES TO [OPEN DATASET] DIALOG BOX

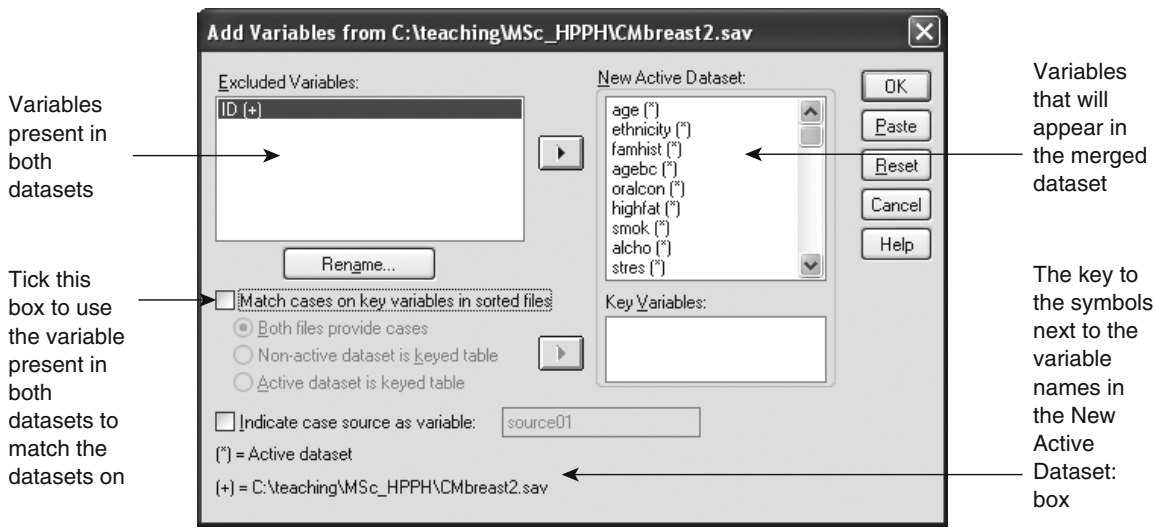

FIGURE 2.24 ADD VARIABLES FROM [SECOND DATASET] DIALOG BOX

match data from the two datasets; so that it can be ensured that data from participants in one dataset are from the same participants in the other dataset. For this to happen, the Match cases on key variables in sorted files box should be ticked, then ensuring the Both files provide cases radio button is selected (this is the default), then move the variable from the Excluded Variables: box to the Key Variables: box. Then click OK. The merge will then be complete and variables from both datasets will be visible in the first dataset. This should be saved so that the results of merging are not lost.

The second possible situation covered by merging is where data from additional participants is added to the main dataset. This is also going to be explained using the student breast cancer awareness study; this time the dataset is bisected horizontally so

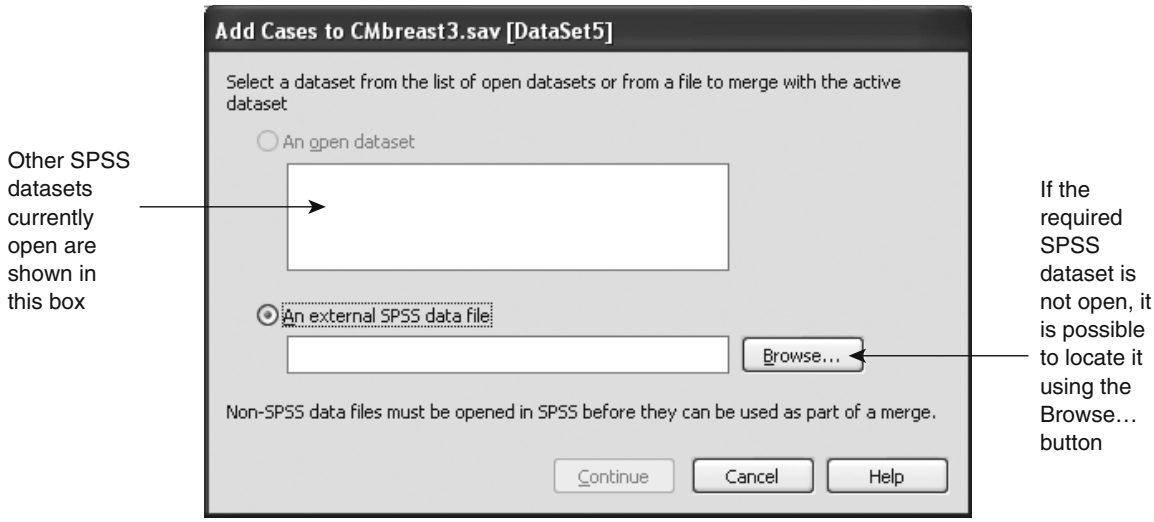

FIGURE 2.25 ADD CASES TO [OPEN DATASET] DIALOG BOX

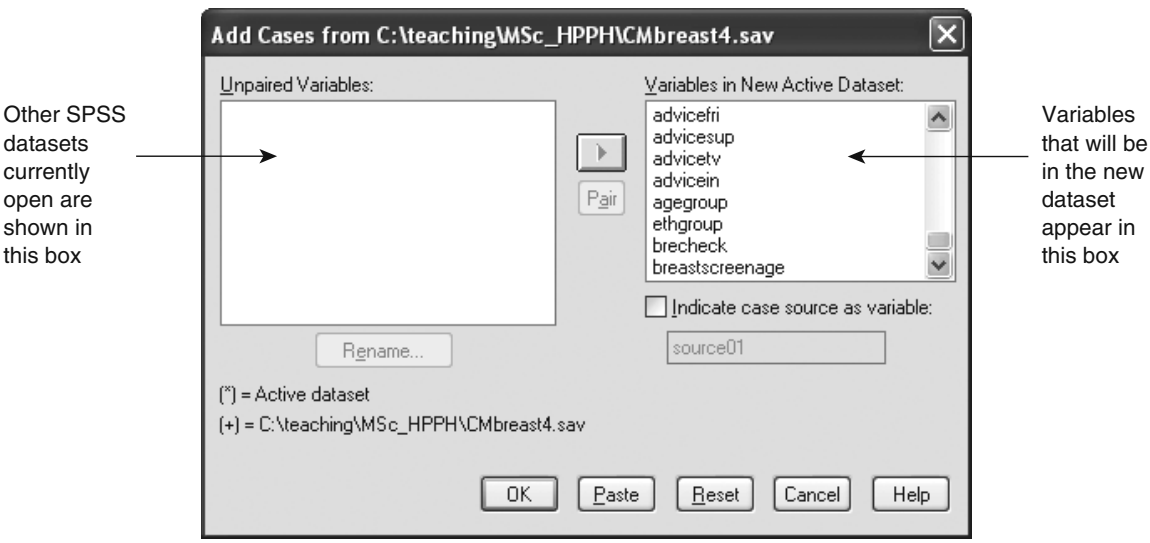

FIGURE 2.26 ADD CASES FROM [SECOND DATASET] DIALOG BOX

that the main dataset contains 100 participants whilst the second dataset contains the remainder of the participants. In SPSS, this merge is invoked by clicking on Data → Merge Files → Add Cases… when the main dataset is open to get the Add Cases to [open dataset] dialog box (Figure 2.25). The aim of this dialog box is to locate the dataset to be merged into the open dataset; this can either be a dataset that is already open (these will be shown in the An open dataset box), or a dataset that is not open, which can be searched for using the Browse… button.

When the dataset has been selected, click Continue to give the Add Cases from [second dataset] dialog box (Figure 2.26). The aim of this dialog box is show the variables

that will be in the new dataset (in the Variables in New Active Dataset: box) and where necessary pair variables with different variable names which are representing the same variable (shown in the Unpaired Variables: box). In the example shown in Figure 2.26 there are no variables where this is the case. When any unpaired variable issues have been resolved, click OK to give the complete dataset. Remember to save the new dataset.

SUMMARY

- Sometimes it is necessary to recode variables for analysis. To do this, click on Transform → Recode into different variables….
- New variables can be computed using Transform → Compute Variable ….
- Cases can be selected on the basis of variable(s) in the dataset using Data → Select Cases….
- The file can be split so that further analysis can be carried out by given variable(s), using Data → Split file….
- Variables within a dataset can be sorted in either ascending or descending order by clicking on Data → Sort Cases….
- Datasets can be merged so that either more variables are added or more cases (participants or subjects) are added. This can be done using Data → Merge Files.

EXERCISES

Open obesity.sav:

1 Recode the variable 'agegroup' into the same variable so that 0 = those aged 11 to 30 years, 1 = those aged 31 to 40 years and 2 = those aged 41+ years. Remember to change the value labels when the recoding is complete.
2 Make a new variable using BMI to produce a dichotomous variable indicating obese (BMI 30+) versus not obese. Remember to consider missing data.
3 Create dummy variables for the variable 'exercise'. Use 'never' as the reference category.
4 Sort the dataset to find the largest and smallest BMI.