

Chapter Three: Transforming Data




This chapter explains how to change, or transform, the values associated with your variables, like the values entered in the data entry process shown in Chapter 2. IBM SPSS can transform the values in several ways. IBM SPSS can:

- Combine values of a variable into several categories
- Create new variables out of old variables
- Select particular cases and analyze only these cases
- Weight cases so that some cases count more heavily than others

IBM SPSS provides many ways to transform data. Covered in this chapter are **recode**, **compute**, and **if**.

Recoding Variables

Recoding is a way of combining the values of a variable into fewer categories. Let me give a hypothetical example. Let's say you have conducted a survey and one of your demographic questions was the age of the respondent. Entering the actual age in years would be the simplest way of working with the data. But let's also say that you want to compare people of different age groupings. In other words, your data would be more useful if it was organized into collapsed categories, like "Young," "Middle age," and "Older." Using IBM SPSS you could reorganize the data so that you had these three groupings. There are two things you need to know before you recode the values. First, you need to decide the number of categories you want to end up with. Generally this will be determined by the way you plan to use the information. If you are going to analyze the data using a table where you crosstabulate two variables (see Chapter 5), you probably want to limit the number of new categories to three or four. The second thing you need to know is which of the old values are going to be combined into a new category. For example, you might do something like this.

The actual age of the respondent as originally recorded in the data file.	The new, collapsed, category.
 18 years through 35 years	Young
 36 years through 55 years	Middle age
 56 years of age or over	Older

Another example might be if respondents were asked how often they prayed, and the original responses were several times a day, once a day, several times a week, once a

week, less than once a week, once a month, once a year, or never. With recode we can combine the people who said “several times a day” with the people who said “once a day” and put all these respondents into a new category which we could call “a lot.” Similarly, we could combine the people who said “several times a week” with those who said “once a week” and call this category “sometimes” and combine those who said “less than once a week” and “once a month” and call this category “infrequently.” Those who said “once a year” or “never” could be combined into a fourth category called “hardly ever.” Recoding is the process in IBM SPSS that will do the above examples.

Starting IBM SPSS the way you were taught in Chapter 1, bring in the GSS10A file, as you did in Chapter 1. Our task is going to be to recode the variable called *age*, which is, of course, the respondent's age.

Click on Transform. Your screen will look like Figure 3-1.



Figure 3-1

Now we have two options: **Recode Into Different Variables** and **Recode Into Same Variables**. It is strongly suggested that the beginning student only use the **Recode Into Different Variables** option. If you make an error, your original variable is still in the file and you can try again. If you make an error using **Recode Into the Same Variables**, you have changed the original variable. If you also saved the file after doing this, and you did not have another copy of the file, you have just eliminated any chance of correcting your error.

Recoding Into Different Variables

The recoding into a different variable starts with giving the new variable a variable name. For example, if we recode into different variables we could combine ages into one set of categories and call this new variable *age1* and then recode ages into a different set of categories called *age2*. To do that, click on **Recode Into Different Variables**. Your screen will look like Figure 3-2.

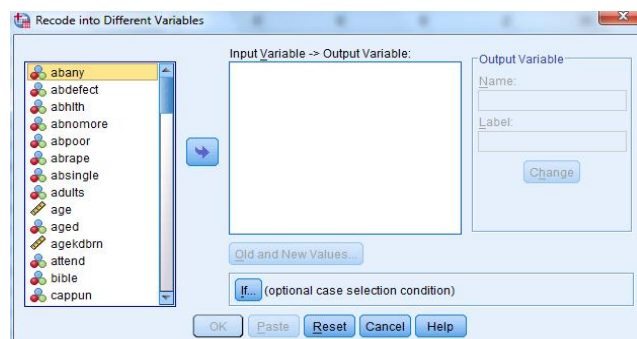


Figure 3-2

Find *age* in the list of variables on the left and click on it to highlight it, and then click on the arrow just to the left of the big box in the middle of the window. This will move *age* into the list of variables to recode.

You want to give a name to this new variable so click in the Name box under Output Variable and type the name **age1** in this box. You can even type a variable label for this new variable in the Label box just below the Name box. Try typing **Age in Four Categories** as your label. Click on the **Change** button to tell IBM SPSS to make these changes. Your screen will look like Figure 3-3.

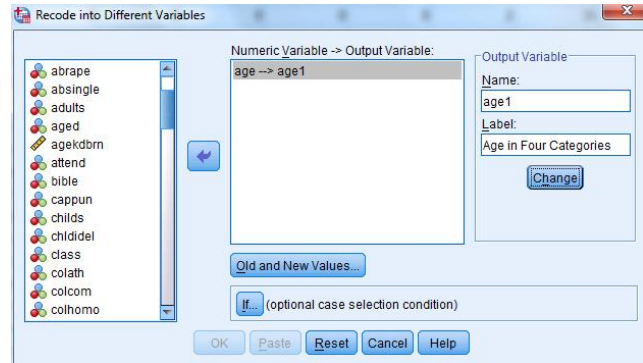


Figure 3-3

Now we have to tell IBM SPSS how to create these categories referred to as values. Click on the **Old and New Values** button at the bottom of the window. The screen will look like Figure 3-4.

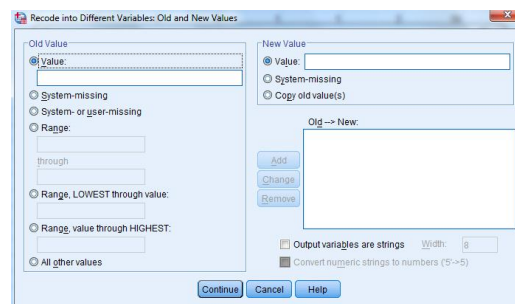


Figure 3-4

There are several options. You can change a particular value into a new value by entering the value to be changed into the Old Value box and the new value into the New Value box and then clicking on **Add**. You will usually change one “real” value to another “real” value. For example, change 18 thru 35 into value 1. (The next paragraph tells you how to do this.) There are also other options¹.

As you can tell from the previous example, you can also change a range of values into a new value. Click on the fourth bubble from the top labeled **Range**. Notice how this marks this choice by filling in the bubble. Then type **18** (the youngest age in the data set) in the box above **Through**, click on the box below **Through**, and type **29** in that box. Then click on **Value** just below **New Value** and type **1** in that box. This will have IBM SPSS combine all ages from 18 through 29 into a single category and give it the value of 1. Then click on **Add**.

Repeat this process for the other categories. Click on the box under **Range** and type **30** in the box above **Through**, click on the box below **Through**, and type **49** in that box. Click on **Value** just below **New Value** and type **2** in that box and click on **Add**. Do the same thing for the category 50 to 69 (give this a new value of **3**) and the category 70 to 89 (the largest age in the data set). Give this last category a new value of **4**. Your screen should look like Figure 3-5.

¹ For example, you can work with what SPSS calls “system-missing” values. All blanks will automatically be changed to system-missing values. You can change these system-missing values into another value, or you can change both the system-missing values and the missing values that you define into another value.

To change one of your categories, highlight that category in the Old->New box, make the changes in the Old or New Value box, and then click on **Change**. The new category should appear in the Old->New box. To remove a category, highlight it and click on **Remove**.

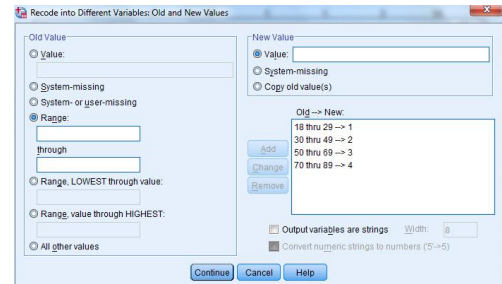


Figure 3-5

Now we want IBM SPSS to carry out the recoding. Click on **Continue** at the bottom of the window. This will take you back to the Record into Different Variables box. Click on **OK** and IBM SPSS will carry out your commands. IBM SPSS will show you the command it just executed. When it says that the IBM SPSS Processor is **Ready** at the bottom of the window you know that IBM SPSS has finished with the recoding.

Statistics

Age in Four Categories

N	Valid	2041
	Missing	3

Age in Four Categories

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid				
1.00	407	19.9	19.9	19.9
2.00	750	36.7	36.8	56.7
3.00	668	32.7	32.8	89.5
4.00	215	10.5	10.5	100.0
Total	2041	99.8	100.0	
Missing	System	3	.2	
Total		2044	100.0	

Figure 3-6

Click on **Analyze**, then point your mouse at **Descriptive Statistics**, and then click on **Frequencies**. Notice that *age1*² appeared in the list of variables on the left. Click on it to highlight it and click on the arrow to move it to the Variables box. Then click on **OK**. An output window will open. Your screen will look like Figure 3-6.

Let's take a look at the data matrix. Click on **Window** in the menu bar. In the box that is opened you will see a list of all the windows you have opened. One of these windows will be called **GSS10A - IBM SPSS Data Editor**. Click on

	spkml	spkrac	thnkself	trust	thhours	vote04	vote08	zodiac	wtss	year	age1	var
1	0	0	1	1	0	3	1	9	46	2010	1.00	
2	0	0	2	3	6	1	1	9	93	2010	3.00	
3	2	2	2	3	-1	2	2	12	93	2010	1.00	
4	2	2	0	0	5	1	1	8	46	2010	4.00	
5	0	0	4	3	5	1	1	10	46	2010	4.00	
6	0	0	1	3	0	1	1	3	46	2010	2.00	
7	0	0	1	1	0	1	2	6	46	2010	2.00	
8	1	2	0	0	5	1	1	3	2.53	2010	3.00	
9	0	0	1	1	0	3	3	5	93	2010	2.00	
10	1	2	5	3	-1	2	1	8	93	2010	1.00	
11	2	2	0	0	3	8	1	3	46	2010	2.00	
12	0	0	1	1	0	1	1	2	93	2010	3.00	
13	1	1	1	1	-1	1	1	5	1.68	2010	3.00	
14	2	2	1	1	-1	3	3	5	93	2010	2.00	
15	2	2	0	0	3	1	1	3	84	2010	4.00	
16	2	1	2	1	-1	1	1	1	93	2010	4.00	
17	2	2	0	0	3	1	1	2	46	2010	3.00	
18	0	0	2	1	2	2	1	5	93	2010	2.00	
19	1	1	1	1	-1	1	1	11	46	2010	4.00	
20	1	1	0	0	2	1	1	3	93	2010	3.00	
21	0	0	1	3	5	1	1	7	46	2010	4.00	
22	1	2	0	0	2	1	1	5	93	2010	3.00	
23	1	1	1	1	-1	1	1	10	93	2010	1.00	
24	2	2	0	0	2	3	2	6	1.68	2010	1.00	
25	0	0	9	1	3	1	1	99	93	2010	3.00	
26	2	2	1	3	-1	2	1	11	93	2010	2.00	
27	0	0	1	3	1	1	1	7	46	2010	2.00	
28	0	0	1	1	4	1	1	12	93	2010	3.00	
29	0	0	1	1	2	1	1	6	93	2010	2.00	
30	1	1	1	1	-1	3	2	7	93	2010	1.00	
31	0	0	4	1	2	1	1	8	93	2010	3.00	
32	0	0	3	1	4	1	1	3	93	2010	4.00	
33	1	2	0	0	2	2	1	8	46	2010	1.00	
34	0	0	2	3	3	2	2	4	93	2010	2.00	

Figure 3-7

that line and the data matrix window will be moved to the foreground and you will see it on your screen. Use the scroll bar in the lower-right part of the window to scroll to the right until you see a column titled *age1*. (It will be the last column in the matrix.) This is the new variable you just created. Your screen should look like Figure 3-7.

² If your list shows labels, you can change the display. Check Footnote 1 in Chapter 1.

Use the other scroll bar to scroll down and see the values in this variable. Look back at Figure 3-6 and you will see that there are no value labels for categories 1 through 4 for the new variable *age1*.

If you want the output to give you more information about what each category means, you need to insert value labels. To do this, point your mouse at the variable name at the top of the column (*age1*) and double click. This will open the Variable View tab in the Data Editor. Now you're going to enter labels for the values in the recoded variable using what you learned in Chapter 2.

Click in the Values box and you will see a small gray button in the right-hand side of the box. Point your mouse at this box and click. This will open the Value Labels box. You will see two more boxes, Value and Label. Click in the Value box and type the value **1**.

Then click in the Label box and type the label for the first category, **under 30**. Then click on **Add** and the new label will appear in another box just to the right of the **Add** button. Then click in the Value box and type the value **2** and type the label for the second category, **30 to 49**, and click on **Add**. Do this for values **3** and **4**.

If you make a mistake you can use the **Change** and **Remove** buttons, which work the same way we just described. Your screen should look like Figure 3-8.

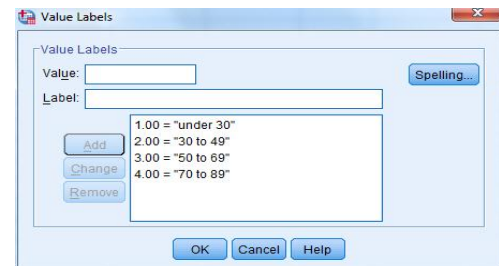


Figure 3-8

Click on **OK**. Now click on **Analyze**, point your mouse at **Descriptive Statistics**, and then click on **Frequencies** and rerun the frequencies distribution for *age1*. This time it should have the value labels you just entered on the output. Your screen should look like Figure 3-9.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	under 30	407	19.9	19.9	19.9
	30 to 49	750	36.7	36.8	56.7
	50 to 69	668	32.7	32.8	89.5
	70 to 89	215	10.5	10.5	100.0
	Total	2041	99.8	100.0	
Missing	System	3	.2		
Total		2044	100.0		

Figure 3-9

We said that recoding into different variables allowed you to recode a variable in more than one way. Let's recode *age* again, but this time let's recode age into three categories—18 through 34, 35 to 59, and 60 and over. Let's call this new variable *age2*. Retracing the steps you used to create *age1*, recode *age* into *age2*.

Be sure to click on **Reset** in the Recode into Different Variables box to get rid of the recoding instructions for *age1*. When you are done, do a frequency distribution for *age2*.

There are two more important points to discuss. Look back at Figure 3-4. It shows the Recode into Different Variables: Old and New Values box. There are three options in the Old Value box that we haven't discussed. Two are different ways of entering ranges.

You can enter the lowest value of the variable through some particular value and you can enter some particular value through the highest value of the variable. Make sure that you do not include your missing values in these ranges or your missing values will become part of that category. For example, if 99 is the missing value for *age*, then recoding 70 through highest would include the missing values with the oldest age category. This is probably not what you want to do. So be careful.

Here is another important point. What happens if you don't recode a particular value? If it is a missing value, it retains its status as a missing value in the new variable. But what if it isn't a missing value? Any value (other than a missing value) that is not recoded is changed into a system-missing value. If you want to leave a value in its original form, then click on All Other Values in the Old Value box and click on Copy Old Value in the New Value box and then click on Add.

Recoding into the Same Variable

Now we are going to recode and have the recoded variable replace the old variable. This means that we will not create a new variable. We will replace the old variable with the recoded variable, but remember the warning given you earlier in this chapter. Click on **Transform** and click on **Recode Into Same Variables**. Let's recode the variable called *pray*. Find *pray* on the list of variables on the left, click on it to highlight it, and then click on the arrow to the left of the Variable box. This will move the variable *pray* into the big box in the middle of the window. Click on the **Old and New Values** button. This will open the Recode into Same Variables: Old and New Values box. Your screen should look like Figure 3-10.

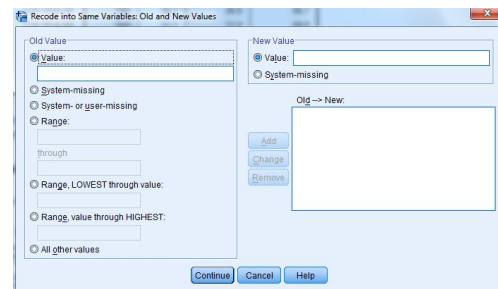


Figure 3-10

This looks very much like the box you just used (see Figure 3-4). Combine the values 1 and 2 by clicking on the fourth circle from the top under **Old Value** and entering a **1** in the box above **through** and a **2** in the box below **through** and then entering a **1** in the New Value box and then clicking on **Add**. Now combine values 3 and 4 into a category called **2**. Then combine values 5 and 6 into a third category called **3**. Since this is not a new variable, it will still be called *pray*.

You will want to change the value labels. Find the variable *pray* in the **Data View** by scrolling to that variable. Point your mouse at the variable name (*pray*) and double click. This will open the **Variable View** tab in the **Data Editor**. Click in the **Values** box and then click on the small gray box and change the labels to an appropriate name for the new variable. You will have to use the **Change** and **Remove** buttons to do this. Follow the instructions we just went through for recoding into different variables.

When you finish, click on **Analyze**, then point your mouse at **Descriptive Statistics**, then click on **Frequencies** and move *pray* over to the Variables box and click on **OK**. Your screen should look like Figure 3-11.

HOW OFTEN DOES R PRAY					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	one or more times a day	1158	56.7	57.2	57.2
	several times a week	381	18.7	18.8	76.1
	less than once a week	484	23.7	23.9	100.0
	Total	2023	99.0	100.0	
Missing	DK	12	.6		
	NA	8	.4		
	Total	21	1.0		
Total		2044	100.0		

Figure 3-11

When you recode into the same variable, a value that is not recoded stays the same as it was in the original variable. If we had decided to keep “never” (value 6) as a separate category, we could have left it alone and it would have stayed a 6. Or we could have changed it to another value such as 4. This is an important difference between recoding into the same and different variables.

Recoding is a very useful procedure and one that you will probably use a lot. It's worth spending time practicing how to recode so you will be able to do it with ease when the time comes.

Creating New Variables Using COMPUTE

You can also create new variables out of old variables using **compute**. There are seven variables in the data set we have been using that ask respondents if they think a woman ought to be able to obtain a legal abortion under various scenarios. These are the variables *abany* (woman wants abortion for any reason), *abdefect* (possibility of serious birth defect in baby), *abhlth* (woman's health is seriously threatened), *abnomore* (woman is married and doesn't want any more children), *abpoor* (woman is poor and can't afford more children), *abrape* (pregnant as result of rape), and *absingle* (woman is not married). Each variable is coded 1 if the respondent says yes (ought to be able to obtain a legal abortion) and 2 if the person says no. The missing values are 0 (not applicable, question wasn't asked), 8 (don't know), and 9 (no answer).

Compute will allow us to combine these seven variables, creating a new variable that we will call *abortion*. If a person said yes to all seven questions, the new variable would equal 7 and if he or she said no to all seven questions, the new variable would equal 14. But what about missing values? If any of the seven variables have a missing value, then the new variable would be assigned a system-missing value.

To use **Compute** to do this, click on **Transform** and then click on **Compute**. Your screen should look like Figure 3-12.

Type the name of the new variable, *abortion*, in the Target Variable box. Then enter the formula for this new variable in the Numeric Expression box. There are two ways to do this. One method is to click on the first of the seven variables, *abany*, in the list of variables on the left, then click on the arrow to the right of this list. This will move *abany* into the Numeric Expression box. Now click on the plus sign and the plus sign moves into the box.

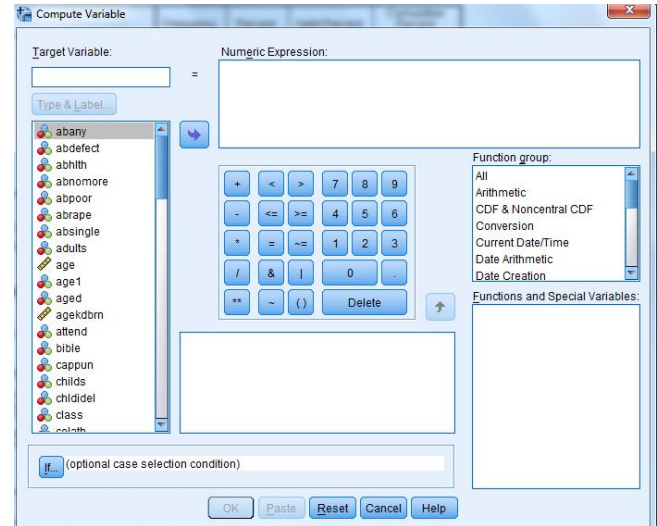


Figure 3-12

Continue doing this until the box contains the following formula: *abany* + *abdefect* + *abhlth* + *abnomore* + *abpoor* + *abrape* + *absingle*. (Don't type the period after *absingle*.) If you make a mistake, just click in the Numeric Expression box and use the arrow keys and the delete and backspace keys to make corrections. Your screen should look like Figure 3-13.

Click on **OK** to indicate that you want IBM SPSS to create this new variable. You can use the scroll bar to scroll to the far right of the matrix and view the variable you just created. A second way to enter the formula in the Numeric Expression box is to click in the box and type the formula directly into the box using the keyboard.

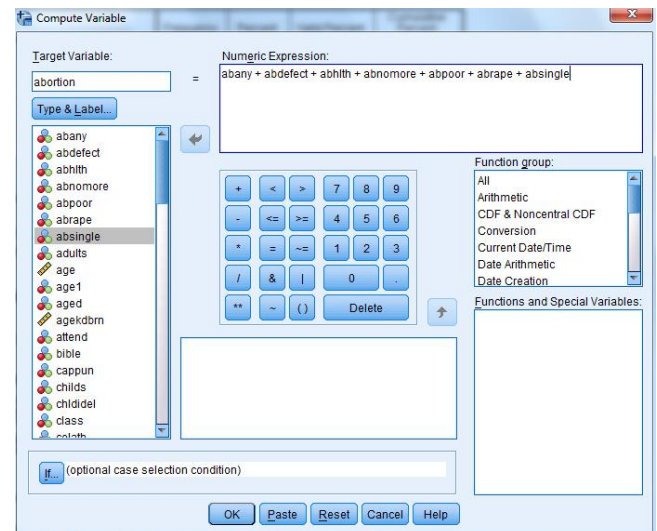


Figure 3-13

You can add variable and value labels to this variable by pointing your mouse at the variable name (*abortion*) at the top of the column in the data matrix and double clicking. This will open the **Variable View** tab in the **Data Editor**. You can enter the variable and value labels the way you were taught earlier in this chapter and in Chapter 2.

Enter the variable label **Sum of Seven Abortion Variables**. Enter the value label **High Approval** for the value 7 and **Low Approval** for the value 14. (Remember that seven means they approved of abortion in all seven scenarios and fourteen means they disapproved all seven times.) Click on **OK**.

You should check your new variable to see that it was calculated correctly. Go to **Analyze**, then **Descriptive Statistics**, and then **Frequencies**. Click on **Reset** to get rid of what is already in the box. Find the variable *abortion*, highlight it and click on the arrow to the left of the Variables box. Then click on **OK**. Your

		abortion			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	high approval	409	20.0	36.9	36.9
	8.00	66	3.2	6.0	42.8
	9.00	64	3.1	5.7	48.5
	10.00	79	3.8	7.1	55.6
	11.00	180	8.8	16.2	71.8
	12.00	122	6.0	11.0	82.9
	13.00	86	4.2	7.8	90.6
	low approval	104	5.1	9.4	100.0
	Total	1109	54.2	100.0	
Missing	System	935	45.8		
Total		2044	100.0		

Figure 3-14

screen should look like Figure 3-14. The lowest number should be 7 and the highest number should be 14. Do you remember why?

One of the problems with this approach is that the new variable (*abortion*) will be assigned a system missing value if one or more of the original variables have a missing value. We can avoid this problem by summing the values of the original variable and dividing by the number of variables with valid values. For example, if six of the seven original variables had valid values, then we would divide the sum by six. We can also tell IBM SPSS to create this new variable only if at least four of the original variables have valid values. If fewer than four of the original variables have valid values, IBM SPSS will assign it a system missing value.

We can do this by clicking on **Transform** and then on **Compute** and entering the new variable name in the Target Variable box. Let's call this variable *abort*. In the Function Group box, scroll down and click on **Statistical**. This will list the statistical functions in the Functions and Special Variables box. Double-click on **Mean**. Your screen should look like Figure 3-15.

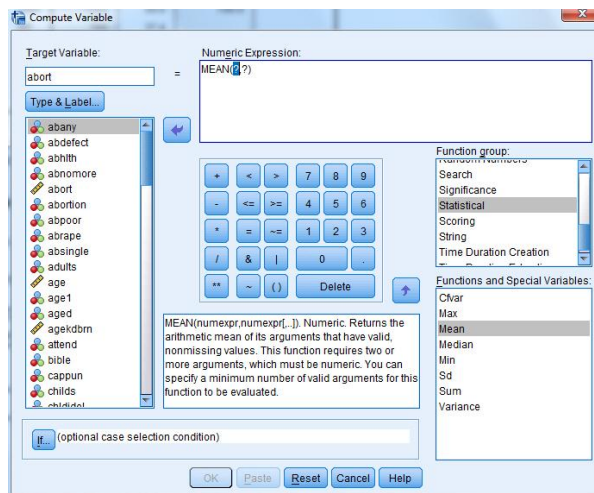


Figure 3-15

Notice that **Mean(?,?)** has been inserted in the Numeric Expression box. What you want to do is to replace the **(?,?)** with the list of the seven original variables. It should now read (*abany, abdefect, abhlth, abnomore, abpoor, abrape, absingle*). All that is left is to tell IBM SPSS that you want to create this new variable only if at least four of the

original variables have valid values. Do this by entering **.4** following Mean so the expression reads **MEAN .4 (abany, abdefect, abhlth, abnomore, abpoor, abrape, absingle)**. Your screen should look like Figure 3-16.

Click on OK and run a frequency distribution to see what your new variable looks like. Your screen should look like Figure 3-17.

Try creating another variable. Two of the variables in the data set are the number of years of education of the respondent's father (*paeduc*) and of the respondent's mother (*maeduc*). If we divide *paeduc* by *maeduc* we will get the ratio of the father's education to the mother's education. Any value greater than one will mean that the father has more education than the mother and any value less than 1 means the mother has more education than the father. Any value close to 1 means that the father and mother have about the same education.

We have a small problem though. If the mother's education is zero, then we will be dividing by zero, which is mathematically undefined. Let's recode any value of zero for *maeduc* so it becomes a one. This will avoid dividing by zero and still give us a useful ratio of father's to mother's education. Click on Transform, then click on Recode into same variables. (You may need to click on Reset to get rid of the recoding instructions used earlier.) Move *maeduc* into the Variables box by highlighting it in the list of variables on the left and clicking on the arrow to the right of this list. Click on Old and New Values and then type **0** into the Value box under Old Value and click in the Value box under New Value. Type **1** in this box and click on Add. Your screen should look like Figure 3-18.

Click on Continue and then on OK in the Recode Variables box. Now we have changed each 0 for *maeduc* into a 1.

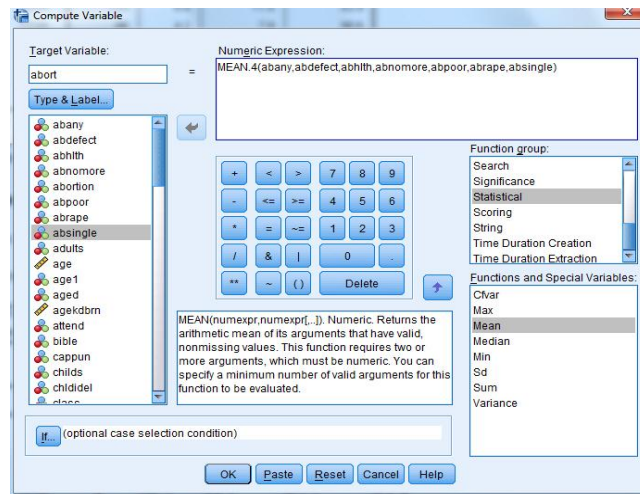


Figure 3-16

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	428	21.0	34.1	34.1
	1.14	66	3.2	5.2	39.3
	1.17	6	.3	.5	39.8
	1.20	1	.1	.1	39.9
	1.25	5	.2	.4	40.3
	1.29	64	3.1	5.1	45.3
	1.33	8	.4	.7	46.0
	1.40	12	.6	.9	46.9
	1.43	79	3.8	6.2	53.2
	1.50	19	.9	1.5	54.7
	1.57	180	8.8	14.3	68.9
	1.60	5	.2	.4	69.3
	1.67	17	.8	1.4	70.7
	1.71	122	6.0	9.7	80.4
	1.75	1	.0	.1	80.5
	1.80	6	.3	.4	80.9
	1.83	19	.9	1.5	82.5
	1.86	86	4.2	6.8	89.3
	2.00	135	6.6	10.7	100.0
Total		1258	61.6	100.0	
Missing	System	786	38.4		
Total		2044	100.0		

Figure 3-17

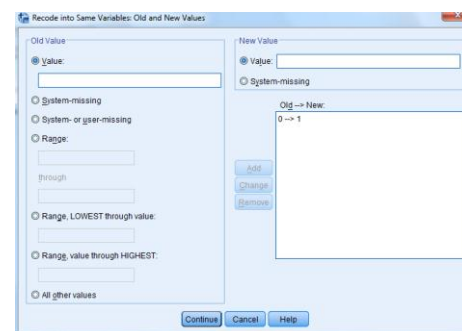


Figure 3-18

To create our new variable, click on **Transform** and then on **Compute**. (If necessary, click on **Reset** to get rid of the formula for the *abort* variable you just created.) Call this new variable *ratio*. So type **ratio** in the Target Variable box. Now we want to write the formula in the Numeric Expression box. Click in the list of variables on the left and scroll down until you see *paeduc*. Click on it to highlight it and click on the arrow to the right of the list to move it into the Numeric Expression box.

IBM SPSS uses the slash (/) to indicate division, so click on the / in the box in the center of the window. Click on the list of variables again and scroll up until you see *maeduc* and click on it to highlight it. Move it to the Numeric Expression box by clicking on the arrow. Your screen should look like Figure 3-19.

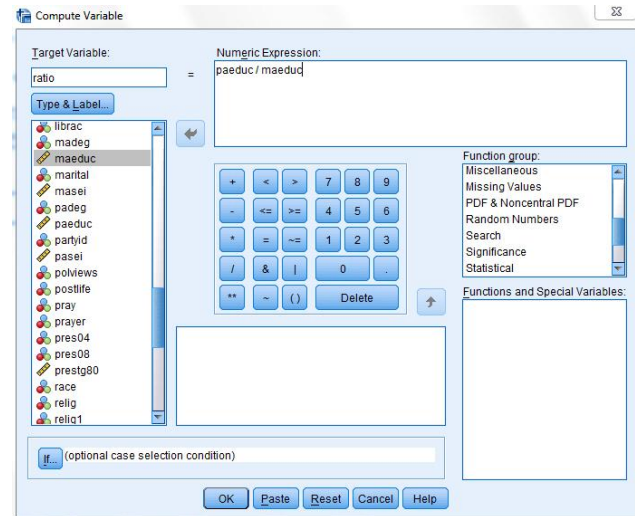


Figure 3-19

Click on **OK** and IBM SPSS will create your new variable. Use the scroll bar to scroll to the right in the data matrix until you can see the new variable you called *ratio*. Scroll up and down so you can see what the values of this variable look like. You may want to do a frequencies distribution as a check to make sure the new variable was created correctly.

After looking at the frequencies distribution, it is obvious that it would be easier to understand if we grouped some of the scores together, so create a new variable by recoding it into a Different Variable. Click on **Transform** and then click on **Recode Into Different Variables**. Find the variable *ratio* in the list of variables on the left and click on it to highlight it. (Again, you may have to click **Reset** if there is old information still in the boxes.) Click on the arrow to the right of this list to move it into the box in the middle of the window. Type **ratio1** in the Name box under Output Variable and type **Recoded Ratio** in the label box. Then click on **Change**.

Click on **Old and New Values** to open the Recode Into Different Variables: Old and New Values box. Click on the fifth bubble from the top under **Old Value** and then type **0.89** in the box to indicate that you want to recode the lowest value through 0.89. Click on the **Value** box under **New Value** and type **1** in that box, and then click on **Add**. Click on the fourth bubble from the top under **Old Value** and type **0.90** in the

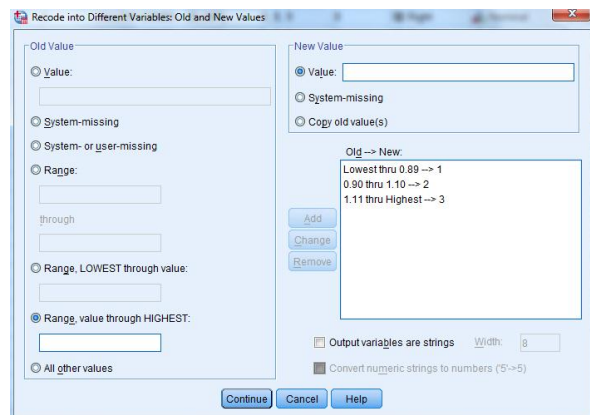


Figure 3-20

box above through and **1.10** in the box below. Then type **2** in the Value box under New Value and click on Add. Finally, click on the sixth bubble from the top under Old Value and type **1.11** in the box. Type **3** in the Value box under New Value and click on Add. Your screen should look like Figure 3-20. Click on Continue and then on OK in the Recode Into Different Variables box.

Let's add value labels to the new values. Find the variable *ratio1* in the data matrix and double click on the variable name, *ratio1*. This will open the Variable View tab in the Data Editor. Click the Values box and then click in the small gray box and enter the labels. Type **1** in the Value box and **under 0.90** in the Value Label box and then click on Add.

Do this twice more to add the label **0.90 through 1.10** to the value 2 and **over 1.10** to the value 3. Your screen should look like Figure 3-21. (This should seem familiar to you now.)

Click on OK in the Value Labels box. Run a frequencies distribution on the new variable to double-check your work. Your screen should look like Figure 3-22.



Figure 3-21

The first category (under 0.90) means that Father's Education was less than 90% of Mother's Education. The second category (0.90 through 1.10) means that Father's and Mother's Education were about the same, while the third category (over 1.10) means that Father's Education was more than 110% of Mother's Education. You can see that about 46% of the respondents have fathers and mothers with similar education, while about 27% have fathers with substantially less education than the mother and another 27% have fathers with substantially more education than the mother.

ratio1					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	under 0.90	386	18.9	27.1	27.1
	0.90 through 1.10	662	32.4	46.5	73.5
	over 1.10	377	18.4	26.5	100.0
	Total	1425	69.7	100.0	
Missing	System	619	30.3		
	Total	2044	100.0		

Figure 3-22

You have already seen that IBM SPSS uses + for addition and / for division. It also uses - for subtraction, * for multiplication, and ** for exponentiation. There are other arithmetic operators and a large number of functions (e.g., square root) that can be used in compute statements.

Creating New Variables Using IF

The IF command is another way to create new variables out of old variables. Perhaps we want to compare the level of education of each respondent's father to that of his or her mother. Now, however, we're not interested in the precise ratio, but just want to know if the father had more education than the mother, the same amount, or less. We'll create a new variable that will have the value 1 when the father has more education than the mother, 2 when both have the same amount of education and 3 when the mother has more education.

Click on Transform and then click on Compute. (You may need to click on Reset to get rid of the instructions for creating *ratio*.) Type the name of the new variable, *compeduc*, in the Target Variable box. Then click on the Numeric Expression box and enter 1. So far, this is just like what you did in the previous section. Your screen should look like Figure 3-23. This time, however, click on IF.

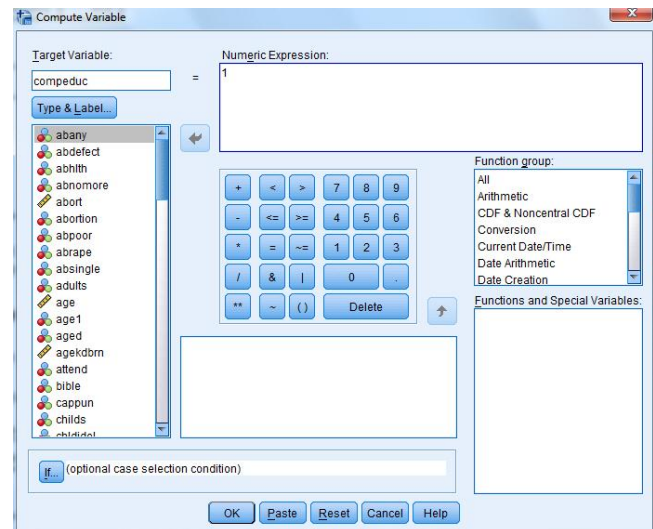


Figure 3-23

Click on Include if case satisfies the condition. Find *paeduc* in the list of variables on the left and click on it to highlight it. Then click on the arrow to the right of this list. This will move *paeduc* into the box to the right of the arrow. Now click on > (greater than). Find *maeduc* in the list of variables on the left, click on it, and click on the arrow to add *maeduc* to the formula. (Alternatively, you could click on the box to the right of the arrow and directly enter the formula, *paeduc* > *maeduc*.) Your screen should look like Figure 3-24. Now click on Continue.

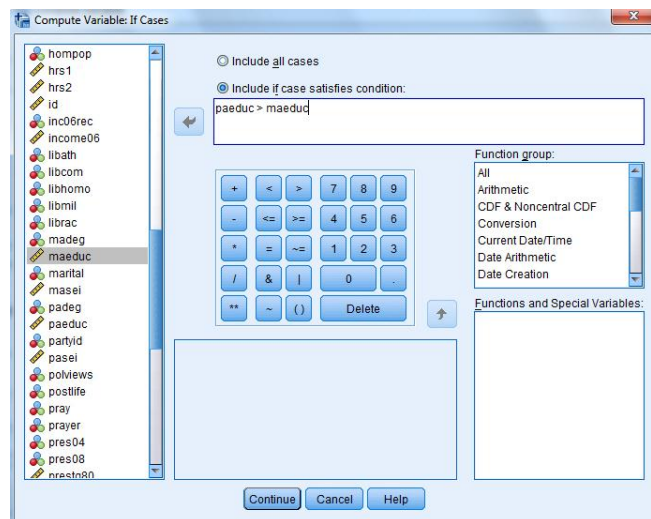


Figure 3-24

Click on OK. Now repeat the same procedures as above, but this time setting the value of *compeduc* to 2 (instead of 1) and the formula to *paeduc* = *maeduc*. When you are asked if you want to Change existing

variable, click on OK. Now repeat the procedures a third time, but change the value of *compeduc* to 3 and the formula to *paeduc < maeduc*.

You can add variable and value labels to this variable, just as you did earlier in this chapter and in Chapter 2. To do this, point your mouse at the variable name at the top of the column (*compeduc*) and double click. This will open the Variable View tab in the Data Editor. Click in the Values box and then in the small gray button in the right-hand side of the box. Point your mouse at this box and click. This will open the Value Labels box. Click in the box next to Value and type 1. Click on the box next to value label (or press the Tab key) and type **Dad More**. Now click on Add. Repeat this procedure for values 2 and 3, labeling them **Same** and **Mom More** respectively. Click on Continue, then on OK. Now run frequencies on your new variable to double-check your work.

Using Select Cases

IBM SPSS can also select subsets of cases for further analysis. One of the variables in the data set is the respondent's religious preference (*relig*). The categories include Protestant (value 1), Catholic (2), Jewish (3), none (4), as well as other categories. The missing values are 98 (don't know) and 99 (no answer). We might want to select only those respondents who have a religious preference for analysis. We can do this by using the Select Cases option in IBM SPSS.

Click on Data and then on Select Cases. This will open the Select Cases box. Your screen should look like Figure 3-25. Notice that All Cases is currently selected. (The circle to the left of All Cases is filled in to indicate that it is selected.) We want to select a subset of these cases so click on the circle to the left of If condition is satisfied to select it. At the bottom of the window it says that unselected cases are filtered out. This means that the cases you do not select can be used later if you click on All cases. If you had selected Delete unselected

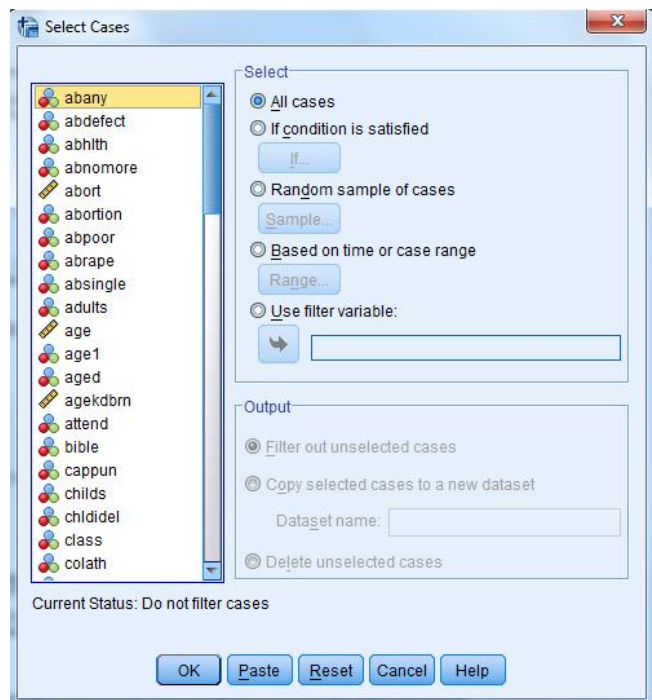


Figure 3-25

cases, these unselected cases could not be used later. You should be very careful about saving a file after you have deleted cases because they are gone forever in that file. (You could, of course, get another copy of the data file by clicking on File and on Open.)

Click on IF (below the button that says If condition is satisfied) and this will open the Select Cases: IF box. Scroll down the list of variables on the left until you come to *relig* and then click on it to highlight it. Click on the arrow to the right of this list to move *relig* into the box in the middle of the window. We want to select all cases that are not equal to 4 so click on the \neq sign. This symbol means “not equal to.” Now click on 4 and the expression in the box will read *relig* \neq 4 which means that the variable *relig* does not equal 4 (the code for no religious preference). Your screen should look like Figure 3-26. Click on Continue and then on OK in the Select Cases box.

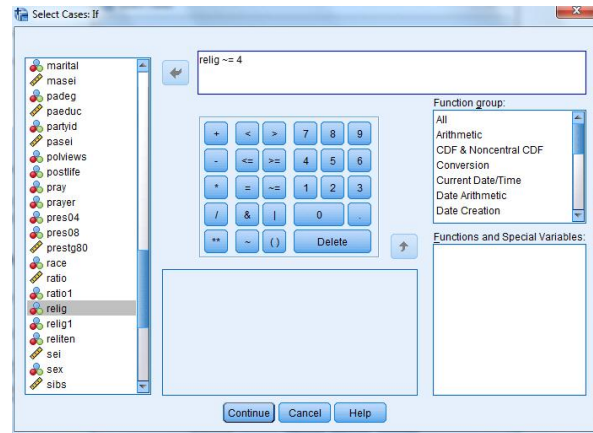


Figure 3-26

Run a frequencies distribution and check that your new variable gives you a reasonable range of values. Your screen should look like Figure 3-27.

RS RELIGIOUS PREFERENCE					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	PROTESTANT	1036	62.1	62.1	62.1
	CATHOLIC	513	30.8	30.8	92.9
	JEWISH	33	2.0	2.0	94.9
	OTHER	85	5.1	5.1	100.0
	Total	1667	100.0	100.0	

Figure 3-27

There are no respondents without a religious preference (value 4) in this table because you selected only those cases with values not equal to four.

What if we wanted to analyze only Protestants and Catholics? Click on Data and then on Select Cases. Click on Reset to eliminate what you had entered previously. Click on IF condition is satisfied and then on IF. Scroll down the list of variables and click on *relig* and then click on the arrow to the right of the list to move it into the box. Click on = and then on 1 so the expression in the box reads *relig* = 1. IBM SPSS uses the symbol & for AND and the symbol | for OR. We want all cases for which *relig* is 1 or 2.

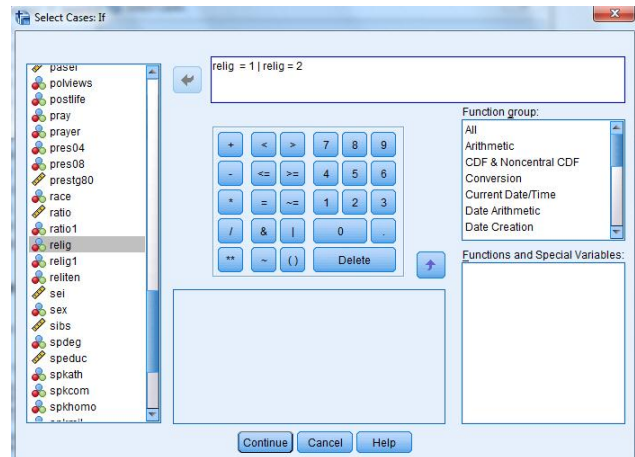


Figure 3-28

Now click on |. Click on *relig* in the list of variables again and on the arrow to move it into the box. Then click on = and then on 2 so the expression in the box reads *relig* = 1 | *relig* = 2 which means that *relig* will equal 1 or 2. Your screen should look like Figure 3-28. Click on Continue and on OK in the Select Cases box.

Run a frequencies distribution on the new variable to see what it looks like. Your screen should look like Figure 3-29. You will only have Protestants (1) and Catholics (2) in your table because you selected only those cases with values one and two on *relig*.

RS RELIGIOUS PREFERENCE					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	PROTESTANT	1036	66.9	66.9	66.9
	CATHOLIC	513	33.1	33.1	100.0
	Total	1549	100.0	100.0	

Figure 3-29

After you have selected cases for analysis, you will probably want to continue your analysis with all the cases. To do this, click on **Data**, then on **Select Cases**, and then click on the circle to the left of **All Cases**. Click on **OK** and IBM SPSS will select all the cases in the data file. This is very important. If you don't do this, you will continue to work with just the cases you have selected. This will work only if you selected filter out unselected cases in the Select Cases box when you began using select cases. If you selected delete unselected cases, you will have to get another copy of the data file by clicking on **File** and then on **Open**.

Conclusion

In this part of the book you have learned how to recode, create new variables using compute and if, and select particular cases for analysis. You can do more complicated things with these commands than we have shown you but these are the basics. In the rest of this book, we will focus on some of the statistical procedures that IBM SPSS can do for you.

Chapter Three Exercises

Use the data set 2010 General Social Survey for all these exercises.

RECODE Exercises

There are two variables that refer to the highest year of school completed by the respondent's mother and father (*maeduc* and *paeduc*). Do a frequency distribution for each of these variables. Now recode each of them (into a different variable) into three categories: under 12 years of school, 12 years, and over 12 years. Create new value labels for the recoded categories. Do a frequency distribution again to make sure that you recoded correctly.

Income06 is the total family income. Do a frequency distribution to see what the variable looks like before recoding. Recode (into a different variable) into eight categories: under \$10,000, \$10,000 to \$19,999, \$20,000 to \$29,999, \$30,000 to \$39,999, \$40,000 to \$49,999, \$50,000 to \$59,999, \$60,000 to \$74,999, and \$75,000 and over. Be very careful that you recode the values, not the labels associated with the values. Call the new variable *income1*. Create new value labels for the recoded categories. Do another frequency distribution to make sure you recoded correctly.

Now recode *income06* again (into a different variable). This time use only four categories: under \$20,000, \$20,000 to \$39,999, \$40,000 to \$59,999, and \$60,000 and over. Call the new variable *income2*. Create new value labels for the recoded categories. Do another frequency distribution to make sure you recoded correctly.

COMPUTE Exercises

In this chapter we created a new variable called *abortion*, which was the sum of the seven abortion variables in the data set. Create a new variable called *ab1*, which is the sum of *abdefect*, *abhlth*, and *abrape*. Do a frequency distribution for this new variable to see what it looks like. How is this distribution different from the distribution for the *abortion* variable based on all seven variables?

There are five variables that measure tolerance for letting someone speak in your community who may have very different views than your own (*spkath*, *spkcom*, *spkhomo*, *spkmil*, and *spkrac*). For each of these variables, 1 means that they would allow such a person to speak and 2 means that they would not allow it. Create a new variable (call it *speak*), which is the sum of these five variables. This new variable would have a range from 5 (would allow a person to speak in each of the five scenarios) to 10 (would not allow a person to speak in any of the five scenarios). Do a frequency distribution for this new variable to see what it looks like.

IF Exercises

There are two variables that describe the highest educational degree of the respondent's father and mother (*pdeg* and *mdeg*). Create a new variable (call it *mapaeduc*) that indicates if the father and mother have a college education. This variable should equal 1 if both parents have a college education, 2 if only the father has a college education, 3 if only the mother has a college education, and 4 if neither parent has a college education. Create new value labels for the recoded categories. Do a frequency distribution for this new variable to see what it looks like.

One variable indicates how often the respondent prays (*pray*) and another variable indicates if the respondent approves or disapproves of the Supreme Court's decision regarding prayer in the public schools (*prayer*). Create a new variable (call it *pry*) that is a combination of these two variables. This variable should equal 1 if the respondent prays a lot (once a day or several times a day) and approves of the Supreme Court's decision, 2 if the respondent prays a lot (once a day or several times a day) and disapproves of the Supreme Court's decision, 3 if the respondent doesn't pray a lot and approves of the Supreme Court's decision, and 4 if the respondent doesn't pray a lot and disapproves of the Supreme Court's decision. Do a frequency distribution for this new variable to see what it looks like.

SELECT IF Exercises

Select all males (1 on the variable *sex*) and do a frequency distribution for the variable *fear* (afraid to walk alone at night in the neighborhood). Then select all females (2 on the variable *sex*) and do a frequency distribution for *fear*. Are males or females more fearful of walking alone at night?

Select all whites (1 on the variable *race*) and do a frequency distribution for the variable *pres04*. Did they vote for Bush, Kerry, or someone else in 2004? Then select all blacks (2 on the variable *race*) and do a frequency distribution for *pres04*. Were whites or blacks more likely to vote for Kerry?