

Chapter 12 “Dealing With Data” Tutorial

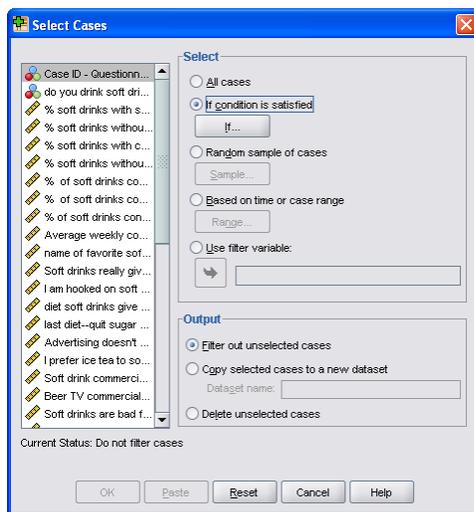
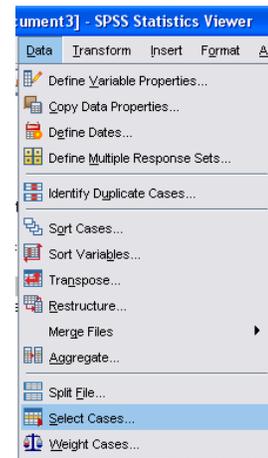
1. **Visit** the student website at for this textbook at www.clowjames.net/students.
2. **Download** the following files:
 “Chapter 12 Dealing with Data” (SPSS data file)
 “Chapter 12 Dealing with Data Survey” (Word file)
3. Open SPSS 17.0.
 The program may appear as an icon on your computer desktop, or you may need to use the START > ALL PROGRAMS command in Windows to find the “SPSS Inc.” folder. Once you have located this folder, open it, then open the “Statistics 17.0” folder. Finally, launch the SPSS Statistics 17.0 program by clicking on that item.
4. Select the “Open an existing data source” option from the resulting pop-up window to open the data file named, “Chapter 12 Dealing with Data”. If the file name does not appear in the window below this option, select the “More Files” option, then click the OK button, and browse to the location where your copy of the data file is saved. (If you have forgotten how to locate and open an SPSS data file, review the Dealing With Data instructions from chapter 2.)
5. This tutorial will show you how to use frequencies, filtering, and crosstabs to help validate data. You’ll also learn how to validate a constant sum scale. Read the Chapter 12 “Statistics Review” before you do anything else.

Validating Data

6. Begin by running frequencies for all variables. From the ANALYZE menu, select DESCRIPTIVE STATISTICS > FREQUENCIES. Highlight all variables except for the CASE ID, and move them to the variable window. Click the OK button to create frequency tables. Switch to the SPSS Output window. Examine each frequency table for obvious data entry errors (see below).

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 0	3	.6	.6	.6
Freshman	46	8.7	8.7	9.2
Sophomore	68	12.8	12.8	22.1
Junior	96	18.1	18.1	40.2
Senior	252	47.5	47.5	87.7
Graduate student	63	11.9	11.9	99.6
7	2	.4	.4	100.0
Total	530	100.0	100.0	

7. While you could create a cross-tabulation table listing the STUDENT CLASSIFICATION variable in columns and the CASEID variable in rows, the resulting table would be very large and time-consuming to search, as the number of rows would be equal to the size of the sample. Instead, it is much easier if you first filter the data so that the only case(s) with valid values remain. In this example, values less than 1 or greater than 6 are not valid. NOTE: Always check the VALUE LABEL information to verify what numbers represent valid responses before filtering the data.



8. From the DATA menu, choose the SELECT CASES option.

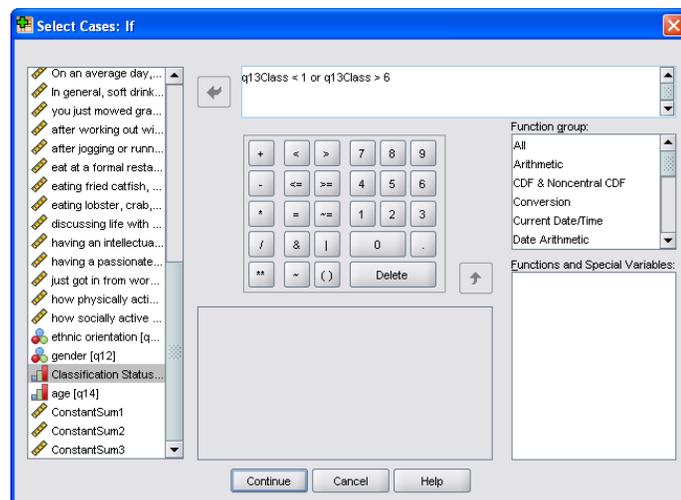
9. The SELECT CASES window (shown at left) will appear.

10. Click the "If condition is satisfied" radio button, then click the IF button underneath this option.

11. The SELECT CASES: IF window will appear. Highlight the name of the variable that you are validating (you may only validate one variable at a time). Click the arrow button to transfer that variable into the expression box.

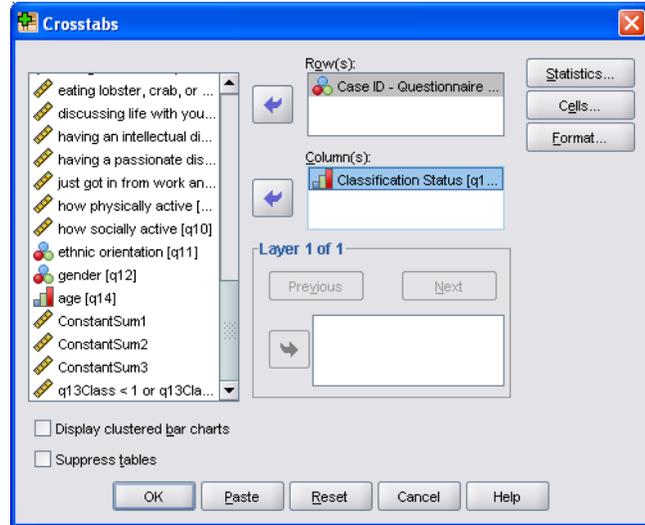
12. What you do from this point will depend upon the nature of the validation problem you are investigating and the range of numbers of that represent legitimate responses.

13. In this example, the frequency table showed values for "0" and "7". The legitimate values for this variable ranged from 1 to 6.



14. The expression entered instructs SPSS to select only those cases in which the CLASSIFICATION STATUS value is either less than 1 or greater than 6. Once you have determined the correct expression for YOUR validation purposes, click the CONTINUE button, then click OK.

15. The data set is now filtered.
16. Now it's time to create a cross tabulation table that will identify the specific cases (or questionnaires) that need to be investigated when cleaning up the data. From the ANALYZE menu, select DESCRIPTIVES > CROSSTABULATIONS.
17. Enter the CASE ID variable as the ROW variable, and the CLASSIFICATION variable as the column variable. Click OK.
18. Switch to the OUTPUT window. A cross tabulation table similar to that shown below will appear.



Case ID - Questionnaire * Classification Status Crosstabulation

Count		Classification Status		Total
		0	7	
Case ID - Questionnaire	18	1	0	1
	230	1	0	1
	283	0	1	1
	410	0	1	1
	480	1	0	1
Total		3	2	5

19. In "real life", the data validator would find the original questionnaires corresponding to the five Case ID numbers shown, then check both the answer and the coding of the CLASSIFICATION question. If the questionnaires were coded correctly, then the error occurred during data entry. The correct value can be entered into the SPSS data file for each CASE ID.

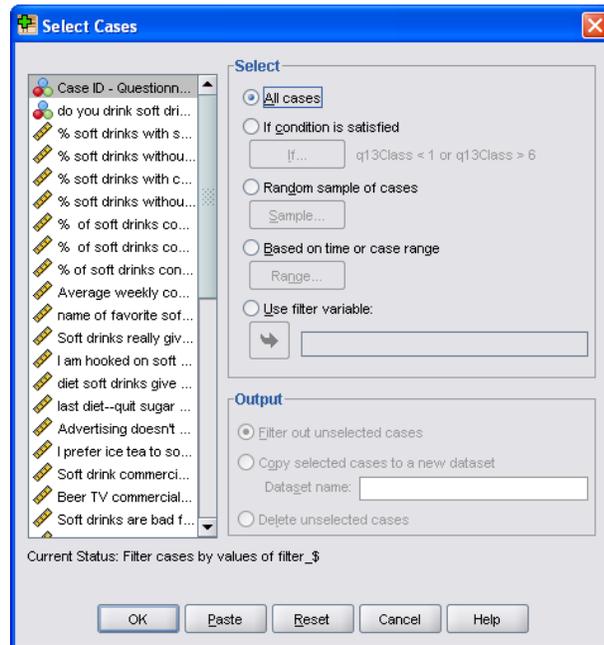
If the question was CODED IMPROPERLY, then the data validator should immediately check the coding for ALL questions on that survey. The coder may have been tired, or extremely careless; regardless, there is a need to further validate his or her work.

Furthermore, it may be beneficial to separate out a sample of the questionnaires coded by that individual for further validation. It is possible that the individual miscoded other values that would not be caught by an examination of the frequency table (e.g., coded an answer as "2" when it should have been a "3", etc.)

20. Often the frequency tables reveal more than one variable with values outside of the acceptable range. The process described can be repeated to identify other problem cases for different variables.

21. **IMPORTANT!** Before attempting any further data validation, you must “turn off” the existing filter. If you don’t remove the existing filter, any actions you attempt will be applied **ONLY** to the cases that remained active in the filtered dataset. You will **NOT** have access to the full dataset.

22. To “turn-off” a filter, use the DATA > SELECT CASES command to return to the SELECT CASES dialogue box. Click the ALL CASES, radio button, then click OK to turn off the filter. **You MUST do this BEFORE attempting to validate any other variable or before attempting any other type of data analysis.**



23. Now you are ready to investigate another variable using the procedure outlined above. Repeat the process discussed as often as needed to find the case numbers matching data that falls outside the range of acceptable responses.

Validating Constant Sum Scales

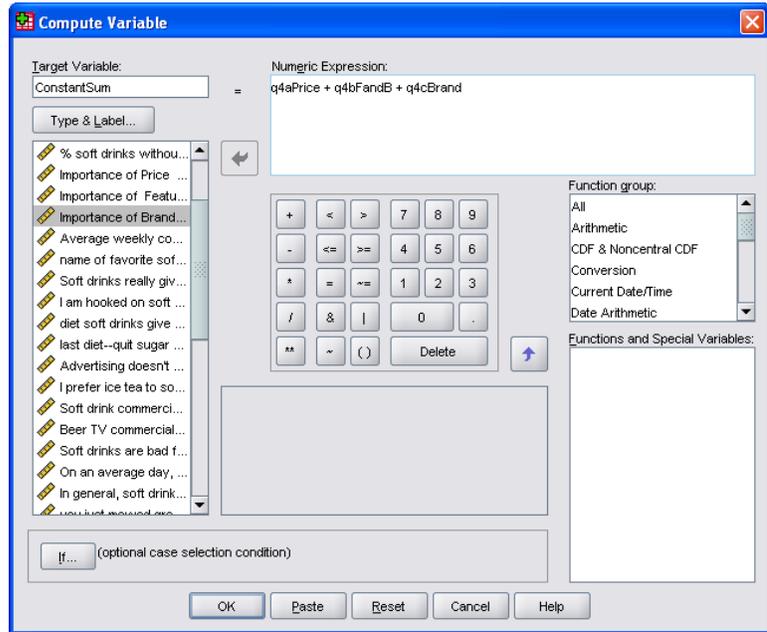
24. The next portion of the validation process requires checking to see whether the values entered for each case in a Constant Sum scale total 100. REMOVE THE FILTER as discussed previously, if you haven’t already done so.

25. In this simplified example, respondents were asked to enter a number between 0 and 100 to indicate the importance of three different factors when making a purchase decision: price, features & benefits, and brand name. The values entered for these three factors should sum to 100.

26. You will begin the validation process by creating a new variable that adds the values of the individual constant sum variables together. From the TRANSFORM menu, select the COMPUTE VARIABLE option. The COMPUTE VARIABLE dialogue box will appear.

27. Enter a name for the variable in the box underneath the “Target Variable:” heading. In this example, the variable is named “ConstantSum”.

28. Highlight the first item in the constant sum scale, click the arrow pointing towards the “Numeric Expression” box, click the “+” sign, then click the second item in the constant sum scale, click the “+” sign and so on until all variables comprising the measure are shown in the numeric expression box. Then click the OK button.



NOTE: If you failed to turn-off the filter, the variable will ONLY be created for those cases which were active during your previous variable validation.

29. Now use the ANALYZE > DESCRIPTIVE STATISTICS > FREQUENCIES command sequence to create a frequency table for the new variable.

ConstantSum

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	100.00	464	87.5	99.6	99.6
	107.00	1	.2	.2	99.8
	110.00	1	.2	.2	100.0
	Total	466	87.9	100.0	
Missing	System	64	12.1		
Total		530	100.0		

30. Next, use the DATA > SELECT CASES IF process previously discussed to identify the specific CASEIDs in which the constant sum scale either falls below 100 or is greater than 100.