Exploring Crime Rates in New South Wales

In this workshop, we will cover the following:

1. How to select cases that meet a user-defined condition.
2. How to categorise continuous variables
3. How to combine two existing variables to create a new variable.

This workshop continues using the data for Crime Rates in Local Government Areas in New South Wales. Through this workshop you will be more familiarised with this dataset.

# Select Cases

## Select Cases

Sometimes we want to carry out an analysis for specific groups (e.g. young people, women, unemployed, etc) rather than the entire sample. In this case, we need to define subset of interest by specifying some conditions, and then select cases that meet the conditions while excluding those not. This procedure is called as “subsetting” and SPSS provides several methods for this procedure with its ***Select Case*s** function. In this workshop, we will focus on how to select cases by specifying a particular value (or range of values) of a variable. Let’s say now we are only interested in urban areas in NSW for our crime rate analyses and thus will select urban LCAs only.

First, go to **Data > Select Cases.**

Let’s take a look on the options. The upper section, *select*, provides several methods for subsetting. The default, All cases, uses all cases. We will use the next option, so **1) choose “*If condition is satisfied*”.** This option uses a conditional expression to select cases. If the result of the conditional expression is true, the case is selected. If the result is false or missing, the case is not selected.



<Figure 1>

Next, look the *output* options in the lower section. This section controls the treatment of unselected cases. **2) make sure** ***“Filter out unselected cases”* is chosen for your output option (the default)**. The second, “*Copy selected cases to a new dataset”* will create a new dataset with only selected cases leaving your original dataset unaffected. The last, “*Delete unselected cases”* will delete unselected cases from the dataset. It changes your original dataset and you can’t undo this operation once it is done. Therefore you should be very careful when you use this option (You will not need to use this option for this class).

Now let’s 3) **click the *If* button** to specify conditions.



<Figure 2>

You will see a white box like <Figure 3> in which you can type an expression specifying the cases you wish to select. **1) Choose the variable *urban* from the variable list and move it into the white box** and **2) type the expression as “urban=1”**. A variable *urban* has two values: 0 equals to ‘rural’ and 1 equals to ‘urban’. Thus, “urban=1” will let SPSS select only urban LGAs. **3)** **click *Continue*** at the bottom. Then, you will be back to the previous window. **4) Click *OK*** at the bottom.



<Figure 3>

It may be useful to learn basic operators of SPSS which you see in the <Figure 3>. The basic operators are mainly used to select cases meeting the ***IF*** condition. You can use a combination of these operators for more complex selection of cases.

|  |  |  |
| --- | --- | --- |
| **Expression** | **Meaning** | **Examples** |
| = | Equal | **urban=1**: if urban equals to 1 |
| ~= | Not equal | **urban~=1**: if urban does not equal to 1 |
| < | Less than | **age< 30**: if age is less than 30 |
| <= | Less than or equal to | **age<=30**: if age is less than or equal to 30 |
| > | Greater than | **age>30** if age is greater than 30 |
| >= | Greater than or equal to | **age>=30**: if age is greater than or equal to 30 |
| & | AND: all the conditions hold | **age>30 & age<40**: if age is greater than 30 AND less 40 → 21 to 29 years  |
| | | OR: either one of the conditions holds | **age<20 | age>80**: if age is less than 20 OR greater than 80 → 0 to 19 years and 81 to maximum years  |

<Table 1>

Selecting cases will create an additional variable called ‘filter\_$’ in the rightmost column (see <Figure 4>). The variable takes the value 1 if the case is selected in the subset and 0 otherwise. But you don’t need to do anything with this filter variable, so you can ignore this variable.



<Figure 4>

Note that any analysis you conduct after this point will be done ONLY for the selected subset. For example, let me make a frequency table for *urban*. As seen in the <Figure 5>, the frequency table lists only urban LGAs as a valid case, because all rural LGAs are unselected by the procedure.



<Figure 5>

## Deselect Cases

If you want to go back to the original dataset instead of the subset, you will need to turn the filtering off as follows: 1) **Go to Data > Select Cases**, again. 2) in the popped-up window, **click *Reset* at the bottom** and you will see the option of “All cases” under *Select* is chosen (see <Figure 6>). 3) **Click *OK*** at the bottom.



<Figure 6>

# Categorising a Continuous Variable

Sometimes we want to transform a continuous variable into a categorical one. Using an example of income, we may find that classifying people into income groups such as high-income, middle-income, and low-income groups is more meaningful than using a continuous income measure in dollars. How can we do this categorisation? You may remember that we can use the ***recode*** function which we learned in the workshop 1 (see pp. 17-21 of the workshop 1).

Suppose that we want to classify LGAs into two categories depending on the average income of households: High 50% LGAs where the average household income is above the median and low 50% LGAs with the average household income being below the median. This is called as median-splitting by which our sample will be split in two having the equal size of cases. Of course, there are other ways to categorize income groups such as using more refined percentiles (e.g. quantiles) or even tax-brackets. But median-splitting is the simplest and easiest method and can be used for preliminary analysis. Therefore, in this workshop, we will use median as cut point to categorize LCAs into two income groups.

## Obtaining the Values that Split the Group

This task requires you to split all the LGAs in two groups depending on their values on the average household income variable (the variable name is *meaninc).* And we want each group to have equal size of cases so that high income LCAs have 50% cases and low income LCAs have another 50% of cases. The first step of this procedure is to calculate the cut points by which all the cases are split into half.

1. **Go to Analyze > Descriptive Statistics > Frequencies**.
2. **Choose a variable by which cases will be split** (in our example, it is *meaninc*)
3. **Click *Statistics*** in the Frequencies window.
4. **Tick “Cut points for \_\_\_\_ equal groups”** under *Percentile Values*. And **type the number of groups you want to make** (in our example it is 2).



<Figure 7>

1. **Click *Continue*** at the bottom.
2. **Click *OK*** in the Frequencies window.

The output shows the cut point which splits all the cases by half (see <Figure 8>). As you may notice, 52,265 is the median value for *meaninc*. The next step is to make a new variable using this cut point for our recoding scheme.



<Figure 8>

## Recoding Existing Variables into Different Variables

This section will use the same steps detailed in pp. 17-21 of the workshop 1.

1. Go to **Transform > Recode into Different Variables**
2. In the popped-up dialog box, **choose *meaninc* in the left pane and move it to the right pane**.
3. **Type “meaninc2” as the *Name* of output variable and “mean income – two groups” as *Label***. And **click *Change***.
4. **Click the button of *Old and New Values***.
5. In the popped-up window, **choose “Range, LOWEST through value”** and **type the value of cut point we calculated in the previous task, which is 52265**. Then, **type 1 in the *New Value* section**. This tells SPSS to recode all the LCAs with mean income below 52265 as 1. **Click *Add***.



<Figure 9>

1. **Choose “Range, value through HIGHEST”** and **again type the value of cut point (52265)**. Then, **type 2 in the *New Value* section**. All the LCAs with mean income higher than 52265 will be coded as 2. **Click *Add***.
2. **Choose “System- or user-missing” under the Old Value section, “system-missing” under the New Value section**.
3. **Click *Continue*** at the bottom.
4. **Click *OK*** in the previous window.



<Figure 10>

1. In the tab of Variable View, you will see a new variable named meaninc2 at the bottom of your list. Assign an appropriate level of measurement (Nominal) and value labels (1 = Bottom 50%; 2 = “Top 50%”).
2. **Make a frequency table of this newly generated variable** to check whether the variable is created correctly.



<Figure 11>

As seen in the <Figure 11>, almost the same number of cases are assigned for each group (The category of bottom 50% include a case with the median value and thus have one more case).

**NOTE)**

**Using the same procedure, make a new variable (*unemploy2*) in the same manner that split all the LGAs into the top and bottom 50% with respect to unemployment rates (*unemploy*)**. We will use this variable in the next section.

# How to Combine Two Existing Variables to Create a New Variable

We can further classify LCAs with a combination of income and unemployment. Suppose that we want to make a variable that combines *meaninc2* and *unemploy2* in the following way:

|  |  |
| --- | --- |
| Old Variables | New Variable (*incunemp4*) |
| *meaninc2* | *employ2* | Value | Label |
| 1 | 1 | 1 | low income, low unemployment  |
| 1 | 2 | 2 | low income, high unemployment |
| 2 | 1 | 3 | high income, low unemployment |
| 2 | 2 | 4 | high income, high unemployment |

<Table 2>

There may be several ways to create the new variable named *incunemp4*. Using the ***Compute*** command is one way, although you may need to change the values of your recoded variables . In this workshop, we will use the ***recode*** command again, but this will involve more complicated recoding schemes using the IF function to select cases, which we just learned today.

1. **Go to Transform > Recode into Different Variables**.
2. In the popped-up dialog box, **choose *employ2* variable** (in our example, *employ2*) in the left pane and move it to the right pane.
3. **Type “*incunemp4*” as the *Name* of output variable and “4 groups by income and unemployment” as *Label***. And **click *Change***.
4. **Instead of clicking Old and New Values right away, Click *If***.



<Figure 12>

1. We will first specify recoding schemes for employ2 but only within low income LCAs (meaninc2=1). So we will select cases having value 1 on our meaninc2. In the popped-up window, **choose “Include if case satisfies condition”**. Then **type “meaninc2=1” in the white box** to select only low income LCAs. Click *Continue*.

**<Figure 13>**

1. In the previous window (as seen in the <Figure 12>), **click “Old and New Values”** and then **type 1 for Old Value and 1 for New Value**. Then **click *Add***. After then, **type 2 for Old Value and 2 for New Value**. Then **click *Add***. **Click *Continue*** at the bottom.
2. You will return to the previous window. **Click OK** at the bottom.



<Figure 14>

What you have done so far recoding for low income LCAs (*meaninc2=1)*. You see from <Figure 15> that cases with meaninc2=2 have all missing values in our new incunempl4 variable. Therefore, now we need to recode those cases in the similar fashion.



**<Figure 15>**

1. **Go to Transform > Recode into Different Variables** again. You will see the previous setting and use the same setting with only change in the IF section.
2. **Click *If***again.
3. In the popped-up window, **change the expression into “meaninc=2”**. Then, **click *Continue***.



<Figure 15>

1. In the previous window, **click “Old and New Values”**.
2. In the popped-up window, **click each previous old and new value**. Then, **click *Remove***. Now, you are ready to input a new coding scheme.
3. Then, we will assign the value 3 and 4 in the <Table 2> . **Type 1 for *Old Value* and 3 for *New Value***. Then, **type 2 for *Old Value* and 4 for *New Value***. **Click *Continue*** at the bottom.
4. **Click OK** in the previous window.



<Figure 16>

Sort all the cases by *twogrpmeaninc* and *twogrpunemploy* in ascending orders. Then check whether the new variable is created correctly.

**Workshop Activities**

Q1. Report a frequency table of *unemploy2*.

Q2. Report a frequency table of *incunemp4*.

Q3. How many LGAs in the Greater Metropolitan Sydney have the value 3 (high income, low unemployment rates) for *incunemp4*. (Tip: a combination of Select Cases and Frequencies would make your task easier)

Q4. Compare the distribution of robbery rates (robbery) by the economic conditions of LGAs (*incunemp4*). The result will show the relationship between economic status (measured by mean household income), unemployment rates and robbery rates. Based on the result, which economic variable do you think is more related with robbery rates, mean household income or unemployment rates? (**Note**: **It is necessary to DESELECT CASES before you answer this question**. Otherwise, your result will be based on just 30 LGAs in Greater Metropolitan Sydney).