Bivariate Association

This week, we will use the 2012 Australian Survey of Social Attitudes (AuSSA). This workshop introduces how to produce a cross-tabulation and how to conduct a Chi-square test of Independence.

1. **Crosstabs and Chi-Square**

We use frequency table to describe one categorical variable. When you want to describe a relationship between two categorical variables, we use a special type of table called cross-tabulation (or crosstab for short).

In a crosstab, the categories of one variable determine the rows of the table, and the categories of the other variable determine the columns. The cells of the table contain the frequency that a particular combination of categories occurred.

Suppose that we are investigating whether there is association between gender of respondents (*sex*) and attitudes toward traditional gender roles (*hubbywk*). hubbywk measures the extent to which respondents agree or disagree with the statement that a man’s job is to earn money, and a woman’s job is to look after the home and family. We assume that gender may influence attitudes toward traditional gender roles. Therefore, we think of gender as an independent variable and attitudes toward traditional gender roles as a dependent variable.

To create a crosstab, click **Analyze > Descriptive Statistics > Crosstabs**. In the popped-up box, 1) move a dependent variable (in this case *hubbywk*) to the box of *Row(s)* and 2) an independent variable (in this case *sex*) to the box of *Column(s)*. 3) Click *Statistics*.



<Figure 1>

In the popped-up window, 1) tick the box of *Chi-square*. Then 2) Click *Continue*.



<Figure 2>

In the previous window (<Figure 1>), click *Cells*. In the popped-up window, 1) tick the box of *Observed* under *Counts* and 2) the box of *Column* under *Percentages* (which will show the column percentages).



<Figure 3>

In the previous window (<Figure 1>), click *OK*. This will show the cross-tabulation and Chi-square statistics.



<Figure 4>

In <Figure 4>, the first table is a summary of valid cases and missing data. The second table is the cross-tabulation of attitudes toward traditional gender roles by gender. Each cell shows the frequency and its column percentage. You see that column percentages are calculated within specific column (in this case, within specific sex of respondents). The third table displays Chi-square statistics and its associated p-value.

Based on <Figure 4>, how would you describe the association between gender and attitudes toward traditional gender roles? Do you think this association is statistically significant? And explain why?

1. **Recoding Variables**

When you examine bivariate association using a crosstab, it can be very daunting if your categorical variable has too many categories or you have a continuous variable. You will need to recode the variable (e.g. combining similar categories) so that there are reduced number of categories, but those are still theoretically meaningful.

For this week’s workshop activities, we will use Education, Age, and Class as three independent variables. As you will see, Education (*degree)* is a categorical variable with 7 categories but we may not need such detailed specifications of degree when you examine the relationship between education and gender role attitudes. And Age and Class are continuous variables and thus categorisation for these two is necessary for a crosstab. Let’s recode these three variables before conducting any bivariate analysis.

1. Generate the frequency table of the variable you want to recode

When you recode a variable, it is always a good idea to produce the frequency table of your original variable. Here is the frequency table of *degree:*

****

<Figure 5>

1. Make the recoding scheme

There are 7 categories for *degree.*  How would you recode the variable? There are no hard-and-fast rules for choosing which categories to combine. Yet, there are two rules of thumb to guide your decision: one theoretical, the other empirical.

First, there is sometimes a logical basis for choosing cutting points at which to divide the resulting categories. Using the degree as an example, it is logical to make one break between high school graduate and post-secondary.

The second guideline is based on the empirical advantage of having sufficient numbers of cases in each of the combined categories. This is because a very small categories will hamper subsequent analyses. Again in our example, those who did not complete high school may be too small to be treated as one single category (But if this group is theoretically important, you still should treat this group as one single group. Theory is more important than empirical design, almost always!)

After some deliberation, the following recoding scheme <Table 1> is designed. To make a recoding scheme like this by yourself, you have to check specific values that are assigned to individual categories. You can do so by clicking the ***Values*** column of your variable in SPSS Variable View as <Figure 6> or looking up the codebook of your data (See the simple codebook under the dataset tab of our iLearn page for this purpose).

* **Education (*degree*)**

|  |  |
| --- | --- |
| Old variable (*degree*) | New variable (*degree\_r*) |
| Value | Value label | Value | Value label |
| **1** | Did not complete High School to Year 10 | **1** | High school or less |
| **2** | Completed High School to Year 10 |
| **3** | Completed High School to Year 12 |
| **4** | Trade qualification or apprenticeship | **2** | Vocational Education & Training (VET) |
| **5** | Certificate or Diploma (TAFE or business college) |
| **6** | Bachelor Degree (including Honors) | **3** | University or more |
| **7** | Postgraduate Degree or Postgraduate Diploma |

<Table 1>

Colours are used to highlight that those with same colours are now combined into the same category in the new degree\_r variable.



**Click**

<Figure 6>

1. Enter your recoding scheme into SPSS using **Recode into Different Variables**.



<Figure 7>

This is a familiar process, so I won’t go into the details. However, I emphasize that you should always include the specification of how to treat missing values and it is always smart to choose the System- or User- missing (MISSING) to System-missing (SYSMIS)!!! If you look at <Figure 6>, there is a category of 99 (No Answer). This is assigned a value of 99, but defined as missing by User. In contrast, a system-missing happens when there is complete absence of data. See <Figure 8> for an illustration.

If you do not include user-missing but only system-missing in your recoding scheme, such as SYSMIS --> SYSMIS, then user-missing values such as 99 could be treated as a valid case in your new recoded variable, which will ruin subsequent analysis. To sum up, **always include MISSING --> SYSMIS when you recode.**

**System-missing**

**User-missing**



<Figure 8>

As you finished the recoding of education, complete recoding of age and class with the following recoding schemes, <Table 2> for age and <Table 3> for class, respectively.

* **Age (*age*)**

|  |  |
| --- | --- |
| Old variable (age) | New variable (age\_r) |
| Value | Value | Value label |
| Lowest - 40 | 1 | 40 or less |
| 41 - 60 | 2 | 41 - 60 |
| 60 - Highest | 3 | 61 or more |

<Table 2>

* **Class (*topbot*)**

|  |  |
| --- | --- |
| Old variable (*topbot*) | New variable (*topbot\_r*) |
| Value | Value | Value label |
| 1 - 5 | 1 | Lower class |
| 6 - 8 | 2 | Middle class |
| 9 - 10 | 3 | Upper class |

<Table 3>

**Workshop Activities**

Q1. Make a crosstab of the variable *degree\_r* (Education) and *hubbywk* (Attitudes toward traditional gender roles). Also, conduct a chi-square test of independence.

1. Write a research and a null hypothesis.
2. Do you find any noticeable associations in the cross-tabulation? If so, describe your findings.
3. Report the chi-square and its associated p-value.
4. Based on the chi-square test, do you think the association found in the table is statistically significant? And explain why.

Q2. Make crosstab of the variable *age\_r* (Age) and *hubbywk* (Attitudes toward traditional gender roles). Also, conduct a chi-square test of independence.

1. Write a research and a null hypothesis.
2. Do you find any noticeable associations in the cross-tabulation? If so, describe your findings.
3. Report the chi-square and its associated p-value.
4. Based on the chi-square test, do you think the association found in the table is statistically significant? And explain why.

Q3. Make crosstab of the variable *topbot\_r* (Class) and *hubbywk* (Attitudes toward traditional gender roles). Also, conduct a chi-square test of independence.

1. Write a research and a null hypothesis.
2. Do you find any noticeable associations in the cross-tabulation? If so, describe your findings.
3. Report the chi-square and its associated p-value.
4. Based on the chi-square test, do you think the association found in the table is statistically significant? And explain why.