**PART B: Correlation and data visualisation for *categorical* data**

1. **‘Quick-and-dirty’ visualisation**

The relationship between two categorical variables may be visualised graphically using a stacked percentage bar-chart:

* *Graphs* > *Chart* *Builder* > *Bar* > double-click on *Stacked Bar* icon 
* Move the ‘outcome’ variable (e.g. *health*) into the *Stack: set color* box
* Move covariate (second categorical variable) into the *x-axis* box
* In the *Element Properties* window use the drop-down *Statistic* menu to select *Percentage()*
* Set *Parameters…* > *Total for each X-axis category* > *Apply* > *OK*

Using the **FRS 2016-17 dataset**, visualise the association between a person’s health (*health*) and each of their:

* Age (*age\_group*)
* Highest educational qualification (*highest\_qual*)
* Marital status (*marital\_status*)
* Socio-economic status (*nssec*)

**Q6. Which of the associations do you think is strongest?**

WRITE ANSWER HERE: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Q7. Which of the associations do you think is weakest?**

WRITE ANSWER HERE: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Correlation**

As before, rather than relying upon an impressionistic view of the strength of the association between two variables, we can measure that association by calculating the relevant correlation coefficient.

Table 3 identifies the statistically appropriate measure of correlation to use between two continuous variables.

**Table 3 Measuring the correlation of two categorical variables**

|  |  |  |
| --- | --- | --- |
| *Variable data type* | *Measure of correlation* | *Range* |
| Both ordinal | Spearman’s Rank | -1 to +1 |
| One or both nominal | Cramer’s V | **0 to +1** |

The value of Spearman’s correlation coefficient is sensitive to category order. This is OK for ordinal variables because they have a fixed category order. However, the category order in a nominal variable is arbitrary. This means that if Spearman’s correlation coefficient was calculated for a variable pair include *sex*, the result would be different if the categories in sex were ordered male/female or female/male. For this reason, Spearman’s correlation coefficient should *not* be calculated for a variable pair involving a nominal variable.

Cramer’s V, on the other hand, can be used to assess the correlation of a variable pair including an ordinal variable, since the value calculated is insensitive to category order. The main downside of using Cramer’s V is that no information is provided on whether the correlation is positive or negative. This is not a problem if the variable pair includes a nominal variable but represents an information loss if the both variables being correlated are ordinal.

Note that, because they are calculated in different ways, the value of Spearman’s and Cramer’s V correlation coefficients for the same variable pair will differ.

As a result of the above:

* If measuring the correlation of *only one variable pair*, you should use the appropriate measure as outlined in Table 3.
* If measuring and *comparing* the correlations of *multiple pairs* of continuous variables, you should use Cramer’s V only

**Research Question #2: Which of our selected person-level variables is most strongly correlated with an individual’s health status?**

To answer this question, complete Table 4. The information required to help you do this is provided below the table.

**Table 4 Person-level correlations with health status**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Covariates** | | **Correlation Coefficient** | | **Statistical Significance** | |
| *Cramer’s V* | *p-value* | |
| *health* | *age\_group* |  |  | |
| *Health* | *highest\_qual* |  |  | |
| *health* | *marital\_status* |  |  | |
| *Health* | *nssec* |  |  | |

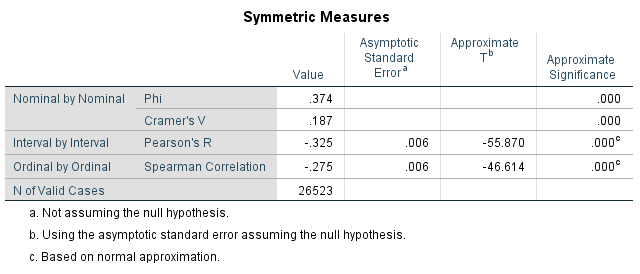
To calculate the Spearman (for categorical data) and Cramer’s correlation coefficents in SPSS:

* *Analyze* > *Descriptive Statistics* > *Crosstabs*
* Move ‘outcome’ variable (e.g. *health*) into *Rows* box
* Move variable(s) it is being correlated with into the *Columns* box
* Turn on the *Suppress tables* option
* *Statistics* > turn on the *Correlations* option to obtain Spearman’s correlation; turn on the *Phi and Cramer’s V* option to obtain Cramer’s V
* *Continue* > *OK*

*Interpreting the output*

There will be one block of output per variable pair.

The block of output below reports the results of correlating *health* with self-rated level of happiness (*happy*):



1. *Correlation*

From the ‘Value’ column we can see that the correlation of *health* with *happy* is 0.187 (Cramer’s V) or -0.325 (Spearman).

1. *Statistical significance*

If the *Approximate Significance* of the measure is less than or equal to 0.05, then we can have 95% confidence that an association exists between the two variables being correlated in the wider population.

If the *Approximate Significance* of the measure is less than 0.01, then we can have 99% confidence that an association exists between these two variables in the wider population.

**Q8. Do the results in Table 4 confirm or call into question the impressions you formed by looking at the stacked percentage bar-charts?**

**Q9. Why might the result of your area- and person-level analyses appear to disagree with one another**? [c.f. this week’s lecture]

**PART C: Mastering the content of this practical**

The only way to master the content in this practical is ‘practise, practise, practise’. Therefore:

1. **Finish working through the practical**
2. **Extension activities**

* Create a scatterplot with your outcome and chosen variables of interest.
* Compute appropriate correlation coefficients between your outcome variable and variables of interest

**APPENDIX: The Practical in a Nutshell**

**SPSS scatter-plot graph**

* *Graphs* > *Chart* *Builder* > *Scatter*/*Dot*
* Double-click on the *Simple Scatter* icon 
* Drag one continuous variable onto the y-axis; drag the other onto the x-axis
* *OK*

**SPSS stacked percentage bar-chart**

* *Graphs* > *Chart* *Builder* > *Bar* > double-click on *Stacked Bar* icon 
* Move the ‘outcome’ variable (e.g. *health*) into the *Stack: set color* box
* Move covariate (second categorical variable) into the *x-axis* box
* In the *Element Properties* window use the drop-down *Statistic* menu to select *Percentage()*
* Set *Parameters…* > *Total for each X-axis category* > *Apply* > *OK*

**Table 1 Measuring the correlation of two continuous variables**

|  |  |  |
| --- | --- | --- |
| *Variable data type* | *Measure of correlation* | *Range* |
| Both symmetrically distributed | Pearson’s | -1 to +1 |
| One or both with a skewed distribution | Spearman’s Rank | -1 to +1 |

**Table 3 Measuring the correlation of two categorical variables**

|  |  |  |
| --- | --- | --- |
| *Variable data type* | *Measure of correlation* | *Range* |
| Both ordinal | Spearman’s Rank | -1 to +1 |
| One or both nominal | Cramer’s V | **0 to +1** |

**Comparing the correlation of multiple variable pairs**

*Continuous*: Calculate Pearson’s and Spearman’s and see if they tell the same story

*Categorical*: Use Cramer’s V

**Calculating correlation coefficients**

*Continuous data*

* *Analyze* > *Correlate* > *Bivariate* > *Pearson* and/or *Spearman*
* Move *all* the variables to be correlated with one another into the *Variables* box

**TIP:** Move the variable you want to correlate with all the other variables (i.e. *pct\_long\_term\_ill*) into the *Variables* box first.

*Categorical data*

* *Analyze* > *Descriptive Statistics* > *Crosstabs*
* Move ‘outcome’ variable (e.g. *health*) into *Rows* box
* Move variable(s) it is being correlated with into the *Columns* box
* Turn on the *Suppress tables* option
* *Statistics* > turn on the *Correlations* option to obtain Spearman’s correlation; turn on the *Phi and Cramer’s V* option to obtain Cramer’s V

PTO