QS101: Introduction to Quantitative Methods in Social Science

Week 14: Assessment 2 & Cross Tabulations

Dr. Florian Reiche
Teaching Fellow in Quantitative Methods
Course Director BA Politics and Sociology
Deputy Director of Student Experience and Progression

January 30, 2015

Introduction to Assessment 2

Introducing the Data Set

Working with Data: Cross Tabulations

Introduction to Assessment 2

Assessment: Task

- Research Question: How do socio-demographic factors influence income in the UK?
- Task: Develop a report of no more than 2,500 words (excluding graphs, tables and footnotes). As a guide, the report is expected to outline the following six issues:

Assessment: Structure

1. Introduce the Research Question explaining how it relates to existing research and how you will test the expectations of theory based on a Literature Review
2. Develop up to three testable hypotheses out of the theoretical framework adopted
3. Operationalisation & measurement of theoretical concepts
4. Methodology. What specific techniques will you use?
5. Analysis of the data, interpretation of results
6. Discussion and Conclusions: What are the implications of your results for the theory? Is it supported/falsified? Recommendations for improving future research?

Introducing the Data Set

The data set – formalities

- The data set we used is called: Understanding Society
- It contains a variety of variables which allows a broad variety of hypotheses to explore
- The data sets are available on the module homepage
- Here you will also find a Word document outlining the variables available, and their labels

Three Waves

- There are three waves available:
  - A: 2009 / 10
  - B: 2010 / 11
  - C: 2011 / 12

Three Waves

- There are three waves available:
  - A: 2009 / 10
  - B: 2010 / 11
  - C: 2011 / 12
- Select ONE wave

Three Waves

- There are three waves available:
  - A: 2009 / 10
  - B: 2010 / 11
  - C: 2011 / 12
- Select ONE wave
- ONLY ONE WAVE!

How do the Waves differ?

- A large amount of variables is available in all three waves.
- If you select one of these variables, then it does not matter, which wave you choose (most recent would be desirable, however)
- Still: Only choose ONE wave!
- Some variables are only available in one particular wave. So if you are interested in the influence of such a variable, this determines the wave.

Notes of Importance

- Due date: 05.05.2015, 2PM

Notes of Importance

- Due date: 05.05.2015, 2PM
- Feedback deadline: 01.05.2015

Notes of Importance

- Due date: 05.05.2015, 2PM
- Feedback deadline: 01.05.2015
- Formatting

Notes of Importance

- Due date: 05.05.2015, 2PM
- Feedback deadline: 01.05.2015
- Formatting
- Spell check

Notes of Importance

- Due date: 05.05.2015, 2PM
- Feedback deadline: 01.05.2015
- Formatting
- Spell check
- Referencing (incl. the data set)

Notes of Importance

- Due date: 05.05.2015, 2PM
- Feedback deadline: 01.05.2015
- Formatting
- Spell check
- Referencing (incl. the data set)
- Proof, correlation & co.

Notes of Importance

- Due date: 05.05.2015, 2PM
- Feedback deadline: 01.05.2015
- Formatting
- Spell check
- Referencing (incl. the data set)
- Proof, correlation & co.
- Advice and Feedback Hours

Registration

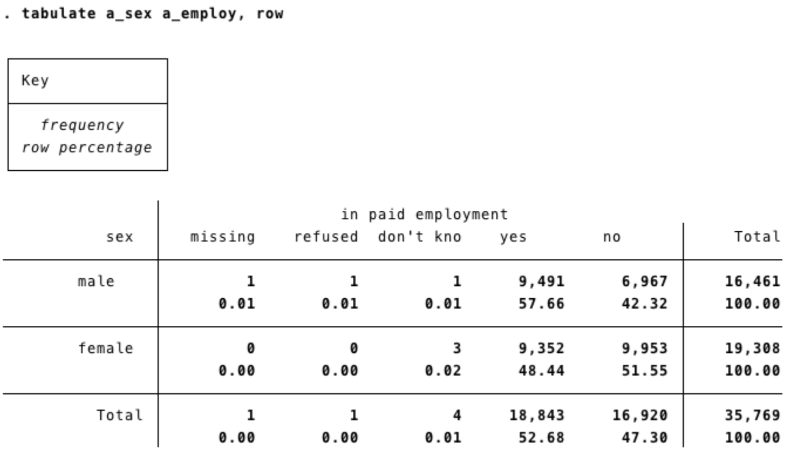
You need to register for the use of the data sets.

Working with Data: Cross Tabulations

How to do a Crosstab in Stata

- The command is beguilingly easy:
    - **tabulate** *rowvar columvar*, **row**
- **row** tells Stata to sum percentages up in the rows

An Example



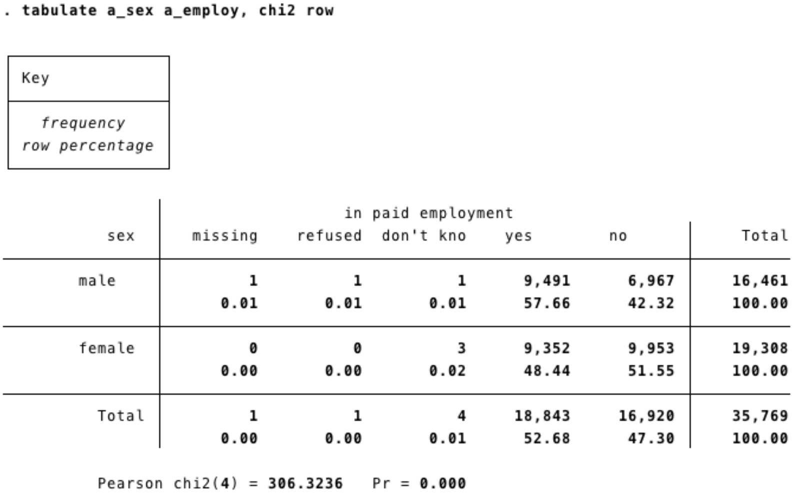
```
. tabulate a_sex a_employ, row

+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
+-------------------+

           |                    in paid employment
       sex |   missing    refused  don't kno        yes         no |     Total
-----------+-------------------------------------------------------+----------
      male |         1          1          1      9,491      6,967 |    16,461 
           |      0.01       0.01       0.01      57.66      42.32 |    100.00 
-----------+-------------------------------------------------------+----------
    female |         0          0          3      9,352      9,953 |    19,308 
           |      0.00       0.00       0.02      48.44      51.55 |    100.00 
-----------+-------------------------------------------------------+----------
     Total |         1          1          4     18,843     16,920 |    35,769 
           |      0.00       0.00       0.01      52.68      47.30 |    100.00 
```

Replicate this result.

And what about χ^2?

- We need to extend the initial command:
  - **tabulate** *rowvar colummvar*, **chi2 row**
- You can also get the expected values by typing **chi2 expected row** after the comma

Example again



```
. tabulate a_sex a_employ, chi2 row

+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
+-------------------+

           |                  in paid employment
       sex |   missing    refused  don't kno        yes         no |     Total
-----------+-------------------------------------------------------+----------
      male |         1          1          1      9,491      6,967 |    16,461 
           |      0.01       0.01       0.01      57.66      42.32 |    100.00 
-----------+-------------------------------------------------------+----------
    female |         0          0          3      9,352      9,953 |    19,308 
           |      0.00       0.00       0.02      48.44      51.55 |    100.00 
-----------+-------------------------------------------------------+----------
     Total |         1          1          4     18,843     16,920 |    35,769 
           |      0.00       0.00       0.01      52.68      47.30 |    100.00 

          Pearson chi2(4) = 306.3236   Pr = 0.000
```

Queries

- Is this statistically significant?

Queries

- Is this statistically significant?
- Is this statistically significant at the 95% level?

Queries

- Is this statistically significant?
- Is this statistically significant at the 95% level?
- Replicate!

Is it always this easy?

- Alas, no...

Is it always this easy?

- Alas, no...
- Sometimes variables have loads of categories

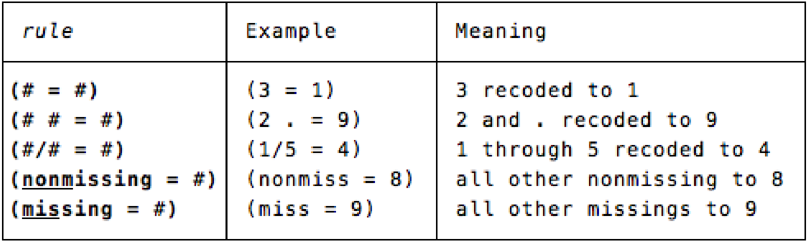
Is it always this easy?

- Alas, no...
- Sometimes variables have loads of categories
- Then we need to recode

How to recode?

- See Section 3.5. in the Acock book
- Book explains how to recode categorical variables
- Example: **`recode a_sex (# = #)`**

Recoding Commands for Categorical Variables

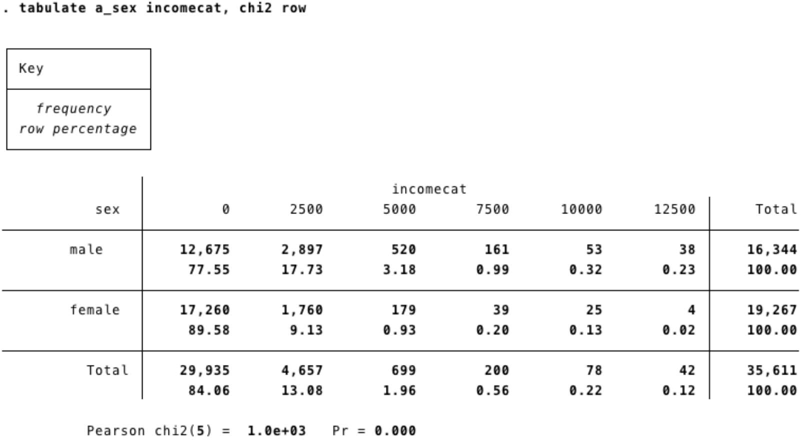


rule	Example	Meaning
(# = #)	(3 = 1)	3 recoded to 1
(# # = #)	(2 . = 9)	2 and . recoded to 9
(#/# = #)	(1/5 = 4)	1 through 5 recoded to 4
(nonmissing = #)	(nonmiss = 8)	all other nonmissing to 8
(missing = #)	(miss = 9)	all other missings to 9

Recoding continuous into categorical variables

- Assume we want to turn the continuous variable on gross personal income into categories
- For this, we use the following command:
- Example: **`egen incomecat = cut(a_fimngrs_dv), at(0,2500,5000,7500,10000,12500,15000)`**
- This creates categories, such as
  - “0 up to (but not including) 2500”
  - “2500 up to (but not including) 5000”
  - etc.

Example



```
. tabulate a_sex incomecat, chi2 row

+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
+-------------------+

           |                          incomecat
       sex |         0       2500       5000       7500      10000      12500 |     Total
-----------+------------------------------------------------------------------+----------
      male |    12,675      2,897        520        161         53         38 |    16,344
           |     77.55      17.73       3.18       0.99       0.32       0.23 |    100.00
-----------+------------------------------------------------------------------+----------
    female |    17,260      1,760        179         39         25          4 |    19,267
           |     89.58       9.13       0.93       0.20       0.13       0.02 |    100.00
-----------+------------------------------------------------------------------+----------
     Total |    29,935      4,657        699        200         78         42 |    35,611
           |     84.06      13.08       1.96       0.56       0.22       0.12 |    100.00

          Pearson chi2(5) =   1.0e+03   Pr = 0.000
```

Queries

- Is this statistically significant at the 95% level?

Queries

- Is this statistically significant at the 95% level?
- Is this statistically significant at the 99% level?

Queries

- Is this statistically significant at the 95% level?
- Is this statistically significant at the 99% level?
- Replicate!