Analysing tabular data
Reducing complexity and increasing information

Ronán Conroy (rconroy@rcsi.ie)

Note
This paper is currently submitted for publication. Comments and suggestions are most welcome. Please do not distribute it widely, as it will be replaced with a revised version in the near future.
Examples are given in **Stata** code. Remember that Stata also has interactive menus which will accomplish the same tasks. Stata commands and output are shown in typewriter type.

Abstract
Surveys frequently require one-way and two-way tabulation of findings. Without some work on the part of the data analyst, large tables can be hard to read and even harder to extract information from. This paper shows two very typical cases and examines ways of making the presentation of results clearer by simplifying categorical variables and by constucting logistic and multinomial logistic models which can test specific associations within the table while controlling for potential confounding variables.

Context
The data are derived from a telephone survey commissioned by a non-governmental organization. Such surveys frequently classify participants using variables which have multiple categories, such as education, social class, job type {etc. These personal characteristics often have to be related to question responses which themselves comprise numerous categories. The result can be very large tables in which it is difficult for anyone to extract all but the most obvious patterns. This paper, based on data from a real survey, shows how large tables may be reduced and made more interpretable.

Problem 1: a table in which one dimension is binary

The first question that the client posed was who had heard of the agency. The table shows the proportion of people who have heard of the agency, categorised by highest completed education level. Tables of this sort are common in survey reports, and there is a temptation just to reproduce the Chi-squared value and the comment that there was no association between education and whether the person had heard of the agency.

```
. tab ever_heard education, col chi

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

Ever heard |
    of the |         Education level
    agency | Primary +  Secondary   Tertiary |     Total
-----------+---------------------------------+----------
        No |        75        161        254 |       490
           |     13.04      16.93      16.10 |     15.79
-----------+---------------------------------+----------
       Yes |       500        790      1,324 |     2,614
           |     86.96      83.07      83.90 |     84.21
-----------+---------------------------------+----------
     Total |       575        951      1,578 |     3,104
           |    100.00     100.00     100.00 |    100.00

          Pearson chi2(2) =   4.3030   Pr = 0.116
```

Slightly more people with the lowest level of education (Primary +) have heard of the agency than in the other two categories, but the Chi-squared test isn't significant. However, there are one or two doubts about this table. The first is that the Chi-squared tests a vague hypothesis: that the proportion of people who have heard of the agency is different in different categories of education. It ignores two things: education is an ordered variable, and that from the standpoint of knowing about the agency, not all the different categories of education may not be different to each other. Finally, there is the suspicion that participants with the lowest level of education may also be older, and so any relationship between education level and having heard of the agency may be confounded by age.

Let's start by treating education as an ordered variable, using logistic regression.

```
. xi:logistic ever_heard i.education

i.education       _Ieducation_1-3     (naturally coded; _Ieducation_1 omitted)

Logistic regression                               Number of obs   =       3104
                                                  LR chi2(2)      =       4.45
                                                  Prob > chi2     =     0.1079
Log likelihood =  -1351.445                       Pseudo R2       =     0.0016

------------------------------------------------------------------------------
  ever_heard | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ieducatio~2 |   .7360248   .1111625    -2.03   0.042     .5474377    .9895784
_Ieducatio~3 |   .7818898   .1106467    -1.74   0.082     .5925027    1.031812
------------------------------------------------------------------------------
```

I have used the lowest education category as a baseline, and used Stata's -xi- command to generate dummy terms for the other categories. After all, there is no expectation that awareness of the agency will be linearly related to education. And it looks like it isn't (confirming what was evident from the table). The odds ratios for secondary and tertiary education look remarkably similar. We can test, formally, to see if they are significantly different.

```
. test _Ieducation_2= _Ieducation_3

 ( 1)  _Ieducation_2 - _Ieducation_3 = 0

           chi2(  1) =    0.30
         Prob > chi2 =    0.5837
```

The important factor governing whether the person has heard of the agency seems to be whether or not the person had a primary education. This is important because we might have thought of education as an ordinal variable (each category represents an increase in the amount of time that the person spend in education). A variable is not *inherently* ordinal - or indeed nominal or metric. Its scaling properties have to be considered in relation to the variable of interest.

All working statisticians (as opposed to mathematical ones...) have encountered this phenomenon. An apparently metric or ordinal variable may exhibit a threshold relationship, where what matters is not the absolute value, but whether this value is above or below a threshold. Likewise, supposedly nominal variables sometimes show ordinal properties. I once analysed a clinical trial in which I correlated the study identifying number with the outcome. There was a relationship, indicating that patients recruited later in the trial had a

different outcome (and, on looking further, they were different in other ways too).
In the present case, education breaks one of the rules for a measurement scale: things in different categories should be different. (The complementary rule is that things in the same category should be the same.) It is clear that people with a primary education are different to everyone else, but people in the other two categories are similar to each other {in respect of whether they have heard of the organisation. Of course, all three categories may be different when we consider the relationship of education to some other variable.

To simplify the table, we can generate a variable that records whether the person had a primary education or not.

```
. gen byte primary_ed=education==1 if ~missing(education)
```

This generates a binary variable quickly. It's useful to remember that when a computer evaluates a logical statement, it evaluates it as true or false, using one for true and zero for false. These ones and zeros are just plain ones and zeros; you can do arithmetic with them, or just use them as the values of a variable as I have done here. Note, too, that I have added the -if- clause to make sure that this variable only has a value if the original education variable has a value. If education is missing, then (education==1) evaluates to false, not missing!

I can now repeat the logistic regression:

```
. logistic ever_heard primary_ed

Logistic regression                               Number of obs   =       3104
                                                  LR chi2(1)      =       4.15
                                                  Prob > chi2     =     0.0415
Log likelihood = -1351.5947                       Pseudo R2       =     0.0015

------------------------------------------------------------------------------
  ever_heard | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  primary_ed |   1.308735   .1766357     1.99   0.046     1.004542    1.705044
------------------------------------------------------------------------------
```

Respondents with a primary education were more likely to have heard of the agency. But is this association perhaps the result of confounding? Could demographic differences between people in different education categories cause an apparent association (or mask an even stronger association)? After all, the least well-educated people in most surveys are also the oldest, and education may also be

related to gender. This is what happens when we add age into the regression:

```
. logistic ever_heard primary_ed age

Logistic regression                               Number of obs   =       3104
                                                  LR chi2(2)      =     305.22
                                                  Prob > chi2     =     0.0000
Log likelihood = -1201.0636                       Pseudo R2       =     0.1127

------------------------------------------------------------------------------
  ever_heard | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  primary_ed |   .8131694   .1182035    -1.42   0.155     .6115742    1.081217
         age |   1.121763   .0080211    16.07   0.000     1.106152    1.137595
------------------------------------------------------------------------------
```

Oops. Our results have changed quite significantly. Older people are more likely to have heard of the agency (the odds ratio for a 1-year difference in age is 1.12) and, adjusted for age, those with a primary education are *less* likely to have heard of the agency, though not significantly so. I am fond of using -adjust- to do a 'typical case' calculation in a case like this.

```
. adjust age=35, by(primary_ed) pr  f(%4.2f)

------------------------------------------------------------------------------
     Dependent variable: ever_heard     Command: logistic
 Covariate set to value: age = 35
------------------------------------------------------------------------------

---------------------
primary_ed|
          |         pr
----------+----------
        0 |       0.90
        1 |       0.88
---------------------
     Key:  pr  =  Probability
```

At age 35, the model predicts that 88% of those with a primary education and 90% of those with a secondary or tertiary education will have heard of the agency. Whether or not this differences is statistically significant (and it might be in a very large sample) it has no real-life importance. It does not matter whether an extra 2% of a population subgroup has heard of your agency. The important finding is that most people have heard of the agency, and there is no association between this and education once we allow for age differences. In practice, I would check other confounders as well, but the point has been amply illustrated.

Key points: 2 x K tables

1. Chi-squared tests on tables bigger than 2 x 2 are vague and rarely test a hypothesis that is of any interest. Just because a Chi-squared test is not statistically significant does not mean that there the table contains no associations. In particular, the Chi-squared test does not test for ordered associations. These can often be detected by eye, or, better still, specified from pre-existing knowledge and tested (with -**ologit**- for example).
2. Even when the Chi-squared is significant, more work will be needed to determine why. Telling the client that there is an association of some sort between two variables tells them almost nothing.
3. Just because a variable is measured using a number of categories, it does not follow that these categories always represent distinct levels of the variable. The analyst should be on the lookout for categories that are equivalent, with a view to pooling them and simplifying the analysis. Of course, just because two categories are equivalent in one analysis does not mean that they will be equivalent in another.
4. More generally, a variable is not inherently metric, ordinal or interval. The character of a scale depends on the relationship in which the variable is being analysed.
5. Logistic regression, combined with -xi- is useful in the analysis of 2 x K tables.
6. A table may be distorted by the effects of confounding factors. The analyst needs to consider the confounding factors such as differences in demographic variables between groups that could distort the relationship between the row and column variables of a table. Regression models allow these to be adjusted for.

Simplifying huge tables

The previous example had the advantage that the predicted variable was binary. In the next example, both the predictor and predicted variables are complicated. The interviewers asked participants to choose their ideal method of receiving information from the agency. Here are the responses

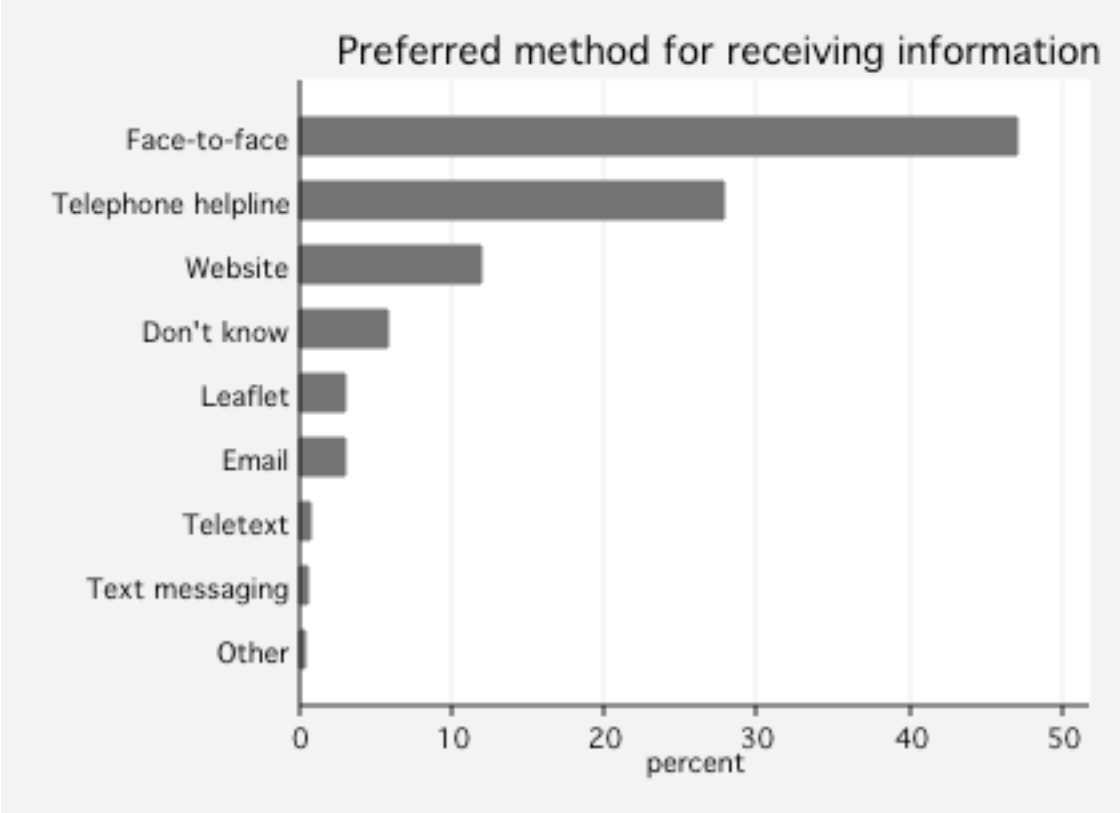
```
. tab info_pref

 Preferred mode of |
       information |      Freq.     Percent        Cum.
-------------------+-----------------------------------
      Face-to-face |      1,525       47.07       47.07
Telephone helpline |        899       27.75       74.81
             Email |         97        2.99       77.81
           Website |        383       11.82       89.63
           Leaflet |         99        3.06       92.69
          Teletext |         25        0.77       93.46
    Text messaging |         16        0.49       93.95
             Other |          9        0.28       94.23
        Don't know |        187        5.77      100.00
-------------------+-----------------------------------
             Total |      3,240      100.00
```

The agency which commissioned the research was very keen on text messaging. Clearly the clients are not - just 16 people picked this as their preferred option. I have retained the ‘don't know' option in the table, because the clients need to know what proportion of their client population prefers each method, not what proportion of people who expressed a clear preference, which would be the interpretation of the table if the people who didn't know were omitted.

The information is a little easier to digest in a graph like this:

```
. catplot hbar info, percent sort descending title("Preferred method for
receiving information")
```



The graph has the distinct advantage that it presents the most common categories first. It underlines the point that there are really only three important modalities: face-to-face (the choice of half of those interviewed), a phone help-line and an internet web-site. This is important, because the commissioning agency was keen to know the relationship between social class and preferred method of receiving information. This could easily form an awful table that looks like this:

```
. tab info_pref social_class, col chi

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

 Preferred mode of |                        Social class
       information | Hi Prof/M  Lo Prof/M  Other Non  Skilled M  Semi-Skil  Unskilled |     Total
-------------------+-----------------------------------------------------------------+----------
      Face-to-face |       295        281        267        243        204         76 |     1,366
                   |     45.88      47.71      49.44      49.29      47.22      47.20 |     47.80
-------------------+-----------------------------------------------------------------+----------
Telephone helpline |       167        163        141        135        137         50 |       793
                   |     25.97      27.67      26.11      27.38      31.71      31.06 |     27.75
-------------------+-----------------------------------------------------------------+----------
             Email |        24         11         11         16          7          9 |        78
                   |      3.73       1.87       2.04       3.25       1.62       5.59 |      2.73
-------------------+-----------------------------------------------------------------+----------
           Website |       103         72         65         44         33         10 |       327
                   |     16.02      12.22      12.04       8.92       7.64       6.21 |     11.44
-------------------+-----------------------------------------------------------------+----------
           Leaflet |        17         16         15         19         16          7 |        90
                   |      2.64       2.72       2.78       3.85       3.70       4.35 |      3.15
-------------------+-----------------------------------------------------------------+----------
          Teletext |         5          3          7          3          5          0 |        23
                   |      0.78       0.51       1.30       0.61       1.16       0.00 |      0.80
-------------------+-----------------------------------------------------------------+----------
    Text messaging |         1          0          1          2          6          0 |        10
                   |      0.16       0.00       0.19       0.41       1.39       0.00 |      0.35
-------------------+-----------------------------------------------------------------+----------
             Other |         2          2          1          3          1          0 |         9
                   |      0.31       0.34       0.19       0.61       0.23       0.00 |      0.31
-------------------+-----------------------------------------------------------------+----------
        Don't know |        29         41         32         28         23          9 |       162
                   |      4.51       6.96       5.93       5.68       5.32       5.59 |      5.67
-------------------+-----------------------------------------------------------------+----------
             Total |       643        589        540        493        432        161 |     2,858
                   |    100.00     100.00     100.00     100.00     100.00     100.00 |    100.00

         Pearson chi2(40) =  72.1881   Pr = 0.001
```

The table contains many cells that are empty, or virtually so. The only excuse for showing it is to demonstrate how difficult it is to extract any useful information, and, indeed, any information at all, from most of it. The salient features are worth pointing out, however: face-to-face seems to be preferred by just under half the respondents in each social class. On the other hand, there seems a clear social class gradient for website preference, starting with 16% of the higher professionals and dropping to six percent in the unskilled manual category. A reverse relationship seems to hold for telephone helpline and for leaflets, with higher preference levels in the lower social classes. But clearly we have some work to do before we can brief the agency on factors affecting preferences for methods of receiving information…

The first decision that I made was to ignore any option chosen by less than 5% of respondents. This is pragmatic. A provider agency needs to concentrate on providing information in a format in which the majority of their clients want it. So the real interest lies in face-to-face, telephone and website. We can create a simpler variable that

pools all other responses (including ‘no preference') into the ‘Other' category. This reduces the table in size:

```
 . tab info_pref_4 social_class, col chi

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

 Information |
 preference: |
        main |                          Social class
  categories | Hi Prof/M  Lo Prof/M  Other Non  Skilled M  Semi-Skil  Unskilled |     Total
-------------+---------------------------------------------------------------------+----------
Face-to-face |       295        281        267        243        204         76 |     1,366
             |     45.88      47.71      49.44      49.29      47.22      47.20 |     47.80
-------------+---------------------------------------------------------------------+----------
   Telephone |       167        163        141        135        137         50 |       793
             |     25.97      27.67      26.11      27.38      31.71      31.06 |     27.75
-------------+---------------------------------------------------------------------+----------
     Website |       103         72         65         44         33         10 |       327
             |     16.02      12.22      12.04       8.92       7.64       6.21 |     11.44
-------------+---------------------------------------------------------------------+----------
       Other |        78         73         67         71         58         25 |       372
             |     12.13      12.39      12.41      14.40      13.43      15.53 |     13.02
-------------+---------------------------------------------------------------------+----------
       Total |       643        589        540        493        432        161 |     2,858
             |    100.00     100.00     100.00     100.00     100.00     100.00 |    100.00

          Pearson chi2(15) =  32.0079   Pr = 0.006
```

The Chi-squared test is not especially useful, here either. All it tells us is that there is an association of some sort between social class and preference. We need to find a more informative model. We can start by observing the obvious: that half of the respondents wanted face to face information. So we could ask how social class determines the choice of other modes of receiving information. This is a case for a multinomial logistic model.

Multinomial logistic regression

Logistic regression models the odds of an outcome happening as opposed to not happening. Multinomial logistic regression extends this to the case where one of several possible outcomes which can happen, but only one. In the usual case, the reference category would consist of cases in which none of the outcomes happened. In this case, we are interested in the role of social class in determining preferences other than face-to-face, which we are treating as the ‘default' choice. This makes sense because a) the agency usually provides information face-to-face, and is interested in developing other strategies of delivery and b) face-to-face seems is the most usual preference.

```
. mlogit info_pref_4 social_class, rrr

Multinomial logistic regression                 Number of obs   =       2858
                                                LR chi2(3)      =      28.23
                                                Prob > chi2     =     0.0000
Log likelihood = -3478.3948                     Pseudo R2       =     0.0040

------------------------------------------------------------------------------
 info_pref_4 |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Telephone    |
social_class |   1.030592   .0296855     1.05   0.295     .9740216    1.090448
-------------+----------------------------------------------------------------
Website      |
social_class |   .8279833   .0347483    -4.50   0.000     .7626037    .8989681
-------------+----------------------------------------------------------------
Other        |
social_class |   1.035096   .0390042     0.92   0.360     .9614036    1.114436
------------------------------------------------------------------------------
(Outcome info_pref_4==Face-to-face is the comparison group)
```

Multinomial logistic regression reports the effect of a predictor as a relative risk ratio (RRR). This measures the effect of social class on the preference for the particular option relative to the preference for the default option (face-to-face).

Social class is related to choice of website rather than face-to-face. So if the agency develops a website as an alternative means of providing information, it will be addressing the preferences of those in the higher social classes (social class 1, remember, is the highest). The choice of telephone rather than face-to-face, which seemed to have a social class gradient, is not statistically significant. And there is no evident social class gradient in the choice of 'other' methods. The agency is going to have to decide how it will develop its services, but the analysis makes it clear that a website and a telephone service are the two most promising avenues, and that developing a website would more meet the wishes of those in higher social classes. The decision rests with them, of course, but this analysis gives them the best supporting information that the analyst can provide.

There are two reasons why this analysis was much more informative than the original Chi-squared results: the multinomial logistic regression treats social class as an ordered predictor (Chi-squared treats the categories as having no inherent order), and it allows us to model the table as departures from some baseline or reference category - in this case the commonest category.

Using multiple logistic regression has also the advantage that all linear models have: we can build up multivariate models to take into

account the effects of confounding variables. You might rightly suspect that apart from social class, age would also influence choice of non-face-to-face methods. Univariate analysis showed that male sex and age were also related to preference for methods. Here is the model extended to include social class, age and gender:

```
 . mlogit info_pref_4 social_class age male_sex, rrr


Multinomial logistic regression                 Number of obs   =       2858
                                                LR chi2(9)      =     119.40
                                                Prob > chi2     =     0.0000
Log likelihood = -3432.8097                     Pseudo R2       =     0.0171

------------------------------------------------------------------------------
 info_pref_4 |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Telephone    |
social_class |   1.022798   .0296546     0.78   0.437     .9662968    1.082603
         age |    .985376   .0058803    -2.47   0.014      .973918    .9969687
    male_sex |   1.073159    .099386     0.76   0.446     .8950216     1.28675
-------------+----------------------------------------------------------------
Website      |
social_class |   .8014223   .0346082    -5.13   0.000     .7363826    .8722064
         age |    .953246   .0080087    -5.70   0.000     .9376777    .9690727
    male_sex |   1.893742   .2375934     5.09   0.000     1.480905    2.421668
-------------+----------------------------------------------------------------
Other        |
social_class |   1.032792    .039573     0.84   0.400      .958071     1.11334
         age |   .9912533   .0077754    -1.12   0.263     .9761302    1.006611
    male_sex |    2.04518   .2420529     6.05   0.000     1.621771    2.579133
------------------------------------------------------------------------------
(Outcome info_pref_4==Face-to-face is the comparison group)
```

Older people are less likely to prefer websites, and also less likely to prefer the phone - the relative risk ratios for age in both of these sections of the model are significantly less than unity. Men are also very much more likely to prefer the internet, and to prefer the various ‘other' methods of getting information. What is emerging is a picture in which preference for traditional face-to-face methods seems to be associated with women, lower social classes and older people. This information is certainly valuable, arguing that if the agency invests in technology, especially in internet technology, they will be meeting the preferences of younger, higher class male clients. Is this what they want to do? That is their decision, but at least it can be an informed one.

Key points: larger tables

1. Large tables are rarely worth publishing. They usually contain cells with data too sparse to be informative, and the task of interpreting them is typically beyond the skills of the intended readership. It is the data analyst's responsibility to extract patterned information that meets the client's needs.
2. A Chi-squared test on a large table never tests a hypothesis of any interest to anyone.
3. The process of extracting patterned information from a large table requires the analyst to know why the research was done and how the results are to be used.
4. Categorical variables need to be examined critically. Categories with few observations are rarely worth analysing, as there will be little statistical power to detect associations. However, some rarely-occurring outcomes may be intrinsically interesting (death, for example). In this case, consider extracting and analysing the outcome separately.
5. Multinomial logistic models offer an avenue for the analysis of categorical dependent variables. They offer a significant advantage of interpretability over loglinear models. They do, however, require the analyst to define a baseline category. Other categories are modelled as departures from the baseline category. A knowledge of the context of the research will help to identify a baseline category, which can represent the optimal outcome (such as no disease), the most common outcome (as in the example), 'standard practice' (also applies to the example) or the outcome which is best understood.
6. Like all tabular data, relationships in complex tables may be distorted by confounding variables. In the case of survey data, demographic variables which are associated either with the predictor or the predicted variables should be checked for their effect on the relationship.

Acknowledgement
I am grateful to Kay Rundle, whose meticulous questions prompted this paper.